A thousand needles in a haystack

The search for invading DNA sequences by the CRISPR immune system

Vink, J.N.A.

DOI
10.4233/uuid:e0fa7522-0380-4ec9-9bc6-7568564b72bf

Publication date
2021

Document Version
Final published version

Citation (APA)
Vink, J. N. A. (2021). *A thousand needles in a haystack: The search for invading DNA sequences by the CRISPR immune system*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:e0fa7522-0380-4ec9-9bc6-7568564b72bf

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

A thousand needles in a haystack: The search for invading DNA sequences by the CRISPR immune system

Dissertation

for the purpose of obtaining the degree of doctor

at Delft University of Technology

by the authority of the Rector Magnificus, Prof.dr.ir. T.H.J.J. van der Hagen

chair of the Board for Doctorates

to be defended publicly on Monday 20 September 2021 at 15:00 o'clock

by

Joachim Nicolaas Alexander VINK.

Master of Science in Molecular Life Sciences,

Wageningen University & Research, the Netherlands

born in Nijmegen, the Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Dr.ir. S.J.J. Brouns	Delft University of Technology, promotor
Dr. J. Hohlbein	Wageningen University & Research, copromoter

Independent members:

Prof.dr. A.M. Dogterom	Delft University of Technology
Prof.dr. J. van der Oost	Wageningen University & Research
Prof.dr. S.M. Depken	Delft University of Technology
Prof.dr. J. Elf	Uppsala Universitet
Dr. U. Endesfelder	Rheinische Friedrich-Wilhelms-Universität Bonn

Casimir PhD series 2021-20

ISBN: 978-90-8593-486-8

An electronic version of this dissertation is available at:

http://repository.tudelft.nl

Voor Will van de Ven

We helpen elkaar stap voor stap

terug te gaan naar hoe het was

Zo wordt het pad, eerst bijna

niet te ontwaren,

aangestampt.

Met elke tred

verstevig je, bevestig je

dat het begaanbaar is, het het

Table of Contents

1

Finding a match

Start looking

Every piece of DNA we find in the natural world around us today has been selected and was maybe not the 'fittest' but at least 'fit enough' at the time of its creation. The genetic code present on this DNA therefore encodes for something that helps this DNA maximize the resources it has to multiply, while minimizing the resources taken from them by other pieces of DNA to do the same. In this context, all biological phenomena we can observe, from tree saplings competing for light, lion packs hunting buffalo, or the occurrence of influenza outbreaks can be summarized as a contest of DNA replication (Dawkins, 1976; Wilson, 1975).

One example in which this competition on the DNA level is very direct and evident, is found in the immune systems of prokaryotic organisms (Bernheim and Sorek, 2019). These organisms are continuously invaded by mobile DNA, mostly found as bacteriophages (prokaryotic viruses) and plasmids. In some cases, foreign DNA invasion can increase the fitness of the host DNA (for example when plasmids encode antibiotic resistance genes (Dimitriu et al., 2016)), but in other cases, this additional DNA is a burden (Millan et al., 2015) or a threat and decreases the replicative success of the host DNA. Therefore, the host DNA encodes a large range of immune systems that recognize and limit foreign DNA entry and replication.

CRISPR, short for Clustered Regularly Interspaced Palindromic Repeats, is an immune system that uses protein complexes that can be programmed to target invading sequences of mobile elements (Barrangou et al., 2007; Francisco J.M. Mojica et al., 2005; Ishino et al., 1987; Jansen et al., 2002) (Figure 1A). It is found in 40% of bacterial genomes and 70% of archaeal genomes (Pourcel et al., 2020). It is furthermore present in phage genomes and on plasmids (Bernheim et al., 2019; Pinilla-Redondo et al., 2019). The diversity of these systems is large, where some systems target DNA (Brouns et al., 2008), and others RNA (Hale et al., 2009). Based on the effector complex, CRISPR systems have been subdivided into two classes (Figure 1B). In Class I systems the effector complex consists of a multi-subunit complex, whereas in Class 2 systems the effector complex is a single protein that binds the crRNA (Mohanraju et al., 2016). A further combination of genes will determine in which of the six described types (e.g. Type I, Type III) and 33 subtypes the locus falls (e.g. Type I-E, Type III-A). The main difference

with other prokaryotic immune systems discovered so far, is its high specificity and its adaptive capabilities.

The high specificity of CRISPR systems also has its cost. First of all, mutating invading DNA elements can obstruct recognition and overcome the immune system (Künne et al., 2018). To combat this, some CRISPR systems have evolved a priming mechanism, in which mutated targets lead to accelerated acquisition of spacers against the same invader (Datsenko et al., 2012; Fineran et al., 2014; Nicholson et al., 2019). Secondly, the number of invaders that can be dealt with is limited by the size of the CRISPR array. As this is in general between 10-100 spacers, this requires the system to discard spacers no longer useful with more recently important invaders (Horvath et al., 2008; Lam and Ye, 2019; Lopez-Sanchez et al., 2012). Lastly, the specificity of CRISPR-Cas also gives it a dynamic disadvantage, because the 30-nt sequence that is recognized by CRISPR-Cas systems occurs much less frequently in an invader sequence compared to a short 6-nt restriction recognition site and therefore takes longer to find. However, the impact of this target search process is not well understood.

The goal of this thesis is to study the impact of target search dynamics on the functioning of CRISPR-Cas immunity. In this introduction, I will describe the molecular organisation of CRISPR-Cas systems, the target search dynamics of proteins in the cell, and I will further describe the experimental biophysical and theoretical bio-informatics techniques that were used to study the dynamic aspects of CRISPR-Cas immune systems.

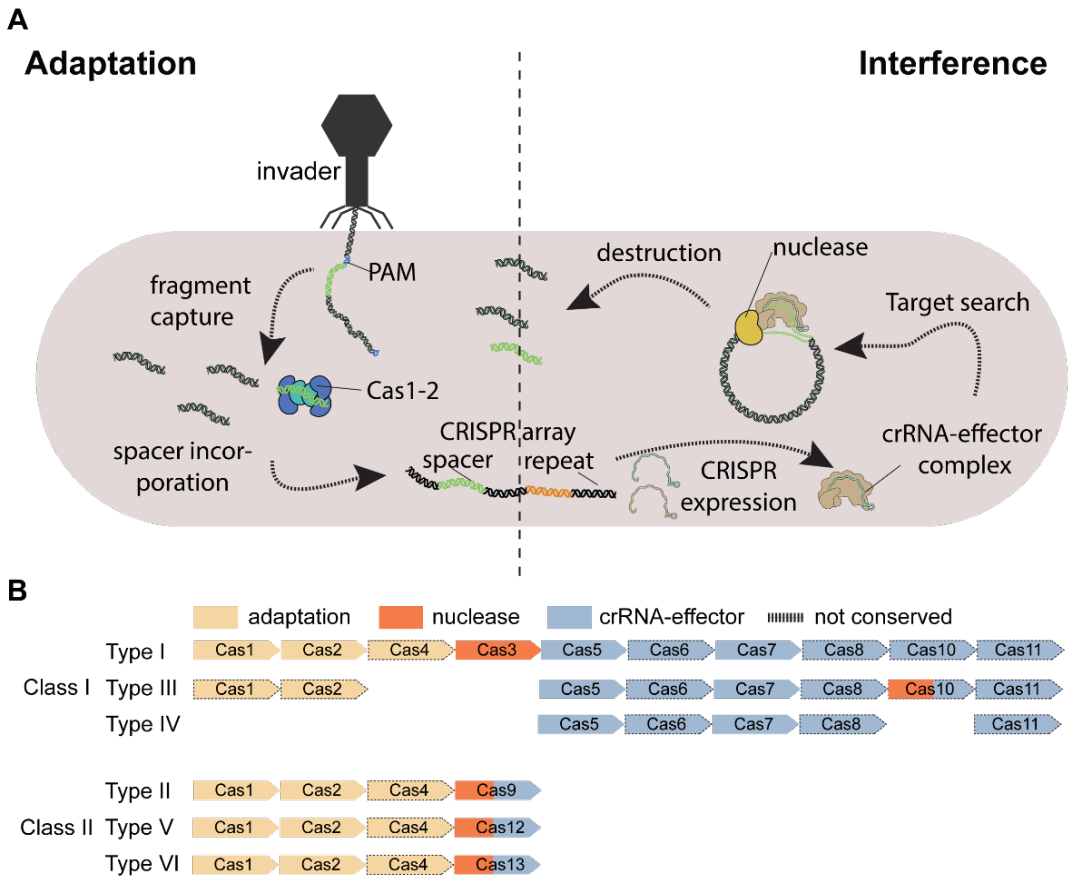


Figure 1. Overview of CRISPR-Cas mechanism and classification. (A) Overview of CRISPR adaptation and interference. The schematic is based on the mechanism of the Type I-E system, other subtypes might have differences in proteins involved and could target RNA instead of DNA. (B) Classification of CRISPR-Cas according to the proteins involved in adaptation, nuclease activity and proteins that form the crRNA-effector complex.

CRISPR-Cas

The CRISPR-Cas system consists of several common modules. An array of CRISPR spacers, where each spacer matches a previously encountered foreign DNA element. For correct processing and storage of these sequences, the spacers are separated by a common repeat. CRISPR-Cas systems also encode an adaptation module, which can integrate new spacers. It furthermore encodes an effector complex, carrying RNA copies of the spacer (crRNA) which scans the cell for a matching invader sequence. Then many systems also contain additional proteins, that either aid the adaptation module, are used in crRNA processing, or nucleases that are recruited by the effector complexes to degrade the DNA.

The CRISPR array, the unit that led to the discovery of the immune system, is the defining feature of CRISPR-Cas mechanism. The array is transcribed from a leader sequence, a usually AT-rich region, which also guides newly acquired spacers to be incorporated at the leader-repeat junction (Kieper et al., 2019; McGinn and Marraffini, 2016; Wei et al., 2015), therefore ordering the spacers chronologically. After transcription, the RNA has to be processed to form separate crRNAs that can be loaded in the effector complex (Brouns et al., 2008; Carte et al., 2008). In many systems the repeat sequences, roughly 30 nt in length, are palindromic, which causes the transcribed crRNA to form hairpin structures. These hairpin structures can help proteins to recognize these crRNAs for processing and effector complex formation (Li, 2015; Niewoehner et al., 2014).

The adaptation module is the most conserved module of the CRISPR-Cas system (Makarova et al., 2019). It almost always consists of a Cas1-Cas2 complex. The function of the adaptation complex is to pick up fragments from the invading DNA and integrate this into the CRISPR array (Jackson et al., 2017). For recognition, the effector complex requires that these fragments (protospacers) are flanked by an oligonucleotide motif called the PAM (Protospacer Adjacent Motif) (Mojica et al., 2009). Therefore, the Cas1-Cas2 complex needs to selectively integrate those spacers in the right orientation, to generate functional spacers (Kim et al., 2020; Yoganand et al., 2019). There are furthermore Cas1-Cas2 associated proteins (both fused and unfused), which can increase incorporation of the right PAM sequence (Cas4; (Kieper et al., 2018)), allow RNA protospacers to be incorporated (Reverse Transcriptase;(Silas et al., 2016)) and form even larger adaptation complexes (Csn2; (Wilkinson et al., 2019)). Also other host factors can play a role in the adaptation process (IHF; (Nuñez et al., 2016)).

The goal of the effector complex is to bind the target matching the crRNA it carries. Effector complexes often contain the nuclease domain that enables destruction of the targeted nucleic acid. But in Type I systems the effector complex recruits a nuclease that destroys the DNA (Sinkunas et al., 2011; Westra et al., 2012). The target nucleic acid is not always the only target: upon binding, Cas13a also starts cleaving other RNA present in the cell, which is believed to help to combat phage spreading through a population (Abudayyeh et al., 2016; Meeske et al., 2019). In Type III RNA-targeting systems, the effector complex also contains a domain producing cyclic oligoadenylates, which are

signaling molecules activating other host defence systems (Kazlauskiene et al., 2017; Niewoehner et al., 2017). Other effector complexes are fused to transposases, which enables the transposon to selectively integrate itself into the genome by using the spacer of the CRISPR array (Klompe et al., 2019; Strecker et al., 2019).

Whatever the type of effector complex, all complexes are programmed to find the target matching the spacer inside the complex environment of the prokaryotic cell. It is thought that the PAM, next to avoiding self-targeting of the CRISPR array, also helps the search process of the effector complexes, by limiting the number of potential target sites. Still, the numbers of potential sites remain high, considering at least 100.000 PAMs in a total prokaryotic genome to be scanned. How does a CRISPR effector complex achieve this?

Target Search

All proteins in the cell work by binding to other molecules, whether they are used for structure, replication, sensing or transport (Alberts, 2002). As DNA and RNA are carriers of the genetic code, it is no surprise that so many proteins act on them. Some proteins are ubiquitous binding proteins. In *E. coli*, the protein HU covers large parts of the chromosome and recognizes AT-rich areas (Bonnefoy and Rouvière-Yaniv, 1992; Luijsterburg et al., 2006). Most proteins have more specific target sites. RNA Polymerases recognize certain sequences in promoter regions of a gene to find the starting place for transcription (Hahn, 2004; Pribnow, 1975). Transcription factors are often targeting a small 9-15 bp sequence close to the start site of transcription (Jayaram et al., 2016). Ribosomes scan along mRNA molecules to find the starting codon which is generally 5'-AUG-3' (Lind and Åqvist, 2016). Often the target search time is crucial for the functioning of the protein in question. For example in homologous repair, the complex needs to sufficiently recruit a homologous piece of DNA before cells are irreparably damaged. It was found that this can occur within 5 minutes (Gynnå et al., 2020; Lesterlin et al., 2014).

How do proteins move around the cell to scan for the cognate sequence in a cell that is so packed with nucleic acids? The most simplistic model would be to consider only 3D-Diffusion (Redding and Greene, 2013). In this scenario protein collides with DNA at random positions and each time checks only that site before returning to the cytoplasm until the next collision with another site (Figure 2A). However, the measured association

rates for some DNA binding proteins were two orders of magnitude faster than predicted by such a model (Hammar et al., 2012; Riggs et al., 1970). An alternative model is to include multiple modes of diffusion where after encountering a non-specific DNA site, the protein does not automatically return to solution but scans the vicinity of the DNA. This was termed facilitated diffusion (Berg et al., 1981) and multiple modes of facilitated diffusion are recognized today. Proteins can undergo *sliding* which is defined as a continuous tight interaction between protein and nucleic acid which contrasts *hopping*, where micro-dissociations take place, but a protein re-associates in the vicinity of the previously bound site and *intersegmental transfer*, where a protein is at one point bound to two segments at the same time before moving to the next site (Cui and Joo, 2019)(Figure 2A). It was suggested that an exact 1:1 ratio between DNA sliding and 3D diffusion will result in optimal target search (Slutsky and Mirny, 2004). However, expanding the model with other forms of motion results in the existence of a different optimum (Klein et al., 2020). In both models a mixture of 1D and 3D motion is required to achieve optimal target search.

Next to moving from site to site, a second important aspect in target search is the way in which a site can be probed. For proteins, in principal there are two distinct read-out modes: base readout and shape readout (Rohs et al., 2010). In base readout, the chemical features of the nucleotide bases are read via hydrophobic contacts or hydrogen bonds within the major or minor groove. The major groove gives a unique readout for all pairs, whereas a readout from the minor groove cannot distinguish between certain pairs (Slattery et al., 2014)(Figure 2B). For shape readout, either local (size of minor/major groove) or global shape (bending) of DNA can be read (Figure 2C). For proteins that carry nucleic acids, such as CRISPR and HDR, the read-out requires base-pairing between the carried and the target nucleic acid. This requires the opening of the double-stranded DNA to form a D-/R-loop (San Filippo et al., 2008; Szczelkun et al., 2014). It is likely that loop formation and base pairing will take considerably longer than reading out the major or minor groove (Globyte et al., 2018) (Figure 2D). Furthermore the binding to bases in the minor groove (entropy driven) also differs thermodynamically from the major groove (enthalpy driven) (Privalov et al., 2007). This demonstrates the role of probing mode in target search times.

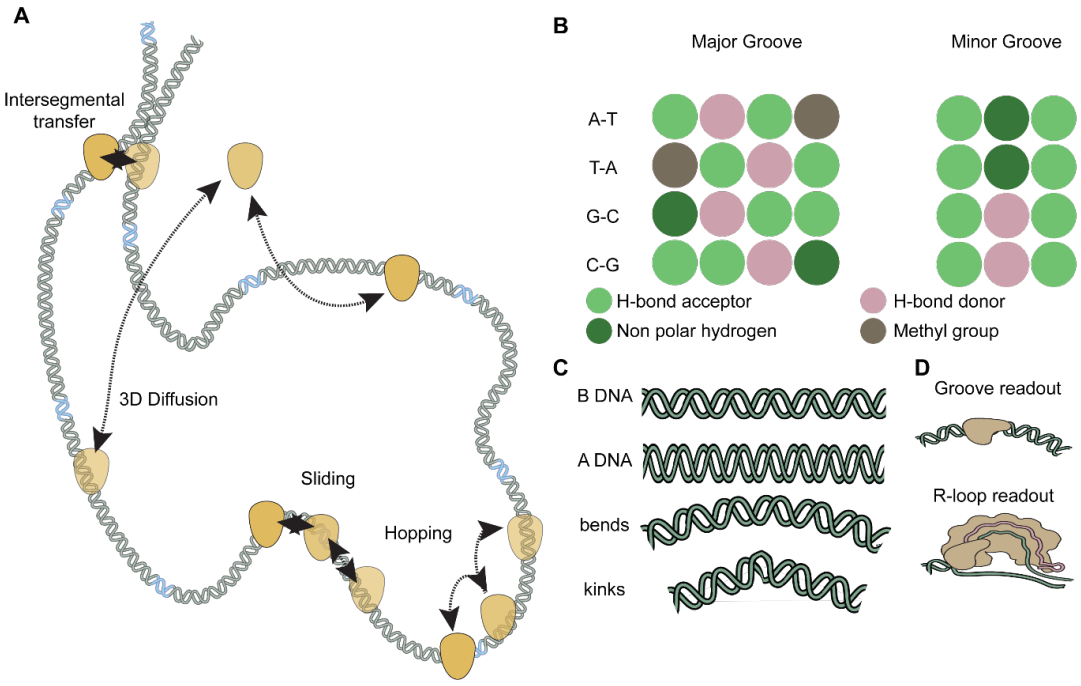


Figure 2. Aspects of DNA target search. (A) Modes of DNA target search. Sliding, hopping and intersegmental transfer are examples of 1D diffusion. (B) Base readout of certain base pairs depends on whether DNA is read from the major or the minor groove. (C) Shape readout can either read out the distance between the grooves, which differs in different confirmations of DNA, or detect bends and kinks in the DNA. (D) Major/minor groove readout occurs without opening DNA strands, whereas readout based on DNA/RNA matching require DNA strands to be separated.

Lastly, the target search can be impacted by copy numbers. Depending on the system, the search for one site can be carried out by multiple proteins at the same time. It was hypothesized that to speed up the search process in HDR, the flanks of the double-stranded break are copied and with many copies the search can be parallelized and sped up (Elf, 2016). Also, polymerases are often present in higher amounts than are involved in repair, indicating that they are required in higher levels to increase the target search speed of mutated DNA regions (Uphoff et al., 2013). Non-coding small RNAs are overabundant in cells as well, which is a likely requirement given the need to provide a rapid regulatory response under stress conditions (Fei et al., 2015).

For CRISPR-Cas systems the target search has been studied in both *in vivo* and *in vitro* systems. It was first observed that undergo solely 3D diffusion target search both *in vitro* (Redding et al., 2015; Sternberg et al., 2014) and *in vivo* (Knight et al., 2015), however other studies under different conditions and with different complexes observed

sliding behaviour (Dillard et al., 2018; Globyte et al., 2018). The PAM interactions are mediated by protein-DNA interactions and either operate on the minor groove (Hayes et al., 2016), major groove (Anders et al., 2014), or both sides (Rollins et al., 2015) dependent on the type of effector complex. The subsequent opening of DNA strand results in base-pairing of the first nucleotides of the crRNA with the potential target in what is called the 'seed region' (Semenova et al., 2011), which are kinetically the most crucial for subsequent target binding (Klein et al., 2018), although PAM-independent strand opening from the opposite direction is also possible at lower rates (Blosser et al., 2015). The multiple modes of diffusion and probing interactions described lead to a qualitative picture of CRISPR target search, but do not explain CRISPR target search in a quantitative context.

The basic search time question can be answered by measuring three quantities. How long does a probing interaction take place? How many interactions are required before finding the target? How many effector complexes are searching? *In vitro* non-specific interactions last between 0.1 and 10 s. (Redding et al., 2015; Xue et al., 2017). These interaction times are unlikely to resemble the interaction times *in vivo*, as this would lead to search times on the order of days, whereas it was found that Cas9 can find a target within the *E. coli* genome in six hours (Jones et al., 2017). This still leaves the question on what timescale do the probing interactions last and how do search times and copy numbers of the effector complexes impact the interference ability of the system? I investigated this question by using *in vivo* single-particle tracking.

Single-particle Tracking

The study of proteins interacting with nucleic acids has undergone a radical transformation in the last 20 years thanks to the development of single-molecule techniques (Candelli et al., 2011; Chaurasiya et al., 2010; Dulin et al., 2013). They allowed the study of heterogeneous dynamics in real-time, which is an important aspect of the functioning of these systems. *In vivo*, the most accessible technique to this date is photoactivatable single-molecule localization microscopy (SMLM) (Kapanidis et al., 2018a; Shashkova and Leake, 2017).

1

SMLM is one of many developed techniques that is potentially able to achieve super-resolution, a better resolution than the diffraction limit of light (Patterson, 2009). SMLM requires that the fluorescence of each emitter is sufficiently separated either in space or time from other nearby emitters to fit the point spread function and determine its location, which increases the resolution depending on the number of photons (Mortensen et al., 2010). The first studies on single-molecule fluorescence were done on low copy molecules, because that ensured this requirement was met (Elf et al., 2007; Yu et al., 2006). However, further developments were made when they made use of photoactivatable fluorescent proteins (Betzig et al., 2006; Patterson and Lippincott-Schwartz, 2002) leading to the development of sptPALM (single-particle tracking photoactivated localization microscopy) (Manley et al., 2008). The stochastic activation of single molecules with light allowed the molecules to be separable in the time dimension and therefore allowed the localization and tracking of any copy number in the cell (Figure 3A).

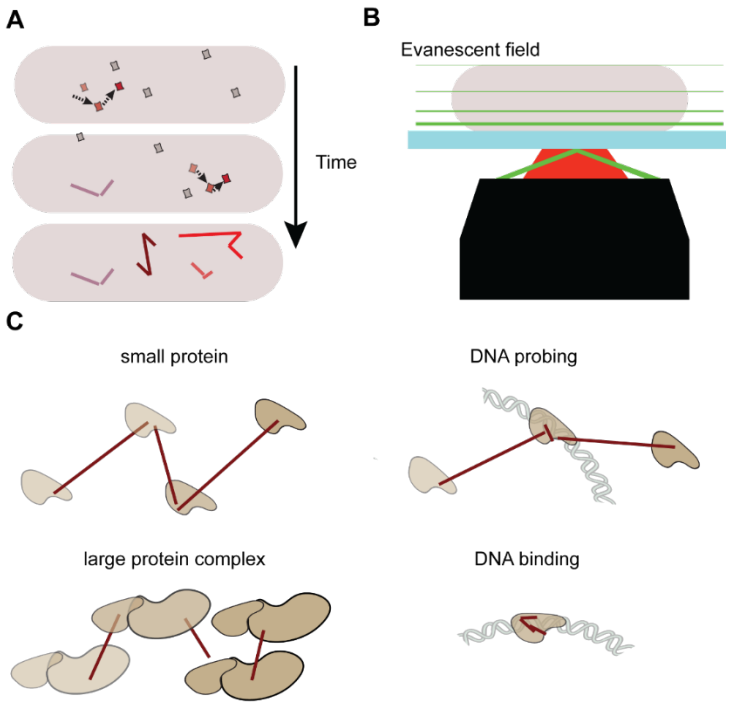


Figure 3. Single-particle tracking in bacteria. (A) sptPALM is based on activating single fluorophores one at a time, and following their position for multiple subsequent frames to form tracks. (B) TIRF microscopy is the most common technique used in combination with sptPALM, because the evanescent field reduces background from sample that is further removed from the glass slide. (C) The distances that protein travel between frames depend on their size and their interactions with other elements in the cell.

A common way to achieve a good signal-to-background ratio from the dim fluorescence of single emitters is to use total internal reflection (TIRF) microscopy (Trache and Meininger, 2008)(Figure 3B). The evanescent field created by the reflecting laser beam only illuminates up to ~ 0.5 μm above the glass surface, removing background coming from any other matter further away from the glass (Kudalkar et al., 2016). Another implementation that can improve image quality is to use stroboscopic illumination, illuminating the sample for only part of the exposure time of the camera each frame, which prevents the point spread function from widening as the particle moves during the exposure further increasing signal-to-background ratio and localization precision (Elf et al., 2007; Hansen et al., 2018).

Single-particle tracking relies on the measurement of traveled distances of single molecules in between frames. As most of the mobility inside cells is governed by Brownian motion (in contrast to active transport), the distances traveled are dependent on the size of the particle (Nenninger et al., 2010). This allows the researcher to establish whether proteins are moving as a monomeric unit, are part of a complex (Sanamrad et al., 2014), or are bound to even larger structures such as the membrane (Torreno-Pina et al., 2016) or the chromosome (Vestergaard et al., 2018) (Figure 3C). To do so, because the tracks contain a limited number of steps (typically 1-10) and there is still a localization error (typically 20-40 nm), the data requires an analysis algorithm to extract useful information. Many algorithms assume that the diffusing species remains in a single state for the whole track (Hansen et al., 2018; Stracy et al., 2015). However, depending on the transition rates between states, this cannot always be safely assumed. Therefore algorithms were designed that can extract the transition rates within the typical datasets retrieved from single-particle tracking (Persson et al., 2013).

The field of single-particle tracking is still evolving rapidly. More recent developments including 3D localization and tracking (Von Diezmann et al., 2017) and systems requiring much fewer photons for similar precision (MINFLUX (Balzarotti et al., 2017) and SIMFLUX (Cnossen et al., 2020)) will contribute to a growing number of biophysical models of how biomolecules carry out their task within the cellular environment. In this thesis, I have developed an algorithm that can extract very fast transition rates, which was needed as the DNA scanning rate of CRISPR systems approach the imaging rate of the camera. Combining the algorithm development with

technological developments can push the boundary towards faster processes and display the kinetics of single molecules in more detail. Even then, the field of *in vivo* single-molecule studies is mostly limited by a few model organisms that can be easily cultivated and modified to perform these challenging measurements (Kapanidis et al., 2018a, 2018b; Vojnovic et al., 2019). To get better insights into the distribution and functioning of these systems outside their model hosts, we need to use a different approach which I briefly describe in the next section.

Bioinformatics

Perhaps the biggest development in biology of the last 20 years is the radical increase in the number of available genome sequences. In the year 2000 13 prokaryote genomes had been sequenced. In 2020 this has increased to 200,000 (full and draft)(Zhang et al., 2020). Furthermore, the field of metagenomics has increased our knowledge of the sequence space outside the cultivatable organisms in the lab. This wealth of information has also significantly expanded the field of bioinformatics.

CRISPR-Cas systems are suitable targets for bio-informatic studies, since the diverse set of systems have some conserved features (Cas1, CRISPR array architecture) that allow new systems to be found and characterized more easily and because CRISPR arrays store a unique history of the encounters that the system in each organism has had. In fact, it was bio-informatics that first showed the presence of this system in many bacteria and archaea (Mojica et al., 2000) and the matching of the spacers to extrachromosomal elements led to the idea that CRISPR could function as an immune system (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005).

The bio-informatic study of CRISPR can be subdivided into two main fields. The first field focuses on the Cas proteins (Figure 4A). Recent studies in this field have expanded the number of CRISPR types tremendously and found evolutionary relationships between the different types (Makarova et al., 2019). A lot of new accessory proteins have been found, indicating there is still a large number of unknown actors in CRISPR biology that await experimental characterization (Shah et al., 2019; Shmakov et al., 2018). Furthermore, machine learning tools are being increasingly applied to detect novel

CRISPR-Cas proteins (Padilha et al., 2020) and anti-CRISPR proteins (Eitzinger et al., 2020; Gussow et al., 2020).

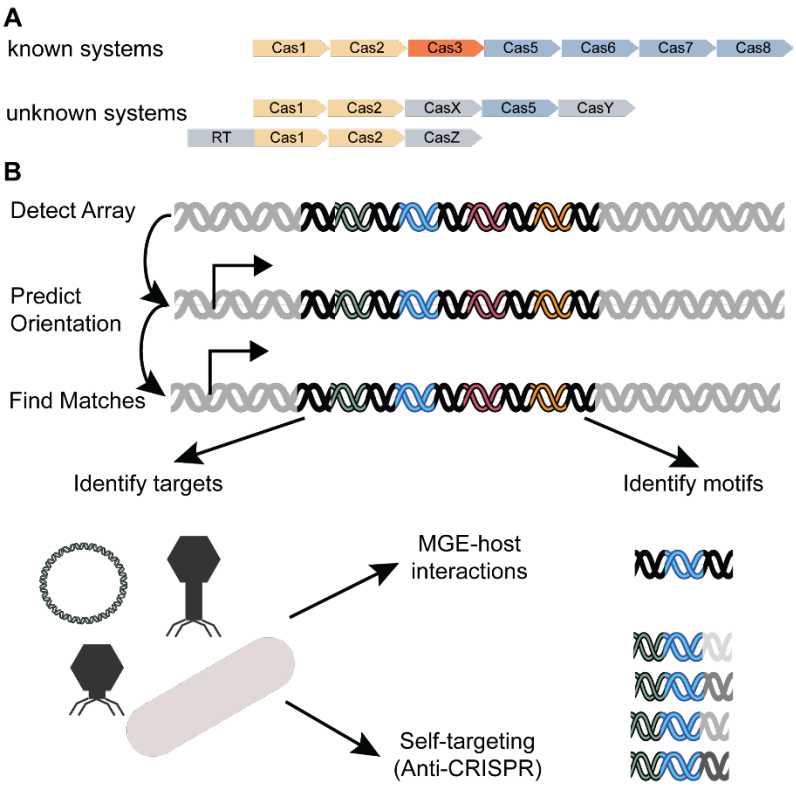


Figure 4. CRISPR bioinformatics. (A) The conserved cas genes of known systems (e.g. Cas1) can be used to find novel CRISPR-associated proteins. (B) The information stored in the array can be used to find information on what targets are being targeted by a specific host, and what (PAM) motifs are used to target these invaders.

The second field uses the information stored in CRISPR arrays (Figure 4B). Detecting CRISPR arrays and extracting the spacers from is not as trivial as it seems and several tools have been developed to accurately find them (Biswas et al., 2016; Couvin et al., 2018; Skennerton et al., 2013) Furthermore, it can be important to know the direction in which the array is transcribed. It is namely this direction that determines the chronological order of acquisition events as the newest spacer tends to be inserted at the leader (promoter) sequence adjacent to the repeat. Furthermore the direction also allows determining the strand bias. This direction can be determined by either looking for the leader sequence (AT-rich) or by comparing the array to a library of repeat sequences that have already been characterized (Alkhnbashi et al., 2016; Biswas et al., 2016), for the repeat sequences are relatively well conserved for each subtype.

The information stored in spacers can be used to verify whether findings in the lab apply in natural settings too. For example, although primed acquisition (accelerated acquisition upon mutation of target sequences) was a commonly studied mechanism under lab conditions (Jackson et al., 2017), it was only recently verified that this process plays an important role in natural systems (Nicholson et al., 2019). Another study has shown that jumbo phages that form nucleoid-like structures are protected from DNA-targeting CRISPR systems and sensitive to RNA-targeting systems, but could also demonstrate that this plays an important role in nature by observing that RNA-targeting Type III systems preferably target these phages (Malone et al., 2020). CRISPR spacers are also a useful tool to identify hosts of putative phages found in metagenomic sequences (Paez-Espino et al., 2016). In this thesis, I have provided a catalog of PAM diversity within CRISPR subtypes and showed general principles of PAM-repeat relationships and spacer sharing based on the spacers stored in genomes across the empire of prokaryotes.

Even though we seem to understand much about CRISPR biology in the lab, I believe we still miss insight into what CRISPR really means in the natural setting. The CRISPR field started, when through bioinformatics these systems were found and it was hypothesized that these might comprise immune systems. Then the main mechanics of CRISPR were revealed, mostly using biochemistry and structural biology. We are now starting to understand the kinetics of CRISPR, revealed through single-molecule and single-cell studies *in vivo* and *in vitro*. I think the future will continue to use bio-informatics to get closer to systems outside our lab, and combined with new ways to probe micro-organisms in their natural environment, really demonstrate when and how CRISPR is useful to their prokaryotic hosts and also under what circumstances it is not.

Outline of Thesis

In the first chapter, the mechanism of target search of the native Type I-E *E. coli* CRISPR system is revealed. By using single-molecule tracking of Cascade complexes, we were able to find the relationship between Cascade copy number and their *in vivo* interference levels. When studying the probing kinetics of Cascade, we find rapid association and dissociation kinetics and a potential optimization of time spent on DNA and freely diffusing in the cytoplasm. We also find evidence for subunit dissociation upon binding

to repeat-like targets. This study led to the hypothesis of a kinetic arms race between invader replication and CRISPR target search, which has implications for the overall functioning of CRISPR in the ecosystem of bacteria and its invaders.

The second chapter explores the target search mechanism of a Type II CRISPR system in *Lactococcus lactis*. Even though this system is very distinct from the I-E system, the probing kinetics are very similar, pointing to a biophysical limit of CRISPR probing speeds. In this study, we also explore the variation in cellular targets and its relationship to target clearance. We also found evidence for inhibition of plasmid replication upon binding of dCas9 to its target.

In the third chapter, we describe the study of the second native E. coli CRISPR system (I-F). We compare its diffusion and subunit kinetics to the previous study described in the first chapter. We found overall faster diffusion rates of the I-F complex in comparison with the I-E complex, perhaps related to the smaller size of the complex. Furthermore, we study the binding to targets containing mismatches in different regions of the target and find that the I-F system is very sensitive to PAM-proximal mutations.

The fourth chapter discusses the extraction of transitioning kinetics of DNA binding proteins which played an important role in the first three chapters. The underlying mathematical model combines diffusion distribution probabilities with PDA statistics used previously in studying transition kinetics of FRET studies. This model enables the study of previously inaccessible transition rates and can therefore shed more light on the biophysics of a wide range of DNA and membrane-binding proteins.

The fifth chapter focuses on the characterization of the spacers found in bacterial and archaeal genomes. We were able to match more than a third of a large set of spacers from complete genomes to a target. With that information, we were able to predict the PAM of all these systems and the orientation of the spacers. These findings brought us to discover prevalent compatibility between Type I and Type III systems, where the spacers match both PAM requirements of the Type I systems and orientation requirements of Type III systems.

References

Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B.T., Shmakov, S., Makarova, K.S., Semenova, E., Minakhin, L., et al. (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science (80-.). *353*.

Alberts, B. (2002). Molecular Biology of the Cell.

Alkhnbashi, O.S., Shah, S.A., Garrett, R.A., Saunders, S.J., Costa, F., and Backofen, R. (2016). Characterizing leader sequences of CRISPR loci. In Bioinformatics, p.

Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature *513*, 569–573.

Balzarotti, F., Eilers, Y., Gwosch, K.C., Gynnå, A.H., Westphal, V., Stefani, F.D., Elf, J., and Hell, S.W. (2017). Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes. Science *355*, 606–612.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science (80-.). *315*, 1709–1712.

Berg, O.G., Winter, R.B., and von Hippel, P.H. (1981). Diffusion-Driven Mechanisms of Protein Translocation on Nucleic Acids. 1. Models and Theory. Biochemistry *20*, 6929–6948.

Bernheim, A., and Sorek, R. (2019). The pan-immune system of bacteria: antiviral defence as a community resource. Nat. Rev. Microbiol.

Bernheim, A., Bikard, D., Touchon, M., and Rocha, E.P.C. (2019). Atypical organizations and epistatic interactions of CRISPRs and cas clusters in genomes and their mobile gemetic elements. Nucleic Acids Res. 1–13.

Betzig, E., Patterson, G.H., Sougrat, R., Lindwasser, O.W., Olenych, S., Bonifacino, J.S., Davidson, M.W., Lippincott-schwartz, J., and Hess, H.F. (2006). Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. *x*, 1642–1646.

Biswas, A., Staals, R.H.J., Morales, S.E., Fineran, P.C., and Brown, C.M. (2016). CRISPRDetect: A flexible algorithm to define CRISPR arrays. BMC Genomics.

Blosser, T.R., Loeff, L., Westra, E.R., Vlot, M., Künne, T., Sobota, M., Dekker, C., Brouns, S.J.J., and Joo, C. (2015). Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. Mol. Cell *58*, 60–70.

Bolotin, A., Quinquis, B., Sorokin, A., and Dusko Ehrlich, S. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology.

Bonnefoy, E., and Rouvière-Yaniv, J. (1992). HU, the major histone-like protein of E. coli, modulates the binding of IHF to oriC. EMBO J. *11*, 4489–4496.

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E. V, and van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. Science *321*, 960–964.

Candelli, A., Wuite, G.J.L., and Peterman, E.J.G. (2011). Combining optical trapping, fluorescence microscopy and micro-fluidics for single molecule studies of DNA-protein interactions. Phys. Chem. Chem. Phys. *13*, 7263–7272.

Carte, J., Wang, R., Li, H., Terns, R.M., and Terns, M.P. (2008). Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes Dev. *22*, 3489–3496.

Chaurasiya, K.R., Paramanathan, T., McCauley, M.J., and Williams, M.C. (2010). Biophysical characterization of DNA binding from single molecule force measurements. Phys. Life Rev. *7*, 299–341.

Cnossen, J., Hinsdale, T., Thorsen, R.Ø., Siemons, M., Schueder, F., Jungmann, R., Smith, C.S., Rieger, B., and Stallinga, S. (2020). Localization microscopy at doubled precision with patterned illumination. Nat. Methods *17*, 59–63.

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., and Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res.

Cui, T.J., and Joo, C. (2019). Facilitated diffusion of Argonaute-mediated target search. RNA Biol. *16*, 1093–1107.

Datsenko, K.A., Pougach, K., Tikhonov, A., Wanner, B.L., Severinov, K., and Semenova, E. (2012). Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat. Commun. *3*, 945.

Dawkins, R. (1976). The Selfish Gene.

Von Diezmann, A., Shechtman, Y., and Moerner, W.E. (2017). Three-Dimensional Localization of Single Molecules for Super-Resolution Imaging and Single-Particle

Tracking. Chem. Rev. *117*, 7244–7275.

Dillard, K.E., Brown, M.W., Johnson, N. V, Xiao, Y., Dolan, A., Hernandez, E., Dahlhauser, S.D., Kim, Y., Myler, L.R., Anslyn, E. V, et al. (2018). Assembly and Translocation of a CRISPR-Cas Primed Acquisition Complex. Cell *175*, 934-946.e15.

Dimitriu, T., Misevic, D., Lotton, C., Brown, S.P., Lindner, A.B., and Taddei, F. (2016). Indirect Fitness Benefits Enable the Spread of Host Genes Promoting Costly Transfer of Beneficial Plasmids. PLoS Biol. *14*.

Dulin, D., Lipfert, J., Moolman, M.C., and Dekker, N.H. (2013). Studying genomic processes at the single-molecule level: Introducing the tools and applications. Nat. Rev. Genet. *14*, 9–22.

Eitzinger, S., Asif, A., Watters, K.E., Iavarone, A.T., Knott, G.J., Doudna, J.A., and Minhas, F.U.A.A. (2020). Machine learning predicts new anti-CRISPR proteins. Nucleic Acids Res.

Elf, J. (2016). Hypothesis: Homologous Recombination Depends on Parallel Search. Cell Syst. *3*, 325–327.

Elf, J., Li, G.W., and Xie, X.S. (2007). Probing transcription factor dynamics at the single-molecule level in a living cell. Science (80-.). *316*, 1191–1194.

Fei, J., Singh, D., Zhang, Q., Park, S., Balasubramanian, D., Golding, I., Vanderpool, C.K., and Ha, T. (2015). Determination of in vivo target search kinetics of regulatory noncoding RNA. Science (80-.). *347*, 1371–1374.

Fineran, P.C., Gerritzen, M.J.H., Suarez-Diez, M., Kunne, T., Boekhorst, J., van Hijum, S. a. F.T., Staals, R.H.J., and Brouns, S.J.J. (2014). Degenerate target sites mediate rapid primed CRISPR adaptation. Proc. Natl. Acad. Sci. *111*, 1629–1638.

Francisco J.M. Mojica, Ce´sar Dı´ez-Villasen˜or, Jesu´s Garcı´a-Martı´nez, and Elena Soria (2005). Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from F...: EBSCOhost.

Globyte, V., Lee, S.H., Bae, T., Kim, J., and Joo, C. (2018). CRISPR/Cas9 searches for a protospacer adjacent motif by lateral diffusion. EMBO J. e99466.

Gussow, A.B., Park, A.E., Borges, A.L., Shmakov, S.A., Makarova, K.S., Wolf, Y.I., Bondy-Denomy, J., and Koonin, E. V. (2020). Machine-learning approach expands the repertoire of anti-CRISPR protein families. Nat. Commun.

Gynnå, A.H., Wiktor, J., Leroy, P., and Elf, J. (2020). RecA mediated homology search

finds segregated sister locus in minutes after a double stranded break. BioRxiv 2020.02.13.946996.

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. Nat. Struct. Mol. Biol. *11*, 394–403.

Hale, C.R., Duff, M.O., Graveley, B.R., Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., and Terns, R.M. (2009). RNA-guided RNA cleavage by a CRISPR RNA- Cas protein complex RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. Cell *139*, 945–956.

Hammar, P., Leroy, P., Mahmutovic, A., Marklund, E.G., Berg, O.G., and Elf, J. (2012). The lac Repressor Displays Facilitated Diffusion in Living Cells. Science *336*, 1595–1598.

Hansen, A.S., Woringer, M., Grimm, J.B., Lavis, L.D., Tjian, R., and Darzacq, X. (2018). Robust model-based analysis of single-particle tracking experiments with Spot-On. Elife *7*.

Hayes, R.P., Xiao, Y., Ding, F., van Erp, P.B.G., Rajashankar, K., Bailey, S., Wiedenheft, B., and Ke, A. (2016). Structural basis for promiscuous PAM recognition in type I–E Cascade from E. coli. Nature *530*, 499–503.

Horvath, P., Romero, D.A., Coûté-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C., and Barrangou, R. (2008). Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. J. Bacteriol. *190*, 1401–1412.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakatura, A. (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isoenzyme conversion in Escherichia coli, and identification of the gene product. J. Bacteriol. *169*, 5429–5433.

Jackson, S.A., McKenzie, R.E., Fagerlund, R.D., Kieper, S.N., Fineran, P.C., and Brouns, S.J.J. (2017). CRISPR-Cas: Adapting to change. Science *356*, eaal5056.

Jansen, R., Van Embden, J.D.A., Gaastra, W., and Schouls, L.M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. Mol. Microbiol. *43*, 1565–1575.

Jayaram, N., Usvyat, D., and Martin, A.C. (2016). Evaluating tools for transcription factor binding site prediction. BMC Bioinformatics *17*.

Jones, D.L., Leroy, P., Unoson, C., Fange, D., Ćurić, V., Lawson, M.J., and Elf, J. (2017).

Kinetics of dCas9 target search in Escherichia coli. Science *357*, 1420–1424.

Kapanidis, A.N., Lepore, A., and El Karoui, M. (2018a). Rediscovering Bacteria through Single-Molecule Imaging in Living Cells. Biophys. J. *115*, 190–202.

Kapanidis, A.N., Uphoff, S., and Stracy, M. (2018b). Understanding Protein Mobility in Bacteria by Tracking Single Molecules. J. Mol. Biol.

Kazlauskiene, M., Kostiuk, G., Venclovas, Č., Tamulaitis, G., and Siksnys, V. (2017). A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. Science (80-.). *357*, 605–609.

Kieper, S.N., Almendros, C., Behler, J., McKenzie, R.E., Nobrega, F.L., Haagsma, A.C., Vink, J.N.A., Hess, W.R., and Brouns, S.J.J. (2018). Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. Cell Rep. *22*, 3377–3384.

Kieper, S.N., Almendros, C., and Brouns, S.J.J. (2019). Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type I-D CRISPR-Cas systems. FEMS Microbiol. Lett. *366*.

Kim, S., Loeff, L., Colombo, S., Jergic, S., Brouns, S.J.J., and Joo, C. (2020). Selective loading and processing of prespacers for precise CRISPR adaptation. Nature *579*, 141–145.

Klein, M., Eslami-Mossallam, B., Arroyo, D.G., and Depken, M. (2018). Hybridization Kinetics Explains CRISPR-Cas Off-Targeting Rules. Cell Rep. *22*, 1413–1423.

Klein, M., Cui, T.J., MacRae, I., Joo, C., and Depken, M. (2020). Skipping and sliding to optimize target search on protein-bound DNA and RNA. BioRxiv 2020.06.04.133629.

Klompe, S.E., Vo, P.L.H., Halpin-Healy, T.S., and Sternberg, S.H. (2019). Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. Nature.

Knight, S.C., Xie, L., Deng, W., Guglielmi, B., Witkowsky, L.B., Bosanac, L., Zhang, E.T., El Beheiry, M., Masson, J.-B.J.-B.J.-B., Dahan, M., et al. (2015). Dynamics of CRISPR-Cas9 genome interrogation in living cells. Science *350*, 823–826.

Kudalkar, E.M., Davis, T.N., and Asbury, C.L. (2016). Single-molecule total internal reflection fluorescence microscopy. Cold Spring Harb. Protoc. *2016*, 435–438.

Künne, T., Zhu, Y., da Silva, F., Konstantinides, N., McKenzie, R.E., Jackson, R.N., and Brouns, S.J.J. (2018). Role of nucleotide identity in effective CRISPR target escape mutations. Nucleic Acids Res. *46*, 10395–10404.

Lam, T.J., and Ye, Y. (2019). Long reads reveal the diversification and dynamics of

CRISPR reservoir in microbiomes. BMC Genomics *20*.

Lesterlin, C., Ball, G., Schermelleh, L., and Sherratt, D.J. (2014). RecA bundles mediate homology pairing between distant sisters during DNA break repair. Nature *506*, 249–253.

Li, H. (2015). Structural Principles of CRISPR RNA Processing. Structure *23*, 13–20.

Lind, C., and Åqvist, J. (2016). Principles of start codon recognition in eukaryotic translation initiation. Nucleic Acids Res. *44*, 8425–8432.

Lopez-Sanchez, M.J., Sauvage, E., Da Cunha, V., Clermont, D., Ratsima Hariniaina, E., Gonzalez-Zorn, B., Poyart, C., Rosinski-Chupin, I., and Glaser, P. (2012). The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome. Mol. Microbiol. *85*, 1057–1071.

Luijsterburg, M.S., Noom, M.C., Wuite, G.J.L., and Dame, R.T. (2006). The architectural role of nucleoid-associated proteins in the organization of bacterial chromatin: A molecular perspective. J. Struct. Biol. *156*, 262–272.

Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P., et al. (2019). Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat. Rev. Microbiol.

Malone, L.M., Warring, S.L., Jackson, S.A., Warnecke, C., Gardner, P.P., Gumy, L.F., and Fineran, P.C. (2020). A jumbo phage that forms a nucleus-like structure evades CRISPR–Cas DNA targeting but is vulnerable to type III RNA-based immunity. Nat. Microbiol.

Manley, S., Gillette, J.M., Patterson, G.H., Shroff, H., Hess, H.F., Betzig, E., and Lippincott-Schwartz, J. (2008). High-density mapping of single-molecule trajectories with photoactivated localization microscopy. Nat. Methods *5*, 155–157.

McGinn, J., and Marraffini, L.A. (2016). CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. Mol. Cell *64*, 616–623.

Meeske, A.J., Nakandakari-Higa, S., and Marraffini, L.A. (2019). Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. Nature *570*, 241–245.

Millan, A.S., Toll-Riera, M., Qi, Q., and MacLean, R.C. (2015). Interactions between horizontally acquired genes create a fitness cost in Pseudomonas aeruginosa. Nat.

Commun. *6*.

Mohanraju, P., Makarova, K.S., Zetsche, B., Zhang, F., Koonin, E. V., and van der Oost, J. (2016). Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science *353*.

Mojica, F.J.M., Díez-Villaseñor, C., Soria, E., and Juez, G. (2000). Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. Mol. Microbiol.

Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., and Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J. Mol. Evol.

Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., and Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology *155*, 733–740.

Mortensen, K.I., Churchman, L.S., Spudich, J.A., and Flyvbjerg, H. (2010). Optimized localization analysis for single-molecule tracking and super-resolution microscopy. Nat. Methods *7*, 377–381.

Nenninger, A., Mastroianni, G., and Mullineaux, C.W. (2010). Size dependence of protein diffusion in the cytoplasm of Escherichia coli. J. Bacteriol. *192*, 4535–4540.

Nicholson, T.J., Jackson, S.A., Croft, B.I., Staals, R.H.J., Fineran, P.C., and Brown, C.M. (2019). Bioinformatic evidence of widespread priming in type I and II CRISPR-Cas systems. RNA Biol. *16*, 566–576.

Niewoehner, O., Jinek, M., and Doudna, J. a. (2014). Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases. Nucleic Acids Res. *42*, 1341–1353.

Niewoehner, O., Garcia-Doval, C., Rostøl, J.T., Berk, C., Schwede, F., Bigler, L., Hall, J., Marraffini, L.A., and Jinek, M. (2017). Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. Nature *548*, 543–548.

Nuñez, J.K., Bai, L., Harrington, L.B., Hinder, T.L., and Doudna, J.A. (2016). CRISPR Immunological Memory Requires a Host Factor for Specificity. Mol. Cell *62*, 824–833.

Padilha, V.A., Alkhnbashi, O.S., Shah, S.A., De Carvalho, A.C.P.L.F., and Backofen, R. (2020). CRISPRcasIdentifier: Machine learning for accurate identification and classification of CRISPR-Cas systems. Gigascience.

Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann,

M., Mikhailova, N., Rubin, E., Ivanova, N.N., and Kyrpides, N.C. (2016). Uncovering Earth's virome. Nature.

Patterson, G.H. (2009). Fluorescence microscopy below the diffraction limit. Semin. Cell Dev. Biol. *20*, 886–893.

Patterson, G.H., and Lippincott-Schwartz, J. (2002). A photoactivatable GFP for selective photolabeling of proteins and cells. Science (80-.). *297*, 1873–1877.

Persson, F., Lindén, M., Unoson, C., and Elf, J. (2013). Extracting intracellular diffusive states and transition rates from single-molecule tracking data. Nat. Methods *10*, 265–269.

Pinilla-Redondo, R., Mayo-Muñoz, D., Russel, J., Garrett, R.A., Randau, L., Sørensen, S.J., and Shah, S.A. (2019). Type IV CRISPR–Cas systems are highly diverse and involved in competition between plasmids. Nucleic Acids Res.

Pourcel, C., Salvignol, G., and Vergnaud, G. (2005). CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology.

Pourcel, C., Touchon, M., Villeriot, N., Vernadet, J.P., Couvin, D., Toffano-Nioche, C., and Vergnaud, G. (2020). CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. Nucleic Acids Res. *48*, D535–D544.

Pribnow, D. (1975). Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. Proc. Natl. Acad. Sci. U. S. A. *72*, 784–788.

Privalov, P.L., Dragan, A.I., Crane-Robinson, C., Breslauer, K.J., Remeta, D.P., and Minetti, C.A.S.A. (2007). What Drives Proteins into the Major or Minor Grooves of DNA? J. Mol. Biol. *365*, 1–9.

Redding, S., and Greene, E.C. (2013). How do proteins locate specific targets in DNA? Chem. Phys. Lett. *570*, 1–11.

Redding, S., Sternberg, S.H.H., Marshall, M., Gibb, B., Bhat, P., Guegler, C.K.K., Wiedenheft, B., Doudna, J.A., and Greene, E.C.C. (2015). Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. Cell *163*, 1–12.

Riggs, A.D., Suzuki, H., and Bourgeois, S. (1970). lac repressor-operator interaction. I. Equilibrium studies. J. Mol. Biol. *48*, 67–83.

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of

Specificity in Protein-DNA Recognition. Annu. Rev. Biochem. *79*, 233–269.

Rollins, M.F., Schuman, J.T., Paulus, K., Bukhari, H.S.T., and Wiedenheft, B. (2015). Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from Pseudomonas aeruginosa. Nucleic Acids Res. *43*, 2216–2222.

San Filippo, J., Sung, P., and Klein, H. (2008). Mechanism of Eukaryotic Homologous Recombination. Annu. Rev. Biochem. *77*, 229–257.

Sanamrad, A., Persson, F., Lundius, E.G., Fange, D., Gynna, A.H., and Elf, J. (2014). Single-particle tracking reveals that free ribosomal subunits are not excluded from the Escherichia coli nucleoid. Proc. Natl. Acad. Sci. *111*, 11413–11418.

Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., Van Der Oost, J., Brouns, S.J.J., and Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc. Natl. Acad. Sci. U. S. A. *108*, 10098–10103.

Shah, S.A., Alkhnbashi, O.S., Behler, J., Han, W., She, Q., Hess, W.R., Garrett, R.A., and Backofen, R. (2019). Comprehensive search for accessory proteins encoded with archaeal and bacterial type III CRISPR-cas gene cassettes reveals 39 new cas gene families. RNA Biol.

Shashkova, S., and Leake, M.C. (2017). Single-molecule fluorescence microscopy review: Shedding new light on old problems. Biosci. Rep. *37*.

Shmakov, S.A., Makarova, K.S., Wolf, Y.I., Severinov, K. V., and Koonin, E. V. (2018). Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. Proc. Natl. Acad. Sci. U. S. A.

Silas, S., Mohr, G., Sidote, D.J., Markham, L.M., Sanchez-Amat, A., Bhaya, D., Lambowitz, A.M., and Fire, A.Z. (2016). Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. Science (80-.). *351*.

Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. EMBO J. *30*, 1335–1342.

Skennerton, C.T., Imelfort, M., and Tyson, G.W. (2013). Crass: Identification and reconstruction of CRISPR from unassembled metagenomic data. Nucleic Acids Res.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: How transcription factors read the genome. Trends

Biochem. Sci.

Slutsky, M., and Mirny, L.A. (2004). Kinetics of Protein-DNA Interaction: Facilitated Target Location in Sequence-Dependent Potential. Biophys. J. *87*, 4021–4035.

Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C., and Doudna, J.A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62–67.

Stracy, M., Lesterlin, C., Garza de Leon, F., Uphoff, S., Zawadzki, P., and Kapanidis, A.N. (2015). Live-cell superresolution microscopy reveals the organization of RNA polymerase in the bacterial nucleoid. Proc. Natl. Acad. Sci. *112*, E4390–E4399.

Strecker, J., Ladha, A., Gardner, Z., Schmid-Burgk, J.L., Makarova, K.S., Koonin, E. V., and Zhang, F. (2019). RNA-guided DNA insertion with CRISPR-associated transposases. Science (80-.). *364*, 48–53.

Szczelkun, M.D., Tikhomirova, M.S., Sinkunas, T., Gasiunas, G., Karvelis, T., Pschera, P., Siksnys, V., and Seidel, R. (2014). Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. Proc. Natl. Acad. Sci. *111*, 9798–9803.

Torreno-Pina, J.A., Manzo, C., and Garcia-Parajo, M.F. (2016). Uncovering homo-and hetero-interactions on the cell membrane using single particle tracking approaches. J. Phys. D. Appl. Phys. *49*.

Trache, A., and Meininger, G.A. (2008). Total Internal Reflection Fluorescence (TIRF) microscopy. Curr. Protoc. Microbiol.

Uphoff, S., Reyes-Lamothe, R., Garza de Leon, F., Sherratt, D.J., and Kapanidis, A.N. (2013). Single-molecule DNA repair in live bacteria. Proc. Natl. Acad. Sci. U. S. A. *110*, 8063–8068.

Vestergaard, C.L., Blainey, P.C., and Flyvbjerg, H. (2018). Single-particle trajectories reveal two-state diffusion-kinetics of hOGG1 proteins on DNA. Nucleic Acids Res. *46*, 2446–2458.

Vojnovic, I., Winkelmeier, J., and Endesfelder, U. (2019). Visualizing the inner life of microbes: Practices of multi-color single-molecule localization microscopy in microbiology. Biochem. Soc. Trans.

Wei, Y., Chesne, M.T., Terns, R.M., and Terns, M.P. (2015). Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus. Nucleic Acids Res. *43*, 1749–1758.

Westra, E.R., van Erp, P.B.G., Künne, T., Wong, S.P., Staals, R.H.J., Seegers, C.L.C., Bollen, S., Jore, M.M., Semenova, E., Severinov, K., et al. (2012). CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. Mol. Cell *46*, 595–605.

Wilkinson, M., Drabavicius, G., Silanskas, A., Gasiunas, G., Siksnys, V., and Wigley, D.B. (2019). Structure of the DNA-Bound Spacer Capture Complex of a Type II CRISPR-Cas System. Mol. Cell *75*, 90-101.e5.

Wilson, E. (1975). Sociobiology: the New Synthesis.

Xue, C., Zhu, Y., Zhang, X., Shin, Y.K., and Sashital, D.G. (2017). Real-Time Observation of Target Search by the CRISPR Surveillance Complex Cascade. Cell Rep. *21*, 3717–3727.

Yoganand, K.N., Muralidharan, M., Nimkar, S., and Anand, B. (2019). Fidelity of prespacer capture and processing is governed by the PAM-mediated interactions of Cas1-2 adaptation complex in CRISPR-Cas type I-E system. J. Biol. Chem. *294*, 20039–20053.

Yu, J., Xiao, J., Ren, X., Lao, K., and Xie, X.S. (2006). Probing gene expression in live cells, one protein molecule at a time. Science (80-.). *311*, 1600–1603.

Zhang, Z., Wang, J., Wang, J., Wang, J., and Li, Y. (2020). Estimate of the sequenced proportion of the global prokaryotic genome. Microbiome.

2

Direct visualization of native CRISPR target search in live bacteria reveals Cascade DNA surveillance mechanism

Published as: J. N. A. Vink, K. J. A. Martens, M. Vlot, R. E. McKenzie, C. Almendros, B. Estrada Bonilla, D. J. W. Brocken, J. Hohlbein, S. J. J. Brouns, Direct Visualization of Native CRISPR Target Search in Live Bacteria Reveals Cascade DNA Surveillance Mechanism. *Mol. Cell*. **77**, 39-50.e10 (2020).

Summary

CRISPR-Cas systems encode RNA-guided surveillance complexes to find and cleave invading DNA elements. While it is thought that invaders are neutralized minutes after cell entry, the mechanism and kinetics of target search and its impact on CRISPR protection levels have remained unknown. Here we visualized individual Cascade complexes in a native type I CRISPR-Cas system. We uncovered an exponential relationship between Cascade copy number and CRISPR interference levels, pointing to a time-driven arms race between invader replication and target search, in which 20 Cascade complexes provide 50% protection. Driven by PAM-interacting subunit Cas8e, Cascade spends half its search time rapidly probing DNA (~30 ms) in the nucleoid. We further demonstrate that target DNA transcription and CRISPR arrays affect the integrity of Cascade and impact CRISPR interference. Our work establishes the mechanism of cellular DNA surveillance by Cascade that allows the timely detection of invading DNA in a crowded, DNA-packed environment.

Introduction

RNA-guided CRISPR-Cas surveillance complexes have evolved to specifically and rapidly recognize sequences of previously catalogued mobile genetic elements (MGEs) (Marraffini, 2015). Target DNA recognition depends on CRISPR RNA (crRNA) – DNA complementarity and on the presence of a protospacer adjacent motif (PAM), a short nucleotide sequence flanking the target site (Deveau et al., 2008; Mojica et al., 2009). To work effectively, the complexes need to find their targets fast enough to prevent an MGE from becoming established in the cell, which can occur within minutes upon cell entry (Shao et al., 2015). Target search inside a cell faces a multitude of challenges: Firstly, cells are packed with DNA, and crRNA surveillance complexes need to find the needle in a haystack before an invading element takes control of the cell. PAM scanning and crRNA-seed interactions with the target have been suggested to speed up the search process by drastically reducing the number of potential target sites in the genome (Gleditzsch et al., 2018; Jones et al., 2017). Several studies have shown that crRNA-effector complexes spend more time probing PAM rich regions, which is indicative of its function as the first recognition site (Globyte et al., 2018; Redding et al., 2015; Sternberg et al., 2014). The *Escherichia coli* K12 genome contains 127.081 preferred PAMs (CTT) that are recognized by the crRNA-effector complex Cascade in the Type I-E CRISPR-Cas system (Leenay et al., 2016). This large number of PAMs suggests that the interaction with the PAM needs to be sufficiently fast to cover enough sequence space to find an invading DNA sequence in time. A second challenge is posed by the action of other proteins present in the cell such as DNA binding proteins, DNA or RNA polymerases that may interfere with target search and formation of target bound crRNA complexes (Jones et al., 2017; Vigouroux et al., 2018). Some invading MGEs even use specialized anti-CRISPR proteins to inhibit crRNA-effector complexes and impair the target search process (Bondy-Denomy et al., 2015; Pawluk et al., 2014). A third challenge that microbes face is to produce appropriate levels of Cascade complexes loaded with one particular crRNA to provide protection against a single invading element. While adding more and more spacers to CRISPR arrays will have the benefit of recognizing many invaders, the tradeoff is that long CRISPR arrays will dilute the number of Cascade complexes loaded with a particular

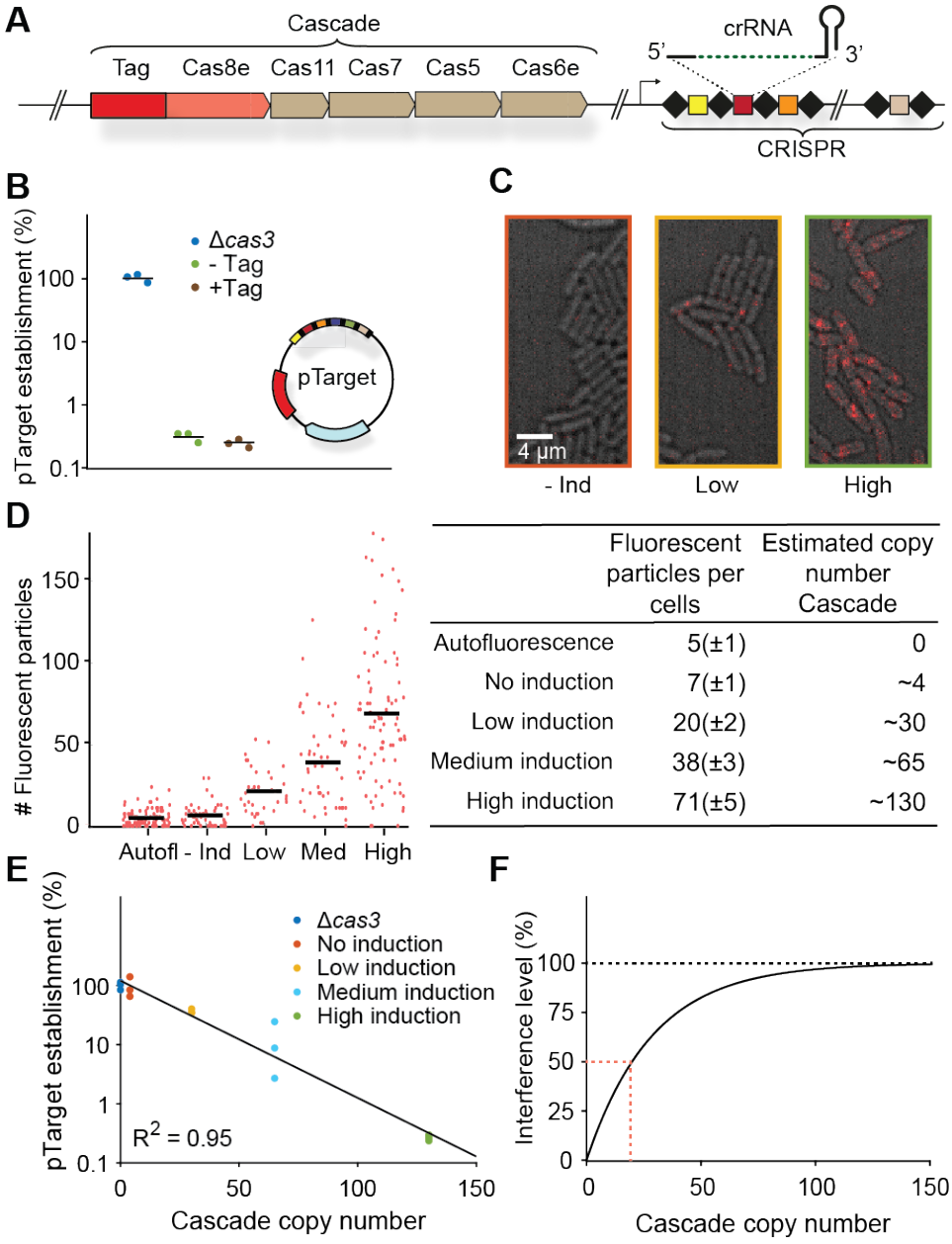
crRNA, potentially decreasing the CRISPR response against that target. These cellular challenges raise the question how Cascade can navigate the crowded cell sufficiently fast to find DNA targets, and how many copies of Cascade are required to do so.

Here, we report the visualization of single-molecule Type I-E Cascade complexes in a native *E. coli* CRISPR-Cas system *in vivo*. We found that the probability of successful CRISPR protection depends exponentially on Cascade copy numbers, which leads to a time-driven arms race model between Cascade target search and invader replication. The localization of Cascade shows the complex is enriched inside the nucleoid. We determined that 60% of the Cas8e subunit is incorporated into Cascade complexes and that Cascade DNA probing interactions are very rapid (~ 30 ms) and are driven by Cas8e. Furthermore, transcription of targets and CRISPR arrays reduce the number of functional complexes in the cell. Our work sheds light on target search and dynamical assembly of Cascade complexes in their native cellular environment, and describes how these processes impact CRISPR protection levels.

Results

Visualizing Cascade abundance and target search at the single-molecule level

To investigate how microbes deal with these challenges at the cellular level we used intracellular single-particle tracking Photo-Activated Localization Microscopy (sptPALM) (English et al., 2011; Manley et al., 2008), a technique capable of following the movement and abundance of individual fluorescently-tagged proteins in cells with high precision. By genetically fusing a photoactivatable fluorescent protein (PAmCherry2, (Subach et al., 2009)) to the N-terminus of Cascade-subunit Cas8e (Figure 1A), which was the only subunit for which labeling had no influence on the CRISPR interference ability of this strain (Figure 1B), we were able to monitor the mobility and abundance of Cascade complexes in *E. coli* cells.



	Fluorescent particles per cells	Estimated copy number Cascade
Autofluorescence	5(±1)	0
No induction	7(±1)	~4
Low induction	20(±2)	~30
Medium induction	38(±3)	~65
High induction	71(±5)	~130

Figure 1: Cascade copy number vs CRISPR protection. (**A**) Chromosomal locus of the Cascade subunits and integration site of the photoactivatable fluorescent protein upstream of *cas8e*. (**B**) pTarget establishment, calculated from the ratio of transformation of pTarget/pGFPuv, is a measure for the interference level of the CRISPR system. To test whether tagged Cascade complexes were able to function normally, we compared the tagged strain to the untagged and the *Δcas3* strain. pTarget (bottom right) contains protospacers for all spacers in the K12 genome (colored, not all depicted) and are flanked by a 5'-CTT-3' PAM (black bars). (**C**) Overlay of brightfield image of cells (grey) and single molecule signal (red) from a single representative frame for different induction levels. (**D**) Number of fluorescent particles measured in each cell plotted for different levels of Cascade expression (left). The mean number of fluorescent particles (± standard

deviation; table left column) was converted to a Cascade copy number (table right column, Methods). (**E**) pTarget establishment plotted for different copy numbers of Cascade. The data points were fitted with an exponential decay function. $pTarget\ establishment = e^{-an}$, where *n* equals Cascade copy number and *a* the fitted coefficient. In our model $a = \bar{t_s}/t_c$. (**F**) The fitted exponential decay on the left converted into an interference level ($Interference\ level = 1 - pTarget\ establishment$). Indicated in red (dashed) is the amount of Cascade copies required for 50% interference.

Twenty Cascade complexes provide 50% CRISPR protection

We first wanted to link the copy number of Cascade to successful target search, and established an assay that measures the level of CRISPR protection in cells at the time of cell entry by a mobile genetic element (MGE). In this assay all Cascade complexes present in the cell must be able to target the incoming MGE and Cascade target search has to be rate limiting. To meet the first requirement, we constructed a high copy plasmid (pTarget; Figure 1B) containing target sites for all 18 spacers found in the genomic arrays of *E. coli* K12, such that all Cascade complexes would be targeting the incoming plasmid. Secondly, we ensured that Cascade copy numbers were rate limiting (Majsec et al., 2016) by equipping cells with a low copy plasmid expressing the nuclease Cas3 (pCas3, adapted from (Westra et al., 2010)). We achieved different expression levels of Cascade in the cell by tuning the expression of the native regulator LeuO (Westra et al., 2010) (Figure 1C). The copy numbers of Cascade under these varying levels of LeuO induction were estimated from the number of fluorescent particles present in the cell, taking complex assembly (see following section), growth rate (Table S1) and maturation time of PAmCherry into account (Figure 1D; Methods). We found that the average number of Cascade complexes per cell in the absence of LeuO induction was low (~4 copies) and that copy numbers increased more than 30-fold for the highest induction level (~130 copies). We measured the interference ability under these conditions by determining the probability that pTarget becomes established in a cell. We observed that establishment of pTarget decreases sharply with increasing copy numbers of Cascade (Figure 1E). However, even with 130 Cascade complexes present, we still observed a level of pTarget survival (~0.5%).

To explain these observations, we modelled the probability that an invading MGE becomes established in the cell depending on the number of Cascade complexes that target this specific MGE. The model is based on multi-copy plasmids and phage systems, where the DNA clearance is most likely to occur when an invader enters as a single copy, as the concentration of invading DNA increases over time. Therefore, depending on the invader and the level of CRISPR interference, there will be a critical time point (t_c) beyond which the invader is permanently established inside the cell and can no longer be cleared (Severinov et al., 2016). Our model describes the probability that it takes a certain copy number of proteins (n) each with an average search time ($\overline{t_s}$) to find the target before t_c is reached.

Our model accurately predicted that pTarget establishment decreases exponentially with increasing copy numbers of Cascade (Figure 1E, Methods). When we translated these establishment probabilities into interference levels, we could deduce that around 20 Cascade complexes are required to reach a CRISPR interference level of 50% (Figure 1F). The exponential relationship further entails every subsequent 20 complexes halve the number of cells not able to achieve interference, which means that 40 Cascade complexes can provide 75% interference; 60 Cascade complexes 87.5%.

It becomes very unlikely for the CRISPR system to destroy multiple genetic copies of the MGE if it has failed to destroy the single copy that was present at the start before replication. Therefore, we can approximate t_c, with the replication time of the plasmid in the absence of copy number control (~3 min, (Olsson et al., 2003a)), which allows us to retrieve an estimated search time of ~90 minutes for one Cascade complex to find a single target in the cell (Methods).

In contrast to pTarget establishment, which decreases exponentially, the average search time decreases linearly with increasing copy numbers of Cascade. Therefore 10 Cascade complexes require approximately 9 minutes to find a single target, while 90 Cascade complexes could achieve this within a minute.

To summarize, we found a direct relation between the number of Cascade complexes and the establishment probability of an MGE. The native *E. coli* system requires 20 Cascade complexes loaded with a cognate crRNA to obtain 50% CRISPR interference levels. This relation depends on the replication rate of the invading

MGE and the average search time of a single complex and demonstrates the importance of rapid target search on CRISPR interference ability.

The majority of Cas8e assembles into the Cascade complex

To quantify the dynamics of target search, we traced the diffusion paths of thousands of individual complexes in the bacterial cell (Figure 2A; Supplementary Video). The apparent diffusion coefficient *D**, a measure for mobility, of Cascade was calculated by extracting the displacement of each fluorescent particle for four consecutive 10 ms steps, allowing us to investigate the abundance, mobility and behavior of individual complexes and subunits in the cell. To minimize the influence of spurious autofluorescent particles in *E. coli* (Floc'h et al., 2018), we used expression levels with the highest estimated Cascade copy numbers (~130 copies, high induction; Figure 1D).

To distinguish diffusion of Cascade complexes from monomeric Cas8e subunits, we first measured the diffusion of the tagged Cas8e fusion protein in a strain lacking genes of the other four Cascade subunits in the genome (Cas11, Cas7, Cas5, and Cas6e). Based on the role of Cas8e in non-specific DNA binding (Brown et al., 2018; Jore et al., 2011; Sashital et al., 2012), we expected to find mobile and DNA-bound populations of Cas8e. However, we were unable to describe the data accurately by static two-state models of non-interconverting fractions (Figure S1). We therefore hypothesized that rapid DNA binding and unbinding events of Cascade on a timescale similar to the framerate (~10-40 ms) would lead to time-averaging of a mobile state (high *D** values) and a DNA-bound state (low *D** values), giving rise to intermediate *D** values (Figure 2A). We accounted for these events by developing a generally applicable analysis method called analytical Diffusion Distribution Analysis (analytical DDA), which is useful for proteins with fast transitioning kinetics between states with different diffusion coefficients, such as DNA-interacting proteins. The distribution of *D** values is not only affected by the fraction of the time spent bound and freely diffusing, but furthermore changes depending on the absolute transition rates (Figure S2). Therefore this method allows us to extract

quantitative information on DNA binding kinetics and enables the study of fast transition rates previously inaccessible to sptPALM (Methods).

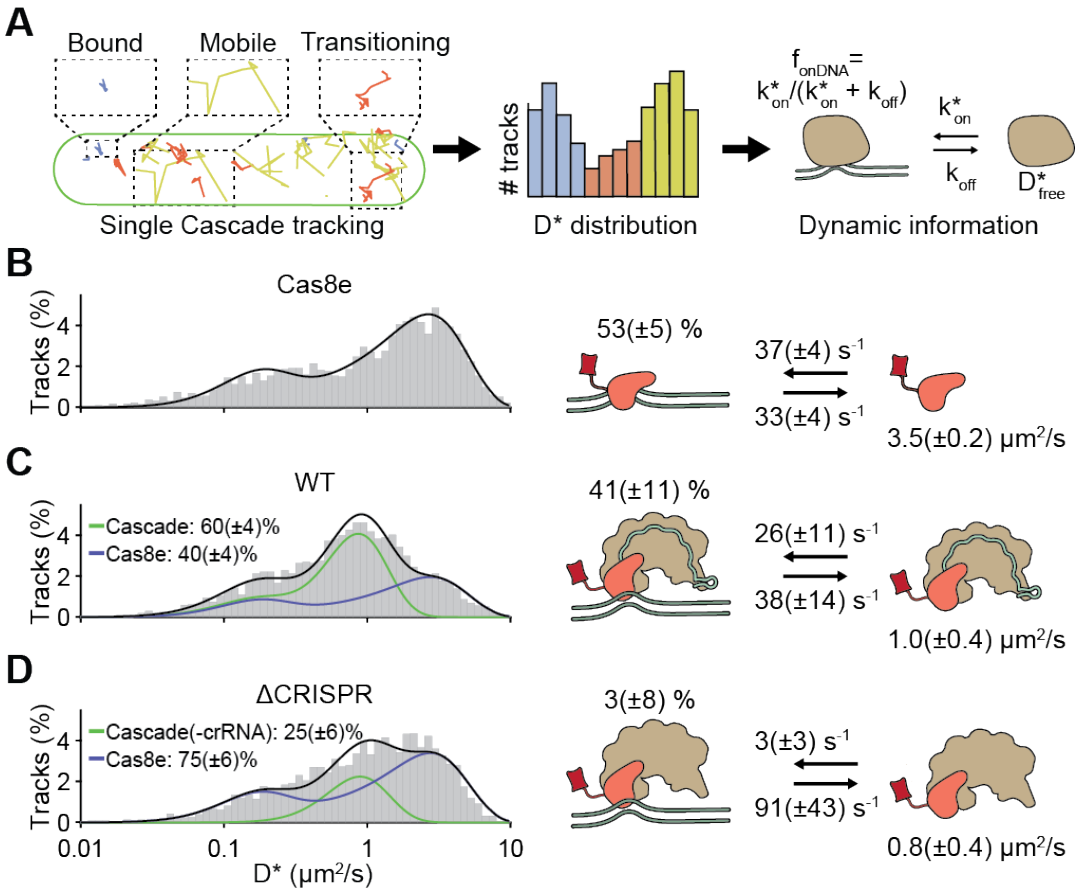


Figure 2: Diffusion behavior of Cas8e and Cascade. (A) Tracks with small (blue), intermediate (orange) and large (yellow) displacements from a single cell of the WT strain (left). The most likely state for three tracks is indicated, although, due to limited track length and fast transitions, states cannot be assigned confidently to every individual track. The *D** distribution (middle), from a large population of tracks, enables reliable extraction of DNA interaction kinetic parameters (pseudo-first order on-rate (k^*_{on}), off-rate (k_{off}) and the apparent free diffusion coefficient (D^*_{free})) by using analytical diffusion distribution analysis (DDA; right). These parameters further allow the calculation of the fraction DNA bound (f_{onDNA}). **(B-D)** *D** distributions for **(B)** Cas8e, **(C)** Cascade and **(D)** ΔCRISPR strain. Total (black), Cas8e (blue) and Cascade (green) fractions fits are indicated by lines. Parameters (right) of Cas8e (B) were used to fit the Cas8e fraction in Cascade (C-D). Error estimation is based on bootstrapping (± standard deviation). See also Figure S1, S2 and S3.

