nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-019-0253-4

Recording transcriptional histories using Record-seq

Tanmay Tanna[1,4], Florian Schmidt[1,4], Mariia Y. Cherepkova[1], Michal Okoniewski[2] and Randall J. Platt[1,3*]

It is difficult to elucidate the transcriptional history of a cell using current experimental approaches, as they are destructive in nature and therefore describe only a moment in time. To overcome these limitations, we recently established Record-seq, a technology that enables transcriptional recording by CRISPR spacer acquisition from RNA. The recorded transcriptomes are recovered by SENECA, a method that selectively amplifies expanded CRISPR arrays, followed by deep sequencing. The resulting CRISPR spacers are aligned to the host genome, thereby enabling transcript quantification and associated analyses. Here, we describe the experimental procedures of the Record-seq workflow as well as subsequent data analysis. Beginning with the experimental design, Record-seq data can be obtained and analyzed within 1–2 weeks.

Introduction

Cellular processes are highly dynamic and complex by nature and are controlled by a plethora of intrinsic and environmental signals. To keep track of such diverse and complex processes in their endogenous environment without disrupting the cellular context remains challenging. Recent advancements in genome editing-based DNA writing technologies have transformed the biological sciences by enabling the encoding and continuous recording of information inside living cells[1–3]. These technologies are poised to overcome limitations in investigating dynamic biological events, such as reconstructing cellular lineage relationships, intra- and extracellular signaling cues, and gene expression dynamics. Motivated by the potential of capturing and recording dynamic biological events, several DNA writing strategies have been developed, including those based on single-stranded DNA (ssDNA) editing[4], recombinases[5,6], polymerases[7], and the CRISPR-associated protein Cas9 (refs. [8–11]) and DNA base editors derived therefrom[12]. More recently, CRISPR spacer acquisition mediated by Cas1 and Cas2 has received increasing attention as a potential DNA writing technology[13–17].

CRISPR spacer acquisition

The microbial adaptive immune system CRISPR–Cas equips bacteria and archaea with a molecular record of previously encountered hostile genetic elements. These molecular records are created by integration of short nucleic acid fragments (protospacers) derived from the RNA or DNA of the invading genetic element. The CRISPR-associated proteins Cas1 and Cas2 form a multimeric protein complex consisting of a central Cas2 dimer bound by a total of four molecules of Cas1 that binds to short fragments of invading nucleic acids and subsequently integrates them into a corresponding CRISPR array[18]. Individual spacers are separated from each other by sequences known as direct repeats. New spacers are preferentially integrated between the leader sequence of the CRISPR array and its first direct repeat.[19] Therefore, CRISPR arrays provide a chronological record of biological events.

CRISPR spacer acquisition from DNA

The type I-E Cas1–Cas2 genome integration complex from *Escherichia coli* continuously acquires spacers derived from double-stranded DNA when heterologously expressed inside living *E. coli* bacteria. This DNA recording and DNA writing property was recently leveraged for digital information storage, analog molecular recording, and lineage tracing[13,15,16]. In particular, digital

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland. [2]Scientific IT Services, ETH Zurich, Zurich, Switzerland. [3]Department of Chemistry, University of Basel, Basel, Switzerland. [4]These authors contributed equally: Tanmay Tanna, Florian Schmidt.
*e-mail: rplatt@ethz.ch

information storage of an image was achieved by converting the bit-based information encoding the position and intensity of each individual pixel within the image into short synthetic fragments of double-stranded DNA[16]. Subsequently, a library of DNA fragments encoding the entire image was synthesized and electroporated into a population of *E. coli* cells overexpressing Cas1 and Cas2. Upon entering the cells, the DNA fragments served as synthetic protospacers, were bound by the CRISPR genome integration complex, and integrated into the corresponding CRISPR array in the *E. coli* genome. Thus, synthetically encoded information of the original image could be recorded into the analog form of DNA, and safely stored in a population of living bacteria, from where it could be recovered by targeted amplification of CRISPR arrays followed by deep sequencing[16].

Analogously, a technology termed TRACE (temporal recording in arrays by CRISPR expansion) leveraged the same *E. coli* Cas1–Cas2 complex to enable analog molecular recording of extracellular stimuli and lineage reconstruction of populations of cells[13]. Analog molecular recording was achieved by placing the RepL protein under transcriptional control of inducible promoters. In the presence of inducing stimuli, RepL is expressed and initiates replication of a copy number-inducible trigger plasmid (pTrig), thereby increasing the copy number of pTrig within the cell and thus changing the intracellular DNA pool. Any DNA within the cell is a potential source of a protospacer, and the Cas1–Cas2 complex will acquire spacers according to abundance. Accordingly, the initial signal can be identified through an increase in spacers aligning to the corresponding trigger plasmid[13]. In addition, as the genomic CRISPR array is passed on to daughter cells, the analog molecular records can be used to reconstruct the lineages of populations of cells.

The CRISPR spacer acquisition from DNA systems discussed above both use the *E. coli* Cas1–Cas2 complex to record and write DNA. Although both are incredibly powerful for recording specific information defined a priori, especially in terms of digital information storage, they are limited in their capacity to record diverse and orthogonal biological events in a scalable way by the need to electroporate synthetic oligonucleotides or by the availability of inducible promoters responsive to a stimulus of interest.

CRISPR spacer acquisition from RNA

To overcome the limitations of current experimental approaches and to enable molecular recording and DNA writing on a massive scale we developed Record-seq[17], an approach that leverages CRISPR spacer acquisition from RNA. This was initially achieved using Cas1 and Cas2 orthologs derived from the type III CRISPR system of the human commensal bacterium *Fusicatenibacter saccharivorans*. The most important hallmark of the Cas1–Cas2 complex in this species is the naturally occurring fusion of Cas1 to a reverse transcriptase (RT-Cas1). Spacer acquisition from RNA presumably occurs through the binding of RT-Cas1–Cas2 to intracellular single-stranded RNA molecules and subsequent ligation of this RNA into the CRISPR array followed by reverse transcription[14].

Owing to the low spacer acquisition frequencies in RNA-adapting CRISPR systems, unlike their DNA adapting counterparts, RNA spacer acquisition had not been used for DNA writing applications before the development of Record-seq. Leveraging overexpression of *Fusicatenibacter saccharivorans* RT-Cas1–Cas2 (*Fs*RT-Cas1–Cas2) in an *E. coli* host and using a newly developed protocol for the selective PCR amplification of expanded CRISPR arrays, Record-seq overcomes this low efficiency limitation and thus enables the direct recording of intracellular RNAs. Using this technology, we demonstrated the recording of stimuli that transiently altered the cellular transcriptome but remained refractory to detection by traditional RNA sequencing (RNA-seq).

Development of the Record-seq protocol

Here we describe a detailed protocol for recording transcriptional events into plasmid-borne CRISPR arrays, named Record-seq. During bacterial growth, RNA-derived spacers are integrated into plasmid-encoded CRISPR arrays and reverse transcribed by the RT domain fused to Cas1, thereby achieving permanent storage of transcriptional events in the form of plasmid DNA (Fig. 1a). The Record-seq protocol entails extracting this plasmid DNA and retrieving the spacer sequences by means of selective amplification, size selection, and deep sequencing (Fig. 1b). As only a small fraction of CRISPR arrays acquire new spacers during the course of the experiment, Record-seq incorporates a newly developed procedure to selectively amplify newly acquired protospacers, named SENECA (selective amplification of expanded CRISPR arrays). Once newly acquired spacers are identified in deep sequencing reads, they are aligned to the reference genome, quantified, and provide a measure of cumulative transcript expression.

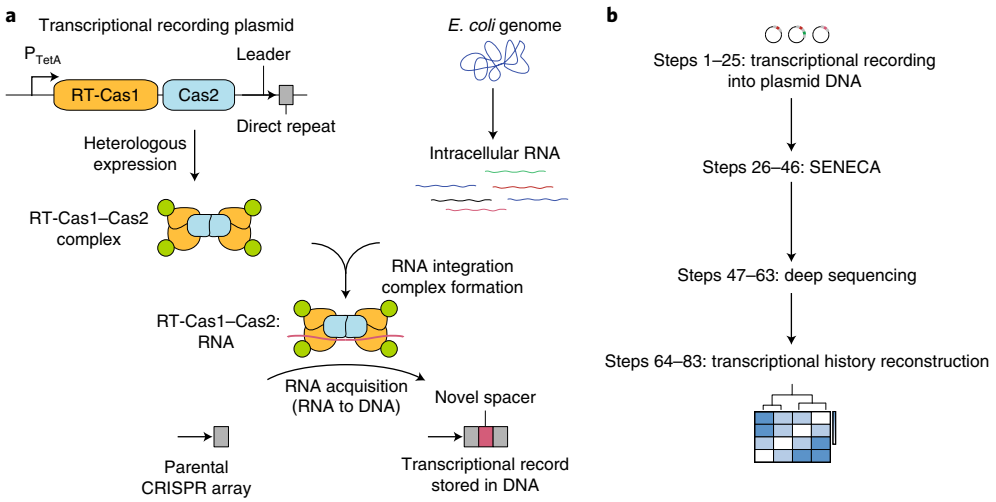


Fig. 1 | Transcriptional recording by CRISPR spacer acquisition from RNA. a, Record-seq uses the RNA-acquiring RT-Cas1–Cas2 complex from *Fusicatenibacter saccharivorans* to encode transcriptional information into plasmid-borne CRISPR arrays. The transcriptional record is generated by CRISPR spacer acquisition directly from intracellular RNAs followed by reverse transcription of RNA protospacers through the RT domain of *Fs*RT-Cas1–Cas2. **b**, Extraction of plasmid DNA followed by selective amplification of expanded CRISPR arrays (SENECA) and deep sequencing enables the reconstruction of transcriptional histories.

Advantages and limitations of Record-seq

Record-seq enables direct recording of transcriptional events into plasmid-borne CRISPR arrays at the transcriptome scale. In contrast to RNA-seq, which measures a snapshot in time, Record-seq encodes information about cumulative transcript expression into a permanent storage record in the form of DNA spacers contained within CRISPR arrays. This record is directly accessible by deep sequencing of plasmid-borne CRISPR arrays. As we have shown, Record-seq, but not RNA-seq, can detect transient exposure to extracellular stimuli such as the bacteriostatic herbicide paraquat[17]. Therefore, one main advantage of Record-seq is its ability to record transient transcriptional responses occurring over long periods of time. Furthermore, as the CRISPR arrays are passed on from mother to daughter cells, the transcriptional information reflects the same population of cells. This also implies that Record-seq is a population-based measurement of global transcription and thus does not account for variation or stochasticity of transcription among individual cells.

The *Fs*RT-Cas1–Cas2 complex, the core of the Record-seq technology, preferentially acquires spacers from abundant transcripts, with a preference for AT-rich regions. Thus, Record-seq enables an unbiased and parallelized quantification of cumulative transcript expression. Accordingly, Record-seq can be applied to study a wide spectrum of transcriptional responses. Unlike previous recording technologies based on the DNA-adapting Cas1–Cas2 complex of *E. coli*, the experimental design does not require the development of specific sensor strains carrying signal-inducible promoters[13].

In addition to being a powerful tool to study transcriptional responses, Record-seq implements SENECA, a new methodology to detect rare acquisition events. This in turn substantially decreases associated sequencing costs and increases the resolution at which spacer acquisition can be investigated, thereby facilitating mechanistic studies of CRISPR–Cas biology, and in particular, systems that are currently refractory to analysis owing to their low natural efficiency (Box 1).

Record-seq opens many opportunities in the fields of genomics and synthetic biology but the method still has some limitations. In particular, the current recording capacity is limited by the efficiency of CRISPR spacer acquisition by *Fs*RT-Cas1–Cas2. In the current state, we require approximately 1.2×10^9 cells per SENECA reaction to quantify a minimum of 10,000 uniquely occurring spacer sequences (see Box 2 for details on spacer quantification), which generates deep-sequencing readouts that encode transcriptomic information at a population scale and can be used for downstream analyses such as differential expression. Improving the efficiency of spacer acquisition could overcome this limitation and might eventually enable recordings from a substantially lower number of cells. Furthermore, Record-seq currently relies on genetic engineering of the CRISPR array to a minimal architecture, namely the leader sequence, a single direct repeat, and a FaqI restriction site downstream of the first direct repeat to be compatible with SENECA.

Box 1 | Use of SENECA for CRISPR biology

SENECA was developed as a custom protocol for the generation of deep sequencing libraries from Record-seq experiments. However, SENECA is generally useful for the detection of CRISPR spacer acquisition, independent from Record-seq, and thus also interesting for those interested in mechanistic investigations of CRISPR–Cas systems. The core of this methodology is the use of a minimal CRISPR array consisting merely of an upstream leader sequence and the first direct repeat. The functionality of such a minimal CRISPR array has been shown previously by our own and independent studies[20] of both RNA- and DNA-acquiring CRISPR–Cas systems. For this method, it is currently essential to include a recognition site for a type IIS restriction enzyme, such as FaqI, cloned immediately downstream of the direct repeat. In the future, the type IIS restriction enzyme approach could be replaced by suitable designer endonucleases to circumvent the need to engineer the CRISPR array a priori. An example of this could be the use of Cas9 with an sgRNA targeted to the direct repeat to truncate the CRISPR array. Thus, each direct repeat would be truncated by Cas9, leading to the detection of only individual spacers. Although this comes at the cost of not being able to investigate multiple acquisition events anymore, it would enable the direct investigation of CRISPR arrays in wild-type cells.

