LETTER

doi:10.1038/nature23017

CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria

Seth L. Shipman[1,2,3], Jeff Nivala[1,3], Jeffrey D. Macklis[2] & George M. Church[1,3]

DNA is an excellent medium for archiving data. Recent efforts have illustrated the potential for information storage in DNA using synthesized oligonucleotides assembled *in vitro*[1–6]. A relatively unexplored avenue of information storage in DNA is the ability to write information into the genome of a living cell by the addition of nucleotides over time. Using the Cas1–Cas2 integrase, the CRISPR–Cas microbial immune system stores the nucleotide content of invading viruses to confer adaptive immunity[7]. When harnessed, this system has the potential to write arbitrary information into the genome[8]. Here we use the CRISPR–Cas system to encode the pixel values of black and white images and a short movie into the genomes of a population of living bacteria. In doing so, we push the technical limits of this information storage system and optimize strategies to minimize those limitations. We also uncover underlying principles of the CRISPR–Cas adaptation system, including sequence determinants of spacer acquisition that are relevant for understanding both the basic biology of bacterial adaptation and its technological applications. This work demonstrates that this system can capture and stably store practical amounts of real data within the genomes of populations of living cells.

By combining the principles of information storage in DNA with DNA-capture systems capable of functioning in living cells, we can create living organisms that capture, store, and propagate information over time. In prokaryotic viral defence, the CRISPR-associated (Cas) proteins, Cas1 and Cas2, function as an integrase complex to acquire nucleotides from invading viruses and store them in the CRISPR array[7,9,10]. In previous work, we found that we could direct the system to acquire synthetic sequences into the CRISPR array if those sequences are supplied as oligonucleotides[8]. Using this approach, we showed simple molecular recordings by supplying different oligonucleotide sequences over time.

Here we markedly scale up this approach to define the information capacity that the system can record, with an eye towards future biological recordings. Rather than arbitrary sequences, we encode real information (images) and optimize the method of delivery, nucleotide content of the sequences, and reconstruction method (for which we use a population of bacteria). In the *Escherichia coli* type I-E CRISPR–Cas system, DNA from invading viruses is inserted into a genomic CRISPR array in 33-base units termed spacers[11]. The sequences from which spacers are derived are termed protospacers[12]. We began with an image (Extended Data Fig. 1a) and stored pixel values in a nucleotide code, distributed over many individual synthetic protospacer oligonucleotides. We electroporated these oligonucleotides into a population of bacteria, each harbouring a functional CRISPR array and overexpressing the Cas1–Cas2 integrase complex, allowing cells to acquire the oligonucleotides into their genome. We recover the information by high-throughput sequencing: newly acquired spacers are decoded to reconstruct the original image.

We first encoded images of a human hand using two different pixel-value-encoding strategies: a rigid strategy (hand^{R}), in which 4 pixel colours were each specified by a different base (Extended Data Fig. 1b, c); and a flexible strategy (hand^{F}), in which 21 possible pixel colours were specified by a degenerate nucleotide triplet table (Fig. 1a, b). To distribute the information across multiple protospacers, we gave each protospacer a barcode that defined which pixel set (denoted as 'pixet') was encoded by the nucleotides in that spacer. Four nucleotides define each pixet, and the pixels of a given pixet are distributed across the image (Fig. 1c, Extended Data Fig. 1d). We included a protospacer adjacent motif (PAM) on each protospacer, which increases the efficiency of acquisition and determines orientation of spacer insertion[8,13–15]. After adding the PAM and pixet, we were left with 28 bases per protospacer to encode pixel values.

For hand^{R}, each of the 28 bases encoded a pixel value, thereby distributing a 4-colour, 56×56 pixel image across 112 oligonucleotide protospacers (total information content of 784 bytes). For hand^{F}, the 28 bases encoded 9 pixels, each specified by a nucleotide triplet. Specific triplet combinations were chosen to build sequences that we hypothesized might increase acquisition efficiency—GCcontent around 50%, no mononucleotide repeats >3 bp, and no internal PAMs. For the hand^{F}, we distributed a 21-colour, 30×30 pixel image across 100 protospacers (total information content of around 494 bytes). Oligonucleotide protospacers were supplied in a minimal hairpin format (design based on insights from the crystal structure[16,17]) to prevent segregation of the two strands into different cells during electroporation (see Supplementary Information, Extended Data Fig. 2).

For each image, we electroporated the pooled oligonucleotides into a population of *E. coli* containing a genomic CRISPR array and expressing the Cas1–Cas2 integrase[18]. Cells were then recovered, passaged overnight, and the next day a sample of the genomic CRISPR arrays were sequenced. Newly acquired spacers were bioinformatically extracted from the arrays, and those that were not derived from the plasmid or genome were analysed. Pixel values were assigned on the basis of the most numerous new spacer with a given pixet. Images reconstructed from the hand^{R} and hand^{F} images are shown in Extended Data Fig. 1e and Fig. 1d, respectively.

Using 655,360 reads, around 88% and around 96% of pixet sequences were accurately recalled from the hand^{R} and hand^{F} images, respectively. We found that hand^{F} was more resistant to errors by under-sampling (Fig. 1e, f, Extended Data Fig. 1f–g). By electroporating subsets of the oligonucleotides, we found that the number of reads required to achieve similar levels of accuracy in recall is linearly related to the number of oligonucleotides electroporated (Fig. 1g, Extended Data Fig. 1h, i), and that it took substantially more reads per oligonucleotide protospacer to reach 80% accuracy from hand^{R} (around 1,580 reads per protospacer) versus hand^{F} (around 150 reads per protospacer).

We also sampled time points of the bacterial culture following the electroporation of hand^{F}. Oligonucleotide-derived spacer acquisitions

[1]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [2]Department of Stem Cell and Regenerative Biology, Center for Brain Science, and Harvard Stem Cell Institute, Harvard University, Bauer Laboratory 103, Cambridge, Massachusetts 02138, USA. [3]Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, Massachusetts 02138, USA.

