THE UNIVERSITY
of EDINBURGH

University of Edinburgh

Doctoral Thesis

Quantification of CRISPR-Cas9 Diffusion Dynamics in *Escherichia coli*

Author:
Xavier Zaoui

Supervisor:
Prof. Meriem El Karoui

THE UNIVERSITY of EDINBURGH

A thesis submitted for the degree of Doctor in Philosophy

In the

School of Biological Science

University of Edinburgh

31 August, 2019

To see a world in a grain of sand
And a heaven in a wild flower,
Hold infinity in the palm of your hand,
And eternity in an hour.

(William Blake, Auguries of Innocence, 1803)

Si le temps n’est qu’illusion, les illusions ne durent qu’un temps.

(Khalil Gibran, Le Prophète, 1923)

Lay Summary

In recent years, a novel technology took the world by storm. Stemming from puzzling repeats of DNA sequences found in bacteria thirty years ago, it was discovered to serve as a powerful defence mechanism against the constant threat of bacterial viruses. In this war of genetic information, bacteria integrated viruses' genetic signatures and used them as targets, effectively turning their enemies' weapons against them. The key element was encoded into the association of a target-finder (CRISPR-RNA) and its target-cleaver (Cas9 protein). Together, they formed CRISPR-Cas9.

Twenty years later, its potential for gene editing came to light and transformed every field of biological sciences, thus empowered by its formidable capacity to rewrite genetic material. However, such power is not without risk. An edition at the wrong place and a whole organism could find itself altered, for the worse. Conversely, one could well protect the source of harm, thereby strengthen and not disarm, the intruder. Thus, questions arose about the reliability of this technology. 'How frequently did CRISPR-Cas9 interact with a wrong target?' 'In fact, how, how much and with what did CRISPR-Cas9 interact, overall?'

To address those questions, we decided to use a direct approach and observe the main actor of this biological system; the Cas9 protein. Using a fluorescent marker, we were able to look at its movements within *Escherichia coli* cells and, with this information, detect when it would stop. Indeed, an arrest would indicate an interaction, and an interaction a potential DNA modification. Taking a step further, we compared the amount of interactions displayed by the protein in different conditions: with and without target-finder (gRNA); with and without target; even with and without DNA. Surprisingly, we found that the protein was predominantly moving away from interactions. Yet, our technique allowed us to detect much larger amounts of non-specific associations than previously reported, highlighting the presence of 'silent' DNA interactions and stressing the need for more extensive studies on their potential effect.

In conclusion, my work not only enriches our comprehension of the CRISPR-Cas9 system, but also reveals the existence of strikingly vast non-specific protein-DNA interactions, whose influence remains to be determined. This technology shows huge promise for research and society in general; let this work be a reminder of the need to keep it under control.

Abstract

What do adaptive immunity, genetic engineering and antimicrobials have in common? CRISPR-Cas9, the popular enzymatic complex that produces DNA double-strand breaks when associated with a guide-RNA. Hundreds of labs routinely use this system to edit genomic DNA; however, some of the mechanisms by which it interacts with nucleic acid remain unclear.

In my lab, we developed an expertise in the study of DNA recombination, DNA repair and DNA interactions in *Escherichia coli*. We use single-molecule fluorescent microscopy to collect images in real time, *in vivo*. During my PhD, I harnessed this expertise to follow the behaviour of the Cas9 protein under different conditions: various expression levels; various gRNAs; and various genomic targets. By observing the diffusion dynamics of the protein, I was able to quantify how different DNA interactions were impacting the motion of the protein in the cytoplasm and inferred that actual ON-target interactions were very rare throughout the lifetime of the protein. In contrast, the protein was mainly involved in non-specific OFF-target DNA interactions, in search of its actual target. Additionally, my results reveal the presence of a large fraction of non-specific interactions, hitherto not reported in the literature, owing to their absence of DNA modification.

In total, this work offers a collection of highly quantitative measurements on the behaviour of a protein whose activity is central to many biologists, while shedding a new light on the importance of Cas9 searching and targeting mechanisms. Finally, it opens a discussion on the role of DNA recognition in the context of gene editing and antimicrobial resistance.

Acknowledgements

Meriem, your supervision, your support and your inspiring enthusiasm taught me that science needs both creativity and rigour to deploy the mind. Thank you! *(Thank you also for the constant input of chocolate, third requirement of good science.)*

Teuta, your acute sense of analysis and your unwavering scientific drive taught me to stay on my toes to keep a good balance! Thank you!

Vincent Danos, Francoise Jacob-Dubuisson, Rebecca Turk-MacLeod and Jeremy Guerin, you probably forgot about me but I remember you. Without you, I'd have never made it to the lab.

Sebastian, Alessia, Lorna, Hannah, Irina K., James H., James B., Ira I. We're more than a team, we're good friends! *(aren't we?)* Thank you for those countless hours of the day and the night where a conversation, a good laugh and a shared pizza made up for bad experiments!

Youssef, Flavia G., Adilah, Joseph, Tony, you were more than ~~free labour~~ internship students; you taught me more than I taught you. (*just kidding*) Let me tell you that my cells and I miss you dearly. If you read those lines, there are still experiments to do! *(please send CV and chocolate)*

Jerko, Leonardo, Dario, Alex, Katya, Tom, Frederic, Smitha, Dima, Vanya, Keiran, Pascoe, Benura, Anton K., Elise, Prabu, Anton D., Ekaterina, Andrei, Lyosha, Mai-Britt, Sophie, Jessica, Andrea, Nahuel, Iannis and so many others that I have to stop somewhere, sorry… Er, I meant, thank you all! This PhD is as much mine as it is yours.

And of course, my loved ones. Papy, Mamie, Papa, Maman. Caro! Et les incorrigibles; Alban, Gauthier, Guillaume, Clara, Renan, Charles. Puis un petit clin d'œil à ma deuxième famille. Et Flavia, la belle? Je ne t'ai pas oubliée!

And finally, my thanks to the Hauts-de-France, to Scotland and to the Darwin Trust. To them, Science knows no boundary; neither does my gratitude.

Oh! and thanks to you, for whichever reason you ended up on those pages! Your attention is like sunshine in a Scottish sky; deeply appreciated.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Xavier Karim Zaoui

Table of Contents

Chapter I – Introduction

Prelude

This work revolves around CRISPR-Cas9, the technology originally designed by bacteria and re-programmed by mankind to target DNA, cleave it and use this open breach to rewrite any type of genetic material.

In the following chapter (Chapter I), I will describe the system as it is used in bacteria, discussing its structures and mechanisms. Then I will highlight the various courses scientists have taken to re-direct this system, either for genetic editing purposes or for antimicrobial ones. The limitations of such endeavours will be discussed, and particularly the discovery of OFF-target events, where the CRISPR system non-specifically interacts with DNA. On this basis, I will present the purpose of this thesis, which will bring us to the study itself; the quantification of CRISPR-Cas9 diffusion dynamics in *Escherichia coli*.

After presenting the Materials and Methods used in this study (Chapter II), I will show how those methods were implemented, tested and employed to give a first overview of CRISPR-Cas9 diffusion (Chapter III). This will be followed by a more comprehensive analysis, based on the multiplication of experimental conditions, allowing us to compare, infer and draw our own interpretations (Chapter IV). Finally, our findings will be more widely discussed in the context of the literature, leading us to our final conclusions (Chapter V).

May this work be clear, honest and as enjoyable to read as it was to write.

1.1 – CRISPR Origin: Structure and Mechanism

To protect themselves against external DNA, every form of life has developed immunity mechanisms, such as RNA-guided systems which purposely target viruses[1]. In bacteria, such adaptive systems have been reported to provide resistance to bacteriophages[2] and even to prevent plasmid conjugation[3]. Indeed, about 40% of bacteria (and 90% of archaea) possess similar mechanisms[2] which in 2002 were officially called **CRISPR**, for **C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats[4]. Many studies soon followed, revealing that relying on endonucleases associated to RNA-DNA homologies, those systems work in three phases.

First, small fragments of external DNA called 'spacers' are captured and incorporated into long genomic sequences called 'CRISPR arrays', where they get interspersed with (up to several hundreds of) DNA repeats[5] (Figure 1.1). Those arrays are, in a second phase, transcribed into long RNA molecules and processed into smaller transcripts ('crRNAs' or CRISPR-RNAs) consisting of single-spacer units, where they associate with Cas complexes; those complexes are formed of multiple subunits, usually transcribed in the vicinity of the CRISPR arrays, which have the ability to bind crRNAs to form Cas-crRNA complexes[6]. Third and final phase, those Cas-crRNA complexes recognise foreign DNA and degrade it through various endonuclease and exonuclease activities[7].

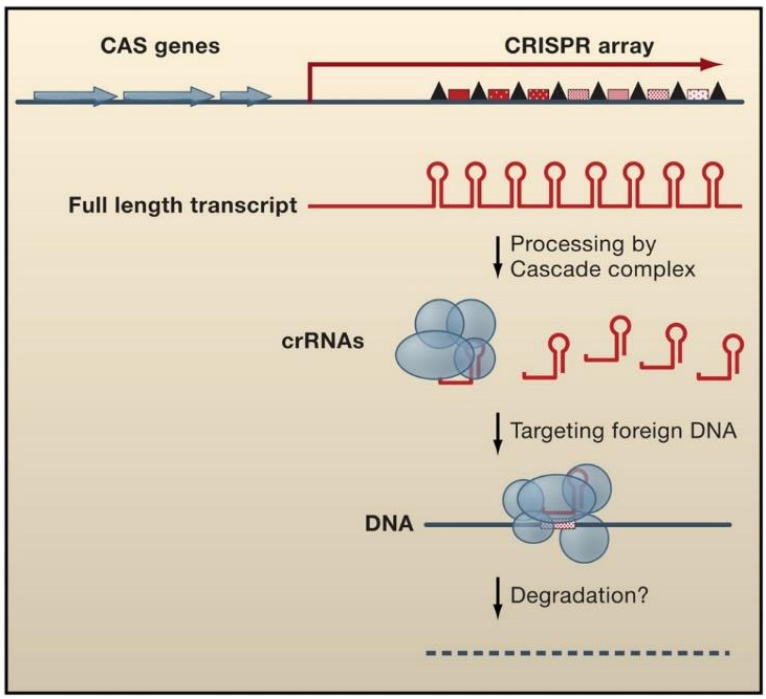

Figure 1.1. CRISPR general mechanism. The CRISPR system is composed of two main parts; the Cas genes (or effectors) and the CRISPR array with repeats (black triangles) and unique spacers (red boxes), resulting in a long transcript which will be further processed by the Cas proteins (or Cascade complex). Together, they will associate and target homologous sequences of external DNA which upon recognition will trigger degradation. Initially, the whole target and subsequent degradation mechanisms were still unknown. Extracted from Waters, et al. Regulatory RNAs in Bacteria. *Cell* **136**, 615–628 (2009)[8].

Many different CRISPR types have been discovered, characterized by their Cas proteins (with more than 40 different genes)[9]. Those proteins present similarities with DNA-binding proteins including endonucleases (*cas9*), exonucleases (*cas4*), helicases (*cas3*) and even polymerases[10,11].

Another feature distinguishing CRISPR types from each other are PAM – or **P**roto **A**djacent **M**otif – sequences, which beautifully ensure that CRISPR-containing cells do not damage their own DNA (Figure 1.2)[12]. Indeed, the first motifs recognised by CRISPR-Cas complexes are those PAM sequences (usually three-nucleotides long), subsequently facilitating direct Watson-Crick pairing between the crRNA and the targeted sequence. CRISPR arrays do not possess such motifs, preventing Cas complexes from cutting their very origin[13].

Due to its simplicity, ease to manipulate and DNA-processing efficiency, one particular type of the CRISPR family attracted most of the attention; the CRISPR type II, also called **CRISPR-Cas9**. Originating from *Streptococcus pyogenes*, this CRISPR complex relies on three agents (Figure 1.2);

(i) The Cas9 protein which possesses two endonuclease domains (HNH and RuvC, a Holliday junction resolvase),

(ii) the crRNA which encodes the specificity of the CRISPR activity, with 20-nucleotide homology with the target sequence,

(iii) the tracrRNA (*trans*-activating crRNA), itself associated to the crRNA and forming, with it, a dual RNA[14].

Effectively acting as a DNA 'handle', the tracrRNA bridges Cas9 (effector protein) and the crRNA (target identifier) together[15]. Much like in other CRISPR families, the type II CRISPR array originates from the insertion of external DNA 'spacers' interspersed by DNA repeats which, after transcription in long RNA molecules, gets processed by the RNAseIII protein to eventually result in several individual Cas9-crRNA-tracrRNA (or CRISPR-Cas9) complexes, ready to 'search and destroy' foreign DNA.

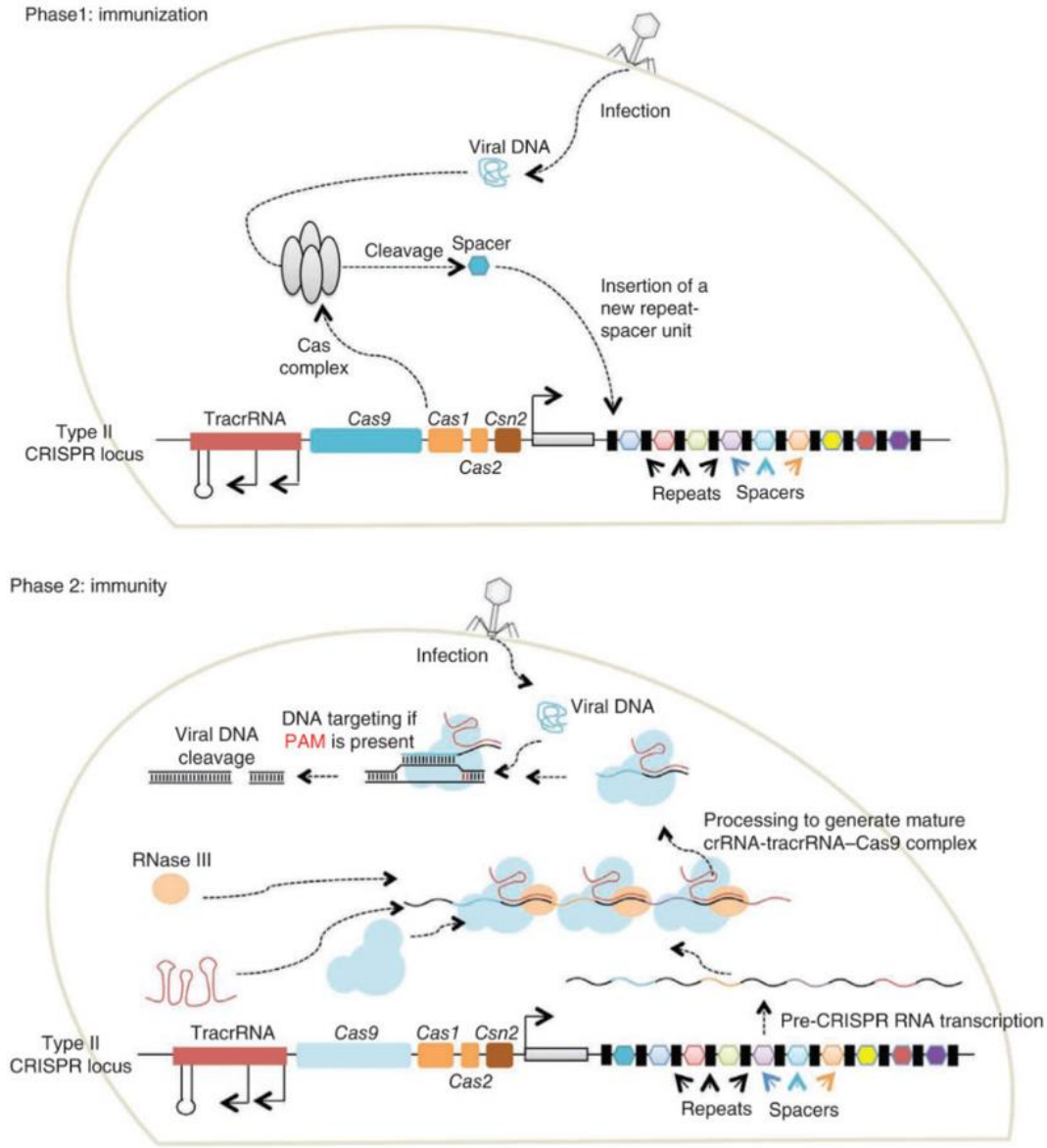

Figure 1.2. CRISPR-Cas9 immunization and immunity. (**Top**) Upon infection, the external (or viral) DNA gets degraded by the Cas complex which results in small sequences (coloured hexagons) inserted in the CRISPR array and interspersed with repeats (black rectangles). (**Bottom**) After downstream transcription, those long crRNAs associate with RNAseIII, tracrRNAs and the Cas9 protein to form individual CRISPR-Cas9 complexes. After PAM sequence recognition, such a complex can unwind the DNA to allow the crRNA to probe for homologies which, if found, triggers Cas9 endonuclease activity. Extracted from Mali, et al. Cas9 as a versatile tool for engineering biology. *Nat Methods* **10**, 957–963 (2013)[16].

In more details, the Cas9 protein has a molecular weight of 160kDa and is itself composed of two main lobes, themselves formed of several domains; the REC lobe (Helicase Recognition lobe) is constituted of three alpha-helical domains whose roles are to associate with the tracrRNA. The NUC lobe (nuclease lobe) contains the two endonuclease domains (RuvC and HNH), bound to each other by two linkers, and a CTD (C-terminal) domain which is involved in PAM sequence recognition. Finally, an arginine-rich domain links the two lobes together (Figure 1.3A, B).

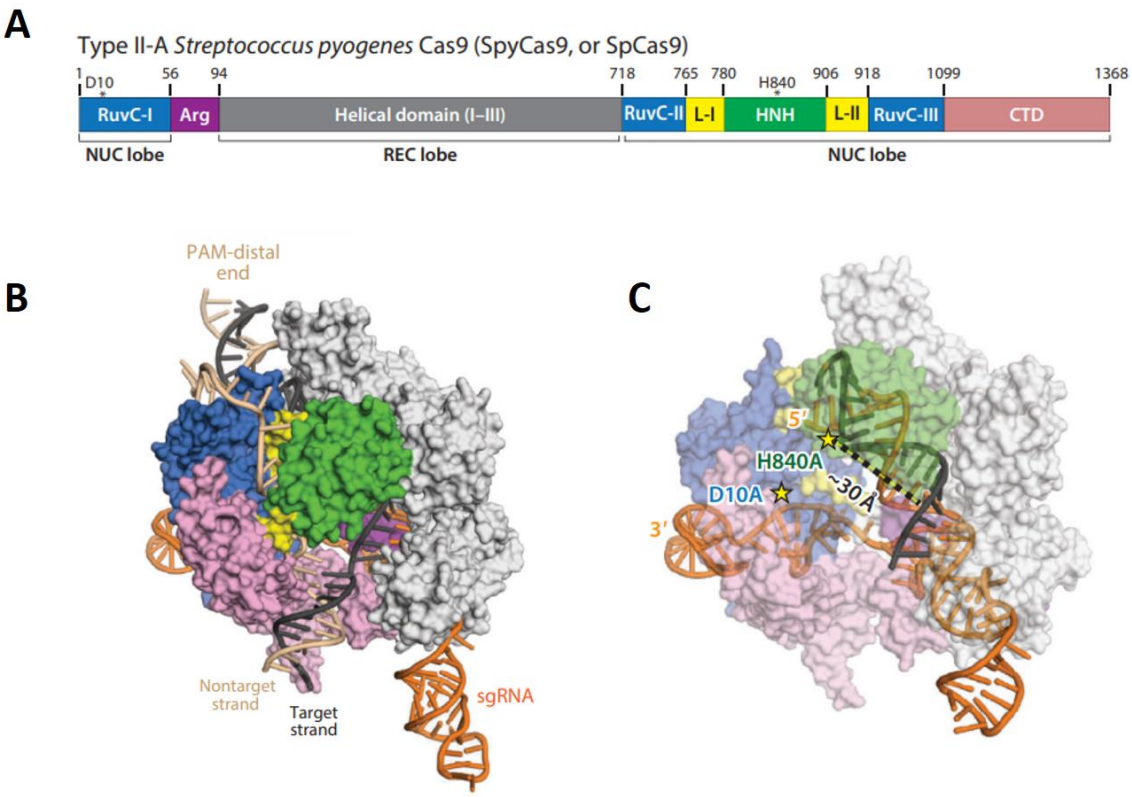

Figure 1.3. Cas9 domain organization and structure. (**A**) Cas9 domains organization. The two endonuclease domains RuvC-I (divided in three parts, in blue) and HNH (green) are associated with two linkers (L-I and L-II, both in yellow). The Helical domains (REC lobe, in grey) allow Cas9 to interact with the sgRNA (association of crRNA and tracrRNA) (orange). The arginine-rich bridge helix (Arg, in purple) is involved in protein conformational change to trigger DNA target (black) cleavage. Finally, C-terminal domain (CTD) relates to PAM recognition. (**B**) With the same previous colour code, this cartoon representation of Cas9 shows how every domain interacts with each other, with the sgRNA and with DNA. (**C**) With added transparency, it is possible to see the two endonuclease sites (yellow stars), and the too-large distance between the HNH active site and its target locus (dashed line), highlighting the necessary change of conformation to be effective. Adapted from Jiang, F. & Doudna, J. A. CRISPR–Cas9 Structures and Mechanisms. *Annu. Rev. Biophys.* **46**, 505–529 (2017).[17]

The dual-RNA formed by the association of the crRNA and the tracrRNA, commonly referred to as the guide-RNA (gRNA), strongly influences the conformation of the Cas9 protein. Indeed, observations of the crystallographic structure of the CRISPR-Cas9 complex showed that several conformational changes were necessary to (i) allow Cas9 to interact and recognise PAM sequences and (ii) to allow the restrictions sites to cleave the target site (Figure 1.3C).

In the apo form (Cas9 without gRNA), the CTD domain (responsible for PAM recognition) is kept partly disorganized and therefore limits Cas9 ability to recognize PAM sequences. As can be seen on Figure 1.4, Cas9 interactions with PAM sequences are facilitated by the association of the protein with the gRNA, as its two recognition lobes (REC and NUC) open up to transition the protein domains from a largely disorganised state to a pre-organized one. After PAM recognition and through additional conformational change, the CRISPR-Cas9 complex unwinds both DNA strands and allow the crRNA to probe the DNA sequence by checking for homology. The specificity of the crRNA rests on its 20-nucleotide sequence, homologous to the target to find. However, a 'seed region' consisting of 10-12 nucleotides towards the PAM sequence (on the 3' end of the spacer) contains most of the target specificity, as mismatches in this region dramatically decrease CRISPR-Cas9 activity[18].

After full recognition of the target sequence, the Cas9 protein uses its two endonuclease domains to produce a double-stranded break (DSB), blunt-ended, 3 nucleotides upstream of the 3′ end of the spacer.

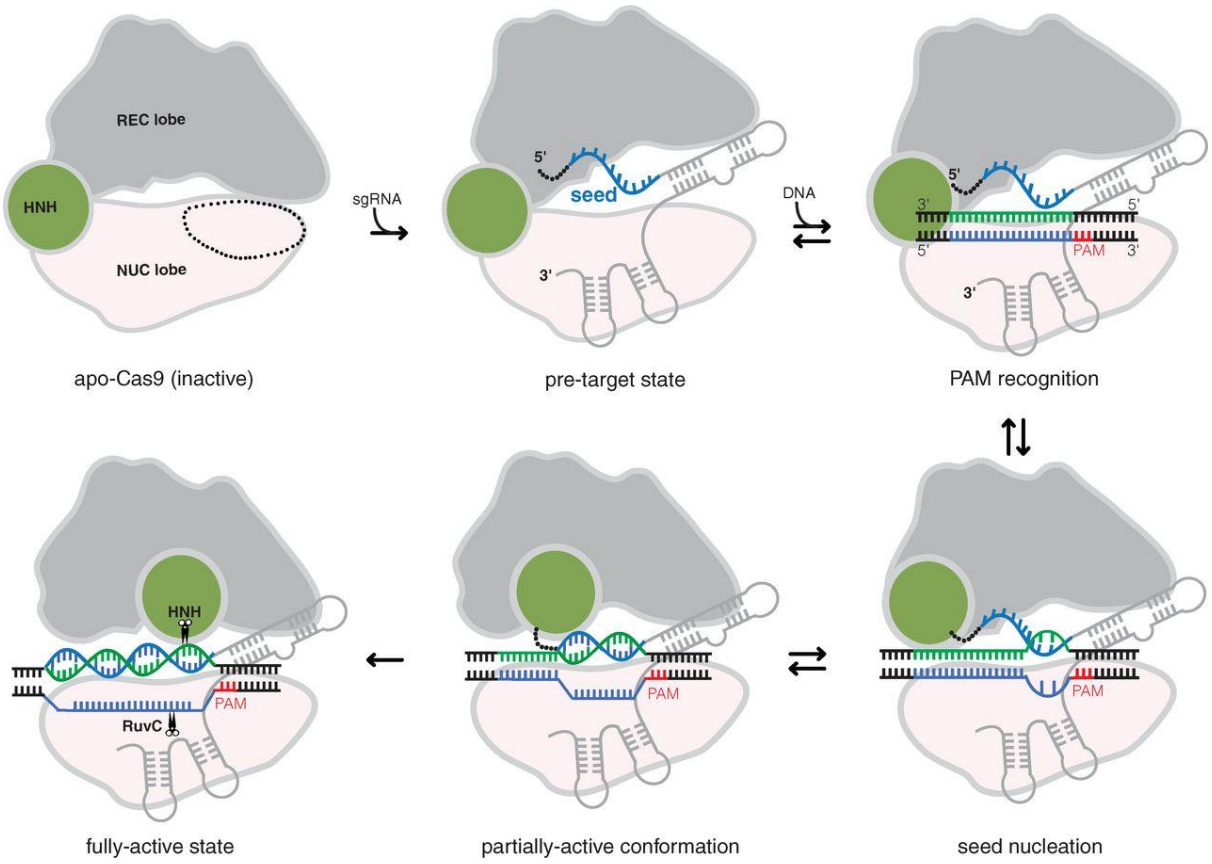

Figure 1.4. CRISPR mechanism for DNA-interaction and cleavage. The Cas9 protein is composed of two main lobes; the REC (Helicase Recognition) one which associates with the gRNA and the NUC (Nuclease) one which activates the endonuclease activity. Upon association with the sgRNA (crRNA-tracrRNA), the PAM-interacting region of the Cas9 protein (dotted circle) opens up to allow for DNA recognition. Through its seed region (blue), the gRNA then probes the DNA sequence for homology (seed nucleation followed by near-active form). After full recognition, the endonuclease domains HNH and RuvC cleave the two strands of target DNA. Extracted from Jiang, F., Zhou, K., Ma, L., Gressel, S. & Doudna, J. A. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science (80-.).* **348**, 1477–1481 (2015)[19].

1.2 – CRISPR Optimization: Genetic Engineering

After a first characterization of the CRISPR complex, a set of additional features artificially designed to further simplify the use of CRISPR-Cas9 and, above all, to optimize its genetic engineering capabilities, were developed.

For example, a chimeric single-guide RNA (sgRNA, or gRNA), made of the fusion of the crRNA (in 3') and of the tracrRNA (in 5') was produced; not only easier to build, this construct also presented the advantage of accelerating the formation of the CRISPR complex, due to its combined expression[15] (Figure 1.5).

Additionally, it was shown that elongating the tail of the tracrRNA (with up to 85 nucleotides) could increase the CRISPR efficiency (up to five times), due to higher RNA stability and stronger Cas9 affinity, thus leading to a new trend of gRNA designs[20].

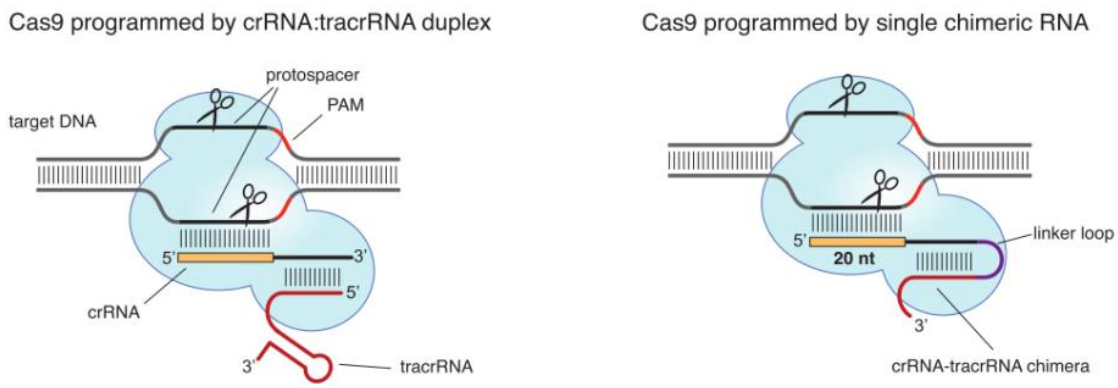

Figure 1.5. Design of the single guide RNA, or sgRNA. Built from the fusion of the crRNA (in 3') and of the tracrRNA (in 5'), associated to a linker loop, the sgRNA keeps the same features and the same affinity for both Cas9 and the DNA target, making it effectively easier and thus more efficient to use. Extracted from Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science (80-.).* **337**, 816–822 (2012)[15].

After CRISPR-Cas9 recognition of the PAM sequence, followed by complete homology with the crRNA, DNA cleavage is triggered and, as we previously mentioned, a DNA Double-Stranded Break (DSB) is produced. In eukaryotes, this break can be faithfully repaired through Homology-Directed Repair (HDR)[21] of which the most common form is Homologous Recombination (HR)[22]. Additionally, it can be repaired by a much less accurate mechanism called Non-Homologous End Joining (NHEJ)[23]. Left unrepaired, such DSB can be lethal.

In *Escherichia coli*, the vast majority of DSBs are repaired through Homologous Recombination, which is triggered by the RecBCD enzyme (homolog of the MRX complex in budding yeast and of the MRN complex in mammals[24]) (Figure 1.6); it recognises DSB ends and processes the two DNA strands until it encounters a Chi site (8-nucleotide sequence), at which point the complex changes its conformation and limits its DNA digestion to a single strand, leaving the other available for the recruitment of a second protein; RecA (homolog of Rad51)[25]. Several RecA proteins then accumulate on the DNA strand and form a nucleoprotein filament, which will drive the homology search for an identical copy of the broken strand, able to act as a template for subsequent repair.

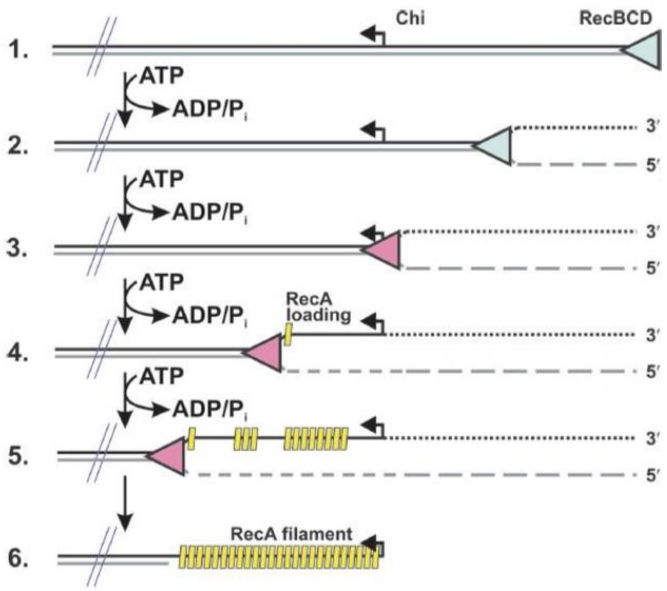

Figure 1.6. Homologous Recombination in *Escherichia coli*. After DNA cleavage, (**1** and **2**) the RecBCD enzyme (triangle) recognises the DNA end and digests both DNA strands (at different speeds), consuming ATP. (**3** and **4**) Upon encountering a Chi site (black arrow), RecBCD changes its conformation and stops degrading one of the strands, leaving the other available for RecA recruitment. (**5** and **6**) A RecA nucleoprotein filament is progressively formed, which will trigger the homology search for an intact copy of the broken DNA. The copy will be used as a repair template to faithfully reconstitute the broken sequence. Extracted from Dillingham, M. S. & Kowalczykowski, S. C. RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol. Mol. Biol. Rev.* **72**, 642–71 (2008).

Importantly, RecA recruitment triggers an additional pathway called the 'SOS response'[26], which sees the overexpression of more than 40 genes in order to support repair mechanisms[27]. In normal conditions, a regulator called LexA specifically binds to those genes (on a SOS box[28]) and represses their expression. Upon DNA damage, the presence of the previously-described RecA nucleoprotein filament interacts with LexA proteins and triggers its auto-catalysis, effectively de-repressing the SOS genes within minutes[29].

Among those genes is the one expressing RecA, thus indirectly over-expressing itself, and many others including SfiA, a division inhibitor, as well as the mutagenic DNA repair polymerase (Pol V) which, as a result, inhibits cell division and facilitates repair while increasing genomic diversity[25]. Finally, as LexA is also repressing itself, the SOS response ensures that its main regulator remains expressed, so that when the damage is repaired, the signal disappears and all those genes go back to normal expression[30].

In contrast, the NHEJ pathway (absent in *E. coli*) uses the Ku protein to allow the recruitment of ligases, which will re-ligate the two DNA ends together[31]. This faster mechanism has the main disadvantage of being much more error-prone which can lead to 'indels' (insertion/deletion), themselves potentially deleterious due to gene loss of function (Figure 1.7).

Now, in the context of CRISPR-Cas9, both repair mechanisms can be harnessed for specific purposes. For instance, with the addition of an appropriate DNA template, HDR can lead to the effective rewriting of virtually any genomic DNA sequence[32] (Figure 1.7). In contrast, NHEJ capacity to produce indels has been used to detect and measure Cas9 activity[20].

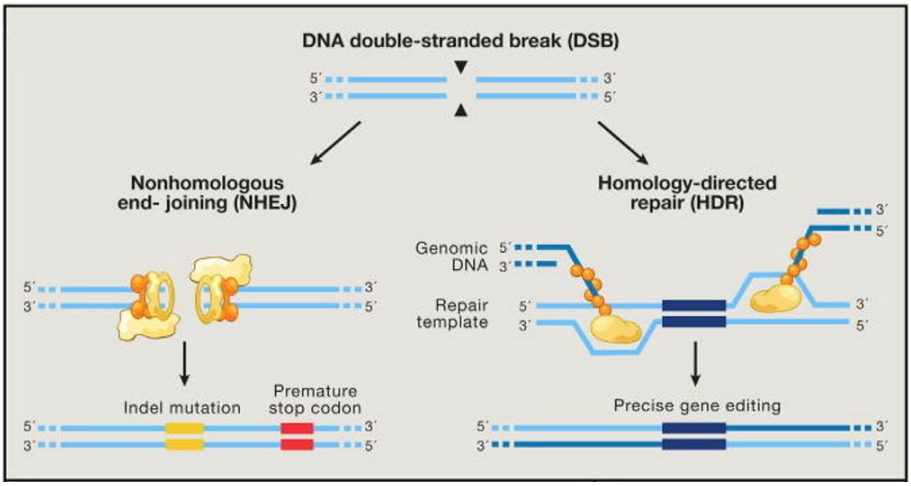

Figure 1.7. Double-Stranded break repair and gene editing. Following Cas9 cleavage, the DNA break is either repaired through non-homologous end joining (NHEJ), or through homology-directed repair (HDR). In the first scenario (NHEJ), the Ku protein triggers an error-prone ligation, which can lead to indel mutations, misalignment (frameshift mutations) and subsequent gene knock-out (represented here by the premature stop codon). In the other case (HDR), Rad51 (in eukaryotes) or RecA (in *E.coli*) uses a DNA template to faithfully repair the break. Consequently, a DNA repair template containing mutations (represented here by the dark rectangles) can be externally added to accurately edit any gene. Extracted from Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).

Note that one (or both) of the endonuclease domains can be voluntarily inactivated (by single-point mutations) so as to prevent Cas9 from cutting the DNA[33,34]. Such catalytically-dead mutant, dubbed 'dead-Cas9' (dCas9) actually offers renewed opportunities such as gene regulation possibilities, illustrated by the CRISPR-interference (CRISPRi) mechanism, where dCas9 binds to promoter regions and prevents RNA polymerases from transcribing neighbouring genes, effectively inhibiting their expression[35] (Figure 1.8).

A similar approach used a dCas9 fusion to the omega subunit of the RNA polymerase (RNAP) to enhance transcription up to 23-fold, by stabilizing the binding of RNAP to the promoter of the nearby dCas9-bound target[36].

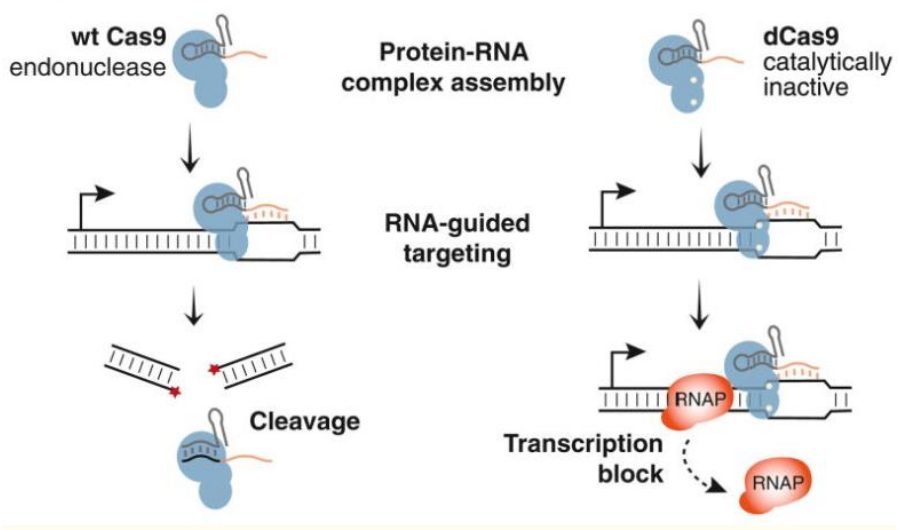

Figure 1.8 CRISPR-interference (CRISPRi) uses dCas9 for gene regulation. As opposed to the original wild-type Cas9 (on the left), dCas9 does not cleave the DNA (on the right). In consequence, it is possible to harness dCas9's binding ability to prevent the RNA polymerase (RNAP) from transcribing any neighbouring gene, effectively interfering with the gene regulation. Extracted from Qi, L. S. *et al.* Repurposing CRISPR as an RNA-γuided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).

Finally, another notable function developed with CRISPR-Cas9 is based on its antimicrobial potential. Indeed, by targeting bacterial genome which, as we previously mentioned, extensively (if not exclusively, for some organisms) relies on Homologous Recombination to repair DNA breaks, it suffices to cleave all available repair templates (which in bacteria are rarely exceeding 10 copies), to irremediably kill the targeted organism[37]. Better still, the specificity embedded in the 20-nucleotides sequence of the crRNA would ensure to only hit designated targets (Figure 1.9). In the context of antimicrobial resistance, this technique would allow the directed targeting of resistant bacteria, leaving the rest to either outcompete the few survivors, or to be treated by usual antibiotic treatment.

Considering that this mechanism was originally designed by bacteria to turn viruses' genetic tools against them, let us appreciate the irony of having it now turned back against bacteria.

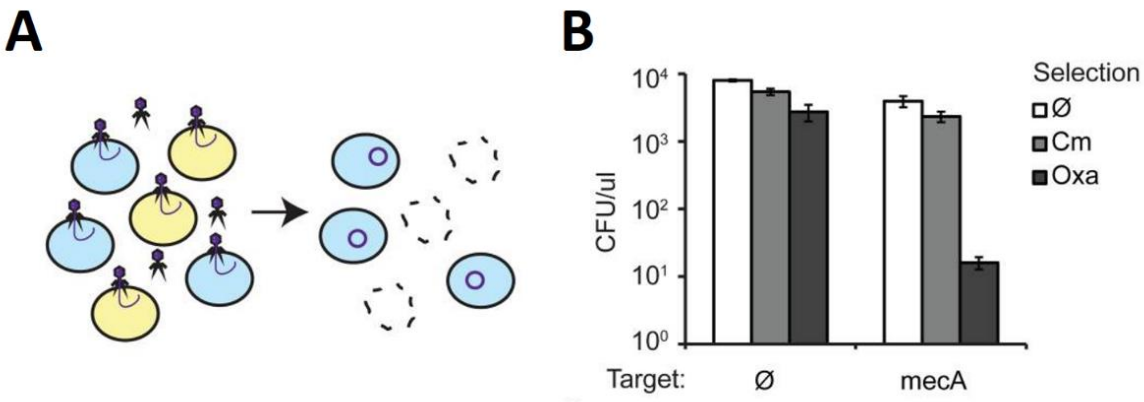

Figure 1.9 CRISPR-Cas9 can be used to selectively kill resistant bacteria. (**A**) Two strains of *Staphilococcus aureus* cells were mixed 1:1 (one being both methicillin and oxacillin resistant) and transformed with a phagemid containing CRISPR-Cas9 targeting the methicillin resistance gene *mecA*. As a result, only the methicillin-resistant cells were killed, while the others were immunized to conjugative plasmids carrying the resistance gene. (**B**) Following transformation with phagemids, cells were either plated on a non-selective medium (ø), either on a medium containing chloramphenicol (Cm) to measure the number of cells carrying the phagemids (chloramphenicol resistant, as a control) or on a medium containing oxacillin (Oxa) to measure the number of resistant cells. In the presence of a spacer targeting the *mecA* gene, resistant cells die about 100-fold, compared to almost no effect on the population of non-resistant bacteria. Measures are in colony forming units (CFU) per microliter, with their respective standard deviations (error bars). Adapted from Bikard, D. *et al.* Development of sequence-specific antimicrobials based on programmable CRISPR-Cas nucleases. *Nat. Biotechnol.* **32**, 1146–1150 (2014)[37].

In conclusion, from this time on, a simple RNA and a protein expressed together could easily recognise, cut and therefore edit any DNA sequence, effectively transforming (or killing) any organism of choice (Figure 1.10). Despite many remaining uncertainties, this discovery was a little revolution[15]…which unleashed a legal battle for its paternity[38].

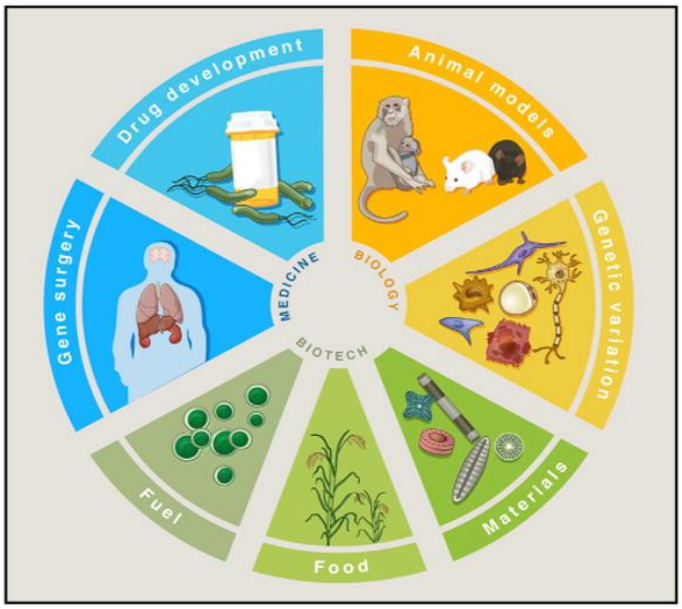

Figure 1.10. A CRISPR revolution. CRISPR-Cas9 genome editing abilities can be harnessed to revolutionize various fields. In Molecular Biology, it can greatly facilitate the use of animal models and the development of different cell lines. In Biotechnology, it can increase fuel and food production. In Medicine, it can open the door to new personalized therapeutics, through gene therapy and new antimicrobials. Extracted from Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).

1.3 – CRISPR Limitations: ON and OFF-target

After validation of the CRISPR-Cas9 potential, a new set of analyses started to highlight limitations in the way this system could be relied upon. For instance, Hsu and colleagues showed that the 'seed' region of the crRNA (spanning 10-12nt) could tolerate mismatches, thus revealing a complex recognition pattern[20] (Figure 1.11).