Following the SENECA protocol, plasmids carrying the minimal CRISPR array are extracted from Cas1–Cas2 expressing cells and then subjected to a golden gate assembly reaction consisting of FaqI, T7 DNA ligase and an adapter sequence to be ligated to the direct repeat of the CRISPR array. Owing to the cleavage pattern of FaqI, which removes the 3′ end of the direct repeat, this step facilitates subsequent use of a PCR primer binding exactly this 3′ cleavage site. Because expanded arrays duplicate the direct repeat upon acquiring a new spacer, the PCR primer can only bind expanded arrays. Accordingly, expanded arrays get exponentially amplified, while plasmids that do not acquire new spacers carry only a binding site for the reverse primer (binding the adapter ligated to the parental direct repeat) and thus undergo linear amplification. Using SENECA, the ratio of expanded over unexpanded arrays in deep sequencing libraries is enhanced substantially compared to classic acquisition readouts for which enrichment for expanded arrays is incomplete even after laborious size selection procedures by DNA gel electrophoresis.

Box 2 | Spacer extractor and Record-seq computation

A python script (*spacerExtractor.py*) is provided (Supplementary Data) to identify newly acquired spacers from deep sequencing reads. Deep sequencing reads obtained from Record-seq are preprocessed and converted to FASTA files, which serve as inputs to this script. The script uses fuzzy string matching to identify direct repeat sequences, and infers that any sequences interspacing direct repeats and falling within a user-specific length range are spacers. It also checks for uniqueness of the discovered spacer sequence, to remove PCR amplification or plasmid replication products of the same spacer, thereby reporting a conservative estimate of the diversity of acquired spacers.

Identified 'unique' spacer sequences are then aligned to the reference genome, which includes the *E. coli* host, the *Fs*RT-Cas1–Cas2 overexpression plasmid, and any other relevant sequence (e.g., MS2 phage). Another python script (*SErmdup.py*) is provided (Supplementary Data) to collapse duplicate alignments as a second filter for amplification products (this accounts for amplification and sequencing errors). This script removes duplicate alignments from a SAM alignment file based on both 5′ and 3′ mapping coordinates for single-end reads (as opposed to '*samtools markdup*' and '*picard markduplicates*', both of which mark duplicate alignments based only on 5′ mapping coordinates for single-end reads). Finally, the number of alignments for each annotated sequence element can be quantified, resulting in a counts matrix analogous to those used for RNA-seq analyses.

We have established Record-seq in multiple strains of *E. coli*, including BL21(DE3), BL21 AI and the K-12 derivative NovaBlue(DE3), thereby demonstrating the approach in the two main research strains of *E. coli*. Some users may wish to apply Record-seq in other bacterial species, and given the abundance of CRISPR systems throughout different bacteria, *Fs*RT-Cas1–Cas2 may work in these as well. However, some RT-Cas1–Cas2 orthologs seem to function only under very specific conditions or require additional bacterial host factors for their function. This is evident from the reported failure of *Marinomonas mediterranea* (MMB1) RT-Cas1–Cas2 to acquire RNA-derived spacers when expressed in an *E. coli* host[14]. We therefore suggest that RNA acquisition and transcriptional dependence of spacer acquisition are confirmed when using Record-seq in bacterial species other than *E. coli* as well as ensuring optimal conditions of spacer acquisition by optimizing the *Fs*RT-Cas1–Cas2 sequence to the codon usage of the intended host. Encouragingly, a screen of orthologous Cas1–Cas2 systems shows that many are functional and thus may provide a starting point[17]. Translating Record-seq to other cell types holds a great potential for a plethora of future applications.

Applications of Record-seq

Spacer acquisition by *Fs*RT-Cas1–Cas2 is a cumulative process that occurs throughout the growth of the bacterial population. The sum of spacers stored within the CRISPR arrays of an *E. coli* population represents its transcriptional history. Thus, Record-seq allows the detection of transient

transcriptional events such as short-term exposure to extracellular stimuli. Furthermore, Record-seq captures cellular RNAs at a transcriptome scale and thus allows for unbiased recording of transcriptional events. Unlike previously published technologies, the researcher does not need to define stimuli of interest a priori and genetically engineer the respective sensor strains detecting these small molecules. This is particularly advantageous when investigating biological phenomena for which optimal conditions or signature genes remain to be discovered; for example, the interaction between bacteria and pathogens in the open environment, and long-term exposure to toxins and chemicals for which genetic sensor systems are not readily available.

We demonstrated previously that Record-seq can recover transcriptional changes in response to extracellular stimuli that provoke a complex cellular response, such as a change in pH or oxidative stress[17]. Signature genes, detected by Record-seq, were consistent with our current knowledge of transcriptional response to these stimuli. In the future, implementing machine learning algorithms could help to develop Record-seq 'signature records' for classifying and deciphering cellular states.

In nature, bacteria form communities and are embedded in diverse environments. In our previous study, we demonstrated the detection of MS2 phage infection as an example of recording an environmental event. This indicates that Record-seq could be applied to study the interaction of bacterial communities in response to pathogen infection. Further increasing the efficiency of spacer acquisition by *Fs*RT-Cas1–Cas2 may also enable us to trace the lineages of subpopulations with bacterial communities. Taken together, Record-seq holds great potential to investigate direct mechanisms of bacterial colonization, persistence, horizontal gene transfer, and quorum sensing in many models, including biofilm formation and the mammalian microbiome.

As Record-seq enables recording, preserving, and reconstructing the transcriptional history of a population of cells, this technology could be employed in 'sentinel cells' embedded within contained (e.g., human gut) or open (e.g., ocean and air) environments for monitoring and tracking biological or environmental phenomena. Such applications hold untapped potential in both basic and applied research.

Experimental design

The Record-seq protocol is divided into three experimental sections: (i) growth of bacteria in the experimental setting of interest, (ii) SENECA and deep sequencing of acquired spacers, and (iii) reconstruction of transcriptional histories (Fig. 2a,b).

Growth of bacteria

Cultures derived from individual bacterial colonies of *E. coli* BL21(DE3) transformed with pFS_0453 are considered independent biological replicates, and we recommend that at least three biological replicates are performed for each treatment condition. Notably, we found an anhydrotetracycline (aTc) concentration of 100 ng/mL in TB medium achieved the highest number of spacers overall. However, under these conditions, many spacers align to the overexpressed *Fs*RT-Cas1–Cas2 sequence itself, while aTc concentrations of 30–50 ng/mL in TB medium proved to be optimal for obtaining high numbers of genome-aligning spacers. When using *Fs*RT-Cas1–Cas2 expressing cells under different culture conditions (i.e., different media or in microbial communities), we recommend establishing the optimal inducer concentration under these conditions to ensure maximum recording efficiency.

Some stimuli, such as chemicals and toxins, can have a substantial impact on protein expression and cell viability, which could also affect the activity of *Fs*RT-Cas1–Cas2. We therefore encourage users to always include a positive control cultured under the conditions described in this standard protocol. Furthermore, some cellular responses to toxic compounds such as paraquat can substantially inhibit *E. coli* proliferation. If the experimenter finds that this is the case for a stimulus of interest, we recommend first growing *E. coli* under *Fs*RT-Cas1–Cas2 expression (i.e., induction by anhydrotetracycline) for 12 h in compound-free media before diluting the culture 1:100 in media containing the stimulus and growing it for another 12 h.

SENECA

SENECA facilitates the specific enrichment and amplification of CRISPR arrays that acquired new spacers during bacterial growth. Owing to this highly sensitive procedure and the relative rarity of spacer acquisition, utmost care must be taken to avoid contamination. Our recommendation is to follow gold-standard deep sequencing procedures, including the strict separation of pre- and

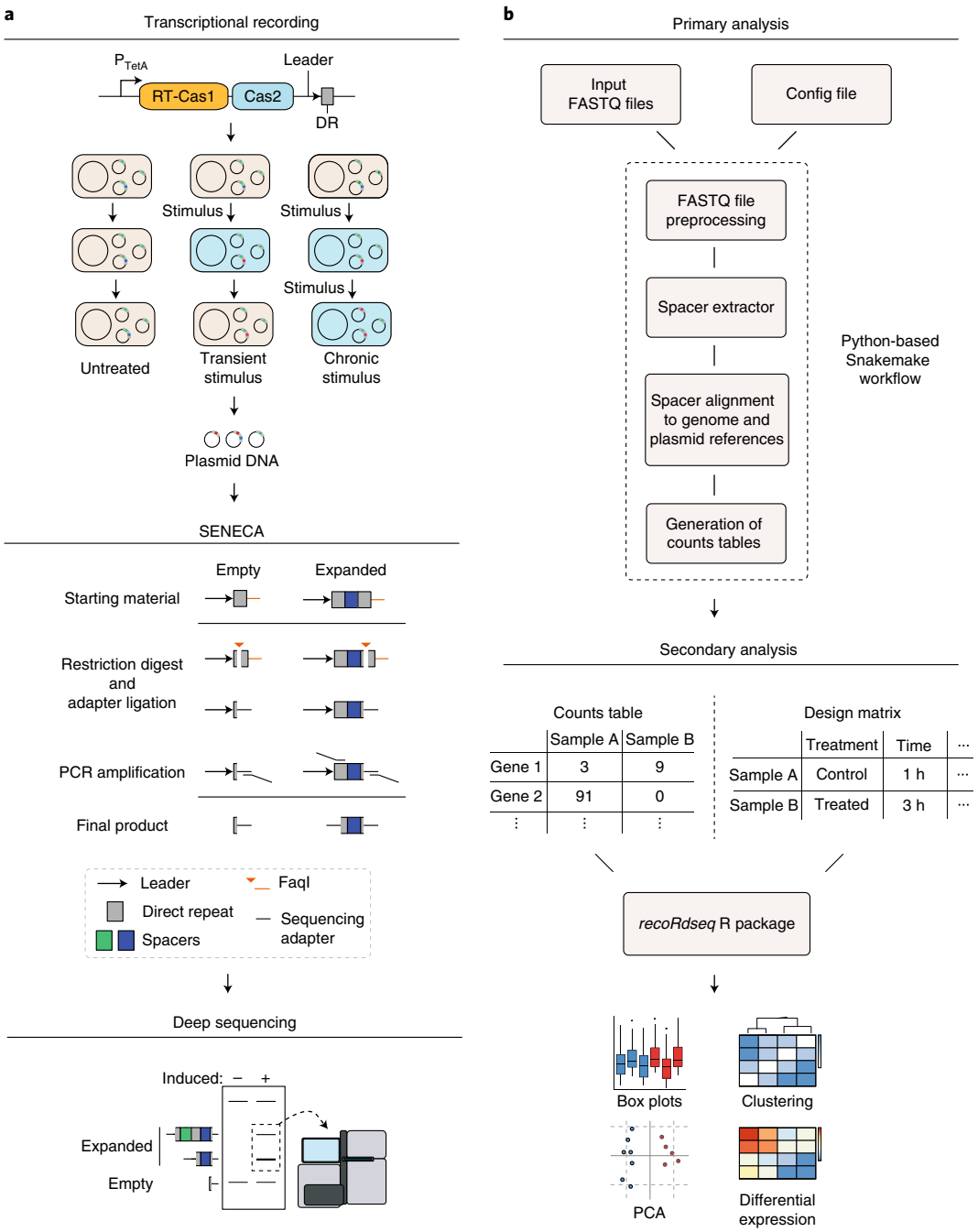


Fig. 2 | Record-seq experimental and computational workflows. a, Experimental workflow for the recording of spacers using *Fs*RT-Cas1–Cas2 followed by extraction of plasmid DNA and selective amplification of expanded CRISPR arrays (SENECA). During SENECA, a one-pot Golden Gate reaction truncates the parental direct repeat of the CRISPR array using the type IIS restriction enzyme FaqI and ligates a sequencing adapter with compatible overhang to this cleavage site. In the second stage, expanded CRISPR arrays are selectively amplified using a forward primer binding the full direct repeat generated during spacer acquisition but not the truncated direct repeat found in parental CRISPR arrays. After size-selection on standard agarose gels, Record-seq libraries are subjected to deep sequencing. **b**, The computational pipeline for processing Record-seq data consists of primary and secondary analysis. The primary analysis is implemented in a python-based Snakemake workflow to process deep sequencing reads, identify spacer sequences, align them to the reference genome and generate count matrices analogous to RNA-seq. The secondary analysis is implemented in R using the *recoRdseq* package, which uses these count matrices along with a design matrix provided by the user (containing information on experimental parameters) to generate descriptive plots, sample clustering and perform differential expression analyses.

post-PCR workspaces, cleaning workspaces with nucleic acid decontamination solutions regularly and before every experiment, and using filter tips. Furthermore, negative control reactions should be included at every point of the procedure (adapter ligation, first round PCR, bead purification, and second round PCR) and continued until the very end of the procedure. Finally, the optimized

SENECA approach described here, and not in the original publication[17], incorporates library barcodes that allow for the detection and computational removal of spacers carried over as contamination from previous library preparations.

