



Published in final edited form as:

Nat Chem Biol. 2021 March ; 17(3): 246–253. doi:10.1038/s41589-020-00711-4.

Robust direct digital-to-biological data storage in living cells

Sung Sun Yim¹, Ross M. McBee^{1,2}, Alan M. Song¹, Yiming Huang^{1,3}, Ravi U. Sheth^{1,3}, Harris H. Wang^{1,4,*}

¹Department of Systems Biology, Columbia University, New York, NY, USA

²Department of Biological Sciences, Columbia University, New York, NY, USA

³Integrated Program in Cellular, Molecular, and Biomedical Studies, Columbia University, New York, NY, USA

⁴Department of Pathology and Cell Biology, Columbia University, New York, NY, USA

Abstract

DNA has been the predominant information storage medium for biology and holds great promise as a next-generation high-density data medium in the digital era. Currently, the vast majority of DNA-based data storage approaches rely on *in vitro* DNA synthesis. As such, there are limited methods to encode digital data into the chromosomes of living cells in a single step. Here, we describe a new electrogenetic framework for direct storage of digital data in living cells. Using an engineered redox-responsive CRISPR adaptation system, we encoded binary data in 3-bit units into CRISPR arrays of bacterial cells by electrical stimulation. We demonstrated multiplex data encoding into barcoded cell populations to yield meaningful information storage and capacity up to 72 bits, which can be maintained over many generations in natural open environments. This work establishes a direct digital-to-biological data storage framework and advances our capacity for information exchange between silicon- and carbon-based entities.

INTRODUCTION

DNA is a ubiquitous molecule in biology that stores life's heritable information. In the digital era, DNA is also poised to become a next-generation universal data medium¹ due to its high-density storage capacity (petabytes per gram)², long-term stability even in harsh environments (half-life of >500 years)³, and low risk of technical obsolescence due to expanding interest in DNA⁴. Data storage in DNA has technologically progressed over the past decade⁴ as strategies to physically isolate and selectively access portions of the stored data^{5,6} as well as algorithmic advances to optimize data encoding and retrieval^{2,7} have

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

*Correspondence to: hw2429@columbia.edu.

Author contributions

S.S.Y., R.U.S., and H.H.W. developed the initial concept. S.S.Y. performed experiments and analyzed the results under the supervision of H.H.W.; S.S.Y., R.M.M., and A.M.S. designed and constructed the electrochemical redox controller setup. S.S.Y. and Y.H. designed the error correction pipeline. S.S.Y. and H.H.W. wrote the manuscript with input from all authors.

Conflict of interest

H.H.W. is a scientific advisor to SNIPR Biome. The authors declare no additional competing interests.

greatly improved the scalability and practicality of DNA information storage. However, current DNA-based data storage methods still rely mainly on *in vitro* iterative chemical or enzymatic synthesis of DNA strands^{4,8}.

At the same time, recent advances in CRISPR and recombinase technologies have led to the development of numerous DNA-based cellular recording systems to interrogate various biological processes^{9,10}, such as lineage tracing for organismal development^{11–13} and real-time recording of horizontal gene transfer events¹⁴. These cellular data recorders offer the capacity to measure biologically relevant signals^{15–19} in places that are otherwise difficult to access, such as inside the body^{20,21}, and over time²². Furthermore, the stored data in DNA can be coupled to gene regulation to directly report cellular states²³ or control cellular logic operations²⁴. These excellent features and inherent compatibility of DNA-based data storage with biological systems have suggested potential use of living cells as a physical medium for data in DNA to provide more protection (e.g. in radiation and heat resistant spores) and enable facile data duplication and amplification (via cell growth and replication)^{4,10}. However, such *in vivo* data storage approaches build largely upon *in vitro* synthesized DNA strands^{25,26}, due to limited capacity in manipulating DNA sequence directly *in vivo*⁴. This challenge motivates the exploration of easy and scalable transmission of digital data into biological systems (i.e. direct encoding) and back (i.e. decoding by sequencing).

Direct information exchange between electronics and biology has tremendous potential to transform our ability to analyze, store, and communicate information^{27–29}. The classic example is the direct electrical stimulation or recording of neurons via ionic potentials and currents²⁷. Beyond ionic potentials, the reduction-oxidation (redox) state of the cell, which is involved in a wide range of biological processes, is also amenable to physiological measurement and perturbation with electronic devices. Recently, redox-responsive biomolecules such as phenazines have been used in several electrochemical strategies to interrogate various biological activities^{30,31} and to control gene expression in living cells^{32,33}, where redox status of the biomolecules could be measured or manipulated by application of electronic potentials. In theory, these approaches could also be used for direct electrochemical encoding of data into DNA in living cells. In practice, however, the utility of such *in vivo* DNA recording systems depends heavily upon the efficiency, robustness, and scalability of the underlying electrogenetic circuits, which may require extensive engineering and optimization.

Here, we describe a scalable and direct strategy, Data Recording *in vivo* by Electrical Stimulation (DRIVES), to encode digital data into the genomes of living cells without the need to synthesize DNA *in vitro* (Fig. 1a). By using electrical signals to tune redox biomolecules and sensors in cells, our framework enables the direct transfer of digital data from a computer to living cells. With a CRISPR-based DNA recorder, we applied this approach to write all possible states of a 3-bit binary data stream into living cells, which can be multiplexed to store larger amounts of information by barcoding cell populations. Data stored in these ‘living hard drives’ are stably maintained and effectively protected over multiple cell generations from external environments, where naked DNA would otherwise be degraded. This study provides a foundation to further advance *in vivo* DNA data storage and direct communication with living cells.

RESULTS

Development of a cellular electrogenetic DNA writer

Previously, we described a directional DNA writing system using CRISPR spacer acquisition to record user-defined signals in bacteria (Fig. 1b)²². We sought to build upon this system for its directional feature in writing information and unique capabilities of temporal recording and multiplexing for scaling. To directly couple an electrical signal for biological recognition, we explored the use of redox molecules and a redox-responsive SoxRS regulon³³ to convert the cell's electrochemical state into a change in gene expression (Extended Data Fig. 1a) thus coupling copy number of the plasmid to oxidative stress. Oxidative stress in the cells can then be induced with phenazine methosulfate (PMS) in a dose-dependent manner (Extended Data Fig. 1b). We further tested ferri/ferrocyanide (oxidized: FCN(O), reduced: FCN(R)) as an alternative electron acceptor and used anaerobic growth conditions to exclude the interference of oxygen in order to improve control of redox conditions (Extended Data Fig. 1c)³³.

To parallelize electrochemical modulation across multiple cell populations, we constructed a 24-chamber electrochemical redox controller that independently delivers a digital electrical pulse (off: 0.0 V or on: +0.5 V) to each chamber (Methods and Extended Data Fig. 2). After optimization of the experimental conditions including inducer concentrations and induction time (Extended Data Fig. 1d–g), we could robustly modulate cell populations using an electrical signal to induce pTrig change. In state 0, the absence of a voltage signal keeps FCN(R) reduced, and therefore pTrig copy number low. In state 1, a +0.5 V signal oxidizes FCN(R) and PMS, which activates the *soxS* promoter to increase pTrig copy number (Fig. 1c). We observed that pTrig copy number increased by more than 400-fold in the presence of the +0.5 V signal (Fig. 1d). Accordingly, newly acquired spacers derived from pTrig were 34-times more prevalent in response to the signal than without, increasing from 0.038(±0.004)% to 1.28(±0.03)% among all arrays in the cell populations (Fig. 1e). Examining the source of newly acquired spacers revealed consistent spacer acquisition across genomic and plasmids regions at each state (Supplementary Fig. 1). These results demonstrate that DRIVES can be wired for direct digital-to-biological encoding in living cells using electronic signals mediated through redox molecules.

Direct encoding of 3-bit digital data into CRISPR arrays

Because spacer acquisition mostly occurs unidirectionally at the 5' position of the expanding CRISPR array, temporal biological events can be recorded over time²². We therefore explored the use of temporal signal induction to encode digital data containing multiple bits in DRIVES as a way to increase data storage capacity of a cell population. To scale from 1 bit to 3 bits, we performed an encoding experiment in which cells were exposed to different electrical signal profiles over three sequential rounds, testing all 8 possible binary induction combinations (Fig. 2a, Extended Data Fig. 3a,b). The pTrig copy number profiles correlated strongly with the 3-bit binary input profiles associated with each cell population (Fig. 2b, Extended Data Fig. 3c). We also observed an increase in CRISPR array expansion over the course of the experiment (Extended Data Fig. 3d) and an increase in the proportion of pTrig-derived spacers as a function of number of electronic signals (Extended Data Fig. 3e).

To better delineate the data structure of the 8 different 3-bit binary data stored in DRIVES, we enriched longer arrays containing more temporal information (Extended Data Fig. 3f) and categorized the observed individual CRISPR arrays in a cell population as a distribution of array-types consisting of either reference (genome- or pRec-derived) or trigger (pTrig-derived) spacers at each positions of an observed array length²² (Methods and Fig. 2c,d). We investigated whether these array-type frequencies could differentiate between different input signal profiles of different cell population by clustering the normalized array-type frequencies (Fig. 2e). Principal component analysis (PCA) on the array-type frequencies revealed 8 distinct clusters that differentiated the 3-bit binary data profiles from each other, although there were some overlaps between the clusters (Extended Data Fig. 3g). Application of our previous classification approach²² using Euclidean distance between observed and predicted (or reference) array-type frequencies failed to return reliable classification results (64.6%) on the test datasets (Supplementary Fig. 2). We suspect that the minimal medium contributed to a weaker pTrig copy number induction and thus a lower array expansion efficiency with pTrig-derived spacers. In turn, the array-type frequencies are more biased towards 'R', 'RR', and 'RRR', which limited the ability of other array-types to contribute to the Euclidean distance metric (Supplementary Fig. 2a). On the other hand, a supervised learning approach might better account for these limitations as well as pleiotropic host responses induced by strong redox stress from electrical stimulations that may introduce variability across datasets³⁴ (Extended Data Fig. 3g).

To leverage the unique patterns of the array-type frequencies, we therefore built classifiers to distinguish the observed CRISPR array data to predict the initial signal profile. From three independent experiments that measured all 3-bit profiles, we first trained a random forest classifier on two randomly selected datasets and tested the model performance on the left-out dataset using L2 and L3 array-types. This initial model yielded an accuracy of 87.5% in profile classification (compared to 12.5% by chance) with 10 iterations of repeated random-subsampling and validation (Extended Data Fig. 4a). Encouraged by these initial classification results, we then retrained the model on all three datasets and then tested its performance on newly acquired datasets from 6 additional independent experiments. This model produced 93.75% accuracy (45 correct classifications out of 48 tested samples) (Fig. 2f) with the 'TRT' array-type frequency as the leading feature for classification (Extended Data Fig. 4b). Approximately, ~10,000 of expanded arrays with uniquely mapping spacers (with ~4,000 L2/L3 arrays) from ~17,000 sequencing reads on size-enriched arrays were sufficient to achieve reasonable classification accuracy (i.e. >90%) (Extended Data Fig. 4c). The number of reads corresponds to ~200,000 cells in the original data-encoded cell population. Using more datasets for training marginally improved the model performance (Extended Data Fig. 4d). Taken together, these results demonstrate that multi-bit digital data can be stored in electrogenetically-actuated CRISPR arrays and the resulting array-type frequencies can be used to recover the stored data from the population.

Scaling data storage capacity with barcoded arrays

To further extend the data storage capacity of DRIVES, we sought to devise a multiplexing strategy to write larger sized binary data across multiple barcoded cell populations in parallel (Fig. 3a). We first generated a library of CRISPR arrays by mutagenizing the distal 8

bp region of the first direct repeat (DR) sequence (Extended Data Fig. 5a) where we previously showed CRISPR arrays could be barcoded²². However, many of the DR variants (72%) exhibited notably lower spacer acquisition rates (i.e. 50% less than that of wild-type DR) (Extended Data Fig. 5b–d), likely due to disrupted interactions between the Cas1-Cas2 complex and the inverted repeats within the first DR sequence at the barcoded region³⁵. We then explored introducing the 8-bp barcode downstream of the first spacer in the CRISPR array. Encouragingly, spacer acquisition efficiencies were consistently high across 24 unique spacer-barcoded variants (Extended Data Fig. 5b,c and Supplementary Table 1). We further assessed CRISPR array expansion for different barcoded cells either individually or as a mixed pool and confirmed that pooling barcoded populations does not significantly affect CRISPR expansion measurements in a multiplex format (Extended Data Fig. 5e). Notably, the pooled arrays could be easily demultiplexed into their associated barcodes through a streamlined Illumina sequencing pipeline that uses each barcode also as a sample index. Barcoding the downstream region of the first spacer also enabled targeted extraction of encoded data belonging to specific barcodes from a mixed population (Supplementary Fig. 3), which was not possible in prior DR barcoding approach. In addition, we performed projections on the scale of DRIVES as a function of Cas1-Cas2 activity, the number of barcodes, and sampling depth (Extended Data Fig. 6). These results demonstrate that this new barcoding strategy can yield active CRISPR array variants with high spacer acquisition efficiencies that can be pooled, thus providing a foundation to scale up DRIVES.