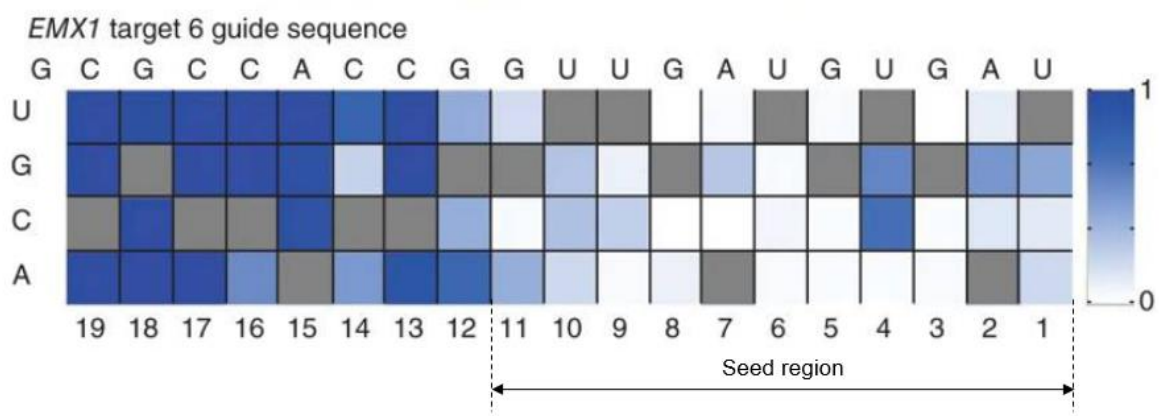

Figure 1.11. gRNA single-mismatches highlight 'seed region' and reveal base-dependent specificity. A set of gRNAs all targeting the human *EMX1* gene and harbouring different single mismatches at different positions from the PAM sequence (starting at number 1) had their respective CRISPR efficiency measured, and normalized to the original target sequence. The heat map represents the CRISPR efficiency (increasing from white to dark blue), characterized by the presence of indel events originating from Cas9 cleavage and subsequent non-homologous end joining (NHEJ). We can observe that within the 'seed' region corresponding to the first 10-12 nucleotides, the CRISPR efficiency really varies from one base to another. Adapted from Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013)[20].

These results shed new light on the interactions of the CRISPR complex with genomic DNA, and especially the secondary structures found at the target site. Indeed, such structures could sometimes improve, and other times entirely remove CRISPR-Cas9's ability to cut the DNA, regardless of the accuracy of the match between the crRNA and its target[33]. Not to mention the additional impact of chromatin types (especially heterochromatin)[39], methylation and other epigenetic DNA modifications[40].

More importantly – and more surprisingly – was the appearance of non-specific activity, where the CRISPR complex was observed interacting with DNA sequences which did not fully match its crRNA[41,42]. By targeting a single sequence with several crRNAs harbouring mismatches with the target, then by measuring gene expression which should be inhibited in the case of effective DSB, studies[41] have noticed the phenomenon, but are yet to understand its cause or even predict its occurrence. Indeed, as opposed to **ON-target** activity where the CRISPR complex recognises the sequence matching its crRNA (*i.e.* its rightful target), **OFF-target** activity relies almost unpredictable on secondary structures which mimic the target sequence and can, potentially, also result in DNA double-stranded breaks[43]. In the context of genetic engineering, those particular effects are non-specific and therefore unwanted, for a reliable system not only requires efficiency, but also accuracy.

As an example, Hsu and colleagues[20] have shown that a single gRNA not only cleaved its target but also other loci across the genome (which happened to be human), going as far as recognising "wrong" PAM sequences (NAG instead of the usual NGG) (Figure 1.12), rendering the whole cleavage prediction process highly difficult and, worse, threatening any attempt at editing human DNA for therapeutic purposes[44]. Indeed, cutting DNA at the wrong locus would open the door to all sorts of unexpected shortcomings; not the least of which, cancer.

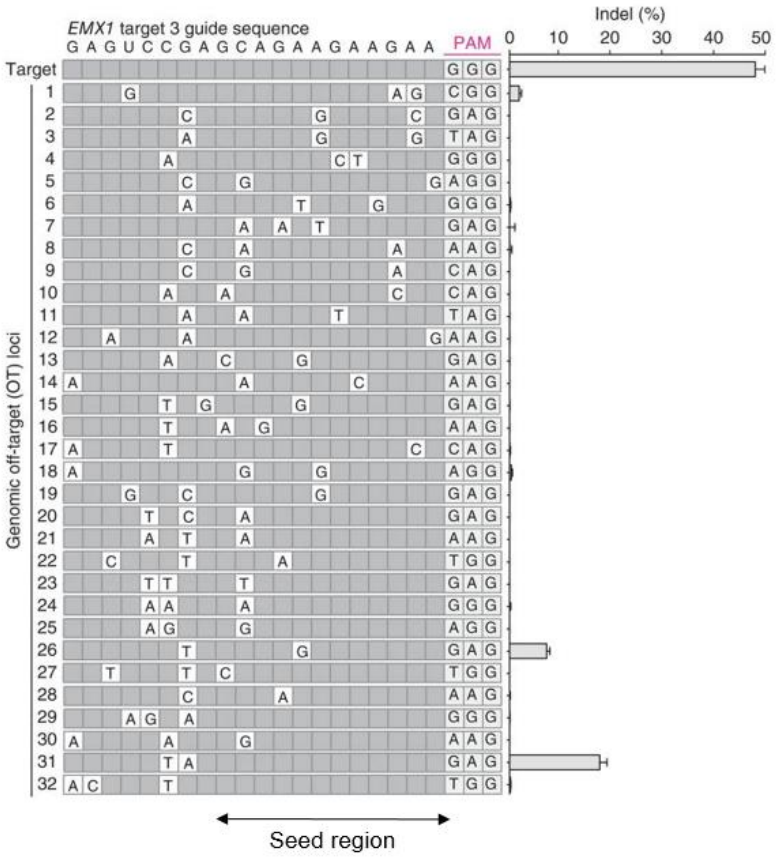

Figure 1.12. OFF-target events are difficult to predict. The activity of the CRISPR-Cas9 complex associated to a gRNA targeting the human *EMX1* gene was measured and compared to its potential activity at 32 different sites harbouring sequence similarities but nucleotide mismatches, either with the target or with the PAM region. The activity was assessed as the frequency of indel events, produced by Cas9 cleavage and subsequent non-homologous end joining. We can observe that, surprisingly, the two putative targets displaying the highest activity (#26 and #31) also contain different PAM sequences (NAG, instead of NGG). In contrast, other sequences with the right PAM and the same nucleotides in the seed region (#29, #32) do not display any activity, making the whole prediction of OFF-target events rather difficult. Adapted from Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).

To adress this challenge, several directions were explored[45], aiming to increase the specificity of the CRISPR recognition mechanisms (*e.g.* so-called 'high-fidelity Cas9' [46]) or by adjusting gRNA sequences [47,48]. Another saw the discovery of a protein preventing the CRISPR complex from binding DNA, so-called 'anti-CRISPR' protein[49]. Despite an elegant perspective, this technique would add the necessity of accurately controling both the 'activation' and 'deactivation' of the CRISPR complex and the complexity of time dynamics.

Another notable idea was double nickases (or 'paired nickases'), as opposed to a single nuclease. Using two Cas9 mutants (Cas9-D10A, with an inactivated RuvC endonuclease and only the HNH domain active), Cho and colleagues[47] reasoned that they could decrease the probability of having non-specific cleavage (due to the necessity of having two proteins present at the site during the same period of time). And indeed, with a single crRNA-target mismatch, not only did they observe as much as a 500-fold decrease in indel frequency (compared to the one without mismatch), but they also managed to keep an efficient ON-target activity.

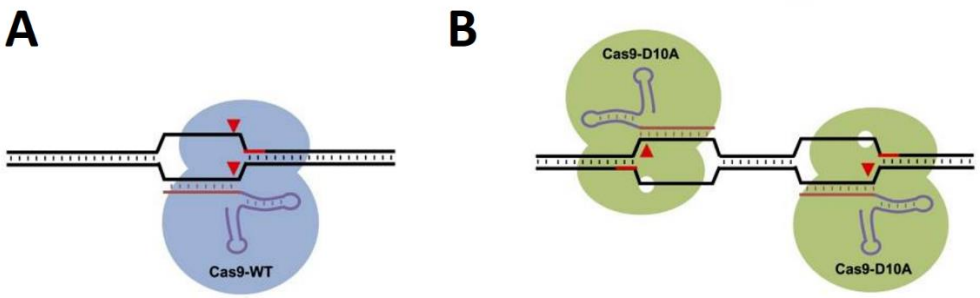

Figure 1.13. Paired nickases can improve Cas9 specificity. (A) In the presence of a gRNA, Cas9 wild-type (WT) uses its two endonuclease domains to produce a DNA double-stranded break. (B) Two Cas9 nickases (Cas9 mutants possessing only one endonuclease domain left, also written Cas9-D10A) can cleave one strand of the DNA at two close positions to effectively trigger Homology Directed Repair mechanisms and subsequent directed editing, without comprising on CRISPR-Cas9 specificity. Adapted from Cho, S. W. *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* **24**, 132–141 (2014)[47].

That said, in order to keep a high level of ON-target events, this system requires a high expression of Cas9 nickases, which also presents its own limitations (in terms of inducible expression system, variability in copy number and, again, increased chances to produce OFF-target interactions).

Facing the necessity of predicting the level of OFF-target events across the genome, teams have developed techniques relying on genomic detection of indels, such as ChIP-seq experiments[50] and other deep sequencing protocols[51,52]. Based on those methods, evidence suggests that significantly low levels of OFF-target cleavage were observed with mismatches superior to one (10^5-time less than ON-target events in human cell lines)[47]. Nevertheless, single-mismatches could still yield levels of OFF-targets 10-time lower than their ON-target counterpart (which is very high)[47].

Moreover, a major hurdle pertains to the limited sensitivity of such detection techniques, which cannot pretend to detect extremely rare events (<0.12%)[51] of the kind produced by an OFF-target CRISPR-Cas9, where a single event could be significant.

Finally, for the detection of OFF-target events, those techniques all rely on actual DNA mutations, thereby entirely ignoring the fraction of CRISPR-DNA interactions which does not result in DNA cleavage. This 'silent' fraction is likely to be vastly larger than what was quantified so far, and could emcompass several other behaviours, such as Cas9 probing non-specific sequences and staying immobile, temporarily blocked on the DNA.

It is indeed reasonable to believe that with near-complete target-gRNA homology, Cas9 would have unwound the two strands of DNA[53] and engaged in a conformational change which, if not brought to completion, may trap the protein between two energetically-unfavorable stages (DNA-cleavage and DNA-release)[54]. Such eventuality may stand as an obstacle to the replication and transcription machineries, thus creating subsequent gene deregulation or replication fork colapse, leading to more unwanted consequences.

An alternative to studying CRISPR OFF-target interactions is to directly observe the protein's movement within the cell, and from its motion infer its interactions with DNA. Indeed, such interactions would slow the protein down, as shown by Elf and colleagues[55] with the example of the LacI repressor. Similarly, as CRISPR-Cas9 specific interactions (ON-target events) would last longer, due to Cas9 breaking DNA, this technique should differentiate between ON and OFF-target events. Accordingly, various levels of protein diffusions would relate to various protein behaviours, some of which may have yet to be determined.

The only way to reach such a detection level is through single-molecule microscopy, where every protein can be illuminated, observed and eventually tracked in time and space.

1.4 – **Protein diffusion: current techniques** and Cas9 past studies

First introduced *in vitro*, the development of single-molecule microscopy techniques was based on a collection of Nobel Prize discoveries (Green Fluorescence Protein[56,57], awarded in 2008; super-resolution microscopy[58,59], awarded in 2014; optical tweezers[60,61], awarded in 2018). It opened the door to many previously-unanswered questions, such as the dynamics of molecular motors[62], of molecular rotors[63] and of DNA replication[64].

However, despite their high temporal and spatial resolution coupled to exquisite experimental controls, those works failed to reproduce the complexity and the unexpected of actual living systems. Indeed, living organisms produce myriads of molecules in diverse quantities which constantly interact with each other, changing their environment and affecting their fate in ways that are impossible to entirely detect, nor predict.

That is why efforts were made to push back the boundaries of both biology and microscopy, resulting in three main complementary approaches.

- The search for bright, photostable and fast-maturing fluorescent proteins, leading to the discovery of the Yellow Fluorescent Protein[65] (YFP), the mCherry protein[66] and the more recent HaloTag protein[67].
- The development of microscopy techniques accommodating the required low number of particles to observe (Photo-activated Localization Microscopy or PALM[59]) and the limited illumination volume necessary to decrease the fluorescent background (Total Internal Reflection Fluorescence or TIRF microscopy[68]).
- The design of fluorescent spot detection algorithms, which could not only accurately localize a protein but also track its trajectory over time[69].

Put together, those studies gradually improved spatial resolution from single cells to single-molecules, eventually permitting the observation and quantification of protein diffusion *in vivo* and in real time[70] (Figure 1.14).

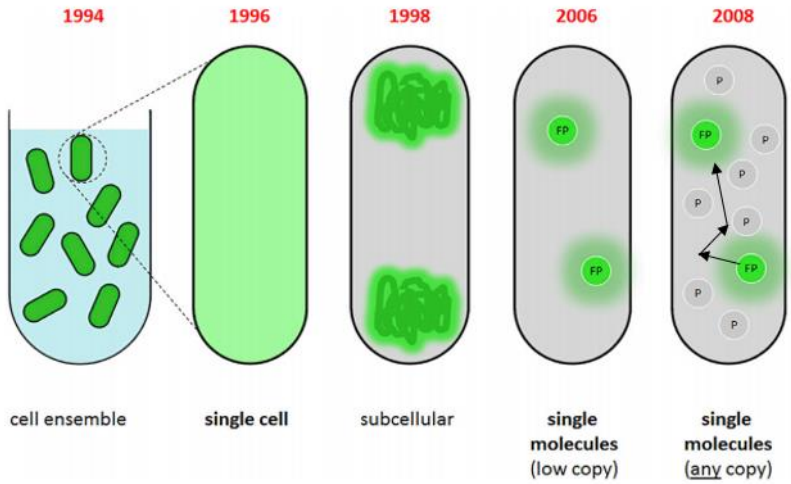

Figure 1.14. The evolution of fluorescent microscopy resolution. (**1994**) GFP was first used to look at cell population and averaged signal and then (**1996**) at single cells. (**1998** and **2006**) The improvement of spatial resolution allowed the detection of subcellular compartments, followed by actual single molecules of YFP, present in low copy numbers. (**2008**) Eventually, the development of photo-activation techniques freed scientists from the necessity of using low copy numbers proteins (P). Adapted from Kapanidis, A. N., Lepore, A. & El Karoui, M. Rediscovering Bacteria through Single-Molecule Imaging in Living Cells. *Biophysical Journal* **115**, 190–202 (2018).[71]

Consequently, it became possible to study the dynamics of proteins such as the *lac* repressor and extract its association constant (with its chromosomal operator), which Elf and colleagues[72] measured as one order of magnitude lower than what was previously measured *in vitro*. These results highlighted the limitation of membrane permeability to the repressor's substrate, IPTG (thus decreasing the speed at which LacI would bind DNA), and the limitations of *in vitro* measurements. Moreover, this team observed that LacI spent on average 90% of its time non-specifically interacting with DNA, and went as far as to estimate its time spent on the DNA to be inferior to 5 ms, thus revealing unprecedented insights into protein-DNA interactions.

An additional interesting feature which pertains to the diffusion of a molecule and which can also be extracted from such techniques is the Mean Square Displacement[73,74,75], a measure of the deviation of the position of a particle with respect to a reference position over time[76]; it was initially developed to study gas particles going through random collisions, propelled by stochastic fluctuations in temperatures[77]. In a relevant example applied to protein diffusion, Billaudeau and colleagues[78] demonstrated that the actin-homologue MreB, involved in the elongation of the bacterial peptidoglycan (PG) was diversely regulated in *E. coli* and in *Bacillus subtilis*, which they accounted on their different cell walls. Furthermore, they observed in *E.coli*

that following nutrient upshift, the fraction of MreB proteins rotating around the cell and following the membrane (and thus synthetizing PG) increased, suggesting that cells would adapt to growth rate upshift by increasing cell wall production, resulting in cell elongation (Figure 1.15).

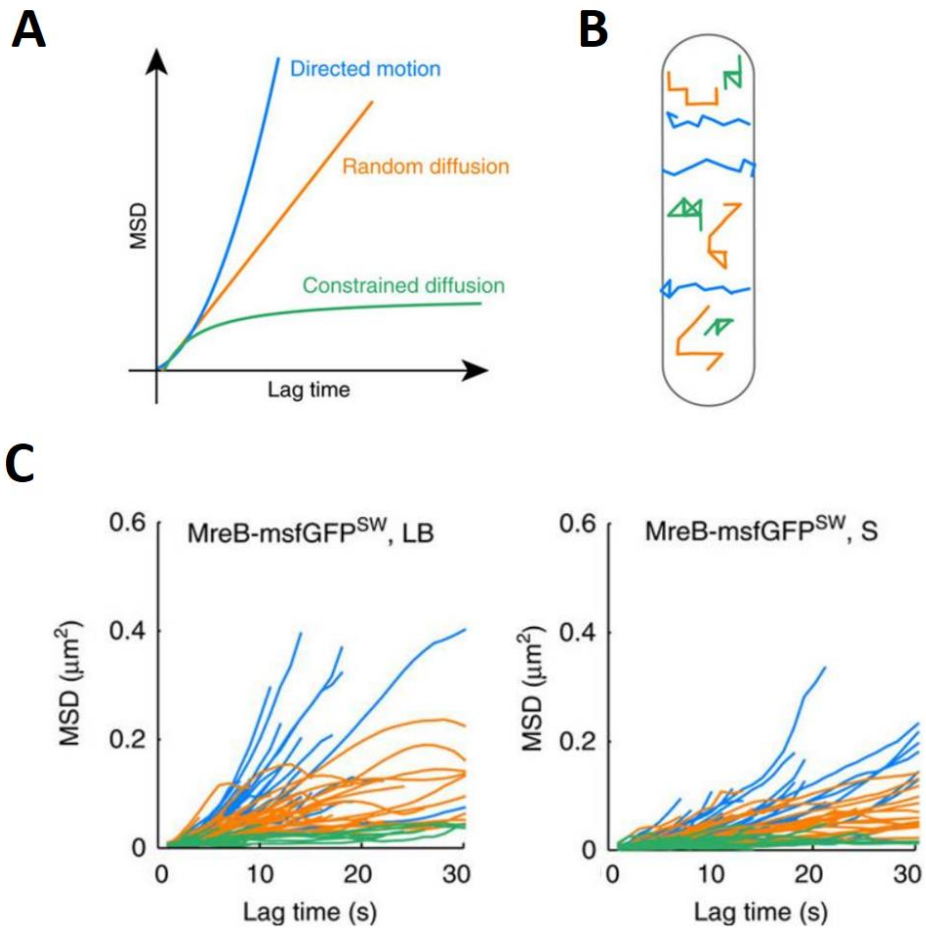

Figure 1.15. Extraction of Mean Square Displacement (MSD) to quantify MreB diffusion behaviours. (**A**) Theoretical representation of the three main modes of protein diffusion and their respective MSD curves; directed motion in blue, random diffusion in orange and constrained diffusion in green. (**B**) Theoretical trajectories associated to their respective MSD, drawn in a rod-shaped bacteria. (**C**) Experimentally extracted MSD of the MreB protein (fused to the monomeric-superfolder-green GFP, msfGFP) observed in *E. coli*, either in LB-rich media (left) or S-poor media (right). Heterogeneous diffusion behaviour can be observed, with the directed-motion fraction increasing when enriching the media, to the expense of the constrained fraction. Adapted from Billaudeau, C. *et al.* Contrasting mechanisms of growth in two model rod-shaped bacteria. *Nat. Commun.* **8**, (2017).[78]

In 2017, the Elf lab applied their single-molecule microscopy expertise to study the catalytically-dead dCas9 mutant[79]. They used an elegant method based on plasmids carrying repeated *lacO* targets to draw dCas9, associated with matching gRNAs. As illustrated in Figure 1.16, upon addition of IPTG (and subsequent release of LacI from *lacO*, thus allowing the CRISPR complex to recognise the targets), they measured the time needed for the fluorescently-labelled protein to find its DNA target (thus yielding a fluorescent focus), eventually resulting in an averaged protein searching time of 6h. Interestingly, such a long search time indicated that the protein was most likely slowed down by other interactions, and the authors suggested that it was probably busy interacting with the numerous PAM sequences found across the genome.

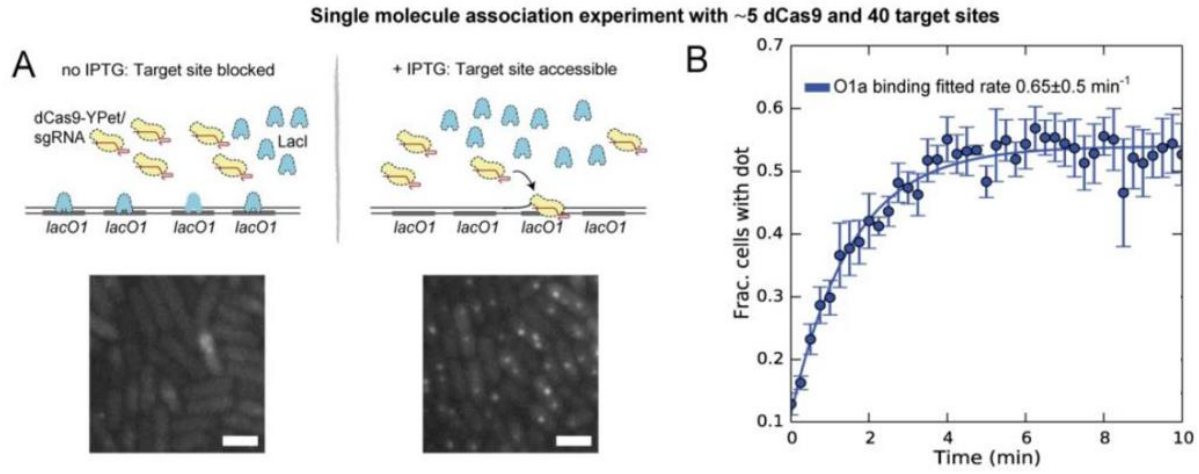

Figure 1.16. dCas9 recruitment after target site made accessible. (**A**) Schematics and associated microscopy images of dCas9-YPet, associated to a gRNA targeting *lacO* repeats present on a plasmid. Upon addition of IPTG, LacI proteins are released from the *lacO* targets, allowing dCas9 recruitment. (**B**) With on average 5 proteins and 40 target sites per cell, the fraction of cells with at least one detected fluorescent spot increases as a function of time, allowing the extraction of the CRISPR-*lacO* binding time of 0.65 min^{-1} which, after renormalization, results in a search time of 6h. Extracted from Jones, D. L. *et al.* Kinetics of dCas9 target search in Escherichia coli. *Science (80-.).* **9**, 1420–1424 (2017) [79].

Despite its excellence, this study neither observed the original Cas9 protein, neither did it study the context of genomic DNA. Rather, it used a catalytically-dead mutant which does not behave as its original counterpart (dCas9 was reported to stay bound to its target for several dozens of minutes[79], as opposed to the wild-type version which was observed for less than a minute[80]). This would indeed affect the interaction time of the protein with DNA, for instead of cleaving the target and leaving the site of the break, the protein stayed bound[80]. Thus limiting the interactions of other dCas9 proteins with the target, it also potentially affected interactions with other DNA-interacting proteins (*e.g* proteins involved in repair, transcription or replication).

Additionally, targeting a plasmid is different from targeting the genome; the search may be facilitated, diffusion is surely higher and there may be less proteins in the vicinity. In conclusion, although this excellent work did improve our understanding of the CRISPR interactions with DNA, a number of uncertainties remains, particularly regarding OFF-target events which were neither detected, nor mentioned.

In 2015, a team led by Jennifer Doudna[81] was the first to investigate Cas9 diffusion in mammalian cells, also using the catalytically-dead dCas9 mutant, fused to the HaloTag protein. Targeting short interspersed nuclear elements (SINEs) of the B2 type present in the cells at about 350,000 copies, their results showed that the protein search was dominated by 3D diffusion and that OFF-target events where short-lived, with an average residence time inferior to 1s. Remarkably, quantifying the protein apparent diffusion coefficient (D, obtained from fitting MSD curves), they observed diffusion shifts towards slower values as a function of gRNA association. Indeed, increasing the gRNA homologies for the SINE target, they saw the fraction of dCas9 freely diffusing decrease (Figure 1.17).

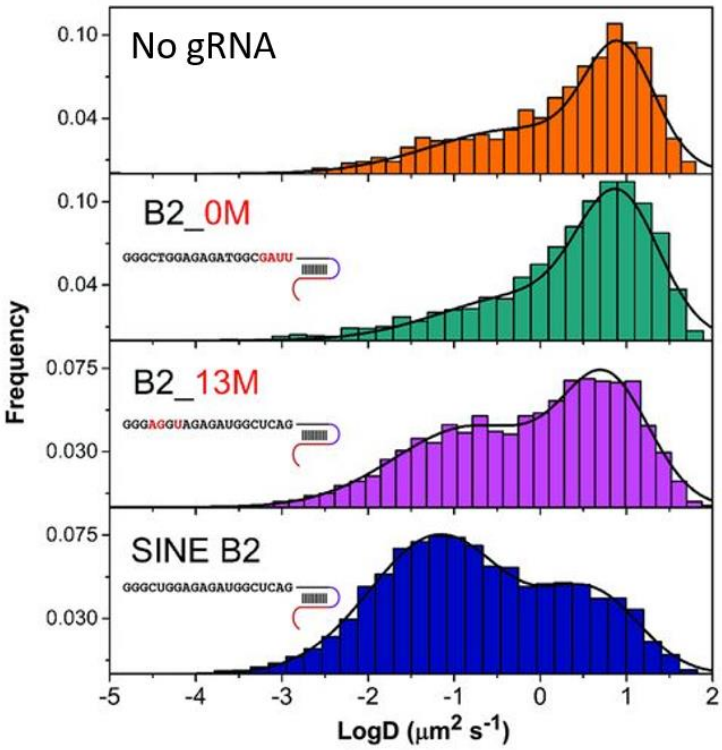

Figure 1.17. The diffusion of dCas9 decreases when associated to a gRNA. Apparent diffusion coefficient distributions of dCas9 either alone (no gRNA, orange), with a gRNA containing mismatches proximal to the PAM region (B2_0M, green), with a gRNA containing mismatches distal from PAM region (B2_13M, purple) and with a gRNA with the correct target sequence (SINE B2, blue). By increasing Cas9 chances to find a target, its diffusion decreases, indicating more interactions with DNA. Log scale. Extracted from Knight, S. C. *et al.* Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science* **350**, 823–6 (2015) [81].

These interesting results seemed to stand in contradiction with the previous study (Elf and colleagues), where dCas9 was observed to predominantly interact with DNA, thus resulting in long searching times. In contrast, Doudna and colleagues suggested that dCas9 spent most of its searching time diffusing, with very few OFF-target events. It is notable that the present study occurred in mammalian cells, where the accessibility to DNA is different from what is found in bacteria; indeed, mammalian DNA is packed in chromatin structures, which can be difficult to access (heterochromatin). Additionally, mammalian cells are far bigger, and their number of targets as well, which makes uneasy the task of comparing one context with the other. Nevertheless, it shows how uncertain this topic remains.

Another recent study harnessed the resolution of High-Speed Atomic Force Microscopy (ATM) to observe the CRISPR complex structure, in real time. Despite the limitations of *in vitro* experiments and the influence of the ATM probe interacting with the protein, Shibata and colleagues[80] managed to uncover striking conformational differences between the Cas9 protein and its catalytically-dead mutant (dCas9). Additionally, they highlighted the importance of the HNH domain, its fluctuations and subsequent stabilization by the R-loop formed by the crRNA and its targeted sequence ('dock-state'), allowing the nuclease to cleave DNA. Yet, they could not explain how the cleaved DNA fragments were released from the protein, itself bound to a mica surface. Regardless, they produced some of the best images ever taken of the Cas9 protein, which is seen in the presence of its gRNA, DNA and Mg^{2+} (which Cas9 needs for DNA cleavage); I take the liberty of reproducing one of their movies, with the following QR code (Figure 1.18).

Figure 1.18. QR-code redirecting to a movie showing Cas9 and its gRNA cutting DNA. Using High-speed AFM, Cas9 is shown in the presence of its gRNA, DNA and Mg^{2+} successfully cleaving DNA, which fragment is then released. This beautiful movie is reproduced from Shibata, M. *et al.* Real-space and real-time dynamics of CRISPR-Cas9 visualized by high-speed atomic force microscopy. *Nat. Commun.* **8**, (2017)[80].

1.5 – Goal of the thesis

We have previously described the outstanding potential of the CRISPR-Cas9 technology, whether in regards to molecular genetics, gene editing or even as an alternative to antibiotics. However, despite the huge amount of already available studies, its main drawback has yet to be fully understood, nor solved. Indeed, OFF-target events remain largely unpredictable and under-studied, stemming from the technical challenges such investigations entail.

So far, OFF-target effects have only been detected when resulting in DNA mutations (indels). However, this is ignoring the potentially vast majority of events where CRISPR-Cas9 does interact with DNA, yet does not cut it, which we could refer as 'silent' non-specific CRISPR-DNA interactions. Those events may be significant, as an immobile DNA-bound Cas9 could represent a major hindrance to other DNA-related mechanisms (replication, transcription, regulation and repair). One way to study these 'silent events' is to use single-molecule microscopy, observing protein diffusion and detecting the different behaviours the protein can adopt; among them, the respective fractions of ON and OFF-targets events.

As shown earlier, previous work found in the literature already used this approach, yet only focusing on the catalytically-dead mutant dCas9, for its easier manipulation, and easier observation[79,81]. However, the protein thus phenotypically different may diffuse significantly differently. Furthermore, its interactions with DNA have shown to be affected, due to the lack of activity of its nuclease domains, crucially impacting the protein conformation[80].

It is in this context that we aim to study the diffusion of the Cas9 wild-type protein, in order to quantify the set of behaviours it can adopt, so as to reveal the fractions of ON and OFF-target events, including the ´silent' OFF-target ones that have not yet been detected. To do so, we want to use single-molecule tracking techniques, *in vivo* and in real time.

For this purpose, we chose our favourite organism, *Escherichia coli*, as its smaller genome and unique chromosome – deprived of chromatin – make it particularly suited to build, study and quantify the CRISPR interactions driven by the various DNA constructs we built for this project. Additionally, being so exhaustively characterized, *E. coli* offers us unparalleled genetic and experimental control.

In the following chapters, you will see described our attempts at harnessing the imaging capabilities of the HaloTag protein and of the HiLo microscopy technique, combined with our custom-made tracking algorithm. I will show how the displacement of the Cas9 protein was successfully tracked, in various conditions, and how it allowed us to infer its different diffusion behaviours, characterized by its range of interactions with the miraculous molecule of DNA.

Chapter II – Materials and Methods

2.1 – Constructs

In this study, I used *Escherichia coli* strains DH5α, MG1655 and BW27783. These strains either came from the lab collection, either from other groups. They are all referenced in Table 3 of the Appendix. Each associated experiment is recorded on the Table 4, also in the Appendix.

CRISPR

Our main protein of interest was *Streptococcus pyogenes* Cas9 expressed under the control of the araBAD promoter (also written pBAD)[82]. The pBAD-Cas9 construct was integrated into *E.coli* genome through clonetegration[83], using the plasmid pOSIP-CT. More information on this construct is present further down and illustrated on Figure 2.3.

Several single guide RNAs (sgRNAs) were used in association with Cas9. In every case, the 20-nt target sequence (crRNA) was flanked by a Cas9-handle chimera (tracrRNA) under the control of the constitutive promoter J23119 (Figure 2.1). Those sgRNAs were on a kanamycin-resistant, low-copy plasmid (sc101 origin of replication). Their target sequences are indicated on Table 6, in the Appendix.

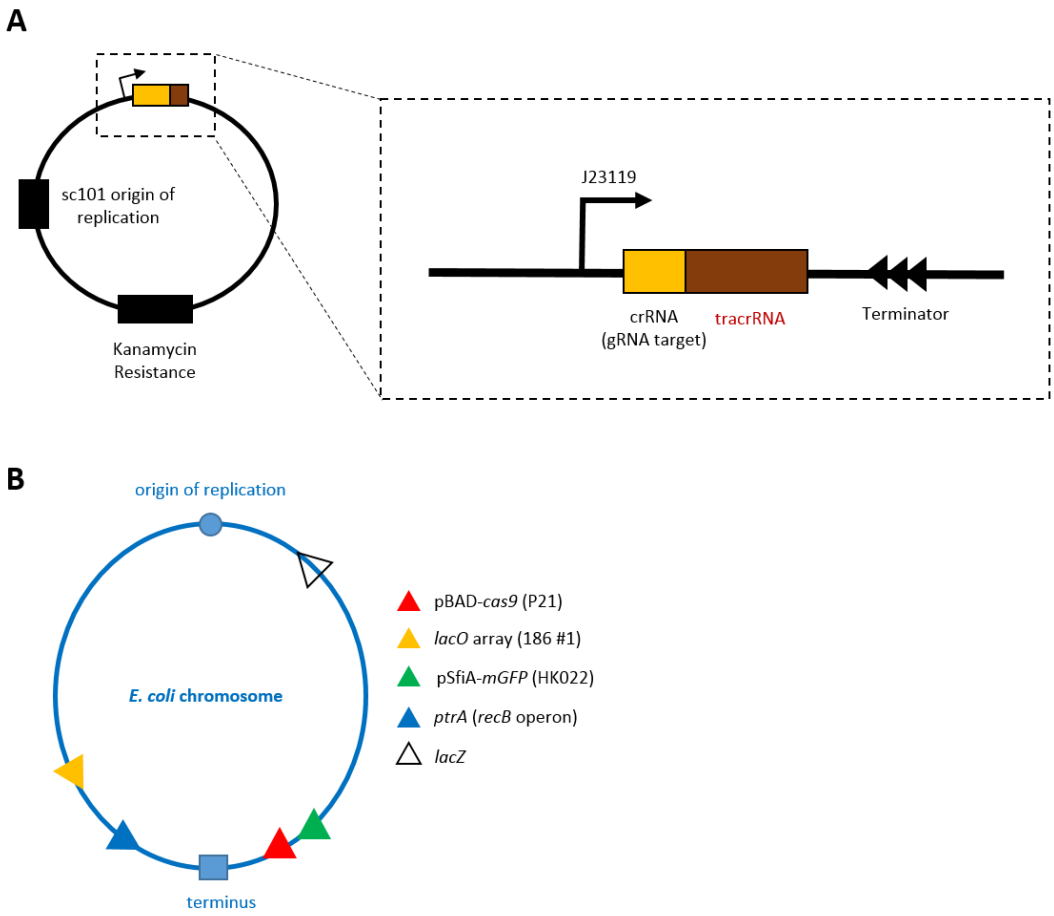

Figure 2.1. sgRNA expression and target. (**A**) Schematic of the plasmids carrying the sgRNAs, with the origin of replication and the kanamycin resistance gene. Zoomed-in, we see that the sgRNAs (combining the crRNA and the tracrRNA) are flanked by the J23119 promoter and a terminator region. (**B**) On this map of the *E.coli* genome are represented the various positions of the sgRNA targets and the integrated constructs with the origin of replication, the terminus and the *lacZ* location as references.

The *ptrA*-sgRNA targets *ptrA*'s promoter (conserved gene within the *recB* operon whose function is unknown). The *lacO*-sgRNA targets *E.coli*'s *lacO* wild-type sequence. Finally, the *lacOA*-sgRNA only targets the *lacO* repeats present on the *lacO* array. All three sequences are shown on Figure 2.2. Note that the *lacOA*-sgRNA is not expected to successfully cut the endogenous *lacO* sequence because of the misalignment between the crRNA and the *lacO* sequence; this was deliberately made this way to ensure the specificity of the *lacOA*-sgRNA for the array. Both sgRNAs have previously been used in the literature[79], which encouraged us to use them as well.

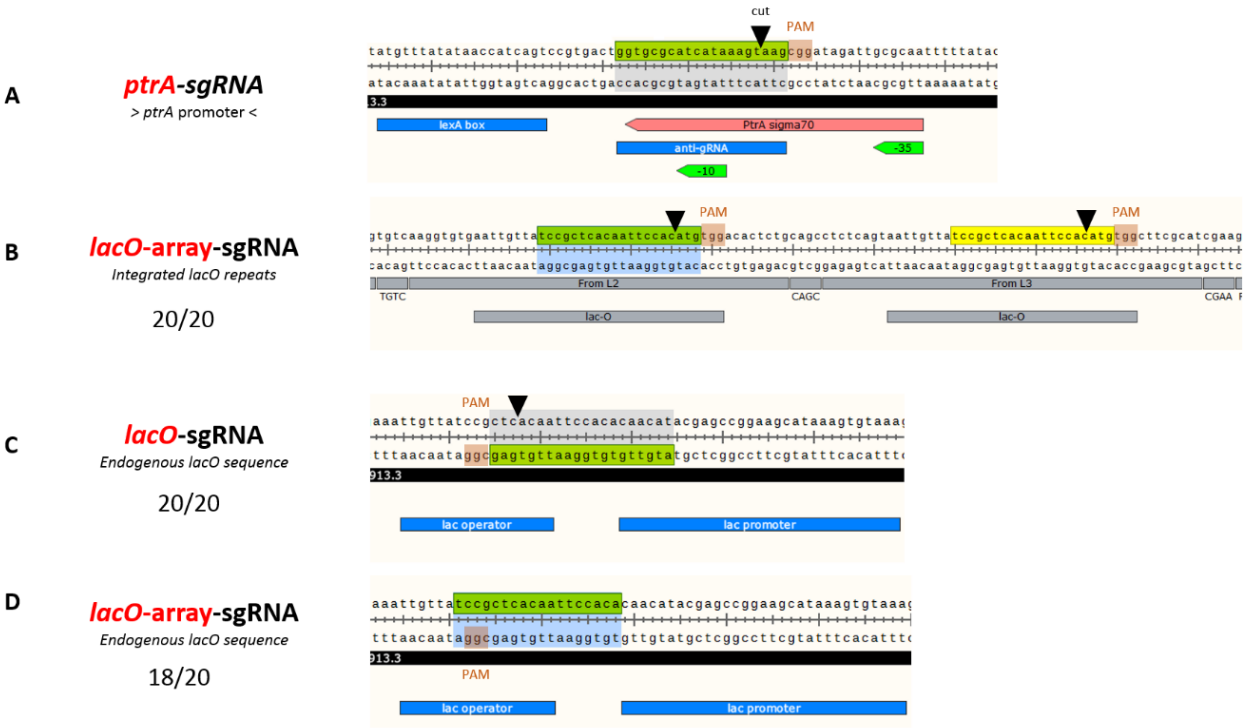

Figure 2.2. sgRNA target sequences. Our three sgRNAs target the *ptrA*, *lacO* and *lacO-array* sequences, the latter being artificially introduced on the chromosome. We see that (**A**) the *ptrA*-sgRNA displays a perfect match with its target sequence and the cut is happening. (**B**) Similarly, the sgRNA targeting the *lacO* array perfectly matches its targets and cuts. (**C**) However, the sgRNA targeting the endogenous *lacO* sequence, although displaying a perfect match, does not appear to successfully cut. See chapter 4 for more details. (**D**) Finally, the sgRNA targeting the *lacO* array was designed to *mis*match the endogenous *lacO* sequence (18 nucleotides match the sequence, but do not properly align with the PAM sequence), and therefore does not successfully cut. The black triangles show where the cuts are happening; in green are highlighted the sgRNA homologies with the target sequences; in red the PAM sequences are displayed and the black triangles show where the cut should be happening.

araBAD promoter

Directly upstream of the pBAD promoter is expressed the AraC protein; in the presence of arabinose, it enhances the recruitment of the RNA polymerase to the pBAD promoter. However, in its absence, it spontaneously prevents its recruitment due to the formation of a DNA loop[84] (Figure 2.3). Additionally, in the absence of glucose, the CAP protein (cAMP receptor Protein) binds to its CAP site, also helping with the recruitment of the RNA polymerase[85]. In *E. coli*, arabinose is transported inside the cell through two types of transporters, low affinity (AraF, G and H) and high affinity (AraE) which synthesis is inducible by arabinose. In the BW27783 strain, *araF, G* and *H* are deleted and *araE* is put under the control of a constitutive promoter to ensure that the pBAD promoter leads to a homogeneous induction.

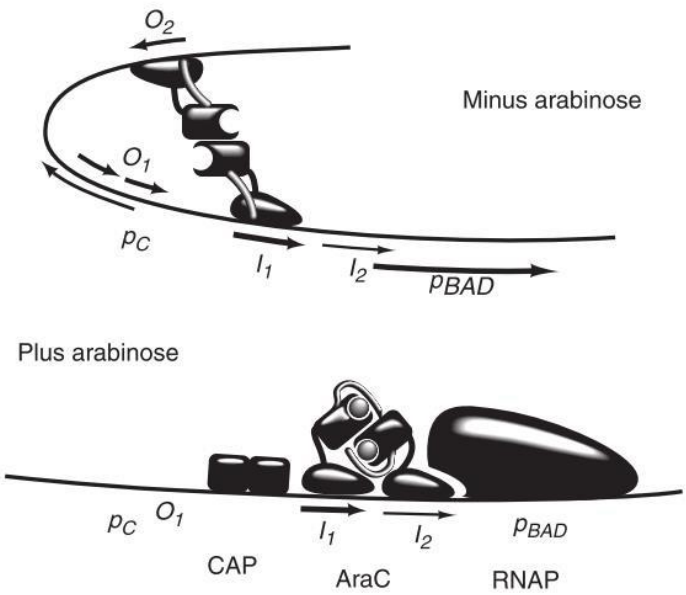

Figure 2.3. araBAD promoter regulation mechanism. In the absence of arabinose (top), the AraC protein forms a dimer that binds together the two operator regions O_2 and I_1, consequently forming a DNA loop which prevents the recruitment of RNA polymerase (RNAP), therefore repressing the expression of the protein (here, Cas9). In the presence of arabinose (bottom), the dimer changes its conformation, this time binding to the I_1 and I_2 operator regions, therefore removing the DNA loop and allowing the recruitment of the RNAP. Additionally, the absence of glucose permits the CAP protein to further increase the recruitment of the RNAP. Extracted from Schleif, R. AraC protein, regulation of the L-arabinose operon in Escherichia coli, and the light switch mechanism of AraC action. *FEMS Microbiol. Rev.* **34**, 779–796 (2010) [82].

Other promoters (pTet, leaky and tightly regulated) have also been tested but left aside in favour of the pBAD promoter, due to its more homogenous expression; they are not described here.

SOS reporter gene

Also integrated into our cells genome is the pSfiA-mGFP construct, acting as a fluorescent reporter of the SOS response. SfiA (also called SulA) inhibits cell elongation by preventing the formation of the division ring. Its expression is induced by SOS activation[86]. Here, we integrated this construct through clonetegration, using the pOSIP-KH plasmid[83].

HaloTag protein, fusion and labelling

The HaloTag protein is a haloalkane dehalogenase[87] which active site has been modified in order to irreversibly bind a chloroalkane linker, itself fused to a molecule called TMR ligand (for tetramethylrhodamine)[88] or a variant called JF549[89] (Figure 2.4). These ligands (also called HaloTag ligands) are highly photostable fluorophores particularly suited for single-molecule fluorescence microscopy. Excited at 555 nm, they emit at 585 nm. With a HaloTag protein fused to our protein of interest (here, Cas9), we make sure that we obtain one-to-one labelling.

Using Gibson Assembly, we fused the HaloTag protein and a 24-nt linker to Cas9, in C-ter. The fusion has been done on the pOSIP plasmid, prior to integration. More information on this construction can be found on Table 2 of the Appendix.

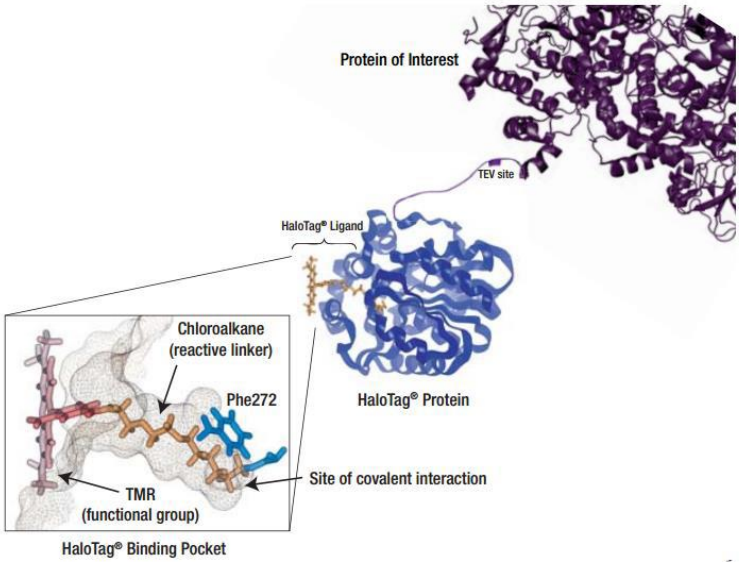

Figure 2.4. HaloTag protein and its ligand. Fused to the protein of interest (here, in C-ter of the Cas9 protein), the HaloTag protein possesses an active site where its HaloTag ligand (TMR or JF549) can covalently bind. The mutated phenylalanine residue (Phe272) which makes the interaction irreversible can be observed in blue. *Taken from HaloTag® Technology, Promega™ website.*

During the expression of the HaloTag-fused protein of interest (usually for one hour), the HaloTag ligand is simultaneously added in excess (5µM final concentration for TMR and 1µM for JF549) for an hour, during which the HaloTag proteins get labelled by HaloTag ligand molecules. After this hour, the unbound ligands are washed out through 4 repeated washes, where cells are spun down, a pump is used to suck up the supernatant, and the cells are resuspended in a new tube with the media used for the experiment. Eventually, the cells are ready to be loaded on an agar-pad for microscopy.

lacO arrays

Several sets of *lacO* repeats were built in association with the Edinburgh Genome Foundry, of which two were successfully integrated into *E.coli*'s genome. They contain 6 and 22 *lacO* repeats respectively. Each repeat is flanked by a PAM sequence, allowing for Cas9 recognition (see Figure 2.5). The arrays were integrated into *E.coli* chromosome through clonetegration, using pOSIP-KO. Those arrays were inspired by work previously done in the Elf lab[79]. The repeats and associated sequences can be found on Table 6 in the Appendix.