Data analysis

The computational pipeline described here implements a two-stage analysis for the standard Record-seq readout (Fig. 2b). The primary analysis involves pre-processing of sequencing reads, identification and extraction of acquired spacer sequences from these reads, alignment of identified spacer sequences to the host reference genome as well as the *Fs*RT-Cas1–Cas2 overexpression plasmid sequence (using a sequence alignment tool), and counting of alignments to each feature in the reference genome and plasmid sequence to generate count matrices. All of these steps are integrated into a unified and scalable Python-based Snakemake workflow, which can be executed on the command line. The secondary analysis is implemented in a dedicated R package (*recoRdseq*), which includes functions for performing differential expression analyses using the count matrices generated by the first stage, sample clustering based on experimental groups, and creating descriptive plots.

Materials

Reagents

Plasmid backbones

- pFS_0453_pET30_pCat-tetR-Term-ptetA-*Fs*RT-Cas1(opt)-Cas2(opt)_T7term_ARRAY2_FaqI (Addgene, cat. no. 117006)
- pFS_0235-pET30-Fusicatenibacter_saccharivorans_ARRAY2_FaqI (Addgene, cat. no. 116954)

General reagents

- UltraPure 1 M Tris-HCl buffer (pH 8.0 and 7.5; Thermo Fisher Scientific, cat. nos. 15568025 and 15567027)
- UltraPure 0.5 M EDTA (pH 8.0; Thermo Fisher Scientific, cat. no. 15575020)
- RNase A (17,500 U; Qiagen, cat. no 19101)
- Sodium hydroxide (5 mol/L) (Merck, cat. no. 1370410503) **! CAUTION** This solution is hazardous when brought in contact with eyes, contact with skin, inhalation, and ingestion. Wear a lab coat, gloves, and goggles and avoid direct skin contact.
- Guanidine hydrochloride (Sigma-Aldrich, cat. no. 50950-2.5KG) **! CAUTION** Harmful if swallowed or inhaled. Hazardous when brought in contact with eyes or with skin. Wear gloves when handling guanidine hydrochloride.
- Potassium acetate (Sigma-Aldrich, cat. no. W292001-5K)
- 2-Propanol (Merck, cat. no. 1096341000)
- Ethanol (Honeywell, cat. no. 02860-2.5L)
- SDS (Sigma-Aldrich, cat. no. 75746-1KG) **! CAUTION** SDS is flammable, harmful if inhaled, causes skin irritation, serious eye damage, respiratory irritation, and is harmful to aquatic life with long-lasting effects. Wear safety goggles, gloves, and mask when preparing buffer P2.
- Glacial acetic acid (Merck, cat. no. 1000632500) **! CAUTION** Glacial acetic acid is highly corrosive to skin and eyes. It can damage internal organs upon ingestions or inhalation of vapors. Use gloves, goggles, and handle the liquid under a fume hood.
- Thymolphthalein (Sigma-Aldrich, cat. no. 114553-10G)
- Stericup filter unit (0.45 μm; Millipore, cat. no. SCHVU02RE)
- SDS solution (Sigma-Aldrich, cat. no. 71736-100ML)
- 2-Mercaptoethanol (Sigma-Aldrich, cat. no. M3148-100ML) **! CAUTION** 2-Mercaptoethanol is toxic if inhaled or swallowed, and irritating to the skin, eyes, and mucosae. Handle 2-mercaptoethanol under a fume hood wearing gloves.
- Formamide (Sigma-Aldrich, cat. no. F9037-100ML) **! CAUTION** Formamide is toxic and mutagenic; avoid contact with skin or eyes. Handle formamide under a fume hood wearing gloves and avoid inhaling fumes. Ensure that formamide waste is collected and disposed of according to institutional rules and regulations.

Record-seq plasmid amplification

- BD Difco dehydrated culture media (LB Broth, Miller; Thermo Fisher Scientific, cat. no. DF0446-07-5)
- Agar bacteriology grade (AppliChem, cat. no. A0949,5000)

- Kanamycin (50 mg/mL in Mili-Q H_2O, sterile filtered; AppliChem, cat. no. A1493.0025)
- 94 × 16 mm PS Petri dish (Greiner Bio-One, cat. no. 633102)
- One Shot Stbl3 chemically competent *E. coli* (Thermo Fisher Scientific, cat. no. C737303)
- SOC outgrowth medium (New England BioLabs, cat. no. B9020S)
- EconoSpin All-In-One Mini Spin Columns (Epoch Life Science, cat. no. 1910-050/250) or QIAprep Spin Miniprep Kit Print (Qiagen, cat. no. 27106)
- TE (pH 8.0; Thermo Fisher Scientific, cat. no. AM9858)

Recording experiment

- BL21-Gold (DE3) competent cells (Agilent Technologies, cat. no. 230132), other strains of B-strains of *E. coli* (deficient in endogenous Cas1–Cas2 proteins) can be used as a substitute
- Tryptone (ingredient) for microbiology (AppliChem, cat. no. A403682.1210)
- Glycerol anhydrous for molecular biology (AppliChem, cat. no. A2926,1000)
- Yeast extract (Sigma-Aldrich, cat. no. Y1625-1KG)
- Anhydrotetracycline hydrochloride (10 mg/mL in DMSO; Cayman Chemicals, cat. no. 10009542-10MG)
- Monopotassium phosphate (Sigma-Aldrich, cat. no. P0662-1KG)
- Dipotassium phosphate (Sigma-Aldrich, cat. no. P3786-1KG)
- Paraquat dichloride hydrate (1 mol/L; Sigma-Aldrich, cat. no. 36541-100MG) **! CAUTION** Paraquat dichloride hydrate is toxic when swallowed, hazardous when brought in contact with skin or eyes, fatal if inhaled, causes damage to organs through prolonged or repeated exposure, and is very toxic to aquatic life with long-lasting effects.

SENECA library preparation

- Quant-iT PicoGreen dsDNA Reagent (Thermo Scientific, cat. no. P7581) **▲ CRITICAL** Use PicoGreen-based quantification of plasmid DNA as opposed to quantification using a UV spectrophotometer (e.g., Nanodrop).
- Eppendorf twin.tec PCR plates LoBind (Eppendorf, cat. no. 0030129512)
- Microplate, 384 well PS, f-bottom, small volume, HiBase, med. binding, black (Greiner Bio-One cat. no. 784076)
- Plasmid DNA standard solution 50 ng/µL. Quantification of plasmid DNA by digital droplet PCR followed by generation of a plasmid DNA standard of defined concentration (50 ng/µL) is performed upon request by various companies, e.g., Eurofins GeneScan **▲ CRITICAL** The use of the linear λ-DNA standard provided with the Quant-iT PicoGreen dsDNA assay kit is not recommended, as PicoGreen shows different binding to circular and linear dsDNA.
- PCR plate (96 well; Eppendorf, cat. no. 0030128648)
- PCR sealing film (Eppendorf, cat. no. 0030127781)
- NEBuffer 2 (New England Biolabs, cat. no. B7002S)
- FastDigest FaqI (Thermo Fisher Scientific, cat. no. FD1814), **▲ CRITICAL** Use FastDigest FaqI, BsmFI (New England Biolabs) or FaqI (BsmFI) 2 U/µL (Thermo Fisher Scientific) recognizing the same sequence cleave less efficiently.
- T7 DNA ligase with 2× rapid ligation buffer (Enzymatics, cat. no. L6020L, or NEB, cat. no. M0318L) **▲ CRITICAL** Use T7 DNA ligase rather than T4 DNA.
- ATP solution (100 mM; Thermo Fisher Scientific, cat. no. R0441)
- DTT (1 M; Thermo Fisher Scientific, cat. no. P2325)
- Eppendorf twin.tec PCR plates (Sigma-Aldrich, cat. no. Z651451-25EA)
- NEBNext High Fidelity PCR Master Mix (2×; New England BioLabs, cat. no. M0541L) **▲ CRITICAL** Use of a high-fidelity polymerase is essential to minimize errors in amplification of high-diversity libraries for deep sequencing. In a comparison of KAPA HiFi (Roche), Phusion Flash (Thermo Fisher Scientific) and NEBNext High-Fidelity PCR Master Mix, we found the latter to work best for amplifying SENECA libraries.
- Agencourt AMPure XP (60 mL; Beckman Coulter, cat. no A63881)
- QIAquick PCR Purification Kit (Qiagen, cat. no. 28106)
- E-Gel 48 agarose gels (2%; Thermo Fisher Scientific, cat. no. G800802)
- E-Gel EX agarose gels (2%; Thermo Fisher Scientific, cat. no. G402002)
- GeneRuler low range DNA ladder (Thermo Fisher Scientific, cat. no. SM1191)
- QIAquick gel extraction Kit (Qiagen, cat. no. 28706)
- Sodium hypochlorite (Sigma-Aldrich, cat. no. 71696-2.5L) **! CAUTION** Causes severe skin burns and eye damage; wear safety goggles and gloves when preparing the 1% (vol/vol) sodium hypochlorite

solution. Do not spray 1% (vol/vol) sodium hypochlorite solution but apply using tissue paper to avoid the creation of an aerosol.

RNA-seq library preparation

- Sodium acetate solution (3 M, pH 5.2; Thermo Fisher Scientific, cat. no. R1181)
- TURBO DNA-free Kit (Thermo Fisher Scientific, cat. no. AM1907)
- Quant-iT RiboGreen RNA Assay Kit (Thermo Fisher Scientific, cat. no. R11490)
- Ribo-Zero rRNA Removal Kit (Bacteria; Illumina, cat. no. MRZB12424)
- Agencourt RNAClean XP (Beckman Coulter, cat. no. A66514)
- Agencourt AMPure XP (Beckman Coulter, cat. no. A63882)
- TruSeq Stranded mRNA Library Prep (96 samples; Illumina, cat. no. 20020595)
- Fragment Analyzer standard sensitivity NGS kit (AATI; Labgene, cat. no. DNF-473-0500)

Targeted deep sequencing

- KAPA Library Quantification Kit for Illumina platforms
- NextSeq 500/550 High Output Kit v2 (150 cycles; Illumina, cat. no. FC-404-2002)
- NextSeq 500/550 High Output Kit v2 (75 cycles; Illumina, cat. no. FC-404-2005)
- PhiX Control Kit v3 (Illumina, cat. no. FC-110-3001)
- Sodium hydroxide solution (10 N; Sigma-Aldrich, cat. no. 72068-100ML) **! CAUTION** Sodium hydroxide solution is hazardous when brought into contact with eyes or skin, or when inhaled or ingested. Wear appropriate personal protective clothing and avoid contact.