Accurate encoding of text directly into living cells

Having established a robust strategy to expand the data storage capacity of DRIVES, we then set out to test the encoding of meaningful information (e.g. a text message) into living cells. To transform text messages into binary code, we utilized an encoding strategy where each ‘byte’ maps to a 6-bit character code (for 2⁶ or 64 possible characters), built from two concatenated 3-bit data units over two barcoded cell populations (Fig. 3b). Because our classifier performance was not equal across all 3-bit data profiles (Fig. 2f, Extended Data Fig. 7a), we examined different encoding schemes to optimize character-to-byte mapping (Extended Data Fig. 7b). Two schemes were explored: (i) the classic DEC 6-bit encoding table for 64 basic ASCII characters, and (ii) an optimized (OPT) 6-bit encoding table designed to take into account the letter usage frequency³⁶ and classifier decoding performance bias. In the OPT encoding scheme, more frequently used characters (based on letter frequency in English text) are assigned to 6-bit bytes with higher decoding performance (Extended Data Fig. 7c). Therefore, OPT encoding was expected to generally outperform DEC encoding for text messages (Extended Data Fig. 7d).

To test the performance of these encoding schemes, we encoded a 12-byte text message ‘hello world!’ using either the DEC or OPT table directly into *E. coli* cells. For each encoding experiment, the text was split into 12 individual 6-bit characters, with each assigned to two barcoded cell populations holding 3-bit data each (Fig. 3b). All 24 barcoded populations were temporally induced with their assigned 3-bit signals in parallel on the multi-channel electrochemical redox controller set-up (Extended Data Fig. 2). During the course of encoding, pTrig copy number profiles exactly matched the binary input profiles for each barcoded population (Fig. 3c). Upon completion of encoding, the resulting 24 cell

populations were pooled and stored at -80°C as a glycerol stock for subsequent analysis by sequencing. From the sequenced spacers, we determined the array-type frequency profiles from these barcoded populations, which were then classified using our pre-trained random-forest model (Fig. 2). Decoding the data from OPT-encoded cells successfully returned the original message ‘hello world!’ (Fig. 3c). On the other hand, decoding from DEC-encoded cells returned ‘xello world!’, due to misclassification of the first 3-bit ‘101’ as ‘111’ (Supplementary Fig. 4). We further examined how data recovery rate depended on the amount of sequencing reads. For the OPT-encoded data, only 1,600 expanded arrays with uniquely mapping spacers for each barcoded cell population (with $\sim 1,000$ L2/L3 arrays) from $\sim 2,600$ sequencing reads on size-enriched arrays were sufficient to correctly classify $\sim 98\%$ of the 72 bits in the data (Extended Data Fig. 8).

Even with OPT encoding where errors are intentionally suppressed toward the least frequently used characters (bottom 14%), this encoding scheme can still suffer from non-negligible error rates (average 12.13%) (Extended Data Fig. 7c). As shown with the DEC-encoded example, a single-bit error can still drastically deteriorate message outcome (Supplementary Fig. 4). To address this shortcoming, we next implemented an error correction strategy using a simple parity check. Given that the last bit of the binary data (generated most recently in the CRISPR array) is always the most reliable for classification, we utilized the last bit of every 6-bit as a checksum for the previous 5 bits (Fig. 2f, Extended Data Fig. 9a,b). After initial classification of an input, the error correction pipeline counts the number of ‘1’ in the first 5 classified bits and then expects ‘0’ or ‘1’ for a checksum value at the sixth bit based on the counts (Extended Data Fig. 9c, Supplementary Table 2). When the classified checksum value does not match the expected value, the classifier flags that an error has occurred during classification of the character and the error is then corrected based on the classifier’s confusion probability. With this error correction pipeline (OPT2), we can only encode up to 32 curated characters but with significantly higher data reconstruction performance (Extended Data Fig. 9d,e). We encoded the text ‘synbio@cu’ using the OPT2 encoding/decoding pipeline into cells and found that 2 out of 54 bits were initially misclassified, but the errors were detected and successfully corrected to return the input message (Extended Data Fig. 9f). Although error correction is still imperfect, the OPT2 strategy significantly reduces error rates to 0.79% on average. Together, these results demonstrate the ability to encode and store meaningful amounts of information directly into living cells using electrical stimulation alone and that careful design of information encoding and error-correction strategies can significantly improve reconstruction accuracy of stored data.

Stability of data in replicating cells

Mutations during DNA replication, genomic recombination, or changes in cellular fitness could all in theory compromise the fidelity of DNA-based data storage in cells. We expect the *E. coli* BL21 genomic CRISPR arrays where data is stored to have neutral cellular fitness since the CRISPR interference machineries is absent and previously showed that the spacers in CRISPR arrays are stable over 50 generations²². Nevertheless, non-negligible off-target spacer integration rates of Cas1-Cas2 complex³⁷ and continuous growth of the pooled population may lead to subtle changes in subpopulations that become magnified over time to

a level that may affect data recovery. To assess the stability of stored digital information within an actively dividing bacterial population, we propagated a cell population containing the OPT-encoded data 'hello world!' for over 16 days (~100 generations) and sampled the population at multiple time points throughout (Extended Data Fig. 10). Although data retrieval efficiency gradually decreased with increasing population generations, we found that the data could still be robustly decoded with >90% accuracy from the population for ~80 generations. A drop in data retrieval efficiency is likely due to fluctuations in the population since the relative abundance of the 24 barcoded subpopulations was stable for ~60 generations before a notable change was observed, suggesting adaptive mutations with fitness effects arising in some of the subpopulation. We further tracked the population in higher resolution using array-type frequencies within each of the barcoded cell populations where encoded information is embedded (Supplementary Fig. 5). While 15 out of 24 barcoded cell populations (62.5%) in the pool stably maintained the data, the array-type frequencies within the rest of the 9 barcoded populations shifted gradually, losing their initially encoded information over time. Nevertheless, a >90% accuracy achieved over ~80 generations highlights that data with 72 bits encoded in living cells can be exponentially and autonomously amplified over 80 iterations to yield $\sim 1.2 \times 10^{24}$ (2^{80}) times more physical copies that can still be robustly decoded.

Integrity of data in natural open environments

The stability and accessibility of DNA are key advantages in data storage⁴. However, there has been limited direct assessment of the fidelity of DNA-based digital information stored in open natural environments, where DNA encounter various degradative factors including DNase enzymes, microorganisms, UV and chemical mutagens (Fig. 4a). To investigate the integrity of data stored in cells in a natural environment, we took a cell population that encoded a text 'synbio@cu' with 54 bits using OPT2 (Extended Data Fig. 9) and challenged it to commercially purchased organic potting soil at concentrations of 10^7 – 10^9 cells per 100 mg of soil. Encouragingly, we could retrieve up to 90% of the data from the data-encoded cells in soil at the highest spike-in ratio (82% of the mixed soil microbial community). However, the decoding accuracy decreased when lower proportions of data-encoded cells were present in the mixed community likely due to missing data from rare array-types (Fig. 4b). To address this, we selectively grew the data-encoded subpopulation from the mixed soil microbial community using LB media supplemented with kanamycin and chloramphenicol to which data-encoded cells are resistant. Efficient enrichment of data-encoded cells yielded >90% accuracy in data reconstruction for spike-in ratios as low as 7%. We further assessed the stability of the data either in cells or in naked DNA in soil over time. In contrast to naked DNA added directly to soil where most of the data degraded during a 6-day incubation period, data stored in cells were robustly protected and could be decoded without any loss of information (Fig. 4c). In addition, beyond the intrinsic layers of data security used to protect the information embedded within cells (e.g. CRISPR array locus, encoding table, etc.), we further envisioned the utility of camouflaging encoded data in a natural microbial community with vast biodiversity and sequence complexity. Metagenomic 16S rRNA sequencing of the mixed soil microbiome revealed diverse taxa (4,083 OTUs) including the data-encoded *Escherichia/Shigella* genus (Supplementary Fig. 6). Natural soil communities with and without hidden data-encoded *E. coli* cells (4% spike-in ratio) showed

highly similar microbial compositions with Pearson's $r > 0.8$ (Fig. 4d), supporting the idea of data concealment in an open setting. Together, these results highlight the relevance of data storage in living cells for protection from natural environments and future steganographic approaches for embedding synthetic data in complex microbiomes.

DISCUSSION

DNA has great potential to become a next-generation data storage medium. While recent DNA-storage efforts have advanced nucleic acid synthesis, manipulation, and sequencing methods, we focused in this study on developing an all-*in vivo* framework for digital-to-biological data encoding directly into the genomes of living cells in a single step. We demonstrated scaling the data storage capacity of DRIVES in two different dimensions: (i) binary data in units of multiple bits by using temporal electronic signals (i.e. from 1 bit to 3 bits with 8 possible states), and (ii) multiplex encoding across many barcoded cell populations (i.e. from 3 bits to 72 bits with 2^{72} possible states). These strategies can be applied to directly write text messages and the stored data can be reliably recovered and physically amplified through multi-generational growth. Furthermore, data can be hidden within a natural microbial community to enable an additional layer of data security by obscurity. Finally, digital data encoded in the genomes of living cells are protected from harsh natural environments where raw DNA would otherwise be damaged or degraded.

With sufficient sequencing depth and read lengths, the data storage capacity in a cell population is in principle governed by the CRISPR array expansion efficiency, the number of barcodes, and the population size, and will require further advancements for practical utility (Extended Data Fig. 6). We chose to use a 3-bit storage unit per barcoded cell population in this study due mostly to a low abundance ($0.173 \pm 0.065\%$) of L3 arrays after three rounds of temporal signal induction. In theory, these rare cells with longer CRISPR arrays contain the most amount of temporal information, but would require larger population sizes to generate at sufficient levels and deeper sequencing coverages (or amplicon size enrichment) for reliable data reconstruction. Current data storage capacity of DRIVES if scaled suggests that more than 5,000 barcoded cell populations could in theory be pooled and decoded using a single Illumina MiSeq sequencing run (Extended Data Fig. 8). Data storage with thousands of barcoded cell populations will require more sophisticated design of multiplexed electrochemical induction setups that leverage microplate, on-chip or microfluidic formats^{30,38}. Other multiplexable induction modalities such as with light or by acoustics could further increase encoding channel capacity across cell populations^{39,40}. We anticipate that improving the CRISPR spacer acquisition system will enable encoding with higher bit units, a faster rate of encoding, and better reconstruction from a smaller cell population size. Metagenomic mining of Cas1-Cas2 orthologs⁴¹ and directed evolution^{42,43} to improve the CRISPR adaptation machinery or other related host factors are promising paths. Other CRISPR-Cas systems with shorter spacer and DR would enable more compact and denser data storage in CRISPR arrays⁴⁴. Employing more efficient size-enrichment methods and long-read sequencing technologies to decode longer arrays would also improve the overall approach.

While data-encoded cells could be passaged for over 80 cell generations and still allow robust data recovery at >90% accuracy, we observed mutations arising over time that altered the relative abundance of subpopulations, which led to loss of some array-types and deterioration of data fidelity. Engineering host strains with lower mutation rates or other biocontainment strategies could reduce these undesired outcomes^{45,46}. Reducing batch-to-batch variability induced by redox-translated electronic signals could improve the reliability of data recovery³⁴. Lyophilization or use of spore-forming bacteria could also extend shelf-life for long-term DNA-based data storage⁴⁷. This digital-to-biological data storage framework could be applied to other microbial systems with unique properties such as native electroactivity⁴⁸, fast growth⁴⁹, or extremotolerance⁵⁰. We anticipate that the technical advances described here can provide a foundation for higher-performance DNA-based cellular memory devices used not only in digital data storage but also in other biological recording applications.