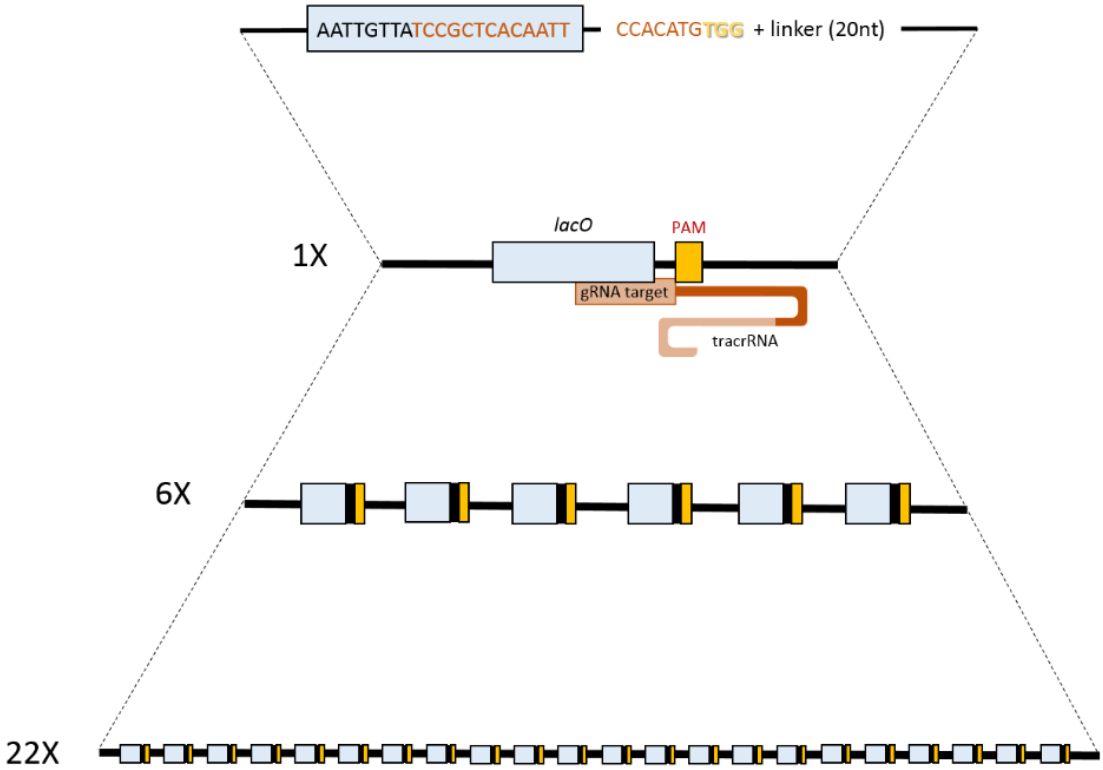

Figure 2.5. LacO arrays architecture. The *lacO* arrays we built and further integrated into *E.coli* chromosome are made of different number of repeats (22X or 6X) which all follow the same pattern; the sgRNA targets 20 nucleotides (in brown) that overlap part of the *lacO* sequence (13 nucleotides) and part of the array (7 nucleotides), flanked by the PAM sequence (TGG, in yellow). Each unit is separated from the next by a unique 20-nucleotide linker, designed to prevent recombination.

ΔRecA strains

Using PMGR (described below), the *recA* gene was knocked out. The ΔrecA phenotype was consequently confirmed by UV testing (see below 2.4 – Viability Tests).

2.2 – Cloning techniques

Gibson Assembly

This technique was consistently used throughout this study to fuse DNA fragments and build plasmid constructs. Following the protocol published by Gibson and colleagues[90], 40-60 nucleotides of homology are needed between the fragments to ligate, either achieved using PCR and overhang primers (which carry the homology), and/or using restriction enzymes (if the homology is already present on the fragments). Briefly, a 5' exonuclease (T5) produces single-stranded DNA allowing homology sequences to anneal, while a polymerase also present in the buffer (Phusion®) fills the gaps, finally allowing a ligase (Taq) to bind the fragments together. After one hour at 50°C, the resulting product is transformed into competent cells, selected with the appropriate antibiotics (Figure 2.6). For more specific details about plasmid construction, please refer to the Table 2 of the Appendix.

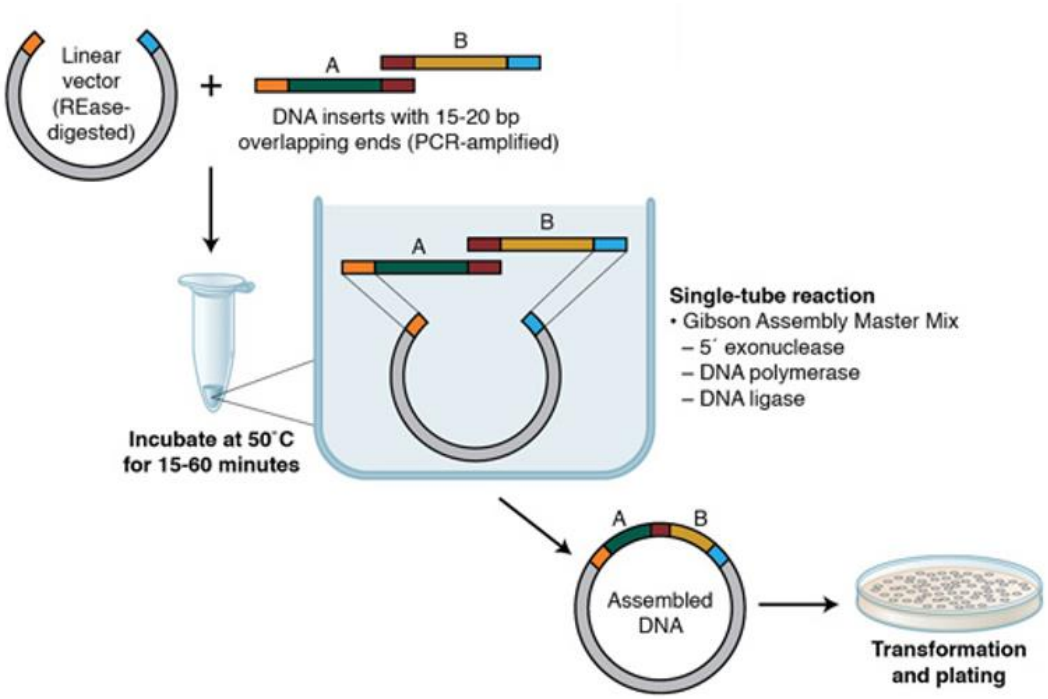

Figure 2.6. Gibson Assembly. The Gibson Assembly technique allows the fusion of several DNA fragments in a single step. In a tube placed at 50°C for an hour are added the fragments to fuse in addition to a 5'-exonuclease, a polymerase and a ligase. The fragments need to possess ends homologous to each other in order to anneal when put in contact. First, the 5'-exonuclease produces single-stranded DNA ends that can be annealed with neighbouring complementary fragments. Second, the polymerase fills the gaps, allowing the ligase to bind the fragments together. Finally, the DNA product is transformed in the presence of the relevant antibiotic to select cells harbouring the resulting plasmid. *Taken from NEB™ website.*

Clonetegration

Developed by St-Pierre and colleagues[83], this technique allows efficient integration of foreign DNA into *E.coli*'s chromosome at various predetermined genomic positions (Figure 2.7B). Briefly, the integration relies on a pOSIP plasmid carrying the sequence to be integrated (flanked by FRT sites) and constitutively expressing a temperature-sensitive phage integrase only active above 30°C. Once the whole plasmid integrated at the genomic position of choice, a second plasmid carrying a flippase (pE-FLP) is transformed to remove the pOSIP backbone, effectively leaving only the sequence of interest in the genome (Figure 2.7A). The resulting pOSIP plasmids and strains are respectively displayed in Table 1 and Table 3 of the Appendix.

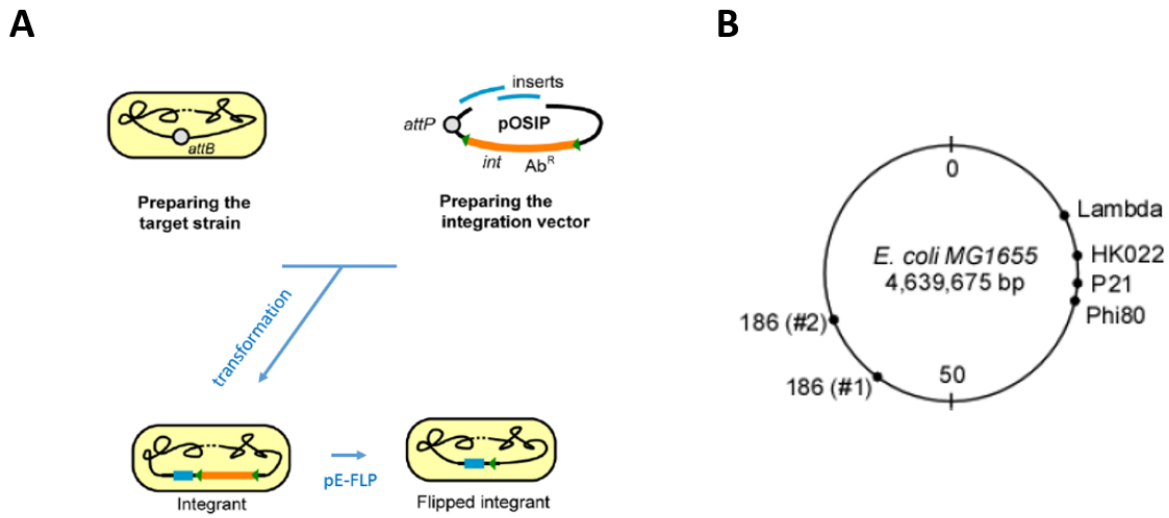

Figure 2.7. Clonetegration. This technique allows the integration of external DNA in *E.coli*'s genome. (A) The pOSIP plasmid containing the sequence to integrate (in blue) is transformed into the recipient strain, previously made competent. Through its phage integrase, the whole plasmid is integrated at the *attB* locus of choice, with its FRT sites (green arrows) flanking the sequence of interest. Later on, the plasmid backbone (in orange) is removed using a helper plasmid pE-FLP containing a flippase. (B) This map of *E.coli* chromosome shows the different sites where clonetegration can be used. Each site requires a different plasmid. Adapted from St-Pierre, F. *et al.* One-step cloning and chromosomal integration of DNA. *ACS Synth. Biol.* **2**, 537–541 (2013).

To integrate the pBAD-Cas9 construct, the P21 site was chosen for historical reasons. Similarly, the SOS marker pSfiA-GFP was integrated at the HK022 site. Finally, the *lacO* arrays were all integrated at the 186 #1 site, due to its proximity to the terminus region as well as its high integration efficiency. Indeed, the proximity with the terminus ensured a low variability in terms of copy numbers. The choice of the highest integration efficiency was due to the challenging nature of the construct, with DNA repeats presenting risks of DNA recombination (Figure 2.8).

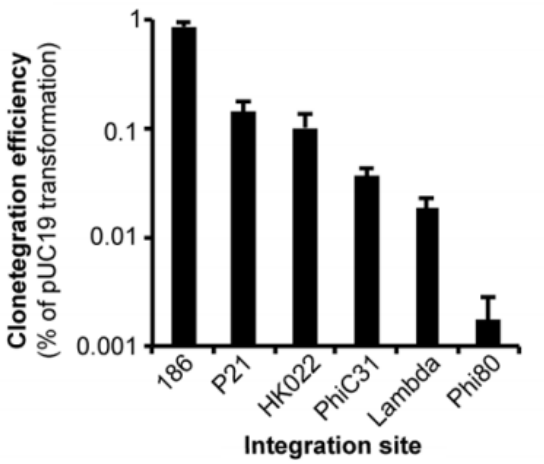

Figure 2.8. Clonetegration efficiency. The efficiency of integration varies with the site used by the phage integrase. We see that 186 gives the best efficiency, while Phi80 gives the worst. Extracted from St-Pierre, F. *et al.* One-step cloning and chromosomal integration of DNA. *ACS Synth. Biol.* **2**, 537–541 (2013).

Plasmid Mediated Gene Replacement

Plasmid Mediated Gene Replacement (PMGR) also allows DNA integration into *E.coli* genome, but relies instead on homologous recombination and a helper plasmid[91] (pTOF). The pTOF plasmid carries the sequence to be integrated, flanked by 100-nt homology arms making recombination possible. Through several selection stages, the first step sees the integration of the whole plasmid, while a second step selects cells that have gone through the recombination an additional time, consequently getting rid of the plasmid backbone (Figure 2.9). As opposed to clonetegration, PMGR is not restricted to a particular locus; however, its efficiency varies significantly from one locus to another. This method was used to build the strains containing the *recB-Halo* construct, as well as those lacking the *recA* gene.

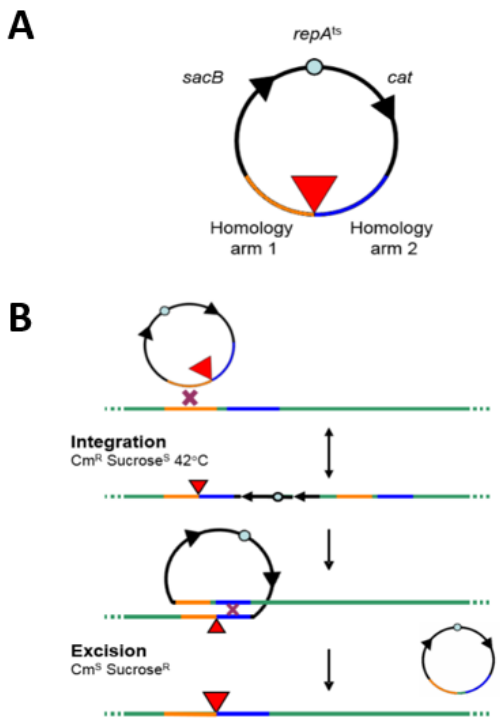

Figure 2.9. PMGR. (**A**) Plasmid Mediated Gene Replacement integrates external DNA into *E.coli*'s genome, using a pTOF helper plasmid where the DNA of interest (red triangle) is flanked by sequences homologous to the locus of integration (orange and blue). (**B**) A first round of homologous recombination integrates the whole plasmid at the locus of interest; the cells are selected at 42°C, where the plasmid cannot replicate (ensuring that only cells with the plasmid successfully integrated can grow on chloramphenicol plates, the associated resistance gene being present on the plasmid). In a second round of selection, the plasmid recombines with itself, thus removing its backbone from the genome and leaving only the sequence of interest; here, the cells are selected with sucrose, because of the presence of the toxic *sacB* gene on the plasmid, consequently selecting cells that have gotten rid of it. *The figure is by courtesy of Benura Azeroglu.*

Cell competency and Transformation

Transformations were carried out following the Chung protocol[92], relying on heat and osmotic shock to transfer DNA. Accordingly, cells were first made competent (grown until OD_{600} 0.4, left on ice for an hour and spun down at 6.2k g to be resuspended in Transformation Buffer), then put in the presence of plasmid DNA for half-an-hour, still on ice, before being heat shocked. They were then allowed to recover for an hour at 37°C (or 30°C), and finally plated with a selection antibiotic. For more challenging constructs, I used the Mix&Go! Kit[93], by Zymo®.

2.3 – Cell cultures

E. coli cells were grown in Luria-Bertani (LB) broth supplemented, when appropriate, with the following antibiotics: kanamycin (50 µg/ml), chloramphenicol (30 µg/ml) and ampicillin (100 µg/ml). For microscopy experiment, Generic Media was used instead, containing M9 minimal media supplemented with 0.2% glucose, 2mM MgSO4, 0.1mM CaCl2, 1X MEM Essential and 1X MEM non-essential Amino Acids (Gibco®). Please refer to Table 7 and 8 of the Appendix for more details.

For microscopy experiments, cells were grown overnight from a frozen stock, in Generic media (with appropriate antibiotics). The overnight cell culture was then diluted a thousand times and grown until OD_{600} 0.1 (measured with a Perkin-Elmer™ Lambda 25 spectrophotometer). Cells were then pelleted (6.2k g), resuspended in clean media to reach OD_{600} 0.2 and labelled with HaloTag ligand in excess for one hour; the ligand was either TMR (5µM final concentration) or JF549[89] (1µM final concentration). If needed, protein expression was simultaneously induced, using arabinose (0.2% w/v). The excess of TMR was then washed 4 times with the same media by spinning them down (3 minutes, 6.2k g) and the cells finally mounted on agar-pads (made of 2% agarose M9 minimal media, with or without 1mM IPTG).

For counting experiments, the protocol is essentially the same, but in order to fix the proteins and count them, I added a step of chemical fixation right after the washes, consisting of the addition of formaldehyde (2.5% formaldehyde, Thermo Scientific, in 1X PBS) for an hour, at room temperature. The cells were washed twice with 1X PBS and finally mounted on agar-pads.

In parallel, microscope coverslips have been treated to remove any dirt or contaminant susceptible to affect our imaging conditions. To do so, coverslips have been put in a 400°C kiln for 12 hours, then kept in KOH (1M) overnight and finally washed in distilled water where they were kept until experiment, where they were used after being dried with nitrogen gas.

2.4 – Viability Tests

Spot tests were performed to assess cell survivability in the presence of a stressing agent. Cells grown overnight from a frozen stock, in LB media, were diluted 250 times in the morning in order to reach exponential growth. When reaching OD_{600} 0.1, cells were washed in fresh media and left shaking for an hour at 37ºC, under the presence of the stressing agent of choice. In the example of Cas9 activity, 0.2% w/v arabinose, or 0.2% w/v glucose were added. Consequently, cells were 10X-serial-diluted on a 96-microwell plate and plated on an agar petri dish supplemented with the relevant antibiotics. The plating was performed with a replica plater (Sigma®).

UV tests were performed in a similar fashion, with the difference that instead of the 1h-stress exposure, cells were directly serial diluted and plated. Afterwards, the plates were exposed to increasing amounts of UV radiation (with a different level of intensity per plate) using a UV Crosslinker[94] (Scientz®). For instance, 0, 1000, 2000 and 3000 µJ/cm^2. Finally, irradiated plates were wrapped in aluminium foil (to prevent a rare photo-induced alternative repair pathway to intrude) and left at 37ºC over-night.

DAPI staining assays have been used to assess the presence or absence of chromosomal DNA within cells, particularly in the context of ΔrecA genotype. Indeed, upon infliction of DNA damage (with cells exposed to 4ng/µl of ciprofloxacin for 2h), cells deprived of RecA proteins can neither repair the damage, nor prevent RecBCD from digesting the entire chromosome. After two hours and the addition of DAPI (1µl in 1mL of washed cells, left for 15min and loaded on agar), it is possible to identify cells without chromosome, visualized by their absence of DAPI signal.

2.5 – Microscopy

Epifluorescence microscopy (conventional fluorescence microscopy) observations were conducted on a Nikon® Ti-E inverted microscope with a 100X Oil-immersion objective lens (N.A = 1.45), a Lumencor® epi-fluorescent light source, a Perfect Focus system and an EMCCD camera (Andor®). Cube filters (Chroma®) were used in front of the camera to select the wavelengths according to the observed signal (GFP, HaloTag-TMR). Images were acquired using the software Metamorph®[95]. GFP proteins were imaged with an exposure time of 80 ms and a camera gain of 4. As for the HaloTag-TMR signal, we used 200 ms exposure time and a camera gain of 10.

Highly inclined and Laminated optical sheet (HiLo) microscopy is a derivative of TIRF microscopy that covers a bigger volume, allowing us to observe the whole bacterial cytoplasm (Figure 2.10A)[96]. Indeed, TIRF microscopy relies on the production of a so-called evanescent wave that is the result of the complete reflection of the incident laser light and subsequently excites the sample with a penetration of 100 nm. In contrast, HiLo microscopy (also called 'dirty TIRF') relies on laser refraction to excite a wider sample volume (up to 1000 nm penetration), in order to be able to capture the whole cell cytoplasm (Figure 2.10B). This technique was performed on the same optical set up as the one used for epifluorescence, at a penetration depth of 1000 nm and in a temperature-controlled chamber, at 37ºC. The TMR was excited using a 561 nm Coherent® laser, controlled through iLas2® system. We used a 0.2 mW/cm^2 laser power and an exposure time of 12 ms. The recorded movies were taken with a Region-of-Interest (ROI) of 256x256 pixels, this in order to match our exposure time with the camera readout time (19 ms with 512x512, 10 ms with 256x256).

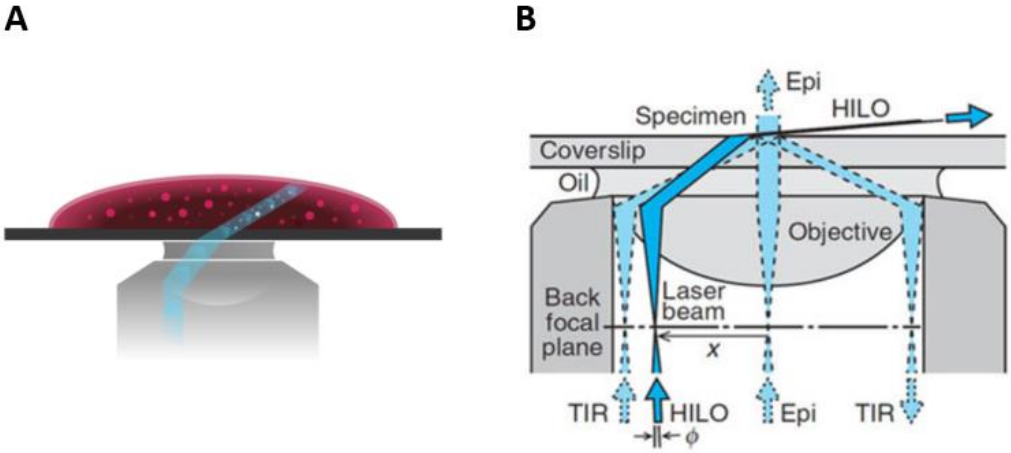

Figure 2.10. HiLo microscopy. (**A**) On this schematic representation, the objective (in grey, below the sample) collects the fluorescent light emitted by the molecules in the sample (red) excited by the refracted laser (blue). (**B**) On this more detailed schematic, we can observe three different microscopy techniques; epifluorescence (Epi) relies on a source of light directly exciting the sample and passing right through it, perpendicular to the coverslip; TIRF microscopy (TIR) relies on complete reflection of the incident laser beam and produces an evanescent wave on a very short distance from the coverslip; finally, HILO uses a refracted laser beam that will excite only part of the sample, decreasing the fluorescent background (as opposed to epifluorescence) while still collecting light from a whole bacterium (unlike TIRF which would only excite part of it). **Figure A** was taken from ONI™ website and **Figure B** from Liu, C., Liu, Y. L., Perillo, E. P., Dunn, A. K. & Yeh, H. C. Single-Molecule Tracking and Its Application in Biomolecular Binding Detection. *IEEE J. Sel. Top. Quantum Electron.* **22**, (2016)[96].

As an illustration, the following QR code redirects towards one of our movies displaying Ca9-HaloTag proteins, labelled with TMR and observed through HiLo microscopy (Figure 2.11).

Figure 2.11. QR code redirecting towards Cas9-HaloTag movie, done with HiLo microscopy. Cell border in green.

2.6 – Image analysis

Cell segmentation and SOS response quantification

Every field of view was observed both within the fluorescent channel and within the Brightfield one, and to both, pictures were taken with Z-stacks (5 stacks separated by 0.4µm reaching a total of 2µm), so as to use the Brightfield images to identify cells edges, a process known as cell segmentation. This particular process was carried out with a custom-made algorithm (Matlab) designed by Sebastian Jaramillo-Riveri[97], itself inspired from a Matlab segmentation routine[98]. Briefly, for each field of view, the algorithm is fed with the brightfield picture displaying the highest contrast (among the Z stacks), then each pixel is scored according to its neighbouring pixel intensity, in order to identify those at the cell edges where indeed, the intensity gradient is highest (because the cell content and the cell membrane are respectively darker and brighter than the image intensity average). A threshold is further applied to select only identified pixels linked to each other, resulting in a binary image (or mask) with each cell individualized with its position, length and fluorescent intensity (Figure 2.12).

Downstream, a curation process manually confirms the quality of the segmentation, while discarding artefacts. For instance, cells in contact with the picture edges are systematically discarded, in order to avoid wrong length measurements. Once the cells properly identified, the program computes the average fluorescent intensity of every cell, normalized by pixel (in order to control for cell shape or length). We can therefore have an accurate measurement of a cell population's SOS induction, with hundreds to thousands of cells per dataset.

We must also note that the choice of segmenting using Brightfield images (and not fluorescent images, which display higher contrast and would therefore be easier to segment) stems from the fact that not all cells display fluorescence, for some do not induce the SOS response (as can be seen on the Figure 2.12). Therefore, segmenting using Brightfield ensures that we measure every cell, and not only those with high SOS induction.

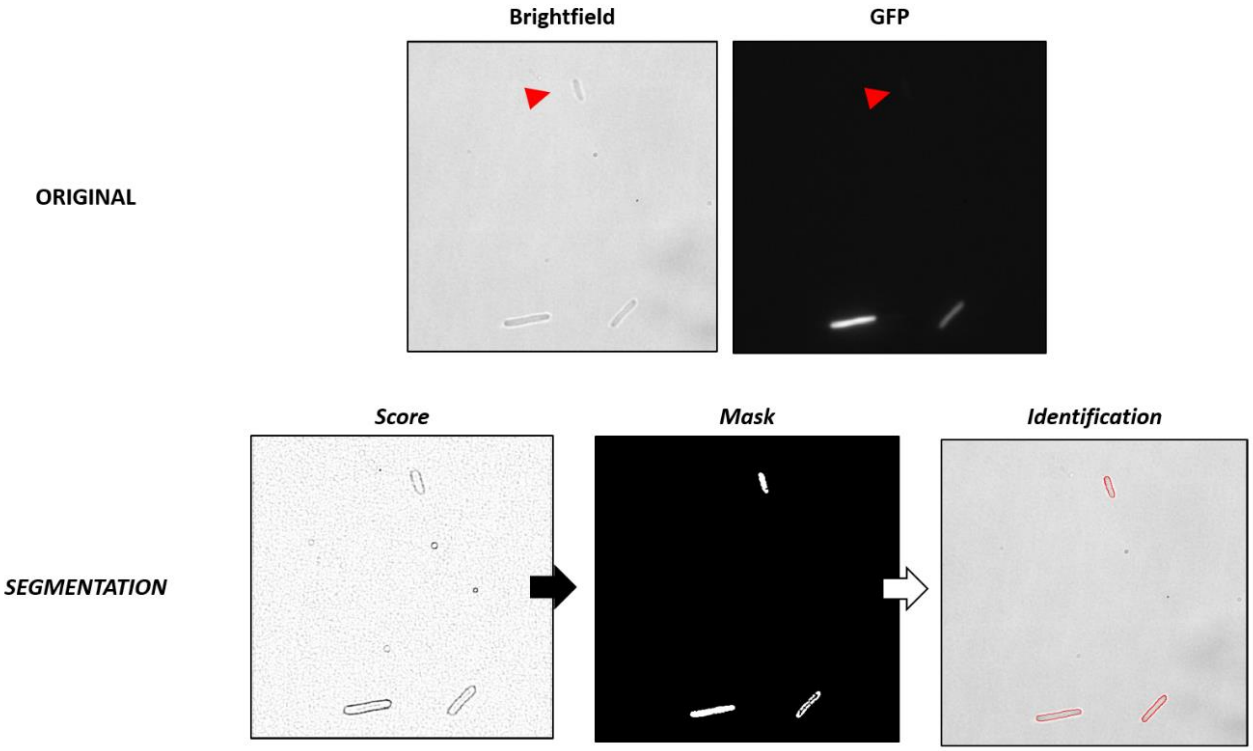

Figure 2.12. Cell segmentation. (**Top**) Brightfield and GFP images of cells expressing pBAD-Cas9-Halo, a sgRNA targeting the *ptrA* gene and the pSfiA-GFP SOS reporter gene. Three cells can be seen, albeit with varying GFP intensities; one is particularly not visible in the GFP channel (red triangle), presumably because it did not induce the SOS response. (**Bottom**) This image is fed into our segmentation algorithm which first identifies the pixels associated to the highest intensity gradients, yielding a score which is then applied to a series of filters (eg. maximum and minimum length) which yield a mask, where the cells are effectively identified, as can be seen on the last picture, where the mask is superimposed with the original Brightfield image.

Fluorescent spot detection and quantification

After chemical fixation, the proteins labelled with TMR were counted in the following way. First, the cells were segmented with their Brightfield signal, as previously described, and then a spot detection algorithm (adapted from Alessia Lepore and colleagues[99], itself inspired from previous protein counting work[100]) was used to identify the fluorescent spots. Briefly, for every field of view, a maximum projection of all Z-stacks images (6) was obtained before going through a band-pass filter[101] supressing the pixel noise, allowing local intensity maxima to be fitted to a 2D Gaussian function so as to accurately identify the position of every fluorescent spot centre (see Figure 2.13).

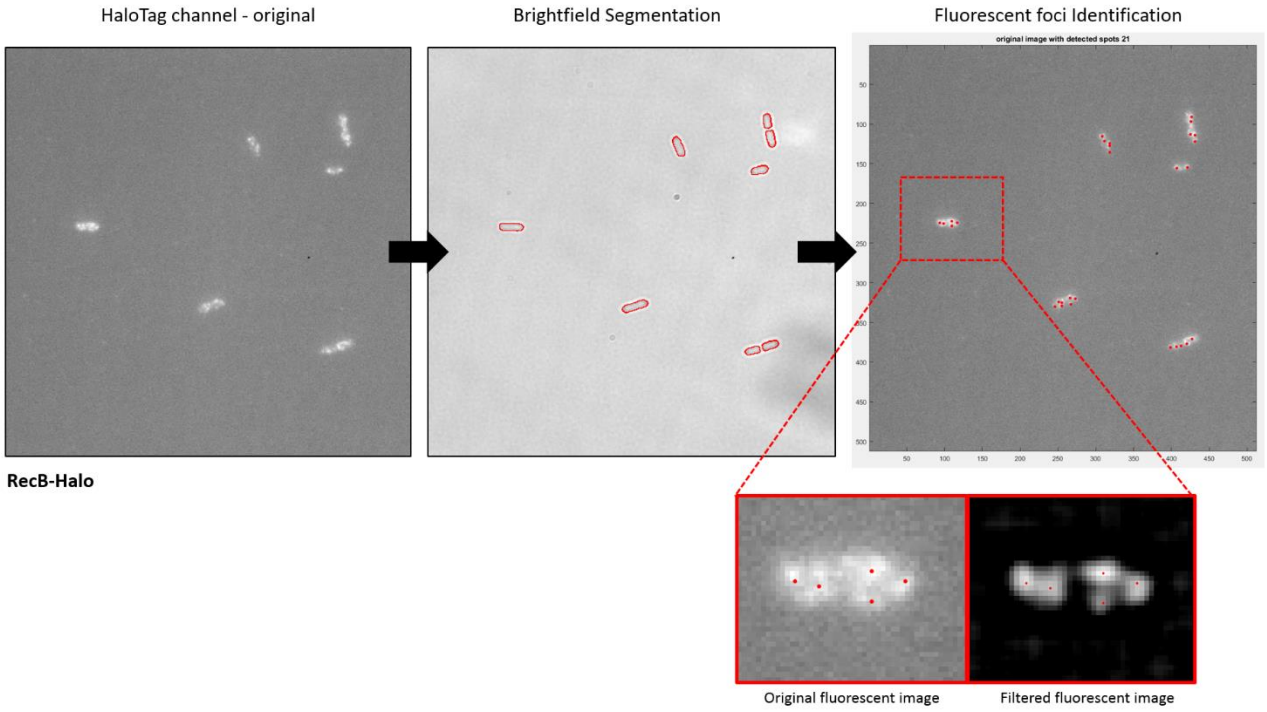

Figure 2.13. Fluorescent spot detection. HaloTag-containing cells can be observed in successive HaloTag and Brightfield channels, with the cell segmentation overlap and spot detection. A zoomed-in picture displays a single cell fluorescent signal, before and after filtering, with the identified foci overlapped (red dots).

Single-Molecule tracking

Our HiLo fluorescent movies were recorded at a rate of 83Hz (a frame every 12 ms), for a total of 600 frames. In order to detect fluorescent proteins moving from one frame to the next, I used a Matlab script written by Alessia Lepore, itself adapted from Crocker and colleagues[102]. Briefly, for every frame, a band-pass filter first supresses the pixel noise, allowing local intensity maxima to be fitted to a 2D Gaussian function so as to accurately identify the position of every fluorescent spot centre. Second, the algorithm links those identified spots between frames, in order to build a trajectory of each particle (Figure 2.14). To be identified as such, those trajectories need to be at least four consecutive frames long, with the maximum distance covered from one frame to the next constrained by the theoretical maximum displacement of the particle (which we empirically defined after analysis of the Cumulative Distributive Function of the particle displacement distribution). When two particles get close to each other, the program follows the particle yielding the lowest displacement. For every movie, we manually optimize the filters in order to maximize the amount of identified tracks without decreasing the tracking accuracy.

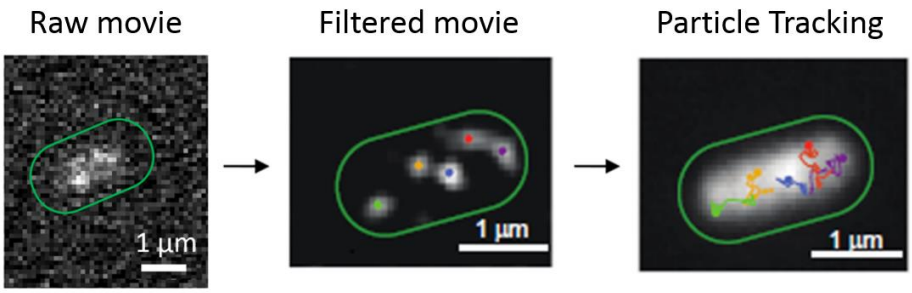

Figure 2.14. Tracking illustration. Each movie is analysed through our custom-made Matlab algorithm where each frame goes through an intensity filter in order to identify the fluorescent spots corresponding to our protein of interest, which can then be followed from one frame to the next, until finally a whole trajectory can be built for every particle. Cell borders in green. Picture by courtesy of Alessia Lepore.

When all the particles have been tracked, we can plot those tracks on the original fluorescent image of the cell and confirm that the tracks have been correctly identified, by checking that they stay within the edges of the cell. Unlike the GFP-SOS analysis, we do not use brightfield segmentation, but rather manually crop the cell using ImageJ and then use the maximum projection (in Z) of its fluorescent intensity to segment it, based on the idea that the particles should move homogeneously within the cell and, coupled to the initial autofluorescence, should make the whole cell area fluorescent when projected over several frames. This successfully allowed us to obtain a resolution differentiating two daughter cells, as can be seen on Figure 2.15.

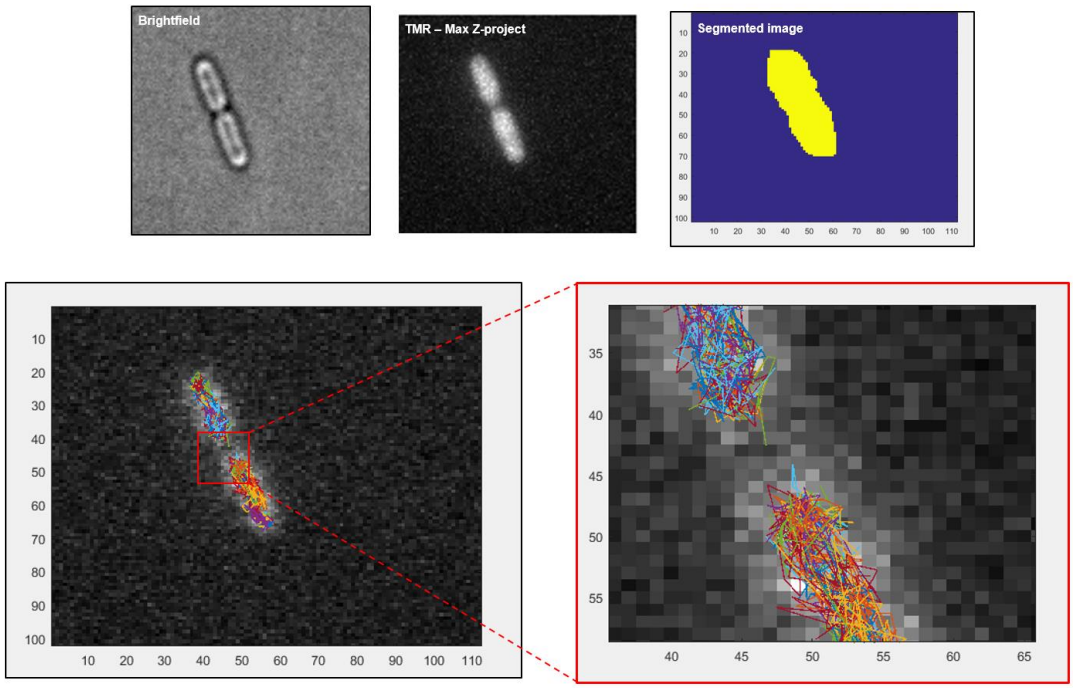

Figure 2.15. Tracking analysis. (**Top**) Each cell is manually cropped on the ImageJ software, as can be seen on the Brightfield picture, while its maximum fluorescent intensity is projected in Z to allow our Matlab software to identify the cell(s) present on the image (segmentation). (**Bottom**) Subsequently, every frame is analysed in order to identify and track the fluorescent spots, built trajectories which can eventually be plotted back on the original fluorescent picture. Our tracking resolution is good enough to differentiate the tracks coming from one cell from its neighbouring sister cell (see zoomed-in subset, in red).

As an example of the tracking process in action, please refer to the QR code below, which redirects to a movie showing Cas9-HaloTag proteins, labelled with TMR and tracked by our algorithm (Figure 2.16).

Figure 2.16. QR code redirecting towards Cas9-HaloTag tracking movie. Labelled with TMR, Cas9-HaloTag proteins cam be seen freely diffusing in the cell (in the absence of gRNA). Overlaid on the original movie, we can observe the tracks detected by our tracking algorithm, following Cas9 over time.

For additional details, a more exhaustive description of the successive steps of the algorithm with associated function names can be found on Table 9, in the Appendix of this document.

Mean Square Displacement (MSD) and apparent Diffusion Coefficient (D_{app})

Using diffusion theories developed to study gas particles randomly moving[103], and considering that freely diffusing proteins adopt a similar random motion (also called Brownian motion), trajectories can be quantitatively analysed. To do so, we applied the same mathematical principles and therefore attributed a Mean Square Displacement (MSD) to every identified particle, which can intuitively refer as a measure of the averaged area explored by a particle during a number of time-intervals. A time-interval represents the time separating one frame to the next; here, one such time interval is equal to our exposure time (12 ms). Two time-intervals therefore represent 24 ms, or the time separating the frame 1 from the frame 3, and so on (Figure 2.17).

The MSD is eventually calculated in the following way:

$$MSD(n) = \frac{1}{N-n}\sum_{i=1}^{N-n}(r_i - r_{i+n})^2$$

with n being the number of time-intervals, N being the total number of frames for which the particle has been tracked, and r_i the position of the molecule at the frame i.

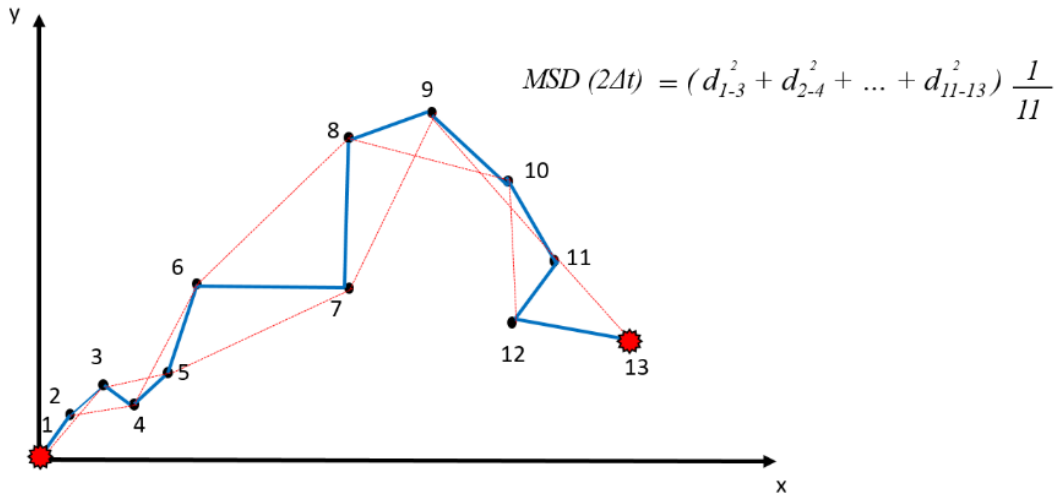

Figure 2.17. Example of MSD calculation. This plot represents a particle (in red) moving in 2D (coordinates X and Y) with its position recorded at every one of the 13 frames within which it has been identified. Its trajectory is highlighted in blue. We can observe that in order to calculate the MSD of this particle for two time-intervals ($2\Delta t$), one has to compute the average of all the squared distances highlighted in dotted red lines, following the written equation.

When the MSD has been computed for every time-interval Δt, it is possible to plot it against the values of those time-intervals and observe the evolution of the particle diffusion over increasing periods of 'time'. When a particle randomly diffuses, its MSD should linearly increase as a function of time-intervals, illustrating that the more time you give a particle to diffuse, the more space it explores (Figure 2.18).

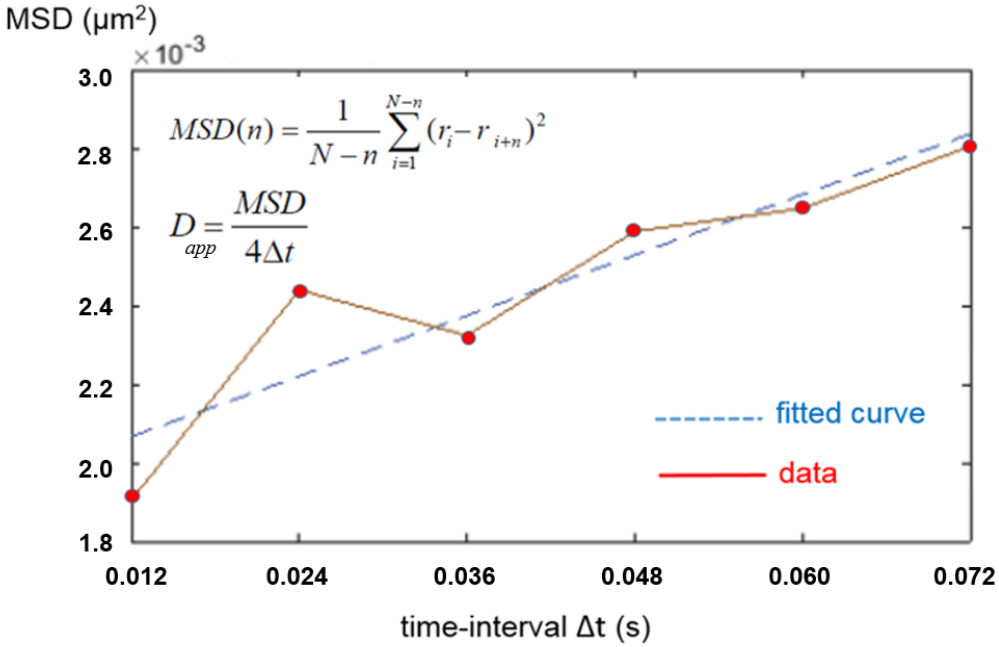

Figure 2.18. MSD plot and extraction of D_{app}. This plot represents the MSD of a particle against 6 time-intervals (up to 72 ms). The experimental data (in red) can be linearly fitted (in blue), illustrating the expected free diffusion of the protein. The equations recall the MSD and D_{app} calculations, the latter being proportional to the MSD gradient.

From this, we deduce the apparent Diffusion Coefficient (D_{app}) of each particle, which is effectively a measure of the diffusion capacity of the particle in its environment, proportionally to the MSD gradient. We can therefore compute it from the first MSD point (being the statistically strongest) and the lag time Δt (here 12 ms), as followed:

$$D_{app} = \frac{MSD}{4\Delta t}$$

The MSD is measured in squared microns (µm^2) whereas the apparent Diffusion Coefficient is measured in squared microns per second (µm^2/s). Following a purely free diffusive behaviour (Brownian motion), a particle MSD should linearly increase as a function of the lag time, however, there are instances where the particle movement is not strictly freely diffusing and that translates into different MSD trends. For instance, a purely one-directional movement (such as a motor protein sliding along a proteofilament) would yield a "super diffusive" behaviour.

On the other hand, a freely diffusing particle constrained in a closed environment would show signs of containment, with its MSD reaching a plateau over longer time intervals Δt (Figure 2.19). The latter case is particularly relevant when one looks at proteins moving inside a cell, for the cell membrane constitutes a barrier through which most proteins can no longer diffuse[104].

