Published in final edited form as:

Nat Nanotechnol. 2018 April ; 13(4): 309–315. doi:10.1038/s41565-017-0004-z.

Genetic programs can be compressed and autonomously decompressed in live cells

Nicolas Lapique¹ and Yaakov Benenson¹

¹Department of Biosystems Science and Engineering, ETH Zurich, Basel 4058, Mattenstrasse 26, Switzerland

Abstract

Fundamental computer science concepts inspired novel information-processing molecular systems in test tubes^{1–13} and genetically-encoded circuits in live cells^{14–21}. Recent research showed that digital information storage in DNA, implemented using deep sequencing and conventional software, can approach the maximum Shannon information capacity²² of 2 bits per nucleotide²³. DNA is used in nature to store genetic programs, but the information content of natural encoding rarely approaches this maximum²⁴. We hypothesize that the biological function of a genetic program can be preserved while reducing the length and increasing the information content of its DNA encoding. Here we support this hypothesis by describing an experimental procedure for compressing a genetic program and its subsequent autonomous decompression and execution in human cells. As a test-bed we choose an RNAi cell classifier circuit²⁵ that comprises redundant DNA sequence and is therefore amenable for compression, as are many other complex gene circuits^{15, 18, 26–28}. In one example, we implement a compressed encoding of a ten-gene four-input AND gate circuit using only four genetic constructs. The compression principles applied to gene circuits can enable fitting complex genetic programs into DNA delivery vehicles with limited cargo capacity, and storing compressed and biologically inert programs *in vivo* for on-demand activation.

The RNAi classifier²⁵ comprises multiple sensor modules that relay the concentration of their cognate miRNA inputs to the computing module, which further integrates the

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 Reprints and permission information is available online at www.nature.com/reprints.

Correspondence to: Yaakov Benenson.

Author Contributions

N.L. conceived research, performed experiments, analysed data, and wrote the paper. Y.B. conceived research, analysed data, supervised the project, and wrote the paper.

Competing financial interests

The original miRNA circuit technology is protected by patents awarded to Y. Benenson and coinventors. The output delay technology is pending, with N. Lapique and Y. Benenson listed as coinventors.

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

Data Availability Statement

The data that support the plots within this paper and other findings of this study are available from the corresponding author upon reasonable request.

information into a single readout (Fig. 1a). The miRNA input targets its sensor to block gene expression from two gene cassettes, the first driving constitutive expression of reverse tetracycline-dependent transactivator (rtTA) and the second containing rtTA-induced, TRE-controlled expression cassette driving a Lac repressor protein (LacI) and intronically-encoded artificial miRNA miR-FF4. In the absence of a miRNA input, rtTA-induced LacI and miR-FF4 components repress the output via Lac operators in the output-driving promoter and miR-FF4 target sites in the output 3'-UTR, respectively, resulting in the *off* output state. In the presence of an input, the expression of rtTA, LacI, and miR-FF4 is greatly reduced, relieving the repression and enabling high output expression (output *on* state) (Fig. 1a). When multiple modules converge on the same output, the output is generated only when the repressor species originating from different modules are blocked simultaneously by their cognate inputs, leading to an AND-like logic between the inputs. Most of the circuit's DNA program encodes these sensor modules that only differ from each other in the sequence of an input-specific miRNA binding site in the 3'-UTR. Accordingly, the DNA sequence of different sensors module is ~95% identical (Fig. 1a). Initially, we considered a two-input miRNA classifier implementing the logic AND gate "miR-21 AND miR-146a". A variant of this circuit based on a published approach²⁵, called the "source" circuit, contains five genes. We developed a compression process to eliminate sequence redundancy, as follows. Consider two different constructs belonging to two different sensor modules that share the coding sequence (*e.g.*, rtTA species) but differ in their miRNA targets. These two constructs are compressed into a single construct that contains the first target in an active orientation (*i.e.*, responding to the first miRNA input) and the second target in an inactive inverted orientation (*i.e.*, insensitive to the second miRNA input). The region containing the target sequences is flanked by a pair of face-to-face recognition sites for a site-specific recombinase, in our case, LoxP sites. The decompression takes place when the iCre recombinase recognizes the cognate LoxP sites and reversibly inverts the intervening sequence. In a single act of inversion, the second target becomes active and the first one inactive (Fig. 1b). Because miRNA sensing in the circuit operates at the mRNA and protein levels, reversible recombination at the DNA level will lead to equal amounts of the species that comprise, respectively, the sensor modules for the first and the second input (Fig. 1c). After full decompression, there is 97% identity at the genetic level between the "source" and the decompressed variants.