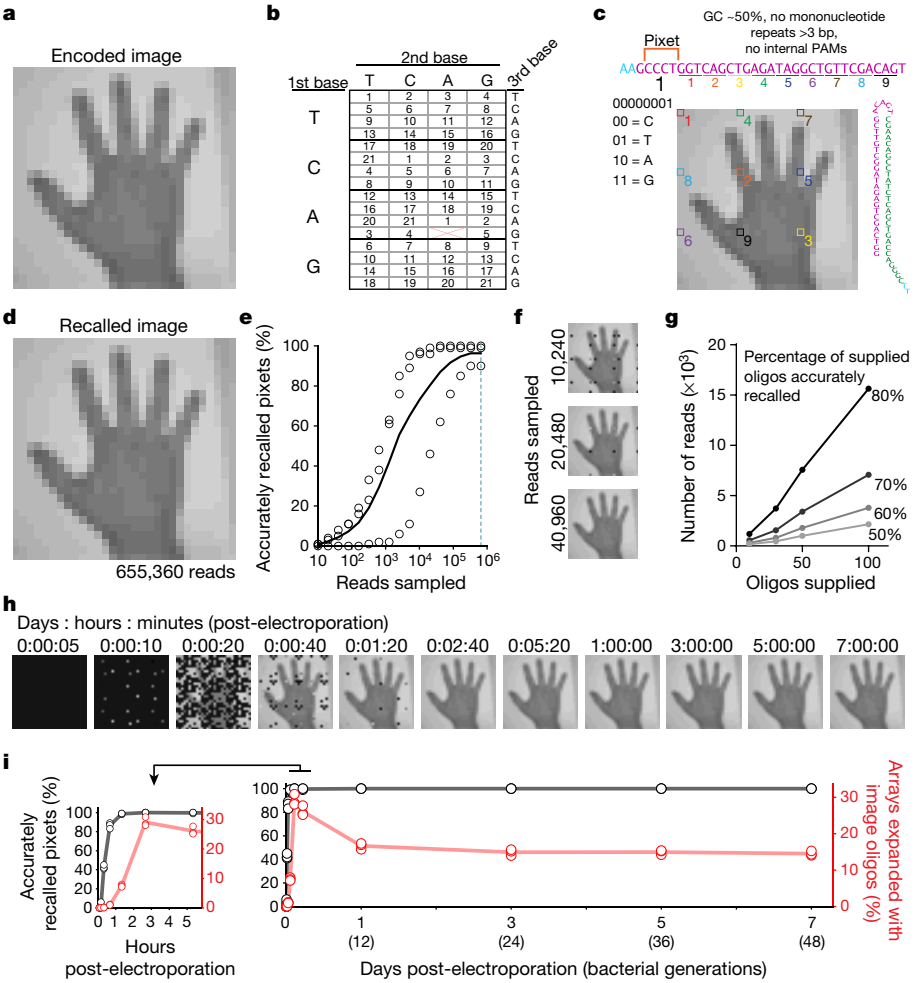


Figure 1 | An image into the genome. a, Hand^{F} image. **b**, Encoding for 21 colours. **c**, Sequence at top shows the linear protospacer with pixet code followed by pixel values (distributed across image). Pixet shown under nucleotides, with binary-to-nucleotide conversion. Small colourful numbers below protospacer indicate individual pixels boxed on the image. Minimal hairpin protospacer shown on the right. **d**, One replicate at 655,360 reads. Black shown if no pixel information recovered. **e**, Accurately recalled pixets by read depth. Unfilled circles indicate points from 3 biological replicates, black line shows the mean. **f**, Result of down-sampling the sequencing reads. **g**, Reads required to reach 50%, 60%, 70%, and 80% accuracy on a given oligonucleotide set as a function of number of oligonucleotides supplied ($n = 3$; linear regression of the 80% curve, $R^2 = 0.9975$; runs test of the 80% curve, $P > 0.99$). **h**, Image recall at time points after electroporation. **i**, Quantification of the percentage of accurately recalled pixets (in black) and percentage of arrays with oligonucleotide-derived spacers (in red) by time point. Unfilled circles represent 3 biological replicates, lines show the mean. Inset graph (left) expands first six hours. Statistical details in Supplementary Table 1.

were detectable ten minutes after the electroporation, and peaked at 2 h 40 min, at which point we could first accurately recall the entire image (Fig. 1h, i). From this peak to 24 h post-electroporation, the percentage of oligonucleotide-expanded arrays declined slightly, then stabilized over the next six days (~48 bacterial generations). Presumably, some cells lose viability following the electroporation and do not contribute to the population after outgrowth (Extended Data Fig. 3, Supplementary Information includes information about the internal integrity of the arrays over time[19–21]).

The total acquisition frequency was higher for hand^{F} than hand^{R}, explaining the improvement in recall (Extended Data Fig. 4a). To test which of the parameters—percentage of GC content (GC%), absence of mononucleotide repeats, or lack of internal PAMs—accounted for this greater acquisition frequency, we designed new sets of oligonucleotide protospacers, systematically testing each parameter (Supplementary Information, Extended Data Fig. 4b–f). GC% had a clear effect on acquisition frequency, with reduced acquisition frequency at low GC%. This effect was most extreme when pools contained a wide range of GC%. In pools with homogeneous percentage, those over 50% were equally effective. Therefore, it is beneficial to limit the range of GC% and keep the percentage at 50% or higher. The protospacers encoding the hand^{R} image had, by chance, an overall lower GC% than those encoding the hand^{F} image (41.8 ± 0.6% versus 50.6 ± 0.6%), which may account for the difference in acquisition frequency. For mononucleotide repeats and internal PAMs, we found that pools of protospacers with substantial numbers of either displayed reductions in total acquisitions.

Despite differences in acquisition frequency between hand^{R} and hand^{F}, a similar range of acquisition frequencies is apparent among individual protospacer sequences from each (Fig. 2a). We compared over-represented protospacers with all protospacers and found a significant motif present in the final two nucleotides in over-represented sequences (Fig. 2b). A similar motif has been previously reported and termed the acquisition affecting motif (AAM)[22]; however, the reported sequence of this motif differs from what we find here. Although we found the motif to be composed of slightly different bases, we believe that these differences probably to arise from the fact that we were able to synthetically control for the presence of the more dominant PAM motif in our sequences, and thus we adopt the previous term AAM. The PAM sequence has been shown to function not only in adaptation but also interference[23] and, given that it lies outside of the acquired spacer, serves as a mechanism of

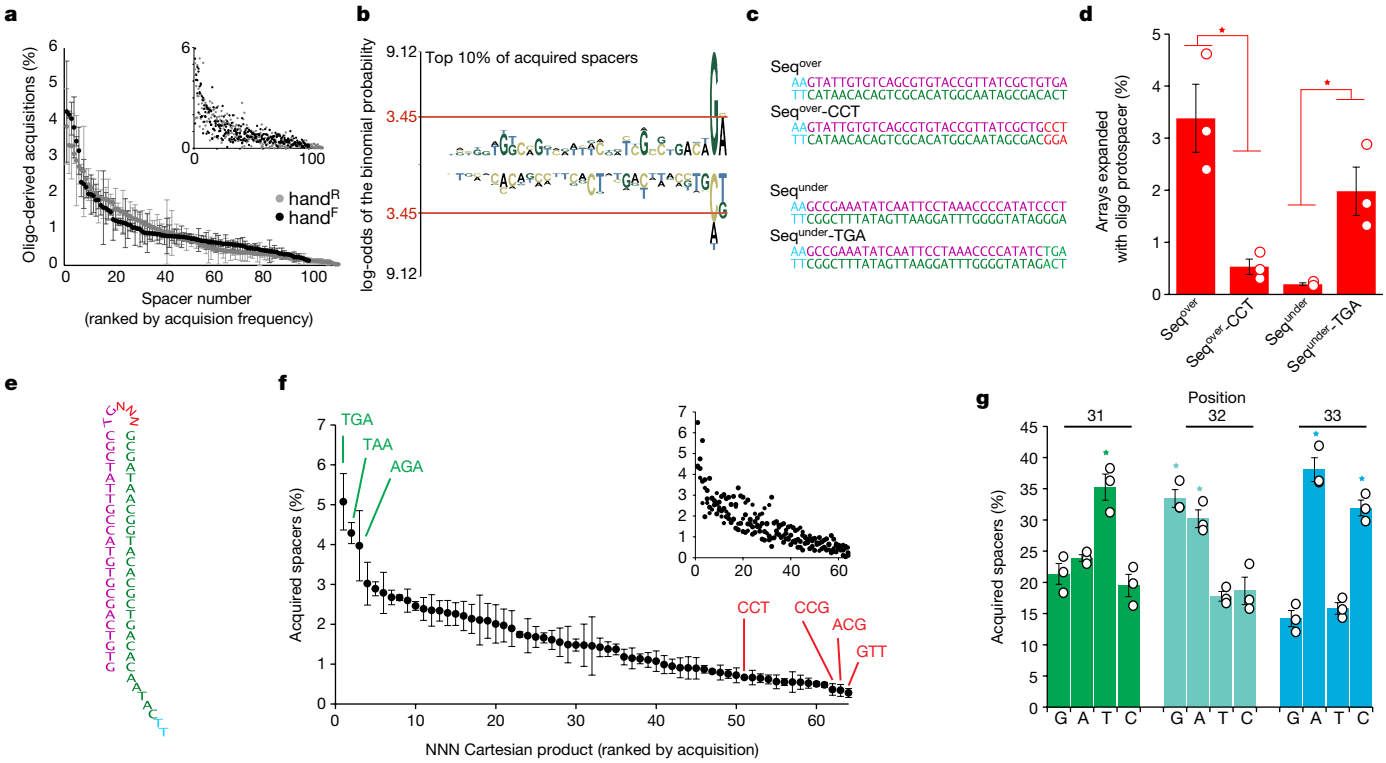


Figure 2 | Sequence determinants of acquisition. a, Acquisition frequency for individual protospacers (of oligonucleotide-derived acquisitions) for both images, ranked by frequency. Main plot circles represent mean ± s.e.m. Smaller inset shows each replicate ($n = 3$). **b**, pLogo[30] of the top 10% of protospacers (all protospacers as background). Red line indicates $P < 0.05$. Over-representation is positive, under-representation is negative ($n = 3$). **c**, Sequences designed to test the motif. **d**, Arrays expanded with the sequences indicated in **c**. Unfilled circles represent individual replicates. Bars show mean ± s.e.m. ($n = 3$; one-way ANOVA on effect of oligo: $P = 0.002$; follow-up Sidak's multiple comparison (corrected), seq^{over} versus seq^{over}-CCT: $P = 0.0023$, seq^{under} versus seq^{under}-TGA: $P = 0.0294$). **e**, NNN-containing oligonucleotide. **f**, Acquisition frequency of protospacers containing each NNN Cartesian product (of oligonucleotide-derived acquisitions), ranked by frequency ($n = 3$). Plots as in **a**. **g**, Representation of nucleotides at positions 31–33 in acquired spacers from the NNN-containing oligonucleotide ($n = 3$; one-way ANOVA on effect of nucleotide, position 31: $P = 0.0006$, position 32: $P = 0.0002$, position 33: $P < 0.0001$; follow-up Tukey's multiple comparison (corrected) see Supplementary Table 1). Plot as in **d**. $^*P < 0.05$. Statistical details in Supplementary Table 1.

self versus non-self discrimination[14]. Although the AAM lies within the acquired spacer (and thus could promote self-targeting) and additionally lies outside the seed region[24], it would still be interesting to directly test whether the presence of an AAM similarly influences interference efficiency.