Finally, we talk of *apparent* Diffusion Coefficient in reference to the raw value of the parameter, being uncorrected for other features such as the localization error or the dynamic error. The localization error stems from the difficulty of accurately identifying the centre of a particle, given the resolution limit of our microscope setup. In contrast, the dynamic error is a consequence of a particle's motion and the blurriness of the fluorescent signal that is associated; indeed, as a particle moves, its signal gets diluted across the space it has covered, hence the apparent blur.

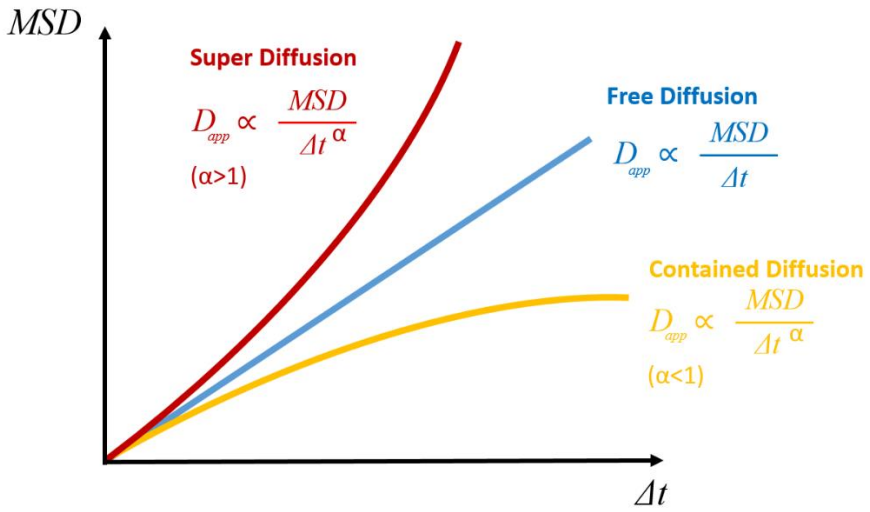

Figure 2.19. Examples of various diffusions. Different diffusion behaviour display different MSD; free diffusion (in blue) yields a linear MSD; super diffusion (eg. directed motion, in red) yields an exponential-like MSD; finally, a contained diffusion (yellow) yields a MSD which ends up with a plateau. All three MSD first display similar values for small time-intervals, and then diverge with bigger ones.

Radius of Gyration (R_g) and Centre of Mass (R)

An alternative to the computation of the MSD is to calculate the Radius of Gyration (R_g), which only relies on a particle's trajectory, that is, without making any assumption on the type of diffusion the particle adopts (Figure 2.20). First, the Centre of Mass (R) is obtained by calculating the average position of the particle across its trajectory, following this equation:

$$R = \frac{1}{N} \sum_{i=1}^{N} r_i$$

where r_i is the position of the particle at the frame i and N is the total number of positions. Then, R_g^2 is obtained by calculating the average squared distance between the particle and the Centre of Mass (R), as can be seen on the following equation:

$$R_g^2 = \frac{1}{N} \sum_{i=1}^{N} (R - r_i)^2$$

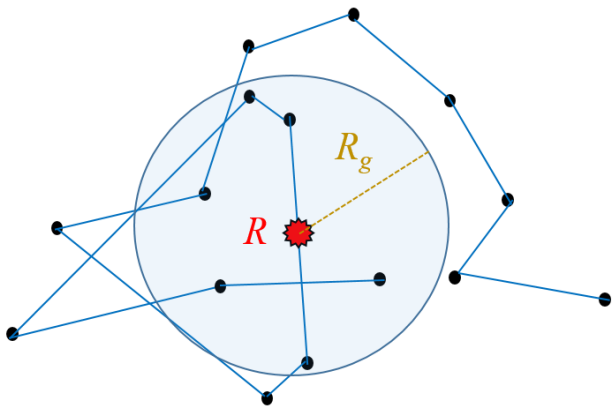

Figure 2.20. Centre of Mass (R) and Radius of Gyration (R_g). Following the trajectory of a particle (blue lines interspersed with black dots), it is possible to compute the particle's centre of mass (R, in red) and its radius of gyration (R_g, in brown).

Chapter III – Results I

CRISPR-Cas9 diffusion: Establishing the Method

3.1 – Introduction

Let us remind the reader of the motivation behind this project; to quantify the diffusion dynamics of the Cas9 protein, so as to unravel how the CRISPR system interacts with DNA. This objective relies on a set of conditions;

- the presence of a strong and stable fluorophore, to observe Cas9 in real time
- the use of a regulated expression system, to express the protein in sufficiently low amounts to perform Single-Particle Tracking
- the use of a suitable microscopy technique, to observe the protein at the single-molecule level
- the use of a reliable image analysis pipeline, to extract quantitative information

All four conditions were met with different techniques which, in the same order, can be summarized as the following;

- the HaloTag protein[67] which, fused to our protein of interest, can covalently bind an external fluorophore, known to be both stable and bright;
- the araBAD promoter[105], which can be induced with arabinose and repressed with glucose, producing a homogeneous expression across the population;
- the HiLo microscopy technique[96], which decreases the fluorescent background due to its narrow illumination, while using a laser for strong excitation;
- custom-made algorithms[103] to detect fluorescent spots with high throughput, build particle trajectories and compute diffusion parameters.

The following chapter is a description of their implementation, development and verification. In parallel to the results obtained through these techniques were also listed the reasons behind their choice, their challenges as well as their limitations.

The first aim of this chapter is to confirm the technical possibility of following the Cas9 protein over time, *in vivo*, through Single-Particle Tracking. The second aim of the chapter is to further quantify the diffusion dynamics of the protein so as to identify the set of behaviours the protein can adopt. Our hypothesis is that quantifying the diffusion behaviour of Cas9 with and without gRNA should highlight the interactions the CRISPR complex has with DNA.

3.2 – Cas9-HaloTag viability and signal specificity

3.2.1 – Cas9-HaloTag looks as active as Cas9 (qualitatively)

The very first thing we needed to control was the ability of our system to be successfully activated and, above all, reliably repressed. Indeed, our main motivation behind a tightly controlled expression system was to avoid the constitutive expression of Cas9 which, in the presence of a gRNA targeting the genome, could very quickly become lethal for the cells. Additionally, single-molecule microscopy requires very small numbers of particles, in order to count them or to track them in time, again highlighting the need for a tightly-controlled expression system.

The araBAD (also called pBAD) promoter is induced by arabinose and repressed by glucose. As a quick reminder, the dimerization of the AraC protein, expressed alongside the promoter and constitutively expressed, encapsulates the regulation by either bending the DNA (and thus inhibiting RNA polymerase recruitment) or either enhancing the recruitment of RNA polymerase. Please refer to Materials and Methods for more details (Figure 2.7).

The second motivation behind the use of this particular system was its ability to yield a homogenous expression across the population. Indeed, such protein expression offers the benefit of decreasing cell-to-cell variability and therefore increasing statistical relevance. Consequently, the pBAD promoter sounded particularly appropriate for our purpose, both because of its regulated induction/repression and its homogeneous expression[105].

We checked the promoter's activity through the expression of the Cas9 gene, placed under its control. For that purpose, cells containing a gRNA targeting the *ptrA* gene (written *ptrA*-gRNA), constitutively expressed from a plasmid (Kanamycin resistant) and pBAD-Cas9, integrated into the genome were exposed to a 1h-pulse of either glucose, arabinose or none of those, then serially diluted and transferred on kanamycin plates (please refer to the 2.4 Viability Test section of the Material And Methods). Cas9's activity being to produce double-strand breaks (DSBs), and those DSBs being hard to repair and thus lethal for the cells, we used spot tests to detect cell survivability, related to Cas9 activity and therefore to pBAD activation. The pulse of protein expression guaranteed the apparition of CRISPR-Cas9 activity at the last stage of the experiment, thus shielding the cells from unwanted DSBs.

In Figure 3.1A, we can observe that cells exposed to increasing concentrations of arabinose increasingly die, which we can compare to the normal growth of cells exposed to glucose. Also, we can observe that cells neither exposed to glucose nor to arabinose seem to grow just as well as those put in the presence of glucose. Indeed, our system presents two levels of repression; an active one in the presence of glucose and a passive one in the absence of arabinose. It seems that the passive repression is enough to prevent cell death.

In other words, there may be several explanations to the lack of cell death in the absence of arabinose; it may indeed be that some Cas9 proteins were still expressed, but not in sufficient numbers to produce DSBs and subsequent cell death. To go further, it was unclear whether a large amount of cells were exposed to repairable DSBs, or whether only a very small numbers of cell did suffer unrepairable breaks. Additionally, when the arabinose concentration reached 10^{-2} %, it was possible to observe some change in colony colour and/or turbidity which we can interpret as the appearance of cell suppressors. Indeed, the high toxicity of the CRISPR-Cas9 system is likely to trigger the appearance of cells harbouring mutations protecting them from the lethality of our system, either by mutating the Cas9 protein, the promoter, the gRNA (on a plasmid) or, most likely, the PAM sequence flanking the target *ptrA* gene.

After confirming the activity of our inducible CRISPR system, we aimed to do the same with our CRISPR-HaloTag system. We thus repeated the previous experiment, this time comparing Cas9-HaloTag to Cas9 wild-type. As can be seen in Figure 3.1B, the system is well repressed, but the HaloTag fusion seems to slightly impair Cas9 activity, since the cells die slightly less than when the wild-type protein is used. Although mostly qualitative, we can make two hypotheses; either the fusion may result in less protein expression (and thus less cell death), either Cas9-HaloTag may be slightly less active than the wild-type protein. If proven to be qualitatively true, those results may affect our future observations of Cas9 diffusion and Cas9 interactions with DNA. Indeed, if the HaloTag fusion were to either affect protein folding, protein conformation with the gRNA or with the DNA, we would consequently underestimate the actual activity of the Cas9 protein. Those observations will be further confronted with single-cell measurements.

But for now, those results confirm that our CRISPR systems are active and their expression successfully controlled.

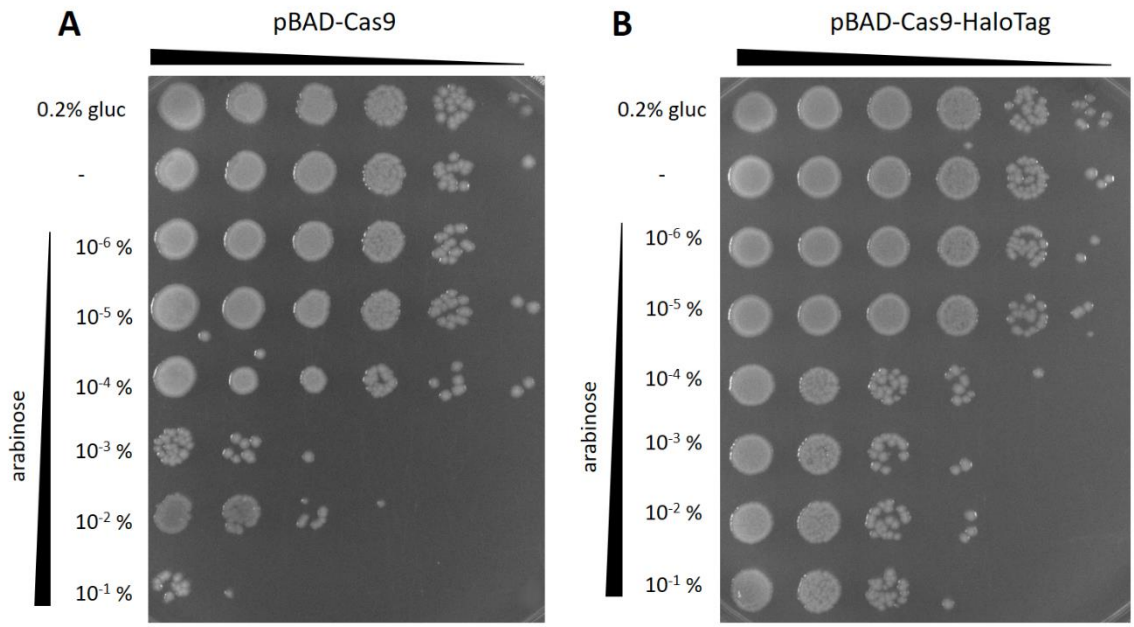

Figure 3.1. pBAD-Cas9 and pBAD-Cas9-HaloTag are operational. Cells containing (**A**) pBAD-Cas9 or (**B**) pBAD-Cas9-HaloTag associated to the *ptrA*-gRNA (kanamycin-resistant plasmid) have been grown over-night in LB supplemented with kanamycin and glucose, diluted then grown until OD_{600nm} 0.1 where they were exposed to either 0.2% w/v arabinose, 0.2% w/v glucose, or nothing (-) for 1h, at 37°C. Finally, they were serially diluted (10X), plated on kanamycin petri dishes and left at 37°C over-night. Those results confirm that in our construct, Cas9 cuts DNA and can be reliably repressed or expressed with varying levels of induction.

In parallel, we also studied the activity of the pTet promoter (inducible with tetracycline or, in our case, anhydrotetracycline). This promoter has previously been reported to yield high induction, but without the possibility of varying expression levels (instead following an ON/OFF or 'bistable' expression). We nonetheless decided to build the constructs and characterize them. Observing the induction of the pTet promoter (Figure 3.2), the activity of Cas9 is maintained but at a lower scale than with the pBAD promoter. Indeed, as quantified on Figure 3.3, the pTet systems yield less cell death than the pBAD systems, which can be attributed to either a lower induction (which could have potentially been solved by increasing the dosage of anhydrotetracycline), or either a heterogenous expression (with only a small proportion of the cells actually expressing Cas9). However, being already satisfied with our pBAD system, we decided to leave the pTet system aside.

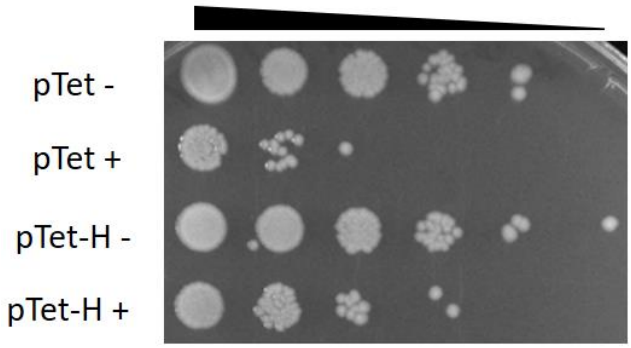

Figure 3.2. pTet promoter is also active, but with a lower range of expression. Cells grown in the presence of the *ptrA*-gRNA and expressing either Cas9 (pTet) or Cas9-HaloTag (pTet-H) under the control of the pTet promoter have been exposed to either 1µM of anhydrotetracycline (+) or nothing (-) for 1h. We see that with this construct, Cas9 is active and can be effectively induced or repressed. However, the pTet promoter does not offer a range of expression as suitable as pBAD does (see Figure 3.1).

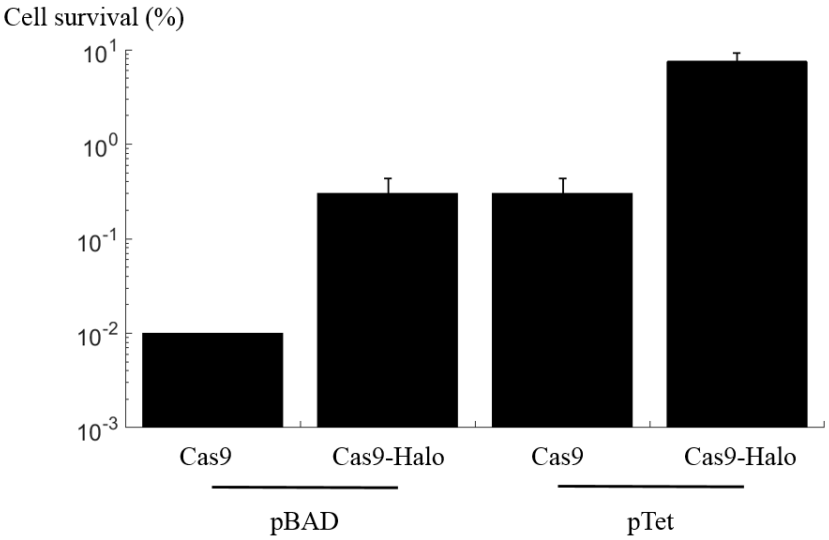

Figure 3.3. pBAD promoter expression is more suitable than pTet's. The results of the previous spot tests are displayed here with their averaged survival rates (expressed as a percentage of cells surviving the fully-induced CRISPR construct, normalized by the amount of cells growing when the system is not induced, in log scale). Standard errors are displayed. With pBAD expression, cells die more than with pTet.

3.2.2 – Cas9-HaloTag is as active as Cas9 (quantitatively)

After we successfully confirmed the activities of both our Cas9 and Cas9-HaloTag constructs, as well as the reliable control we could achieve on their respective expression and activity, we decided to further characterize them by pursuing a more quantitative approach, using single-cell measurements. To do so, we used our SOS-response reporter gene, namely, a GFP under the control of the pSfiA promoter, itself activated by the SOS response. Cells containing the pBAD-Cas9 construct supplemented with the *ptrA*-gRNA on a plasmid were therefore exposed to pulses of either arabinose, glucose or nothing and their fluorescent signal was monitored on the microscope.

In the first experiment we performed, we observed a significant shift in GFP signal (*ie*, SOS response) intensity between the repressed (and leaky) expression conditions and the arabinose-induced one (see Figure 3.4, low dilution). Confirming our previous observations, our CRISPR system did trigger DSBs and subsequent SOS response. However, it was also noted that in the induced condition, a remaining proportion of cells was still displaying very low GFP intensity.

This observation could be interpreted in two ways; either those cells had fail to express Cas9 (for a variety of reasons, ranging from a mutation in the CRISPR system to a lack of access to the arabinose input) or they were either expressing Cas9, but too late in regard to the microscopy acquisition, again for a variety of reasons. This second hypothesis stems from the fact that Cas9 may have taken too much time between its expression, its target recognition and its cutting activity, or perhaps more simply that cells may have started their metabolic activity later than other neighbouring cells.

Indeed, the state of balanced growth is obtained when all the cells in a population have reached an identical metabolic rate. Before this state, the metabolic rates are heterogeneous and therefore lead to some exponentially-growing cells (expressing proteins) and others, still in a state of dormancy (metabolically inactive). As a result, the first would display high SOS intensity while the second would keep it low.

To test this last hypothesis, we decided to increase the initial dilution of the overnight inoculum from 1/250 to 1/1000, as we reasoned that allowing more cell generations to reach the final OD_{600nm} would allow more time for the cells to reach balanced growth. And indeed, we observed a disappearance of those previously 'not-induced' cells and managed to reach a more homogeneous distribution of GFP signal, direct measurement of Cas9 activity (see Figure 3.4, high dilution).

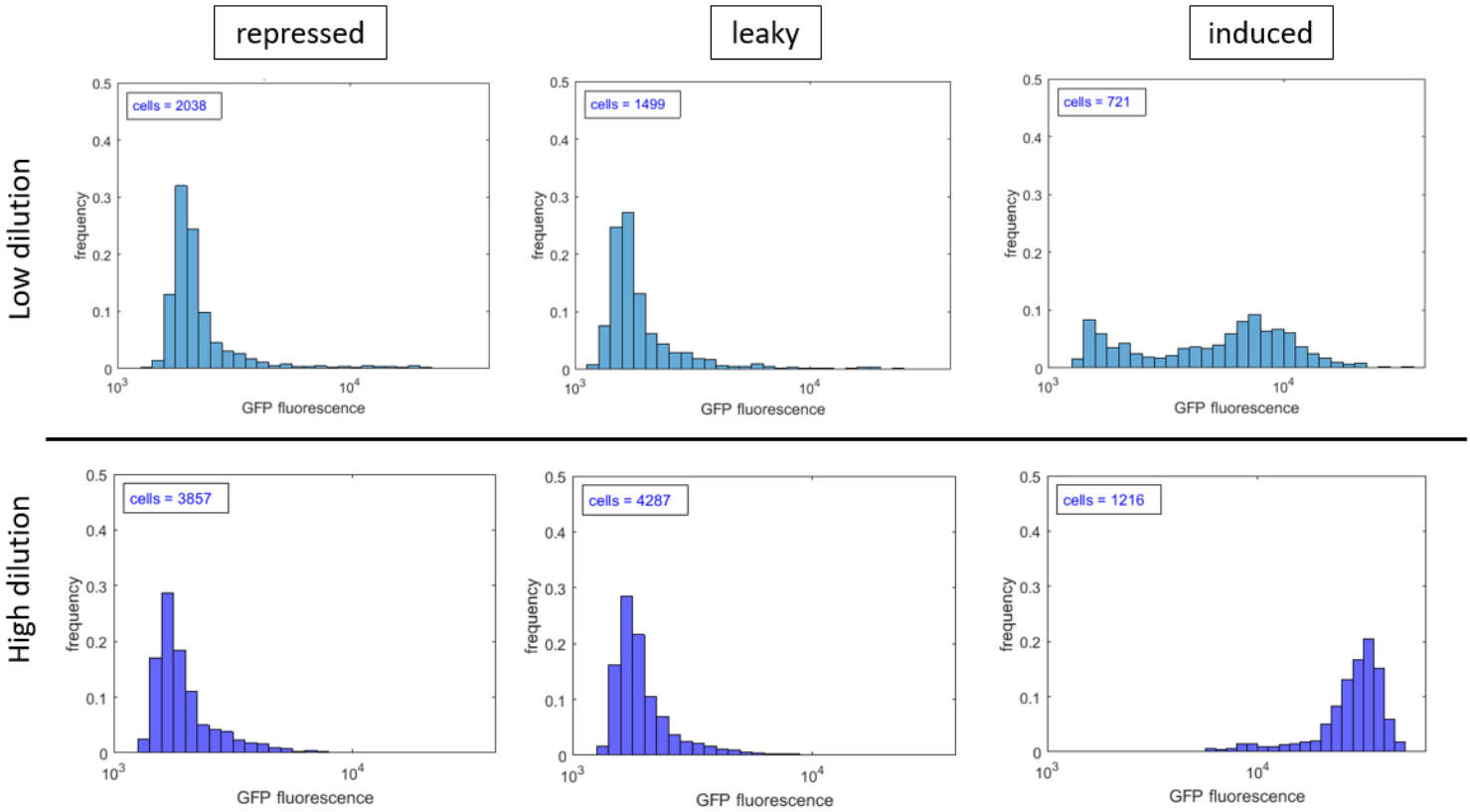

Figure 3.4. In balanced growth, our CRISPR system can be successfully expressed or repressed. SOS response distributions of cells containing pBAD-Cas9 and the *ptrA*-gRNA, exposed for 1h to either 0.2% glucose (repressed), 0.2% arabinose (induced) or nothing (leaky). On the top row, the initial overnight dilution was of 1 in 250 (low dilution), whereas on the bottom row, the dilution was of 1 in 1000 (high dilution, closer to balanced growth). The SOS response is plotted as the average GFP fluorescence intensity per cell, per pixel. The total numbers of cells are indicated.

At this point, we repeated this experiment with our CRISPR-HaloTag system. Figure 3.5 shows how Cas9-HaloTag cells reproducibly displayed significant SOS activity difference depending on the presence or absence of gRNA, following the wild-type Cas9 phenotype.

A longer tail of low SOS-activity cells with the Cas9-HaloTag construct could be observed, which was attributed to a particularly heterogeneous replicate (see Figure 3.6). Following the same previous reasoning, we concluded that this particular dataset was brought to the microscope a little too early, depriving Cas9 from enough time to reach its target, cut and trigger the SOS response.

Nevertheless, we can confidently conclude that the HaloTag fusion does not hinder Cas9 activity, which we can regulate, and that we can now proceed with the study of HaloTag-TMR fluorescent signal.

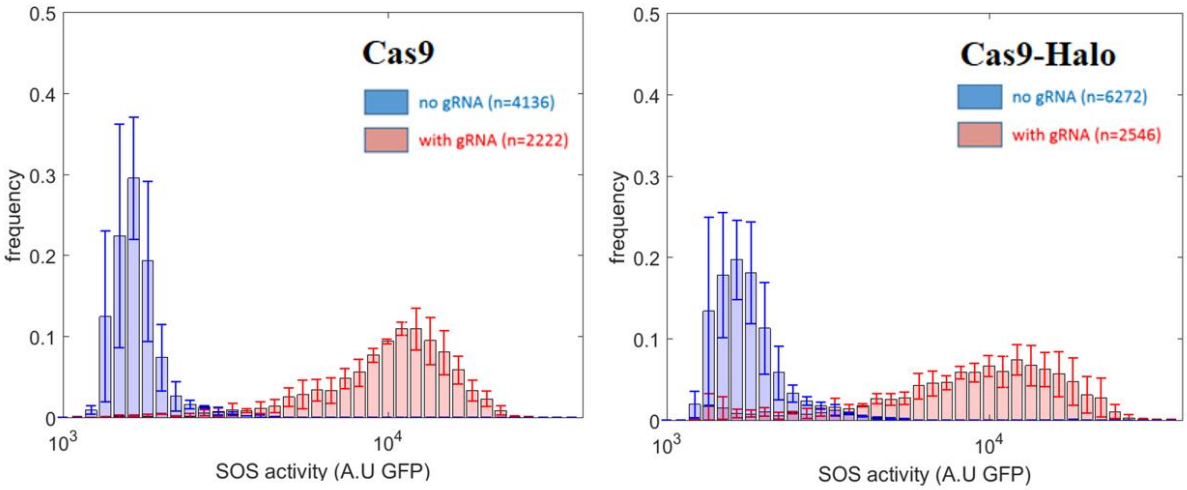

Figure 3.5. Our CRISPR constructs induce SOS. Following the same protocol used for spot tests, cells over-expressing either Cas9 (left) or Cas9-HaloTag (right) were grown either in the presence (red) or absence (blue) of the *ptrA*-gRNA. Their SOS response was monitored through our GFP reporter gene. Here are shown distributions of average GFP intensity per cell, per pixel. Total numbers of cells (n) and standard errors (from duplicates for Cas9 and triplicates for Cas9-HaloTag) are indicated.

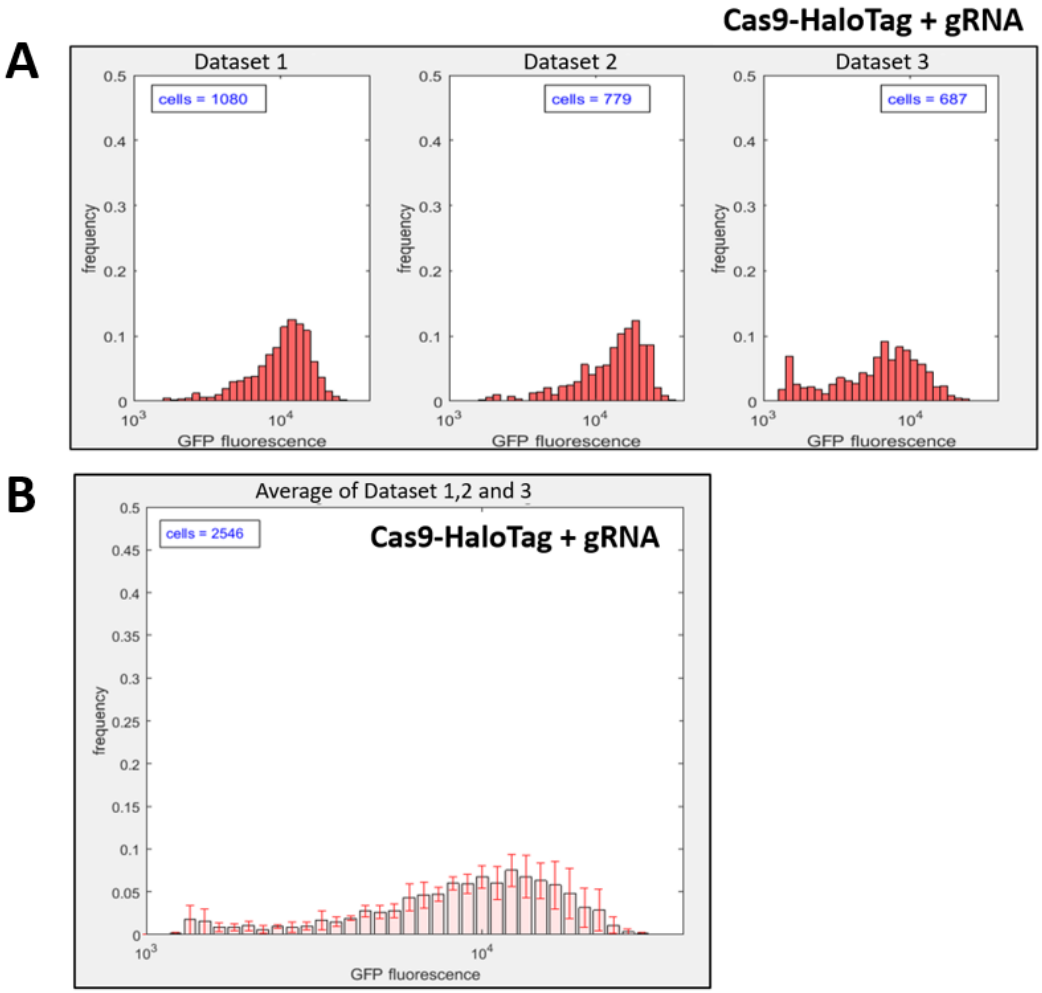

Figure 3.6. One of Cas9-HaloTag triplicates shows a second population. s (**A**) Three replicate datasets are shown monitoring Cas9-HaloTag SOS response, after Cas9 induction, in the presence of the *ptrA*-gRNA. (**B**) Here is displayed the averaged distributions of all three previous datasets, with associated standard errors. We can observe that the third dataset displays an additional peak of cells with low SOS response. Total numbers of cells are indicated.

3.2.3 – Cas9-HaloTag signal is specific

After confirmation of the cutting activity of our CRISPR-HaloTag system, we proceeded to look at the fluorescent signal we could obtain from the HaloTag-TMR complex. As a brief reminder, the TMR molecule is a bright and stable fluorophore that has the capacity to covalently bind the HaloTag protein. However, it needs to be added externally, usually in excess, which requires further washing steps in order to remove the unbound TMR molecules.

As opposed to the previous SOS measurements, based on several thousands of GFP per cell, we were interested in expressing Cas9-HaloTag in relatively low copy numbers; barely 10 molecules per cell, in order to be as close as possible to the single-particle tracking conditions. This fact is worth noting, as cells tend to display autofluorescence, which despite its lower level in the red wavelength (TMR) than in the green wavelength (GFP), can still affect our ability to confidently record HaloTag-TMR signal.
To confirm the specificity of the TMR fluorescent signal to our Cas9-HaloTag protein as well as our ability to detect it, we compared the fluorescence displayed by cells expressing the HaloTag fusion to that of wild-type cells, both following induction with arabinose, TMR labelling and chemical fixation. Using the analysis pipeline previously described to quantify the SOS response (with Brightfield segmentation and quantification of average GFP fluorescence intensity per pixel, per cell), we quantified the fluorescent signal obtained in the TMR emission wavelength.

Our first attempts revealed that HaloTag cells indeed displayed higher TMR signal than wild-type ones, albeit with a rather low signal-to-noise ratio, as can be seen on the example shown in Figure 3.7. This could be explained by a variety of factors, including the accuracy of the focus, the level of cleanliness of the coverslip and the quality of the TMR batch. That said, after optimization of our imaging conditions (use of a 400°C kiln followed by overnight wash in KOH 1M to clean the coverslips; use of a pump to wash the excess of TMR, instead of a pipette) and scrupulous validation of our successive TMR samples (consistently using negative controls without HaloTag fusion to check the level of unbound TMR molecules), we managed to reach levels of signal-to-noise ratios high enough to both confirm the specificity of our HaloTag-TMR construct as well as our ability to confidently record it, as illustrated in Figure 3.8.

Finally, despite this successful qualitative confirmation, this analysis does not permit to quantitatively discuss the expression of our Cas9 protein. We will therefore rely on actual protein counting to obtain absolute numbers of Cas9 proteins per cell, in various expression conditions.

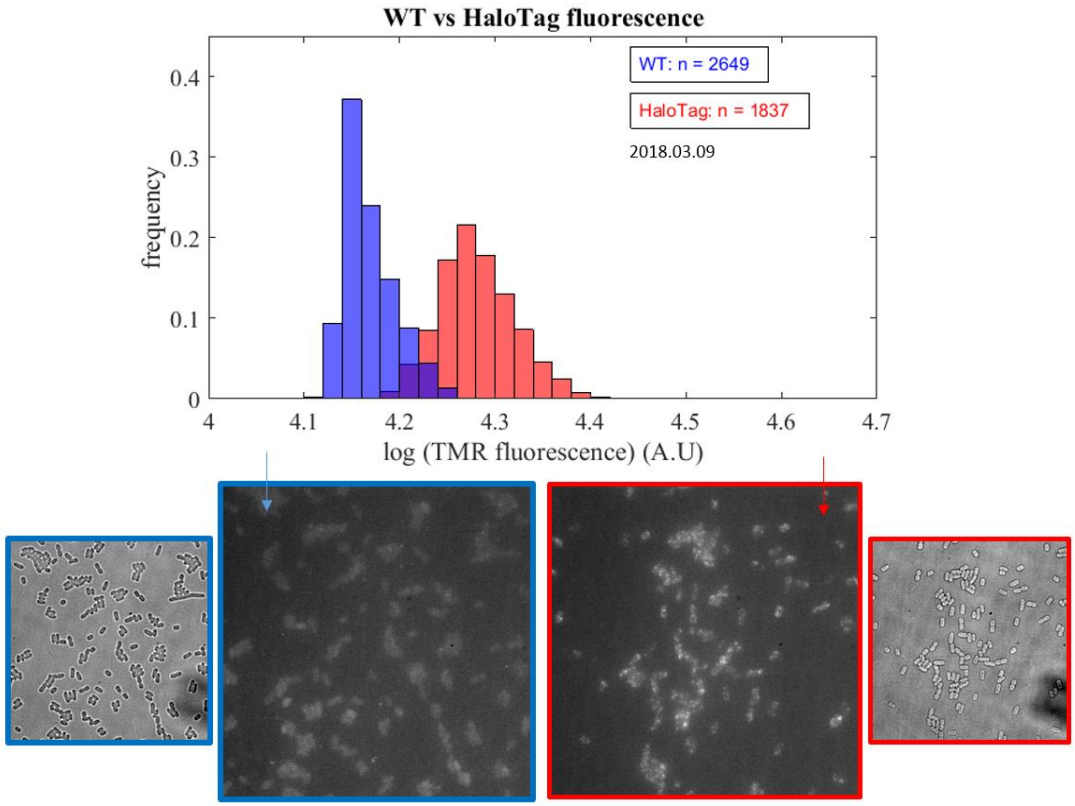

Figure 3.7. In poor imaging conditions, the HaloTag-TMR signal is not specific enough. Cells containing the Cas9-HaloTag fusion (HaloTag, in red) and cells only expressing Cas9 (WT, in blue) were exposed to 0.2% arabinose for 1h, then labelled with excess of TMR, washed and fixed with formaldehyde, before being mounted on agarose-pads to be observed on the microscope. **Top**. The average TMR fluorescent intensity per pixel per cell is displayed. **Bottom**. Associated are images taken in the TMR channel (maximum Z projection, with the same contrast) and in the Brightfield channel. Total numbers of cells (n) are indicated. The two distributions overlap, indicating a low signal-to-noise ratio.

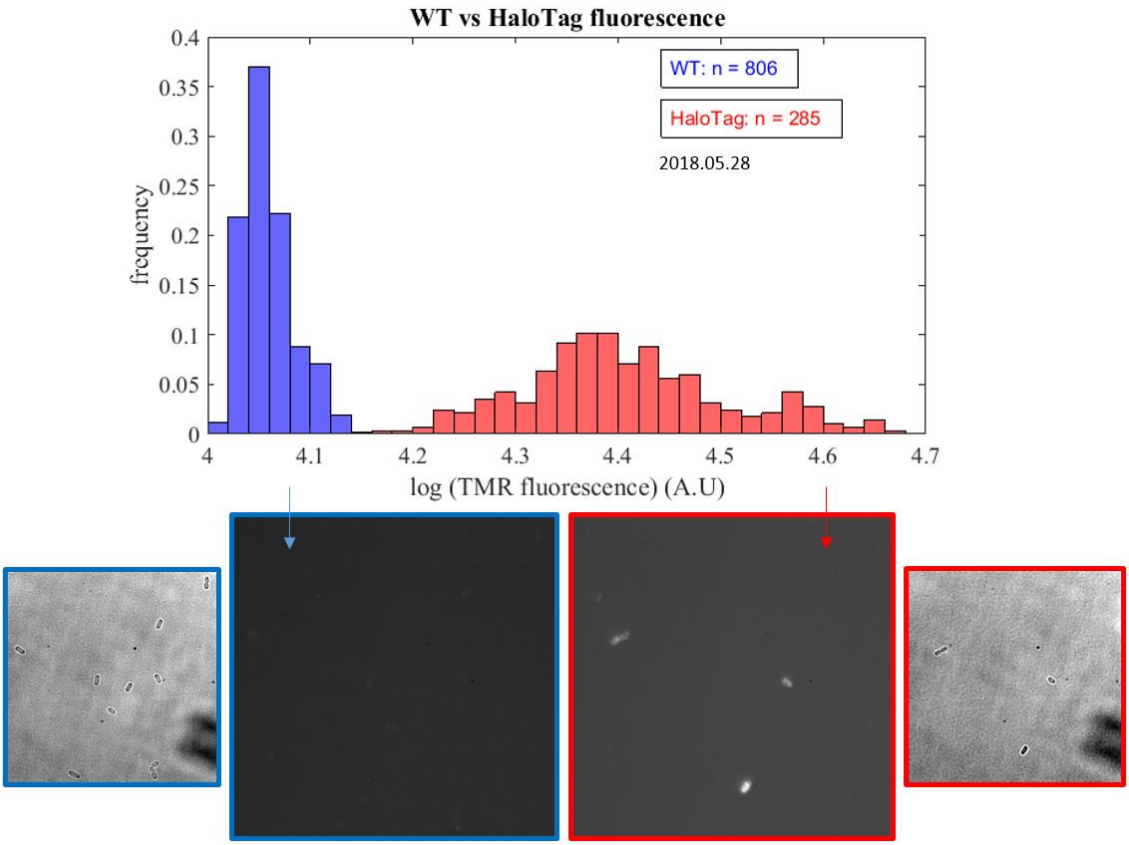

Figure 3.8. After optimization, HaloTag-TMR gives specific signal. Here is a different dataset showing cells expressing the same constructs and exposed to the same conditions as presented in the precedent figure. **Top**. The average TMR fluorescent intensity per pixel per cell is displayed (log scale). **Bottom**. Associated images taken in the TMR channel (maximum Z projection, with the same contrast) and in the Brightfield channel. Total numbers of cells (n) are indicated. Here, the signal-to-noise ratio is much better and the signal specificity confirmed.

3.2.4 – Cas9-HaloTag expression

Now that the specificity of our fluorescent signal was confirmed and in order to get a better characterization of the expression level of our araBAD constructs, I further analysed the previously mentioned datasets (Cas9 with and without HaloTag fusion, and with and without gRNA) and proceeded to count the number of Cas9-HaloTag proteins per cell. Importantly, my lab previously reported[99] that this technique yielded a detection efficiency of labelled proteins superior to 80%, thus adding confidence to our protein quantification.

To count the number of fluorescent proteins per cell, I used a custom-made algorithm which first relied on Brightfield images for cell segmentation, and which then produced Z-maximum-projections (for every field of view, images were taken with varying heights, so called 'Z-stacks', which maximum intensity per pixel were projected onto one resulting image called 'maximum projection'). Those maximum projections were essential to cover the entirety of the cellular content and thus ensure that we would detect fluorescent particles present on different focal planes. The resulting images were filtered to remove much of the fluorescent background noise and then run through another algorithm aimed at identifying fluorescent spots (more details are in Materials and Methods, Chapter II).

We therefore observed that in those conditions (0.2% glucose media, 1h exposure to 0.2% arabinose), cells displayed an average of 2.93 Cas9-HaloTag proteins, which conveniently matched the low number of proteins necessary for Single-Particle Tracking (Figure 3.9). We also observed a homogeneous protein expression, with only a few cells (~15%) exhibiting higher amounts of proteins.
In addition, I looked at the same construct, but in the absence of arabinose and therefore in the absence of araBAD induction, resulting in only twice less Cas9-HaloTag proteins per cell (1.52 spots/cell, Figure 3.10). This surprisingly small discrepancy can be explained by the constant presence of glucose in the media which applied its inhibitory activity and therefore limited the arabinose induction. Interestingly, we could observe that in repressed conditions, 80% of the cells still expressed proteins, albeit in very low amounts.

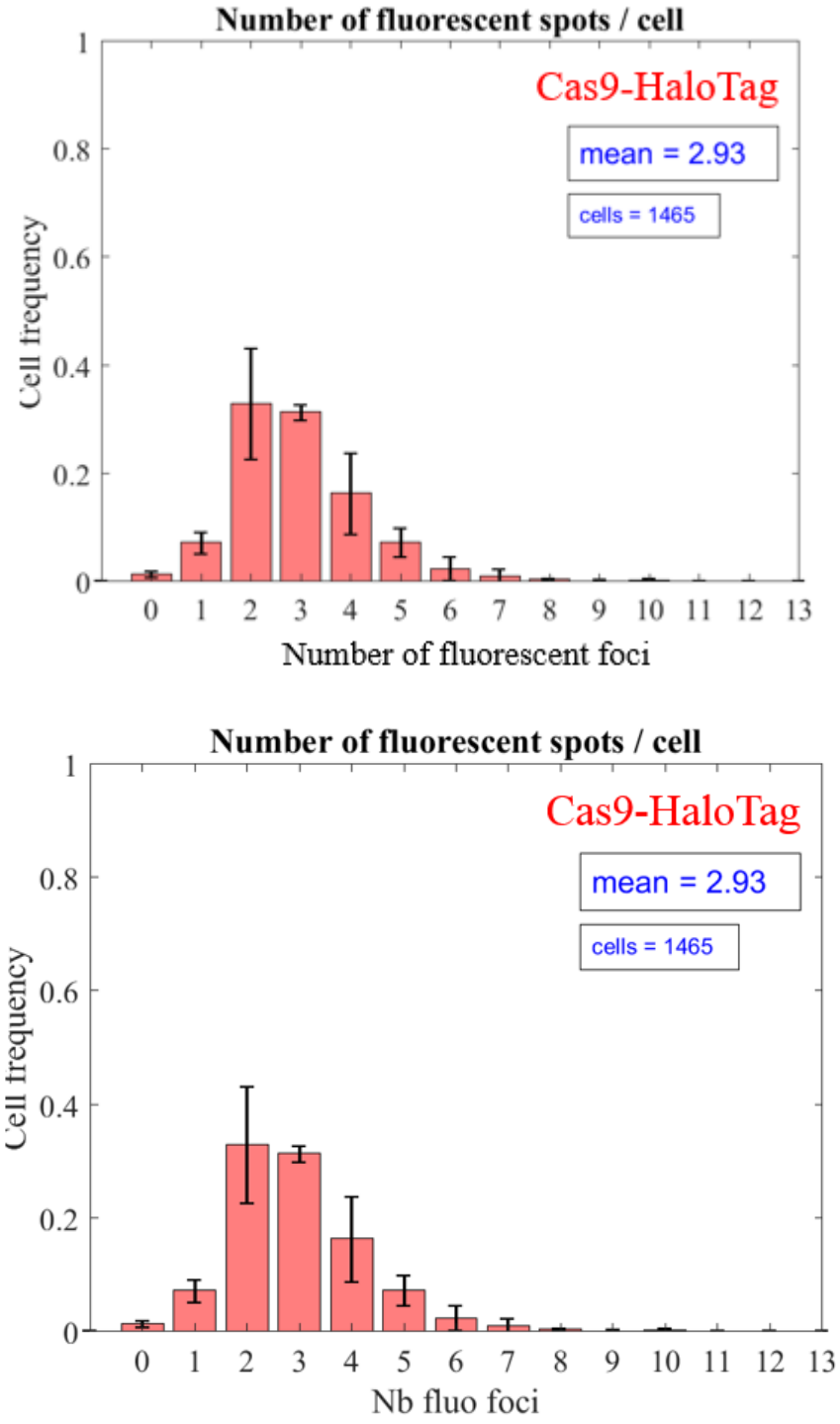

Figure 3.9. In our expression conditions, Cas9-HaloTag is expressed at a suitable amount for single-particle tracking. Cells containing the Cas9-HaloTag construct (under the control of the araBAD promoter) were grown in the presence of 0.2% glucose and then exposed to 0.2% arabinose for 1h (during the TMR labelling), chemically-fixed, washed and finally mounted on agar-pads to be observed on the microscope. Both Brightfield and TMR signals were recorded, the first for cell segmentation and the second to count protein numbers. Total numbers of cells and standard errors (from duplicate experiments) are indicated. The average number of fluorescent foci per cell amounts to 2.93, which is appropriate for single-particle tracking.