In the *in situ* decompression experiments, the compressed circuit's DNA encoding and the gene encoding constitutively-expressed iCre recombinase are transfected into cultured mammalian cells. The decompression and subsequent deployment of the functional circuit therefore takes place autonomously in a living cell (these experiments are labelled "compressed circuit" below and in the figures). In order to assess the decompression efficiency, compressed circuit's performance in the cell classification task is compared to the control circuit's performance. The latter is composed of a pre-made mixture of genetic constructs emulating the outcome of perfect decompression (Fig. 2a); its performance is likewise measured in the presence of iCre. Each circuit is evaluated separately in a cell classification task using a metric based on ROC (Receiver Operating Characteristic) curve analysis²⁹ (false positive rate vs. true positive rate) (Material and Methods "ROC curve analysis" and Supplementary Fig. 1). There is one input combination that should be

classified positively ($\text{miR-21}^{\text{POS}}\text{miR-146a}^{\text{POS}}$) and three input combinations to be classified negatively ($\text{miR-21}^{\text{NEG}}\text{miR-146a}^{\text{NEG}}$, $\text{miR-21}^{\text{POS}}\text{miR-146a}^{\text{NEG}}$ and $\text{miR-21}^{\text{NEG}}\text{miR-146a}^{\text{POS}}$). While the circuit's false-negative rate is measured for the single *on* state, the false-positive rates differ for different *off* states. Accordingly, we calculate not one but three separate ROC curves corresponding to each of the *off* states and calculate their respective area under the curve (AuROC) values. The ratio of the control AuROC value to the compressed circuit's AuROC indicates how well the compressed variant performs in relation to the uncompressed control.

The initial variant of the compressed circuit “miR-21 AND miR-146a” and its control, together with the decompression “hardware” gene constitutively expressing iCre recombinase, were characterized in Human Embryonic Kidney 293 (HEK293) cells that are naturally negative for miR-21 and miR-146a (Supplementary Fig. 2a). Different input combinations were generated by cotransfecting corresponding miRNA mimics. Overall, the compressed circuit behaves qualitatively similar to the control (Supplementary Fig. 2bc), albeit with two deficiencies. First, the presence of LoxP sites and inverted targets in the 3'-UTR of the sensor module genes alters sensor sensitivity to the miRNA inputs (Supplementary Fig. 3a). Second, out of the three *off* states, the input combination $\text{miR-21}^{\text{POS}}\text{miR-146a}^{\text{NEG}}$ shows the largest discrepancy between compressed and control circuits (Supplementary Figs. 2bc and 3b). In the compressed circuit, the miR-146a sensor is initially absent, the production of repressor species by this sensor is delayed relative to the control, and the circuit implements a single-input YES gate “miR-21” rather than “miR-21 AND miR-146a” gate during the initial time window following plasmid delivery. Because the $\text{miR-21}^{\text{POS}}\text{miR-146a}^{\text{NEG}}$ input configuration triggers positive output in the “miR-21” YES gate, the false-positive rate of the compressed circuit is the highest in this case among all *off* states (Supplementary Fig. 3b). Both deficiencies, namely, reduced *on* state as well as high *off* state with $\text{miR-21}^{\text{POS}}\text{miR-146a}^{\text{NEG}}$ input configuration, were partially addressed by adding a 250 bp sequence containing a polyadenylation (polyA) signal between the targets in the 3'-UTR of the transactivator rtTA (Supplementary Figs. 4, 5, and Supplementary Text “Optimization of the two-input circuit”).

Following this optimization, the control and the compressed circuits still suffered from substantial leakage in the *off* states and high false-positive rate. As we showed previously, the leakage in the *off* states is caused by the time lag required to generate high levels of repressor species, LacI and miR-FF4³⁰. The compressed circuit needs even more time to deploy all its components, resulting in further increase in leakage (Supplementary Fig. 6). This lag-phase leakage can be reduced by delaying output availability³⁰ with the same iCre recombinase used to decompress the circuit. Together with the polyA feature, the output-delayed circuit achieved highly reliable decompression whose classification performance is identical to the control (Supplementary Figs. 7–11 and Supplementary Text “Optimization of the two-input circuit”). The iterative circuit optimization process is summarized in Fig. 2b.

In order to confirm that the compression and decompression operate according to the designed strategy, we performed extensive validation experiments (Supplementary Text “Mechanism validation experiments”). In these controls we show that (i) the mechanism of compression/decompression is symmetric (Supplementary Fig. 10); (ii) deletion of a single

LoxP site prevents the decompression from happening and the circuit permanently implements YES gate with miR-21 input (Supplementary Fig. 12); (iii) the inverted targets are not responding to their cognate miRNA inputs (Supplementary Fig. 13); and (iv) the compressed circuit can also implement “miR-21 AND miR-20a” logic (Supplementary Fig. 14), illustrating that the approach is programmable.

