Rapidly evolving homing CRISPR barcodes

Reza Kalhor[1], Prashant Mali[2] & George M Church[1,3]

We present an approach for engineering evolving DNA barcodes in living cells. A homing guide RNA (hgRNA) scaffold directs the Cas9–hgRNA complex to the DNA locus of the hgRNA itself. We show that this homing CRISPR–Cas9 system acts as an expressed genetic barcode that diversifies its sequence and that the rate of diversification can be controlled in cultured cells. We further evaluate these barcodes in cell populations and show that they can be used to record lineage history and that the barcode RNA can be amplified *in situ*, a prerequisite for *in situ* sequencing. This integrated approach will have wide-ranging applications, such as in deep lineage tracing, cellular barcoding, molecular recording, dissecting cancer biology, and connectome mapping.

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 the understanding of human development and on the ability to restore normal function in damaged or diseased tissues. One such parameter that can provide important insights into developmental processes is the lineage history of cells[1,2]. Contemporary lineage-tracing approaches, however, do not readily scale to the model organisms, such as mice, that are most relevant to human development[3–7]. Precise mapping of lineage history in these organisms may be facilitated by combining modern genome engineering and DNA sequencing technologies[8–12]: if every cell in an organism contained a unique and easily retrievable DNA sequence—a barcode—that encompassed its lineage relationship with other cells, this barcode could be probed to delineate the precise lineage history of all cells in the organism.

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 (**Fig. 1**). This approach entails an array of genomically integrated sites that are stochastically targeted by a nuclease. During each cell cycle, the nuclease targets a random subset of these sites where the process of non-homologous end-joining (NHEJ) introduces insertions, deletions, or other mutations, leading to a unique sequence that is related to its parent sequence and may further evolve in subsequent rounds. At the end of the developmental process, single-cell assaying technologies can be applied to each cell to decipher its unique barcode—the sequences of its nuclease-site array—and delineate its lineage history (**Fig. 1**).

Such an array of evolving barcodes can in theory fulfill the requirements of precise lineage tracing. However, for practical implementation into an animal model such as mouse, an array will have to meet several critical criteria. Specifically, 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 independent array elements (i.e., without crosstalk) is 'n', then the system allows the creation of m^n possible signatures. Therefore, any system that generates higher values of m and n would be highly desirable. Furthermore, the system should continue to generate diversity throughout the development of the animal and should present tractable options for stable animal lines. Finally, it should be scalably readable at a single-cell level.

With above concepts in mind, we present homing guide RNAs (hgRNAs), a modified CRISPR–Cas9 system that targets the DNA locus of the guide RNA itself. We show that this simple system generates more diversity than canonical CRISPR–Cas9 (has higher m). We further show that it is consistent with deployment in an array format with independently acting barcoding elements (higher n), and that the rate of diversification can be regulated to match the requisite pace for most model organisms. Additionally, we show that these barcodes are appropriate for lineage-tracing applications *in vitro*, and their corresponding small RNAs can be assayed as single molecules *in situ*. We propose that these properties make hgRNAs an excellent candidate to integrate into animal models for barcoding and lineage-tracing purposes.

RESULTS

The canonical CRISPR–Cas9 system can introduce mutations at a target locus via the process of non-homologous end joining (NHEJ). This system involves three components: a single guide RNA (sgRNA), the Cas9 protein, and a target site that includes a protospacer-adjacent motif (PAM) that is directly recognized by the Cas9 protein. As sgRNAs do not have a PAM, their loci are not targeted by the Cas9–gRNA complex despite containing the spacer sequence (**Fig. 2a**). We sought to engineer a homing CRISPR system that directs Cas9–gRNA nuclease activity to the gRNA locus itself, thus simplifying the system by combining the guide RNA

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [2]Department of Bioengineering, University of California San Diego, La Jolla, California, USA. [3]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts, USA. Correspondence should be addressed to G.M.C. (gchurch@genetics.med.harvard.edu) or P.M. (pmali@ucsd.edu).

