Rapidly evolving homing CRISPR barcodes

Reza Kalhor[1], Prashant Mali[2], George M. Church[1]

[1]*Department of Genetics, Harvard Medical School, Boston, MA 02115.*
[2]*Department of Bioengineering, University of California San Diego, La Jolla, CA 92093.*
Correspondence should be addressed to G. M. C. (gchurch@genetics.med.harvard.edu) or to P. M. (pmali@ucsd.edu).

Keywords: DNA barcodes, lineage tracing, CRISPR, Cas9, homing CRISPR, homing guide RNA, brain mapping, fluorescent in situ sequencing, FISSEQ, genome engineering.

Abstract: We present here an approach for engineering evolving DNA barcodes in living cells. The methodology entails use of a homing guide RNA (hgRNA) scaffold that directs the Cas9-hgRNA complex to target the DNA locus of the hgRNA itself. We show this homing CRISPR-Cas9 system acts as an expressed evolving genetic barcode, and corresponding small RNAs can be assayed as single molecules *in situ*. This integrated approach will have wide ranging applications, such as in deep lineage tracing, cellular barcoding, molecular recording, dissecting cancer biology, and connectome mapping.

Main Text

A single totipotent zygote has the remarkable ability to generate an entire multicellular organism. Methodologies to comprehensively map and modulate the parameters that govern this transformation will have far ranging impact on our understanding of human development and our ability to restore normal function in damaged or diseased tissues. Precise lineage history of cells during development is one of the parameters that can shed important insights into developmental processes (Sulston et al. 1983; Kretzschmar and Watt 2012). Contemporary lineage-tracing approaches, however, do not scale readily to the model organisms, such as mice, that are most relevant to human development (Weisblat, Sawyer, and Stent 1978; Dymecki and Tomasiewicz 1998; Walsh and Cepko 1992; Porter et al. 2014; Lu et al. 2011). Precise mapping of lineage history in these organisms may be facilitated by combining modern genome engineering and DNA sequencing technologies (Mali, Esvelt, and Church 2013; Lee et al. 2014; Church, Marblestone, and Kalhor, 2015.): if every cell in an organism contained a

unique and easily retrievable DNA sequence - a barcode - that encompassed its lineage relationship with other cells, the precise lineage history of the cells that constitute the organism could be delineated by probing this barcode, a process analogous to deriving the tree of life by sequencing loci such as 16S ribosomal RNA in different organisms.

To this end, we propose here the concept of 'evolving' genetic barcodes that are embedded in cells and change their genetic signature progressively over time (Figure 1). In a generalizable manifestation, this approach entails an array of genomically integrated sites that are stochastically targeted by a nuclease. Each site bears a unique identifier and can be expressed through transcription. Before each round of cell division, the nuclease targets a random subset of these sites where the process of non-homologous end joining (NHEJ) introduces insertions, deletions, or other mutations. At each site, these mutations lead to a unique sequence that is related to its parent sequence and may further evolve in subsequent rounds. At the end of the process, single-cell assaying technologies can be applied to all sites in each cell to decipher its unique barcode - the array of all its evolving nuclease sites - and decipher its corresponding lineage-history.

Such an array of evolving barcodes can in theory fulfill the requirements of precise lineage tracing. However, it must create a total diversity commensurate with the total number of cells being targeted. If the diversity of possible mutations in each barcode element is 'm' and the number of array elements is 'n', then the system allows for creation of n^m possible signatures.

With this concept in mind, we utilized the CRISPR-Cas9 system to introduce mutations at a target locus via the process of non-homologous end joining (NHEJ), which simulates partially randomization of barcodes. In its original form, even limited mutagenesis disrupts the ability of the Cas9-gRNA complex to bind its target locus. Therefore, we sought to engineer a homing CRISPR system that directs Cas9-gRNA nuclease activity to the gRNA locus itself, thus enabling retargeting and evolvability of barcodes. A canonical CRISPR system (Figure 2A) involves a guide RNA (gRNA) that forms a complex with the Cas9 protein to target a DNA sequence that matches the gRNA for digestion. In general terms, the gRNA is comprised of two parts: a constant scaffold, which establishes the gRNA's interaction with the Cas9 protein through its primary and secondary structures, and a spacer, which determines the target DNA site and varies for different gRNAs. For a DNA locus to be targeted by a CRISPR system there are two requirements. First, its sequence has to complement the spacer sequence of the gRNA. Second, it has to contain a protospacer adjacent motif (PAM) at a specific position relative to the spacer sequence. The PAM is recognized by the Cas9 protein

and not the gRNA. As gRNAs do not have a PAM, loci that code for standard gRNAs are not targeted by the Cas9:gRNA complex (Figure 2A). We sought to create a homing gRNA that can target its own locus in addition to other target loci (Figure 2B). For that, in the *S. pyogenes CRISPR* gRNA, we mutated the sequence immediately downstream of the spacer sequence from 'GUU' to 'GGG' (Figure 2C), so it matches the requisite 'NGG' PAM sequence of *S. pyogenes* Cas9. These bases are a part of a helix in the secondary structure of the wild type gRNA (Figure 2D, left). To preserve the helix and minimize adverse structural impacts from our mutations, we further introduced compensatory mutations in the hybridizing nucleotides (Figure 2D, right). We then evaluated the functionality of this homing gRNA using a homologous recombination (HR) based reporter assay (Fig. 1E). In this assay, a 'broken' GFP gene is targeted by CRISPR-Cas9 in presence of a repair template(Mali, Yang, et al. 2013). Successful targeting of the broken GFP gene results in its repair through HR and the ensuing restoration of fluorescence can be detected. The results show that our homing gRNA is functional.

Next, we evaluated the targeting of the homing gRNA locus by its gRNA product. For this, we created a HEK/293T clonal cell line genomically integrated with the humanized *S. pyogenes* Cas9 under an inducible Tet-On promoter (293T-iCas9 cells). We introduced the homing gRNA locus into the genome of these cells using lentiviral integration. We then induced Cas9 expression and harvested DNA samples from the cell population at various intervals after induction. As control, we sampled non-induced batches of the same engineered cell line in parallel. Sequencing the homing gRNA locus in the induced cells clearly revealed accumulation of various mutations in its sequence through time, above background sequencing error (Figure 2F). No significant mutagenesis above background sequencing error was observed without inducing Cas9 expression (Figure 2F). These results show that our homing gRNA, or hgRNA, which is altered to contain a PAM sequence adjacent to its spacer sequence, mutates the DNA locus encoding itself.