Software

- The computational methods described in this protocol have been implemented for a Unix-like operating system with a bash shell. The first stage of the computational pipeline is meant to be run in a terminal window. To follow these steps on a Microsoft Windows machine, we recommend using an environment capable of running Unix-like programs, such as Cygwin (http://www.cygwin.com/)
- Python > v3.0 for implementing the Snakemake pipeline, downloadable from https://www.python.org/downloads/
- The *fuzzysearch* python package for fuzzy string matching (https://pypi.org/project/fuzzysearch/)
- Snakemake > v5.3.0 for executing the computational workflow (https://snakemake.readthedocs.io/en/stable/getting_started/installation.html)[21]
- trimmomatic v0.36 for read trimming (http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.36.zip)[22]
- fastQC for quality control (https://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc)
- The FASTX-toolkit for converting FASTQ files to FASTA files (http://hannonlab.cshl.edu/fastx_toolkit/download.html)
- bowtie2 to align spacer sequences to reference genomes (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)[23]
- The samtools program (http://samtools.sourceforge.net/) for manipulation of alignment files[24]
- featureCounts from the Subread package (http://subread.sourceforge.net/) for counting of mapped spacers[25]
- R > 3.5.0 (http://www.r-project.org/)
- (Optional) RStudio (https://www.rstudio.com/products/rstudio/download/)
- The Snakemake workflow for the primary analysis and associated scripts (provided in Supplementary Data) are downloadable from https://github.com/plattlab/Transcriptional-Recording
- The *recoRdseq* R package and dependencies, including *DESeq2*, *edgeR*, *baySeq*, *ggplot2*, *VennDiagram*, *pheatmap*, *ggfortify*, *cluster*, *readxl*, *reshape2*, *rgl*, *gplots*, *RColorBrewer* (downloadable from https://github.com/plattlab/Transcriptional-Recording)
- (Optional) If the user is downloading example deep sequencing reads from the Sequence Read Archive (SRA), the SRA toolkit is recommended (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software)[26]

Annotation files and dataset

- *E. coli* BL21-Gold(DE3)pLysS AG reference genome FASTA sequence and GFF3 annotation file, downloadable from Ensembl (http://bacteria.ensembl.org/Escherichia_coli_bl21_gold_de3_plyss_ag_/Info/Index)

- *Fs*RT-Cas1–Cas2 overexpression plasmid FASTA sequence and GFF3 annotation files (Supplementary Data: 'pFS_0235.fasta', 'pFS_0235.gff3', 'pFS_0453.fasta', 'pFS_0453.gff3')
- (Optional) Example deep sequencing data are accessible on the SRA (PRJNA510019), at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA510019
- (Optional) If the user wishes to skip the primary analysis for the example data, an example counts dataset is accessible at https://github.com/plattlab/Transcriptional-Recording/tree/master/example-data

Equipment

- 1.5 and 2.0 mL Microcentrifuge tubes (Eppendorf, cat. nos. 0030120086 and 0030120094)
- Bacterial culture tubes (Sigma-Aldrich, cat. no. Z642878)
- 1.5 mL DNA LoBind tubes (Eppendorf, cat. no. 0030108051)
- Falcon tubes, polypropylene, 15 and 50 mL (Corning, cat. nos. 352097 and 352070)
- Filtered sterile pipette tips
- Glass bottle (2 L)
- Tecan Infinite M1000 PRO (Tecan, cat. no. 30063849)
- Thermocycler with programmable temperature stepping functionality (96-well; Eppendorf Mastercycler gradient pro, cat. no. 6321 000.515)
- Benchtop microcentrifuges (Eppendorf, cat. nos. 5424 and 5804)
- Incubator for bacterial plates (Memmert, cat. no. IN55)
- Bacterial shaker (New Brunswick Innova 44/44R, cat. no. M1282-0002)
- Vacuum manifold (QIAvac 24 Plus, cat. no. 19413)
- Vortex-Genie 2 (Scientific Industries, cat. no. SI-0236)
- Magnetic separator for 96-well PCR plate (DynaMag-96 Side Skirted Magnet; Thermo Fisher Scientific, cat. no. 12027)
- Eppendorf ThermoMixer C (Eppendorf, cat. no. 5382000015) and 1.5 mL SmartBlock (Eppendorf, cat. no. 5360000038)
- Digital gel imaging system (GelDoc EZ; Bio-Rad, cat. no. 170-8270) and blue sample tray (Bio-Rad, cat. no. 170-8273), and UV sample tray (Bio-Rad, cat. no. 1708271)
- Blue-light transilluminator and orange filter goggles (SafeImager 2.0; Invitrogen, cat. no. G6600)
- UV spectrophotometer (NanoDrop OneC; Thermo Fisher Scientific)
- TapeStation 4200 (Agilent Technologies, cat. no. G2991AA)
- qPCR (LightCycler 96 Instrument; Roche, cat. no. 05815916001)
- 10 mM Tris-HCl with 0.1% Tween 20 (Teknova, cat. no. T7724)
- NextSeq 500/550 System (Illumina, cat. nos. SY-415-1001 and SY-415-1002)

Reagent setup

LyseBlue solution

Dissolve 2.15 g thymolphthalein in 50 mL ethanol absolute. Store at room temperature (18–24 °C) for up to 6 months.

Buffer P1 (500 mL)

Mix 25 mL 1 M Tris-HCl (pH 8) and 10 mL 0.5 M EDTA, fill up to 500 mL using MiliQ-H_2O. Filter through 0.45 µm filter into a 500 mL glass bottle. Store at room temperature for up to 1 year. Before use, prepare an aliquot of 50 mL and add 50 µL of 100 µg/mL RNAse A and 50 µL of LyseBlue solution. ▲ **CRITICAL** Upon addition of RNAse A and LyseBlue solution, store buffer P1 at 4 °C for a maximum of 4 months.

Buffer P2 (500 mL)

Dissolve 5 g SDS pellets in 1 L MiliQ-H_2O, add 20 mL of 5 M NaOH and fill up to 1,000 mL using MiliQ-H_2O. Stir for 15 min using a magnetic stirrer to dissolve SDS. Filter through 0.45 µm filter into a 500 mL glass bottle. Store at room temperature for up to 1 year.

Buffer N3 (500 mL)

Dissolve 200.61 g of guanidine hydrochloride and 44.16 g potassium acetate in 150 mL MiliQ-H_2O. Fill up to 450 mL using MiliQ-H_2O. Adjust pH to 4.8 using glacial acetic acid (~20 mL). Fill up to 500 mL using MiliQ-H_2O and filter through 0.45 µm filter into a 500 mL glass bottle. Store at room temperature for up to 1 year.

Buffer PB (500 mL)
Dissolve 238.83 g of guanidine hydrochloride in 150 mL MiliQ-H_2O and add 150 mL 2-propanol. Fill up to 500 mL using MiliQ-H_2O and filter through 0.45 µm filter into a 500 mL glass bottle. Store at room temperature for up to 1 year.

Buffer PE (500 mL)
Mix 5 mL of TRIS-HCl (pH 7.5) and 95 mL of MiliQ-H_2O. Fill up to 500 mL using absolute ethanol and filter through 0.45 µm filter into a 500 mL glass bottle. Store at room temperature for up to 1 year.

RNASnap solution (10 mL)
Mix 360 µL 0.5 M EDTA (pH 8.0), 25 µL 10% SDS, 100 µL 2-mercaptoethanol, 15 µL Ultrapure water, and 9.5 mL formamide. ▲ **CRITICAL** Prepare RNASnap solution freshly before use.

Procedure

Acquisition procedure; transformation of recording plasmid into expression strain ● Timing 40 min, 5 min hands-on

1 Thaw one aliquot of BL21-Gold (DE3) chemically competent cells by placing on ice. Wait until the aliquot is fully thawed, then transfer 25 µL of competent cells per transformation into a fresh, pre-chilled 1.5 mL tube.
2 Add 30 ng of pFS_0453 plasmid DNA in 0.5 µL to the tube containing the competent cells, gently flick the tube and place it back on ice for 3 min.
3 Heat shock the cells by holding the tube in a water bath heated to 42 °C for 30 s. Place the tube back on ice for 1 min.
4 Add 120 µL of SOC media to the tube and place it in a thermomixer at 37 °C and 800 r.p.m. for 30 min to enable cell recovery and expression of the antibiotic resistance gene.
5 Pre-warm 2 LB agar plates (10 cm, kanamycin) per transformation at 37 °C.
6 After 30 min of recovery, mix the cells by inverting and plate 70 µL of cell suspension on each of the two agar plates.
7 Incubate the LB agar plates overnight at 37 °C for 12–15 h.

Acquisition procedure; recording of transcriptional histories ● Timing: 1 d, 90 min hands-on

8 The following morning, take the LB agar plates out of the incubator and put them at 4 °C or room temperature until the evening.
■ PAUSE POINT LB agar plates with colonies of BL21-Gold (DE3) containing pFS_0453 can be kept at 4 °C for up to 1 week before starting the recording experiment.
9 For each replicate of the recording experiment, add 3 mL of TB media containing kanamycin (50 µg/mL) and aTc (30 ng/mL) into a bacterial culture tube and inoculate each tube with two bacterial colonies from the plate in Step 8. Perform at least three biological replicates per stimulus condition. Stimuli can be applied from the start of the experiment (i.e., culturing of *E. coli* cells in different media) or later during the experiment. Transfer tubes to a bacterial shaker set to 37 °C, 250–300 r.p.m.
10 Incubate cultures for 14–24 h. Sufficient density for plasmid preparation is usually achieved 14 h after inoculation of colonies for the untreated condition. For stimuli that have an adverse effect on cell growth, incubation time may be increased to up to 24 h to reach the same density. For some toxic stimuli, such as the exposure to paraquat, we recommend first growing *E. coli* under FsRT-Cas1–Cas2 expression before diluting the culture in media containing the compound (see 'Experimental design' section).
▲ CRITICAL STEP A density that is too low or high will affect the amount of plasmid DNA isolated from the cells. Too few input cells result in low yields. Too high amounts of bacteria will lead to insufficient lysis during plasmid preparation and thus poor yield and DNA quality. In case the desired treatment is observed to drastically affect cell growth when starting immediately from colonies inoculated into aTc-containing media, it is possible to grow a seed culture overnight in kanamycin containing media in the absence of aTc. The next morning this seed culture can then be diluted to experimental cultures containing aTc.
11 If intending to prepare RNA-seq libraries in parallel with Record-seq libraries (i.e., to compare transcriptional histories obtained by Record-seq with RNA-seq snapshots at defined timepoints),

transfer 100 µL from each liquid culture into a 1.5 mL microcentrifuge tube, pellet bacteria by spinning at 21,000*g* for 1 min at 4 °C, aspirate and discard supernatant and immediately freeze the cell pellet by submerging the closed tube in liquid nitrogen. Frozen cell pellets can be stored at −80 °C for several weeks.

12 Transfer 2 mL from each liquid culture into a 2.0 mL microcentrifuge tube, pellet bacteria by spinning at 21,000*g* for 3 min at room temperature. Plasmid DNA can be isolated using commercially available kits such as QIAprep Spin Miniprep Kit using the volumes of buffers P1, P2, and N3 suggested below or using homemade buffers P1, P2, and N3 along with separately purchased columns from other vendors.

13 Discard the supernatants and resuspend each bacterial pellet in 300 µL of buffer P1.

14 Lyse the bacterial cells by adding 300 µL of buffer P2, gently invert the tubes 5–10 times to achieve an equal mixture.

▲ **CRITICAL STEP** The solution should turn uniformly blue without visible solid pieces of bacteria. Solid pieces or a gooey mixture are indicators of insufficient lysis and will result in poor plasmid yield.

15 Neutralize the mixture by adding 420 µL of buffer N3, again gently invert the tubes until the mixture turns completely colorless.

16 Pellet bacterial debris by centrifugation at 21,000*g* for 12 min at room temperature in a microcentrifuge.

17 Apply the supernatant of each sample to a plasmid mini-prep column using the vacuum manifold. Once the supernatant has passed through the column (supernatant flow-through is discarded), wash each column, first with 1 mL of buffer PB followed by 1 mL of buffer PE (wash flow-through is discarded).

▲ **CRITICAL STEP** Alternatively, all column steps can be performed using a microcentrifuge with centrifugation steps carried out at 21, 000*g* for 1 min at room temperature. Apply a maximum of 800 µL solution per column during each spin. Perform multiple spins, if necessary, to achieve the volumes described in the protocol.

18 Remove the columns from the manifold and centrifuge them at 21, 000*g* for 1 min at room temperature to remove residual wash buffer.

19 Place each column into a fresh 1.5 mL microcentrifuge tube and apply 50 µL of buffer EB prewarmed to 55 °C to the center of the membrane. Transfer the tubes with the columns into a thermomixer and incubate them at 55 °C, 800 r.p.m. for 3 min. This increases the yield of plasmid DNA compared with immediate elution with buffer EB at room temperature.

20 Elute the DNA from the columns by centrifuging the tubes at 16, 000*g* for 1 min at room temperature. Remove the column and transfer the eluted plasmid DNA samples into the individual wells of a 96-well PCR plate.

? TROUBLESHOOTING

■ **PAUSE POINT** plasmid DNA can be stored frozen at −20 °C for at least 6 months.

Preparation of Record-seq libraries for deep sequencing; quantification of plasmid DNA using PicoGreen ● Timing: 1 h hands-on

21 Prepare sample dilution by mixing 1 µL of plasmid DNA from Step 20 with 149 µL of TE buffer in a 96-well DNA LoBind PCR plate, vortex to mix, then centrifuge the plate at 2,200*g* for 1 min at room temperature.

22 Perform a serial dilution of the plasmid DNA standard (50 ng/µL) as outlined below.

Standard	Amount of previous standard (µL)	Amount of TE buffer (µL)	Final concentration of DNA (ng/µL)
Standard 0	—	—	50
Standard 1	10	240	2
Standard 2	100	100	1
Standard 3	100	100	0.5
Standard 4	100	150	0.2
Standard 5	100	100	0.1
Standard 6	100	100	0.05
Standard 7	25	100	0.01
Standard 8 (blank)	0	100	0

23 Prepare PicoGreen solution by diluting PicoGreen stock 1:200 in buffer TE. Each measurement requires 5 µL of PicoGreen solution. Prepare sufficient PicoGreen solution to measure each standard in triplicate and each sample in duplicate.

24 Transfer 3 × 5 µL of DNA standard solution for each concentration of the standard curve and 2 × 5 µL of sample dilution for each sample into a black 384-well microplate.