To test whether the AAM motif that we found is responsible for the difference in acquisition efficiency, we tested individual protospacers, using nucleotides from the over-represented motif, ending in 'TGA', to define one protospacer (seq^{over}), and nucleotides drawn from the under-represented motif, ending in 'CCT', to define another (seq^{under}). We also swapped the final three nucleotides from these two sequences to create two more protospacers (seq^{over}-CCT and seq^{under}-TGA) (Fig. 2c). We found that the final three nucleotides determined acquisition frequency, with 'TGA' yielding high efficiency and 'CCT' yielding low efficiency, regardless of the rest of the sequence content (Fig. 2d). Because these nucleotides are in the loop region of the hairpin protospacer, we also tested these sequences as complementary single-stranded oligonucleotides and found an identical dependence on the final three nucleotides (Extended Data Fig. 5).

Because we identified this motif using sequence-constrained protospacers, we tested a hairpin protospacer with random nucleotides (NNN) in the final three positions (Fig. 2e). Although we observed acquisition events with every possible NNN Cartesian product in the three variable nucleotides, their efficiencies varied and allowed us to define the ideal AAM (Fig. 2f, g).

We next applied our better understanding of protospacer sequence determinants of acquisition to encode multiple images over time within a single population of bacteria, generating a short movie (GIF). We moved the pixet to the final nucleotides of the protospacer, where a reduced sequence space was employed, limited to the most efficient eight AAM triplets from Fig. 2f (Fig. 3a). We again used the flexible 21-colour code from the hand^F image and chose to encode five frames of a galloping mare from Eadweard Muybridge's *Human and Animal Locomotion* at 36 × 26 pixels. Frames were each represented by a unique oligonucleotide set of 104 protospacers, for an overall information content of around 2.6 kilobytes. Pixet codes were reused between frames, and no nucleotides were used to identify frame order. Rather, each frame was electroporated successively over five days into a single population (Fig. 3b). Because new spacers are almost always acquired adjacent to the leader sequence in the CRISPR array[18], pushing previously acquired spacers away from the leader, the order of frames within the GIF can be reconstructed on the basis of the pairwise order of spacers among many individual arrays.

Following electroporation, we found that the protospacers were efficiently acquired from each frame, and populated the first three sequenced positions of CRISPR arrays (Fig. 3c). We extracted all new spacers from the arrays, then analysed the pixet nucleotides to recover the spacers assigned to each unique pixet, but in this case, captured the five most frequently acquired spacers with each pixet—one for each frame.

To order the frames over time, we used positioning within individual arrays to reconstruct the electroporation order of the protospacers. Ordering information can only be recovered from single cells, in which spacers further from the leader within a single array must have been acquired earlier than spacers closer to the leader in that same array. However, the GIF information is widely distributed among a

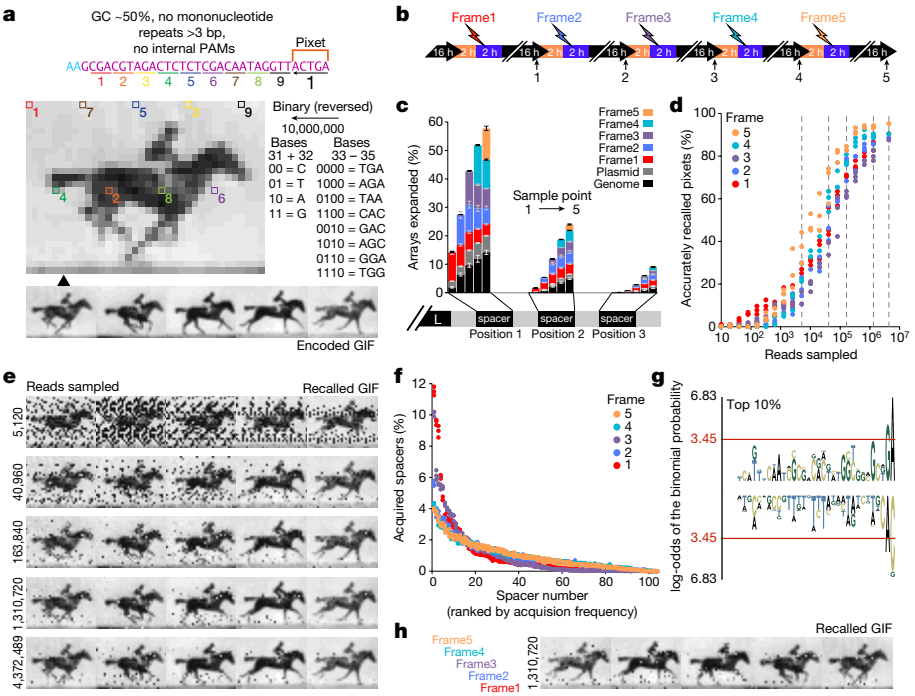


Figure 3 | Encoding a GIF in bacteria. a, GIF to be encoded (adapted from Eadweard Muybridge, *Human and Animal Locomotion*, plate 626, thoroughbred bay mare 'Annie G.' galloping, Wikimedia Commons), along with an example of one pixet protospacer. **b**, Schematic of recording process. **c**, Percentage of arrays with expansions in the first three positions, by protospacer origin, at each sample point. Bars show mean ± s.e.m. ($n = 3$). **d**, Accurately recalled pixets as a function of reads (on the *x* axis) and frame (denoted by colour). Points show individual biological replicates. **e**, Examples of the result at different sequence depths (see dotted grey lines in **d**). **f**, Protospacer acquisition frequency for individual protospacers (of oligonucleotide-derived acquisitions) by frame, ranked by acquisition frequency. Points show 3 biological replicates. **g**, pLogo[30] of the top 10% of protospacers (all protospacers as background). Red line indicates $P < 0.05$. Over-representation is positive, under-representation is negative. **h**, Result of electroporating the same oligonucleotides in the reverse order. Statistical details in Supplementary Table 1.

population of bacteria — no individual cell can be used to reconstruct the entire image series. Therefore, we leveraged many single-cell ordering comparisons among the population of bacteria to reconstruct the entire GIF (see Supplementary Information, Extended Data Fig. 6 for detail).

We found that we could reconstruct each frame and the order of frames (Fig. 3d), and that increasing read depth aided the accuracy of the reconstruction (to >90% overall accuracy) (Fig. 3e). Despite optimization of the protospacer sequence, we still found a range of efficiencies between the protospacers of any given frame (Fig. 3f). We again found a sequence motif at the AAM location, suggesting that we allowed for too large a range of nucleotide triplets in the final position (Fig. 3g) or this range may reflect an inherent competition among protospacers, either for Cas1–Cas2 or the genomic array. As the protospacers themselves contain no code to specify frame position, we tested the robustness of our reconstruction strategy by delivering the oligonucleotide frame sets in reverse order. We were able to accurately reconstruct the reversed GIF, demonstrating reconstruction of an otherwise ambiguous signal based on time (Fig. 3h).

In summary, we found that not all protospacer sequences are equally effective at transferring data into the genome, and for this reason advocate for the use of a flexible encoding scheme to allow optimization of sequence content. We found that sequences with controlled GC content, a lack of mononucleotide repeats, and no internal PAMs outperformed those that lacked such optimization. Further, the inclusion of invariant nucleotides at both the leading (AAG) and trailing (GA) end of the protospacer has large effects on the frequency of acquisition. We were able to track the presence of 104 separately barcoded sequence elements over five time points (520 unique sequence elements), yielding confidence that this system will be capable of recording multidimensional biological information (see Supplementary Information, Extended Data Figs 7, 8 for discussion into error-correction/compression, obstacles to single-cell storage, and a comparison of information storage in DNA versus silicon[25–29]).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 August 2016; accepted 2 June 2017.
Published online 12 July 2017.

1. Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. *Science* **337,** 1628 (2012).
2. Goldman, N. *et al.* Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494,** 77–80 (2013).
3. Gibson, D. G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329,** 52–56 (2010).
4. Clelland, C. T., Risca, V. & Bancroft, C. Hiding messages in DNA microdots. *Nature* **399,** 533–534 (1999).
5. Adleman, L. M. Molecular computation of solutions to combinatorial problems. *Science* **266,** 1021–1024 (1994).
6. Davis, J. Microvenus. *Art J.* **55,** 70–74 (1996).
7. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315,** 1709–1712 (2007).
8. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science* **353,** aaf1175 (2016).
9. Amitai, G. & Sorek, R. CRISPR–Cas adaptation: insights into the mechanism of action. *Nat. Rev. Microbiol.* **14,** 67–76 (2016).
10. Sternberg, S. H., Richter, H., Charpentier, E. & Qimron, U. Adaptation in CRISPR–Cas Systems. *Mol. Cell* **61,** 797–808 (2016).
11. van der Oost, J., Jore, M. M., Westra, E. R., Lundgren, M. & Brouns, S. J. CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem. Sci.* **34,** 401–407 (2009).
12. Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J. Bacteriol.* **190,** 1390–1400 (2008).
13. Paez-Espino, D. *et al.* Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nat. Commun.* **4,** 1430 (2013).
14. Westra, E. R. *et al.* Type I-E CRISPR–Cas systems discriminate target from non-target DNA through base pairing-independent PAM recognition. *PLoS Genet.* **9,** e1003742 (2013).

15. Shmakov, S. *et al.* Pervasive generation of oppositely oriented spacers during CRISPR adaptation. *Nucleic Acids Res.* **42,** 5907–5916 (2014).
16. Nuñez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N. & Doudna, J. A. Foreign DNA capture during CRISPR–Cas adaptive immunity. *Nature* **527,** 535–538 (2015).
17. Wang, J. *et al.* Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR–Cas systems. *Cell* **163,** 840–853 (2015).
18. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40,** 5569–5576 (2012).
19. Diez-Villasenor, C., Almendros, C., Garcia-Martinez, J. & Mojica, F. J. Diversity of CRISPR loci in *Escherichia coli*. *Microbiology* **156,** 1351–1361 (2010).
20. Weinberger, A. D. *et al.* Persisting viral sequences shape microbial CRISPR-based immunity. *PLOS Comput. Biol.* **8,** e1002475 (2012).
21. Held, N. L., Herrera, A., Cadillo-Quiroz, H. & Whitaker, R. J. CRISPR associated diversity within a population of *Sulfolobus islandicus*. *PLoS One* **5,** e12988 (2010).
22. Yosef, I. *et al.* DNA motifs determining the efficiency of adaptation into the *Escherichia coli* CRISPR array. *Proc. Natl Acad. Sci. USA* **110,** 14396–14401 (2013).
23. Westra, E. R. *et al.* CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by cascade and Cas3. *Mol. Cell* **46,** 595–605 (2012).
24. Semenova, E. *et al.* Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl Acad. Sci. USA* **108,** 10098–10103 (2011).
25. Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. Nucleic acid memory. *Nat. Mater.* **15,** 366–370 (2016).
26. Hsiao, V., Hori, Y., Rothemund, P. W. & Murray, R. M. A population-based temporal logic gate for timing and recording chemical events. *Mol. Syst. Biol.* **12,** 869 (2016).
27. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353,** aaf7907 (2016).
28. Frieda, K. L. *et al.* Synthetic recording and *in situ* readout of lineage information in single cells. *Nature* **541,** 107–111 (2017).
29. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355,** 950–954 (2017).
30. O'Shea, J. P. *et al.* pLogo: a probabilistic approach to visualizing sequence motifs. *Nat. Methods* **10,** 1211–1212 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements S.L.S. is a Shurl and Kay Curci Foundation Fellow of the Life Sciences Research Foundation. The project was supported by grants from the National Institute of Mental Health (5R01MH103910), National Human Genome Research Institute (5RM1HG008525), and Simons Foundation Autism Research Initiative (368485) to G.M.C., the National Institute of Neurological Disorders and Stroke (5R01NS045523) to J.D.M and an Allen Distinguished Investigator Award from the Paul G. Allen Frontiers Group to J.D.M. We thank G. Kuznetsov for comments on the manuscript.

Author Contributions S.L.S. and J.N. conceived the study. S.L.S. designed the work, performed experiments, analysed data, wrote custom Python analysis software, and wrote the manuscript with input from J.N., J.D.M. and G.M.C. S.L.S. J.N., J.D.M. and G.M.C. discussed results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to G.M.C. (gchurch@genetics.med.harvard.edu).

Reviewer Information *Nature* thanks R. Barrangou and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. Colony counts were performed blind to experimental condition using Amazon's Mechanical Turk, the investigators were otherwise not blinded to allocation during experiments and outcome assessment.

Bacterial strains and culturing conditions. All experiments were carried out in BL21-AI *E. coli* (Thermo Fisher), containing an integrated, arabinose-inducible T7 polymerase, an endogenous CRISPR array, but no endogenous Cas1 and Cas2. The strain was authenticated based on the known endogenous CRISPR array and expression of T7 polymerase with arabinose induction. A plasmid encoding inducible (T7/lac) Cas1+2 (K-strain origin, pWUR1+2 a.k.a. pCas1+2) was transformed into cells before each experiment. Cells containing the plasmid were maintained in colonies on a plate at 4 °C for up to three weeks. Bacterial cultures were used under antibiotic conditions, and mycoplasma testing was not necessary.

Oligonucleotide protospacer electroporation. Protospacer electroporations were performed as previously described[8]. Briefly, after overnight outgrowth from a single colony, Cas1 and Cas2 were induced in a 3 ml dilution of the culture (containing 80ul of the overnight), and grown at 37 °C for 2 h (L-arabinose 0.2% w/w, Sigma-Aldrich; isopropyl-β-D-thiogalactopyranoside 1 mM, Sigma-Aldrich). For a given condition, 1 ml of the induced culture was spun down and washed with water three times at 4 °C, then resuspended in 50 μl of a 6.25 μM solution (unless other concentration is noted) of either a single protospacer or set of multiple protospacers and electroporated in a 1 mm gap cuvette using a Bio-Rad gene pulser set to 1.8 kV and 25 μF. Only those conditions with an electroporation time constant >4.0 ms were carried through to analysis. After electroporation, cells were recovered in 3 ml LB at 37 °C for 2–3 h, then diluted (50 μl) into a fresh 3 ml culture and grown overnight. Cells were collected for analysis the following morning (unless otherwise noted). For checking the maintenance of the 21-colour image over time, cells were passaged daily (50 μl into 3 ml) after the first 24 h and grown at 30 °C. To estimate the number of bacterial generations, we calculated the number of doublings required take the starting dilution (50 μl into 3 ml) to saturation at 1×10^9 cells per ml (note that the first dilution was not from a saturated culture; in this case empirically determined cell numbers were taken from Extended Data Fig. 3). The oligonucleotide protospacers used can be found in Supplementary Table 3. To estimate the number of cells surviving electroporation, cells were serially diluted 1:300 (normalized to 1 ml of starting culture), then 1:400 before plating on spectinomycin-containing plates. The resulting colonies were imaged using a commercial document scanner. To obtain colony counts blinded to experimental condition, partial or complete plate images were uploaded to Amazon's Mechanical Turk workplace where remote workers were asked to count the number of 'dots' per image. The answers provided by 10 workers were averaged for each plate image, from which the colony-forming units per millilitre of starting volume were calculated.

Analysis of spacer acquisition. To analyse spacer acquisition, bacteria were lysed by heating to 95 °C for 5 min, then subjected to PCR of their genomic arrays using primers that flank the leader-repeat junction and additionally contain Illumina-compatible adapters. Libraries of up to 96 dual-indexed samples were sequenced on a MiSeq sequencer (Illumina) to read up to three spacer positions in from the leader on each array. Spacer sequences were extracted bioinformatically on the basis of the presence of flanking repeat sequences, and compared against pre-existing spacer sequences to determine the percentage of expanded arrays and the position and sequence of newly acquired spacers. New spacers were blasted (NCBI) against the genome and plasmid sequences to determine the origin of the protospacer, with those sequences not derived from the genome or plasmid assumed to be oligonucleotide-derived. As each cell contains a single array, the read depth is roughly equivalent to the number of cells analysed (including both expanded and unexpanded arrays). This and all subsequent image analysis was performed using custom written scripts in Python.