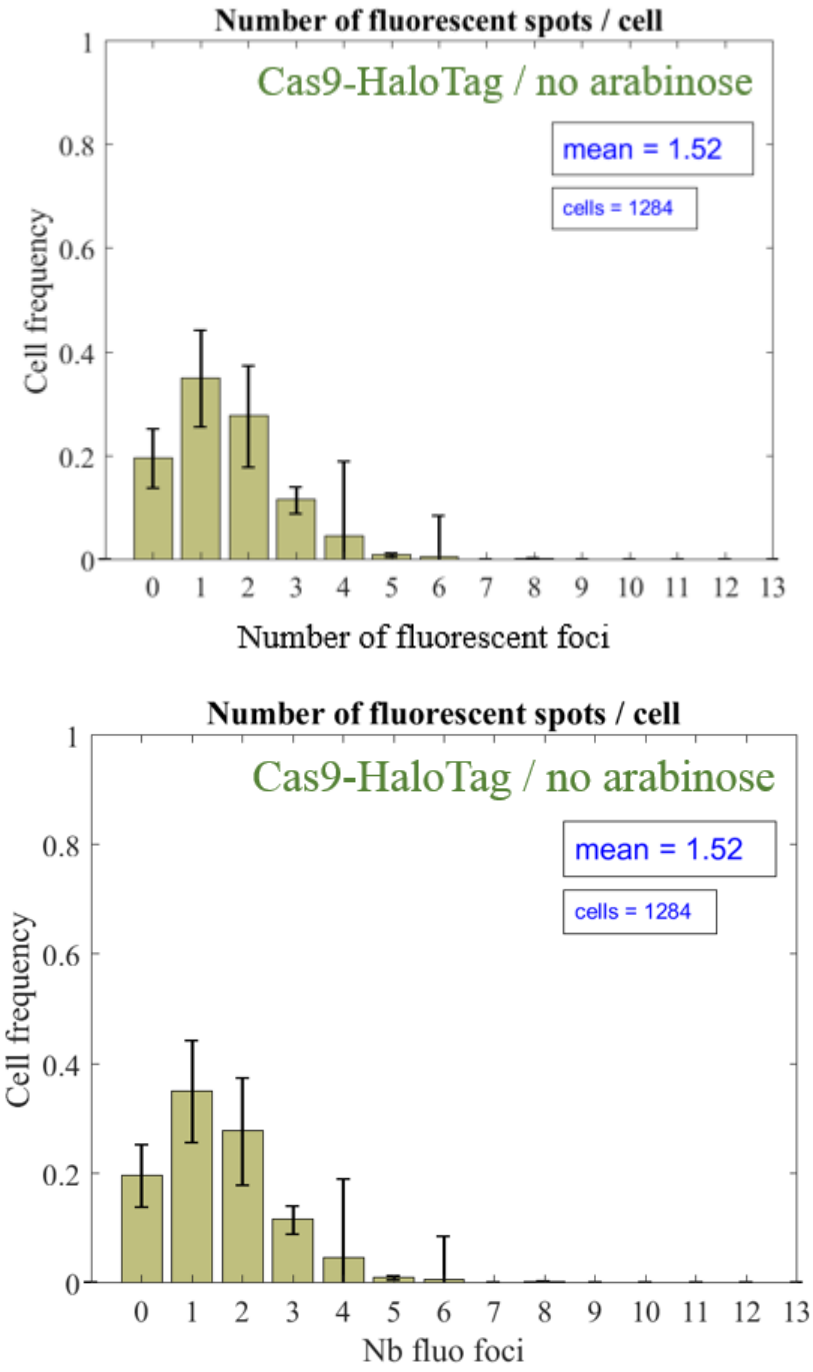

Figure 3.10. Without arabinose, Cas9-HaloTag expression is even lower. Cells containing the Cas9-HaloTag construct were grown in a media containing 0.2% glucose but without addition of any arabinose. They were then labelled, chemically-fixed, washed and finally mounted on agar-pads to be observed on the microscope. Both Brightfield and TMR signals were recorded. Total numbers of cells and standard errors (from duplicate experiments) are indicated. As expected, we observe a diminution of the number of fluorescent foci per cell, compared to cells exposed to arabinose induction.

We repeated this experiment in the presence of a gRNA (targeting *ptrA*) and observed an increase in the number of fluorescent foci per cell (4.01 spots/cell, Figure 3.11). However, this increase does not appear to originate from a change in protein expression, but rather from the elongation of the cells exposed to CRISPR-Cas9 activity which produces DNA double-strand breaks, induces SOS response and subsequent cell elongation, as can be seen on Figure 3.12.

Indeed, elongated cells do not divide and thus do not dilute their labelled proteins by transferring them to daughter cells, effectively displaying more of them. Consistent with that hypothesis, after normalizing the number of fluorescent foci per cell with the cell length, we observe no significant difference between the samples with and without gRNA (Figure 3.13).

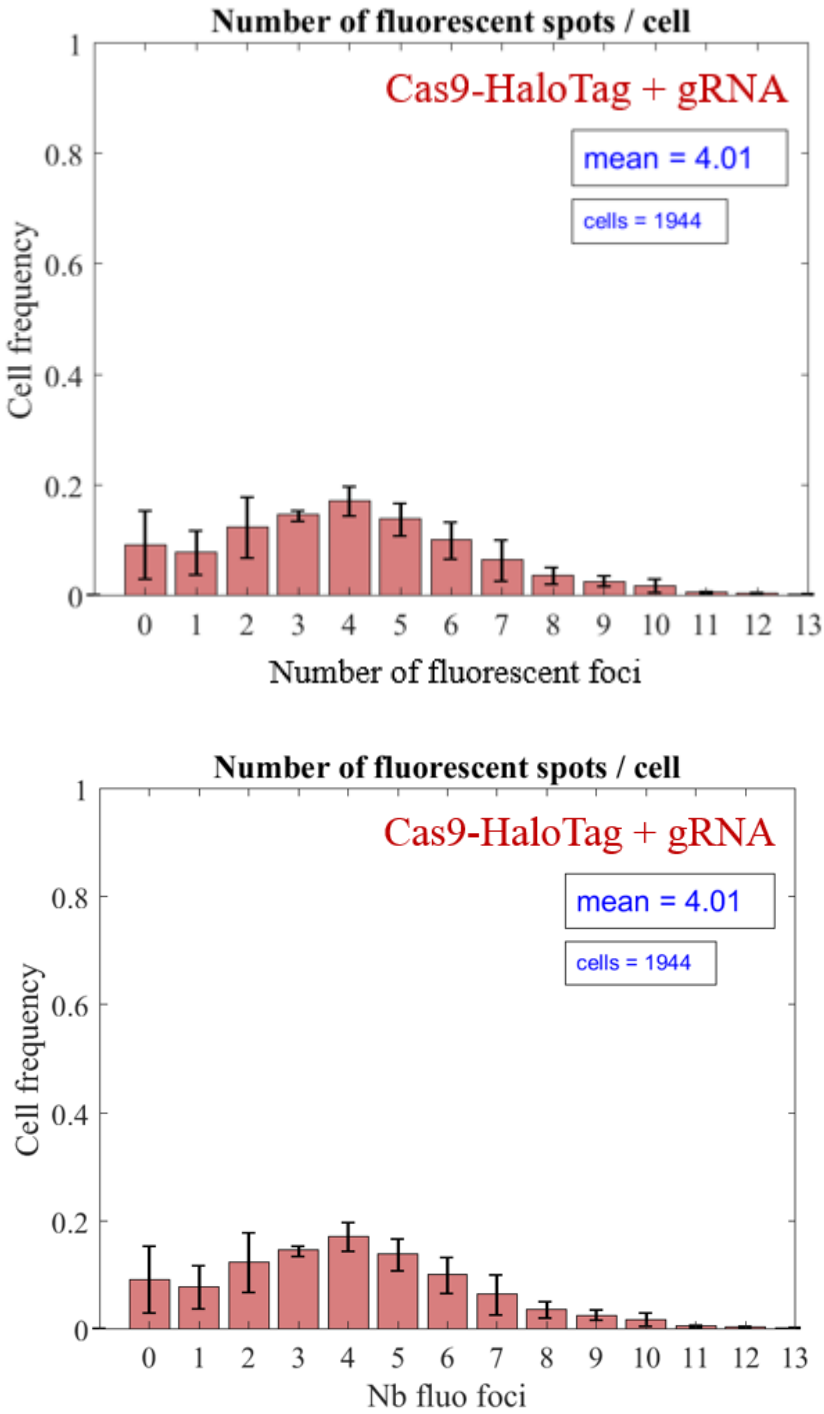

Figure 3.11. In the presence of a gRNA, the number of detected Cas9-HaloTag per cell increases. Cells containing the Cas9-HaloTag construct and the *ptrA*-gRNA were exposed to the same conditions stated above (Figure 3.9). Total numbers of cells and standard errors (from duplicate experiments) are indicated. The average number of fluorescent foci per cell amounts to 4.01, which is higher than what has been observed in cells containing Cas9-HaloTag but without gRNA (2.93 foci/cell).

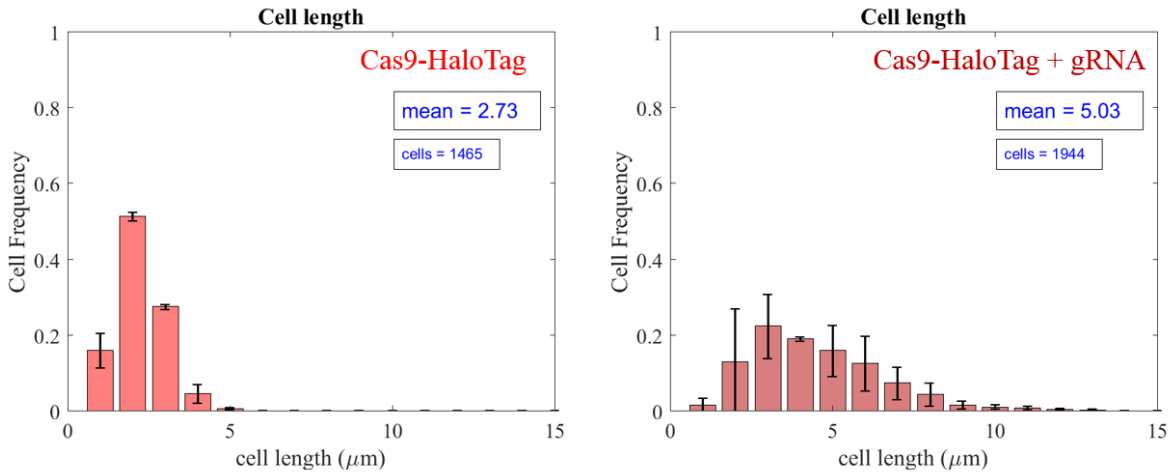

Figure 3.12. In the presence of a gRNA, cells elongate due to the activation of the SOS response. Cells containing the Cas9-HaloTag construct with (right) or without (left) the *ptrA*-gRNA were exposed to the same conditions stated above (Figure 3.9). Total numbers of cells and standard errors (from duplicate experiments) are indicated. With the gRNA, cells experience DNA damage and subsequent SOS response activation which make them elongate.

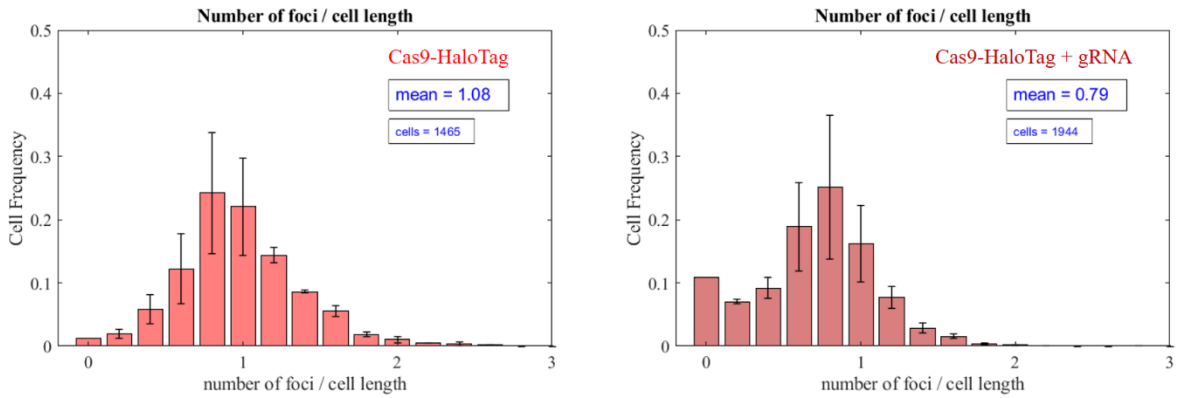

Figure 3.13. Normalized by cell length, the presence of a gRNA does not seem to affect the number of detected proteins per cell. Cells containing the Cas9-HaloTag construct with (right) or without (left) the *ptrA*-gRNA were exposed to the same conditions stated above (Figure 3.9). Total numbers of cells and standard errors (from duplicate experiments) are indicated. With the gRNA, cells display slightly less fluorescent foci than without, probably due to very long cells.

In order to quantify the specificity of our signal, I performed the same experiment with cells lacking the HaloTag fusion and counted an average of 0.69 fluorescent spots/cell, which is fairly high (see Figure 3.14). When comparing with cells not exposed to TMR at all, I only detected 0.08 fluorescent spots/cell (Figure 3.15), indicating that (i) the autofluorescent background was very low and that (ii) our coverslips were very clean. Put together, these results indicate that after TMR labelling and subsequent washes, a non-negligible proportion of unbound – or free – TMR molecules most likely stays within the cells. This could be addressed by increasing the number of washes, which would however decrease the number of cells available for microscopy.

In the context of absolute protein quantification, this issue would be slightly inconvenient, but our aim is different; it is to characterize the optimal expression regime for Single-Particle Tracking, which fits our present conditions. Furthermore, if those unbound TMR molecules can be observed with chemical fixation, they would presumably disappear in live cell imaging, due to their very small size, very high diffusion and subsequent lack of concentrated signal.

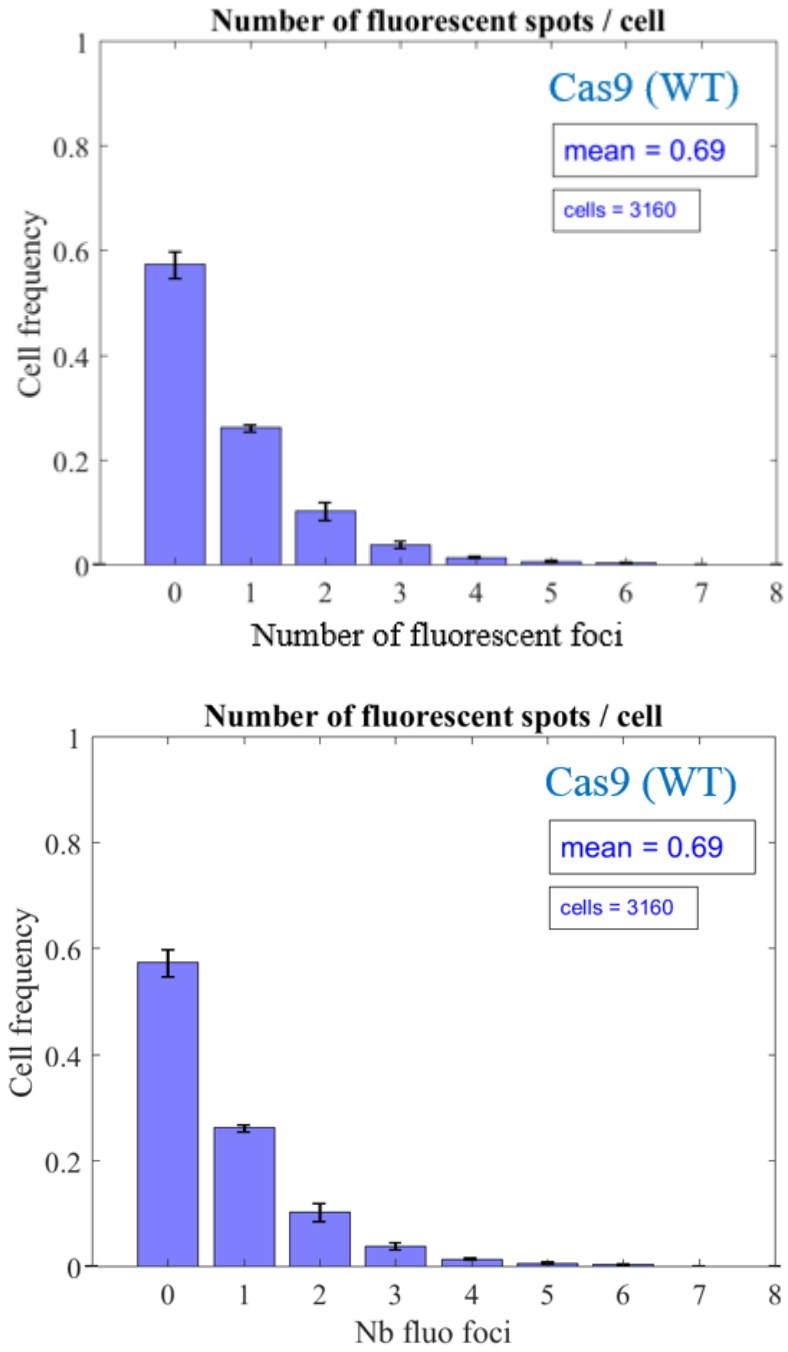

Figure 3.14. Without the HaloTag fusion, very little fluorescent foci are visible. Cells containing the Cas9 (WT) construct were exposed to the same conditions stated above (Figure 3.9). Total numbers of cells and standard errors (from duplicate experiments) are indicated. In the absence of HaloTag fusion, we detect much less fluorescent foci, indicating that our signal is specific and our analysis reliable.

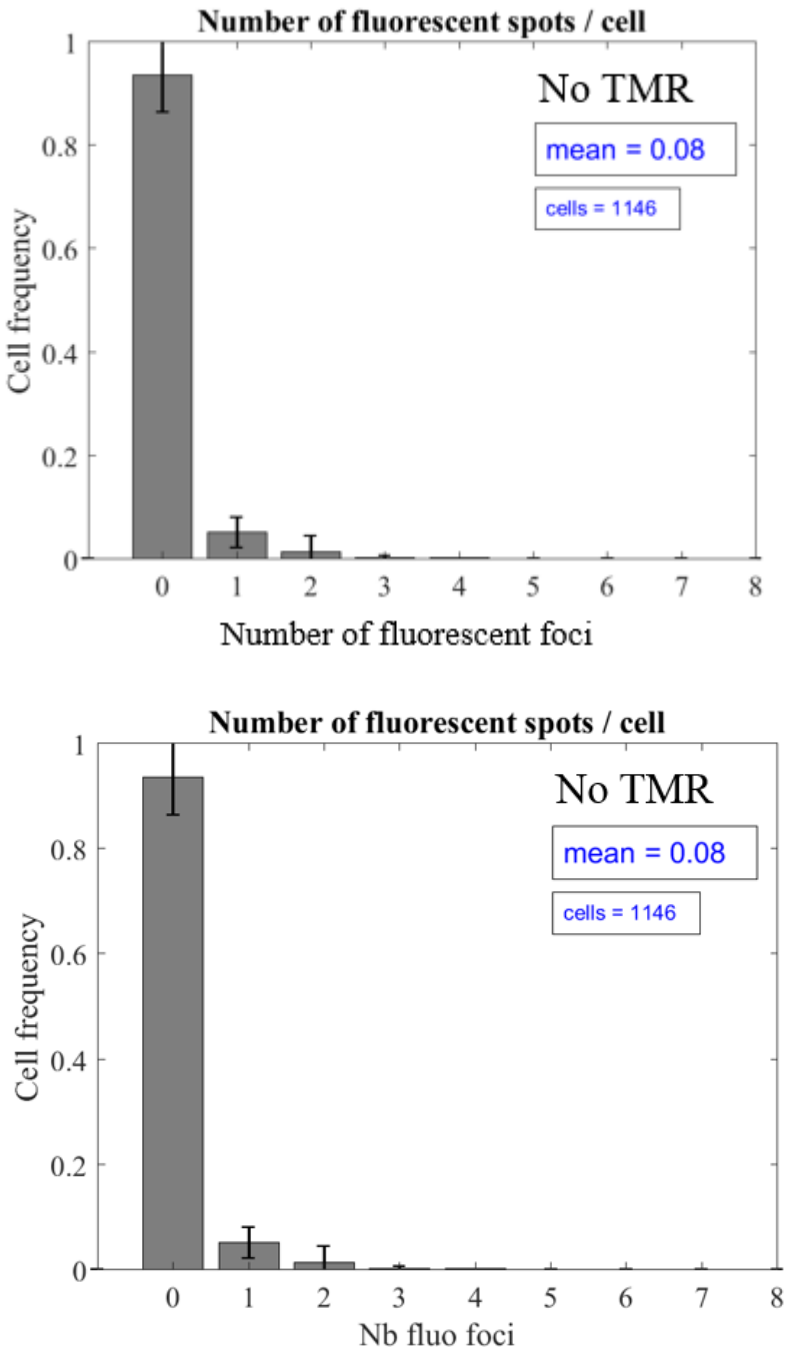

Figure 3.15. Without TMR, no fluorescence is observed. Cells containing the Cas9 (WT) construct were exposed to the same conditions stated above (Figure 3.9). Total numbers of cells and standard errors (from duplicate experiments) are indicated. In the absence of TMR (fluorophore), there is no detected fluorescence, indicating that our coverslips are clean.

Finally, comparing the maximum fluorescent intensity of detected spots, with and without HaloTag, we observe much brighter spots with the protein fusion (Figure 3.16), indicating that the spots detected without it are actually dim artefacts. Consequently, an appropriate intensity threshold would easily differentiate between the two; technique that we will apply to the following tracking analyses.

Overall, we were satisfied with our expression conditions and ready to proceed with the main experiment of this chapter; Single-Particle Tracking.

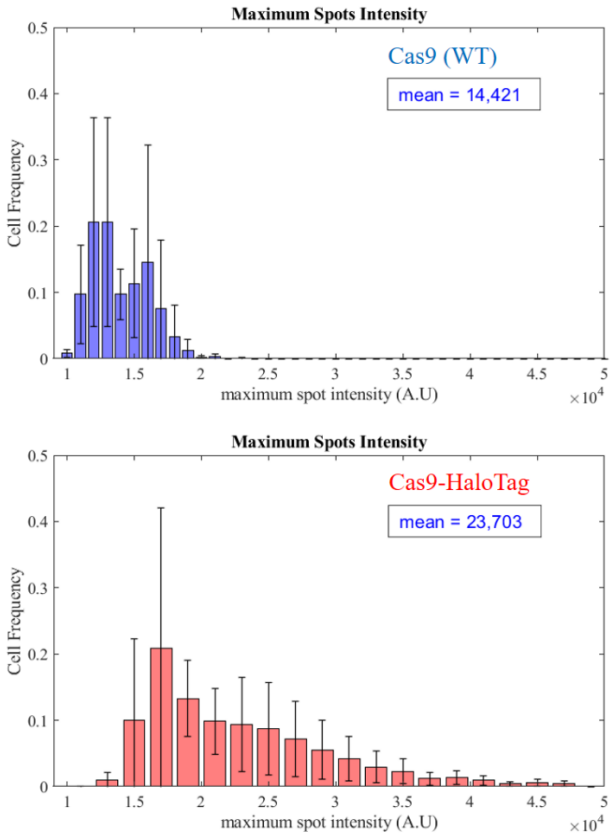

Figure 3.16. With HaloTag, detected spots have a higher fluorescent intensity. Cells containing the Cas9 (WT) or Cas9-HaloTag (HaloTag) construct were exposed to the same conditions stated above (Figure 3.9). Distributions of maximum spot intensities. Mean and standard errors (from duplicate experiments) are indicated. In the presence of HaloTag proteins, the detected spots display higher intensity than without, indicating that a proper threshold could easily differentiate between the two.

3.3 – Cas9-HaloTag tracking with one genomic target

3.3.1 – HiLo quality for tracking

In the previous parts of this chapter, we confirmed the activity of our CRISPR-Cas9 constructs as well as the specificity of our HaloTag fluorescent signal. We now wish to determine whether Single-Particle Tracking is applicable to our system.

As opposed to our previous microscopy experiments which either dealt with very high numbers of fluorescent proteins (pSOS-GFP) or either chemically-fixed molecules (fixed Cas9-HaloTag), Single-Particle Tracking aims to observe very small numbers of proteins in living cells (so without fixation), in real time (with a resolution of dozens of milliseconds, compared to previous 200 ms), therefore relying on much less photons to detect. This creates two technical challenges; the first is to produce a high enough and stable enough fluorescent signal which can be detected with such short exposure times, and which will not photobleach throughout the acquisition. The second is to sufficiently decrease the fluorescent background of the sample in order to highlight the signal of interest.

The HaloTag-TMR complex does provide a strong and photostable signal. However, in those conditions it needs to be even stronger, and above all, without fluorescent background. Those two challenges can be addressed with one technique; HiLo microscopy.

As a reminder, HiLo stands for Highly Inclined Laminated Optical sheets, a technique derived from TIRF microscopy which also relies on laser illumination[96]. A laser triggers much stronger fluorescent excitation than traditional light sources, thus strongly uplifting the number of emitted photons.
Additionally, both techniques dramatically decrease fluorescent background. TIRF microscopy relies on total reflection of the laser at the interface with the sample, in turn producing a so-called ‘evanescent’ wave which illuminates the sample on a scale of several nanometres. In contrast, the HiLo technique uses a laser beam which is not reflected but refracted throughout the sample and illuminates only part of it, therefore limiting the fluorescent background to the area of interest (with a depth of 1 µm), while ensuring that the whole content of an *E.coli* cell can be illuminated. For more details, please refer to the Materials and Methods.

Finally, a more photostable version of the HaloTag ligand (TMR) was used; JF549 (see Material and Methods for more details). Throughout our experiments, a partner laboratory worked with us to provide the new ligand, allowing us to improve the quality of our movies. Added in the same concentration and for the same time period as TMR, our tracking experiments remained identical. From this point on, every experiment was carried on with JF549 instead of TMR, but to facilitate the reading, we chose to keep the same "TMR" label.

Therefore, following the same concentration and time periods as in the previous "protein counting" experiments (to ensure the results between the two types of experiments would remain consistent), I collected movies (100 frames, 20 ms exposure time) capturing the HaloTag signal emitted from our cells.

As can be observed in Figure 3.17, the resulting images we obtained with the HaloTag construct clearly displayed fluorescent signal. With maximum projections executed on all the frames of the movies (here, 100), even some foci can be seen, and the background looks sufficiently dark to indicate a good signal-to-noise ratio. Moreover, we can compare this result to the wild-type dataset which further confirms that the signal is specific to the HaloTag-TMR complex, for without it, there is no fluorescent signal to be seen.

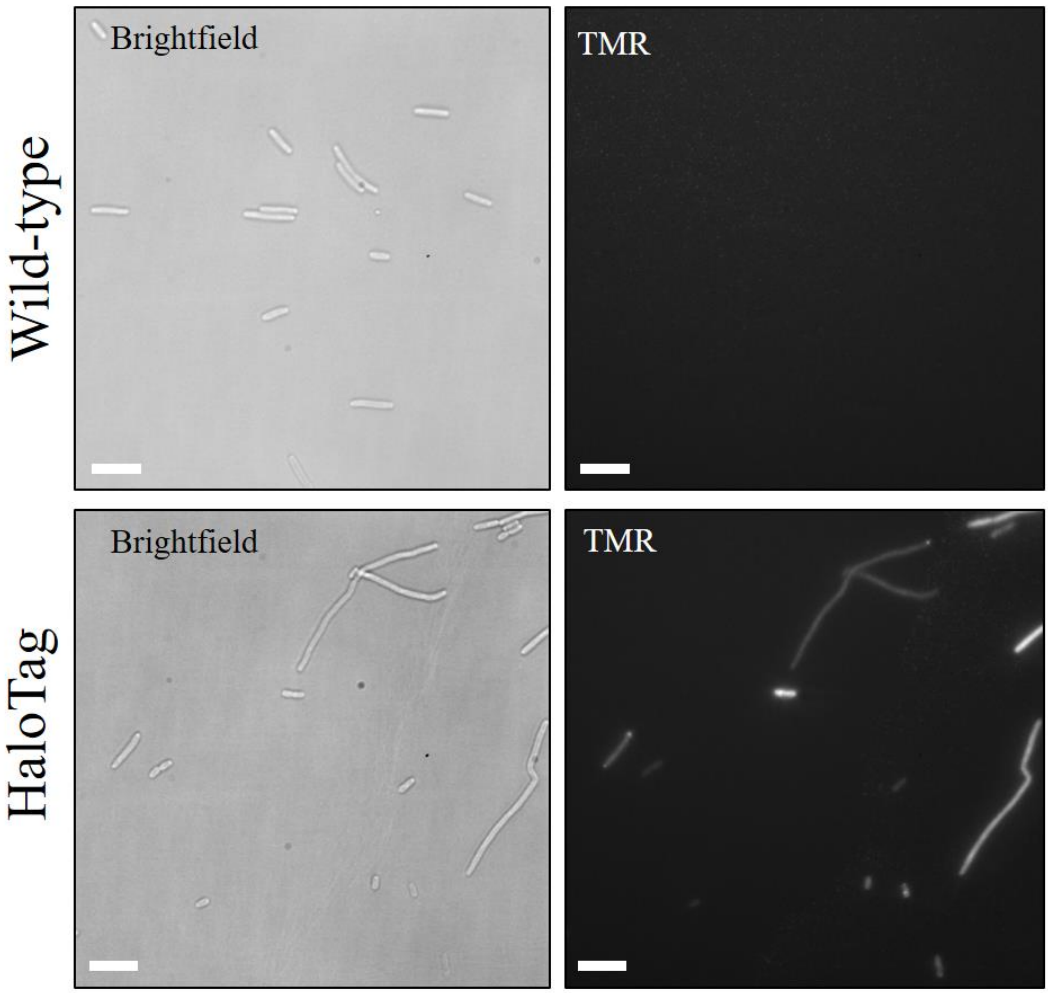

Figure 3.17. With HiLo microscopy, the HaloTag-TMR signal is visible and specific. (**Top**) Wild-type cells and (**Bottom**) cells containing the pBAD-Cas9-HaloTag construct were grown in the presence of 0.2% glucose, labelled and simultaneously exposed to 0.2% arabinose for 1h, washed and finally mounted on agar-pads to be observed through HiLo microscopy. In the TMR channel (laser excitation at 561 nm), images were taken with an exposure time of 20 ms. Both Brightfield and TMR signals were recorded. The TMR pictures are time-projections (all the frames' maximum intensity pixels were projected on those resulting images). Both datasets TMR images are using the same contrast. The scale bars indicate 5µm. We can detect fluorescent signal from the HaloTag samples, with a good signal-to-noise ratio. In the absence of HaloTag fusion, we do not detect any fluorescent signal in those conditions.

However, an artefact's low signal is not a guarantee of not being captured by a spot-detection algorithm, as was previously shown in Figure 3.14, were even wild-type cells displayed a few detected spots. Consequently, to shield our tracking analysis from false positives, we used intensity threshold parameters to analyse both WT and HaloTag datasets, resulting in not even one track detected among wild-type cells (Figure 3.18). In contrast, the HaloTag-fusion-containing cells did yield an average of 53 detected tracks per cell.

Furthermore, about 87% of those cells did emit enough signal to be detected by our tracking algorithm, compared to 0% without the HaloTag fusion (Figure 3.19). Together, those results indicate that our microscopy setup coupled to our genetic constructs yield enough specific fluorescent signal to perform Single-Particle Tracking experiments.

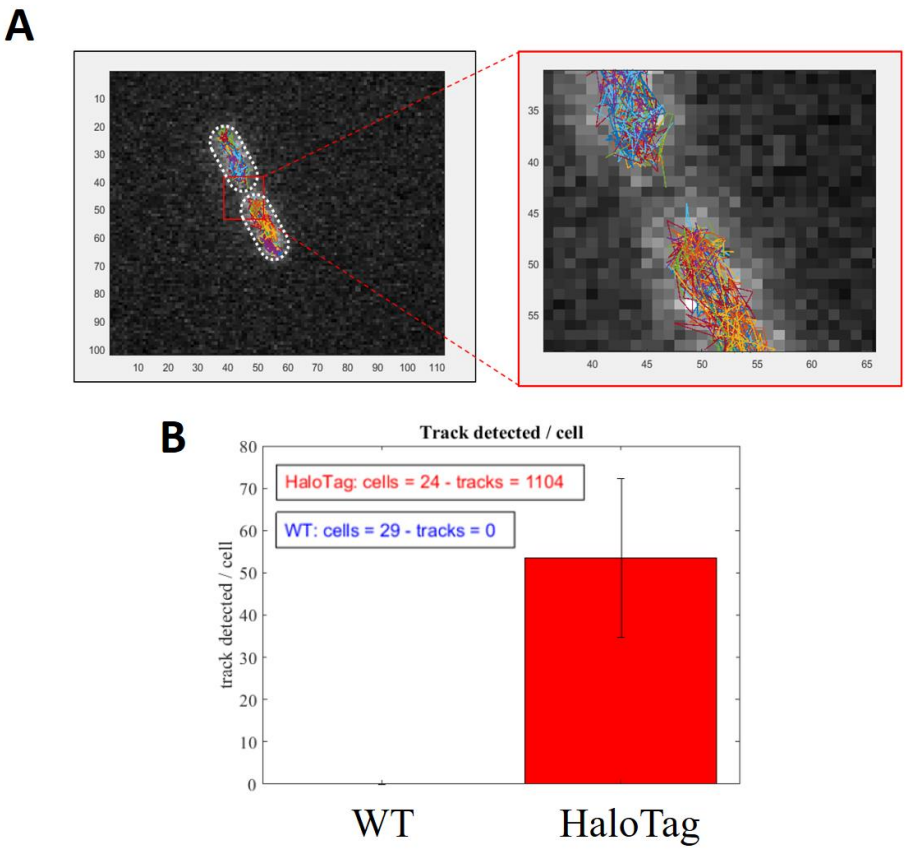

Figure 3.18. Without HaloTag fusion, our algorithm does not detect any track. (**A**) Example of tracks detected from two cells in a HaloTag dataset. Cell borders are highlighted in white. On the right, a zoomed-in image shows that with our tracking parameters, tracks are not detected outside of the cells, illustrating our good tracking resolution. (**B**) Cells containing the wild-type (WT) or the HaloTag construct were grown and prepared in the same conditions as described in Figure 3.17. Both Brightfield and TMR signals were recorded and run through our tracking algorithm. Plotted here is the number of detected tracks per cell. Total numbers of cells, of tracks and standard errors (from duplicate experiments) are indicated. In the absence of HaloTag fusion, the tracking algorithm detects no track. With the HaloTag fusion, we observe an average of 53.4 tracks/cell.

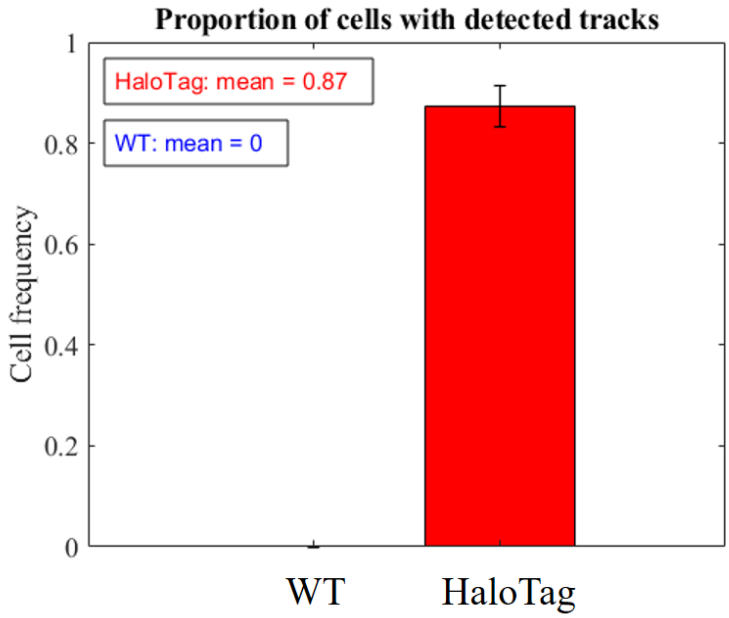

Figure 3.19. With HaloTag fusion, most of the cells display detected tracks. Same cells as in Figure 3.17. Standard errors (from duplicate experiments) and averages are indicated. In the presence of HaloTag fusion, 87% of the cells show enough fluorescent signal to be tracked over time.

3.3.2 – Tracking of HaloTag protein

We have shown that we can reliably detect the fluorescent signal emitted by our HaloTag-TMR constructs and follow their trajectory through Single-Molecule Tracking. Now, in order to acquire a more quantitative understanding of the diffusion of our protein of interest (Cas9-HaloTag), we decided to start by analysing the diffusion of a simpler protein; HaloTag. Indeed, the HaloTag protein expressed on its own is free to diffuse in the cell without any interaction, thus limiting our observations to a single diffusive behaviour.

Therefore, from the trajectories of our HaloTag proteins identified through our tracking algorithm, we computed the respective Mean Square Displacements (MSDs), which have been extensively used in the literature to characterize the diffusion of proteins. Indeed, following the theory developed to study freely-diffusing gas particles, it is possible to apply similar principles to freely-diffusing cytoplasmic proteins, such as our current HaloTag protein, which should not interact with any cellular component.

In theory, freely diffusing particles display a MSD which, as a function of 'lag time' (here proportional to the exposure time, 10 ms), should follow a linear trend. However, we also know that due to spatial constraints generated by the enclosed cellular compartment, the MSD should also display a so-called 'containment' effect and therefore follow a plateau, direct consequence of the protein's incapacity to diffuse more largely in its environment. As an illustration, please refer to Figure 2.19, in the Materials and Methods section of this work.

However, when plotting the averaged MSD of the tracked HaloTag proteins, we observed no such trends (Figure 3.20). In fact, a closer look indicates that if most of the curve does not follow a linear trend, the first part actually does (Figure 3.21). Interestingly, the computation of the MSD reveals a statistical bias that increases as a function of time; for every lag-time, the associated MSD corresponds to the (squared) displacement of a particle, averaged on all the displacements sharing the same duration throughout the track. It soon became apparent that for longer lag-times, much less displacements emerged, thus decreasing the averaging statistical strength and eventually making noise and variability more apparent. As an illustration, please refer to Figure 2.17 found in the Materials and Methods.

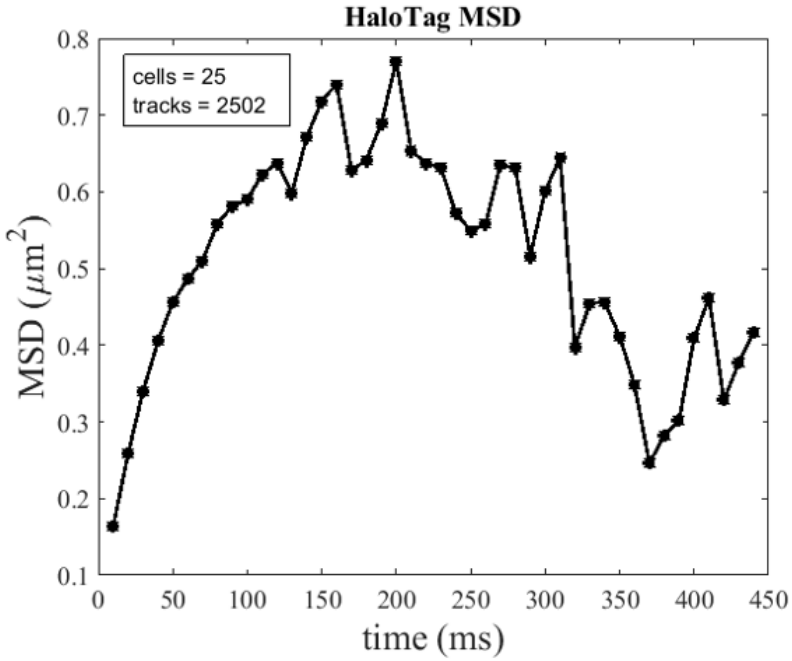

Figure 3.20. Mean Square Displacement (MSD) of the HaloTag protein. Cells containing the HaloTag protein (on a plasmid) were observed through HiLo microscopy. In the TMR channel (laser excitation at 561 nm), images were recorded with an exposure time of 10 ms. Those images were run through our tracking algorithm which computed their respective MSD, which average (over 25 cells) is shown here. Time (in ms) refers to the lag-time Δt. Total numbers of cells and tracks are indicated.

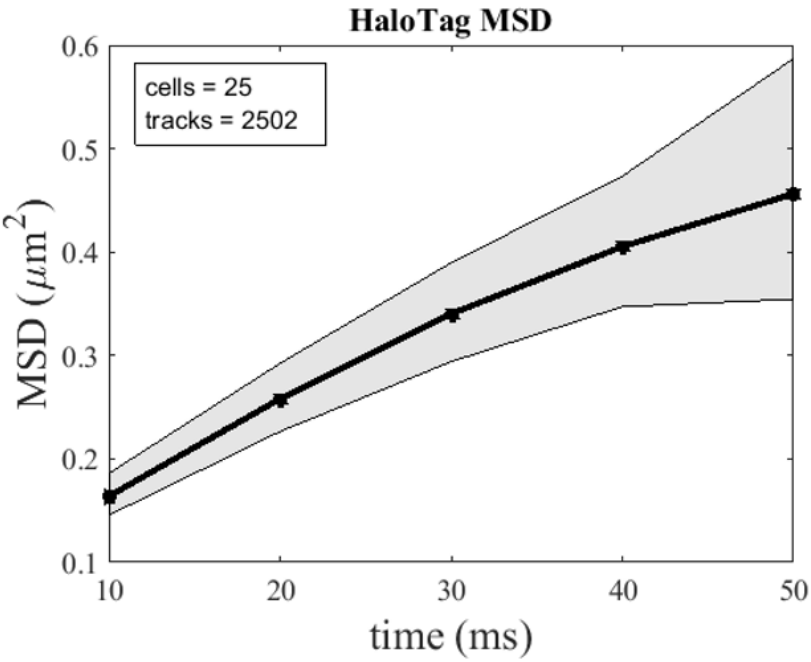

Figure 3.21. Linear trend of the HaloTag protein MSD. Only the first five timepoints of the previous plot are shown here, highlighting the linear trend of the MSD as a function of lag times. Time (in ms) refers to the lag-time Δt. Total numbers of cells and tracks are indicated. The black shade represents the standard deviation over the 25 datasets.

Looking at the MSD in this way, we can therefore confirm the freely diffusive nature of our HaloTag protein, associated to an initial linear trend followed by the emergence of a plateau, matching our expectations. We can now go one step further and compute the apparent Diffusion Coefficient (D_{app}) of our protein.

There are two ways to do this. The first one relies on the MSD curve's gradient and is therefore obtained from linear fitting. However, this technique has several disadvantages; (i) it is strongly dependent on the accuracy of the fit and therefore leads to large parts of the data being discarded (around 50%) for not sufficiently matching our accuracy threshold ($r^2 > 0.8$). Additionally, (ii) variability was further introduced in the analysis due to the wide range of track lengths (Figure 3.22), amplifying the statistical bias previously mentioned.
In contrast, the second technique is more straightforward; it consists in computing D_{app} from the first MSD timepoint. We still filter the tracks by selecting those longer than four frames (to discard fluorescent artefacts) but we are no longer dependent on additional fitting thresholds, thus preventing significant loss of (precious) data. Moreover, the use of a single timepoint limits the impact of the statistical dilution affecting the MSD over longer timepoints.

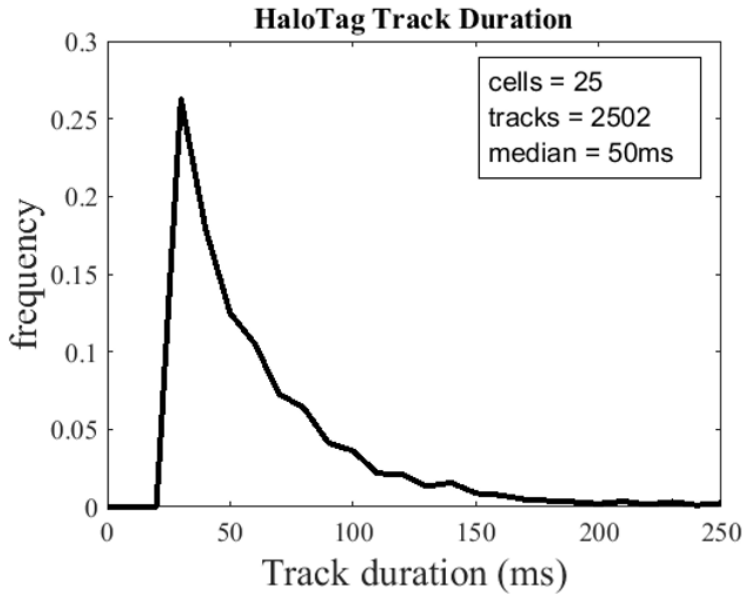

Figure 3.22. HaloTag track length. Distribution of the duration of tracks (in ms) identified from the analysis of the datasets introduced in Figure 3.21. Total numbers of cells, of tracks and median of the distribution are indicated.

The quantification of D_{app}, seen on Figure 3.23, displays a log-normal distribution centred on 4.1 µm^2/s, which does fit our expectations. Indeed, such a fast diffusion constant corresponds to proteins of similar sizes[106].