ONLINE METHODS

Electrochemical set-up

Electrochemical set-up was based on the work by Bentley and colleagues³³ with minor modifications (Extended Data Fig. 2). Briefly, 12-cm long platinum wires (0.5 mm diameter, 99.99% purity) were wound and used for both working and counter electrodes. For agar salt bridges, 12-cm clear PVC tubings (2 mm ID, 4 mm OD) were filled with heated 3% agar with 1 M KCl solution and stored in 3 M KCl at 4 °C. Typical electrochemical set-up for FCN(R/O) conversion and encoding experiments were performed as follows: the working electrode was placed in 2-mL tube (working chamber) with 1.5 mL of M9 minimal medium supplemented with 100 ng/mL anhydrotetracycline (aTc), 1.56 mM ferrocyanide (FCN(R), reduced), 100 µM phenazine methosulfate (PMS), 20 µg/mL chloramphenicol, and 50 µg/mL kanamycin; the counter electrode was placed in another 2-mL tube (counter chamber) with 1.5 mL of M9 minimal medium supplemented with 1.56 mM ferricyanide (FCN(O), oxidized) and 100 µM PMS, unless otherwise stated. A pair of working and counter chamber was connected by a PVC salt bridge.

Electronic control of recordings

E. coli BL21 strain was transformed with pRec and pTrig that are modified from our previous work²² by replacing *lacI* gene with *soxR* gene on pRec and *lac* promoter with *soxS* promoter on pTrig (Extended Data Fig. 1a). The transformed strain was inoculated into a culture tube with 3 mL LB medium supplemented with 20 µg/mL chloramphenicol, and 50 µg/mL kanamycin and grown overnight in a shaking incubator at 37 °C aerobically. The culture was diluted 1:30 into a new culture tube with 3 mL LB medium supplemented with 20 µg/mL chloramphenicol, and 50 µg/mL kanamycin and grown for 2 hours aerobically to bring the culture into exponential growth phase. The culture was then moved to an anaerobic chamber and diluted 1:30 into a working chamber prepared as above. For induction by electronic signals, +0.5 V (on, 1) or 0.0 V (off, 0) was applied to the chamber for 14 hours. The solutions in the working and counter chambers were mixed by pipetting every 1–2 hours to facilitate electrochemical conversion and gene expression induction. Subsequently, 500 µL of the culture was collected to assess pTrig copy number by qPCR. The remaining cell

culture was diluted 1:100 into a new culture tube with 3 mL LB medium supplemented with 20 µg/mL chloramphenicol, and 50 µg/mL kanamycin and grown for overnight aerobically. 500 µL of the cell culture was collected for subsequent analysis of CRISPR arrays for this round of encoding. For multi-round encoding, the remaining cell culture was diluted again into 3 mL LB medium supplemented with 20 µg/mL chloramphenicol, and 50 µg/mL kanamycin and other steps were repeated for the next round.

Barcoding of CRISPR arrays

To facilitate CRISPR array barcoding, endogenous CRISPR array I in *E. coli* BL21 genome was removed by homologous recombination using pSim6 plasmid⁵¹. Both direct repeat (DR) barcoding and spacer barcoding were done by one step cloning and integration protocol based on a bacteriophage integrase⁵². For DR barcoding, the 8 bp of distal end of the first DR of the minimal CRISPR array (80-bp leader sequence + DR + the first spacer of the original CRISPR array I) was diversified using degenerate oligonucleotides and the barcoded arrays were inserted into pOSIP-CH backbone plasmid. For spacer barcoding, we added the 8 bp sequence of illumina i7 indexes to the downstream region of the first spacer of the minimal CRISPR array on pOSIP-CH backbone plasmid. After integration of the plasmids into the genomes of *E. coli* BL21 strain without endogenous CRISPR array I, the backbone part of the plasmids was excised by introducing pE-FLP plasmid for FLP recombinase expression which was then removed using a temperature-sensitive replicon. The sequences of the spacer barcoded CRISPR arrays are listed in Supplementary Table 1.

Array sequencing and data analysis

CRISPR arrays were sequenced using our established sequencing pipeline²² with minor modifications for the barcoded CRISPR arrays. Briefly, cells were lysed using prepGEM bacteria kit (MicroGEM) for amplification of input CRISPR array sequences from the DNA. After PCR amplification of CRISPR arrays, samples were pooled, and for selected libraries, magnetic bead-based size enrichment was performed using AMPureXP beads (Beckman Coulter A63881) as previously described. Sequencing was performed on Illumina MiSeq platform (MiSeq v2 300 cycle) with additional spike-in of custom sequencing primers. The primers sequences are listed in Supplementary Table 3. The raw sequencing data was processed using our established CRISPR spacer extraction and mapping pipeline which can be found at <https://github.com/ravisheth/trace> with minor modifications. Briefly, raw sequencing reads were subjected to spacer extraction (spacer_extraction.py), the extracted spacers were mapped to their sources (blast_search.sh), then uniquely mapping spacers were determined (unique_spacers.py). Further analysis and data visualization were performed mostly in Python with the numpy, scipy, pandas, scikit-learn, matplotlib, and seaborn packages. We considered only arrays with uniquely mapping spacers at all positions within the array, determined if each spacer was either from reference (genome or pRec) or trigger (pTrig), and determined the frequency of each array-type normalized across all possible combinations for the given array length.

qPCR assay for pTrig copy number

pTrig copy number of a cell culture was assessed by qPCR. Briefly, 5 µL of 2x KAPA SYBR Fast qPCR master mix, 0.5 µL of 10 µM forward and reverse primers, 3 µL of

nuclease free water and 1 μ L of cell lysate prepared using prepGEM bacteria kit (MicroGEM) were mixed in each well of a 96 well qPCR plate. Two qPCRs were performed to quantify pTrig and genomic DNA present in each sample. The primers sequences are listed in Supplementary Table 3.

DNA extraction from soil

DNA extraction from soil was performed using our established protocol with Qiagen MagAttract PowerMicrobiome DNA/RNA Kit (Qiagen 27500–4-EP)⁵³. Briefly, 100 mg of soil samples mixed with data-encoded cell population at various ratios were added to the plate. 200 μ L of 0.1-mm Zirconia Silica beads (BioSpec 11079101Z) and 750 μ L of lysis solution (90 mL master mix: 9 mL of 1 M Tris-HCl, pH 7.5, 9 mL of 0.5M EDTA, pH 8.0, 11.25 mL of 10% SDS, 22.5 mL of Qiagen lysis reagent, 38.25 mL of nuclease-free water) were added to each well of the plate. The plate was then subjected to bead beating for 2.5 min followed by 7.5 min of cooling on a bead beater (BioSpec 1001). This bead beating cycle was repeated for 4 times. The plate was centrifuged for 5 min at $4,300 \times g$ and 150 μ L of supernatant was transferred to a V-bottom microplate. 35 μ L of Qiagen inhibitor removal solution was added and the plate was centrifuged for 5 min at $4,300 \times g$. 100 μ L of supernatant was transferred to a round-bottom plate (Corning 3795) on a robotic liquid handler (Biomek 4000) for magnetic bead purification according to the manufacturer's recommendations, but at a scaled volume. The final elute was further diluted 10-fold with nuclease-free water to minimize the effect from any residual PCR inhibitors from soil.

16S rRNA sequencing and data analysis

The V4 region of 16S rRNA gene was sequenced using our established sequencing pipeline⁵³. After PCR amplification of the 16S rRNA V4 regions from the soil DNA, the resulting ~390 bp amplicon was gel-purified and sequenced on Illumina MiSeq platform (MiSeq v2 300 cycle) with additional spike-in of custom sequencing primers. The sequencing data was processed using USEARCH v11.0.667⁵⁴. Reads were merged (-fastq_mergepairs), filtered (-fastq_filter -fastq_maxee 1.0 -fastq_minlen 240), and 100% ZOTUs were generated (-unoise3) and OTU table was created (-otutab). Taxonomy was assigned to ZOTUs using the RDP classifier⁵⁵. Phylogenetic tree was constructed using The Interactive Tree of Life (<https://itol.embl.de>)⁵⁶.

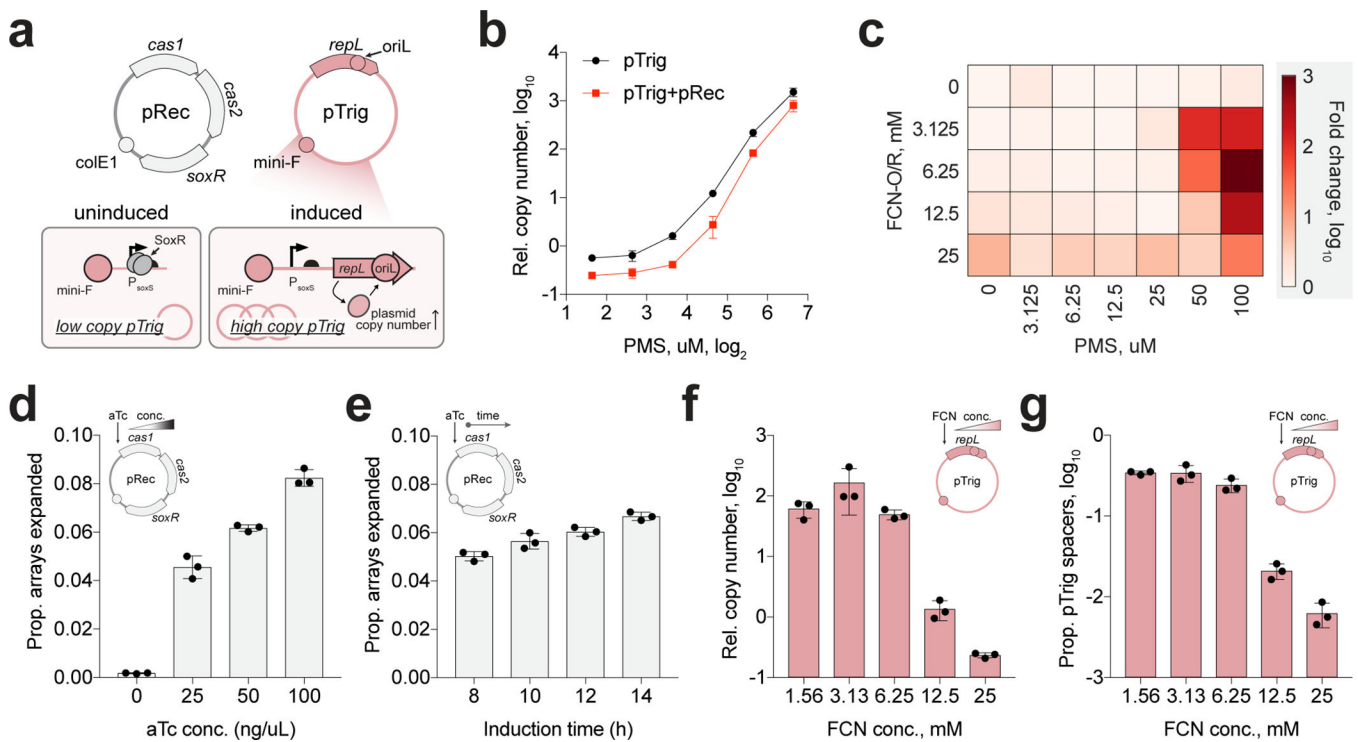
Data availability

All data supporting the findings of this study are available within the article and its supplementary information, or are available from the authors upon request. Sequencing data associated with this study is available at NCBI SRA under PRJNA625964.

Code availability

All of the CRISPR spacer extraction and mapping software can be accessed at <https://github.com/ravisheth/trace>, or are available from the authors upon request.