RECEIVED 28 JUNE; ACCEPTED 6 NOVEMBER; PUBLISHED ONLINE 5 DECEMBER 2016; DOI:10.1038/NMETH.4108

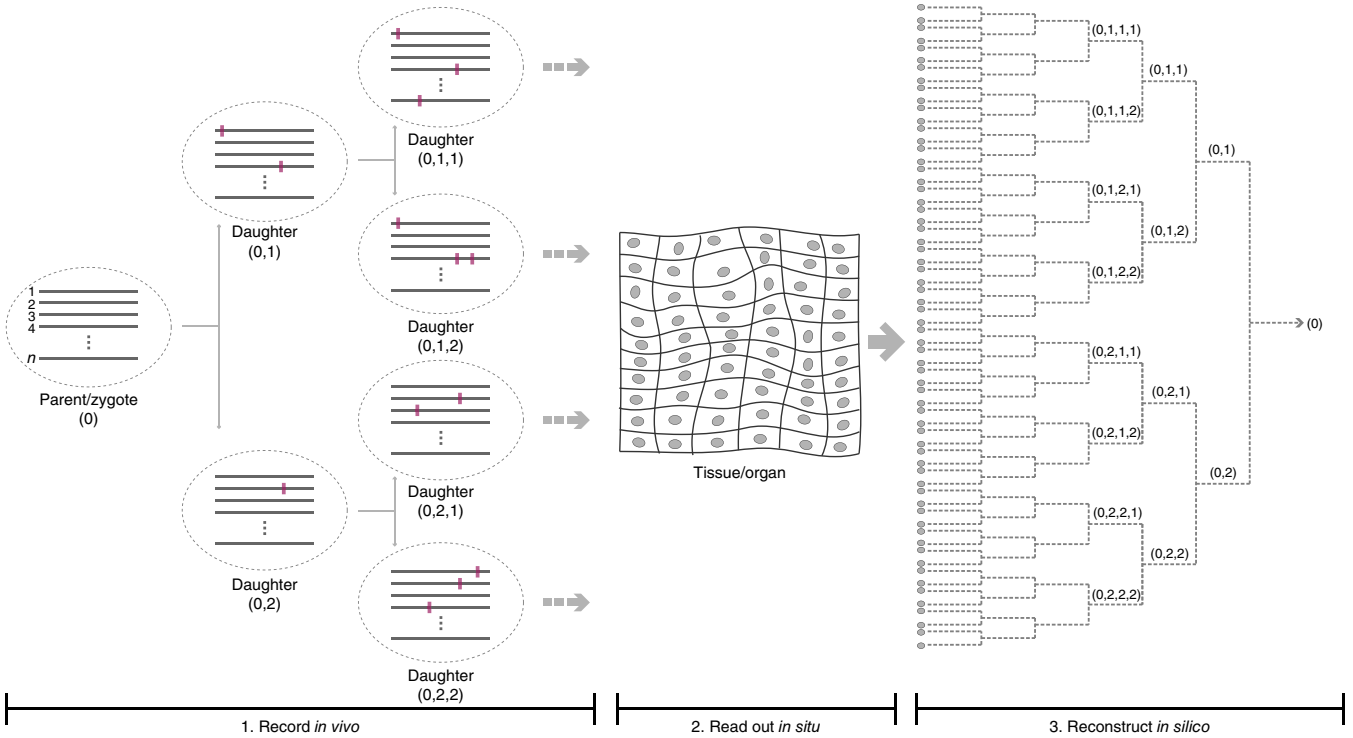

Figure 1 | Schematic representation of a lineage-tracing approach in multicellular eukaryotes using genome engineering and single cell DNA sequencing. Left, lineage recording during development. Array of *n* barcoding sites are represented by gray lines, cells are represented by dashed ovals, and mutations are represented by the purple rectangles. Center, *in situ* readout of barcode sequences in tissues and organs. Right, *in silico* reconstruction of the lineage dendrogram based on the *in situ* sequencing readout of each cell.

locus and the target site and enabling retargeting and evolvability of barcodes (**Fig. 2b**). To engineer a homing gRNA, we mutated the sequence immediately downstream of the *Streptococcus pyogenes* sgRNA spacer from 'GUU' to 'GGG' (**Supplementary Fig. 1a**), so that 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 canonical sgRNA. To preserve the helix and minimize adverse structural impacts of our mutations, we also introduced compensatory mutations in the hybridizing nucleotides. We then evaluated the functionality of this homing gRNA using an assay based on homologous recombination (**Supplementary Fig. 1b**). In this assay, a 'broken' GFP gene is targeted by the Cas9–gRNA complex in the presence of a repair template[13]. Successful targeting of the broken GFP gene results in its repair through homologous recombination, and the ensuing restoration of fluorescence can be detected. The results showed that our homing gRNA, or hgRNA, can digest a target sequence.

To validate that homing gRNAs target their own locus, we created a HEK/293T clonal cell line genomically integrated with the humanized *S. pyogenes* Cas9 under the control of an inducible Tet-On promoter (293T-iCas9 cells). We introduced both the hgRNA locus and its target into the genome of these cells using lentiviral integration. As a control, we assembled the same system but with the canonical non-homing version of the hgRNA. We then induced Cas9 expression, harvested DNA samples at various intervals, and sequenced the guide RNA locus and its target in each system. The results showed cutting of the target locus by both the hgRNA and its sgRNA counterpart (**Fig. 2c**,**d**). sgRNA was more efficient in mutating the target locus; however, only hgRNA induced mutations in its own DNA (**Fig. 2d**, **Supplementary Fig. 1c**). In fact, the hgRNA was as efficient in mutating its own DNA as the sgRNA was in mutating the target locus, both resulting in mutations in almost all cells at 5 d after induction. These mutations were above background sequencing error rates that we measured by sequencing fragments of a similar size from the hygromycin-resistance gene upstream of the gRNA locus on the lentiviral backbone and the 16S rRNA gene on the genomic DNA of the cells.

We next compared the diversity generated by hgRNAs to that generated by sgRNAs. Two factors are important in considering this diversity: the number of different variants and the frequency distribution of those variants. The ideal barcoding locus would generate a very high number of variants, all with an equal likelihood. As a proxy for both these factors, we measured the Shannon entropy of the frequencies of all the variants generated by both our hgRNA and its corresponding sgRNA in their gRNA and target loci (**Fig. 2e**,**f**). The results show that hgRNA can generate about 5 bits of diversity in its locus after Cas9 expression. This amount is a substantial improvement over the 2 bits generated by the sgRNA counterpart in its target and does not appear to come at a substantial cost to cell viability (**Supplementary Fig. 1d**). The more diverse output, which is likely due to the evolution of hgRNAs past the first set of NHEJ products, suggests that hgRNAs are more suitable than sgRNAs for barcode generation.