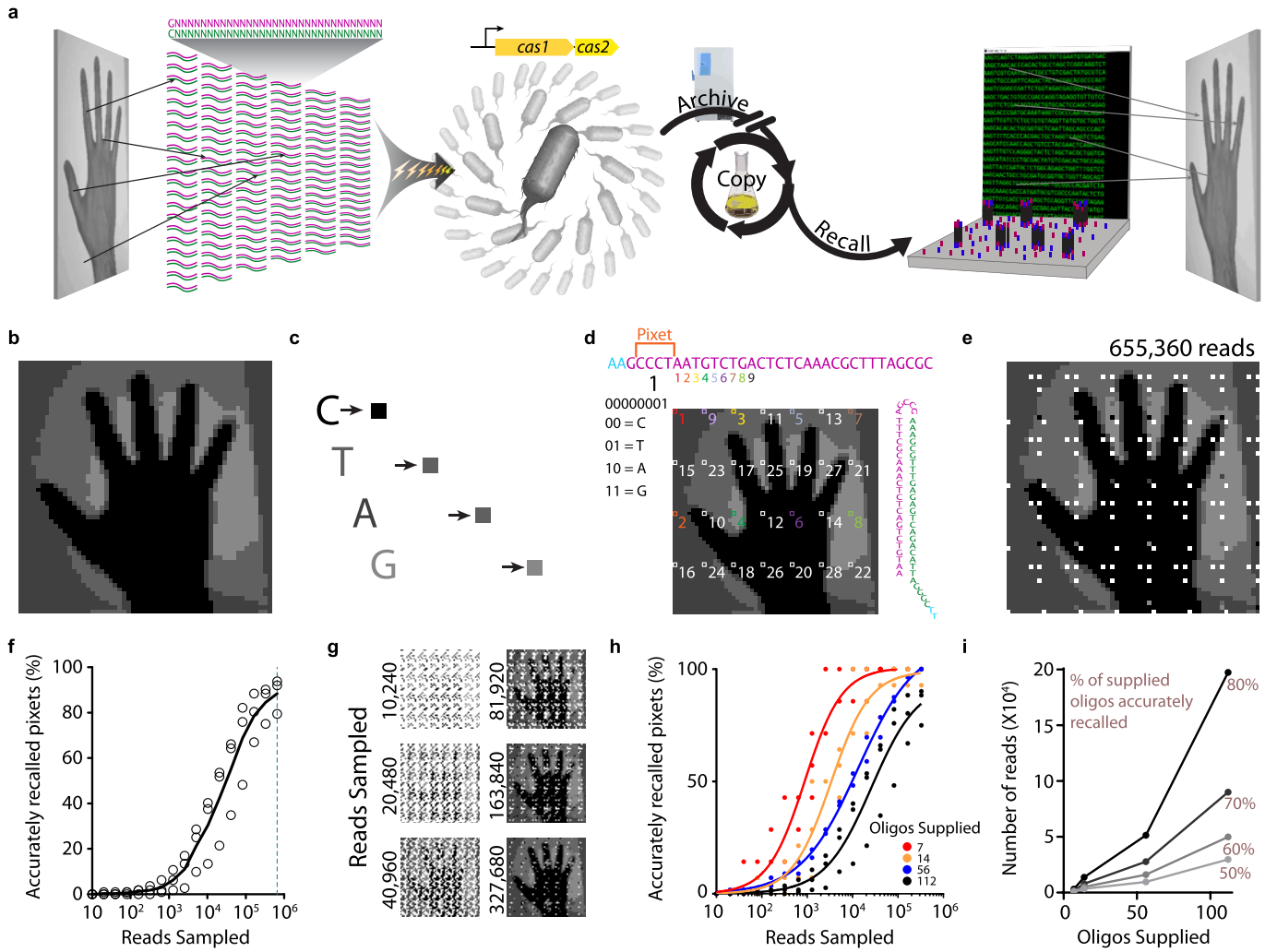
Image coding and decoding. Image protospacer sets were created using a custom Python script to first open and read the pixel values of a previously created image. Each protospacer was given a pixet code by a binary-to-nucleotide conversion, and populated by nucleotides encoding the pixel values according to the scheme detailed in the text. For the single images, the pixet code was interleaved between ascending and descending numbers to introduce more sequence diversity in neighbouring pixet protospacers. In the case of the flexible code used in Figs 2 and 3, the protospacer was built sequentially. For each new pixel value the three possible nucleotide sequences were ranked according to which triplet would best push GC% of the resulting sequence towards 50%, then tested for whether the addition of the triplet would create either an internal PAM or a mononucleotide repeat >3. If such a situation was created, the next triplet in the list was tested until an acceptable triplet was identified (Extended Data Fig. 8a–d). For the hand^F image, the final base was assigned to the least numerous base in the rest of the spacer. We did not attempt to actively exclude sequences that matched the plasmid or genome, as this would be an exceedingly unlikely event given our library sizes. Finally, the sequences were re-formatted to match the minimal hairpin structure and written to a spreadsheet for synthesis by Integrated DNA Technologies. For the GIF, this process was repeated for each frame.

To reconstruct the single images, newly acquired oligonucleotide-derived spacers (plasmid- and genome-derived spacers were set aside before this analysis) were ranked according to frequency of acquisition, then the most frequent spacer sequences for each pixet (by the reversed nucleotide to binary conversion) were assigned to that pixet. Pixel values were extracted from the remaining spacer sequence according to the schemes outlined in the text, and figures and used to populate an image. The more complicated reconstruction of the GIF is described in detail in Supplementary Information as are the calculations of information content.

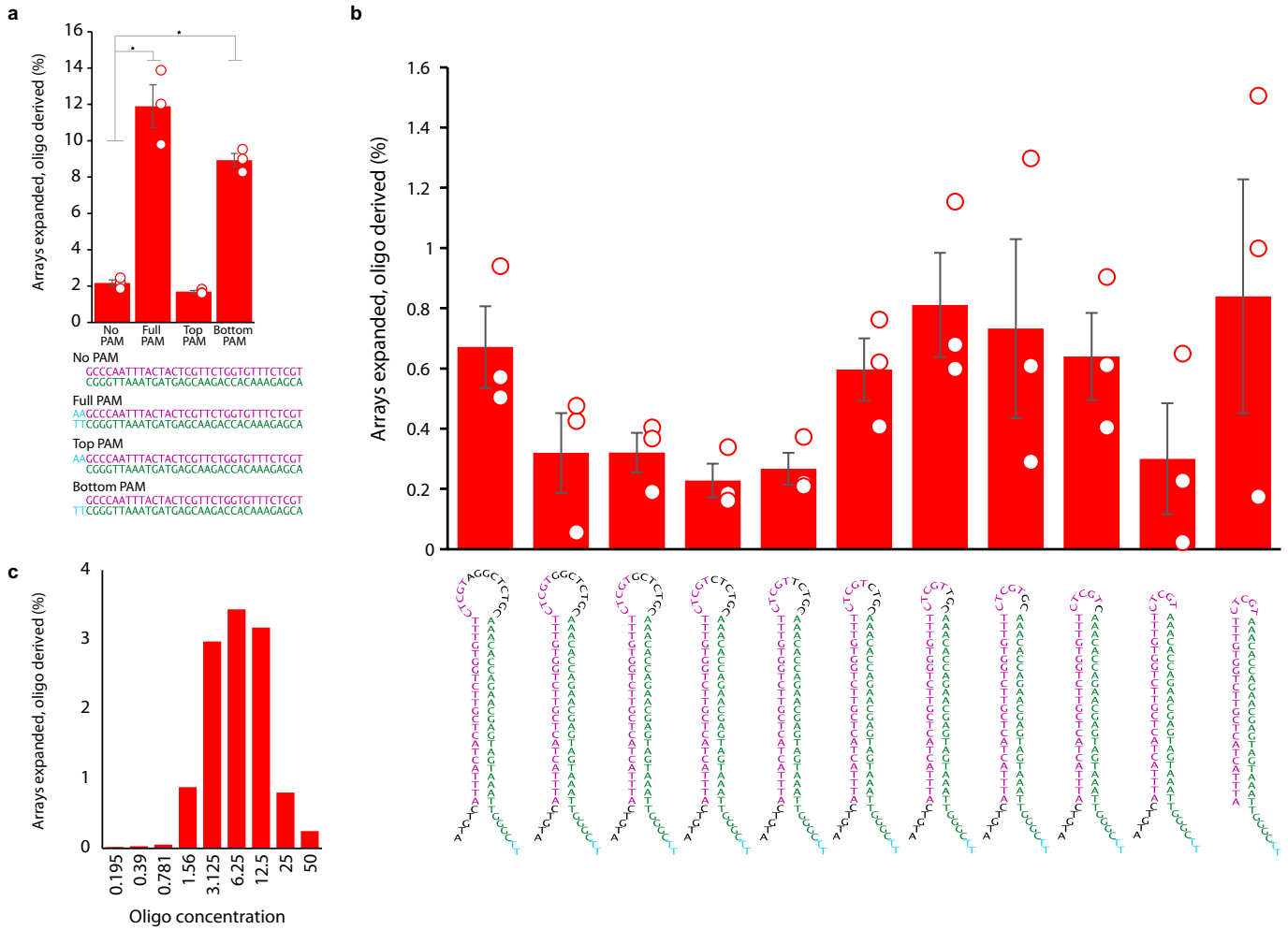
Statistics. A list of statistical tests can be found in Supplementary Tables 1 and 2.

Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request.

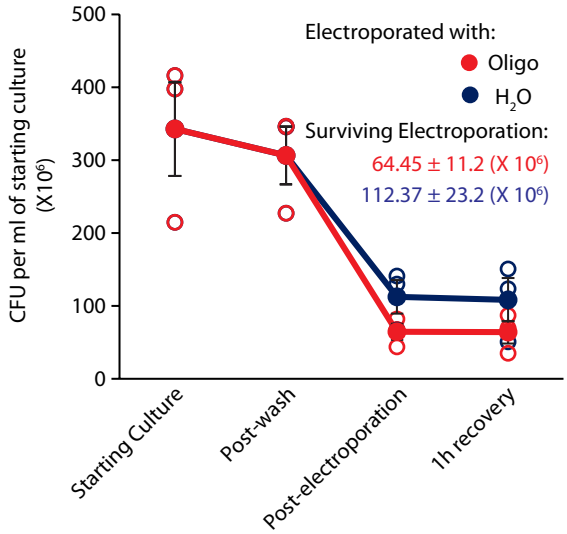
Code availability. Custom code used in this study can be accessed at https://github.com/churchlab/crispr-images.



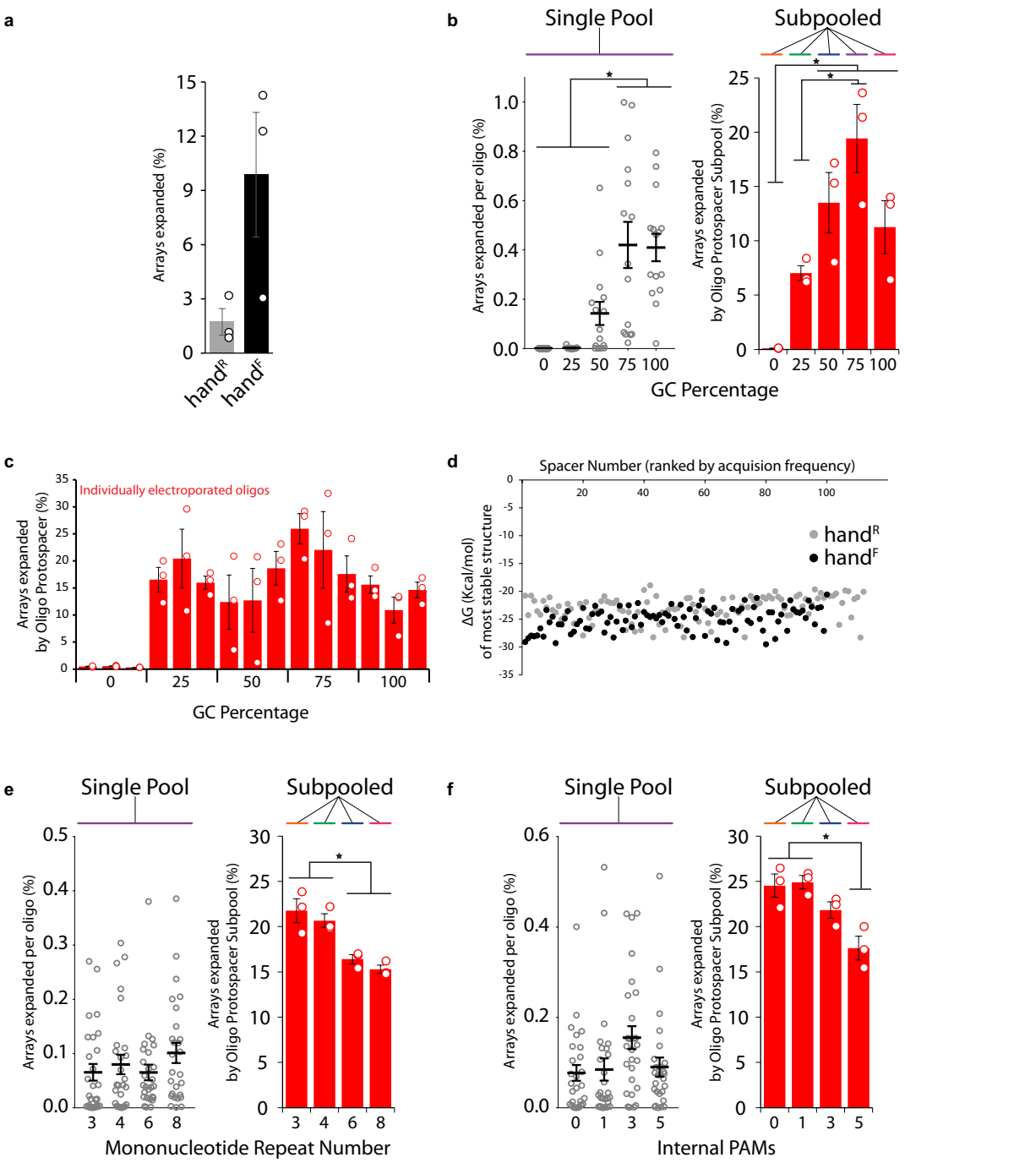
Extended Data Figure 1 | Recording images into the genome. a, Pixel values are encoded across many protospacers, which are electroporated into a population of bacteria that overexpress Cas1 and Cas2 to store the image data. These bacteria can be archived, propagated, and eventually sequenced to recall the image. **b**, Initial image to be encoded. **c**, Nucleotide-to-colour encoding scheme. **d**, Example of the encoding scheme. Sequence at top shows the protospacer linear view with pixet code (specifying a pixel set) followed by pixel values, which are distributed across the image. Pixet number is shown under the pixet nucleotides, with the binary-converted pixet and binary-to-nucleotide conversion reference below that. Small numbers (in colour) below the protospacer indicate individual pixels, identified by boxes on the image. Protospacer in minimal hairpin format for electroporation is shown on the right. **e**, Results of one replicate at a depth of 655,360 reads. White is shown if no information was recovered about the pixel value (owing to a pixet protospacer not being recovered after sequencing). **f**, Percentage of accurately recalled pixets as a function of read depth. Unfilled circles indicate points derived from 3 biological replicates. The black line is the mean of the replicates. **g**, Examples of the images that result from down-sampling the sequencing reads. **h**, Effect of supplying fewer oligonucleotides on recall accuracy as a function of reads sampled when smaller pools of oligonucleotides are supplied and recalled. Individual points show 3 biological replicates, lines are the means of the replicates. **i**, Number of reads required to reach 50%, 60%, 70%, and 80% accuracy on a given oligonucleotide set as a function of oligonucleotides supplied ($n = 3$; linear regression of the 80% curve, R^2=0.9466; runs test of the 80% curve, $P > 0.99$). Additional statistical details in Supplementary Table 2.