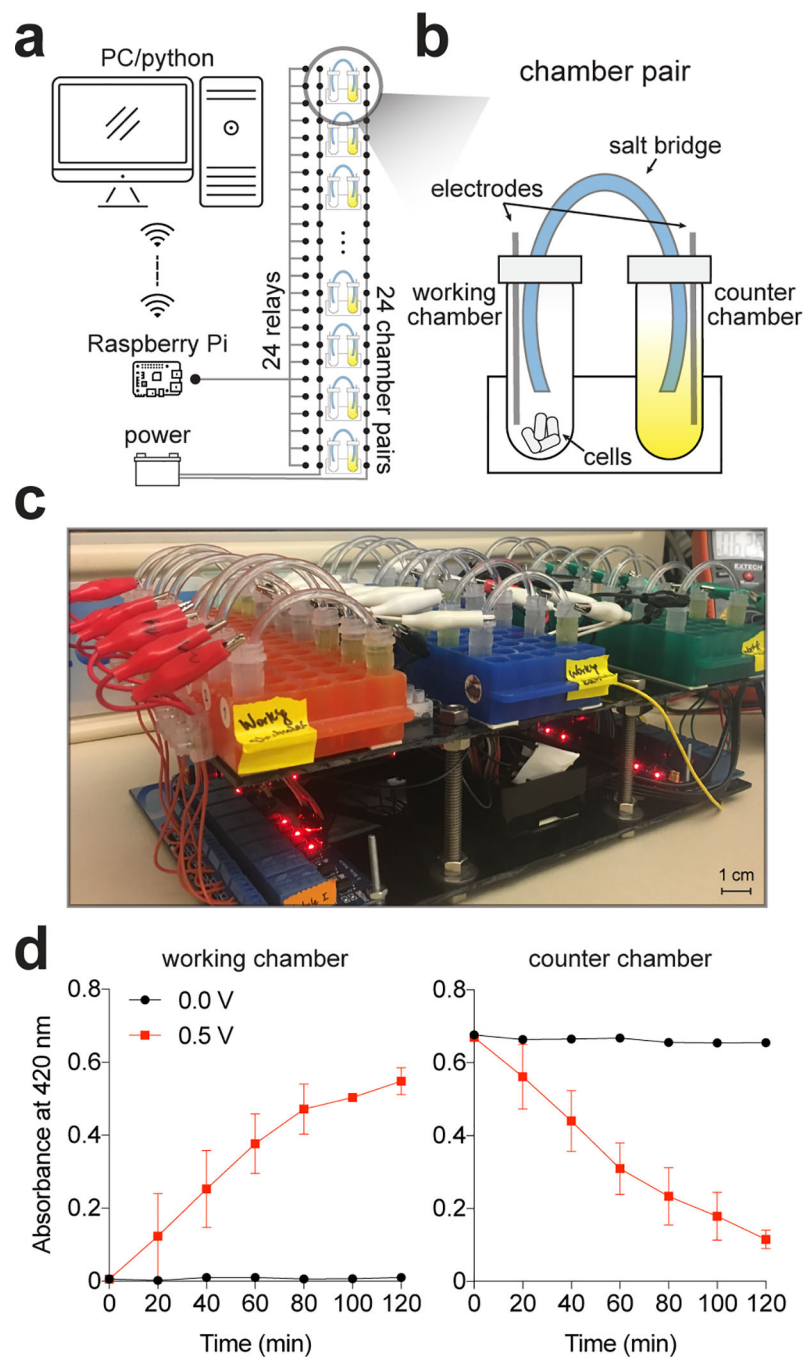
Extended Data



Extended Data Fig. 1. Development of a redox-sensing DNA-based cellular recorder for direct digital-to-biological data storage.

This system is composed of two distinct modules: (i) a ‘sensing module’ that converts a desired biological signal into a change in copy number of a trigger plasmid (pTrig), and (ii) a ‘writing module’ that overexpresses Cas1-Cas2 from a recording plasmid (pRec) to unidirectionally expand genomic CRISPR arrays with novel ~33 bp spacers acquired from genomic or plasmid DNA sources in the cell. In the presence of the desired signal, cells experience a shift in their intracellular DNA pool, driven by an increase in pTrig copy number, which results in an acquisition bias for pTrig-derived spacers amongst expanding CRISPR arrays. **(a)** The *lacI* gene in the previous pRec²² was replaced with *soxR* gene from *E. coli*, and the *lac* promoter in the previous pTrig²² was replaced with *soxS* promoter from *E. coli*. P1 replication system is inactive in the absence of oxidative stress, and a mini-F origin keeps the pTrig plasmid copy number low. Upon induction with oxidative stress, SoxR detaches from *soxS* promoter and activates the P1 replication system to increase the copy number of the plasmid. **(b)** pTrig copy number in the presence of various concentrations of phenazine methosulfate (PMS) in aerobic condition. pRec (with an additional copy of *soxR* gene) helps get higher fold-change of pTrig copy number by more efficient repression in absence of the inducer. **(c)** pTrig copy numbers in the presence of pRec and various concentrations of PMS, and FCN(R) or FCN(O) in anaerobic condition. Fold change of the pTrig copy numbers at the given concentrations of FCN(R) or FCN(O) were plotted. **(d)** Various aTc concentrations and **(e)** induction time for the expression of *cas1* and *cas2* genes were tested for CRISPR array expansion. **(f)** Various FCN(R) and FCN(O) concentrations were tested for pTrig copy number induction and **(g)** pTrig-derived

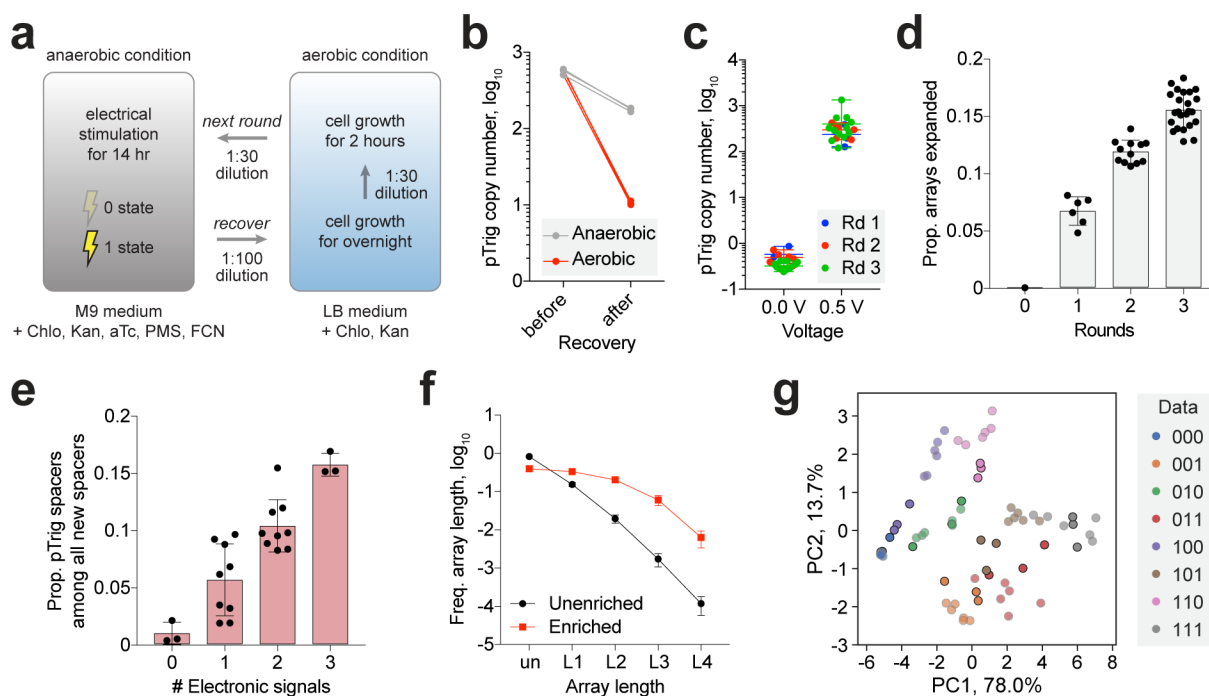
spacer incorporation. The proportions of pTrig-derived spacers among all newly incorporated spacers are displayed. All measurements are based on three biological replicates. Error bars represent standard deviation of three biological replicates.



Extended Data Fig. 2. Construction of a multi-channel electrochemical redox controller.

(a) In an anaerobic chamber, a Raspberry Pi controls 3 of 8-channel relay modules (total 24 relays), which turn on or off electrical signals into each chamber pair from a power supply, based on a python script running on a wirelessly connected PC. (b) A pair of working and

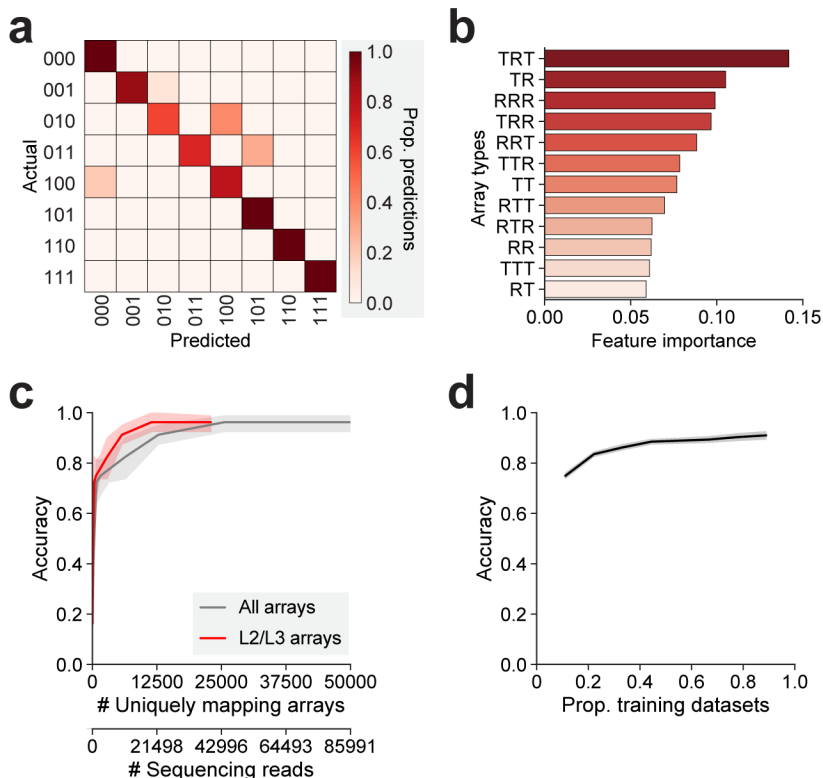
counter chambers is connected by an agar salt bridge. In a working chamber, cells are incubated in M9 minimal medium supplemented with antibiotics, aTc, FCN(R) and PMS. M9 minimal medium supplemented with FCN(O) and PMS is filled in another chamber (counter). **(c)** A photograph of the multi-channel electrochemical redox controller in an anaerobic chamber. **(d)** Changes in electrochemical redox states of FCN(R) in a working chamber (left) and FCN(O) in a counter chamber (right) measured by absorbance at 420 nm with (0.5 V) and without (0.0 V) electronic signals. All measurements are based on three replicates. Error bars represent standard deviation of three biological replicates.



Extended Data Fig. 3. Encoding of 3-bit binary data profiles.

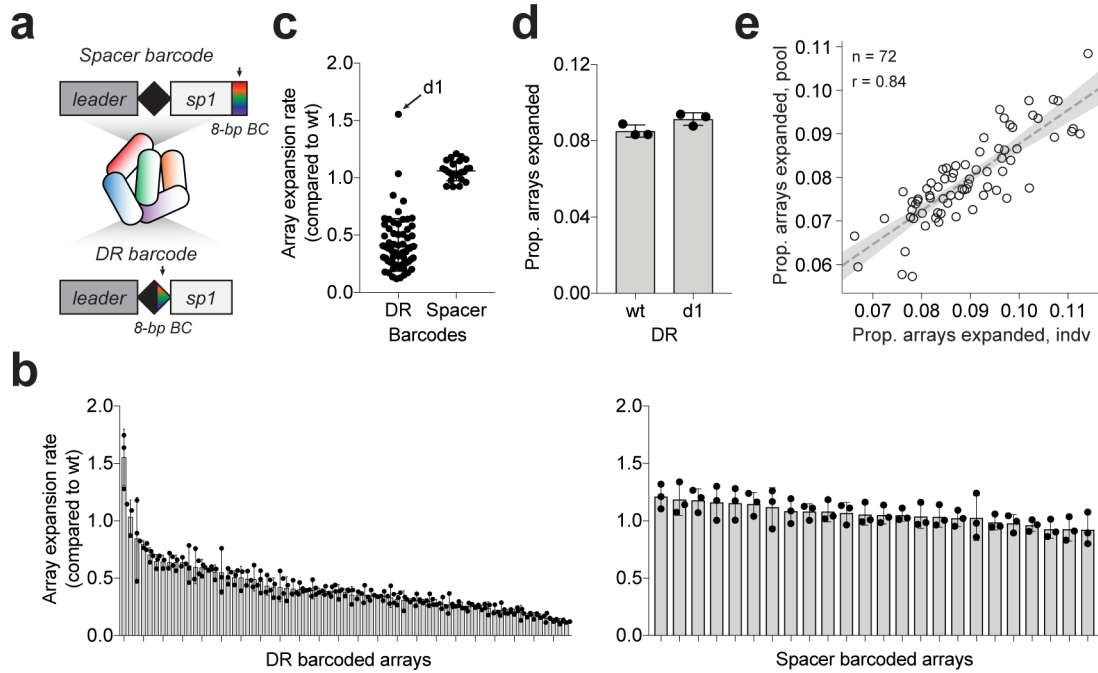
(a) Schematic diagram of experimental steps for multi-round encoding. After each round of electrical stimulation, the cell population was recovered in the rich medium (LB) aerobically so that the induced/uninduced plasmid copy number in the previous encoding round can be diluted out and reset low. **(b)** To determine the recovery condition, anaerobic and aerobic conditions were compared. **(c)** Overlaid distributions of the plasmid copy numbers with/without signals at each round over the course of the multi-round encoding (Figure 2b). **(d)** CRISPR array expansion over the course of the experiment. **(e)** The 3-bit binary data profiles are grouped by the number of electronic signals, and the proportions of pTrig-derived spacers among all newly incorporated spacers are displayed. **(f)** To enrich the sequencing reads for expanded arrays with more new spacers (longer arrays), the magnetic bead-based size enrichment was performed. Frequency of arrays of different lengths (unexpanded and L1-L4) with and without size enrichment are plotted. **(g)** Principal component analysis on the array-type frequency profiles for the 3-bit digital data profiles. All 9 independent biological replicates are shown for each 3-bit digital data profiles. The first three independent datasets used for training of the Random Forest classifier are

highlighted. All measurements are based on two or more biological replicates. Error bars represent standard deviation of three or more biological replicates.