The amount of diversity generated by a single hgRNA suggests that uniquely barcoding all neurons in a mouse brain requires an array of at least six hgRNAs per cell (see Discussion). To assess whether hgRNAs with a different spacer sequence show a similar behavior, we created six additional hgRNAs (B21, C21, D21, E21, F21, and G21) (**Fig. 3a**, **Supplementary Note**). We assayed these

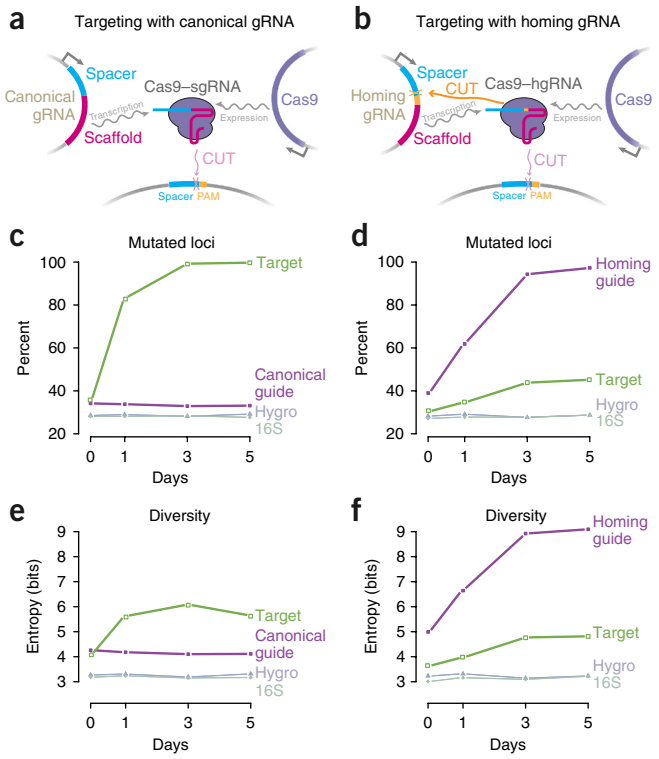

Figure 2 | Comparison between standard and homing CRISPR–Cas9 systems. (**a**) Canonical CRISPR–Cas9 system, in which Cas9 and sgRNA are expressed from their respective loci, form a complex, and cut their target that both matches the spacer sequence of the gRNA and contains a PAM. (**b**) Homing CRISPR–Cas9 system, in which the Cas9–hgRNA complex, in addition to cutting their target sequence, also targets the locus encoding the hgRNA itself. (**c**,**d**) Accumulation of mutations in the gRNA, the target, a fragment of hygromycin, and a fragment of the 16S ribosomal RNA loci upon Cas9 expression in cells with either canonical (**c**) or homing (**d**) versions of gRNA-A21. (**e**,**f**) Generated diversity in experiments in **c** and **d** measured as the Shannon entropy of the frequency vector of all variants that were observed in each condition. Data points are means (n = 2, biological replicates; s.e.m. small and not distinguishable on the plot scale).

hgRNAs in 293T-iCas9 cells (**Fig. 3b**,**c**, **Supplementary Table 1**) and saw that five of the six are highly active in targeting their parent loci and generate a similar amount of diversity, ranging from 5 to 6 bits. hgRNA-B21, which showed much lower activity, has a spacer with multiple 'GG' dinucleotides, a feature that has previously been shown to inhibit gRNA function[14]. These observations suggest that hgRNAs are generally functional irrespective of their spacer sequence and can thus operate as barcoding loci with minimal crosstalk between them.

Our first hgRNA set generated diversity for only a short time after induction of Cas9 expression before being inactivated (**Fig. 3d**). While some were inactivated as a result of deletions that removed the PAM from their scaffold, others were rendered inactive by a truncation of their spacer below the 16–18 nucleotides necessary for Cas9–gRNA cleavage (**Supplementary Table 1**). As our hgRNA set had only 21 total bases between the RNA transcription start site and the scaffold (**Fig. 3a**), 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 variants that were 5, 30, 55, and 80 bases longer than hgRNA-A21 but had a similar initial spacer sequence (**Fig. 3e**, **Supplementary Fig. 2**, **Supplementary Note**). These hgRNAs were all active in our standard assay (**Fig. 3f**, **Supplementary Fig. 2b**), with the mutation and diversification rates decreasing with increasing hgRNA length (**Fig. 3f**,**g**, **Supplementary Fig. 2b**,**c**). Furthermore, unlike A21, these longer hgRNAs continued to generate diversity for a few weeks after induction (**Fig. 3h**, **Supplementary Fig. 2d**). These results show that hgRNA activity level can be regulated to accommodate developmental processes on the scale of weeks as well as those on the scale of days.

We further assessed whether hgRNAs can fulfill lineage-tracing schemes similar to those illustrated in **Figure 1**. To simplify the experiment, instead of using multiple hgRNAs in a single parent cell and establishing the lineage relationship of its daughter cells, we used a single hgRNA in a parent cell population and attempted to decipher the lineage relationship of subpopulations that were derived from this parental population (**Fig. 4a**,**b**). Specifically, we created one cell line with hgRNA-A21 (**Fig. 4a**) and another with hgRNA-C21 (**Fig. 4b**). From each of these parental lines, we established a first generation of daughter subpopulations using ~100 cells and a brief Cas9 induction to generate diversity in the hgRNA locus. In a similar fashion, a second generation of daughter subpopulations was created from the first and a third from the second (**Fig. 4a**,**b**). In the end, the hgRNA locus was sequenced in each subpopulation to profile its variants. The presence of non-shared variants between subpopulations was used as a measure of distance between them (**Supplementary Table 2**, Online Methods). On the basis of these distances, we clustered the second (**Fig. 4c**,**d**, left) and third (**Fig. 4c**,**d**, right) generations of subpopulations and found that their lineage relationship could be reconstituted from sequencing data. These results confirm that the diversity generated in hgRNA loci can be used to inform their lineage relationship.