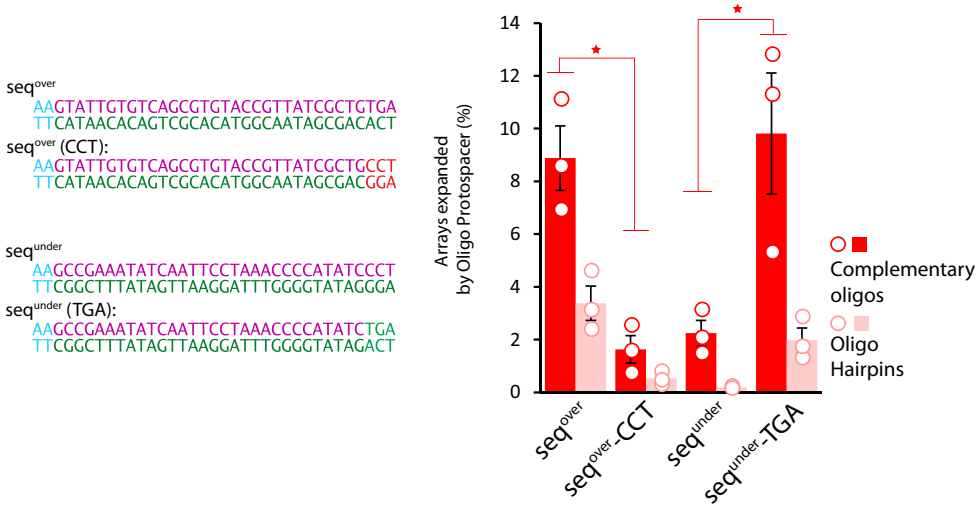
Extended Data Figure 2 | Testing a minimal hairpin protospacer. **a**, Percentage of arrays expanded with oligonucleotide-supplied spacers following electroporation of the sequences indicated below, aimed at testing PAM inclusion on both the top and bottom strands. Unfilled circles indicate biological replicates, bars are mean ± s.e.m ($n = 3$; one-way ANOVA: $P < 0.0001$; follow-up Dunnett's multiple comparison (corrected), no PAM versus full PAM: $P = 0.0001$, no PAM versus bottom PAM: $P = 0.0002$). *$P < 0.05$. Oligonucleotides supplied at 3.125 μM each. **b**, Percentage of arrays expanded with oligonucleotide-supplied spacers following electroporation of the sequences indicated to the left, right, and below aimed at finding a minimal functional hairpin protospacer. Unfilled circles indicate individual biological replicates, bars are mean ± s.e.m ($n = 4$; one-way ANOVA effect of protospacer: $P > 0.05$). Oligonucleotides supplied at 3.125 μM. **c**, Percentage of arrays expanded following electroporation of different concentrations of the minimal hairpin oligonucleotide protospacer ($n = 1$). Additional statistical details in Supplementary Table 2.



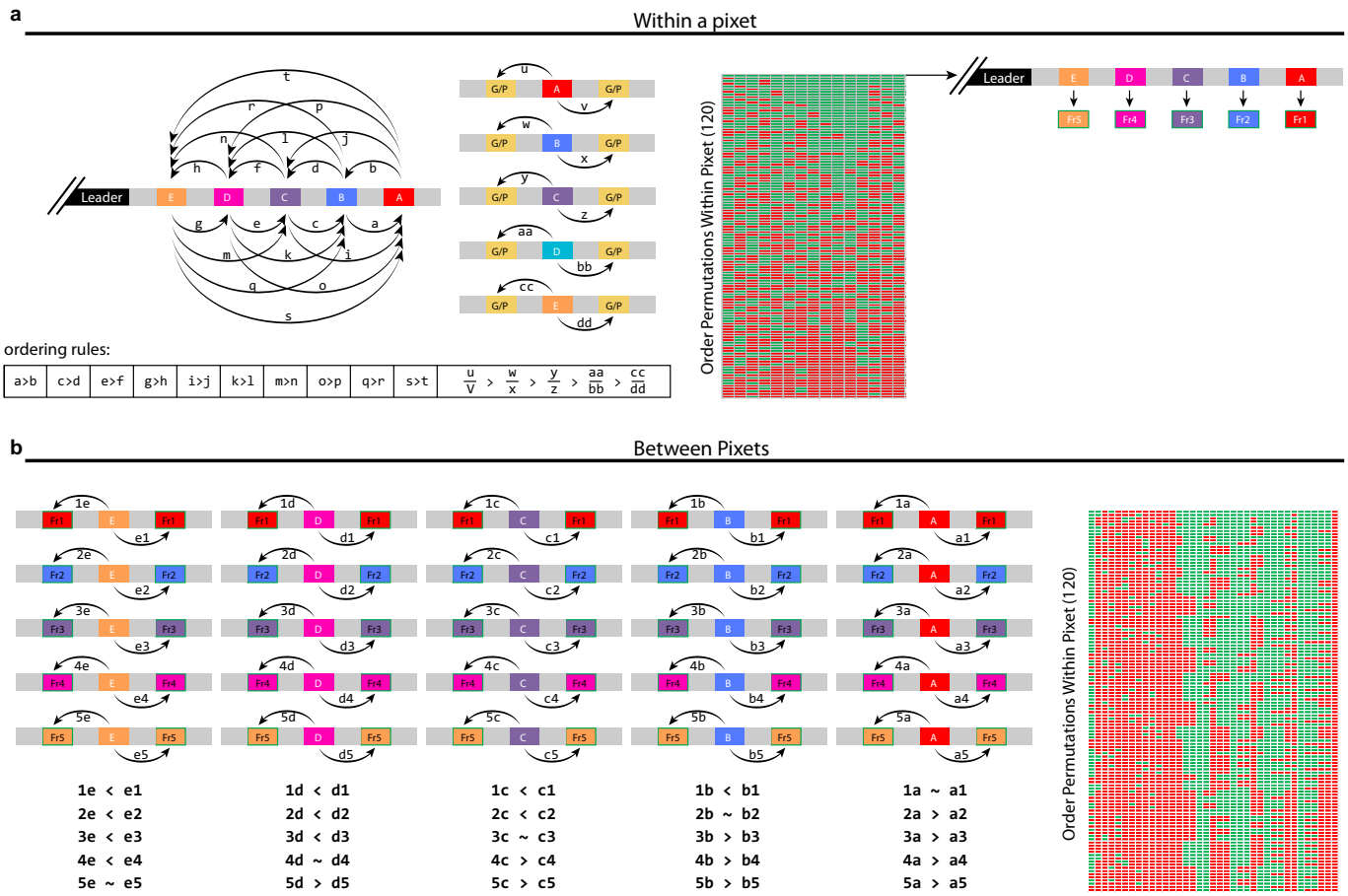
Extended Data Figure 3 | Cell surviving electroporation. Colony-forming units per millilitre of starting culture before beginning electroporation, after pre-electroporation washes, immediately post-electroporation, and after 1 h of recovery. Cells in red were electroporated with a minimal hairpin oligonucleotide, those in blue were electroporated in water alone. Unfilled circles represent individual biological replicates ($n = 3$), filled circles are mean ± s.e.m.



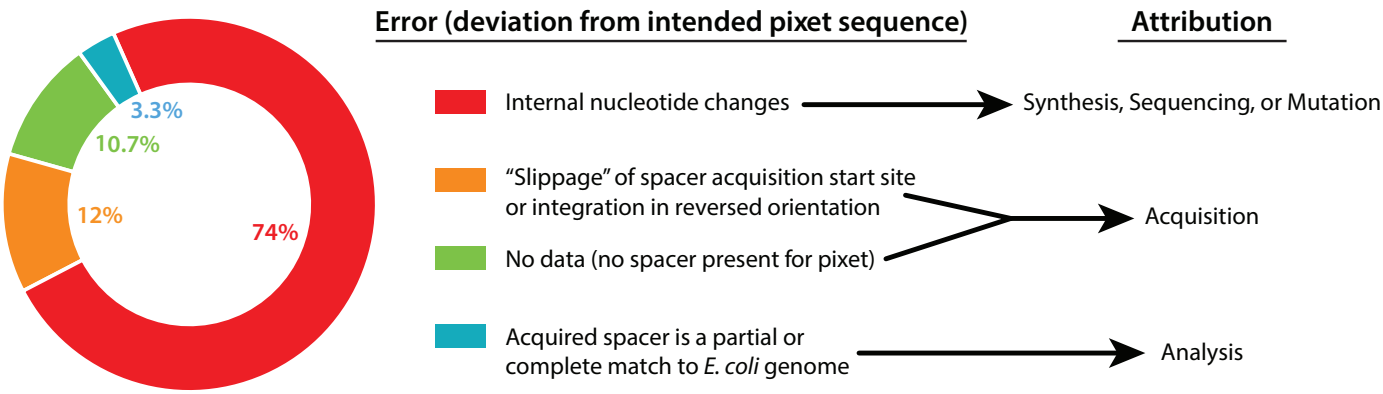
Extended Data Figure 4 | Optimization of protospacer sequence parameters. a, Comparison of the percentage of arrays that were expanded after encoding hand^R and hand^F images ($n = 3$). **b**, Percentage of arrays expanded per oligonucleotide (single pool) or per subpool (subpooled) across a range of GC percentages. Unfilled black circles to the left represent individual oligonucleotide protospacer sequences (three biological replicates each), while black line shows mean ± s.e.m. Unfilled red circles to the right represent individual biological replicates. Bars are mean ± s.e.m ($n = 3$; one-way ANOVA on effect of GC percentage, single pool: $P < 0.0001$, subpooled $P = 0.0011$; follow-up testing with Tukey's multiple comparison (corrected), see Supplementary Table 2). **c**, Percentage of arrays expanded per oligonucleotide electroporated individually across a range of GC percentages. Unfilled red circles are individual biological replicates. Bars show mean ± s.e.m ($n = 3$; one-way ANOVA on effect of GC percentage: $P = 0.0001$; follow-up testing with Tukey's multiple comparison (corrected), see Supplementary Table 2). **d**, Gibbs free energy of minimal hairpin protospacers structures for each of the images, with protospacers ranked by overall acquisition frequency ($n = 3$; linear regression, hand^R: $P = 0.0089$, hand^F: $P = 0.0004$). **e**, Percentage of arrays expanded per oligonucleotide (single pool) or per subpool (subpooled) with different numbers of mononucleotide repeats ($n = 3$; one-way ANOVA on effect of mononucleotide repeats, single pool: $P = 0.3843$, subpooled: $P = 0.0015$; follow-up testing with Tukey's multiple comparison (corrected), see Supplementary Table 2). Panel attributes as in **b**. **f**, Percentage of arrays expanded per oligonucleotide (single pool) or per subpool (subpooled) with different numbers of internal PAMs ($n = 3$; one-way ANOVA on effect of internal PAMs, single pool: $P = 0.0565$, subpooled: $P = 0.0052$; follow-up testing with Tukey's multiple comparison (corrected), see Supplementary Table 2). Panel attributes as in **b**. $^*P < 0.05$. Additional statistical details in Supplementary Table 2.