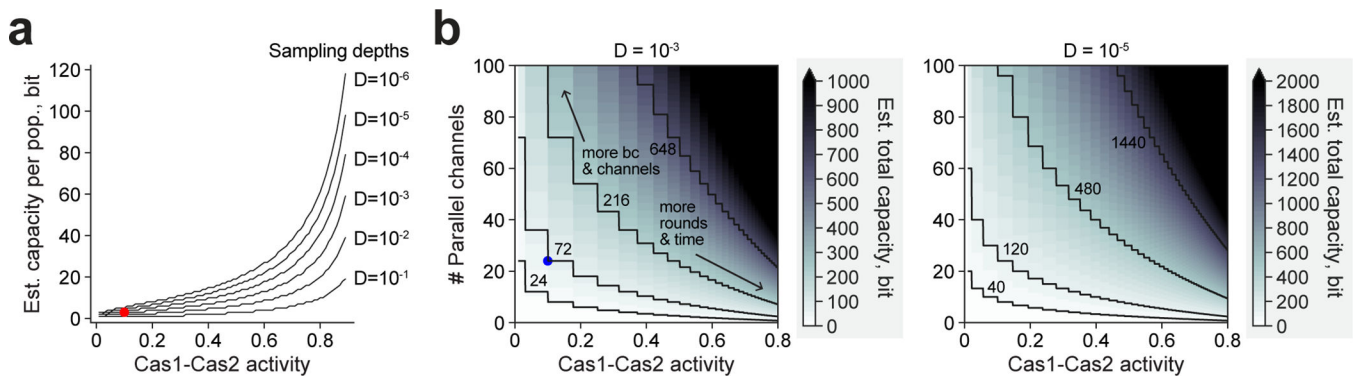
Extended Data Fig. 4. Performance of a Random Forest classifier for data reconstruction.

(a) Confusion matrix from cross validation of the Random Forest classifier for 10 times by training on randomly selected 2 datasets for each 3-bit digital data profile from the 3 independent experiments and testing the trained model on the left-out 1 dataset. **(b)** Importance of features (array-types) for the Random Forest classifier in Figure 2f. **(c)** Classification performance for the number of CRISPR arrays. CRISPR arrays with new uniquely mapping spacers were randomly subsampled to the various numbers for the 3-bit digital data profiles and classifications were performed. Recall accuracies for distinguishing 8 different types of 3-bit digital data profiles were displayed as a function of the number of expanded arrays with uniquely mapping spacers (grey: all arrays, red: L2/L3 arrays). The number of sequencing reads corresponding to the number of expanded arrays with uniquely mapping spacers (grey: all arrays) is also provided as an additional x-axis. Shaded regions represent 95% confidence interval of 10 iterations of subsampling and classification. **(d)** Recall accuracies for distinguishing 8 different types of 3-bit digital data profiles with varying proportions of randomly selected training datasets for each 3-bit digital data profile. Shaded regions represent 95% confidence interval of 100 iterations of subsampling and classification.



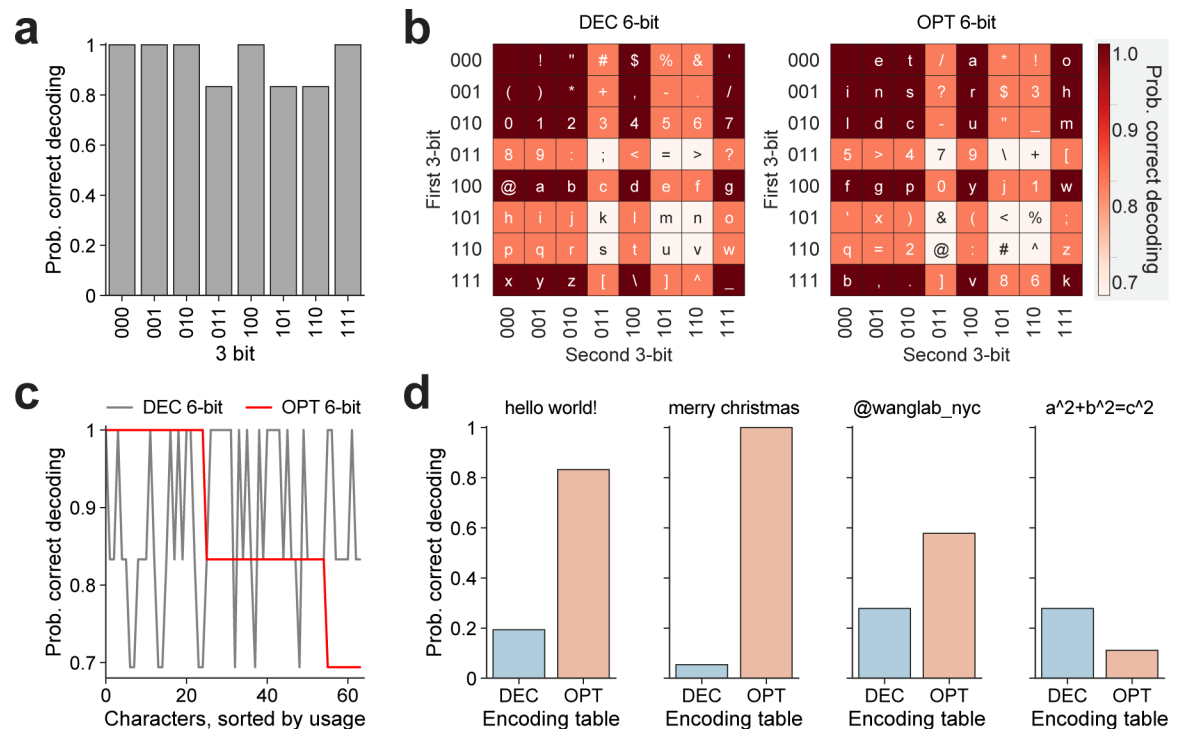
Extended Data Fig. 5. Barcoding CRISPR arrays for multiplexed encoding.

(a) CRISPR arrays can be barcoded with 8-bp unique sequences either downstream of the 1st spacer region or within direct repeat (DR) region. (b) CRISPR array expansion rates (relative to wild-type array) of 69 DR-barcoded CRISPR arrays and 24 spacer-barcoded CRISPR arrays. (c) Distribution of array expansion rates of spacer-barcoded CRISPR arrays is much more uniform and consistent than that of DR-barcoded CRISPR arrays. A DR variant (d1) that was more efficient than the wild-type DR sequence in the initial 96-well plate-based test is highlighted. (d) The d1 DR variant was tested again in tube culture condition. In tube culture condition, however, the DR variant did not show significantly higher activity than that of the wild-type DR sequence. (e) Comparison of CRISPR array expansion rates measured individually or in pool. Shaded region represents 95% confidence interval for linear regression (dashed grey line). Sample sizes (n) and Person correlation coefficient (r) are shown. All measurements are based on three biological replicates. Error bars represent standard deviation of three biological replicates.



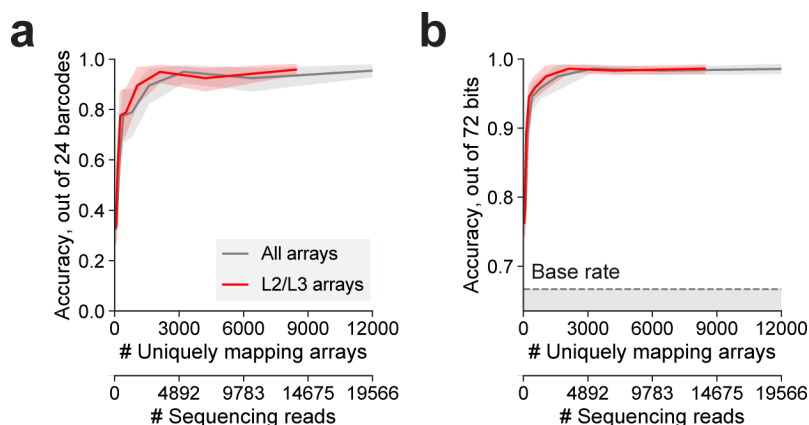
Extended Data Fig. 6. Projections on the scale of DRIVES.

(a) Data storage capacity ('n' bits of information or 'n' rounds of encoding) per cell population is estimated as a function of Cas1-Cas2 activity ('X' proportion of the cell population expanded arrays with a new spacer after a single round of encoding). Here, 'Xⁿ' proportion of the cell population would have expanded arrays every round resulting 'n' new spacers (Ln arrays) after 'n' rounds of encoding, and we assumed that the sampling capacity for the Ln array population governs the data storage capacity. We considered various sampling depths 'D', where 'D' proportion of the cell population can be sufficiently sampled. This 'D' could be affected by many factors including the sequencing depth and size enrichment efficiency. We assumed that if the 'Xⁿ' is same or higher than the given sampling depth constraint 'D', 'n' bits can be stored and reliably decoded. For example, when 0.001 of the cell population can be sufficiently sampled (D=0.001), maximum data storage capacity would be 3 bits (n=3) with the current Cas1-Cas2 activity level (X=0.1) as in our current experimental dataset (highlighted in red in the plot). And when 0.0001 of the cell population can be sufficiently sampled (D=0.0001), maximum data storage capacity would be 4 bits (n=4) with the current Cas1-Cas2 activity level (X=0.1). Although the Illumina MiSeq v2 300 cycles kit used in this study can read only up to 5 new spacers, we assumed that sequencing read length is not the limiting factor in this projection as other long read sequencing technologies could be employed. **(b)** Estimated total data storage capacity across barcoded cell populations as a function of Cas1-Cas2 activity and the number of parallel channels in the culture platform at two different sampling depths (D=0.001 and D=0.00001). A larger data per cell population would require more rounds of encoding which takes longer time, and a larger number of parallel channels would require more barcoded cell populations and more sophisticated design of the culture platform. Current capacity of the system with 24 channels in the culture platform is highlighted in blue in the plot.