Finally, we turned our attention to barcode readout strategies, as retrievability is a requirement for effective lineage tracing. In this regard, *in situ* sequencing is a highly desirable method for barcode retrieval as it allows lineage information to be extracted without loss of histological information such as position and cell type[9,15]. One limitation of fluorescence *in situ* sequencing (FISSEQ) technologies is their short read length of only 20–30 base pairs (bp). Our approach is uniquely appropriate for FISSEQ as the spacer sequence of the hgRNA loci where the barcode is generated is 20 bp in length. However, available FISSEQ technologies probe arbitrary regions in a stochastic fraction of longer transcripts in a cell; they face difficulty in both targeted sequencing and detection of smaller RNA molecules[9,15,16]. We therefore assessed whether hgRNAs, which are small, can be probed in a targeted fashion using FISSEQ. FISSEQ can be divided into two stages: amplicon generation and amplicon sequencing. The difficulties associated with targeted probing of small RNA molecules relate to the amplicon generation step[9,15,16]; thus, we used an assay to specifically address the amplification step for hgRNAs (**Supplementary Fig. 3a**). We then executed the assay on representative hgRNA constructs with specific reverse-transcription (RT) primers that would target the barcoded region. Since RT and rolling-circle amplification (RCA) primers that aim to amplify the entire hgRNA could not generate amplicons above the background level (**Supplementary Fig. 3b**, middle), we

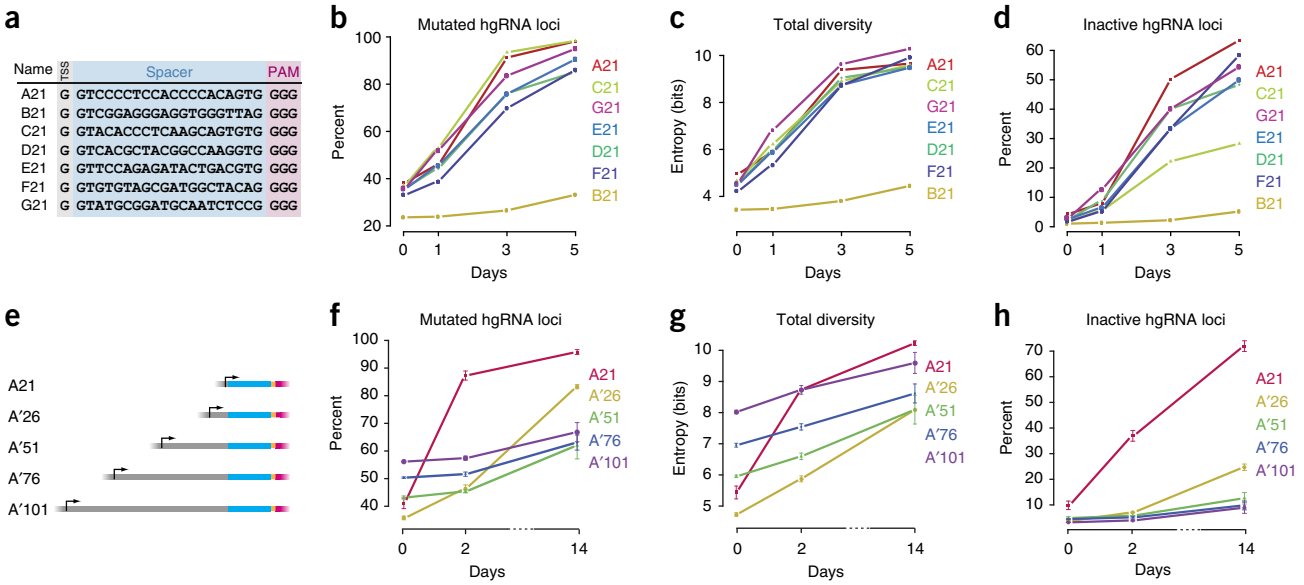

Figure 3 | Performance of various hgRNAs. (**a**) The sequence of seven different hgRNAs. Transcription start site (TSS), spacer sequence, and PAM are marked by gray, blue, and pink boxes, respectively. (**b–d**) Cas9 is induced in cell lines with the hgRNAs shown in **a** integrated genomically. DNA samples harvested before (t = 0 days) and at various points after induction are characterized by high-throughput sequencing to quantify mutations, functionality, and generated diversity for each hgRNA (**Supplementary Table 1**). Data points represent single replicates. (**e**) Design of four longer variants of hgRNA-A21. Stuffer sequences of 5, 30, 55, or 80 base pairs were added upstream of a spacer very similar to the A21 spacer to obtain the four increasingly lengthy A′ variants. (**f–h**) Cas9 is induced in cell lines with the hgRNAs shown in **e** integrated genomically. DNA samples harvested before (t = 0 days) and at various points after induction are characterized by high-throughput sequencing to quantify mutations, functionality and diversity of hgRNA loci. Data points are mean ± s.e.m. (n = 2, biological replicates).

inserted the RT primer docking sequence inside the hgRNA scaffold at positions previously shown to be tolerant of insertions[17,18]. This new arrangement resulted in robust *in situ* amplification and detection of the hgRNA spacer region in a target-specific fashion (**Supplementary Fig. 3b**, right). These results address the challenges associated with *in situ* amplification step of hgRNA barcodes and, assuming 100 amplicons per cell and 10 amplicons per barcode, allow for reliable detection of up to 10 barcodes per cell. Further improvements in imaging capabilities[19] and automation of *in situ* sequencing will pave the way for ultra-dense readout of cellular barcodes.