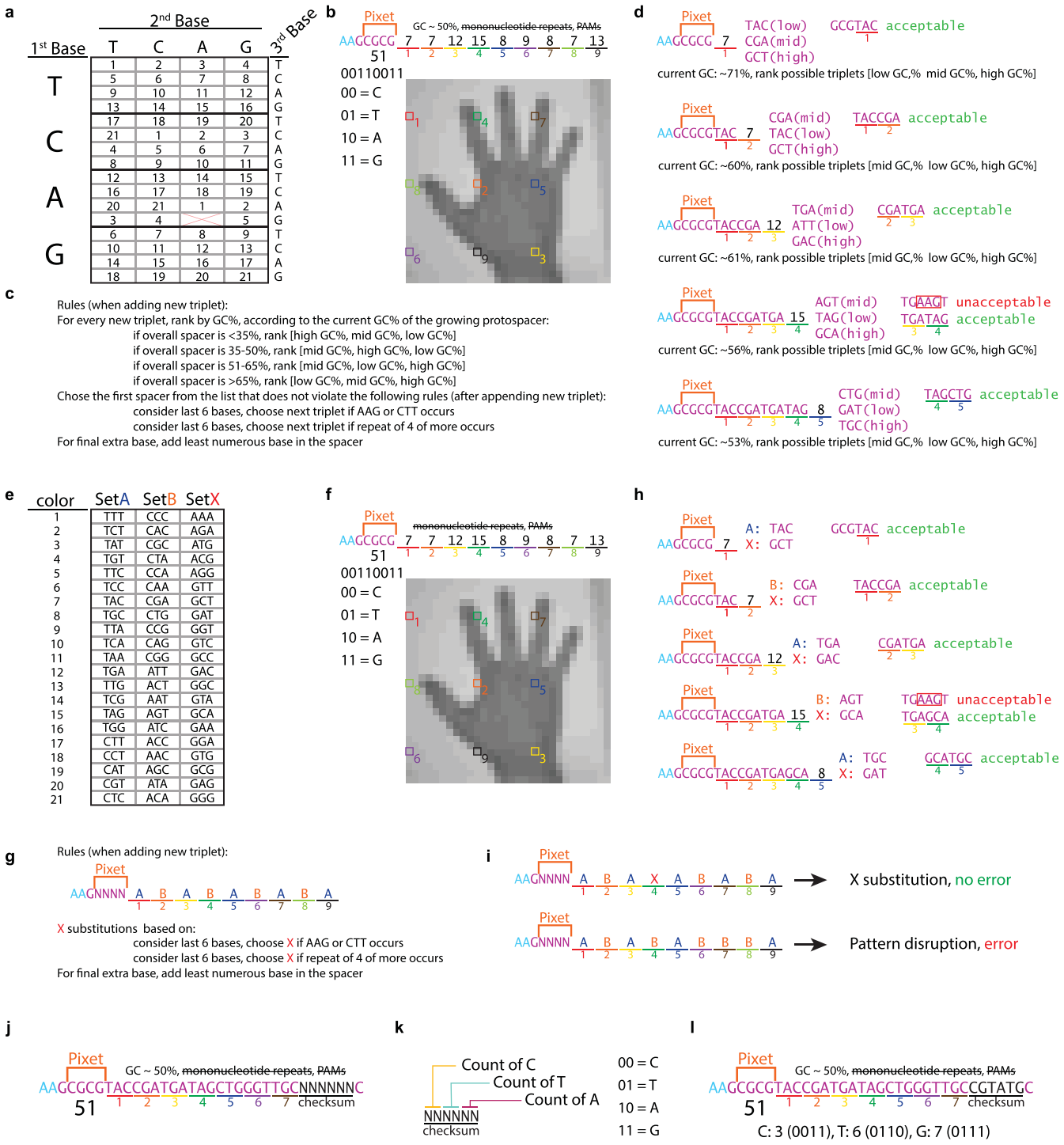
Extended Data Figure 5 | Effect of the 3′ motif on protospacer acquisition when supplied as two complementary oligonucleotides. Individual sequences designed to directly test the motif identified in Fig. 2b shown to the left. To the right, percentage of arrays expanded following electroporation of the sequences indicated as two complementary oligonucleotides (in dark red), rather than a minimal oligonucleotide hairpin (shown for comparison in pink). Unfilled circles indicate individual biological replicates. Bars show mean ± s.e.m. ($n = 3$; one-way ANOVA on effect of oligonucleotide: $P = 0.0041$; follow-up testing with Sidak's multiple comparison (corrected), seq^{over} versus seq^{over}-CCT: $P = 0.0103$, seq^{under} versus seq^{under}-TGA: $P = 0.0081$). *$P < 0.05$. Additional statistical details in Supplementary Table 2.



Extended Data Figure 6 | Recall of frame order over time based on position in the CRISPR array. a, Initial set of rules to test the order of spacers within a pixet. Every time two spacers from the same pixet are found in a single array, their relative physical location (with respect to the leader) is extracted. As is the location of each spacer relative to spacers drawn from the genome or plasmid (G/P). The actual sequence of electroporated protospacers should occupy arrays in a predictable physical arrangement, as described by these ordering rules. Every possible permutation of spacers within a pixet is tested against each of these rules and, if a permutation satisfies all the rules, spacers are assigned to frame. **b**, Second set of tests to compare between pixets. If no permutation satisfies all of the tests in **a**, spacers are compared to previously assigned spacers from other pixets pairwise when found in the same array. A larger set of rules will hold true for the actual sequence of electroporated protospacers when compared against previously assigned spacers. Again, all possible order permutations are tested, and order is assigned based on the best overall satisfaction of these ordering rules.



Extended Data Figure 7 | Quantification of errors by source. Includes any instance of a called spacer that does not match the supplied protospacer.



Extended Data Figure 8 | Methods of image encoding for error-correction. a–d, Method used in Fig. 1. **a**, Triplet code to flexibly specify 21 colours. **b**, Example of a pixet to be encoded into nucleotide space with pixel values marked. **c**, Rules specifying how the protospacer will be built. **d**, Example of the build of the protospacer. The AAG introduced by the addition of pixel 4 is unacceptable and invokes the flexible switch to another triplet. In a test of the extendibility of this encoding scheme, we ran three random sets of 100 million different nine-colour orderings through the sequence build and found that 99.86 ± 0.07% of colour orders were able to satisfy the requirements we set out without optimization by hand. **e–i**, Method of alternating clusters for error correction. **e**, Triplet assignment to clusters A, B, and X. **f**, Example of a pixet to be encoded into nucleotide space with pixel values marked. **g**, Rules for adding new triplets in this scheme. **h**, Example of the build of the protospacer. The AAG introduced by the addition of pixel 4 is unacceptable and invokes the flexible switch to cluster X. **i**, Example of an error signal. **j–l**, Method of checksum error correction. **j**, Annotation of protospacer with the addition of a checksum. **k**, Annotation of the checksum itself. **l**, Full protospacer with checksum implemented.