25 Quickly add 5 µL of PicoGreen solution to each well containing DNA solution. Seal 384 wells using a PCR sealing film, vortex briefly and centrifuge at 800*g* for 1 min at room temperature. Then incubate the assay plate for 5 min at room temperature before measuring the fluorescence using a Tecan Infinite M1000 PRO microplate reader (temperature, 25 °C; excitation wavelength, 480 nm; emission wavelength, 520 nm; excitation bandwidth, 9 nm; emission bandwidth, 20 nm; mode, top; gain, optimal; flash frequency, 400 Hz; number of flashes, 50; settle time, 0 ms; lag time, 0 ms).

Preparation of Record-seq libraries for deep sequencing; SENECA ● Timing: 2 d, 3 h hands-on

26 Anneal oligonucleotides containing the Illumina adapters with matching library barcodes (listed in Table 1) by mixing as follows. Here we suggest eight library barcodes (corresponding to SENECA_Lib._BC_1_fw/rv to SENECA_Lib._BC_8_fw/rv). Use an individual library barcode for each SENECA library prepared in your own laboratory.

Component	Amount per reaction (µL)	Final concentration
NEBuffer 2, 10×	5	1×
SENECA_Lib._BC_X_fw	2.5	5 µM
SENECA_Lib._BC_X_rv	2.5	5 µM
UltraPure water	40	—
Total volume	50	—

Anneal the oligonucleotides in a thermocycler, by first heating to 95 °C for 5 min followed by cooling down to 12 °C at a ramp rate of 0.12 °C/s. Dilute the annealed oligonucleotides 1:100 by mixing 10 µL of annealed oligonucleotides with 990 µL of buffer TE.

27 *Golden Gate assembly of DNA oligonucleotides from Step 26 to plasmid DNA from recording experiment*. We recommend first dispensing 40 fmol of plasmid DNA from Step 20 in 12.5 µL of UltraPure water for each sample, and then preparing a master mix of all other components below. 40 fmol of plasmid DNA is equal to 168 ng of pFS_0453.

Component	Amount per reaction (µL)	Final concentration
Plasmid DNA from recording experiment (Step 20)	40 fmol	2 fmol/µL
Annealed, diluted oligonucleotides from Step 26	1	—
ATP/DTT, 10 mM	2	1 mM
T7 DNA ligase	0.5	—
FastDigest FaqI	1	—
SAM, 20×	1	1×
FastDigest Buffer, 10×	2	1×
UltraPure water	Fill up to 20	—
Total volume	20	—

28 Perform a Golden Gate assembly reaction using the following cycling conditions:

Cycle number	Condition
1–99	37 °C for 3 min, 20 °C for 3 min
100	55 °C for 15 min
Hold	12 °C

■ **PAUSE POINT** Adapter-ligated plasmid DNA can be stored frozen at −20 °C for at least 1 month.

Table 1 | Oligonucleotide sequences for SENECA

Oligonucleotide	Sequence (5′–3′)	Purpose	Procedure step
SENECA_Lib._BC_1_fw	AAAGATTGAAAGCCATAAGCGCTTGATCGGAAGAGCACACGTCTGAACTCCAGTCAC	Adapter ligation	26
SENECA_ Lib._BC_1_rv	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAAGCGCTTATGGCTTTCAAT	Adapter ligation	26
SENECA_ Lib._BC_2_fw	AAAGCACTTTGGTTATAGAAGAGGGATCGGAAGAGCACACGTCTGAACTCCAGTCAC	Adapter ligation	26
SENECA_Lib._BC_2_rv	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCCTCTTCTATAACCAAAGTG	Adapter ligation	26
SENECA_ Lib._BC_3_fw	AAAGCCAAATCTTCCACTTGCAAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC	Adapter ligation	26
SENECA_ Lib._BC_3_rv	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGCAAGTGGAAGATTTGG	Adapter ligation	26
SENECA_ _Lib._BC_4_fw	AAAGTCCCATGAATGTTCCACATGATCGGAAGAGCACACGTCTGAACTCCAGTCAC	Adapter ligation	26
SENECA_ Lib._BC_4_rv	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATGTGGAACATTCATGGGA	Adapter ligation	26
SENECA_ Lib._BC_5_fw	AAAGTTCTTGCCATCAATCTCTTCGATCGGAAGAGCACACGTCTGAACTCCAGTCAC	Adapter ligation	26
SENECA_ _Lib._BC_5_rv	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAAGAGATTGATGGCAAGAA	Adapter ligation	26
SENECA_ Lib._BC_6_fw	AAAGTTTCAGGGAGTTATTCAGTAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC	Adapter ligation	26
SENECA_ Lib._BC_6_rv	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACTGAATAACTCCCTGAAA	Adapter ligation	26
SENECA_ Lib._BC_7_fw	AAAGAAAGCTATCCACGAATGTGATCGGAAGAGCACACGTCTGAACTCCAGTCAC	Adapter ligation	26
SENECA_ Lib._BC_7_rv	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCACATTCGTGGATAGCTTT	Adapter ligation	26
SENECA_ Lib._BC_8_fw	AAAGATTAGCTCTAGGAAACACAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC	Adapter ligation	26
SENECA_ _Lib._BC_8_rv	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTGTTTCCTAGAGCTAAT	Adapter ligation	26
SENECA_1st_rd_PCR_1_fw	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNCCTAAAAGGAATTGAAAC	SENECA first round PCR	29
SENECA_1st_rd_PCR_2_fw	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNCCTAAAAGGAATTGAAAC	SENECA first round PCR	29
SENECA_1st_rd_PCR_3_fw	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCCTAAAAGGAATTGAAAC	SENECA first round PCR	29
SENECA_1st_rd_PCR_4_fw	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNCCTAAAAGGAATTGAAAC	SENECA first round PCR	29
SENECA_1st_rd_PCR_5_fw	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNCCTAAAAGGAATTGAAAC	SENECA first round PCR	29
SENECA_1st_rd_PCR_6_fw	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNCCTAAAAGGAATTGAAAC	SENECA first round PCR	29
SENECA_1st_rd_PCR_7_fw	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNCCTAAAAGGAATTGAAAC	SENECA first round PCR	29
SENECA_1st_rd_PCR_rv	GTGACTGGAGTTCAGACG	SENECA first round PCR	29
SENECA_2nd_rd_PCR_BC_1_fw	AATGATACGGCGACCACCGAGATCTACACAAGTAGAGACACTCTTTCCCTACACGACGCTCTTCCGATCT	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_2_fw	AATGATACGGCGACCACCGAGATCTACACCATGCTTAACACTCTTTCCCTACACGACGCTCTTCCGATCT	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_3_fw	AATGATACGGCGACCACCGAGATCTACACGCACATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_4_fw	AATGATACGGCGACCACCGAGATCTACACTGCTCGACACACTCTTTCCCTACACGACGCTCTTCCGATCT	SENECA second round PCR	40

Table continued

Table 1 (continued)

Oligonucleotide	Sequence (5′–3′)	Purpose	Procedure step
SENECA_2nd_rd_PCR_BC_5_fw	AATGATACGGCGACCACCGAGATCTACACAGCAATTCACACTCTTTCCCTACACGACGCTCTTCCGATCT	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_6_fw	AATGATACGGCGACCACCGAGATCTACACAGTTGCTTACACTCTTTCCCTACACGACGCTCTTCCGATCT	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_7_fw	AATGATACGGCGACCACCGAGATCTACACCCAGTTAGACACTCTTTCCCTACACGACGCTCTTCCGATCT	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_8_fw	AATGATACGGCGACCACCGAGATCTACACTTGAGCCTACACTCTTTCCCTACACGACGCTCTTCCGATCT	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_9_fw	AATGATACGGCGACCACCGAGATCTACACACACGATCACACTCTTTCCCTACACGACGCTCTTCCGATCT	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_10_fw	AATGATACGGCGACCACCGAGATCTACACGGTCCAGAACACTCTTTCCCTACACGACGCTCTTCCGATCT	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_11_fw	AATGATACGGCGACCACCGAGATCTACACGTATAACAACACTCTTTCCCTACACGACGCTCTTCCGATCT	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_12_fw	AATGATACGGCGACCACCGAGATCTACACTTCGCTGAACACTCTTTCCCTACACGACGCTCTTCCGATCT	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_1_rv	CAAGCAGAAGACGGCATACGAGATCATGATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_2_rv	CAAGCAGAAGACGGCATACGAGATAGGATCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_3_rv	CAAGCAGAAGACGGCATACGAGATGACAGTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_4_rv	CAAGCAGAAGACGGCATACGAGATCCTATGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_5_rv	CAAGCAGAAGACGGCATACGAGATTCGCCTTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_6_rv	CAAGCAGAAGACGGCATACGAGATATAGCGTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_7_rv	CAAGCAGAAGACGGCATACGAGATGAAGAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_8_rv	CAAGCAGAAGACGGCATACGAGATATTCTAGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_9_rv	CAAGCAGAAGACGGCATACGAGATCGTTACCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_10_rv	CAAGCAGAAGACGGCATACGAGATGTCTGATGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_11_rv	CAAGCAGAAGACGGCATACGAGATTTACGCACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	SENECA second round PCR	40
SENECA_2nd_rd_PCR_BC_12_rv	CAAGCAGAAGACGGCATACGAGATTTGAATAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	SENECA second round PCR	40

29 Preparation of first round PCR primer mix is as follows:

Component	Amount (µL)	Final concentration (µM)
SENECA_1st_rd_PCR_1_fw, 100 µM	14.29	1.43
SENECA_1st_rd_PCR_2_fw, 100 µM	14.29	1.43
SENECA_1st_rd_PCR_3_fw, 100 µM	14.29	1.43
SENECA_1st_rd_PCR_4_fw, 100 µM	14.29	1.43
SENECA_1st_rd_PCR_5_fw, 100 µM	14.29	1.43
SENECA_1st_rd_PCR_6_fw, 100 µM	14.29	1.43
SENECA_1st_rd_PCR_7_fw, 100 µM	14.29	1.43
SENECA_1st_rd_PCR_rv, 100 µM	100	10
TE buffer	800	—
Total volume	1,000	—

Prepare aliquots of 100 µL and store frozen at −20 °C, which can last for at least 1 year.

30 *First round PCR amplification*. This step specifically amplifies expanded CRISPR arrays.

We recommend preparing a PCR master mix as follows, dispensing 19.62 µL into each sample well of a 96-well PCR plate and adding 10.38 µL of adapter-ligated template DNA from Step 28 in the end.

Component	Amount per reaction (µL)	Final concentration
NEBNext High Fidelity PCR Master Mix, 2×	15	1×
First round PCR primer mix from Step 29, 10 µM	1.5	500 nM
UltraPure water	3.12	—
Adapter ligated template DNA from Step 28	10.38	—
Total volume	30	—

31 Perform a PCR reaction using the following cycling conditions.

Cycle number	Denature	Anneal	Extend
1	98 °C, 30 s	—	—
2–23	98 °C, 10 s	57 °C, 30 s	72 °C, 20 s
24	—	—	72 °C, 5 min

■ **PAUSE POINT** First round PCR product can be stored frozen at −20 °C for at least 2 months.

32 *Purification of PCR products using AMPure XP beads*. Add 36 µL of AMPure XP beads per sample well of the PCR plate and mix by pipetting up and down 10 times.

33 Place the PCR plate on the magnetic separator and allow the beads to migrate to the magnet (2 min).

34 Keeping the PCR plate on the magnetic separator, aspirate, and discard the supernatant.

35 Freshly prepare 80% ethanol by mixing 40 mL ethanol absolute and 10 ml UltraPure water in 50 mL conical tube. Keeping the PCR plate on the magnetic separator, add 150 µL of 80% ethanol to each sample well of the PCR plate and incubate for 30 s.

36 Aspirate and discard the 80% ethanol, allow the beads to air-dry for 6–8 min.

▲ **CRITICAL STEP** Make sure all ethanol is removed. Residual ethanol can inhibit subsequent PCR reactions.

37 Add 25 µL of TE buffer to each sample well of the PCR plate to elute the DNA from the beads, close the plate using a PCR sealing foil and vortex it until the beads in each well are homogenously resuspended in the buffer.

38 Spin the PCR plate for 20 s at 500*g* at room temperature to collect the liquid at the bottom of each well.

39 Place the PCR plate back on the magnetic separator and allow the beads to migrate to the magnet (3 min), then transfer the supernatant to a fresh 96-well PCR plate.

▲ **CRITICAL STEP** Avoid carry-over of beads after elution of DNA. Residual AMPure XP beads can inhibit subsequent PCR reactions.