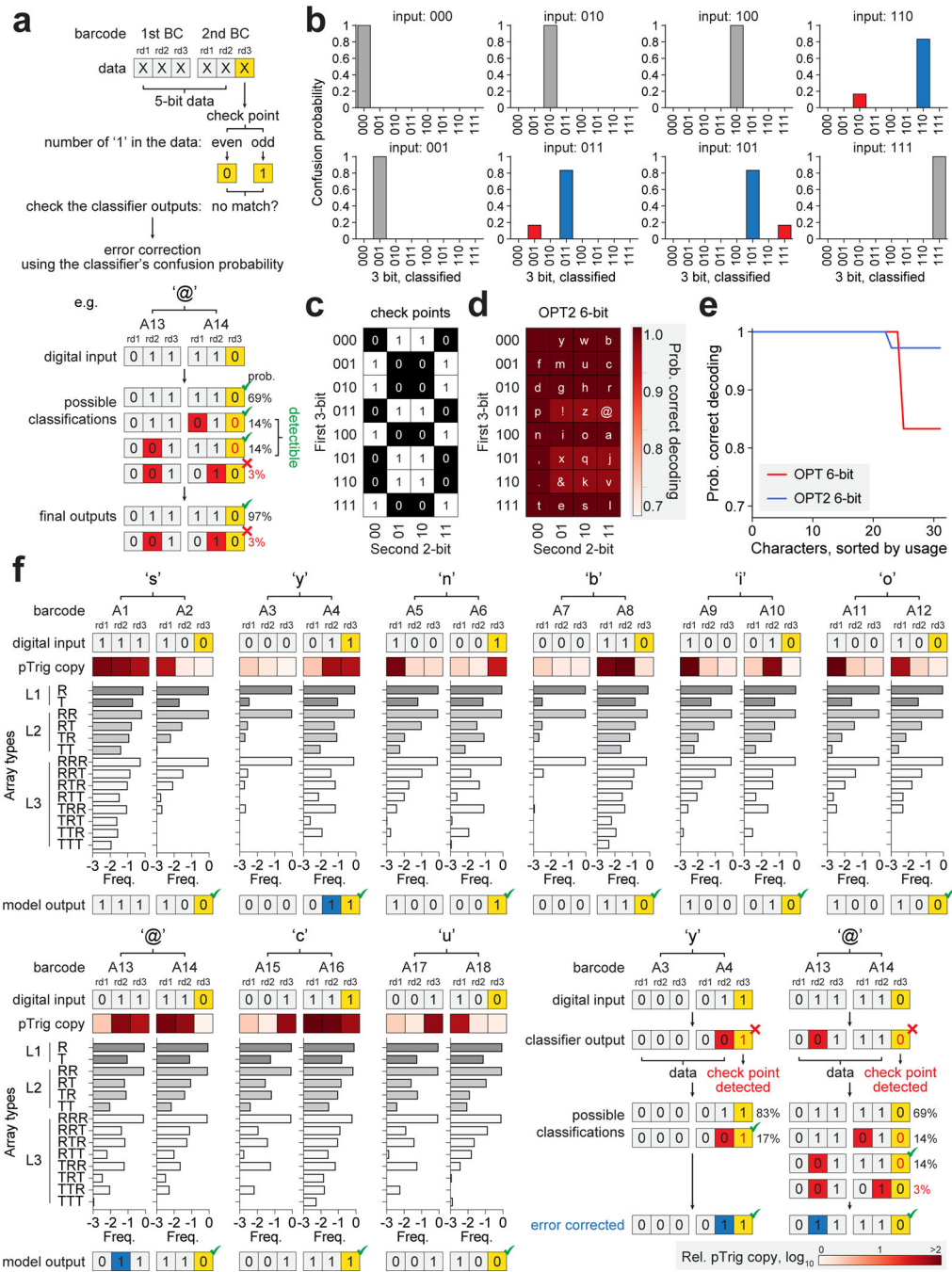


Extended Data Fig. 7. Design of 6-bit encoding tables for text messages.

(a) Probability of correct classification for each of the 3-bit digital data profiles by the Random Forest classifier on the newly generated independent datasets is calculated based on the result in Figure 2f. (b) DEC and OPT encoding tables with estimated probabilities of correct classification for the 64 characters. OPT 6-bit encoding table was designed by considering the correct classification probability and the usage frequency of the characters (https://mdickens.me/typing/letter_frequency.html). (c) Probability of correct decoding for the 64 character (ordered by usage) with DEC and OPT 6-bit encoding tables. (d) Comparison of predicted probabilities of correct decoding for various text messages based on the two encoding tables. The predicted probabilities of correct decoding for each character or text message were calculated by multiplying the correct decoding probability values of each 3-bit digital data profile units.

**Extended Data Fig. 8. Reading 'hello world!' from subsampled sequencing reads.**

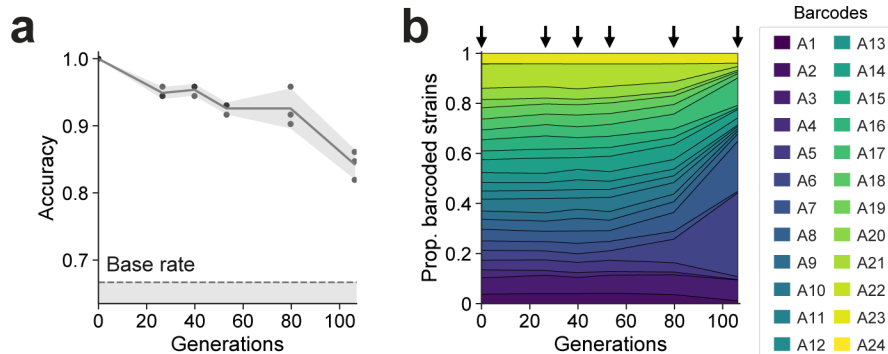
Sequencing reads from each barcode in the 'hello world!'-encoded cell population using OPT table were randomly subsampled to the various numbers and classifications were performed. Recall accuracies for (a) distinguishing 3-bit digital data profiles for 24 barcoded populations or for (b) calling correct bits out of 72 bits were displayed as a function of the number of expanded arrays with uniquely mapping spacers (grey: all arrays, red: L2/L3 arrays). The number of sequencing reads corresponding to the number of expanded arrays with uniquely mapping spacers (grey: all arrays) is also provided as an additional x-axis. Shaded regions represent 95% confidence interval of 10 iterations of subsampling and classification.



Extended Data Fig. 9. Improving data reconstruction with error correction.

(a) By using every sixth bit as a check point (checksum) for the first 5 bits, errors in data reconstruction can be detected and corrected for the selected 32 combinations of 6-bit digital data profiles based on the classifier's confusion probability in Figure 2f and Extended Data Fig. 9b. For example, for a digital input '011110' could be classified as '011110', '011010', '001110', or '001010' with the probabilities of 69%, 14%, 14%, or 3%, respectively. Out of these 4 possible initial classifications, the last 3 are wrong and the 2 wrong classifications with a single bit error can be detected by the check point values and fixed. However, the

classification result with 2 bits error cannot be detected by the check point value and therefore cannot be fixed. For all 32 combinations of 6-bit digital data profiles, possible classification results, their probabilities, and whether they can be fixed or not are summarized in Supplementary Table 2. **(b)** Confusion probability for each of the 3-bit digital data profiles based on Figure 2f. **(c)** The check point values for each combination of eight 3-bit and four 2-bit digital data profiles. **(d)** OPT2 encoding table with the estimated probabilities of correct classification for the 32 characters. **(e)** Probability of correct decoding for the 32 characters (ordered by usage) for OPT and OPT2 6-bit encoding tables. **(f)** ‘synbio@cu’ encoded in the genomes of barcoded *E. coli* populations using the OPT2 error correction strategy. Two errors from the initial classification were detected using the check points and successfully corrected as described in the figure. For classification of each barcoded cell population, an average of 492,289 total sequencing reads with 268,066 reads of expanded arrays (or 106,242 of L2/L3 arrays) that uniquely map spacers were used. Bead-based size enrichment was performed to enrich for expanded arrays and deplete unexpanded arrays. Frequencies of array-types are in \log_{10} scale. All measurements are based on a single experimental study.



Extended Data Fig. 10. Data stability in replicating cells.

A mixed pool of 24 barcoded cell population encoded with a 72-bits text message ‘hello world!’ in Figure 3 was subsequently diluted 1:100 every 24 hours into 3 mL fresh LB media with antibiotic for a total of 16 days (~106 generation, ~6.6 generations per day). **(a)** Data stability in the propagating cell population over 100 generations. Accuracy indicates the proportion of bits that are correctly classified. >90% of the 72 bits could be correctly retrieved up to ~80 generations. Shaded region represents standard deviation of three biological replicates. For classification of each barcoded cell population, an average of 82,860 of total sequencing reads with 40,502 reads of expanded arrays (or 17,139 of L2/L3 arrays) that uniquely map spacers were used. Bead-based size enrichment was performed to enrich for expanded arrays and deplete unexpanded arrays. **(b)** Gradual changes in the relative abundance of 24 barcoded cell population over time suggests adaptive mutations with fitness effects arising in some of the subpopulation. Samples were collected at the time points indicated by arrows (day 0, 4, 6, 8, 12, and 16). All measurements are based on three biological replicates.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank A.Kaufman and members of the Wang lab for advice and comments on the manuscript. H.H.W. acknowledges funding support from ONR (N00014-17-1-2353), NSF (MCB-1453219), NIH (1R01AI132403-01), and the Burroughs Wellcome Fund (PATH1016691). S.S.Y. thanks support from Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (NRF-2017R1A6A3A03003401). R.U.S. was supported by a Fannie and John Hertz Foundation Fellowship and an NSF Graduate Research Fellowship (DGE-1644869).

REFERENCES

1. Church GM, Gao Y & Kosuri S Next-generation digital information storage in DNA. *Science* 337, 1628, doi:10.1126/science.1226355originally (2012). [PubMed: 22903519]
2. Erlich Y & Zielinski D DNA Fountain enables a robust and efficient storage architecture. *Science* 355, 950–954 (2017). [PubMed: 28254941]
3. Allentoft ME et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc Biol Sci* 279, 4724–4733, doi:10.1098/rspb.2012.1745 (2012). [PubMed: 23055061]
4. Ceze L, Nivala J & Strauss K Molecular digital data storage using DNA. *Nat Rev Genet* 20, 456–466, doi:10.1038/s41576-019-0125-3 (2019). [PubMed: 31068682]
5. Newman S et al. High density DNA data storage library via dehydration with digital microfluidic retrieval. *Nat Commun* 10, 1706, doi:10.1038/s41467-019-09517-y (2019). [PubMed: 30979873]
6. Organick L et al. Random access in large-scale DNA data storage. *Nat Biotechnol* 36, 242–248, doi:10.1038/nbt.4079 (2018). [PubMed: 29457795]
7. Anavy L, Vaknin I, Atar O, Amit R & Yakhini Z Data storage in DNA with fewer synthesis cycles using composite DNA letters. *Nat Biotechnol* 37, 1229–1236, doi:10.1038/s41587-019-0240-x (2019). [PubMed: 31501560]
8. Lee HH, Kalhor R, Goela N, Bolot J & Church GM Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nat Commun* 10, 2383, doi:10.1038/s41467-019-10258-1 (2019). [PubMed: 31160595]
9. Farzadfard F & Lu TK Emerging applications for DNA writers and molecular recorders. *Science* 361, 870–875 (2018). [PubMed: 30166483]
10. Sheth RU & Wang HH DNA-based memory devices for recording cellular events. *Nat Rev Genet* 19, 718–732, doi:10.1038/s41576-018-0052-8 (2018). [PubMed: 30237447]
11. Chan MM et al. Molecular recording of mammalian embryogenesis. *Nature* 570, 77–82, doi:10.1038/s41586-019-1184-5 (2019). [PubMed: 31086336]
12. Kalhor R et al. Developmental barcoding of whole mouse via homing CRISPR. *Science* 361, doi:10.1126/science.aat9804 (2018).
13. McKenna A et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* 353, aaf7907, doi:10.1126/science.aaf7907 (2016). [PubMed: 27229144]
14. Munck C, Sheth RU, Freedberg DE & Wang HH Recording mobile DNA in the gut microbiota using an *Escherichia coli* CRISPR-Cas spacer acquisition platform. *Nat Commun* 11, 95, doi:10.1038/s41467-019-14012-5 (2020). [PubMed: 31911609]
15. Farzadfard F & Lu TK Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science* 346, 1256272, doi:10.1126/science.1256272 (2014). [PubMed: 25395541]
16. Loveless TB et al. DNA writing at a single genomic site enables lineage tracing and analog recording in mammalian cells. *bioRxiv*, doi:10.1101/639120 (2019).
17. Schmidt F, Cherepkova MY & Platt RJ Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* 562, 380–385, doi:10.1038/s41586-018-0569-1 (2018). [PubMed: 30283135]