We then created six additional hgRNAs (B21, C21, D21, E21, F21, and G21), each with a different spacer sequence (Figure 3A). We assayed these hgRNAs in 293T-iCas9 cells (Figure 3B). The results showed that five of the six hgRNAs are highly active in targeting their parent loci, with the sixth (hgRNA-B21) showing a much lower activity level. hgRNA-B21 has a spacer with multiple 'GG' dinucleotides, a feature which has previously been shown to reduce gRNA activity. These observations suggest that hgRNAs are generally functional irrespective of their spacer sequence.

When tracking the fraction of hgRNA loci that remain capable of encoding an active hgRNA after multiple rounds of targeting, we observed that they become quickly inactivated after induction of Cas9 expression (Figure 3C). While some were inactivated due to a deletion that removed the PAM or a part of the scaffold, most were rendered inactive by truncation of their spacer below the 16-18 nucleotides necessary for Cas9-gRNA cleavage. As our hgRNA set had only 21 total bases between the RNA transcription start site and the hgRNA scaffold, even small deletions in the spacer would lead to truncated hgRNAs. We therefore sought to evaluate whether hgRNAs' active lifespan can be prolonged by increasing their lengths. As such, based on hgRNA-A21, we created four more variants that were 25, 50, 75, and 100 bases longer than hgRNA-A21 but had the same initial spacer sequence (Figure 3E). These hgRNAs were all active in our standard assay, albeit at lower levels than hgRNA-A21 (Figure 3F), with activity levels dropping with increasing hgRNA length (Figure 3F). This reduced activity can be a combination of several factors, including lower expression of longer transcripts from the U6 promoter, lower stability or activity of the Cas9:hgRNA complex, or disruptive secondary structure in the RNA. Despite the reduced activity, the longer hgRNA loci are less likely to be inactivated as they mutate themselves (Figure 3G), and thus have extended life spans.

We next analyzed the suitability of hgRNAs for barcode generation in lineage tracing by estimating its diversity after inducing mutations. Two factors are important in considering this diversity: first, the number of different variants, and second, the frequency distribution of those variants. The ideal barcoding locus would generate a very high number of variants and all with equal likelihood. As a proxy for both these factors, we measured the Shannon entropy of the frequencies of all the variants generated by each hgRNA in different samples (Figures 3D and 3H). The results indicate that our six more active hgRNAs, A21, C21, D21, E21, F21, and G21, can generate at least 5 bits of diversity each in addition to the background diversity created by sequencing and other processes (see Materials and Methods). While longer hgRNAs should have higher ceilings for generating diversity, our observation times were not long enough to observe such an effect (Figure 3H). This data suggests that uniquely barcoding all neurons in a mouse brain requires at least 6 hgRNA loci per cell (see discussion).

Since easy retrievability is a requirement for effective lineage tracing using evolving barcodes, *in situ* sequencing is the most desirable method for barcode retrieval as it would allow lineage to be deciphered without losing histological information such as position and cell type (Ke et al. 2013; Lee et al. 2014). However, available fluorescent *in situ* sequencing (FISSEQ) technologies currently face challenges in targeted

sequencing of specific loci and have biases still to be resolved(Lee et al. 2014; Lee et al. 2015; Ke et al. 2013). We therefore assessed whether gRNAs that are small and expressed via polymerase III promoters can be probed in a targeted fashion using FISSEQ. FISSEQ can be generally divided into two stages: amplicon generation and amplicon sequencing. As it is the amplicon generation step that may not translate adequately between various transcript types and challenges targeted sequencing(Lee et al. 2014; Lee et al. 2015; Ke et al. 2013), we used an assay to specifically address this step (Figure 4A). We then executed the assay on representative gRNA constructs with reverse-transcription (RT) and rolling-circle amplification (RCA) primers for their targeted detection in FISSEQ. Amplicons of a generic gRNA construct with RT and RCA primers flanking the entire gRNA could not be successfully generated above background level (Figure 4B, middle). We surmised that possible reasons for failure were: 1) the strong secondary structure of the intervening gRNA, and 2) the long distance (128bp) between the RT and RCA primers. We thus created a second version of the gRNA construct by repositioning the RT primer further upstream inside the gRNA at a position previously shown to be tolerant of small insertions(Mali, Aach, et al. 2013). This new arrangement resulted in robust *in situ* amplification and detection of gRNAs in a target-specific fashion (Figure 4B, right). These results show that it is in fact possible to read out a barcoding gRNA locus using available *in situ* sequencing technologies. We anticipate that insertions of this primer binding site in the gRNA stem loops can be utilized in future designs to further enhance the functionality of this approach and also retain efficacy of gRNA function(Mali, Aach, et al. 2013)(Konermann et al. 2015).

Discussion

We describe a new homing CRISPR system that enables engineering of evolving barcodes (two other recent reports similarly demonstrate the utility of nuclease mediated diversity generation for barcoding (McKenna et al. 2016) and recording (Perli, Cui, and Lu 2016)). We evaluate several important parameters in the implementation of this system, including the effect of spacer sequence and gRNA length. We show that self-targeting is a general property of hgRNA loci and that both their life-span and mutation rate can be customized by adjusting their transcripts' lengths.