The circuits can be scaled by adding sensor modules in order to compute increasingly complex logic with miRNA inputs. We tested the optimized compression approach with a three-input logic AND gate between the inputs miR-21, miR-146a, and miR-20a. The circuit is compressed by augmenting the constructs with miR-21 target in the forward orientation followed by an inverted miR-146a (T21-T146a^{Rev}) or miR-20a target (T21-T20a^{Rev}) (Fig. 3a). The data show similar output levels between the compressed and the control circuits (Fig. 3a and Supplementary Fig. 15). This setup, where multiple sensors are compressed on more than one compressed sensor precursor, is called “split module” approach. The three-input circuit can be further compressed via appropriate configuration of heterospecific recombination sites such that only one construct generates all three sensor genes that share a coding frame (Fig. 3b and Supplementary Fig. 16a). We call this configuration a “single module” approach. To ensure that there is only one active miRNA target per transcript following decompression, a strong transcriptional terminator was added between the target sequences (Supplementary Fig. 16a). Optimization of the terminator sequence and of the recombination sites, as well as repositioning the intron expressing synthetic miR-FF4 from the 3'-UTR to the coding sequence of LacI, were required to improve the dynamic range of certain input combinations (Supplementary Figs. 16, 17 and Supplementary Text “Optimization of the three-input circuit”). Due to the time required to generate all the circuit variants *in situ*, the classification performance of the compressed circuit using a single module approach differs significantly from the control (Supplementary Fig. 18), which is not the case for the compression using a split module approach (Supplementary Fig. 15). Nevertheless, the single module compression shows a dynamic range above 6-fold for all inputs configurations (Fig. 3b).

Four-input AND gate was designed for miR-21, -20a, -141, and -146a inputs. Using the split module method we engineered compressed sensors with the initial configurations of T21-T146a^{Rev} and T20a-T141^{Rev}. All possible 16 input states were evaluated and found to behave consistently with the truth table and in a digital manner, be it with a compressed or the control circuit (Fig. 4a and Supplementary Fig. 19). Similarly to the three-input circuit, the four-input counterpart can be compressed using a single module method by an appropriate implementation of three pairs of heterospecific recombination sites and a terminator sequence (Fig. 4b, Supplementary Figs. 20, 21 and Supplementary Text “Optimisation of the four-input circuit”). The results obtained with the single sensor module show the expected trend (Fig. 4b and Supplementary Fig. 22); a large improvement can be expected with more efficient recombination sites.

The compression and decompression approach described here can be generalized to cases where a circuit comprises genetic constructs that contain identical parts coupled to variable components (Supplementary Fig. 23 and Supplementary Text “Generalization of the compression and decompression procedure”). The mechanism depends on the availability of

multiple heterospecific recombinase recognition sites that flank the variable regions. The method can be applied to components that comprise identical promoter and coding sequences but have variable 3'-UTR (Supplementary Fig. 23a), as is the case with our circuits, but also to components that use identical promoters to drive different coding sequences (Supplementary Fig. 23b), or constructs where the same coding sequence is controlled by different promoters (Supplementary Fig. 23c). We illustrate possible application of the general compression strategy to various published circuits from our and other groups implementing diverse logic (Supplementary Figs. 24 and 25). We note that for our particular approach, the number of recombination steps it takes to reach a given decompressed species doubles when the number of compressed components is increased by 2, although for realistic numbers of 4 to 6 the median number of steps is moderate (Supplementary Fig. 26). Moreover, this number drops quickly when multiple identical compressed constructs are present, each executing its own stochastic decompression (Supplementary Fig. 26g). This can happen when multiple delivery vectors enter the same cell, a common occurrence. It could also be engineered intentionally if the delivered construct is designed in a viral-like fashion to replicate in a cell or when it is embedded in a naturally-replicating delivery vector, such as an oncolytic virus.

The compression ratios that can be obtained in the general case depend on multiple parameters (Supplementary Fig. 27), reaching around 4-fold in the best cases. In general, the higher is the number of compressed constructs, the higher is the compression ratio; it also grows with the increase in the length of the repeated sequence. The recombinase adds a fixed cost to the compressed version that in general decreases the ratio but still makes it favourable in most cases; in a specific case when the recombinase is already a part of the source circuit, as is the case in our logic gates with delayed outputs³⁰, the recombinase cost is added to both the source and the compressed version and thus the compression is always beneficial.

With the current circuit architecture and compression approach, substantial size reduction can be achieved without loss of classification performance, while above a certain degree of compression the circuit performance is impaired (Supplementary Fig. 28). Greater compression ratios are theoretically possible, as shown by the results obtained with *in silico* algorithms (Supplementary Fig. 28). Recombinases allow complex series of permutations^{17, 31, 32}, and future research might result in compression and decompression methods with improved scalability and faster decompression compared to the method reported here. In addition, recombinases cannot be employed to compress certain type of sequence repetition such as tandem repeats. Shannon's landmark publication²² demonstrated that for any symbol sequence, there exists an encoding that can remove all redundancy without loss of information. Additional DNA manipulation technologies could play a decisive role in constructing compact genetic encoding that approaches the maximum Shannon information capacity of a DNA sequence.