18. Tang W & Liu DR Rewritable multi-event analog recording in bacterial and mammalian cells. *Science* 360, doi:10.1126/science.aap8992 (2018).
19. Yang L et al. Permanent genetic memory with >1-byte capacity. *Nat Methods* 11, 1261–1266, doi:10.1038/nmeth.3147 (2014). [PubMed: 25344638]
20. Mimee M, Tucker AC, Voigt CA & Lu TK Programming a human commensal bacterium, *Bacteroides thetaiotaomicron*, to sense and respond to stimuli in the murine gut Microbiota. *Cell Syst* 1, 62–71, doi:10.1016/j.cels.2015.06.001 (2015). [PubMed: 26918244]
21. Riglar DT et al. Engineered bacteria can function in the mammalian gut long-term as live diagnostics of inflammation. *Nat Biotechnol* 35, 653–658, doi:10.1038/nbt.3879 (2017). [PubMed: 28553941]
22. Sheth RU, Yim SS, Wu FL & Wang HH Multiplex recording of cellular events over time on CRISPR biological tape. *Science* 358, 1457–1461 (2017). [PubMed: 29170279]
23. Roquet N, Soleimany AP, Ferris AC, Aaronson S & Lu TK Synthetic recombinase-based state machines in living cells. *Science* 353, aad8559, doi:10.1126/science.aad8559 (2016). [PubMed: 27463678]
24. Farzadfard F et al. Single-nucleotide-resolution computing and memory in living cells. *Mol Cell* 75, 769–780 e764, doi:10.1016/j.molcel.2019.07.011 (2019). [PubMed: 31442423]
25. Akhmetov A, Ellington AD & Marcotte EM A highly parallel strategy for storage of digital information in living cells. *BMC Biotechnol* 18, 64, doi:10.1186/s12896-018-0476-4 (2018). [PubMed: 30333005]
26. Shipman SL, Nivala J, Macklis JD & Church GM CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* 547, 345–349, doi:10.1038/nature23017 (2017). [PubMed: 28700573]
27. Liu Y et al. Connecting Biology to Electronics: Molecular Communication via Redox Modality. *Adv Healthc Mater* 6, doi:10.1002/adhm.201700789 (2017).
28. Mimee M et al. An ingestible bacterial-electronic system to monitor gastrointestinal health. *Science* 360, 915–918 (2018). [PubMed: 29798884]
29. Weber W et al. A synthetic mammalian electro-genetic transcription circuit. *Nucleic Acids Res* 37, e33, doi:10.1093/nar/gkp014 (2009). [PubMed: 19190091]
30. Bellin DL et al. Electrochemical camera chip for simultaneous imaging of multiple metabolites in biofilms. *Nat Commun* 7, 10535, doi:10.1038/ncomms10535 (2016). [PubMed: 26813638]
31. VanArsdale E et al. A co-culture based tyrosine-tyrosinase electrochemical gene circuit for connecting cellular communication with electronic networks. *ACS Synth Biol*, doi:10.1021/acssynbio.9b00469 (2020).
32. Gordonov T et al. Electronic modulation of biochemical signal generation. *Nat Nanotechnol* 9, 605–610, doi:10.1038/nnano.2014.151 (2014). [PubMed: 25064394]
33. Tschirhart T et al. Electronic control of gene expression and cell behaviour in *Escherichia coli* through redox signalling. *Nat Commun* 8, 14030, doi:10.1038/ncomms14030 (2017). [PubMed: 28094788]
34. Bhokisham N et al. A redox-based electrogenetic CRISPR system to connect with and control biological information networks. *Nat Commun* 11, 2427, doi:10.1038/s41467-020-16249-x (2020). [PubMed: 32415193]
35. Nunez JK, Lee AS, Engelman A & Doudna JA Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* 519, 193–198, doi:10.1038/nature14237 (2015). [PubMed: 25707795]
36. Michel JB et al. Quantitative analysis of culture using millions of digitized books. *Science* 331, 176–182 (2011). [PubMed: 21163965]
37. Nivala J, Shipman SL & Church GM Spontaneous CRISPR loci generation in vivo by non-canonical spacer integration. *Nat Microbiol* 3, 310–318, doi:10.1038/s41564-017-0097-z (2018). [PubMed: 29379209]
38. Din MO, Martin A, Razinkov I, Csicsery N & Hasty J Interfacing gene circuits with microelectronics through engineered population dynamics. *Science Advances* 6, eaaz8344, doi:10.1126/sciadv.aaz8344 (2020). [PubMed: 32494744]

39. Fernandez-Rodriguez J, Moser F, Song M & Voigt CA Engineering RGB color vision into *Escherichia coli*. *Nat Chem Biol* 13, 706–708, doi:10.1038/nchembio.2390 (2017). [PubMed: 28530708]
40. Piraner DI, Abedi MH, Moser BA, Lee-Gosselin A & Shapiro MG Tunable thermal bioswitches for in vivo control of microbial therapeutics. *Nat Chem Biol* 13, 75–80, doi:10.1038/nchembio.2233 (2017). [PubMed: 27842069]
41. Makarova KS et al. Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol* 18, 67–83, doi:10.1038/s41579-019-0299-x (2020). [PubMed: 31857715]
42. Heler R et al. Mutations in Cas9 Enhance the Rate of Acquisition of Viral Spacer Sequences during the CRISPR-Cas Immune Response. *Mol Cell* 65, 168–175, doi:10.1016/j.molcel.2016.11.031 (2017). [PubMed: 28017588]
43. Shipman SL, Nivala J, Macklis JD & Church GM Molecular recordings by directed CRISPR spacer acquisition. *Science* 353, aaf1175, doi:10.1126/science.aaf1175 (2016). [PubMed: 27284167]
44. Wright AV et al. A Functional Mini-Integrase in a Two-Protein Type V-C CRISPR System. *Molecular Cell* 73, 727–737.e723, doi:10.1016/j.molcel.2018.12.015 (2019). [PubMed: 30709710]
45. Blazejewski T, Ho H-I & Wang HH Synthetic sequence entanglement augments stability and containment of genetic information in cells. *Science* 365, 595–598 (2019). [PubMed: 31395784]
46. Deatherage DE, Leon D, Rodriguez AE, Omar SK & Barrick JE Directed evolution of *Escherichia coli* with lower-than-natural plasmid mutation rates. *Nucleic Acids Res* 46, 9236–9250, doi:10.1093/nar/gky751 (2018). [PubMed: 30137492]
47. Cox JP Long-term data storage in DNA. *Trends Biotechnol* 19, 247–250 (2001). [PubMed: 11412947]
48. Li F et al. Modular engineering to increase intracellular NAD (H⁺) promotes rate of extracellular electron transfer of *Shewanella oneidensis*. *Nat Commun* 9, 3637 (2018). [PubMed: 30194293]
49. Lee HH et al. Functional genomics of the rapidly replicating bacterium *Vibrio natriegens* by CRISPRi. *Nat Microbiol* 4, 1105–1113, doi:10.1038/s41564-019-0423-8 (2019). [PubMed: 30962569]
50. Davis J et al. In vivo multi-dimensional information-keeping in *Halobacterium salinarum*. *bioRxiv*, doi:10.1101/2020.02.14.949925 (2020).
51. Sharan SK, Thomason LC, Kuznetsov SG & Court DL Recombineering: a homologous recombination-based method of genetic engineering. *Nat Protoc* 4, 206–223, doi:10.1038/nprot.2008.227 (2009). [PubMed: 19180090]
52. St-Pierre F et al. One-step cloning and chromosomal integration of DNA. *ACS Synth Biol* 2, 537–541, doi:10.1021/sb400021j (2013). [PubMed: 24050148]
53. Ji BW et al. Quantifying spatiotemporal variability and noise in absolute microbiota abundances using replicate sampling. *Nat Methods* 16, 731–736, doi:10.1038/s41592-019-0467-y (2019). [PubMed: 31308552]
54. Edgar RC Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461, doi:10.1093/bioinformatics/btq461 (2010). [PubMed: 20709691]
55. Wang Q, Garrity GM, Tiedje JM & Cole JR Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73, 5261–5267, doi:10.1128/AEM.00062-07 (2007). [PubMed: 17586664]
56. Letunic I & Bork P Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 47, W256–W259, doi:10.1093/nar/gkz239 (2019). [PubMed: 30931475]

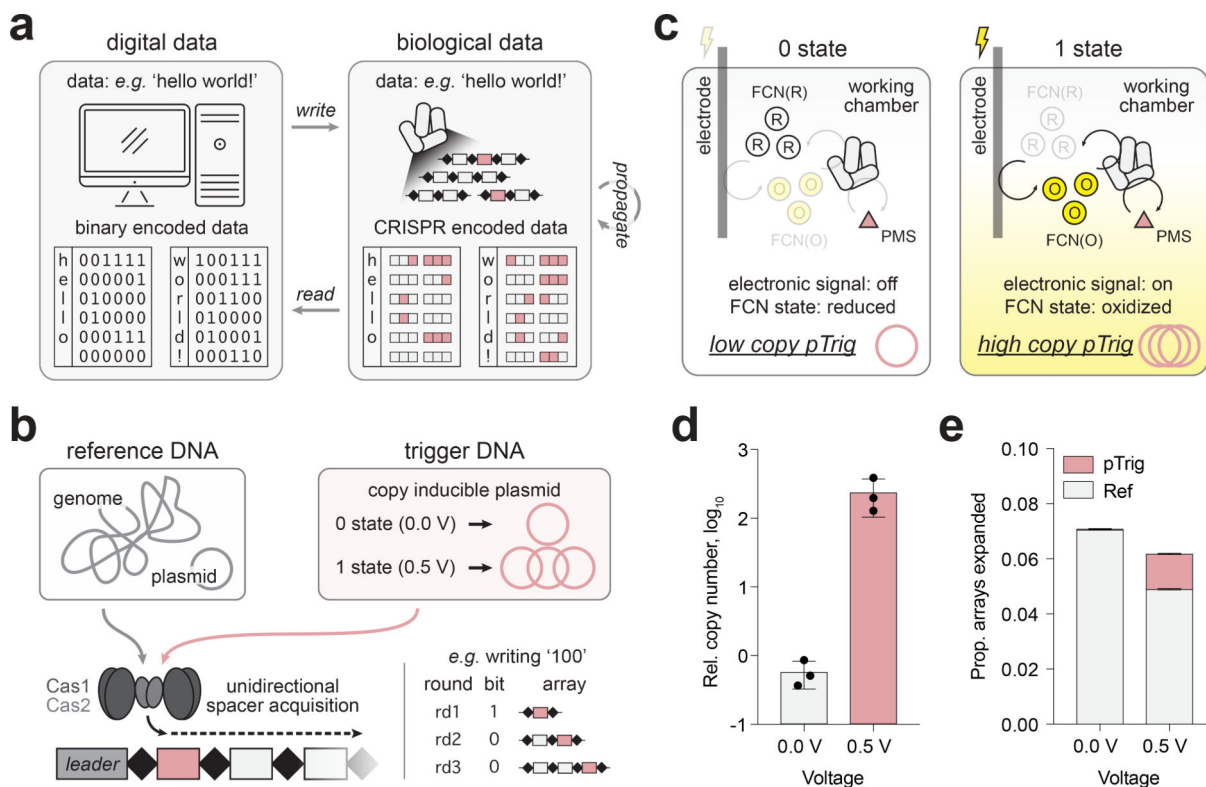


Figure 1. Direct digital-to-biological data storage into CRISPR arrays.

(a) Digital information can be directly encoded into CRISPR arrays of a bacterial population using electronic signals. The cell population can then be archived for long-term storage, propagated for data amplification, and sequenced for data retrieval. (b) Overexpression of the Cas1-Cas2 complex results in constant incorporation new spacers into CRISPR arrays of a cell population. Electronic signals induce a change in abundance of a copy number inducible plasmid (pTrig) and thus the proportion of pTrig-derived spacers. (c) At 0 state, the electrical signal is not applied (0.0 V) to keep FCN(R) and PMS reduced and pTrig copy number is low. At 1 state, the electrical signal (0.5 V) oxidizes FCN(R) and PMS, which activates the *soxS* promoter to increase pTrig copy number. FCN(R), ferrocyanide; FCN(O), ferricyanide; PMS, phenazine methosulfate. (d) Relative copy number of pTrig, (e) proportion of expanded CRISPR arrays and source of the new spacers without (0.0 V) and with (0.5 V) electrical signal for 14 hours. Ref, genome- and pRec-derived spacers; pTrig, pTrig-derived spacers. All measurements are based on three biological replicates. Error bars represent standard deviation of three biological replicates.