In addition, we did expect a single diffusive behaviour – freely diffusing – for HaloTag should neither interact with DNA, nor with other proteins. We can also observe that our observations are consistent across datasets (according to the standard deviation), thus confirming the robustness of our microscopy setup, the accuracy of our image analysis and the relevance of our diffusion computation.

In conclusion, we are now sufficiently confident in our materials and methods, both from the hardware and software standpoints, to proceed with the observation and analysis of the Cas9-HaloTag protein.

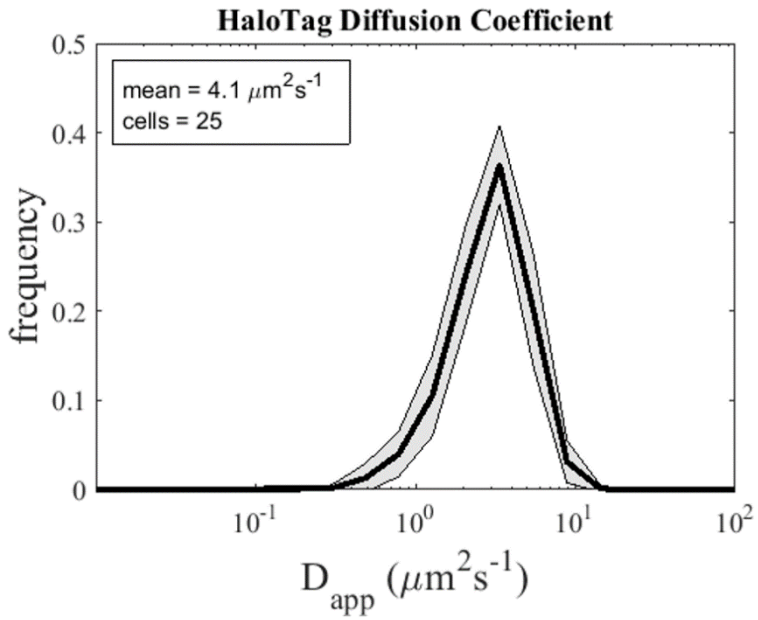

Figure 3.23. Apparent Diffusion Coefficient of the HaloTag protein. Cells containing the HaloTag protein (on a plasmid) were observed through HiLo microscopy. In the TMR channel (laser excitation at 561 nm), images were recorded with an exposure time of 10 ms. The Apparent Diffusion Coefficient of each track was computed from the first timepoint of their associated MSD. Total numbers of cells and mean of the distribution are indicated. The black shade represents the standard deviation over 25 datasets.

3.3.3 – With gRNA, Cas9 diffusion slows down

After confirmation of the robustness of our analysis, we could proceed with the analysis of our protein of interest, Cas9-HaloTag. On a first approach, we quantified the diffusion of the protein in the absence of gRNA, and then compared it to its behaviour in its presence.

As shown previously with freely-diffusing HaloTag, we extracted the Mean Square Displacement (MSD) of Cas9-HaloTag across 28 datasets (each one being a different cell, see Figure 3.24 and remarked a much slower MSD trend than with the previous HaloTag protein. Example movies can be found in the Materials and Methods chapter (Figure 2.11 and 2.16). Again, those observations match our expectations, and for two reasons. First, because of the size difference between the two proteins (34kDa compared to 194kDa); a longer protein should indeed diffuse slower. The second reason lies in their respective interactions; where HaloTag alone should not interact with anything, Cas9-HaloTag should interact with the genome, even in the absence of gRNA. Consequently, not only Cas9-HaloTag tracks should be slower than HaloTag's, but some of them should be even more so.

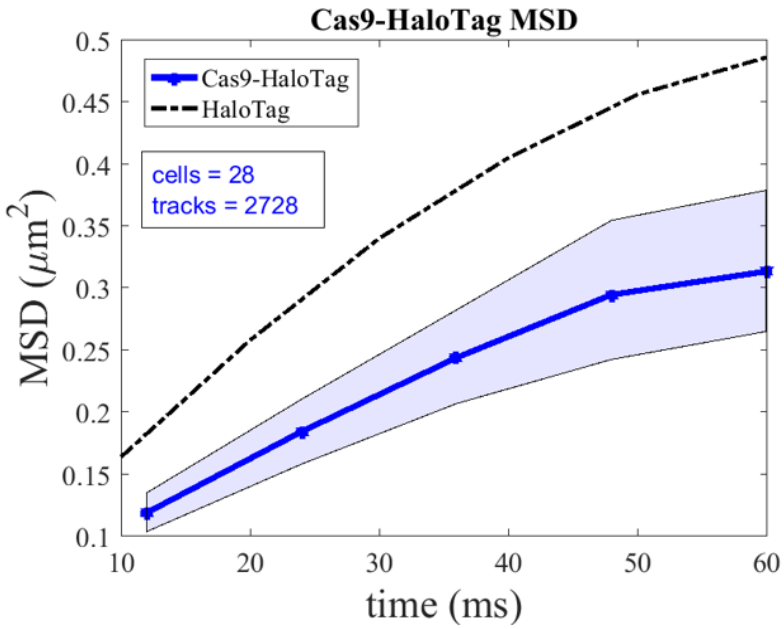

Figure 3.24. MSD of Cas9-HaloTag (first timepoints). Cells containing the Cas9-HaloTag protein were grown in the presence of 0.2% glucose, labelled and simultaneously exposed to 0.2% arabinose for 1h, washed and finally mounted on agar-pads to be observed through HiLo microscopy. Here, we can observe the MSD of the proteins, averaged over 28 datasets, with the blue shade representing the standard deviation, compared to previous HaloTag measurements (black). Time (in ms) refers to the lag-time Δt. Total numbers of cells and of tracks are indicated.

A further quantification reveals that Cas9-HaloTag apparent Diffusion Coefficient (D_{app}) is also lower than HaloTag's, which naturally follows our previous interpretations (Figure 3.25). More importantly, it appears impossible at this stage to be able to differentiate more than one diffusive behaviour, despite our previous hypothesis that the protein should either be freely diffusing, either interacting with the genomic DNA.

Already, we can assume various explanations for this phenomenon; (i) Cas9 may either spend most of its time interacting with DNA, or either none of it, therefore resulting in a single diffusive behaviour. Or (ii), our present setup may not be able to detect such differences, again for several potential reasons. However, we shall leave those considerations aside for the time being, and rather proceed with further analyses which should shed more light on this discussion.

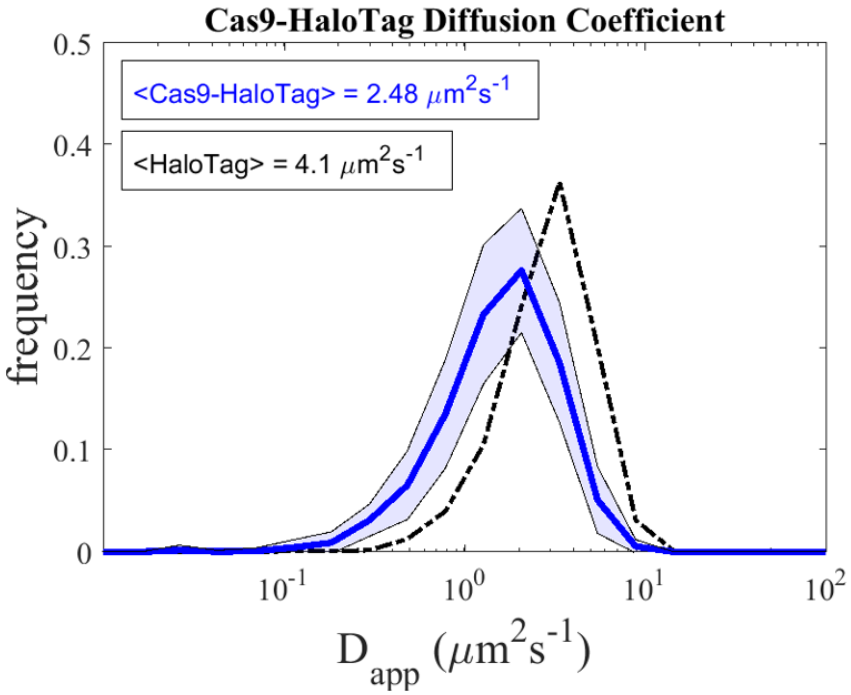

Figure 3.25. Apparent Diffusion Coefficient (D_{app}) of the Cas9-HaloTag protein. Cells containing the Cas9-HaloTag protein were prepared as described in Figure 3.24 and observed through HiLo microscopy. Here is the distribution of D_{app}, averaged over 28 datasets, with the blue shade representing the standard deviation, compared to previous HaloTag measurements (black). Total numbers of cells and mean of the distributions are indicated.

After addition of a gRNA targeting the *ptrA* gene, also written *ptrA*-gRNA or more simply gRNA, we did observe a difference in terms of MSD, but rather small (Figure 3.26). These results suggest that most of the CRISPR complex (Cas9 with its gRNA) diffusive behaviours are in common with Cas9's, diffusing on its own.

Another observation relates to the weight of the construct. Indeed, as opposed to our previous comparison (Cas9-HaloTag and free-HaloTag, with a 5.7-fold weight difference), the CRISPR complex is only 1.9% (or 1.01-fold) heavier than the Cas9 protein. This would also explain why no large diffusion difference has been observed by the addition of the gRNA.

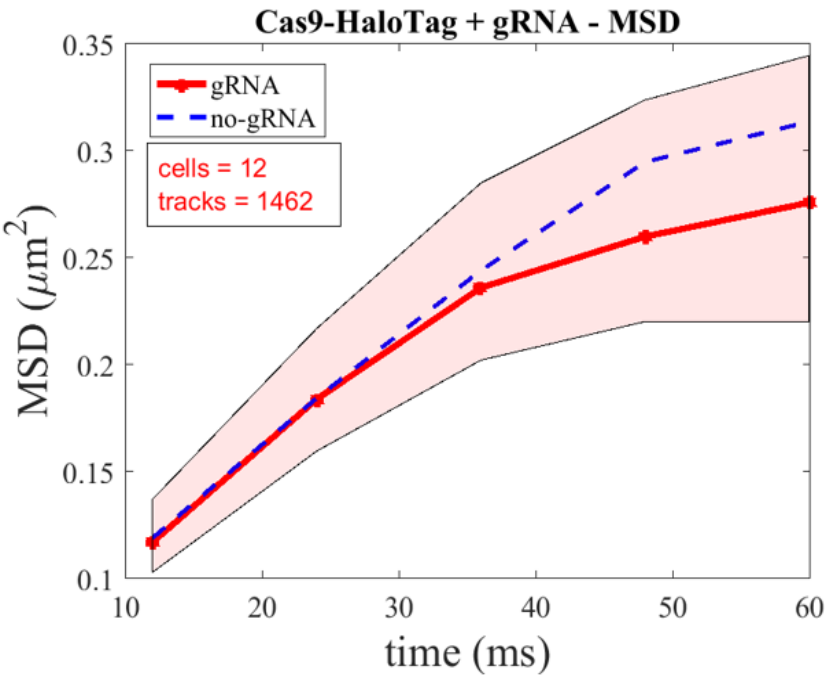

Figure 3.26. The MSD of Cas9-HaloTag is slower with a gRNA. Cells containing the Cas9-HaloTag protein with the *ptrA*-gRNA were prepared as described in Figure 3.24 and observed through HiLo microscopy. Here, we can observe the MSD of the protein, averaged over 12 datasets, with the red shade representing the standard deviation. Time (in ms) refers to the lag-time Δt. We can compare to previous measurements without gRNA (blue dotted line). With the gRNA, Cas9-HaloTag MSD slows down and reaches a plateau quicker.

Looking at their respective D_{app}, additional observations can be made. The two distributions overlap (Figure 3.27). However, with the gRNA, there seems to be an additional mode, slower, which could very well represent the new diffusion behaviour adopted by Cas9. Indeed, the presence of a gRNA may be characterized by increasing interactions with genomic DNA, probably leading to DSBs. Also, the mean of the distribution reflects this shift with a 1.6% decrease.

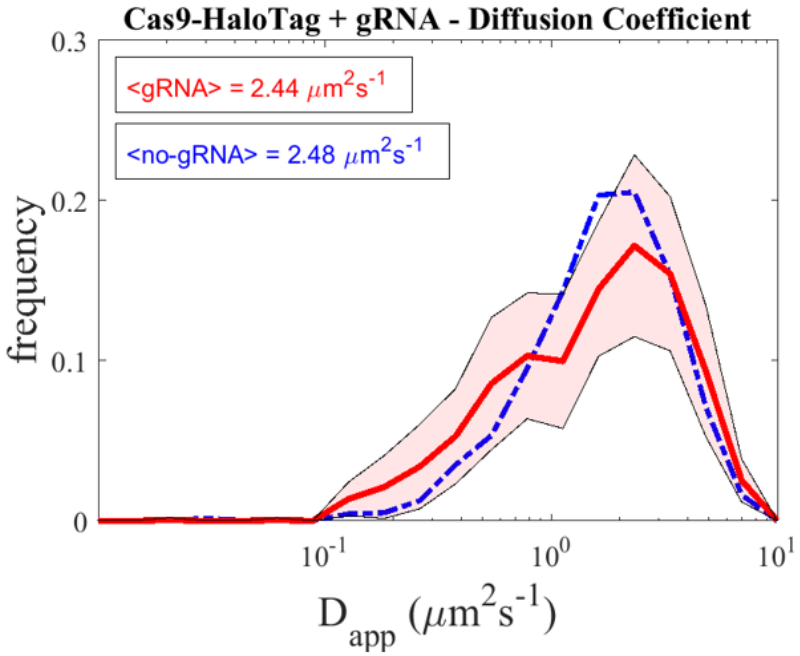

Figure 3.27. With a gRNA, Cas9-HaloTag displays two D_{app} populations. Cells containing the Cas9-HaloTag protein and the *ptrA*-gRNA were prepared as described in Figure 3.24 and observed through HiLo microscopy. Here is the distribution of D_{app}, averaged over 12 datasets, with the red shade representing the standard deviation. We can compare to previous measurements without gRNA (blue dotted line). Mean of the distributions are indicated. Two populations seem to appear, one that overlaps with the no-gRNA distribution and a slower one.

Finally, we can confirm that the presence of a gRNA does decrease Cas9-HaloTag diffusion, but it remains unclear whether this slow-down stems from a simple weight increase, or actual interactions with DNA. It is also worth noting that in those conditions, the gRNA only targets one particular target, *ptrA*, which can reach a maximum of two copies per cell (according to our cellular growth rate), thus limiting the chances of the CRISPR complex to find it and with that, our chances of observing the protein bound to its target.

3.4 – Summary and Future Work

In this chapter, we have shown that Cas9-HaloTag can be reliably expressed without impairing the original endonuclease activity of the protein, and at levels both suitable for phenotypic observation (high expression) and Single-Particle Tracking (SPT) (low expression).

We took advantage of our tightly-controlled expression system to count the number of proteins per cell, after fluorescent labelling and chemical fixation. Using our glucose-containing media and a 1h-exposure to arabinose, we observed an average of 2.93 Cas9-HaloTag per cell, very suitable number of proteins for SPT (<5 proteins/cell). In parallel, in the absence of arabinose and therefore in complete inhibitory conditions, we could still observe an average of 1.52 proteins/cell, illustrating the extent of the repression and of the leakiness of our promoter. Finally, in the presence of a gRNA (and after addition of arabinose), we could count on average 4.01 proteins/cell, highlighting the cell elongation provoked by CRISPR DNA double-strand breaks and subsequent SOS response activation, preventing cell division and corresponding labelled-protein dilution.

I then proceeded to actual SPT using HiLo microscopy, confirmed the validity of the technique (in my hands), both in terms of hardware (microscopy images quality) and software (downstream image analysis), first with the freely-diffusing HaloTag protein and then with our protein of interest, Cas9-HaloTag.

By comparing the diffusion of the protein both in the presence and absence of a gRNA, I showed that our setup could reliably distinguish two diffusion behaviours, albeit with a set of limitations. Indeed, despite the emergence of a new diffusion behaviour corresponding to the addition of the gRNA, we were unable to identify whether this new behaviour appeared in response to a heavier protein weight (that we could call ‘passive’ diffusion shift) or rather in response to different protein-DNA interactions (or ‘active’ diffusion shift), or even both.

To address these limitations, it is possible to either increase or decrease those interactions and observe the associated protein diffusion, therefore validating one hypothesis over the other. Indeed, if such variations do affect the resulting protein diffusion, it will prove that our previous observations did originate from protein-DNA interactions rather than from simple molecular weight variations.

One way to decrease those interactions can be to remove the target of the gRNA (either directly from the genome or by using a different gRNA without an associated genomic target). Another way, more drastic, is to completely remove the DNA content of the cell, for example by triggering RecB digestion (with DNA damage) in a ΔrecA cellular background[107].

Conversely, it is possible to increase the amount of protein-DNA interactions, either by increasing the amounts of targets recognised by a gRNA (for instance by introducing repeated targets in the genome) or by increasing the time during which the protein stays bound to the genome (for example, by using the catalytically-dead mutant dCas9, shown to stay immobile on DNA for several tens of minutes).

Those examples will precisely be explored in the next chapter, which will see the introduction of different gRNAs of which targets will or will not be present in the cells.

In conclusion, this third chapter has presented the elaboration, the development and the confirmation of our method, of our techniques and of our constructs which can now be extended to further characterize the diffusion behaviour of the CRISPR-Cas9 complex.

Chapter IV – Results II

Quantification of CRISPR-Cas9 diffusion behaviours

4.1 – Introduction

The previous chapter confirmed our ability to qualitatively and quantitatively study Cas9 diffusion through Single-Particle Tracking (SPT). I showed how the technique was developed, implemented and how our CRISPR constructs activity was controlled.

Our first study dealt with Cas9 diffusion differences with and without a gRNA targeting the *ptrA* gene. We showed that the addition of the gRNA slowed down Cas9 diffusion, albeit mildly, for the apparent Diffusion Coefficient (D_{app}) decreased by a mere 1.6%.

However, this decrease looking particularly small, it limited our ability to identify further diffusion behaviours of the CRISPR-Cas9 complex. We thus decided to improve our sensitivity by increasing the number of targets recognised by the CRISPR complex. Doing so, we reasoned, should push Cas9's D_{app} towards lower values and highlight additional diffusion behaviours.

Additionally, one remaining uncertainty was tied to the complex molecular weight increase when associated to a gRNA. Surely, adding more targets should address this question in the following way; if the aforementioned D_{app} decrease observed with the addition of the gRNA was solely the result of a weight increase, then adding more targets should not affect D_{app} any further. Conversely, observing D_{app} decreasing as a function of a target number increase would irremediably prove that Cas9 slowing down originates from actual protein-DNA interaction.

To achieve this task, we chose to use two arrays of *lacO* repeats; one with 6 repeats and a second with 22. Every repeat would be flanked by PAM sequences (to be recognised by Cas9) and integrated into *E.coli*'s chromosome, at the same locus (see Materials and Methods). In addition to this, two gRNAs were designed; one targeting the array, called *lacOA*-gRNA, and a second targeting the *lacO* endogenous sequence, labelled *lacO*-gRNA. Please refer to Figure 2.9 in the Materials and Methods, Chapter II.

The choice of *lacO* sequences as CRISPR targets was not arbitrary. Importantly, *lacO* sequences are usually bound by LacI proteins, which in our case offers the benefit of having LacI effectively protecting the DNA from Cas9 recruitment, until addition of IPTG (which binds to LacI and releases it from its *lacO* binding). Additionally, the *lacO*-LacI interaction is already well characterized and can be used in several ways (*e.g.* LacI fused to a fluorescent protein to localize the site of the array), as was demonstrated by Johann Elf who used similar repeats as dCas9 plasmid targets[79].

In this chapter, I will show how those newly-built CRISPR-*lacO* constructs successfully work, I will examine the diffusion results obtained through SPT and finally discuss their implications in the light of Cas9 interactions with DNA reported in the literature.

4.2 – *lacO* arrays: viability and target recognition

4.2.1 – CRISPR activity with *lacO* arrays

After integration of the *lacO* repeats into cells containing the pBAD-Cas9-HaloTag construct, our first objective was to test the efficiency of our new sgRNAs. Indeed, having a perfect homology between a gRNA and its target is not a guarantee that the CRISPR complex will cut the DNA. As an illustration, the secondary structure of the target sequence may prevent the correct recruitment of the Cas9 protein, its activity or even the recognition of the gRNA. Conversely, some gRNAs are so effective that targeting the genome becomes too lethal for the cells and it is impossible to express both Cas9 and the gRNA without mutating and losing the CRISPR activity altogether. For example, in our hands, the crRNA-lacZ1[108] would not damage the cells whereas the crRNA-lacZ2[108] would be so damaging that mutants having lost the construct were systematically selected (data not shown).

An appropriate way to test the activity of the CRISPR complex in the presence of our new gRNAs is to perform a transformation assay[108]; namely, to transform cells already containing the Cas9 construct, overexpressed with arabinose, with plasmids containing different gRNAs targeting the genome. Successful CRISPR complexes would prevent the survival of cells that managed to receive the gRNA-plasmids. Comparing those transformations to the same ones performed in the absence of Cas9 overexpression, it is possible to identify which gRNAs effectively drive Cas9 to cut the DNA.

We performed this assay and observed that our two new gRNAs were behaving very differently (see Figure 4.1). First, after normalizing our results to the negative control (deprived of any gRNA), we observed that the *ptrA*-gRNA successfully prevented the transformation, as expected. Similarly, the *lacOA*-gRNA (targeting the *lacO*-array sequences) prevented the survival of any transformants, validating the efficiency of both our constructs: the gRNA's for driving Cas9 breakage and the *lacO* array's for being recognised by the CRISPR system.

In contrast, the *lacO*-gRNA (targeting the endogenous *lacO* sequence) did not seem to affect the transformation efficiency of Cas9-overexpressed cells, showing a defect in CRISPR activity, of unknown origin, much like the previous example of crRNA-lacZ1.

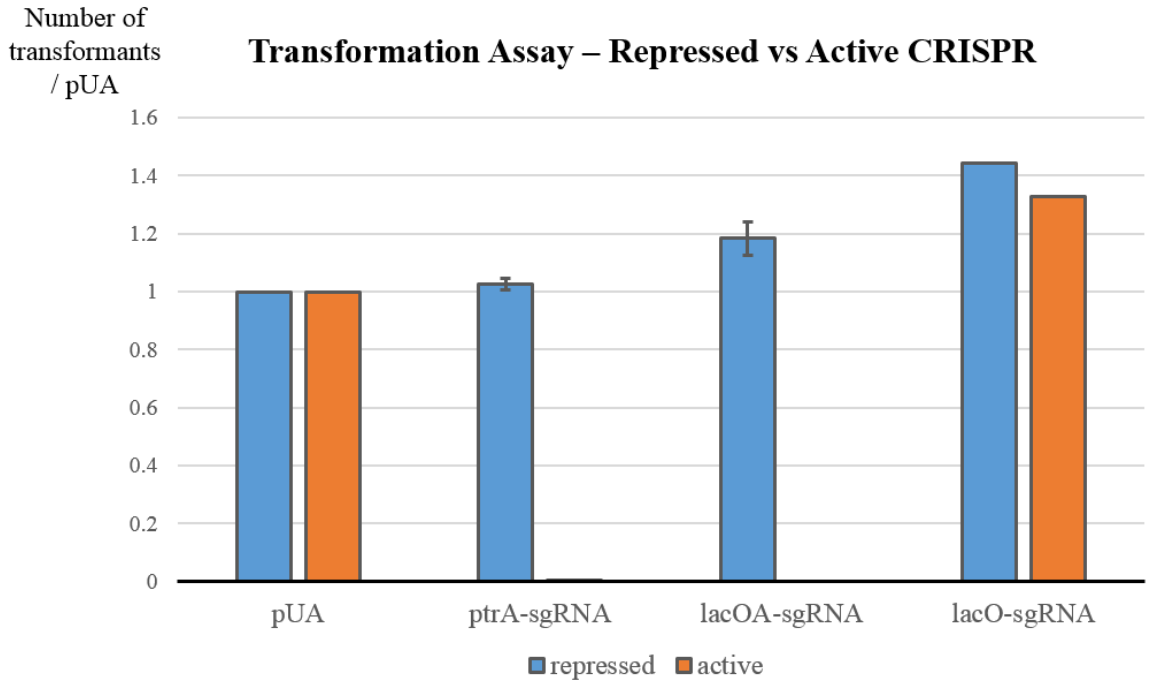

Figure 4.1. With CRISPR induction, cells containing the *lacO* array cannot be transformed with the lacOA-gRNA. Cells containing pBAD-Cas9-HaloTag and the *lacO6*-array have been grown in LB supplemented with either 0.2% glucose (repressed, blue) or 0.2% arabinose and 1mM IPTG (active, orange), and transformed with a plasmid containing either of the following; *ptrA*-sgRNA, *lacOA*-sgRNA, *lacO*-sgRNA or empty (pUA). Transformed colonies were then counted and normalized by the number of colonies obtained with the pUA transformation performed in the same condition (repressed or active). Samples were duplicated (except for lacO-sgRNA). Mean values and standard errors are displayed. With CRISPR active, both the *ptrA*-sgRNA and the *lacOA*-sgRNA fail to give transformants. However, the *lacO*-sgRNA does not seem to affect transformation.

4.2.2 – *lacO*-array-gRNA viability and specificity

As mentioned earlier, an effective CRISPR system may prove too effective and therefore present too much of a burden for the cells. This may in turn trigger selection to get rid of the construct altogether, which would prove particularly detrimental to our experiments; first because it would make our construct highly unstable (and thus unreliable) and second because of the technical difficulty of working with cells constantly on the brink of death. Therefore, to confirm our ability to work with the CRISPR-*lacOA* construct, I performed a spot test with cells previously transformed with the plasmids carrying the gRNAs (in Cas9-repressed conditions). The cells were exposed for 1h to either 0.2% of glucose (to keep repressing Cas9, as a negative control), either 0.2% arabinose (to overexpress Cas9), with or without 1mM of IPTG (which releases LacI from the *lacO* sequences and therefore leaves the target free for Cas9 recognition). If the phenotypes previously observed during the transformation assay could be repeated with the spot test, that would validate our ability to conserve the constructs and with them, the CRISPR activity.

Our results indicate that for both the *lacO6* and *lacO22* constructs, we do conserve CRISPR-*lacOA* activity across several generations (in Cas9-repressed conditions). Indeed, Figure 4.2 shows that the cells transformed with the *lacOA*-gRNA and grown for 16-hours (in glucose) displayed successful CRISPR activity in the form of cell death. After an hour of exposure to 0.2% arabinose, we start to observe cell death. In the presence of additional IPTG (1mM), the cell death increases to an additional log. In contrast, kept with glucose, the cells fail to show signs of CRISPR activity.

Put together, those results indicate that our CRISPR-*lacOA* system can be submitted to three levels of regulation; full repression (i) with glucose, where Cas9 is not expressed and therefore does not cut. Half-repression (ii) where Cas9 is expressed but where LacI protects the *lacO* targets by staying on the DNA. And full induction (iii), where Cas9 is expressed and where the *lacO* targets are free for CRISPR recognition, due to IPTG releasing LacI from DNA. Also, we confirm that even in the presence of arabinose and IPTG, the gRNA targeting the endogenous *lacO* sequence (*lacO*-gRNA) does not lead to any observable CRISPR activity (see Figure 4.2A).

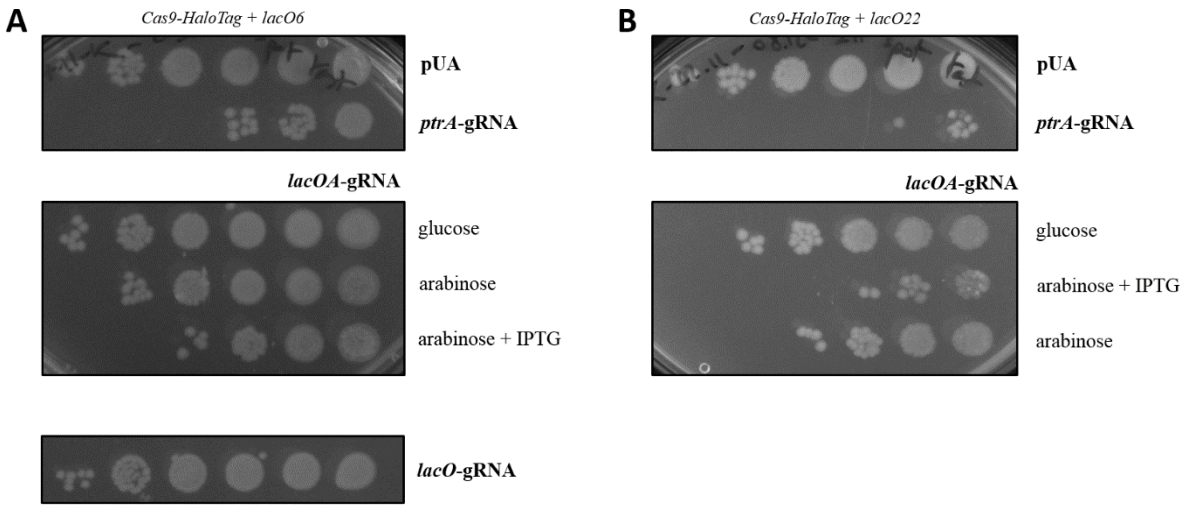

Figure 4.2. Cells containing the *lacO* array and the *lacOA*-gRNA conserve the CRISPR phenotype. Cells containing pBAD-Cas9-HaloTag and the *lacO6*-array (**A**) or the *lacO22*-array (**B**) have been grown in LB supplemented with 0.2% glucose and kanamycin (to keep the plasmids carrying the different gRNAs, or the empty plasmid pUA), then exposed for 1h to either 0.2% arabinose (with or without 1mM IPTG) or nothing, then serially plated on kanamycin petri dishes. *ptrA*-gRNA and *lacOA*-gRNA contain cells with the eponymous gRNAs. In those cells, CRISPR is normally activated (and repressed) with the *ptrA*-gRNA as well as with the *lacOA*-gRNA (targets the *lacO* array). However, there is no sign of CRISPR activity with the *lacO*-gRNA (targets the *lacO* endogenous sequence). pUA rows correspond to the same cells containing the gRNA-containing plasmids, but without the gRNA.

In addition to the previous confirmation of the successful activity of our CRISPR-*lacOA* system, we tested whether it displayed any non-specific activity. To do so, we performed the same transformation assay, but in cells lacking the *lacO* repeats. As a result, there should not be any CRISPR recognition nor DNA break, unless another genomic locus were non-specifically cut. We observed that in the absence of the proper targets (*lacO* array), the transformation was just as efficient as in the absence of gRNA (empty plasmid) and therefore confirmed the strong specificity of our system for its rightful target (see Figure 4.3). This was very important as this further confirmed our ability to reliably express our system and observe effects stemming for its activity, and from it alone.

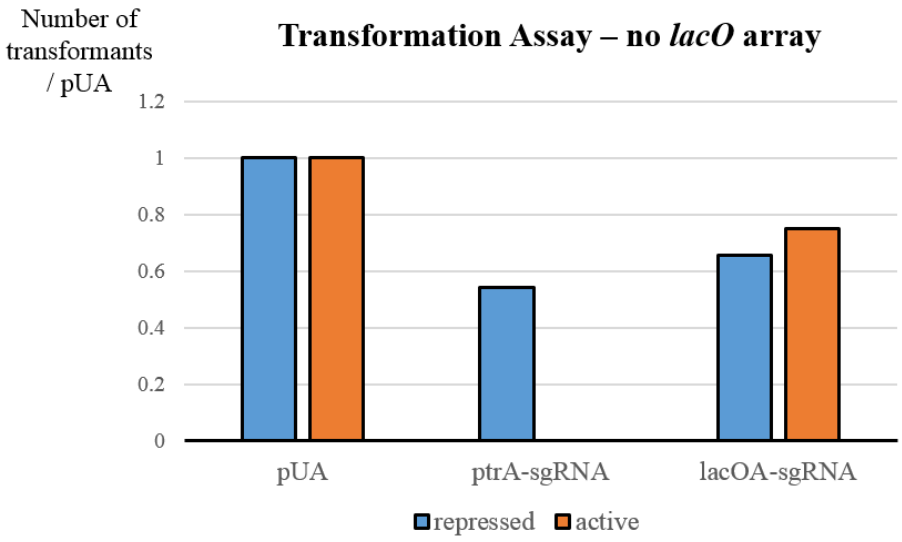

Figure 4.3. Without *lacO* array, the lacOA-sgRNA does not produce any damage. Cells containing pBAD-Cas9-HaloTag and no integrated *lacO*-array have been grown in LB supplemented with either 0.2% glucose (repressed, blue) or 0.2% arabinose and 1mM IPTG (active, orange), and transformed with a plasmid containing either of the following; *ptrA*-sgRNA, *lacOA*-sgRNA, or empty (pUA). Transformed colonies were then counted and normalized by the number of colonies obtained with the pUA transformation performed in the same condition (repressed or active). With CRISPR active, the *ptrA*-sgRNA fails to give transformants, confirming Cas9 cutting. However, the *lacO*-sgRNA does not affect transformation, indicating that in the absence of integrated *lacO* array, the sgRNA will not drive Cas9 to cut.

4.2.3 – SOS dynamics with CRISPR-*lacO*

Despite our confirmation, at the population level, of the successful activity of our *lacOA*-CRISPR constructs (*i.e.* the association of pBAD-Cas9-HaloTag, the *lacOA*-gRNA and the *lacO*-repeats), we still lacked the insight provided by single-cell measurements; the level of CRISPR activity's heterogeneity among the population, the dynamics of the activity and more simply, the differences between the three levels of regulation; full repression, full activation and intermediary state.

To address these questions, I measured the SOS activity of single cells exposed to the three aforementioned levels of regulation (through our pSfiA-GFP construct, see the Materials and Methods chapter). Fully-repressed cells were constantly exposed to glucose. Fully-activated cells were grown in glucose but exposed to both arabinose (0.2% w/v) and IPTG (1mM) for an hour. Finally, the intermediary state of regulation was triggered by a pulse of arabinose, but without IPTG, the idea being to have Cas9-HaloTag expressed (in the presence of the *lacOA*-gRNA), but with LacI still sitting on the *lacO* sequences, thus protecting them from Cas9 recruitment.

Figure 4.4 clearly shows those three states of regulation, with three populations that progressively shift towards higher SOS activity following increasing CRISPR activation. Interestingly, we reproduce the results previously obtained through the spot test experiment, but this time with the additional insight of the shape of each distribution. When fully repressed, the great majority of the cells show no sign of SOS induction. When fully activated (arabinose and IPTG), most of the cells switch to a high-SOS condition. Finally, the intermediary state (with arabinose but without IPTG) does show an intermediary result, with a heterogeneous population overlapping both previously-mentioned ones.

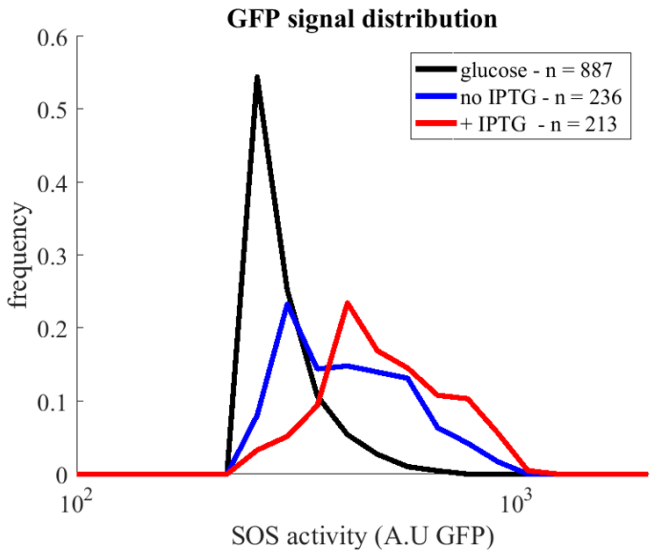

Figure 4.4. araBAD and LacI give two layers of CRISPR/*lacO* repression. Cells containing pBAD-Cas9-HaloTag, the *lacO22*-array and the *lacOA*-sgRNA have been grown in Generic media (supplemented with 0.2% glucose) and then exposed for 1h to either 0.2% arabinose with 1mM IPTG (+ IPTG, red), only arabinose (no IPTG, blue) or nothing (glucose, black). Then the cells were mounted on agar-pads for microscopy. Their respective distributions of GFP signal (fluorescence intensity/pixel/cell) are shown here. Number of cells (n) are indicated. We can see three populations, displaying three levels of regulation; fully repressed with glucose, half-repressed with only arabinose and fully induced with both arabinose and IPTG.

However at this point, it was impossible to say whether the intermediary state originated from a combination of cells – those with CRISPR fully working, and the others with LacI completely preventing Cas9 recruitment – or whether this state was just a snapshot of a more dynamic picture, with an equilibrium between Cas9 and LacI recruitment at the *lacO* sites not yet reached.

To answer this question, I imaged the same sample (intermediary dataset, with arabinose but no IPTG) 105 minutes after the first acquisition (from which the previous results have been extracted). After such a period of time (corresponding to roughly 4 generations), we could observe a clear shift of the SOS expression towards very high SOS activity (Figure 4.5), indicating that during the previous acquisition, the cells were simply waiting for Cas9 to bind the target sites. This further shows that the main difference presented by the absence of IPTG is a delay in CRISPR-Cas9 recruitment, due to LacI protecting the *lacO* targets, but not a complete repression.

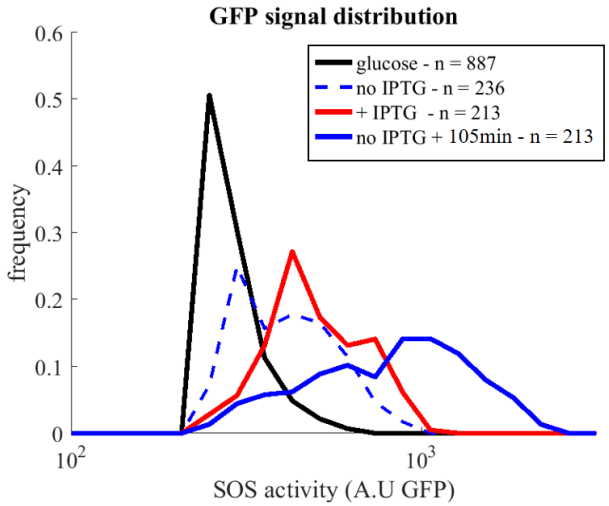

Figure 4.5. LacI repression deceases over time. The GFP signal of the same cells described in Figure 4.4 are displayed here. In addition, cells exposed to arabinose but without IPTG have been observed 105 minutes after the first acquisition (no IPTG + 105min, blue). In comparison, the first acquisition is shown in dashed blue line. Numbers of cells are indicated (n). After 105 minutes, cells display much more GFP signal (SOS induction).

This interpretation is confirmed by the fact that after 105 minutes, the amount of cells highly inducing SOS is actually higher than what could be observed in the sample exposed to IPTG in the first place. Indeed, in order to obtain additional quantification, I counted the number of cells displaying SOS activity above a certain threshold, which has been chosen with the glucose dataset as a reference (in normal conditions, about 10% of the cells in a population should spontaneously induce SOS[109]). Figure 4.6 clearly shows the three levels of regulation (full repression, intermediary and full activation), with respective 10%, 57% and 83% of cells displaying SOS activity higher than the chosen threshold. Interestingly, after 105 minutes, the intermediary state displayed 91% of its cells expressing very high SOS, confirming that over time, the vast majority of the cells temporarily protected by LacI did suffer a Cas9 break.

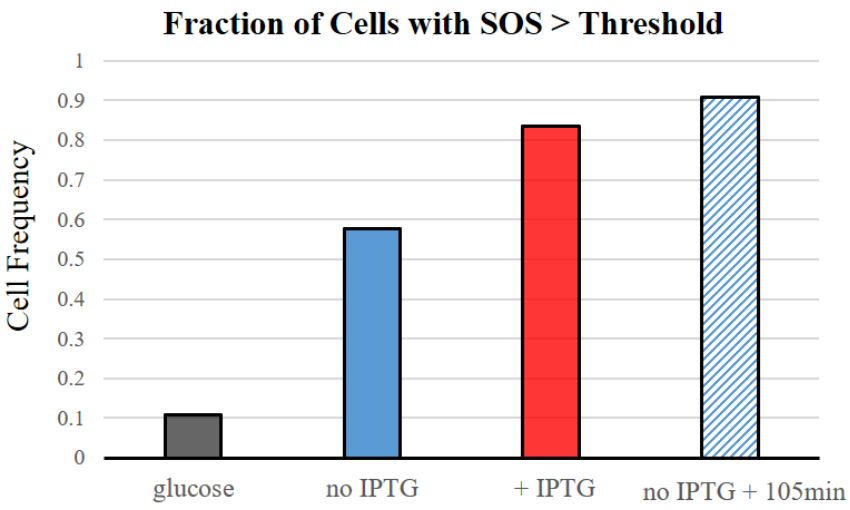

Figure 4.6. Quantification of CRISPR/*lacO* repression with araBAD and LacI. An intensity threshold has been applied to the GFP signal distributions shown on the previous figure (4.5), and the resulting proportions of cells displaying GFP fluorescent above this threshold are shown here. Compared to CRISPR full repression (glucose, black, 10%), we see an increasing CRISPR activation with arabinose (blue, 57%) and arabinose with IPTG (red, 83%). Finally, after 105 minutes, even in the absence of IPTG, the cells show a high SOS induction (blue stripes, 91%).

In conclusion, we have shown that we can successfully use our *lacOA*-sgRNA to drive Cas9-HaloTag, itself under the control of the araBAD promoter, to the *lacO* array we integrated into the chromosome. Furthermore, we have also shown that this system offers three levels of regulation which allow us to obtain different shades of CRISPR activity, which, as we will see further along, will prove particularly useful.

4.3 – Cas9-HaloTag tracking with several genomic targets

4.3.1 – Tracking with *lacO6*

Of the two *lacO* arrays that we have at our disposal, respectively containing 6 and 22 repeats, let us have a look at the first one, which we simply write *lacO*6.

Harbouring six-times more targets than our previous construct targeting the *ptrA* gene, we were expecting to increase the chances of Cas9-HaloTag to be recruited on the DNA and thus observe a slower protein diffusion. And indeed, Figure 4.7 shows exactly that, with the Mean Square Displacement (MSD) curve clearly below its *ptrA* counterpart, indicating a much smaller diffusion area, on average. A closer look at the shape of the curves, however, indicate similar slopes and therefore points at similar behaviours, which again fits our expectations, as the protein should still behave similarly (with the same cellular containment), yet slower.

Quantifying the duration of the tracks detected in our Cas9-HaloTag movies, we observe that our *lacO6* construct does yield longer tracks, on average (116 ms as opposed to 80 ms with the *ptrA*-gRNA, Figure 4.8). Indeed, due to a more concentrated signal, slower molecules are easier to track and therefore yield longer trajectories. This is yet another indication that the protein interacts more with DNA, presumably due to the higher number of targets.

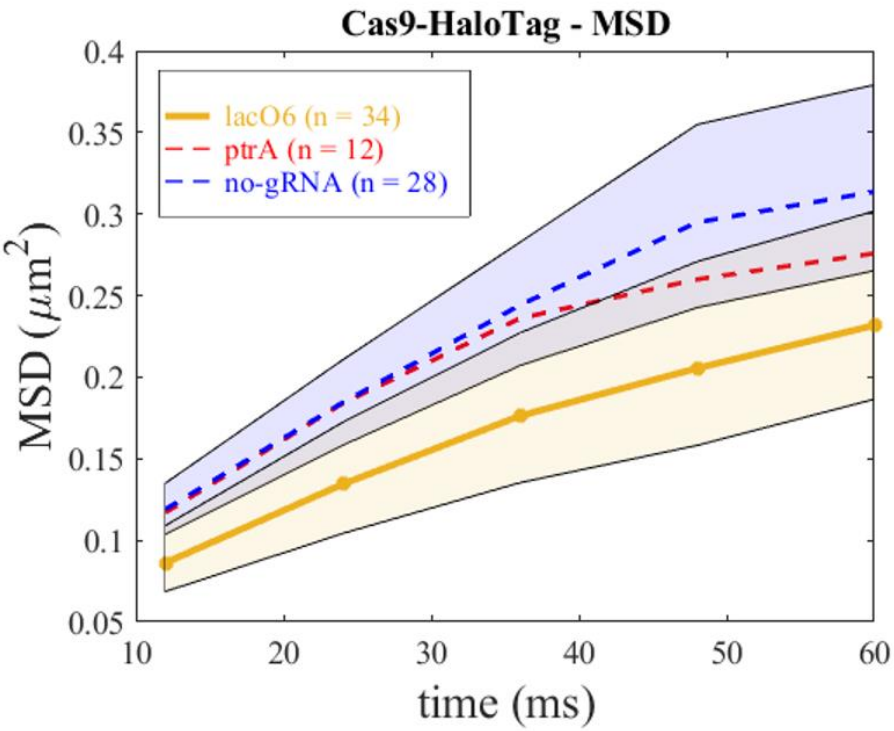

Figure 4.7. With the *lacO6* array and the *lacOA*-gRNA, Cas9-HaloTag displays a lower Mean Square Displacement (MSD). Cells containing the Cas9-HaloTag construct, the *lacO6* array and the *lacOA*-gRNA were grown in the presence of 0.2% glucose, labelled (with TMR) and simultaneously exposed to 0.2% arabinose for 1h, washed and finally mounted on agar-pads to be observed through HiLo microscopy. In the TMR channel (laser excitation at 561 nm), images were recorded with an exposure time of 12 ms. The MSD is plotted as a function of lag-times (or time, in ms), compared to previous HaloTag measurements (*ptrA*-gRNA in red and no-gRNA in blue). Total numbers of cells are indicated. The yellow shading represents the standard deviation of the *lacO6* dataset (over 34 cells) and the blue one recalls the pUA dataset standard deviation (over 28 cells). In the presence of the *lacO6* array and of the *lacOA*-gRNA (targeting the array), Cas9-HaloTag appears slower.