40 *Second round PCR amplification*. This step amplifies the individual samples appending a unique, sample-specific barcode for subsequent deep sequencing. Prepare the reaction as follows:

Component	Amount per reaction (µL)	Final concentration
NEBNext High Fidelity PCR Master Mix, 2×	10	1×
SENECA_2nd_rd_PCR_BC_X_fw (4 µM)	2.5	0.5 µM
SENECA_2nd_rd_PCR_BC_Y_rv (4 µM)	2.5	0.5 µM
First round PCR product from Step 39	5	—
Total volume	20	—

41 Perform PCR reaction using the following cycling conditions

Cycle number	Denature	Anneal	Extend
1	98 °C, 30 s	—	—
2–9	98 °C, 10 s	65 °C, 30 s	72 °C, 20 s
10	—	—	72 °C, 5 min

42 Dilute 3 µL of second round PCR product into a fresh 96-well PCR plate, containing 17 µL of UltraPure water.

43 Load 20 µL of the diluted PCR product from Step 42 on an E-Gel 48 agarose gel 2%. Use GeneRuler low range DNA ladder as a size marker and run the gel for 12 min. Validate the correct size of the PCR product (252 bp for a single acquired spacer).

44 Pool all the PCR reactions from Step 41 and purify the PCR product using the QIAquick PCR Purification Kit according to the manufacturer's directions. Elute the DNA from each pool in 20–40 Buffer EB and use a NanoDrop UV spectrophotometer to quantify the product. Dilute the product to 25 ng/µL using buffer EB.

▲ **CRITICAL STEP** If extreme differences in the intensity of PCR product bands were observed in Step 43, preparing multiple pools by binning the samples by similar band intensity and purifying them separately is an option to ensure equal representation in the final sequencing library.

45 Run the pools of PCR-purified and diluted amplicons from Step 44 on an E-Gel EX agarose gel, 2% along with GeneRuler low range DNA ladder. Load 20 µL of sample per well and run the gel for 13 min.

? TROUBLESHOOTING

46 Gel-extract the PCR products corresponding to expanded CRISPR arrays (236–302 bp) using the QIAQuick Gel Extraction Kit eluting in 25 µL into a 1.5 mL DNA LoBind tube according to the manufacturer's directions and quantify the final product using a Nanodrop UV spectrophotometer.

PCR product size (bp)	Amplicon species
170	Unexpanded CRISPR array
236	Single acquisition event
302	Double acquisition event

Deep sequencing of Record-seq libraries ● Timing: 1 d, 3 h hands-on

47 Determine the concentration of each individual library from Step 46 using the KAPA Library Quantification Kit for Illumina platforms according to the manufacturer's protocol.

48 Using the individual libraries quantified by qPCR in Step 47, prepare a final sequencing pool with a concentration of 2 nM using 10 mM Tris-HCl with 0.1% Tween 20 as a diluent. During this step, normalize the concentrations of each individual library from Step 47 such that in the final pool each library is present at a concentration of 2 nM.

49 Sequence the samples on an Illumina NextSeq500/550 platform according to the Illumina user manual and at a depth of at least 1 million reads per sample. Use the NextSeq 500/550 High Output Kit v2.5 (150 cycles) with 151 cycles of read 1 (forward), 8 cycles of index 1 and 8 cycles of index 2. We recommend sequencing with a PhiX concentration of 1–3% and a loading concentration of 1.6–2.0 pM on the NextSeq to achieve optimal diversity and cluster density on the flowcell, respectively. The Illumina adapters featured in this protocol support sequencing of Record-seq libraries on other Illumina machines (e.g., MiSeq, NovaSeq). Note that optimal cluster densities and therefore loading concentrations vary.

Preparation of RNA-seq libraries for deep sequencing ● Timing: 3 d, 5 h hands-on

50 Extract the RNA from the frozen cell pellets from Step 11 using the RNASnap[27] reagent and solution prepared as follows:

Component	Amount per reaction (µL)	Final concentration
0.5 M EDTA	3.6	18 mM
10% SDS solution	0.25	0.025%
Formamide	1	1%
UltraPure water	95	—
Total volume	100	—

51 Resuspend each pellet from Step 11 in 100 µL RNASnap solution prepared in Step 50 in a 1.5 mL reaction tube, vortex vigorously and immediately place into a heating block set to 95 °C. Incubate for 7 min at 95 °C.

52 Centrifuge at 16,000*g* at room temperature for 5 min in a microcentrifuge.

53 Transfer 95 µL of supernatant into a fresh tube and add 400 µL of UltraPure water as well as 10 µL of 3 M sodium acetate solution. Mix by pipetting up and down several times. Then add 1 mL of pure ethanol and incubate the samples at −80 °C for 60 min.

54 Centrifuge at 16,000*g*, 4 °C for 30 min in a microcentrifuge. Aspirate and discard the supernatant and wash each pellet by adding 250 µL of 75% ethanol.

55 Centrifuge at 8,000*g* 4 °C for 5 min in a microcentrifuge. Aspirate and discard the supernatant. Spin again to collect residual ethanol and remove it completely by pipetting. Then dry the pellet for 10–15 mins at room temperature. Resuspend each pellet in 50 µL of UltraPure water.

56 Treat the extracted nucleic acids with the TURBO DNA-free Kit as follows:

Component	Amount per reaction (µL)	Final concentration
RNA sample in water from Step 55	43.5	—
10× TURBO DNase buffer	5	1×
TURBO DNase	1.5	0.0612 U/µL
Total volume	50	—

57 Mix gently, spin down briefly and incubate at 37 °C for 30 min.

58 Inactivate the TURBO DNase by adding 5 µL of DNase inactivation reagent per sample, incubate at room temperature for 5 min and mix 2–3 times during this incubation by gently flicking the tube.

59 Centrifuge at 21,000*g*, room temperature for 90 s in a microcentrifuge, then transfer the supernatant of each tube into separate wells of a skirted 96-well PCR plate.

60 Measure the concentration of the RNA using the Quant-iT RiboGreen RNA Assay Kit according to the manufacturer's instructions. Conduct the readout using a Tecan Infinite M1000 PRO microplate reader analogously to the plasmid DNA quantification described in Steps 23–25 with the standard curve provided from the RNA standard included in the Quant-iT RiboGreen RNA Assay Kit.

61 Prepare RNA-seq libraries for Illumina sequencing using the TruSeq Stranded mRNA Library Prep (96 Samples) kit according to the manufacturer's protocol. Use 200 ng of total RNA input material for depletion of ribosomal RNAs using the Ribo-Zero rRNA Removal Kit (bacteria).

62 Following the manufacturer's protocol, clean the depleted RNA using RNAClean XP beads and prepare cDNA libraries using the TruSeq Stranded mRNA Library Prep - perform all DNA purification steps during the protocol using AMPure XP beads.

63 Sequence the samples on an Illumina NextSeq500/550 platform according to the Illumina user manual and at a depth of at least 3 million reads per sample. Use the NextSeq 500/550 High Output Kit v2.5 (75 cycles) with 76 cycles of read 1 (forward), 8 cycles of index 1, and 8 cycles of index 2. We recommend sequencing with a PhiX concentration of 1% and a loading concentration of 1.4 pM on the NextSeq to achieve optimal diversity and cluster density on the flowcell, respectively. RNA-seq libraries prepared using the TruSeq Stranded mRNA Library Prep (96 samples) kit may be sequenced on other Illumina machines (e.g., MiSeq, NovaSeq). Note that optimal cluster densities and thus loading concentrations vary.

Data analysis ● Timing: 1 h per sample, 30 min hands-on

▲ **CRITICAL** The computational analysis pipeline is implemented for a Unix-like operating system with installed dependencies. We provide instructions on how to install dependencies mentioned in the software requirements at https://github.com/plattlab/Transcriptional-Recording

64 Using a bash shell, go to the desired path and create a directory to store reference genomes, a directory to store scripts and a directory for the experiment:

```
$ mkdir genomes scripts experimentDirectory
```

65 Download and save the *spacerExtractor.py* and *SEdedup.py* scripts (see Supplementary Data), as well as the JAR file for trimmomatic v0.36 (stored in the trimmomatic directory downloaded as mentioned in 'Software' section) in the 'scripts' directory.

66 Create a sub-directory to store sequencing results and move FASTQ files (gzip compressed) to this sub-directory:

```
$ mkdir experimentDirectory/fastqs
```

67 (Optional) If using example deep sequencing data, download the SRA accession list for the project —'SraAccessionList.txt' (Supplementary Data) and save it in the experiment directory, then use the following command to download the example FASTQ files:

```
$ while read -r line; do SRR=$(echo $line | awk '{print $1}'); SAMPLE=
$(echo $line | awk '{print $2}'| tr -d '[:space:]'); fastq-dump —gzip -Z
$SRR > fastqs/"$SAMPLE".fastq.gz; done < SraAccessionList.txt
```

68 Add all the paths for directories containing executable software tools used in the Snakefile (fastQC, FASTX-toolkit, bowtie2, samtools, featureCounts) to the PATH environment variable using:

```
$ export PATH=$PATH:/path/to/program/directory
```

This makes sure that these tools can be executed without referring to their paths on the system. To enable this in the default bash environment, add this command to your '~/.bash_profile' file as follows (for each program):

```
$ echo "export PATH=${PATH}:/path/to/program/directory" » ~/.bash_
profile
```

! CAUTION Make sure you use the '»' symbol and not a single '>' in the above command, as using a single '>' will overwrite your '.bash_profile' file.

69 In the reference genome directory, create a sub-directory for the reference *E. coli* BL21-Gold (DE3) pLysS genome named as the provided genome name (e.g., 'NC_012947.1').

```
$ mkdir genomes/NC_012947.1
```

Download and move the FASTA sequence and GFF3 annotation file for the reference genome to this sub-directory, and also name these files using the same genome name (e.g., 'NC_012947.1.fasta' and 'NC_012947.1.gff3').

70 In the reference genome directory, create a sub-directory for each plasmid construct.

```
$ mkdir genomes/pFS_0453
```

Download and move the plasmid FASTA and GFF3 to this subdirectory, and name the sub-directory and files contained therein using the provided plasmid name (e.g., 'pFS_0453')

71 For each plasmid construct, create a sub-directory for merged references (*E. coli* genome and plasmid) and name it using a concatenation of genome name and plasmid names separated by a single dot—genomename.plasmidname (e.g., 'NC_012947.1. pFS_0453').

```
$ mkdir genomes/NC_012947.1.pFS_0453
```

Create merged reference FASTA and GFF3 files and store them within this sub-directory following the same naming scheme. These merged FASTA and GFF3 files can be created by simply concatenating genome files with the respective plasmid references as follows:

```
$ cat genomes/NC_012947.1/NC_012947.1.fasta genomes/pFS_0453/pFS_
0453.fasta > genomes/NC_012947.1.pFS_0453/NC_012947.1.pFS_0453.fasta
```

72 Build bowtie2 indices within all these sub-directories (genome, plasmid, and merged) using the `bowtie2-build` command on the command line, as follows:

```
$ bowtie2-build genomes/NC_012947.1/NC_012947.1.fasta genomes/
NC_012947.1/NC_012947.1
$ bowtie2-build genomes/pFS_0453/pFS_0453.fasta genomes/pFS_0453/
pFS_0453
$ bowtie2-build genomes/NC_012947.1.pFS_0453/NC_012947.1.pFS_0453.
fasta genomes/ NC_012947.1.pFS_0453/NC_012947.1.pFS_0453
```

73 Change the current working directory to the experiment directory:

```
$ cd experimentDirectory
```

74 Download the provided supplementary files 'Snakefile' and 'config.yml' (Supplementary Data) and set up the following parameters to be used by Snakemake in the 'config.yml' file (using any text editor):

(i) 'datafolder': system path to sub-directory containing FASTQ file
(ii) 'scriptPath': system path to the 'scripts' directory
(iii) 'genomepath': system path to directory containing reference genomes
(iv) 'genome': genome name
(v) 'plasmid': plasmid name
(vi) 'dr1', 'dr2', and 'drfull': partial and complete direct repeat sequences
(vii) 'library_barcode': sequence of library identifier in case library identifiers are used (change it to 'null' in case no library identifiers are used)
(viii) 'infofile': Boolean (True or False) value determining whether files with information about spacer lengths and GC content are to be generated

Move the 'config.yml' file and the 'Snakefile' to the experiment directory

75 Perform primary data analysis for raw sequencing reads (FASTQ files) using Snakemake from within the experiment directory:

```
snakemake --cores no_of_cores
```

Alternatively, for an overview of the sequence of commands to be executed by the Snakemake workflow before running it, perform a 'dry' run of Snakemake using:

```
snakemake --dryrun
```

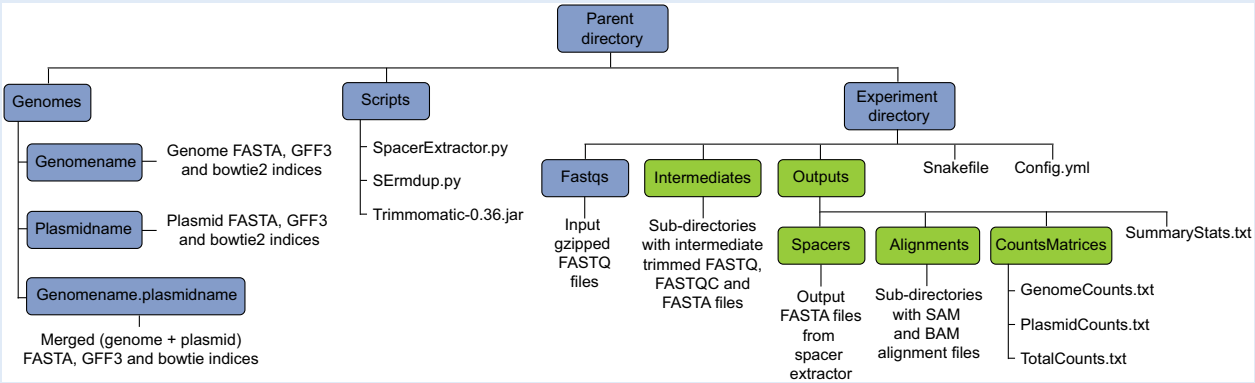
Box 3 | Overview of input requirements and output files for the primary analysis