The strength of this system lies in its potential for informative mutations to accrue over cell divisions or in response to an external stimulus. Consequently, we also evaluate the suitability of this system for barcoding and lineage tracing applications. We show that a homing gRNA locus is capable of storing about 5 bits of information, which is enough to distinguish 2^5=32 different states. This measurement suggests that uniquely barcoding the roughly 12 billion cells in a mouse will require at least 7 such hgRNA loci per cell (

(2^5)^7 > 12E9). Uniquely barcoding the estimated 75 million neurons in a mouse brain will require at least 6 such hgRNA loci per neuron ((2^5)^6 > 75E6). These are minimum requirements; the actual number of barcoding loci needed in practice will depend on the tolerance of the experiment to duplicate barcodes. Nonetheless, they suggest that adequate barcoding of mouse neurons, which is essential for some of the proposed brain mapping projects in the brain initiative (A. H. Marblestone et al. 2013; Adam H. Marblestone et al. 2013; Church, Marblestone, and Kalhor, 2015.), may be within reach of our current strategy.

We assess the feasibility of reading out small gRNA-based barcodes using current in situ sequencing technologies. We find that an extended gRNA sequence presents limitations for in situ amplification, a prerequisite of in situ sequencing technologies, likely due to long amplicon length and RNA secondary structure. We assessed new versions of gRNA by inserting the primer-binding site into the scaffold and show that the amplification challenge may be addressed with modifying the hgRNA backbone to include effective amplification primers.

We note three key limitations in our current implementation: one, the limited duration of evolvability; two, the limited diversity generated by each element; and three, the difficulty to assay genotypes at the single-cell level. We hope that the use of improved and orthogonal inducible systems(Platt et al. 2014), coupled with arrays of evolving barcodes and enhanced imaging capabilities(Liu and Keller 2016), and also use of molecules that can modulate NHEJ outcomes (such as end processing enzymes, polymerases and terminal transferases(Certo et al. 2011)) will significantly enhance the scalability and broad applicability of this approach. The ability to sequence DNA barcodes *in situ* will also pave the way for ultra-dense cellular barcoding. Such advances can help leverage the remarkable diversities that longer DNA barcodes can generate. Taken together this approach will have broad applications in developmental biology, molecular recording(Farzadfard and Lu 2014), cancer biology, and in mapping neural connectivity(A. H. Marblestone et al. 2013).

Materials and Methods

Vector construction. The Cas9 vectors used in the study are based on earlier published work (Yang et al. 2013; Mali, Yang, et al. 2013). The hgRNA vectors were constructed by incorporating corresponding gBlock (IDT DNA) synthesized DNA fragments (spacers and scaffolds) into the pLKO.1 lentiviral backbone (MISSION shRNA vectors via SIGMA) which was modified to have Hygromycin B resistance.

HEK/293T cells with inducible *S. pyogenes* Cas9. Humanized *S. pyogenes* Cas9 (hCas9) under the control of a Tet-On 3G inducible promoter and carrying the puromycin resistance gene was genomically integrated in HEK/293T cells (ATCC CRL-11268) using a PiggyBac transposon system. Multiple clonal lines were derived from the transduced population and doxycycline-induced expression of hCas9 was measured in each line using reverse-transcription followed by quantitative PCR. The best line showed low baseline levels of hCas9 expression and about 300 fold enrichment of hCas9 upon induction (data not shown). This line was used in all ensuing experiments and will be referred to as 293T-iCas9. These cells were cultured on poly-D-lysine coated surfaces and in DMEM with 10% Fetal Bovine Serum and 1μg/ml Puromycin in all experiments.

Lentivirus production. Lentiviruses were packaged in HEK/293T cells using a second generation system with VSV.G as the envelope protein. Viral particles were purified using polyethylene glycol precipitation and resuspension in PBS. They were stored at -80C until use.

Transduction of 293T-iCas9-puro cells with lentiviral vectors carrying hgRNAs. 293T-iCas9 cells were grown to 70% confluency at which point they were infected with 0.3-0.5 MOI of lentiviral particles in presence of 6μg/ml polybrene. About 3 days after infection, cells were placed under selection with 200μg/ml Hygromycin B. Cells were retained under selection for at least 1 week before any experiments to assure genomic integration. In all cases, loss of about half the entire cell population after selection was used to confirm single infection with lentiviruses. Cells were maintained under Hygromycin B selection throughout subsequent experiments.

Induction of hCas9 in cells with hgRNAs. For each cell line transduced with a hgRNA construct, a sample was harvested before induction to represent the state of the non-induced population. Cells were then induced with 2μg/ml doxycycline. At various time points after induction, as indicated for each experiment, a sample of the cells was harvested during a passage to represent different time points after induction. In one experiment, multiple samples were harvested from of a non-induced cell line at various time points.

High-throughput DNA sequencing and analysis. Genomic DNA was extracted from each sample using the Qiagen DNAeasy Blood&Tissue kit. To amplify the hgRNA locus, each sample was subjected to PCR amplification with the following primers for about 20 cycles:

SBS3_scSp_F1 acactctttccctacacgacgctcttccgatct atggactatcatatgcttaccgt
SBS9_scSp_R1 tgactggagttcagacgtgtgctcttccgatct ttcaagttgataacggactagc

These primers amplify a fragment starting 81bp upstream of the transcription start site for the U6 promoter and ending 52bp after the PAM site in hgRNA. The total fragment length varies for various hgRNA constructs, but in its shortest form is 157bp in length. These PCR reactions were carried out in a real-time setting, and stopped in mid-exponential phase.

To add the complete Illumina sequencing adaptors, the above PCR product was diluted and used as a template for a second PCR reaction with NEBNext Dual Indexing Primers, with each sample receiving a different index. Once again, PCR was carried out in a real-time setting and stopped in mid-exponential phase, which was after about 15 cycles in all cases. Samples were then combined and sequenced using Illumina MiSeq with reagent kit v3. Sequencing was done in one direction, starting from the U6 promoter and in the direction of the hgRNA scaffold, and for 170bp on average, but longer for libraries with longer hgRNA constructs.

The frequency of each hgRNA variant, which contained at least one mismatch, deletion, or insertion compared to the sequence of the original hgRNA, was determined for each sample. The diversity of the hgRNA library that was produced as the result of induction of hCas9 expression in the cells was represented by the Shannon Entropy of the vector of all variants' frequencies in each condition. The generated diversity by each hgRNA (5 bits for most) was calculated as the difference between the highest observed diversity after induction and the diversity observed before induction. The background amount of diversity that is observed before induction for each hgRNA, which is expected to have no diversity at that point, is likely due to sequencing based errors. Combined with the fact that the entire sequenced region of each construct was considered with the analysis, this factor creates a background level of diversity in hgRNAs. As this level is independent of Cas9 induction and hgRNA function, and as all time-points corresponding to each line in our diversity plots (Figures 3D, 3H) are the result of the same lentiviral production and sequencing experiments, this background can be simply subtracted.