DISCUSSION

Recent reports demonstrate the utility of nuclease-mediated diversity generation in biological systems[20–22]. For instance, one study utilizes canonical sgRNA–target pairs with decreasing affinities for lineage tracing in zebrafish[20], and another uses homing CRISPR systems for recording analog cellular signals[21]. We apply the homing CRISPR system to engineer evolving barcodes for barcoding and lineage-tracing purposes. We evaluate several important parameters of this system that reflect on its potential for barcoding and lineage-tracing applications.

It is perhaps instructive to estimate the minimum number of barcoding elements that would yield a useful animal model. A homing gRNA locus is capable of storing about 5 bits of information, which is enough to distinguish $2^5 = 32$ different states. Accordingly, uniquely barcoding the roughly 12 billion cells in a mouse will require at least 7 such hgRNA loci per cell ($(2^5)^7 > 12 \times 10^9$). Uniquely barcoding the estimated 75 million neurons in a mouse brain will require at least 6 such hgRNA loci per neuron ($(2^5)^6 > 75 \times 10^6$). While the actual number of barcoding loci needed in practice will depend on the tolerance of an experiment to duplicate barcodes, these estimates suggest that adequate barcoding of mouse neurons, which is essential for some of the proposed brain-mapping projects in the brain initiative[10,23,24], may be within reach of our current strategy. Our preliminary experiments show that a cross between a mouse bearing hgRNA loci and a mouse expressing Cas9 generates offspring with a diversity of barcodes that awaits further follow up.

An important strength of this hgRNA-based barcoding system lies in its potential for establishing a stable transgenic line in a model organism, such as mouse. hgRNA barcoding elements are single-component genetic loci. Canonical sgRNAs, in contrast, require paired gRNA and target elements (**Fig. 2a**) to be integrated into model organisms, a problem that becomes increasingly complex with the requirement for arrays of multiple independent barcoding elements. Furthermore, while hgRNAs are tolerant of mutations in their spacer region, canonical sgRNAs are susceptible to inactivation by mutations in their gRNA spacer or the target protospacer, presenting further challenges for creating stable transgenic lines. Finally, as compared to non-CRISPR nuclease systems, a barcoding model organism with hgRNAs offers all the flexibility and versatility of Cas9 without a need to recreate the transgenic line that carries the hgRNA array: any modified or tissue-specific Cas9 can be delivered by viral vectors or through crossing with a transgenic animal.

We also note key limitations in our current implementation: first, the limited duration of evolvability for the fastest-evolving hgRNAs, and second, the still limited diversity generated by each element; we hope that the use of improved and orthogonal inducible systems[25], coupled with large arrays of evolving barcodes, as