The suggested setup for directories to be used for primary analysis is described in the figure below:
The directories in the blue boxes are to be set up by the user, while the directories in green (and contents) are generated by executing the Snakemake workflow. The 'genome' directory containing the reference genome files, and the 'scripts' directory containing the necessary scripts can also be set up elsewhere, with the corresponding path mentioned in the 'config.yml' file within the 'experimentalDirectory'. The user should also store the 'Snakefile' and create a 'fastqs' subdirectory with the input gzipped FASTQ files within the 'experimentalDirectory'.
On execution, in addition to count matrices and summary files, the Snakemake workflow creates intermediates such as trimmed FASTQ files, FASTA files, FASTQC analyses, and alignment outputs, and stores them in an 'intermediates' sub-directory within the experiment directory. It also generates an 'outputs' sub-directory, containing all useful output files.
'outputs' contains three further sub-directories – 'spacers', 'alignments' and 'countsMatrices'.
The 'spacers' directory contains the following output FASTA files generated by *spacerExtractor.py* for each sample:

1 A FASTA file containing all identified spacer sequences, suffixed by '.all.fasta'
2 A FASTA file containing unique identified spacers by sequence, suffixed by '.unique.fasta'
3 A FASTA file containing all reads having double acquisitions (two spacers acquired on the same array), suffixed by '.doubleAcquisitions.fasta'
4 A FASTA file containing all spacers from unique and complete double acquisitions (both spacers present), suffixed by '.doubleAcquisitions.paired.fasta'
5 A FASTA file containing all reads having multiple acquisitions (more than two spacers acquired on the same array), suffixed by '.multipleAcquisitions.fasta'
6 A FASTA file containing all spacers from unique and complete multiple acquisitions (all spacers present), suffixed by '.multipleAcquisitions.complete.fasta'
7 (optional) A text file containing unique spacer sequences alongside their GC content and length

The spacer extractor also generates a 'summaryStats.txt' file and stores it within the 'outputs' directory. This file contains the following quantities for each sample:
'totalReads', 'nonLibraryBarcodeRead%', 'allSpacers', 'uniqueSpacers', 'allDoubleAcquisitions', 'allDoubleAcquisitions.paired', 'allMultipleAcquisitions', 'allMultipleAcquisitions.complete', 'uniqueSingleAcquisitions', 'uniqueDoubleAcquisitions', 'uniqueMultipleAcquisitions'
For these parameters, 'all' refers to the total number of spacers identified in sequencing reads, while 'unique' refers to the number of unique spacers after collapsing by spacer sequence (Box 2). As an example, 'allSpacers' refers to the total number of spacers identified by the script for a particular sample, while 'uniqueSpacers' refers to the number of unique spacer sequences identified by the script for that sample.
'nonLibraryBarcodeRead%' is the percentage of reads identified for each sample that do not contain the library barcode, which could be an estimate of contamination by Record-seq libraries previously prepared in the laboratory. It should be noted that this metric also includes reads that are too short to capture the library barcode located in the 3′ of the amplicon. 'outputs/alignments' contains sub-directories with BAM and SAM alignment files, while 'outputs/countsMatrices' contains the following files:

1 A 'genomeCounts.txt' file, which contains genome-aligning spacer counts by gene for each sample.
2 A 'plasmidCounts.txt' file, which contains plasmid-aligning spacer counts by gene for each sample.
3 A 'totalCounts.txt' file, which contains the total number of genome-aligning and plasmid-aligning counts for each sample.



Snakemake can also be executed on a high-performance computing cluster environment, which significantly reduces runtime and parallelizes the processing of samples. A cluster config file has to be set up to submit jobs to a batch queue. Refer to the Snakemake user manual for instructions on how to implement this (https://snakemake.readthedocs.io/en/stable/executable.html#cluster-execution).

? TROUBLESHOOTING

76 The Snakemake run creates RNA-seq analogous count matrices (stored in 'experimentDirectory/outputs/countsMatrices') and a summary file with statistics on spacer abundances ('experimentDirectory/outputs/summaryStats.txt'). In addition to this, it creates many useful intermediate output files and stores them within sub-directories of the experiment directory (Box 3).

77 The second stage of the computational pipeline is implemented within R (RStudio recommended) using the *recoRdseq* R package. To install this package, start an R session and run the following commands:

```
> install.packages("devtools")
> library(devtools)
>   install_github("plattlab/Transcriptional-Recording",   subdir=
"recoRdseq")
```

78 The subsequent analysis requires a design matrix with information on the experimental design (Fig. 2b). Here, we integrate the design matrix as an excel (.xlsx) file for user convenience. A design matrix is provided for the count matrix generated by the Snakemake run (Supplementary Data: 'DesignMatrix.xslx'), which can also be used for the example counts dataset, with the same sample names as provided on SRA. This can be used as a template for other experimental designs. The design matrix can also be used for sample filtering—any samples that are to be excluded from secondary analysis can be left out of the design matrix. The first column of the design matrix contains sample names, subsequent column(s) contain information on experimental parameters (such as 'treatment', 'time', 'batch', etc.), hereafter referred to as design factors. The first row of the design matrix serves as a header for annotation.

▲ **CRITICAL STEP** The names of samples in the design matrix have to be the same as sample names used for the input gzipped FASTQ files. For example, if a sample input file is 'sample1.fastq.gz', the corresponding sample name in the design matrix should be 'sample1'. Also, ensure that control samples precede treated samples in the design matrix (if applicable).

79 Within RStudio (or R), load the installed Record-seq package with all dependencies as follows:

```
> library("recoRdseq")
```

The *recoRdseq* package has two main encapsulated functions:
recoRdSeqAnalysis(), which is the main function performing differential expression analyses and sample clustering.
spacerStats(), which outputs descriptive plots for spacer abundance and properties.

80 As an example, using the genome-aligning spacer counts generated by the Snakemake pipeline or the example counts dataset, the *recoRdSeqAnalysis* function can be called as follows:

```
> recoRdSeqAnalysis(countsMatrix = "experimentDirectory/outputs/
countsMatrices/genomeCounts.txt", designMatrix = "path/to/design
Matrix.xlsx", outPath = "path/for/output")
```

'countsMatrix' is used to specify the path to the genome-aligning (or plasmid-aligning) counts file generated as an output of the Snakemake run.
'designMatrix' is the path to the design matrix file in excel (.xlsx) format.
'outPath' is the desired path for the output directory.
Optionally, to generate descriptive bar plots for plasmid and genome-aligning counts grouped by experimental design(s) contained in the design matrix, the 'totalCountsFile' argument can be used in the *recoRdSeqAnalysis()* function call to supply the path to total counts file generated by the Snakemake workflow (e.g., 'experimentDirectory/outputs/unique/countsMatrices/totalCounts.txt'). The other arguments offer control over the processing and analysis of data and can be used to generate different types of outputs. Detailed information about these can be found in the *recoRdseq* user manual or R vignette.

? TROUBLESHOOTING

81 When the *recoRdSeqAnalysis()* function is executed as above, it loops through the design factors, sorts the samples into groups for each design factor, and generates the following plots based on these groups:

(i) Box and whisker plots for top 'nGenes' genes by expression (Fig. 3a), where 'nGenes' is a user-specified argument for the *recoRdSeqAnalysis()* function (50 by default)
(ii) Principal component analyses (PCA) plots for top 'nGenes' genes by variance across samples (Fig. 3b)

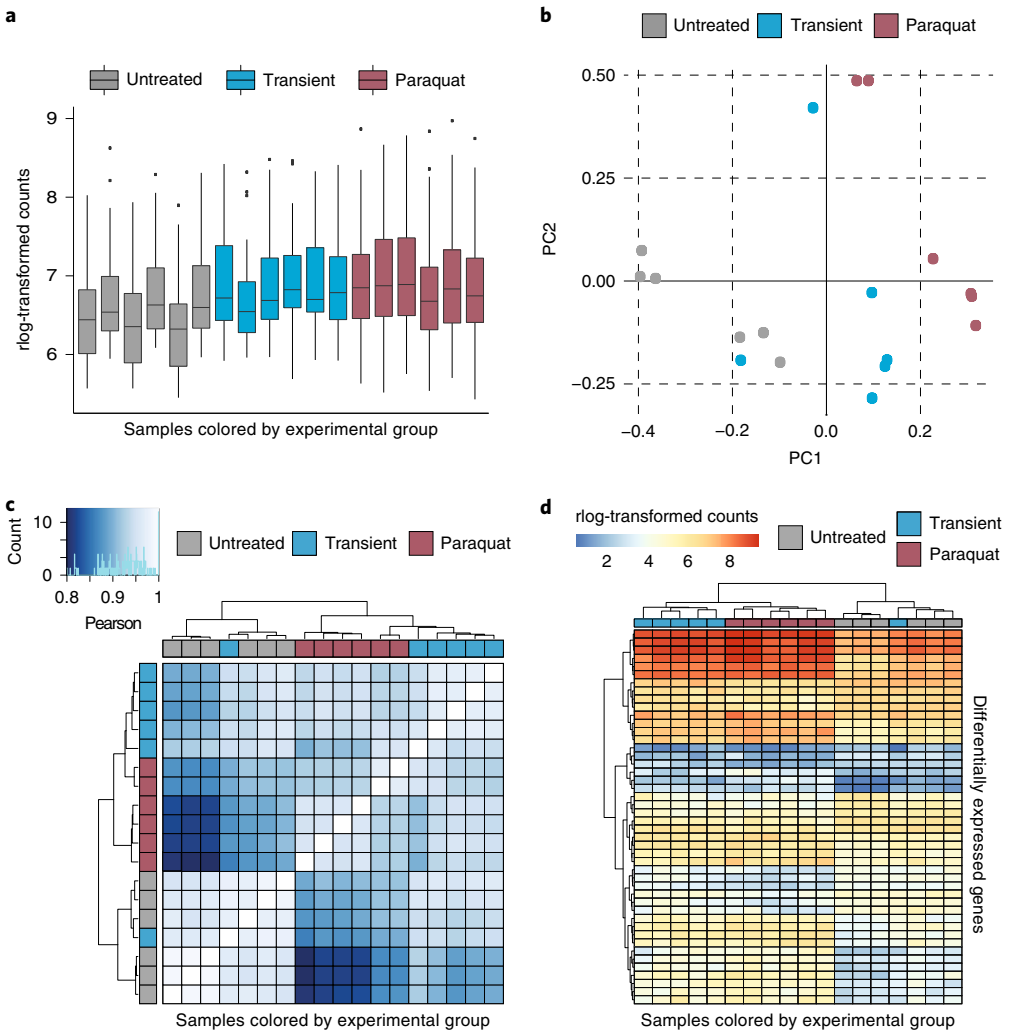


Fig. 3 | Illustrative output plots generated by the *recoRdseq* R package, with samples color-coded by experimental group specified in the design matrix. a, Boxplots for regularized log (rlog) transformed genome-aligning spacer counts for the top 50 genes by expression. **b**, PCA plots for rlog-transformed genome-aligning spacer counts for the 50 most variable genes. **c**, Correlation heatmap showing sample clustering performed using Pearson correlation. **d**, Heatmap showing unsupervised hierarchical clustering of samples based on rlog transformed genome-aligning spacer counts for the union of the top 20 DE genes identified by *DESeq2, edgeR and baySeq* (padj < 0.05). For all panels, $n = 6$ independent biological replicates.

(iii) PCA plots with sample clustering into k clusters based on k-means (Supplementary Fig. 1b) and Fuzzy Analysis (FANNY) approaches, where k is a user-specified argument for the *recoRdSeqAnalysis()* function (2 by default)

(iv) Correlation heatmap showing hierarchical clustering based on Pearson correlation between different samples (Fig. 3c)

(v) Individual CSV files with significant DE genes identified by the DE packages *DESeq2*, *edgeR*, and *baySeq* as well as the intersect and union of these genes. Genes identified by all 3 tools are called 'signature' genes

(vi) Venn diagram illustrating the overlap of DE genes identified by the three DE packages (Supplementary Fig. 1c)

(vii) Heatmaps based on unsupervised hierarchical clustering of samples using the union or intersect of DE genes identified by the 3 DE packages (Fig. 3d and Supplementary Fig. 1d)

(viii) (optional) Bar plots for total genome and plasmid-aligning counts based on sample groups (Supplementary Fig. 1a)

82 In case the differential expression analysis is to be performed for a particular design factor or combination of design factors instead of performing it individually for each design factor, the

'designFormula' argument can be supplied to the *recoRdSeqAnalysis()* function. For example, if the analysis is to be performed for a combination of the 'treatment' and 'time' design factors, the following function call can be executed:

```
> recoRdSeqAnalysis(countsMatrix = "experimentDirectory/outputs/ counts-
Matrices/genomeCounts.txt", designMatrix = "path/to/designMatrix.xlsx",
outPath = "path/for/output", designFormula = "~treatment+time")
```

The design formula can contain any combination of different design factors and arithmetic operators; however, it must start with a tilde (~) symbol, and only use design factors described in the design matrix without whitespaces. For further information on design formulas, refer to the *DESeq2* or *edgeR* user manuals.