To obtain fast alignment of a large number of reads to a short template we used two-step approach. First, we aligned all reads to their expected template using blat, which was run on an in-house server. Blat helped determine insertions and deletions, while keeping mismatches intact. In cases where blat results indicated a deletion that was partially or entirely overlapping with an insertion, we used a dynamic programming

algorithm with a match score of 5 and mismatch and deletion scores of 0 to optimize the alignment further. Based on the alignment results, we annotated the spacer, the PAM, and the sequenced portions of the promoter and scaffold from each sequenced hgRNA. A sequenced hgRNA was deemed functional, or capable of re-cutting itself, if it had a PAM and a functional spacer, as well as correct promoter and scaffold. A promoter was annotated as correct if 80% of its sequenced and non-primer overlapping bases correctly matched the consensus promoter. The scaffold was annotated as correct if 90% of its sequenced and non-primer overlapping bases, excluding the PAM bases, correctly matched the expected scaffold. PAM was considered as correct if it matched the NGG sequence (though it was noticed that non-NGG PAMs, such as NHG and NGG, showed significant activity). The spacer was considered functional when it was longer than 17 bases (deletions often lead to inactive hgRNAs with shortened spacers if the distance between the U6 transcription start site and PAM is short).

In situ amplification and detection. HEK/293T cells were seeded at 10,000 per well in 96-well polystyrene dishes coated with poly-D-lysine. 12 hours later, each well was transfected with 100ng of plasmid a plasmid DNA packaged with 0.5μL of Lipofectamin 2000 reagent (ThermoFisher Scientific) accordingly to the manufacturer protocol. Positive samples received plasmids encoding for Design 1 or Design 2 constructs. Negative control samples received a GFP plasmid. 24 hours after transfection cell were subjected to in situ amplification and detection of the gRNA transcripts.

In situ detection was carried out according to the previously described sequencing in situ sequencing protocol(Lee et al. 2014; Lee et al. 2015). In brief, cells were fixed using formalin and permeabilized. Reverse-transcription was then carried out using a target-specific primer (5P-tcttctgaaccagactcttgtcattggaaagttggtataagacaacagtg) in present of aminoallyl-dUTP. Nascent cDNA strands were crosslinked by treatment with BS(PEG)9 (ThermoFisher Scientific) and RNA was degraded by RNaseA and RNaseH treatment. cDNA was circularized using CircLigaseII (Epicentre). Rolling circle amplification (RCA) was carried out with Phi29 polymerase using a target-specific primer (gcctggagcaattccacaacac) overnight in presence of aminoallyl-dUTP. Nascent amplicons or ‘rolonies’ were crosslinked by treatment with BS(PEG)9. Target amplicons were labeled with a fluorescent target-specific detection probe (5Cy5-tcttctgaaccagactcttgt) which recognizes the reverse-transcription primer and nuclei were stained with DAPI. Samples were imaged with a Zeiss Observer.Z1 inverted microscope using a 20X magnification objective in the DAPI and Cy5 channels.

References

Certo, Michael T., Byoung Y. Ryu, James E. Annis, Mikhail Garibov, Jordan Jarjour, David J. Rawlings, and Andrew M. Scharenberg. 2011. “Tracking Genome Engineering Outcome at Individual DNA Breakpoints.” *Nature Methods* 8 (8): 671–76.

Church, George M., Adam H. Marblestone, and Reza Kalhor. 2015. “Rosetta Brain.” In *The Future of the Brain: Essays by the World’s Leading Neuroscientists*, edited by Gary Marcus and Jeremy Freeman, 50–66. Princeton: Princeton University Press.

Dymecki, S. M., and H. Tomasiewicz. 1998. “Using Flp-Recombinase to Characterize Expansion of Wnt1-Expressing Neural Progenitors in the Mouse.” *Developmental Biology* 201 (1): 57–65.

Farzadfard, Fahim, and Timothy K. Lu. 2014. “Synthetic Biology. Genomically Encoded Analog Memory with Precise in Vivo DNA Writing in Living Cell Populations.” *Science* 346 (6211): 1256272.

Ke, Rongqin, Marco Mignardi, Alexandra Pacureanu, Jessica Svedlund, Johan Botling, Carolina Wählby, and Mats Nilsson. 2013. “In Situ Sequencing for RNA Analysis in Preserved Tissue and Cells.” *Nature Methods* 10 (9): 857–60.

Konermann, Silvana, Mark D. Brigham, Alexandro E. Trevino, Julia Joung, Omar O. Abudayyeh, Clea Barcena, Patrick D. Hsu, et al. 2015. “Genome-Scale Transcriptional Activation by an Engineered CRISPR-Cas9 Complex.” *Nature* 517 (7536): 583–88.

Kretzschmar, Kai, and Fiona M. Watt. 2012. “Lineage Tracing.” *Cell* 148 (1-2): 33–45.

Lee, Je Hyuk, Evan R. Daugharthy, Jonathan Scheiman, Reza Kalhor, Thomas C. Ferrante, Richard Terry, Brian M. Turczyk, et al. 2015. “Fluorescent in Situ Sequencing (FISSEQ) of RNA for Gene Expression Profiling in Intact Cells and Tissues.” *Nature Protocols* 10 (3): 442–58.

Lee, Je Hyuk, Evan R. Daugharthy, Jonathan Scheiman, Reza Kalhor, Joyce L. Yang, Thomas C. Ferrante, Richard Terry, et al. 2014. “Highly Multiplexed Subcellular RNA Sequencing in Situ.” *Science* 343 (6177): 1360–63.

Liu, Zhe, and Philipp J. Keller. 2016. “Emerging Imaging and Genomic Tools for Developmental Systems Biology.” *Developmental Cell* 36 (6): 597–610.