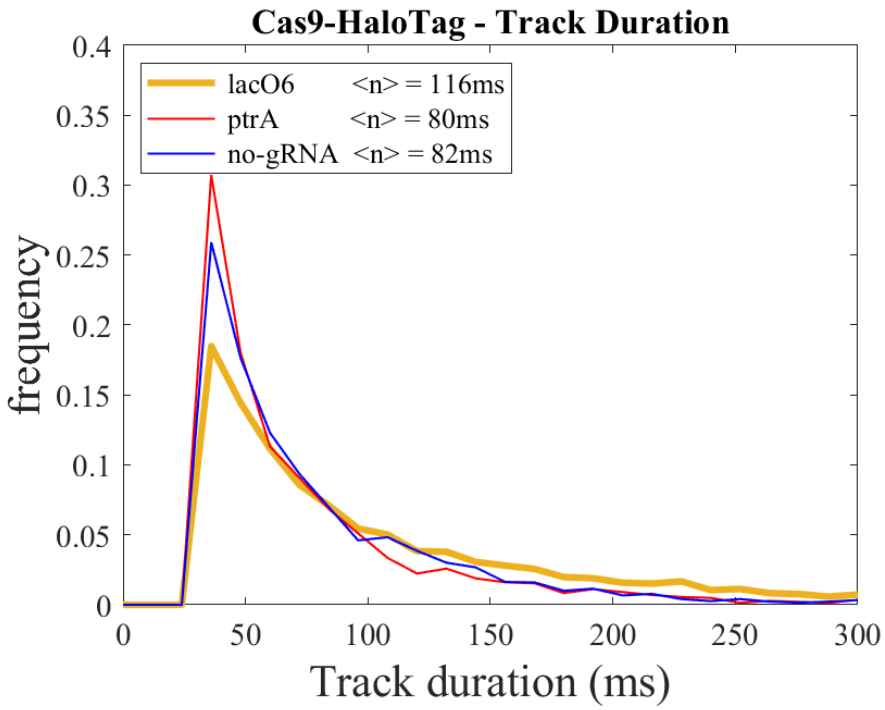

Figure 4.8. With the *lacO6* array and the *lacOA*-gRNA, Cas9-HaloTag displays longer tracks. Cells containing the Cas9-HaloTag construct, the *lacO6* array and the *lacOA*-gRNA were prepared as described in Figure 4.7. Here, we can observe the distribution of track duration of the aforementioned cells (in yellow), compared to previous datasets (*ptrA*-gRNA in red and no-gRNA in blue). Mean of the distributions are indicated. In the presence of the *lacO6* array and of the *lacOA*-gRNA (targeting the array), Cas9-HaloTag tracks appear longer.

The distribution of apparent Diffusion Coefficient (D_{app}) of the CRISPR-*lacO6* construct reproduces a clear shift from the no-gRNA form, and goes much further left than the CRISPR-*ptrA* did (Figure 4.9). Indeed, where the CRISPR-*ptrA* shows two modes, one overlapping the no-gRNA and one shifting away towards slower values, the CRISPR-*lacO6* only shows one mode, leftwards, displaying higher homogeneity and slower diffusion.

The distributions' mean follow the same trend; where CRISPR-*ptrA* only displayed an overall 1% decrease from the no-gRNA construct's D_{app}, CRISPR-*lacO6* shows a 27% decrease. Interestingly, this 27-times difference is much higher than the expected 6-times one (for 6-times more targets). At this point, we could even hypothesize the presence of a form of facilitated recruitment, where several targets located at short distance from each other may in fact increase the recruitment of the Cas9 protein. Indeed, after a first recruitment, the protein may stay in the vicinity of the array and therefore find it easier to run into another neighbouring target.

Another round of quantification was produced by counting the proportion of tracks in each dataset displaying an apparent Diffusion Coefficient (D_{app}) inferior to 1 µm^2/s (chosen as an indicator of slow molecules). This threshold was chosen in accordance with previous measurements found in the literature[81], where the D_{app} of the catalytically-dead Cas9 mutant (dCas9) was observed to vary between 10 µm^2/s (for very fast molecules) and 0.1 µm^2/s (for immobile molecules); therefore, they observed that slow molecules would appear so below 1 µm^2/s. Additionally, another study[106] studying D_{app} as a function of molecular weight allows us to estimate that Cas9 should freely diffuse above 1 µm^2/s, reinforcing our choice of threshold.

Consequently, this additional quantification resulted in a similar trend, yet not as dramatic (Figure 4.10); without any gRNA, 18% of the detected tracks displayed 'slow' diffusion, as opposed to 26% and 31%, respectively for CRISPR-*ptrA* and CRISPR-*lacO6*. This time, the increase difference between the two CRISPR constructs is only of 1.6-times.

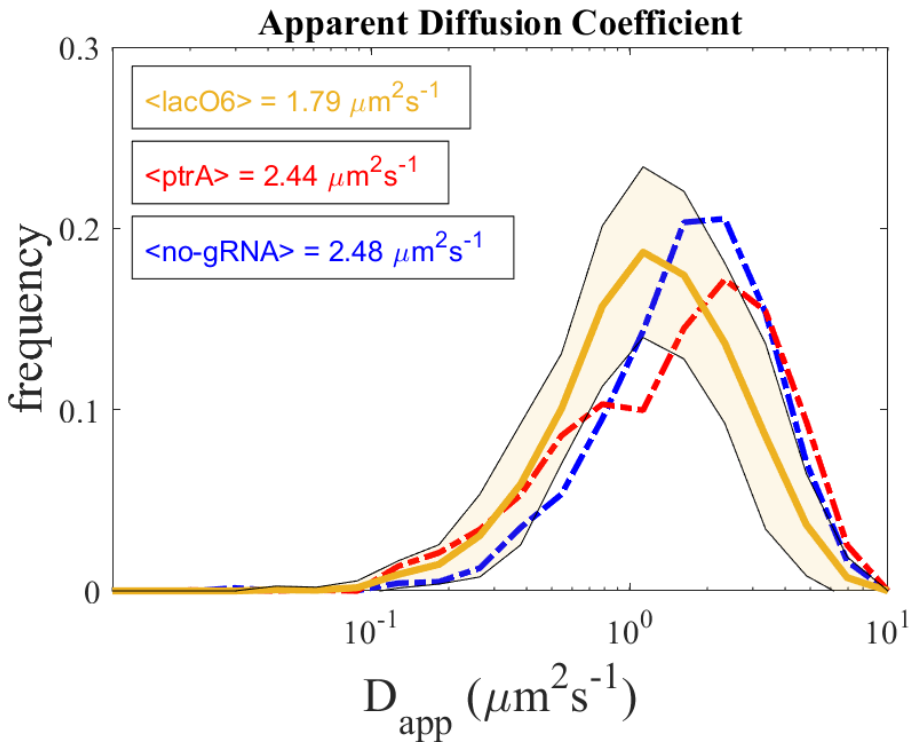

Figure 4.9. With the *lacO6* array and the *lacOA*-gRNA, Cas9-HaloTag appears slower than previously observed. Cells containing the Cas9-HaloTag construct, the *lacO6* array and the *lacOA*-gRNA were prepared as described in Figure 4.7. Here, we can observe the distribution of apparent Diffusion Coefficient (D_{app}) of the aforementioned cells (in yellow), compared to previous datasets (*ptrA*-gRNA in red and no-gRNA in blue). Mean of the distributions are indicated. In the presence of the *lacO6* array and of the *lacOA*-gRNA (targeting the array), Cas9-HaloTag detected molecules appear slower than the protein diffusing in the absence of a gRNA, and the protein diffusing with the *ptrA*-gRNA.

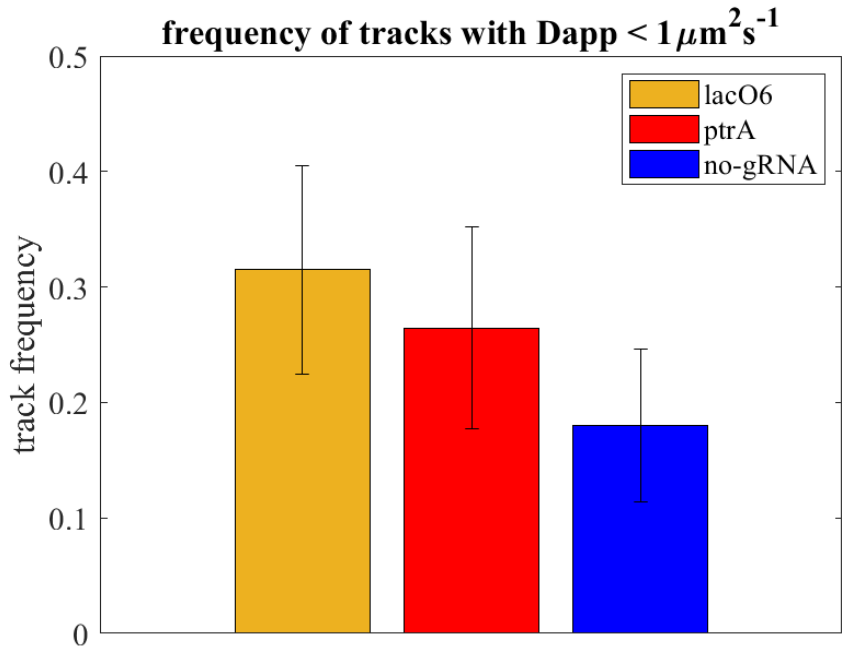

Figure 4.10. With the *lacO6* array and the *lacOA*-gRNA, Cas9-HaloTag displays very slow tracks. Cells containing the Cas9-HaloTag construct, the *lacO6* array and the *lacOA*-gRNA were prepared as described in Figure 4.7. Here, we quantified the proportion of detected tracks with $D_{app} < 1$ µm²/s. The *lacO6* dataset is in yellow (31%), the *ptrA*-gRNA is in red (26%) and the one without a gRNA is in blue (18%). In the presence of the *lacO6* array and of the *lacOA*-gRNA, Cas9-HaloTag displays more slow tracks than in the absence of a gRNA, and with the *ptrA*-gRNA. Error bars indicate the standard deviation calculated on their respective number of cells (34, 12, 28).

In conclusion, our CRISPR-*lacO6* construct did significantly slow down Cas9-HaloTag, and its comparison with our previous *ptrA* construct seems to indicate that this D_{app} decrease stems from an increased number of targets. Repeating these experiments with our second array containing 22 *lacO* targets, we will investigate whether this assumption is further confirmed, and D_{app} further slowed down.

4.3.2 – Tracking with *lacO22*

After studying the protein diffusion observed with the *lacO6* construct, let us examine the behaviour displayed with the CRISPR-*lacO22* construct, and see if another increase in CRISPR targets does yield a similar slow-down in protein diffusion. A first glance at the MSD of the CRISPR-*lacO22* construct shows a very similar behaviour to the *lacO6*'s, albeit slightly slower (3%) (Figure 4.11A).

This trend is conserved with the apparent Diffusion Coefficient (Figure 4.11B), although in both cases, the area representing the standard deviation of the *lacO22* dataset covers both construct averages, therefore making it hard to assess the significance of such a difference. That said, this difference hardly reaches the expected 3.6-times increase due to the higher number of targets (from 6 to 22).

The additional quantification of the proportion of slow tracks in each dataset showed a 20% increase in slow tracks with the *lacO22* construct (compared to the *lacO6* one) (Figure 4.12A). This indicates that a wider look at the population does not show much behaviour change, while a closer look to the slow part of it does highlight such a difference, pointing at the particularly infrequent occurrences of Cas9 interacting with its target.

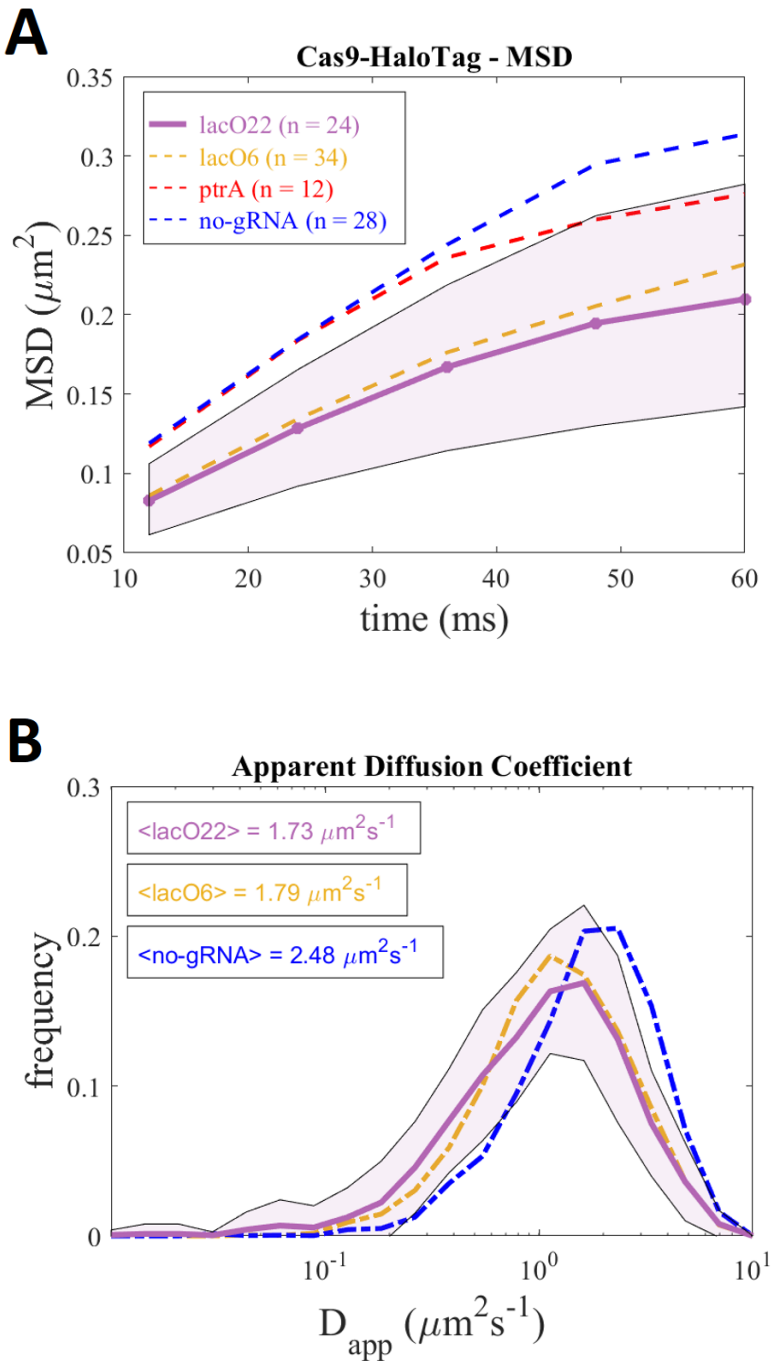

Figure 4.11. With the *lacO22* array and the *lacOA*-gRNA, Cas9-HaloTag appears slower than with *lacO6* array. Cells containing the Cas9-HaloTag construct, the *lacO22* array and the *lacOA*-gRNA were prepared as described in Figure 4.7. (**A**) The MSD is plotted as a function of lag-times (or time, in ms) (in purple), compared to previous HaloTag measurements (*lacOA6*-gRNA in yellow, *ptrA*-gRNA in red and no-gRNA in blue). (**B**) The distribution of apparent Diffusion Coefficient (D_{app}) of the aforementioned cells shows a similar trend. Mean of the distribution and total numbers of cells are indicated. The purple shadings represent the standard deviation (over 24 cells). In the presence of the *lacOA*-gRNA and the *lacO22* array, Cas9-HaloTag appears slightly slower than with 6 *lacO* targets, making it the slowest condition.

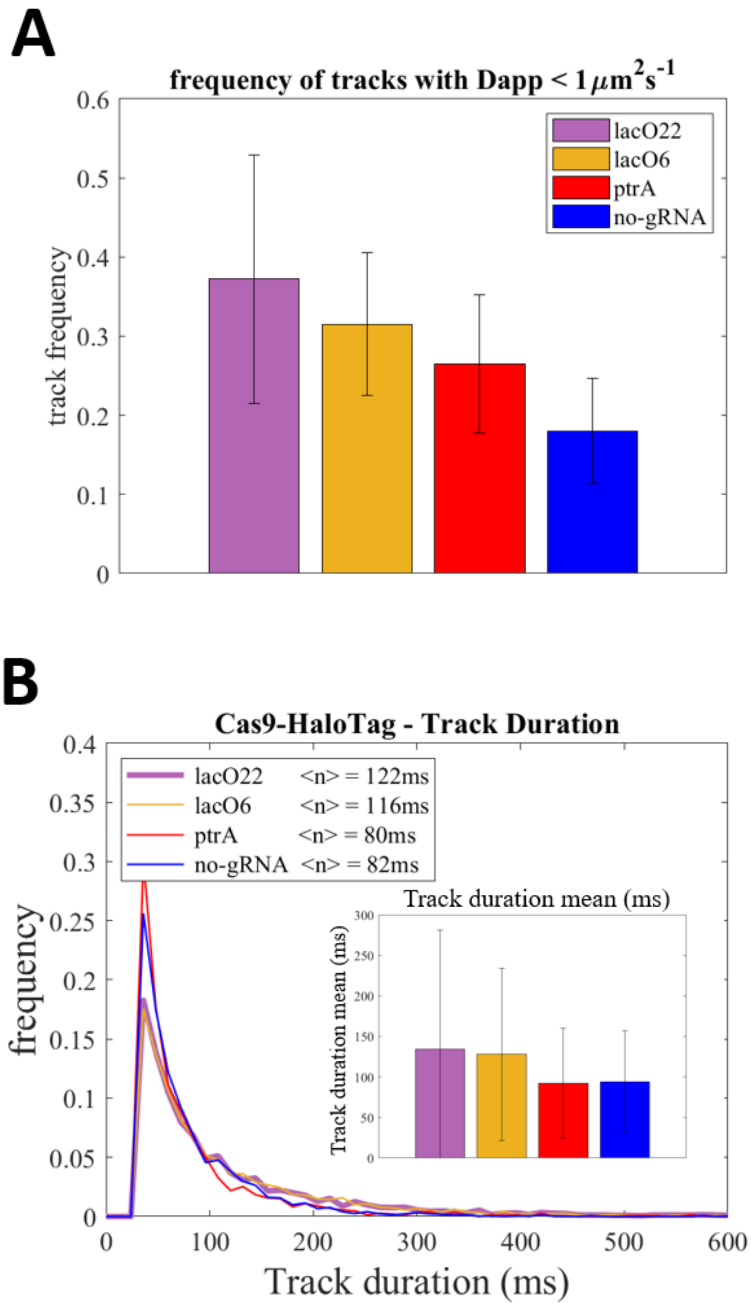

Figure 4.12. With the *lacO22* array and the *lacOA*-gRNA, Cas9-HaloTag displays slower and longer tracks than with the *lacO6* array. Cells containing the Cas9-HaloTag construct, the *lacO6* array and the *lacOA*-gRNA were prepared as described in Figure 4.7. (**A**) Here, we quantified the proportion of detected tracks with D_{app} < 1 µm²/s. The *lacO22* dataset is in purple (37%), the *lacO6* dataset is in yellow (31%), the *ptrA*-gRNA is in red (26%) and the one without a gRNA is in blue (18%). In the presence of the *lacO22* array and of the *lacOA*-gRNA, Cas9-HaloTag displays more slow tracks than in the absence of a gRNA, and with the *ptrA*-gRNA. Error bars indicate the standard deviation calculated on their respective number of cells (24, 34, 12, 28). (**B**) The distribution of track duration of the aforementioned cells follows a similar trend, with increasing longer tracks across datasets. The means of the distributions are indicated and plotted (inlet) with their respective standard deviation (error bars).

Now, CRISPR-Cas9 interactions with DNA can be described in two ways; **ON-target** interactions which are characterized by the CRISPR complex finding the target encoded in the gRNA; and **OFF-target** ones where the CRISPR complex does interact with DNA, but not at its right target. Both processes slow down Cas9, which must momentarily stop its course to unwind the DNA. That is the reason why our measure of slow tracks (with $D_{app} < 1$ µm^2/s) can effectively be a first approach to describe the fraction of the population combining both ON-target and OFF-target events.

A reasonable assumption is that within this combination of ON and OFF-target events, only the proportion of ON-target ones varies as a function of the number of targets present in the cell, because it originates from specific interactions with targets. In contrast, OFF-target events originate from non-specific interactions throughout the genome, which should not be affected by the number of targets.

Bearing those hypotheses in mind, and assuming that the fraction of ON-target events linearly increases as a function of the number of targets (as was also hypothesised in the literature[79]), then we can build an easy set of equations, with a_i and the b_i the respective fractions of OFF and ON-target events with the *lacOi* construct:

$$\begin{cases} a_6 + b_6 = 31\% \\ a_{22} + b_{22} = 37\% \\ a_{22} = a_6 \\ b_{22} = b_6 \times 22/6 = b_6 \times 3.66 \end{cases} \quad \text{resulting in:} \quad \begin{cases} a_6 = 28.75\% \\ a_{22} = 28.75\% \\ b_6 = 2.25\% \\ b_{22} = 8.25\% \end{cases}$$

The result of these equations shows that only 2.25% of all the tracks with the *lacO6* construct would actually be ON-target, which would increase to 8.25% with the *lacO22* construct (Figure 4.13). Even more interesting, in this model, about 29% of the tracks would come from non-specific (OFF-target) DNA interactions, so respectively 13 and 3.5-times more than the specific (ON-target) interactions.

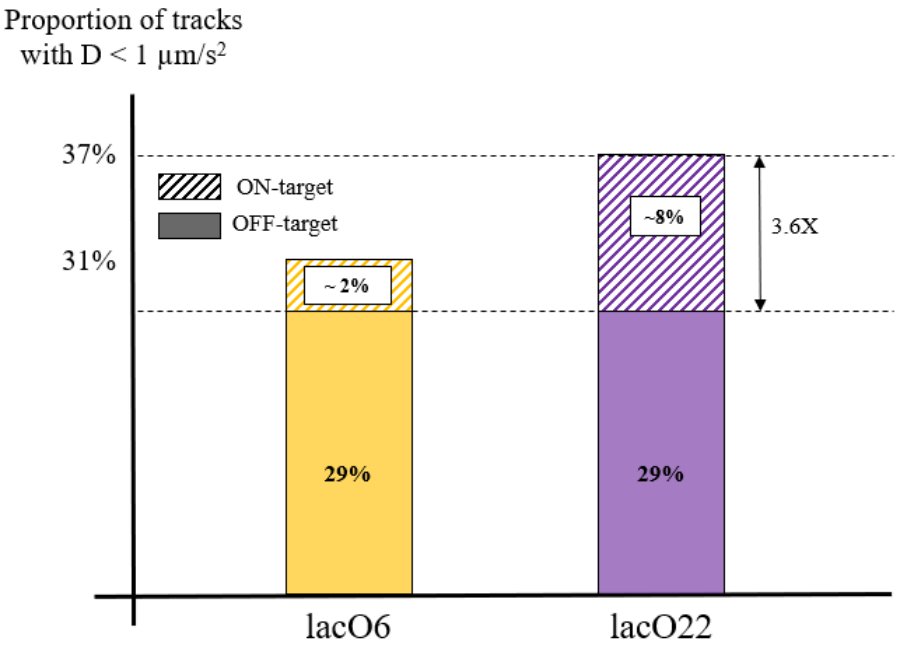

Figure 4.13. We hypothesize the respective proportions of ON and OFF-target tracks with the *lacO6* and *lacO22* CRISPR constructs. Cells containing the Cas9-HaloTag construct, the *lacOA*-gRNA and either the *lacO6* (yellow) or *lacO22* (purple) arrays were prepared as described in Figure 4.12. After quantification of the proportion of detected tracks with D_{app} < 1 µm²/s, we show that respectively 31% and 37% of *lacO6* and *lacO22* tracks are slow. Making the assumption (i) that in both cases, the proportion of OFF-target interactions is the same, and the assumption (ii) that with 22 targets, there should be (22/6 =) 3.6-times more ON-target interactions than with 6 targets, we find that with *lacO6*, only about 2% of the overall tracks actually come from ON-target interactions, compared to about 7% with *lacO22* (and opposed to 29% of OFF-target interactions).

Despite those promising results, the coarse simplicity of our model should be further supported by a more thorough analysis, where quantitative data would be directly extracted from D_{app} distributions. Nevertheless, we can still observe a diffusion shift from 6 targets to 22, thus confirming that the consecutive diffusion contractions we have observed so far are indeed the result of protein-DNA interactions, and not of the gRNA's putative weight increase (see Chapter 3, Discussion).

Finally, after studying the effect of increased amounts of targets recognised by the CRISPR complex and confirming an increase in protein-DNA interactions, it would be particularly useful to compare those previously-showed results to the opposite situation; a CRISPR-Cas9 complex deprived of a genomic target. This time, we should expect a decrease of slow tracks and an acceleration of the protein.

4.3.3 – Tracking without gRNA target

We have just shown that increasing the number of targets recognised by the CRISPR complex increases the amount of slow trajectories, and with it the amount of protein-DNA interactions. However, would the opposite be true? Could we remove any form of target and observe faster protein diffusion? We decided to test this hypothesis by observing cells containing our usual pBAD-Cas9-HaloTag construct, in association with the *lacOA*-gRNA (recognising the *lacO* arrays), but without any integrated array, hence without any proper target.

As mentioned, we expected to see Cas9-HaloTag diffusing faster, freed from all its ON-target interactions (specific target recognition) and only subjected to OFF-target (non-specific) ones. And indeed, Figure 4.14 matches our expectations. Without target, the protein looks slightly faster than with 22 targets, both with MSD and D_{app}.

Further quantifications confirm these observations (Figure 4.15), with the proportion of CRISPR-*lacOA* slow tracks falling to 30% (as opposed to 36%, with 22 targets, as can be seen on Figure 4.15A). This result loosely fits our previous model of ON/OFF-targets, which indeed estimated the proportion of OFF-target activity to be of 29% (Figure 4.13).

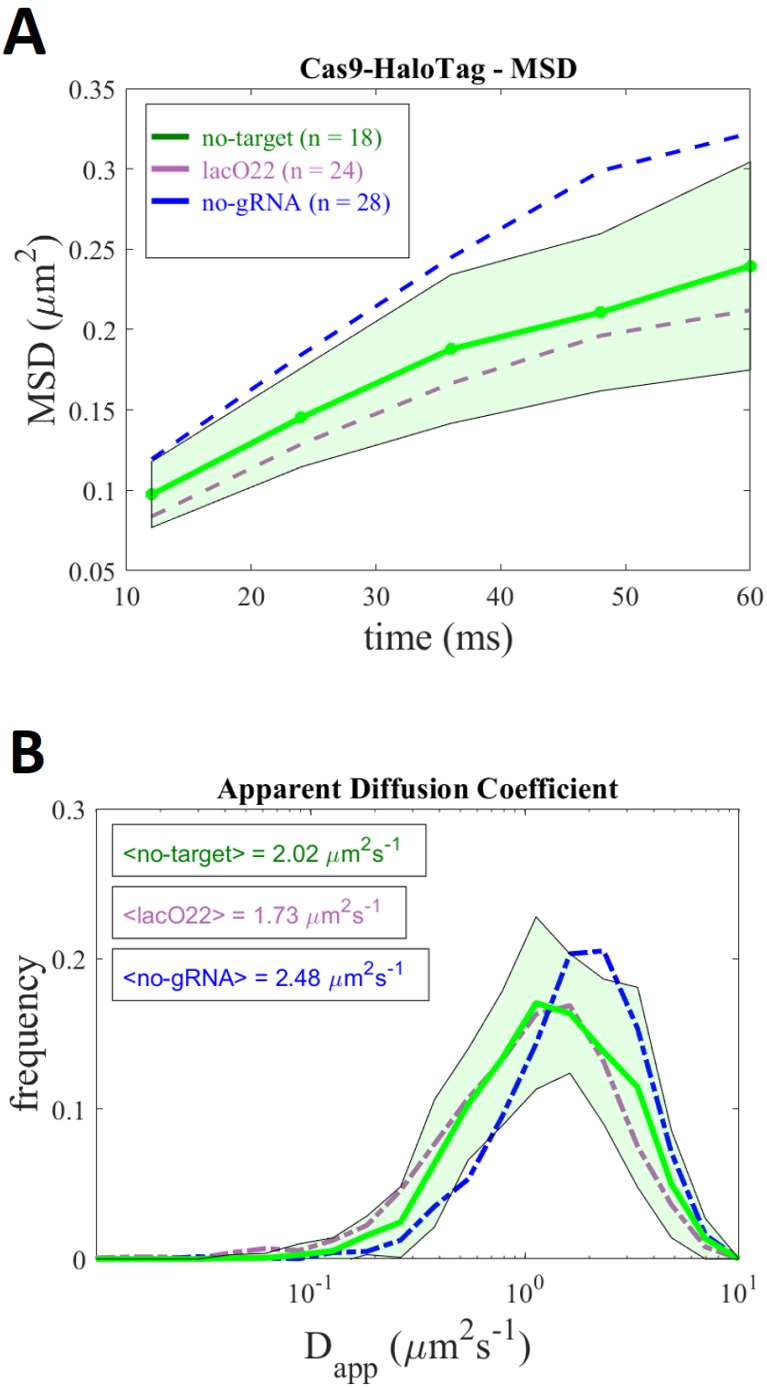

Figure 4.14. Without a target, the CRISPR complex diffuses faster than with a target. Cells containing the Cas9-HaloTag construct and the *lacOA*-gRNA - but without a target (*lacO* array) - were prepared as described in Figure 4.7. (**A**) The MSD is plotted as a function of lag-times (or time, in ms) (in green), compared to previous HaloTag measurements (*lacOA22*-gRNA in purple and no-gRNA in blue). (**B**) Distribution of apparent Diffusion Coefficient (D_{app}) of the aforementioned cells. Mean of the distributions and total numbers of cells (n) are indicated. The green shadings represent the standard deviation (over 18 cells). In the absence of a target (*lacO* array), Cas9-HaloTag and its gRNA appear faster than with the 22 *lacO* targets, and slower than Cas9-HaloTag on its own (without gRNA).

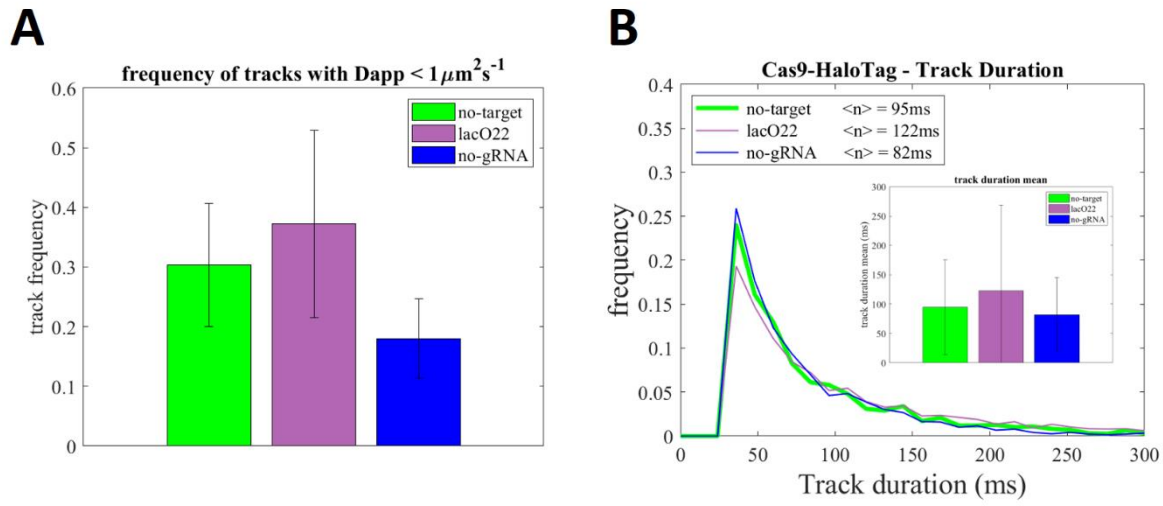

Figure 4.15. Without a target, the CRISPR complex displays an intermediate diffusive behaviour. Cells containing the Cas9-HaloTag construct and the *lacOA*-gRNA - but without a target (*lacO* array) - were prepared as described in Figure 4.7. (**A**) Here, we quantified the proportion of detected tracks with $D_{app} < 1$ $\mu m^2/s$. The aforementioned dataset (no-target) is in green, the *lacO22* dataset is in purple and the one without a gRNA is in blue. In the absence of a target, the CRISPR complex displays more slow tracks than in the absence of a gRNA, but less than in the presence of a target (*lacO22*), which fits our expectations. Error bars indicate the standard deviation calculated on their respective number of cells (18, 24, 28). (**B**) The distribution of track duration of the aforementioned cells follows a similar trend. The mean of the distributions are indicated and plotted (inlet) with their respective standard deviation (error bars).

However, the comparison with other constructs (*lacO6* and *ptrA*) offers a more complex picture. Figures 4.16, 4.17 and 4.18 all indicate that without target, the CRISPR-*lacOA* complex diffuses slower than the CRISPR-*ptrA* complex, despite the clear phenotypic evidence that no DNA double-strand break (DSB) happens in the first case, but does happen in the second (Figure 4.3). This is rather counter-intuitive, as we would expect DSBs to significantly slow down the CRISPR complex, and yet, we seem to see just the opposite.

So then, what is the origin of this discrepancy?

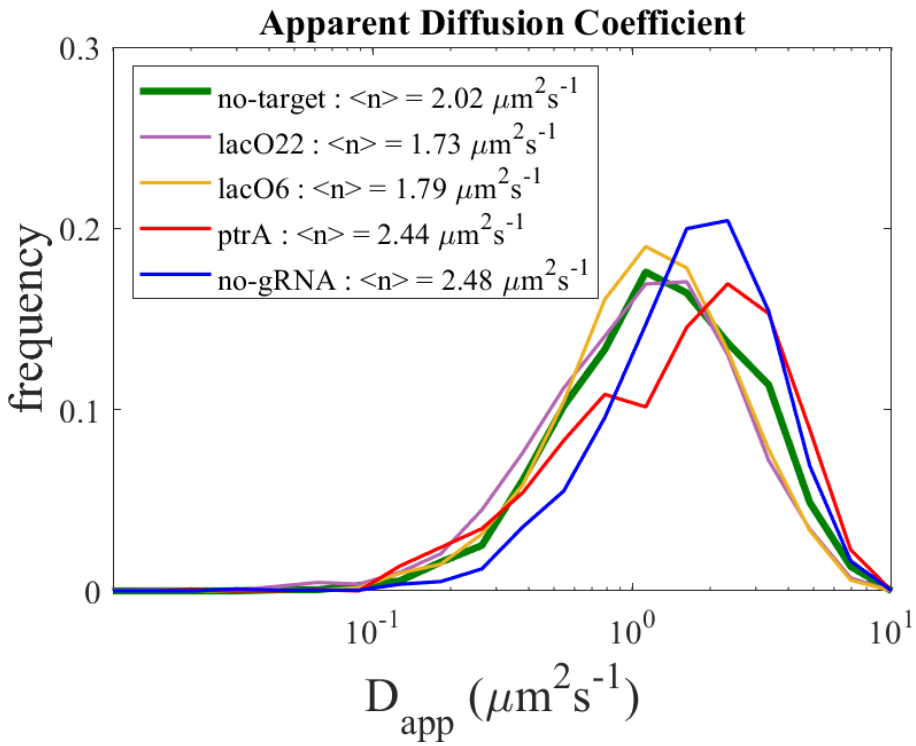

Figure 4.16. Cas9-HaloTag seems to diffuse as a function of its gRNA type, regardless of the presence of a target. Cells containing the Cas9-HaloTag protein were observed as described in Figure 4.7, with either of those constructs; the *lacO22* array and *lacOA*-gRNA (purple), the *lacO6* array and *lacOA*-gRNA (yellow), the *ptrA*-gRNA (red), the *lacOA*-gRNA without target (green) and finally no gRNA at all (blue). The distributions of apparent Diffusion Coefficient (D_{app}) show that Cas9-HaloTag diffusions with the *lacOA*-gRNA cluster together, regardless of the presence of a target, as opposed to the diffusions with *ptrA*-gRNA and without gRNA. Mean of the distributions are indicated.

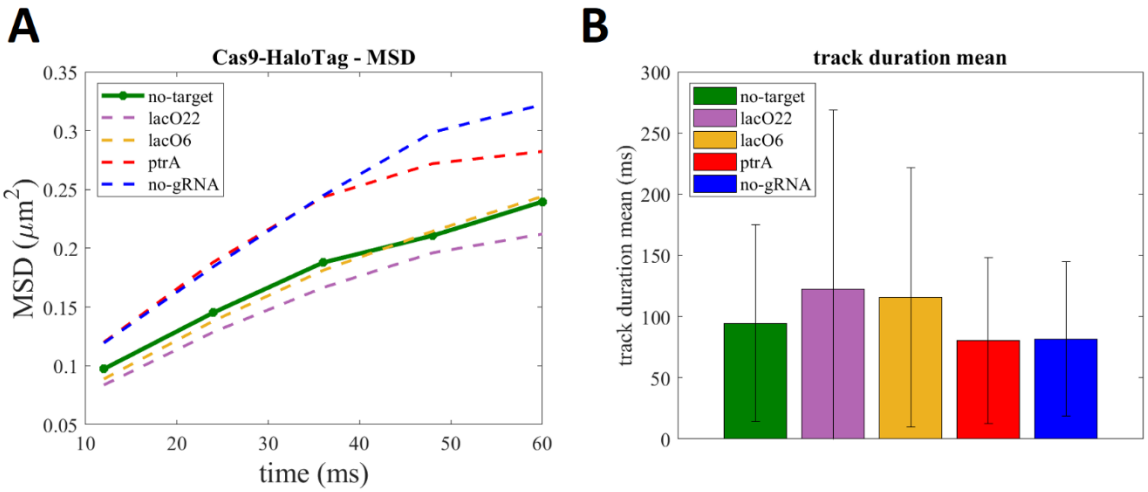

Figure 4.17. Cas9-HaloTag MSD confirms that the protein seems to diffuse as a function of its gRNA type, regardless of the presence of a target. Cells containing the Cas9-HaloTag protein were observed as described in Figure 4.7, with either of those constructs; the *lacO22* array and *lacOA*-gRNA (purple), the *lacO6* array and *lacOA*-gRNA (yellow), the *ptrA*-gRNA (red), the *lacOA*-gRNA without target (green) and finally no gRNA at all (blue). (**A**) The MSD is plotted as a function of lag-times (in ms). (**B**) Track duration means with standard deviations (error bars). CRISPR-*lacOA* constructs similarly diffusive (with or without target), away from the CRISPR-*ptrA* one.

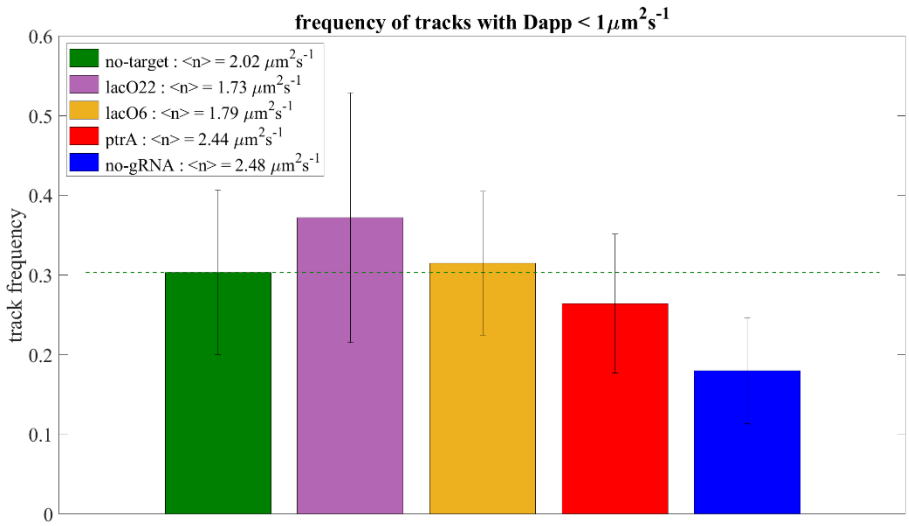

Figure 4.18. With the *ptrA*-gRNA, Cas9-HaloTag displays less slow tracks than the *lacOA*-gRNA and no target. Cells containing the Cas9-HaloTag protein were observed as described in Figure 4.7, with either of those constructs; the *lacO22* array and *lacOA*-gRNA (37%, purple), the *lacO6* array and *lacOA*-gRNA (31%, yellow), the *ptrA*-gRNA (26%, red), the *lacOA*-gRNA without target (30%, green) and finally no gRNA at all (18%, blue). After quantification of the proportion of detected tracks with D_{app} < 1 µm²/s, we can observe that Cas9-HaloTag associated to its *lacOA*-gRNA displays as much slow tracks with and without *lacO* targets. The CRISPR-*ptrA* construct looks faster than all CRISPR-*lacOA* ones. Error bars represent standard deviations.

So, why does the CRISPR-*lacOA* complex diffuse slower than the CRISPR-*ptrA* complex? Could the level of CRISPR-*ptrA's* OFF-target interactions be lower than that of CRISPR-*lacOA's*? A first explanation could stem from a different GC content, which may facilitate the interactions of one gRNA with genomic DNA over the other. Unfortunately, the *ptrA* target sequence shows 45% GC content, compared to 50% with the *lacOA* sequence, thus showing easier DNA interactions with *ptrA*, in opposition to the faster diffusion we observed with this construct. We can therefore refute this hypothesis.

A second explanation relates to the level of nucleic acid similarities found across the genome, which change from one gRNA to another, due to their different target sequences. Indeed, more similarities means more interactions; could the *lacOA* target have more similarities than *ptrA*?

BLASTing the two target sequences against *E.coli*'s genome, we first found no significant differences (Figure 4.19). However, restricting our observation to long mismatches (longer than 15bp), the *lacOA* sequence displayed twice more hits than its *ptrA* counterpart (Figure 4.20). Indeed, long mismatches are more likely to stabilize the CRISPR complex on the DNA, in accordance with the gRNA 'seed region' (of 12 nucleotides[19]). Potentially, such interactions may even block Cas9 on the DNA as it neither gets the signal to cut (without full recognition, Cas9 does not rearrange itself to DNA-cleavage conformation[19]), neither the signal to move out (because of crRNA-DNA base-pairing).

Therefore, we can now confirm that the *lacOA*-gRNA most likely triggers significantly more CRISPR OFF-target (non-specific) interactions than its *ptrA* counterpart.

Additionally, as we estimated the level of ON-target interactions stemming from the *lacOA*-complex to reach a mere 2% of the total amounts of tracks, with 6 targets (Figure 4.13), it is reasonable to assume that the level of ON-target interactions reached with the unique target of the *ptrA*-complex would certainly not exceed this amount.

Put together, those observations show that the CRISPR-*ptrA* construct displays both small amounts of OFF-target interactions (indicated by its few long genomic mismatches), and small amounts of ON-target interactions (due to its unique target sequence) which explain that despite the presence of a target, the CRISPR-*ptrA* construct still appears faster than the CRISPR-*lacOA's* lack of one.

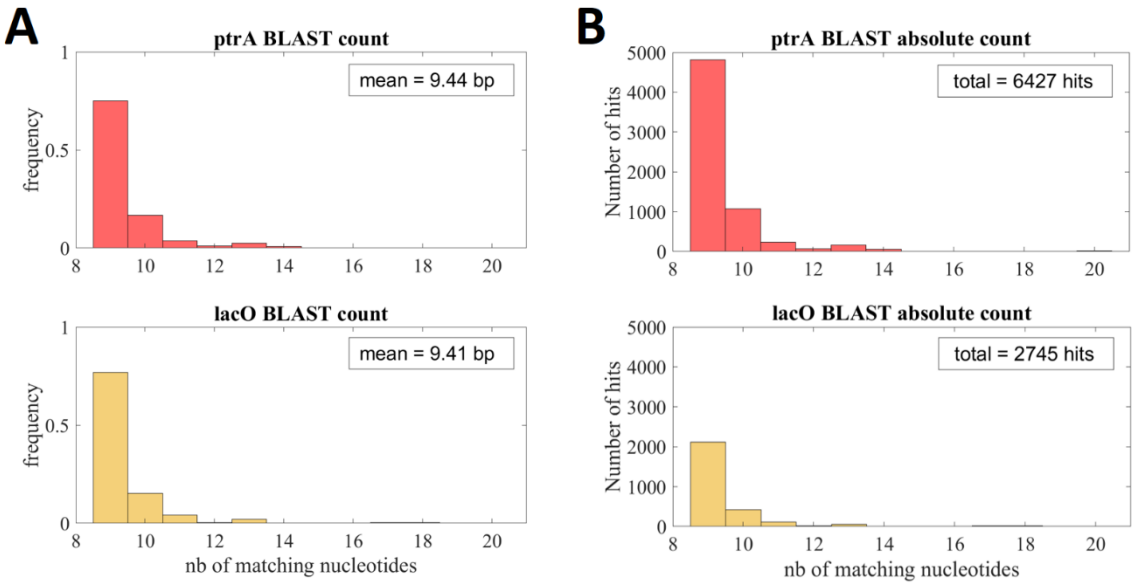

Figure 4.19. Genomic mismatches of the *ptrA*- and *lacOA*-gRNA target sequences. The 20-nucleotides target sequences of the *ptrA*-gRNA (red) and of the *lacOA*-gRNA (yellow) were BLASTed against *E.coli*'s genome. (**A**) Distribution of numbers of matching nucleotides, normalized to 1. (**B**) Distribution of numbers of matching nucleotides, in absolute numbers. Mean of the distributions and total numbers of BLAST hits are indicated.