83 To generate descriptive boxplots for spacer abundances grouped by experimental design(s) contained in the design matrix, the *spacerStats()* function can be called as follows:

```
> spacerStats(summaryStats = "experimentDirectory/outputs/summary-
Stats.txt", outPath = "path/for/output", designMatrix = "path/to/
designMatrix.xlsx")
```

'summaryStats' is used to specify the path to the'summaryStats.txt' file generated by *spacerExtractor.py*
'designMatrix' and 'outPath' are specified above
Alternatively (or in addition to this), if the user wants to generate barplots for spacer GC content and spacer lengths for samples grouped by experimental design(s) contained in the design matrix, the *spacerStats()* function can be called as follows:

```
> spacerStats(SpacerInfoFiles_Path = "experimentDirectory/outputs/
spacers", outPath = "path/for/output", designMatrix= "path/to/design-
Matrix.xlsx")
```

'SpacerInfoFiles_Path' is the path to the sub-directory containing the output files generated by the spacer extractor

▲ **CRITICAL STEP** To generate spacer GC content and length plots using *spacerStats()*, the 'infoplot' argument in the 'config.yml' file for Snakemake must be set to 'True', so that spacer info files (with the suffix '.info.txt') are generated by spacerExtractor.py and stored in the 'outputs/spacers' directory.

Troubleshooting

Troubleshooting advice can be found in Table 2.

Table 2 | Troubleshooting table

Step	Problem	Possible reason	Solution
20	Poor DNA yield	Insufficient number of input cells; experimental conditions affect plasmid DNA replication	Increase the volume of the culture or time of recording under your specific conditions; optimize plasmid extraction procedure
	Poor DNA yield	Too much cellular material leading to insufficient cell lysis	Decrease the volume of the culture used for plasmid extraction
45	No band after second round PCR of SENECA NGS library preparation	Insufficient input for SENECA adapter ligation; wrong primers for first round PCR; problems during purification with AMPure XP beads	Make sure that DNA concentration is correct; to confirm proper reaction conditions add a positive control (*E. coli* BL21(DE3) transformed with pFS_0453, grown in TB media containing aTc for 14 h) to help you to test for successful spacer acquisition
	Intensity of PCR bands varies within one biological replicate	Differences in spacer acquisition rate; errors in measurements of plasmid DNA concentration; poor library preparation technique	Verify plasmid DNA concentration and ensure DNA quality (e.g., by running a small amount of DNA on a 1% agarose gel to check DNA integrity)

Table continued

Table 2 (continued)

Step	Problem	Possible reason	Solution
	180 bp PCR product corresponding to non-expanded array band	Inefficient spacer acquisition or adapter ligation during SENECA	Include a positive control as described above; ensure reagents used for adapter ligation (FaqI, T7 DNA Ligase, DTT solution) are functional
75	Snakemake run is not executed successfully	One of the steps in the computational workflow had an error	Snakemake generates log files that help you identify the source of the error, which can then be resolved
	featureCounts generates empty count matrices	Inaccurate feature annotation or attributes	Ensure that the sequence annotations used in the FASTA and GFF3 files are identical. Also, ensure that an exon feature is present for each gene and specify the 'exon_id' parameter in the 'attributes' column
	Snakemake run on a cluster environment exits before completion	Output file of a job took too long to generate—this is usually true for the *spacerExtractor.py*	Use the `—latency-wait` flag with a sufficiently high wait time while executing Snakemake to make the program wait longer for the outputs at each stage
76	Samples provided substantially fewer deep sequencing reads than desired	Wrong pooling strategy	Prepare pools of samples based on similar band intensity and quantify these pools separately by qPCR before assembly the final 2 nM library for sequencing
	A large fraction of reads does not have the expected library barcode	Contamination by previous library preparation; poor quality of reads	Usually we have 1–5% reads not carrying library barcode; rigorous cleaning procedures using 1% sodium hypochlorite decrease the risk of contamination by previous library preparations
80	*recoRdseq* R package functions show error for sample naming	Sample names for FASTQ files and sample names specified in design matrix are not identical; sample names start with numbers	The names of FASTQ files (e.g., 'sample1.fastq.gz') and sample names specified in the design matrix (e.g., 'sample1') should be identical (capitalization matters); there should be a FASTQ file corresponding to each sample name in the design matrix; sample names should not start with numbers

Timing

Steps 1–7, transformation of recording plasmid into expression strain: 40 min, 5 min hands-on
Steps 8–20, recording of transcriptional histories: 1 d, 90 min hands-on
Steps 21 - 25, quantification of plasmid DNA using PicoGreen: 1 h hands-on
Steps 26–46, selective amplification of expanded CRISPR arrays (SENECA): 2 d, 3 h hands-on
Steps 47–49, deep sequencing of Record-seq libraries: 1 d, 3 h hands-on
Steps 50–63, preparation of RNA-seq libraries for deep sequencing: 3 d; 5 h hands-on
Steps 64–83, data analysis: ~1 h per sample; 30 min hands-on

Anticipated results

As a reference for Record-seq results, we provide data for the transient exposure of *E. coli* BL21 cultures to the herbicide paraquat. This dataset was obtained using plasmid isolations identical to those presented in our original publication[17] and was processed using the updated version of the computational pipeline described above.

As seen in the PCA plot (Fig. 3b), samples that were transiently exposed to paraquat form a distinct cluster closer to samples chronically exposed to the herbicide. *k*-means or FANNY clustering using two centroids reveals that transiently exposed samples, albeit with one outlier, cluster with chronically exposed samples while unexposed samples form a distinct cluster (Supplementary Fig. 1b). Furthermore, clustering Record-seq data using Pearson correlation reveals distinct clusters for samples chronically exposed, transiently exposed, and unexposed to paraquat highlighting the detection of different transcriptional states of populations of cells by Record-seq (Fig. 3c). Differential expression analysis revealed 111 signature genes, and unsupervised hierarchical clustering of samples for these genes also confirmed the finding that transiently exposed samples, albeit with one outlier, cluster with chronically exposed samples and unexposed samples form a distinct cluster (Fig. 3d). Thus, Record-seq (but not RNA-seq) was capable of detecting cultures transiently exposed to paraquat.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Deep sequencing data are available in the National Center for Biotechnology Information Sequence Read Archive (PRJNA510019). An example counts dataset can be accessed from https://github.com/plattlab/Transcriptional-Recording/tree/master/example-data.

Code availability

The python code is available in the Supplementary Data. The latest versions of the code can be accessed at the Platt lab github (https://github.com/plattlab/Transcriptional-Recording). The code in this protocol has been peer reviewed.

References

1. Schmidt, F. & Platt, R. J. Applications of CRISPR-Cas for synthetic biology and genetic recording. *Curr. Opin. Syst. Biol.* **5**, 9–15 (2017).
2. Farzadfard, F. & Lu, T. K. Emerging applications for DNA writers and molecular recorders. *Science* **361**, 870–875 (2018).
3. Esvelt, K. M. & Wang, H. H. Genome-scale engineering for systems and synthetic biology. *Mol. Syst. Biol.* **9**, 641 (2013).
4. Farzadfard, F. & Lu, T. K. Synthetic biology. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science* **346**, 1256272 (2014).
5. Roquet, N., Soleimany, A. P., Ferris, A. C., Aaronson, S. & Lu, T. K. Synthetic recombinase-based state machines in living cells. *Science* **353**, aad8559 (2016).
6. Weinberg, B. H. et al. Large-scale design of robust genetic circuits with multiple inputs and outputs for mammalian cells. *Nat. Biotechnol.* **35**, 453–462 (2017).
7. Zamft, B. M. et al. Measuring cation dependent DNA polymerase fidelity landscapes by deep sequencing. *PloS ONE* **7**, e43876 (2012).
8. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
9. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
10. Frieda, K. L. et al. Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).
11. Perli, S. D., Cui, C. H. & Lu, T. K. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science* **353**, https://doi.org/10.1126/science.aag0511 (2016).
12. Tang, W. & Liu, D. R. Rewritable multi-event analog recording in bacterial and mammalian cells. *Science* **360**, https://doi.org/10.1126/science.aap8992 (2018).
13. Sheth, R. U., Yim, S. S., Wu, F. L. & Wang, H. H. Multiplex recording of cellular events over time on CRISPR biological tape. *Science* **358**, 1457–1461 (2017).
14. Silas, S. et al. Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science* **351**, aad4234 (2016).
15. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science* **353**, aaf1175 (2016).
16. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* **547**, 345–349 (2017).
17. Schmidt, F., Cherepkova, M. Y. & Platt, R. J. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* **562**, 380–385 (2018).
18. Nuñez, J. K. et al. Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528–534 (2014).
19. Jackson, S. A. et al. CRISPR-Cas: adapting to change. *Science* **356**, https://doi.org/10.1126/science.aal5056 (2017).
20. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.* **40**, 5569–5576 (2012).
21. Koster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
22. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
23. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
24. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
25. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**, e108 (2013).

26. Leinonen, R., Sugawara, H. & Shumway, M., International Nucleotide Sequence Database, C. The sequence read archive. *Nucleic Acids Res.* **39**, D19–21 (2011).
27. Stead, M. B. et al. RNAsnap: a rapid, quantitative and inexpensive, method for isolating total RNA from bacteria. *Nucleic Acids Res.* **40**, e156 (2012).

Acknowledgements
We thank C. Beisel, E. Burcklen, K. Eschbach, T. Schär, and I. Nissen from the Genomics Facility Basel for assistance in Illumina sequencing. F.S, M.Y.C, T.T., and R.J.P. are supported in part by funds from the Swiss National Science Foundation, ETH domain Personalized Health and Related Technologies, Brain and Behavior Research Foundation, and the National Centres of Competence–Molecular Systems Engineering. Plasmid reagents are available through Addgene. Correspondence and requests for materials should be addressed to R.J.P. (rplatt@ethz.ch).

Author contributions
F.S., M.Y.C., T.T., and R.J.P. designed the study. F.S. performed experiments, T.T. and F.S. analyzed data, T.T. created the python pipeline for primary analysis, T.T. and M.O created the *recoRdseq* library for secondary analysis, F.S., T.T, M.Y.C., and R.J.P. wrote the manuscript with input from all authors.

Competing interests
R.J.P. and F.S. are inventors on a patent application filed by ETH Zurich relating to work in this article. All other authors have no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41596-019-0253-4.

Correspondence and requests for materials should be addressed to R.J.P.

Peer review information *Nature Protocols* thanks Stan J. J. Brouns, Simon Jackson and Malcolm F. White for their contribution to the peer review of this work.

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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 18 April 2019; Accepted: 8 October 2019;
Published online: 10 January 2020

Related link
Key references using this protocol
Schmidt, F. et al. *Nature* **562**, 380–385 (2018): https://doi.org/10.1038/s41586-018-0569-1

nature research

Corresponding author(s): Randall J. Platt

Last updated by author(s): Aug 29, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	*Provide a description of all commercial, open source and custom code used to collect the data in this study, specifying the version used OR state that no software was used.*
Data analysis	Python > 3.0, Snakemake > 5.3.0, trimmomatic 0.36, fastQC > 0.11, FASTX-toolkit 0.0.14, bowtie2, samtools, Subread > 1.6.0, R >3.5.0, RStudio, R packages: recoRdseq, DESeq2, edgeR, baySeq, ggplot2, VennDiagram, pheatmap, ggfortify, cluster, readxl, readr, reshape2, rgl, gplots, RColorBrewer

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data and code availability
Deep sequencing data are available in the National Center for Biotechnology Information Sequence Read Archive (PRJNA510019). The python code is available in the supplementary data. The latest versions of the code can be accessed at the Platt lab github (https://github.com/plattlab/Transcriptional-Recording).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We reasoned that a sample size of at least 3 (n equal to or greater than 3) would be sufficient to measure the effect size as is standard practice.
Data exclusions	no data were excluded from the analyses
Replication	experimental findings were replicated in independent cultures picked from individual bacterial colonies,
Randomization	This is not relevant to molecular/cell biology studies and samples were not randomized.
Blinding	This is not relevant to molecular/cell biology studies and the investigators were not blinded

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging