

Forming molecular memories; Record-seq

Technical Journal Club

Daniel Heinzer

January 29th 2019

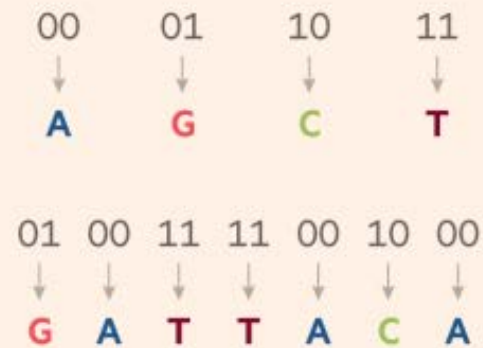
DNA as data storage medium

Encoding data in DNA

How a digital file's binary code can be converted into a 'genetic file' and stored as strands of DNA.

1. Coding

A digital file's binary code is translated into pairings of DNA bases, abbreviated A (adenine), C (cytosine), G (guanine) and T (thymine). These form the rungs that make up the DNA strands



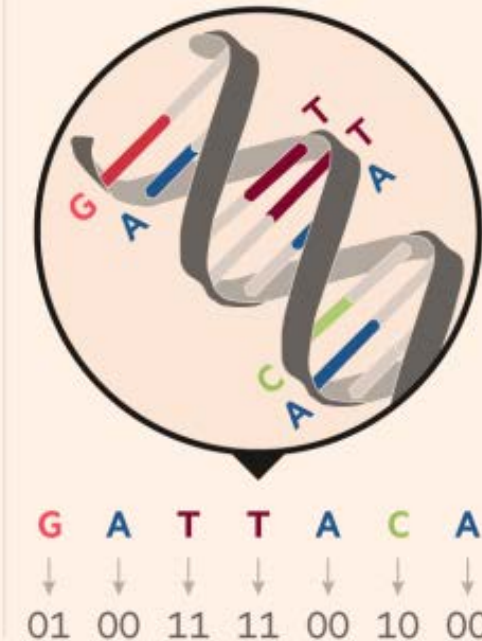
2. Synthesis and storage

A synthetic biological engineering company builds DNA strands matching the sequence of digital code. These can be held indefinitely in cold storage



3. Retrieval and decoding

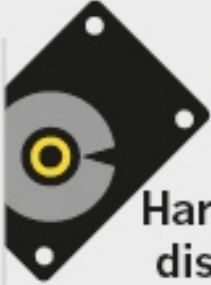
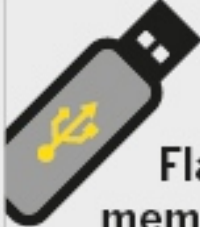


DNA is run through a sequencer which returns the generic code. This is then translated back to binary



DNA as data storage medium

STORAGE LIMITS

Estimates based on bacterial genetics suggest that digital DNA could one day rival or exceed today's storage technology.

	 Hard disk	 Flash memory	 Bacterial DNA	WEIGHT OF DNA NEEDED TO STORE WORLD'S DATA  ~1 kg
Read-write speed (μ s per bit)	> ~3,000–5,000	~100	<100	
Data retention (years)	>10	>10	>100	
Power usage (watts per gigabyte)	~0.04	~0.01–0.04	<10 ⁻¹⁰	
Data density (bits per cm ³)	~10 ¹³	~10 ¹⁶	~10 ¹⁹	

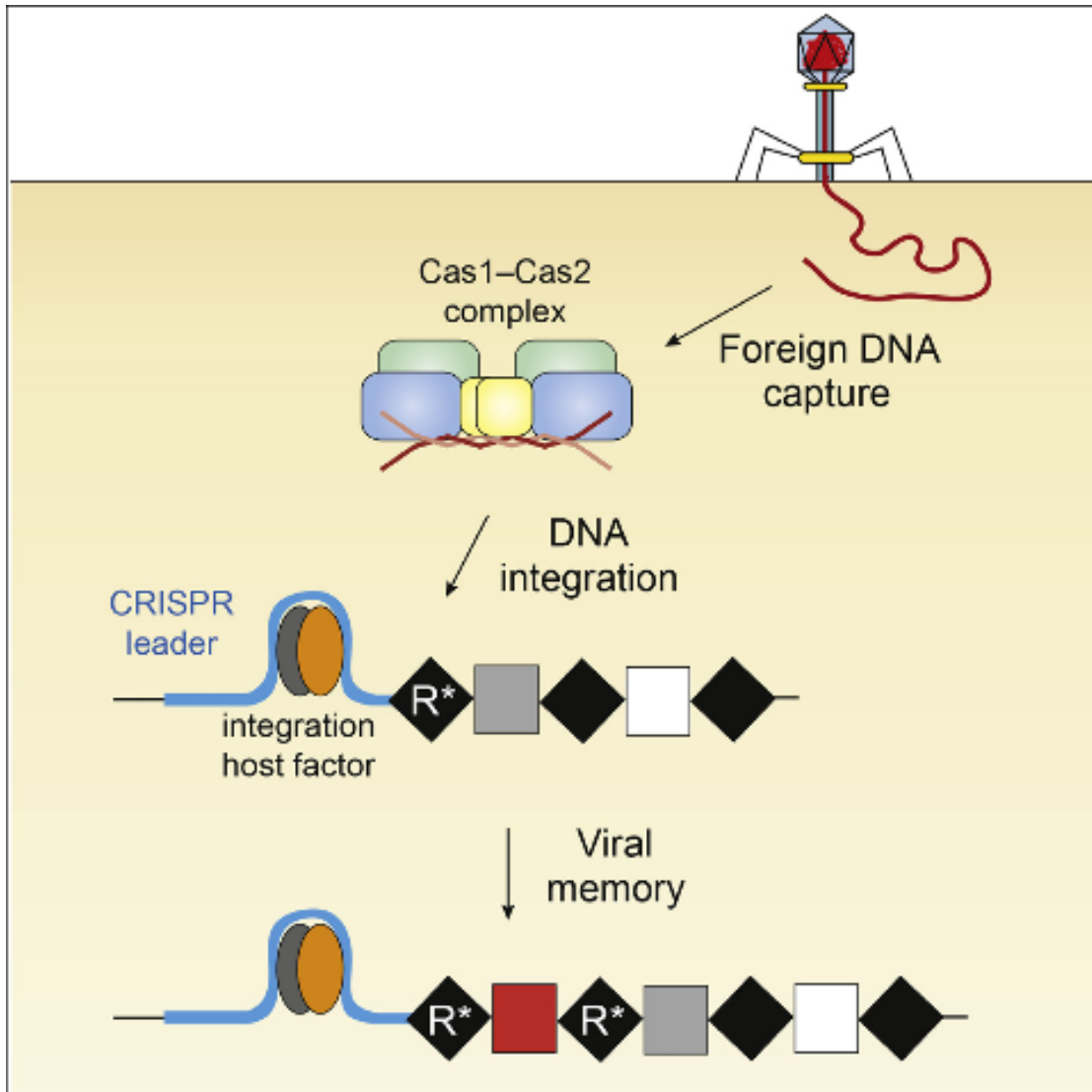
©nature

Nature 537, 22–24 (01 September 2016) | doi:10.1038/537022a

DNA as data storage medium

- In vitro synthesized DNA can be used to store information
- Can the system be adapted to organisms?

DNA based data storage in organisms



Cas1-Cas2 protein complex captures 30 – 40 base pair segments of foreign DNA and catalyzes their integration into the host genome (of *E. Coli*) in the CRISPR array as unique spacer sequences.

Papers

1.) CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria

2.) Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein

3.) Transcriptional recording by CRISPR spacer acquisition from RNA

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

Seth L. Shipman^{1,2,3}, Jeff Nivala^{1,3}, Jeffrey D. Macklis² & George M. Church^{1,3}

2017

- Use of Cas1–Cas2 system to encode pixel values of black and white images and a short movie into the genome of a population of living bacteria
- Uncovering underlying principles of the CRISPR-Cas adaption system, including sequence determinants of spacer acquisition

Image pixel values stored in a nucleotide code

Image pixel values were stored in a nucleotide code as synthetic oligonucleotides and electroporated into a population of bacteria (**overexpressing Cas1-Cas2** and harbouring a functional CRISPR array)

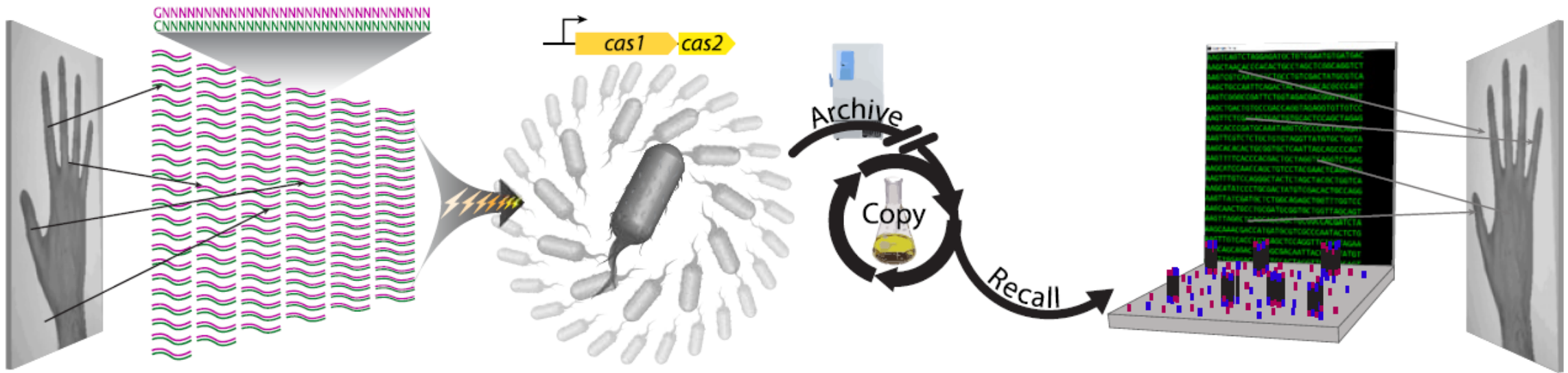


Image pixel values stored in a nucleotide code

Protospacer set up:

- PAM to increase efficiency of acquisition and define direction of spacer insertion
- Pixet (consisting of 4 nucleotides) serves as a barcode defining a set of pixels in the image.
- Each following nucleotide (28 per protospacer) encodes another colour of a pixel, distributing a 56x56 pixel image across 112 oligonucleotides

b



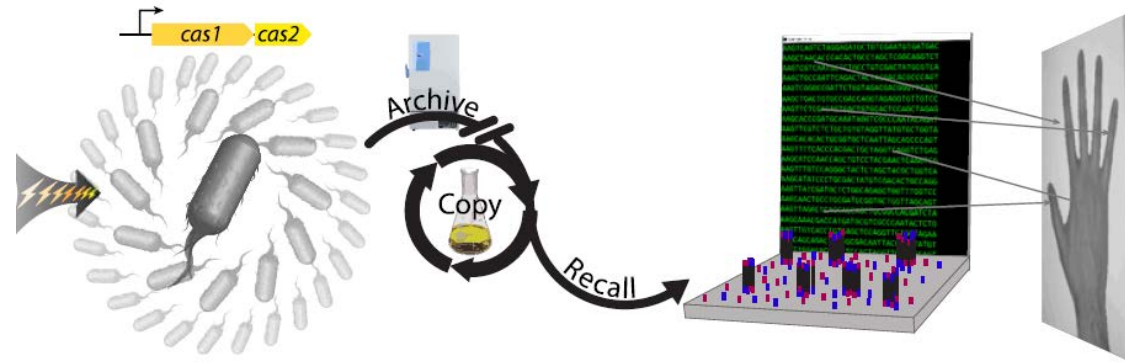
c



d



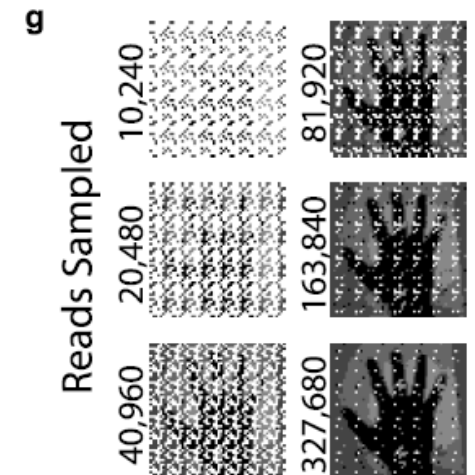
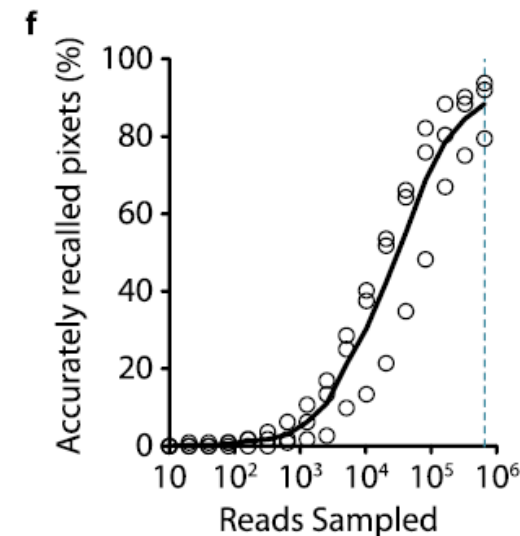
Image pixel values stored in a nucleotide code



Pooled oligonucleotides were electroporated into a population of *E. Coli*, grown overnight and a sample was sequenced

Newly acquired spacers were bioinformatically extracted

Recovered image



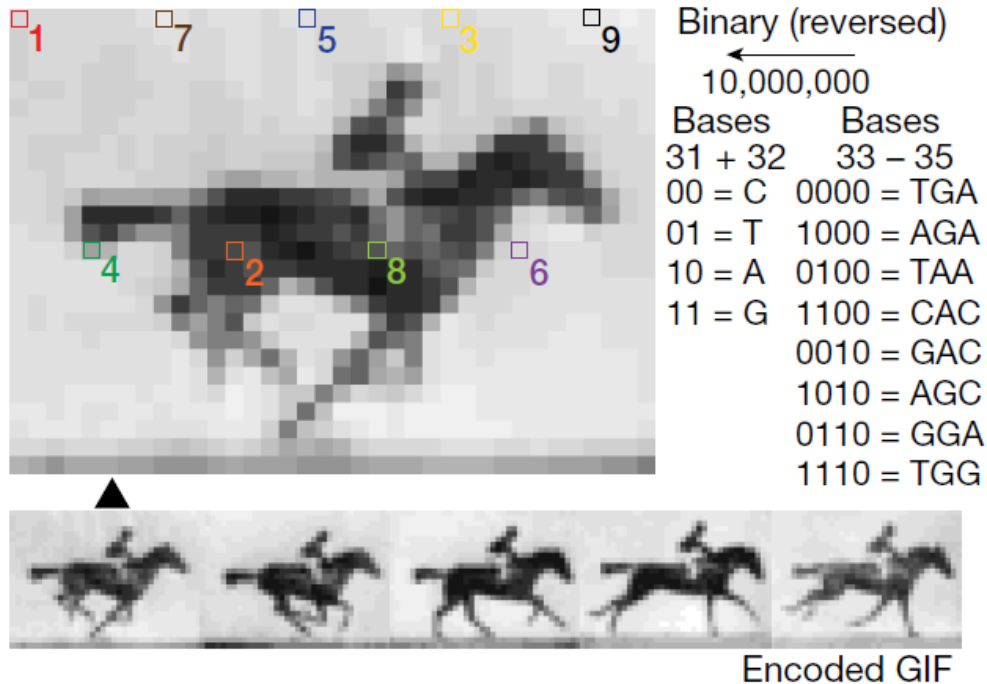
GIF pixel values stored in a nucleotide code

Using the same principle, slightly differently encoding and optimized Protospacer sequences, they tried to store the information of a GIF into a population of bacteria

a

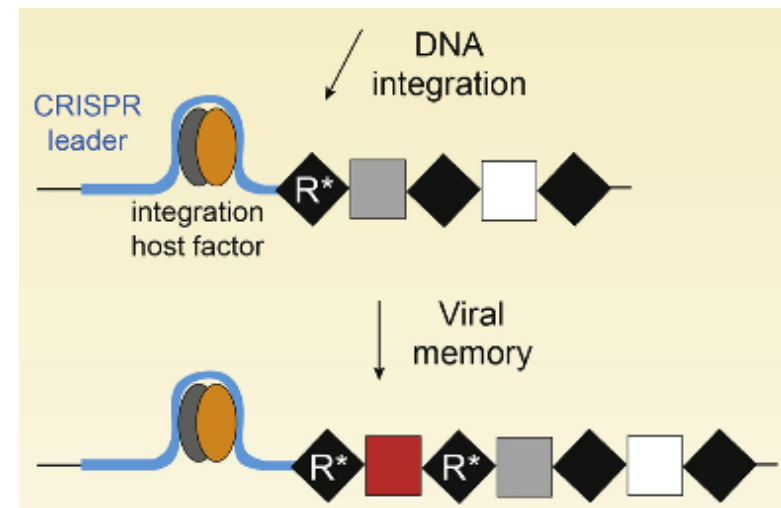
GC ~50%, no mononucleotide
repeats >3 bp,
no internal PAMs

Pixel
AAGCGACGTAGACTCTCTCGACAATAGGTTACTGA
1 2 3 4 5 6 7 8 9 1



However, no encoding of the frame

Trick: Electroporate the bacteria over 5 days, each day using protospacer for another frame

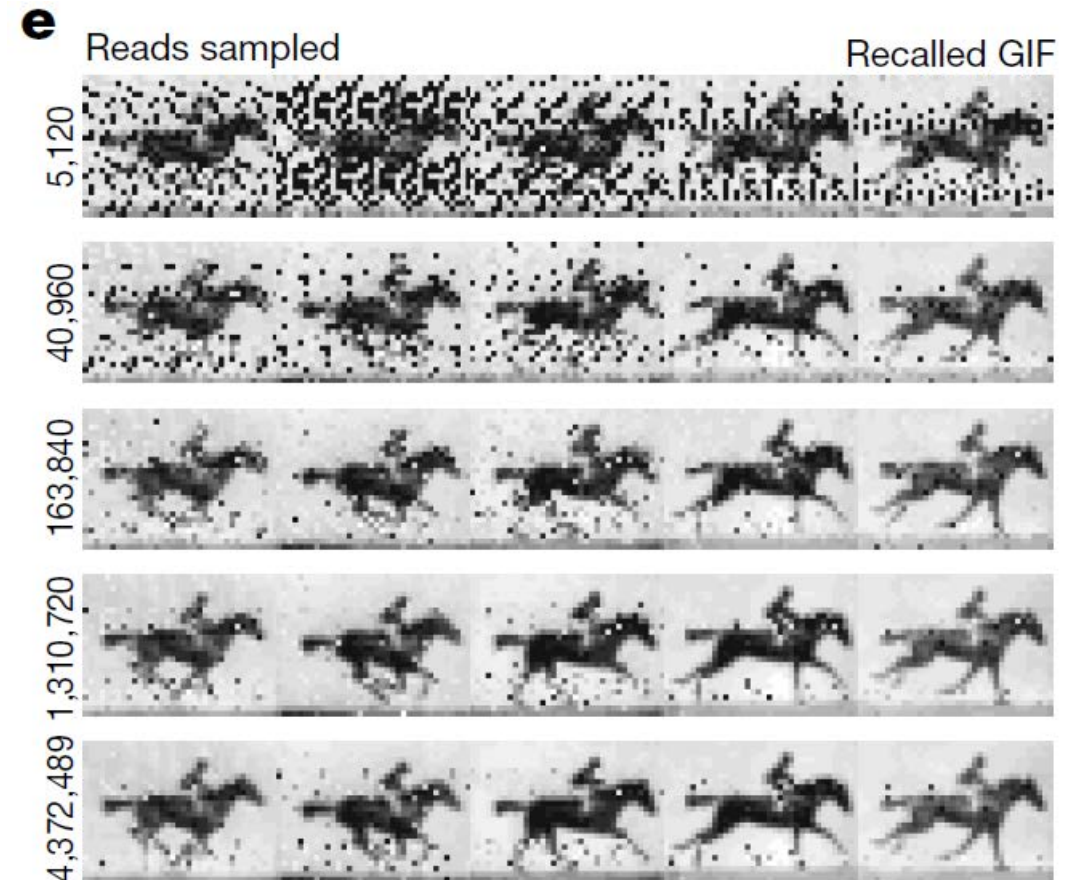


GIF pixel values stored in a nucleotide code

Again, sample was sequenced, and a complex analysis based on permutation for each position was applied to recover the GIF:

Again, depending on the read depth, they were capable to recall the GIF with a <90% accuracy.

Insertion of spacers on different time points led to a physical arrangement of temporal information about the frame



Key points Paper 1

- system can capture and stably store practical amounts of real data within the genomes of populations of living cells.
- CRISPR arrays are capable of inserting DNA snippets, yielding temporal information about an event.

Papers

1.) CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria

2.) Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein

3.) Transcriptional recording by CRISPR spacer acquisition from RNA

GENE EDITING

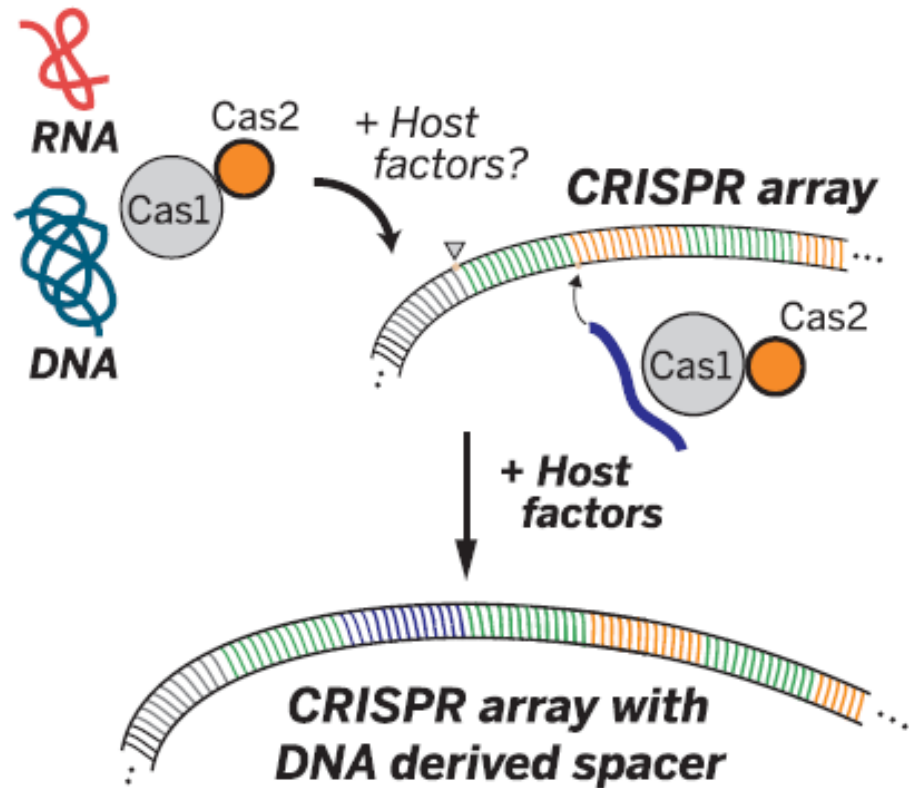
Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase–Cas1 fusion protein

Sukrit Silas,^{*} Georg Mohr,^{*} David J. Sidote, Laura M. Markham, Antonio Sanchez-Amat, Devaki Bhaya, Alan M. Lambowitz,[†] Andrew Z. Fire[†]

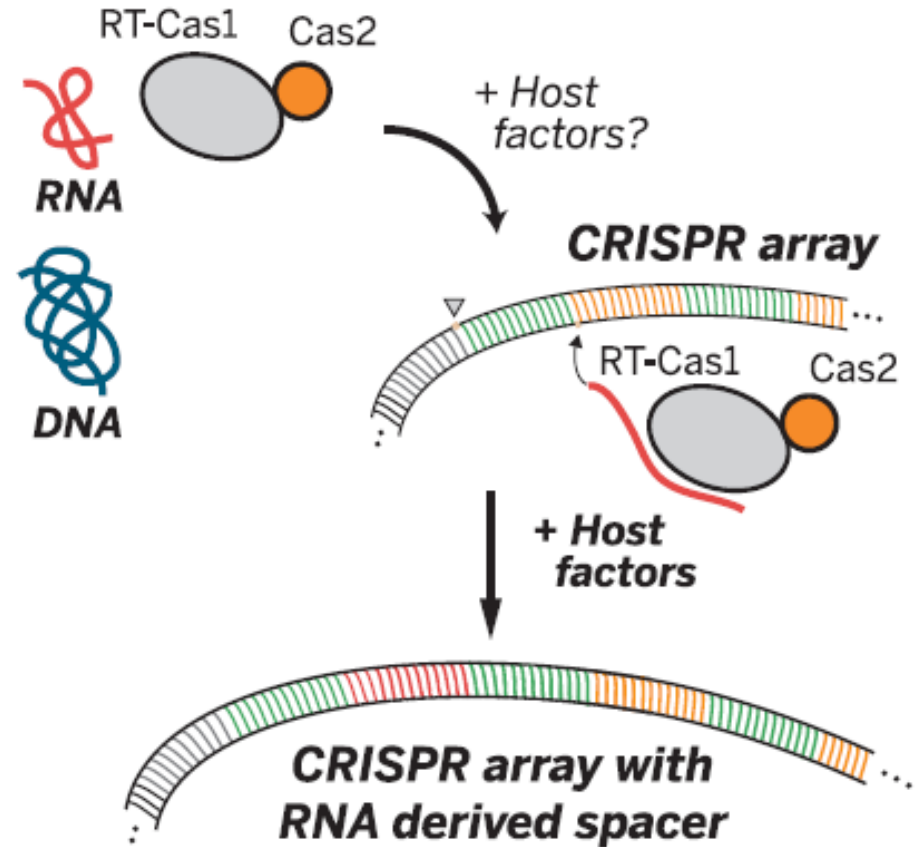
- sought to determine whether some CRISPR-Cas systems build CRISPR arrays through the acquisition of spacer sequences from RNA rather than DNA
- in some CRISPR systems, Cas1 is naturally fused to a reverse transcriptase (RT), suggesting the possibility of a concerted spacer integration mechanism involving Cas1 integrase activity and the reverse transcription of RNA to DNA.
- This would enable the acquisition of new spacers from RNA and outline a host-mediated mechanism for reverse information flow from RNA to DNA.

CRISPR array with RNA derived spacers?

DNA spacer integration by Cas1/Cas2



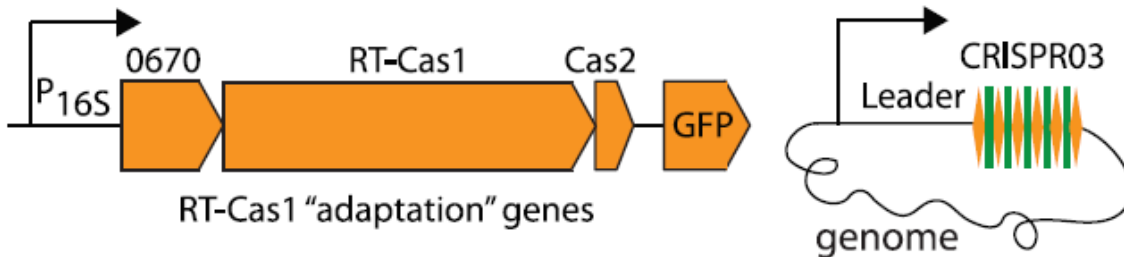
RNA spacer integration by RT-Cas1/Cas2



CRISPR array with RNA derived spacers?

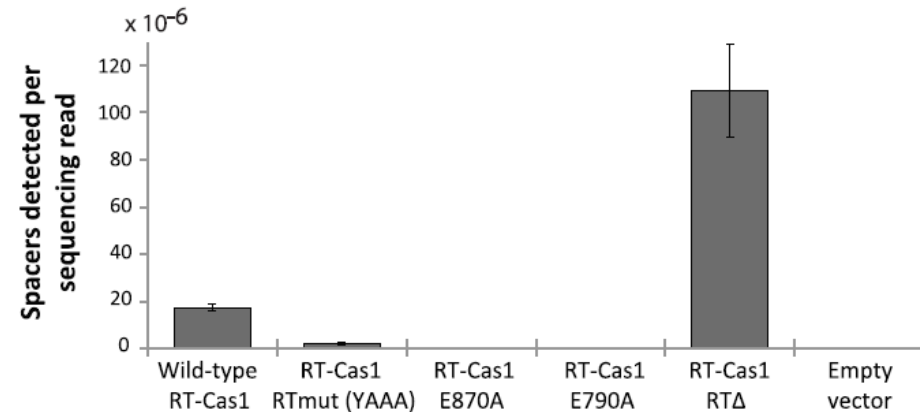
Subject of study: Type III-B CRISPR locus in *M. mediterranea* (MMB-1), an easily cultured, nonpathogenic organism that contains a RT-Cas1–encoding gene.

A Overexpression constructs for MMB-1 Type III-B CRISPR adaptation genes

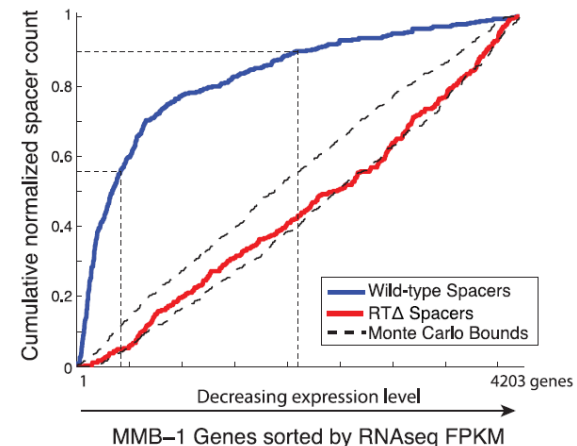


Overexpression of the Cas1RT-Cas2 complex and analysis of spacer acquisition in the genomic CRISPR array CRISPR03 by means of amplification of the region by PCR using primers for the leader sequence and the first native spacer followed by high throughput sequencing

B New spacer detection frequency



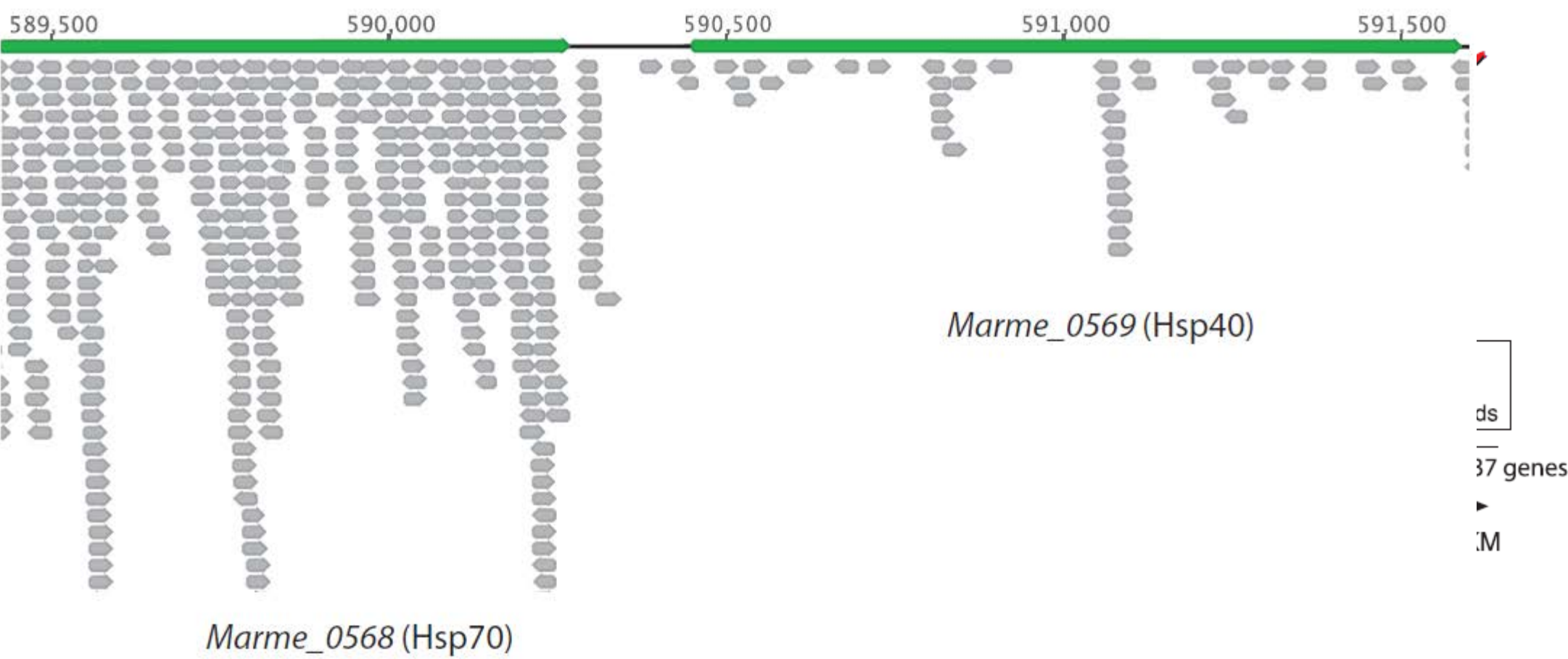
E Protospacer association with transcription level



B. Spacers acquired from a representative genomic locus in *E. coli*

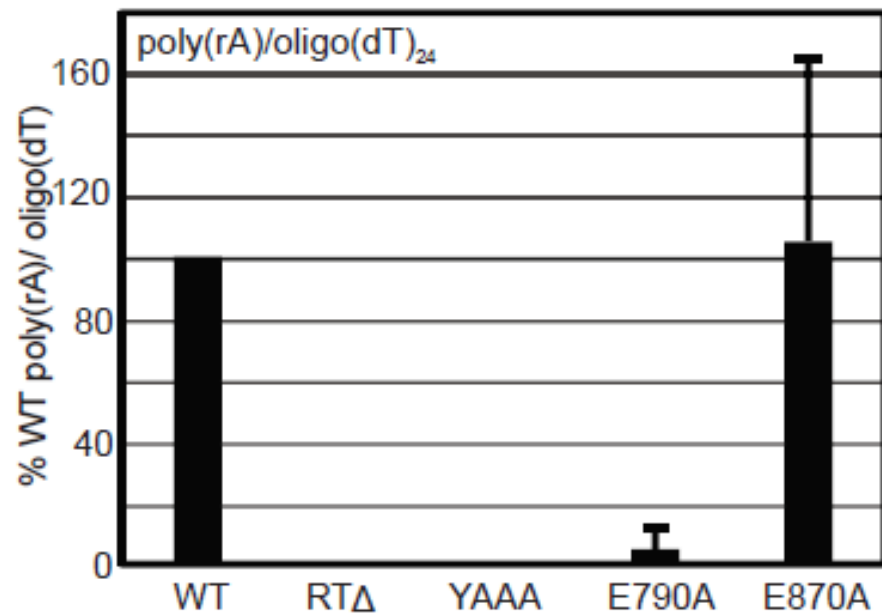


C. Spacers acquired from a representative genomic locus in MMB-1

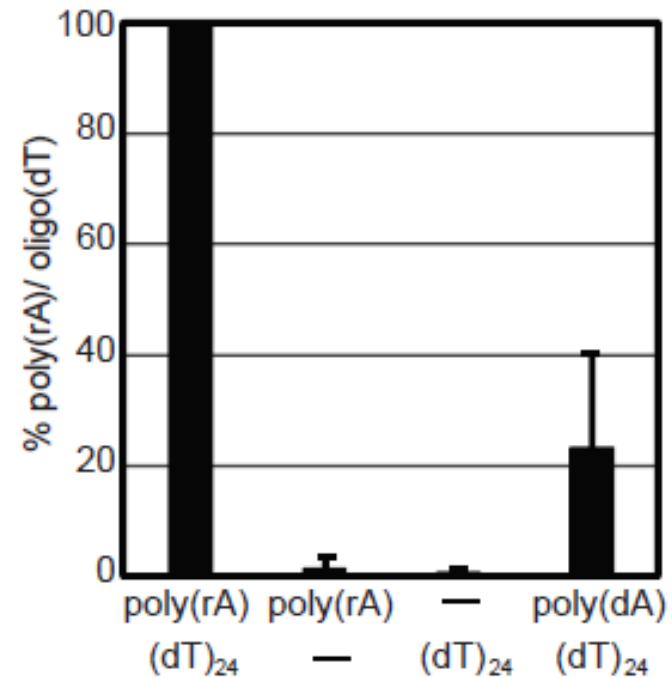


RT-Cas1 is an active reverse transcriptase

A. RT activity of wild-type and mutant RT-Cas1 proteins

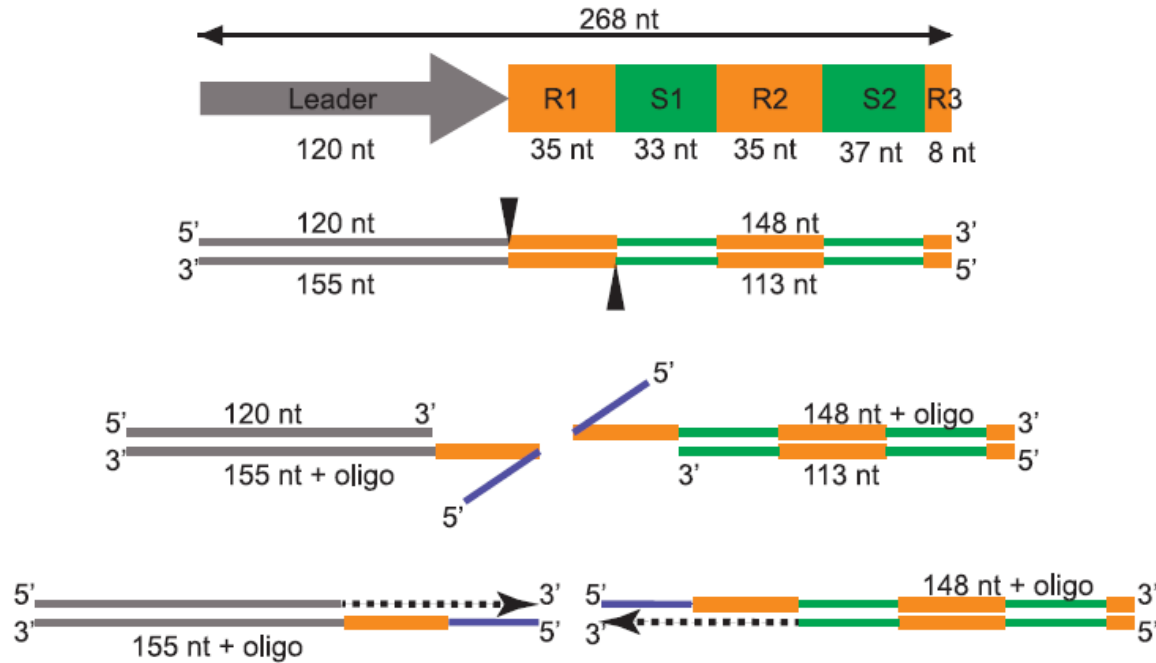


B. RT activity of RT-Cas1 with various primers/templates



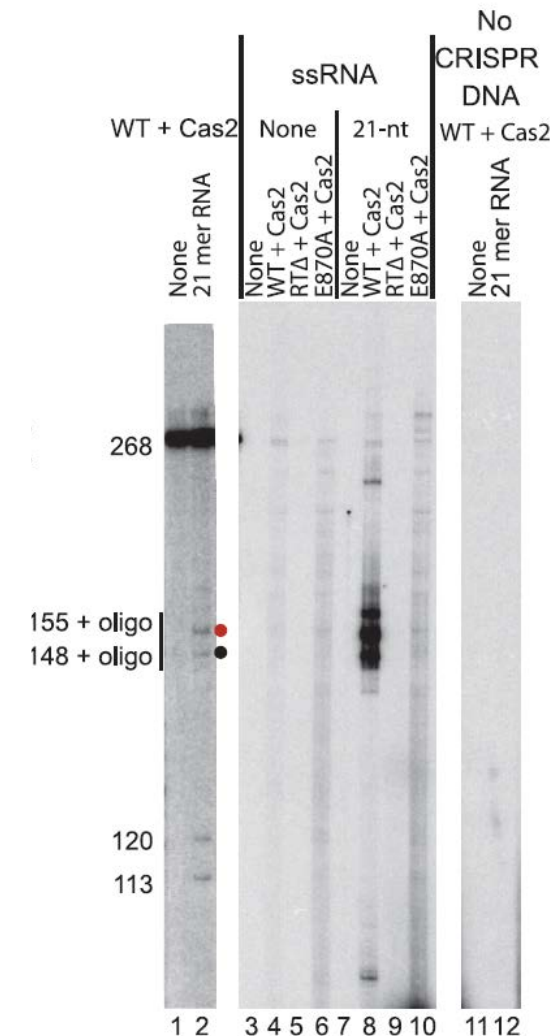
Integration of 21nt ssRNA

A Schematic of cDNA synthesis



B cDNA synthesis using cleaved

CRISPR DNA ligated to 21-nt ssRNA



WT or mutant RT-Cas1 plus Cas2 proteins were incubated with 268-bp CRISPR DNA in the presence of 21-nt RNA oligonucleotide, labeled dCTP, and unlabeled dATP, dGTP, and dTTP.

Key points Papers 1 & 2

- system can capture and stably store practical amounts of real data within the genomes of populations of living cells.
- CRISPR arrays are capable of inserting DNA snippets, yielding temporal information about an event.
- MMB1 RT-Cas1 fusion protein can mediate the direct acquisition of spacers from donor RNA, using the Cas1 integrase activity to directly ligate an RNA protospacer into CRISPR DNA repeats.

Papers

1.) CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria

2.) Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein

3.) Transcriptional recording by CRISPR spacer acquisition from RNA

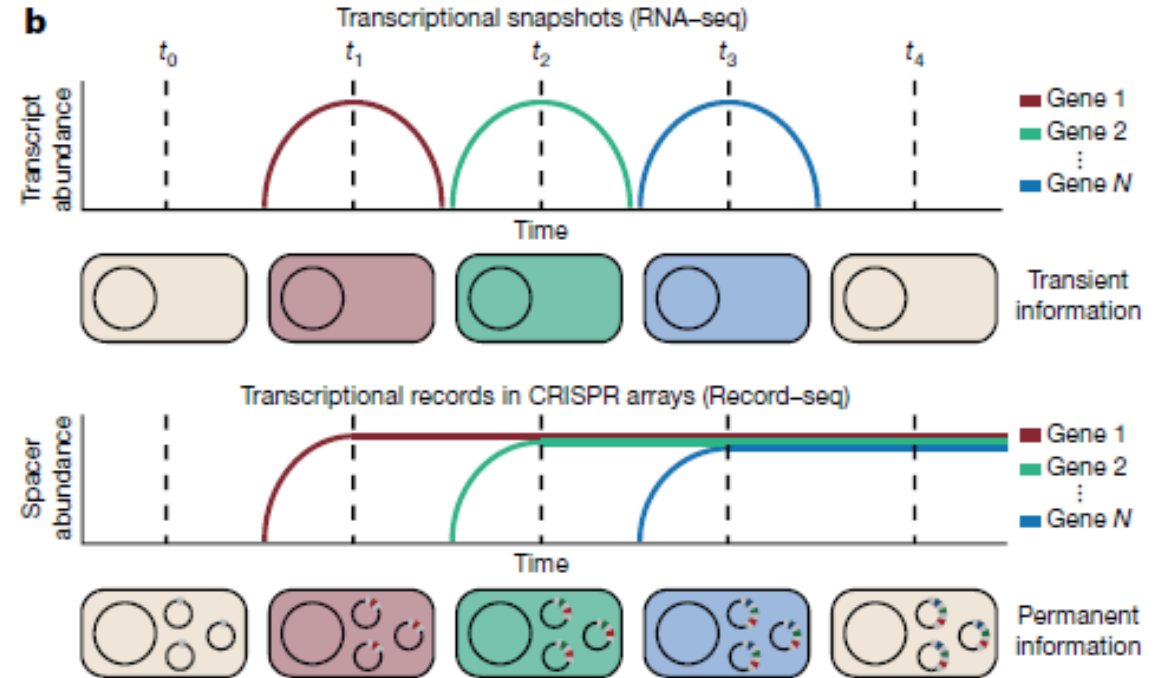
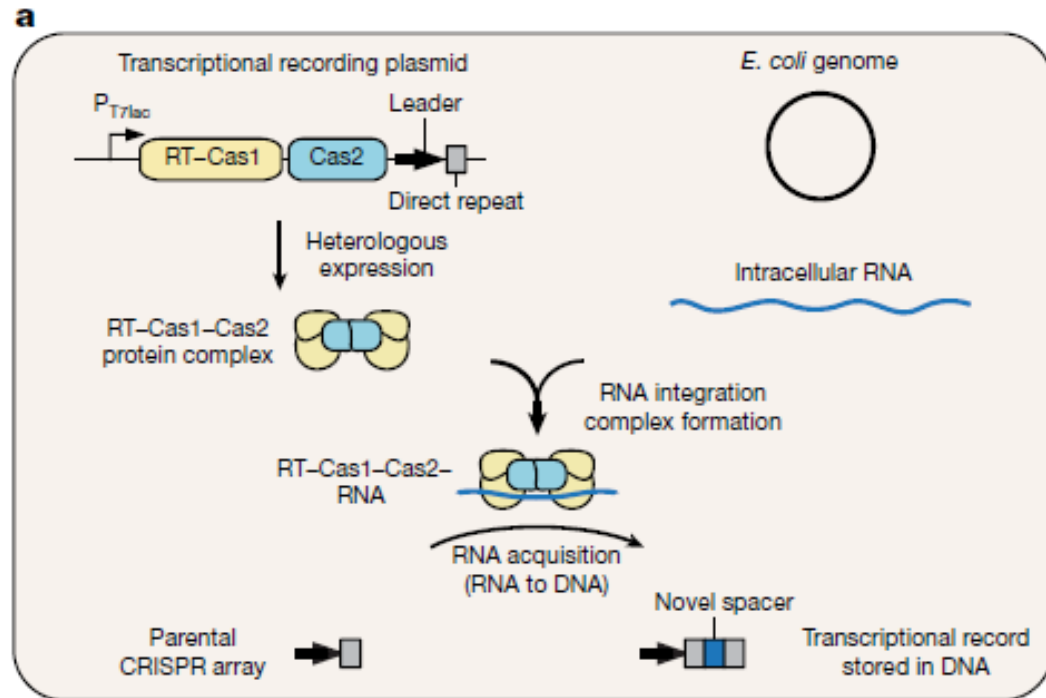
Transcriptional recording by CRISPR spacer acquisition from RNA

Florian Schmidt¹, Mariia Y. Cherepkova¹ & Randall J. Platt^{1,2*}

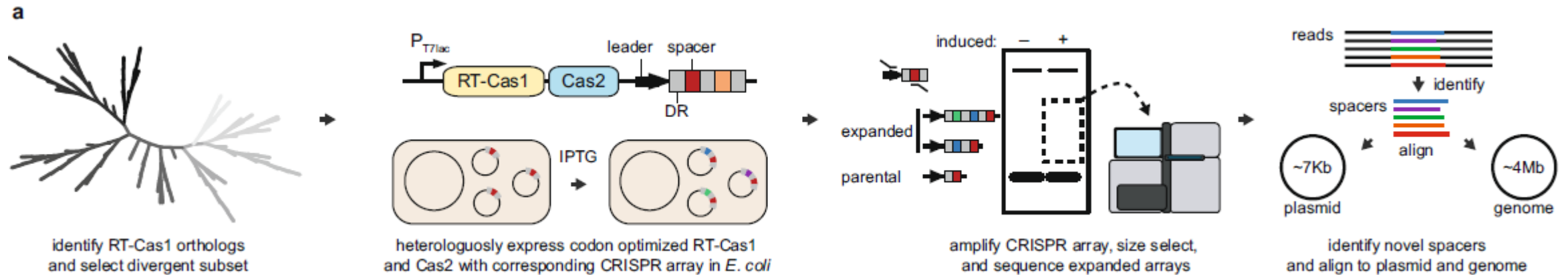
2018

- Can the presented systems be repurposed to acquire spacers from endogenous RNA allowing to store transcriptional information over time in CRISPR arrays?

RNA derived spacer acquisition?



MMB-1 did not work in E.Coli, alternatives?

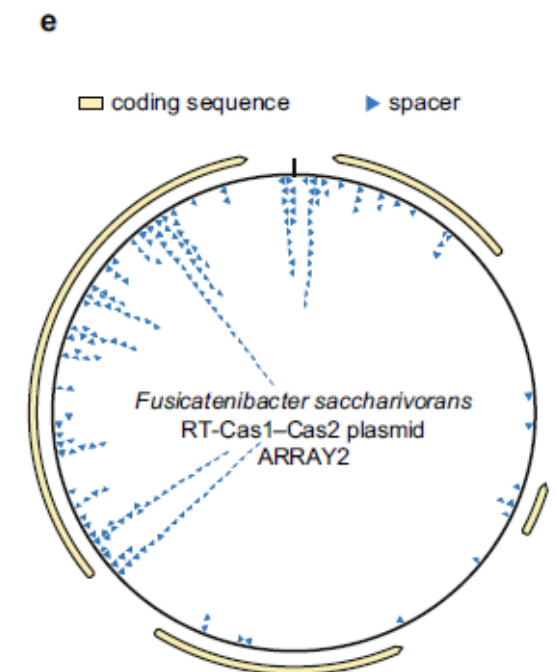
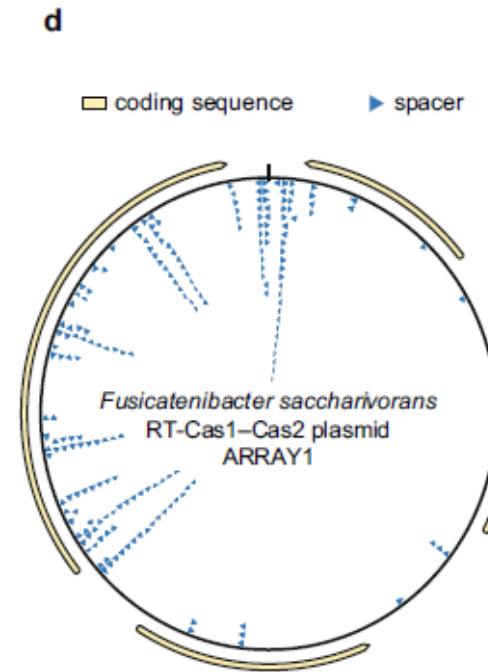
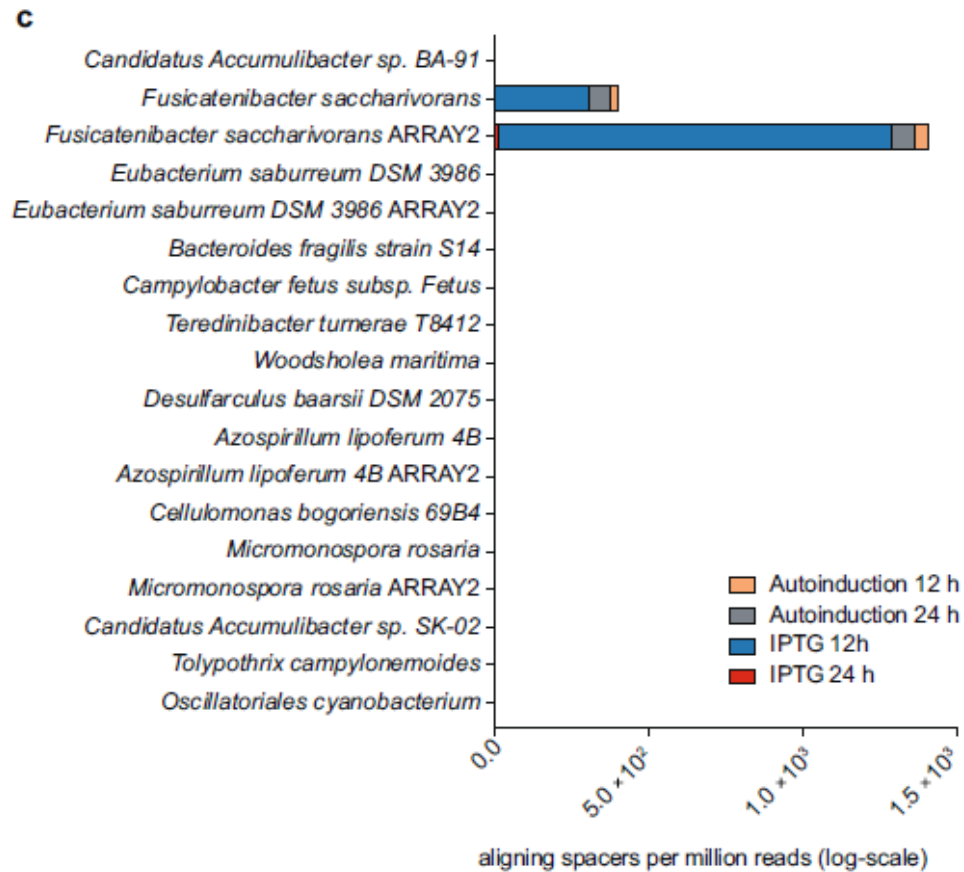


121 RT-Cas1 orthologs were identified, 14 of which were chosen for functional characterization

According RT-Cas1-Cas2 was overexpressed, and expression was induced

Spacer acquisition was determined by amplification, size selection and sequencing of expanded arrays

MMB-1 did not work in E.Coli, alternatives?

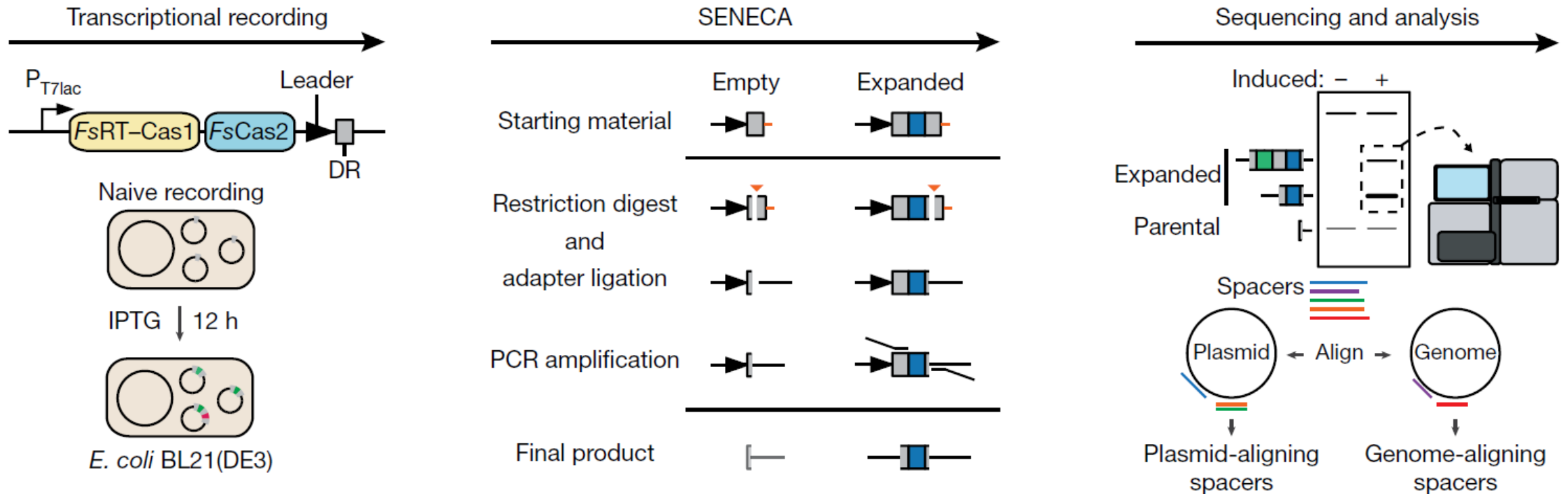


RT-Cas1-Cas2 of *Fusicateribacter saccharivorans* was the only Ortholog that acquired Spacers in E.Coli.

=>

FsRT-Cas1-FsCas2

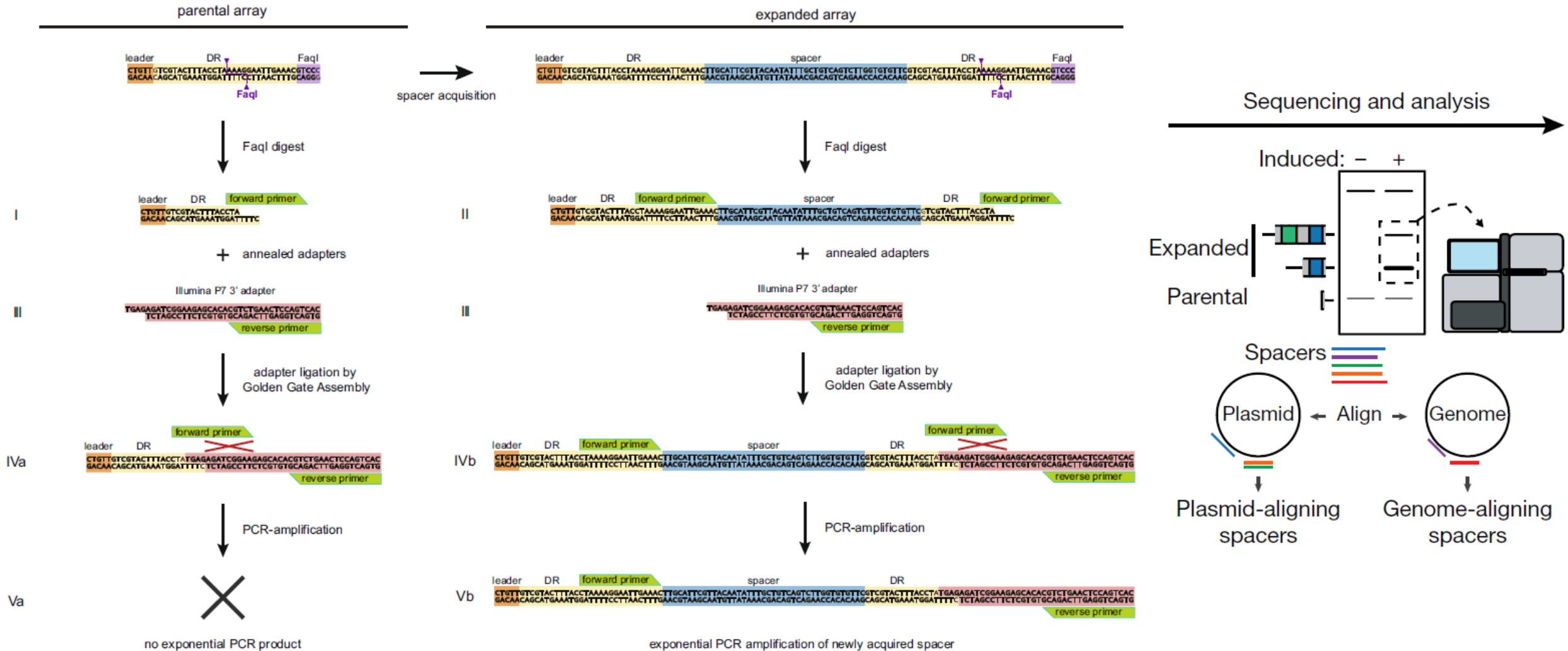
Workflow of Record-seq



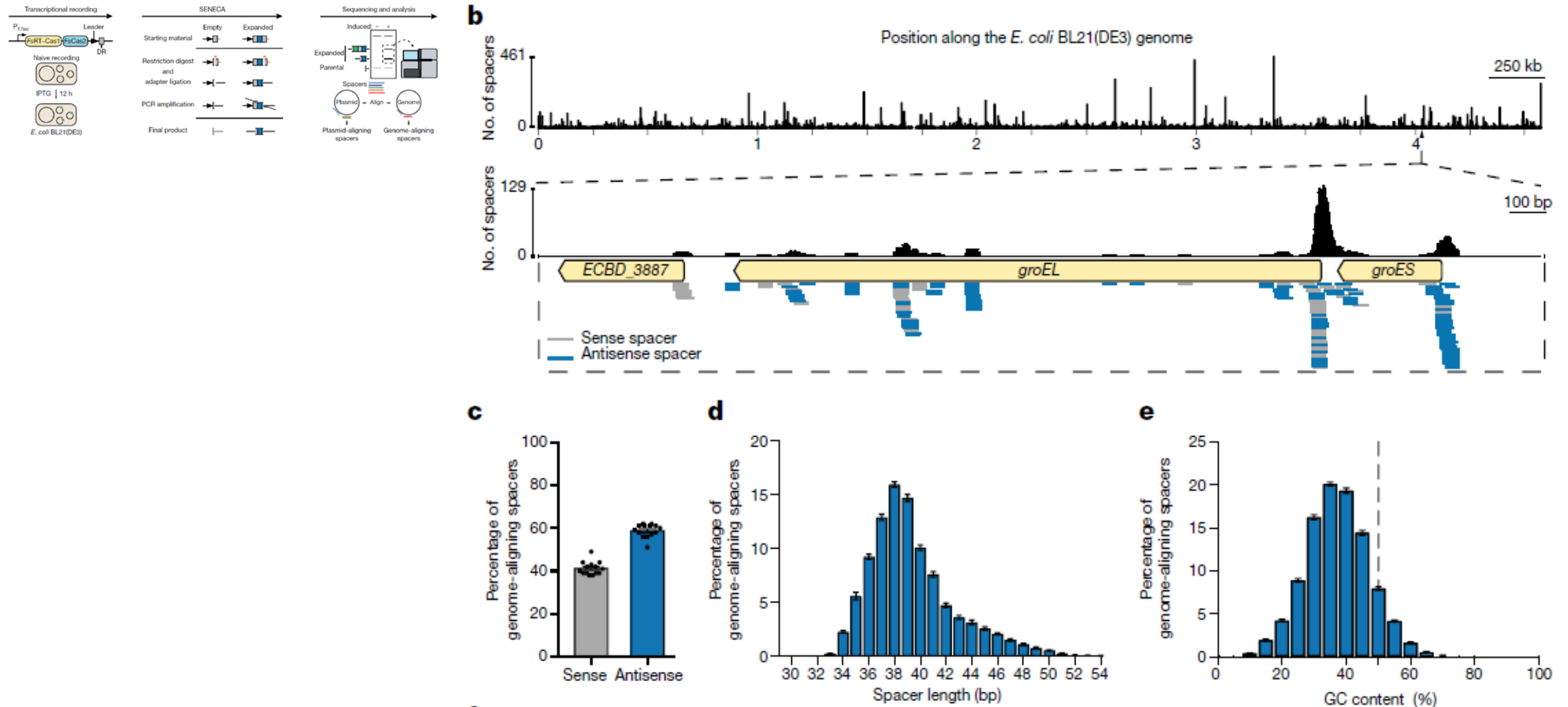
Workflow of Record-seq:

- 1.) Induce expression of *FsRT-Cas1-Cas2* over 12h
- 2.) SENECA ('selective amplification of expanded CRISPR arrays; see next slide)
- 3.) Size selection, sequencing and analysis by alignment to available transcripts

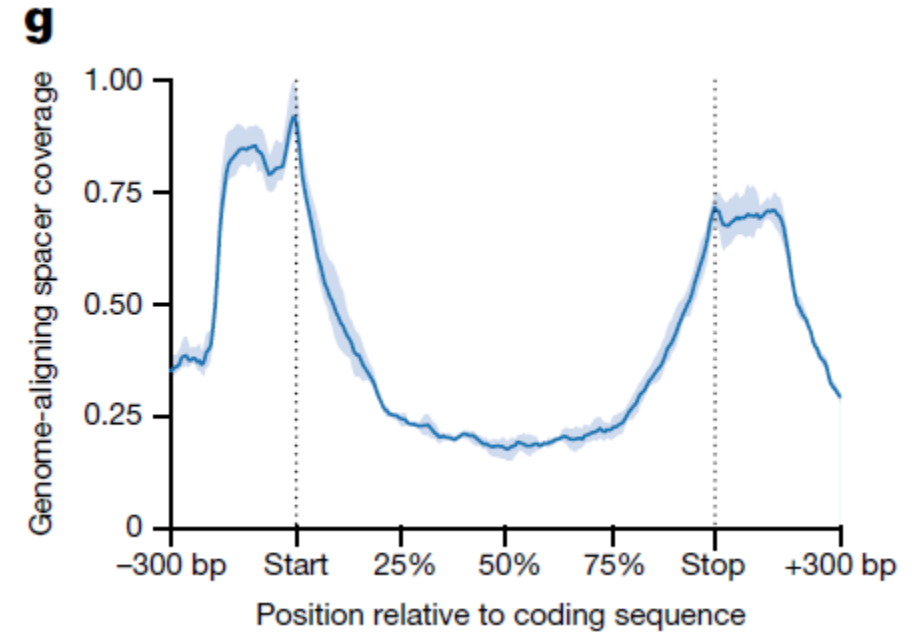
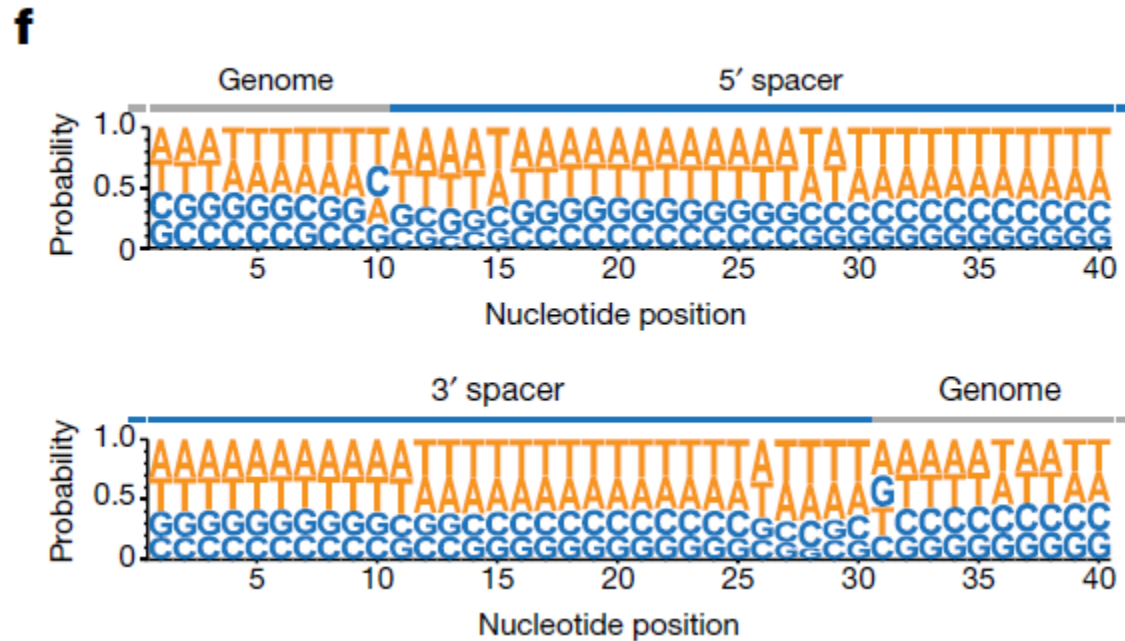
SENECA



Characteristics of FsRT–Cas1–Cas2 spacer acquisition

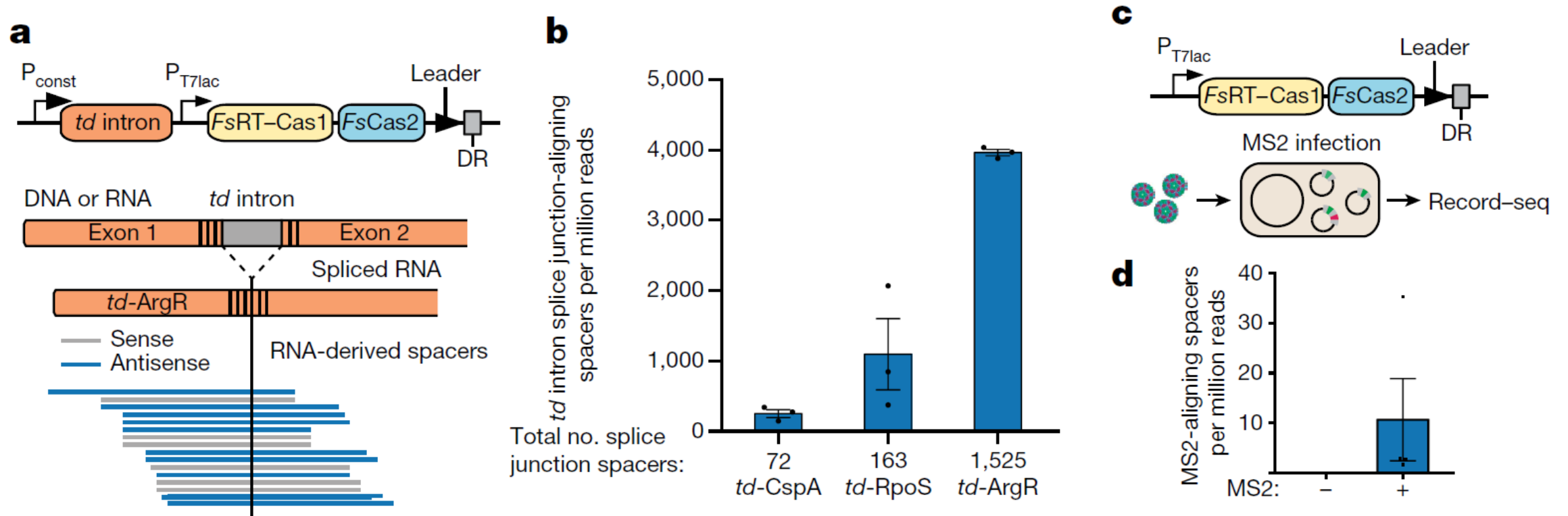


Characteristics of FsRT–Cas1–Cas2 spacer acquisition



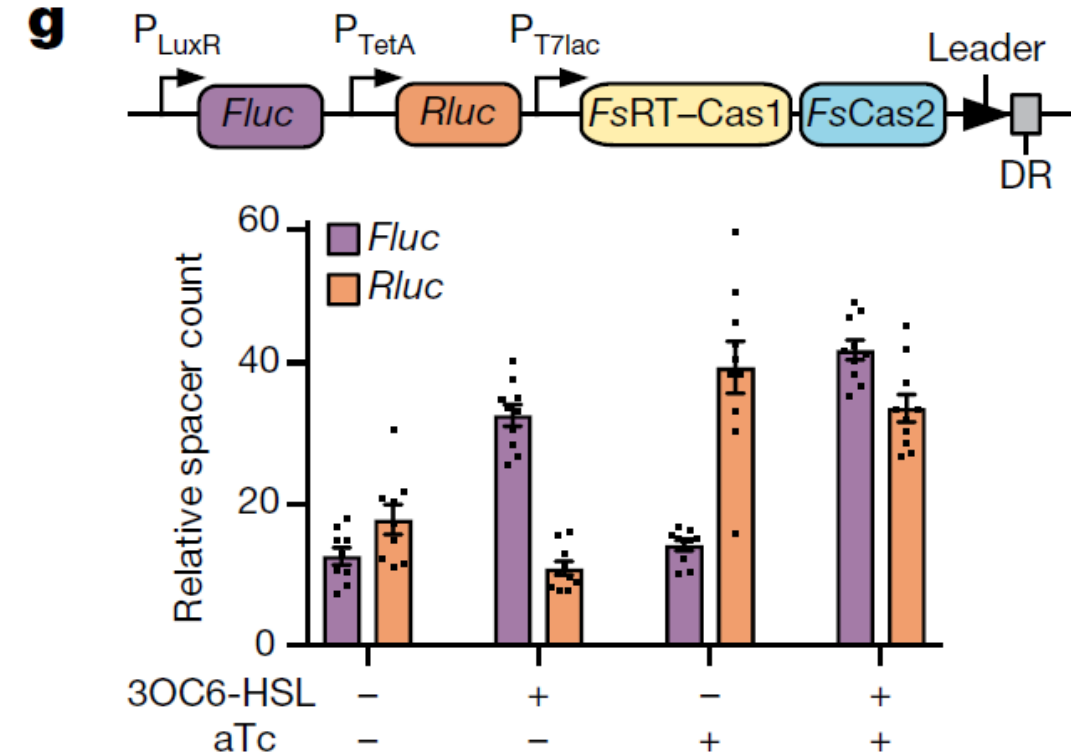
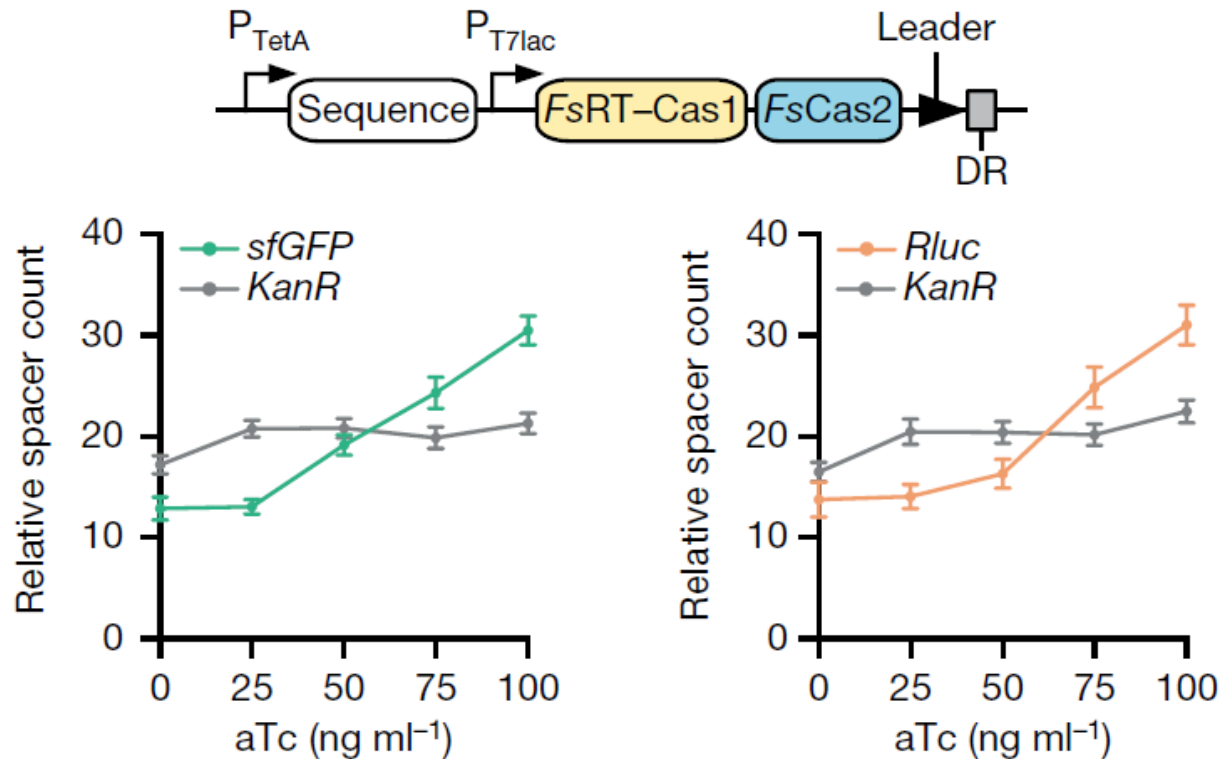
No PAM necessary
(has been previously described for type III CRISPRs systems)

FsRT–Cas1–Cas2 acquires spacers directly from RNA



Infection with RNA virus
With no sequence similarity
to the plasmid or the host genome