Methods

Synthetic microRNA mimics and LNA

miRNA mimics were purchased from Dharmacon RNAi Technology (Thermo Scientific). miR-21 mimic (Cat No C-300492-03-0005) is a dsRNA that mimics the function of the human miRNA-21 (miRBase entry MI0000077). miR-146a mimic (Cat No C-300630-03-0005) is a dsRNA that mimics the function of the human miRNA-146a (miRBase entry MI0000477). miR-20a mimic (Cat No C-300491-03-0002) is a dsRNA that mimics the function of the human miRNA-20a (miRBase entry MI0000076). miR-20a mimic (Cat No C-300491-03-0002) is a dsRNA that mimics the function of the human miRNA-20a (miRBase entry MI0000076). miR-141 mimic (Cat No C-300608-03-0005) is a dsRNA that mimics the function of the human miRNA-141 (miRBase entry MI0000457). Negative Control mimic (Cat No CN-001000-01-05) is based on a mix of *C. Elegans* miRNA sequences. LNA microRNA power inhibitor of miR-20a (Cat No 426943-00) was purchased from Exiqon.

Cell culture and transfection

HEK293 (293-H) cell line was purchased from Invitrogen (Cat No 11631-017) and authenticated by the supplier. HEK293 cells were cultured in DMEM medium (Life Technologies GIBCO) supplemented with 10% FBS (Sigma), 0.045 g/mL of penicillin and 0.045 g/mL streptomycin at 37 °C, 100% humidity and 5% CO₂. Mycoplasma tests were routinely performed using PCR-based assay and found to be negative. Lipofectamine 2000 transfection reagent (Life Technologies) was used in HEK293 experiments. 6.5×10^4 HEK293 cells were seeded in 0.5 mL DMEM medium into each well of 24-well uncoated plastic plate (Thermo Scientific Nunc) and grown for ~24 h. 1.5 µL of Lipofectamine 2000 were added to each sample as described in the manual using Optimem (Life Technologies) to resuspend DNA (50 µL/sample) and incubation reagent (50 µL/sample). DNA plasmids were mixed according to Supplementary Tables 1–26. Growth medium was replaced before transfection with identical medium formulation containing doxycycline (Fluka) at a final concentration of 1 µg/mL. Transfected cells were incubated for 3 to 4 days before flow cytometry characterization. All reported data are arithmetic means of three to five biological replicas. The error bars represent $\pm$ one standard deviation. All key experiments were reliably reproduced at least two-three times.

Flow cytometry measurements

All samples were measured ~72 hours after transfection (4 days for three- and four-input logic circuits' characterization) with Fortessa flow analyzer (BD Biosciences). DsRed was measured using 561 nm laser, a 600 nm long pass filter and a 610/20 emission filter with a PMT voltage of 310 V. AmCyan was measured using 445 nm laser and 473/10 emission filter with a PMT voltage of 300 V. At least 100000 events were collected per sample and live cell gating was done using forward and side scatter channels.

Analysis of flow cytometry measurements

Scatter plots and bar charts in all the main and supplementary figures were generated as follows. Gating for and determination of DsRed-positive cells frequency (%DsRed⁺) was done using AmCyan single color transfection with 99.9% DsRed⁺ cells outside the gate. Gating for and determination of AmCyan-positive cells frequency (%AmCyan⁺) was done using DsRed single color transfection with 99.9% AmCyan⁺ cells outside the gate. For each sample, we measured the frequency of DsRed-positive cells, the arithmetic mean DsRed value in DsRed⁺ cells, mean(DsRed⁺), the frequency of AmCyan positive cells and the arithmetic mean AmCyan value in AmCyan⁺ cells, mean(AmCyan⁺). The average signal per transfected cell, denoted as **DsRed/Cell, abs. u.** in the charts, was calculated as:

$$\frac{\text{DsRed}}{\text{Cell}}, \text{abs. u.} = \frac{\text{mean(DsRed}^+) \times (\% \text{ DsRed}^+)}{\% \text{ AmCyan}^+} \quad (1)$$

When the analysis required the comparison of datasets with very different transfection efficiency, the normalization is taking into account the mean expression value of the transfection control and was denoted as **DsRed/Cell, rel. u.** in the charts.

$$\frac{\text{DsRed}}{\text{Cell}}, \text{rel. u.} = \frac{1000 \times \text{mean(DsRed}^+) \times (\% \text{ DsRed}^+)}{\text{mean(AmCyan}^+) \times (\% \text{ AmCyan}^+)} \quad (2)$$

Dynamic range is not sensitive to the quantification method (i.e. DsRed/Cell, abs. u. or DsRed/Cell, rel. u.)³⁰.

Error propagation was used to estimate the standard error of the dynamic range estimate and calculated as:

$$\left\{ \sqrt{\left(\frac{SD(\text{dataset 1})}{\text{mean}(\text{dataset 1})} \right)^2 + \left(\frac{SD(\text{dataset 2})}{\text{mean}(\text{dataset 2})} \right)^2} \right\} * \frac{\text{mean}(\text{dataset 1})}{\text{mean}(\text{dataset 2})} \quad (3)$$

SD denotes standard deviation.