Lu, Rong, Norma F. Neff, Stephen R. Quake, and Irving L. Weissman. 2011. “Tracking Single Hematopoietic Stem Cells in Vivo Using High-Throughput Sequencing in Conjunction with Viral Genetic Barcoding.” *Nature Biotechnology* 29 (10): 928–33.

Mali, Prashant, John Aach, P. Benjamin Stranges, Kevin M. Esvelt, Mark Moosburner, Sriram Kosuri, Luhan Yang, and George M. Church. 2013. “CAS9 Transcriptional Activators for Target Specificity Screening and Paired Nickases for Cooperative Genome Engineering.” *Nature Biotechnology* 31 (9). Nature Publishing Group: 833–38.

Mali, Prashant, Kevin M. Esvelt, and George M. Church. 2013. “Cas9 as a Versatile Tool for Engineering Biology.” *Nature Methods* 10 (10): 957–63.

Mali, Prashant, Luhan Yang, Kevin M. Esvelt, John Aach, Marc Guell, James E. DiCarlo, Julie E. Norville, and George M. Church. 2013. “RNA-Guided Human Genome Engineering via Cas9.” *Science* 339 (6121). American Association for the Advancement of Science: 823–26.

Marblestone, Adam H., Bradley M. Zamft, Yael G. Maguire, Mikhail G. Shapiro, Thaddeus R. Cybulski, Joshua I. Glaser, Dario Amodei, et al. 2013. "Physical Principles for Scalable Neural Recording." *Frontiers in Computational Neuroscience* 7 (October): 137.

Marblestone, A. H., E. R. Daugharthy, R. Kalhor, I. D. Peikon, J. M. Kebschull, S. L. Shipman, Y. Mishchenko, et al. 2013. "Conneconomics: The Economics of Dense, Large-Scale, High-Resolution Neural Connectomics." doi:10.1101/001214.

McKenna, Aaron, Gregory M. Findlay, James A. Gagnon, Marshall S. Horwitz, Alexander F. Schier, and Jay Shendure. 2016. "Whole Organism Lineage Tracing by Combinatorial and Cumulative Genome Editing." *Science*, May. American Association for the Advancement of Science, aaf7907.

Perli, Samuel, Cheryl Cui, and Timothy K. Lu. 2016. "Continuous Genetic Recording with Self-Targeting CRISPR-Cas in Human Cells." *bioRxiv*. doi:10.1101/053058.

Platt, Randall J., Sidi Chen, Yang Zhou, Michael J. Yim, Lukasz Swiech, Hannah R. Kempton, James E. Dahlman, et al. 2014. "CRISPR-Cas9 Knockin Mice for Genome Editing and Cancer Modeling." *Cell* 159 (2): 440–55.

Porter, Shaina N., Lee C. Baker, David Mittelman, and Matthew H. Porteus. 2014. "Lentiviral and Targeted Cellular Barcoding Reveals Ongoing Clonal Dynamics of Cell Lines in Vitro and in Vivo." *Genome Biology* 15 (5): R75.

Sulston, J. E., E. Schierenberg, J. G. White, and J. N. Thomson. 1983. "The Embryonic Cell Lineage of the Nematode Caenorhabditis Elegans." *Developmental Biology* 100 (1): 64–119.

Walsh, C., and C. L. Cepko. 1992. "Widespread Dispersion of Neuronal Clones across Functional Regions of the Cerebral Cortex." *Science* 255 (5043): 434–40.

Weisblat, D. A., R. T. Sawyer, and G. S. Stent. 1978. "Cell Lineage Analysis by Intracellular Injection of a Tracer Enzyme." *Science* 202 (4374): 1295–98.

Yang, Luhan, Marc Guell, Susan Byrne, Joyce L. Yang, Alejandro De Los Angeles, Prashant Mali, John Aach, et al. 2013. "Optimization of Scarless Human Stem Cell Genome Editing." *Nucleic Acids Research* 41 (19). Oxford University Press: 9049–61.

Acknowledgements

The authors would like to acknowledge Wei Leong Chew, John Aach, Susan Byrne, Evan Daugharthy, Thomas Ferrante, Je Hyuk Lee, Mark Moosburner, Ian Peikon, Henry Lee, Alex Ng, Javier Fernandez Juarez, Adam Marblestone, Alex Chavez, Yoav Mayshar, Jonathan Scheiman, Garry Cuneo, Ting Wu, Jay Shendure and Tim Lu for helpful comments or discussions. This work has been supported by funding from NIH grants MH103910, HG005550 and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00008, and UCSD new faculty startup funds.

Competing financial interests

The authors declare no competing financial interests. RK, PM and GMC have filed patent applications based on this study.