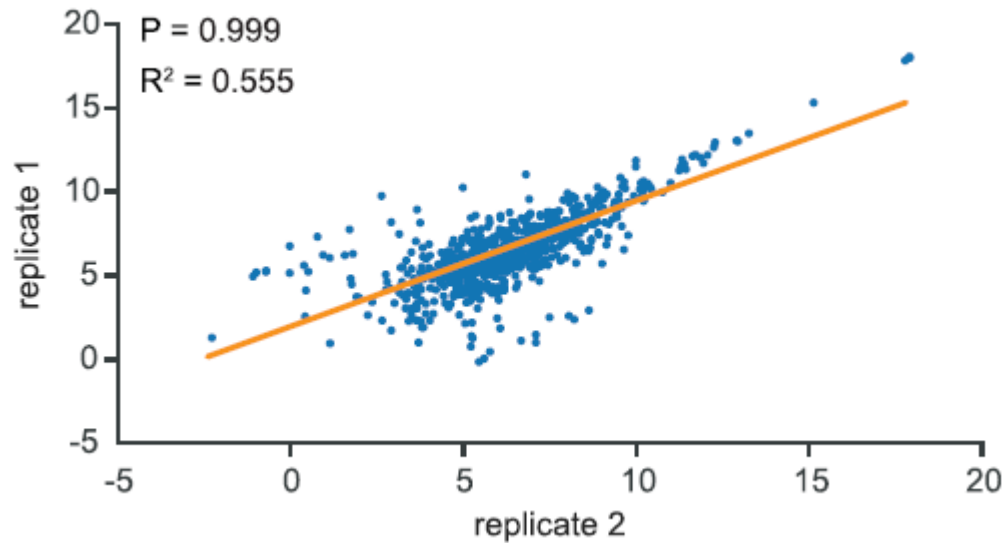
Recording of arbitrary transcripts using Record-seq



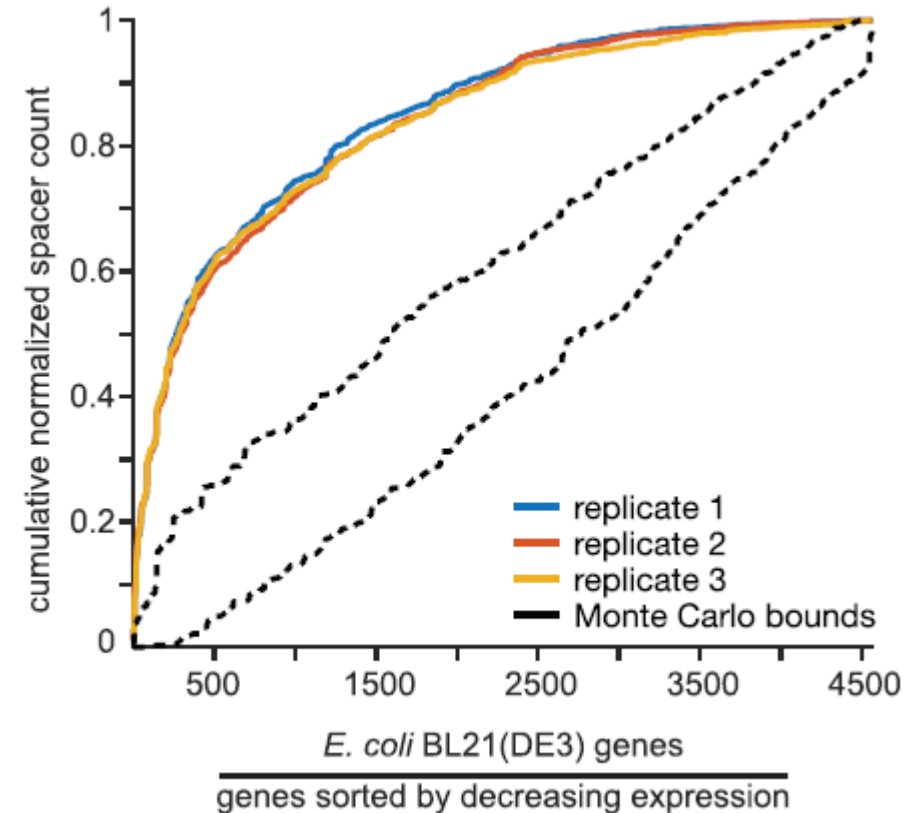
To assess the potential of *FsRT-Cas1-Cas2* for quantitatively recording transcriptional events, they used an inducible expression systems to determine whether spacers were being acquired according to RNA abundance.

Results show that CRISPR spacer acquisition from RNA can generate a quantifiable record of cumulative transcript abundance

Record-seq shows cumulatively highly expressed genes

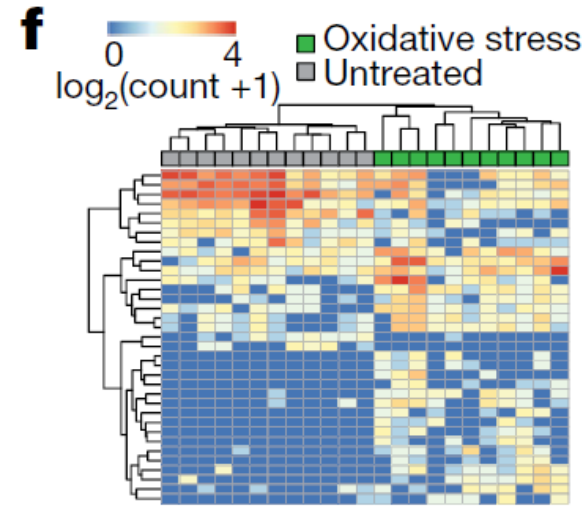
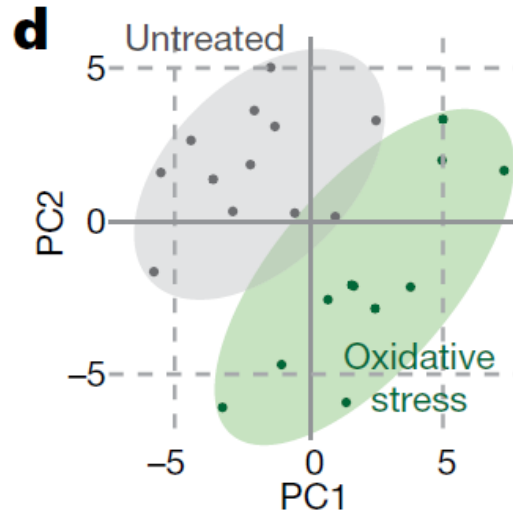
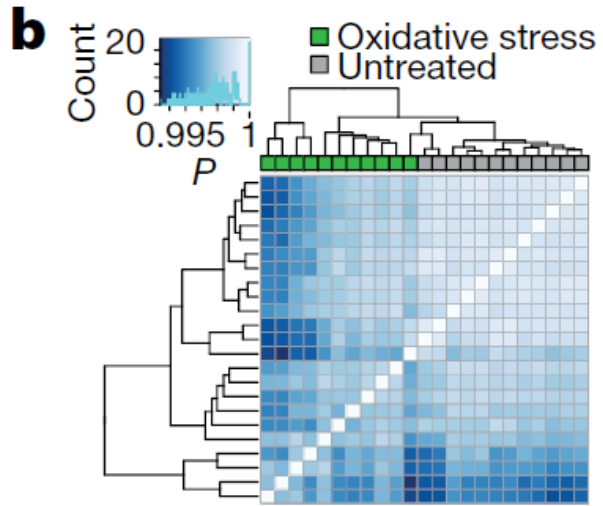


Replicate correlation between two biological replicates



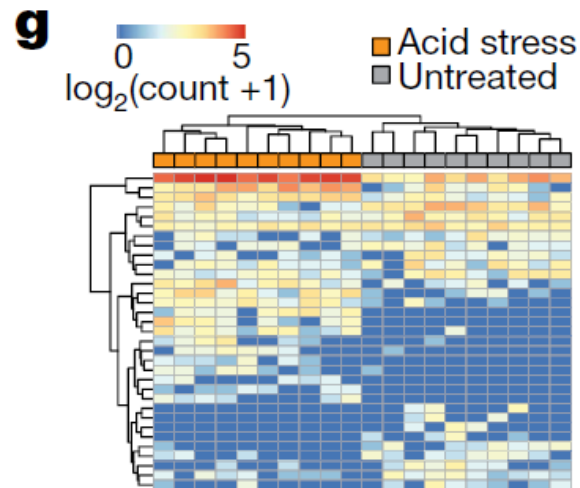
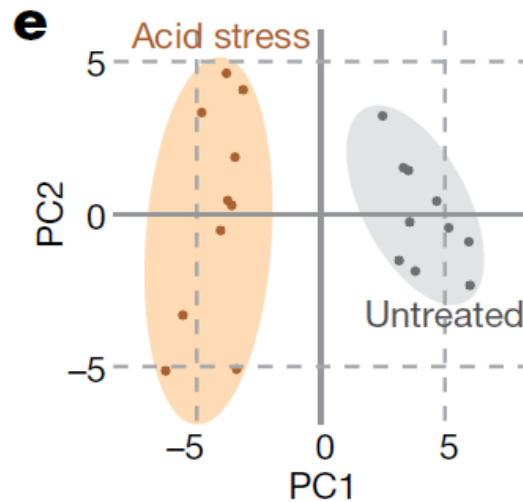
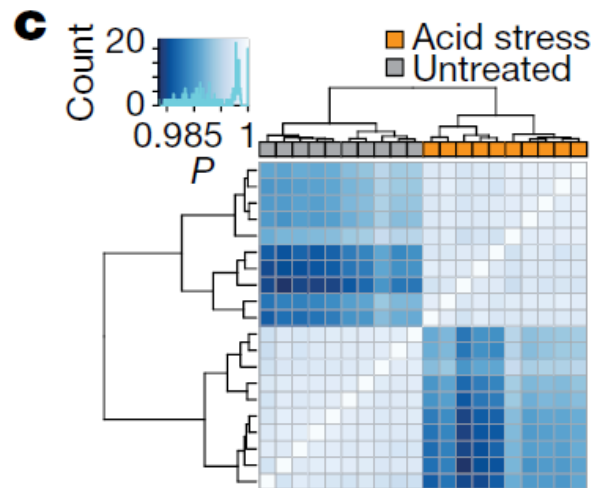
Spacers are preferentially acquired from highly expressed genes

Transcriptome-scale recording reveals cell behaviours



Oxidative stress exposure

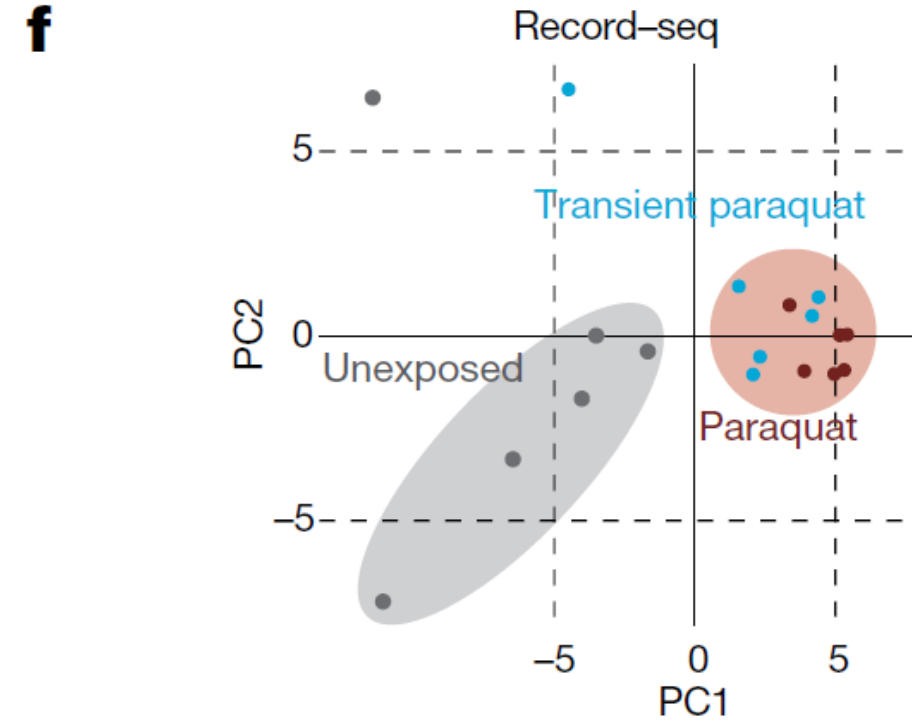
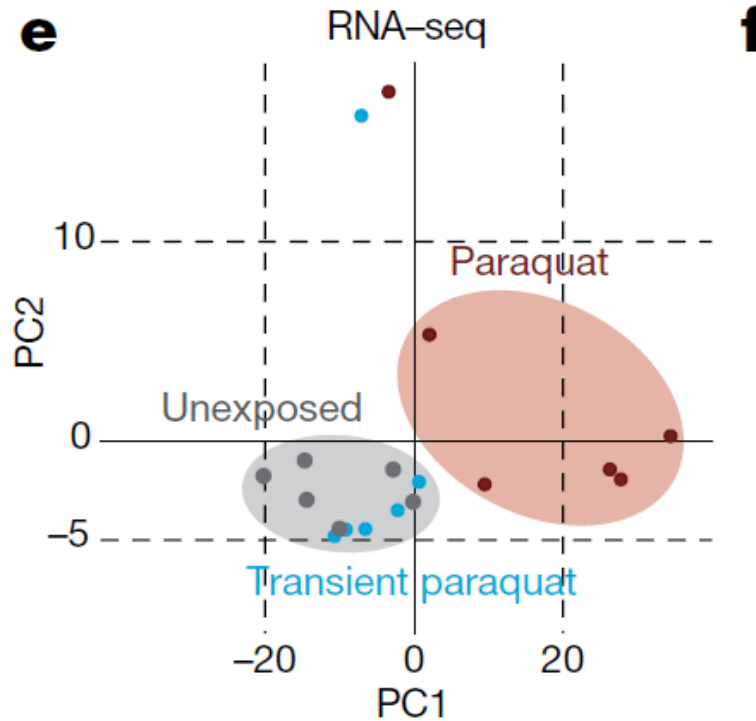
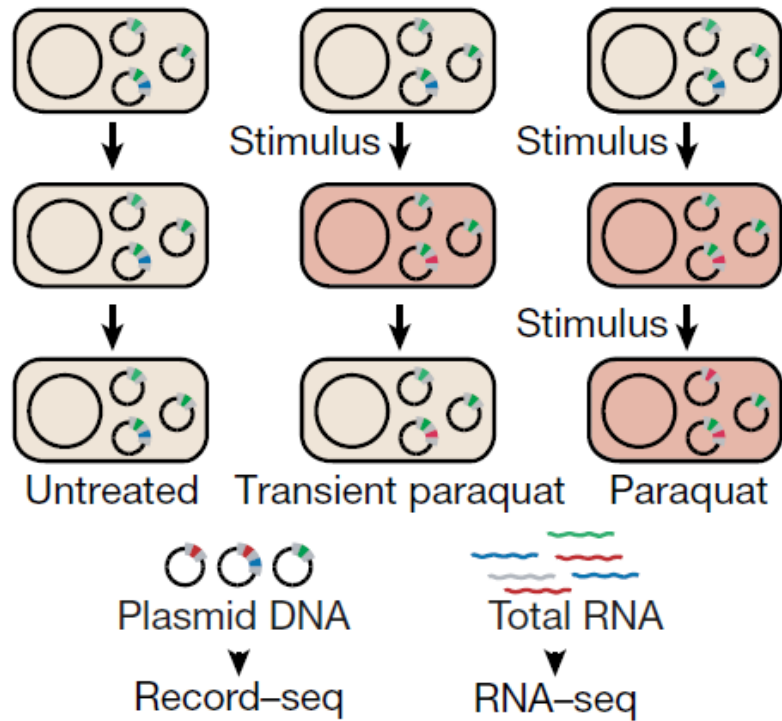
Signature genes for clustering
received via RNA-seq