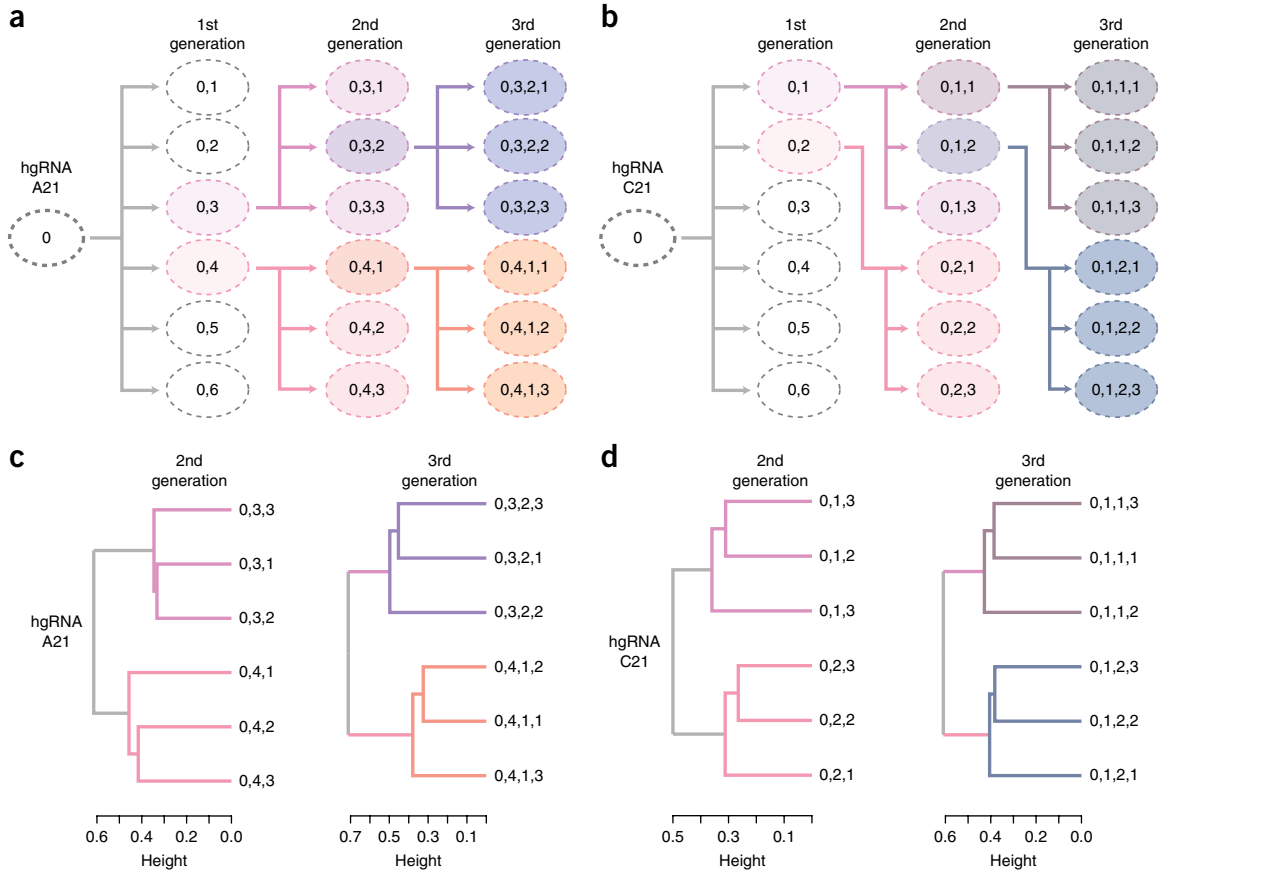

Figure 4 | Lineage tracing in cultured cell populations. (**a**,**b**) Passaging scheme: cells carrying hgRNA-A21 (**a**) or hgRNA-C21 (**b**) were induced to briefly express Cas9, generating a parent population ('0') with a diverse array of hgRNA barcodes. As shown by arrows, the parent population ('0') was passaged into multiple first-generation subpopulations (0,*N*), some of which were then further passaged into second- and third-generation subpopulations (0,*N*,*N* and 0,*N*,*N*,*N* respectively). The ovals represent populations, and the numbers in the ovals represent the label of each line while indicating its relationship to the other lines. Each daughter subpopulation was seeded from about 100 cells that were briefly induced to express Cas9 immediately before seeding. (**c**,**d**) Clustering of subpopulations in **a** and **b** on the basis of the hgRNA variants observed in each (**Supplementary Table 2**).

well as the use of molecules that modulate NHEJ outcomes (such as end-processing enzymes, polymerases and terminal transferases[26]) or base-editing Cas9s[27], can substantially enhance both the durability of hgRNAs and the amount of diversity they generate. Third, the exact effects on cell viability and developmental pathways of integrating an array of nuclease sites into the genome are unclear. Substantial work is required to determine if and to what extent such effects exist and how they can be mitigated.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors would like to acknowledge W.L. Chew, J. Aach, S. Byrne, E. Daugharthy, T. Ferrante, J.H. Lee, M. Moosburner, I. Peikon, H. Lee, A. Ng, J. Fernandez Juarez, A. Marblestone, A. Chavez, Y. Mayshar, J. Scheiman, K. Kalhor, T. Wu, J. Shendure, and T. Lu for helpful comments or discussions and the Biopolymers Facility at HMS for technical assistance. This work has been supported by funding from NIH grants MH103910 and HG005550 (G.M.C.) and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00008 (G.M.C.), and by UCSD new faculty startup funds (P.M.).

AUTHOR CONTRIBUTIONS

R.K., P.M., and G.M.C. conceived the study. R.K. and P.M. carried out the experiments. R.K. analyzed the data. R.K. and P.M. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the online version of the paper.

1. Sulston, J.E., Schierenberg, E., White, J.G. & Thomson, J.N. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100**, 64–119 (1983).
2. Kretzschmar, K. & Watt, F.M. Lineage tracing. *Cell* **148**, 33–45 (2012).
3. Weisblat, D.A., Sawyer, R.T. & Stent, G.S. Cell lineage analysis by intracellular injection of a tracer enzyme. *Science* **202**, 1295–1298 (1978).
4. Dymecki, S.M. & Tomasiewicz, H. Using Flp-recombinase to characterize expansion of Wnt1-expressing neural progenitors in the mouse. *Dev. Biol.* **201**, 57–65 (1998).
5. Walsh, C. & Cepko, C.L. Widespread dispersion of neuronal clones across functional regions of the cerebral cortex. *Science* **255**, 434–440 (1992).
6. Porter, S.N., Baker, L.C., Mittelman, D. & Porteus, M.H. Lentiviral and targeted cellular barcoding reveals ongoing clonal dynamics of cell lines in vitro and in vivo. *Genome Biol.* **15**, R75 (2014).

7. Lu, R., Neff, N.F., Quake, S.R. & Weissman, I.L. Tracking single hematopoietic stem cells *in vivo* using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* **29**, 928–933 (2011).
8. Mali, P., Esvelt, K.M. & Church, G.M. Cas9 as a versatile tool for engineering biology. *Nat. Methods* **10**, 957–963 (2013).
9. Lee, J.H. *et al.* Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).
10. Church, G.M., Marblestone, A.H. & Kalhor, R. in *The Future of the Brain: Essays by the World's Leading Neuroscientists* (eds. Marcus, G. & Freeman, J.) 50–66 (Princeton University Press, 2016).
11. Peikon, I.D., Gizatullina, D.I. & Zador, A.M. In vivo generation of DNA sequence diversity for cellular barcoding. *Nucleic Acids Res.* **42**, e127 (2014).
12. Naik, S.H., Schumacher, T.N. & Perié, L. Cellular barcoding: a technical appraisal. *Exp. Hematol.* **42**, 598–608 (2014).
13. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
14. Malina, A. *et al.* PAM multiplicity marks genomic target sites as inhibitory to CRISPR-Cas9 editing. *Nat. Commun.* **6**, 10124 (2015).
15. Ke, R. *et al.* *In situ* sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013).
16. Lee, J.H. *et al.* Fluorescent *in situ* sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10**, 442–458 (2015).
17. Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31**, 833–838 (2013).
18. Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583–588 (2015).
19. Liu, Z. & Keller, P.J. Emerging imaging and genomic tools for developmental systems biology. *Dev. Cell* **36**, 597–610 (2016).
20. Junker, J.P. *et al.* Massively parallel whole-organism lineage tracing using CRISPR/Cas9 induced genetic scars. Preprint at bioRxiv http://dx.doi.org/10.1101/056499 (2016).
21. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
22. Perli, S.D., Cui, C.H. & Lu, T.K. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science* **353**, aag0511 (2016).
23. Marblestone, A.H. *et al.* Physical principles for scalable neural recording. *Front. Comput. Neurosci.* **7**, 137 (2013).
24. Marblestone, A.H. *et al.* *Conneconomics: The Economics of Dense, Large-Scale, High-Resolution Neural Connectomics*. Preprint at bioRxiv http://dx.doi.org/10.1101/001214 (2013).
25. Platt, R.J. *et al.* CRISPR-Cas9 knockin mice for genome editing and cancer modeling. *Cell* **159**, 440–455 (2014).
26. Certo, M.T. *et al.* Tracking genome engineering outcome at individual DNA breakpoints. *Nat. Methods* **8**, 671–676 (2011).
27. Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A. & Liu, D.R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).

ONLINE METHODS

Vector construction. The Cas9 vectors used in the study are based on earlier published work[11,28]. The hgRNA vectors and the sgRNA counterpart of hgRNA-A21 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 carry Hygromycin B resistance. The target locus of sgRNA-A21 and hgRNA-A21 were constructed by incorporating corresponding gBlock (IDT DNA) synthesized DNA fragment into the pLKO.1 lentiviral backbone (MISSION shRNA vectors via SIGMA) which was modified to carry Blasticidin resistance. This target locus was designed such that its sequence is identical to that of sgRNA-A21 in the region that is subjected to sequencing, however, it cannot act as a hgRNA because it lacks the U6 promoter and contains only a truncated scaffold. The sequence of these inserts is available in the **Supplementary Note**.

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 Human Embryonic Kidney 293T cells (HEK/293T, 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 1 μg/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 −80 °C until use.

Transduction of 293T-iCas9 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 6 μg/ml polybrene. About 3 days after infection, cells were placed under selection with 200 μg/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.

Double transduction of 293T-iCas9 cells with lentiviral vectors carrying standard or homing A21 guide RNA with and lentiviral vectors carrying their target. For the experiments where both the guide RNA and its target were expressed (**Fig. 2**), 293T-iCas9 cells were grown to 70% confluency, at which point they were infected either with sgRNA-A21-Hygromycin and A21-Target-Blasticidin (**Fig. 2a**,**c**,**e**) or with hgRNA-A21-Hygromycin and A21-Target-Blasticidin (**Fig. 2b**,**d**,**f**) with 0.3–0.5 MOI of lentiviral particles in presence of 6 μg/ml polybrene. 3 days after infection, cells were placed under double selection with 200 μg/ml Hygromycin B and 10 μg/ml Blasticidin. Cells were retained under double selection for at least 1 week before any experiments to assure genomic integration. Cells were maintained under Hygromycin B and Blasticidin 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 2 μg/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 (**Supplementary Fig. 1c**).

Lineage tracing in cultured cell populations. A 293T-iCas9 cell line, carrying either hgRNA-A21 or hgRNA-C21, was subjected to two hours of induction with doxycycline to generated limited initial diversity in the hgRNA loci, thus creating each founder cell population, here referred to as '0'. After doxycycline was removed, about 100 cells from each '0' populations were used to seed six subpopulations, '0,1' through '0,6', in 6.5-mm wells. In the course of about ten days, each subpopulation was expanded into larger wells and a sample was taken for sequencing analysis. Two of the subpopulations were then randomly selected for further passaging. About 100 cells from each selected subpopulations were induced with doxycycline for two hours and used to seed the next group of daughter subpopulations, for example, '0,4,1' through '0,4,3'. This passaging scheme was repeated for one more round to create the third generation of daughter subpopulations (**Fig. 4a**,**b**).

High-throughput DNA sequencing. Genomic DNA was extracted from each sample using the Qiagen DNAeasy Blood&Tissue kit. To amplify gRNA loci (canonical or homing) the following primer pair was used:

SBS3_Guide_F acactctttccctacacgacgctcttccgatct atggactatcat atgcttaccgt

SBS9_Guide_R tgactggagttcagacgtgtgctcttccgatct ttcaagttga taacggactagc

These primers amplify a fragment starting 81 bp upstream of the transcription start site for the U6 promoter and ending 55 bp after the start of gRNA scaffold. The total fragment length varies for various hgRNA constructs, but in its shortest form (for example, A21) is 157 bp.