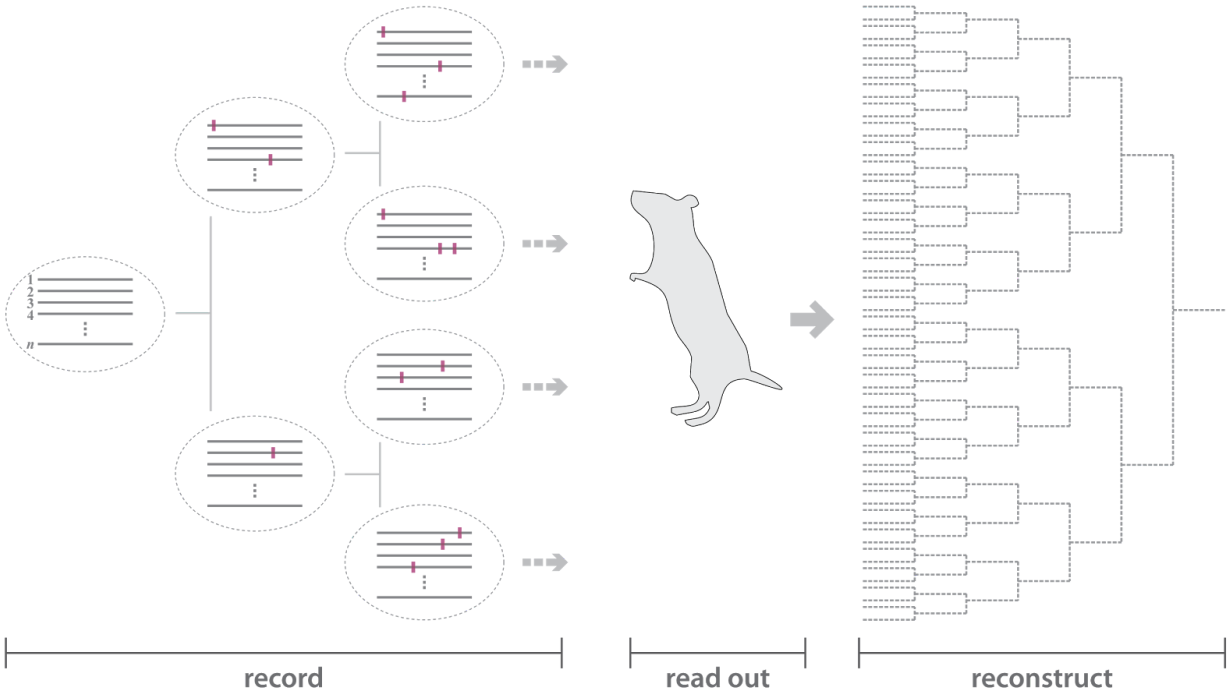

Figure 1. Schematic representation of a lineage tracing approach in multi-cellular eukaryotes using genome engineering and DNA sequencing. (Left) An array of *n* barcoding sites, represented by lines, are introduced into the genome of the parent cell of a lineage (e.g., a zygote or stem cell), represented by dashed ovals. The sites are designed to be stochastically targeted and mutated during development to confer each cell a unique set of target sites, or barcode, which encompasses the lineage history of each cell. (Center) Once the developmental process is finished, the array is sequenced in all cells to read out its information content. (Right) The sequence of the array in each cell is used to compute a dendrogram of their lineage history.

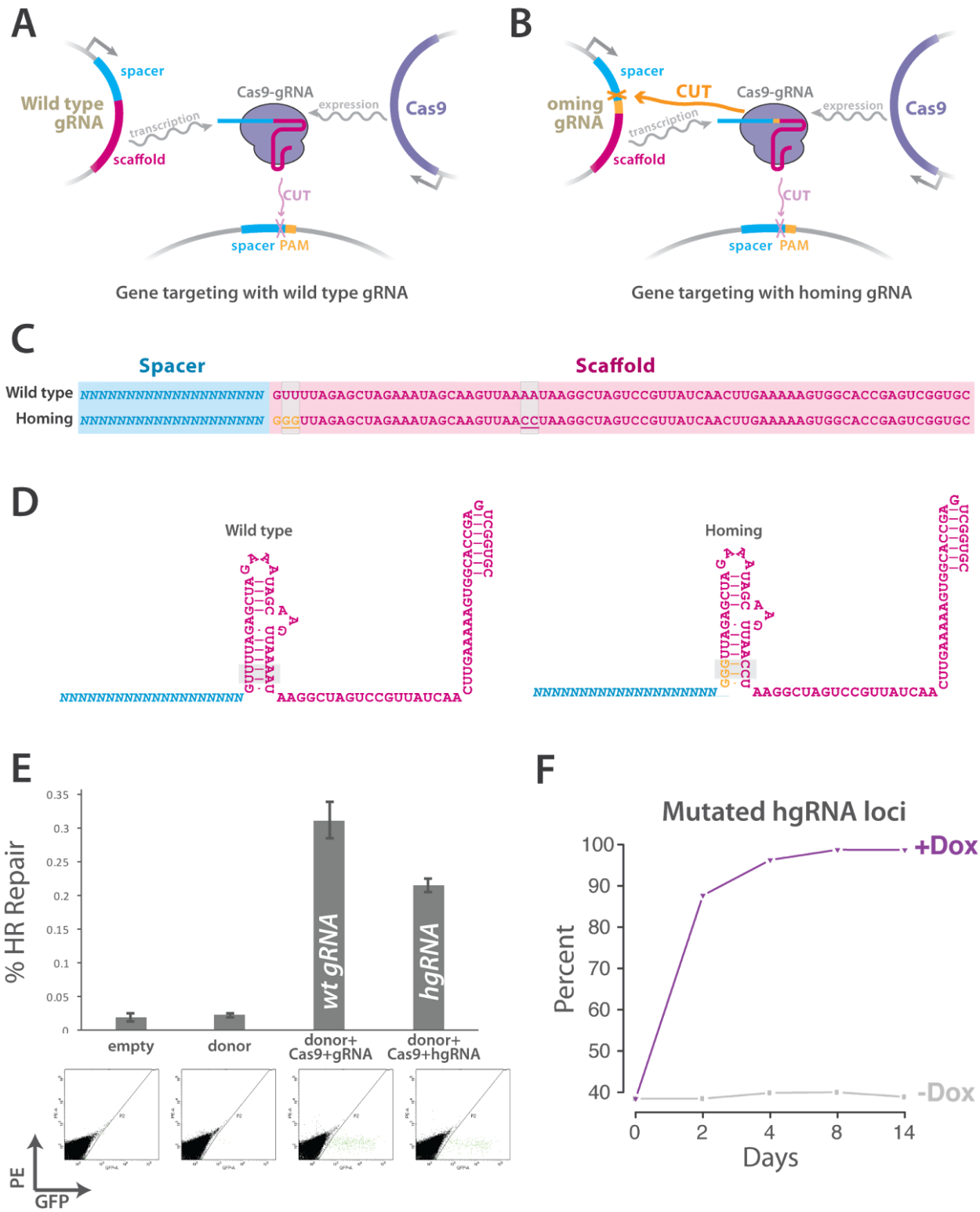

Figure 2. Homing CRISPR-Cas9 system. (**A**) In a canonical CRISPR-Cas9 system, Cas9 and gRNA are expressed from their respective loci. They form a complex and cut their target that contains both the spacer sequence of the gRNA and the PAM. (**B**) In a

homing CRISPR-Cas9 system, the Cas9-gRNA complex also targets the locus encoding the gRNA in addition to cutting its target sequence. (**C**) Primary sequences of wild type and homing gRNAs are shown and the positions that were mutated to create a hgRNA are underlined. Orange bases mark the position of PAM in the hgRNA. Grey boxes mark the positions where hgRNA is mutated compared to the wild type gRNA. (**D**) The predicted secondary structures of wild type and homing gRNAs are depicted. The compensatory mutations made in the homing gRNA enable it maintain the RNA helix structure while creating a PAM for Cas9 binding to the parent DNA of the gRNA. (**E**) Results of a HR-based assay to evaluate to functionality of hgRNAs. A genomically integrated GFP coding sequence is disrupted by the insertion of a stop codon and a 68-bp genomic fragment from the AAVS1 locus. Restoration of the GFP sequence by HR with an appropriate donor sequence results in GFP+ cells that can be quantified by FACS. AAVS1 locus contains a site known as T1 which matches the spacer sequence of hgRNA-A21. Bar graph on top depicts HR efficiencies induced by wild type and homing T1 guide RNAs, as measured by FACS. Representative FACS plots of the targeted cells are depicted below. Data are shown as means ± SEM (N = 4). (**F**) Sequencing results showing above-background accumulation of mutations in hgRNA locus upon Cas9 expression. Cas9 expression is induced in cells with a lentivirally integrated hgRNA-A21. DNA samples harvested before (t=0 days) and at various points after induction are characterized by high-throughput sequencing to quantify mutations in the DNA locus that encodes hgRNA-A21. Any mismatch, deletion, or insertions compared to the original locus is considered a mutation. The high initial fraction of mutants is due to sequencing error (materials and methods) as the steady level of mutations in the non-induced sample suggests there is no significant Cas9 expression leakage.

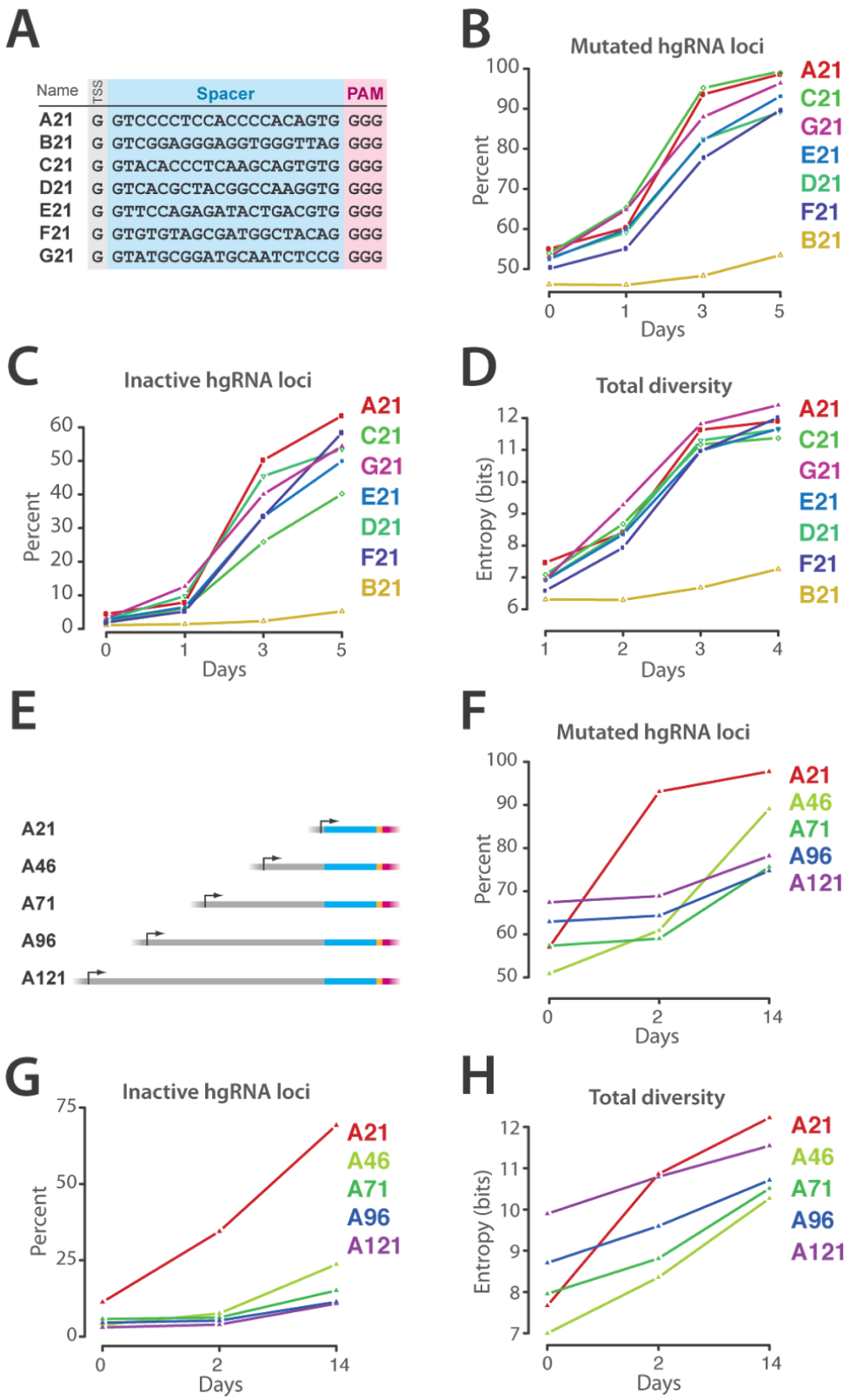

Figure 3. Performance of various hgRNAs. (**A**) Design of seven different hgRNAs is shown. Transcription start site (TSS), Spacer sequence, and PAM are marked with

grey, blue, and pink boxes, respectively. (**B,C,D**) Cas9 is induced in cell lines with the hgRNAs from panel A integrated lentivirally. DNA samples harvested before (t=0 days) and at various points after induction are characterized by high-throughput sequencing to quantify mutations and functionality of hgRNA loci. Any mismatch, deletion, or insertion compared to the original locus is considered a mutation (B). The high pre-induction fraction of mutants is due to sequencing error (materials and methods). Any hgRNA lacking PAM or a spacer shorter than 17 bases is considered inactive (C). Amount of sequence diversity generated by hgRNAs is measured as the Shannon entropy of the frequency vector of all variants that were observed in each sample (D). (**E**) Design of four longer variants of hgRNA-A21 is shown. Stuffer sequences of 25, 50, 75, or 100 base pairs were added upstream of the hgRNA-A21 spacer to obtain the four increasingly lengthy variants. (**F,G,H**) Cas9 is induced in cell lines with the hgRNAs from panel E integrated lentivirally. DNA samples harvested before (t=0 days) and at various points after induction are characterized by high-throughput sequencing to quantify mutations and functionality of hgRNA loci. Any mismatch, deletion, or insertion compared to the original locus is considered a mutation (F). The high pre-induction fraction of mutants is due to sequencing error (materials and methods). Any hgRNA lacking PAM or a spacer shorter than 17 bases is considered inactive (G). Amount of sequence diversity generated by hgRNAs is measured as the Shannon entropy of the frequency vector of all variants that were observed in each sample (H).

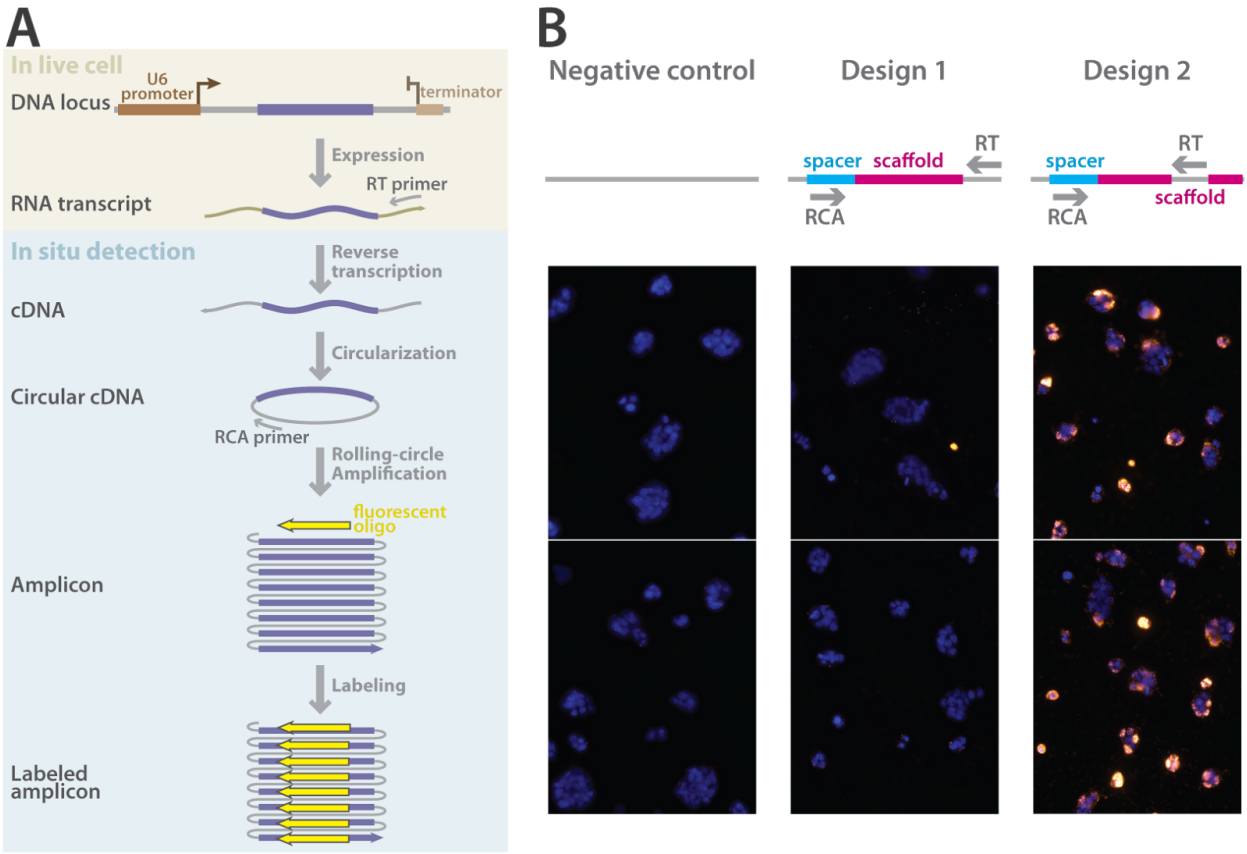

Figure 4. Target-specific in situ detection of gRNA molecules expressed by RNA polymerase III promoter. (**A**) A schematic of our in situ amplification and detection assay based on FISSEQ is shown. A DNA locus expressing a gRNA (purple) under the U6 promoter (brown) is introduced into cells. The construct also contains designed primer binding sites both downstream and upstream of the gRNA in grey-colored regions. A terminator (light brown) is placed after the second primer binding region. Cells containing this locus, and thus expressing its RNA transcript, are fixed for in situ amplification and detection. In the fixed cells, the RNA transcript is reverse-transcribed using a locus-specific RT (reverse-transcription) primer to obtain a cDNA which is then circularized. The circular cDNA is amplified by the rolling circle amplification (RCA) using a second locus-specific RCA primer, producing a concatemerized amplicon that is confined to a small space in the hydrogel matrix of the experiment. The amplicon is then labeled by a fluorescent oligonucleotide. (**B**) Results of target-specific in situ amplification and detection for two different gRNA loci and a negative control locus are depicted. The schematic on top shows the positions of the specific primers in each design. The bottom panels show the results for cells containing those loci. Amplicons are labeled with Cy3 (yellow) and nuclei are labeled with DAPI (blue). The amplicon is only detectable in cell transfected with the Design 2 constructs, whereas Design 1 only

shows very few labeled amplicons, at a level similar to the false positive amplicons in the negative control.