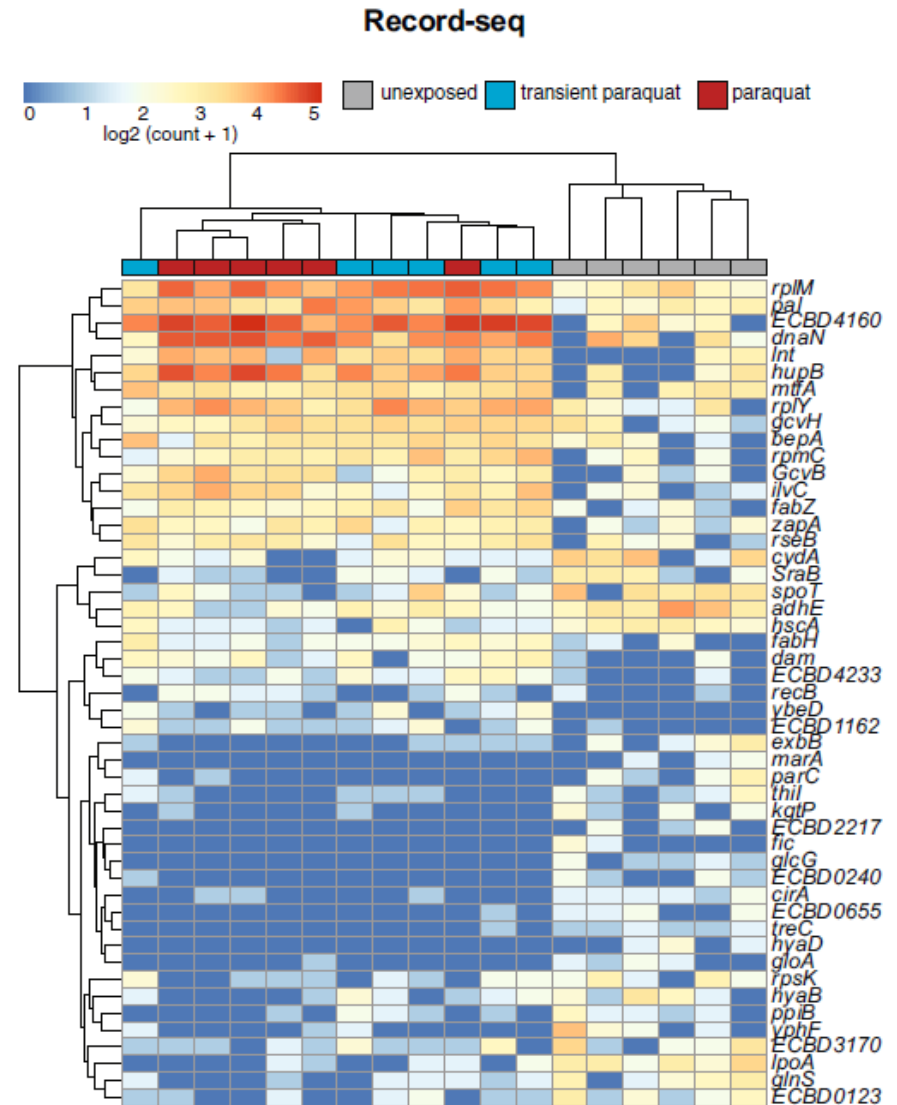
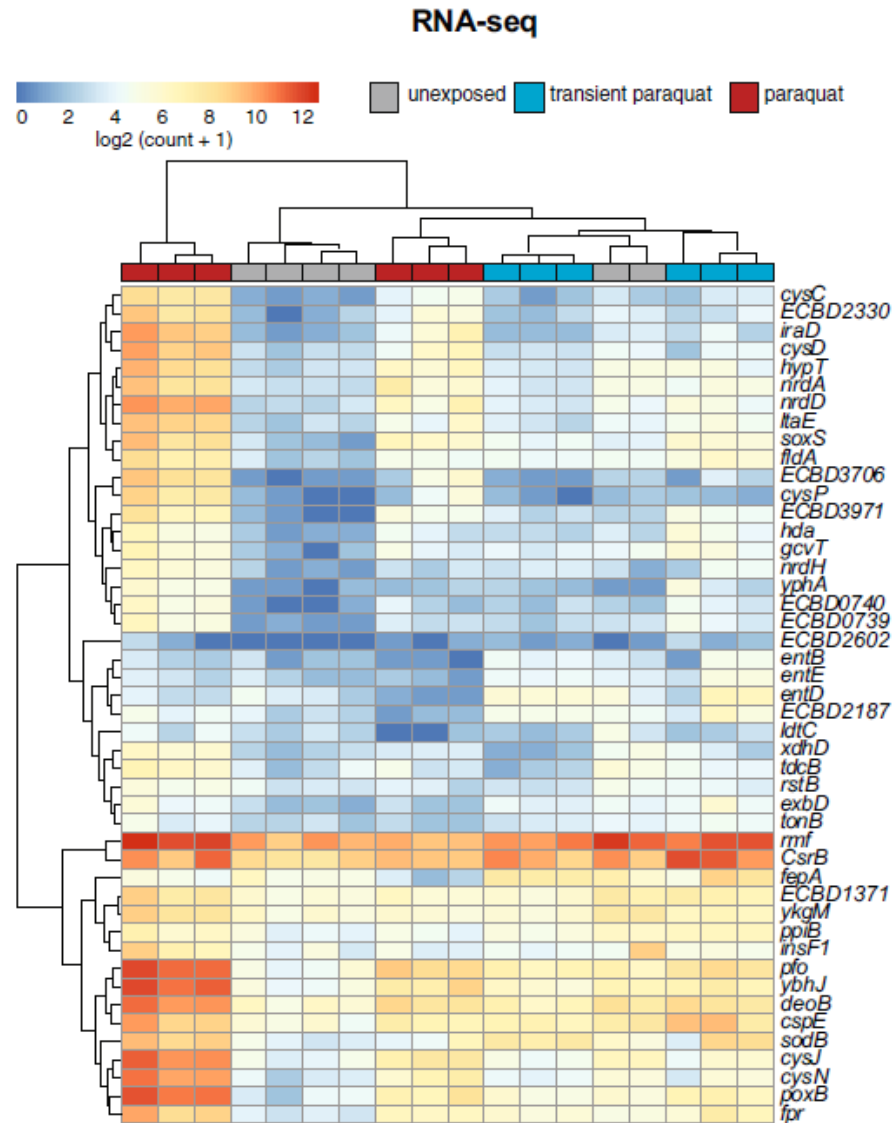
Acid stress exposure

Signature genes for clustering
received via RNA-seq

Sentinel cells encode transient herbicide exposure



Sentinel cells encode transient herbicide exposure



Benefits of Record-seq

- Ability to heterologously express orthologous RT-Cas1-containing CRISPR acquisition systems in order to capture and store RNA species within DNA in an abundance-dependent process;
- Capacity to efficiently and scalably read out molecular histories permanently stored in DNA and reconstruct transcriptome-scale events;
- Application of this technology for recording specific inputs, such as virus infection or any single or orthogonal set of inducible expression system
- Potential applications of this system for creating 'sentinel' cells for medical or biotechnology applications. Even if specific external stimuli cannot be recorded directly, the transcriptome scale molecular signatures recorded within a bacterial population may be sufficient to report meaningful physiological states

Remaining technical challenges of Record-seq

- Majority of spacers are acquired from highly overexpressed plasmid-borne genes, necessitates deeper sequencing when interested in transcriptome-scale events.
- Low efficiency of F_sRT–Cas1–Cas2 CRISPR spacer acquisition from RNA necessitates the use of populations on the order of ten million bacteria for Record-Seq, thereby precluding applications in single cells
- Method is currently semi-quantitative, which could be improved through the implementation of unique molecular identifiers and spike-ins enabling absolute quantification.
- Low efficiency of Type III CRISPR spacer acquisition in general also leads to only a minor fraction of CRISPR arrays acquiring more than a single spacer, and thus valuable temporal information is currently not preserved.
- Limitations of host cells to tolerate and maintain large DNA records of dynamic transcriptome-scale information within single cells and the computational framework to reconstruct meaningful transcriptional and lineage histories.

These challenges currently preclude transcriptome-wide recordings within single cells akin to the current state of RNA sequencing technologies. Despite these challenges, Record-Seq facilitates transcriptome-scale recordings within a population of bacteria.

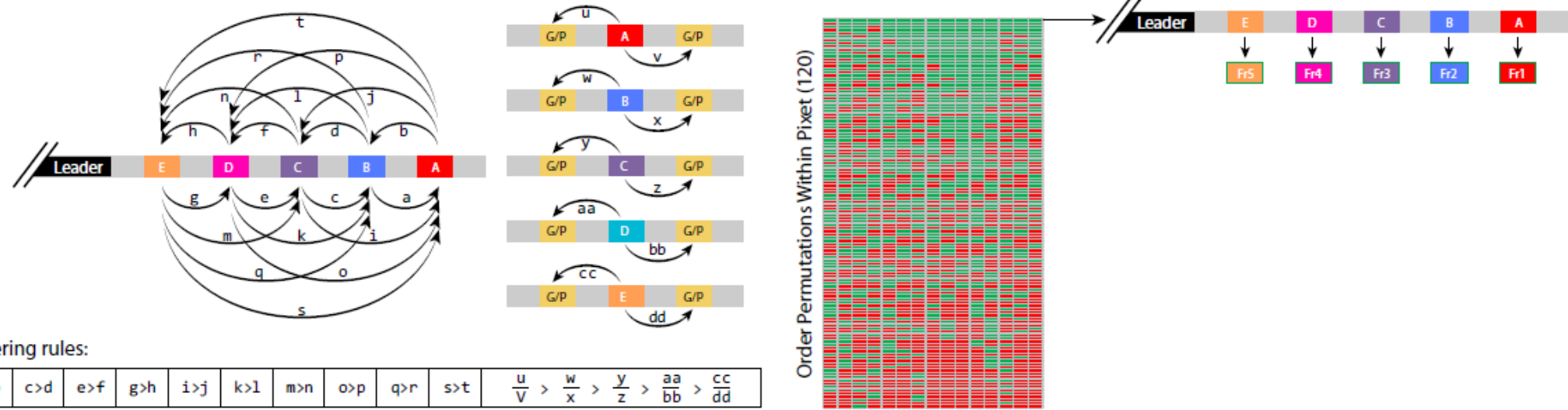
Outlook Record-seq

- CRISPR spacer acquisition components could be introduced into other cell types (including eukaryotic cells) to record the molecular sequences of events, and lineage paths, that gives rise to particular cell behaviours, cell states and types
- Usage of cells engineered to perform Record-seq to monitor gene expression in difficult-to-access environments, such as the human gut, or to identify gene-expression profiles that are a signature of disease or abnormality

Thank you for your attention!

a

Within a pixet



b

Between Pixels