ROC curve analysis

In this study, ROC curves are calculated to evaluate classification performance between the *on* state *versus* each one of the *off* states separately. Further, each of these comparisons is performed for different multiplicities of transfection (*i.e.*, the copy number of circuit plasmids delivered to cell nucleus) judged from the intensity of the transfection control. Because the measurements are performed in triplicate, we calculate nine ROC curves comparing every measurement in an *on* state with every measurement in a given *off* state. The following process is also illustrated in Supplementary Fig. 1. Multiplicity of transfection is judged indirectly through intensity percentile of the transfection marker AmCyan. Depending on the transfection percentile we are interested in (*e.g.*, top 10%, ... 100%) we

create a dataset filtered according to the transfection marker readout AmCyan. This percentile is denoted P_{Cyan} and 100% correspond to all transfected cells (not all cells) in the dataset. Then, single-cell flow cytometry data are ranked by their output (DsRed) channel value in descending order. Afterwards, we iterate through the data and for each cell in the *on* state define a threshold as being equal to the DsRed value of this cell. For every threshold (T_i), true positive rate (sensitivity) is calculated from the total number of cells in the *on* state and false positive rate (specificity⁻¹) is calculated from the total number of cells in the *off* state. The analysis is run with a Perl script available on request. The formulas below summarize the process.

For circuit measurement in an *on* input state, given a transfection percentile cutoff P_{Cyan} :

$$\text{True positive rate of } T_i(P_{Cyan}) = \frac{\text{number of cells with } AmCyan > P_{Cyan} \text{ and } DsRed \geq T_i}{\text{number of cells with } AmCyan > P_{Cyan}} \quad (4)$$

For circuit measurement in a particular *off* input state Off_i , given a transfection percentile cutoff P_{Cyan} :

$$\text{False positive rate of } T_i(P_{Cyan}, Off_i) = \frac{\text{number of cells with } AmCyan > P_{Cyan} \text{ and } DsRed \geq T_i}{\text{number of cells with } AmCyan > P_{Cyan}} \quad (5)$$

All false positive vs. true positive values are plotted on a graph and the area under the curve is calculated by summation of trapezes. The area under the curve is reduced by 0.5 to give a value that indicates classifier advantage over a random classifier; we term this difference “advantage value”. Finally, we compute the ratio between advantage values of a control circuit and a compressed circuit that is decompressed in cells with iCre. The ratio of the control AuROC value to the compressed circuit’s AuROC indicates which variant has higher classification power: a value above 1 means that the control outperforms the compressed circuit, a value of 1 indicates similar performance, and a value below 1 shows that the compression improves performance relative to the control. The ratio is calculated for different cell subpopulations to evaluate performance dependency on the multiplicity of transfection.

Fluorescent microscopy

Fluorescent images were acquired before flow cytometry measurements by an inverted Fluorescent Microscope (Nikon Eclipse Ti) using a Fiber Illuminator (Nikon Intensilight C-HGFI), optimized optical filtersets (Semrock) and a Digital Camera System (Hamamatsu ORCA R2). DsRed was measured with the filter combination TxRed HC (HC 624/40, HC 562/40, BS 593) with an exposure time of 2 s. AmCyan was measured with the filter combination CFP HC (HC 438/24, HC 483/32, BS 458) with an exposure time of 200 ms. For all experiments a Plan Fluor x10 Ph1 DLL objective was used and 2×2 or 3×3 frames were acquired. The microscopy images were contrast-enhanced for better visualization by ImageJ software using LUT image intensity. In AmCyan channel LUT values were set to 2300–7000 in Fig. 3a and Supplementary Fig. 19 and to 6000–50000 in Fig. 3b and

Supplementary Fig. 22. In DsRed channel, LUT values were set to 2300–4000 in Fig. 3a and Supplementary Fig. 19 and set to 8000–20000 in Fig. 3b and Supplementary Fig. 22.

Microscopy time course

One image (2×2 frames) was acquired for each well every 2 h with the same settings as described in Fluorescent microscopy section. The images were processed with MATLAB as described³³. Briefly a background correction was applied for each pixel using a baseline correction of signal with peaks (MATLAB's `msbackadj` function). Average pixel signal of DsRed output was normalized for transfection efficiency using the average pixel signal of AmCyan in Supplementary Fig. 6b.

In vitro recombination rate assay

500 ng of plasmids CMV-rtTA-LoxP-T21-SV40_PolyA-T146a^{Rev}-LoxP-SV40_PolyA (pNL228) or CMV-rtTA-LoxP-T21-T146a^{Rev}-LoxP-SV40_PolyA (pNL306) were incubated for 45 min or 120 min with 1.5 units of iCre (NEB) in a total volume of 30 µl, followed by heat inactivation at 70°C for 20 minutes. 10 µl of each solution was transformed into chemically competent *E. coli* DH5a cells and spread on kanamycin agar plates. Single colonies were picked and incubated overnight in 5 ml of LB with Kanamycin followed by plasmid purification (Sigma). The recombined clones were identified by digestion and/or sequencing.

Calculation of compression ratio and performance recovery

The calculation of the compression ratio achieved with our biomolecular compression method is described in the Supplementary Text “Generalization of the compression and decompression procedure”. The *in silico* compression ratio was calculated using Lempel-Ziv 77 (LZ77) algorithm, which is best adapted to compress genomic sequences²⁴. LZ77 compression algorithm was written in Perl (script available on request). The four nucleotide types that compose the genetic programs were converted to four UTF-8 symbols (1 symbol = 1 octet). To calculate the ratio, the size in octets of the source circuit is divided by the size in octets of the circuit compressed with LZ77. A minimal decompression program of 142 octets was written in Perl (script available on request).

The performance recovery of the compressed genetic circuits was calculated from the average of the advantage values (see ROC curve analysis section) of all the input combinations. The average advantage value of the compressed circuit was then divided by the average advantage value of the control circuit.

Simulations

The data for Supplementary Figs. 26 and 27 was obtained with Matlab (Mathworks) scripts (available upon request). For simulating state transitions, the initial configuration was encoded as a sequence of labels corresponding to structure [1] in Supplementary Fig. 23a. The time sequence of sequence states was simulated by random choice of a pair of engaged recombinase sites (from 1 to N-1 for N miRNA target sites). Then, the sequence of labels was updated to reflect the inversion between the chosen site pair. The sequence was repeated up to 1000 times for N=10 and N=8, and less for smaller N's. This process was itself

repeated 10,000 times. For each 1000-state sequence, we searched for the first appearance of targets ($T_2 \dots T_N$) in an active position. This was termed “wait time”. The histograms of wait times for a given target obtained from 10,000 simulations are shown in Supplementary Fig. 26 together with the plots of median wait times for each target. To calculate the number of steps to reach a given configuration for a number of decompression processes running in parallel (Supplementary Fig. 26g), the 10,000 simulated sequences of state transitions were randomly sampled for a required number of simulated sequences (from 2 to 10 in the Supplementary Fig. 26g), with each sequence representing one construct. For each target, we chose among these sequences the one with the shortest time to reach this target. We repeated this 10,000 times and calculated the median shortest times, which were then presented in the figure. For heat maps in Supplementary Fig. 27, equations 8 and 9 in Supplementary Text “Generalization of the compression and decompression procedure” were encoded in Matlab scripts and the ratios were calculated for the conditions described in the Supplementary Text and figure legend.

Recombinant DNA

The plasmids were cloned using standard Restriction-ligation and Gibson assembly. The detailed protocols for each construct are in the Supplementary Information, section **Plasmid construction**.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The research was funded by the National Institutes of Health award 5R01CA155320 and by ETH Zürich. We thank B. Angelici for discussions and E. Shapiro for commenting on the manuscript.

References

1. Adleman LM. Molecular computation of solutions to combinatorial problems. *Science*. 1994; 266:1021–1024. [PubMed: 7973651]
2. Mao CD, LaBean TH, Reif JH, Seeman NC. Logical computation using algorithmic self-assembly of DNA triple-crossover molecules. *Nature*. 2000; 407:493–496. [PubMed: 11028996]
3. Benenson Y, et al. Programmable and autonomous computing machine made of biomolecules. *Nature*. 2001; 414:430–434. [PubMed: 11719800]
4. Qian L, Winfree E, Bruck J. Neural network computation with DNA strand displacement cascades. *Nature*. 2011; 475:368–372. [PubMed: 21776082]
5. Liu Q, et al. DNA computing on surfaces. *Nature*. 2000; 403:175–179. [PubMed: 10646598]
6. Rothmund PWK. Folding DNA to create nanoscale shapes and patterns. *Nature*. 2006; 440:297–302. [PubMed: 16541064]
7. Phillips A, Cardelli L. A programming language for composable DNA circuits. *J Royal Soc Interface*. 2009; 6:S419–S436.
8. Gehani A, LaBean T, Reif J. DNA-based cryptography. *Lecture Notes in Comp Sci*. 2003; 2950:167–188.
9. Chen YJ, et al. Programmable chemical controllers made from DNA. *Nat Nanotech*. 2013; 8:755–762.

10. Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA. *Science*. 2012; 337:1628. [PubMed: 22903519]
11. Padirac A, Fujii T, Rondelez Y. Bottom-up construction of in vitro switchable memories. *Proc Natl Acad Sci USA*. 2012; 109:E3212–3220. [PubMed: 23112180]
12. Weitz M, et al. Diversity in the dynamical behaviour of a compartmentalized programmable biochemical oscillator. *Nat Chem*. 2014; 6:295–302. [PubMed: 24651195]
13. Stojanovic MN, Stefanovic D. A deoxyribozyme-based molecular automaton. *Nat Biotech*. 2003; 21:1069–1074.
14. Gardner TS, Cantor CR, Collins JJ. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*. 2000; 403:339–342. [PubMed: 10659857]
15. Rinaudo K, et al. A universal rna-based logic evaluator that operates in mammalian cells. *Nat Biotech*. 2007; 25:795–801.
16. Regot S, et al. Distributed biological computation with multicellular engineered networks. *Nature*. 2011; 469:207–211. [PubMed: 21150900]
17. Bonnet J, Yin P, Ortiz ME, Subsoontorn P, Endy D. Amplifying genetic logic gates. *Science*. 2013; 340:599–603. [PubMed: 23539178]
18. Daniel R, Rubens JR, Sarpeshkar R, Lu TK. Synthetic analog computation in living cells. *Nature*. 2013; 497:619–623. [PubMed: 23676681]
19. Nielsen A, et al. Genetic circuit design automation. *Science*. 2016:352.
20. Green, Alexander A., Silver, Pamela A., Collins, James J., Yin, P. Toehold switches: De-novo-designed regulators of gene expression. *Cell*. 2014; 159:925–939. [PubMed: 25417166]
21. Anderson JC, Voigt CA, Arkin AP. Environmental signal integration by a modular and gate. *Mol Syst Biol*. 2007; 3:133. [PubMed: 17700541]
22. Shannon CE. A mathematical theory of communication. *The Bell System Technical Journal*. 1948; 27:379–423. 623–656.
23. Erlich Y, Zielinski D. DNA fountain enables a robust and efficient storage architecture. *Science*. 2017; 355:950–954. [PubMed: 28254941]
24. Kuruppu, S., Puglisi, S., Zobel, J. Optimized relative lempel-ziv compression of genomes. *Proc. 34th Australas. Comp. Sci. Conf*; 2011. p. 91-98.
25. Xie Z, Wroblewska L, Prochazka L, Weiss R, Benenson Y. Multi-input rna-based logic circuit for identification of specific cancer cells. *Science*. 2011; 333:1307–1311. [PubMed: 21885784]
26. Xie M, et al. B-cell-mimetic designer cells provide closed-loop glycemic control. *Science*. 2016; 354:1296–1301. [PubMed: 27940875]
27. Prochazka L, Angelici B, Haefliger B, Benenson Y. Highly modular bow-tie gene circuits with programmable dynamic behaviour. *Nat Commun*. 2014; 5:4729. [PubMed: 25311543]
28. Schreiber J, Arter M, Lapique N, Haefliger B, Benenson Y. Model-guided combinatorial optimization of complex synthetic gene networks. *Mol Syst Biol*. 2016; 12:899. [PubMed: 28031353]
29. Heagerty P, Lumley T, Pepe M. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*. 2000; 56:337–344. [PubMed: 10877287]
30. Lapique N, Benenson Y. Digital switching in a biosensor circuit via programmable timing of gene availability. *Nat Chem Biol*. 2014; 10:1020–1027. [PubMed: 25306443]
31. Bhatia S, LaBoda C, Yanez V, Haddock-Angelli T, Densmore D. Permutation machines. *ACS Syn Biol*. 2016; 5:827–834.
32. Ham TS, Lee SK, Keasling JD, Arkin AP. Design and construction of a double inversion recombination switch for heritable sequential genetic memory. *PLoS One*. 2008; 3:e2815. [PubMed: 18665232]
33. Haefliger B, Prochazka L, Angelici B, Benenson Y. Precision multidimensional assay for high-throughput microRNA drug discovery. *Nat Commun*. 2016; 7:10709. [PubMed: 26880188]

Figure 1. Mechanism of compression and decompression of a multi-input miRNA logic circuit
a, On the left, general schematics of a two-input circuit. miRNA input is depicted as a squiggly line and DNA molecules represent different genetic components of the circuit. X and Y are two arbitrary miRNA inputs. On the right, detailed mechanism of the Y-sensor. Pointed arrows indicate activation and blunted arrows denote repression. Rectangles targeted by miRNA represent four identical sites fully-complementary to miRNA inputs, and three identical sites complementary to the synthetic miR-FF4. **b**, Detailed schematics of the reversible inversion of a DNA fragment containing active and inactive miRNA target sites producing two different mRNA species. Different sequence moieties are indicated in the scheme. At the bottom, shorthand representations used in other figures are shown, with the box-like target sites in the top and the bottom strands representing, respectively, a target in

the active and inactive orientation. TX is a target for miR-X, and TX with a bar on top indicates that the target is inverted and inactive. The black triangles denote Cre-specific recombination sites and their orientation. **c**, Compressed circuit diagram (left) and the process of circuit decompression illustrated with a two-input logic AND gate. Blunt arrows between the miRNA input and the target sites indicate sensory interaction, while the lack of thereof (as with the input miR-Y) shows lack of interaction due to absence of properly-oriented sites. Small shaded insets represent the shorthand notation corresponding to the shaded part of the gene circuit. pTRE, TRE promoter. CMV, Early-late cytomegalovirus promoter. LacI, Lac repressor. rtTA, reverse tetracycline-responsive transcriptional activator. CAGop, CAG promoter with two LacO sites in the intron. miR-FF4, a synthetic microRNA spliced from LacI transcript.