To amplify the target locus the following primer pair was used:

SBS3_Target_F acactctttccctacacgacgctcttccgatct aagaggatggt gcagcaaccaag

SBS9_Target_R tgactggagttcagacgtgtgctcttccgatct tcaatctgacag gtgcctctcac

These primers amplify a fragment starting 82 bp upstream of the spacer sequence and ending 53 bp after the PAM site. The total fragment length is 158 bp.

To amplify the Hygromycin B resistance locus as a control, the following primer pair was used:

SBS3_Hyg_F acactctttccctacacgacgctcttccgatct gtcgatgcgacg caatcgtc

SBS9_Hyg_R tgactggagttcagacgtgtgctcttccgatct ttcctttgccctcg gacgag

These primers amplify a fragment starting 881 bp downstream of the Hygromycin gene start codon and ending just before the stop codon—1,032 bp after the start codon. The total fragment length is 152 bp.

To amplify a part of the 16S rRNA locus as another control the following primer pair was used:

SBS3_16S_F acactctttccctacacgacgctcttccgatct atgcatgtctgag tacgcac

SBS9_16S_R tgactggagttcagacgtgtgctcttccgatct ccgaggttatcta gagtcac

These primers amplify a 259-bp fragment from the human 16S rRNA locus.

PCR reactions with above primer pairs were carried out in a real-time setting and stopped in mid-exponential phase, typically about 20 cycles. To add the complete Illumina sequencing adaptors, this first 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 forward (F) primer in each sample and for 170 bp on average, but longer for libraries with longer hgRNA constructs.

High-throughput DNA sequencing analysis. For each gRNA locus that was subjected to sequencing, only the fragment starting 4 bp before the expected transcription start site for U6 promoter and ending 32 bp into the scaffold was considered during all below analyses. For hgRNA-A21, for instance, this fragment is 57 bp in total length. For the A21 Target locus only the 57-bp fragment exactly equivalent to that of the hgRNA-A21 or sgRNA-A21 was considered. For the Hygromycin and 16S rRNA loci, which acted as controls, only a 57-bp fragment with the same relative start and end positions in the sequencing read as hgRNA-A21 were considered to control for potential positional variations in sequencing accuracy.

The frequency of each hgRNA variant or mutant that contained at least one mismatch, deletion, or insertion compared to the sequence of the original hgRNA was determined for each sample. For the 57-bp fragment considered for hgRNA-A21, sgRNA-A21, Hygromycin and 16S rRNA loci (**Fig. 2c**,**d**), an average sequencing error rate of 0.0064, characterized under standard conditions[29], results in 30% of all fragments to appear as mutants [1 – (1 – 0.0064)[57]]. This rate is in agreement with our observed values for Hygromycin and 16S rRNA loci and shows that the observed background mutant levels are largely a result of sequencing error. For longer construct, hgRNAs A′26, A′51, A′76, A′101, fragments of 62, 87, 112, and 137 bp were considered in the analysis (**Fig. 3f**–**h**). Considering an average sequencing error rate of 0.0064, the expected background mutant levels for these longer fragments would be 33%, 43%, 51%, 59% respectively. These background levels are also in agreement with baseline mutant fractions observed before Cas9 induction in **Figure 3f**.

The diversity of the hgRNA library that was produced as the result of Cas9 expression is represented by the Shannon Entropy of the vector of all variants' frequencies in each condition. Because sequencing error produces some "apparent" diversity in each library, as a measure of true diversity we used the difference between the observed diversities after Cas9 induction and the diversity observed before induction in the same experiment. Separate experiments confirmed that mutation levels in non-induced samples remain steady in the course of our experimental times (**Supplementary Fig. 1c**).

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, NHG, or NGH (it was noticed that non-NGG PAMs, such as NHG and NGH, showed substantial 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).

For lineage tracing in cultured cell, first all hgRNA variants that were present in each subpopulation were determined. For that, the observed frequency of each gRNA variant was measured. Any variant with an observed frequency of at least 0.01% in a subpopulation was considered present in that subpopulation (setting this cutoff to as high as 1% and as low as 0.001% did not change downstream results), otherwise it was considered absent. As such, for each subpopulation, a binary vector was obtained with 1 for present hgRNA variants and 0 for absent ones. A binary distance matrix was constructed from these vectors (**Supplementary Table 2**). The vectors for the subpopulations at the same level were clustered hierarchically with a complete agglomeration strategy[30].

***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 h later, each well was transfected with 100 ng of plasmid a plasmid DNA packaged with 0.5 μL of Lipofectamine

doi:10.1038/nmeth.4108

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 h 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[9,16]. 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 RNase A and RNase H treatment. cDNA was circularized using CircLigaseII (Epicentre). Rolling circle amplification (RCA) was carried out with Phi29 polymerase using a target-specific primer (ggtggagcaattccacaacac) 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 20× magnification objective in the DAPI and Cy5 channels.

Data availability statement. The sequencing data that support the findings of this study are available in the Sequence Read Archive with the identifier SRA SRP092492. All data sets generated and analyzed during the current study are available from the corresponding authors on reasonable request. Source data for **Figures 2** and **3** are available online.

28. Yang, L. *et al.* Optimization of scarless human stem cell genome editing. *Nucleic Acids Res.* **41**, 9049–9061 (2013).
29. Schirmer, M. *et al.* Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **43**, e37 (2015).
30. Becker, R.A., Chambers, J.M. & Wilks, A.R. *The New S Language: A Programming Environment for Data Analysis and Graphics* (Chapman & Hall/CRC, 1988).