When we applied the analytical DDA on the Cas8e diffusional data, we retrieved an average residence time of ~30 ms on DNA and a similar average time spent (~30 ms) rapidly diffusing (*D** ~3.5 μm²/s, as expected for a protein of 82 kDa; Methods),

indicating that Cas8e is bound to DNA for ~50% of the time (Figure 2B). The D^* distribution of Cas8e then allowed us to extract the diffusion behavior of the Cascade complex as a whole. We estimated the fraction of free Cas8e and Cascade-containing Cas8e at 40% and 60%, respectively (Figure 2C). This finding suggests that Cas8e is produced in excess (Westra et al., 2010) or somehow involved in a dynamic interaction with the core Cascade subunits (crRNA, Cas11, Cas7, Cas5, Cas6e) (Jore et al., 2011; Sashital et al., 2012).

Surprisingly, we found that the DNA binding kinetics of Cascade were similar to Cas8e alone, indicating that Cas8e is an important driver of DNA probing characteristics of the Cascade complex. Furthermore, the DNA residence times are on average ~30 ms and are thereby considerably shorter than the 0.1-10 s that have been reported for *in vitro* studies previously (Brown et al., 2018; Redding et al., 2015; Xue et al., 2017). As expected, we found a smaller diffusion coefficient for unbound Cascade complexes (~1.0 $\mu m^2/s$) (Methods) due to their larger size. Together, our analysis shows that more than half of the Cas8e proteins are part of intact Cascade complexes, and that the DNA interacting behavior of Cascade is largely determined by the properties of Cas8e.

The probing kinetics that we measured determine the number of sites Cascade can scan every minute. The total time Cascade needs to probe a single site includes the average time the complex is bound to a DNA site and the average time it requires to find the next DNA site. The Cascade probing time *in vivo* sums up to roughly 60 ms ($1/k_{\mathrm{off}} + 1/k^*_{\mathrm{on}}$), which implies that the complex is able to scan approximately 1000 DNA sites per minute. The probing kinetics of single sites are furthermore linked to the distributions of target search times, and with simulations we could verify that our model of Cascade DNA scanning indeed leads to the expected distribution of interference levels (Figure S3). Using our previous estimate of the overall target search time for a single Cascade of ~ 90 min, we calculate that the complex scans 90.000 DNA sites in the cell before finding a target (Methods).

To investigate the role of crRNAs in Cascade complex assembly, we deleted all CRISPR arrays in the K12 genome (ΔCRISPR). The resulting diffusion behavior can be described by fractions of free Cas8e and with Cascade-like diffusion behavior (Figure 2D) that almost entirely lacks interaction with DNA (f_{onDNA} = 3%). This

indicates that although Cascade (sub)complex formation does not strictly require the presence of crRNA (Beloglazova et al., 2015; Brouns et al., 2008), Cascade assembly is greatly enhanced by crRNA. Taken together, the majority of Cas8e proteins are incorporated in Cascade complexes in the presence of crRNA, and this gives Cascade DNA interacting properties.

Cascade is enriched but not exclusively present in the nucleoid

Not all potential DNA interaction sites in the host chromosome might be accessible to Cascade. The host DNA is concentrated in the middle of the cell in the nucleoid and is very compact which excludes large complexes such as ribosomes (Mondal et al., 2011). Nucleoid exclusion would reduce the amount of DNA available for scanning and increase the amount of freely diffusing Cascade complexes. To investigate whether the DNA-bound fraction is governed by affinity properties of Cascade for DNA rather than a restricted search space outside the DNA-containing nucleoid region, we studied the spatial distribution of Cascade localizations. Nucleoid-excluded ribosomes are enriched away from the central long axis of the cell (Sanamrad et al., 2014). For Cascade, we found a homogeneous spatial distribution throughout the cell (Figure 3A), indicating that Cascade is small enough to freely scan the nucleoid for target sites.

We furthermore used the spatial distribution of Cascade to extract quantitative information on the DNA-bound fraction. To that purpose, we created a DNA-free environment in the cell by adding cephalexin (Reyes-Lamothe et al., 2014). This antibiotic affects cell wall synthesis and causes cells to elongate, forming DNA-free cytoplasmic space between nucleoids without condensing the nucleoid (Figure 3B). The time Cascade is bound to DNA is inherently linked to the relative amount it spends in DNA-free and DNA containing regions. Therefore, by calculating the relative amount of localizations in both regions (Enrichment Factor; *EF*) we can extract the fraction of time spent on DNA independently from the DDA analysis. Cascade was only moderately enriched (*EF* of 1.8 ± 0.2 fold) in the nucleoid regions (Figure 3C), indicating that Cascade spends a considerable amount of time diffusing in the cytoplasm while not associated with DNA. From the enrichment factor, the

fraction of Cascade complexes bound to DNA can be approximated to 45% (Figure 3D; for derivation see Methods). This value is consistent with the ~ 50% value we extracted from the DDA distribution of Cascade (Figure 2C).

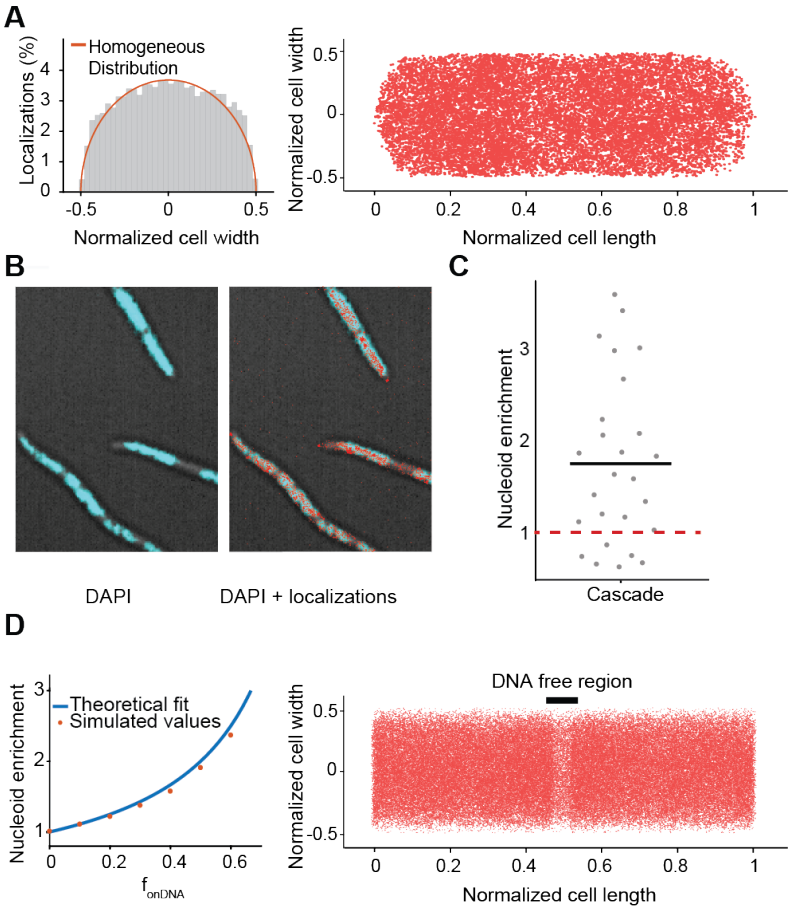


Figure 3: Cascade localization inside the cell. (A) Localization of Cascade in the cell. Left: Distribution of Cascade over the cell width (n = 33 cells; 15428 localizations): in orange is indicated the expected distribution in case of a homogeneous localization within the cell. Right: same localizations plotted within dimensions of single cell in which the cell length and cell width of each cell was normalized. **(B)** Overlay of DAPI fluorescence and brightfield image (left) with Cascade localizations (right) in cephalexin treated cells. **(C)** The nucleoid enrichment in the WT strain (27 subregions in 18 cells). The average ratio is indicated with a black bar. The expected ratio if Cascade has no interaction with DNA is indicated in red (dashed). **(D)** Relation between DNA bound fraction and nucleoid enrichment. Left: A theoretical relation between nucleoid enrichment and DNA bound fraction was derived (Methods) and compared to simulated values for different amounts of f_{onDNA}. Right: Localizations of simulated Cascade proteins (n =50.000) diffusing through part of an elongated cell are plotted on top of long cell axis. A DNA-free region (black bar) is visible due to enrichment of Cascade binding to DNA in nucleoid regions. Simulations of particles were performed with off-rate of 38 s^{-1} and an on-rate of 26 s^{-1} to reach a nucleoid enrichment of 1.8, similar to the average that was found for Cascade.

However, it strongly contrasts other DNA binding proteins such as Fis and RNA polymerase, which show a much higher nucleoid enrichment (Reyes-Lamothe et al., 2014; Stracy et al., 2015). The above findings indicate that Cascade inherently

spends more time freely diffusing the cell and that this is caused by the nature of DNA-Cascade interactions and not by size-based nucleoid exclusion, as is the case for ribosomes (Sanamrad et al., 2014). Therefore, we decided to study the nature of the DNA interactions in more depth.

Cascade-DNA interactions are not only PAM-dependent

Next, we assessed how PAM interactions contributed to DNA binding by introducing mutation G160A in the Cas8e subunit which abolishes the interaction with the PAM (Hayes et al., 2016). This G160A mutation decreased the fraction of DNA-bound Cascade from 41 ± 11 to 28 ± 6% (Figure 4A) without fully inhibiting DNA binding, suggesting that PAM-independent interactions (Van Erp et al., 2015; Hayes et al., 2016; Xiao et al., 2017) play a role in DNA probing as well. To assess the contribution of these different types of interactions to the average DNA residence time found previously, we measured the persistence of Cascade-DNA interactions by increasing the dark time between exposures (Figure 4B). Our data showed that sustained binding events at longer time scales (100 – 250 ms) were more frequently observed for WT Cascade than for the PAM binding mutant complex Cascade-Cas8e$_{G160A}$ (Figure 4C). Together with the increased off-rate of the mutated complex (Figure 4A), this finding demonstrates that PAM-dependent interactions of Cascade with DNA last longer than PAM-independent interactions.

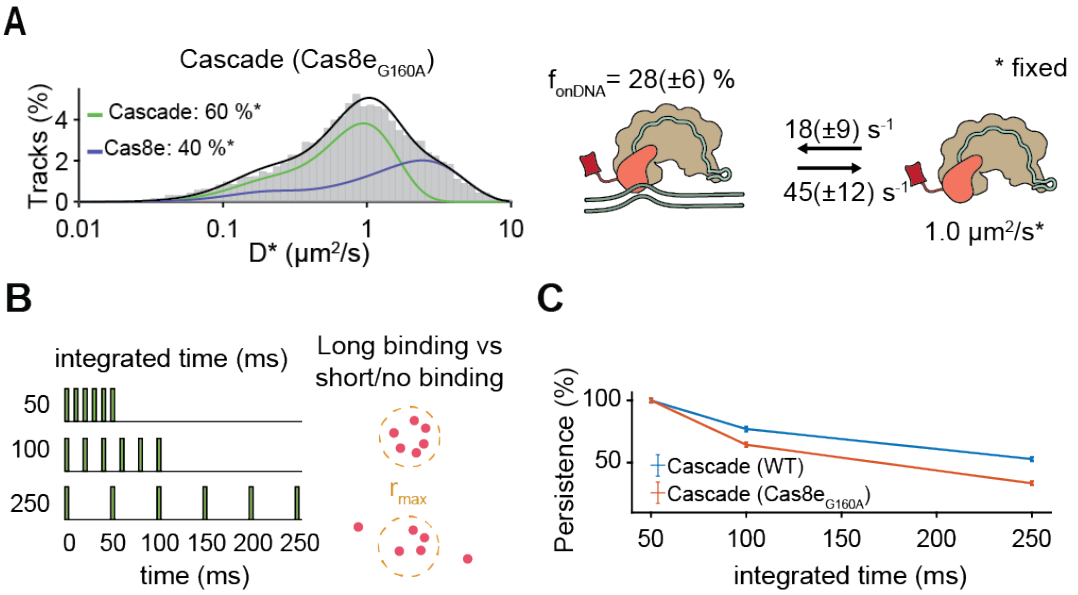
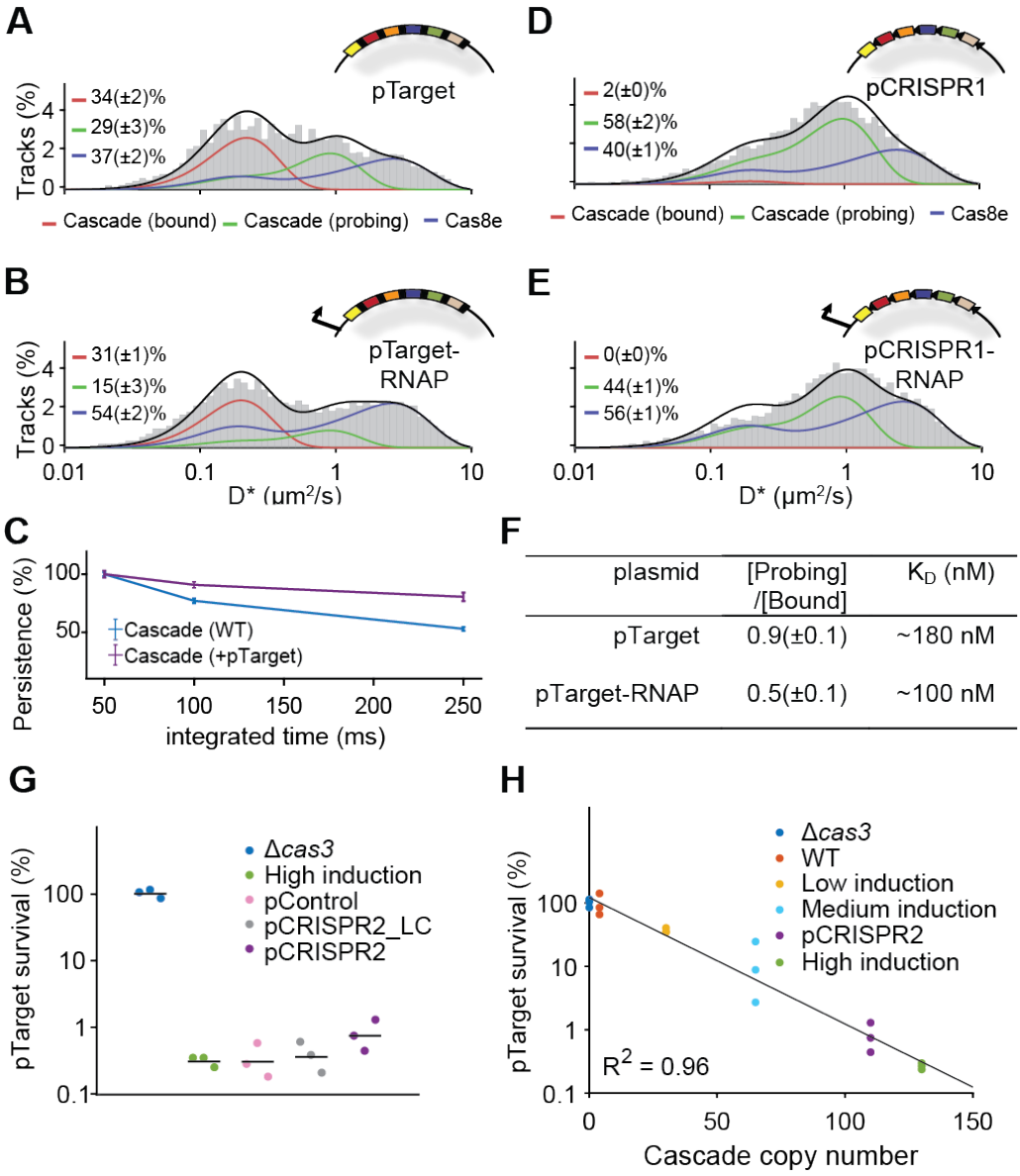


Figure 4: PAM-dependent and PAM-independent DNA probing. (A) *D** distributions for Cascade and Cas8e with a mutation (G160A) deficient in PAM binding. To compare kinetic rates, we assumed that the relative Cas8e-Cascade fractions and the diffusion of free Cascade and Cas8e were not altered by the mutation and those values were fixed. (**B**) Depiction of persistence analysis. Increasing the integration time while keeping exposure time constant and counting the number of localizations within a certain radius allow the calculation of the persistence of binding events. **(C)** The relative amount of long binding events (6 consecutive localizations within r_{max}: 1 pixel (0.128 μm) of the mean position) for WT and PAM binding mutant Cascade normalized to 50 ms integration time. Error estimation in (A) and (C) is based on bootstrapping (± standard deviation).

Target DNA binding is influenced by the cellular environment

After establishing intrinsic DNA probing characteristics of Cascade, we next investigated its diffusion behavior in the presence of targets (Figure 5). To prevent target DNA degradation by Cas3 nucleases, we deleted the *cas3* gene and verified that the deletion did not alter Cascade diffusion behavior (Figure S4). To verify that all Cascade complexes could bind a target, we measured the copy number of pTarget to be ~ 400 copies/cell (Figure S5). As the native *E. coli* CRISPR arrays contain 18 spacers, this resulted in ~7000 target sites per cell which far outnumbers Cascade copy numbers under our growth conditions (~130, Figure 1D).

Compared to a non-targeted control plasmid (Figure S4), the introduction of pTarget in cells decreased the fraction of free Cascade complexes (from 60 ± 4 to 29 ± 3%), and gave rise to a 34 ± 2% immobile, target-bound Cascade fraction ($D^*_{\text{Cascade(bound)}}$ = 0.06 μm^2/s) (Figure 5A). As expected, addition of pTarget increased the persistence of sustained binding events, indicating specific DNA target binding (Figure 5C). The combined information of plasmid copy number and the ratio of probing to target bound Cascade enabled us to determine a cellular K_D value for the affinity of Cascade for targets of ~180 nM (Figure 5F; Methods), indicating that the affinity *in vivo* is around 10 times lower than what has been observed *in vitro* (Hayes et al., 2016).



plasmid	[Probing] /[Bound]	K_D (nM)
pTarget	0.9(±0.1)	~180 nM
pTarget-RNAP	0.5(±0.1)	~100 nM

Figure 5: Cascade - DNA interactions in the presence of targets. (**A** and **B**), *D** distribution for the Δ*cas3* strain carrying pTarget (A) and pTarget-RNAP (B). pTarget contains protospacers for all spacers in the K12 genome (colored, not all depicted) and are flanked by a 5'-CTT-3' PAM (black bars). Cascade (probing) (green) and Cas8e (blue) fractions were fitted with parameters from Figure 1C and 1D, and a new target-bound fraction (Cascade (bound)) was introduced as a single diffusion state (D^* = 0.06 $\mu m^2/s$ ($+\sigma^2/t$); red). (**C**) The abundance of sustained binding events as in Figure 3C, but for WT and pTarget-carrying cells. (**D** and **E**), *D** distribution for the Δ*cas3* strain carrying pCRISPR1 (D) and pCRISPR1-RNAP (E). pCRISPR1 contains the same protospacers as pTarget that are now flanked by repeat PAMs. (**F**) In *vivo* K_D estimates based on the ratio between Probing/Bound Cascade and the plasmid copy number (Figure S5; Methods). (**G**) pTarget establishment for *Δcas3* (blue), WT (high induction; green), an empty high copy plasmid (pControl; pink), and low or high copy plasmids carrying CRISPR arrays (pCRISPR2_LC/pCRISPR2; grey/purple). Each dot represents an independent biological replicate. (**H**) pTarget establishment plotted for different copy numbers of Cascade. Same as Figure 1E but with addition of pCRISPR2. The Cascade copy number of the pCRISPR2 strain was estimated from the relative abundance of the Cascade (probing) fraction in the WT (high induction; Figure 2C) and pCRISPR2 (Figure S4) strain. Each dot represents an independent biological replicate. Error estimation in (A-F) is based on bootstrapping (± standard deviation). See also Figure S4, S5 and S6.

We hypothesized that transcription of DNA along target sites would be one of the main factors influencing Cascade target DNA binding. To investigate the effects of transcription by host RNA polymerase (RNAP), we introduced a (lac) promoter in front of the pTarget sequence. To our surprise, we observed that the affinity of Cascade for target sites that undergo transcription (~100 nM) was higher than for non-transcribed target sites (~180 nM; probing/target bound Cascade from 0.5 ± 0.1 to 0.9 ± 0.1). In addition, we observed an increased fraction of free Cas8e subunits (from 37 ± 2% to 54 ± 2%) in the strain containing transcribed pTarget (Figure 5B). Collectively, these findings suggest that transcription of a target DNA sequence somehow facilitates target search and increases the affinity of a target. In addition, it appears that collisions of RNAP with target-bound Cascade result in changes in the Cascade assembly, likely by dissociation of the Cas8e subunit from the complex upon collision with RNA polymerase, which potentially dissociates Cascade from the target.

The relatively dynamic association of Cas8e within the Cascade complex has been observed previously *in vitro* (Jore et al., 2011) and was more recently also observed upon binding to the CRISPR array (Jung et al., 2017). We hypothesized that this dynamic behavior might be a functional characteristic and will also occur upon encountering CRISPR arrays inside the cell. To test this hypothesis, we made a variant of pTarget where all 18 interference PAMs were replaced by the trinucleotide sequence matching the repeats of the CRISPR array (pCRISPR1). Cascade did not show any interaction with the non-transcribed pCRISPR1 plasmid (Figure 5D). However, when we added a promoter sequence in front of the pCRISPR1 array of targets, we observed moderately enhanced levels of free Cas8e (from 40 ± 1 to 56 ± 1%) (Figure 5E), reminiscent of Cas8e expulsion from the complex upon collision with RNA polymerase, or from targets with repeat like PAMs (Jung et al., 2017). Effectively this shows that transcribed CRISPR arrays may function as target decoys in the cell and can therefore potentially influence the levels of functional Cascade complexes in the cell.

To test whether CRISPR array really form decoys in the cell and could impact interference levels, we constructed a compatible high copy number plasmid pCRISPR2 containing a normal CRISPR array (Figure S6). While the introduction of pCRISPR2 into cells containing pTarget only led to a small decrease in the number of Cascade complexes (15% less) (Figure S4), the CRISPR interference levels were reduced by as much as 50% (Figure 5G). This effect was not observed with low copy variant of pCRISPR2 (pCRISPR2_LC) or with a high copy plasmid lacking CRISPR arrays (pControl), indicating that this effect comes from the presence of a large number of CRISPR arrays in the cell (Figure 5G). We further found that the observed impact of CRISPR arrays on Cascade copy number and interference level fits well with our previously predicted relation between Cascade copy numbers and probability of successful MGE establishment (Figure 5H). It furthermore demonstrates how relatively small changes in Cascade copy numbers (15%) can have a big impact on CRISPR interference levels (50%). Taken together, our data indicate that Cascade target search and binding is strongly influenced by the action of RNA polymerase and that CRISPR arrays form target decoys in the cell, which can affect CRISPR interference levels.

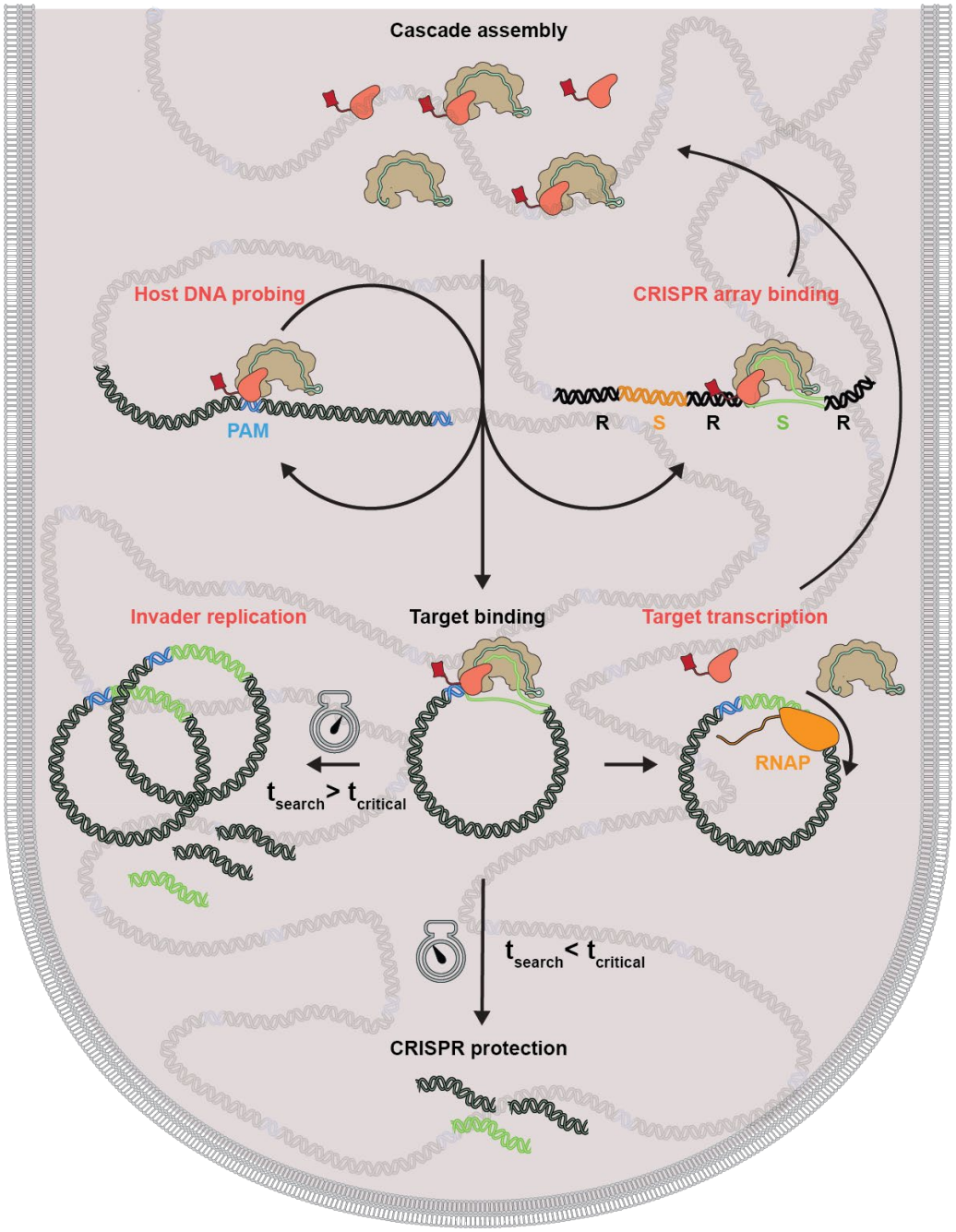


Figure 6: Model of how Cascade protects the cell. Successful protection against an invader requires Cascade target search to circumvent several potential diversions (red). After Cascade is assembled, the complex probes the host DNA by rapidly binding and dissociating. It uses PAM-dependent and PAM-independent DNA interactions and scans the entire nucleoid region. If it binds to a CRISPR array (S: spacer; R: Repeat), the complex disintegrates. When it has found its target, it depends on the search time (t_{search}) and the critical time ($t_{critical}$) whether the invader is cleared and the cell protected, or the invader can replicate and establish itself in the cell. Moreover, transcription by RNA polymerase (RNAP) can still remove bound complexes, compromising CRISPR protection.

Discussion

How crRNA-effector complexes can achieve timely detection of incoming mobile genetic elements in the crowded environment of the cell is an intriguing aspect of CRISPR biology that remains poorly understood. We provide first insights into the fundamental kinetics of the surveillance behavior of type I crRNA-effector complexes in their native cellular environment. We determined how many copies of Cascade are required to establish effective immunity and uncovered how Cascade complexes navigate the crowded bacterial cell packed with DNA. Our results indicate that Cascade does not restrict its search space to parts of the cell, for example the nucleoid-free periphery, but instead is occupied scanning the entire host nucleoid for a match. Depending on genome size of a microbe and the number of copies of the genome in the cell, the nucleoid size may vary widely. To cover this vast sequence space sufficiently fast, the Cascade complex interrogates DNA sequences by using a combination of PAM-dependent and PAM-independent interactions which on average last only 30 ms. This probing interaction is much faster than previously reported interaction times determined of type I Cascade complexes by *in vitro* methods, which range between 0.1 and 10 s (Brown et al., 2018; Redding et al., 2015; Xue et al., 2017). The ability to rapidly probe DNA sequences for potential matches with the crRNA, and to move from one place in the nucleoid to the next, may explain how a relatively low number of Cascade complexes in *E. coli* may still confer CRISPR immunity. Interestingly, the average probing time of 30 ms for Cascade matches values found for *Streptococcus pyogenes* dCas9 in *E. coli* (Jones et al., 2017; Martens et al., 2018), suggesting that DNA probing interactions of crRNA-effector complexes from both Class I and II systems may have evolved independently to take place at this time scale.

The probing kinetics we measured for Cascade will allow the complex to scan 1000 DNA sites per minute. Given the abundance of PAMs in the host DNA, this interaction time would lead to a search time in the order of hours. This value matches our independently calculated estimate of 1.5 hours for a single Cascade to find a single DNA target in the cell, which is four times faster than dCas9 search time estimates of 6 hours (Jones et al., 2017). However, our data also indicates that Cascade not only probes PAMs, the complex also spends a considerable amount of

time engaged in PAM-independent DNA interactions. These might be constituted by direct crRNA – DNA interactions (Blosser et al., 2015; Xue et al., 2016), or electrostatic interactions of Cascade with the DNA (Van Erp et al., 2015; Hochstrasser et al., 2014). This suggests an even larger DNA sequence space needs to be covered, creating the need for even more efficient and functionally flexible surveillance solutions. This more flexible probing behavior would be required to recognize targets with mutations in the PAM or protospacer in order to trigger a CRISPR memory update pathway called priming (Datsenko et al., 2012; Jackson et al., 2017), which appears to be unique for type I CRISPR-Cas systems.

One possibility to reconcile Cascade DNA probing characteristics to the overall search time could be that Cascade undergoes facilitated 1D DNA sliding, where Cascade probes multiple sites per DNA binding event. We have shown that Cascade spends 50% of its search time on DNA, and the other 50% diffusing to a new site in the cytoplasm. This value may seem low compared to other DNA interacting proteins such as transcription factor LacI, which is DNA bound for 90% of the time (Elf et al., 2007). However, 50% has been theoretically derived as the optimum for a target search process involving one-dimensional DNA sliding and 3D translocation/hopping (Slutsky and Mirny, 2004). Indeed, recently it has been shown *in vitro* that Cascade and Cas9 can slide along the DNA in search of targets (Brown et al., 2018; Globyte et al., 2018). If this also occurs *in vivo*, this would be a striking example of a DNA binding protein having an optimized time division between DNA-bound and freely mobile states to survey the DNA content of the cell.

The relatively high abundance (50%) of freely diffusing Cascade complexes may have benefits as well, as this will lead to more Cascade complexes in the periphery of the cell outside of the nucleoid. By surveying these peripheral regions more frequently, Cascade may be able to detect incoming bacteriophage or plasmid DNA more rapidly when these genetic elements enter the cell.

Besides the chromosomal host DNA, other cellular constituents also affect target DNA binding properties. We found a much higher K_D value *in vivo* (180 nM) than was reported earlier using *in vitro* methods (20 nM) (Hayes et al., 2016). The discrepancy in binding affinity between *in vivo* and *in vitro* measurements may be caused by an increase in target search time (i.e. a lower on-rate) or an increase in

target dissociation rate (i.e. a higher off-rate) *in vivo*. In any scenario, this discrepancy highlights the strong role of the crowded cellular environment on target binding.

Counterintuitively, we have found that Cascade binds transcribed target sites with higher affinity (100 nM) than non-transcribed target sites (180 nM). Previous studies have shown that negative-supercoiling is required for Cascade binding (Westra et al., 2012), and that increased negative super-coiling accelerates the rate of R-loop formation (Szczelkun et al., 2014). As transcribed regions cause more negative supercoiled regions in the DNA (Ma and Wang, 2016), this could explain the increase in the affinity for transcriptionally active sites. Rates of spacer acquisition were also found to be higher for transcriptionally active regions (Staals et al., 2016), so together these effects may influence the abundance and effectivity of spacers in nature.

Next to the positive effect of transcription on target search, we have also found that collisions between RNAP and target-bound Cascade lead to Cascade disassembly, where the Cas8e subunit is expelled from the Cascade core. Furthermore, CRISPR arrays themselves can trigger Cascade disassembly, indicating they form target decoys in the cell. When present at high copy number, CRISPR arrays can even impact CRISPR interference levels (Fig. 5G). The loose association of Cas8e with the core Cascade complex as observed *in vitro* (Jore et al., 2011), might serve a biological role in cells to recycle Cascade from off-targets including the CRISPR array, and may prevent Cas3 recruitment and subsequent self-targeting (Xiao et al., 2018).

By measuring cellular copy numbers, and accurately measuring CRISPR interference levels, we could uncover an exponential relationship between the number of Cascade complexes in the cell and CRISPR interference. This relationship describes that every 20 Cascade complexes loaded with one crRNA can provide 50% more protection from an invading DNA element (i.e. 20 copies provide 50%, 40 copies 75% protection). Therefore at constant Cas protein production and degradation levels, the effective concentrations of Cascade complexes loaded with one type of crRNA will become diluted when CRISPR arrays become longer. The size of the CRISPR array is therefore a tradeoff between the higher protection levels

of a few spacers, and lower protection levels of many spacers. With our findings we can test optimality of this tradeoff under different conditions and help explain the observed sizes of CRISPR arrays found in nature (Martynov et al., 2017).

The initial entry is the most vulnerable time for the invader, but invading MGEs have the possibility to outrun CRISPR-Cas immunity by replicating faster than being found. In the native cellular environment, we have found that scanning of host DNA, binding to CRISPR arrays and encountering transcribing RNA polymerases can prevent Cascade from finding the target before the critical time (t_c) is reached and the invader is permanently established (Figure 6). We therefore hypothesize the presence of a kinetic arms race, in which invaders have evolved to replicate increasingly fast upon cell entry, while CRISPR-systems have evolved to increase the rate at which they are able to find the target. A recent study has indeed shown that the replication rate of foreign elements affects CRISPR interference levels (Høyland-Kroghsbo et al., 2018). Many bacteriophages use a two-stage injection (Chen et al., 2018; Davison, 2015), which may have evolved to limit the amount of time their DNA is exposed to intracellular defense mechanisms, while already allowing the production of proteins to replicate phage DNA, control host takeover, or to inhibit host defense (e.g. anti-CRISPR proteins) (De Smet et al., 2017). It has been previously shown that the host can counter this strategy by selectively targeting early injected DNA regions, maximizing the time available to look for targets (Modell et al., 2017).

The specificity and kinetics of the CRISPR-system inside the crowded cellular environment is remarkable. Our study has observed very rapid scanning of DNA sites by Cascade complexes and our model predicts the impact of probing kinetics and copy numbers of Cascade on protection levels of CRISPR-Cas systems. We believe that not only specificity and evasion strategies such as anti-CRISPRs but also target search and infection kinetics have played an important role in the evolution of this immune system. The target search equations established here could be expanded to the population level, allowing to model how individual variability in Cascade expression levels and replication rates can impact the survival of entire populations. Therefore, our data provides an important framework for further

quantitative cellular studies that will address how CRISPR systems optimally deal with the challenges of cost-effective and rapid target search.

Acknowledgements The authors thank Dr. A. Košmrlj (Princeton University) for deriving equation 29 presented in Methods. We acknowledge S. Creutzberg for supplying plasmid pSC020 and M. Siliakus for supplying plasmid pMS011 and all members of the Hohlbein and the Brouns groups for input during group discussions. We thank Jaap Keijsers, Fiona Murphy and Stan van de Wall for providing preliminary measurements and scripts for data analysis. S.B. is supported by the European Research Council (ERC) Stg grant 639707, and by a Vici grant of the Netherlands Organisation for Scientific Research (NWO). R.M is supported by the Frontiers of Nanoscience (NanoFront) program from NWO/OCW.

Author contributions

S.B. and J.H. conceived and supervised the project; J.V., M.V. R.M., C.A., D.B., B.B. did the experimental work; J.V. and J.H. derived the theory; J.V. and K.M. wrote analysis scripts; J.V., K.M. and J.H. established microscopy workflow, J.V., J.H. and S.B. wrote the manuscript with input from all authors.

Declaration of Interests

The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.B. (stanbrouns@gmail.com) or J.H. (johannes.hohlbein@wur.nl).

Star Methods

Lead Contact and Materials Availability

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Stan Brouns (stanbrouns@gmail.com).

Cloning

The inserts to create pTarget and pCRISPR1 plasmids were purchased as synthetic constructs from Gen9 (pTarget insert and pCRISPR1 insert; Table S3). To increase the copy number of targets in the cell, the constructs were cloned into a pUC19 backbone with XbaI and KpnI restriction sites, yielding pTarget-RNAP and pCRISPR1-RNAP. The lac promoter was removed for both plasmids by digestion with SalI and PciI, creating blunt ends with Klenow Fragment and subsequently religated to yield pTarget and pCRISPR1. CRISPR arrays were amplified from the K12 BW25113 strain (primers BN383 and BN384; BN370 and BN385 for CRISPR array 2.1 and 2.3 respectively) and cloned into pJPC-12 plasmid containing the pSC101 ori with KpnI and SalI sites (for CRISPR array 2.1) and SalI and EcoRV sites (for CRISPR array 2.3). The copy number of the plasmid could be varied by introducing mutations in the *repA* gene with site-directed mutagenesis PCR (BN373-375). The E96R mutation of RepA yields a reported copy number of ~240/cell (pCRISPR2) compared to the WT RepA (pCRISPR2_LC) copy numbers of ~7/cell (Peterson and Phillips, 2008). A plasmid was made from the high copy-variant that did not contain any CRISPR arrays (pControl). All constructs were verified by sequencing.

Recombination

The strains used in this study were created by using Lambda red recombineering (Datsenko and Wanner, 2000). Strains harbouring the pSC020 plasmid that contains both the Lambda red recombinase and Cre-recombinase were grown at 30 °C. Before transformation of an insert containing an antibiotic resistance marker, the expression of Red recombinase was induced with 0.2% L-Arabinose. Colonies on the specific

antibiotic plate were verified with PCR and sequencing and subsequently Cre recombinase expression was induced with 1 mM IPTG at 37 °C to promote plasmid and antibiotic resistance gene loss. The strain was subsequently patch plated to screen for resistance sensitivity due to plasmid loss.

If the scar that is left after lox-site recombination is directly upstream or downstream of a gene it might influence gene transcription/termination. In the design of constructs for *pamcherry2* (Subach et al., 2009) the lox-*cat*-lox sequence was placed upstream of the IGR (Intergenic region) that is present between *cas3* and *cas8e*. To allow for correct termination of *cas3,* a part of the IGR was also added at the 5' end of the antibiotic resistance marker. The 3' flank of the constructs overlapped with the *cas8e* gene. The 5' flank of the constructs matched a sequence upstream and downstream of *cas3* (PAmCherry ins; Table S3). Amplification of the constructs with a forward primer matching the downstream region kept *cas3* intact upon insertion (BG7128), whereas a primer matching the upstream region deleted the *cas3* gene allowing measurements in the presence of targets (BG7129). The insert also contained a part of the *cas8e* sequence containing a G160A mutation. This mutation could be introduced into the gene simultaneously with the fluorescent protein, depending on the reverse primer that was used for insert amplification (BG7130 for WT, BG7131 for G160A).

Knockouts of the CRISPR arrays and Cas gene subunits of the K12 strain were made by amplifying a lox-*kan*-lox or lox-*cat*-lox sequence with flanks matching the specific sequences and introducing them into the strain as described above (BG7366+BG7367 for CRISPR array 2.1; BG7368+BG7369 for CRISPR array 2.2+2.3; BG8366+BG8367 for Δ(*cas11-cas6e*)). A full overview of the sequences of these inserts is given in Table S3.

Growth conditions

To prevent the high-copy target plasmids from influencing the growth rate of the strains and therefore changing the fraction of matured PAmCherry complexes we used a rich defined medium with minimal autofluorescence. Strains were grown in M9 minimal medium containing the following supplements: 0.4% glucose, 1x EZ

amino acids supplements (M2104 Teknova), 20 µg/ml uracil (Sigma-Aldrich), 1mM $MgSO_4$ (Sigma-Aldrich) and 0.1 mM $CaCl_2$ (Sigma-Aldrich) (further referred to as M9 medium). Strains were inoculated o/n from glycerol stocks and 200x diluted in fresh medium the next day. Cells were always grown with the required antibiotics. The expression level of Cascade for strains carrying the pKEDR13 plasmid could be tuned by different expression levels of LeuO. The expression level referred to in the text as low induction was achieved by leaky expression of LeuO (no addition of IPTG), medium induction was achieved by addition of low levels of IPTG (0.01 mM), whereas high induction was achieved by addition of 1 mM IPTG upon dilution of the o/n culture. For all sptPALM measurements the high induction condition was used. The cells were grown for ~2.5 hours to an OD of 0.1 before use. For enforced elongation of cells, cephalexin (40 µg/ml) was added 0.5 hour after fresh inoculation and grown for two more hours. When required, DAPI for staining of DNA was added right before imaging (0.5 mg/ml).

Transformation assay

Each culture was grown under conditions described above and 30 ml were used to create competent cells. Cells were washed 3 times in ice-cold 10% glycerol solution and the final culture was reduced to 250 µl. The cells were aliquoted and stored at -80 °C. A mixture of pTarget (10 pg/µl) and pGFPuv (10pg/µl) was transformed into 40 µl of culture. In case of strong interference levels, the ratio was adjusted to a 100:1 (pTarget (100 pg/µl):pGFPuv (1 pg/µl)). The transformability of strains was linear in these concentration regimes, allowing these different relative concentrations to be used.

Electroporated cells were immediately plated in two dilutions on plates containing ampicillin (100 µg/ml) and glucose (0.4%). Glucose was added to prevent premature expression of GFPuv which would cause a decrease in fitness of cells containing this plasmid. The next day, 96 colonies from each replicate were reinoculated in 96-wells plate with LB containing ampicillin (100 µg/ml) and IPTG (1 mM). After overnight incubation, the 96 well colonies were analysed in a plate-reader (Synergy H1, Biotek). pTarget establishment was defined as

$$p_{\text{establishment}} = \frac{\#\text{ pTarget colonies}}{\#\text{ GFPuv colonies}} \frac{[\text{pGFPuv Transformed}]}{[\text{pTarget Transformed}]} \quad (1)$$

pTarget establishment was further normalized to the interference level of a Δ*cas3* strain.

qPCR

Each culture grew under conditions described above and 2 ml were used to extract the DNA. DNA was isolated with the Genejet Genomic DNA kit (Thermo Scientific) and concentrations were measured with the Qubit dsDNA HS Assay kit (Thermo Scientific). qPCR was performed with primers that have been used before in plasmid copy determination (BG8677-BG8680) (Reyes-Lamothe et al., 2014). The Ct value of the PCR amplifying the *dxs* gene and the *bla* gene was a measure for the ratio between chromosomal and plasmid DNA. 1 ng of genomic DNA and 0.5 µM of each primer were added to the iTaq^{TM} SYBR Green SYBR Green PCR reaction mixture. A standard curve for the amplification efficiency was made by a dilution series of pMS011, a plasmid containing one copy of the *dxs* and the *bla* gene.

Slide preparation

In order to work with very clean slides, an extensive cleaning procedure was used (modified from (Chandradoss et al., 2014)). Slides were burned in the oven at 500 °C for two hours, and stored in aluminium foil until the day of usage. Slides were subsequently sonicated in MilliQ, Acetone and KOH, incubated in Piranha Solution (75% H_2SO_4, 7.5% H_2O_2) and afterwards rinsed with MilliQ. 1% Agarose slabs containing the growth medium were hardened between two cleaned glass slides, spaced slightly apart using parafilm. After hardening, a concentrated culture of cells was added in between the slab and one of the slides. The agarose slab was always prepared within 20 minutes of the measurement to prevent desiccation.

Microscope set-up

For the acquisition of microscopy data, a home-build TIRF microscope was used, which is described in more detail elsewhere (Martens et al., 2018). Briefly, four lasers with different wavelengths (405, 473, 561 and 642 nm) are situated in a Lighthub laser box (Omicron, Germany), and are transformed in a collimated beam via a reflective collimator and an optical fibre. Stroboscopic illumination was used to allow for 2 ms excitation in the temporal middle of the captured 10 ms long frame (Farooq and Hohlbein, 2015). The excitation laser is focused on the backfocal plane of a 100x oil immersion SR/HP objective (NA = 1.49, Nikon, Japan), and the emission is captured on a Zyla 4.2 plus sCMOS camera (Andor, UK). 2x2 pixel binning was used, resulting in 128x128 nm pixels. Data acquisition was performed using MicroManager (Edelstein et al., 2010). Measurements were performed at room temperature (21 °C)

Single-molecule Measurements

The cells were imaged with a brightfield light and 405 and 561 nm lasers. First brightfield images were taken to find contours of the cells. The 405 nm laser was used to stochastically activate PAmCherry and the laser intensity was slowly increased during the measurement up to 10 μW. The laser intensities were measured directly after the reflective collimator. With increasing the laser intensity of the 405 nm laser during the measurements, we aimed at keeping the number of activated molecules relatively constant (~1-10 per FOV). The 561 nm laser was used to excite the fluorescent protein tags (40 mW pulses with 2ms pulse width, leading to average exposure intensity of 8 mW).

To measure Cascade localization in cephalexin-treated cells that were stained with DAPI, we took an alternative approach. To prevent DAPI fluorescence from influencing the fluorescence measurements of the single molecules, we briefly activated a subset of particles with the 405 nm laser and subsequently tracked Cascade for a couple of frames with 561 nm excitation, repeatedly doing this, until most fluorescent proteins were photobleached.

Analysis

Detection, localization and tracking

Analysis was done with home-built software, adapted from (Holden et al., 2010; Uphoff et al., 2013). The sCMOS camera we used has pixel dependent offset, gain and variance, which we took into account to minimize the detection of false positive localisations. We estimated these parameters by measuring 60.000 dark frames and 20.000 homogeneously illuminated frames with increasing levels of intensity (Vliet et al., 1998). To further optimize our detection, we implemented a temporal median filter (time window 400 frames) for background estimation (Hoogendoorn et al., 2015). The background estimate was not directly subtracted from the image, but photon statistics were incorporated in a likelihood-ratio test that calculated the probability of a scenario with and without an emitter for each pixel in every frame. Briefly, a raw image was first converted into photon counts by using the camera offset and gain maps. Subsequently for every pixel the intensity (I_{tot}) of a potential emitter was estimated by Gaussian-weighted (σ =1 pixel) summation of a 7x7 window to a background subtracted image. Subsequently, potential emitters of more than 50 photons were preselected and were further subjected to a ratio test. The ratio test uses the probability defined for pixel *i* to have a transformed value *v* in the 7x7 region around the preselected pixels as previously described (Huang et al., 2013):

$$p_{\mathrm{scmos}}(v = [(d_i - o_i)/g_i + var_i/g_i^2]|\mu_i, var_i, g_i, o_i) = \frac{e^{-(\mu_i + var_i/g_i^2)}(\mu_{i+} var_i/g_i^2)^v}{\Gamma(v+1)} \quad (2)$$

Where d_i is the raw image value, g_i is the gain, var_i the variance and o_i the offset for pixel *i*. The ratio test calculates the product of the probability of all pixels in the subregion in case of an emitter $\mu_i = b_i + I_i$, where b_i is the estimated background an I_i is the estimated intensity of the emitter at pixel *i* (which was estimated by a Gaussian from the centre of the 7x7 subregion with emitter intensity I_{tot}) divided by the product of the probability of all pixels in the subregion in case of absence of an emitter $\mu_i = b_i$.

We set the likelihood to a level that achieved approximately one false positive per frame of 512 x 512 pixels. This method allowed the detection efficiency to be more robust across and between FOVs and independent of manual thresholding for each

measurement. Detected particles were subsequently localized with MLE-sCMOS software as previously described (Huang et al., 2013).

The localized particles were subsequently linked. Localizations in subsequent frames which were closer to each other than 6 pixels in length (0.78 μm) were assigned as a track. Particles were allowed to disappear for one frame (due to blinking/moving out of focus), but these steps were not used in the calculation of the apparent diffusion coefficient, *D**.

Determination of diffusion coefficients

Several methods were employed to extract diffusion states and their abundances from the analysed tracks. The distribution of the apparent diffusion coefficients can be fitted to an analytical equation as reported earlier (Stracy et al., 2015; Vrljic et al., 2002). These equations depend on the number of steps that is used to generate the average diffusion coefficients of each particle. We used tracks containing a minimum of four steps and only four steps were used in longer tracks.

For a single diffusion coefficient fitting becomes:

$$f_D(x; D, n) = \frac{\left(\frac{n}{D + \sigma^2/\mathrm{dt}}\right)^n x^{n-1} e^{-\frac{nx}{D + \sigma^2/\mathrm{dt}}}}{(n-1)!} \quad (3)$$

With multiple states this equation becomes:

$$f_D(x; A_i, D_i, n) = \sum_{i=1}^{N} A_i \frac{\left(\frac{n}{D_i + \sigma^2/\mathrm{dt}}\right)^n x^{n-1} e^{-\frac{nx}{D_i + \sigma^2/\mathrm{dt}}}}{(n-1)!} \quad (4)$$

Where A_i are the fractions ($\sum A_i = 1$), D_i* are the apparent diffusion coefficients of the different states and *n* are the number of steps. The localization error (σ) was found to be 40 nm, based on the apparent diffusion of the slowest moving fraction in our global data set and similar to other studies using the same fluorescent protein (Stracy et al., 2015; Uphoff et al., 2013) or set-up (Martens et al., 2018). This equation was fitted to our track distributions with a Maximum Likelihood Estimation algorithm. The uncertainty in the fit was estimated with Bootstrap resampling. The

list of *D** values was resampled 20.000 times with replacement to the size of the original data set. Each resample was then fitted with the same Maximum Likelihood Estimation algorithm.

Analytical Diffusion Distribution Analysis (DDA)

*D** Distributions have been fitted in numerous studies of DNA binding proteins (see above) (Stracy et al., 2015; Vrljic et al., 2002), making use of distributions developed by Qian *et al.* (Qian et al., 1991). The goal is to find the distribution of measured *D** values (x), for a certain number of underlying states that each have a probability A_i and a diffusion coefficient D_i. It is derived from repeated convolution of the exponential distribution of displacement, resulting in a gamma function for each state. These distributions assume, however, that there is no transitioning occurring between states.

In order to incorporate dynamics of state transitions into our fitting, we incorporated statistics coming from photon distribution analysis (PDA) that is used for single molecule FRET diffusion coefficient distributions (Antonik et al., 2006; Kalinin et al., 2008; Palo et al., 2006). This method, that we term Diffusion Distribution Analysis (DDA), describes the distribution of time spent in each state given a certain k^*_{on}, k_{off} and the integrated time t_{int}. Here we discuss the analytical way to find this distribution.

Firstly, the probability distribution function for time can be calculated by three equations corresponding to 0, an odd and an even number of transitions (Palo et al., 2006):

$$W_{\mathrm{cont}S1}(t_{S1} = t_{\mathrm{int}} | k_{\mathrm{off}}, t_{\mathrm{int}}) = e^{-k_{\mathrm{off}} t_{\mathrm{int}}} \tag{5}$$

$$W_{\mathrm{odd}S1}(t_{S1} |\, k_{\mathrm{off}}, k^*_{\mathrm{on}}, t_{\mathrm{int}}) = k_{\mathrm{off}} e^{-k_{\mathrm{off}} t_{S1} - k^*_{\mathrm{on}} t_{S2}} I_0\left(2\sqrt{k_{\mathrm{off}} k^*_{\mathrm{on}} t_{S1} t_{S2}}\right) \tag{6}$$

$$W_{\text{even}S1}(t_{S1}|k_{\text{off}}, k^*_{\text{on}}, t_{\text{int}}) = \sqrt{k_{\text{off}}k^*_{\text{on}}t_{S1}/t_{S2}}e^{-k_{\text{off}}t_{S1}-k^*_{\text{on}}t_{S2}}I_1(2\sqrt{k_{\text{off}}k^*_{\text{on}}t_{S1}t_{S2}}) \quad (7)$$

Where t_{S1} and t_{S2} are times spent in state *S1* and state *S2* and I_0 and I_1 are Bessel functions of order zero and one respectively. Note that $t_{S1} + t_{S2} = t_{\text{int}}$. Equations for starting in state 2 ($W_{\text{cont}S2}$, $W_{\text{odd}S2}$ and $W_{\text{even}S2}$), can be found by exchanging k_{off} for k^*_{on} and t_{S1} for t_{S2} and vice versa in equations 5-7.

We can convert the time spent in the mobile state (t_{S2}) to the diffusion coefficient by the following equation:

$$D = \frac{D_{\text{free}}t_{S2}}{t_{\text{int}}} \quad (8)$$

It follows that the probability distribution functions can be converted by:

$$W(D) = W\left(t_{S2} = \frac{Dt_{\text{int}}}{D_{\text{free}}}\right) \quad (9)$$

Furthermore, the chance that the particle at the start is in state 1 or state 2 is provided by:

$$p_{S1} = \frac{k^*_{\text{on}}}{k^*_{\text{on}} + k_{\text{off}}} \quad (10)$$

$$p_{S2} = \frac{k_{\text{off}}}{k^*_{\text{on}} + k_{\text{off}}} \quad (11)$$

To correctly describe the distribution over a certain number of frames, we first calculated the distribution over a single time frame t_f. Within a single frame, a particle started in that state can either end in the same state or in a different state. Therefore, in a two-state system the probability function for four scenarios have to be calculated:

$$W(D|k_{\text{off}}, k^*_{\text{on}}, t_f)_{\text{S1}\rightarrow\text{S1}} = W_{\text{even}S1}(D) + W_{\text{cont}S1} \quad (12)$$

$$W(D|k_{\text{off}}, k^*_{\text{on}}, t_f)_{\text{S1}\rightarrow\text{S2}} = W_{\text{odd}S1}(D) \quad (13)$$

$$W(D|k_{\text{off}}, k^*_{\text{on}}, t_f)_{\text{S2}\to\text{S1}} = W_{\text{odd}S2}(D) \tag{14}$$

$$W(D|k_{\text{off}}, k^*_{\text{on}}, t_f)_{\text{S2}\to\text{S2}} = W_{\text{even}S2}(D) + W_{\text{cont}S2} \tag{15}$$

Subsequently the probability to find a certain diffusion coefficient (x) for a single time step given the underlying average diffusion coefficient (D) is given by $f_D(x|D,1)$ (Eq. 3). Then we find the distribution of measured diffusion coefficients for a single frame by:

$$W\left(x\middle|k_{\text{off}}, k^*_{\text{on}}, t_f\right)_{\text{Si}\to\text{Sj}} = f_D(x|D,1)\, W\left(D\middle|k_{\text{off}}, k^*_{\text{on}}, t_f\right)_{\text{Si}\to\text{Sj}} \tag{16}$$

$i = j = 1,2$

Now that we have the distribution for a single time step, we need to find the distribution for the average of multiple frames. For this we use the same method as Qian *et al.* (Qian et al., 1991), namely repeated convolution of the distribution for a single frame, while keeping track of the start and end state. The probability distributions are therefore:

$$W(x|2t_f)_{\text{S1}\to\text{S1}} = \sum_{i=1,2} (W(x|t_f)_{\text{S1}\to\text{Si}} * W(x|t_f)_{\text{Si}\to\text{S1}}) \tag{17}$$

$$W(x|2t_f)_{\text{S1}\to\text{S2}} = \sum_{i=1,2} (W(x|t_f)_{\text{S1}\to\text{Si}} * W(x|t_f)_{\text{Si}\to\text{S2}}) \tag{18}$$

$$W(x|2t_f)_{\text{S2}\to\text{S1}} = \sum_{i=1,2} (W(x|t_f)_{\text{S2}\to\text{Si}} * W(x|t_f)_{\text{Si}\to\text{S1}}) \tag{19}$$

$$W(x|2t_f)_{\text{S2}\to\text{S2}} = \sum_{i=1,2} (W(x|t_f)_{\text{S2}\to\text{Si}} * W(x|t_f)_{\text{Si}\to\text{S2}}) \tag{20}$$

For 4 frames, the distributions found for 2 frames can be convoluted again. The full distribution is then found by summing up each of the partial distributions multiplied by the chance they start in *S1* or *S2*:

$$\begin{aligned} W_{\text{tot}} = {} & p_{S1}(W(x|4t_f)_{\text{S1}\to\text{S2}} + W\left(x\middle|4t_f\right)_{\text{S1}\to\text{S1}}) \\ & + p_{S2}(W(x|4t_f)_{\text{S2}\to\text{S1}} + W(x|4t_f)_{\text{S2}\to\text{S2}}) \end{aligned} \tag{21}$$

We then have to further correct for the broadening of the distribution of immobile particles where the apparent step size comes from localization error (Figure S2). As localization error, in contrast to diffusion, is correlated (Michalet, 2010), the distribution is not described by a gamma distribution, or any other known exact solution. We find very close agreement with simulations when we subtract the fraction of immobile particles after four time steps ($W_{contS1}(t_{S1} = 4t_f)$, Eq.5) multiplied with the distribution of expected D^* for four time steps $f_D(x|0,4)$ (Eq. 3) and replace it with the same fraction of immobilized particles multiplied with the distribution of expected D^* for 2.9 time steps $f_D(x|0,2.9)$. This value stems from the variance found for correlated MSD values due to localization error (Michalet, 2010).

The assumptions that underlie this model are as follows:

- Each diffusing species can be in two states, namely an immobile and a mobile state.
- The immobile state in our case includes all species bound to chromosomal DNA, including potential 1D sliding events, for which the diffusion is at such a low relative speed that we can consider them as immobile. Our model therefore cannot distinguish between bound and 1D sliding species.
- The immobile state is still perceived as diffusing due to a localization error, σ, which in our case is 40 nm. As the distribution of sequential localization errors differs from sequential diffusion steps we correct for this (Figure S2).
- The mobile state is defined by the parameter D_{free}, which is the diffusion coefficient of a species in the absence of interactions with DNA. All slowing down in the motion because of transient DNA interaction are captured in our model by the introduction of transitions and do not affect the value found for this parameter.
- The transition between the two states for each species is Markovian, meaning that transition rates are independent of past or future states.

For each species that you fit there are four degrees of freedom, namely the abundance of the species in the total population and the three kinetic parameters k^*_{on}, k_{off}, and D_{free}. However because the sum of all fractions of species is one ($\sum c = 1$) and the sum of the average time spent freely diffusing multiplied by the free diffusion coefficient for each species is equal to the average measured Diffusion Coefficient ($< D > = \sum c_i \frac{k_{\mathrm{off,i}}}{k_{\mathrm{off,i}}+k^*_{\mathrm{on,i}}} D_{\mathrm{free},i}$), the amount of free fitting parameters is reduced by two. This means that for a single diffusing species (in our case monomeric Cas8e) we only need to fit two parameters and for a two-species distribution (in our case Cascade) for which one is already known (Cas8e) we need to fit three parameters (8 degrees of freedom – 3 already known Cas8e kinetic parameters – 2 from the above described equations). We found that the uncertainty of our fit, determined by bootstrapping and simulations, is reasonable up to three fitting parameters, therefore we designed our experiments in a way, that in the presence of multiple species (such as pTarget (Fig. 5)) we already predetermined the kinetic parameters for most species to limit the required fitting parameters to three.

Copy number determination

The copy number of the Cascade complex was determined by generating cell outlines from brightfield images (only well separated cells were chosen). The cell outlines were made with the Oufti software (Paintdakhi et al., 2016). The total number of tracks that were found in the outlined cells generated a copy number (Figure 1D). Because single localization events can partly stem from false positives, the total amount of tracks was estimated based on the distribution of tracks longer than 1 step and subsequently this distribution was fitted with an exponential to calculate the amount of particles that only had a single localization before bleaching. Similarly, as we know the false positive rate was approximately one per frame, we could also subtract the number of frames from the single step tracks and in this way estimate the total number of tracks. This approach yielded comparable results.

The copy number of proteins in cells are hard to quantify (Lee et al., 2012). Currently, protein copy numbers can be estimated either by western blot or by single-

molecule fluorescence based methods both of which have specific drawbacks. Although single molecule studies are regarded as the most accurate method, especially at low copy numbers (Huang et al., 2007), there are a lot of variables that can lead to over- or underestimation. Underestimation can originate from maturation time of the protein, misfolded/inactivated protein, false negative detections, overlap of PSFs and linking of two separate molecules in a single track. Overestimation can come from failed linking of tracks, false positive detections and blinking fluorescent proteins.

As has been done in previous studies, we take the underestimations stemming from maturation time (23 min for PAmCherry (Subach et al., 2009)), close to growth rate of 31 min) and estimated *in vivo* folding efficiency (50% (Durisic et al., 2014)) into account (Uphoff et al., 2013). We also consider that an estimated 40% of the particles we observed come from Cas8e subunits not active complexes. Taken together, the number of particles we observe are subtracted by the amount of estimated autofluorescent particles and subsequently multiplied by a compensation factor of two to reach our estimated copy number values.

We believe that the assumptions made in this study could maximally lead to over- or underestimating our estimated copy numbers by two to three-fold. We note that the relative amounts we observed between the different expression levels will be independent of these assumptions.

Cascade in DNA-containing/DNA-free regions

To get an independent measure of the total time fraction spent probing DNA, Cascade was visualized in cells that were elongated by addition of cephalexin. The drug cephalexin disabled the ability of the cells to divide, creating elongated cells where nucleoids were separated by DNA-free spaces (Reyes-Lamothe et al., 2014). Subregions of cell outlines were manually selected and further refined with the Oufti software (Paintdakhi et al., 2016). The relative amount of localizations of DNA-free and DNA-containing regions was not calculated for entire cells, as differences in illumination intensity between parts of the FOV could also change the amount of localizations detected for different parts of the cell. Each subregion contained one

nucleoid free region, flanked by two nucleoid containing regions with a total length of around 4 μm. Segments of 0.1 μm divided along the long axis of the cell are separated into nucleoid or DNA-free segments based on the sum of the DAPI fluorescence within each segment. The average number of localizations of Cascade molecules in nucleoid segments divided by the average number of localizations Cascade molecules in DNA-free segments could be used to infer the DNA bound time fraction (see below, f_{onDNA} from nucleoid enrichment).

Persistence sustained binding events for different integrated times

To estimate how long binding events last, one could plot the number of particles remaining within a certain radius from the first frame position for different number of steps. However, particles can diffuse away when they are released from DNA or be lost due to photobleaching. To account for bleaching rates, previous studies increased dark time between exposures, while keeping exposure times the same (Ho et al., 2018; Knight et al., 2015). This approach uses the data of all time steps, including only single time steps.

As we are investigating lifetime of binding events on a subsecond timescale this approach fails, as single steps of slow-moving particles, which can be clearly separated from bound particles on larger timescales (t_{int} > 1 s), will be counted as bound particles leading to overestimated off-rates. At these timescales, it is more reliable to use tracks of at least 5 steps to distinguish bound from moving particles. As we are interested in how many of these events we observe, depending on the framerate, normalization is required.

For this we cannot use the sum of all tracks observed at each frametime, as a larger amount of fast moving molecules diffuse further than the maximum tracking distance of 0.78 μm between two exposures, and are also more affected by confinement with increasing integrated time. Therefore, the number of moving particles of certain track length is not an accurate normalization when comparing different frame times. However, as we used similar exposure for all frame times, the number of detected localizations per protein is unaffected. Furthermore, bound

molecules are not affected by confinement or linking errors with increasing frame rates.

The most robust normalization procedure was therefore to normalize the number of localizations within sustained bound tracks (all localizations within 1 pixel of the mean location of the track) to the total number of localizations, as those do not depend on the length of introduced dark time between exposures. A further increase of the dark time was not possible as on longer time scales the movement of the plasmid (D^*_{free} = 0.06 μm^2/s) made plasmid bound particles diffuse further than 1 pixel.

Confinement and localization error simulation

To verify whether our new transitional *D** analysis yielded accurate parameter predictions and investigate the influence of localization error and confinement on the parameters of the fit, we simulated particles moving and transitioning between bound and free moving states within the dimensions of an *E. coli* cell, adapted from methodology used in (34). At every time step particles were simulated to be either in a bound state *S1* (D = 0 μm^2/s), or a mobile state *S2* ($D = D_{\text{free}}$). At the starting time point, states were assigned to each particle according to the equilibrium probability p_{S1} and p_{S2} (Eq. 10 + 11). Subsequently, at following time steps of 0.1 ms, particles in state *S1* were assigned to *S2* with a probability of $p_{S1\rightarrow S2} = k_{\text{off}} t_{\text{step}}$ (where $t_{\text{step}} = 0.0001\ s$) and particles in state *S2* were assigned to *S1* with a probability of $p_{S2\rightarrow S1} = k^*_{\text{on}} t_{\text{step}}$. Displacements in three dimensions at each time step were taken from a standard normal distribution multiplied with $\sqrt{2D t_{\text{step}}}$ (where D is either 0 for particles in state *S1* or D_{free} for particles in state *S2*). Steps beyond the boundaries of a cell were rejected and new displacements were randomly drawn.

The 2D projection of five localizations at 10 ms time intervals for each molecule was generated as output and was analysed in our tracking software. Localization error was included in the simulation by addition of a random displacement for each

position taken from a Gaussian distribution (σ = 40 nm). It was found that changes in outcome of the simulation were not sensitive to cell length in the range of our bacteria (3-6 μm), decreasing less than 5% for the smallest size. Most of the confinement effect is caused by the cell width, which was relatively constant between all the cells measured.

Cascade nucleoid enrichment simulation

The simulation above was adapted to simulate the movement in DNA-free and DNA-containing regions. Particles were simulated to move inside of a cell of 10 μm in length and 1 μm in width consisting of 100 segments without endcaps (0.1 μm per segment). Five segments were modelled as DNA-free segments and the rest of the segments as DNA-containing segments.

Cascade molecules were randomly placed throughout the cell and subsequently were simulating with similar time steps as described above, except that moving particles were only allowed to transition to *S*1 (bound state) inside of the nucleoid containing regions. Before recording the position of the simulated particles, the simulation ran for 100.000 time steps (10 s) so that equilibrium was reached. Localization error was added in the same way as described above.

Expected free diffusion coefficients

The diffusion coefficient of molecules in classic (Newtonian) fluids can generally be estimated by the Stokes-Einstein equation. A study measuring the diffusion of GFP multimers inside the *E. coli* cytoplasm has shown good agreement with the predictions of this equation (Nenninger et al., 2010), whereas a second study found a different relation attributed to the complex nature of the cytoplasmic fluid (Mika and Poolman, 2011). To compare our findings of the apparent free diffusion coefficient of Cas8e (~3.5 $\mu m^2/s$) and Cascade (~ 1.0 $\mu m^2/s$), we therefore looked for reported free cytoplasmic diffusion coefficient values of proteins of similar size inside *E. coli* cells. For Cas8e, two proteins have been studied with a similar size to

PamCherry-Cas8e (82 kDa), namely CFP-CheR-YFP (86 kDa) (Kumar et al., 2010) and TorA-GFP3 (84 kDa) (Nenninger et al., 2010), which have reported values of 1.7 $\mu m^2/s$ and 6 $\mu m^2/s$. Our estimate for Cas8e lies within the range of these values. For Cascade (430 kDa), the closest reported protein in size is RNA polymerase, for which the D^*_{free} was found to be 1.1 $\mu m^2/s$ (400 kDa core enzyme, 470 kDa holoenzyme) (Stracy et al., 2015). Furthermore larger proteins such β-Gal-GFP_4 (582 kDa; 0.6 $\mu m^2/s$) (Mika et al., 2010), and 30S ribosome subunits (900 kDa 0.4 $\mu m^2/s$) (Sanamrad et al., 2014) were reported with lower diffusion coefficients as expected. These findings support the free apparent diffusion value we found for Cascade (~ 1.0 $\mu m^2/s$).

f_{onDNA} from nucleoid enrichment

The distribution of Cascade in nucleoid-free and nucleoid containing regions depends on the time Cascade spends on DNA. We divided the cell up along the long axis into segments of 100 nm wide. During the time Cascade is bound to DNA it can only be inside of the nucleoid regions whereas, when it is not bound to DNA Cascade can be anywhere within the cell. Therefore, the average number of particles in a DNA-containing segment is given by:

$$\overline{N_{\text{DNA}}} = \left(\frac{f_{\text{onDNA}}}{sm_{\text{DNA}}} + \frac{1 - f_{\text{onDNA}}}{sm_{\text{tot}}}\right) N_{\text{tot}} \tag{22}$$

and the average number of particles in a DNA-free segment is given by

$$\overline{N_{\text{DNA-free}}} = \frac{1 - f_{\text{onDNA}}}{sm_{\text{tot}}} N_{\text{tot}} \tag{23}$$

Where f_{onDNA} is the fraction of time bound to DNA, sm_{DNA} and sm_{tot} are the number of DNA segments and the total number of segments respectively and N_{tot} is the total number of particles in a cell. The ratio, which is equal to the enrichment factor *EF*, can then be expressed as:

$$EF = \frac{\overline{N_{\text{DNA}}}}{\overline{N_{\text{DNA-free}}}} = \left(\frac{f_{\text{onDNA}}}{sm_{\text{DNA}}} + \frac{(1-f_{\text{onDNA}})}{sm_{\text{tot}}}\right) / \frac{1-f_{\text{onDNA}}}{sm_{\text{tot}}} \tag{24}$$

If the number of DNA-free segments is much less than the number of DNA segments $sm_{\mathrm{DNA}} \approx sm_{\mathrm{tot}}$ the expression above can be simplified to:

$$EF = \frac{1}{1-f_{onDNA}} \tag{25}$$

This equation allows extraction of f_{onDNA} from EF directly and implies that this value does not depend on the diffusion coefficients of the mobile population.

In vivo K_D values

The K_D value is a commonly calculated affinity constant used for binding kinetics of proteins and assembly of multicomponent systems (McGuigan et al., 2006), but the K_D has also been used as an estimate for in vivo binding affinity (Zawadzki et al., 2015). In the reaction scheme A + B ⇄ AB, the K_D is calculated as

$$K_D = [A][B]/[AB] \tag{26}$$

For Cascade the reaction scheme is as follows: [Cascade (probing)] + [free target sites] ⇄ [Cascade (bound)]. The concentration of a single entity inside of a cell of length 4 μm and width 1 μm with hemispherical endcaps is approximately 0.5 nM. The copy number for pTarget was estimated by qPCR to be approximately 100 plasmids per chromosome. As the number of chromosomes in actively dividing cells is generally higher than one, we used literature values for the number of chromosomes/cell found in (Wallden et al., 2016), providing 4/cell which also used a glucose and amino acid enriched M9 medium as growth medium. This brings the copy number of pTarget to 400/cell, which is equal to 200 nM. For a Cascade complex carrying one of several crRNAs in the cell, the amount of free target sites is equal to the copy number of the plasmid pTarget minus the amount of already occupied target sites of that crRNA, but as the copy number of each target (400) is much higher than the number of Cascade complexes potentially carrying that crRNA (on average 130/18 ≈ 7), [free targets] ≈ [pTarget].The K_D value was then calculated as:

$$\begin{aligned} K_D &= [\text{pTarget}][\text{Cascade(probing)}]/[\text{Cascade(bound)}] \\ &= 200\ \text{nM}\ [\text{Cascade(probing)}]/[\text{Cascade(bound)}] \end{aligned} \tag{27}$$

Theoretical model interference level vs copy number

In the case where the interference level is limited by the target search of the proteins, we can model the relation based on the distribution of search times of single proteins. The search time for a single protein, because it is the arrival time of a recurring independent random event, is exponentially distributed and characterized by the average search time, $< t_s >$:

$$p_1(t_s) = 1/\overline{< t_s >}\, e^{-t_s/\overline{<t_s>}} \tag{28}$$

We have verified given our kinetic model of Cascade with simulations that this is the case (Figure S3). The chance that one of *n* proteins finds the target at search time t_s while the other proteins have not yet found the target is:

$$p_n(t_s) = np_1(t_s)(\int_{t_s}^{\infty} p_1(t)dt)^{n-1} = n/\overline{< t_s >}\, e^{-nt_s/\overline{<t_s>}} \tag{29}$$

We have verified this derivation with simulations (Figure S3). The establishment probability of the plasmid is equal to the likelihood for all search times larger than t_{critical} (t_c), the time point at which the cell can no longer clear the invader. Therefore:

$$p_{\text{establishment}}(t_c) = \int_{t_c}^{\infty} p_n(t)dt = \; e^{-nt_c/\overline{<t_s>}} \tag{30}$$

As the chance of targeting after replication is low, we assume in our model that Cascade is only able to clear the foreign DNA before replication. Therefore t_c is equal to the replication time of the plasmid t_R.

As we found that 20 copies of Cascade reduce interference level by half, this leads to

$$\ln(0.5) = -20t_R/\overline{< t_s >} \tag{31}$$

or

$$t_R/\overline{< t_s >} = 0.035 \tag{32}$$

Right after transformation, the negative regulators of copy numbers are absent, so replication in that instant is faster than the growth rate of the cell. Replication time of pTarget has not been measured so far, but by using a temperature-dependent ori,

Olsson *et al.* measured a replication time of 3 min for a slightly larger plasmid in the absence of copy number control (Olsson et al., 2003b). If we assume pTarget replication occurs on a similar time scale, we get an estimated search time for one Cascade to find a single target of ~90 minutes.

We can further describe the relationship between the average search time $<t_s>$ and the k_{off} and k^*_{on} that were measured for Cascade. This relationship is found by multiplying the amount of time spent for each binding event times the average amount of binding events required to find the target. The amount of time spent for each binding event is equal to the sum of the time spent on binding ($1/k_{\text{off}}$) and the time spent on diffusing to the next site ($1/k^*_{\text{on}}$). Therefore the average search time is:

$$\overline{< t_s >} = \left(\frac{1}{k_{\text{off}}} + \frac{1}{k^*_{\text{on}}}\right) \frac{\#DNA\ binding\ events}{\#DNA\ target\ sites} \qquad (33)$$

We have again verified this description by using our simulations of our kinetic model of Cascade target search (Figure S3).

It must be emphasized that the number of binding events is different from the number of binding sites in the fact that if a single binding event scans multiple sites (during 1D sliding), the number of binding sites probed per event are more than one. Using Eq. 30 and 33, the chance of establishment of a single invader in the cell with multiple Cascade copies is therefore as follows:

$$p_{\text{establishment}} \sim e^{-n/\#DNA\ binding\ events \left(\frac{1}{k_{\text{off}}} + \frac{1}{k^*_{\text{on}}}\right)} \qquad (34)$$

Simulation Cascade search times

To see whether the above described theoretical model was compatible with our kinetic model of Cascade search, we simulated the search times of Cascade. To do so, we simulated Cascade probing DNA sites as was described above (See

Simulation Localization and Confinement). Subsequently every time Cascade changed from a mobile state to a bound state we added with a certain probability that the newly probed site is the target (1/90.000). When Cascade located the target the simulation for that particle was stopped and the search time was recorded for each individual Cascade complex. To simulate the search time for 5 Cascades, we grouped the single search times in multiples of five and took the fastest search time of 5 Cascades.

Data and Code availability

The data and code that support the findings of this study are available upon request to the Lead Contact, Stan Brouns (stanbrouns@gmail.com).

Supplementary figures

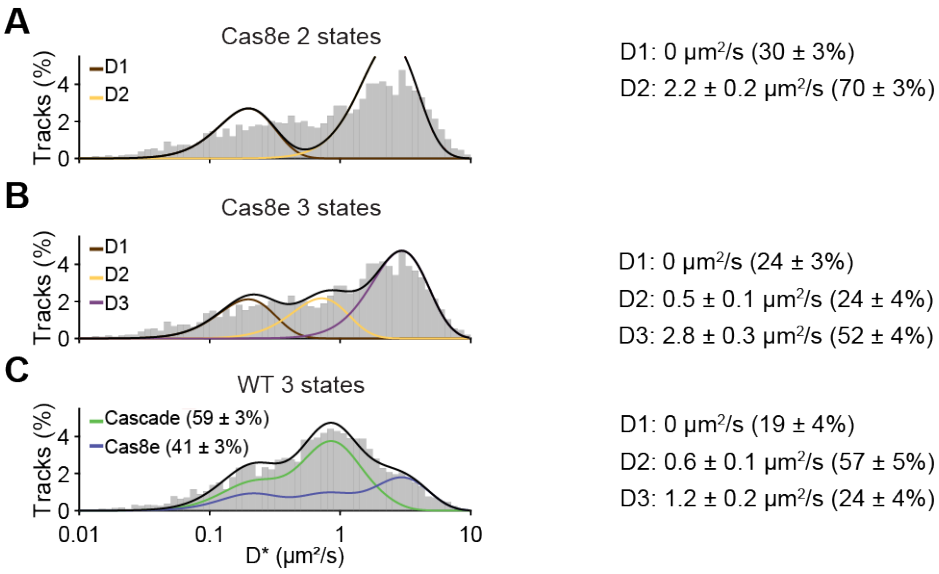


Figure S1. Static *D fitting, Related to Figure 2. (A)** *D** distribution (left) of the Cas8e strain (Figure 2B) fitted with two static states with extracted *D** value of each fraction on the right (relative abundance). The slowest state (D1; brown) was fixed to 0 $\mu m^2/s$. **(B)** Same as (A) but then for three static states. **(C)** *D** distribution (left) of the WT strain (Figure 2C). Cas8e distribution from Figure S1B was taken and used to fit the distribution with additional three states for Cascade diffusion. The relative abundance of Cas8e and Cascade estimated from static *D** fitting is similar to that found for dynamic fitting (60 and 40%), even though the distributions of Cascade and Cas8e are different. Error estimation in (A-C) is based on bootstrapping (± standard deviation).