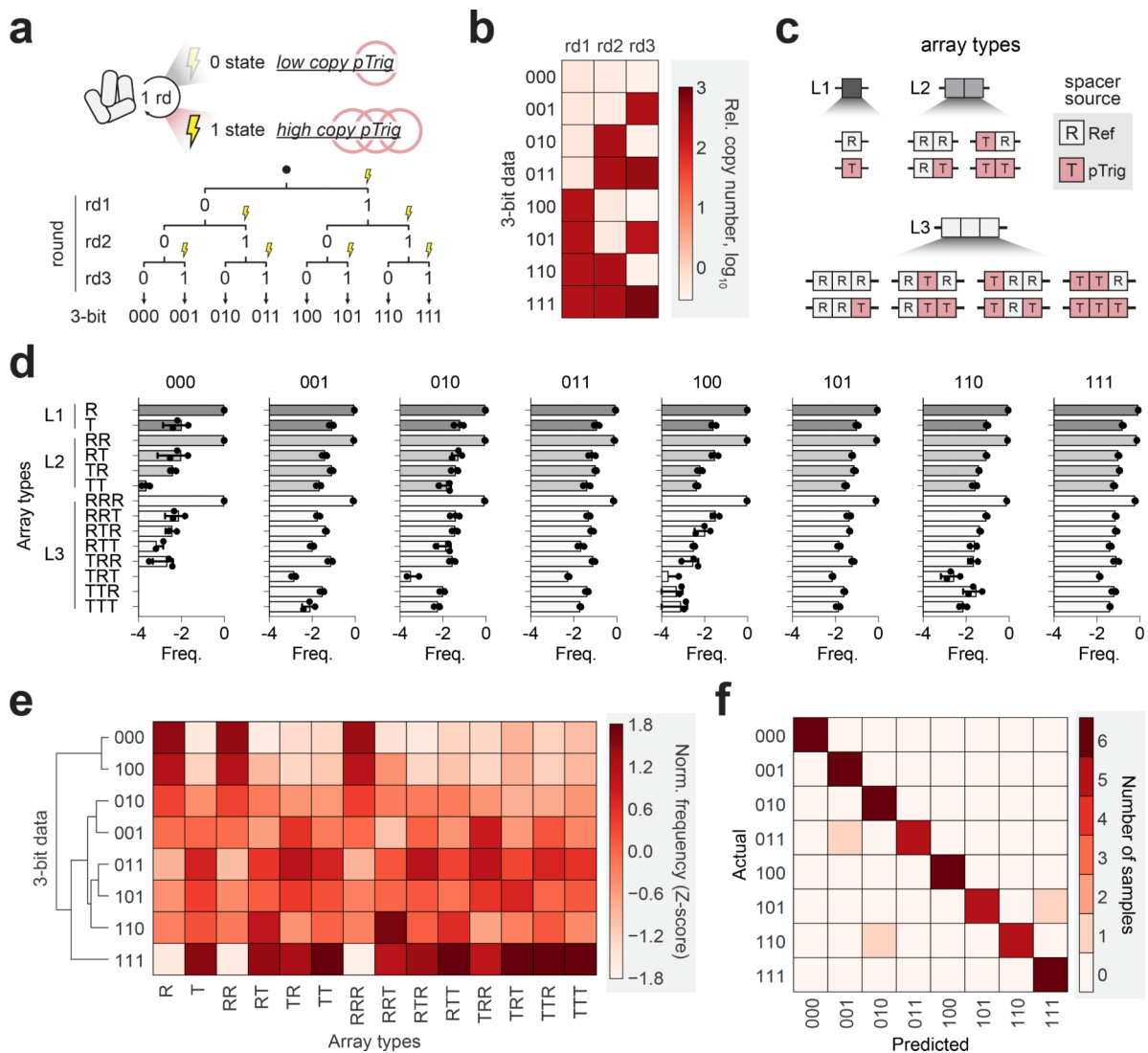


Figure 2. Encoding 3-bit binary data into *E. coli* populations.

(a) Cells were subjected to electrical signals over three sequential rounds, constituting all 8 possible 3-bit binary data profiles. (b) pTrig copy number profiles for each round of the 3-bit binary data profiles. (c) CRISPR array populations can be described as a frequency distribution constituting of all permutations of reference spacers (R, grey) derived from genome or pRec and trigger spacers (T, red) derived from pTrig for a given array length (L). (d) Frequencies of array-types in \log_{10} scale for each array lengths for the 3-bit data encoded CRISPR array populations. (e) Clustering CRISPR arrays based on their array-type frequency profiles normalized to Z-score across all 3-bit binary profiles. (f) Performance of a Random Forest classifier trained on data from 3 independent experiments and tested on data from 6 subsequent independent experiments. For classification of each sample, an average of 172,788 total sequencing reads with 89,928 reads of expanded arrays (or 38,295 of L2/L3 arrays) that uniquely map spacers were used. Bead-based size enrichment was performed to enrich for expanded arrays and deplete unexpanded arrays (see Methods). All

measurements are based on three or more biological replicates. Error bars represent standard deviation of three biological replicates.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

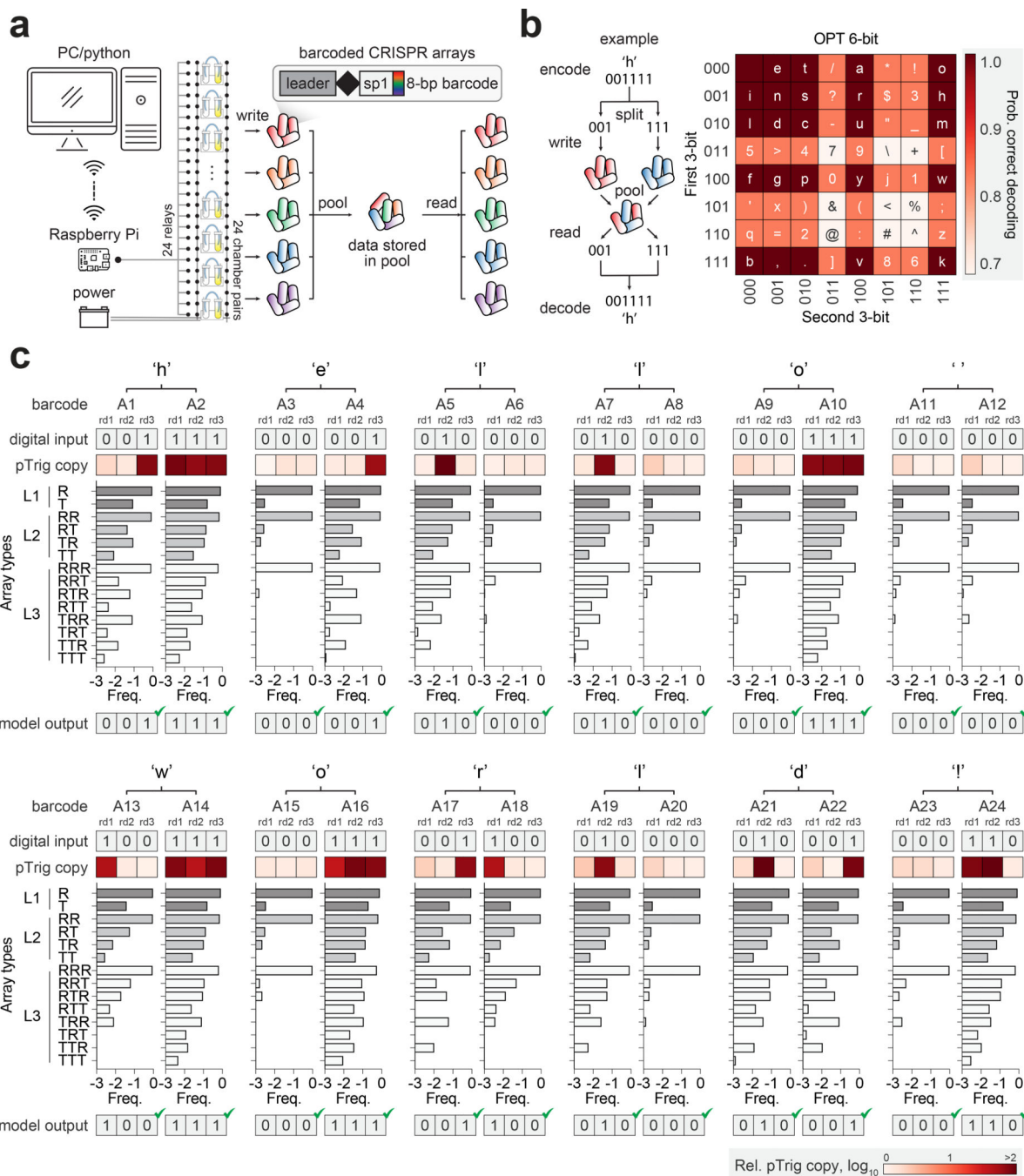


Figure 3. Writing the text message ‘hello world!’ containing 72 bits into barcoded *E. coli* cells. (a) Uniquely barcoded cell populations in each chamber on the multi-channel electrochemical redox controller can receive and store 3-bit binary profiles in parallel split from an original data. The 3-bit encoded cells in each chamber can be pooled and stored. Data can be retrieved by sequencing and demultiplexing barcode sequences for data reconstruction. (b) The OPT 6-bit character table that leverages letter usage frequency and retrieval bias is shown. The 6-bit binary data for each character is split into two barcoded cell populations. An example of encoding ‘h’ is shown. (c) Array-type frequencies (in log₁₀

scale) from a ‘hello world!’ encoded cell population is shown. For classification of each barcoded cell population, an average of 443,051 total sequencing reads with 271,725 reads of expanded arrays (or 179,174 of L2/L3 arrays) that uniquely map spacers were used. Bead-based size enrichment was performed to enrich for expanded arrays and deplete unexpanded arrays (see Methods). All measurements are based on a single encoding experiment.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

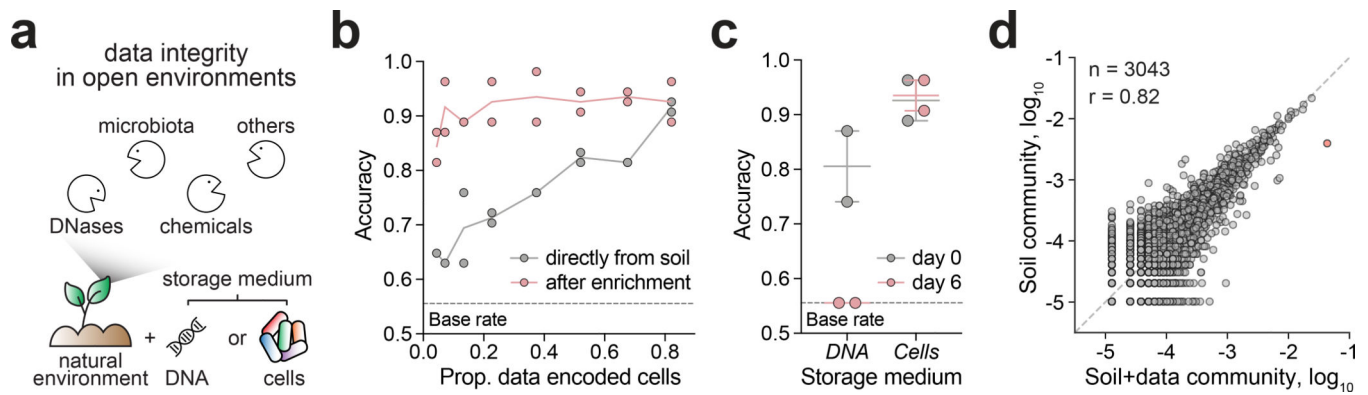


Figure 4. Cell envelope as a physical barrier to protect data.

(a) A data-encoded cell population or naked genomic DNA extracted from the same amount of the data-encoded population was challenged to a natural soil environment. **(b)** Retrieval of a text message ('synbio@cu') from a mixed microbial community of data-encoded *E. coli* cells and natural soil microbiota with and without selective growth enrichment. Accuracy is defined as the proportion of bits that are correctly classified. For classification of each barcoded cell population, an average of 41,740 total sequencing reads with 20,811 reads of expanded arrays (or 9,821 of L2/L3 arrays) that uniquely map spacers were used. Bead-based size enrichment was performed to enrich for expanded arrays and deplete unexpanded arrays. **(c)** Retrieval of text message 'synbio@cu' stored in naked DNA or in encoded cells after exposure to soil for 0 or 6 days. Accuracy is defined as the proportion of bits that are correctly classified. The plot displays the mean values and whiskers span the highest and lowest points. For classification of each barcoded cell population, an average of 20,542 total sequencing reads with 7,868 reads of expanded arrays (or 2,692 of L2/L3 arrays) that uniquely map spacers were used. Bead-based size enrichment was performed to enrich for expanded arrays and deplete unexpanded arrays. **(d)** Comparison of microbial compositions of a natural soil community with and without hidden data-encoded *E. coli* cells (*Escherichia/Shigella* genus is highlighted red, 4% spiked-in). OTUs (n) and Pearson correlation coefficient (r) are shown. Dashed line represents $y=x$. All measurements are based on two biological replicates.