Figure 2. Optimisation of the decompression process

a, The illustration of the experimental setup comparing the compressed and the control circuits. The diagrams represent time-dependent evolution of the different species, shown in the shorthand notation, for the two cases as indicated. TX is a target for miR-X, and TX with a bar on top indicates that the target is inverted and inactive. **b**, The comparison of decompression efficiency between four different circuit variants, represented by four heat maps. Each variant is characterized by a particular miRNA target arrangement (rows) and the delay in output availability (columns). Within each heat map, the rows correspond to the three different *off* states (shown in the captions on the left), and the columns to the percentile

of transfected cells (shown at the top), that were used to calculate the classification performance ratio (see main text) between the control and the compressed circuits. The ratio is colour coded, with the value of 1 (orange) indicating identical classification performance between the two circuit variants. A heat map with a uniform orange hue would indicate identical performance between the compressed and the control circuits across all conditions.

c, Experimental data showing classification performance of the best circuit variant. On the left, population-averaged output values for different input states, indicated below. The bars show the arithmetic mean of three to five biological replicates with the error bars showing ± 1 standard deviation. All the observed differences between the *on* and the *off* states in the compressed circuits are statistically significant with a p-value < 0.001 calculated using a 2-sided t-test. On the right, representative flow cytometry plots and transfection percentiles used for ROC analysis. Note that “compressed circuit” refers to a compressed circuit that is co-delivered with iCre recombinase and decompressed *in situ* in live cells. The results were consistently reproduced at least two-three times. The abbreviations used are identical to Figure 1.

Figure 3. Compression of three-input AND gate circuits

Logic computation using compressed “miR-21 AND miR-146a AND miR-20a” gate circuits. **a**, Compression using a split module approach. The diagram on the left illustrates the process of sensor decompression using shorthand sensor notations (See Fig. 1c for the full depiction of genetic constructs corresponding to a shorthand notation). TX is a target for miR-X, and TX with a bar on top indicates that the target is inverted and inactive. Genetically distinct sensors that nevertheless respond to the same input are placed on a shared shaded background; their cognate input (*e.g.*, miR-21) is indicated on top. The 3D bar chart on the right shows the normalized output levels achieved with different input combinations as indicated, in an experiment comparing control and compressed circuits. The fold-change is the worst-case on/off ratio. At the bottom, the micrographs show the

expression of AmCyan (transfection marker, green pseudocolor) and DsRed (circuit output, red pseudocolor) for all eight input combinations, as indicated. **b**, Compression using a single module approach. Left, the diagram showing the process of decompression using the shorthand sensor notations. Black and grey triangles correspond, respectively, to LoxP and Lox5171 sites. Similarly color-coded arrows show the consequence of LoxP or Lox5171 site engagement at the initial stage (top) and during the ensuing equilibrium (bottom). Genetically-distinct sensors that respond to the same input are boxed together on a shared grey background, with the cognate input name indicated. TX is a target for miR-X, and TX with a bar on top indicates that the target is inverted and inactive. The 3D bar chart on the right shows the normalized output levels achieved with different input combinations as indicated, in an experiment comparing control and compressed circuits. The fold-change is the worst-case *on:off* ratio. At the bottom, the micrographs show the expression of AmCyan (transfection marker, green pseudocolor) and DsRed (circuit output, red pseudocolor) for all eight input combinations, as indicated. Each bar shows the arithmetic mean of a biological triplicate with the error bars indicating 1 standard deviation. All the observed differences between the *on* and the *off* states in the compressed circuits are statistically significant with a p-value <0.001 calculated using a 2-sided t-test. The experiments were consistently reproduced at least two-three times. “STOP” octagons denote GAPDH polyA transcriptional terminator.

single module approach. The diagram shows the process of decompression using the shorthand sensor notations. Black, grey, and turquoise triangles correspond, respectively, to LoxP, Lox5171, and LoxFAS sites. Identically color-coded arrows show the result of these sites' engagement by the recombinase at the first stage of decompression, and later during the ensuing equilibrium. Note that the original compressed construct on top is regenerated among the products. Genetically-distinct sensors that respond to the same input are boxed together on a shared grey background, with the cognate input name indicated. The 3D bar chart on the right shows the normalized output levels achieved with different input combinations as indicated, in an experiment comparing control and compressed circuits. The fold-change is the worst-case *on:off* ratio. Each bar shows the arithmetic mean of a biological triplicate with the error bars showing 1 standard deviation. All the observed differences between the *on* and the *off* states in the compressed circuits are statistically significant with a p-value <0.001 calculated using a 2-sided t-test. The experiments were consistently reproduced at least two times. Red octagons denote GAPDH polyA transcriptional terminator.