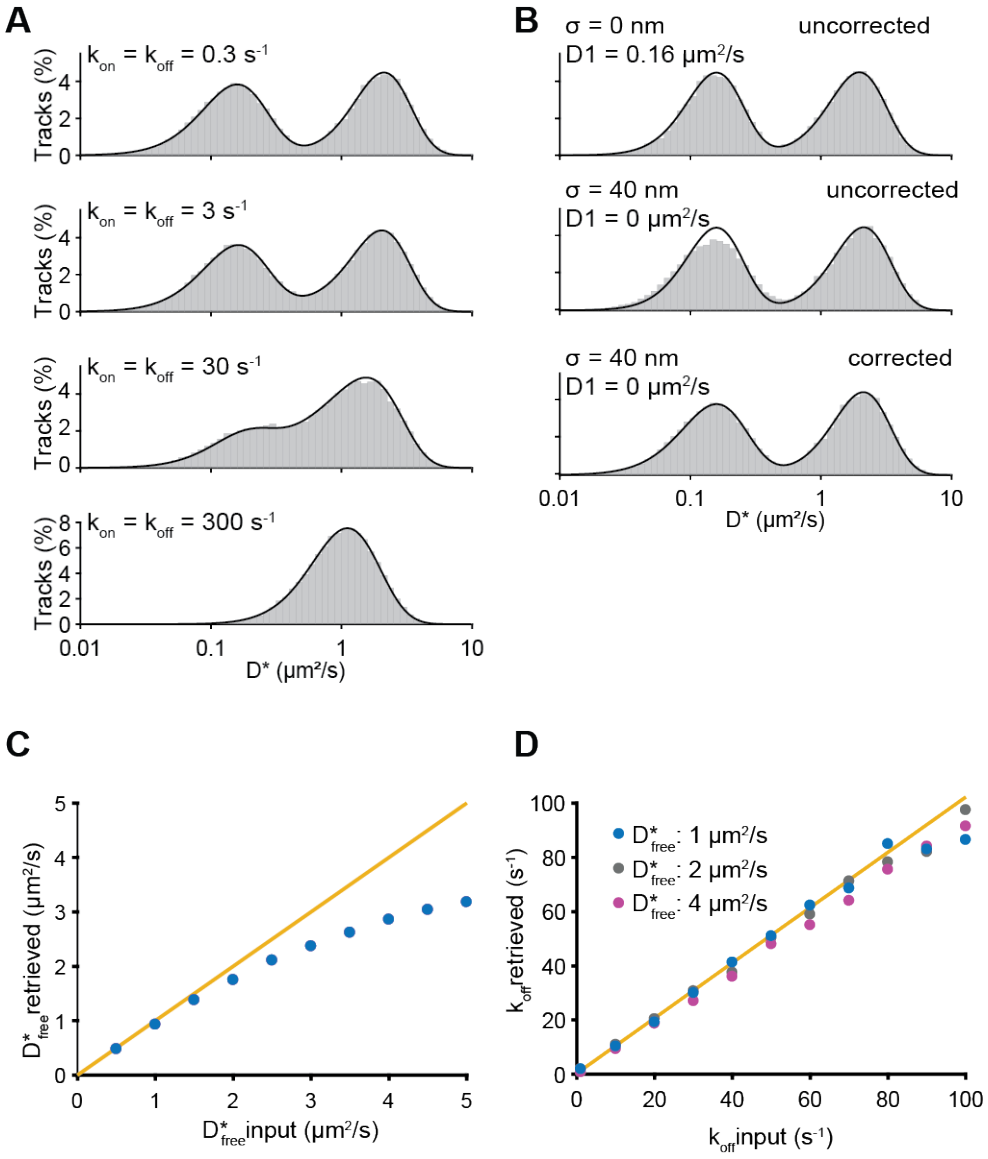


Figure S2. Performance of analytical DDA, Related to Figure 2. (A) Comparison of simulation to the theoretical distribution (black line) found with the newly developed analysis method. 50.000 particles were simulated to move without boundaries and position was recorded for 4 consecutive steps. Particles were simulated with $\boldsymbol{D^*_{free}}$= 2 μm²/s and increasing on- and off-rates (from 0.3 to 300 s^{-1}). The theoretical model (black line) is directly plotted on top of the histogram of simulated *D** values. A localization error drawn from a Gaussian distribution with σ = 40 nm was added to both the model and the simulation. (B) Influence of localization error. Distribution of an average of consecutive displacements that are offset by a localization error are correlated, which is why in the absence of localization error in the simulation (top) there is no requirement for correction. However immobile particles offset by localization error with the same mean apparent diffusion coefficient are slightly differently distributed (middle). Correction (described in Methods) for the immobile particles is sufficient to restore the fit (bottom). (C)

Influence of confinement. Particles were simulated inside of a cell 4 μm long and of 1 μm diameter. Simulations were run through analysis software to retrieve parameters. $\boldsymbol{D}^{*}_{\mathbf{free}}$ estimates are influenced by confinement where fast moving particles appear to be slower. (D) The off-rate is not as influenced by effects of confinement and stays the same even for the fastest moving particles (purple). Estimates become more unreliable for much faster or slower transitions than are measured in the integrated time of typical tracks.

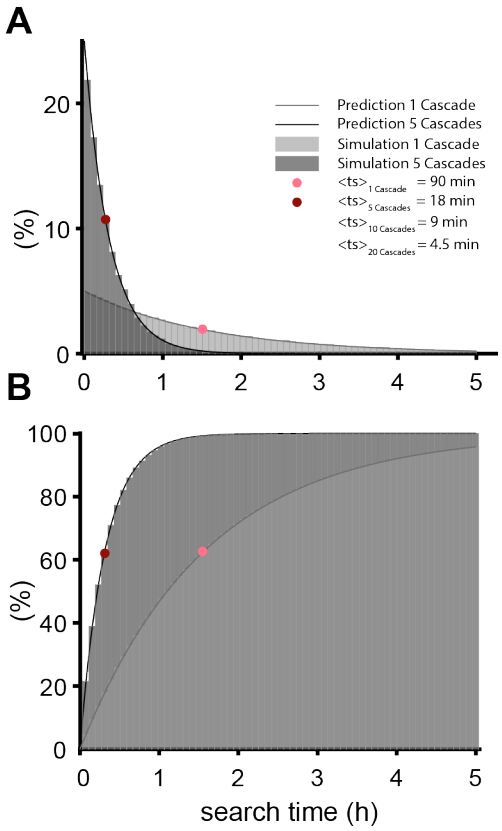


Figure S3. Comparison between theoretical prediction and independent simulations of Cascade search times, Related to Figure 2. (A) Probability density function (PDF) and (B) cumulative distribution function (CDF) of the distribution of search times. The equations of search time distributions for single and multiple (5) copies (Eq. 28 and Eq. 29) were tested against a simulation of Cascade search times using parameters of the kinetic model that were found experimentally (k_{off} = 38 s^{-1} ; k^{*}_{on}= 26 s^{-1}) and an estimated 90.000 DNA sites to be scanned before reaching the target (See Methods, theoretical model interference level vs copy number). For the theoretical prediction, the average search time was calculated by using Equation 33. Both the theoretical prediction and simulations indicate that the average search time (indicated by a dot), decreases sharply from 90 minutes for a single Cascade copy to 4.5 minutes for 20 Cascade copies.

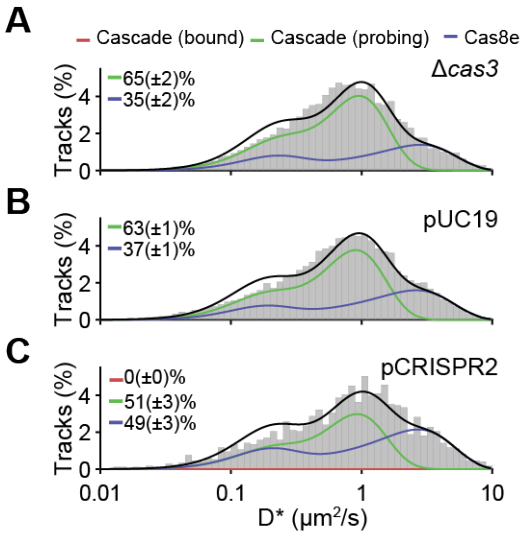
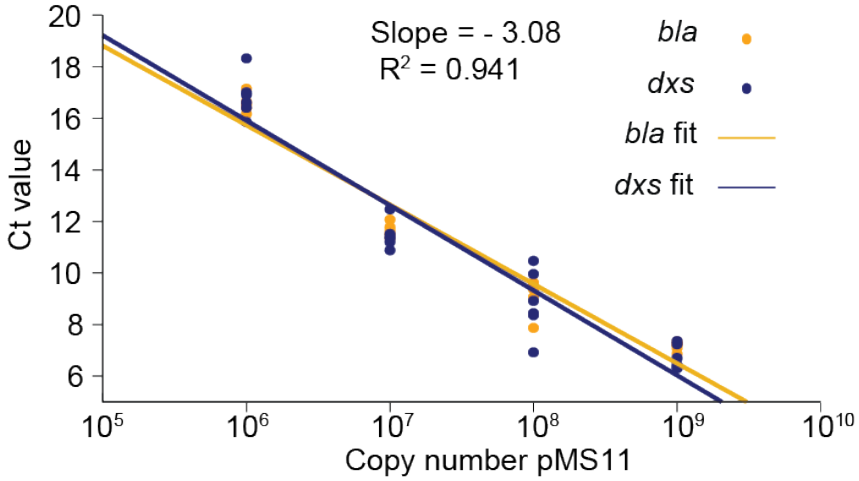


Figure S4. *D Histograms other conditions, Related to Figure 5.** *D** distributions for (A) Δ*cas3* strain, (B) Δ*cas3* strain + pUC19, the empty variant of pTarget-RNAP and pCRISPR1-RNAP and (C) Δ*cas3* strain + pCRISPR2. The amount of available Cascade complexes in the interference assay for strain pCRISPR2 targeting (Figure 5H) were extracted from the relative amount of Cascade complexes in this strain (51%) divided by the number of complexes in the WT strain (60%).

A



B

	Ct (CN) *bla*	Ct (CN) *dxs*	PCN
Δ*cas3* +pTarget	8.82 ± 0.09 ($10^{-5.75}$)	15.15 ± 0.11 ($10^{-7.76}$)	102 ± 10
Δ*cas3* +pCRISPR1	8.85 ± 0.07 ($10^{-5.76}$)	15.20 ± 0.22 ($10^{-7.78}$)	104 ± 15

Figure S5. Plasmid copy number determination, Related to Figure 5. (A) Calibration curve of *dxs* and *bla* primer amplification with dilution series of pMS11 (plasmid containing both *dxs* and *bla* gene). The regression of six technical replicates was used to make the calibration curve for both primer sets (regression parameters of *bla* and *dxs* gene in orange and purple respectively). (B) The Ct values of *bla* and *dxs* gene amplifications were calculated from biological triplicates. These Ct values were converted to absolute copy numbers (CN) by using the regression values from the calibration curve. The plasmid copy number per chromosome (PCN/chromosome) was calculated by dividing the copy number of the *bla* gene by the copy number of the *dxs* gene. The plasmid copy number per cell was estimated by multiplying PCN/chromosome by the expected number of chromosomes per cell (4) based on a literature value (Wallden et al., 2016).

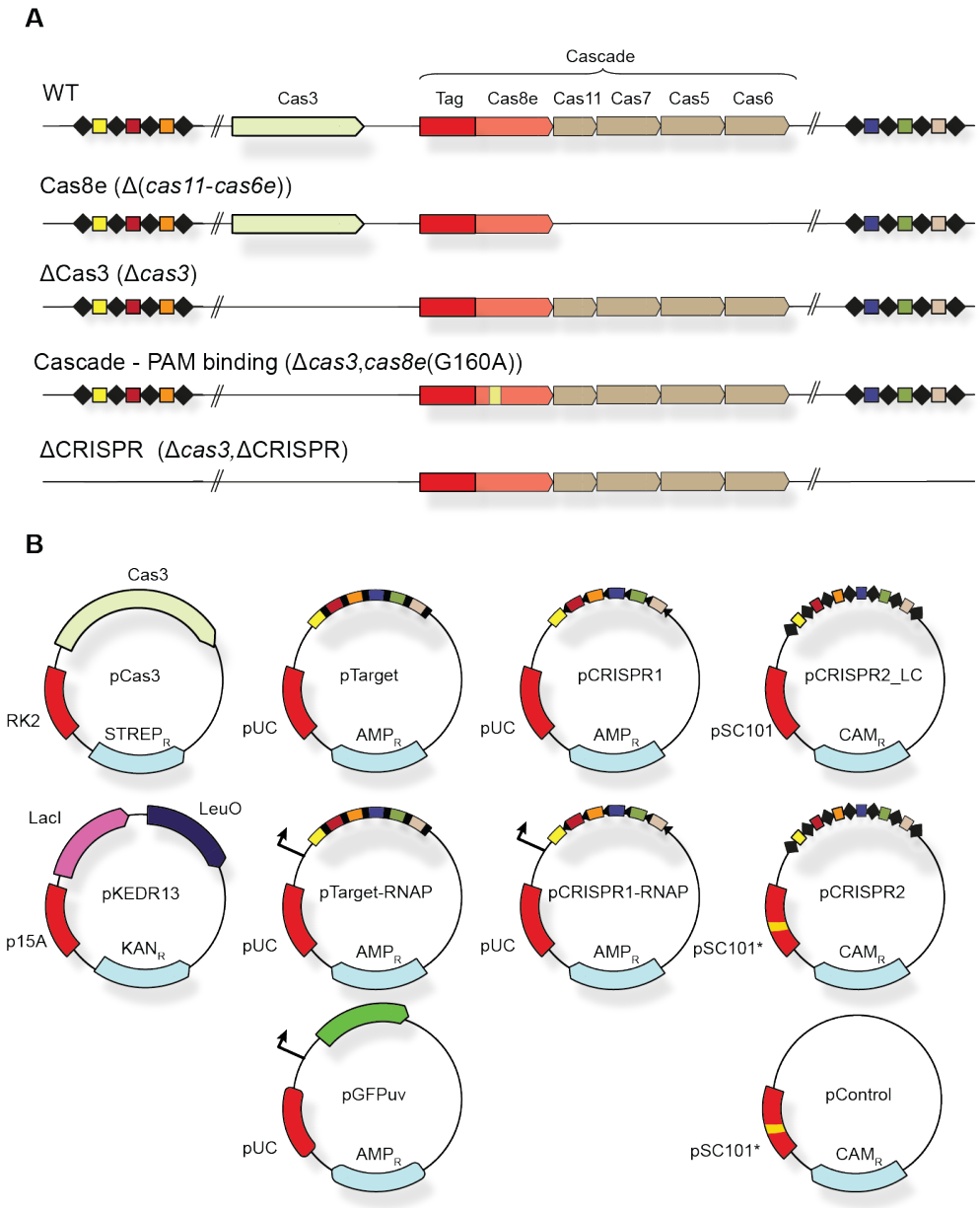


Figure S6. Strains and plasmids used, Related to Figure 5. (**A**) Strains used in this study, strains were constructed with lambda recombination and verified by sequencing. Only part of each CRISPR array indicated (total 18 spacers). (**B**) Plasmids used in this study. Indicated are the ori (red), antibiotic resistance marker (light blue) and other components on the plasmid. Only part of the total 18 spacers are indicated for pTarget, pCRISPR1 and pCRISPR2. For sequences and descriptions see Table S3 and S4.

Supplementary tables

	Doubling time
K12 BW25113	24.9 ± 0.1 min
WT + pKEDR13	24.5 ± 0.4 min
WT + pKEDR13 + 1 mM IPTG	31.7 ± 0.6 min
WT + pCas3 + pKEDR13 + 1 mM IPTG	33.3 ± 0.2 min
ΔCas3 + pKEDR13 + pTarget + 1 mM IPTG	31.8 ± 0.4 min

Table S1. Growth rate of *E. coli* strains used in this study, Related to STAR Methods. Growth rates were determined in a plate reader where cells were inoculated in similar conditions as described in Methods. The instantaneous growth rate was determined at t = 2.5 hours, which represented the growth rate at the time of the microscope studies. Three independent cultures were measured to get the mean and standard error values.

Name	Description	Sequence (5'-3')
BG7128	PAmCherry (lox-*cam*-lox) insert fw (WT)	GGAGGCTATTAAAGGTGCACAAT
BG7129	PAmCherry (lox-*cam*-lox) insert fw (Δ*cas3*)	GTCTCTTCTTTGCAGGGAGG
BG7130	PAmCherry (lox-*cam*-lox) insert rv (WT)	TATCGTCACGGGGCAAACT
BG7131	PAmCherry (lox-*cam*-lox) insert rv (G160A)	AGCAGGTATAGACTCATTGGACT
BG7366	ΔCRISPR1 insert (lox-*kan*-lox) fw	GCAGAGGCGGGGGAACTCCAAGTGATATCCATCATCGCATCCAGTGCGCCGGTGTCTTTTTTACCTGTTTGACC
BG7367	ΔCRISPR1 insert (lox-*kan*-lox) rv	GGTTGTTTTTATGGGAAAAAATGCTTTAAGAACAAATGTATACTTTTAGATTCCTACCTCTGGTGAAGGAGTTG
BG7368	ΔCRISPR2+3 insert (lox-*kan*-lox) fw	TAAGTGAGAAGGCCGGGCGGGAAACTGCCCGGCCTGAACATACCTGAATTAGAGTCGGACTTCGCGTTCGC
BG7369	ΔCRISPR2+3 insert (lox-*cam*-lox) rv	GATTGTGACTGGCTTAAAAAATCATTAATTAATAATAGGTTATGTTTAGAGCTAGTTATTGCTCAGCGGTGG
BG8366	Δ(*cas11-cas6e*) insert (lox-*kan*-lox) fw	TTGAGTGGAATGGGATTAAGGGGAAGCCAGGTCATTTTATTACACCTCAAGGTGTCTTTTTTACCTGTTTGAC
BG8367	Δ(*cas11-cas6e*) insert (lox-*kan*-lox) rv	ACAAACATTTACGGGAGTTAAAACCGCAAGGAGGGCCATCAAATGGCTGATTCCTACCTCTGGTGAAGGAGTTG
BG8677	qPCR *bla* fw	CTACGATACGGGAGGGCTTA

BG8678	qPCR *bla* rv	ATAAATCTGGAGCCGGTGAG
BG8679	qPCR *dxs* fw	CGAGAAACTGGCGATCCTTA
BG8680	qPCR *dxs* rv	CTTCATCAAGCGGTTTCACA
BN370	pCRISPR2 (array2.3) rv	GTGAGCTGATACCGCTCGCCTGAACCTCTCTGGCATGGA
BN383	pCRISPR2 (array2.1) fw	TGCTTTAAGAACAAATGTATACTTTTAG
BN384	pCRISPR2 (array2.1) rv	TCTAAACATAACCTATTATTACCAAGTGATATCCATCATCGC
BN385	pCRISPR2 (array2.3) fw	GCGATGATGGATATCACTTGGTAATAATAGGTTATGTTTAGA
BN373	Site-directed mutagenesis RepA HC fw	TGGTTAAAGGCTTTCGGATCTTCCAG
BN374	Site-directed mutagenesis RepA LC fw	TGGTTAAAGGCTTTGAGATCTTCCAG
BN375	Site-directed mutagenesis RepA HC+LC rv	AAGGATTCCTGATTTCCACAGTTC

Table S2. Primers used in this study, Related to STAR Methods.

Description	Sequence (5'-3')
PAmCherry ins	TGGCGTTAAGCATTCGCGAGGTTCCAGATGGACAAAAGCCCC AGGCGATATTTCTATCAACCTGAGGCCAGCGTTCGAACCCAAA CAATTCGAATGTTAGTCTCTTCTTTGCAGGGAGGCAAGACATG TGTATATCACTGTAATTCGATATTTATGAGCAGCATCGAAAAA TAGCCCGCTGATATCATCGATAATACTAAAAAAACAGGGAGG CTATTAAAGGTGCACAATGTACATCTTCTTTTAATTTCCCGGTA TGAGATTTTATATTCACAGTATGAATATTTTATGTAATAAAATT CATGGTAATTATTATAACTAAAAGTTTCTTTAATAATAAGGCG CCCCTAGGTACCGTTCGTATAATGTATGCTATACGAAGTTATG AGCTGTTGACAATTAATCATCGGCTCGTATAATGTGTGGGCAA TGAGCTTGCACTGCAGAACTTTGATATACCATGGAGAAAAAA ATCACTGGATATACCACCGTTGATATATCCCAATGGCATCGTA AAGAACATTTTGAGGCATTTCAGTCAGTTGCTCAATGTACCTA TAACCAGACCGTTCAGCTGGATATTACGGCCTTTTTAAAGACC GTAAAGAAAAATAAGCACAAGTTTTATCCGGCCTTTATTCACA TTCTTGCCCGCCTGATGAATGCTCATCCGGAATTCCGTATGGC AATGAAAGACGGTGAGCTGGTGATATGGGATAGTGTTCACCCT TGTTACACCGTTTTCCATGAGCAAACTGAAACGTTTTCATCGCT CTGGAGTGAATACCACGACGATTTCCGGCAGTTTCTACACATA TATTCGCAAGATGTGGCGTGTTACGGTGAAAACCTGGCCTATT TCCCTAAAGGGTTTATTGAGAATATGTTTTTCGTCTCAGCCAAT CCCTGGGTGAGTTTCACCAGTTTTGATTTAAACGTGGCCAATA TGGACAACTTCTTCGCCCCCGTTTTCACTATGGGCAAATATTAT ACGCAAGGCGACAAGGTGCTGATGCCGCTGGCGATTCAGGTT CATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTTA ATGAATTACAACAGTACTGCGATGAGTGGCAGGGCGGGGCGT AAATAACTTCGTATAATGTATGCTATACGAACGGTATCTAGAC TTCGGGAATGATTGTTATCAATGACGATAATAAGACCAATAAC GGTTTATCCCTACTTAAGTAGGGAAGGTGCACAATGTACATCT TCTTTTAATTTCCCGGTATGAGATTTTATATTCACAGTATGAAT ATTTTATGTAATAAAATTCATGGTAATTATTATAACTAAAAGT TTCTTTAATAATAAAACGAATAACTTGCAGATTTGAAATGCAT GCATTATTGTCTTTAAACAATTCAACACATCTTAATATATGTAT AGGTTAATTGTATTAAACCAATGAATATATTTTTGCAGTGAAT GTGATTATTGAATTAATTACGCCGTATTTTTTCTTTGTTTTTACC

	GATAACGGAAGTGTGCCGACGTATAGAAATGCAGGAGAAATG TCGGAGCATATGAAGGAGAACAAATGGTGAGCAAGGGCGAGG AGGATAACATGGCCATCATCAAGGAGTTCATGCGCTTCAAGGT GCACCTGGAGGGGTCCGTGAACGGCCACGAGTTCGAGATCGA GGGCGAGGGCGAGGGCCGCCCCTACGAGGGCACCCAGACCGC CAAGCTGAAGGTGACCAAGGGTGGCCCCTTGCCCTTCGCCTGG GACATCCTGTCCCCTCAGTTCATGTACGGCTCCAATGCCTACG TGAAGCACCCCGCCGACATCCCCGACTACTTTAAGCTGTCCTT CCCCGAGGGCTTCAAGTGGGAGCGCGTGATGAACTTCGAAGA CGGCGGCGTGGTGACCGTGACCCAGGACTCCTCCCTGCAGGAC GGCGAGTTCATCTACAAGGTGAAGCTGCGCGGCACCAACTTCC CCTCCGACGGCCCCGTAATGCAGAAGAAGACCATGGGCTGGG AGACCCTCTCCGAGCGGATGTACCCCGAGGACGGCGCCCTGA AGGGAGAGCTCAAGGCGAGGACGAAGCTGAAGGACGGCGGC CACTATGACACTGAGGTCAAGACCACCTACAAGGCCAAGAAG CCCGTGCAGTTGCCCGGCGCCTACAACGTCAACCGCAAGTTGG ATATCACCTCCCACAACGAGGACTACACCATCGTGGAACAGTA CGAACGTGCCGAGGGCCTCCACTCCACCGGCGGCATGGACGA GCTGTACAAGCCCGGGGCGCTCATGGCTAATTTGCTTATTGAT AACTGGATCCCTGTACGCCCGCGAAACGGGGGGAAAGTCCAA ATCATAAATCTGCAATCGCTATACTGCAGTAGAGATCAGTGGC GATTAAGTTTGCCCCGTGACGATATGGAACTGGCCGCTTTAGC ACTGCTGGTTTGCATTGGGCAAATTATCGCCCCGGCAAAAGAT GACGTTGAATTTCGACATCGCATAATGAATCCGCTCACTGAAG ATGAGTTTCAACAACTCATCGCGCCGTGGATAGATATGTTCTA CCTTAATCACGCAGAACATCCCTTTATGCAGACCAAAGGTGTC AAAGCAAATGATGTGACTCCAATGGAAAAACTGTTGGCTGGG GTAAGCGGCGCGACGAATTGTGCATTTGTCAATCAACCGGGGC AGGGTGAAGCATTATGTGGTGGATGCACTGCGATTGCGTTATT CAACCAGGCGAATCAGGCACCAGGTTTTGGTGCCGGTTTTAAA AGCGGTTTACGTGGAGGAACACCTGTAACAACGTTCGTACGTG GGATCGATCTTCGTTCAACGGTGTTACTCAATGTCCTCACATTA CCTCGTCTTCAAAAACAATTTCCTAATGAATCACATACGGAAA ACCAACCTACCTGGATTAAACCTATCAAGTCCAATGAGTCTAT ACCTGCTTCGTCAATTGGGTTTGTCCGTGGTCTATTCTGGCAAC CAGCGCATATTGAATTATGCGATCCCATTGGGATTGGTAAATG TTCTTGCTGTGGACAGGAAAGCAATTTGCGTTATACCGG

pTarget insert	TCTAGAGAATTCGACAGAACGGCCTCAGTAGTCTCGTCAGGCT CCTTCTGTTTTCGCAAATCTATGGACTATTGCTATTCTTGGGCG CACGGAATACAAAGCCGTGTATCTGCTCTTTGGCTCTGCAACA GCAGCACCCATGACCACGTCTTGAAATGCTGGTGAGCGTTAAT GCCGCAAACACCTTATTACGCCTTTTTGCGATTGCCCGGTTTTT GCCTTCCATGGCAGCGTCAGGCGTGAAATCTCACCGTCGTTGC CTTTCGGTTCAGGCGTTGCAAACCTGGCTACCGGGCTTGTAGT CCATCATTCCACCTATGTCTGAACTCCCTTCCGGGGGATAATG TTTACGGTCATGCGCCCCCCTTTGGGCGGCTTGCCTTGCAGCC AGCTCCAGCAGCTTAAGCTGGCTGGCAATCTCTTTCGGGGTGA GTCCTTTAGTTTCCGTATCTCCGGATTTATAAAGCTGACTTGCA GGCGGCGACGCGCAGGGTATGCGCGATTCGCTTGCGACCGCTC AGAAATTCCAGACCCGATCCAAACTTTCAACATTATCAATTAC AACCGACAGGGAGCCCTTAGCGTGTTCGGCATCACCTTTGGCT TCGGCTGCTTTGCGTGAGCGTATCGCCGCGCGTCTGCGAAAGC TTGGTACC
pCRISPR1 insert	TCTAGAGAATTCGACAGAACGGCCTCAGTAGTCTCGTCAGGCT CCGGCTGTTTTCGCAAATCTATGGACTATTGCTATTCGGGGGC GCACGGAATACAAAGCCGTGTATCTGCTCGGTGGCTCTGCAAC AGCAGCACCCATGACCACGTCGGGAAATGCTGGTGAGCGTTA ATGCCGCAAACACCGGATTACGCCTTTTTGCGATTGCCCGGTT TTTGCCGGCCATGGCAGCGTCAGGCGTGAAATCTCACCGTCGT TGCCGGTCGGTTCAGGCGTTGCAAACCTGGCTACCGGGCGGGT AGTCCATCATTCCACCTATGTCTGAACTCCCGGCCGGGGGATA ATGTTTACGGTCATGCGCCCCCCGGTGGGCGGCTTGCCTTGCA GCCAGCTCCAGCAGCGGAAGCTGGCTGGCAATCTCTTTCGGGG TGAGTCCGGTAGTTTCCGTATCTCCGGATTTATAAAGCTGACG GGCAGGCGGCGACGCGCAGGGTATGCGCGATTCGCGGGCGAC CGCTCAGAAATTCCAGACCCGATCCAAACGGTCAACATTATCA ATTACAACCGACAGGGAGCCCGGAGCGTGTTCGGCATCACCTT TGGCTTCGGCTGCGGTGCGTGAGCGTATCGCCGCGCGTCTGCG AAAGCGGGGTACC

Table S3. Synthetic DNA inserts used in this study, Related to STAR Methods

Name in study	Name in storage	Description	Source
pKEDR13	pKEDR13	Expression plasmid LeuO	(Westra et al., 2010)
pGFPuv	pGFPuv	Expression plasmid GFPuv	Clontech
pMS011	pMS011	Plasmid containing *bla* and *dxs* gene (qPCR)	(Caforio et al., 2018)
pSC020	pSC020	Plasmid containing Cre and lambda recombinase	S. Creutzberg (unpublished)
pTarget	pTU256	Target plasmid containing all 18 potential protospacers for flanked by 5'-CTT-3'	This study
pTarget-RNAP	pTU150	Target plasmid containing all 18 potential protospacers for flanked by 5'-CTT-3' and p*lac* upstream	This study
pCRISPR1	pTU258	Target plasmid containing all 18 potential protospacers for flanked by 5'-CGG-3'	This study
pCRISPR1-RNAP	pTU152	Target plasmid containing all 18 potential protospacers for flanked by 5'-CGG-3' and p*lac* upstream	This study
pCRISPR2_LC	pTU158	Plasmid containing all 18 potential protospacers for flanked by repeat	This study

		sequences; low copy backbone variant of pSC101	
pCRISPR2	pTU160	Plasmid containing all 18 potential protospacers for flanked by repeat sequences; high copy backbone variant of pSC101	This study
pControl	pTU254	High copy backbone variant of pSC101	This study
pCas3	pTU255	Expression plasmid Cas3	This study

Table S4. Plasmids used in this study, Related to STAR Methods

Glossary

Full name	Symbol	Description
Apparent diffusion coefficient	D^*	Apparent due to confinement
Bound state	$S1$	
Dissociation constant	K_D	Constant which is a measure for the binding affinity of two objects with each other
DNA segments	sm_{DNA}	Amount of segments defined as containing DNA by DAPI staining
Enrichment Factor	EF	The number of localizations in DNA-containing segments divided by the number of localizations in DNA-free segments
Fraction DNA bound	f_{onDNA}	Fraction of the time DNA binding proteins spend on DNA is calculated from the off- and on-rate (Figure 1).
Frametime	t_f	Positions of simulated/measured particles are recorded for each frametime
Free diffusion coefficient	D^*_{free}	Diffusion coefficient in the absence of DNA binding. Apparent due to confinement.
Integrated time	t_{int}	Overall timescale: can be one or multiple frametimes
Localization error	σ	Average error in determination of particle position

Mobile state	*S2*	
off-rate	k_{off}	Rate DNA bound protein is released from DNA. Inverse of residence time
pseudo-first order on-rate	k_{on}^{*}	Rate mobile protein is binding to DNA. As the amount of potential DNA probing sites is very large, on-rate is independent of DNA concentration (pseudo-first order)
pTarget establishment	$p_{establishment}$	Measure for interference level calculated from the transformation ratio of pTarget and pGFPuv (Eq. 1)
Time step	t_{step}	Displacements of simulated particles are calculated for each time step
Total number of segments	sm_{tot}	Total number of segments in the cell

Movie S1 Cascade diffusion through the cell. Real-time data from WT strain where on the left cumulative overlay of tracks (each track differently coloured) are plotted on top of a brightfield image showing the outline of the cells and on the right the fluorescent signal of the single molecules are depicted. Scale bar and time are indicated at the bottom (total duration 50 s). The movie shows only a small part of a normal FOV.

References

Antonik, M., Felekyan, S., Gaiduk, A., and Seidel, C.A.M. (2006). Separating structural heterogeneities from stochastic variations in fluorescence resonance energy transfer distributions via photon distribution analysis. J. Phys. Chem. B *110*, 6970–6978.

Beloglazova, N., Kuznedelov, K., Flick, R., Datsenko, K. a., Brown, G., Popovic, a., Lemak, S., Semenova, E., Severinov, K., and Yakunin, a. F. (2015). CRISPR RNA binding and DNA target recognition by purified Cascade complexes from Escherichia coli. Nucleic Acids Res. *43*, 530–543.

Blosser, T.R., Loeff, L., Westra, E.R., Vlot, M., Künne, T., Sobota, M., Dekker, C., Brouns, S.J.J., and Joo, C. (2015). Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. Mol. Cell *58*, 60–70.

Bondy-Denomy, J., Garcia, B., Strum, S., Du, M., Rollins, M.F., Hidalgo-Reyes, Y., Wiedenheft, B., Maxwell, K.L., and Davidson, A.R. (2015). Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. Nature *526*, 136–139.

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E. V, and van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. Science *321*, 960–964.

Brown, M.W., Dillard, K.E., Xiao, Y., Dolan, A., Hernandez, E., Dahlhauser, S., Kim, Y., Myler, L.R., Anslyn, E., Ke, A., et al. (2018). Assembly and translocation of a CRISPR-Cas primed acquisition complex. Cell *175*, 1–13.

Caforio, A., Siliakus, M.F., Exterkate, M., Jain, S., Jumde, V.R., Andringa, R.L.H., Kengen, S.W.M., Minnaard, A.J., Driessen, A.J.M., and van der Oost, J. (2018). Converting Escherichia coli into an archaebacterium with a hybrid heterochiral membrane. Proc. Natl. Acad. Sci. U. S. A. *115*, 3704–3709.

Chandradoss, S.D., Haagsma, A.C., Lee, Y.K., Hwang, J.-H., Nam, J.-M., and Joo, C. (2014). Surface Passivation for Single-molecule Protein Studies. J. Vis. Exp. *86*, 4–11.

Chen, Y.J., Wu, D., Gelbart, W., Knobler, C.M., Phillips, R., and Kegel, W.K. (2018). Two-Stage Dynamics of in Vivo Bacteriophage Genome Ejection. Phys. Rev. X *8*.

Datsenko, K.A., and Wanner, B.L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc. Natl. Acad. Sci. U. S. A. *97*, 6640–6645.

Datsenko, K.A., Pougach, K., Tikhonov, A., Wanner, B.L., Severinov, K., and Semenova, E. (2012). Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat. Commun.

Davison, J. (2015). Pre-early functions of bacteriophage T5 and its relatives. Bacteriophage *5*, e1086500.

Deveau, H., Barrangou, R., Garneau, J.E., Labonte, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., and Moineau, S. (2008). Phage Response to CRISPR-Encoded Resistance in Streptococcus thermophilus. J. Bacteriol. *190*, 1390–1400.

Durisic, N., Laparra-Cuervo, L., Sandoval-Álvarez, Á., Borbely, J.S., and Lakadamyali, M. (2014). Single-molecule evaluation of fluorescent protein photoactivation efficiency using an in vivo nanotemplate. Nat. Methods *11*, 156–162.

Edelstein, A., Amodaj, N., Hoover, K., Vale, R., and Stuurman, N. (2010). Computer Control of Microscopes Using μManager. In Current Protocols in Molecular Biology, (Hoboken, NJ, USA: John Wiley & Sons, Inc.), p. Unit14.20.

English, B.P., Hauryliuk, V., Sanamrad, A., Tankov, S., Dekker, N.H., and Elf, J. (2011). Single-molecule investigations of the stringent response machinery in living bacterial cells. Proc. Natl. Acad. Sci. *108*, 365–373.

Van Erp, P.B.G., Jackson, R.N., Carter, J., Golden, S.M., Bailey, S., and Wiedenheft, B. (2015). Mechanism of CRISPR-RNA guided recognition of DNA targets in Escherichia coli. Nucleic Acids Res. *43*, 8381–8391.

Farooq, S., and Hohlbein, J. (2015). Camera-based single-molecule FRET detection with improved time resolution. Phys. Chem. Chem. Phys. *17*, 27862–27872.

Floc'h, K., Lacroix, F., Barbieri, L., Servant, P., Galland, R., Butler, C., Sibarita, J.-B., Bourgeois, D., and Timmins, J. (2018). Bacterial cell wall nanoimaging by autoblinking microscopy. Sci. Rep. *8*, 14038.

Gleditzsch, D., Pausch, P., Müller-Esparza, H., Özcan, A., Guo, X., Bange, G., and Randau, L. (2018). PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures. RNA Biol. 15476286.2018.1504546.

Globyte, V., Lee, S.H., Bae, T., Kim, J., and Joo, C. (2018). CRISPR/Cas9 searches for a protospacer adjacent motif by lateral diffusion. EMBO J. e99466.

Hayes, R.P., Xiao, Y., Ding, F., van Erp, P.B.G., Rajashankar, K., Bailey, S., Wiedenheft, B., and Ke, A. (2016). Structural basis for promiscuous PAM recognition in type I–E Cascade from E. coli. Nature *530*, 499–503.

Ho, H.N., Van Oijen, A.M., and Ghodke, H. (2018). The transcription-repair coupling factor Mfd associates with RNA polymerase in the absence of exogenous damage. Nat. Commun. *9*, 1570.

Hochstrasser, M.L., Taylor, D.W., Bhat, P., Guegler, C.K., Sternberg, S.H., Nogales, E., and Doudna, J. a. (2014). CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. Proc. Natl. Acad. Sci. *111*, 6618–6623.

Holden, S.J., Uphoff, S., Hohlbein, J., Yadin, D., Le Reste, L., Britton, O.J., and Kapanidis, A.N. (2010). Defining the Limits of Single-Molecule FRET Resolution in TIRF Microscopy. Biophys. J. *99*, 3102–3111.

Hoogendoorn, E., Crosby, K.C., Leyton-Puig, D., Breedijk, R.M.P., Jalink, K., Gadella, T.W.J., and Postma, M. (2015). The fidelity of stochastic single-molecule super-resolution reconstructions critically depends upon robust background estimation. Sci. Rep. *4*, 3854.

Høyland-Kroghsbo, N.M., Muñoz, K.A., and Bassler, B.L. (2018). Temperature, by Controlling Growth Rate, Regulates CRISPR-Cas Activity in Pseudomonas aeruginosa. MBio *9*.

Huang, B., Wu, H.K., Bhaya, D., Grossman, A., Granier, S., Kobilka, B.K., Zare, R.N., Huang, B., Wu, H.K., Bhaya, D., et al. (2007). Counting low-copy number proteins in a single cell. Science *315*, 81–84.

Huang, F., Hartwich, T.M.P., Rivera-Molina, F.E., Lin, Y., Duim, W.C., Long, J.J., Uchil, P.D., Myers, J.R., Baird, M.A., Mothes, W., et al. (2013). Video-rate nanoscopy using sCMOS camera-specific single-molecule localization algorithms. Nat. Methods *10*, 653–658.

Jackson, S.A., McKenzie, R.E., Fagerlund, R.D., Kieper, S.N., Fineran, P.C., and Brouns, S.J.J. (2017). CRISPR-Cas: Adapting to change. Science (80-.).

Jones, D.L., Leroy, P., Unoson, C., Fange, D., Ćurić, V., Lawson, M.J., and Elf, J. (2017). Kinetics of dCas9 target search in Escherichia coli. Science *357*, 1420–1424.

Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, Ü., Wurm, R., Wagner, R., et al. (2011). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat. Struct. Mol. Biol. *18*, 529–536.

Jung, C., Hawkins, J.A., Jones, S.K., Xiao, Y., Rybarski, J.R., Dillard, K.E., Hussmann, J., Saifuddin, F.A., Savran, C.A., Ellington, A.D., et al. (2017). Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips. Cell *170*, 35–47.e13.

Kalinin, S., Felekyan, S., Valeri, A., and Seidel, C.A.M. (2008). Characterizing Multiple Molecular States in Single-Molecule Multiparameter Fluorescence Detection by Probability Distribution Analysis. J. Phys. Chem. B *112*, 8361–8374.

Knight, S.C., Xie, L., Deng, W., Guglielmi, B., Witkowsky, L.B., Bosanac, L., Zhang, E.T., El Beheiry, M., Masson, J.-B.J.-B.J.-B., Dahan, M., et al. (2015). Dynamics of CRISPR-Cas9 genome interrogation in living cells. Science *350*, 823–826.

Kumar, M., Mommer, M.S., and Sourjik, V. (2010). Mobility of cytoplasmic, membrane, and DNA-binding proteins in Escherichia coli. Biophys. J. *98*, 552–559.

Lee, S.-H., Shin, J.Y., Lee, A., and Bustamante, C. (2012). Counting single photoactivatable fluorescent molecules by photoactivated localization microscopy (PALM). Proc. Natl. Acad. Sci. *109*, 17436–17441.

Leenay, R.T., Maksimchuk, K.R., Slotkowski, R.A., Agrawal, R.N., Gomaa, A.A., Briner, A.E., Barrangou, R., and Beisel, C.L. (2016). Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. Mol. Cell *62*, 137–147.

Ma, J., and Wang, M.D. (2016). DNA supercoiling during transcription. Biophys. Rev. *8*, 75–87.

Majsec, K., Bolt, E.L., and Ivančić-Baće, I. (2016). Cas3 is a limiting factor for CRISPR-Cas immunity in Escherichia coli cells lacking H-NS. BMC Microbiol. *16*, 28.

Manley, S., Gillette, J.M., Patterson, G.H., Shroff, H., Hess, H.F., Betzig, E., and Lippincott-Schwartz, J. (2008). High-density mapping of single-molecule trajectories with photoactivated localization microscopy. Nat. Methods *5*, 155–157.

Marraffini, L.A. (2015). CRISPR-Cas immunity in prokaryotes. Nature *526*, 55–61.

Martens, K.J.A., Beljouw, S. van, Els, S. van der, Baas, S., Vink, J.N.A., Brouns, S.J.J., Baarlen, P. van, Kleerebezem, M., and Hohlbein, J. (2018). An open microscopy framework suited for tracking dCas9 in live bacteria. BioRxiv 437137.

Martynov, A., Severinov, K., and Ispolatov, I. (2017). Optimal number of spacers in CRISPR arrays. PLoS Comput. Biol. *13*.

McGuigan, J.A.S., Kay, J.W., and Elder, H.Y. (2006). Critical review of the methods used to measure the apparent dissociation constant and ligand purity in Ca2+ and Mg2+ buffer solutions. Prog. Biophys. Mol. Biol. *92*, 333–370.

Michalet, X. (2010). Mean square displacement analysis of single-particle trajectories with localization error: Brownian motion in an isotropic medium. Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys. *82*, 041914.

Mika, J.T., and Poolman, B. (2011). Macromolecule diffusion and confinement in prokaryotic cells. Curr. Opin. Biotechnol. *22*, 117–126.

Mika, J.T., Van Den Bogaart, G., Veenhoff, L., Krasnikov, V., and Poolman, B. (2010). Molecular sieving properties of the cytoplasm of Escherichia coli and consequences of osmotic stress. Mol. Microbiol. *77*, 200–207.

Modell, J.W., Jiang, W., and Marraffini, L.A. (2017). CRISPR-Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. Nature *544*, 101–104.

Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., and Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology *155*, 733–740.

Mondal, J., Bratton, B.P., Li, Y., Yethiraj, A., and Weisshaar, J.C. (2011). Entropy-based mechanism of ribosome-nucleoid segregation in E. coli Cells. Biophys. J. *100*, 2605–2613.

Nenninger, A., Mastroianni, G., and Mullineaux, C.W. (2010). Size dependence of protein diffusion in the cytoplasm of Escherichia coli. J. Bacteriol. *192*, 4535–4540.

Olsson, J.A., Berg, O.G., Dasgupta, S., and Nordström, K. (2003a). Eclipse period during replication of plasmid R1: contributions from structural events and from the copy-number control system. Mol. Microbiol. *50*, 291–301.

Olsson, J.A., Berg, O.G., Dasgupta, S., and Nordström, K. (2003b). Eclipse period during replication of plasmid R1: contributions from structural events and from the copy-number control system. Mol. Microbiol. *50*, 291–301.

Paintdakhi, A., Parry, B., Campos, M., Irnov, I., Elf, J., Surovtsev, I., and Jacobs-Wagner, C. (2016). Oufti: An integrated software package for high-accuracy, high-throughput quantitative microscopy analysis. Mol. Microbiol. *99*, 767–777.

Palo, K., Mets, Ü., Loorits, V., and Kask, P. (2006). Calculation of photon-count number distributions via master equations. Biophys. J. *90*, 2179–2191.

Pawluk, A., Bondy-Denomy, J., Cheung, V.H.W., Maxwell, K.L., and Davidson, A.R. (2014). A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR-Cas system of pseudomonas aeruginosa. MBio *5*.

Peterson, J., and Phillips, G.J. (2008). New pSC101-derivative cloning vectors with elevated copy numbers. Plasmid *59*, 193–201.

Qian, H., Sheetz, M.P., and Elson, E.L. (1991). Single particle tracking. Analysis of diffusion and flow in two-dimensional systems. Biophys. J. *60*, 910–921.

Redding, S., Sternberg, S.H.H., Marshall, M., Gibb, B., Bhat, P., Guegler, C.K.K., Wiedenheft, B., Doudna, J.A., and Greene, E.C.C. (2015). Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. Cell *163*, 1–12.

Reyes-Lamothe, R., Tran, T., Meas, D., Lee, L., Li, A.M., Sherratt, D.J., and Tolmasky, M.E. (2014). High-copy bacterial plasmids diffuse in the nucleoid-free space, replicate stochastically and are randomly partitioned at cell division. Nucleic Acids Res. *42*, 1042–1051.

Sanamrad, A., Persson, F., Lundius, E.G., Fange, D., Gynna, A.H., and Elf, J. (2014). Single-particle tracking reveals that free ribosomal subunits are not excluded from the Escherichia coli nucleoid. Proc. Natl. Acad. Sci. *111*, 11413–11418.

Sashital, D.G., Wiedenheft, B., and Doudna, J.A. (2012). Mechanism of Foreign DNA Selection in a Bacterial Adaptive Immune System. Mol. Cell *46*, 606–615.

Severinov, K., Ispolatov, I., and Semenova, E. (2016). The Influence of Copy-Number of Targeted Extrachromosomal Genetic Elements on the Outcome of CRISPR-Cas Defense. Front. Mol. Biosci. *3*.

Shao, Q., Hawkins, A., and Zeng, L. (2015). Phage DNA Dynamics in Cells with Different Fates. Biophys. J. *108*, 2048–2060.

De Smet, J., Hendrix, H., Blasdel, B.G., Danis-Wlodarczyk, K., and Lavigne, R. (2017). Pseudomonas predators: Understanding and exploiting phage-host interactions. Nat. Rev. Microbiol.

Staals, R.H.J., Jackson, S.A., Biswas, A., Brouns, S.J.J., Brown, C.M., and Fineran, P.C. (2016). Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. Nat. Commun. *7*.

Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C., and Doudna, J.A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62–67.

Stracy, M., Lesterlin, C., Garza de Leon, F., Uphoff, S., Zawadzki, P., and Kapanidis, A.N. (2015). Live-cell superresolution microscopy reveals the organization of RNA polymerase in the bacterial nucleoid. Proc. Natl. Acad. Sci. *112*, E4390–E4399.

Subach, F. V., Patterson, G.H., Manley, S., Gillette, J.M., Lippincott-Schwartz, J., and Verkhusha, V. V. (2009). Photoactivatable mCherry for high-resolution two-color fluorescence microscopy. Nat. Methods *6*, 153–159.

Szczelkun, M.D., Tikhomirova, M.S., Sinkunas, T., Gasiunas, G., Karvelis, T., Pschera, P., Siksnys, V., and Seidel, R. (2014). Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. Proc. Natl. Acad. Sci. *111*, 9798–9803.

Uphoff, S., Reyes-Lamothe, R., Garza de Leon, F., Sherratt, D.J., and Kapanidis, A.N. (2013). Single-molecule DNA repair in live bacteria. Proc. Natl. Acad. Sci. U. S. A. *110*, 8063–8068.

Vigouroux, A., Oldewurtel, E., Cui, L., Bikard, D., and van Teeffelen, S. (2018). Tuning dCas9's ability to block transcription enables robust, noiseless knockdown of bacterial genes. Mol. Syst. Biol. *14*, e7899.

Vliet, L. Van, Sudar, D., and Young, I. (1998). Digital fluorescence imaging using cooled charge-coupled device array cameras. Cell Biol. *III*, 109–120.

Vrljic, M., Nishimura, S.Y., Brasselet, S., Moerner, W.E., and McConnell, H.M. (2002). Translational diffusion of individual class II MHC membrane proteins in cells. Biophys. J. *83*, 2681–2692.

Wallden, M., Fange, D., Lundius, E.G., Baltekin, Ö., and Elf, J. (2016). The Synchronization of Replication and Division Cycles in Individual E. coli Cells. Cell *166*, 729–739.

Westra, E.R., Pul, Ü., Heidrich, N., Jore, M.M., Lundgren, M., Stratmann, T., Wurm, R., Raine, A., Mescher, M., Van Heereveld, L., et al. (2010). H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. Mol. Microbiol. *77*, 1380–1393.

Westra, E.R., van Erp, P.B.G., Künne, T., Wong, S.P., Staals, R.H.J., Seegers, C.L.C., Bollen, S., Jore, M.M., Semenova, E., Severinov, K., et al. (2012). CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. Mol. Cell *46*, 595–605.

Xiao, Y., Luo, M., Hayes, R.P., Kim, J., Ng, S., Ding, F., Liao, M., and Ke, A. (2017). Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. Cell *170*, 48–60.

Xiao, Y., Luo, M., Dolan, A.E., Liao, M., and Ke, A. (2018). Structure basis for RNA-guided DNA degradation by Cascade and Cas3. Science *361*, eaat0839.

Xue, C., Whitis, N.R., and Sashital, D.G. (2016). Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity. Mol. Cell *64*, 826–834.

Xue, C., Zhu, Y., Zhang, X., Shin, Y.K., and Sashital, D.G. (2017). Real-Time Observation of Target Search by the CRISPR Surveillance Complex Cascade. Cell Rep. *21*, 3717–3727.

Zawadzki, P., Stracy, M., Ginda, K., Zawadzka, K., Lesterlin, C., Kapanidis, A.N., and Sherratt, D.J. (2015). The Localization and Action of Topoisomerase IV in Escherichia coli Chromosome Segregation Is Coordinated by the SMC Complex, MukBEF. Cell Rep. *13*, 2587–2596.

2

3

Visualisation of dCas9 target search *in vivo* using an open-microscopy framework

Published as: K. J. A. Martens, S. P. B. van Beljouw, S. van der Els, J. N. A. Vink, S. Baas, G. A. Vogelaar, S. J. J. Brouns, P. van Baarlen, M. Kleerebezem, J. Hohlbein, Visualisation of dCas9 target search in vivo using an open-microscopy framework. *Nat. Commun.* **10**, 3552 (2019).

Abstract

CRISPR-Cas9 is widely used in genomic editing, but the kinetics of target search and its relation to the cellular concentration of Cas9 have remained elusive. Effective target search requires the constant screening of the protospacer adjacent motif (PAM) and an upper limit for PAM screening (<30 ms) was recently found. To quantify the rapid switching between DNA-bound and freely-diffusing states of dCas9 further, we developed an open-microscopy framework that combines straightforward installation with high spatiotemporal resolution and introduce Monte-Carlo diffusion distribution analysis (MC-DDA). Our analysis revealed that dCas9 is screening PAMs 40% of the time in Gram-positive *Lactoccous lactis*, averaging just 17 ± 4 ms per binding event. Using heterogeneous expression of dCas9, we further determined the number of cellular target-containing plasmids and modelled the expected cleavage efficiency. We found that dCas9 is not irreversibly bound to target sites but can still interfere with plasmid replication. Taken together, our quantitative data will facilitate further optimization of the CRISPR-Cas toolbox.

Introduction

The discovery of clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated proteins (Cas) as a microbial defence mechanism triggered an ongoing scientific revolution, as CRISPR-Cas can be adapted to perform sequence-specific DNA modification in prokaryotes, archaea, and eukaryotes(Jiang et al., 2013; Komor et al., 2017; Liu et al., 2019; Qi et al., 2013). *Streptococcus pyogenes* Cas9 is a widely used variant(Sapranauskas et al., 2011) and an endonuclease activity-deficient version, termed 'dead' Cas9 (dCas9), has been used to visualise endogenous genomic loci in living cells(Chen et al., 2013). The biochemical interaction mechanisms of Cas9 are well understood(Anders et al., 2014; Bonomo and Deem, 2018; Globyte et al., 2018; Knight et al., 2015; Singh et al., 2016; Sternberg et al., 2014). The DNA-binding protein domain probes the DNA for a specific protospacer adjacent motif (PAM; 5'-NGG-3') via a combination of 3-dimensional diffusion and 1-dimensional sliding on the DNA(Globyte et al., 2018). Upon recognition of the PAM, the enzyme starts unwinding the DNA double helix to test for complementarity with a 20 nucleotide-long single guide RNA (sgRNA; R-loop formation). If full complementarity is found, Cas9 continues to cleave the DNA at a fixed position 3 nucleotides upstream of the PAM(Gasiunas et al., 2012).

Optimization of Cas9-mediated genomic engineering in a desired incubation time whilst minimizing off-target DNA cleavage requires exact kinetic information. In the Gram-negative bacterium *E. coli*, an upper limit for the binding time (30 ms) of dCas9 with DNA has been determined *in vivo*(Jones et al., 2017), but it is unknown if such binding times are ubiquitous in prokaryotes. In addition, there is a limited understanding of the spatiotemporal relationship between cellular copy numbers of Cas9 proteins, the number of DNA target sites and the duration and dissociation mechanisms of target-bound dCas9. Since genomic engineering of food-related microbes such as Gram-positive lactic acid bacteria(Machielsen et al., 2011) is becoming increasingly valuable(Hidalgo-Cantabrana et al., 2017; Zhang et al., 2016), it is important to assess whether previously determined dCas9 kinetic information can be transferred to food-related microbes.

To study the behaviour of dCas9 *in vivo* with millisecond time resolution, we used single-particle tracking photo-activated localisation microscopy (sptPALM)(Manley et al., 2008; Uphoff et al., 2013). In sptPALM, a photo-activatable fluorescent protein, which is by default not fluorescently active but can be activated via irradiation, is fused to the gene of interest, and the fusion protein is expressed in living cells. By stochastically activating a subset of the available chromophores, the signal of a single emitter is localized with high precision (~ 30 - 40 nm(Rieger and Stallinga, 2014; Smith et al., 2010)) and, by monitoring its position over time, the movement of the protein fusion is followed and analysed(Shen et al., 2017).

However, sptPALM mostly provides quantitative information if the protein of interest remains in a single diffusional state for the duration of a track (e.g. > 40 ms using at least 4 camera frames of 10 ms). As this temporal resolution is insufficient to elucidate *in vivo* Cas9 dynamic behaviour (< 30 ms)(Jones et al., 2017), we developed a Monte-Carlo based variant of diffusion distribution analysis (MC-DDA, for analytical DDA see(Vink et al.)) to extract dynamic information on a timescale shorter than the duration of a single track.

For the experimental realisation, we refined existing single-molecule microscopy frameworks by designing the super-resolution microscope miCube. This in-house built microscope is constructed from readily available parts, ensuring accessibility for interested laboratories. We then used MC-DDA in combination with the miCube in an assay that employs a heterogeneous expression system in order to explore the dynamic nature of DNA-dCas9 interactions in live bacteria and their dependency on (d)Cas9 protein copy numbers. In particular, we assessed dCas9 fused to photo-activatable fluorophore PAmCherry2 in in the lactic acid bacterium *L. lactis*, in the presence or absence of DNA targets. The results were combined in a model that predicts Cas9 cleavage efficiencies in prokaryotes.

Results

Monte-Carlo diffusion distribution analysis (MC-DDA) to elucidate sub 30 ms dynamic interactions with sptPALM

In the absence of cellular target sites, dCas9 is expected to be present in either one of two states (Fig.1a): bound to DNA (red), which results in low diffusion coefficients (~ 0.2 $\mu m^2/s$); or freely diffusing in the microbial cytoplasm (yellow), which results in high diffusion coefficients (~ 2.2 $\mu m^2/s$). If the transitioning between these states is slow compared to the length of each track (here: 40 ms), diffusion coefficient histograms can be fitted with two static states (Fig. 1b, top, Supplementary Fig. 1).

However, if transitioning between the states is on a similar or shorter timescale as the length of sptPALM tracks, these transient interactions of dCas9 with DNA (orange) will result in temporal averaging of the diffusion coefficient obtained from a single track. Therefore, we developed a Monte-Carlo diffusion distribution analysis (MC-DDA; Fig. 1b, bottom, Methods, with an analytical approach available elsewhere(Vink et al.) that used the shape of the histogram of diffusion coefficients to infer transitioning rates between diffusional states. The analysis is based on similar approaches used to describe dynamic conformational changes observed with single molecule Förster resonance energy transfer(Farooq and Hohlbein, 2015; Santoso et al., 2010b, 2010a). Briefly, MC-DDA consists of simulating the movement and potential interactions of dCas9 inside a cell with a Monte-Carlo approach: the simulated protein is capable of interchanging between interacting with DNA and diffusing freely, defined by $k_{\text{bound}\rightarrow\text{free}}$ and $k_{\text{free}\rightarrow\text{bound}}$. The MC-DDA diffusional data is compared with the experimental data, and by iterating on the kinetic rates and diffusion coefficients a best fit is obtained.

miCube: an open framework for single-molecule and super-resolution microscopy

For MC-DDA to deduce high kinetic rates, experimental data with high spatiotemporal resolution (< ~ 50 nm, < ~ 20 ms) is required. This is challenging, as individual fluorescent proteins have a limited photon budget (< 500 photons(Subach

et al., 2009)), and substantial background fluorescence is introduced by the living cells in which the fluorescent proteins are embedded. While suitable commercial microscopes are available, they often lack accessibility or are prohibitively expensive. This has led to the creation of a plethora of custom-built microscopes in the recent past(Aristov et al., 2018; Arsenault et al., 2015; Auer et al., 2018; Babcock, 2018; Diederich et al., 2018; Diekmann et al., 2017; Holm et al., 2014; Kwakwa et al., 2016; Ma et al., 2017; Nicovich et al., 2017; Zhang et al., 2015), ranging from simplified super-resolution microscopes(Auer et al., 2018; Babcock, 2018; Diekmann et al., 2017; Holm et al., 2014; Ma et al., 2017) to additions to commercial microscopes(Kwakwa et al., 2016) or extremely low-cost microscopes(Diederich et al., 2018; Zhang et al., 2015).

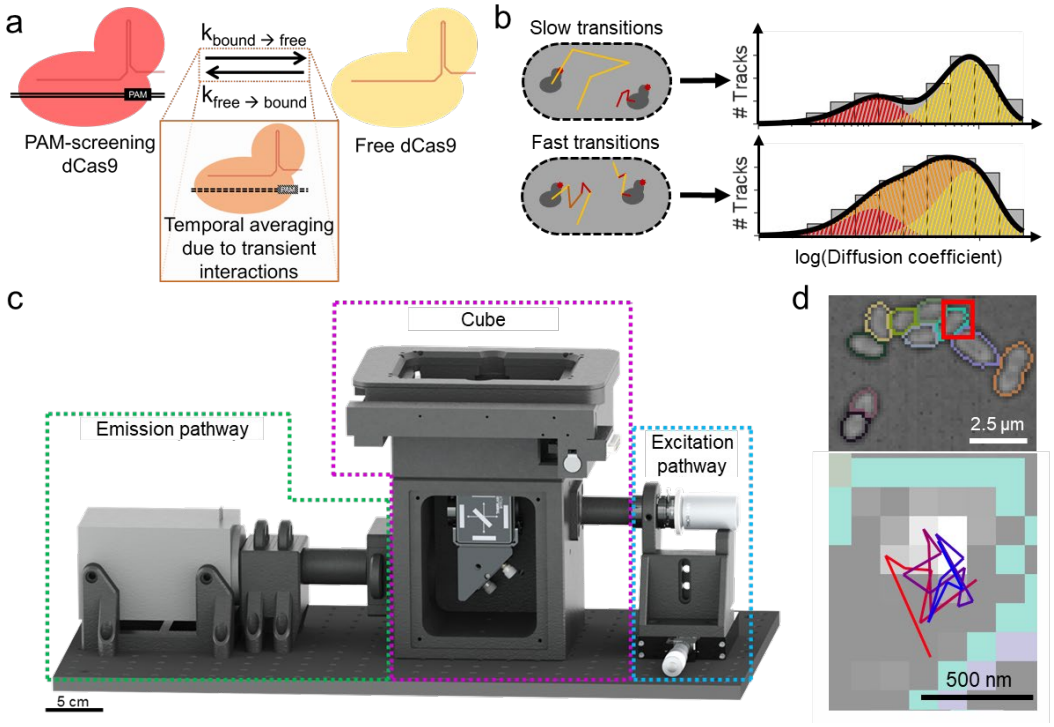


Fig. 1 Probing cellular dynamics of dCas9 on an open-source microscope using sptPALM. a Simplified expected dynamic behaviour of dCas9 in absence of DNA target sites. The protein can be temporarily bound to DNA (PAM screening), or diffuse freely in cytoplasm, with two kinetic rates governing the dynamics. If the interaction is on a similar timescale as the detection time, a temporal averaging due to transient interactions is expected. **b** If the dynamic transitions are slow with respect to the camera frame time used in sptPALM, the obtained diffusional data can be fitted with a static model (top), which assumes that every protein is either free (yellow) or DNA-bound (red), but does not interchange. If the dynamic transitions are as fast or faster than the frame time used, Monte-Carlo diffusion distribution analysis (MC-DDA; bottom) can fit the diffusional

data. In MC-DDA, dCas9 can interchange between the two states, resulting in a broader distribution. **c** Render of the open-source miCube super-resolution microscope. The excitation components, main cube, and emission components are indicated in blue, magenta, and green, respectively. Details are provided in Methods. **d** Brightfield images of *L. lactis* used for computationally obtaining the outline of the cells via watershed (top), and raw single molecule data (bottom; red outline in top is magnified) as obtained on the miCube as part of a typical experiment, overlaid with the determined track where this single molecule belongs to (starting at red, ending at blue).

To increase accessibility of single-molecule microscopy with high spatiotemporal resolution further, we developed the miCube, an open-source, modular and versatile super-resolution microscope, and provide details to allow interested researchers to build their own miCube or a derivative instrument (Fig. 1c, Supplementary Fig. 2, Methods, https://HohlbeinLab.github.io/miCube). We used 3D printed components where possible, surrounding a custom aluminium body to minimize thermal drift and provide rigidity. All custom components are supported by technical drawings (Appendix), along with STL files for direct 3D printing. We provide full details on the chosen commercial components, such as lenses, mirrors, and the camera. A detailed description on building a functioning miCube, along with rationale of the design choices, is given in the Methods section. Moreover, we discuss additional options for replacing expensive components with cheaper options.

To facilitate straight-forward installation and flexible usability of the miCube, we simplified the alignment of the excitation module by decoupling the movement in the three spatial dimensions (Supplementary Fig. 2e). A variety of imaging modalities are possible on the miCube; super-resolution microscopy in 2D and 3D(Martens et al., 2017), total internal reflection fluorescence (TIRF) microscopy, and LED-based brightfield microscopy. In its current version, the sample area fits a 96-wells plate. The excitation and illumination pathways of the microscope are fitted with 3D-printed enclosures, allowing the instrument to be used under ambient light conditions (including single-particle microscopy). Lastly, we restrained the footprint of the microscope to a 600 x 300 mm breadboard (excluding lasers; Supplementary Fig. 2b), further improving accessibility.

Linear drift calculations indicate that the system experiences a drift of 13 ± 12 nm/min in the lateral plane and 25 ± 15 nm/min in the axial plane without active drift-suppressions systems in place(Coelho et al., 2018) (average of three super-resolution measurements performed on three different days). A typical drift measurement is shown in Supplementary Fig. 3.

In vivo sptPALM in *L. lactis* on the miCube

For our sptPALM assay(Beljouw et al., 2019), we introduced dCas9 fused to the photo-activatable fluorophore PAmCherry2(Subach et al., 2009) in *L. lactis* under control of the inducible and heterogeneous *nisA* promotor(Mierau and Kleerebezem, 2005) (pLAB-dCas9, Methods). On the same plasmid, a sgRNA with no fully matching targets in the genome is constitutively expressed. We immobilized the *L. lactis* cells on agarose, and using diffused brightfield LED illumination we computationally separated the cells via the ImageJ watershed(Vincent and Soille, 1991) plugin (Fig. 1d top). Single-particle microscopy was performed with low induction levels (0.1 ng/mL Nisin) and low activation intensities (3 – 620 μW/cm^2, 405 nm) to obtain on average PAmCherry2 activation of < 1 fluorophore/frame/cell to avoid overlapping tracks (Fig. 1d, bottom). Single-particle tracks were limited to individual cells by using the previously obtained cell outlines.

dCas9 is PAM-screening for 17 ms

We first assessed the diffusional behaviour of dCas9-PAmCherry2 (hereafter described as dCas9, unless specifically mentioned) in *L. lactis* in the absence of target sites (pNonTarget plasmid; Methods). Under these conditions, dCas9 is expected to diffuse freely around the cytoplasm and screen PAM sites on the DNA for under 30 ms(Jones et al., 2017). Under this assumption, diffusion ranges from completely immobile (and thereby fully determined by the localization uncertainty: ~ 40 nm leads to ~ 0.16 μm^2/s) to freely-moving dCas9. The expected free-moving diffusion coefficient can be theoretically described: the fusion protein has a hydrodynamic radius of 5 - 6 nm(Nishimasu et al., 2015; Subach et al., 2009), resulting in a diffusion coefficient of 36 – 43 μm^2/s(Edward, 1970). Cytoplasmic

retardation of ~ 20x due to increased viscosity and crowding effects reduces this to ~ 1.8 – 2.2 μm^2/s(Trovato and Tozzini, 2014). We obtained diffusion coefficients in the range of ~ 0 – 3 μm^2/s (Fig. 2a), which is within the expected range.

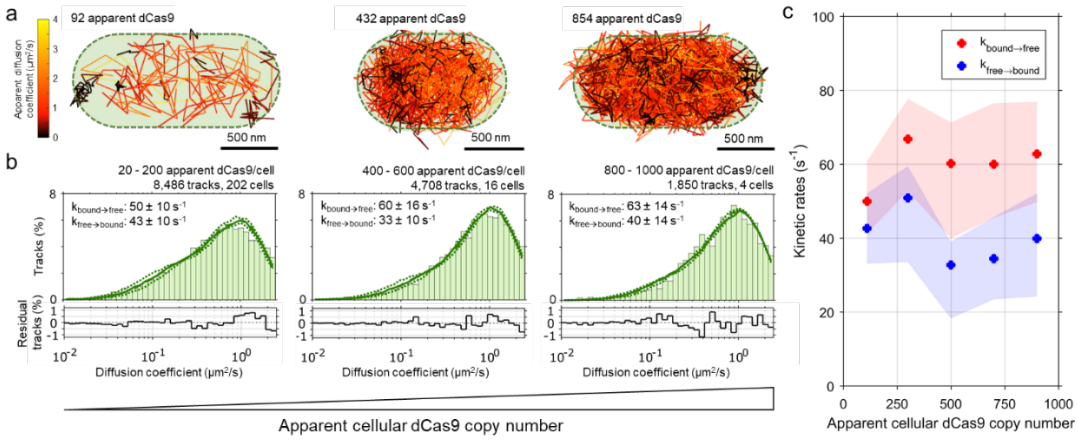


Fig. 2 sptPALM of dCas9-PAmCherry2 in pNonTarget *L. lactis* with increasing dCas9 concentration. a Identified tracks in single pNonTarget *L. lactis* cells. Tracks are colour-coded based on their diffusion coefficient. Three separate cells are shown with increasing cellular concentration of dCas9. Green dotted outline is an indication for the cell membrane. **b** Diffusion coefficient histograms (light green) belonging to 20 - 200, 400 - 600, and 800 - 1000 dCas9 copy numbers, from left to right. Histograms are fitted (dark green line) with a theoretical description of state-transitioning particles between a mobile and immobile state (dashed line represents 95% confidence interval based on bootstrapping the original data). Five diffusion coefficient histograms (Supplementary Fig. 4) were globally fitted with a single free diffusion coefficient (2.0 ± 0.1 μm^2/s), a single value for the localization error (σ = 38 ± 3 nm = 0.15 ± 0.03 μm^2/s), and 5 sets of $k_{bound \rightarrow free}$ and $k_{free \rightarrow bound}$ values (indicated in the figures). Residuals of the fit are indicated below the respective distribution. **c** $k_{bound \rightarrow free}$ (red) and $k_{free \rightarrow bound}$ (blue) plotted as function of the apparent cellular dCas9 copy number. Solid dots show the fits of the actual data; filled areas indicate the 95% confidence intervals obtained from the bootstrapped iterations of fitted MC-DDAs with 20.000 simulated proteins.

We used a heterogeneous promotor (*nisA*, Methods), causing the apparent cellular dCas9 copy numbers to vary between 20 and ~ 1000 (Fig. 2a, Supplementary Fig. 4; cells with less than 20 copies were excluded as we corrected for ~ 7 tracks (~ 14 apparent dCas9) found in non-induced cells). The value of the cellular dCas9 is an

approximation (Discussion), but a relative increase in cellular dCas9 copy number is certain. We then created five diffusional histograms belonging to cells with a particular apparent dCas9 copy number range (ranges of ~ 200 dCas9 copy number intervals; Fig. 2b, Supplementary Fig. 4). These diffusional histograms are fitted with the aforementioned MC-DDA, where the shape of the MC-DDA is governed by the localization uncertainty, the free-moving diffusion coefficient, and the kinetic rates of PAM-screening. The localization uncertainty and free-moving diffusion coefficient are independent of cellular dCas9 copy number, since they are determined by the number of photons and a combination of hydrodynamic radius and cytoplasm viscosity, respectively. Therefore, the histograms were globally fitted with a combination of 5 MC-DDAs, each consisting of 20.000 simulated dCas9 proteins, containing a single value for free-moving diffusion coefficient (D_{free} = 2.0 ± 0.1 μm^2/s, in agreement with the theoretical expectation of ~ 1.8 – 2.2 μm^2/s), a single value for localization uncertainty (σ = 38 ± 3 nm, or $D_{immobile}$* = 0.15 ± 0.03 μm^2/s, expected for fluorescent proteins illuminated for 2 ms(Beljouw et al., 2019; Martens et al., 2017)), and five pairs of $k_{free \rightarrow bound}$ and $k_{bound \rightarrow free}$ (specified in Fig. 2b and 2c).

The obtained kinetic constants of $k_{free \rightarrow bound}$ and $k_{bound \rightarrow free}$ were 40 ± 12 s^{-1} and 60 ± 13 s^{-1} (mean ± 95% CI), respectively, and did not show a significant dependence on apparent cellular dCas9 copy number (Fig. 2c). This indicates that dCas9 is PAM-screening for 17 ± 4 ms in *L. lactis*, consisting of screening 1 or more PAMs via 1D diffusion. This value is in the same order of magnitude as the upper limit of 30 ms reported earlier for PAM-screening in *E. coli*(Jones et al., 2017), suggesting that these PAM-screening kinetics are a general feature of dCas9. Additionally, dCas9 is on average diffusing within the cytoplasm for 25 ± 8 ms before finding a new site for PAM screening. This duration is governed by the diffusion coefficient of the fusion protein, along with the average distance between DNA PAM sites. These results also entail that dCas9 is diffusing in the cytoplasm ~ 60% of the time, while interacting with the DNA ~ 40% of the time. Removal of the sgRNA resulted in similar diffusional data, which agrees with PAM-screening being a solely protein-DNA interaction ($k_{free \rightarrow bound}$: 34 ± 16 s^{-1}; $k_{bound \rightarrow free}$: 62 ± 21 s^{-1}; diffusion time on average 29 ± 18 ms; PAM-screening time on average 16 ± 6 ms; Supplementary Fig.

5). This also indicates that partial sgRNA-DNA matching of dCas9 with non-targets is not prevalent enough in our assay to affect the screening time significantly.

Target-binding of dCas9 can be observed with sptPALM

We then investigated the effect of DNA target sites complementary to the sgRNA loaded dCas9. To this end, we introduced 5 target sites on a plasmid (pTarget; Methods), which replaced the pNonTarget plasmid used so far. Qualitative visualisation of diffusion in the *L. lactis* bacteria shows tracks with small diffusion coefficients (Fig. 3a, black tracks), indicative of target-bound dCas9. This immobile population can be observed throughout the dCas9 copy number range but is more prevalent in cells with lower cellular dCas9 copy numbers.

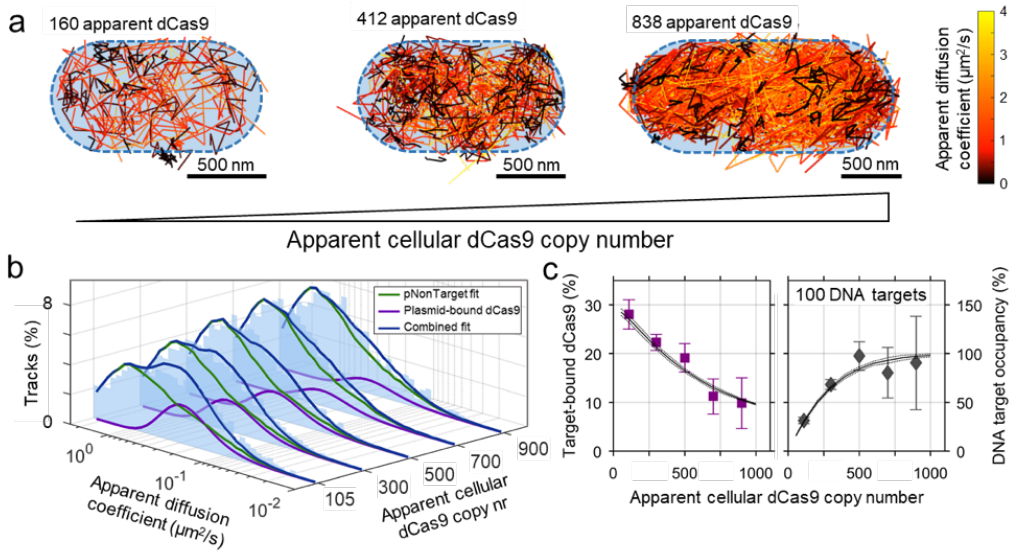


Fig. 3: sptPALM of dCas9-PAmCherry2 in pTarget *L. lactis* shows target-binding behaviour of dCa9s. a Identified tracks in individual pTarget *L. lactis* cells. Tracks are colour-coded based on their diffusion coefficient. Three separate cells are shown with increasing dCas9 concentration. Blue dotted outline is an indication for the cell membrane. **b** Diffusion coefficient histograms (light blue) are fitted (dark blue line) with a combination of the respective fit of pNonTarget *L. lactis* cells (green line), along with a single globally fitted population corresponding to target-bound dCas9 (purple) at 0.38 ± 0.04 $\mu m^2/s$. **c** Left: The population size of the plasmid-bound dCas9 decreases as a function of the cellular dCas9 copy number. The error bar of the measurement is based on the 95% confidence interval determined by bootstrapping); the solid line is a model fit with 20 plasmids, with a 95% confidence interval determined by repeating the model simulation.

Right: Occupancy of DNA targets by dCas9 based on 20 target plasmids (100 DNA target sites), based on the same data as presented in the left figure.

We expect target-bound dCas9 to move with a diffusion coefficient determined by the plasmid size, which is independent on the cellular dCas9 copy number. Therefore, we globally fitted the pTarget-obtained diffusional histograms with a combination of the corresponding pNonTarget MC-DDA fit and an additional single diffusional state belonging to target-bound dCas9 (Fig3b, $D_{plasmid}$* = 0.38 ± 0.04 μm^2/s = $D_{immobile}$* + 0.23 μm^2/s, which agrees with the expected diffusion coefficient from plasmids of similar size in bacterial cytoplasm(Prazeres, 2008; Trovato and Tozzini, 2014; Vos and M, 1987)). The plasmid-bound dCas9 population decreases with increasing apparent cellular dCas9 copy numbers from 28 ± 3% at 105 (20 – 200) copies to 10 ± 5 % at 900 (800 – 1000) copies (Fig. 3c left, purple squares; mean ± 95% CI). No target-binding behaviour was observed when the sgRNA was removed (Supplementary Fig. 5).

dCas9 is bound to target DNA for ~ 1.6 minutes and interferes with plasmid replication

This anti-correlation between dCas9 copy number and the size of the plasmid-bound population is indicative of competition for target sites by an increasing amount of dCas9 proteins. To evaluate this hypothesis, we consecutively simulated dCas9 proteins until the cellular dCas9 copy number was reached (Methods). In the simulation, every protein binds or dissociates from a PAM with the kinetic constants determined previously, and will instantly bind to a target site if it binds to a PAM directly adjacent to it. We thus disregard effects of 1D sliding on the DNA, but we believe these effects are limited, as 1D sliding between PAM sites has a low probability when PAMs are randomly positioned on the DNA (<< ~ 5% at 32 bp distance average(Globyte et al., 2018)). A k_{off} is introduced which dictates removal of dCas9 from the target sites.

This model fully explained the dependency of the target-bound dCas9 fraction on the cellular dCas9 copy number (Fig. 3c left, black line). The slope of the curve towards low cellular dCas9 concentration is dependent on the total cellular number of PAM sites and k_{off}. Assuming on average 1.5 genome's worth of DNA (haploid

genome replicated in half the cells) present in the cell, the k_{off} is ~ 0.01 ± 0.003 s^{-1}. The number of DNA target sites determines the lower bound of the model, and ~ 100 ± 50 DNA target sites (~ 20 ± 10 plasmids) led to the observed bound fraction at 900 cellular dCas9 proteins. The fit of the number of target sites at high cellular dCas9 concentration is independent of k_{off}, since at the modelled concentrations and PAM-screening kinetic parameters, the target sites are essentially fully occupied (Fig. 3c, right). It thus follows that the used pTarget plasmid, a derivative of pNZ123, is present at a lower copy number than expected (~ 60 - 80) during measurements(Vos and M, 1987). This could hint towards interference of plasmid replication due to dCas9 binding(Whinn et al., 2018). We investigated this with quantitative polymerase chain reaction (qPCR)(Tal and Paulsson, 2012), and we indeed observed a decrease in the amount of pTarget DNA with dCas9 production (Supplementary Fig. 6).

These collective results lead to the model presented in Fig. 4a. dCas9 diffuses freely in the cytoplasm for 25 ± 8 ms on average, and will then interact with a PAM site for 17 ± 4 ms. If the PAM site is not directly adjacent to a target site, dCas9 will move back to freely diffusing in the cytoplasm. However, if the PAM is directly followed by a target site, dCas9 will be bound to this site for 1.6 minutes on average, before it is removed by intrinsic or extrinsic factors.

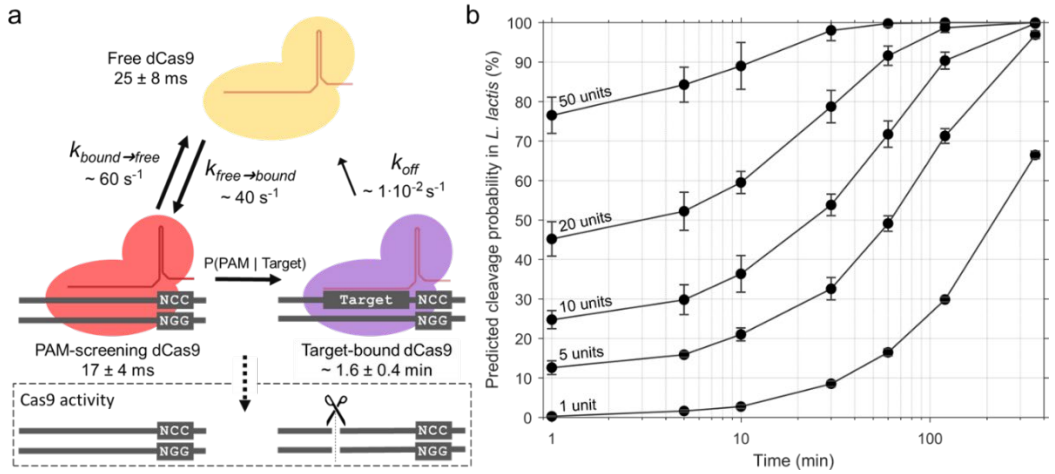


Fig. 4: Extrapolation of the dCas9 dynamic model to assess single target cleavage by Cas9 a The proposed model surrounding dCas9 interaction with the obtained kinetic rates. Free dCas9 (yellow) in the cytoplasm interact with PAM sequences (5'-NGG-3') on average

every 25 ms. If the PAM is not in front of a target sequence (red), only PAM-screening will occur for on average 17 ms. If the PAM happens to be in front of a target, the dCas9 will be target-bound (purple). We extend this model to predict Cas9 cleavage under conditions where target-bound Cas9 will always cleave the target DNA. **b** Calculated predicted probability that a single target in the *L. lactis* genome is cleaved after a certain period of time with a certain cellular Cas9 copy number, based on the model shown in **a**. Error bars indicate standard deviation calculated from iterations of the model.

A single copy of Cas9 find a single DNA target in ~ 4 hours

We adapted the computational target-binding model to predict Cas9 cleavage in *L. lactis* and other prokaryotes with similar DNA content. We assume that all DNA is accessible to Cas9 and that Cas9 behaves identical to dCas9, but will cleave a target directly after binding. Our proposed Cas9 kinetic scheme depends only on PAM-screening kinetic rates and the ratio of total PAM sites to target sites. We predicted the incubation time-dependent probability that a certain number of cellular Cas9 proteins will bind a single target site on the *L. lactis* genome (Fig. 4b).

The model shows that a single Cas9 protein can effectively find a single target with 50% probability in ~ 4 hours. It also shows that an increasing cellular Cas9 copy number quickly decreases this search time: With 10 cellular copies of Cas9, the search time is reduced to ~ 25 minutes, and 20 copies reduce the search time to ~ 3 minutes. Therefore, a single target is almost certainly found within a typical prokaryotic cell generation time (> ~ 20 min). This agrees with *in vivo* data of Cas9(Jones et al., 2017) (accounting for *E. coli*'s larger genome (~ 4.6 mbp versus ~ 2.5 mbp)) and with *in vivo* data of Cascade in *E. coli*(Vink et al.), though in different organisms or with different CRISPR-Cas systems.

Discussion

We have designed a sptPALM assay to probe DNA-protein interactions *in vivo*, and assessed the kinetic behaviour of dCas9 in *L. lactis* on the open-hardware, super-resolution microscope miCube. The high spatiotemporal resolution of the experimental data along with the heterogeneity of the used induction protocol

allowed us to develop a Monte-Carlo diffusion distribution analysis (MC-DDA) of the diffusional equilibrium.

The obtained dCas9 PAM-screening kinetic rates ($k_{free \rightarrow bound} = 40 \pm 12\ s^{-1}$, $k_{bound \rightarrow free} = 60 \pm 13\ s^{-1}$) indicate that non-target binding of dCas9 has a mean lifetime of 17 ± 4 ms, and spends ~ 40% of its time on PAM screening. In fact, a 1:1 ratio between diffusing and binding was shown to be optimal for target search time of DNA binding proteins(Slutsky and Mirny, 2004). The MC-DDA further suggests that the kinetic rates governing PAM-dCas9 interactions do not depend on cellular copy number levels of dCas9.

We observed target-binding of dCas9, and showed that higher cellular dCas9 copy numbers resulted in lower probabilities of target-bound dCas9, although absolutely more targets were occupied by dCas9. We linked this finding to the previously found $k_{free \rightarrow bound}$ and $k_{bound \rightarrow free}$ rates and postulate that dCas9 dissociation from target sites is responsible for the obtained probabilities of target binding by dCas9. We made two assumptions when obtaining absolute cellular dCas9 copy numbers. Firstly, we assumed that measurements directly end after all fluorophores in the centre of the microscopy field of view have been imaged once. Secondly, we assumed a maturation grade of 50% (identical to that of PAmCherry1 in *Xenopus*(Durisic et al., 2014)). Although an exact determination is possible(Durisic et al., 2014; Nagai et al., 2002), this is beyond the scope of this study.

We obtained a dCas9-target k_{off} rate of ~ 0.01 s^{-1} that is dependent on the exact cellular dCas9 copy number and total *L. lactis* genomic content. The biological cause of dissociation of target bound dCas9 from DNA remains speculative: it could be an intrinsic property, resulting in spontaneous release from target-sites, or it could be caused by an extrinsic factor, such as RNA transcription or DNA replication. We do not expect RNA polymerase activity on the DNA target sites, although we did not actively block transcription. It is currently unknown whether genomic target-bound dCas9 dissociates from the DNA due to DNA replication, with studies contradictory showing that dCas9 is removed during cell duplication(Jones et al., 2017) and that dCas9 is hindering genomic DNA replication(Whinn et al., 2018). We note that genomic DNA replication substantially differs from the rolling-circle DNA replication of pTarget(Khan, 1997).

Our data indicates that dCas9 binding to plasmid DNA hinders DNA rolling-circle replication. The pNZ123 plasmid, of which pTarget is a derivative, is believed to be high-copy(Vos and M, 1987) (60 – 80 plasmids per cell), although the quantification of plasmid copy numbers is challenging (discussed for the single-cell level in reference (Tal and Paulsson, 2012)). Our model suggests that pTarget is present in only ~ 20 copy numbers during our measurements. Although we saw an effect of dCas9 production on pTarget copy number via qPCR, the obtained decrease (~ 20%) is not as large as observed with sptPALM (~ 70%). The median cellular dCas9 copy number, however, is low (~ 40; Supplementary Fig. 6) compared to most of the dCas9 copy number bins evaluated with MC-DDA. Therefore, using the averaged cellular community, not all pTarget (60 – 80 cellular plasmids containing 300 – 400 target sites), are occupied by a dCas9 protein, which would affect the ensemble qPCR results. The sptPALM plasmid copy number determination, on the other hand, is mostly determined by the *L. lactis* sub-population with high dCas9 copy numbers, for which pTarget replication is restricted more strongly.

We used our model to make predictions about Cas9 cleavage probabilities, based on kinetic values extracted from the MC-DDA, which are not influenced by the approximated cellular dCas9 copy number. The kinetic parameters of dCas9-PAmCherry2 provide estimates for those of Cas9. We reason that $k_{\text{bound}\rightarrow\text{free}}$ will be unchanged, since this rate is based on the duration of the PAM screening, while $k_{\text{free}\rightarrow\text{bound}}$ will be slightly lower for Cas9 compared to dCas9-PAmCherry2, due to the relatively higher diffusion coefficient of Cas9. The model can be expanded to incorporate a protein diffusion coefficient to obtain a modified $k_{\text{free}\rightarrow\text{bound}}$ rate, and to include accessibility of the DNA. These additions would allow the model to predict Cas9 behaviour in more diverse environments such as eukaryotic cells. Other computational models have taken these parameters into account(2016), but these models were not based on experimental *in vivo* data, and were based on different assumptions.

Our open microscopy framework enables the study of *in vivo* protein-DNA interactions with high spatiotemporal resolution, here shown for CRISPR-Cas9 target search, and improves the general accessibility of super-resolution microscopy. Our data shows that heterogeneity in an expression system can be used to obtain new

insights in any protein-DNA or protein-protein interaction *in vivo*, here indicating that target-bound dCas9 interferes with rolling-circle DNA replication. The derived kinetic parameters and information on target search times provides valuable practical insights in CRISPR-Cas engineering and gene silencing in lactic acid bacteria specifically, and suggest to reflect prokaryotic Cas9 search times in general.

Methods

Detailed description miCube

We designed the miCube to be easy to set up and use, while retaining a high level of versatility. The instrument and its design choices will be described in three parts: the excitation path; the emission path, and the 'cube' connecting the sample with the excitation and emission paths. Throughout this description, we will refer to numbered parts as shown in Supplementary Fig. 2a and c and described in Supplementary Table 1. The information on the miCube presented here can also be found on https://HohlbeinLab.github.io/miCube/component_table.html. The instrument is fully functional in ambient light, due to a fully enclosed sample chamber, illumination pathway and emission pathway. Moreover, the miCube has a small footprint: the final design of the miCube, excluding the lasers and controllers, fits on a 300 x 600 mm Thorlabs breadboard. We placed the whole ensemble in a transparent polycarbonate box (MayTec Benelux, Doetinchem, The Netherlands) to minimize airflow disturbing the setup during experiments.

miCube excitation path

The excitation path is designed to be both robust and easy to align and adjust. The four laser sources located in an Omicron laser box are combined and guided via a single mode fibre towards a reflective collimator (nr. 18) ensuring a well-collimated beam. The reflective collimator is attached directly to an aperture (nr. 17), a focusing lens (nr. 16, 200 mm focal length), and an empty spacer (nr. 12). This excitation ensemble is placed in the 3D-printed piece designed to hold the assembly into place (nr. 13). This holder is then attached to a right-angled mounting plate (nr. 14), which is placed on a 25mm translation stage (nr. 15). The translation stage should be placed at such a position on the breadboard that the focusing lens (nr. 16) is exactly 200

mm separated from the back-focal plane of the objective when following the laser path.

Easy alignment and adjustment are ensured by isolating the three axes of movement of this excitation ensemble (Supplementary Fig. 2e). Adjustments of distance from objective is achieved by moving the collimator ensemble (nrs. 12, 16-18) inside its holder (nr. 13). Height of the path can be adjusted via a bracket clamp that supports the collimator ensemble (nrs. 13 and 14), and the horizontal alignment can be adjusted via a translation stage where the bracket clamp rests on (nr. 15). We note that the excitation pathway is uncoupled from any laser source due to the fibre-connection, allowing for freedom of choice for the excitation laser unit.

Additionally, the translation stage (nr. 15) can be used to enable highly inclined illumination (HiLo) or total internal reflection (TIR). The stage allows fine and repeatable adjustment of the excitation beam position on the back focal plane of the objective. By aligning the excitation beam in the centre of the objective, the microscope will act as a standard epifluorescence instrument. If the excitation beam is aligned towards the edge of the back focal plane, the miCube will operate in HiLo or TIR.

miCube cube and sample mount

The central component of the miCube is the cube (nr. 5) that connects excitation path, emission path, and the sample. The cube is manufactured out of a solid aluminium block maximising stability and minimising effects of drift due to thermal expansion. Black anodization of the block prevents stray light and unwanted reflections. The illumination path is further protected from ambient light by screwing a 3D-printed cover (nr. 11) on the side of the cube, as well as a door to close the cube off.

Next, the ‘dichroic mirror – full mirror’ part is assembled (nrs. 6-10). The dichroic mirror unit (nr. 7) consists of a dichroic mount that is magnetically attached to an outer holder. On the side of the dichroic mirror unit, opposing the excitation path, a neutral density filter (nr. 6) is placed to prevent scattered non-reflected light entering the miCube thereby minimizing background signal being recorded by the camera. At the bottom of the dichroic mount assembly, a TIRF filter (nr. 8) is placed to

remove scattered back-reflected laser light from entering the emission pathway. This ensembled dichroic mirror unit is screwed via a coupling element (nr. 9) to a mirror holder containing a mirror placed at a 45° angle (nr. 10), which reflects the emission light from the objective to the camera. This completed 'dichroic mirror – full mirror' part is screwed into the backside of the miCube via two M6 screws, which hold the ensemble into place and directly in line with the excitation path (nrs. 12-18), the objective (nr. 3), and the tube lens (nr. 30).

Then, an objective (nr. 3) (Nikon 100x oil, 1.49 NA, HP, SR) is directly screwed into an appropriate thread on top of the cube. Around the objective, a sample mount (nr. 4) is screwed on top of the cube, which is designed to ensure correct height of the sample with respect to the parfocal distance of the chosen objective. We opted for using a sample mount, as it can be easily swapped for another to retain freedom in peripherals. For example, only the height of the sample mount has to be changed if an objective has a different parfocal distance as the one used here. We designed two different sample mounts (nr. 4a, 4b). The first one can hold an xy-translation stage with z-stage piezo insert (nr. 2), to enable full spatial control of the sample (nr. 4a). The second one only holds the z-stage piezo insert, which decreases instrument cost (nr. 4b). In any case, the xy-translation stage with z-stage piezo insert, or only the z-stage piezo insert is screwed in place into corresponding threaded holes in the sample mount. A glass slide holder (nr. 1) is made from aluminium to fit inside a 96-wells-holder like the *z*-stage (nr. 2).

miCube detection path

A tube lens ensemble is made (nrs. 27-30) which houses a 200 mm focal length tube lens (Thorlabs) in a 3D-printed enclosure which provides space to slot in an emission filter (nrs. 27,28). This ensemble is then attached directly to the miCube by screwing it into place with four M6 screws. The alignment of the tube lens is therefore exactly in line with the emission light, as the centre of the full mirror (nr. 10) is at the same height of the tube lens. The direction of the emission light can be aligned, which can simply be achieved by tuning the angle of the full mirror (nr. 10).

A cover (nr. 25) is attached to the tube lens ensemble to ensure darkness of the emission path, which is connected to the tube lens by a 3D-printed connector piece

(nr. 26). On the other end of the cover, a 3D-printed holder for 2 astigmatic lenses (nr. 21) is placed and screwed into place in the breadboard. Astigmatic lenses (nrs. 22-24) can optionally be used to enable 3D super-resolution microscopy(Huang et al., 2008). They can be easily changed for lenses with a different focal length or with empty holders. With this, astigmatism can be enabled or disabled, and a choice between more accurate z-positional information with a lower total z-range, or less accurate information with a larger range can be made. The Andor Zyla 4.2 PLUS camera (nr. 19) is placed behind the astigmatic lens holder, and is positioned in a 3D-printed camera mount (nr. 20) to ensure correct height and position of the camera, so that the focus of the tube lens is at the camera chip. We chose for a scientific Complementary Metal-Oxide Semiconductor (sCMOS) camera to take advantage of a larger field of view and increased temporal resolution compared to the more traditional electron-multiplying charge coupled device (EMCCD) cameras(Almada et al., 2015).

Note that the length of the cover (nr. 25) and the alignment of the holes at the feet of the 3D-printed astigmatic lens holder (nr. 21) are dependent on the focal length of the tube lens, and of the position of the chosen camera chip with regards to the 3D-printed mount for the camera. The pieces used here were designed for the Andor Zyla 4.2 PLUS, a 200 mm focal length tube lens, and the used custom-designed camera mount (nr. 20).

Biological methods

Strain preparation and plasmid construction

Lactococcus lactis NZ9000 was used throughout the study. NZ9000 is a derivative of *L. lactis* MG1363(Kuipers et al., 1998) in which the chromosomal *pepN* gene is replaced by the *nisRK* genes that allow the use of the nisin-controlled gene expression system(Mierau and Kleerebezem, 2005). Cells were grown at 30°C in GM17 medium (M17 medium (Tritium, Eindhoven, The Netherlands) supplemented with 0.5% (w/v) glucose (Tritium, Eindhoven, The Netherlands) without agitation.

DNA manipulation and transformation

Vectors used in this study are listed in Supplementary Table 2. Oligonucleotides (Supplementary Table 3) and primers Supplementary Table 4) were synthesised at Sigma-Aldrich (Zwijndrecht, The Netherlands). Plasmid DNA was isolated and purified using GeneJET Plasmid Prep Kits (Thermo Fisher Scientific, Waltham, MA, USA). Plasmid digestion and ligation were performed with Fast Digest enzymes and T4 ligase respectively, according to the manufacturer's protocol (Thermo Fisher Scientific, Waltham, MA, USA). DNA fragments were purified from agarose gel using the Wizard SV gel and PCR Clean-Up System (Promega, Leiden, The Netherlands). Electro competent L. lactis NZ9000 cells were generated using a previously described method(Wells et al., 1993). Prior to electro-transformation, ligation mixtures were desalted for one hour by drop dialysis on a 0.025 µm VSWP filter (Merck-Millipore, Billerica, US) floating on MQ water. Electro-transformation was performed with GenePulser XcellTM (Bio-Rad Laboratories, Richmond, California, USA) at 2 kV and 25 µF for 5 ms. Transformants were recovered for 75 minutes in GM17 medium supplemented with 200 mM MgCl2 and 2 mM CaCl2. Chemically competent E. coli TOP10 (Invitrogen, Breda, The Netherlands) were used for transformation and amplification of the Pnis-dCas9-PAmCherry2-containing pUC16 plasmid (Supplementary Fig. 7). Antibiotics were supplemented on agar plates to facilitate plasmid selection: 10 µg/ml chloramphenicol (for pTarget/pNonTarget) and 10 µg/ml erythromycin (for pLAB-dCas9). Screening for positive transformants was performed using colony PCR with KOD Hot Start Mastermix according to the manufacturer's instructions (Merck Millipore, Amsterdam, the Netherlands). Electrophoresis gels were made with 1% agarose (Eurogentec, Seraing, Belgium) in tris-acetate-EDTA (TAE) buffer (Invitrogen, Breda, The Netherlands). Plasmid digestions were compared with *in silico* predicted plasmid digestions (Benchling; https://benchling.com).

pLAB-dCas9 plasmid construction

Construction of the pLAB-dCas9 plasmid is described in detail elsewhere(Beljouw et al., 2019). Briefly, the fragment containing the sequence of Pnis-dCas9-PAmCherry2 is flanked by XbaI/SalI restriction sites (Supplementary Fig. 7, Supplementary Note) and was synthesized by Baseclear (Baseclear B.V., Leiden,

The Netherlands), and cloned in a pUC16 plasmid. After transformation in *E. coli*, the plasmid was isolated and digested with XbaI and SalI to obtain the Pnis-dCas9-PAmCherry2 fragment. The Cas9 expression module was removed from the pLABTarget expression vector(Els et al., 2018) by digestion with XbaI and SalI and replaced by the XbaI-SalI fragment containing Pnis-dCas9-PAmCherry2. The single-stranded guide RNA (sgRNA) for targeting *pepN* was constructed and inserted according to earlier described protocol(Els et al., 2018) to yield the pLAB-dCas9 vector. The plasmids used in this study, and vector maps for pLABTarget and pLAB-dCas9 are available upon request. pLAB-dCas9-PAmCherry2 was sequenced, and was confirmed to be intact in the used strains.

pLAB-dCas9 no-sgRNA

The pLAB-dCas9-nosgRNA plasmid was constructed by BoxI/SmaI digestion of the pLAB-dCas9-PAmCherry2 plasmid, and subsequent self-ligation. This resulted in deletion of the sgRNA handle and transcriptional terminator, successfully removing the functional sgRNA. The resulting pLAB-dCas9-nosgRNA plasmid was confirmed via sequencing.

pTarget and pNonTarget plasmid construction

The plasmid with binding sites for dCas9 (pTarget) was established by engineering five *pepN* target sites in the pNZ123 plasmid(van Asseldonk et al., 1990). To this end, two single-stranded oligonucleotides (10 µl of 100 µM, each, Supplementary Table 3) that upon hybridization form the a single target sequence for the *pepN*-targeting sgRNA were incubated in 80 µl annealing buffer (10 mM Tris [pH = 8.0] and 50 mM NaCl) for 5 minutes at 100°C, followed by gradual cooling to room temperature. The annealed mixed multiplexed oligonucleotides were cloned in HindIII-digested pNZ123. Afterwards we selected a derivative that contains five *pepN* target sites via colony PCR (Supplementary Table 4). HindIII re-digestion was prevented by flanking the *pepN* DNA product by different base pairs, changing the HindIII site. Plasmids with five *pepN* target sites were designated pTarget (Supplementary Fig. 8). Plasmids without the pepN target sites (the original pNZ123 plasmids) were designated pNonTarget. The vector maps for pTarget and

pNonTarget are shown in Supplementary Fig. 8. Correct insertion of the five *pepN* target sites was confirmed via sequencing.

Construction of strains harbouring both pLAB-dCas9 and pTarget/pNonTarget

Electro competent *L. lactis* NZ9000 cells(Wells et al., 1993) harbouring pLAB-dCas9 were transformed with pTarget or with pNonTarget and subsequently used for sptPALM or stocked at -80°C.

Quantitative Polymerase Chain Reaction (qPCR) for plasmid copy number detection

Both *L. lactis* strains containing pLAB-dCas9 and pTarget or pNonTarget were grown under similar lab conditions as the imaging experiments performed in this study (N=2). After 3 hours of growth, the cultures were split and dCas9 was induced (0 ng/ml Nisin, 0,4 ng/ml Nisin and 2 ng/ml Nisin). The cells were then harvested after 12 hours of growth by centrifugation. The cell pellets were washed, and DNA was extracted using InstaGene Matrix (Bio-Rad Laboratories, Richmond, California, USA).

Oligonucleotides were designed to amplify a region of spanning approximately 1000 base pairs on both pTarget and pNonTarget, and a region of similar length on the NZ9000 chromosome (Q3 + Q4 and Q7+Q8; Supplementary Table 4). These oligonucleotides were used in a PCR reaction to generate templates which were diluted to function as a calibration curve in the following qPCR. Both qPCR reactions were performed on each isolated DNA sample (6 technical replicates) and the ratio between measured chromosomal amplicons (Q5+Q6) and plasmid amplicons (Q1+Q2) was determined. The samples which were uninduced with Nisin were used to standardize the estimated pTarget and pNonTarget copy numbers.

Single molecule microscopy

Strain preparation

The strains to be used for single molecule microscopy were grown o/n from glycerol stocks at 30°C in chemically defined medium for prolonged cultivation(CDMPC)(Goel et al., 2011). Then, they were sub-cultured at 5% v/v and grown for 3 hours (average duplication time in CDMPC is ~ 90 minutes (determined via OD600 measurements)), before induced with 0.1 ng/ml Nisin. 90 minutes later, the sample preparation began (see below).

Sample preparation

Samples were prepared as described previously(Beljouw et al., 2019). Briefly, after culturing of the cells, 0.5 µg/mL ciprofloxacin (Sigma-Aldrich, Zwijndrecht, The Netherlands) was added to slightly inhibit further cell division and DNA replication for sgRNA-pTarget and sgRNA-pNonTarget experiments(Drlica et al., 2008). Then, cells were centrifuged (3500 RPM for 5 minutes; SW centrifuge (Froilabo, Mayzieu, France) with a CENSW12000024 swing-out rotor fitted with CENSW12000006 15 ml culture tube adaptors) and washed two times by gentle resuspension in 5 mL phosphate-buffered saline (PBS; Sigma-Aldrich, Zwijndrecht, The Netherlands). After removal of the supernatant, cells were resuspended in ~ 10-50 µL PBS from which 1-2 µL was immobilized on 1.5% 0.2 µm-filtered agarose (Certified Molecular Biology Agarose; BioRad, Veenendaal, The Netherlands) pads between two heat-treated glass coverslips (Paul Marienfeld GmbH & Co. KG, Lauda-Königshofen, Germany; #1.5H, 170 µm thickness). Heat treatment of glass coverslips involves heating the coverslips to 500°C for 20 minutes in a muffle furnace to remove organic impurities.

Experimental settings

All imaging was performed on the miCube as described at 20°C. A 561 nm laser with ~ 0.12 W/cm^2 power output was used for HiLo-to-TIRF illumination with 4 ms stroboscopic illumination(Farooq and Hohlbein, 2015) in the middle of 10 ms frames. Low-power UV illumination (µW/cm^2 range) was used and increased during experiments to ensure a low and steady number of fluorophores in the sample until exhaustion of the fluorophores. A UV-increment scheme was consistently used for all experiments (Supplementary Table 5). No emission filter was used except for the

TIRF filter (Chroma ZET405/488/561m-TRF). The raw data was acquired using the open source Micro-Manager software(Edelstein et al., 2014). During acquisition, 2x2 binning was used, which resulted in a pixel size of 128x128 nm. The camera image was cropped to the central 512x512 pixels (65.64 x 65.64 μm) or smaller. For sptPALM experiments, frames 500-55,000 were used for analysis, corresponding to 5-550 seconds. This prevented attempted localization of overlapping fluorophores at the beginning, and ensured a set end-time. 200-300 brightfield images were recorded by illuminating the sample at the same position as during the measurement. For the brightfield recording, we used a commercial LED light (INREDA, IKEA, Sweden) and a home-made diffuser from weighing paper.

Localization

To extract single molecule localizations, a 50-frame temporal median filter (https://github.com/marcelocordeiro/medianfilter-imagej) was used to correct background intensity from the movies(Hoogendoorn et al., 2014). In short, the temporal median filter determines the median pixel value over a sliding-window of 50 pixels to determine the median background intensity value for a pixel at a specific position and time point. This value is subtracted from the original data, and any negative values are set to 0. In the process, all pixels are scaled according to the mean intensity of each frame to account for shifts in overall intensity. The first and last 25 frames from every batch of 8096 frames are removed in this process.

Single particle localization was performed via the ImageJ(Schneider et al., 2012)/Fiji(Schindelin et al., 2012) plugin ThunderSTORM(Ovesný et al., 2014) with added phasor-based single molecule localization algorithm (pSMLM(Martens et al., 2017)). Image filtering was done via a difference-of-Gaussians filter with Sigma1 = 2 px and Sigma2 = 8 px. The approximate localization of molecules was determined via a local maximum with a peak intensity threshold of 8, and 8-neighbourhood connectivity. Sub-pixel localization was done via phasor fitting(Martens et al., 2017) with a fit radius of 3 pixels (region-of-interest of 7-by-7 pixels). Custom-written MATLAB (The MathWorks, Natick, MA, USA) scripts were used to combine the output files from ThunderSTORM.

Cell segmentation

A cell-based segmentation on the localization positions was performed. First, a watershed was performed on the average of 300 brightfield-recorded frames of the cells. The watershed was done via the Interactive Watershed ImageJ plugin (http://imagej.net/Interactive_Watershed). Second, the localizations were filtered whether or not they fall in a pixel-accurate cell outline. If they do, a cell ID is added to the localization information.

Estimating the copy number of dCas9

The total copy number of dCas9 in a cell is not identical to the number of tracks found in each cell. Firstly, the UV illumination (405 nm wavelength) on the miCube required to photo-activate PAmCherry2 is not homogeneous over the complete field of view. To correct for this, a value for the average UV illumination experienced by each *L. lactis* cell is calculated. For this, a map of the UV intensity is made by placing a mirror on top of the objective and measuring the reflected scattering of the UV signal. Then, the mean UV intensity in the pixels corresponding to a cell according to the segmented brightfield images is stored. The cellular apparent dCas9 copy number is corrected for this normalized mean cellular UV intensity. Moreover, the cellular apparent dCas9 copy number was corrected for the average maturation grade of PAmCherry1, which is ~ 50%(Durisic et al., 2014) (shown schematically in Supplementary Fig. 9). We assume the maturation grades of PAmCherry1 and PAmCherry2 to be similar.

Tracking and fitting of apparent diffusion coefficient histogram

A tracking procedure was performed in MATLAB, using a modified Particle Point Analysis script(Crocker and Grier, 1996) (https://nl.mathworks.com/matlabcentral/fileexchange/42573-particle-point-analysis) with a tracking window of 8 pixels (1.0 μm) and no memory. Localizations corresponding to different cells were excluded from being part of the same track. As the tracking window is of similar size as the cells itself, in practice all localizations in a cell are linked together in a track if they appear in successive frames.

An apparent diffusion coefficient, D^*, was then calculated for each track from the mean-squared displacement (MSD) of single-step intervals(Stracy and Kapanidis, 2017). In short, for every track with at least 4 localizations, the D^* was calculated by calculating the mean square displacement between the first four steps and taking the average of that. Qualitative tracking information in cells (Fig. 2a, Fig. 3a) shows that diffusion coefficients up to ~ 4 $\mu m^2/s$ are obtained. These high diffusion coefficient tracks are caused by including false positive localizations in tracks, Therefore, tracks with a diffusion coefficient clearly caused by inclusion of false positive localizations ($D^* > 2.5$ $\mu m^2/s$) were excluded from further analysis; we binned the diffusion coefficients in 40 logarithmic-divided bins from $D^* = 0.01$ to $D^* = 2.5$ $\mu m^2/s$. The pNonTarget diffusional information was first corrected for the diffusion histogram obtained from a non-induced sample, subtracting the non-induced histogram from the pNonTarget histogram based on the approximated relative size of the non-induced histogram (~ 7.2 tracks per cell were found in non-induced cells).

Then, a Monte-Carlo diffusion distribution analysis (MC-DDA; described below) consisting of 20.000 dCas9 proteins was fitted via a general Levenberg-Marquardt fitting procedure in MATLAB. The error of this fit was determined via a general bootstrapping approach, where a D^*-list with the same length as the original, but randomly filled with values from the original (allowing for more than one entry of the same data), was fitted via the same procedure. For the pTarget diffusional information, the pNonTarget best fitted model (calculated via the same model, but with 100.000 dCas9 proteins) was fitted and smoothed via a Savitzky-Golay filter with order 3 and length 7, to reduce noise on the fit, alongside a single population following the following equation:

$$y = \frac{\left(\frac{n}{D_{plasmid}}\right)^n \cdot x^{(n-1)} \cdot e^{-n\frac{x}{D_{plasmid}}}}{(n-1)!} \quad (1)$$

Where $D_{plasmid}$ is the D^*-value corresponding to plasmid-bound dCas9, n the number of steps in the trajectory (set to four in this study), y the count of the histogram, and x the D^*-value of the histogram. $D_{plasmid}$ was kept constant in the global fit, while the size of this population and the size of the pNonTarget model were allowed to

vary between apparent cellular dCas9 copy number bins. Again, the error of this fit was determined via a general bootstrapping approach.

pNonTarget Monte-Carlo diffusion distribution analysis

The pNonTarget data is fitted with a Monte-Carlo diffusion distribution analysis (MC-DDA), in which a variable D_{free}, localization error, $k_{free\rightarrow bound}$, and $k_{bound\rightarrow free}$ need to be provided. A set number of dCas9 proteins are simulated (20.000 for the fit, 100.000 for visualising the fit). These proteins are then randomly placed in a cell, which is simulated as a cylinder with length 0.5 µm and radius 0.5 µm, capped by two half-spheres with radius 0.5 µm, and the current state of the proteins is set to free or immobile, based on the respective kinetic rates ($c_{bound} = k_{free\rightarrow bound} /(k_{bound\rightarrow free} + k_{free\rightarrow bound})$, $c_{free} = 1\text{-}c_{bound}$). Moreover, the proteins are given a time before they are changed between states (log(rand)/-k, where rand is a random number between 0 and 1, and k is the respective kinetic rate). Then, the movement of the proteins is simulated with over-sampling with regards to the frame time (0.1 ms). The free proteins will move a distance equal to a randomly sampled normal distribution with $\sigma = \sqrt{2 \cdot D_{free} \cdot steptime}$, where steptime is 0.1 ms. Then, it checked if this position is still within the cell outline. If not, a new location will be pulled from the distribution and checked against the cell outline. Every time-step, the time until state-change is subtracted with the time-step, and if this value becomes ≤ 0, the proteins will switch states, getting a new diffusion coefficient and state-change time. Every 10 ms after an initial equilibrium time of 200 ms, the current location of the proteins is convoluted with a random localization error, from a randomly sampled normal distribution with σ = localization error. The simulation is ended after 5 localization points are acquired for every protein. Further tracking and diffusion coefficient calculations are done the same as the experimental data.

Target simulation

For the target simulation, a certain number of dCas9 are simulated (similar to the average of the bins used in experiments), alongside a variable total DNA content (~ 7.5mln base pairs, or 1.5x double-stranded *L. lactis* genome(Linares et al., 2010)), plasmid copy number, target sites (5 per plasmid), incubation time (90 minutes),

fluorophore maturation time (20 minutes(Subach et al., 2009)), and a k_{off} rate. The dCas9 proteins are simulated one by one. The first dCas9 will have access to all target sites, and will be simulated for [incubation time], assuming the first dCas9 was made exactly at the start of the Nisin incubation. Subsequent dCas9 proteins will have access to fewer target sites, depending on whether or not earlier dCas9 proteins have ended the simulation bound to target sites. Subsequent dCas9 proteins will also be simulated for a shorter time, linearly scaling from [incubation time] to [fluorophore maturation time], which assumes that dCas9 proteins are steadily produced throughout the incubation time, but allowing for the fact that dCas9 proteins that do not yet have a matured PAmCherry2 are not visible during sptPALM.

Then, the dCas9 proteins randomly start in the free, PAM-probing, or target-bound state, based on the previously determined kinetic constants, similarly as in the pNonTarget simulation. The proteins are also given a time until state change, as was done in the pNonTarget simulation. Next, the simulation time of a single dCas9 protein was decreased by this time until state change, whereupon a new state was given to the protein: free proteins changed to PAM-probing or target-bound, with the target-bound chance being equal to $^{\text{nr target sites}}/_{\text{total nr of PAM sites}}$; PAM-probing or target-bound proteins were changed to free proteins. This was continued until the end of the simulation, after which the final state was determined. If the dCas9 was bound to a target, the available target sites were decreased by 1 for the next simulated dCas9. The reported values are the mean of 50 repetitions of the simulation, with the 95% confidence interval determined via the standard deviation of these repetitions.

For simulating Cas9 cleavage rates, it was assumed that a single target site was present and that a dCas9 would never be removed from a target site. By then analysing the ‘bound’ dCas9, it indicates whether the target site has been cleaved by Cas9. The other simulation parameters were kept constant.

miCube drift quantification

We characterised the positional stability of the miCube via super-resolution measurements of GATTA-PAINT 80R DNA-PAINT nanorulers (GATTAquant

GmbH, Germany). We imaged the nanorulers in total internal reflection (TIR) mode using a 561 nm laser (~ 7 mW) with a frame time of 50 ms using 2x2 pixel binning on the Andor Zyla 4.2 PLUS sCMOS. Astigmatism was enabled by placing a 1000 mm focal length astigmatic lens (Thorlabs) 51 mm away from the camera chip. A video of 10.000 frames was recorded via the MicroManager software(Edelstein et al., 2014).

After recording the movie, we first localized the *x, y,* and *z*-positions of the point spread functions of excited DNA-PAINT nanoruler fluorophores with the ThunderSTORM software(Ovesný et al., 2014) for ImageJ(Schneider et al., 2012) with the phasor-based single molecule localization (pSMLM) add-on(Martens et al., 2017). The ThunderSTORM software was used with the standard settings, and a 7 by 7 pixel region of interest around the approximate centre of the point spread functions was used for pSMLM. To determine the *z*-position, we compared the astigmatism of the point-spread function to a pre-recorded calibration curve recorded using immobilized fluorescent latex beads (560 nm emission peak, 50 nm diameter).

After data analysis we performed drift-correction in the lateral plane using the cross-correlation method of the ThunderSTORM software. The cross-correlation images were calculated using 10x magnified super-resolution images from a sub-stack of 100 original frames. The fit of the cross-correlation was used as drift of the lateral plane. Drift of the axial plane was analysed by taking the average *z*-position of all fluorophores, assuming that all DNA-PAINT nanorulers are fixed to the bottom of the glass slide.

Code and data availability

All code necessary to perform this study is made available as Supplementary Data, along with an accompanying programming flowchart. Raw data of part of the experiments is available upon reasonable request.

Acknowledgments

K.J.A.M. is funded by a VLAG PhD-fellowship grant awarded to J.H. J.H. acknowledges funding from the Innovation Program Microbiology Wageningen

(IPM-3). S.v.d.E is funded by the BE-Basic R&D program, which was granted a FES subsidy from the Dutch Ministry of Economic affairs. The authors thank the WOSM (Warwick Open Source Microscope, see www.wosmic.org) for inspiration.

Author contributions

K.J.A.M., S.B., and J.H. designed, built and characterised the miCube setup. K.J.A.M., S.P.B.v.B. and G.A.V recorded and analysed the experimental single molecule data. J.H., S.v.d.E and P.v.B. envisioned using *L. lactis*, dCas9, fluorescent proteins and p(Non-)Target cells to conduct super-resolution single molecule studies. S.P.B.v.B, S.v.d.E., P.v.B., and M.K. designed the DNA vectors used in this study. S.P.B.v.B. and S.v.d.E. assembled the DNA vectors and transformed cells. K.J.A.M, J.N.A.V., and J.H developed DDA. K.J.A.M. and J.N.A.V wrote software for data analysis. J.N.A.V and S.J.J.B. provided reagents and expertise for setting up the single molecule assays. K.J.A.M. and J.H. wrote the manuscript with input from all authors. J.H. initialised the study and the collaborations, and supervised all aspects of the study.

Competing interests

None to declare.

References

Almada, P., Culley, S., and Henriques, R. (2015). PALM and STORM: Into large fields and high-throughput microscopy with sCMOS detectors. Methods *88*, 109–121.

Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature *513*, 569–573.

Aristov, A., Lelandais, B., Rensen, E., and Zimmer, C. (2018). ZOLA-3D allows flexible 3D localization microscopy over an adjustable axial range. Nature Communications *9*, 2409.

Arsenault, A., Leith, J.S., Henkin, G., McFaul, C.M., Tarling, M., Talbot, R., Berard, D., Michaud, F., Scott, S., and Leslie, S.R. (2015). Open-frame system for single-molecule microscopy. Review of Scientific Instruments *86*, 033701.

van Asseldonk, M., Rutten, G., Oteman, M., Siezen, R.J., de Vos, W.M., and Simons, G. (1990). Cloning of usp45, a gene encoding a secreted protein from Lactococcus lactis subsp. lactis MG1363. Gene *95*, 155–160.

Auer, A., Schlichthaerle, T., Woehrstein, J., Schueder, F., Strauss, M., Grabmayr, H., and Jungmann, R. (2018). Nanometer-scale multiplexed super-resolution imaging with an economic 3D-DNA-PAINT microscope. ChemPhysChem.

Babcock, H.P. (2018). Multiplane and Spectrally-Resolved Single Molecule Localization Microscopy with Industrial Grade CMOS cameras. Scientific Reports *8*, 1726.

Beljouw, S.P.B. van, Els, S. van der, Martens, K.J.A., Kleerebezem, M., Bron, P.A., and Hohlbein, J. (2019). Evaluating single-particle tracking by photo-activation localization microscopy (sptPALM) in Lactococcus lactis. Phys. Biol. *16*.

Bonomo, M.E., and Deem, M.W. (2018). The physicist's guide to one of biotechnology's hottest new topics: CRISPR-Cas. Phys. Biol. *15*, 041002.

Campelo, A.B., Roces, C., Mohedano, M.L., López, P., Rodríguez, A., and Martínez, B. (2014). A bacteriocin gene cluster able to enhance plasmid maintenance in Lactococcus lactis. Microbial Cell Factories *13*, 77.

Chen, B., Gilbert, L.A., Cimini, B.A., Schnitzbauer, J., Zhang, W., Li, G.-W., Park, J., Blackburn, E.H., Weissman, J.S., Qi, L.S., et al. (2013). Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. Cell *155*, 1479–1491.

Coelho, S., Baek, J., Graus, M.S., Halstead, J.M., Nicovich, P.R., Feher, K., Gandhi, H., and Gaus, K. (2018). Single molecule localization microscopy with autonomous feedback loops for ultrahigh precision. BioRxiv 487728.

Crocker, J.C., and Grier, D.G. (1996). Methods of Digital Video Microscopy for Colloidal Studies. Journal of Colloid and Interface Science *179*, 298–310.

Diederich, B., Diederich, B., Then, P., Jügler, A., Forster, R., Forster, R., Heintzmann, R., and Heintzmann, R. (2018). cellSTORM — Super-Resolution on a

Cellphone. In Frontiers in Optics / Laser Science (2018), Paper FM4E.7, (Optical Society of America), p. FM4E.7.

Diekmann, R., Till, K., Müller, M., Simonis, M., Schüttpelz, M., and Huser, T. (2017). Characterization of an industry-grade CMOS camera well suited for single molecule localization microscopy–high performance super-resolution at low cost. Scientific Reports *7*, 14425.

Drlica, K., Malik, M., Kerns, R.J., and Zhao, X. (2008). Quinolone-Mediated Bacterial Death. Antimicrobial Agents and Chemotherapy *52*, 385–392.

Durisic, N., Laparra-Cuervo, L., Sandoval-Álvarez, Á., Borbely, J.S., and Lakadamyali, M. (2014). Single-molecule evaluation of fluorescent protein photoactivation efficiency using an in vivo nanotemplate. Nature Methods *11*, 156.

Edelstein, A.D., Tsuchida, M.A., Amodaj, N., Pinkard, H., Vale, R.D., and Stuurman, N. (2014). Advanced methods of microscope control using μManager software. Journal of Biological Methods *1*, e10.

Edward, J.T. (1970). Molecular volumes and the Stokes-Einstein equation. Journal of Chemical Education *47*, 261.

Els, S. van der, James, J.K., Kleerebezem, M., and Bron, P.A. (2018). Development of a versatile Cas9-driven subpopulation-selection toolbox in Lactococcus lactis. Appl. Environ. Microbiol. AEM.02752-17.

Farooq, S., and Hohlbein, J. (2015). Camera-based single-molecule FRET detection with improved time resolution. Physical Chemistry Chemical Physics *17*, 27862–27872.

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. PNAS *109*, E2579–E2586.

Globyte, V., Lee, S.H., Bae, T., Kim, J.-S., and Joo, C. (2018). CRISPR/Cas9 searches for a protospacer adjacent motif by lateral diffusion. The EMBO Journal e99466.

Goel, A., Santos, F., Vos, W.M. de, Teusink, B., and Molenaar, D. (2011). A standardized assay medium to measure enzyme activities of Lactococcus lactis while mimicking intracellular conditions. Appl. Environ. Microbiol. AEM.05276-11.

Hidalgo-Cantabrana, C., O'Flaherty, S., and Barrangou, R. (2017). CRISPR-based engineering of next-generation lactic acid bacteria. Current Opinion in Microbiology *37*, 79–87.

Holm, T., Klein, T., Löschberger, A., Klamp, T., Wiebusch, G., van de Linde, S., and Sauer, M. (2014). A Blueprint for Cost-Efficient Localization Microscopy. ChemPhysChem *15*, 651–654.

Hoogendoorn, E., Crosby, K.C., Leyton-Puig, D., Breedijk, R.M.P., Jalink, K., Gadella, T.W.J., and Postma, M. (2014). The fidelity of stochastic single-molecule super-resolution reconstructions critically depends upon robust background estimation. Scientific Reports *4*, 3854.

Huang, B., Wang, W., Bates, M., and Zhuang, X. (2008). Three-Dimensional Super-Resolution Imaging by Stochastic Optical Reconstruction Microscopy. Science *319*, 810–813.

Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L.A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nature Biotechnology *31*, 233–239.

Jones, D.L., Leroy, P., Unoson, C., Fange, D., Ćurić, V., Lawson, M.J., and Elf, J. (2017). Kinetics of dCas9 target search in Escherichia coli. Science *357*, 1420–1424.

Khan, S.A. (1997). Rolling-circle replication of bacterial plasmids. Microbiol. Mol. Biol. Rev. *61*, 442–455.

Knight, S.C., Xie, L., Deng, W., Guglielmi, B., Witkowsky, L.B., Bosanac, L., Zhang, E.T., Beheiry, M.E., Masson, J.-B., Dahan, M., et al. (2015). Dynamics of CRISPR-Cas9 genome interrogation in living cells. Science *350*, 823–826.

Komor, A.C., Badran, A.H., and Liu, D.R. (2017). CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes. Cell *168*, 20–36.

Kuipers, O.P., de Ruyter, P.G.G.A., Kleerebezem, M., and de Vos, W.M. (1998). Quorum sensing-controlled gene expression in lactic acid bacteria. Journal of Biotechnology *64*, 15–21.

Kwakwa, K., Savell, A., Davies, T., Munro, I., Parrinello, S., Purbhoo, M.A., Dunsby, C., Neil, M.A., and French, P.M. (2016). easySTORM: a robust, lower-cost approach to localisation and TIRF microscopy. Journal of Biophotonics *9*, 948–957.

Linares, D.M., Kok, J., and Poolman, B. (2010). Genome Sequences of Lactococcus lactis MG1363 (Revised) and NZ9000 and Comparative Physiological Studies. Journal of Bacteriology *192*, 5806–5812.

Liu, J.-J., Orlova, N., Oakes, B.L., Ma, E., Spinner, H.B., Baney, K.L.M., Chuck, J., Tan, D., Knott, G.J., Harrington, L.B., et al. (2019). CasX enzymes comprise a distinct family of RNA-guided genome editors. Nature 1.

Ma, H., Fu, R., Xu, J., and Liu, Y. (2017). A simple and cost-effective setup for super-resolution localization microscopy. Sci Rep *7*.

Machielsen, R., Siezen, R.J., Hijum, S.A.F.T. van, and Vlieg, J.E.T. van H. (2011). Molecular Description and Industrial Potential of Tn6098 Conjugative Transfer Conferring Alpha-Galactoside Metabolism in Lactococcus lactis. Appl. Environ. Microbiol. *77*, 555–563.

Manley, S., Gillette, J.M., Patterson, G.H., Shroff, H., Hess, H.F., Betzig, E., and Lippincott-Schwartz, J. (2008). High-density mapping of single-molecule trajectories with photoactivated localization microscopy. Nature Methods *5*, 155–157.

Martens, K.J.A., Bader, A.N., Baas, S., Rieger, B., and Hohlbein, J. (2017). Phasor based single-molecule localization microscopy in 3D (pSMLM-3D): An algorithm for MHz localization rates using standard CPUs. The Journal of Chemical Physics *148*, 123311.

Mierau, I., and Kleerebezem, M. (2005). 10 years of the nisin-controlled gene expression system (NICE) in Lactococcus lactis. Applied Microbiology and Biotechnology *68*, 705–717.

Nagai, T., Ibata, K., Park, E.S., Kubota, M., Mikoshiba, K., and Miyawaki, A. (2002). A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. Nature Biotechnology *20*, 87.

Nicovich, P.R., Walsh, J., Böcking, T., and Gaus, K. (2017). NicoLase—An open-source diode laser combiner, fiber launch, and sequencing controller for fluorescence microscopy. PLOS ONE *12*, e0173879.

Nishimasu, H., Cong, L., Yan, W.X., Ran, F.A., Zetsche, B., Li, Y., Kurabayashi, A., Ishitani, R., Zhang, F., and Nureki, O. (2015). Crystal structure of Staphylococcus aureus Cas9. Cell *162*, 1113–1126.

Ovesný, M., Křížek, P., Borkovec, J., Švindrych, Z., and Hagen, G.M. (2014). ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. Bioinformatics *30*, 2389–2390.

Prazeres, D.M.F. (2008). Prediction of diffusion coefficients of plasmids. Biotechnology and Bioengineering *99*, 1040–1044.

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell *152*, 1173–1183.

Rieger, B., and Stallinga, S. (2014). The Lateral and Axial Localization Uncertainty in Super-Resolution Light Microscopy. ChemPhysChem *15*, 664–670.

Santoso, Y., Joyce, C.M., Potapova, O., Le Reste, L., Hohlbein, J., Torella, J.P., Grindley, N.D., and Kapanidis, A.N. (2010a). Conformational transitions in DNA polymerase I revealed by single-molecule FRET. Proceedings of the National Academy of Sciences *107*, 715–720.

Santoso, Y., Torella, J.P., and Kapanidis, A.N. (2010b). Characterizing Single-Molecule FRET Dynamics with Probability Distribution Analysis. ChemPhysChem *11*, 2209–2219.

Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Res. *39*, 9275–9282.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nature Methods *9*, 676–682.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nature Methods *9*, 671.

Shen, H., Tauzin, L.J., Baiyasi, R., Wang, W., Moringo, N., Shuang, B., and Landes, C.F. (2017). Single Particle Tracking: From Theory to Biophysical Applications. Chem. Rev. *117*, 7331–7376.

Singh, D., Sternberg, S.H., Fei, J., Doudna, J.A., and Ha, T. (2016). Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. Nature Communications *7*, 12778.

Slutsky, M., and Mirny, L.A. (2004). Kinetics of Protein-DNA Interaction: Facilitated Target Location in Sequence-Dependent Potential. Biophysical Journal *87*, 4021–4035.

Smith, C.S., Joseph, N., Rieger, B., and Lidke, K.A. (2010). Fast, single-molecule localization that achieves theoretically minimum uncertainty. Nat Meth *7*, 373–375.

Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C., and Doudna, J.A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62–67.

Stracy, M., and Kapanidis, A.N. (2017). Single-molecule and super-resolution imaging of transcription in living bacteria. Methods *120*, 103–114.

Subach, F.V., Patterson, G.H., Manley, S., Gillette, J.M., Lippincott-Schwartz, J., and Verkhusha, V.V. (2009). Photoactivatable mCherry for high-resolution two-color fluorescence microscopy. Nat Methods *6*, 153–159.

Tal, S., and Paulsson, J. (2012). Evaluating quantitative methods for measuring plasmid copy numbers in single cells. Plasmid *67*, 167–173.

Trovato, F., and Tozzini, V. (2014). Diffusion within the Cytoplasm: A Mesoscale Model of Interacting Macromolecules. Biophysical Journal *107*, 2579–2591.

Uphoff, S., Reyes-Lamothe, R., Leon, F.G. de, Sherratt, D.J., and Kapanidis, A.N. (2013). Single-molecule DNA repair in live bacteria. PNAS *110*, 8063–8068.

Vincent, L., and Soille, P. (1991). Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Transactions on Pattern Analysis & Machine Intelligence 583–598.

Vink, J.N.A., Martens, K.J.A., Vlot, M., McKenzie, R.E., Almendros, C., Estrada Bonilla, B., Brocken, D.J.W., Hohlbein, J., and Brouns, S.J.J. In preparation: Direct visualization of CRISPR target search in live bacteria.

Vos, D., and M, W. (1987). Gene cloning and expression in lactic streptococci. FEMS Microbiol Rev *3*, 281–295.

Wells, J.M., Wilson, P.W., and Le Page, R.W.F. (1993). Improved cloning vectors and transformation procedure for Lactococcus lactis. Journal of Applied Bacteriology *74*, 629–636.

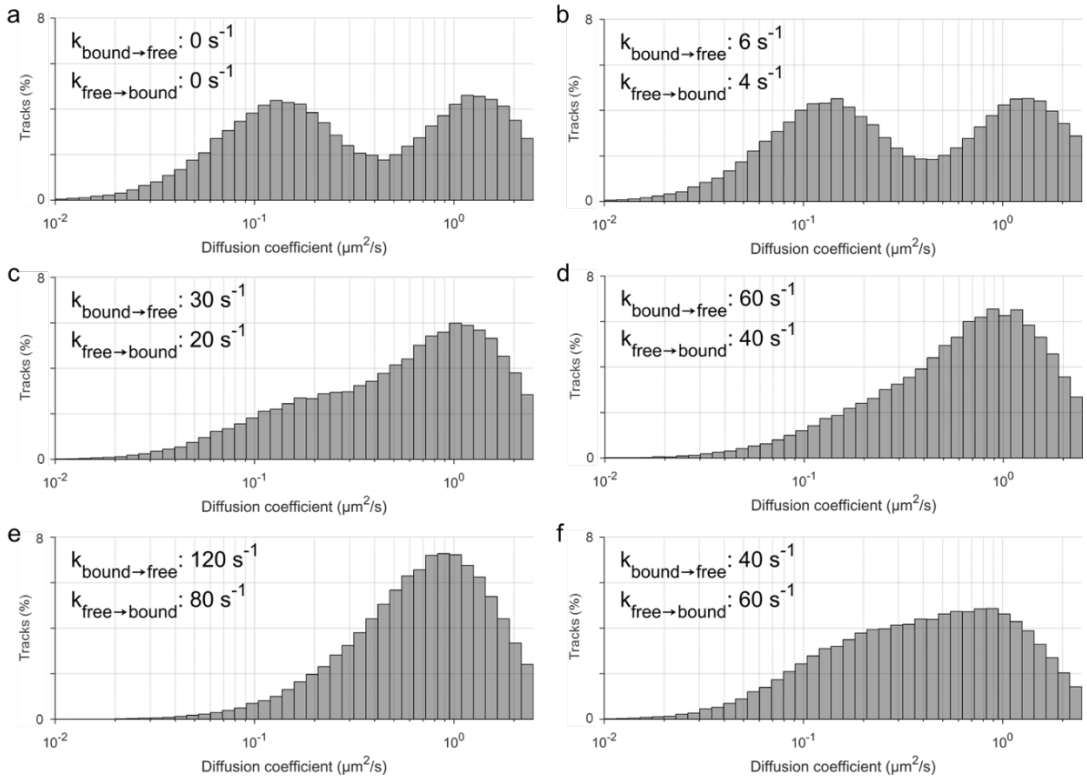
Whinn, K., Kaur, G., Lewis, J.S., Schauer, G., Müller, S., Jergic, S., Maynard, H., Gan, Z.Y., Naganbabu, M., Bruchez, M.P., et al. (2018). Nuclease dead Cas9 is a programmable roadblock for DNA replication. BioRxiv 455543.

Zhang, C., Wohlhueter, R., and Zhang, H. (2016). Genetically modified foods: A critical review of their promise and problems. Food Science and Human Wellness *5*, 116–123.

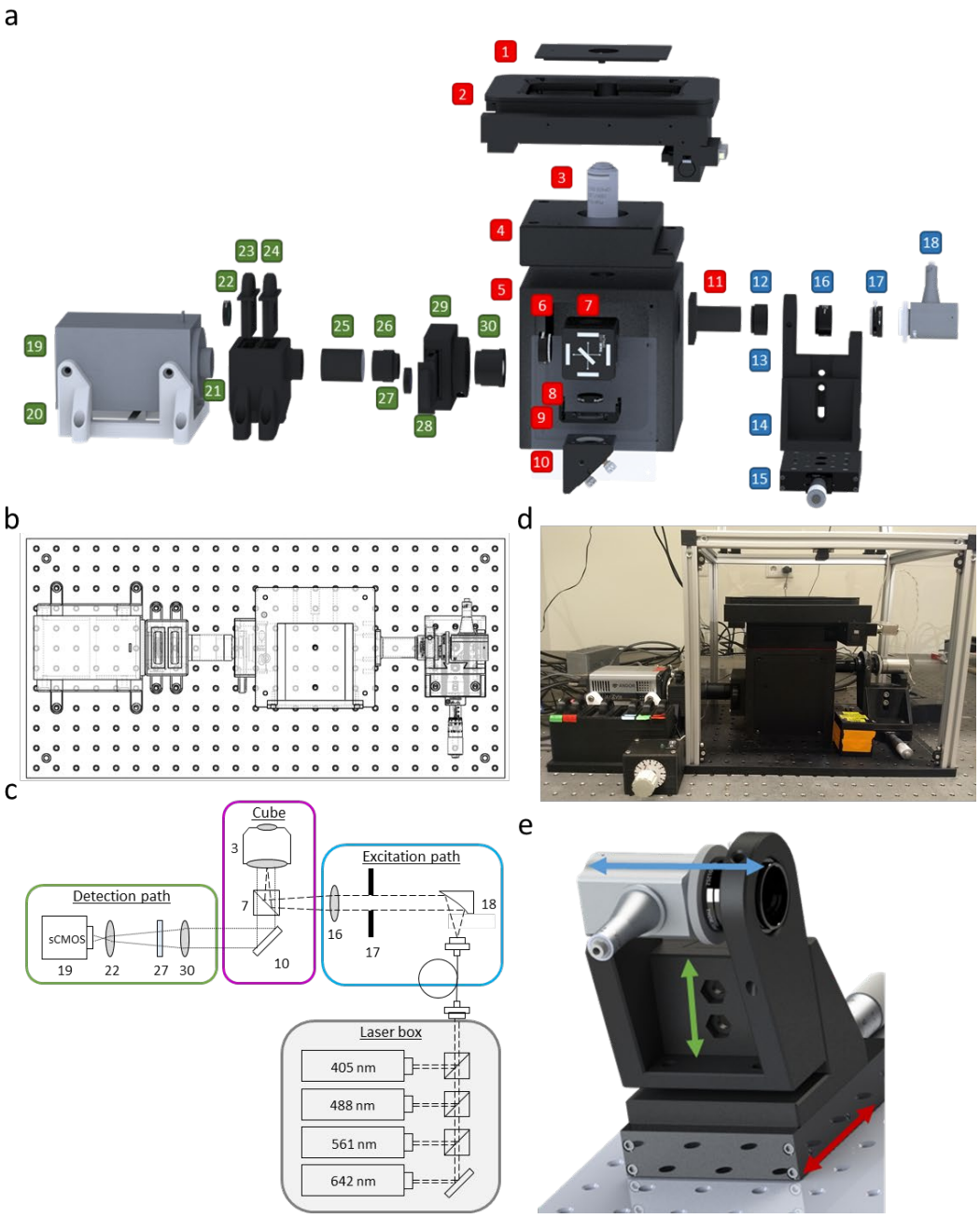
Zhang, Y.S., Ribas, J., Nadhman, A., Aleman, J., Selimović, Š., Lesher-Perez, S.C., Wang, T., Manoharan, V., Shin, S.-R., and Damilano, A. (2015). A cost-effective fluorescence mini-microscope for biomedical applications. Lab on a Chip *15*, 3661–3669.

(2016). A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation. PLOS Computational Biology *12*, e1004724.

Supplementary information

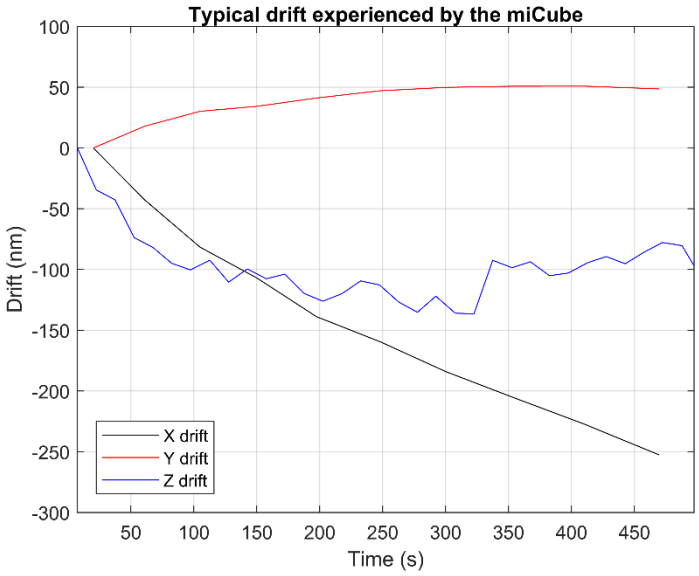


Supplementary Fig. 1 Effect of state transitions on diffusion coefficient histogram. The pNonTarget model as described in Methods was ran with varying $k_{bound→free}$ and $k_{free→bound}$ values as indicated in the figure, while keeping the localization error and D_{free} constant at the values determined while fitting the actual data (38 nm and 2.0 μm²/s, respectively). **a** Diffusion coefficient histogram if no state transitions would be present. **b to e** Diffusion coefficient histograms with the same $k_{bound→free}$: $k_{free→bound}$ ratio as the determined best-fitting values of ~ 3:2, while varying the absolute values of the two. **f** Diffusion coefficient histogram if the kinetic parameters were swapped.

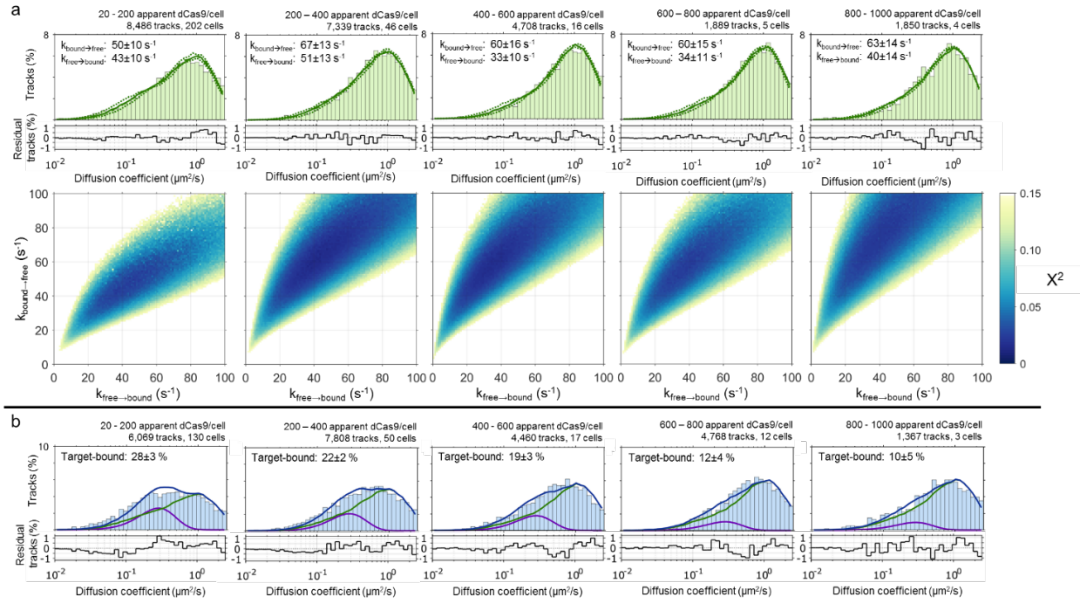


Supplementary Fig. 2 The open-source miCube single-particle microscope. a Exploded render of the miCube highlighting individual components. A full list of components indicated by the numbered items can be found in Supplementary Table 5. **b** Top-down schematic view of the miCube on the breadboard, allowing clear view of mounting positions. Distance between mounting holes on the breadboard is 25 mm. **c** Schematic overview of the miCube instrument. Numbered items correspond to the items in **a** and Supplementary Table 5. The excitation path is visualized with dashed lines, the emission path is visualized with dotted lines. **d** Photograph of the fully assembled miCube as used for measurements in this manuscript. **e** Detailed view of the miCube excitation path. This sub-assembly is comprised of numbers 12-18. Arrows indicate isolated movement in the

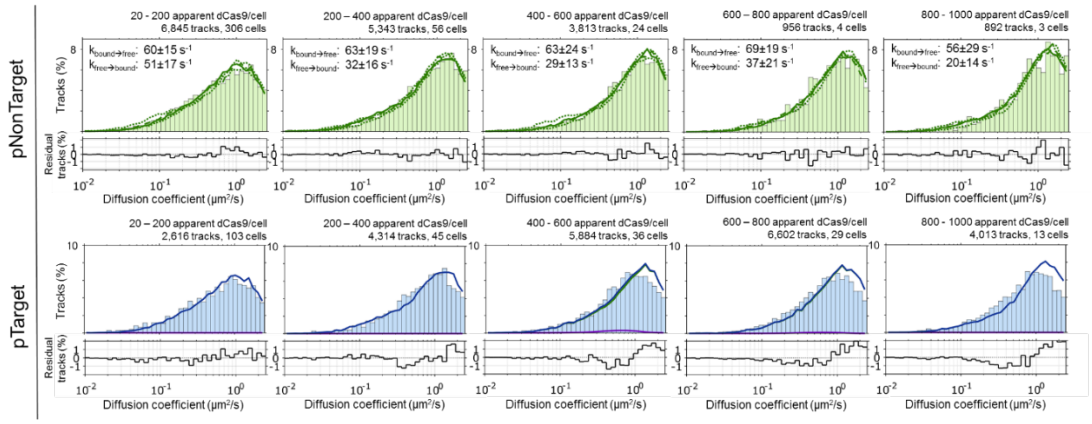
three spatial dimensions: distance from objective (blue), height of excitation unit (green), and horizontal position with respect to the objective (red).



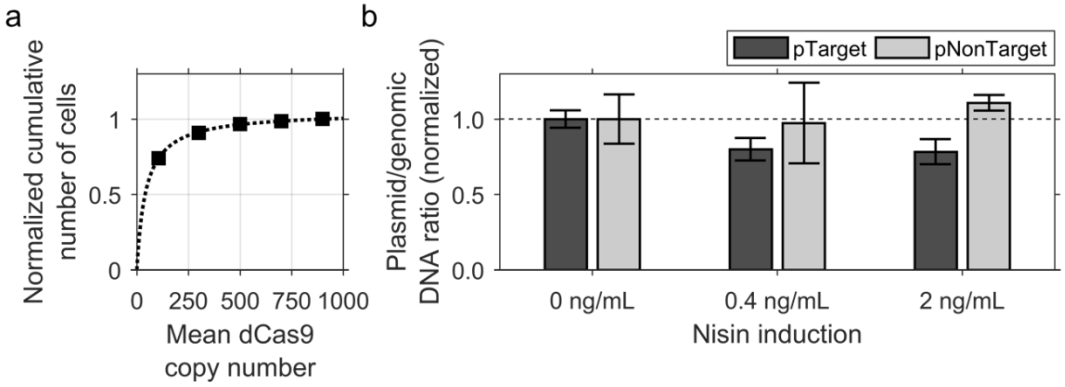
Supplementary Fig. 3 Typical drift experienced by the miCube. Typical drift in X (black), Y (red), and Z (blue) as experienced by the miCube used throughout this study. Repetition of this experiment led to the values specified in the main text.



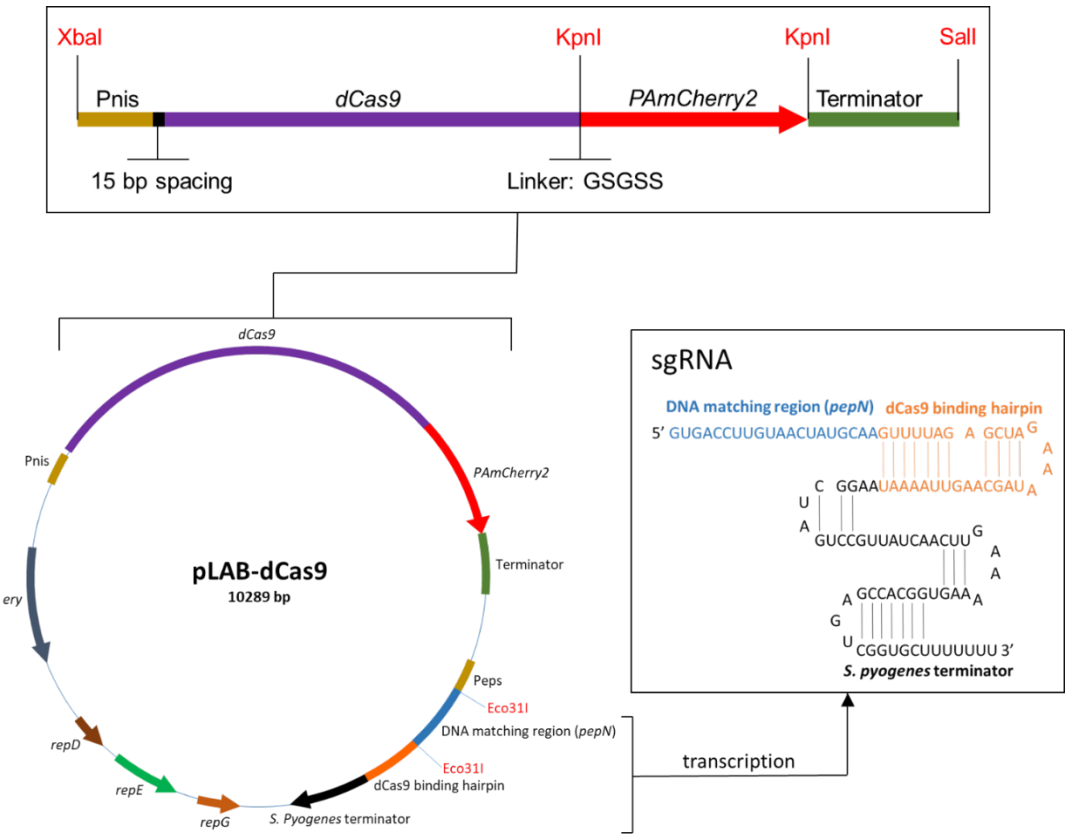
Supplementary Fig. 4 Individual pNonTarget and pTarget distributions. a All five pNonTarget diffusional distributions fitted with MC-DDA, as explained in the main text, Methods section, and Fig. 2. At the bottom, the Chi-squared value is plotted for a range of MC-DDAs (100k simulated proteins) with different $k_{free \rightarrow bound}$ and $k_{bound \rightarrow free}$. **b** All five pTarget diffusional distributions fitted with the computational target-binding model, as explained in the main text, Methods section, and Fig. 3.



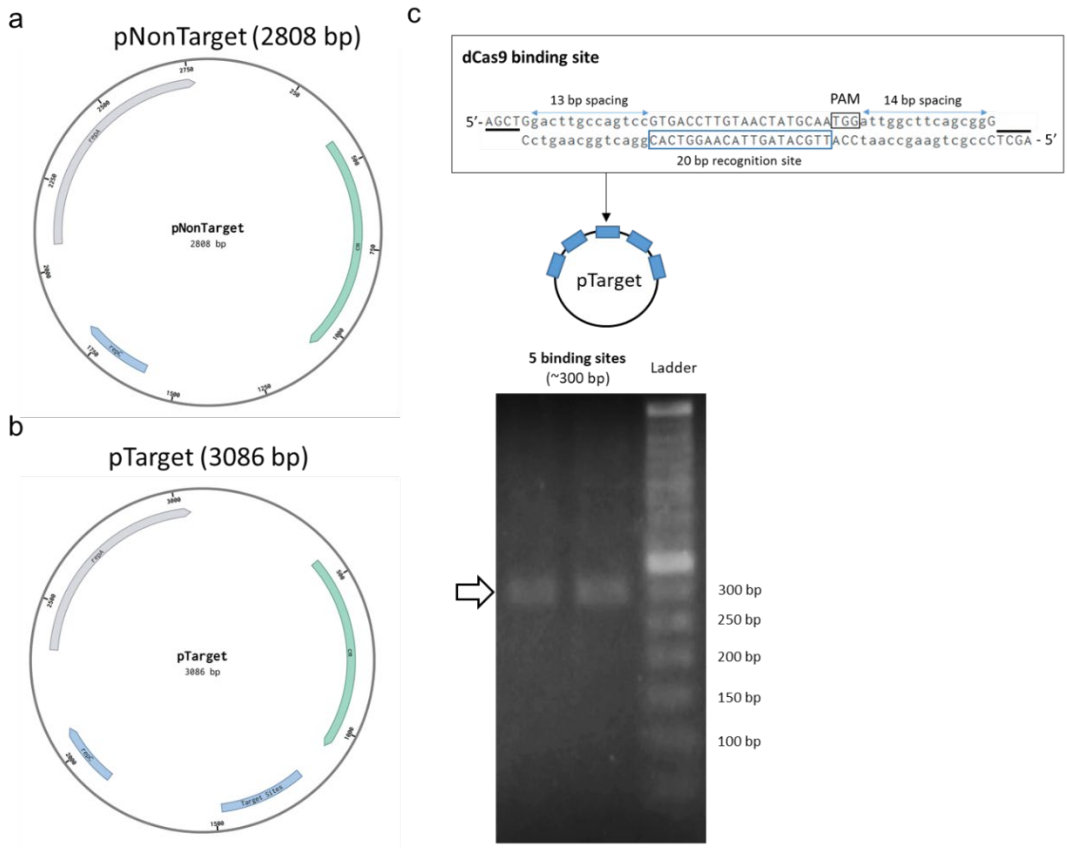
Supplementary Fig. 5 no-sgRNA distributions fitted with MC-DDA or the target-binding model. The fitting of diffusional data was performed identically as for samples with sgRNA. We note that no target-bound dCas9 diffusional data could be fitted with the same fitting algorithm as used for the sgRNA-pTarget diffusional data.



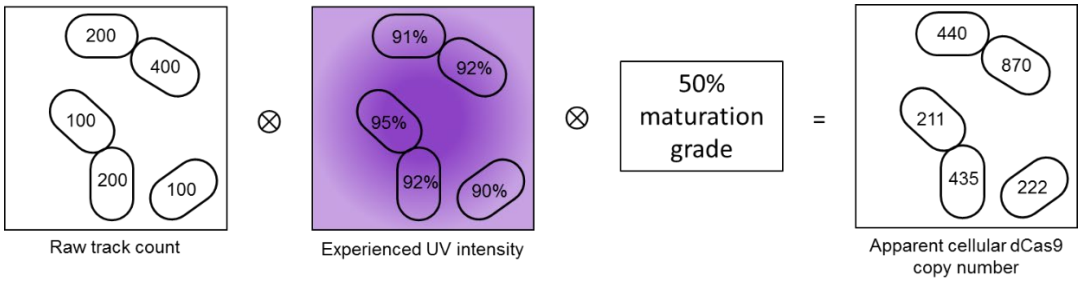
Supplementary Fig. 6 Effect of dCas9 on pTarget copy number. a Representative normalized cumulative number of cells that have certain mean dCas9 copy numbers. Black squares are values taken from pNonTarget dataset, the dotted line is a fitted curve with equation $1.05 \cdot [\text{dCas9 copy number}] / 44 \cdot [\text{dCas9 copy number}]$. **b** Normalized qPCR-determined ratio of plasmid:genome DNA for pTarget and pNonTarget for different Nisin induction. Error bars are the standard deviation determined from the average of two biological replicates (both averaged on two technical replicates).



Supplementary Fig. 7 Outline of the pLAB-dCas9 vector.Top: The sequence encoding dCas9(Qi et al., 2013) (*S. pyogenes*; AddGene.org plasmid #44249) is fused to the sequence encoding PAmCherry2(Subach et al., 2009) (AddGene.org plasmid #31932) with a flexible linker (amino acid sequence GSGSS), downstream of the *nisA*-promoter (Pnis) with an ribosomal binding site (15 bp spacing) and ending with a transcriptional terminator sequence derived from a lactococcal *pepN* gene. PAmCherry2 is flanked by two KpnI sites which should allow for interchanging fluorophores. The whole sequence is flanked by XbaI and SalI restriction sites to allow convenient cloning into a (expression) vector of choice. **Bottom:** The pLAB-dCas9 expression vector consists of PAmCherry2-labelled dCas9, an erythromycin resistance marker (Ery) and replication genes (*repD*, *repE* and *repG*)(Campelo et al., 2014). The *pepN* DNA matching region together with the dCas9 binding hairpin and the *S. pyogenes* terminator form the sgRNA, which is expressed under a constitutive promoter (Peps). Once the sgRNA molecule is transcribed, it folds to form the secondary structure that allows complex formation with dCas9.



Supplementary Fig. 8 pNonTarget and pTarget construction and verifiction. a,b Vector maps of pNonTarget and pTarget. Both targets contain *repA* and *repC* (DNA replication initiators) and a chloramphenicol-resistance marker (*cm*). Moreover, pTarget contains 5 target sites specified at 'Target Sites'. **c** dCas9 binding sites consisting of a 20 base pairs *pepN* recognition site, a 5'-NGG-3' PAM sequence, and spacing and overhang sequence motifs that are complementary to each other (indicated with black stripes) were annealed and ligated. This formed an array of five dCas9 binding sites in pNZ123, resulting in pTarget. Digestion and subsequent gel electrophoresis of plasmids isolated from two colonies revealed the expected length of the binding array in pTarget. One binding site is 54 base pairs in length, the final array of five binding sites is 278 base pairs (the expected PCR amplicon is 300 base pairs).



Supplementary Fig. 9 Schematic representation of obtaining cellular dCas9 copy number from number of tracks. The raw track count (first subfigure) is convoluted with the experience UV intensity that the cell on average experienced (second subfigure; deduced via reflective scattering of excitation lasers), and with the expected maturation grade of PAmCherry2 (Methods).

Supplementary Note: DNA and amino acid sequences

> dCas9-PAmCherry2 DNA sequence

ATGGACAAAAAATACAGCATCGGCCTTGCCATCGGCACCAACAGCGTCGGGTGGGCCGTCATCACTGACGAGTATAAAGTGCC
TAGTAAGAAGTTCAAGGTGCTAGGCAACACTGACCGACACAGCATTAAGAAGAACCTGATCGGTGCTCTGCTATTCGATAGTG
GTGAGACGGCCGAAGCTACAAGATTAAAAAGAACTGCAAGACGTAGATATACAAGACGTAAAAATCGTATTTGTTATTTACAG
GAAATTTTTAGTAATGAGATGGCTAAGGTTGATGATAGTTTCTTTCATAGATTAGAAGAATCATTTTTAGTAGAAGAAGATAA
AAAACATGAACGACATCCTATATTTGGAAATATAGTAGATGAAGTAGCTTATCATGAAAAATATCCTACTATTTATCATTTAC
GTAAAAAATTAGTTGATAGTACTGATAAAGCTGATTTAAGATTGATATATTTAGCATTAGCACACATGATTAAATTTCGTGGT
CATTTTCTAATTGAAGGAGATTTAAATCCTGATAACTCTGATGTTGATAAACTTTTTATTCAATTAGTCCAAACTTATAATCA
ATTATTTGAAGAAAATCCAATTAATGCTAGCGGTGTAGATGCTAAAGCTATTTTATCAGCTAGATTAAGTAAAAGCAGAAGAC
TAGAAAATTTAATTGCACAACTTCCTGGTGAGAAAAAGAATGGTTTATTTGGAAATTTGATTGCACTTAGTTTAGGTTTAACA
CCTAATTTTAAAAGTAATTTTGATTTAGCTGAAGATGCAAAACTTCAATTGAGTAAAGATACATATGATGATGATTTAGATAA
TTTATTAGCTCAAATTGGTGATCAATATGCAGATTTATTTTTAGCTGCCAAAAATTTATCTGATGCTATTTTATTAAGTGATA
TATTACGTGTAAATACTGAAATTACTAAAGCACCTTTATCAGCATCTATGATTAAAAGATATGATGAACATCATCAAGACTTA
ACATTATTAAAAGCATTAGTTAGACAACAATTACCAGAAAAATATAAAGAAATTTTCTTTGATCAATCAAAAAATGGATATGC
TGGATATATTGATGGTGGAGCATCACAAGAAGAATTTTATAAATTTATAAAACCTATTTTAGAAAAAATGGATGGAACTGAAG
AATTACTTGTTAAACTTAATAGAGAAGATTTATTAAGAAAACAAAGAACATTCGATAATGGATCAATCCCACATCAAATTCAT
TTAGGTGAATTGCATGCTATTTTACGTAGACAAGAAGATTTTTATCCATTCTTGAAAGATAATAGAGAAAAAATTGAAAAAAT
TTTAACTTTTAGAATTCCATATTATGTAGGACCTTTAGCACGAGGTAATTCTCGATTTGCATGGATGACACGTAAATCTGAAG
AAACAATTACACCATGGAATTTTGAAGAAGTTGTTGATAAAGGTGCTAGTGCACAATCTTTTATTGAAAGAATGACTAATTTT
GATAAAAATTTACCTAATGAAAAAGTATTACCAAAACATTCTTTATTATATGAATATTTTACTGTTTATAATGAACTTACAAA
AGTAAAATATGTTACTGAAGGAATGAGAAAACCAGCATTTTTATCAGGTGAACAAAAGAAAGCAATAGTTGATTTGTTATTTA
AAACAAATCGTAAAGTTACTGTTAAACAACTTAAAGAAGATTATTTTAAGAAAATTGAATGTTTTGATAGTGTTGAAATTTCT
GGAGTTGAAGATAGATTTAATGCTAGTTTAGGTACATATCATGATTTATTAAAAATTATTAAAGATAAAGATTTTCTTGATAA
TGAAGAAAATGAAGATATTTTAGAAGATATTGTTTTAACATTAACATTATTTGAAGATCGTGAAATGATTGAAGAACGTTTAA
AAACATATGCACATTTATTTGATGATAAAGTAATGAAACAATTAAAAAGACGTAGATATACTGGATGGGGACGTTTATCTCGT
AAATTAATTAATGGTATTAGAGATAAACAATCTGGTAAAACAATTTTAGACTTTCTAAAATCTGATGGATTCGCTAATCGTAA
TTTTATGCAATTAATTCATGATGATTCATTAACTTTTAAAGAAGATATTCAAAAAGCTCAGGTTAGTGGTCAAGGTGATAGCC
TTCATGAACATATAGCTAACCTAGCTGGTAGTCCAGCAATTAAAAAAGGTATTTTGCAAACAGTGAAAGTAGTTGATGAACTT
GTTAAAGTTATGGGTCGTCATAAACCTGAAAACATTGTTATTGAAATGGCACGAGAAAATCAAACTACACAAAAAGGACAAAA
GAATTCACGTGAACGTATGAAACGTATTGAAGAAGGTATTAAAGAACTAGGTAGTCAAATTCTTAAAGAACATCCAGTTGAAA
ATACACAATTACAAAATGAGAAATTATATTTATATTATTTACAAAATGGTCGTGATATGTATGTTGATCAAGAATTAGATATA
AATCGCTTGTCAGATTATGATGTAGATGCAATTGTTCCTCAATCATTTTTGAAAGATGATTCAATTGATAATAAAGTTTTGAC
ACGTAGTGATAAAAATCGTGGTAAAAGTGATAATGTTCCTAGTGAAGAAGTTGTCAAGAAAATGAAAAATTATTGGAGACAAT
TACTTAATGCTAAATTAATTACTCAACGTAAATTTGATAATTTAACAAAAGCAGAACGGGGAGGATTAAGTGAACTTGATAAA
GCTGGTTTTATAAAACGTCAATTAGTTGAAACAAGACAAATTACTAAACATGTAGCTCAAATATTAGATTCGCGTATGAATAC
TAAATATGATGAAAATGATAAATTAATTAGAGAAGTTAAAGTTATAACATTAAAATCTAAATTAGTTAGTGATTTTAGAAAAG
ATTTTCAATTTTATAAAGTTCGTGAAATAAATAATTATCATCATGCTCATGATGCCTATCTTAATGCAGTAGTTGGAACAGCT
TTAATTAAAAAATATCCAAAACTTGAAAGTGAATTTGTTTATGGTGATTATAAAGTCTATGATGTTCGCAAAATGATTGCTAA
ATCTGAACAAGAAATTGGTAAAGCTACAGCTAAATATTTCTTTTATAGTAATATTATGAATTTCTTTAAAACTGAAATTACTT
TAGCAAATGGAGAAATTAGAAAAAGACCATTAATTGAAACTAATGGTGAAACTGGAGAAATTGTTTGGGATAAAGGAAGAGAC
TTTGCAACAGTACGTAAAGTGTTATCTATGCCTCAAGTAAATATAGTTAAGAAAACTGAAGTTCAAACAGGCGGATTTAGTAA
AGAATCTATCTTACCAAAAAGAAATAGTGATAAATTAATTGCTCGTAAGAAAGATTGGGACCCTAAAAAATATGGTGGTTTTG
ATTCTCCAACTGTCGCTTATTCGGTCTTAGTTGTTGCTAAAGTAGAAAAAGGTAAAAGTAAAAAATTAAAATCAGTTAAAGAA
TTGTTAGGTATTACTATTATGGAAAGAAGTTCATTTGAAAAGAATCCTATTGACTTTTTAGAAGCCAAAGGTTACAAAGAGGT
CAAGAAAGACCTGATCATCAAACTGCCAAAGTACTCTCTCTTTGAATTAGAAAATGGACGTAAAAGAATGTTAGCATCTGCTG
GTGAATTGCAAAAAGGAAATGAATTAGCATTACCTAGTAAATATGTAAATTTCTTATACTTAGCTTCTCATTATGAAAAATTA
AAAGGTTCACCGGAGGACAACGAGCAGAAGCAACTTTTCGTGGAGCAACATAAACACTACCTCGACGAGATAATCGAACAAAT

TAGCGAGTTTTCAAAACGCGTCATCCTTGCCGATGCCAATTTAGATAAAGTTTTATCAGCTTATAATAAACATAGAGATAAAC
CTATTAGAGAACAAGCTGAAAATATTATTCATTTATTTACTTTAACTAATTTAGGTGCACCAGCTGCATTTAAATATTTCGAT
ACAACAATTGATCGAAAAAGATATACATCAACTAAAGAAGTTTTAGATGCAACATTAATACATCAATCAATTACAGGATTATA
TGAAACACGTATTGATTTATCTCAATTAGGTGGTGATGGATCCGGAAGTTCAGCTATTATTAAAGAATTTATGCGTTTTAAAG
TTCATTTAGAAGGTAGCGTTAATGGTCATGAATTTGAAATTGAAGGAGAAGGTGAAGGTAGACCATATGAAGGTACACAAACA
GCTAAATTAAAAGTTACAAAAGGTGGTCCATTACCTTTTGCTTGGGATATTTTGTCACCACAATTTATGTATGGTTCAAATGC
TTATGTTAAACATCCAGCTGATATTCCAGATTATTTTAAATTATCATTTCCAGAGGGTTTTAAATGGGAAAGAGTTATGAATT
TTGAAGATGGTGGTGTTGTAACAGTTACACAAGATTCATCTTTACAAGATGGTGAATTTATTTATAAAGTTAAATTAAGAGGT
ACTAATTTCCCTAGTGACGGACCGGTGATGCAAAAGAAAACCATGGGATGGGAGACATTAAGTGAACGTATGTATCCTGAAGA
TGGTGCGTTGAAAGGTGAGCTGAAAGCTAGAACTAAATTGAAAGATGGAGGCCACTATGATACTGAAGTAAAAACAACGTATA
AGGCTAAGAAACCCGTTCAGTTACCAGGGGCATATAACGTTAATCGTAAACTAGACATAACCTCTCATAATGAGGATTACACG
ATAGTTGAGCAATATGAACGAGCTGAAGGGCTTCATAGCACAGGTGGAATGGATGAACTTTATAAAtaa

> dCas9-PAmCherry2 amino acid translation

MDKKYSIGLAIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTARRRYTRRKNRICYLQ
EIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRG
HFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGLT
PNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNTEITKAPLSASMIKRYDEHHQDL
TLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIH
LGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFIERMTNF
DKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEIS
GVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSR
KLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDEL
VKVMGRHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMYVDQELDI
NRLSDYDVDAIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDK
AGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGTA
LIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRD
FATVRKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKE
LLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALPSKYVNFLYLASHYEKL
KGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFD
TTIDRKRYTSTKEVLDATLIHQSITGLYETRIDLSQLGGDGSGSSAIIKEFMRFKVHLEGSVNGHEFEIEGEGEGRPYEGTQT
AKLKVTKGGPLPFAWDILSPQFMYGSNAYVKHPADIPDYFKLSFPEGFKWERVMNFEDGGVVTVTQDSSLQDGEFIYKVKLRG
TNFPSDGPVMQKKTMGWETLSERMYPEDGALKGELKARTKLKDGGHYDTEVKTTYKAKKPVQLPGAYNVNRKLDITSHNEDYT
IVEQYERAEGLHSTGGMDELYK*

The asterisk (*) represents the stop codon ‘taa’.

Supplementary Table 1: Descriptive list of components miCube.

Numbers are in accordance with Supplementary Fig. 1. Entities marked with 'custom-built' have their complete technical drawings present in the Appendix. A more exhaustive list can be found on https://HohlbeinLab.github.io/miCube/component_table.html.

Main cube			
Nr	**Description**	**Details**	**Manufacturer**
1	Glass plate insert		Custom built – CNC milled
2	ASI XYZ stage	MS-2000 stage with Piezoconcept Z-insert	Applied Scientific Instrumentation, Eugene, OR, USA; and Piezoconcept, Lyon, France
3	Objective	TIRF 1.49NA HP SR objective	Nikon, Amsterdam, The Netherlands
4	TopCover		Custom built – CNC milled
5	miCube block		Custom built – CNC milled
6	Neutral density filter	NE60A-A	Thorlabs GmbH, Dachau/Munich, Germany
7	Dichroic mirror holder	DFM1/M	Thorlabs GmbH, Dachau/Munich, Germany
	Dichroic mirror	ZT405/488/561rpc-UF2 or ZT405/488/561/640rpc-UF2	Chroma, Bellows Falls, VT, USA

8	TIRF filter	ZET405/488/561m-TRF or ZET405/488/561/640m-TRF	Chroma, Bellows Falls, VT, USA
9	Connector	C4W-CC	Thorlabs GmbH, Dachau/Munich, Germany
10	45˚ elliptical mirror	KCB1E/M and BBE1-E02	Thorlabs GmbH, Dachau/Munich, Germany
11	Cover		Custom built – 3D printed

Excitation path			
Nr	**Description**	**Details**	**Manufacturer**
12	Spacer		Custom built – 3D printed
13	Reflective collimator holder		Custom built – 3D printed
14	Right-angle mounting plate		Custom built – 3D printed
15	25 mm Translation Stage	PT1/M	Thorlabs GmbH, Dachau/Munich, Germany
16	TIRF lens	AC254-200-A-ML	Thorlabs GmbH, Dachau/Munich, Germany
17	Aperture	SM1D12SZ	Thorlabs GmbH, Dachau/Munich, Germany
18	Reflective collimator	RC12APC-F01	Thorlabs GmbH, Dachau/Munich, Germany

-	Laser box	Lighthub 6 containing 4 lasers at 405 nm (60 mW), 488 nm (200 mW), 561 nm (500 mW), and 642 nm (2x 200 mW).	Omicron, Rodgau-Dudenhofen, Germany

Emission path			
Nr	**Description**	**Details**	**Manufacturer**
19	Camera	Zyla 4.2 PLUS	Andor, Belfast, Northern Ireland
20	Camera Mount		Custom built – 3D printed
21	Astigmatism block		Custom built – 3D printed
22	Astigmatism lens	LJ1516RM or LJ1144RM	Thorlabs GmbH, Dachau/Munich, Germany
23	Astigmatism lens holder		Custom built – 3D printed
24	Astigmatism lens holder		Custom built – 3D printed
25	Cover	SC600, cut to length	Thorlabs GmbH, Dachau/Munich, Germany
26	Connector		Custom built – 3D printed
27	Emission filter	ET525/50m or ET595/50m or ET700/75m	Chroma, Bellows Falls, VT, USA
28	Emission filter holder		Custom built – 3D printed
29	Tube lens holder		Custom built – 3D printed
30	Tube lens	ITL200	Thorlabs GmbH, Dachau/Munich, Germany

3D printing (Ultimaker 2+) settings							
Nr	**Material**	**Support**	**Adhesion**	**Layer height (µm)**	**Top / Bottom**	**Wall Thickness (mm)**	**Infill %**

					Thickness (mm)		
13/14	PLA	Yes	Brim (5)	200	1.0	1.2	20
20/26	ABS	Yes	Brim (5)	200	0.8	0.8	20
21/23/24/28	PLA	Yes	Brim (5)	200	1.0	1.6	20

Supplementary Table 2: List of vectors

Vector	Relevant properties	Size (kb)	Selection marker	Reference
pNZ123	Vector replicating in *L. lactis.*	2.8	Chloramphenicol	de Vos, 1987
pLABTarget	Encoding functional Cas9 expression system	10	Erythromycin	van der Els et al., 2018
pNonTarget	pNZ123 without binding sites	2.8	Chloramphenicol	This study
pTarget	pNZ123 containing five binding sites cognate to *pepN* sgRNA	3.1	Chloramphenicol	This study
pLAB-dCas9-nosgRNA	Cas9 module of pLABTarget replaced with Pnis – dCas9 – PAmCherry2	10.2	Erythromycin	This study
pLAB-dCas9	pLAB-dCas9-nosgRNA with added *pepN* sgRNA under constitutive promotor	10.3	Erythromycin	This study

Supplementary Table 3: List of oligonucleotides.

Experiment	Construct	Oligonucleotide (5'- 3')
pTarget construction	Forms *pepN* dCas9 binding site with complementary overhangs upon annealing.	AGCTGGACTTGCCAGTCCGTGACCTTGTAACTATGCAATG GATTGGCTTCAGCGG
		AGCTCCCGCTGAAGCCAATCCATTGCATAGTTACAAGGT CACGGACTGGCAAGTCC
pLAB-dCas9 construction	Forms *pepN* sgRNA upon annealing.	TGATGTGACCTTGTAACTATGCAA
		AAACTTGCATAGTTACAAGGTCAC

Supplementary Table 4: List of primers.

Experiment	Construct	Primer sequence (F=forward, R=reverse)
PCR insert validations and colony PCR	For pNZ123 (to check *pepN* binding site insertion)	F: TGAGATAATGCCGACTGTAC R: CATTCAGTCATCGGCTTTCA
	For pLAB-dCas9 (to check *pepN* sgRNA insertion)	F: TGATGTGACCTTGTAACTATGCAA R: TTGAAGAACCCGATTACATGG
qPCR	Q1, Q2: pTarget/pNonTarget	F: ACGAAAGTCGACGGCAATAGTT R: CGTTTGTTGAACTAATGGGTGC
	Q3, Q4: Nested pNonTarget	F: GGGAGCGGAGTTTGGAATTT R: ATAACCTAACTCTCCGTCGC
	Q5, Q6: ColonyCount	F: TCGATATGCACGTTGTCACC R: CCCTCTCAGCTGCAATCTCT
	Q7, Q8: Nested qPCR Colony	F: GTGCTGAACCAGCGATTACA R: TTGCTTTCACGTCAAGTTGG

Supplementary Table 5: Adjustment of the 405 nm laser power during sptPALM experiments.

Time (s)	Approximate 405 nm laser power for photo activation (μW/cm^2)
0	2.7
30	2.8
50	3.2
70	3.8
100	4.7
130	6.2
150	8.9
180	13.7
200	20.1
220	27.6
240	36.2
260	46.1
280	62.1
300	78.1
320	83.8
340	97.0
360	113.9
380	126.3
400	146.2
430	194.4
450	260.7
470	320.9
490	383.6
510	455.4
530	508.4
540	619.0

4

Comparing the target search mechanisms of native type I-E and type I-F CRISPR systems in live E. coli

Manuscript in preparation

Abstract

CRISPR-Cas systems confer adaptive immunity in prokaryotes and feature a large diversity of subtypes. Currently, it is not known, how this diversity translates into differences in the kinetics of target search and under which conditions one subtype could be more advantageous than another for the survival of the cell. In this study, we compare the target search efficiency and dynamics of two native CRISPR-Cas systems found in *E. coli*, type I-E and type I-F. We show that their interference efficiencies, which represent the number of effector complexes required to clear an invader, are very similar (20 complexes offering 50% protection for I-E, 25 complexes for I-F). We found that the free diffusion of the I-F Cascade complex is considerably faster than for the I-E Cascade complex (~3.5 $\mu m^2/s$ versus 1 $\mu m^2/s$), making it experimentally difficult to differentiate the I-F complex diffusion from the free I-F Cas8f subunit under the same conditions. We could show, however, that the DNA binding interactions of I-F Cas8f are slower compared to the I-E Cas8e subunit, which could be a result of (fast) PAM-independent interactions in the case of the I-E subunit. Furthermore, we reveal that *in vivo* target binding still occurs with many PAM-distal mutations, but is abolished with PAM-proximal mutations for Cascade I-F. This study is a first step in unraveling the different ways in which CRISPR-Cas systems balance the tradeoffs of the target search process that are crucial to successful CRISPR immunity

Introduction

The CRISPR system is the only known adaptive immune system in prokaryotes and can be found in a large percentage of bacterial and archaeal genomes (Koonin et al., 2017). CRISPR-Cas systems are currently classified into 6 main types and 33 subtypes (Makarova et al., 2020). This rich diversity represents differences in the adaptation machinery (either containing or missing Cas4), the surveillance complexes (either composed of a single or of multiple complexes), and the target molecule (DNA or RNA).

Approximately 20% of completely sequenced genomes with a CRISPR-Cas system contain multiple system types (Pourcel et al., 2020). This co-occurrence in a single organism could have several reasons. First, since immune systems are shared via horizontal gene transfer at a high rate relative to other genes (Puigbò et al., 2017), redundant systems in bacterial genomes could accumulate in the absence of natural selection. Second, natural selection could favor the co-occurrence of CRISPR-Cas systems, if they provide benefits to host survival.

One way of providing a benefit would be direct cooperation. Examples among CRISPR-Cas systems were demonstrated in *Pyrococcus furiosus* (Subtype I-A, I-B and III-B) (Majumdar et al., 2015), *Marinomonas mediterranea* (Subtype I-F and III-B) (Silas et al., 2017) and *Flavobacterium columnare* (Subtype II-C and V-A) (Hoikkala et al., 2020). In these examples, either crRNAs were shared between surveillance complexes or adaptation machineries were shared between different CRISPR array types. In all cases, it allows the host to prevent phage escape via type-specific anti-CRISPRs or PAM mutations.

A second benefit that multiple systems could provide is if each CRISPR-Cas system type is specialized in a certain type of invader. It was shown that some CRISPR types have preferential targets. The type IV system, for example, was found to mainly target plasmids (Pinilla-Redondo et al., 2019) and the RNA-targeting Type III systems was found to target jumbo phages that evade DNA-targeting systems (Malone et al., 2020; Mendoza et al., 2020). Specialization can occur at the adaptation stage, where acquisition machineries would specialize in recognizing and preferentially incorporating a certain feature of the invader, such as demonstrated in I-E Cas1-2, with preferential acquisition spacers from replication forks (Levy et al., 2015). In some systems, Cas1 is naturally fused to a reverse transcriptase which allows these systems to incorporate spacers from RNA molecules (Mohr et al., 2018; Silas et al., 2016; Toro et al., 2019). These systems can therefore specialize towards RNA invaders or invaders that lead to a large expression of RNA inside the cell.

Another way of specialization can occur at the interference stage at which the target search of surveillance complexes plays a crucial role in the interference efficiency (Vink et al., 2020a). Recent single-molecule studies revealed kinetic details on the

target search mechanism of CRISPR systems *in vivo* (Jones et al., 2017; Knight et al., 2015; Martens et al., 2019; Vink et al., 2020a). Surveillance complexes of DNA-targeting CRISPR systems spend most of their time scanning the PAMs (protospacer adjacent motif) that occur throughout the host chromosome. This oligonucleotide motif differs between CRISPR types, whereas some are short (CC; *Pseudomonas* I-F) or promiscuous (AWG; *Escherichia* I-E), others can be long (NAAAAY; *Treponema* II-C) (Gasiunas et al., 2020; Leenay et al., 2016). More stringent PAMs would require fewer strand openings and would therefore potentially find their target faster, but more permissive PAMs would allow for more functional spacers to be incorporated and reduce the chances of invader escape via point mutations. This trade-off raises the question whether CRISPR systems have different target search kinetics and whether this could allow for subtypes to specialize towards certain targets.

In this study, we report the target search kinetics of the native CRISPR I-F system of *E. coli* Ed1A, which contains a smaller surveillance complex (350 kDa vs 405 kDa) and shorter dinucleotide PAM (CC) than the PAM (AAG) of the previously studied I-E system of *E. coli* K12 (Vink et al., 2020a). The target efficiency, expressed as the number of complexes required to find a target within the time of replication, were found to be comparable. We then measured the DNA probing speeds by following single I-F Cascade complexes with single particle tracking photo-activation light microscopy and found similarities and differences between I-E and I-F target search kinetics for both free Cas8 subunits as full complexes (Manley et al., 2008). We furthermore investigated mismatch tolerance for the I-F Cascade complex and found a high tolerance for PAM-distal mutated segments, but complete abolishment of binding with mutations in PAM-proximal segments, which contrasts *in vitro* observations on I-E Cascade with similarly mutated targets. This work demonstrates that the diversity of CRISPR systems not only occurs on the genetic or molecular level, but that there is also a diversity in target search kinetics between these systems arising from these molecular differences and this might indicate that each subtype performs best under a different set of conditions in the host and invader.

Materials and Methods

Cloning

The insert to create pTarget was purchased as synthetic constructs from Gen9 (pTarget insert; Table S3). To increase the copy number of DNA targets in the cell, the constructs were cloned into a pUC19 backbone with XbaI and KpnI restriction sites. Subsequently, pTarget variants were created in which several regions of each protospacer regions were systematically mutated. To maximize the effect of mutated regions on target binding we based the nucleotide mutation rules on Künne *et al.,* 2018, where it was found that guanine – adenosine base mismatches clash thereby significantly reducing the binding of the Cascade complex. In the mutated sections of the array, every thymine base was changed to a guanine base to induce a G – A clash. Every cytosine base was changed to an adenosine base to induce a A – G clash. In addition, every adenosine was changed to a thymine and every guanine was changed to a cytosine for further mismatching. Since every sixth nucleotide in the target site is not involved in Cascade complex target binding (Jung et al., 2017), the 32 to 33 base pair targets were mutated in segments of 5 nucleotide (except for the last segment of two to three base pairs) in a scheme from (Blosser et al., 2015) (Figure 9). The mutated arrays were ordered as genomic blocks from Twist Bioscience, with XbaI and KpnI sites included on the ends. The backbone of the mutant arrays was amplified through PCR from pTarget using primers BG5825 and BG7848. Subsequently the amplified backbones and ordered constructs were restricted and ligated.

Cas2/3 was amplified from the *E.coli Ed1A* genome using primers BN1387 and BN1386 which added restriction sites SpeI and SphI. The backbone was amplified from an existing pCas3 plasmid (Vink et al., 2020a) using primers BN1388 and BN1390. The restriction, ligation, and transformations were carried out as described above. Cas1-Cas2/3 was amplified from the *E.coli Ed1A* genome using primers BN1437 and BN1386 which added restriction sites SpeI and SphI. The backbone was amplified from the existing pCas3 plasmid using primers BN1388 and BN1389. All constructs were verified by sequencing.

Recombination

The knockout insert containing the kanamycin resistance cassette flanked by lox sites was amplified using primers BG8387 and BG 8388. To ensure continued transcription/termination of the *cas8f* gene, the 5' flank of the insert overlapped with the *cas8f* gene and the 3' flank overlapped with the downstream region of the *cas6f* gene. The *E.coli* Ed1A strain harboring the temperature sensitive pSC020 plasmid was grown in LB with ampicillin at 30°C and made competent. After transformation, the cells were plated on LBA with ampicillin, kanamycin, and glucose (to suppress Cre recombinase expression). Colonies were verified with PCR and sequencing. Subsequently, the Cre recombinase was induced with 1mM IPTG at 37°C to promote plasmid and antibiotic resistance loss. The strain was then sequence verified to ensure antibiotic resistance loss. No mutations occurred within the *cas8f* gene.

The *pamcherry2* gene containing the chloramphenicol resistance (cat) cassette flanked by lox sites was amplified using primers BN43 and BN291. Due to the possibility of the lox-site recombination scar influencing gene transcription/termination, the chloramphenicol resistance cassette was placed upstream of the IGR (Intergenic region) present between *cas2-3* and *cas8f*. Part of the IGR was also added to the 5' flank of the insert to insure correct termination of the *cas2-3*. The 3' flank was constructed to overlap with the beginning of the *cas8f* gene. The further recombination procedure was performed as described above (*Δcas5-7*).

Growth conditions

To prevent the high-copy target plasmids from influencing the growth rate of the strains and therefore changing the fraction of matured PAmCherry complexes we used a rich defined medium with minimal autofluorescence. Strains were grown in M9 minimal medium containing the following supplements: 0.4% glucose, 1x EZ amino acids supplements (M2104 Teknova), 20 μg/ml uracil (Sigma-Aldrich), 1mM $MgSO_4$ (Sigma-Aldrich) and 0.1 mM $CaCl_2$ (Sigma-Aldrich) (further referred to as M9 medium). Strains were inoculated o/n from glycerol stocks and 200x diluted in fresh medium the next day. Cells were always grown with the required antibiotics. The expression level of I-F and I-E Cascade complexes for strains carrying the

pKEDR13 plasmid could be tuned by different expression levels of LeuO. The expression level referred to in the text as low induction was achieved by leaky expression of LeuO (no addition of IPTG), whereas high induction was achieved by addition of 1 mM IPTG upon dilution of the o/n culture. For all sptPALM measurements we used the high induction condition. The cells were grown for ~2.5 hours to an OD of 0.1 before use.

Transformation assay

Transformation assay was performed as described previously (Vink et al., 2020a). In brief, a mixture of pTarget and pGFPuv was electroporated and establishment probabilities were calculated from the formula below:

$$p_{establishment} = \frac{\#\ \text{pTarget colonies}}{\#\ \text{GFPuv colonies}} \times \frac{[\text{pGFPuv Transformed}]}{[\text{p Target Transformed}]}$$

Slide preparation

In order to work with very clean slides, an extensive cleaning procedure was used (modified from (Chandradoss et al., 2014)). Slides were burned in an oven at 500 °C for two hours, and then wrapped in aluminum foil until the day of usage. Slides were subsequently sonicated in MilliQ, Acetone and KOH, incubated in Piranha Solution (75% H_2SO_4, 7.5% H_2O_2) and afterwards rinsed with MilliQ. 1% Agarose slabs containing the growth medium were hardened between two cleaned glass slides, spaced slightly apart using parafilm. After hardening, a concentrated culture of cells was added in between the slab and one of the slides. The agarose slab was always prepared within 20 minutes of the measurement to prevent desiccation.

Microscope set-up and imaging

The image acquisition of the fluorescently labeled samples was carried using the Nikon Ti2E TIRF microscope at Delft University of Technology. The Gataca iLAS TIRF illumination module (used in HILO mode) was used to image the cells. A 100x Nikon 100x/1.49NA TIRF objective lens was used for visualization. The 405 nm laser was used to photo-activate PAmCherry and the 561 nm laser was used for

fluorescence excitation. Stroboscopic illumination was used to allow for 2 ms of excitation within each 10 ms frame. The camera filter used during image capture was a band pass filter of 609 nm ± 27 nm. The Andor iXon Ultra 897 EM-CCD camera (fluorescence) had a pixel size of 125 nm in the sample plane and was maintained at -70°C during use. The EM gain was set to 300. Samples were imaged for 4000 frames three times at three different locations using the acquisition software Metamorph. The pixel size of the Retiga R1 CCD from QImaging (phase contrast) camera is 64 nm.

Detection, localization and tracking

Analysis was done with home-built software described in our previous study (Vink et al., 2020a) and adapted from (Holden et al., 2010; Uphoff et al., 2013). Different from the previous study, measurements here were performed from an EM-CCD camera. Therefore, localizations were detected by different filtering and thresholding, considering the noise characteristics of these type of cameras. After finding all localizations in each frame, localizations in subsequent frames that were closer to each other than 6 pixels in length (0.78 μm) were assigned as a track. Particles were allowed to disappear for one frame (due to blinking/moving out of focus), but these steps were not used in the calculation of the apparent diffusion coefficient, D^*.

Analytical Diffusion Distribution Analysis (anaDDA)

Analytical Diffusion Distribution Analysis was previously developed and described (Vink et al., 2020b).

Copy number determination

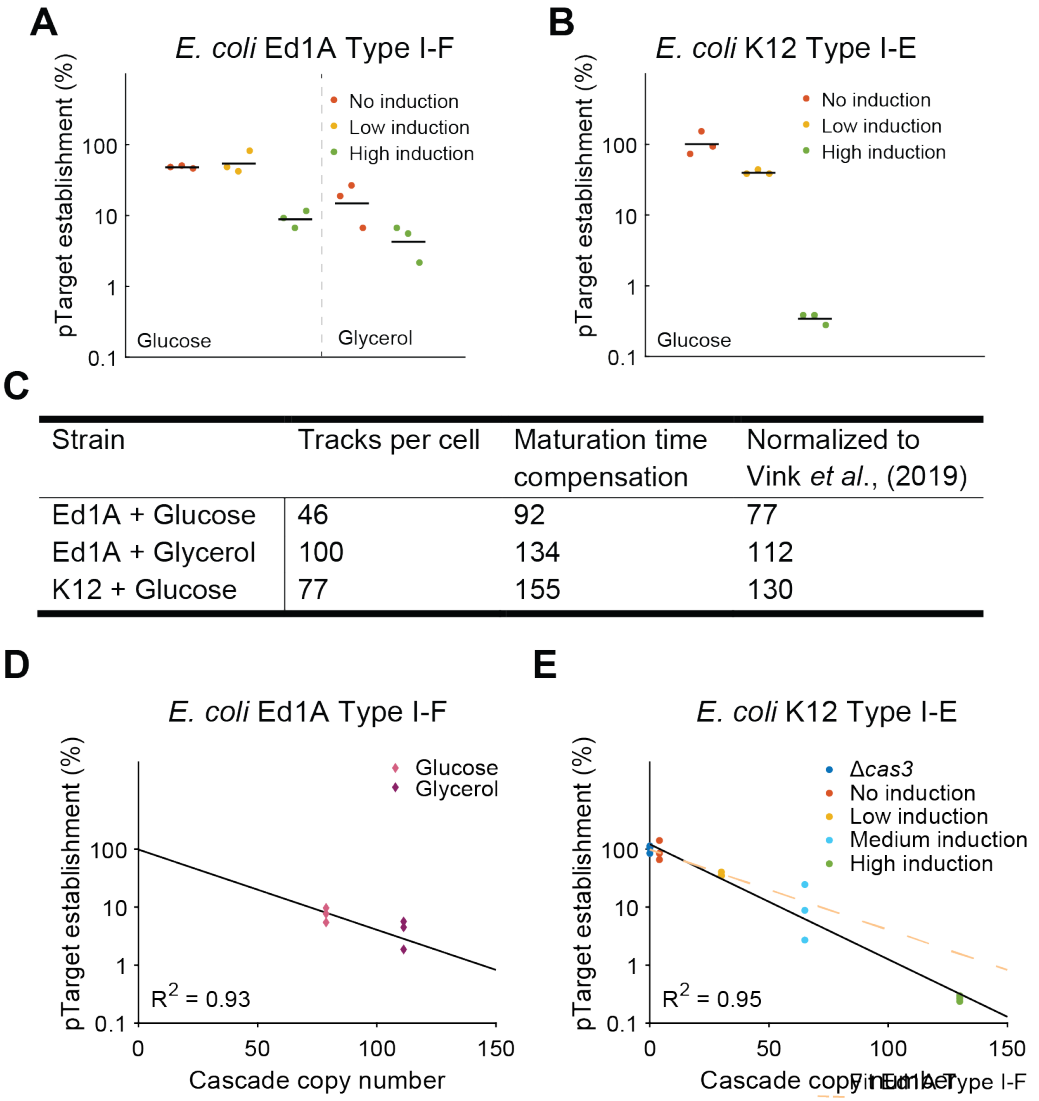
Cells were imaged for 20000 frames of 10 ms each. The 405 nm laser intensity was increased by 0.3% every 10 seconds to induce new PAmCherry fluorescence as old ones were photobleached. The number of fluorescence particles with a minimum number of steps of two or greater were counted. A distribution of step sizes from those particles was fitted to an exponential decay (the number of particles decreases

exponentially as the step size increases) and the number of complexes of step size one and zero was determined. The sum of all tracks was divided by the number of cells in the field of view to determine the average number of complexes per cell. The effect of differences in growth rate in different media on the fraction of matured proteins was taken into account during calculations (50% more complexes in Ed1A pKEDR13 induced in glucose and 75% more complexes in Ed1A induced in glycerol than visible at the moment of imaging (unmatured)). The number of complexes was then normalized to the number of complexes in strain K12 pKEDR13 induced grown in glucose from our previous study (Vink et al., 2020a).

Results

Interference efficiency in E. coli Type I-F can be measured with LeuO induction

The interference efficiency, the number of proteins required to achieve a certain protection level, can be quantified by performing a transformation assay and a single-molecule copy number measurement in parallel as described in our previous study of the Type I-E system (Vink et al., 2020a). In that study, we found an exponential relation between the probability of MGE establishment and the number of crRNA-guided complexes. To test whether Type I-F systems also follow the same relationship, we first looked for ways to modulate the expression levels of these complexes. Regulation of Type I-F systems have been known to be regulated by CRP-cAMP (Hampton et al., 2019; Patterson et al., 2015), Quorum Sensing (Høyland-Kroghsbo et al., 2017; Patterson et al., 2016) and iron depletion (Ahator et al., 2020), but no studies have described regulation of Type I-F systems in *E. coli*. To be able to predict potential regulators of Type I-F in *E. coli*, we looked for regulator binding sites in the promoter region of the *cas8f* gene. We found a potential CRP binding site, indicating that CRP-cAMP could function as a regulator (Supplementary Figure 1A). We furthermore found a putative binding site of LeuO, a known regulator of the type I-E *E. coli* CRISPR system (Supplementary Figure 1B) (Westra et al., 2010). This was surprising, since the *leuO* gene in strain Ed1A is disrupted by a transposon (Supplementary Figure 1C).



C

Strain	Tracks per cell	Maturation time compensation	Normalized to Vink *et al.*, (2019)
Ed1A + Glucose	46	92	77
Ed1A + Glycerol	100	134	112
K12 + Glucose	77	155	130

Figure 1. pTarget survival levels and interference efficiency of *E. coli* Ed1A Type I-F and *E. coli* K12 Type I-E. (A) pTarget contains 30 protospacers against all spacers in the Ed1A Type I-F system (pTargetF). pTarget establishment, calculated from the ratio of the transformation of pTarget/pGFPuv, is a measure of the interference level of the CRISPR system. The system was induced with different levels of LeuO and strains were grown on two different carbon sources (glycerol or glucose). (B) Same as Figure 1A but for data from the K12 Type I-E system. Modified from (Vink et al., 2020a). (C) The number of tracks is divided by the number of cells to determine the average number of tracks per cell. The growth rate to fluorescent protein maturation time difference is compensated for (Materials and Methods) and the results are then normalized to the values determined in Vink et al., (2019) to be able to compare results obtained with different microscopy set-ups. (D) The combination of the observed pTarget establishment levels (Figure 1A; High induction) and normalized copy numbers (Figure 1C) were fitted with an exponential decay function. (E) The same fit as Figure 1D but for data from the K12 Type I-E system, modified from (Vink et al., 2020a).

To test whether these predicted regulators were functional, we performed transformation assays with pTargetF, a high copy plasmid containing target sites for all 30 spacers found in the genomic arrays of *E. coli* Ed1A. We found that the pTargetF survival probability when challenged by the I-F system in Ed1A grown in glucose M9 media is lower (48 ± 1%; Figure 1A) than pTarget survival challenged by the I-E system in K12 (101 ± 8%; Figure 1B). This baseline protection level has been previously observed in Ed1A (Vorontsova et al., 2015) and markedly differs from the almost fully repressed I-E CRISPR system in K12. The pTarget survival probability further decreased when Ed1A was grown in glycerol (17 ± 6%), suggesting that CRP-cAMP induces the I-F CRISPR system in *E. coli* in a similar fashion to the previously studied *P. atrosepticum* system (Patterson et al., 2015).
We previously found that for the I-E CRISPR system in *E. coli* K12 increasing levels of LeuO, led to increasing levels of protection (Figure 1B; from (Vink et al., 2020a)). For the I-F system we found that low levels of LeuO (without addition of IPTG) did not significantly change the survival probability of pTargetF, but that high levels of LeuO induction (1 mM IPTG) lead to a significantly lower the survival probability was in glucose and glycerol growth medium (9 ± 1%, and 5 ± 1% respectively; Figure 1A).
Altogether, these findings indicate that increased levels of both CRP-cAMP and LeuO enhance expression of the Cascade complex in *E. coli* Type I-F systems.

25 Cascade complexes provide 50% CRISPR protection

Being able to regulate the copy numbers through LeuO expression and glucose levels and having measured different levels of MGE establishment, we set out to determine the interference efficiency by measuring the single-molecule copy numbers under the different conditions previously described. We fused the *pamcherry2* gene to the N terminus of *cas8f*. Subsequently we determined the copy number by counting the number of tracks and compensating for maturation time of the fluorophore and growth rate of cells (Materials and Methods). Furthermore, we calibrated the Nikon

Ti2E TIRF microscope set-up used in this study to our previous copy number measurements on the miCube (Martens et al., 2019; Vink et al., 2020a), by measuring the copy number of the I-E Cascade in *E. coli* K12 on both systems. We found that the average Cascade copy number measured on the Nikon Ti2E TIRF set-up (~155 copies) was close to the copy number under similar conditions in the miCube (~130 copies) enabling the comparison of I-F and I-E interference efficiency.

We subsequently measured the levels of I-F Cascade complexes and found that in presence of LeuO the complex copy number was higher when cells were grown in glycerol culture media (~110 copies; Figure 1C) compared to glucose (~75 copies), which matches the trend observed in the interference levels (Figure 1A). When we combined these observations with protection levels found in the transformation assays, we found that the relation between MGE establishment and complex copy number follow a similar exponential relation (Figure 1D) as described for the Type I-E system (Figure 1E). We deduced that ~25 copies of I-F Cascade complexes were required to provide 50% CRISPR protection which is similar to the ~20 complexes that were required to provide the same protection levels in the Type I-E *E. coli* system.

The I-F Cascade complex diffuses faster in the cell than the I-E Cascade complex

After finding a similar interference efficiency, we wondered whether the target search strategies of the Type I-E and Type I-F were also alike. We studied the target search by following the diffusion of thousands of fluorescently labeled Cas8f proteins as previously described for the Type I-E system (Vink et al., 2020a). Subsequent analysis with anaDDA (Vink et al., 2020b), enabled us to extract the free diffusion coefficient, the rate of diffusion in the absence of DNA binding, and the on and off-rates of DNA binding events from the distribution of measured displacements. To distinguish the diffusion behavior of Cas8f assembled within I-F Cascade complexes from monomeric Cas8f proteins, we first studied the diffusion

of the PAmCherry-Cas8f fusion in a strain lacking the other three subunits of the I-F Cascade complex.

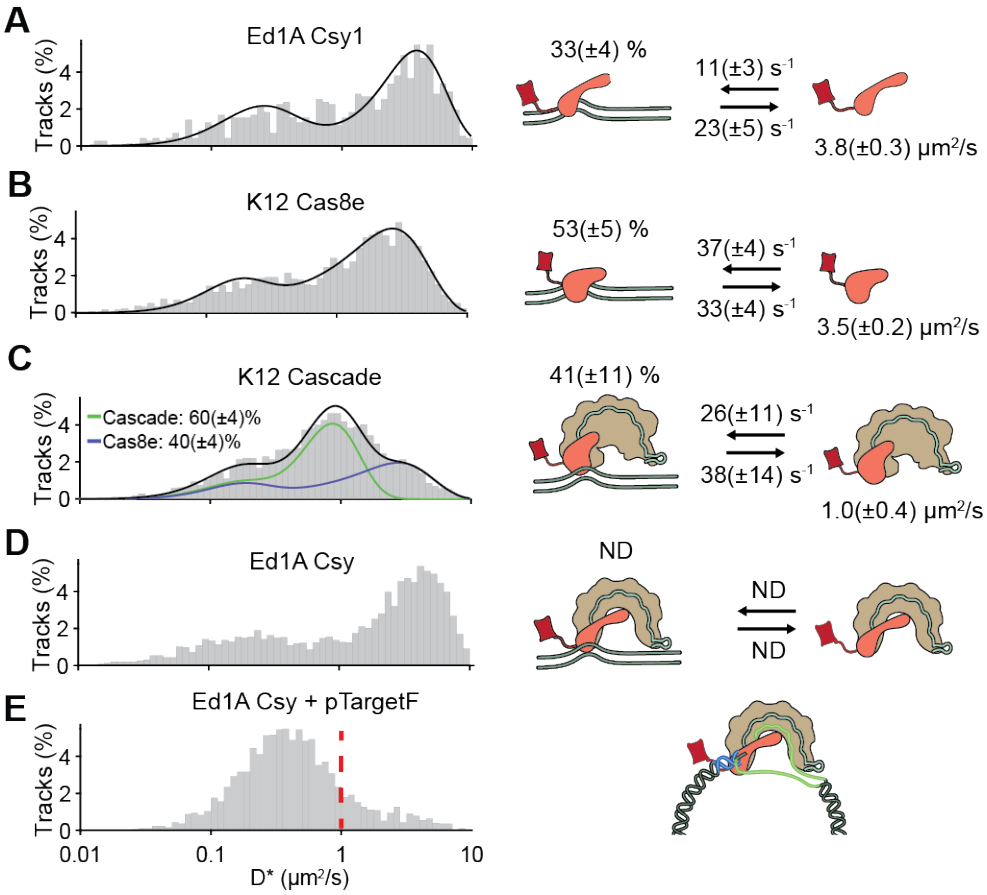


Figure 2. Diffusion dynamics of Cas8f subunit and complex in the absence and presence of targets. *D** distributions for (A) Cas8f, (B) Cas8e, (C) I-E Cascade and (D) I-F Cascade complex. The fraction of complex and free subunit could not be determined for the I-F Cascade complex as the *D** distribution was not significantly different from the Cas8f diffusion (ND: Not Determined). (E) Due to the lack of resolution between subunit and complex, the fraction of complexes bound in the presence of pTargetF could not be accurately determined. Instead, we used the threshold of 1 µm²/s (red line) to approximate the fraction slow/moving immobile vs mobile I-F Cascade complexes. Error estimation is based on bootstrapping (±SD). Figure 2B-C were taken from (Vink et al., 2020a).

The diffusion behaviour of PAmCherry-Cas8f (Figure 2A) was distinct from the diffusion behavior of PAmCherry-Cas8e (Figure 2B). When we analyzed the distribution with anaDDA, we found that even though the free diffusion coefficients

were relatively the same (3.8 ± 0.3 μm^2/s for Cas8f and 3.5 ± 0.3 μm^2/s for Cas8e), the on and off-rated of PamCherry-Cas8f on DNA (k_{off} = 23 ± 5 s^{-1}; k^*_{on}= 11 ± 3 s^{-1}) were slower than for PAmCherry-Cas8e (k_{off} = 33 ± 4 s^{-1} ; k^*_{on}= 37 ± 4 s^{-1}). Also, the fraction of time spent on DNA, calculated from the ratio between the on and off-rates, is different, where Cas8e spends roughly half its time on DNA (53 ± 5%), Cas8f only spends 33 ± 4% of its time on DNA.

When we subsequently measured PAmCherry-Cas8f in the presence of all subunits, we expected to observe slower average diffusion due to complex formation, which we previously observed in the Type I-E system (Figure 2C). However, in this case the distribution stayed relatively the same (Figure 2D). This made it impossible to distinguish the diffusion of Cas8f and the I-F Cascade complex and therefore the fraction of subunits that assembled into a complex could not be determined.

4

To exclude the possibility that the I-F Cascade complex was not able to assemble in the presence of a fluorescent tag, we studied the diffusion distribution in the presence of pTargetF. We performed the measurement in a *Δcas2-3* background to prevent degradation of the target plasmid. In the presence of pTargetF, the number of immobile/slow-moving tracks (D^* < 1 μm^2/s) increased (Figure 2E). As we only expect assembled complexes containing matching crRNA to be able to tightly bind to pTargetF, this indicated that a significant fraction of PamCherry-Cas8f was part of a complex, but that it was moving too fast to accurately resolve from the Cas8f subunit.

Despite the absence of a detailed quantification of complex formation and target search kinetics of the I-F Cascade complex, we can thus far determine that the I-F Cascade complex is formed in the presence of a fluorescent tag and able to bind to pTargetF and that the free diffusion coefficient of the I-F Cascade complex is higher than of the I-E Cascade complex. Given their overlapping distributions, it is likely that Cas8f and I-F Cascade complex have similar DNA probing dynamics. In that case, the interaction kinetics with DNA of the I-F complex are slower and less frequent than of its I-E counterpart Cascade.

The I-F Cascade complex only tolerates PAM-distal mutations

So far we have studied the target search kinetics of the Type I-E and Type I-F systems by investigating the interference efficiency and the diffusion behavior in the cell in the absence and presence of targets. Another feature that is important for the functioning of immune system is its tolerance to escape mutations. An *in vitro* study has shown that Cascade in the I-E system has two binding modes, the canonical binding mode, occurring through PAM-proximal RNA-DNA interactions and the non-canonical PAM-distal RNA-DNA interactions (Blosser et al., 2015; Jung et al., 2017). The non-canonical binding mode, would still allow partial binding of targets mutated in the region close to the PAM. We wanted to test whether the Cascade complex of the Type I-F system also can bind targets through these two binding modes. We made variants of pTargetF where each certain segments of each target on the plasmid was systematically mutated following the scheme of previously published *I-E* in vitro work (Figure 4A). The protospacers (32 nt length) were subdivided in 6 segments (S1-6) of 6 nucleotides long (except S6, 2 nt long) and segments were either mutated starting from the PAM-proximal or PAM-distal segment.

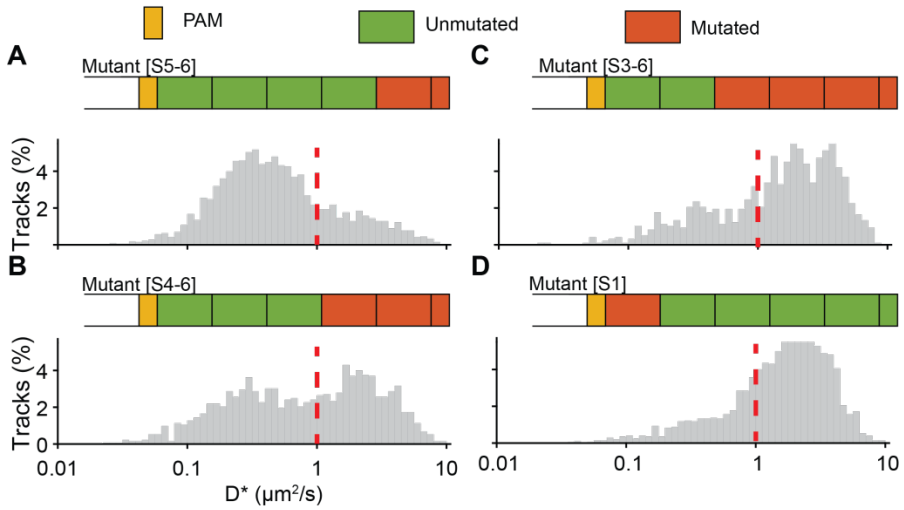


Figure 3. Diffusion dynamics of I-F Cascade complex in the presence of mutated targets. Each of 30 targets was divided in six segments which were systematically mutated (red) or kept intact (green) either starting from segments far away from the PAM (yellow; PAM-distal) or

close to the PAM (PAM-distal). (**A-C**) Cascade diffusion in the presence of targets with PAM-distal mutated segments. **(D)** Cascade diffusion in the presence of targets with PAM-proximal mutated segments. The threshold of 1 $\mu m^2/s$ (red line) was used to approximate the fraction slow/moving immobile vs mobile Cascade complexes.

The presence of PAM-distal mutated segments increased the fraction of unbound Cascade complexes ($D^*>1$) compared to the bona fide target plasmid pTargetF (9%, Figure 2E) progressively from 24% for mutated segments S5-6, 54% for mutated segments S4-6 and 71% for mutated segments S3-6 (Figure 3A-C). However, even with many mutated segments (S3-S6), we still observed slightly higher levels of immobile/slow-moving tracks (29%) compared to the absence of targets (21%, Figure 3C) indicating that complexes were able to bind to targets only containing the first 12 nucleotides of the protospacer. In contrast, mutating a single segment on the PAM-proximal side led to almost complete abolishment of binding (Figure 3D). This data suggest that *in vivo* I-F Cascade exclusively uses the canonical binding mode, which entails directional R-loop formation from a PAM-proximal seed sequence (Guo et al., 2017; Tuminauskaite et al., 2020) and does not utilize the PAM-distal RNA-DNA interactions as an initial step as was shown for I-E Cascade *in vitro* (Blosser et al., 2015).

Altogether, this data suggests that unlike I-E Cascade, the I-F Cascade does not contain multiple binding modes, but that it can accommodate many mutations in the PAM-distal region of the protospacer.

Discussion

In this study, we have investigated, whether the native I-F *E. coli* CRISPR-Cas system uses similar or different target search strategies and whether this process attains similar efficiencies compared to the native I-E *E. coli* CRISPR-Cas system.

Our data indicated that the Cascade complex diffuses as fast as the subunit, which is unlikely given the large differences in size. We think that measured apparent diffusion coefficient of the subunit is limited by the temporal resolution of our setup and our tracking algorithm and therefore higher than what we can currently measure.

However, even though we are unable to quantify the diffusion coefficient further, we can conclude that the diffusion of the I-F complex is faster (~ 3.5 $\mu m^2/s$) than the I-E Cascade (1.0 $\mu m^2/s$) in native *E. coli* cells. The fast diffusion coefficient of the I-F system is unexpected, given their mass is not so different (I-E Cascade 405 kDa; I-F Cascade: 350 kDa). However, the shape of the I-F Cascade complex is very different, the helical pitch of Cas7 is much tighter in I-F compared to I-E, resulting in an almost complete ring shape (Chowdhury et al., 2017), compared to the more seahorse shape of I-E Cascade. However, software (Fleming and Fleming, 2018) predicting the hydrodynamic radius of structures only show a minor decrease in the hydrodynamic radius (65 Å for I-F Cascade (Guo et al., 2017); 68 Å for I-E Cascade (Xiao et al., 2018)), making it unlikely that it could have a large impact on diffusion (Erickson, 2009). Another potential explanation that could account for the difference is that for the I-E system it was shown *in vitro* that Cas3 (101 kDa) and Cas1-2 (154 kDa) are associated with Cascade (Dillard et al., 2018). This would make the combined size much larger and therefore their diffusion slower. For the I-F Cascade complex it was shown that Cas2-3 is not associated to Cascade in the cell (Govindarajan et al., 2020).

Apart from the free diffusion rate, if we assume the binding kinetics of the complex are comparable to the Cas8f subunit, which is the case for the Cascade I-E system (Vink et al., 2020a), we can also conclude that the DNA binding and unbinding occur less frequently in the I-F system. The less frequent binding kinetics are also contradictory to what you would expect, given that the smaller CC PAM occurs more frequently (~4x) in the genome than the consensus PAM of the I-E Cascade (AAG). However, a recent study of genome surveillance demonstrated that the I-F Cascade complex only has PAM-dependent interactions (Govindarajan et al., 2020), whereas our previous I-E study demonstrated both PAM-dependent and PAM-independent interactions (Vink et al., 2020a). The less frequent DNA interactions could be explained by the lack of PAM-independent interactions in the case of the I-F system. We found that the interference efficiencies are roughly similar between the two systems, while their binding and diffusion kinetics are more distinct. The faster diffusion of the I-F complex would make it more efficient, however it could be

counterbalanced by the less frequent DNA binding interactions. It will be interesting to add more systems to this comparison to find out whether this efficiency is a general physical limit of DNA-targeting CRISPR systems.

It was previously shown that the I-F system of *E. coli* preferentially targets plasmids, whereas the I-E system has more spacers targeting phages (Díez-Villaseñor et al., 2010). The findings of this study show similar efficiencies when given a transformed plasmid, which cannot explain the higher preference for I-F systems in targeting. It could be that conjugation, with its distinct single stranded DNA entry and different replication kinetics, does confer a specific advantage with I-F systems. Secondly, this preference for plasmids has however not been described in other organisms and, given the low number of spacer matches in that study (Díez-Villaseñor et al., 2010), could be coincidental. Thirdly, the main cause of different subsystem usage might not only be attributed to efficiency and speed, but also to their robustness in dealing with mutated targets.

In this study, we tested the robustness of I-F Cascade *in vivo* given PAM-proximal and PAM-distal mutations. Compared to a previous *in vitro* study of I-E, we did not observe Cascade binding to targets with PAM-proximal (seed) mutations. This could mean that the I-E Cascade offers more robust binding, whereas the I-F Cascade might show less off-target binding by rejecting mismatches more easily. However, for a more accurate comparison, I-E Cascade mutated target binding should be assessed under similar *in vivo* conditions. In our interference efficiency test, these aspects are not studied. Therefore, we suggest testing interference efficiencies of these systems within a population of mutating phages that will then provide an estimate of the escape rate, the number of phages that, by acquiring a mutationm can escape CRISPR interference, and could further provide more insight into the preferred use of one CRISPR-Cas system compared to another given a certain invader.

In this study we have outlined ways in which the target search processes of CRISPR systems can be compared in native settings. So far, we have studied two of the most closely related CRISPR systems, namely the I-E and the I-F system. Expanding this research to more distantly related systems could help researchers to decide which CRISPR systems are most suitable given a certain application, e.g. precise vs

promiscuous editing, large vs small genomes, and to explain the wide diversity of CRISPR-Cas effector complexes and their co-occurring usage in many prokaryotes. Overall, our study demonstrates the many trade-offs that have to be considered for CRISPR-Cas systems such as search efficiency, accuracy, robustness and speed and the diversity of ways to balance them.

References

Ahator, S. Dela, Jianhe, W., and Zhang, L.-H. (2020). The ECF sigma factor PvdS regulates the type I-F CRISPR-Cas system in Pseudomonas aeruginosa BioRxiv 2020.01.31.929752.

Blosser, T.R., Loeff, L., Westra, E.R., Vlot, M., Künne, T., Sobota, M., Dekker, C., Brouns, S.J.J., and Joo, C. (2015). Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. Mol. Cell *58*, 60–70.

Chandradoss, S.D., Haagsma, A.C., Lee, Y.K., Hwang, J.-H., Nam, J.-M., and Joo, C. (2014). Surface Passivation for Single-molecule Protein Studies. J. Vis. Exp. *86*, 4–11.

Chowdhury, S., Carter, J., Rollins, M.F., Golden, S.M., Jackson, R.N., Hoffmann, C., Nosaka, L., Bondy-Denomy, J., Maxwell, K.L., Davidson, A.R., et al. (2017). Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. Cell *169*, 47-57.e11.

Díez-Villaseñor, C., Almendros, C., García-Martínez, J., and Mojica, F.J.M. (2010). Diversity of CRISPR loci in Escherichia coli. Microbiology.

Dillard, K.E., Brown, M.W., Johnson, N. V, Xiao, Y., Dolan, A., Hernandez, E., Dahlhauser, S.D., Kim, Y., Myler, L.R., Anslyn, E. V, et al. (2018). Assembly and Translocation of a CRISPR-Cas Primed Acquisition Complex. Cell *175*, 934-946.e15.

Erickson, H.P. (2009). Size and Shape of Protein Molecules at the Nanometer Level Determined by Sedimentation, Gel Filtration, and Electron Microscopy. Biol. Proced. Online *11*, 32.

Fleming, P.J., and Fleming, K.G. (2018). HullRad: Fast Calculations of Folded and Disordered Protein and Nucleic Acid Hydrodynamic Properties. Biophys. J.

Gasiunas, G., Young, J.K., Karvelis, T., Kazlauskas, D., Urbaitis, T., Jasnauskaite, M., Grusyte, M.M., Paulraj, S., Wang, P.H., Hou, Z., et al. (2020). A catalogue of biochemically diverse CRISPR-Cas9 orthologs. Nat. Commun.

Govindarajan, S., Borges, A., and Bondy-Denomy, J. (2020). Direct observation of genome surveillance by CRISPR-Cas in bacteria. BioRxiv 2020.09.29.318501.

Guo, T.W., Bartesaghi, A., Yang, H., Falconieri, V., Rao, P., Merk, A., Eng, E.T., Raczkowski, A.M., Fox, T., Earl, L.A., et al. (2017). Cryo-EM Structures Reveal

Mechanism and Inhibition of DNA Targeting by a CRISPR-Cas Surveillance Complex. Cell.

Hampton, H.G., Patterson, A.G., Chang, J.T., Taylor, C., and Fineran, P.C. (2019). GalK limits type I-F CRISPR-Cas expression in a CRP-dependent manner. FEMS Microbiol. Lett. *366*.

Hoikkala, A.V., Ravantti, J.J., Díez-villaseñor, C., Tiirola, M., Rachel, A., Hoikkala, V., Ravantti, J.J., Diez-Villasenor, C., Tiirola, M., Conrad, R., et al. (2020). Cooperation between CRISPR-Cas types enables adaptation in an RNA- targeting system. BioRxiv.

Holden, S.J., Uphoff, S., Hohlbein, J., Yadin, D., Le Reste, L., Britton, O.J., and Kapanidis, A.N. (2010). Defining the Limits of Single-Molecule FRET Resolution in TIRF Microscopy. Biophys. J. *99*, 3102–3111.

Høyland-Kroghsbo, N.M., Paczkowski, J., Mukherjee, S., Broniewski, J., Westra, E., Bondy-Denomy, J., and Bassler, B.L. (2017). Quorum sensing controls the pseudomonas aeruginosa CRISPR-Cas adaptive immune system. Proc. Natl. Acad. Sci. U. S. A. *114*, 131–135.

Jones, D.L., Leroy, P., Unoson, C., Fange, D., Ćurić, V., Lawson, M.J., and Elf, J. (2017). Kinetics of dCas9 target search in Escherichia coli. Science *357*, 1420–1424.

Jung, C., Hawkins, J.A., Jones, S.K., Xiao, Y., Rybarski, J.R., Dillard, K.E., Hussmann, J., Saifuddin, F.A., Savran, C.A., Ellington, A.D., et al. (2017). Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips. Cell *170*, 35-47.e13.

Knight, S.C., Xie, L., Deng, W., Guglielmi, B., Witkowsky, L.B., Bosanac, L., Zhang, E.T., El Beheiry, M., Masson, J.-B.J.-B.J.-B., Dahan, M., et al. (2015). Dynamics of CRISPR-Cas9 genome interrogation in living cells. Science *350*, 823–826.

Koonin, E. V., Makarova, K.S., and Zhang, F. (2017). Diversity, classification and evolution of CRISPR-Cas systems. Curr. Opin. Microbiol. *37*, 67–78.

Leenay, R.T., Maksimchuk, K.R., Slotkowski, R.A., Agrawal, R.N., Gomaa, A.A., Briner, A.E., Barrangou, R., and Beisel, C.L. (2016). Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. Mol. Cell *62*, 137–147.

Levy, A., Goren, M.G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U., and Sorek, R. (2015). CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature *520*, 505–510.

Majumdar, S., Zhao, P., Pfister, N.T., Compton, M., Olson, S., Glover, C.V.C., Wells, L., Graveley, B.R., Terns, R.M., and Terns, M.P. (2015). Three CRISPR-Cas immune effector complexes coexist in Pyrococcus furiosus. RNA.

Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P., et al. (2020). Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat. Rev. Microbiol.

Malone, L.M., Warring, S.L., Jackson, S.A., Warnecke, C., Gardner, P.P., Gumy, L.F., and Fineran, P.C. (2020). A jumbo phage that forms a nucleus-like structure evades CRISPR–Cas DNA targeting but is vulnerable to type III RNA-based immunity. Nat. Microbiol.

Manley, S., Gillette, J.M., Patterson, G.H., Shroff, H., Hess, H.F., Betzig, E., and Lippincott-Schwartz, J. (2008). High-density mapping of single-molecule trajectories with photoactivated localization microscopy. Nat. Methods *5*, 155–157.

Martens, K.J.A., van Beljouw, S.P.B., van der Els, S., Vink, J.N.A., Baas, S., Vogelaar, G.A., Brouns, S.J.J., van Baarlen, P., Kleerebezem, M., and Hohlbein, J. (2019). Visualisation of dCas9 target search in vivo using an open-microscopy framework. Nat. Commun. *10*, 3552.

Mendoza, S.D., Nieweglowska, E.S., Govindarajan, S., Leon, L.M., Berry, J.D., Tiwari, A., Chaikeeratisak, V., Pogliano, J., Agard, D.A., and Bondy-Denomy, J. (2020). A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. Nature *577*, 244–248.

Mohr, G., Silas, S., Stamos, J.L., Makarova, K.S., Markham, L.M., Yao, J., Lucas-Elío, P., Sanchez-Amat, A., Fire, A.Z., Koonin, E. V., et al. (2018). A Reverse Transcriptase-Cas1 Fusion Protein Contains a Cas6 Domain Required for Both CRISPR RNA Biogenesis and RNA Spacer Acquisition. Mol. Cell *72*, 700-714.e8.

Patterson, A.G., Chang, J.T., Taylor, C., and Fineran, P.C. (2015). Regulation of the type I-F CRISPR-Cas system by CRP-cAMP and GalM controls spacer acquisition and interference. Nucleic Acids Res. *43*, 6038–6048.

Patterson, A.G., Jackson, S.A., Taylor, C., Evans, G.B., Salmond, G.P.C., Przybilski, R., Staals, R.H.J., and Fineran, P.C. (2016). Quorum Sensing Controls Adaptive Immunity through the Regulation of Multiple CRISPR-Cas Systems. Mol. Cell *64*, 1102–1108.

Pinilla-Redondo, R., Mayo-Muñoz, D., Russel, J., Garrett, R.A., Randau, L., Sørensen, S.J., and Shah, S.A. (2019). Type IV CRISPR–Cas systems are highly diverse and involved in competition between plasmids. Nucleic Acids Res.

Pourcel, C., Touchon, M., Villeriot, N., Vernadet, J.P., Couvin, D., Toffano-Nioche, C., and Vergnaud, G. (2020). CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. Nucleic Acids Res. *48*, D535–D544.

Puigbò, P., Makarova, K.S., Kristensen, D.M., Wolf, Y.I., and Koonin, E. V. (2017). Reconstruction of the evolution of microbial defense systems. BMC Evol. Biol.

Silas, S., Mohr, G., Sidote, D.J., Markham, L.M., Sanchez-Amat, A., Bhaya, D., Lambowitz, A.M., and Fire, A.Z. (2016). Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. Science (80-.). *351*.

Silas, S., Lucas-Elio, P., Jackson, S.A., Aroca-Crevillén, A., Hansen, L.L., Fineran, P.C., Fire, A.Z., and Sánchez-Amat, A. (2017). Type III CRISPR-Cas systems can provide redundancy to counteract viral escape from type I systems. Elife.

Toro, N., Mestre, M.R., Martínez-Abarca, F., and González-Delgado, A. (2019). Recruitment of Reverse Transcriptase-Cas1 Fusion Proteins by Type VI-A CRISPR-Cas Systems. Front. Microbiol. *10*.

Tuminauskaite, D., Norkunaite, D., Fiodorovaite, M., Tumas, S., Songailiene, I., Tamulaitiene, G., and Sinkunas, T. (2020). DNA interference is controlled by R-loop length in a type I-F1 CRISPR-Cas system. BMC Biol.

Uphoff, S., Reyes-Lamothe, R., Garza de Leon, F., Sherratt, D.J., and Kapanidis, A.N. (2013). Single-molecule DNA repair in live bacteria. Proc. Natl. Acad. Sci. U. S. A. *110*, 8063–8068.

Vink, J.N.A., Martens, K.J.A., Vlot, M., McKenzie, R.E., Almendros, C., Estrada Bonilla, B., Brocken, D.J.W., Hohlbein, J., and Brouns, S.J.J. (2020a). Direct

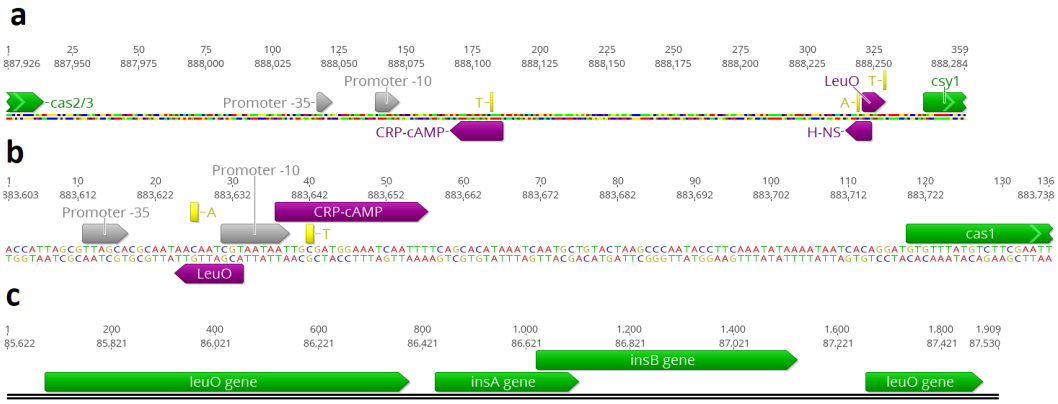
Visualization of Native CRISPR Target Search in Live Bacteria Reveals Cascade DNA Surveillance Mechanism. Mol. Cell *77*, 39-50.e10.

Vink, J.N.A., Brouns, S.J.J., and Hohlbein, J. (2020b). Extracting Transition Rates in Particle Tracking Using Analytical Diffusion Distribution Analysis. Biophys. J.

Vorontsova, D., Datsenko, K.A., Medvedeva, S., Bondy-Denomy, J., Savitskaya, E.E., Pougach, K., Logacheva, M., Wiedenheft, B., Davidson, A.R., Severinov, K., et al. (2015). Foreign DNA acquisition by the I-F CRISPR-Cas system requires all components of the interference machinery. Nucleic Acids Res. *43*, 10848–10860.

Westra, E.R., Pul, Ü., Heidrich, N., Jore, M.M., Lundgren, M., Stratmann, T., Wurm, R., Raine, A., Mescher, M., Van Heereveld, L., et al. (2010). H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. Mol. Microbiol. *77*, 1380–1393.

Xiao, Y., Luo, M., Dolan, A.E., Liao, M., and Ke, A. (2018). Structure basis for RNA-guided DNA degradation by Cascade and Cas3. Science *361*, eaat0839.



Supplementary figure 1. ***cas8f* (*E. coli* Ed1a), *cas1* (*E. coli* K12) promoter elements and the truncated *leuo* gene (*E. coli* Ed1a)**. The highlighted (yellow) nucleotides indicate binding mismatches with previously determined binding sites in other strains. (A) Promoter region of the *cas8f* gene (green). Potential CRP-cAMP, H-NS, and LeuO binding sites (purple) are present downstream of the promoter -10 and -35 regions (grey). (b) Promoter region of the *cas1* gene (green). A potential CRP-cAMP binding site (purple) is present downstream of the promoter -10 and -35 regions (grey) while a potential LeuO binding site (purple) is present between them. (c) The nonfunctional *leuo* pseudogene (green) truncated by transposon elements (green).

5

Extracting transition rates in particle tracking using analytical diffusion distribution analysis

Published as: J. N. A. Vink, S. J. J. Brouns, J. Hohlbein, Extracting Transition Rates in Particle Tracking Using Analytical Diffusion Distribution Analysis. *Biophys. J.* (2020), doi:10.1016/j.bpj.2020.09.033.

Abstract

Single-particle tracking is an important technique in the life sciences to understand the kinetics of biomolecules. The analysis of apparent diffusion coefficients *in vivo*, for example, enables researchers to determine whether biomolecules are moving alone, as part of a larger complex or are bound to large cellular components such as the membrane or chromosomal DNA. A remaining challenge has been to retrieve quantitative kinetic models especially for molecules that rapidly switch between different diffusional states. Here, we present analytic diffusion distribution analysis (anaDDA), a framework that allows extracting transition rates from distributions of apparent diffusion coefficients calculated from short trajectories that feature less than 10 localisations per track. Under the assumption that the system is Markovian and diffusion is purely Brownian, we show that theoretically predicted distributions accurately match simulated distributions and that anaDDA outperforms existing methods to retrieve kinetics especially in the fast regime of 0.1-10 transitions per imaging frame. AnaDDA does account for the effects of confinement and tracking window boundaries. Furthermore, we added the option to perform global fitting of data acquired at different frame times, to allow complex models with multiple states to be fitted confidently. Previously, we have started to develop anaDDA to investigate the target search of CRISPR-Cas complexes. In this work, we have optimized the algorithms and reanalysed experimental data of DNA polymerase I diffusing in live *E. coli*. We found that long-lived DNA interaction by DNA polymerase are more abundant upon DNA damage, suggesting roles in DNA repair. We further revealed and quantified fast DNA probing interactions that last shorter than 10 ms. AnaDDA pushes the boundaries of the timescale of interactions that can be probed with single-particle tracking and is a mathematically rigorous framework that can be further expanded to extract detailed information about the behaviour of biomolecules in living cells.

Statement of significance

Fluorescence-based single-particle tracking is an important tool to study the dynamics of biological systems. The rate at which biomolecules move and interact is ideally inferred from their positional trajectories. Currently, however, extraction of these kinetic parameters remains challenging, especially with short trajectories. We have developed an analytical framework (anaDDA) that extracts transition rates directly from the distribution of apparent diffusion coefficients. AnaDDA outperforms existing tools, especially in regimes in which transition rates approach the data acquisition rate. We demonstrate its general applicability by re-analysing previously published data on DNA polymerase diffusion and find fast DNA interactions previously unobserved. AnaDDA is computationally fast, easy to use and allows researchers to reveal detailed information about the behaviour of biomolecules in living cells.

Introduction

Single-molecule studies have greatly expanded our knowledge of the mode of action and kinetics of DNA-protein interactions at the nanoscale (Miller et al., 2018). Single-molecule Förster resonance energy transfer (smFRET) and optical/magnetic tweezers, for example, are well suited techniques to study forces, conformational changes and displacements of DNA-binding proteins such as DNA and RNA polymerases(Hodges et al., 2009; Hohlbein et al., 2013), helicases (Craig et al., 2017; Rothenberg and Ha, 2010) and CRISPR-Cas proteins (Blosser et al., 2015; Rutkauskas et al., 2017) *in vitro* with high spatiotemporal resolution (Blouin et al., 2015; Heller et al., 2014; Hohlbein et al., 2014; Lerner et al., 2018). *In vivo*, however, single-particle tracking (SPT) remains the most convenient choice to study dynamic interactions (Kapanidis et al., 2018). For performing SPT, a gene of interest is fused to a gene expressing either a fluorescent protein or a protein tag (HaloTag/SnapTag) that can be later labelled with an organic fluorophore (Banaz et al., 2018; Jradi and Lavis, 2019). To avoid the temporal overlapping of emitters moving in the confined volume of (bacterial) cells, two strategies can be pursued. Either the expression level of the protein of interest is kept sufficiently low, or the emission signal from different proteins has to be separated in time which can be achieved using photoswitchable or photoactivatable fluorescent proteins or equivalent organic fluorophores enabling single-particle tracking photoactivation light microscopy (sptPALM) (English et al., 2011; Garza de Leon et al., 2017; Manley et al., 2008; Uphoff et al., 2013). After linking subsequent localizations of tagged proteins into tracks, the apparent diffusion coefficient D_j^* for each track *j* is calculated from the average of *n* squared displacements ($D_j^* = \frac{\sum_{i=1}^{n} r_i^2}{4nt}$), where *n* represents a given step number. Summing over all tracks *j* will lead to a distribution of D^* values, even if the motion of each particle is governed by a single diffusive state with diffusion coefficient *D* (Rocha et al., 2019).

The different mobilities of proteins switching between a DNA-bound state, in which proteins diffuse very slowly, and a DNA-free state, in which proteins diffuse through the cytoplasm, can provide kinetic information on the frequency and longevity of DNA-protein interactions.

For tracking applications in which the number of localisations per track is large (>50 localisations), previous studies have demonstrated the reliable extraction of diffusion and transition kinetics (Akimoto and Yamamoto, 2017; Lanoiselée and Grebenkov, 2017). For sptPALM with fluorescent proteins, however, the ability to extract this information is severely compromised by pre-mature photo bleaching often limiting the length of each track to a few localisations (Van Beljouw et al., 2019). Furthermore, the limited localization precision increases the apparent diffusion of immobile states. Therefore, measured displacements cannot be unambiguously assigned to either a bound or a diffusing state. As a consequence, histograms of D^* values are often rather broad making a clear distinction between two diffusional states of a single species impossible. For the special case of non-interconverting D^* distributions, the shape of distributions can be calculated for a fixed number of analysed steps (Qian et al., 1991; Saxton, 1997) and, via fitting of the experimental data, used to extract the fractions of mobile and immobile proteins. Another factor that can increase the overlap between two states in D^* distributions are state transitions occurring within single tracks. Using a typical frame time of 10 ms and a typical track length of 40 ms, any transition occurring within that track length will average out (Figure 1A). The framework described in references (Qian et al., 1991; Saxton, 1997) does not account for the possibility of transitions within a track. Consequently, the overlap can lead to overfitting, as an increase of intermediate values would necessitate the addition of more states, which are not necessarily biologically relevant. *In vitro* smFRET measurements have encountered a similar challenge, in which the interchanging of conformational states within single bursts or within single frames resulted in the averaging of FRET values. By implementing probability distribution analysis (PDA) (Kalinin et al., 2008; Palo et al., 2006) previous studies were able to extract kinetic information and fit the entire FRET distribution (Farooq and Hohlbein, 2015; Nir et al., 2006; Santoso et al., 2010).

In this study, we aim to incorporate the statistical framework of PDA into D^* fitting based on averaging single-frame displacements in individual sptPALM tracks, which will allow us to directly extract biologically relevant parameters such as on-

and off-rate next to the free diffusion coefficient and the total DNA-bound fraction. This method, which we call analytical diffusion distribution analysis (anaDDA), finds the kinetic parameters by implementing maximum likelihood estimation (MLE) and uses the probability to find *D** for all tracks between 1-8 steps long (where step number is the number of localizations-1), present in the data set (Figure 1B). We benchmark this analysis method, with simulation of transitioning particles and implement modifications that account for specific experimental challenges, such as varying tracking windows and confinement effects within the cell. Furthermore, we compare anaDDA to a different kinetic analysis tools that use Bayesian statistics or unsupervised Gibbs sampling to infer state transitions from the data (Karslake et al., 2020; Persson et al., 2013). We study the effects of confinement and tracking parameters on the fitting of the distribution coefficient distribution and although anaDDA was not designed to automatically determine the number of states, we discuss ways to manually assess the number of states required to fit the data. We furthermore re-analyse previously published sptPALM data of DNA interacting proteins, obtain their kinetic parameters, and reveal that fast DNA probing interactions were hidden in the published data. Using anaDDA on short trajectories, we demonstrate the fast and accurate analysis of transient DNA-protein interactions in the millisecond time range, a range that was previously only accessible in slimfield microscopy (Plank et al., 2009). In addition, anaDDA allows users to quickly check whether any tracking data that would imply the existence of either many static states or non-Brownian diffusion can be reduced to a simple Brownian diffusion model with dynamic state transitions.

Methods

*D** fitting with transitioning states

Distributions of *D** have been fitted in numerous studies of DNA binding proteins (Stracy et al., 2015; Vrljic et al., 2002) using an formalism derived by Qian *et al.*(Qian et al., 1991) from repeated convolution of the exponential distribution of

displacement, resulting in a gamma function for each state. The formalism was later expanded by Michalet to account for localization errors (Michalet, 2010) leading to

$$f_D(x; D, n) = \frac{\left(\frac{n}{D + \sigma^2/t}\right)^n x^{n-1} e^{-\frac{nx}{D + \sigma^2/t}}}{(n-1)!}, \quad (35)$$

where x is the measured displacement, D is the diffusion coefficient, n is the number of steps per track, t is the frame time, σ is the localization error and $f_D(x; D, n)$ is the probability to find a measured displacement given D and n. For multi-state (or multi-species) systems, terms can be added with different values of D_i and normalised by probability coefficients A_i. The goal is to find the distribution of apparent D^* values (x), for a certain number of underlying states that each have a probability A_i and a diffusion coefficient D_i. These distributions assume, however, that there is no dynamic transitioning occurring between diffusional states of one species.

In order to account for dynamics of state transitions in a two state system, we incorporated a statistical framework derived for probability distribution analysis (PDA) that is used to analyse single-molecule FRET distributions (Antonik et al., 2006; Kalinin et al., 2008; Palo et al., 2006). This method describes the distribution of time spent in each state given a certain k_{on}^*, k_{off} and the integrated time t_{int}.

The probability of staying in an initially occupied state S1 for an occupation time t_{S1} without transition is

$$W_{contS1}(t_{S1} = t_{int} | k_{off}, t_{int}) = e^{-k_{off} t_{int}}. \quad (36)$$

The probability density functions describing t_{S1} for an odd or an even number of transitions starting from state S1 are given by(Palo et al., 2006)

$$W_{oddS1}(t_{S1} | k_{off}, k_{on}^*, t_{int}) = k_{off} e^{-k_{off} t_{S1} - k_{on}^* t_{S2}} I_0\left(2\sqrt{k_{off} k_{on}^* t_{S1} t_{S2}}\right), \quad (37)$$

$$W_{evenS1}(t_{S1} | k_{off}, k_{on}^*, t_{int}) = \sqrt{\frac{k_{off} k_{on}^* t_{S1}}{t_{S2}}} e^{-k_{off} t_{S1} - k_{on}^* t_{S2}} I_1\left(2\sqrt{k_{off} k_{on}^* t_{S1} t_{S2}}\right). \quad (38)$$

Here, t_{S1} and t_{S2} are times spent in state $S1$ and state $S2$ and I_0 and I_1 are Bessel functions of order zero and one, respectively. Note that $t_{S1} + t_{S2} = t_{int}$. Equations

for starting in state *S*2 (W_{contS2}, W_{oddS2} and W_{evenS2}), can be found by exchanging k_{off} for k^*_{on} and t_{S1} for t_{S2} and vice versa in equations 2-4.

To correctly describe the distribution over a certain number of frames, we first calculated the distribution over a single time frame t_f. Within a single frame, a particle started in that state can either end in the same state or in a different state. Therefore, in a two-state system the probability functions for four scenarios have to be calculated

$$W(t_{Si}|k_{off}, k^*_{on}, t_f)_{S1\to S1} = W_{evenS1}(t_{Si}) + W_{contS1}, \quad (39)$$

$$W(t_{Si}|k_{off}, k^*_{on}, t_f)_{S1\to S2} = W_{oddS1}(t_{Si}), \quad (40)$$

$$W(t_{Si}|k_{off}, k^*_{on}, t_f)_{S2\to S1} = W_{oddS2}(t_{Si}), \quad (41)$$

$$W(t_{Si}|k_{off}, k^*_{on}, t_f)_{S2\to S2} = W_{evenS2}(t_{Si}) + W_{contS2} \quad (42)$$

for i = 1,2.

To link the distribution of times spent in a state to the distribution of measured displacements (x), we can convert the time spent in each state and its diffusion coefficient to the average diffusion coefficient by the following equation

$$D = D_{S2}\frac{t_{S2}}{t_{int}} + D_{S1}\frac{t_{S1}}{t_{int}}. \quad (43)$$

In case of the transition between an immobile bound state *S*1 (D_{S1}= 0) and a mobile state with diffusion coefficient $D_{S2} = D_{free}$we can modify the above equation to

$$D = D_{free}\frac{t_{S2}}{t_{int}}. \quad (44)$$

AnaDDA is able to fit systems with two mobile states, but for the rest of the manuscript (except Figure S3) we analyse systems with an immobile state and use equation 10.

Using equation 10, the probability distribution function (equation 1) can be modified according to

$$f_D(x; t_{S2}, D_{free}, n) = \frac{\left(\frac{n}{D_{free}\frac{t_{S2}}{t_{int}} + \sigma^2/t_{int}}\right)^n x^{n-1} e^{-\frac{nx}{D_{free}\frac{t_{S2}}{t_{int}} + \sigma^2/t_{int}}}}{(n-1)!}. \quad (45)$$

Subsequently, the probability to find a certain diffusion coefficient (x) for a single time step given the time spent in the mobile state is given by $f_D(x|t_{S2}, 1)$. We can

then find the distribution of measured diffusion coefficients for a single frame by integrating over all possible times spent in the mobile state

$$W\left(x\middle|k_{off}, k_{on}^{*}, D_{free}, t_f\right)_{Si\to Sj} = \int_0^{t_f} f_D(x|t_{S2}, 1)\, W\left(t_{S2}\middle|k_{off}, k_{on}^{*}, t_f\right) dt_{S2}\,{}_{Si\to Sj} \quad (46)$$

i = j = 1,2.

Now that we have the distribution for a single time step, we need to find the distribution for the average of multiple frames. For this we use the same method as Qian *et al.*(Qian et al., 1991), namely repeated convolution of the distribution for a single frame, while keeping track of the start and end state. The probability distributions are therefore

$$W(x|2t_f)_{S1\to S1} = \sum_{i=1,2} (W(x|t_f)_{S1\to Si} * W(x|t_f)_{Si\to S1}), \quad (47)$$

$$W(x|2t_f)_{S1\to S2} = \sum_{i=1,2} (W(x|t_f)_{S1\to Si} * W(x|t_f)_{Si\to S2}), \quad (48)$$

$$W(x|2t_f)_{S2\to S1} = \sum_{i=1,2} (W(x|t_f)_{S2\to Si} * W(x|t_f)_{Si\to S1}), \quad (49)$$

$$W(x|2t_f)_{S2\to S2} = \sum_{i=1,2} (W(x|t_f)_{S2\to Si} * W(x|t_f)_{Si\to S2}). \quad (50)$$

For a track consisting of 4 frames, the distributions found for 2 frames can be convoluted again. The full distribution is then found by summing up each of the partial distributions multiplied by the chance they start in *S*1 or *S*2:

$$W_{tot} = p_{S1}(W(x|4t_f)_{S1\to S2} + W(x|4t_f)_{S1\to S1}) + p_{S2}(W(x|4t_f)_{S2\to S1} + W(x|4t_f)_{S2\to S2}), \quad (51)$$

with p_{S1} and p_{S2} defined in equations 18 and 19, respectively:

$$p_{S1} = \frac{k_{on}^{*}}{k_{on}^{*} + k_{off}}, \quad (52)$$

$$p_{S2} = \frac{k_{off}}{k_{on}^{*} + k_{off}}. \quad (53)$$

Localization error

As two consecutive steps share at least one localisation, the localisation error of this localisation leads to a correlation between the measured displacements (Michalet, 2010). Only in the special case of the localisation error being zero, the measured displacements are uncorrelated. The distribution of the sum of displacements for a certain number of steps is therefore not described by a gamma distribution, which is the sum of independent variables. However, as each step separately is a gamma random variable, we calculate the summation of correlated gamma random variables to describe the distribution of localization error analytically for different number of steps. For derivations, see supplemental section *Derivation of D* distributions of localization error*.

Tracking window

In order to the prevent the accidental linking of different diffusing particles, many tracking algorithms use a certain cut-off, in which steps longer than a certain distance are not allowed (Hansen et al., 2018; Lee and Park, 2018; Uphoff et al., 2014). However, this tracking window can influence the distribution of *D* values recovered. In analytical DDA, we correct for this by setting $f_D(x > \mathrm{maxD}|D_i, 1) = 0$, where *maxD* is the maximum *D** value that can be obtained given the tracking window.

Confinement

To take the effects of geometrical confinement within the cell into account, we implemented an analytical way to calculate the effective diffusion coefficient given the geometry and the real diffusion coefficient. Most boundary geometries encountered in *in vivo* settings are either spherical or rod-shaped. For a spherical geometry, the effective measured MSD given a diffusion coefficient *D* and a timestep Δt have been previously derived for multiple dimensions (Bickel, 2007). We have used these equations to find $D_{obs} = f_{boundary}(r, t, D)$, which is the observed diffusion coefficient given a certain boundary condition (spherical/rod-

shaped), the boundary radius r, the frametime t and the real diffusion coefficient D. For derivations, see supplemental section *Derivation of confinement corrections*.

Maximum likelihood estimation (MLE)

To find the underlying parameters of experimental data and simulations, we use MLE which maximizes the joint probability of observing by iteration through the parameter space. Generally, MLE requires a probability density function to calculate and sum up all probabilities of each observed data point. The benefit of the method is that it does not require any binning, compared to other optimization methods. However, MLE does require the exact probability for each data point to be calculable. Because we use numerical convolution (for increasing the performance of the algorithm, we implemented a FFT convolution (Smith, 2003)), we will only get the probability at discrete points within the probability density function. Therefore, to calculate the probabilities for the points of our data set, we use spline interpolation.

Because MLE is known to be affected by local minima (Myung, 2003), we use a number of cycles (generally four) in which we generate random starting parameters and run the algorithm several times after which we select the end parameter set with the maximal likelihood. Those parameters are then used as starting parameters for bootstrapping in which we run the analysis through a number of subsets of the data to get an estimate of the standard deviations of our parameter estimates.

Plotting of diffusion distribution histograms

With the parameter sets used in our simulations, the diffusion histograms are visually more distinguishable when $\log(D^*)$ is plotted compared to D^*. We therefore integrated the linear density function with widths specified by the bin size of the logarithmic scale to calculate the probability density function for $\log(D^*)$ instead of D^*.

Results

AnaDDA generates *D** distributions equal to the ground truth of simulated distributions

AnaDDA allows calculating the shape of the *D** distribution, depending on the free diffusion coefficient (the diffusion coefficient in absence of binding interactions) and the transition rates. As this shape depends on the number of localizations per track, we separate the tracks according to their respective length and fit each data point to the distribution that matches their step length. To benchmark our new analysis method, we first compared our theoretical predictions of the *D** distribution to data in which we simulated the diffusional characteristics of a particle that dynamically switches between a (DNA-) bound state and a freely diffusing state without including any boundary conditions for diffusion (see section below for confinement within cells). With increasing number of tracks, the predicted *D** distribution increasingly resembles the predicted theoretical distribution (Figure 1C). To test whether our theoretical distributions differed from the simulated ground truth, we performed Kolmogorov-Smirnov tests. We found that the test statistic D_{KS} converged to zero for larger number of tracks analysed and was on average smaller than the critical value required to reject the null-hypothesis (D_{KS} = 0.004 for $p < 0.05$), indicating that the ground-truth simulations and our theoretical predictions come from the same distribution (Figure 1D).

We varied the range of transition timescales (Figure 1E) ranging from 0.01 to 10 transitions per frame (at 0.01 s frame time) at all different step numbers per track included in this analysis (1-8; Figure S1A) and compared a range of frame times (20-100 Hz) and experimentally realistic localization errors (20-50 nm) (Figure S1B). Under all these conditions, the ground truth simulations (N = 100.000 tracks) and the anaDDA generated distributions showed very close agreement ($D_{KS} < 0.004$). As this analysis involved a direct comparison between the predicted and simulated distribution without fitting the data or any optimization of parameters, it can be concluded that our theoretically predicted distributions are similar to the ground truth distributions.

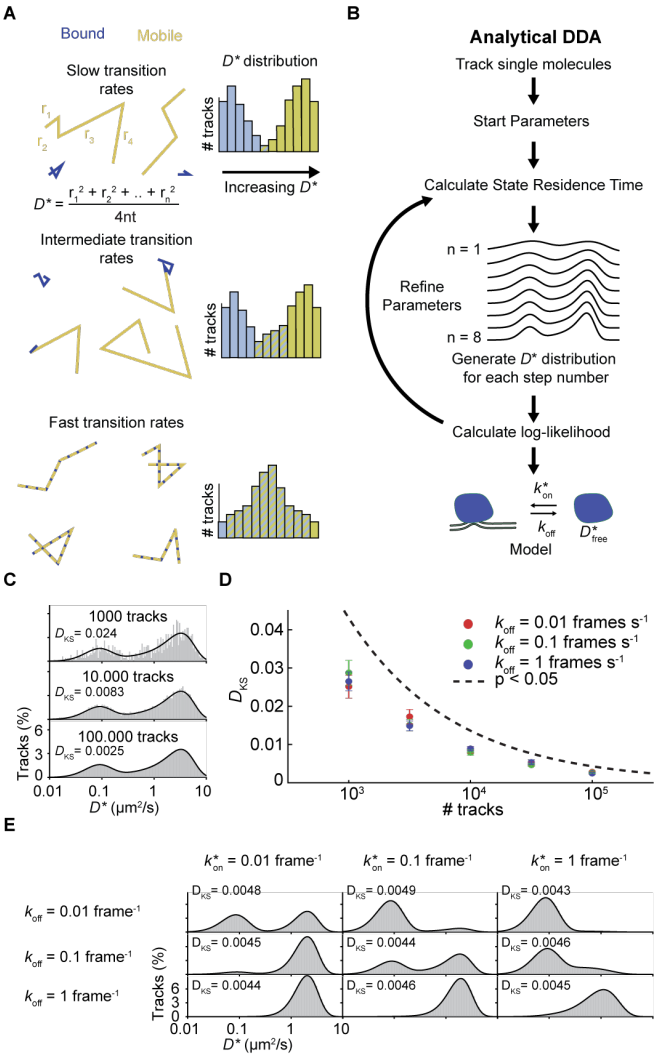


Figure 1. Analytical DDA (A) The effect of transition rates on *D** distributions is depicted with simulated tracks of four steps and different transition rates. With increasing transition rates relative to the frame rate, the bound and unbound distributions start merging towards an intermediate apparent diffusion speed diffusivity (right). The distribution of apparent diffusion coefficients *D**, calculated from the mean jump distances of a track, originates from the finite number of steps (n) that are measured for each particle and allows the extraction of the underlying diffusion coefficient and transition kinetics of the states. **(B)** Procedure of analytical DDA: The *D** values from tracked single particles are run into an MLE optimization program which refines a set of start parameters based on the likelihood to find a certain value given the number of steps (all tracks longer than 8 are reduced to the first 8 steps). **(C)** Comparison of simulated (grey bars) and theoretically predicted (black line) distribution with different amount of tracks and the following starting parameters: k^*_{on} = 0.2 frame^{-1}, k^*_{on} = 0.2 frame^{-1}, D_{free} = 4 μm^2/s and σ = 30 nm (localisation precision), step number = 4 steps. Tracks are simulated without any confinement boundaries. The Kolmogorov-Smirnov test statistic (D_{KS}) is indicated at each histogram. **(D)** The Kolmogorov-Smirnov test statistic compared to the threshold for statistically distinguishable distributions. Values above the threshold line indicate that two distributions significantly differ from each other. Error bars indicate S.E.M. of three independent simulations. **(E)** Comparison of simulated *D** distributions (grey) and the distributions calculated with analytical DDA for different transition rates (black). The shape of the distributions depends on both the ratio between k^*_{on} and k_{off} and the absolute values of these parameters. In this example D_{free} = 4 μm^2/s and σ = 30 nm. For more tested parameters see Figure S1.

AnaDDA can extract transition rates from tracks with more than one transition per frame

With data from experimental measurements, the ground truth is unknown, and parameters have to be inferred by fitting. First, we tested via simulations how reliably parameters can be extracted over a large dynamic range of transitions. We compared the input parameters to the extracted ones with Maximum Likelihood Estimation (MLE). To benchmark the performance of extraction we calculate the accuracy through the geometric mean and the precision through the geometric standard deviation of 10 independent simulations. For all tested data sizes (5000-100.000 tracks) and transition rates (0.001-10 transitions per frame), the analysis method is accurate (< ± 5% of input parameters). The precision decreased slightly with decreasing data size and for small/large transition rates (Figure 2). Furthermore, the precision at high transition rates (>1 transition per frame) is lower for k_{on}^{*} than k_{off} (Figure 2A-B). In general, the highest precision is found for tracks between 0.1 and 1 transition per frame. With 50.000 tracks per simulation, the transition rates over three orders of magnitude (0.002-2 transitions per frame) were determined with an error smaller than 20% of the actual value (Figure 2A-C).

We compared our method with a previously published framework that used Bayesian statistics to infer transition and diffusion dynamics (vbSPT) (Persson et al., 2013) and a framework that used unsupervised Gibbs sampling for similar purposes (SMAUG) (Karslake et al., 2020). As vbSPT and SMAUG deduce the number of states from the data, we limited the amount of states in this analysis software to two to achieve a fair comparison. For slow transitions (<0.01 transition per frame) both anaDDA and vbSPT were able to extract the correct kinetic parameters for data sets containing 50.000 tracks (<20% error; Figure 2D-F), whereas SMAUG overestimated the transition rates. At faster transitions (> 0.02 transitions per frame), however, we observed a decrease in the extracted free diffusion coefficient and a decrease in the extracted on and off-rates for both vbSPT and SMAUG. A similar trend was observed for data sets containing only 1000 tracks (Figure S2A).

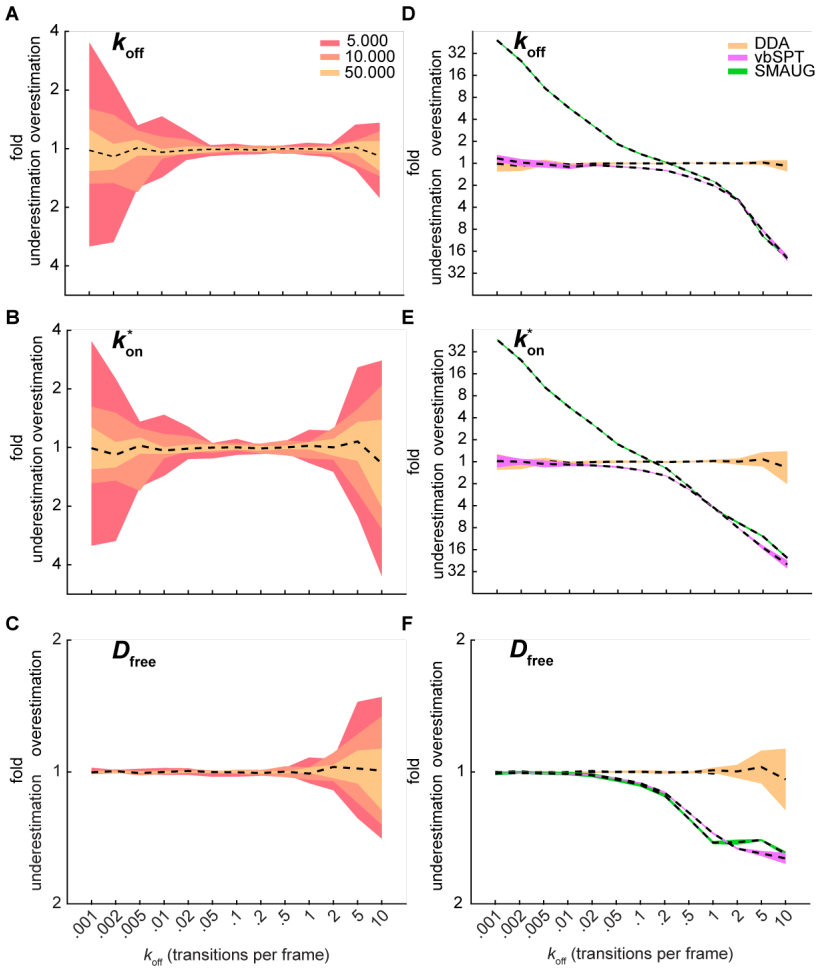


Figure 2. MLE extraction of parameters. The accuracy is calculated through the value of the geometric mean (dashed black line) and the precision is calculated through the geometric standard deviation of 10 independent simulations. The step number per track was exponentially distributed with a mean of three steps and a cut off at 8 steps (D_{free} = 4 μm²/s, σ = 30 nm). **(A-C)** Effect of data size on accuracy and precision of extraction of (A) k_{off}, (B) k^*_{on} and (C) D_{free} for *N* = 5.000 tracks (red), 10.000 tracks (orange) and 50.000 tracks (yellow). (**D-F**) Comparison of anaDDA versus vbSPT and SMAUG on accuracy and precision of extraction of (D) k_{off}, (E) k^*_{on} and (F) D_{free}. 50.000 tracks were used for both methods.

We furthermore compared the different analysis methods in the presence of tracking errors, arising from high density measurements, where tracks from different particles are erroneously linked. We simulated tracks occurring simultaneously in increasing densities (0.01 to 0.25 particles per μm²). Subsequently, we linked the localisations using a previously described tracking algorithm that uses minimization of the total

squared displacement of all possible trajectories within a given tracking window (Crocker and Grier, 1996). For most timescales, anaDDA can still extract the correct parameters, but for low and high transition rates, the extraction is sensitive to the tracking errors occurring at high densities (0.1 to 0.25 particles per μm^2; Figure S2B). At low transition rates (0.001-0.05 transitions per frame), the transition rates were overestimated and at high transition rates (2 – 10 transitions per frame) the on-rate and free diffusion coefficient were overestimated. When vbSPT and SMAUG were tested with simulations at the highest densities (0.25 particles per μm^2), the extracted kinetic parameters were even further away from the ground truth. Our simulation shows that in order to robustly extract kinetic parameters, localization densities should be kept low (<0.1 per μm^2).

When we removed the restriction of a two-state model, vbSPT started introducing multiple false states (Figure S3A). Already at low transition rates (0.01 transitions per frame), vbSPT suggests the presence of a false third state. At this transition rate, two states (0.06 and 0.11 $\mu m^2/s$) were close to the expected average diffusion coefficient of the simulated immobile state (σ^2/t = 0.09 $\mu m^2/s$). The highest number of predicted states (4 states) was found for transition rates between 0.05 and 0.5 transitions per frame. To see whether anaDDA also would fit more false states, we tried to force a second dynamic species (Figure S3B). In this case, the second species fraction was found to have zero amplitude, indicating that under the tested conditions anaDDA would not introduce a false state.

So far, we have limited the analysis to systems for which one of the states is immobile, but anaDDA can also be applied to systems with two mobile states. We expected that the extraction of parameters would be less accurate for these systems, firstly because a new parameter needs to be extracted from the data and secondly because the overlap of D^* distributions from two mobile distributions tend to overlap more than distributions of a mobile and an immobile state (Figure S3C). We found that under these conditions anaDDA still performs well in the range 0.01 to 2 transitions per frame (less than 20% error with 50.000 tracks) but that parameters extracted from lower or higher transition rate simulations are less accurate compared to systems with an immobile state (Figure S3D-G). Under the same simulation conditions, vbSPT and SMAUG overestimate the transition rates at low transition

rates (>4x at 0.001 transitions per frame) and underestimate at high transition rates (>20 x at 10 transitions per frame).

Our findings suggest that vbSPT and SMAUG fail to account for the increasing occurrence of multiple transitions within a single frame at fast transition rates. Our analysis software is distinctive in its ability to extract kinetic parameters when multiple transitions are likely to occur within a single track. In fact, anaDDA can validate whether a simple two-state model with fast transitions is sufficient to explain the data, whereas vbSPT and SMAUG would introduce virtual static or slowly interconverting states. To further improve the applicability of anaDDA to real experimental sptPALM data, we wanted to correct for artefacts that can influence diffusion distribution analysis, namely confined diffusion within cells and application of tracking windows.

AnaDDA corrects for confinement within cells and restricted tracking windows

To study the effect of geometrical confinement, we simulated diffusive particles within the confined boundaries of different cell shapes. We previously showed that confinement only has a very small effect on observed transition rates in bacterial cells (Vink et al., 2020). However, as the measured diffusion coefficient can be greatly affected by confinement, we implemented an algorithm based on previously developed derivations (Bickel, 2007) (for details see Materials and Methods) to account for confinement in both rod-shaped (e.g. *E. coli* cells) and spherical-shaped boundaries (e.g. eukaryotic nuclei) (Figure 3A).

For both spherical and rod-shaped cells (cell length : radius = 8:1) we found that our theoretical predictions for varying cell sizes ($r^2 = (2, 5, \text{or } 20)D_{free}t$) match well with simulated data (Figure 3B-C; $D_{KS} < 0.006$) in contrast to uncorrected distributions for which the predicted distributions are statistically different from the simulated distributions ($D_{KS} > 0.04$). In an *E. coli* cell ($r = 0.5$ μm) and under standard measurement frame times (0.01 s), these confinement regimes ($D_{free}t$) would be reached with D_{free} values of 12.5 $\mu m^2/s$ respectively, which matches the values found for small single fluorescent proteins

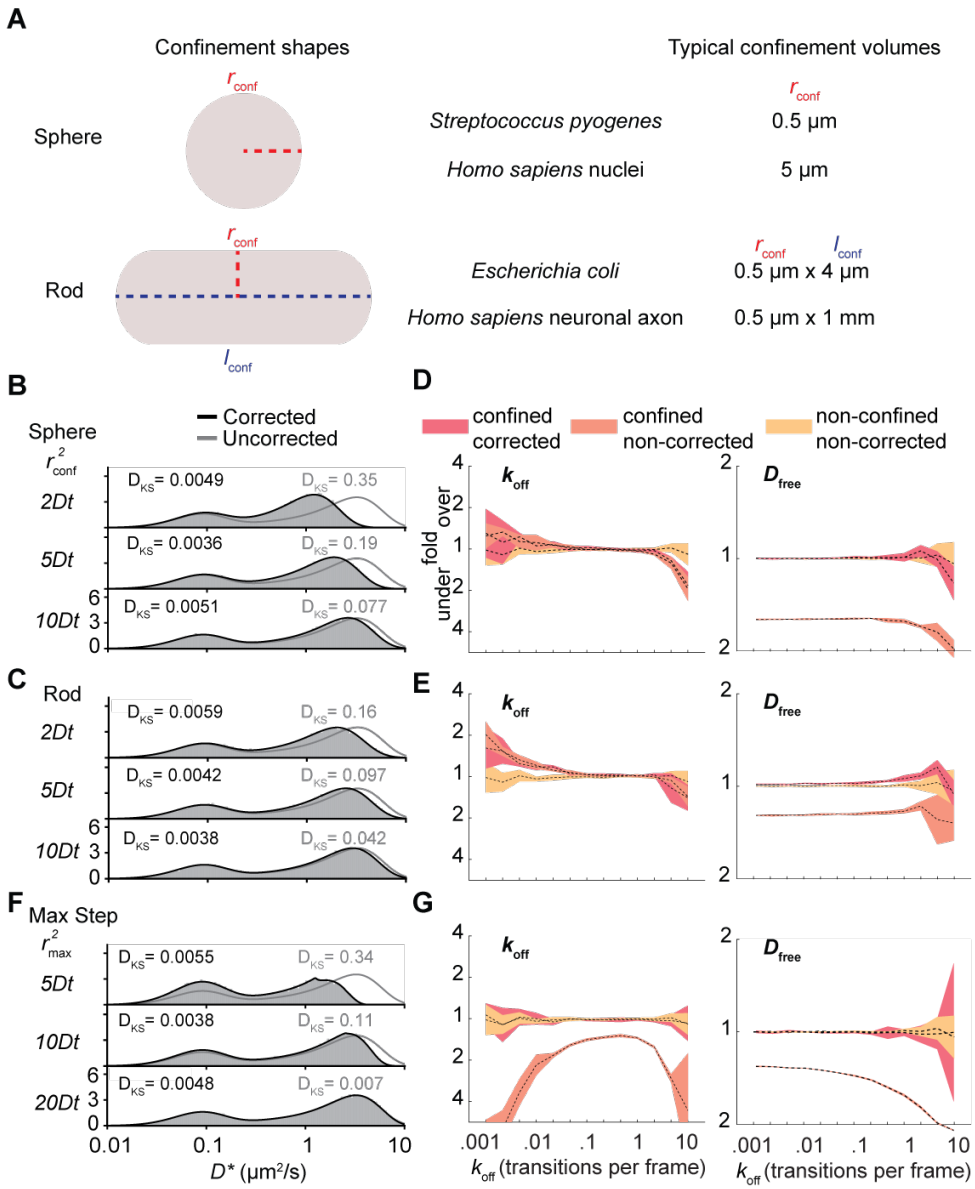


Figure 3. Effects of geometrical confinement and the length of the tracking window. **(A)** Typical confinement shapes within cells. The boundary shape of spherical cells is defined by a single parameter (radius; $r_{conf.}$), whereas rod-shaped cells are defined by two parameters (radius and length; $r_{conf.}$ and $l_{conf.}$). **(B-C)** Influence of spherical and rod-shaped boundaries on the distribution of simulated (grey box) and uncorrected DDA (grey line) and corrected DDA (black line) distributions. (k_{off} = 0.2 frame^{-1}, k^{*}_{on} = 0.2 frame^{-1}, step number = 4 steps**)** **(D-E)** Influence of spherical and rod-shaped cells on the estimation of parameters of DDA on unconfined simulated trajectories (yellow), uncorrected DDA on confined simulated trajectories (orange) and corrected DDA on confined trajectories (red). **(F-G)** Same as B-E except for simulated trajectories with a maximum step size. Simulation parameters: D_{free} = 4 µm²/s and σ = 30 nm (localisation precision), *N* = 50.000 tracks. The Kolmogorov-Smirnov test statistic (D_{KS}) is indicated in each histogram.

(Woodside et al., 2006). In a eukaryotic nucleus ($r = 5$ μm), these regimes would correspond to D_{free} values up to 750 $\mu m^2/s$ which is generally much faster than any reported literature values. This finding indicates that geometrical confinement by cell boundaries is mostly limiting in prokaryotic studies. However, at longer frame times (0.1 s) confinement effects will play a role when studying diffusion within eukaryotic nuclei.

As not every cell in a population is the same size, the distribution might be further affected by a variation of cell sizes. We therefore analysed a mixture of three different simulated cell sizes and found that the distributions remained statistically indistinguishable from a uniform population of the same cell size (Fig. S4; $D_{KS} < 0.006$). This shows that the correction method remains valid as long as the average dimensions of the cell boundaries are known.

To further test our ability to infer parameters from the data in a system where diffusion is geometrically confined, we performed MLE with and without corrections for confinement. We observe that the incorporation of our confinement corrections increases the accuracy and precision of the estimation of D_{free} (Fig. 3D-E). Compared to unconfined diffusion, there is a bias in recovered transition rates at very small and large transition rates, as these regimes are most sensitive to small deviations of the predicted distribution to the ground truth. These minor deviations are most likely caused by a correlation which occurs for diffusing particles within boundaries, where particles that are close to the boundary in one frame, are again likely to encounter the boundary in the next frame. That effect is not taken into account in our current implementation. However, for most transition regimes (0.01-2 transitions per frame), the error of the estimated parameters falls within 20%.

Another type of analysis artefact comes from the settings for tracking windows. When the density of labelled fluorophores is higher than 1 per cell, different molecules can be falsely assigned to the same track. To prevent this effect, multiple tracking software algorithms set a limit to the maximum step length that individual tracks are allowed to have. Although this is sometimes unavoidable, the absence of the largest steps can severely affect the MLE fitting parameters. AnaDDA is able to correct for this, by integrating this max displacement in the probability distribution

(see Materials and Methods). The effect of this correction was tested for a range of radii of tracking windows ($r^2 = (5, 10, \text{or } 20)D_{free}t$) and in all cases the D_{KS} of the corrected distributions were below the threshold for significantly different distributions (D_{KS} = 0.006), whereas for small and intermediate tracking windows ($r^2 = (5 \text{ and } 10)D_{free}t$) uncorrected distributions were significantly different (D_{KS} = 0.34 and D_{KS} = 0.11; Figure 3F-G). The tracking window also had a large effect on both the predicted transition rates and free diffusion coefficients from MLE, where in the absence of corrections all parameters were significantly underestimated (>1.5x). With the correction, the estimations were again unbiased and very similar to the accuracy and precision of estimations in the absence of tracking windows.

Taken together, anaDDA can correct the distributions for measurements that are affected by confinement within spherical and rod-shaped boundaries and by the application of a maximum step size within tracking algorithms. Because these artefacts cause a non-linear relationship between the MSD and the timestep in a similar fashion as anomalous diffusion (Robson et al., 2013), it allows the user to validate whether a simple Brownian model with confinement is able to explain the data before assuming more complex modes of diffusion.

AnaDDA can be expanded for multiple states and can integrate multiple frame times

So far, we have discussed the presence of one diffusing species converting between two diffusional states. In the following, we will expand the DDA-fitting to account for more species and states.

Many DNA binding proteins contain both non- and target-specific interactions with DNA. Therefore, it is likely that the kinetics of these two interactions are different, which would require the model to be expanded beyond a two-state model. PDA statistical analysis currently does not incorporate more than two dynamic states. However, it is possible to incorporate more states by assuming that their dynamics are much slower than the non-specific DNA interactions, which would result in a negligible amount of transitions in the timeframe studied. Then these states can be approximated by separate static (non-interchanging) species (Figure 4A). Generally the specific interactions are much longer lasting than the non-specific interactions (Slutsky and Mirny, 2004), so in many cases this assumption would be valid.

To test how well this approximation works and how well the method can distinguish this model from a simple two-state model, we simulated a linear (A ⇄ B ⇆ C) three-state model containing one slow transitioning bound state ($k^*_{on,1}$ = 0.005 frame^{-1}, $k_{off,1}$ = 0.01 frame^{-1}) and one fast transitioning bound state ($k^*_{on,2}$ = 0.2 frame^{-1}, $k_{off,2}$ = 0.2 frame^{-1}). We compared this simulation to our theoretically predicted distribution where we approximated the slower transitioning state as a separate immobile species and the faster transitioning state as a separate species (Figure 4B). The fraction of the approximated immobile species (20 %) and transitioning species (80 %) can be calculated from the ratio of the on- and off-rates (Figure 4B). We found very good agreement between the theoretical prediction and the simulation (D_{KS} < 0.006) indicating that this approximation can be applied in this case.

We then tried to find whether a single species two-state model could also fit the distribution of the three-state model (Figure 4C). We found that although for smaller tracks there are parameters that can fit the distribution quite well (D_{KS} = 0.0078 for step number of 4 steps), the distribution for larger tracks significantly deviated from the ground truth (D_{KS} = 0.0149 for step number of 8 steps). Therefore, with a sufficient number of longer tracks two-state and three-state models are clearly distinguishable.

We then tested under which conditions the parameters can be reliably extracted from the data. To this end, we varied the transition rates of the fast-bound state ($k^*_{on,2}$ and $k_{off,2}$) while keeping the slower bound state fixed. We observed that under all transition rates tested (0.1-10 transitions per frame), the error of the estimated parameters falls within 25% and that with increasing rates of the fast-bound state, the extraction of the fraction parameter became more reliable (Figure 4D). This finding indicates that as long as the transition rates associated with the different bound states are different enough (>10 fold), with one of them being significantly slower than the frame time used in the measurements, parameters for three state models can be reliably extracted with anaDDA.

More complex models with larger number of species, each having up to three states and meeting the requirements described above can also be fitted using anaDDA but are prone to increased uncertainty and under-/overfitting as many parameters in

these models could give rise to similar distributions. We therefore advise users to fit models with a maximum of four free parameters, when the data was recorded using a single frame time. To overcome this limitation, we implemented the ability to use data acquired at different frame times into a single global fit. By fitting data from multiple frame times simultaneously, the number of potential parameters that can fit all the data decreases, leading to more accurate and precise fitting for more complex models.

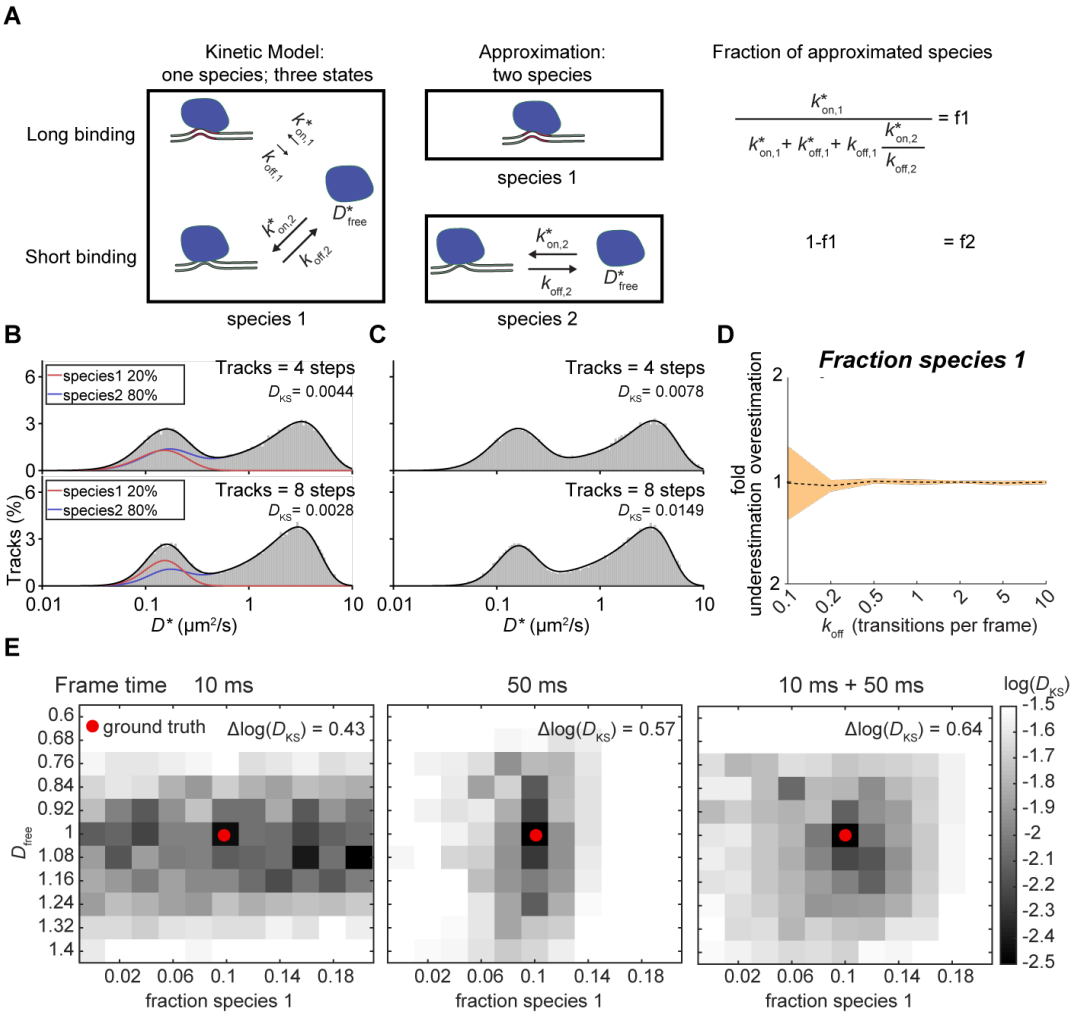


Figure 4 Three-state models and multiple frame times (A) Three-state models cannot be directly described with PDA statistics. If some interactions are slower than the typical frame time, however, the approximation can be made that they belong to a non-transitioning separate species. The expected fraction of each of this species can be calculated from the on- and off-rates of all states (right). (**B**) Comparison of a simulated three-state model ($k_{off,1}$ = 0.01 frame^{-1}, $k^*_{on,1}$ = 0.005 frame^{-1}, $k_{off,2}$ = 0.2 frame^{-1}, $k^*_{on,2}$ = 0.2 frame^{-1}) with a predicted theoretical approximated two-species model, where the slower transitioning state is approximated as a separate immobile species

(red) and the other species (blue) still contains two states with $k_{off,2}$ and $k^{*}_{on,2}$ as transition rates. Upper panel step number of 4 steps, lower panel step number of 8 steps. (**C**) Best fit of the simulated three-state model from (B) with a single-species two-state model. (**D**) MLE extraction of the expected fraction of the first approximated species for different values of $k_{off,2}$ (**E**) Heat map of the log(D_{KS}) between a simulated distribution (D_{free}= 1, fraction immobile = 0.1; ground truth (red dot)) and a theoretical predicted distributions with varying parameters around the parameters used for the simulation, where the simulation consisted either of 100.000 tracks at 10 ms frame time (left), 100.000 tracks at 50 ms frame time (middle) or 50.000 tracks at 10 ms frame time and 50.000 tracks at 50 ms frame time respectively (right). The discrete Laplacian Δlog(D_{KS}) calculated from the ground truth coordinate, is the sum of the second derivatives in both dimensions and indicates how quickly log(D_{KS}) increases with parameter values slightly different than the ground truth. The Kolmogorov-Smirnov test statistic (D_{KS}) is indicated in each histogram.

As an example, we simulated a two-species (one immobile, one transitioning) model and calculated the Kolmogorov-Smirnov test statistic (D_{KS}) for a range of parameters around the input parameters for a simulated dataset consisting of tracks either measured at a single frame time (10 or 50 ms) or a combined set where halve of the dataset contained simulated tracks from each frame time (Figure 4E). If there are other closely related parameters with similar D_{KS} values to the ground truth, the fit can converge to these values as well. Therefore, the uncertainty is linked to the parameter space with D_{KS} similar to the D_{KS} of the ground truth. We observed that different frame times perform better on different parameters. In particular, short frame times led to more uncertainty in the determination of the fraction of each species, whereas long frame times gave more uncertainty in the determination of the free diffusion coefficient. When data recorded at different frame times is combined, there is only a single set of parameters that give rise to a similar D_{KS} as the ground truth. To quantify the benefit of the combination of frame times, we calculated the discrete Laplacian score (Δlog(D_{KS})) from the ground truth coordinate. The score is the sum of the second derivatives in both dimensions and indicates how quickly log(D_{KS}) increases when moving away from the ground truth. We found that the combined frame times of 10 and 50 ms data had a higher score (0.64), compared to datasets from either frametime alone (0.57 for 50 ms and 0.43 for 10 ms datasets), indicating that datasets with more than one frame time outperform datasets recorded

at a single frame time. In conclusion, the benefit of gathering data with different frame times is that it reduces the parameter space that can simultaneously fit multiple distributions and therefore offers better performance with the same number of data points.

E. coli DNA polymerase I undergoes rapid DNA interactions

To test the applicability of our analysis method to experimental data, we re-analysed previously published data on the diffusion of DNA polymerase I in *E. coli* (Uphoff et al., 2013). In this study, the diffusion distribution of PAmCherry-Pol1 was grouped into immobile and mobile diffusing particles by simple thresholding without determination of any transition kinetics. The authors found that under normal conditions only 4-5% of the proteins were immobile. However, they found that even the mobile tracks were mostly located within the nucleoid, which may suggest that these tracks represent transient DNA binding, probably probing the DNA for repair sites. We therefore hypothesized that the previously assigned mobile fraction is also undergoing rapid transitions between DNA bound and freely diffusing states.

We decided to fit the data with two species, one belonging to proteins involved in repair (a species with a single bound state) and one to probing (a species with a bound and a freely diffusing state). When we fitted this model (two species and three states; Figure 5A) we found a similar percentage of proteins involved in repair as described in the previous study (4%; Figure 5B). Furthermore, we found that the probing species had a free diffusion speed of 2.8 (± 0.2) $\mu m^2/s$ in the cytoplasm and that it is involved in vary rapid DNA probing events (k_{off} 137 ± 7 s^{-1}; k^*_{on} 155 ± 25 s^{-1}). Based on the on and off-rates we calculated that the probing species spends more than half the time (~ 55%) bound to DNA. Altogether, DNA polymerase spends approximately ~ 60% bound to DNA either in repair (4%) or probing for mismatch sites (55%).

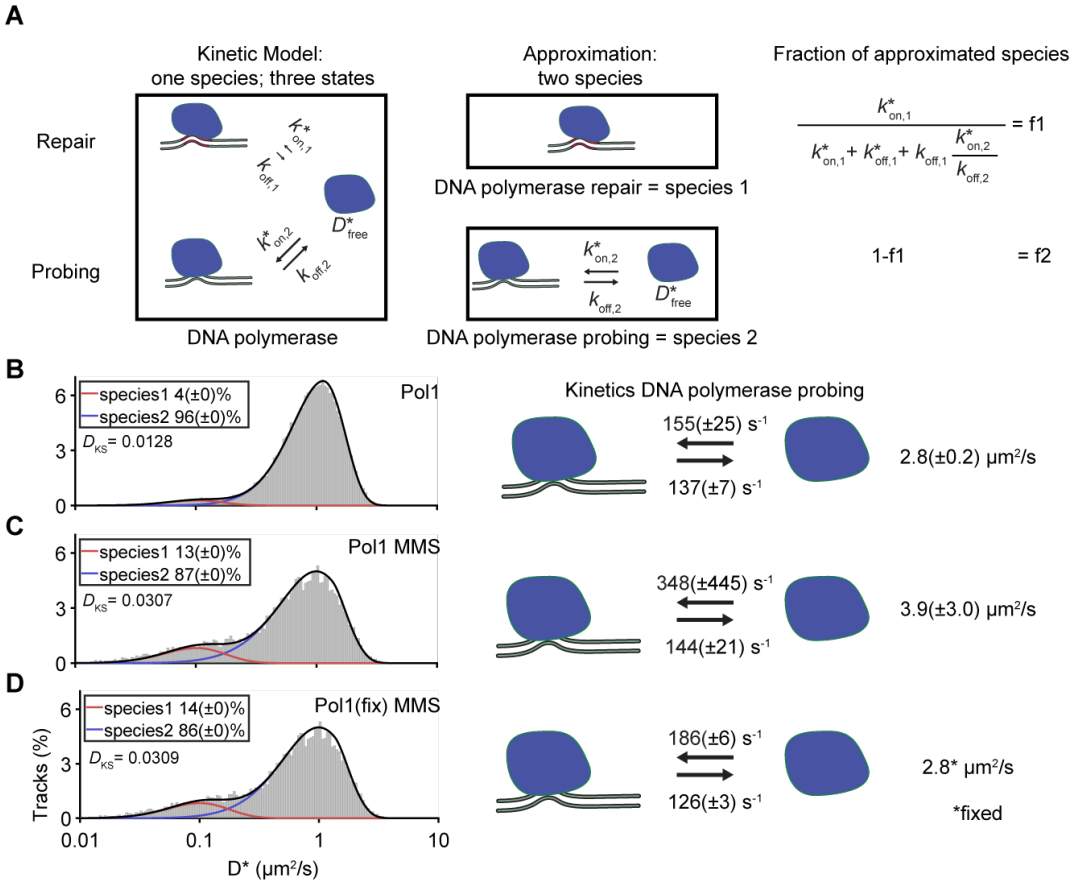


Figure 5. Extracting kinetic information of DNA polymerase I diffusing in live *E. coli* cells. (A) Approximated model of the kinetic model of DNA polymerase diffusion containing a DNA repair and DNA probing state (left). These states were separated into a single-state repair species (species 1; middle) and a probing species with two states (species 2; middle) The fraction of the two species are caused by the underlying ratios of the on- and off rates (right; Fig. 4A) **(B)** Fit of DNA polymerase I in untreated cells (n = 179.511 tracks). The *D** was fit with two species, one species involved in repair (red line) with a single state (immobile) and one species involved in scanning DNA with two states (mobile and immobile; blue line). The transition between the latter two states and the free diffusion coefficient of the mobile state are depicted. Fit was performed on all different step numbers (1-8 steps) and histograms are only shown for a single step number (4 steps). Tracks with 8 or more steps were truncated to 8 steps for the entire fit. For the histogram, *D** calculated from tracks truncated to 4 steps are shown. **(C)** Same as B but performed on data of DNA polymerase in cells treated with MMS. **(D)** Same as C except that the free diffusion coefficient of the mobile state was fixed to the same value as was found for polymerase in untreated cells (B). The Kolmogorov-Smirnov test statistic (D_{KS}) is indicated in each histogram and uncertainty in parameters were estimated with bootstrapping (±SD). Experimental data was taken from a previous study (Uphoff et al., 2013).

The study also measured the diffusivity of DNA polymerase in presence of the DNA damaging agent MMS. Using anaDDA, we found that the immobile species increased to 13% which matches the findings in the publication (13 ± 0.2%; Figure 5C). The transition rates and diffusion coefficients under this condition could not be assigned with confidence based on the bootstrap values (k_{off} 137 ± 7 s^{-1}; k^*_{on} 348 ± 25 s^{-1}). values. We hypothesized that this is caused by the lower number of available tracks (41.415 tracks) compared to the untreated dataset (142.178 tracks).

To quantitatively assess the transition kinetics in presence of DNA damage, we made the assumption that DNA damage would not alter the free diffusion behaviour of DNA polymerase in the cytoplasm but only the kinetics of the interactions with DNA. We therefore fixed D_{free} to the value found for DNA polymerase in untreated cells (2.8 $\mu m^2/s$; Figure 5D) which caused the fitting to converge to a narrow range of transition rates. We observed that although the k_{off} remained the same (126 ± 3 s^{-1}), the on-rate increased in the presence of damaged DNA (185 ± 6 s^{-1}) indicating that more DNA polymerases were bound to DNA in long-term repair events (from 4 to 13%) and that also the polymerases engaged in probing spent more time bound to DNA. Altogether, these numbers would indicate that DNA polymerase in the presence of MMS spent ~ 75% of its time to DNA either at a repair site (13%) or while probing the DNA (60%).

We further found that the maximum step size of 5 pixels used in the original analysis significantly affected the distribution of observed *D** values (Figure S5). AnaDDA was able to correctly predict and take this effect into account. Overall, the transition rates between bound and unbound polymerase found under both conditions are high compared to the frame rate (>1 transition per frame), which demonstrates the applicability of anaDDA to quantify very fast transition kinetics *in vivo*.

Discussion

Analytical diffusion distribution analysis (anaDDA) is able to accurately extract kinetics occurring within 4 orders of magnitude with around 10 to 0.01 transitions per frame. With conventional camera frame rates of 100 Hz, this range translates to interaction kinetics of 1 ms to 1 s even if the mean track length is as short as 3-4 frames. Furthermore, anaDDA is able to account for confinement and tracking window effects and has the possibility to fit data acquired at multiple frame times into a single global model. The re-analysis of previously published data on DNA polymerase I in *E. coli* suggests that this protein complex uses rapid probing of DNA and therefore spends more than 50% of its time bound to DNA, a value previously hypothesized based on its preferred localization in the nucleoid but not quantified up to now. These new insights into the biology of DNA polymerase *in vivo*, can experimentally be further tested. The predicted times spent on DNA in the absence (60%) and presence of MMS (75%) can be independently quantified by measuring the ratio of polymerases in DNA-containing and DNA-free segments of cells elongated by cephalexin as was done previously for CRISPR-Cas complexes in *E. coli* (Vink et al., 2020).

Compared to other simulation-based frameworks for estimating transition rates (Martens et al., 2019; Rocha et al., 2019; Wieser et al., 2008), anaDDA holds several advantages. First, the distributions of simulations are not exact as they are generated from a limited number of particles and therefore do not allow for using an MLE approach, which requires convergence based on exact probability even for small changes in the parameter space. Secondly, since analysis methods can only be verified by knowing the ground truth, these algorithms can only be tested with and against simulations itself. Consequently, the analysis and verification data are not independent, which could lead to unobservable errors. Furthermore, our analysis method is computationally significantly faster. MLE takes just around 15 s to find the optimal parameter set for a global fit to a 50.000 tracks dataset with a step number range of 1-8 steps (Intel Core i7), whereas a simulation estimating three parameters with a global fit of all step numbers, required around 10 hours to find an optimal set of parameters.

Despite the new possibilities that anaDDA offers to analyse complex sptPALM data, a number of challenges remain. Firstly, our transition rate analysis is limited to Markovian processes, which assume that the transition rates are independent of past events. This assumption seems to be valid for protein binding kinetics *in vivo* (Ho et al., 2019; Persson et al., 2013; Slutsky and Mirny, 2004), but might not be generalizable for all biological systems (Morimatsu et al., 2007; Talaga, 2007). Secondly, macromolecules such as DNA binding proteins, potentially have many different binding sites and therefore would have many different k_{on} and k_{off} values. The transition rates extracted with anaDDA do not fully capture this complex biological behaviour and therefore should be interpreted as an average timescale at which these transitions take place. Thirdly, the number of states cannot, unlike Bayesian methods, be automatically extracted from the dataset. However, given the complexity of sptPALM data, Bayesian algorithms are prone to overfit the data (Figure S3A). A more robust way for model selection can be achieved by incorporating experimental controls (e.g. mutants or subunits which reduce complexity) and measurements at multiple frame times (Figure 4B). Fourthly, potential effects of finite exposure times (Berglund, 2010; Goulian and Simon, 2000) on the measured displacements have not been yet incorporated in anaDDA. These effects, however, can be minimized by using stroboscopic illumination (Elf et al., 2007; Hansen et al., 2018). Fifthly, anaDDA assumes Brownian motion and does not incorporate anomalous diffusion, which has been observed in some *in vivo* systems (Bohrer and Xiao, 2020; Höfling and Franosch, 2013). Our method can be adapted to incorporate anomalous diffusion, once it is clear which of the many potential models (Metzler et al., 2014) is suited best for the observed anomalous diffusion (Barkai et al., 2012). Again, care should be taken as these more complex models are more easily overfitted. Lastly, for performance reasons the tracks that we analyse in anaDDA are currently limited to a maximum step number of 8 steps, which under most experimental conditions represent more than 90% of the tracks and longer tracks are truncated to the maximum step number of 8.

In our current implementation it is possible to include two transitioning state into the direct fitting. We have shown, however, that when transition rates are slow compared to the frame time of the measurement, states can be treated as separate

species. Further development of the underlying master equations of PDA statistics could allow direct implementation of multistate models.

With the increasing use of brighter and more stable organic fluorophore (Banaz et al., 2018; Los et al., 2008) or low photon flux measurements (Balzarotti et al., 2017) for single-particle tracking, the resulting increase of the step number per track and the decrease of the localization error will enable further improvements in the precision of extracted kinetic parameters. Currently, we have implemented the software for tracking in two dimensions, but the algorithms can be further modified towards tracking in three dimensions. Using the estimated error for each individual localization can further improve the robustness of the analysis as has been demonstrated previously (Lindén and Elf, 2018). Another improvement which can be incorporated in our framework and has already been developed is to take the effect of particles moving out-of-focus, and the recovery of localizations depending on diffusion coefficients into account (Hansen et al., 2018; Rocha et al., 2019).

Our analysis method allows the quantification of fast kinetic transitions inside living cells with state lifetimes in the 1 ms to 1 s range opening a temporal range at which many DNA screening interactions are expected to take place (Elf et al., 2007). So far however, quantifying these interactions has been limited due to a lack of appropriate analytic and experimental methods. We are convinced that anaDDA will offer the means to determining fast kinetics *in vivo* which will be key to uncover and understand the behaviour of biomolecular complexes in cells.

References

Akimoto, T., and Yamamoto, E. (2017). Detection of transition times from single-particle-tracking trajectories. Phys. Rev. E *96*.

Antonik, M., Felekyan, S., Gaiduk, A., and Seidel, C.A.M. (2006). Separating structural heterogeneities from stochastic variations in fluorescence resonance energy transfer distributions via photon distribution analysis. J. Phys. Chem. B *110*, 6970–6978.

Balzarotti, F., Eilers, Y., Gwosch, K.C., Gynnå, A.H., Westphal, V., Stefani, F.D., Elf, J., and Hell, S.W. (2017). Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes. Science *355*, 606–612.

Banaz, N., Mäkelä, J., and Uphoff, S. (2018). Choosing the right label for single-molecule tracking in live bacteria: Side-by-side comparison of photoactivatable fluorescent protein and Halo tag dyes. J. Phys. D. Appl. Phys.

Barkai, E., Garini, Y., and Metzler, R. (2012). Strange kinetics of single molecules in living cells. Phys. Today *65*, 29–35.

Bausch, J. (2013). On the efficient calculation of a linear combination of chi-square random variables with an application in counting string vacua. J. Phys. A Math. Theor.

Van Beljouw, S.P.B., Van Der Els, S., Martens, K.J.A., Kleerebezem, M., Bron, P.A., and Hohlbein, J. (2019). Evaluating single-particle tracking by photo-activation localization microscopy (sptPALM) in Lactococcus lactis. Phys. Biol.

Berglund, A.J. (2010). Statistics of camera-based single-particle tracking. Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys. *82*.

Bickel, T. (2007). A note on confined diffusion. Phys. A Stat. Mech. Its Appl. *377*, 24–32.

Blosser, T.R., Loeff, L., Westra, E.R., Vlot, M., Künne, T., Sobota, M., Dekker, C., Brouns, S.J.J., and Joo, C. (2015). Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. Mol. Cell *58*, 60–70.

Blouin, S., Craggs, T.D., Lafontaine, D.A., and Penedo, J.C. (2015). Functional studies of DNA-protein interactions using FRET techniques. In Methods in Molecular Biology, pp. 115–141.

Bohrer, C.H., and Xiao, J. (2020). Complex Diffusion in Bacteria. In Physical Microbiology, G. Duménil, and S. van Teeffelen, eds. (Cham: Springer International Publishing), pp. 15–43.

Craig, J.M., Laszlo, A.H., Brinkerhoff, H., Derrington, I.M., Noakes, M.T., Nova, I.C., Tickman, B.I., Doering, K., De Leeuw, N.F., and Gundlach, J.H. (2017). Revealing dynamics of helicase translocation on single-stranded DNA using high-resolution nanopore tweezers. Proc. Natl. Acad. Sci. U. S. A. *114*, 11932–11937.

Crocker, J.C., and Grier, D.G. (1996). Methods of digital video microscopy for colloidal studies. J. Colloid Interface Sci. *179*, 298–310.

Elf, J., Li, G.W., and Xie, X.S. (2007). Probing transcription factor dynamics at the single-molecule level in a living cell. Science *316*, 1191–1194.

English, B.P., Hauryliuk, V., Sanamrad, A., Tankov, S., Dekker, N.H., and Elf, J. (2011). Single-molecule investigations of the stringent response machinery in living bacterial cells. Proc. Natl. Acad. Sci. *108*, 365–373.

Farooq, S., and Hohlbein, J. (2015). Camera-based single-molecule FRET detection with improved time resolution. Phys. Chem. Chem. Phys. *17*, 27862–27872.

Garza de Leon, F., Sellars, L., Stracy, M., Busby, S.J.W., and Kapanidis, A.N. (2017). Tracking Low-Copy Transcription Factors in Living Bacteria: The Case of the lac Repressor. Biophys. J. *112*, 1316–1327.

Goulian, M., and Simon, S.M. (2000). Tracking single proteins within cells. Biophys. J. *79*, 2188–2198.

Hansen, A.S., Woringer, M., Grimm, J.B., Lavis, L.D., Tjian, R., and Darzacq, X. (2018). Robust model-based analysis of single-particle tracking experiments with Spot-On. Elife *7*.

Heller, I., Hoekstra, T.P., King, G.A., Peterman, E.J.G., and Wuite, G.J.L. (2014). Optical tweezers analysis of DNA-protein complexes. Chem. Rev. *114*, 3087–3119.

Ho, H.N., Zalami, D., Köhler, J., van Oijen, A.M., and Ghodke, H. (2019). Identification of Multiple Kinetic Populations of DNA-Binding Proteins in Live Cells. Biophys. J. *117*, 950–961.

Hodges, C., Bintu, L., Lubkowska, L., Kashlev, M., and Bustamante, C. (2009). Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. Science *325*, 626–628.

Höfling, F., and Franosch, T. (2013). Anomalous transport in the crowded world of biological cells. Reports Prog. Phys. *76*.

Hohlbein, J., Aigrain, L., Craggs, T.D., Bermek, O., Potapova, O., Shoolizadeh, P., Grindley, N.D.F., Joyce, C.M., and Kapanidis, A.N. (2013). Conformational landscapes of DNA polymerase I and mutator derivatives establish fidelity checkpoints for nucleotide insertion. Nat. Commun. *4*, 2131.

Hohlbein, J., Craggs, T.D., and Cordes, T. (2014). Alternating-laser excitation: Single-molecule FRET and beyond. Chem. Soc. Rev. *43*, 1156–1171.

Jradi, F.M., and Lavis, L.D. (2019). Chemistry of Photosensitive Fluorophores for Single-Molecule Localization Microscopy. ACS Chem. Biol. *14*, 1077–1090.

Kalinin, S., Felekyan, S., Valeri, A., and Seidel, C.A.M. (2008). Characterizing Multiple Molecular States in Single-Molecule Multiparameter Fluorescence Detection by Probability Distribution Analysis. J. Phys. Chem. B *112*, 8361–8374.

Kapanidis, A.N., Lepore, A., and El Karoui, M. (2018). Rediscovering Bacteria through Single-Molecule Imaging in Living Cells. Biophys. J. *115*, 190–202.

Karslake, J.D., Donarski, E.D., Shelby, S.A., Demey, L.M., DiRita, V.J., Veatch, S.L., and Biteen, J.S. (2020). SMAUG: Analyzing single-molecule tracks with nonparametric Bayesian statistics. Methods.

Lanoiselée, Y., and Grebenkov, D.S. (2017). Unraveling intermittent features in single-particle trajectories by a local convex hull method. Phys. Rev. E *96*.

Lee, B.H., and Park, H.Y. (2018). HybTrack: A hybrid single particle tracking software using manual and automatic detection of dim signals. Sci. Rep. *8*.

Lerner, E., Cordes, T., Ingargiola, A., Alhadid, Y., Chung, S.Y., Michalet, X., and Weiss, S. (2018). Toward dynamic structural biology: Two decades of single-molecule förster resonance energy transfer. Science *359*.

Lindén, M., and Elf, J. (2018). Variational Algorithms for Analyzing Noisy Multistate Diffusion Trajectories. Biophys. J.

Los, G. V., Encell, L.P., McDougall, M.G., Hartzell, D.D., Karassina, N., Zimprich, C., Wood, M.G., Learish, R., Ohana, R.F., Urh, M., et al. (2008). HaloTag: A novel

protein labeling technology for cell imaging and protein analysis. ACS Chem. Biol. *3*, 373–382.

Manley, S., Gillette, J.M., Patterson, G.H., Shroff, H., Hess, H.F., Betzig, E., and Lippincott-Schwartz, J. (2008). High-density mapping of single-molecule trajectories with photoactivated localization microscopy. Nat. Methods *5*, 155–157.

Martens, K.J.A., van Beljouw, S.P.B., van der Els, S., Vink, J.N.A., Baas, S., Vogelaar, G.A., Brouns, S.J.J., van Baarlen, P., Kleerebezem, M., and Hohlbein, J. (2019). Visualisation of dCas9 target search in vivo using an open-microscopy framework. Nat. Commun. *10*, 3552.

Martos-Naya, E., Romero-Jerez, J.M., Lopez-Martinez, F.J., and Paris, J.F. A MATLAB TM program for the computation of the confluent hypergeometric function Φ 2.

Metzler, R., Jeon, J.H., Cherstvy, A.G., and Barkai, E. (2014). Anomalous diffusion models and their properties: Non-stationarity, non-ergodicity, and ageing at the centenary of single particle tracking. Phys. Chem. Chem. Phys. *16*, 24128–24164.

Michalet, X. (2010). Mean square displacement analysis of single-particle trajectories with localization error: Brownian motion in an isotropic medium. Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys. *82*, 041914.

Miller, H., Zhou, Z., Shepherd, J., Wollman, A.J.M., and Leake, M.C. (2018). Single-molecule techniques in biophysics: A review of the progress in methods and applications. Reports Prog. Phys. *81*.

Morimatsu, M., Takagi, H., Ota, K.G., Iwamoto, R., Yanagida, T., and Sako, Y. (2007). Multiple-state reactions between the epidermal growth factor receptor and Grb2 as observed by using single-molecule analysis. Proc. Natl. Acad. Sci. U. S. A. *104*, 18013–18018.

Myung, I.J. (2003). Tutorial on maximum likelihood estimation. J. Math. Psychol. *47*, 90–100.

Nir, E., Michalet, X., Hamadani, K.M., Laurence, T.A., Neuhauser, D., Kovchegov, Y., and Weiss, S. (2006). Shot-noise limited single-molecule FRET histograms: Comparison between theory and experiments. J. Phys. Chem. B *110*, 22103–22124.

Palo, K., Mets, Ü., Loorits, V., and Kask, P. (2006). Calculation of photon-count number distributions via master equations. Biophys. J. *90*, 2179–2191.

Paris, J.F. (2011). A Note on the Sum of Correlated Gamma Random Variables.

Persson, F., Lindén, M., Unoson, C., and Elf, J. (2013). Extracting intracellular diffusive states and transition rates from single-molecule tracking data. Nat. Methods *10*, 265–269.

Plank, M., Wadhams, G.H., and Leake, M.C. (2009). Millisecond timescale slimfield imaging and automated quantification of single fluorescent protein molecules for use in probing complex biological processes. Integr. Biol.

Qian, H., Sheetz, M.P., and Elson, E.L. (1991). Single particle tracking. Analysis of diffusion and flow in two-dimensional systems. Biophys. J. *60*, 910–921.

Robson, A., Burrage, K., and Leake, M.C. (2013). Inferring diffusion in single live cells at the single-molecule level. Philos. Trans. R. Soc. B Biol. Sci. *368*.

Rocha, J., Corbitt, J., Yan, T., Richardson, C., and Gahlmann, A. (2019). Resolving Cytosolic Diffusive States in Bacteria by Single-Molecule Tracking. Biophys. J. *116*, 1970–1983.

Rothenberg, E., and Ha, T. (2010). Single-molecule FRET analysis of helicase functions. Methods Mol. Biol. *587*, 29–43.

Rutkauskas, M., Krivoy, A., Szczelkun, M.D., Rouillon, C., and Seidel, R. (2017). Single-Molecule Insight Into Target Recognition by CRISPR–Cas Complexes. In Methods in Enzymology, pp. 239–273.

Santoso, Y., Torella, J.P., and Kapanidis, A.N. (2010). Characterizing Single-Molecule FRET Dynamics with Probability Distribution Analysis. ChemPhysChem *11*, 2209–2219.

Saxton, M.J. (1997). Single-particle tracking: The distribution of diffusion coefficients. Biophys. J. *72*, 1744–1753.

Slutsky, M., and Mirny, L.A. (2004). Kinetics of Protein-DNA Interaction: Facilitated Target Location in Sequence-Dependent Potential. Biophys. J. *87*, 4021–4035.

Smith, S.W. (2003). FFT Convolution. In Digital Signal Processing, pp. 311–318.

Stracy, M., Lesterlin, C., Garza de Leon, F., Uphoff, S., Zawadzki, P., and Kapanidis, A.N. (2015). Live-cell superresolution microscopy reveals the

organization of RNA polymerase in the bacterial nucleoid. Proc. Natl. Acad. Sci. *112*, E4390–E4399.

Talaga, D.S. (2007). Markov processes in single molecule fluorescence. Curr. Opin. Colloid Interface Sci. *12*, 285–296.

Uphoff, S., Reyes-Lamothe, R., Garza de Leon, F., Sherratt, D.J., and Kapanidis, A.N. (2013). Single-molecule DNA repair in live bacteria. Proc. Natl. Acad. Sci. U. S. A. *110*, 8063–8068.

Uphoff, S., Sherratt, D.J., and Kapanidis, A.N. (2014). Visualizing protein-DNA interactions in live bacterial cells using photoactivated single-molecule tracking. J. Vis. Exp.

Vink, J.N.A., Martens, K.J.A., Vlot, M., McKenzie, R.E., Almendros, C., Estrada Bonilla, B., Brocken, D.J.W., Hohlbein, J., and Brouns, S.J.J. (2020). Direct Visualization of Native CRISPR Target Search in Live Bacteria Reveals Cascade DNA Surveillance Mechanism. Mol. Cell *77*, 39-50.e10.

Vrljic, M., Nishimura, S.Y., Brasselet, S., Moerner, W.E., and McConnell, H.M. (2002). Translational diffusion of individual class II MHC membrane proteins in cells. Biophys. J. *83*, 2681–2692.

Wieser, S., Axmann, M., and Schütz, G.J. (2008). Versatile analysis of single-molecule tracking data by comprehensive testing against Monte Carlo simulations. Biophys. J. *95*, 5988–6001.

Woodside, M.T., Anthony, P.C., Behnke-Parks, W.M., Larizadeh, K., Herschlag, D., and Block, S.M. (2006). Direct measurement of the full, sequence-dependent folding landscape of a nucleic acid. Science *314*, 1001–1004.

Acknowledgements The authors thank Dr. S. Uphoff for supplying the experimental data on DNA polymerase and K. Martens for feedback and co-development of the Monte Carlo DDA implementation (MC-DDA) and all members of the Hohlbein and the Brouns groups for input during group discussions. This work was supported by the European Research Council (ERC) Stg grant 639707 and by

the Netherlands Organisation for Scientific Research (NWO/OCW), as part of the Frontiers of Nanoscience (NanoFront) program.

Software The latest version of the software is available on GitHub: https://github.com/HohlbeinLab/anaDDA

Author contributions S.B. and J.H. conceived and supervised the project; J.H. provided the initial idea, J.V. developed the framework, derived the equations and wrote the analysis scripts, J.V. and J.H. wrote the manuscript and S.B. provided feedback on the manuscript.

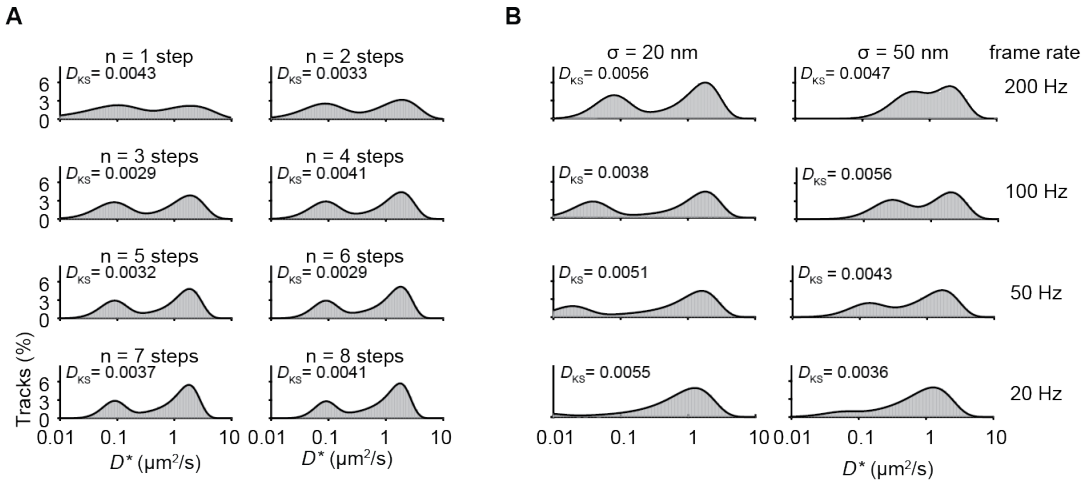


Figure S1. The effect of track length, localization error and frame rate on the shape of *D distributions. (A)** *D** distributions for different number of steps within a single trajectory, simulated (grey boxes) and DDA predicted distributions (black line) for 1-8 number of steps. Simulation parameters: k^*_{on} = 0.2 frame^{-1}, k^*_{on} = 0.2 frame^{-1}, D_{free} = 4 μm^2/s, σ = 30 nm (localisation precision) and *n* = 50.000. **(B)** *D** distributions for different localization errors and frame rates. Increasing the framerate shifts the peak of the bound population left to lower *D** values and increased localization errors shift this peak right to higher *D** values. At low frame rates, due to time averaging, the combined distribution of two states merge towards a single peak. Simulation parameters: k^*_{on} = 0.02 frame^{-1}, k^*_{on} = 0.02 frame^{-1}, D_{free} = 2 μm^2/s and *n* = 50.000. The Kolmogorov-Smirnov test statistic (D_{KS}) is indicated at each histogram.

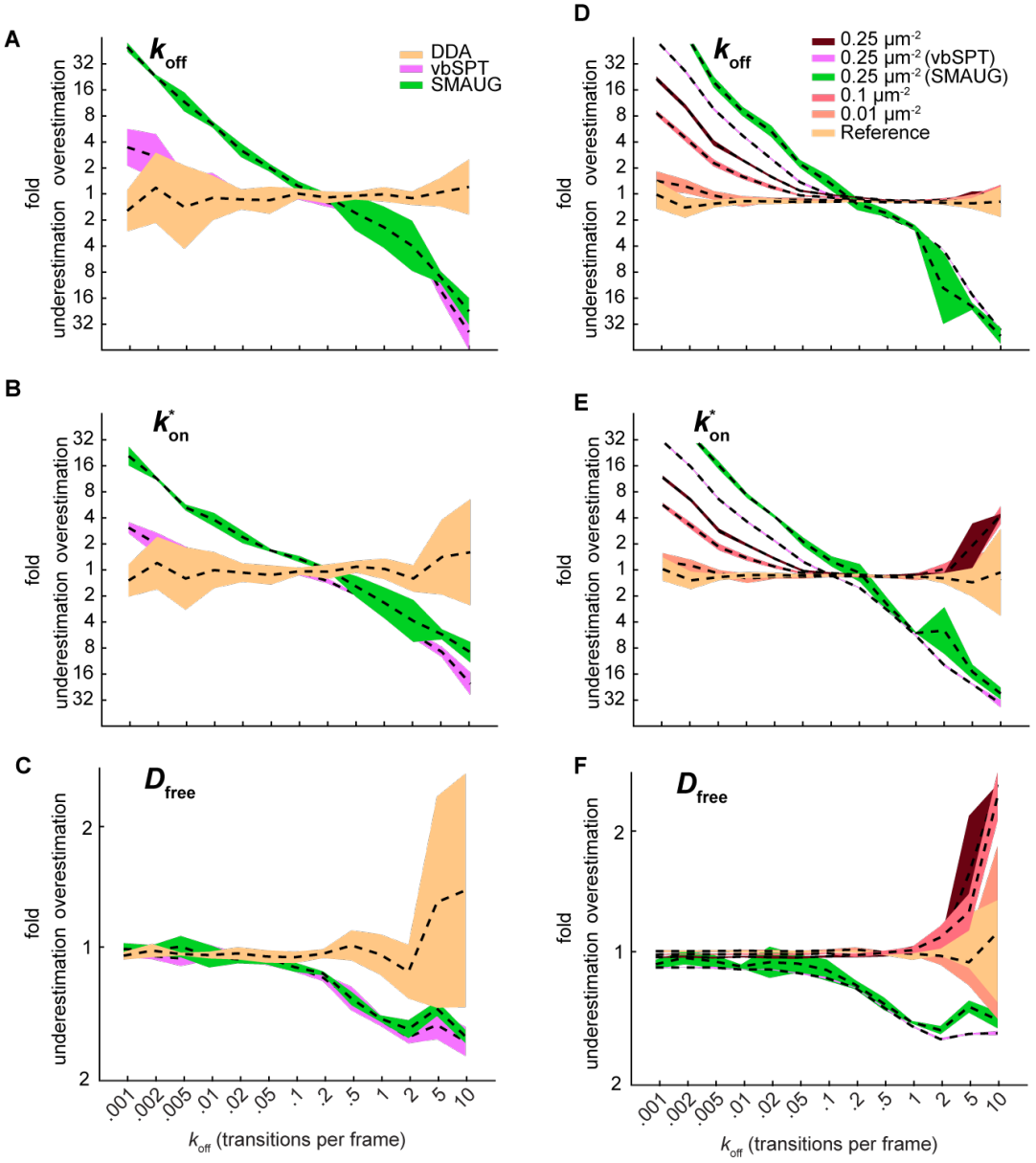


Figure S2. The effect of small data size and tracking errors on the extraction of parameters (A-C) Same as Figures 2D-F but with 1.000 tracks instead of 50.000 tracks. **(D-F)** Same as Figures 2D-F but with increasing particle densities. The simulations were modified in order that multiple particles were present in the same area at the same time point and therefore had a probability of being erroneously linked. The extraction accuracy was also compared to SMAUG and vbSPT for the highest tested density (0.25 localizations/μm^2). The reference (yellow) indicates the standard simulation (without tracking errors) run elsewhere in the manuscript. Tracking was done with algorithms described previously (Crocker and Grier, 1996) and a tracking window of 0.8 μm. Other parameters included in all figures were: $k^*_{on} = k_{off}$, D_{free} = 4 μm^2/s and σ = 30 nm.

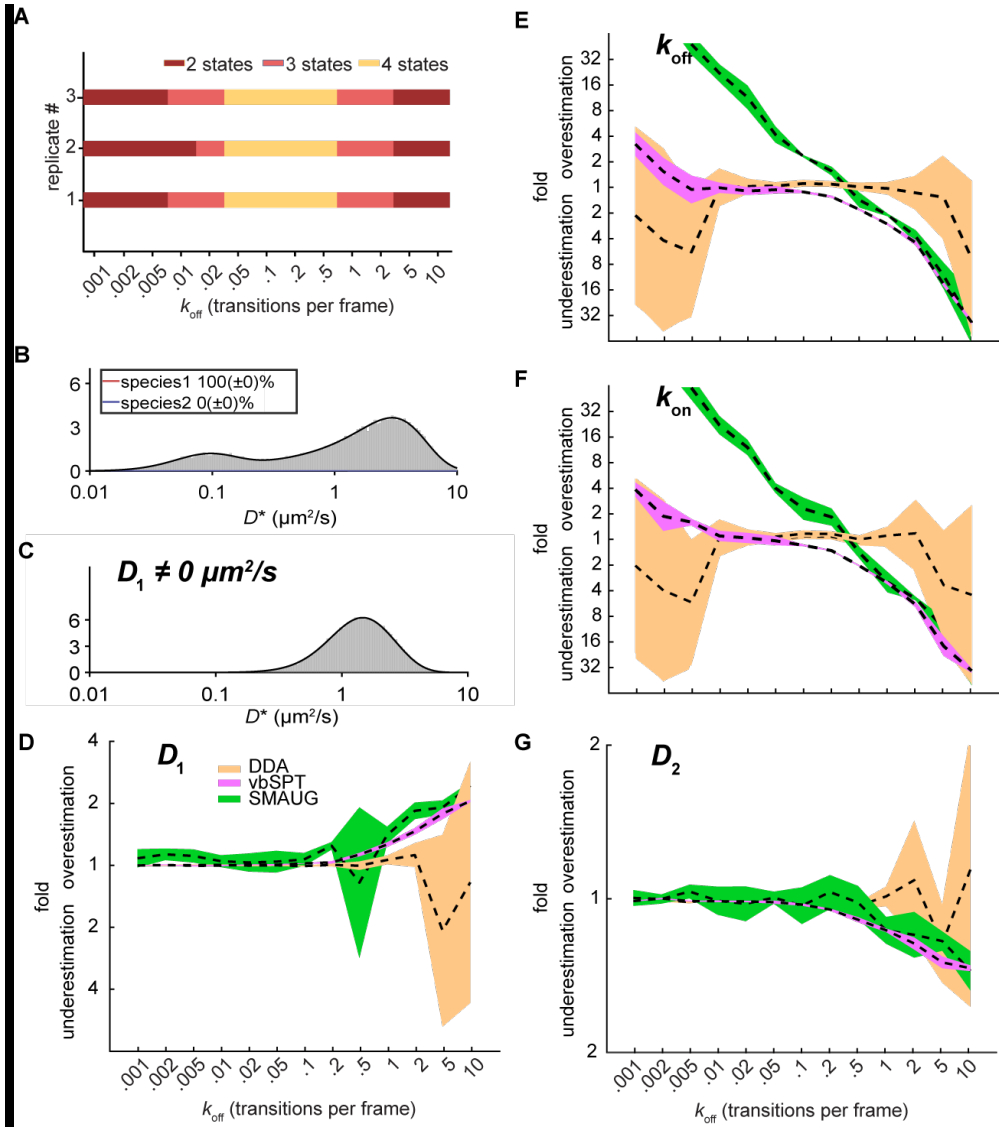


Figure S3. Model selection for multiple states and parameter extraction of systems with more than one mobile state. (A) The number of states extracted from vbSPT in a simulated two-state system. After removing restrictions on the maximum amount of states in vbSPT, the number of states fitted under some conditions differed from the amount of states modelled (2 states). Indicated are which simulated replicates contained which number of states (dark red: 2 states, light red: 3 states and yellow 4 states. **(B)** Result of fit with two species on a simulation containing only a single species $k^*_{on} = k_{off}$ = 0.2 frame^{-1}, D_{Free} = 4 μm²/s, σ = 30 nm and step number 4. **(C)** *D** distribution for a system with two mobile states, simulated (grey boxes) and DDA predicted distributions (black line). Parameters used were: $k^*_{on} = k_{off}$ = 0.2 frame^{-1}, D_1 = 1 μm²/s, D_2 = 4 μm²/s, σ = 30 nm and step number 4. **(D-G)** Same as Figures 2D-F but with two mobile states (which both are estimated from the data). Parameters used were $k^*_{on} = k_{off}$, D_1 = 1 μm²/s, D_2 = 4 μm²/s and σ = 30 nm.

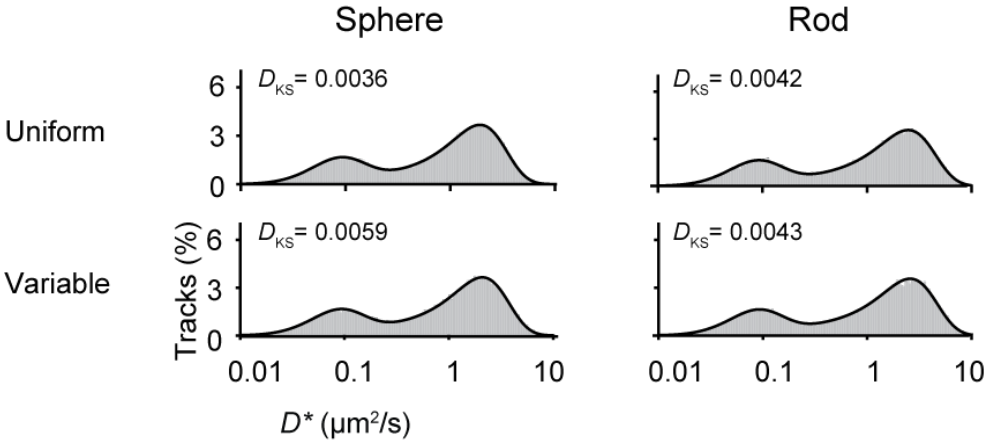


Figure S4. Distributions of a population of cells with uniform and variable cell size. The cell shapes were either spherical (left) or rod-shaped (right; radius to length ratio is 1:8). The average radius for uniform (upper row) and variable (lower row) cell sizes was the same: $r_{confined} = \sqrt{5D_{free}t}$. For the variable cell size, 60% of the cells were simulated with the same size as the average, whereas for 20% of the cells were simulated 25% smaller and for 20% of the cells were simulated with 25% larger cells. Further parameters used were: k_{on}^{*}= 0.02 frame^{-1}, k_{on}^{*} = 0.02 frame^{-1}, D_{free} = 4 μm^2/s, σ = 30 nm and *dt* = 0.01 s. The Kolmogorov-Smirnov test statistic (D_{KS}) is indicated at each histogram.

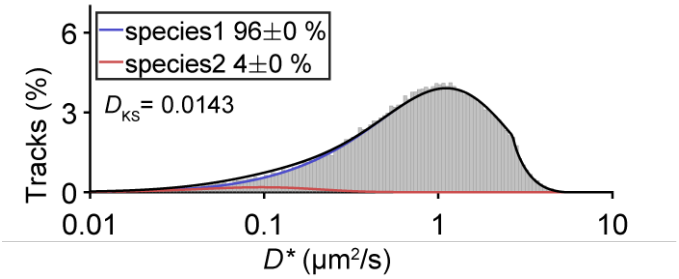


Figure S5. DNA polymerase histogram for tracks with a step number of 2 steps. The same condition and fitting parameters as for Figure 5A were used except that a two-step tracks are shown here. The maximum step size 5 pixels (0.6 μm) that was initially applied to this dataset results in a discontinuous distribution, which is correctly captured by the ana-DDA fit. Data from previous study (Uphoff et al., 2013).

Derivation of *D** distributions of localization error

As mentioned in the methods section, the *D** distribution of localization error can be described by a summation of correlated gamma random variables. The extend by which the localization error affects the correlation of sequential steps can be quantified by calculating the correlation coefficient $\rho_{ij} = \langle x, y\rangle/\sigma_x\sigma_y$ and the covariance of sequential steps as derived by Berglund(Berglund, 2010)

$$\langle \Delta x_i, \Delta x_j\rangle = 2DR\Delta t - \sigma^2 \quad \text{for } |i-j| = 1, \tag{54}$$

where *R* is the motion blur coefficient caused by movement of the particle during the illumination time and *D* is the diffusion coefficient. We assume further that measurements were taken with very short illumination pulses leading to $R \rightarrow 0$. We further convert equation 20 to MSD notation

$$\langle {\Delta x_i}^2, {\Delta x_j}^2\rangle = \langle \Delta x_i \Delta x_j\rangle^2 = \sigma^4. \tag{55}$$

After converting to two dimensions and assuming that Δx and Δy are independent, we get

$$\langle {\Delta x_i}^2+{\Delta y_i}^2, {\Delta x_j}^2+{\Delta y_j}^2\rangle = 4\langle \Delta x_i \Delta x_j\rangle^2 = 4\sigma^4. \tag{56}$$

To calculate the correlation coefficient ρ_{ij}, we use the following expression for the standard deviation of the MSD in two dimensions(Michalet, 2010)

$$\sigma_{MSD} = 4D\Delta t + 4\sigma^2, \tag{57}$$

leading to

$$\rho_{ij} = \frac{\langle {\Delta x_i}^2+{\Delta y_i}^2, {\Delta x_j}^2+{\Delta y_j}^2\rangle}{\sigma_{MSD,i}\,\sigma_{MSD,j}} = \frac{4\sigma^4}{(4D\Delta t + 4\sigma^2)^2}. \tag{58}$$

For most applications $D\Delta t > \sigma^2$ and ρ can be neglected. However, for immobile particles $D\Delta t = 0$, and $\rho = \frac{4\sigma^4}{(4\sigma^2)^2} = 1/4$. For a number of *n* measured steps with localization error of an immobile particle, the correlation matrix ρ is therefore given by

$$\rho = \begin{bmatrix} 1 & 1/4 & 0 & 0 & \dots \\ 1/4 & 1 & 1/4 & 0 & \dots \\ 0 & 1/4 & 1 & 1/4 & \dots \\ 0 & 0 & 1/4 & 1 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}. \tag{59}$$

The summation of gamma random variables given a certain correlation matrix has been previously derived in terms of confluent Lauricella series(Paris, 2011). Using the definitions above, this equation can be written as

$$f_D(x|0,n) = \Phi_2(1,..,..,1;n;-\frac{y}{\lambda_1},..,..,-\frac{y}{\lambda_n})\, x^{-1+n}/\det(A)\Gamma(n), \quad (60)$$

where Φ_2 is the confluent Lauricella function, λ_1-λ_n are the eigenvalues of the matrix $A = B \cdot B$, where *B* is an *n* x *n* matrix with diagonal values σ^2, and *C* is an *n* x *n* matrix with values $C_{ij} = \sqrt{\rho_{ij}}$.

This summation, for each number of measured steps *n* is the modified distribution for immobile particles taking into account the correlation between sequential measured displacements. To implement this distribution in the calculation of our total *D* distributions, we subtract the fraction of immobile particles after *n* time steps ($W_{contS1}(t_{S1} = 4t_f)$, Eq.5) multiplied with the distribution of expected *D** for *n* time steps $f_D(x|0,n)$ (Eq. 1) and replace it with the same fraction of immobilized particles multiplied with the distribution calculated based on the Lauricella series. The calculation of confluent Lauricella series was implemented from MATLAB code described in Martos-Naya et al. (2016)(Martos-Naya et al.).

The equation above can be further refined to experimental data, if there is a large difference in localization error between particles. In that case, there is another correlation factor due to the difference in brightness/focus of particles. As some particles might show a dynamic brightness, e.g. by diffusing in and out the excitation/detection focus, localizations of this track will have a higher precision the brighter the emission of the particle is, altering the correlation matrix to

$$\rho_{ij} = 1 \text{ for } i = j,$$

$$\rho_{ij} = \frac{1}{4} + \frac{3}{4}r \text{ for } |i-j| = 1,$$

$$\rho_{ij} = r \text{ for } |i-j| > 1,$$

where *r* is the correlation coefficient between two steps within the same track not sharing any localizations ($|i-j| > 1$). We found that this correlation coefficient can

be experimentally determined by measuring correlation of displacement of immobilized particles, or by measuring the correlation of estimated localization errors within tracks. This can be done by making a matrix in MATLAB where the rows are the different tracks and the columns are either the different step size of immobilized particles or the estimated localization errors. The built-in function *coerrcoef* then automatically calculates the correlation coefficient of this dataset.

Derivation of confinement corrections

The effective measured MSD given a diffusion coefficient *D* and a timestep Δt have been previously derived for a spherical geometry in multiple dimensions(Bickel, 2007), from which we derived the effective diffusion coefficient given the geometry and the real diffusion coefficient for spherical or rod-shaped geometries. First, the authors defined the zeros α_m at which $j_1'(\alpha_m) = 0$, with j_1' being the derivative of the spherical Bessel function of the first kind. This can be rewritten as

$$(\alpha_m{}^2\text{-}2)\sin(\alpha_m) + 2\alpha_m\cos(\alpha_m) = 0. \tag{61}$$

Subsequently the effective measured MSD within a spherical confined space of radius *r* is equal to

$$\mathrm{MSD} = r^2\left(\frac{6}{5}\text{-}12\sum_{m=1}^{\infty} e^{-\frac{\alpha_m{}^2 tD}{r^2}}\frac{1}{\alpha_m{}^2(\alpha_m{}^2\text{-}2)}\right). \tag{62}$$

This infinite series converges to zero. We therefore used the first 10.000 terms for calculation as a reasonable approximation. Because the previous equation refers to the three-dimensional MSD we use the following relation to calculate the observed diffusion coefficient we divide by 6*t*,

$$D_{obs} = \frac{\mathrm{MSD}}{6t} = \frac{r^2}{6t}\left(\frac{6}{5}\text{-}12\sum_{m=1}^{\infty} e^{-\frac{\alpha_m{}^2 tD}{r^2}}\frac{1}{\alpha_m{}^2(\alpha_m{}^2\text{-}2)}\right). \tag{63}$$

We can define the above equation as a function to calculate the observed diffusion given a certain radius, frame time and diffusion coefficient in the presence of spherical confinement

$$D_{obs} = f_{spherical}(r, t, D). \tag{64}$$

We then substitute $f_D(x|D, 1)$ for $f_D(x|D_{obs}, 1)$ and use equation 12 to calculate the distribution under any number of steps and given a k_{off} and k_{on}^*.

For rod geometries, there is no analytically derived solution available. However, we can combine the spherical derivation with a derivation in the same study for circular 2D geometries. In this geometry, the authors defined zeros β_m of the function $J_1'(\beta_m) = 0$, where J_1' is the derivative of the Bessel function of the first kind.

Subsequently, the effective measured MSD within a circular confined space of radius *r* is equal to

$$\mathrm{MSD} = \mathrm{r}^2\left(1\text{-}8\sum_{\mathrm{m}=1}^{\infty} \mathrm{e}^{-\frac{\alpha_\mathrm{m}{}^2\mathrm{tD}}{\mathrm{r}^2}} \frac{1}{\alpha_\mathrm{m}{}^2(\alpha_\mathrm{m}{}^2\text{-}1)}\right), \tag{65}$$

which we can again convert to a function to calculate the observed diffusion coefficient, but now as the MSD is two-dimensional, we divide by 4*t*

$$\mathrm{D_{obs}} = \frac{\mathrm{MSD}}{4\mathrm{t}} = \frac{\mathrm{r}^2}{4\mathrm{t}}\left(1\text{-}8\sum_{\mathrm{m}=1}^{\infty} \mathrm{e}^{-\frac{\alpha_\mathrm{m}{}^2\mathrm{tD}}{\mathrm{r}^2}} \frac{1}{\alpha_\mathrm{m}{}^2(\alpha_\mathrm{m}{}^2\text{-}1)}\right) = \mathrm{f_{circular}}(\mathrm{r},\mathrm{t},\mathrm{D}). \tag{66}$$

To calculate the effective measured MSD in a rod-shaped geometry, we split the cell in two parts: the hemispherical (consisting of two hemi-spheres) and the cylindrical part. If the cell is much longer than it is wide the cylindrical part dominates. For diffusion within a cylinder, movement along the cell length is not restricted, whereas movement along the width of the cell is constrained as given by equation 31. If the cell is as long as wide, we have a spherical cell for which the diffusion is described by equation 28. For cells featuring intermediate aspect ratios, we can calculate the ratio of these two domains via the ratio of their volumes

$$\mathrm{V_{total}} = \mathrm{V_{cylindrical}} + \mathrm{V_{hemi\text{-}sphere}} = \pi\mathrm{lr}^2 + \frac{4}{3}\pi\mathrm{r}^3, \tag{67}$$

where *r* is the radius of the cell width and *l* the length of the cylindrical part of the rod-shaped cell. The observed diffusion coefficient along the cell length, $\mathrm{D_{obs,x}}$ is not being restricted in the cylindrical part. The observed diffusion coefficient along the cell width $\mathrm{D_{obs,y}}$ on the contrary is restricted within the cylindrical part. Therefore, we separately calculate these two observed diffusion coefficients

$$\mathrm{D_{obs,x}}(\mathrm{r},\mathrm{t},\mathrm{D}) = \frac{\mathrm{V_{spherical}}}{\mathrm{V_{total}}}\mathrm{f_{spherical}}(\mathrm{r},\mathrm{t},\mathrm{D}) + \frac{\mathrm{V_{cylindrical}}}{\mathrm{V_{total}}}\mathrm{D}, \tag{68}$$

$$\mathrm{D_{obs,y}}(\mathrm{r},\mathrm{t},\mathrm{D}) = \frac{\mathrm{V_{spherical}}}{\mathrm{V_{total}}}\mathrm{f_{spherical}}(\mathrm{r},\mathrm{t},\mathrm{D}) + \frac{\mathrm{V_{cylindrical}}}{\mathrm{V_{total}}}\mathrm{f_{circular}}(\mathrm{r},\mathrm{t},\mathrm{D}). \tag{69}$$

For the last step, we require a probability distribution function of the sum of the two distributions $\mathrm{D_{obs,x}}$ and $\mathrm{D_{obs,y}}$ and go back to the distribution of *x*

$$\mathrm{x} \sim \mathrm{N}\left(0, \sqrt{2\mathrm{D_{obs,x}}\mathrm{t}}\right), \tag{70}$$

representing a normal distribution with mean of zero and $\sigma = \sqrt{2D_{obs,x}t}$. The distribution of the squared displacement is therefore a chi-square distribution

$$X^2/_{2D_{obs,x}t} = X^2/_{\sigma^2} \sim \chi_1^2. \quad (71)$$

The same holds for *y* with

$$Y^2/_{2D_{obs,y}t} \sim \chi_1^2. \quad (72)$$

To get to the distribution of D_{obs} we calculate

$$D_{obs} = {D_{obs,x}}/_2\left(X^2/_{2D_{obs},xt}\right) + {D_{obs,y}}/_2\left(Y^2/_{2D_{obs},yt}\right) = \frac{X^2+Y^2}{4t}. \quad (73)$$

Consequently, the distribution of D_{obs} is a summation of two chi-square variables weighted by the different diffusion coefficients. The formula for this summation was given in a previous study(Bausch, 2013) for the following case: Let $X, Y \sim \chi_k^2$ two independent and identically distributed chi-square random variables with *k* degrees of freedom. Let $Z \coloneqq aX + bY$, then the density function f_z is given by:

$$f_z = \theta(z)\frac{1}{(4ab)^{\frac{k}{2}}}\left(\frac{a-b}{8ab}\right)^{\frac{1}{2}\frac{k}{2}}\frac{\Gamma\left(\frac{1}{2}+\frac{k}{2}\right)}{\Gamma(k)}e^{-\frac{a+b}{4ab}z}x^{\frac{k}{2}\frac{1}{2}}I_{\frac{k}{2}\frac{1}{2}}\left(\frac{b-a}{4ab}z\right) \quad (74)$$

In our case, where *k = 1*, this equation is simplified to

$$f_z = \frac{1}{(4ab)^{\frac{1}{2}}}\exp\left(-\frac{a+b}{4ab}z\right)I_0\left(\frac{b-a}{4ab}z\right), \quad (75)$$

where I_0 is the zeroth order modified Bessel function of the first kind. When we substitute ${D_{obs,x}}/_2$ and ${D_{obs,y}}/_2$ for *a* and *b* respectively (combine equation 39 and 41), we obtain the following equation for the summation of two diffusion coefficients in two dimensions

$$f_D = \frac{1}{\left(D_{obs,x}D_{obs,y}\right)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(D_{obs,y}-D_{obs,x})x\right)I_0\left(\frac{1}{2}(D_{obs,y}-D_{obs,x})x\right) \quad (76)$$

where *x* is the measured displacement as in equation 1. This distribution can then be used as substitution for $f_D(x|D,1)$ and we can use equation 12 to solve the distribution under any number of steps and given a k_{off} and k_{on}^*..

6

PAM-repeat associations and spacer selection preferences in single and co-occurring CRISPR-Cas systems

Published as: J. N. A. Vink, J.H.L. Baijens, S. J. J. Brouns, PAM-repeat associations and spacer selection preferences in single and co-occurring CRISPR-Cas systems. *Genome Biology* (2021), accepted.

Abstract

Background

The adaptive CRISPR-Cas immune system stores sequences from past invaders as spacers in CRISPR arrays and thereby provides direct evidence that links invaders to hosts. Mapping CRISPR spacers has revealed many aspects of CRISPR-Cas biology, including target requirements such as the protospacer adjacent motif (PAM). However, studies have so far been limited by a low number of mapped spacers in the database.

Results

By using vast metagenomic sequence databases, we map approximately one-third of more than 200,000 unique CRISPR spacers from a variety of microbes and derive a catalog of more than two hundred unique PAM sequences associated with specific CRISPR-Cas subtypes. These PAMs are further used to correctly assign the orientation of CRISPR arrays, revealing conserved patterns between the last nucleotides of the CRISPR repeat and PAM. We could also deduce CRISPR-Cas subtype-specific preferences for targeting either template or coding strand of open reading frames. While some DNA-targeting systems (Type I-E and Type II systems) prefer the template strand and avoid mRNA, other DNA- and RNA-targeting systems (Type I-A, I-B and Type III systems) prefer the coding strand and mRNA. In addition, we find large-scale evidence that both CRISPR-Cas adaptation machinery and CRISPR arrays are shared between different CRISPR-Cas systems. This could lead to simultaneous DNA- and RNA targeting of
invaders, which may be effective at combating mobile genetic invaders.

Conclusions

This has broad implications for our understanding of how CRISPR-Cas systems work in a wide range of organisms that have never previously been studied.

Background

The adaptive CRISPR-Cas immune system provides heritable defence in the form of spacers, which are short nucleic acid sequences (28-36 bp) obtained from previous encounters with mobile genetic elements (MGE). These are stored in the bacterial or archaeal chromosome in CRISPR arrays (Jackson et al. 2017). CRISPR arrays contain spacers flanked on both sides by repeat sequences (~30 bp) and are transcribed as a single RNA, and subsequently processed into multiple crRNAs. crRNAs can be loaded into effector complexes formed by Cas proteins, that subsequently scan the cell for nucleic acid targets. Base pairing between the spacer and target nucleic acids (protospacer) allows the specific binding of effector complexes to targets, which are then destroyed (Brouns et al. 2008; Marraffini 2015). CRISPR-Cas systems are widespread in bacteria and archaea, with 42% of bacterial and 85% of archaeal genomes containing a CRISPR-Cas system (Makarova et al. 2020).

Both acquisition of new spacers (CRISPR adaptation) and target inactivation (CRISPR interference) are carried out by specialized sets of Cas proteins. *Cas* genes likely have originated from Casposons (Krupovic et al. 2014), a family of self-replicating transposons, and have since evolved many new genes and gene variants (Makarova et al. 2020). Based on the evolutionary classification of their *cas* genes, there are two classes of CRISPR-Cas systems. Class I systems contain crRNA-effector complexes made up of multiple subunits, while effector complexes of Class II systems are encoded by a single *cas* gene (Makarova et al. 2020). The two classes are further divided into six types, where each type is further divided into subtypes. The different types and subtypes do not occur homogeneously in nature, with Class II systems being nearly exclusive to bacteria (Makarova et al. 2020). More than 95% of CRISPR-Cas systems found in complete genomes are one of the first three types: Type I, II or III (Pourcel et al. 2020).

CRISPR-Cas systems can be studied on a mechanistic or on a functional level. Mechanistic features describe how CRISPR-Cas systems are able to fulfil their role. The mechanisms through which CRISPR-Cas systems operate are diverse. For example, some CRISPR-Cas systems defend the cell by targeting DNA (e.g. Type

I, II, IV and V), whereas other CRISPR-Cas types target invader RNA (e.g. Type III and VI) (Makarova et al. 2020). Another important mechanistic feature is the presence of a protospacer adjacent motif (PAM), which DNA-targeting systems require to differentiate self from non-self (Gleditzsch et al. 2019; Hale et al. 2009; Mojica et al. 2009). Furthermore, the PAM is an important feature in the target search process of DNA-targeting systems within the cell (Vink et al. 2020; Xue et al. 2017). This motif sequence flanking the crRNA-pairing site, between one and five nucleotides long, not only differs between subtypes, but can also differ between *cas* gene orthologs within the same subtype, for example Cas9 variants (Gasiunas et al. 2020).

An important aspect of the PAM is the moment of selection. While a more stringent PAM selection is achieved during the adaptation stage by Cas1-Cas2 and sometimes Cas4 (Kieper et al. 2018; Lee et al. 2018; Shiimori et al. 2018), , PAM selection during the CRISPR interference phase by the crRNA-effector complex will also occur (Cooper, Stringer, and Wade 2018; Hayes et al. 2016; Musharova et al. 2019). This led to the distinction of PAM into SAM (Spacer acquisition motif) and TIM (target interference motif) (Shah et al. 2013). In the above case where acquisition modules are more stringent, the PAMs that are observed are usually mostly determined by the acquisition machinery (SAM). However, in other situations the observed patterns might have been the result of selection for a working TIM. For example most of the spacers selected for in RNA targeting systems were found to be acquired at random(Artamonova et al. 2020), even though spacers present in natural CRISPR arrays often show a bias towards the coding strand (Cao et al. 2016; Goldberg et al. 2014), suggesting that the bias emerged from effective interference spacers through natural selection. On the other hand, there are systems that contain a reverse transcriptase fused to Cas1 (RT-Cas1) (Silas et al. 2016a) which can already select spacers from the correct strand. In experimental settings, these effects can be separated, but in bio-informatic analyses of natural spacers, the resulting effect is a combination of acquisition selection and interference selection.

Functional features describe what purposes CRISPR-Cas systems fulfil within the cell. There is evidence for some CRISPR-Cas functioning beyond adaptive immunity (Westra, Buckling, and Fineran 2014; Wimmer and Beisel 2020), however

even within the context of an adaptive immune system, CRISPR-Cas systems can serve different roles (e.g. as a first line of defence, or as an activator of other immune system pathways). This can be a reason why 23% of genomes with CRISPR-Cas systems contain more than one subtype (Bernheim et al. 2020), despite their costs (Nobrega et al. 2020; Vale et al. 2015). There are preferred combinations of certain subtypes, suggesting that there is an added benefit of having a specific combination of different subtypes present in the cell. The added benefit might consist of cooperativity between systems by formation of different lines of defence, avoidance of type-specific CRISPR inhibition by MGE or coupling of abortive infections mechanisms (Bernheim et al. 2020; Hoikkala et al. 2021; Pawluk, Davidson, and Maxwell 2017; Silas et al. 2017). On the other hand, some CRISPR-Cas systems are specialized to protect from certain invaders, which may require multiple co-occurring systems to be present in a single genome to protect from different types of invaders. Type IV systems that co-occur together with Type I systems primarily target plasmids (Pinilla-Redondo et al. 2020) and Type III systems target a class of phages that other Type I and V systems cannot (Malone et al. 2020; S. D. Mendoza et al. 2020), indicating that specialization in targets is a potential reason for co-occurrence of different subtypes. Through cooperation and specialization, co-occurring subtypes can function complementarily.

The functional and mechanistic features described above have been demonstrated experimentally for several microbial model systems, and these are often of specific interest to applications such as genome-editing. High-throughput assays to identify the PAM of CRISPR-Cas systems have been developed, but remain limited compared to the total range of CRISPR-Cas systems accessible bio-informatically (Gasiunas et al. 2020; Marshall et al. 2018; Walton et al. 2021). The full diversity of PAM and other mechanistic and functional features of CRISPR-Cas systems in nature remain understudied. To improve our knowledge on mechanistic and functional features of single and co-occurring CRISPR-Cas systems beyond the model organisms, we relied on vast metagenomic sequence databases to computationally find targets for spacers from diverse bacteria and archaea. This approach was recently taken to study phage-host interactions (Camarillo-Guerrero

et al. 2021; Dion et al. 2021; Soto-Perez et al. 2019). We mapped a third of the unique spacers to a target in publicly available metagenome sequence databases. We used the flanking regions of found spacer targets to build an initial PAM catalog of more than two hundred unique PAMs, and for more than half of the spacers in CRISPRCasdb (Pourcel et al. 2020). This was then employed to assign the correct orientation of transcription of CRISPR arrays, giving access to target strand information of invaders, and uncovering conserved links between repeat ends and PAM. Through the quantification of the spacers targeting template or coding strands we found that the preference for one of these strands is subtype-specific and indicates that some DNA-targeting systems (Type I-E, Type II-A and Type II-C) avoid RNA while other DNA- and RNA-targeting systems preferentially target RNA (Type I-A, Type I-B and Type III systems). We found spacers in co-occurring CRISPR-Cas systems to be compatible with both PAM and strand requirements, indicating that they may be shared between systems and will lead to both DNA and RNA targeting. Lastly, we identified three categories of multi-effector compatible spacers, which meet the PAM and strand requirements of co-occurring DNA and RNA-targeting systems.

Results

Blast analysis finds matches for 32% of spacers from CRISPRCasdb

The first step in our analysis was to select a set of CRISPR spacers and find potential matches to these sequences in DNA sequence databases. To this end, we selected the previously described CRISPRCasdb, which contained all spacers from 4266 complete bacterial and archaeal genomes (Pourcel et al. 2020). The spacers from CRISPRCasdb were then mapped to sequences from the NCBI nucleotide database as well as metagenomic databases with a high number of prokaryotic and virus sequences. Matches between spacers and sequences from the databases were found using BLASTn (Altschul et al. 1990). The matches were then filtered using an optimized approach which increased the number of matches while keeping the false

positives to a minimum (Methods, Additional file 1: Fig. S1A). As an indication of the false positive rate, we determined that for the matches found in the NCBI nucleotide database, 1% were eukaryotic or eukaryotic viral sequences, 10% were prokaryotic viral sequences and the majority (88%) corresponded to prokaryotic genome sequences (Additional file 1: Fig. S1A). This specificity towards prokaryotic sequences in a database that contains predominantly (83%) eukaryotic sequences shows that even though false positive hits cannot be excluded, the false positive rate is low.

From the 221,850 total unique spacers analysed, this optimized filtering approach resulted in 72,099 spacers (32% of total) with at least one match (Figure 1A), of which 31,327 spacers (15% of total) had a match in the NCBI nucleotide database (Figure 1B). For more than 25,000 of these, the best hit was completely identical to the spacer and for the vast majority (60,294) the total number of mismatched nucleotides was three or less (Figure 1C). Also in most cases more than one hit was found per spacer (Figure 1D).

The fraction of spacers with matches differed greatly between different genera, with *Streptococcus*, *Pseudomonas* and *Staphylococcus* among the genera with the highest fraction of matches (77%, 69% and 64% respectively) and *Calothrix*, *Nostoc* and *Thermosipho* among the lowest (4%, 4% and 3% respectively) (Figure 1E). Genera with high spacer matches typically occurred in well-sampled environments (human-associated), whereas the genera with lower matches occurred in what appear to be poorly sampled environments (soil, oceanic). A previous study (Shmakov et al. 2017) which looked for spacer matches in the NCBI nucleotide database found matches for 7% of spacers, using a more stringent 95% sequence identity and 95% coverage cut off as filtering thresholds. This difference in the fraction of spacers with matches in the NCBI nucleotide database indicates the added benefit and importance of our more sensitive filtering process. Additionally, the number of sequences in the database has increased in recent years from ~230 billion to ~700 billion bases. The most important factor for the increase in the number of spacers with matches however was the use of metagenomic databases, as the majority of

unique spacer matches derived from these databases (Figure 1B, Additional file 1: Fig. S1B).

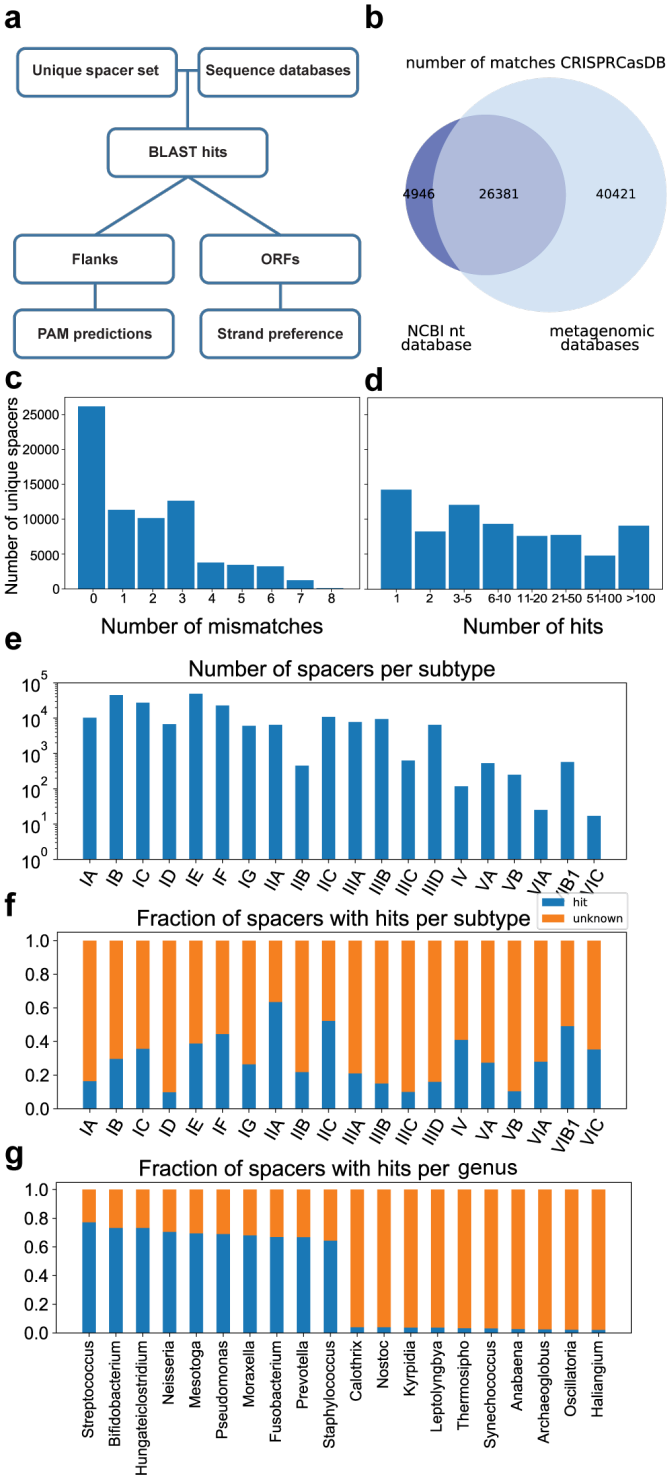


Figure 1. Spacer targets found with BLAST. (A) Computational pipeline for finding spacer targets. Targets of 72 099 spacers were found using blastn and filtered based on the fraction of spacer nucleotides matching a target sequence (See methods). (B) Venn diagram of spacers with matches in the NCBI nucleotide database versus metagenomic databases. (C) Plotted is the number of unique spacers (total 72,099) for which a match was found. Generally spacers < 4 mismatches fall within >90% identity threshold and are selected directly, and spacers with 4 or more mismatches generally within the >80% and <90% threshold and were selected in case another spacer from the same genus targeted the same sequence. (D) Number of sequences targeted by each spacer. Due to redundancy in the datasets, some of these sequences can be identical. (E) Fraction of spacers with hits for the ten genera with the highest and ten genera with the lowest fraction of hits. Only genera with at least 500 spacers are shown. (F) Number of spacers per subtype. The subtype of a spacer was predicted based on similarity of the repeat sequence to repeats with a known subtype (See methods). (G) Fraction of spacers with hits per subtype.

To find the subtypes of the spacers, we aligned the CRISPR repeat sequences to repeat sequences with known subtypes, based on the method described by Bernheim et al., 2020. Except for subtype II-B for which we extracted 453 spacers, all analysed subtypes from Type I, II and III systems contained more than a thousand spacers (Figure 1F). An exceptionally high fraction of spacers with matches was found for subtypes II-A (63%) and II-C (53%), while subtype I-A, subtype I-D, and Type III subtypes had notably lower fractions of spacer matches than average (15%, 11% and 20% respectively; (Figure 1G). The differences in fractions of matches found between subtypes may be due to their phylogenetic distributions, where well-sampled genera have different subtypes than poorly sampled genera (see above). However, even within well-sampled genera the fraction of spacers with matches differs between subtypes, with Type III subtypes having fewer hits on average (22%) than other subtypes (38%). The biases that we observed for both the fraction of hits in certain genera and subtypes remained true when we only used perfectly matching spacers (Additional file 1: Fig. S2). Overall, the large number of spacers with matches revealed sets of sequences that were targeted by each CRISPR-Cas subtype, which were then used to study mechanistic and functional aspects of CRISPR-Cas defence.

Alignment of protospacer flanks reveals 220 unique subtype-specific PAMs covering 55% of spacers

One of the important mechanistic features of CRISPR-Cas defence for DNA targeting systems (type I, II, IV and VI) is PAM recognition (Deveau et al. 2008; Horvath et al. 2008; Mojica et al. 2009; Shah et al. 2013). The first PAM was discovered in the alignment of bacteriophage sequences that were targeted by *Streptococcus* spacers (Bolotin et al. 2005). Later studies revealed more PAMs or the effect of mutant versions of the PAM (Anders et al. 2014; Fischer et al. 2012; Leenay et al. 2016; Musharova et al. 2019). We expand on these known PAMs that are limited to well-studied organisms by predicting new PAMs based on the alignment of the flanks of spacer matches (protospacers). The potential of this

method for large-scale PAM predictions was shown in a previous bioinformatics study (Mendoza & Trinh, 2018), with a key limiting factor being the number of spacers with matching targets. It was also previously shown that PAMs, acquisition machinery and repeat clusters co-evolve (Shah et al. 2013). We therefore increased the number of spacers with matches within one group by clustering spacers based on repeat similarity (>90% nucleotide identity and same repeat length). The sensitivity of PAM detection depends on the information content of the nucleotide positions of the PAM (signal) compared to the information content of the other flanking positions (noise). We found that clustering based on repeat similarity increased signal to noise ratio for PAM detection compared to clustering based on species-subtype (e.g. *Escherichia coli* I-E) or genus-subtype (e.g. *Pseudomonas* I-F). We furthermore found that spacers originating from organisms with very high or low GC-contents, displayed increased noise. We thus further increased the signal-to-noise ratio by adjusting the expected frequency of flanking nucleotides based on the average GC-content of the spacers within the cluster (Additional file 1: Fig. S3A). The flanks of unique hits within each cluster can subsequently be aligned, and with enough spacer hits, the information content reliably reveals the PAM sequence and position relative to the protospacer (Figure 2).We further checked whether our filtering approach leads to optimal PAM prediction and found that with stricter hit requirements (95-100% identity), the signal-noise ratio of PAM prediction decreased (Additional file 1: Fig. S3B). This was caused by a lower number of spacers per cluster and the number of hits per spacer.

This clustering approach together with our large number of hits led to a PAM prediction for 123,144 spacers (55% of all spacers; Additional file 2 and 3). For Type I and Type IV the PAM is known to occur in the 5' (upstream) flank of the protospacer, while Type II systems have their PAM in the 3' (downstream) flank of the protospacer (Jackson et al. 2017) (Figure 2A). This well-characterized feature of the PAM therefore allows the unique possibility to correctly orient CRISPR arrays given the rules described above. The orientation of arrays is an important feature to properly identify the chronology of acquisition events, the CRISPR leader sequence and potential RNA targeting. Tools have been developed to predict these bio-

informatically (Alkhnbashi et al. 2016; Biswas, Fineran, and Brown 2014; Milicevic et al. 2019). However, these tools in some cases contradict each other, implying that this prediction is not straightforward and fully accurate (Houenoussi et al. 2020; Milicevic et al. 2019).

To measure the accuracy of CRISPR array orientation predictions, we compared predictions to experimentally determined orientations from a recent study using transcriptome sequencing (TOP) to determine the direction of transcription of arrays (Houenoussi et al., 2020). The 7968 experimentally inferred spacer orientations were the same as our predictions in 85% of cases, while only 33% of TOP predicted spacer orientations were the same as the CRISPRCasdb prediction (Additional file 2) which is a combination of CRISPRdirection and a GC-content based leader prediction tool (Biswas, Fineran, and Brown 2014; Couvin et al. 2018). For the 15% where TOP did not match our predictions, both CRISPRCasdb and our PAM based orientations predicted the same orientation, indicating that some of the TOP orientation prediction based on transcription data might not have been correct. When we compared the predictions of CRISPRCasdb with our PAM-based orientations directly, we found a 88% match between all spacers. We furthermore found that many Type I and Type III repeats for which we predicted the orientation based on the PAM, contained the 3'-end motif ATTGAAAC of their repeat (Additional file 1: Fig. S4) described previously (Lange et al. 2013). This conserved motif is transcribed and forms the 5' handle of the crRNA and is held by crRNA-effector complexes. Altogether, these findings indicate that the position of the PAM is a reliable indicator for the orientation of the CRISPR array, and can be used to annotate CRISPR array information, giving access to features such as spacer acquisition chronology and strandedness.

Type I PAMs are conserved within repeats, Type II PAMs are genome specific

Sequence logos of alignments of Type I (Figure 2A) recover previously known PAMs including the subtype I-E AWG PAM found in *Escherichia* and subtype I-F CC PAM found in *Pseudomonas* (Leenay and Beisel 2017), but also previously undescribed PAMs. Out of the 43 unique PAM-subtype combinations, 25 were not found in previous publications (Table 1). Interesting examples of novel PAMs include a CTT PAM in I-C systems (compared to the more canonical TTC) and a CCA PAM in I-F systems (compared to the more canonical CC). They are generally short (2-3 nt) and are well-defined (high information content/bit score). Diversity is highest in I-B systems (11 unique PAMs) and lowest in I-F systems (3 unique PAMs).

For Type II PAMs, we found both short, well defined PAM motifs (such as *Streptococcus* II-A) as well as longer PAMs with less conserved PAM motifs (Figure 2B). Poorly conserved PAM motifs could be caused by a variation of PAMs used within the same repeat cluster or by the promiscuity of PAM recognition in Type II systems (Crawley et al. 2018). In previous work, it was shown that in some cases Cas9 proteins that use the same repeat can have different PAMs (Gasiunas et al. 2020; Magadán et al. 2012). We questioned whether our clustering of spacer hits based on repeat sequence would result in the low conservation scores in some PAM motifs. When we based our PAM motif predictions on spacers coming from a single genome, we recovered different PAMs for Type II systems that use the same repeat (Additional file 1: Fig. S5), whereas for Type I systems we always recovered the same PAM for each genome within a repeat cluster. We conclude that repeat sequence clustering is not an option and therefore only determined PAMs from spacers from individual genomes in Type II systems. From this genome-based clustering of spacers, unique PAM sequences were recovered from 302 genomes in Type II systems. For the few examples where the genomes from our database were characterized, the PAM prediction matched previous reported PAMs.

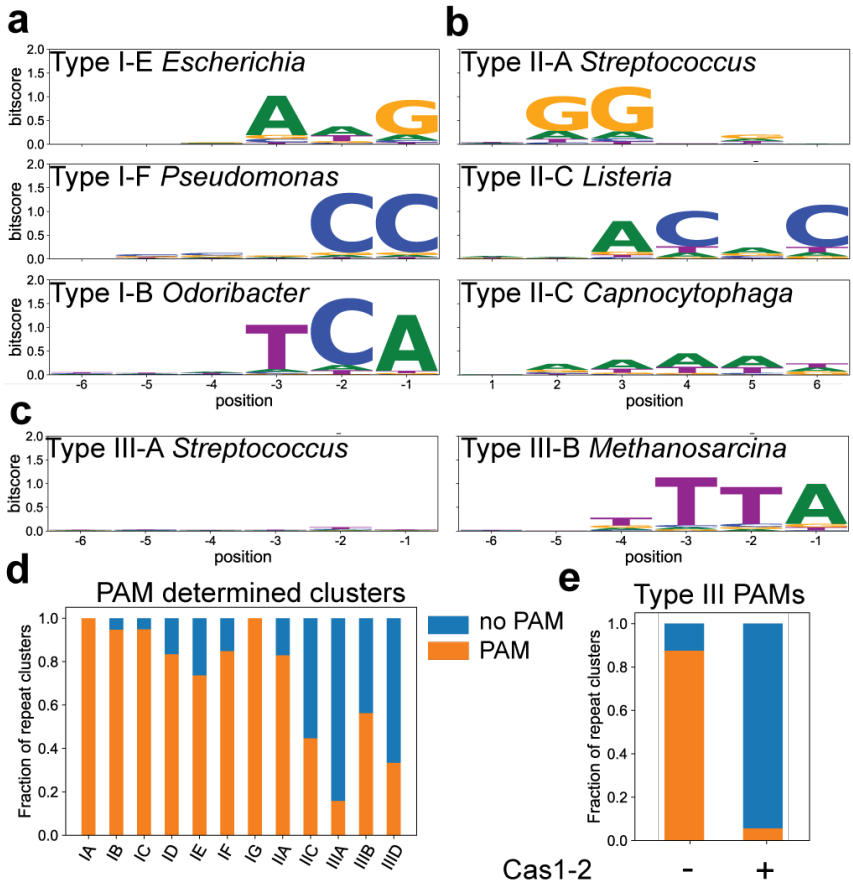


Figure 2. PAM determination of repeat clusters. (A) Sequence logos of upstream flank of hits to spacers from Type I repeat clusters. Sequence logos of protospacer flanking regions per repeat cluster. Y-axes show information content per nucleotide position. Label includes subtype of the repeat cluster and a representative genus in which this repeat cluster is found. To (B) Same as (A) but for downstream flanks of spacers from Type II repeat clusters. (C) Same as (A) but for upstream flanks from Type III repeat clusters. (D) Frequency of PAM determined repeat clusters with more than 25 hits. Nucleotide positions were considered part of PAM with a bitscore of at least 0.4 and 10 times above the median bitscore of the 23 nucleotides surrounding the hits. PAM size was at least 2 nucleotides. (E) Frequency of PAM determined repeat clusters for Type III systems that contain Cas1-2 vs Type III systems that lack Cas1-2. Additional file 5 contains the PAM for each strain-subtype combination.

Overall, the diversity in PAM motifs in Type II systems is higher than in Type I systems. For Type I, we found hits for 56.026 spacers, from 588 different genera in 34 different phyla. For Type II systems we found hits for 9883 spacers from 149 different genera in 14 different phyla. Based on these numbers you would expect the

number of unique Type I PAMs we recovered to be higher than Type II PAMs. However, in total we find 43 unique PAM-subtype combinations for Type I compared to 134 unique PAM-subtype combinations for Type II systems.

43% of Type III repeat clusters contain a PAM

When investigating Type III repeat clusters, we found many devoid of a PAM. This is expected, as RNA-targeting systems do not require a PAM to find a target (Figure 2C), and rely on the Protospacer Flanking Sequence (PFS) to avoid self targeting (Deng et al. 2013; Elmore et al. 2016). Interestingly, other repeat clusters contained PAMs that appeared to be the same as Type I PAMs, which raised the question, why these clusters contained a PAM. We compared the PAM detection frequency for clusters with at least 25 unique spacer hits (Figure 2D). For Type I subtypes whereas for Type III systems the number of PAM-containing repeat clusters was lower, with Type III-A having the lowest (16%) and III-B the highest (56%) fraction of PAM-containing repeat clusters in Type III systems. As it was previously shown that Type III systems often lack their own acquisition machinery (Makarova et al. 2015), we hypothesized that the PAM found in Type III repeat clusters originates from the spacer acquisition machinery that Type I systems share with Type III systems. We observed that the PAM frequency in Type III clusters that lack their own acquisition machinery is high (95%; Figure 2E), whereas the PAM frequency is low in Type III clusters that contain their own *cas1-cas2* genes (8%). This supports the hypothesis that the PAM in Type III arrays originates from Type I spacer acquisition modules functioning in *trans*. Genomes with PAM-containing Type III systems can be found in Additional File 5.

Conserved patterns between PAM and repeats

PAMs usually differ from the ends of CRISPR repeats, which allows for self-nonself discrimination (Leenay et al. 2016; Mojica et al. 2009; Westra et al. 2013). Type III and other RNA-targeting CRISPR-Cas systems do not require a PAM, but many do require mismatching between the repeat end and the protospacer flanking sequence (PFS) (Johnson et al. 2019; Marraffini and Sontheimer 2010). Given these previous

observations, we wanted to investigate if there are conserved links between repeat ends and PAM of individual systems (Figure 3A), and whether Type III PAMs that originate from Type I spacer acquisition modules are also compatible with Type III PFS requirements.

We collected all unique repeat-PAM sequence combinations in our dataset and compared the repeat nucleotide with the corresponding PAM nucleotide in each position. For Type I systems (Figure 3B) we found that the -3 and -2 nucleotide of the repeat can be a strong predictor of the corresponding PAM nucleotide, where a -3C in the repeat would lead to a -3A in the PAM, -3G to -3T, -3T to -3A. At the middle position, a -2C would lead to a -2A in the PAM. (Figure 3B). The most common -2 and -3 repeat nucleotide is an A, in which case the PAM nucleotide mostly is either a T or a C. For the -1 position, the nucleotide identity of the PAM sequence cannot be predicted directly from the repeat sequence.

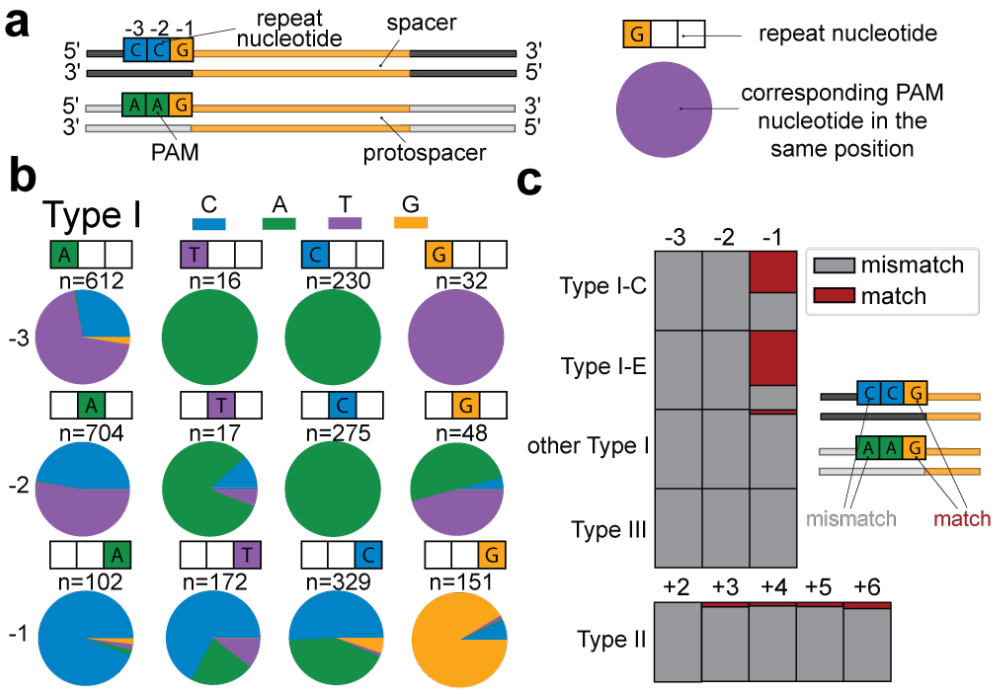


Figure 3. Relationship between repeat and PAM sequence. (A) Schematic of the analysis of PAM and repeat sequence. The nucleotide identity of the PAM in each position is compared to the nucleotide of the repeat. (B). PAM nucleotide frequency for Type I repeats. For each given repeat nucleotide position (indicated with coloured boxes) the PAM nucleotide (pie chart) for each unique PAM-repeat combination of our database. Number of occurrences is indicated above the pie chart

(n). (C). The frequency of matches (red) and mismatches (grey) between the PAM and the corresponding repeat nucleotide for each position in relation to the spacer. For Type II, the positions are compared on the other side of the spacer.

For Type II systems, most nucleotide positions can accommodate two or three PAM nucleotides (Additional file 1: Fig. S6A). In +2 and +3 positions, the most common repeat nucleotide (T), accommodates either an A or G PAM nucleotide, which is analogous to the most common nucleotide in Type I systems (-3 and -2 adenine), which tends to co-occur with a C or T PAM nucleotide. For Type III systems, the variation of repeat nucleotides is smaller, but generally similar combinations are found as in Type I systems (Additional file 1: Fig. S6B). Overall, the most conserved repeat-PAM co-occurrence patterns are found in the -2 and -3 positions of the Type I and Type III arrays.

These co-occurrence patterns suggest that in most cases the PAM that is used and selected for differs from the repeat. This holds true for most of the experimentally determined and previously predicted PAM sequences (Almendros et al. 2019; Garcia-Heredia et al. 2012; Kieper et al. 2018; Lillestøl et al. 2009; Lopatina et al. 2019; Manica et al. 2011; Mojica et al. 2009). However previous studies have shown that in some cases, part of the repeat sequence is PAM-derived (Swarts et al. 2012). We then asked in what CRISPR-Cas subtypes the PAM matches the corresponding repeat nucleotide for each of the spacer flanking positions. When we counted the occurrence of a matching PAM, we found that this only occurred frequently in the -1 position of Type I-C (35%) and Type I-E (48%; Figure 3C). We found that these matches are associated with repeats that have TTC PAMs in Type I-C and AAG PAMs in Type I-E, which could indicate that the C of Type I-C repeat sequences is PAM-derived, as was similarly demonstrated for the G of AAG PAMs in Type I-E (Swarts et al. 2012).

In other positions and CRISPR-Cas types, >98% of the repeat-PAM combinations did not match each other, which shows that the general patterns between repeats and PAMs, and perhaps mechanism of self- vs non-self discrimination is conserved in all subtypes. In Type III systems all cases demonstrate mismatches between PAM and repeat, which is a requirement of functional Type III spacers (Johnson et al. 2019; Marraffini and Sontheimer 2010). This finding demonstrates that the PAMs

of Type III array spacers acquired with Type I acquisition modules are compatible with PFS requirements of Type III systems.

Strand bias for the template or coding strand is subtype-specific

Our method has revealed a large number of newly identified PAMs and has shown that Type III systems which lack their own acquisition machinery and co-occur with Type I systems, almost always contain a PAM. The presence of a PAM in these systems could enable Type I systems to use the spacers stored in Type III arrays as they are compatible with the PAM requirements of Type I effector complexes. Furthermore, Type III effector complexes could benefit from a PAM-selecting acquisition module, as it excludes spacers with repeat-PAM matches (Figure 3C).
Besides the PFS, another requirement for type III spacers is that the spacer comes from the correct strand, as these complexes can only bind to the RNA transcripts. We wondered whether some species indeed use Type I and III dual functionality CRISPR arrays, as PAM-dependent DNA targeting and PAM-independent mRNA targeting are not mutually exclusive. We therefore asked whether spacers of DNA-targeting systems are also compatible with Type III surveillance complexes, if they happened to be picked from the correct strand.
To determine the potential ability of crRNA to target RNA, we measured the strand bias by counting the spacers that targeted the coding or template strand of predicted open reading frames (ORFs) (Figure 4A). As spacers targeting the template strand are unable to base pair the transcribed RNA, the fraction of spacers targeting the coding strand serves as an estimate of the RNA targeting ability of the crRNA. For example, in *Moraxella* IIIB arrays, a significant bias for the coding strand was found (88%, $p<e^{-11}$) (Figure 4B). This bias allows Type III effectors carrying crRNA from those spacers to bind to their target RNA. However, also I-C spacers in *Moraxella*, for whose effectors this is not strictly required, show significant bias for the coding strand ($p<e^{-3}$), indicating a selection for RNA-targeting spacers.

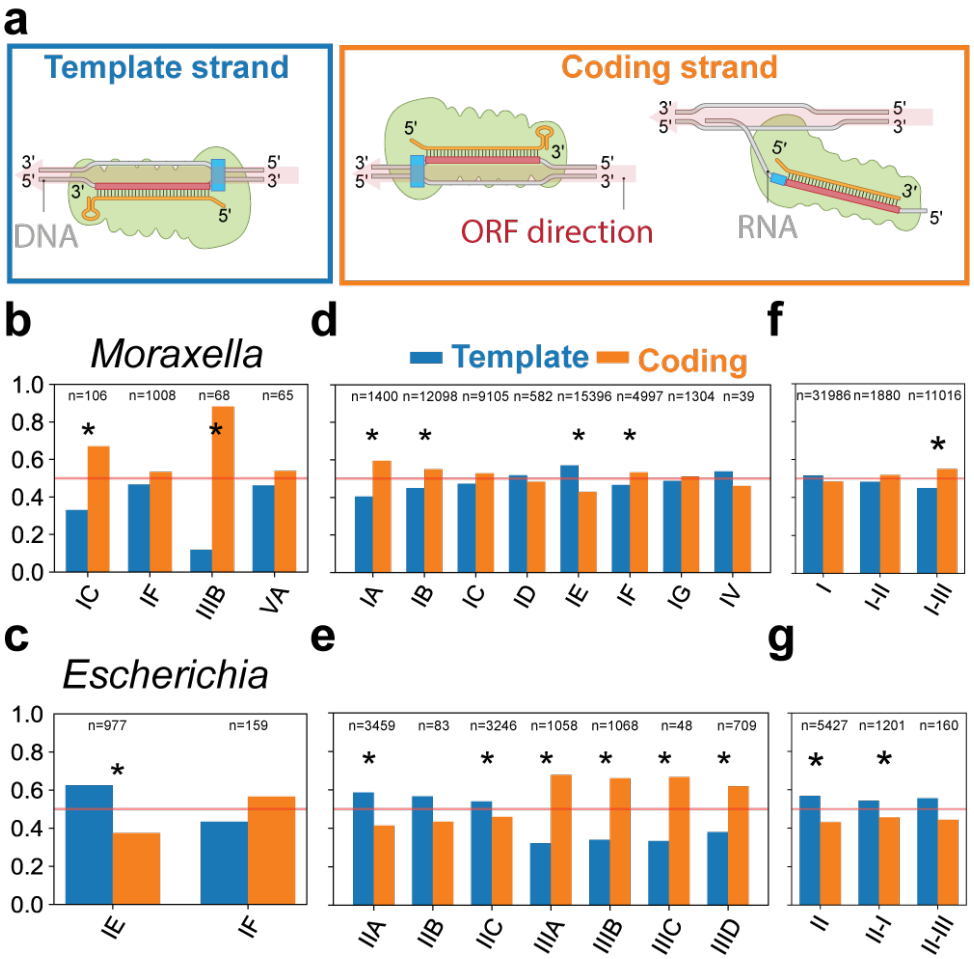


Figure 4. Template and coding strand targeting of spacers. (A) Schematic representation of a spacer targeting the template strand and a spacer targeting the coding strand inside an ORF. Spacers targeting the coding strand are also able to base pair with and target transcribed RNA. (B) Fraction of *Escherichia* spacers targeting template (blue) and coding (orange) strand by subtype. (C) Fraction of *Moraxella* spacers targeting template and coding strand by subtype. (D) Fraction of spacers targeting template and coding strand for Type I and Type IV subtypes. (E) Fraction of spacers targeting template and coding strand for Type II and Type III subtypes. (F) Fraction of spacers targeting template and coding strand for Type I. Spacers are grouped based on which other type of Cas effector genes are present in the genome. (G) Same as (F) but for Type II spacers. Significance of strand bias is calculated with a binomial test and a p-value<0.01 is indicated with an asterisk. Additional file 2 contains the strand targeted of each spacer and allows to extract the strand bias for each taxon.

For *Escherichia* subtype I-E, 977 spacer matches inside ORFs were found, of which 611 (63%) targeted the template strand (Figure 4C), showing a significant bias for targeting the template strand ($p<e^{-14}$) potentially avoiding RNA. No significant

strand bias was found for *Escherichia* subtype I-F (43% template strand, p=0.11), suggesting that strand bias is CRISPR-Cas subtype-specific.

Analysis of our complete dataset revealed general trends in the strand preferences for each subtype (Figure 4D, E). The strongest strand bias was found in Type III systems with an average of 65% of the spacers matching the coding strand (coding strand : template strand ~ 2:1). This result demonstrates that there is selection in Type III systems for spacers to target the transcribed RNA. This selection can originate at the adaptation stage by dedicated adaptation machinery selecting from RNA/coding strands such as RT-Cas1 (Silas et al. 2016b) or at the interference stage, where only functional RNA-targeting spacers are retained in the population (Artamonova et al. 2020).

The strand biases we found are consistent with our curated CRISPR array orientation predictions, because an incorrect CRISPR array orientation prediction would obscure strand-specific targeting. Type I-A and Type I-B also displayed significant strand bias for the coding strand although at lower levels (60% and 55%; $p<e^{-9}$ and $p<e^{-14}$ respectively).

Contrary to the Type III, Type I-A and I-B systems, we found a significant strand bias towards the template strand in in subtype I-E, Type IV and Type II systems, with the strongest bias found in subtype II-A (59%) and subtype I-E (57%). Given the high number of spacers in these groups the chance of observing this bias by chance is small ($p <e^{-23}$ and $p <e^{-69}$ respectively), again suggesting avoidance of RNA.

Co-occurrence of Type I and Type III systems lead to PAM and strand targeting compatibility

As we noticed that Type III spacers were compatible with Type I PAMs in multiple cases, we next asked whether Type I spacers are compatible with RNA targeting in microbes with co-occurring Type I and III systems. We measured the strand bias of Type I spacers in genomes containing either combination of Type I, Type II and Type III surveillance complexes (Figure 4F). No significant strand bias was found

for Type I spacers in the presence of Type I and/or Type II surveillance complexes. However, in the presence of Type I and Type III surveillance complexes, Type I spacers had a slight but significant coding strand bias (55%, $p<e^{-14}$). This might be caused by increased selection pressure to keep RNA targeting spacers in the presence of RNA targeting surveillance complexes. This would suggest that spacers are selected to be compatible for both Type I and Type III effector complexes in such situations. For Type II spacers, the presence of Type III did not significantly change the strand bias (Figure 4G). Given the natural tendency of Type II spacers to bias towards the template strand (Figure 4E), these findings suggest that Type II spacers are less compatible with co-occurring Type III effector complexes than Type I spacers.

Three distinct categories of co-occurring multi-effector compatible arrays exist

The findings above indicate that subtype-specific preferences exist for either the template or coding strand of the DNA. These preferences might enable or preclude compatibility between the spacers of co-occurring subtypes. The subtype-specific preference of template targeting (e.g. in Type I-E and Type II) will reduce the number of effective spacers that can be used by co-occurring RNA-targeting systems, whereas subtypes with a preference for the coding strand (Type I-A, Type I-B) might make their spacers more compatible for RNA-targeting systems. We categorised all multi-effector compatible arrays that can be used by effector complexes from different subtypes. This means for co-occurring DNA-targeting systems these arrays need to have a PAM that can be used in both systems, whereas for co-occurrence of a DNA- target CRISPR-Cas system with an RNA targeting system, the arrays present in the genome need to both have the correct PAM and have a bias for the coding strand.

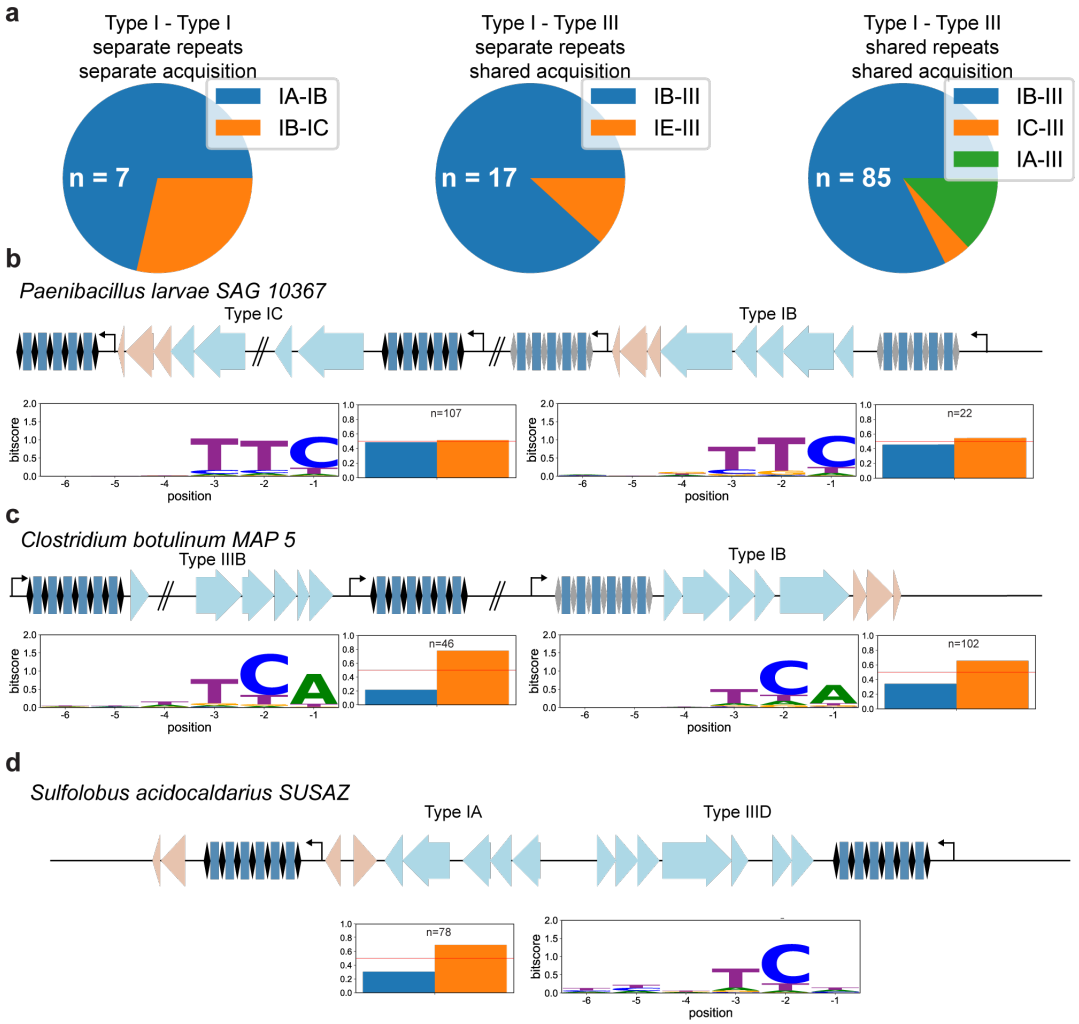


Figure 5. Different organisations of subtypes containing compatible spacer sequences. (A) Pie chart of frequency of genomes each category of organisation, based on the subtype combination involved. Total number of genomes for which this category was found (n) is noted in each chart (n). (B-D) Genome representations of examples for the different organisation categories, (b) Type I-Type I compatibility, (c) Type I- Type III compatibility (different repeat sequences), (d) Type I- Type III compatibility (same repeat sequences). Genes involved in interference (blue) and adaptation (red) are shown for the different subtypes within the genome. PAM logo and strand bias of each associated repeat cluster is depicted below the genomic representations.

Overall, we can distinguish three main categories of co-occurring CRISPR-Cas systems in which spacers are compatible for multiple effectors (Figure 5A, Additional file 4). Firstly, two co-occurring DNA-targeting systems which have their own adaptation machinery and their own repeat sequences (Figure 5B; n=7;

Type I-A-Type I-B: 5, Type I-B-Type I-C: 2). Secondly a co-occurring DNA-targeting and RNA-targeting system, with distinct repeat sequences but a commonly shared acquisition machinery (Figure 5C; n=17; Type I-B-Type III: 15, Type I-E-Type III: 3). Thirdly, a co-occurring DNA-targeting and RNA-targeting system, with shared repeat sequences and shared acquisition machinery (Figure 5D; n=85; Type I-B-Type III 71; Type I-A-Type III 11; Type I-C-Type III 3).

Taken together, our data indicate that multi-effector compatible arrays are most prevalent between Type I and Type III systems. Within the Type I systems, the most common subtype to use multi-effector compatible arrays is Type I-B, but also Type I-A, Type I-C and Type I-E use these arrays. The Type III systems that use compatible arrays lack their own adaptation machinery, however repeat clusters in these co-occurring systems display a strand bias that suggests selection for RNA-targeting spacers. The information content is similarly strong for PAMs in Type III arrays as in Type I arrays, which demonstrates that the PAM is selected to the same extent for Type I as shared Type III arrays.

Discussion

In this study we have matched CRISPR spacers of complete genomes of bacteria and archaea with their targets in (meta)genome databases and subsequently analysed the genomic flanks of the protospacers. We computationally found targets for 32% of CRISPR spacers from thousands of bacterial and archaeal genomes. This is a major increase in spacer targets compared to previous studies and is due to our sensitive filtering process and use of metagenomic databases (Shmakov et al. 2017). We found that Type III spacers had the highest fraction of unknown targets of any CRISPR-Cas type. This was not solely caused by the phylogenetic or environmental occurrence of Type III systems, because the fraction of Type III spacers with unknown targets within a genus was typically higher than that of other types. This means that the targets of Type III systems are either under-sampled, or that Type III spacers contain more mismatches to their targets, making them harder to find computationally. Recently, a single new study doubled the number of known RNA viruses including phages (Wolf et al. 2020), while another study greatly increased the number of known single-stranded RNA phages (Callanan et al. 2020), indicating

that RNA phages have been poorly sampled. We predict the fraction of spacers with matches to increase with increasing numbers of available metagenomic data, especially including more RNA viruses and more data from poorly sampled environments.

By analysing the flanks of the spacer hits in great depth, we have generated a catalog of PAM sequences for each CRISPR repeat cluster. The repeat sequence is a good predictor of parts of the PAM sequence in Type I, and outperformed clustering based on genus-subtype classifications. This finding is corroborated by the position-wise comparison of PAM and repeat nucleotides, which shows certain repeat nucleotides predict PAM nucleotides. This may be helpful to either predict the PAM from scratch, or to further experimentally determine the PAM while reducing the degeneracy at certain positions, limiting the predicted PAM sequence space. However, for Type II systems, this repeat based PAM prediction does not work, because PAM motifs are not conserved within each repeat sequence. Instead, PAMs in type II systems seem to be conserved within a CRISPR-Cas system combined with a certain repeat in individual strains (Additional file 1: Fig. S5C-D). Strain specific PAM analysis in Type II systems uncovered a large diversity of PAM sequences, much larger than the PAM diversity in Type I systems. Further analysis could perhaps base clustering on the PAM-interacting domains of Cas9 protein sequences, which might serve as a better predictor for PAM sequence conservation than the repeat sequence (Gasiunas et al. 2020; Qiuyan Wang et al. 2021).

The mismatch between repeat and PAM nucleotides generally holds, except for the Type I-E and Type I-C, where for some repeat clusters the repeat nucleotide matches the PAM at the -1 position. The most common PAMs of these systems (TTC for I-C; AAG for I-E) are also complementary to each other. These findings indicate Type I-C systems could have a similar mechanism of spacer acquisition with a PAM-derived last repeat nucleotide as in Type I-E (Swarts et al. 2012), even though these systems do not share related Cas1 proteins (Makarova et al. 2011) or repeat structures (Lange et al. 2013). The crystal structures of the Cas1-Cas2 adaptation machinery from both systems (Lee, Dhingra, and Sashital 2019; Wang et al. 2015) indicate that the same strand is probed for the PAM (5'GAA3' in I-C and 5'CTT3'

in I-E), which demonstrates that this phenomenon has not arisen from complementary strand selection.

The PAM catalog can be used to predict the PAM for arrays in newly sequenced genomes and metagenomic contigs in Type I and Type III systems if they contain repeats that are closely related to the repeats in our database. We furthermore have uncovered novel PAMs in Type II systems and together these developments give access to unexplored mechanistic and biotechnological potential. For repeats that are not in our database, the nucleotide identities of the repeat in the spacer flanking positions can be used to predict, with lesser certainty, which PAM it could have and select certain CRISPR-Cas systems of interest for further study.

Furthermore, the position of the PAM in the target is a reliable indicator for the orientation of transcription of CRISPR arrays. Correct prediction of transcription of CRISPR arrays gives access to measuring chronology of invader encounters and strand-specific targeting of CRISPR-Cas systems, which is especially relevant for RNA targeting CRISPR-Cas. The spacers of Type III systems, which target RNA, have a bias towards targeting coding strands, making them capable of base pairing and thereby targeting RNA. Unexpectedly we also found several subtypes with a preference for the template strand (I-E and Type II). The reason for this type of strand bias is not yet clear, but we pose that this could be caused by a selection for spacers that do not target RNA (RNA avoidance), as DNA-targeting with these spacers might be impacted by inactivating complementary RNA (Jore et al. 2011). In addition, there might be a difference in binding or dislodging of crRNA effector complexes from the template strand vs coding strand by RNA polymerase (Clarke et al. 2018; Vink et al. 2020). Lastly, we cannot exclude the possibility that DNA replication might cause the observed strand bias for some subtypes, as transcription and replication are often co-oriented in prokaryotes, plasmids and phages (Brewer 1988; Srivatsan et al. 2010).

We have categorized multi-effector compatible CRISPR arrays whether they share the same repeats and/or acquisition machinery and whether only DNA, or both DNA and RNA are targeted. DNA-targeting systems that use multi-effector compatible arrays generally have their own acquisition machinery and the low frequency of this co-occurrence in nature might indicate that this is not actively selected for. It needs

to be experimentally verified whether the spacers in these compatible arrays are actually shared between complexes. However, some crRNA sharing between DNA systems has already been observed experimentally, so it's therefore likely to be found for more systems (Majumdar et al. 2015).

Multi-effector compatible arrays are much more common in co-occurring DNA- and RNA-targeting systems. The strand bias that occurs in Type I arrays indicates that Type III effector complexes are using these spacers and thereby creating selection pressure on the RNA binding potential of the transcribed crRNA. It also seems that the most commonly co-occurring Type I systems (I-A, I-B and I-C) that use compatible arrays, also have the largest coding strand bias. Whether this strand bias is induced by the presence of Type III or whether these subtypes by their nature have a strand preference and therefore became more commonly compatible with Type III systems is not yet clear. Interestingly, many of the subtype combinations that share PAMs also co-occur more often than expected by chance, suggesting they have positive epistatic interactions (Bernheim et al. 2020). Furthermore, repeat sequences of type I-A and I-B are in the same repeat families as Type III repeats, providing further indications of their compatibility (Lange et al. 2013).

The experimentally determined spacer sharing in *Marinomonas mediterranea* (Silas et al. 2017) described previously does not fall within the categories in this study as the Type III system has its own adaptation machinery. In this case, the systems are not mutually compatible because the Type I systems cannot use the Type III spacers due to a lack of PAM, which we have not further investigated in this study. Also the other previously experimentally described spacer sharing systems in *Pyrococcus* (Majumdar et al. 2015) and *Flavobacterium* (Hoikkala et al. 2021) were not found due to a lack of sufficient hits, which demonstrates that this bio-informatic analysis likely underestimates the number of systems that can cooperate.

The discovery of multi-effector spacer compatibility in a large number of genomes in this study together with previous experimental evidence of spacer sharing of RNA and DNA-targeting systems (Deng et al. 2013; Majumdar et al. 2015; Silas et al. 2017) shows that there is selection pressure to share spacers cooperatively within arrays. The evolutionary benefits of such cooperativity could be profound. Firstly,

as two subtypes generally have different mismatch tolerance (Anderson et al. 2015; Fineran et al. 2014; Manica et al. 2013), targeting the same sequence with two subtypes can reduce the probability of escape mutation. Secondly, a combination of RNA and DNA targeting systems can provide multiple layers of defence, where RNA-targeting might give more time for DNA-targeting systems to destroy the invader before the cell is taken over (Vink et al. 2020). Thirdly the length of arrays in a genome has recently been shown to be limited by auto-immunity (H. Chen, Mayer, and Balasubramanian 2021). By sharing spacers, each subtype is supplied with a maximum diversity of spacers while self-targeting costs are minimized. Lastly, the different mechanisms these systems use allows for complementary and distinct benefits. The priming mechanism (Datsenko et al. 2012; Nicholson et al. 2019), unique to DNA targeting systems can accelerate spacer acquisition for both systems, whereas cOA signaling pathways (Kazlauskiene et al. 2017; Niewoehner et al. 2017), unique to Type III, could activate defence systems that benefit both systems.

Conclusion

Altogether this study highlights the wealth of information that can be retrieved by analysing the targets of CRISPR spacers on a large scale. It furthermore demonstrates under what conditions CRISPR-Cas systems can cooperate and provides a catalog of PAM predictions and targeted MGEs awaiting further study.

Methods

CRISPR spacers and sequence data

221 089 spacers along with information on *cas* gene presence, genome and repeat sequence were obtained from CRISPRCasdb (Pourcel et al. 2020) in February 2020 and the taxonomy of the genomes was obtained from NCBI Taxonomy database (Federhen 2012). We created our own sequence database by combining all sequences from the NCBI nucleotide database (Benson et al. 2018; Pruitt, Tatusova, and Maglott 2005), environmental nucleotide database (Sayers et al. 2009), PHASTER (Arndt et al. 2016), Mgnify (Mitchell et al. 2020), IMG/M (I. M. A. Chen et al. 2017), IMG/Vr (Paez-Espino et al. 2019), HuVirDb (Soto-Perez et al. 2019), HMP database (Peterson et al. 2009), and data from Pasolli et al., 2019. All databases were accessed in February 2020.

Subtypes were predicted based on the repeat sequences using the subtype predictions and method described by Bernheim et al., 2020, where the subtype of a spacer was inferred by the similarity of its repeat sequence to repeat sequences with known subtype (74% identity threshold to infer subtype).

Blast hits and filtering

Hits between spacers and sequences from the aforementioned databases were obtained using the command line blastn program (Altschul et al. 1990) version 2.10.0, which was run with parameters word_size 10, gapopen 10, penalty 1 and an e-value cutoff of 1, to find as many potential targets as possible. These blast hits were then filtered to remove hits of spacers inside CRISPR arrays and false positive hits found by chance. Hits inside CRISPR arrays were detected by aligning the repeat sequence of the spacer to the flanking regions of the spacer hit (23 nucleotides on both sides). This alignment was done using the globalxs function from the Biopython pairwise2 package (Cock et al. 2009) with -3 gap open and -3 gap extend parameters. If more than 13 nucleotides were identical in the alignment of at least one flank, the hit was suspected to fall inside a CRISPR array and was filtered out.

To minimize the number of hits found by chance, we filtered hits based on the fraction of spacer nucleotides that hit the target sequence, as this metric considers both the sequence identity and the coverage of the spacer by the blast hit. In a first step, only hits with this fraction higher than 90% were kept. To find targets for even more spacers while keeping the number of false positives low, we included a second step where hits with a fraction higher than 80% were kept if another spacer from the same genus hit the same contig or genome in the first step. This second step did not introduce hits on any new contigs or genomes and was based on the assumption that multiple spacers from the same genus hitting the same contig or genome is unlikely to be caused by chance. Finally, we removed spacers that were shorter than 27 nucleotides (54 spacers) and removed 7 spacers that were hitting aspecifically, such as inside ribosomal RNAs or tRNAs. This left 72,099 unique spacers with target hits for downstream analysis.

Protospacer flank alignment for orientation and PAM predictions

The PAM is known to occur on the 5' end of the protospacer for Type I, Type IV and V CRISPR-Cas systems, and on the 3' end for Type II systems (Collias and Beisel 2021; Jackson et al. 2017). We used this property to predict the orientation of transcription of CRISPR arrays and sequence of crRNA. The PAM sides were compared to the nucleotide conservation in the flanking regions of the spacer hits and the spacer orientations were predicted such that the flank with the greater conservation matched the known PAM side.

To measure the nucleotide conservation in the flanking regions, data from multiple spacers was combined based on the subtype and repeat sequences of the spacers. Highly similar repeat sequences from the same subtype were clustered using CD-HIT (Fu et al. 2012) with a 90% identity threshold. We hypothesized that similar repeat sequences would be used in a similar orientation and would utilize the same PAM sequences, as coevolution of PAM, repeat and Cas1 and Cas2 sequences has been shown previously (Alkhnbashi et al. 2014; Lange et al. 2013). For each repeat cluster the flanking regions of the spacer hits were aligned. To equally weigh each

spacer within the repeat cluster, irrespective of the number of blast hits, consensus flanks were obtained per spacer. These consensus flanks contained the most frequent nucleotide per position of the flanking regions. From the alignment of consensus flanks the nucleotide conservation, or information content, in each flank was calculated in bitscore (Schneider and Stephens 1990) using the Sequence logo python package. We corrected for GC-content of the targeted sequences by calculating the expected occurrences of each nucleotide based on the GC-content of the spacer sequences. To minimize the number of orientation predictions based on little or noisy data, we only predicted the orientation for repeat clusters when the alignment of consensus flanks consisted of at least 10 unique protospacers. Furthermore, the information content of at least two positions was higher than 0.3 bitscore and higher than 5 times the median bitscore calculated from 23-nt flanks on both sides. These parameters were chosen as strictly as possible, while still yielding orientation predictions for the highest number of spacers.

Using the orientation predictions described above, we predicted the PAMs for each repeat cluster by checking which nucleotide positions were conserved. To minimize PAM predictions based on noise, we only predicted the PAM for repeat clusters where the alignment of consensus flanks consisted of at least 10 unique protospacers. A nucleotide position was predicted to be part of the PAM when higher than 0.5 bitscore and higher than 10 times the median bitscore. These parameters were chosen as strictly as possible, while maximizing the number of repeat clusters with PAM predictions and minimizing the number of unique PAMs predicted.

We subsequently categorized and counted multi-effector compatible spacers in the following ways. Firstly by an occurrence of multiple repeat clusters with different subtype classification that both contained the same PAM, either two DNA targeting clusters (category I) or a DNA and an RNA targeting cluster (category II). Secondly, if multiple *cas* gene clusters from different subtypes were in the vicinity of a single repeat cluster and their genomes did not further contain other arrays linked to these *cas* gene clusters they were counted as a third category multi-effector compatible array.

Coding versus template strand targeting analysis

For each spacer target inside an open reading frame (ORF), we determined if the spacer targets the coding (DNA and RNA) or template strand (DNA-only) during transcription. The ORFs and their orientation were predicted using Prodigal (Hyatt et al. 2010) for one target sequence per spacer. The target sequence of each spacer was selected as the longest hit sequence in the NCBI nucleotide database, excluding 'other sequences', or, if no such sequence was hit, the longest hit sequence in metagenomics database. Using our spacer orientation predictions for Type I, II and IV spacers, and the orientation predictions from CRISPRCasdb for the other spacers, we checked if the spacer target (blast hit orientation) was on the coding or template strand of the predicted ORF. To test for significant bias towards either the temperate or the coding strand, a two-sided tailed binomial test was performed with an expected probability of 0.5.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets on which the analysis is based have been submitted as Additional Files. Scripts to reproduce figures are available on request.

Acknowledgements

The authors thank Christine Pourcel and Pierre-Albert Charbit for supplying the CRISPRCasdb in a spacer-based format and all members of the Brouns groups for input during group discussions.

Author contributions

S.B. and J.V. conceived and supervised the project; J.V. gathered databases; J.V. and J.B. wrote analysis scripts; J.V., J.B. and S.B. wrote the manuscript.

Competing interests

The authors declare no competing financial interests.

Funding

S.B. is supported by a Vici grant of the Netherlands Organisation for Scientific Research (VI.C.182.027; NWO).

Additional file description

Additional file 1:
Supplementary figures
Additional file 2:
CSV file containing for each unique spacer in the CRISPRCasDB the following columns:
Spacers: spacer sequence; **Repeats:** repeat sequence in host(s) (can be multiple if multiple genomes contain same spacer)**; Accessionnrs:** accession number of host(s)**; Subtype:** subtype of array; **cas_genes**: Cas_genes present in host(s)**; hit**: if match found in (meta)genomic database equals 1 (else 0)**; consensus**_flanks: consensus sequence of left and right flank from flanks of all the hits in databases to this spacer**; repeat_cluster:** id of repeat cluster generated with CD-hit**; strandbias:** Orientation of hit in reference to ORF (1 coding strand 0 template strand, -1 undetermined)**; type:** Type of CRISPR array**; orientation_CRISPRCasdb:** Orientation of spacer determined in CRISPRCasDB (Pourcel et al., 2020)**; orientation_PAMbased:** Orientation of spacer determined in this study based on PAM**; orientation_TOPbased:** Orientation of spacer determined with TOP

(Houenoussi et al., 2020)**; PAM:** PAM sequence of repeat cluster (if predicted)**; Genus, Family, Order ….:** Taxonomy of host**; Type I, TypeII, Type III…**: Whether host genomes contain genes related to specific Type (1 yes, 0 no)**; Subtypesingenomes:** Which subtypes are in genomes**; Subtypesinproximity:** Which subtypes are in proximity (<25000 bp from spacer)**; Proximity_subtypes:** Distance of spacer to gene cluster of specific subtype**; subtypesCas1**: Which subtypes are in genomes that contain a Cas1 protein

Additional file 3:

CSV containing PAM catalog (each unique repeat for which PAM was determined) with following columns: repeat, PAM and subtype

Additional file 4:

CSV containing genomes for which compatible arrays were found with following columns: accession number genome, compatible subtypes of array, PAM, category

Additional file 5:

CSV containing genomes for which PAM was predicted with following columns: PAM, accession number, subtype.

References

Alkhnbashi, Omer S. et al. 2014. "CRISPRstrand: Predicting Repeat Orientations to Determine the CrRNA-Encoding Strand at CRISPR Loci." In *Bioinformatics*,.

———. 2016. "Characterizing Leader Sequences of CRISPR Loci." In *Bioinformatics*,.

Almendros, Cristóbal et al. 2012. "Target Motifs Affecting Natural Immunity by a Constitutive CRISPR-Cas System in Escherichia Coli." *PLoS ONE*.

Almendros, Cristóbal, Franklin L. Nobrega, Rebecca E. McKenzie, and Stan J.J. Brouns. 2019. "Cas4-Cas1 Fusions Drive Efficient PAM Selection and Control CRISPR Adaptation." Nucleic Acids Research.

Altschul, Stephen F. et al. 1990. "Basic Local Alignment Search Tool." Journal of Molecular Biology.

Anders, Carolin, Ole Niewoehner, Alessia Duerst, and Martin Jinek. 2014. "Structural Basis of PAM-Dependent Target DNA Recognition by the Cas9 Endonuclease." Nature 513(7519): 569–73.

Anderson, Emily M. et al. 2015. "Systematic Analysis of CRISPR-Cas9 Mismatch Tolerance Reveals Low Levels of off-Target Activity." Journal of Biotechnology.

Arndt, David et al. 2016. "PHASTER: A Better, Faster Version of the PHAST Phage Search Tool." Nucleic acids research.

Artamonova, Daria et al. 2020. "Spacer Acquisition by Type III CRISPR–Cas System during Bacteriophage Infection of Thermus Thermophilus." Nucleic Acids Research.

Benson, Dennis A. et al. 2018. "GenBank." Nucleic Acids Research.

Bernheim, Aude, David Bikard, Marie Touchon, and Eduardo P.C. Rocha. 2020. "Atypical Organizations and Epistatic Interactions of CRISPRs and Cas Clusters in Genomes and Their Mobile Genetic Elements." Nucleic Acids Research.

Biswas, Ambarish, Peter C. Fineran, and Chris M. Brown. 2014. "Accurate Computational Prediction of the Transcribed Strand of CRISPR Non-Coding RNAs." Bioinformatics.

Bolotin, Alexander, Benoit Quinquis, Alexei Sorokin, and S. Dusko Ehrlich. 2005. "Clustered Regularly Interspaced Short Palindrome Repeats (CRISPRs) Have Spacers of Extrachromosomal Origin." Microbiology.

Boudry, Pierre et al. 2015. "Function of the CRISPR-Cas System of the Human Pathogen: Clostridium Difficile." mBio.

Brewer, Bonita J. 1988. "When Polymerases Collide: Replication and the Transcriptional Organization of the E. Coli Chromosome." Cell.

Brouns, Stan J J et al. 2008. "Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes." Science (New York, N.Y.) 321(5891): 960–64. http://www.ncbi.nlm.nih.gov/pubmed/18703739 (February 21, 2018).

Callanan, J. et al. 2020. "Expansion of Known SsRNA Phage Genomes: From Tens to over a Thousand." Science Advances.

Camarillo-Guerrero, Luis F. et al. 2021. "Massive Expansion of Human Gut Bacteriophage Diversity." Cell.

Cao, Linyan et al. 2016. “Identification and Functional Study of Type III-A CRISPR-Cas Systems in Clinical Isolates of Staphylococcus Aureus.” International Journal of Medical Microbiology.

Chen, Hanrong, Andreas Mayer, and Vijay Balasubramanian. 2021. “A Scaling Law in CRISPR Repertoire Sizes Arises from Avoidance of Autoimmunity.” bioRxiv: 2021.01.04.425308. http://biorxiv.org/content/early/2021/01/04/2021.01.04.425308.abstract.

Chen, I. Min A. et al. 2017. “IMG/M: Integrated Genome and Metagenome Comparative Data Analysis System.” Nucleic Acids Research.

Clarke, Ryan et al. 2018. “Enhanced Bacterial Immunity and Mammalian Genome Editing via RNA-Polymerase-Mediated Dislodging of Cas9 from Double-Strand DNA Breaks.” Molecular Cell.

Cock, Peter J.A. et al. 2009. “Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics.” Bioinformatics.

Collias, Daphne, and Chase L Beisel. 2021. “CRISPR Technologies and the Search for the PAM-Free Nuclease.” Nature Communications 12(1): 555. https://doi.org/10.1038/s41467-020-20633-y.

Cooper, Lauren A, Anne M Stringer, and Joseph T Wade. 2018. “Determining the Specificity of Cascade Binding, Interference, and Primed Adaptation In Vivo in the Escherichia Coli Type I-E CRISPR-Cas System.” mBio 9(2): e02100-17. http://www.ncbi.nlm.nih.gov/pubmed/29666291 (December 28, 2018).

Couvin, David et al. 2018. “CRISPRCasFinder, an Update of CRISRFinder, Includes a Portable Version, Enhanced Performance and Integrates Search for Cas Proteins.” Nucleic Acids Research.

Crawley, Alexandra B. et al. 2018. “Characterizing the Activity of Abundant, Diverse and Active CRISPR-Cas Systems in Lactobacilli.” Scientific Reports.

Datsenko, Kirill A. et al. 2012. “Molecular Memory of Prior Infections Activates the CRISPR/Cas Adaptive Bacterial Immunity System.” Nature Communications.

Deng, Ling et al. 2013. “A Novel Interference Mechanism by a Type IIIB CRISPR-Cmr Module in Sulfolobus.” Molecular Microbiology.

Deveau, Hélène et al. 2008. "Phage Response to CRISPR-Encoded Resistance in Streptococcus Thermophilus." Journal of Bacteriology 190(4): 1390–1400. http://jb.asm.org/cgi/doi/10.1128/JB.01412-07.

Dion, Moïra B. et al. 2021. "Streamlining CRISPR Spacer-Based Bacterial Host Predictions to Decipher the Viral Dark Matter." Nucleic Acids Research.

Elmore, Joshua R. et al. 2016. "Bipartite Recognition of Target RNAs Activates DNA Cleavage by the Type III-B CRISPR–Cas System." Genes and Development.

Federhen, Scott. 2012. "The NCBI Taxonomy Database." Nucleic Acids Research.

Fineran, P. C. et al. 2014. "Degenerate Target Sites Mediate Rapid Primed CRISPR Adaptation." Proceedings of the National Academy of Sciences 111(16): 1629–38. http://www.pnas.org/cgi/doi/10.1073/pnas.1400071111.

Fischer, Susan et al. 2012. "An Archaeal Immune System Can Detect Multiple Protospacer Adjacent Motifs (PAMs) to Target Invader DNA." Journal of Biological Chemistry.

Fu, Limin et al. 2012. "CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data." Bioinformatics.

Garcia-Heredia, Inmaculada et al. 2012. "Reconstructing Viral Genomes from the Environment Using Fosmid Clones: The Case of Haloviruses." PLoS ONE.

Gasiunas, Giedrius et al. 2020. "A Catalogue of Biochemically Diverse CRISPR-Cas9 Orthologs." Nature Communications.

Gleditzsch, Daniel et al. 2019. "PAM Identification by CRISPR-Cas Effector Complexes: Diversified Mechanisms and Structures." RNA Biology 16(4): 504–17. https://www.tandfonline.com/doi/full/10.1080/15476286.2018.1504546.

Goldberg, Gregory W., Wenyan Jiang, David Bikard, and Luciano A. Marraffini. 2014. "Conditional Tolerance of Temperate Phages via Transcription-Dependent CRISPR-Cas Targeting." Nature.

Hale, Caryn R et al. 2009. "RNA-Guided RNA Cleavage by a CRISPR RNA- Cas Protein Complex RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex." Cell 139(5): 945–56. http://dx.doi.org/10.1016/j.cell.2009.07.040.

Hayes, Robert P. et al. 2016. “Structural Basis for Promiscuous PAM Recognition in Type I–E Cascade from E. Coli.” Nature 530(7591): 499–503. http://www.ncbi.nlm.nih.gov/pubmed/26863189 (January 17, 2017).

Hoikkala, Ville et al. 2021. “Cooperation between Different CRISPR-Cas Types Enables Adaptation in an RNA-Targeting System” ed. Peter Charles Parkhill Fineran Julian. mBio 12(2): e03338-20. http://mbio.asm.org/content/12/2/e03338-20.abstract.

Horvath, Philippe et al. 2008. “Diversity, Activity, and Evolution of CRISPR Loci in Streptococcus Thermophilus.” Journal of Bacteriology 190(4): 1401–12.

Houenoussi, Kimberley et al. 2020. “TOP the Transcription Orientation Pipeline and Its Use to Investigate the Transcription of Non-Coding Regions: Assessment with CRISPR Direct Repeats and Intergenic Sequences.” bioRxiv: 2020.01.15.903914. http://biorxiv.org/content/early/2020/01/15/2020.01.15.903914.abstract.

Hyatt, Doug et al. 2010. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.” BMC Bioinformatics.

Jackson, Simon A. et al. 2017. “CRISPR-Cas: Adapting to Change.” Science.

Johnson, Kaitlin, Brian A Learn, Michael A Estrella, and Scott Bailey. 2019. “Target Sequence Requirements of a Type III-B CRISPR-Cas Immune System.” Journal of Biological Chemistry 294(26): 10290–99. https://doi.org/10.1074/jbc.RA119.008728.

Jore, Matthijs M et al. 2011. “Structural Basis for CRISPR RNA-Guided DNA Recognition by Cascade.” Nature Structural and Molecular Biology 18(5): 529–36. http://www.nature.com/doifinder/10.1038/nsmb.2019 (January 17, 2017).

Kazlauskiene, Migle et al. 2017. “A Cyclic Oligonucleotide Signaling Pathway in Type III CRISPR-Cas Systems.” Science 357(6351): 605–9.

Kieper, Sebastian N. et al. 2018. “Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation.” Cell Reports 22(13): 3377–84.

Krupovic, Mart et al. 2014. “Casposons: A New Superfamily of Self-Synthesizing DNA Transposons at the Origin of Prokaryotic CRISPR-Cas Immunity.” BMC Biology.

Lange, Sita J. et al. 2013. “CRISPRmap: An Automated Classification of Repeat Conservation in Prokaryotic Adaptive Immune Systems.” Nucleic Acids Research.

Lee, Hayun, Yukti Dhingra, and Dipali G. Sashital. 2019. “The Cas4-Cas1-Cas2 Complex Mediates Precise Prespacer Processing during CRISPR Adaptation.” eLife.

Lee, Hayun, Yi Zhou, David W. Taylor, and Dipali G. Sashital. 2018. “Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays.” Molecular Cell.

Leenay, Ryan T. et al. 2016. “Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems.” Molecular Cell 62(1): 137–47.

Leenay, Ryan T., and Chase L. Beisel. 2017. “Deciphering, Communicating, and Engineering the CRISPR PAM.” Journal of Molecular Biology.

Li, Ming, Rui Wang, Dahe Zhao, and Hua Xiang. 2014. “Adaptation of the Haloarcula Hispanica CRISPR-Cas System to a Purified Virus Strictly Requires a Priming Process.” Nucleic Acids Research.

Lillestøl, Reidun K. et al. 2009. “CRISPR Families of the Crenarchaeal Genus Sulfolobus: Bidirectional Transcription and Dynamic Properties.” Molecular Microbiology.

Lin, Jinzhong et al. 2020. “DNA Targeting by Subtype I-D CRISPR-Cas Shows Type i and Type III Features.” Nucleic Acids Research.

Lopatina, Anna et al. 2019. “Natural Diversity of CRISPR Spacers of Thermus: Evidence of Local Spacer Acquisition and Global Spacer Exchange.” Philosophical Transactions of the Royal Society B: Biological Sciences.

Magadán, Alfonso H., Marie Ève Dupuis, Manuela Villion, and Sylvain Moineau. 2012. “Cleavage of Phage DNA by the Streptococcus Thermophilus CRISPR3-Cas System.” PLoS ONE.

Majumdar, Sonali et al. 2015. “Three CRISPR-Cas Immune Effector Complexes Coexist in Pyrococcus Furiosus.” RNA (New York, N.Y.).

Makarova, Kira S. et al. 2015. “An Updated Evolutionary Classification of CRISPR–Cas Systems.” Nature Reviews Microbiology 13(11): 722–36. http://www.nature.com/doifinder/10.1038/nrmicro3569 (January 17, 2017).

———. 2020. “Evolutionary Classification of CRISPR–Cas Systems: A Burst of Class 2 and Derived Variants.” Nature Reviews Microbiology.

Makarova, Kira S et al. 2011. “Evolution and Classification of the CRISPR–Cas Systems.” Nature Publishing Group 9(6): 467–77. http://dx.doi.org/10.1038/nrmicro2577.

Malone, Lucia M. et al. 2020. “A Jumbo Phage That Forms a Nucleus-like Structure Evades CRISPR–Cas DNA Targeting but Is Vulnerable to Type III RNA-Based Immunity.” Nature Microbiology.

Manica, Andrea, Ziga Zebec, Julia Steinkellner, and Christa Schleper. 2013. “Unexpectedly Broad Target Recognition of the CRISPR-Mediated Virus Defence System in the Archaeon Sulfolobus Solfataricus.” Nucleic Acids Research.

Manica, Andrea, Ziga Zebec, Daniela Teichmann, and Christa Schleper. 2011. “In Vivo Activity of CRISPR-Mediated Virus Defence in a Hyperthermophilic Archaeon.” Molecular Microbiology.

Marraffini, Luciano A. 2015. “CRISPR-Cas Immunity in Prokaryotes.” Nature 526(7571): 55–61. http://www.nature.com/doifinder/10.1038/nature15386 (January 19, 2017).

Marraffini, Luciano A, and Erik J Sontheimer. 2010. “Self versus Non-Self Discrimination during CRISPR RNA-Directed Immunity.” Nature 463(7280): 568–71. https://doi.org/10.1038/nature08703.

Marshall, Ryan et al. 2018. “Rapid and Scalable Characterization of CRISPR Technologies Using an E. Coli Cell-Free Transcription-Translation System.” Molecular Cell.

Mendoza, Brian J., and Cong T. Trinh. 2018. “In Silico Processing of the Complete CRISPR-Cas Spacer Space for Identification of PAM Sequences.” Biotechnology Journal.

Mendoza, Senén D. et al. 2020. “A Bacteriophage Nucleus-like Compartment Shields DNA from CRISPR Nucleases.” Nature 577(7789): 244–48.

Milicevic, Ognjen et al. 2019. “A Simple Criterion for Inferring CRISPR Array Direction.” Frontiers in Microbiology.

Mitchell, Alex L. et al. 2020. “MGnify: The Microbiome Analysis Resource in 2020.” Nucleic acids research 48(D1): D570–78. https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz1035/5614179.

Mojica, F. J M, C. Díez-Villaseñor, J. García-Martínez, and C. Almendros. 2009. "Short Motif Sequences Determine the Targets of the Prokaryotic CRISPR Defence System." Microbiology 155(3): 733–40.

Musharova, Olga et al. 2019. "Systematic Analysis of Type I-E Escherichia Coli CRISPR-Cas PAM Sequences Ability to Promote Interference and Primed Adaptation." Molecular Microbiology.

Nicholson, Thomas J. et al. 2019. "Bioinformatic Evidence of Widespread Priming in Type I and II CRISPR-Cas Systems." RNA Biology 16(4): 566–76.

Niewoehner, Ole et al. 2017. "Type III CRISPR-Cas Systems Produce Cyclic Oligoadenylate Second Messengers." Nature 548(7669): 543–48.

Nobrega, Franklin, Hielke Walinga, Bas Dutilh, and Stan Brouns. 2020. "Prophages Are Associated with Extensive, Tolerated CRISPR-Cas Auto-Immunity." bioRxiv: 2020.03.02.973784.

Paez-Espino, David et al. 2019. "IMG/VR v.2.0: An Integrated Data Management and Analysis System for Cultivated and Environmental Viral Genomes." Nucleic Acids Research.

Pan, Meichen, Matthew A. Nethery, Claudio Hidalgo-Cantabrana, and Rodolphe Barrangou. 2020. "Comprehensive Mining and Characterization of Crispr-Cas Systems in Bifidobacterium." Microorganisms.

Pasolli, Edoardo et al. 2019. "Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle." Cell 176(3): 649-662.e20.

Pawluk, April, Alan R Davidson, and Karen L Maxwell. 2017. Anti-CRISPR: Discovery, Mechanism and Function.

Peterson, Jane et al. 2009. "The NIH Human Microbiome Project." Genome Research.

Pinilla-Redondo, Rafael et al. 2020. "Type IV CRISPR-Cas Systems Are Highly Diverse and Involved in Competition between Plasmids." Nucleic Acids Research.

Pourcel, Christine et al. 2020. "CRISPRCasdb a Successor of CRISPRdb Containing CRISPR Arrays and Cas Genes from Complete Genome Sequences, and Tools to

Download and Query Lists of Repeats and Spacers." Nucleic Acids Research 48(D1): D535–44.

Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. 2005. "NCBI Reference Sequence (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins." Nucleic Acids Research.

Pujato, Silvina et al. 2021. "Analysis of CRISPR Systems of Types II-A, I-E and I-C in Strains of Lacticaseibacillus." International Dairy Journal.

Qiuyan Wang, Xiaomeng Bian et al. 2021. "PAM-Interacting Domain Swapping Is Extensively Utilized in Nature to Evolve CRISPR-Cas9 Nucleases with Altered PAM Specificities." bioRxiv.

Sayers, Eric W. et al. 2009. "Database Resources of the National Center for Biotechnology Information." Nucleic Acids Research.

Schneider, Thomas D., and R. Michael Stephens. 1990. "Sequence Logos: A New Way to Display Consensus Sequences." Nucleic Acids Research.

Shah, Shiraz A., Susanne Erdmann, Francisco J.M. Mojica, and Roger A. Garrett. 2013. "Protospacer Recognition Motifs: Mixed Identities and Functional Diversity." RNA Biology.

Shiimori, Masami, Sandra C. Garrett, Brenton R. Graveley, and Michael P. Terns. 2018. "Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci." Molecular Cell.

Shmakov, Sergey A. et al. 2017. "The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes." mBio.

Silas, Sukrit et al. 2016a. "Direct CRISPR Spacer Acquisition from RNA by a Natural Reverse Transcriptase-Cas1 Fusion Protein." Science.

———. 2016b. "Direct CRISPR Spacer Acquisition from RNA by a Natural Reverse Transcriptase-Cas1 Fusion Protein." Science 351(6276).

———. 2017. "Type III CRISPR-Cas Systems Can Provide Redundancy to Counteract Viral Escape from Type I Systems." eLife.

Soto-Perez, Paola et al. 2019. "CRISPR-Cas System of a Prevalent Human Gut Bacterium Reveals Hyper-Targeting against Phages in a Human Virome Catalog." Cell host & microbe 26(3): 325-335.e5. https://linkinghub.elsevier.com/retrieve/pii/S1931312819304172.

Srivatsan, Anjana, Ashley Tehranchi, David M. MacAlpine, and Jue D. Wang. 2010. "Co-Orientation of Replication and Transcription Preserves Genome Integrity." PLoS Genetics.

Swarts, Daan C, Cas Mosterd, Mark W J van Passel, and Stan J J Brouns. 2012. "CRISPR Interference Directs Strand Specific Spacer Acquisition." PLOS ONE 7(4): e35888. https://doi.org/10.1371/journal.pone.0035888.

Vale, Pedro F. et al. 2015. "Costs of CRISPR-Cas-Mediated Resistance in Streptococcus Thermophilus." Proceedings of the Royal Society B: Biological Sciences.

Vink, Jochem N.A. et al. 2020. "Direct Visualization of Native CRISPR Target Search in Live Bacteria Reveals Cascade DNA Surveillance Mechanism." Molecular Cell 77(1): 39-50.e10. https://linkinghub.elsevier.com/retrieve/pii/S1097276519307993.

Walker, Julie E. et al. 2020. "Development of Both Type I–B and Type II CRISPR/Cas Genome Editing Systems in the Cellulolytic Bacterium Clostridium Thermocellum." Metabolic Engineering Communications.

Walton, Russell T, Jonathan Y Hsu, J Keith Joung, and Benjamin P Kleinstiver. 2021. "Scalable Characterization of the PAM Requirements of CRISPR-Cas Enzymes Using HT-PAMDA." Nature protocols.

Wang, Jiuyu et al. 2015. "Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems." Cell: 1–14. http://dx.doi.org/10.1016/j.cell.2015.10.008.

Westra, Edze R. et al. 2013. "Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base Pairing-Independent PAM Recognition" ed. Patrick H. Viollier. PLoS Genetics 9(9): e1003742. http://dx.plos.org/10.1371/journal.pgen.1003742 (January 17, 2017).

Westra, Edze R., Angus Buckling, and Peter C. Fineran. 2014. "CRISPR-Cas Systems: Beyond Adaptive Immunity." Nature Reviews Microbiology.

Wimmer, Franziska, and Chase L. Beisel. 2020. "CRISPR-Cas Systems and the Paradox of Self-Targeting Spacers." Frontiers in Microbiology.

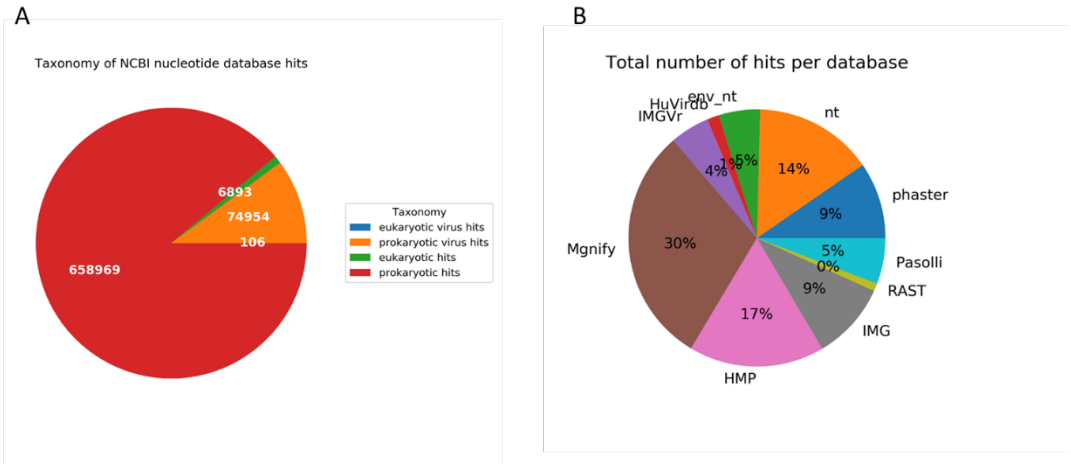
Wolf, Yuri I. et al. 2020. “Doubling of the Known Set of RNA Viruses by Metagenomic Analysis of an Aquatic Virome.” Nature Microbiology.

Xiao, Guohui et al. 2019. “Characterization of CRISPR-Cas Systems in Leptospira Reveals Potential Application of CRISPR in Genotyping of Leptospira Interrogans.” APMIS.

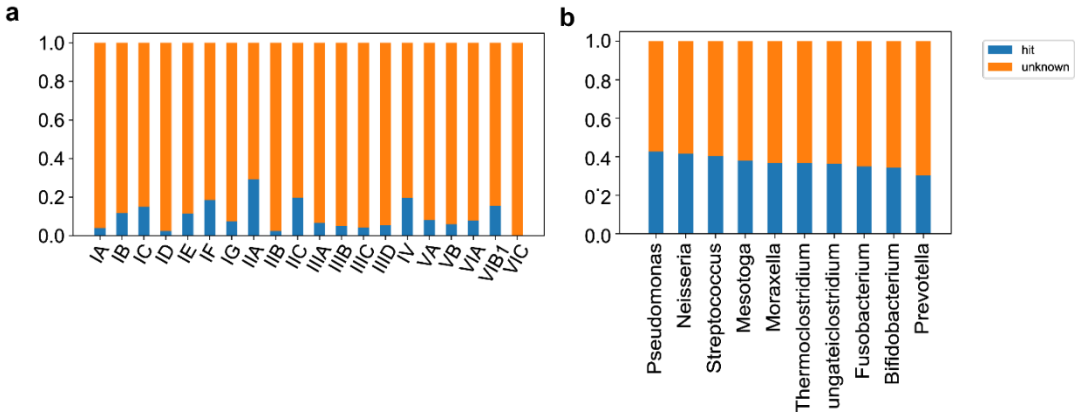
Xue, Chaoyou et al. 2017. “Real-Time Observation of Target Search by the CRISPR Surveillance Complex Cascade.” Cell Reports 21(13): 3717–27.

Type	PAM	Genus	Ref	Type	PAM	Genus	Ref	Type	PAM	Genus	Ref
I-A	ATG	*Leptospira*	**1**	**I-B**	TTTA	*Petrimonas*		**I-E**	AAG	*Geobacter*	12
	CCN	*Acidianus*	2		TTG	*Thermobacillus*			AAN	*Klebsiella*	
	TTA	*Thermodesulf-obacterium*		**I-C**	CTN	*Anaerobutyricum*			AAT	*Lactobacillus*	
	TCN	*Sulfurisphaera*	3		CCN	*Porphyromonas*			AAA	*Kosakonia*	13
	ATN	*Aminobacterium*			CTT	*Ruminococcus*			AC	*Corynebacterium*	
I-B	CCA	*Moorella*	**4**		TTC	*Geobacillus*	8		AG	*Xenorhabdus*	
	CCN	*Clostridium*	4		TTN	*Acidovorax*			AWG	*Escherichia*	
	CCT	*Ureibacillus*	4		TTT	*Lachnoclostridium*		**I-F**	ACC	*Aeromonas*	
	TAC	*Halorubrum*		**I-D**	GCN	*Haloquadratum*			CC	*Pseudomonas*	14
	TCA	*Campylobacter*			GGTG	*Halorubrum*			CCA	*Pseudomonas*	
	TCN	*Campylobacter*			GTN	*Methanotrix*	9	**I-G**	TAC	*Rothia*	
	TTA	*Methanosarcina*	**5**		GTT	*Microcystis*			TAN	*Propionibacterium*	11
	TTC	*Halobacterium*	**6**		GTG	*Methanospirillum*	10		TTN	*Pseudopropionibacterium*	15
	TTN	*Novibacillus*	**7**	**I-E**	AAC	*Bifidobacterium*	11		AAN	*Acidipropionibacterium*	
									TTC	*Rhodothermus*	

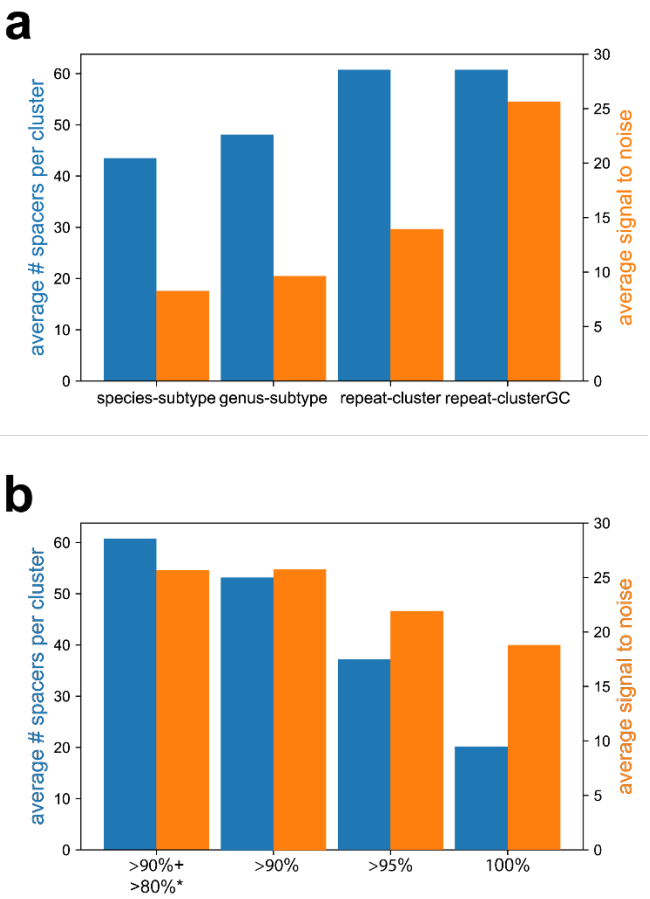
Table 1. Unique Type I PAM sequences. Table of all unique Type I PAMs found for the different subtypes and representative genera that contain the repeat cluster for which each PAM was determined. For previously described PAMs a reference ID has been added which correspond to the following: 1: (Xiao et al. 2019); 2: (Lillestøl et al. 2009), 3: (Manica et al. 2011), 4: (Boudry et al. 2015), 5: (Fischer et al. 2012), 6: (Li et al. 2014), 7: (Walker et al. 2020), 8: (Leenay et al. 2016), 9: (Kieper et al. 2018), 10: (Lin et al. 2020), 11: (Pan et al. 2020), 12: (Swarts et al. 2012), 13: (Pujato et al. 2021), 14: (Almendros et al. 2012), 15: (Almendros et al. 2019).



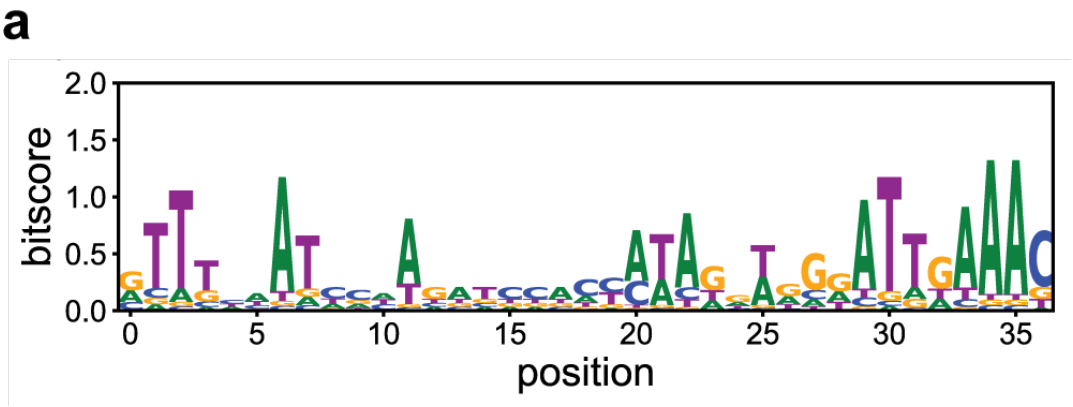
Supplementary figure 1. Taxonomy of spacer targets and number of found targets per database. (A) The taxonomy of targeted sequences of the NCBI nucleotide database was obtained from the NCBI taxonomy database. For hits in viral sequences, the taxonomy of known hosts was used to label the virus as a eukaryotic or prokaryotic virus. (B) The contribution of each database to the total number of hits after filtering. All databases were accessed in February 2020.



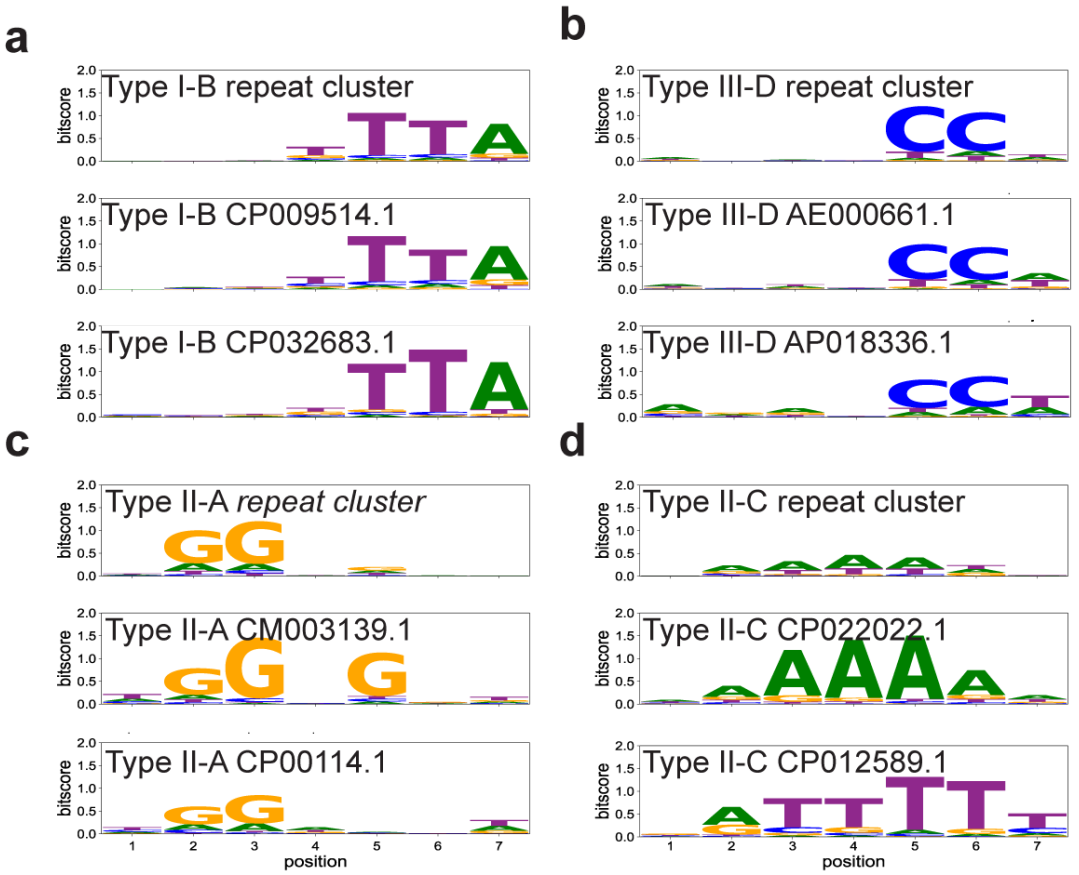
Supplementary figure 2. Perfect match statistics. (A) Distribution of fraction of hits per subtype as in Figure 1G but only for perfect matching (100% identity) spacers. (B) Distribution of fraction of hits for the ten highest scoring genera as in Figure 1E (with at least 500 spacers) but only for perfect matching spacers.



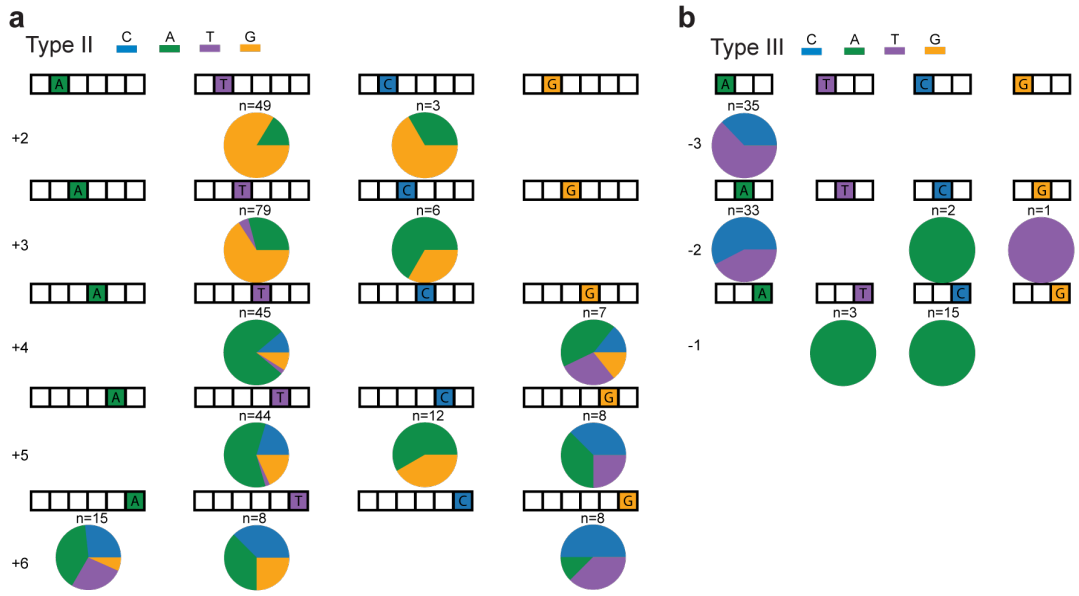
Supplementary figure 3. Average number and signal-to-noise ratio of clustered hits. Different clustering methods were compared for their average number of unique hits (blue) and average signal-to-noise ratio (orange). The signal-to-noise ratio was calculated by dividing the average information content of the two top positions in the flank (potential PAM nucleotides) by the median information content in sequence logos generated from flanks of hits. (B) The clustering categories are based on whether spacers come from same species and subtype (species-subtype), from same genus and subtype (genus-subtype), from clusters of repeat sequences with 90% identity (repeat-cluster) or clusters of repeat sequences with 90% identity and additionally compensation for GC-content of spacers within the cluster (see Materials and Methods, repeat-clusterGC). (B) The average number of hits and signal-to-noise ratio in case of using increasing levels of nucleotide identity to filter hits. The >80%* indicates hits that have >80% identity but they are only accepted in case a different spacer within the same genus targets the same sequence.



Supplementary figure 4. Sequence logo Type III repeats. ClustalW alignment of Type III repeats for which orientation was determined based on presence of PAM (n = 21 unique repeats). The 3'end of the repeat, which is the 5' handle of the transcribed crRNA, has a conserved motif (ATTGAAAC).



Supplementary figure 5. Sequence logo of protospacer flanking regions of Type I and Type II systems based on clustering approach. For each subtype ((A) Type I-B; (B) Type III-D; (C) Type II-A and (D) Type II-C) a representative cluster was chosen and a sequence logo was made as in Figure 2 for either all spacers within that cluster (top), or two representative genomes within that cluster (middle, bottom). For Type I and Type III system, the PAM is conserved within a repeat cluster (A) and (B) whereas for Type II systems, the PAM can differ within a cluster (C,D).



Supplementary Figure 6. Relationship between repeat and PAM sequence of Type II and Type III systems. Same as Figure 3B except for Type II (A) and Type III (B) systems

Summary

Viruses are, as we have seen over the past year, very proficient at invading a host and subsequently reproducing rapidly. Our adaptive immune system, after a first encounter with a virus, can store information about the outside protein shell and use this information to destroy the virus in later encounters. Bacteria have an adaptive immune system that works on the nucleic acid level (DNA/RNA). The right functioning of the system demands that a fragment of a virus is incorporated into the bacterial genome correctly (adaptation). Subsequently, proteins carrying copies of this fragment have to find and destroy the virus quickly enough after it enters the cell (interference). An obstacle in the interference process is that the cell is filled with host DNA, which has to be scanned to differentiate it from viral DNA. How bacteria are still able to find these viruses fast enough is the subject of my thesis.

In **chapter 1** I introduce three important fields of research that are needed to follow the rest of my thesis: CRISPR-Cas, single-particle tracking and bioinformatics. In **chapter 2** we follow Cascade, a CRISPR-Cas complex of proteins and RNA that naturally occurs in *E. coli*, searching for invasive DNA molecules. We have determined the efficiency of the immune system, by quantifying the number of Cascade complexes needed for a certain level of protection and we found that roughly 20 Cascade complexes are required to have a 50% of clearing the invader. We subsequently investigated the scanning speed and time and found that Cascade spends half its time on DNA and the other half moving to the next site on the DNA. The transition between these two states occurs 30 times a second. We furthermore observed that complexes remained bound to bona fide targets for a long time, but when bound to bacterial CRISPR arrays, some of the complexes disintegrated. This study showed what kinetic challenges the CRISPR immunity faces in the context of a native immune system.

In **chapter 3** we follow Cas9, a protein with a similar function as the Cascade complex, but made up of only a single protein. We have engineered a strain of

Lactoccocus lactis, an important model organism in the food industry. We find that the scanning speed also occurs roughly 30 times a second in this system, which indicates a shared biophysical characteristic of CRISPR systems. In this chapter we have also looked into the influence of Cas9 copy number on target binding and find that at high copies there is a relatively lower fraction of Cas9 bound, which suggests saturation of potential binding sites.

In **chapter 4** we follow a different Cascade variant that occurs naturally in *E. coli*, namely the I-F system. As these two different systems occur in the same species, and in some cases the same individual, we were curious what the similarities and differences between this system (I-F Cascade) and the previously studied I-E Cascade are. We found a similar protective efficiency (25 complexes provided a 50% chance of invader clearance), and similar binding rates, even though the kinetics of diffusion in the cell were very different. We furthermore investigated the influence of mutations on the binding of I-F Cascade to its target sequence. This study brought forward what similarities and differences occur between different native CRISPR-Cas systems.

In **chapter 5** I discuss a mathematical framework we used to study the interaction kinetics of CRISPR-Cas proteins with DNA. Our model enabled us to study much faster interactions that other models could not detect. To demonstrate the applicability of this analysis method, we re-analyzed published data on DNA polymerase, the DNA replication and repair protein, and find that it undergoes extremely fast DNA interactions (on the order of 100 times per second) in its search for replication and repair sites.

In **chapter 6** we explore the functioning of CRISPR-Cas proteins outside of the model organisms discussed in preceding chapters. We have done this by searching for DNA sequences that matched a large database of spacers, fragments of past invaders that were incorporated in the CRISPR array. By studying these matching DNA sequences we could predict on a large scale what requirements CRISPR spacers have for each subtype of CRISPR. The requirements we investigated were the presence or absence of a sequence motif adjacent to the viral DNA matching the spacer (PAM) and the presence or absence of a strand preference in reference to the orientation of transcription. When we tested these requirements for co-occurring

systems, multiple CRISPR-Cas systems occurring in a single genome, we found that many spacers are able to fit the requirement for both DNA and RNA-targeting systems, which enables these systems to cooperate by sharing their spacers.

Samenvatting

Virussen zijn, zoals we de afgelopen tijd gemerkt hebben, erg goed in staat een gastheer binnen te dringen en zich vervolgens snel te vermenigvuldigen. Ons adaptief immuun systeem, na een eerste kennismaking met een virus, kan informatie over de buitenkant van dat virus, een capside van eiwitten, opslaan en vervolgens die informatie gebruiken om een virus onschadelijk te maken. Bacteriën hebben een adaptief immuun systeem, dat werkt op het niveau van het nucleïnezuur (DNA en RNA). De juiste werking van het systeem vereist dat een fragment van een virus als informatie wordt ingebouwd in het genoom van de bacterie (adaptatie). Vervolgens moeten andere eiwitten met kopieën van die fragmenten snel genoeg het virus vinden en vernietigen (interferentie). Een obstakel in het interferentieproces is dat de cel gevuld is met DNA van de gastheer, wat gescand moet worden om onderscheid met virus DNA te kunnen maken. Hoe bacteriën toch in staat zijn om snel genoeg deze virussen te herkennen is het onderwerp van mijn proefschrift.

In **hoofdstuk 1** introduceer ik drie belangrijke onderzoeksvelden die nodig zijn om de rest van mijn proefschrift te kunnen volgen: CRISPR-Cas, single-particle tracking en bioinformatica. In **hoofdstuk 2** volgen we de zoektocht van Cascade, een CRISPR-Cas complex van eiwitten en RNA die van nature voorkomt in *E. coli* in de cel, op zoek naar indringende DNA moleculen. We hebben de efficiëntie van het immuunsysteem bepaald, door te kwantificeren hoeveel Cascade complexen nodig zijn voor een bepaalde bescherming en vonden dat ongeveer 20 Cascade complexen nodig zijn voor een 50% kans om de indringer uit te schakelen. Vervolgens hebben we gekeken naar de scansnelheid en vonden dat Cascade de helft van de tijd doorbrengt op DNA en de andere helft onderweg is naar een volgende locatie op het DNA. De omschakeling tussen deze twee toestanden vindt 30 x per seconde plaats. Verder zagen we dat complexen lang gebonden bleven aan doelwitsequenties, maar complexen gebonden aan de bacteriële sequentie uit elkaar vielen. Deze studie liet

zijn welke kinetische uitdagingen het CRISPR immuunsysteem moet overwinnen in de context van een natief immuun systeem.

In **hoofdstuk 3** volgen we Cas9, een eiwit dat dezelfde functie heeft als Cascade, maar alleen maar bestaat uit een enkel eiwit. Wij hebben dit eiwit tot expressie gebracht in *Lactococcus lactis*, een belangrijk modelorganisme in de voedingsindustrie. Wij vinden dat ook hier de scanningssnelheid grofweg 30 x per seconde plaatsvindt, wat duidt op een gedeelde biofysische eigenschap van CRISPR systemen. In dit hoofdstuk hebben we ook afzonderlijk nog gekeken naar de invloed van de hoeveelheid kopieën van Cas9 in de cel op de binding aan het doelwit en vinden bij hoge kopieën een relatief lager aandeel aan gebonden Cas9, wat duidt op saturatie van de mogelijke bindingsplekken.

In **hoofdstuk 4** volgen we een andere Cascade die in *E. coli* van nature voorkomt, namelijk het I-F systeem. Sinds deze twee verschillende systemen beiden voorkomen in dezelfde soort, en in sommigen gevallen hetzelfde individu, vroegen we ons af wat de overeenkomsten en verschillen zijn tussen het eerder bestuurde I-E Cascade en het hier bestudeerde I-F Cascade. Wij hebben gevonden dat dit systeem een soortgelijke efficiëntie van bescherming kent (met 25 complexen is er 50% kans om een indringer uit te schakelen) en soortgelijke bindingssnelheid aan DNA, maar dat de kinetiek van de diffusie in de cel anders verloopt. Verder hebben we hier ook nog onderzocht wat de invloed van mutaties is op de binding van I-F Cascade aan zijn doelsequentie. Deze studie bracht naar voren welke overeenkomsten en verschillen er bestaan tussen genetisch verschillende CRISPR-Cas systemen.

In **hoofdstuk 5** behandel ik een wiskundig model die we gebruikt hebben om de interactiekinetiek van CRISPR-eiwitten met DNA. Ons model is in staat om veel kortere interacties te herkennen dan eerdere modellen. Om de algemenere toepasbaarheid van dit model te laten zien hebben we eerder gepubliceerd werk over DNA polymerase opnieuw geanalyseerd. Daar vonden we dat DNA polymerase, het eiwit verantwoordelijk voor de replicatie en reparatie van DNA, erg snelle DNA interacties gebruikt om de plek voor replicatie en reparatie te vinden (100x per seconde).

In **hoofdstuk 6** verkennen we de werking van CRISPR-Cas systemen buiten de modelorganismen beschreven in de eerdere hoofdstukken. Dit hebben we gedaan door voor een groot aantal spacers, de fragmenten van virussen die ingebouwd worden door CRISPR systemen, een overeenkomende DNA sequenties te vinden in grote databases van DNA (meta)genomen. Deze overeenkomende sequenties komen van virussen en ander DNA dat de cel heeft geprobeerd binnen te dringen. Door deze overeenkomende sequenties verder te bestuderen konden we op grote schaal voorspellen aan welke eisen de fragmenten moesten voldoen die geselecteerd werden in elk systeem. De eisen zijn het al dan niet aanwezig zijn van een bepaald aangrenzend motief in het viraal DNA en het al dan niet aanwezig zijn van een voorkeur voor een bepaalde streng van het DNA, afhankelijk van de richting van transcriptie. Toen we deze eisen onderzochten voor gelijktijdig voorkomende systemen, meerdere CRISPR-Cas systemen in een enkel genoom, vonden we dat veel spacers aan zowel de eisen van een RNA en een DNA bindend systeem voldoen, wat het mogelijk maakt dat deze spacers gedeeld worden door samenwerkende CRISPR-Cas systemen.

About the author

Jochem was born on the 14th of July 1991 and defended his PhD thesis on the 20th of September 2021. He likes to keep things simple.

List of publications

J. N. A. Vink, J.H.L. Baijens, S. J. J. Brouns, PAM-repeat associations and spacer selection preferences in single and co-occurring CRISPR-Cas systems. *Genome Biology* (2021), accepted.

J. N. A. Vink, S. J. J. Brouns, J. Hohlbein, Extracting Transition Rates in Particle Tracking Using Analytical Diffusion Distribution Analysis. *Biophys. J.* (2020), doi:10.1016/j.bpj.2020.09.033.

J. N. A. Vink, K. J. A. Martens, M. Vlot, R. E. McKenzie, C. Almendros, B. Estrada Bonilla, D. J. W. Brocken, J. Hohlbein, S. J. J. Brouns, Direct Visualization of Native CRISPR Target Search in Live Bacteria Reveals Cascade DNA Surveillance Mechanism. *Mol. Cell*. **77**, 39-50.e10 (2020).

K. J. A. Martens, S. P. B. van Beljouw, S. van der Els, J. N. A. Vink, S. Baas, G. A. Vogelaar, S. J. J. Brouns, P. van Baarlen, M. Kleerebezem, J. Hohlbein, Visualisation of dCas9 target search in vivo using an open-microscopy framework. *Nat. Commun.* **10**, 3552 (2019).

R. E. McKenzie, E. M. Keizer, J. N. A. Vink, J. van Lopik, F. Büke, V. Kalkman, C. Fleck, S. J. Tans, S. J. J. Brouns, Single Cell Variability of CRISPR-Cas Interference and Adaptation. *bioRxiv*, in press, doi:10.1101/2021.07.21.453200.

S. P. B. van Beljouw, A. C. Haagsma, A. Rodríguez-Molina, D. F. van den Berg, J. N. A. Vink, S. J. J. Brouns, The gRAMP CRISPR-Cas effector is an RNA endonuclease complexed with a caspase-like peptidase. *Science (80-.).*, eabk2718 (2021).

C. Hu, C. Almendros, K. H. Nam, A. R. Costa, J. N.A. Vink, S. R. Bagde, S. J.J. Brouns, A. Ke, A High-resolution Mechanism for Cas4-assisted PAM-selection and directional spacer acquisition in CRISPR-Cas. *Nature* (2021), accepted.

S. N. Kieper, C. Almendros, J. Behler, R. E. McKenzie, F. L. Nobrega, A. C. Haagsma, J. N. A. Vink, W. R. Hess, S. J. J. Brouns, Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. *Cell Rep.* **22**, 3377–3384 (2018).

I. Iermak, J. Vink, A. N. Bader, E. Wientjes, H. Van Amerongen, Visualizing heterogeneity of photosynthetic properties of plant leaves with two-photon fluorescence lifetime imaging microscopy. *Biochim. Biophys. Acta - Bioenerg.* (2016), doi:10.1016/j.bbabio.2016.05.005.

L. Marivaux, R. Salas-Gismondi, J. Tejada, G. Billet, M. Louterbach, J. Vink, J. Bailleul, M. Roddaz, P.-O. Antoine, A platyrrhine talus from the early Miocene of Peru (Amazonian Madre de Dios Sub-Andean Zone). *J. Hum. Evol.* **63**, 696–703 (2012).

Acknowledgements

Getting your doctorate originally meant getting a license to teach (licentia docendi). Being able to teach should imply you have learnt something that you can then pass on. The most important things I have learnt are however not in the first 293 pages of this book. For that you need to be here (probably you already managed to skip the rest to get here). Considering this is still a very serious thesis, I will focus on those elements of each of you that I have learnt from and in the end helped me to finish this piece of work.

Stan, here we are, finally. My first day in Wageningen as a starting master student, is when I met you (8 years ago!). I was immediately impressed by this new CRISPR field that you had started in. I sometimes wonder how different my life would have been, if I had not taken the introductory tour of Anne Vogel through campus, or if she had chosen to show different faculties than the one she was working in. But that is what happened. The first thing you taught me, is how cool bacterial immune systems are! Thank you for taking me on this great adventure. You also taught me how to write, how to make the story as simple as possible (even though sometimes it wasn't really possible). And I learnt a lot about the best fishing spots in the Netherlands, maybe will try them out some day. Thanks for your support and the space I got during the hardest moment in my life. And maybe the most important way in which you changed my life is gathering this great group of people together. Those I will discuss in a moment.

Johannes, I was jealous when Jaap managed to get that internship with you and I had to change plans and go for photosynthesis studies (although we had some great coffee chats together with Arjen). Never thought to still be around almost 8 years later. I have learnt too much to mention here, but some of the things will include, that I learnt how to phrase things more precisely, how microscopes actually work, how to systematically optimize a setup. You really pushed the project further in crucial moments. Thanks for all those years of involvement, all the rides back to the west of the country, all the coffee and beer moments together.

Carina, ondanks dat het alweer 10 jaar geleden is dat ik mijn bachelor project bij jou volgde, moet ik je toch ook hier vermelden. Want het was bij jou dat ik leerde hoe die

wonderlijke academische wereld in elkaar steekt, hoe ik een onderzoek moest opzetten en hoe ik die moest afronden. Bedankt voor alle tijd en steun die je bereid was te stoppen in dat een beetje uit de hand gelopen bachelor project.

Van mijn studiegenootjes in Amsterdam, **Sjoerd** en **Lizzy**, leerde ik dat de beste presentaties over de meest willekeurige onderwerpen gaan. Hoe je liedjes kunt maken uit het niets en hoe je met zijn vieren in één bed slaapt. En **Guus** jij leerde mij dat je altijd je pad weer kan hervormen en je niet gebonden bent aan de keuzes in het verleden, hoe enorm tof Chili is en hoeveel verschillende kapsels 1 persoon kan hebben. **Dirk,** van jou leerde ik de soms bizarre wereld van de chemische industrie kennen en leerde ik hoe het is om dat leven te combineren met dat van een topsporter. Sorry dat ik jullie zo weinig heb gezien de afgelopen tijd, het heeft niet aan jullie gelegen.

Dan voor alle lieve mensen van familie Vink, **Heleen**, **Bart**, **Dirk**, **Cees**, **Dick**, **Thomas**, **Steven**, **Renée**, **Dezsö**, **Lennox**, **Wil**, **Annet**, **Martie**, **Ton, Loes**. Van jullie leerde ik dat roodborstjes toch magische vogeltjes zijn. Dat risk of poker uit kunnen lopen tot jarenlange familievetes. Dat je mag zijn wie je maar wilt zijn.

Hetzelfde geldt voor de familie van de Ven, **Suzan**, **Dennis**, **Rian**, **Rik, Anny**, **Nico**, **Tinie**, **Gerrit**, hartelijk dank voor het altijd warme ontvangst en de gezellige familieweekenden. Van jullie leerde ik dat de enige echte erwtensoep toch echt varkenstaarten bevat, dat mensen met een kantoorbaantje wat harder mogen werken en mocht ik het mooiste festival van Nederland meemaken (Effe noar Geffe). Bedankt voor al jullie steun in de moeilijke tijden rond mijn moeders ziekte.

Bas, van jou heb ik toch wel geleerd hoe ik moest voetballen en leerde ik nog meer over mijn moeder en jouw moeder. All the best with **Magda** en **Rosa**!

I am happy to have found another side of my family at the other end of the world! Dear **Alan**, **Prue**, **Callum** and **Catherine**. You taught me how to have the best Christmas bbq, how those cricket rules exactly work, and how to ship the most amazing birthday packages across the world! I am looking forward to learn and see more of you and your beautiful country very soon! He aroha whakato, he aroha puta mai.

Kelly en **Joel**, de Hoveniersstraat is niet thuis zonder jullie! Wat fijn om zulke buren te hebben gehad. Jullie hebben mij geleerd dat kanaries wel de allerslimste dieren zijn die bestaan en waar de beste reisbestemmingen op dit moment te vinden zijn. Het allerbeste in jullie nieuwe huis!

Marijn, zonder jou was mijn kledingkast maar half zo groot, bedankt voor al je gulle donaties. Jij leerde mij dat kubben een heel tof spel is en dat ik het woord patent nooit meer mag gebruiken (oeps).

Imre, jij hebt mij geleerd waar je de beste berenklauw met satésaus kan vinden in Nijmegen, hoe je huizen moet verbouwen, hoe je muziek moet maken, hoe de nieuwste blue-ray high definition en de 5.3 surround system toch echt een verschil kan maken. Samen met **Marleen** en **Ezra** hebben jullie een prachtig gezinnetje. Ik denk nog vaak terug aan onze eindeloze gesprekken en grappen in en rondom het voetbalveldje op de Hengstdalseweg. Ik heb enorm veel steun van je gevoeld, toen het slecht ging met mam. Ik voel me altijd thuis als ik bij je langskom.

De **leesclub** heeft mij geleerd hoeveel boeken ik nog niet gelezen heb, en hoe boeken kunnen leiden tot allerlei andere creatieve kunstwerken en hersenspinsels. Dankzij jullie is het leesplezier weer aangewakkerd, na gesust te zijn in de stapel wetenschappelijke literatuur!

Jitka, **Tove** en **Kess**, jullie hebben mij geleerd dat zussen net zo leuk kunnen zijn als een broer. En hoe om te gaan met verliezen van bordspelletjes (en winst i.g.v. Cluedo 😊) en hoe de politie te ontwijken tijdens graffiti-rondje (**Jitka**), hoe om te gaan met de meest fascinerende figuren in de pannekoekenbusiness (**Tove**), hoe kom ik nou mijn scheikundelessen en uiteindelijk PhD door (**Kess**)? **Hans** en **Resi** hebben mij geleerd dat de meest zachtaardige en lieve mensen antropologie studeren.

Peter and **Asli**, you taught me that biology is the real frontier of science and that Sinterklaas actually exists!

Dan de mensen van Duivenrust, die mij adopteerden als een dakloze in Delft. **Luka** en **Anouk**, van jullie heb ik geleerd dat Belgische bierborrels moeilijk te combineren zijn met vluchten naar Nieuw-Zeeland en dankzij jullie eindelijk mijn Koningsdagboottochtje beleefd. Veel succes en plezier met al wat jullie te wachten staat op de korte termijn! **Guus** en **Enza** van jullie heb ik de mooiste wandelroutes in de op-een-na oudste stad van Nederland geleerd, dat Premiere Pro het beste programma is om videos te editen en dat wereldwijde pandemieën toch makkelijk op te lossen zijn als je maar een goed team hebt. Hopelijk zetten we onze missie voort in Nieuw-Zeeland.

Then we move to the wonderful crew of Wageningen UR. **Koen**, you taught me that phasors are the answer to everything and sorry for all the times I had to borrow your campus card. **Alex**, you taught me that Slovenia is one of the most beautiful countries

in the world, what a beautiful trip! **Nico**, you taught me that you should always keep your eyes out for seemingly lost students that might need a hand in the lab, but are too shy to ask. Thank you for helping me out. **Yifan,** taught me the best whiskeys in the world are actually from Japan. **Prarthana**, I have learnt what the best curry in the world should taste like and how to shift your working day to 8 hours later. I am still hoping to learn what an Indian wedding is like one day! **Wen**, I have learnt how to communicate science with awesome pictures and of course your crazy enthusiasm!

Then on to the people of Delft. First of all, the corona cooking club! **Alberto** and **Mireia,** you taught me that when you cut vegetables really small, the sauce becomes amazing and I also learnt that some campsites do not go well with loud game nights! **Fede** and **Costi**, you taught me that panna cotta is the easiest dish ever and thanks for showing us around beautiful le Marche! **Anthony** and **Paola**, you taught me that old bread is the most valuable item in the kitchen and I have learnt the exact ratios required to make the perfect gin and tonic. Thanks guys for those great cooking sessions and also all the midweek drinks/foodies that made me feel part of the whole department.

Nicole, you taught me how to make the most amazing PhD videos and **Mehran** you taught me how to interact with people from BN when you were my first contact in this new world, **Helena** taught me crazy dancing and **Alicia** taught me where the best pancakes can be found in Castricum. **Johannes** taught me how to host great house parties and **Mathia** taught me to how to make the world more joyful and colourful, from **Carsten** I learnt all the behind-the-scenes of life as a teacher and **Lisa** taught me that I wasn't the only one suffering from all these MATLAB tracking algorithms.

Tanja and **Nils**, I am very grateful to have met such like-minded people as me (there are not that many out there!). You taught me how to make the Feuerzangenbowle, turn waste into fertile worm compost and how to grow the best tomatoes (although I have yet to achieve this myself). **Lola** has taught me that bearded men with red hair can be quite scary. But also that things that seem scary at first, can turn out fine later.

The newcomers **Sam**, you taught me discussions about fundamental science are more fun with someone that doesn't agree with you and that there is a fantastic spoonbill colony in Delft, **Lucia** taught me how to sell a house and how awful it is when Argentina loses a soccer match, **Marina** taught me that the Swiss are not as peaceful as they sometimes may seem and that it is important to check the timeslots when you book tickets for Keukenhof ;). Hope to see you soon in NZ! **Jelger** taught me that it is always

good to bring extra beer and chocolate to the lab**.** It's nice to finally see the lab growing again and a new atmosphere emerging. Good luck on continuing the Brouns group!
Then to my students, **Daan**, **Fiona**, **Alexander** and **Jan**, you guys taught me how to work in a team and trust the work other people do. Thank you for all the hard work you put into this, I would have needed 6 more years to finish if it wasn't for you! I wish you guys all the best in your future careers!
Boris, I still like to think of you as my student, even though you have outgrown the master very soon after that. Thanks for teaching me my soccer skills suck and I need to go back and train harder!
Seb, you taught me that Hannover is not at all the most boring city in Germany and also that some jobs in industry can actually be fun (jealous)! Also all the best for **Matilda** and **Viktorija** (who taught me the magic of Lithuanian potato dishes) and thanks for the journey we started together in the Dreijen basement. **Tobal** and **Patri**, you taught me how to make excellent sangria, paint Delft blue pottery and how to countdown in the right way at New Years. Great that we are still in touch. **Teunke**, you taught me how to bake the most amazing cakes (or at least how to eat them). **Anna**, I learnt how it is to work in a well-organized environment because of you, thanks for all the support with the daily issues of ordering, shipping etc and sorry for all the mess I made! **Rita**, you taught me that a PostDoc is able to manage twenty projects at the same time and I learnt how to treat everyone equally and fairly. **Franklin,** you taught me how to pay more attention to the people around you and how to escape from the scariest escape room ever! I am ready to receive my 100 bucks you owe me over our bet. **Benjamin**, you taught me how to make great wiener schnitzel and Knodel, and how to get super muscular only by pushing against a side wall, fascinating! It was great to have you around for a while, I hope you move back in with us one day! **Cristian**, you taught me how to be way better at bouldering than myself. And you taught me some great dance moves (although I think I need a refresher soon!). I'm sorry I won't be here for your defense, but I will party and drink on your behalf from the other side of the world.
Dan de mensen van de Brusselreis, de VSL of de house of the rising sun. Hoewel de naamgeving continu verandert is deze harde kern nog steeds hetzelfde. Als groep heb ik van jullie geleerd dat Belgisch bier beter voor mij is dan shisha en de vele politieke/economische/wetenschappelijke discussies tot laat bij Samson of waar dan ook zal ik nooit vergeten. **Roelant**, van jou heb ik geleerd hoe ik Nederlandse muziek moet

herwaarderen en waarom de Ooij toch echt de beste plek is om Oud en Nieuw te vieren. Ik hoop de komende weken nog wat te leren over groenten en huizen verbouwen in Wageningen Hoog! **Bram**, van jou heb ik vele Vercammen regels geleerd, over wanneer te plassen bij bierdrinken enzovoorts. Daarnaast ook welke 90 hits op welk moment in te zetten op dansfeestjes. Nog bedankt voor de fantastische avonturen die we in Berkeley en Yellowstone beleefd hebben. **Maarten**, jij leerde mij hoe ver het wandelen is als je op de midzomerdag de Veluwe probeert over te steken en dat saucijzenbroodjes goed moeten worden opgewarmd voordat je ze kan eten. **Manu**, van jou leerde ik hoe ik mijn data beter moet encrypten op mijn computer en hoe het is om zonder enig geheugen plotseling te ontwaken in het ziekenhuis (sorry moest het toch even vermelden). Leuk dat we nog even langskomen in der Schweiz! **Wouter** van jou heb ik geleerd dat de jongen die op de eerste dag op school eten op je hoofd smijt later alsnog een van je beste vrienden kan worden en hoe je je voorbereidt op de triatlon. Ik ga jullie missen guys.

Dan de twee dames die al sinds 1F, 18 jaar geleden, een belangrijk onderdeel in mijn leven zijn. **Lotna**, van jou heb ik geleerd hoe je geobsedeerd kan raken door bitterballen, hoe je de beste slaapfeestjes organiseert, hoe om te gaan met diarree op reis, hoe ik minder dik moet zijn (😊). Ben benieuwd waar onze volgend avontuur ons naartoe brengt. En natuurlijk heel veel succes (en een fijne tijd met **Luuk**) zodra je weer terug naar de Limbo's gaat! **Anna**, van jou heb ik geleerd dat slapen in de auto soms de betere optie is, hoe het is om iemand te hebben die er echt altijd voor mij zal zijn, hoe ik me beter moet kleden, wat de beste hits van il Divo zijn, hoe je heel bewust jezelf nieuwe dingen kan leren. Ik herinner me nog de vele gesprekken die we elke week hadden met de loempia (half pikant of zoetzuur) in de hand. Bedankt voor het opzetten van onze weekendtrips, oorspronkelijk om ons 10 jarig jubileum te vieren, een mooie jaarlijke traditie. Ik heb ook geleerd dat Belgen gewoon mensen zijn. **Bart**, vanaf dag 1 dat we elkaar ontmoetten klikte het gelijk. Van jou heb ik geleerd dat er veel gelegenheden zijn om augurken te eten. Het gaat lastig zijn zo ver van jullie te zijn, maar we zagen elkaar eerder wel eens voor een hele tijd niet en onze sterke vriendschap heeft dat altijd weten op te vangen.

Dan naar de belangrijkste twee steunpilaren van vandaag. Niet voor niks zijn jullie allebei de beste pubquizzers die ik ooit heb gezien, zodat als ik het antwoord niet weet, ik er vertrouwen in heb dat jullie dat wel weten. **Patrick**, wij hebben samen dit PhD pad

bewandeld. En hoewel jij de uitgang eerder had weten te vinden dan ik, heb ik nu toch het gevoel dat we dat pad nu samen uitgelopen hebben. Van jou heb ik geleerd hoe je hetzelfde verhaal honderd keer opnieuw kan vertellen en het nooit verveelt, hoe de Haïtiaanse geschiedenis nou ook alweer precies in elkaar zat, hoe ver je kan komen vanuit een MBO opleiding, hoe je zo integer mogelijk met je collega's om moet gaan. De momenten dat er overal gedoe was in het lab en we daarvan even ontsnapten met een kop koffie zijn ontzettend belangrijk geweest. Het gevoel dat ik een rots was voor jou, is dus wederzijds. En natuurlijk waren de paastripjes onvergetelijk. **Lucy,** who taught me all about how to get rich and about very complex , and you are a great match, I am so happy to be at your wedding. And then of course we hope to also be there for the Indonesian edition 😉!

Steef, het is moeilijk na te gaan wat ik niet van jou heb geleerd. Onze levens zijn zo vervlochten geraakt, dat ik soms niet meer weet wat nou mijn eigen ontwikkeling is geweest of wat ik stiekem van jou heb afgekeken. Ik kan in ieder geval de volgende leermomenten met zekerheid aan jou toeschrijven: hoe houd ik mijn adem zo lang mogelijk onder water, hoe versier ik een meisje die ik leuk vind, hoe versier ik meisjes die ik niet leuk vind, hoe maak ik een kampvuur, hoe win ik een onderzoeksprijs op school, hoe kom ik op voor wat ik belangrijk vind, hoe word ik een goede begeleider, hoe zorg ik dat mijn spullen langer meegaan, hoe kan ik mijn angsten in de ogen aankijken. Bedankt voor al die wijsheden, die we in de late uurtjes in bushokjes, onderweg in Engeland of gewoon al doende vergaarden en uitwisselden. Natuurlijk ook bedankt voor al die keren dat je er voor me was. Ik zal mijn hele leven van jou blijven leren. En ook veel geluk met **Ceci**, die mij leerde dat het niet per se een verloren zaak is, mocht ik ooit een van mijn ledematen verliezen en hoe het is om succesvol een enorme grote overstap te maken binnen de wetenschap.

Dan **pap**, dertig jaar hard werken om mij in leven te houden. Elke ochtend nog even zwaaien, als ik op de fiets zat naar school. Elke keer meedenken hoe we probleem X of Y moeten oplossen. Dan zijn deze 300 pagina’s daarvan het eindresultaat. Ik hoop dat ik je hiermee niet teleurgesteld heb. Maar daar ben ik niet zo bang voor, want jij hebt mij altijd het gevoel gegeven dat wat ik ook doe, jij mij daarin steunt en er voor me bent. Van jou heb ik geleerd dat de snelste manier van aardappelschillen is, als je rond de aardappel draait in een spiraal. Ik heb geleerd dat hoe slecht de situatie ook is, er altijd een lichtpuntje is, waar je dan ook flink de nadruk op kan leggen (dit vinden andere

mensen soms vervelend, maar ik blijf het toch maar doen). Ik heb geleerd hoe ik om moet gaan met het afscheid nemen van mam, en hoe een nieuwe liefde weer kan opbloeien na het verlies van een ander. **Corrie**, ook jij hebt bijgedragen aan dit proefschrift, want ik weet niet wat ik had moeten doen als pap nog steeds alleen zou zijn geweest. Jij hebt mij geleerd dat Utrecht toch wel de op-een-na leukste stad van Nederland genoemd kan worden. Bedankt voor jullie warme ontvangst deze weken!

Dan mijn grote broer, **Marnix**, je bent altijd mijn grote voorbeeld geweest. Hoe (ver-)koop ik een huis, hoe trim (vroeger het liefst scheer) ik mijn baard, hoe strik ik mijn stropdas, hoe rijd ik auto (daar valt nog verbetering te boeken), maar ook hoe drink je Bacardi breezers, hoe speel je het beste vals tijdens bordspelletjes, hoe creëer je spanning en sensatie in de James Bond club. Het is nog steeds een mysterie dat je nooit boos werd op je kleine broertje en je er altijd voor hem was (ondanks dat je een keer probeerde van me af te komen door me te laten verdwalen op die camping van Texel ;)). Je bent een lucky man dat je **Nicole** zo gek hebt weten te krijgen een Vink te trouwen. Van haar heb ik geleerd dat kerstmis zeker niet ongevierd moet blijven en hoe handig het is een dokter in de familie te hebben. Jullie hebben een zeer enerverende tijd voor de boeg samen, geniet ervan!

Then on to the last two, most important women in my life. First of all **Becca**. I still don't understand how it can be, that all these years felt so natural with you, every day biking to work together, sitting next to each other in the office, biking back home and spending an insane amount of time together. But it did and it still does. And I could not have spent so much time with anyone else (no offense to everyone above). I am learning every day from you, how to put the cutlery in the dishwasher, how to be honest with people, how to replace a flat tyre in the car, how to keep in touch with people. I wonder if it wasn't for you, whether I would have turned into a hermit, hiding in a dark corner of the lab. You helped me get through the hardest of times, I am forever grateful for that. And I am ready for some new adventures in our life! E kore e ea i te kupu taku aroha mōu.

En als laatste, **mam**, want aan jou draag ik dit proefschrift op. Niet alleen omdat ik je zo erg mis, of omdat ik zoveel van je houd, maar omdat je echt daadwerkelijk deze scriptie samen met mij hebt geschreven. Ten eerste omdat je mijn hele leven lang de weg hiernaartoe geplaveid hebt, door mij het plezier van het lezen te leren, de precisie van het schrijven, dat ik de tijd kreeg om alles over de wereld te leren terwijl jij ons huishouden organiseerde. En sinds je er niet meer bent, wijs je me nog steeds de weg.

Als ik vastzat in mijn project en ik niet meer wist hoe ik verder moest, dacht ik, hoe zou mam dit oplossen. Als ik er helemaal doorheen zat of doodmoe was, dacht ik, zou mam nu moe zijn? En omdat jij altijd zo sterk was, kreeg ik daar vanzelf weer energie van om door te gaan. Nu gaan we door naar een volgende stap en ik draag je weer met me mee. Ik ben je nog elke dag dankbaar dat ik dankzij jou weet hoe ik verder moet gaan.