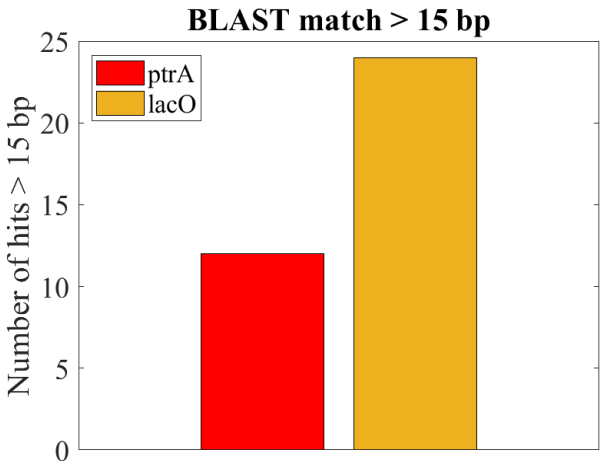

Figure 4.20. The *lacOA*-gRNA displays more long mismatches than the *ptrA*-gRNA. The 20-nucleotides target sequences of the *ptrA*-gRNA (red) and of the *lacOA*-gRNA (yellow) were BLASTed against *E.coli*'s genome. Here are displayed the numbers of hits with at least 15-base-pair (15bp) homology, for each gRNA. The *ptrA* target sequence displays 12 hits and the *lacOA* one displays twice more (24 hits).

In conclusion, we have shown that increasing the number of targets that a gRNA can recognise increases the CRISPR construct's chances to (specifically) interact with DNA. However, surprisingly, we have also shown that *non-specific* DNA interactions, based on the gRNA target sequence itself, are actually more relevant when it comes to protein diffusion.

Now that we have identified this discrepancy, we aim to further quantify the ON/OFF balance of CRISPR interactions by dramatically increasing the numbers of targets, no longer by a factor of 10, but by a factor of 100.

4.3.4 – Tracking with large amounts of plasmid targets

In the previous parts of this chapter, we have shown that Cas9-HaloTag can adopt several diffusion behaviours which are affected by the gRNAs with which the protein associates. We have also shown that in those conditions, the strongest feature of the CRISPR complex's diffusion was not its specific interactions with DNA, but rather its non-specific ones.

However, these observations have been made with small numbers of targets; in the case of our *lacOA*-gRNA, we have built up to 22 actual targets, as opposed to 24 long non-specific interactions, and almost 3,000 short non-specific ones (Figure 4.19B). Therefore, our question is the following; can we tilt the balance of CRISPR ON/OFF interactions by adding much higher numbers of targets? To address this question, we built a plasmid containing the *lacO22* array, sitting on a high-copy origin of replication (pUC) reaching hundreds of copies per cell. In effect, this new system should produce, on average, around 2,000 *lacO* targets per cell, thus reaching a level similar to the amount of short genomic sequences non-specifically recognised by the *lacOA*-CRISPR complex.

A first observation shows that despite much larger numbers of *lacO* targets, both MSD (Figure 4.21) and D_{app} (Figure 4.22) seem rather unperturbed. The MSD curve almost entirely overlaps with *lacO22*'s, as if no new targets had been added. Similarly, the apparent Diffusion Coefficient only recorded a meagre 1.7% decrease from the previous one obtained with solely 22 genomic targets.

Considering the very small effect of the addition of the plasmid, I decided to assess the recognition of the targets present on the plasmid by the *lacOA*-CRISPR complex. To do so, I tracked Cas9-HaloTag in cells containing the *lacOA*-gRNA, with targets only present on the plasmid (and not on the genome).

Figure 4.21 shows that despite the absence of genomic targets, Cas9-HaloTag diffusion is still very slow, and even slower than its non-specific diffusion behaviour (where there is no target at all). These results indicate that our CRISPR complex recognises targets from plasmid DNA (as it was originally meant to), albeit not as strongly as we expected, given the considerable amount of targets coming from the plasmids. Indeed, this last experiment confirms that specific targets are, even in great number, surprisingly not the driving force of the CRISPR complex diffusion.

Additionally, in the *absence* of genomic targets and in the *presence* of plasmid ones, no SOS response was triggered (data not shown), despite Cas9-HaloTag diffusion slowing down, confirming that previous effects observed on protein diffusion did not stem from SOS activation (whether we talk of protein regulation, division arrest or even crowding effects), but really from protein-DNA interactions.

Another thing to consider is the actual presence of the plasmids, for indeed, due to homologies with either the endogenous *lacO* sequence, or with the integrated *lacO* arrays, recombination events may have provoked significant variations in target numbers, or even their complete removal. Such eventualities could be investigated by plasmid extraction and further verification.

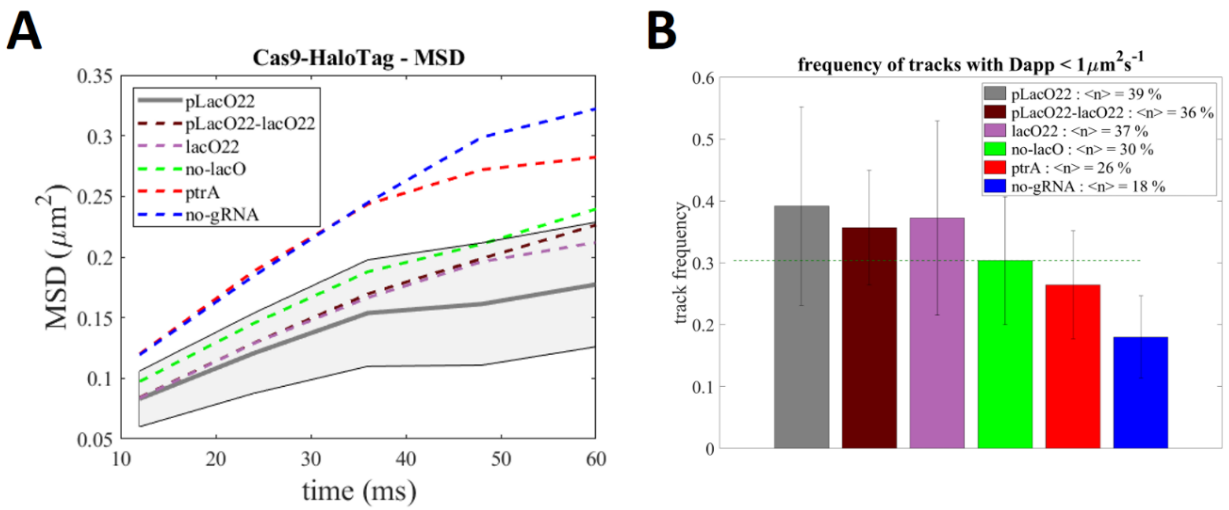

Figure 4.21. The *lacO* targets on the plasmid are recognised by Cas9-HaloTag. Cells containing the Cas9-HaloTag protein and the *lacOA*-sgRNA were observed as described in Figure 4.7, with either of the following constructs; the *lacO22* array on a (high-copy) plasmid (pLacO22, grey), the *lacO22* array both on a plasmid and on the genome (pLacO22-lacO22, brown), the *lacO22* array only on the genome (*lacO22*, purple) and finally, no *lacO* array on the genome (green). As a comparison, diffusion from cells with the *ptrA*-gRNA (red) and without any gRNA (blue) are also displayed. (**A**) The MSD is plotted as a function of lag-times (or time, in ms). The grey shading represents the standard deviation of the pLacO22 dataset (over 16 cells). (**B**) Quantification of the proportion of detected tracks with D_{app} < 1 µm²/s (individual numbers are indicated in the inlet). We can observe that when the CRISPR complex only targets plasmid DNA, Cas9-HaloTag diffusion slows down to a similar level as when the complex only targets genomic DNA, which is slower than its OFF-target behaviour (no-*lacO*, highlighted by the green dashed line). Error bars represent standard deviations.

Moreover, we will remember that in those conditions, Cas9-HaloTag is expressed in very low amounts (barely more than 5 proteins/cell, as showed in Chapter 3, Figure 3.12 and 3.14), which could in turn produce a saturation effect, where not enough proteins would be available to bind additional targets. This hypothesis would explain why above 22 *lacO* targets, Cas9 diffusion barely changes.

Initially, the reason behind this very low expression level was precisely to perform Single-Particle Tracking (SPT, see Materials and Methods), but in the light of those last observations, we could increase the expression level of our protein and adapt our imaging conditions by labelling less proteins (under-labelling). That way, we would aim to obtain a similar number of labelled proteins per cell (as required by SPT), while providing suitable conditions for the detection of differences between varying numbers of targets.

That said, in those conditions where the number of targets is maximal, Cas9 diffusion reached its slowest point (with, on average, D_{app} = 1.76 µm^2/s and the fraction of slow tracks reaching 39%). Conversely, the fastest point was reached with Cas9 diffusing without gRNA (averaged D_{app} = 2.48 µm^2/s with slow tracks = 18%). Put together, those results indicate that the range of diffusion variations we can detect barely covers a 2-fold difference. This is not great sensitivity.

Now, as highlighted on Figure 4.22, our protein diffusion behaviour is very much affected by the target sequence of the gRNA involved in the CRISPR complex; we clearly see the *lacOA*-CRISPR constructs clustering together, away from the *ptrA*-CRISPR one. As discussed before, this feature originates from the gRNAs' respective OFF-target interactions, with *lacOA*-gRNA interacting much more with genomic DNA than its *ptrA* counterpart.

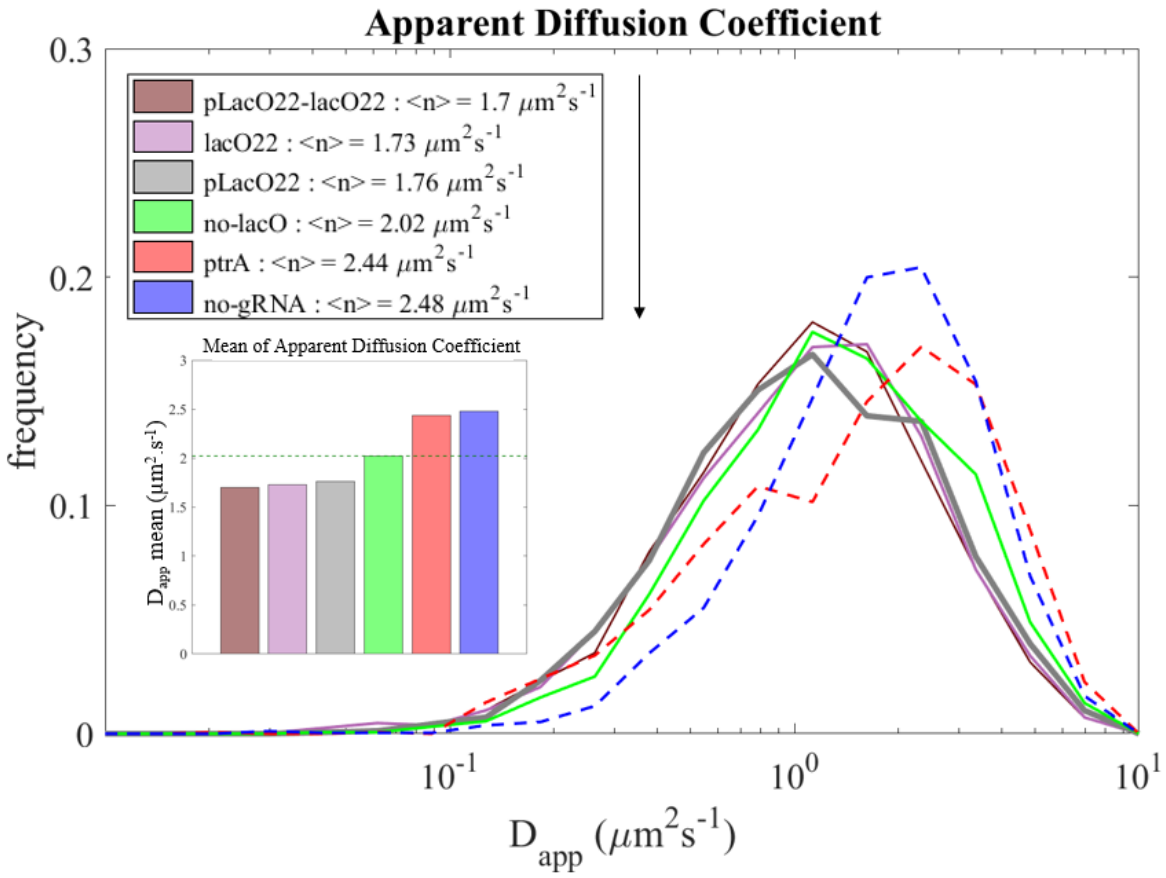

Figure 4.22. Cas9-HaloTag diffusion in various conditions. Cells containing the Cas9-HaloTag protein and the *lacOA*-sgRNA were observed as described in Figure 4.7, with either of the following constructs, in order of increasing apparent Diffusion Coefficient (D_{app}); the *lacO22* array both on a plasmid and on the genome (pLacO22-lacO22, brown), the *lacO22* array only on the genome (*lacO22*, purple), the *lacO22* array only on a plasmid (pLacO22, grey) and finally, no *lacO* array at all (green). As a comparison, diffusion from cells with the *ptrA*-gRNA (dotted red) and without any gRNA (dotted blue) are also displayed. Here, we can see the distributions of D_{app} with their respective means (indicated in the legend and plotted in bottom inlet). The dotted green line indicates the separation of the *lacO* constructs with the two other constructs.

Therefore, an additional direction to explore to further quantify CRISPR-Cas9 diffusion behaviours may be to look for a gRNA which triggers less OFF-target (non-specific) interactions, to start with a faster profile which would then get progressively slowed down by the addition of increasing numbers of targets. In other words, the slowest D_{app} would remain the same (corresponding to the maximal amount of targets, and its maximal amount of ON-target interactions), but CRISPR-Cas9 non-specific diffusion (OFF-target) would be faster, allowing a gradual increase in target numbers and a correspondingly gradual decrease in diffusion measure.

Another possibility could be to increase the range of diffusion behaviours by suppressing OFF-target interactions entirely, either using anti-CRISPR proteins (preventing DNA interactions)[49], or by removing all genomic content. Such goal can be achieved in a *ΔrecA* background, where cells put in the presence of DNA Double-Strand Breaks (DSBs) cannot stop the RecBCD enzyme from digesting their entire chromosome. Deprived of genomic DNA to interact with, Cas9 is left entirely free to diffuse in the cell[104]. In that scenario, one could also study ON-target interactions, using plasmid targets; indeed, the comparatively short plasmid sequences would deprive the gRNA from any unwanted homology with its target, leaving only a few available PAM sequences for Cas9 to interact with (which again would be far less numerous than in the genome).

As a matter of fact, we tested this last hypothesis by observing Cas9-HaloTag in such *ΔrecA* background and induced DSBs using ciprofloxacin (4ng/µl), thus removing the chromosome through the aforementioned mechanism. I checked the absence of DNA content with DAPI staining and analysed the tracks detected from such cells. In Figure 4.23A, we can observe that such tracks are indeed 24% faster than those displayed by Cas9-HaloTag with intact genomic DNA (and without gRNA).

Additionally, Figure 4.23B shows that the Mean Square Displacement of the protein without genomic DNA reaches a stronger plateau much sooner (containment effect), again matching our expectation of a freely diffusion protein, reaching the edges of the cell much faster.

Finally, some 8% of total tracks can be considered “slow”, characterized by an apparent Diffusion Coefficient (D_{app}) slower than 1 µm^2/s. That is interesting, considering that those proteins should not have any genomic DNA to interact with anymore. We can assume that those remaining slow tracks emerge from some proteins diffusing closer to the cell membrane, or perhaps trapped in inclusion bodies (due to potential misfolding). However, we can also note that 92% of the tracks are actually faster than 1 µm^2/s, confirming the chosen threshold as a good indicator of the difference between slow and fast Cas9-HaloTag proteins.

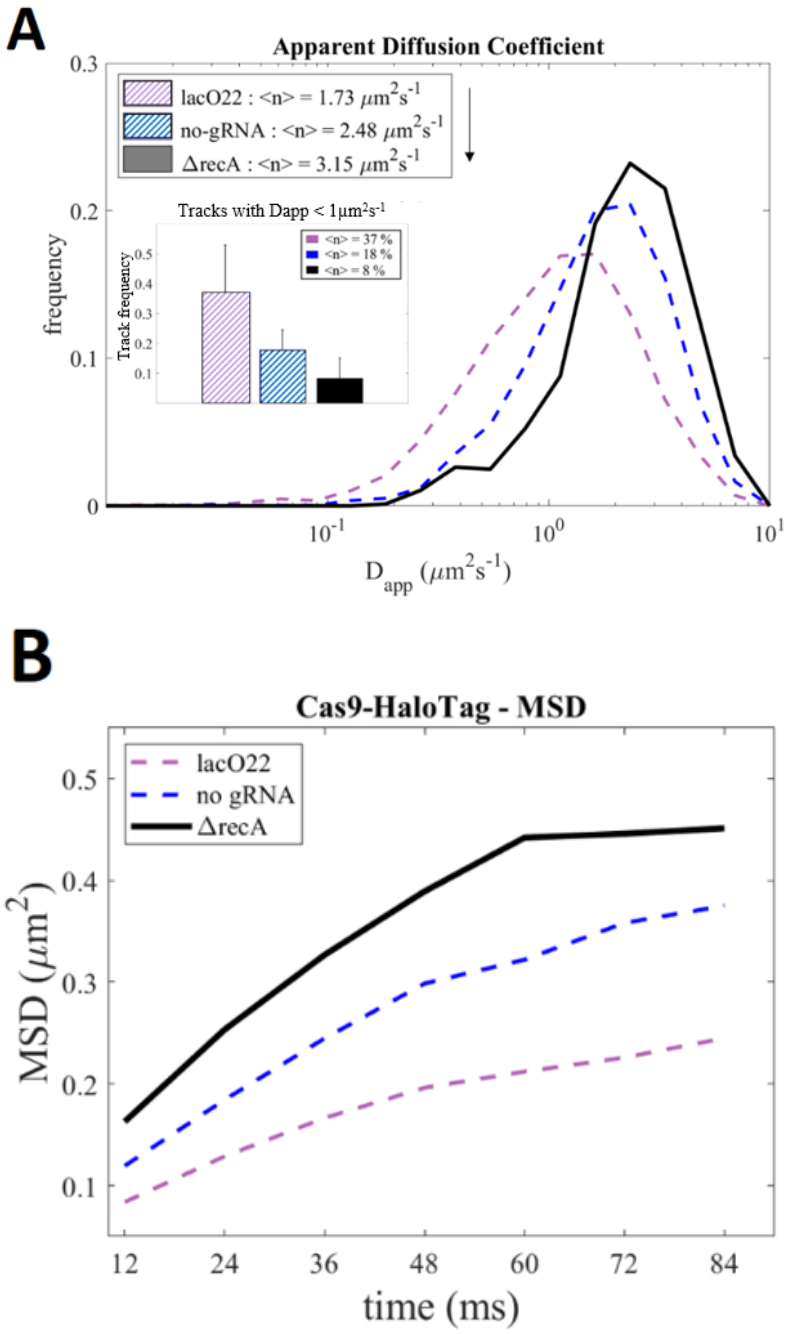

Figure 4.23. Cas9-HaloTag diffusion without genomic DNA is faster. Cells containing the Cas9-HaloTag protein in ΔrecA background were exposed to ciprofloxacin (4ng/µl) for one hour so as to get rid of their genomic DNA. Consequently, Cas9-HaloTag expression was induced and the protein labelled. Finally, the cells were observed as described in Figure 4.7, while a DAPI staining allowed the identification of cells deprived of DNA content. Those cells are here represented in black, while as a comparison, are also displayed cells containing the lacO22 array with the associated lacOA-sgRNA (in purple) and those not containing any gRNA at all (in blue). (**A**) The distributions of D_{app} with their respective means (indicated in the legend) and their respective proportions of slow tracks (with D_{app} < 1 µm²/s) (**inlet**, with indicated frequencies) show that without genomic DNA (ΔrecA), Cas9-HaloTag diffuses faster. (**B**) Likewise, this construct displays the fastest MSD.

4.4 – Discussion and Future Work

A model of Cas9 diffusion behaviours

We can now associate the data we have collected and quantified from Cas9-HaloTag diffusion in different conditions and explore whether our previous model can be further improved. Putting side-by-side the proportions of slow tracks we detected from Cas9-HaloTag with-or-without gRNA, with-or-without target and with-or-without genomic DNA, it is possible to infer the respective fractions of protein behaviours eventually contained in a tracking dataset.

Indeed, when associated to a gRNA and expressed in the presence of a genomic target, Cas9-HaloTag can adopt four diffusion behaviours. The first and simplest behaviour is 'free diffusion', where the protein can randomly diffuse in the cell and does not interact with DNA, which we can attribute to our ΔrecA dataset. The second behaviour describes the interaction Cas9 has on its own with DNA, through PAM recognition, but without gRNA. The third behaviour relates to so-called 'OFF-target' interactions, when Cas9-HaloTag is in a complex with a gRNA, interacts with DNA but not at its rightful target; these interactions can be observed in the 'no-target' dataset, when Cas9-HaloTag is associated with the *lacOA*-gRNA, but when no *lacO* target is present in the cell. Finally, the last behaviour appears when Cas9 does find its target or 'ON-target' interaction; here, we can use any such dataset which has both a gRNA and the target.

As mentioned earlier, we speculated on a simple model of ON/OFF-target quantification, assuming that among CRISPR-*lacO* datasets with varying number of targets, the proportion of OFF-target interactions should not vary (Figure 4.13). As a reminder, we reasoned that the OFF-target interactions being a result of genomic homologies, and not of actual targets, the number of targets should not impact the OFF-target fraction of total detected tracks.

Based on this assumption, we decided to compare our CRISPR-*lacO* datasets, making the additional hypothesis that all four diffusion behaviours were independent from each other, thus mutually exclusive. In other words, taking the example of the *lacO22* dataset, we reasoned that its proportion of slow tracks (36%) included four smaller fractions; (a) one of proteins diffusing in the cytoplasm but confined to a small area (*e.g.* the vicinity of the membrane), (b) another of proteins merely interacting with PAM sequences, (c) one characterized by OFF-target interactions and (d) a last one with ON-target events.

Applying the same reasoning to the *no-target*, the *no-gRNA* and the *no-genomic-DNA* datasets, respectively containing three (a, b & c), two (a & b) and one (a) of the four previous behaviours, it is possible to identify their respective fractions of slow tracks.

Of course, this model only constitutes a first approach towards the proper quantification of Cas9 diffusion behaviours and should be supported by a statistical analysis of our D_{app} distributions, much like Kapanidis and colleagues[110] did.

Nevertheless, this approach still offers a first quantitative glance, as illustrated on Figure 4.24, where our model infers that those four behaviours respectively occupy 8% (confined diffusion), 10% (Cas9-PAM interactions), 11% (OFF-target) and 7% (ON-target) of the total amounts of tracks detected by Single-Particle Tracking.

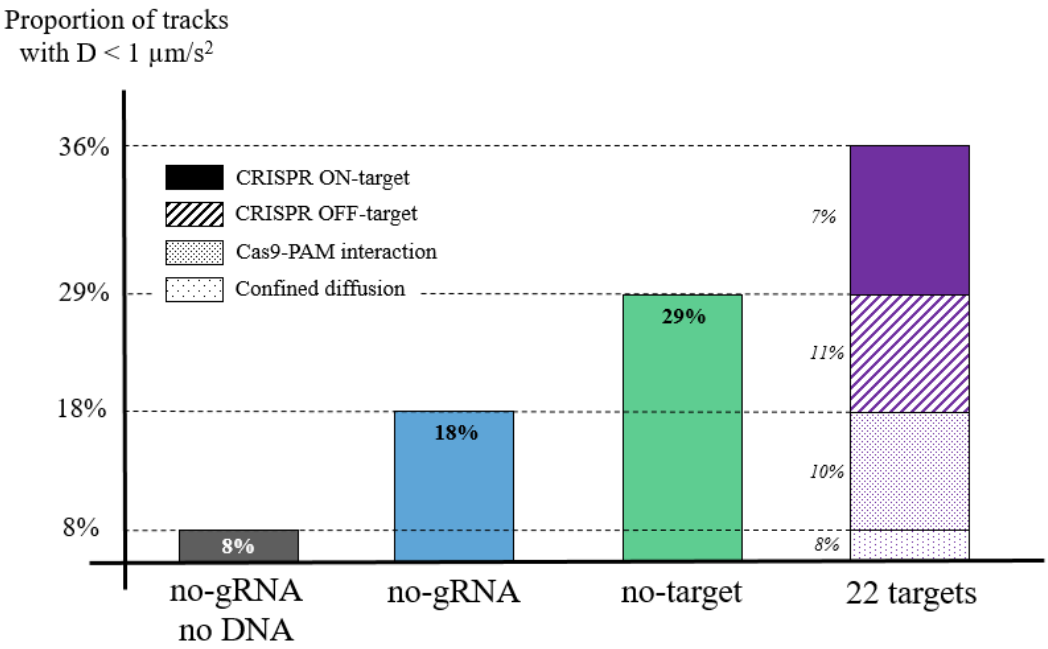

Figure 4.24. A quantification of Cas9 various diffusion behaviours. Here displayed are the proportions of slow tracks (with D_{app} < 1 µm²/s) detected from cells containing the Cas9-HaloTag protein with either of the following constructs; no gRNA and no genomic DNA (ΔrecA dataset, black), no gRNA (blue), the *lacOA*-gRNA without *lacO* target (green) and the *lacOA*-gRNA with the *lacO22* array (purple). On the *lacO22* dataset (purple) are displayed the different diffusion behaviours the protein can adopt (see legend), deduced from the three other datasets, with their respective quantifications. The plot is not to scale.

Cas9 diffuses in the cytoplasm, searching for DNA before searching for a target

Another observation pertains to the 'non-visible' part of our previous plot, the 64% of *lacO22* tracks appearing to be faster than our 1 $\mu m^2/s$ threshold, thus displaying freely-diffusing behaviour.

This observation seems to stand in contradiction with our previous hypothesis of Cas9-HaloTag reaching target saturation, due to insufficient protein expression. Our previous hypothesis stated that because of a lack of Cas9 proteins in the cell, increasing the number of targets would not affect Cas9 apparent diffusion, as most of the proteins would already be interacting with their targets. And yet, even with thousands of targets (on plasmids), we still observed a significant portion of freely-diffusing Cas9 proteins, confirming that our initial hypothesis of Cas9 saturating was incorrect.

Nevertheless, this still does not explain why more than half of Cas9 proteins freely diffusive in the cytoplasm, when plenty of targets are available.

As a matter of fact, before searching for a target on the genome, Cas9 proteins first need to find the genome…! We can imagine Cas9, in complex with its gRNA, diffusing in the cytoplasm and eventually encountering genomic DNA. However, we ignore how long this simple process takes. Previous studies[79] have measured Cas9 searching time to be of about 6 hours; and yet of those 6 hours, we do not know what actual time Cas9 spent diffusing in the cytoplasm, and how much time it spent on the DNA. Our results suggest that of those 6 hours, a lot of time is spent in the cytoplasm, where Cas9 most likely enters in cycles of quick DNA-interactions followed by free-diffusion steps.

Our measurements do not differentiate between ON and OFF-target events

We have just shown that the first reason why we do not observe a dramatic diffusion difference when increasing target numbers is probably due to Cas9's predominant genomic-searching time, where the protein freely diffuses in the cytoplasm, unaffected by DNA and thus unaffected by target numbers. However, every time Cas9 encounters the genome, it probes for targets, and this type of interaction should be affected by target number. So then again, should we not expect a larger diffusion difference with larger target numbers?

A second answer to this question relates to one of our previous assumptions, the 'independence of OFF-target interactions from varying target numbers', due to OFF-target events originating from non-specific DNA interactions across the genome (thus unrelated to specific ones). We also initially considered the respective fractions of OFF-target and ON-target events to be independent from each other, but are they really?

In fact, we can imagine that if the number of targets increases, then every time Cas9 interacts with DNA, it has more chances to encounter a target – or 'specifically' interact – and thus less to 'non-specifically' do so. Practically speaking, this means that increasing the fraction of ON-target should decrease the fraction of OFF-target.

Moreover, our method of choice to identify slow tracks (with a D_{app} threshold) cannot differentiate between specific and non-specific interactions, which put in relation with our previous reasoning, would mean that increasing target numbers may very well increase the amount of ON-target interactions, but would correspondingly decrease the amount of OFF-target ones, which would result in an apparently unaffected fraction of slow tracks, effectively dissimulating any significant change, as illustrated on Figure 4.25.

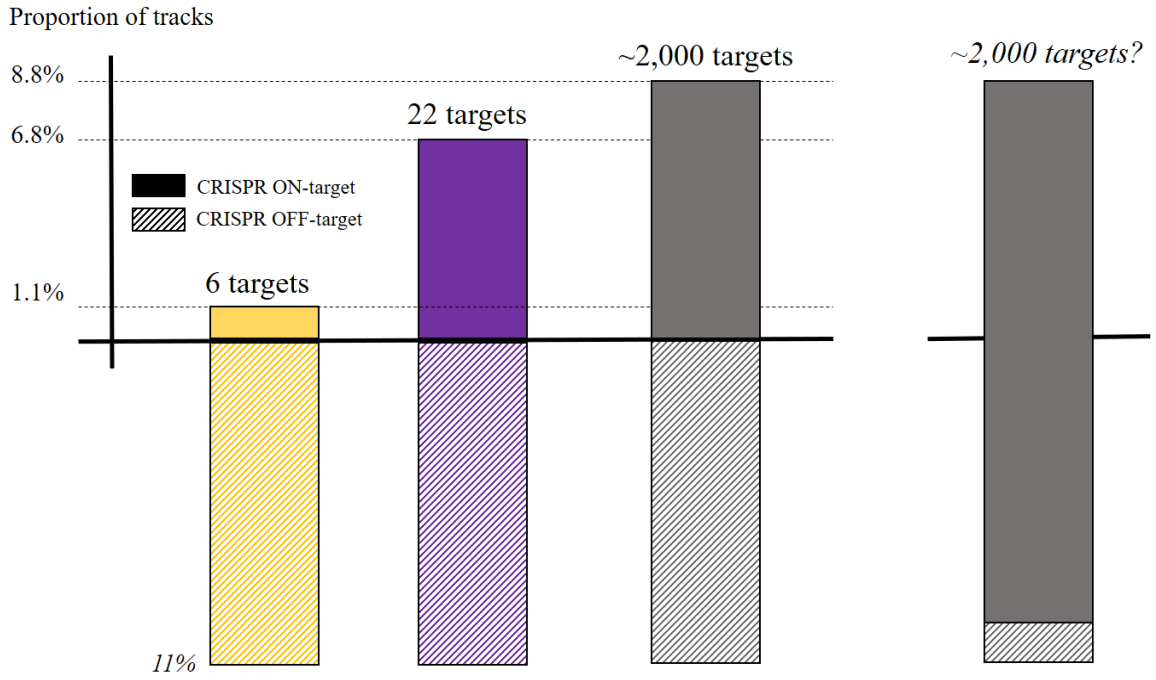

Figure 4.25. 'ON+OFF'-target quantification. Here displayed are the proportions of 'ON+OFF' target tracks, identified as the proportions of slow tracks (with D_{app} < 1 µm²/s) with gRNA, thus above the 'no-gRNA' fraction quantified in Figure 4.24 (>18%). The datasets come from Cas9-HaloTag cells with *lacOA*-gRNA and the following constructs; the *lacO6* array (yellow), the *lacO22* array (purple) and the plasmid carrying the *lacO22* array (black). Plain rectangles represent the estimated proportions of ON-target tracks, whereas the striped rectangles represent the OFF-target tracks. If we consider the proportion of OFF-target tracks identical in all datasets, we do observe an increase in ON-target tracks as the number of targets increases. However, the right-hand panel shows how an increase in the fraction of ON-target events may actually decrease the proportion of OFF-target ones, without affecting the overall proportion of slow tracks.

Put together, these two reasons can explain why in the present conditions, we do not observe striking diffusion differences between samples with varying target numbers. One is that Cas9 appears to spend most of its searching time away from DNA, freely diffusing. The second is that when Cas9 does find DNA, our current method cannot detect the difference between specific and non-specific interactions; we thus remain oblivious to changes in ON/OFF-target variations, actual key of specific diffusion change.

In order to single out ON and OFF-target behaviours, one potential method relies on the assumption that the fraction of ON-target tracks linearly increases as a function of target number, before reaching a plateau, due to actual saturation. This hypothesis, already described in the literature[79], assumes that the interactions of the CRISPR-complex with different targets are independent from each other and constrained by both the number of proteins and the number of targets. Following this putative behaviour, we mathematically derived (data not shown) that an additional array with a different number of targets would prove essential to truly extract the respective fractions of ON and OFF-target tracks from our current quantifications.

Our measurements can still quantify Cas9 behaviours

Despite our inability to quantify absolute fractions of "ON *and* OFF"-target interactions, we can still quantify the fractions of "ON *or* OFF" interactions, which indeed show differences between samples with various amounts of targets, and notably between 6 and 22 targets (Figure 4.25). We observe that *lacO6* "ON or OFF" fraction is of 12%, and the *lacO22* one of 18%, resulting in a 50% increase, showing that Cas9 did increase its proportion of CRISPR-DNA interactions as a result of a higher target number.

Also, we still have the ability to identify other fractions of diffusion behaviour. For instance, in the *lacO22* dataset, we identify 64% of freely-diffusing protein, 8% of confined proteins, 10% of Cas9 interacting with PAM sequences and 18% of CRISPR-Cas9-gRNA interactions ("ON *or* OFF" interactions, Figure 4.26). Put together, those results indicate that more than half of Cas9's lifetime is spent freely diffusing in the cytoplasm. Then when Cas9 interacts with DNA, at least half of those interactions relate to the CRISPR complex inserting the gRNA to check for targets, while a third relates to the protein simply recognizing PAM sequences[111].

Regarding diffusion absolute quantification, I have also shown how freely diffusing Cas9-HaloTag displays an averaged D_{app} of 3.15 $\mu m^2/s$, first ever measure of Cas9-HaloTag diffusion coefficient *in vivo*. Another experiment would consist in observing Cas9-HaloTag in a ΔrecA background, this time with the addition of a plasmid containing the targets, so as to ensure that the majority of Cas9 interactions with DNA would be the result of ON-target interactions. Such an experiment would have the advantage of quantifying the diffusion of the protein bound to its DNA target which, as of now, has never been accurately quantified.

Finally, further fitting and statistical work should refine the parameters (fraction and D_{app}) of its additional DNA-interacting behaviours, following previous work done on other proteins, found in the literature[71,112].

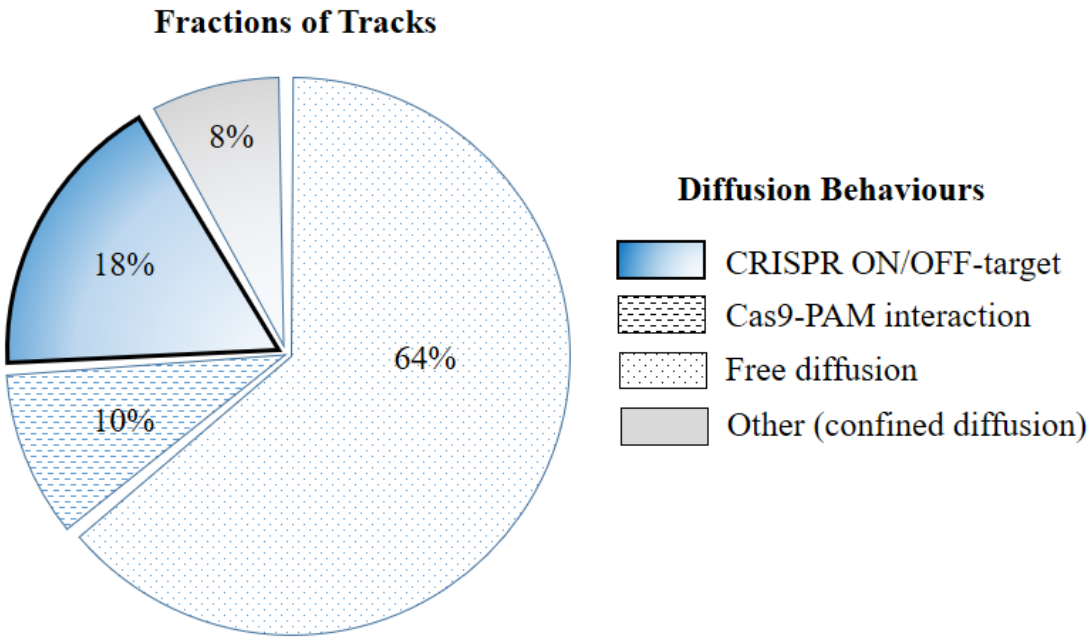

Figure 4.26. Fractions of Cas9 diffusion behaviours. Here displayed are the proportions of tracks quantified from the *lacO22* dataset, detected from Cas9-HaloTag cells with *lacOA*-gRNA and the *lacO22* array, with their respective diffusion behaviours (described in the legend).

In conclusion, this chapter shows how one can successfully use different arrays of targets to draw the CRISPR complex to cutting DNA, while offering a range of different conditions to study protein diffusion. Indeed, this setup allowed me to quantify Cas9 diffusion with-and-without gRNA, with-and-without target and, eventually, with-and-without genomic DNA, so as to as identify, study and quantify its four main diffusion behaviours.

Chapter V – Discussion

Summary

The aim of this project was to study the interactions of CRISPR-Cas9 with DNA, both specific (ON-target) and non-specific (OFF-target). For this purpose, we used single-molecule microscopy and single-particle tracking to follow Cas9 displacement over time, within *E. coli* cells.

Using a range of genetic constructs integrated into the chromosome (and on plasmids), we managed to detect, identify and quantify the various diffusion behaviours the protein can adopt, themselves associated to determined types of protein interactions with DNA.

Those behaviours can be summarized in five main types:

- **Free diffusion**, where the protein randomly diffuses in the cytoplasm.
- **Constrained diffusion**, where the protein diffuses in a limited area, such as the vicinity of the cell membrane.
- **Cas9-PAM** interactions, where apo-Cas9 (without gRNA) interacts with DNA, only recognising PAM sequences.
- **OFF-target** interactions, where the Cas9-gRNA complex recognises PAM sequences, unwinds the DNA but not at the right target site encoded in the gRNA sequence.
- **ON-target** interactions, where the Cas9-gRNA complex interacts with its rightful target and subsequently cleaves the DNA.

Quantifications

Throughout this work, we managed to measure absolute values of apparent diffusion coefficients (D_{app}) of the Cas9-HaloTag protein, in different conditions. The first one corresponds to Cas9-HaloTag freely diffusing, which we measured by observing the protein on its own, in the absence of genomic DNA (ΔrecA background), to be of 3.15 $\mu m^2/s$, on average. Interestingly, a comparison with the freely diffusion HaloTag protein (4.10 $\mu m^2/s$) shows that Cas9-HaloTag was still slower even in the absence of DNA and gRNA, which matches our expectations as Cas9-HaloTag is significantly heavier than HaloTag (Figure 5.1).

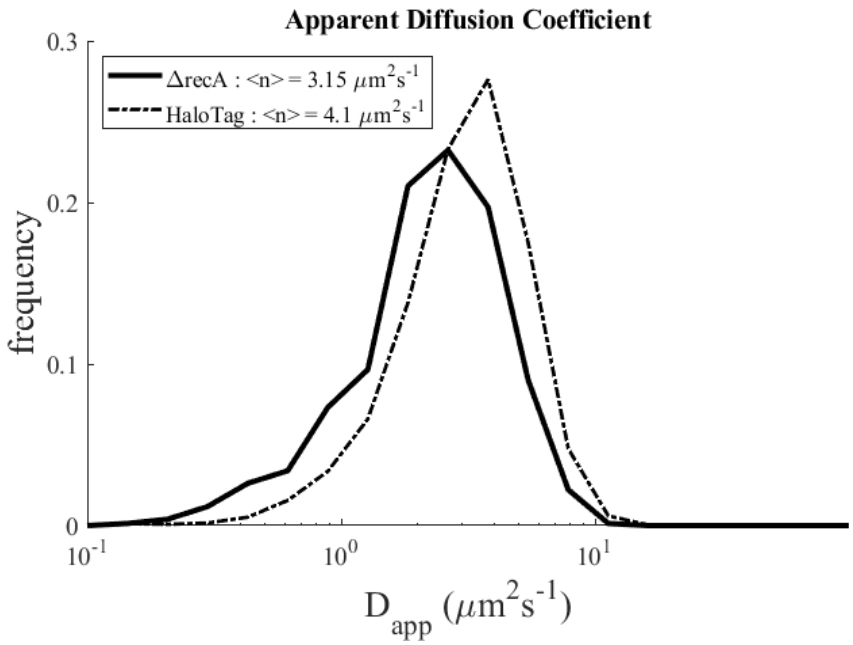

Figure 5.1. HaloTag diffuses faster than freely-diffusing Cas9-HaloTag. Apparent Diffusion Coefficient distributions of HaloTag (dotted line) and Cas9-HaloTag (in ΔrecA background, thus without genomic DNA; plain line). Despite Cas9 free diffusion, HaloTag remains faster.

HaloTag is indeed 30% faster, despite being 82% smaller. This observation corresponds to previously reported non-linear relationship between molecular mass and diffusion[106]. Notably, whereas HaloTag very well matched the model described in those studies, Cas9-HaloTag appeared faster (Figure 5.2), which we can account on (i) the protein's conformation potentially facilitating diffusion across the cytoplasm and (ii) the fact that those studies were performed in the presence of genomic DNA. Consistent with that second hypothesis, our HaloTag measurements were also performed in the presence of genomic DNA, whereas our Cas9-HaloTag ones were not (as free diffusing Cas9-HaloTag corresponds to our ΔrecA sample, with genomic DNA digested). Indeed, genomic DNA may preferentially interfere with large proteins, due to increased surface interactions.

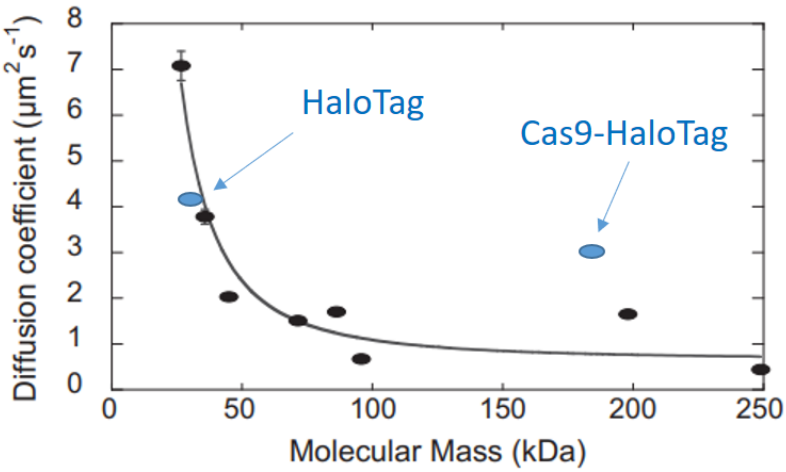

Figure 5.2. The relationship between protein weight and diffusion does not appear linear. Experimentally, the diffusion coefficients of proteins of different molecular masses were measured and plotted against each other, highlighting a non-linear relationship between the two. As reference, the blue arrows indicate the relative positions of both HaloTag (34kDa, 4.10 μm²/s) and Cas9-HaloTag (194kDa, 3.15 μm²/s) on the plot. Adapted from Kumar, M., Mommer, M. S. & Sourjik, V. Mobility of cytoplasmic, membrane, and DNA-binding proteins in Escherichia coli. *Biophys. J.* **98**, 552–559 (2010) [106].

Another important observation relates to the difference between tracks and proteins. Indeed, all our tracking data originate from tracks, not proteins. The reason for this discrepancy is that a single protein may be detected several times by our tracking algorithm and thus produce several tracks. Furthermore, fast proteins are more difficult to track and therefore lead to more tracks/protein, as the algorithm loses sight of the protein and, once found back, simply starts a new trajectory. In contrast, slow molecules are easier to follow and produce less and longer trajectories, as can be observed in Figure 4.15B (Chapter 4).

Indeed, in the presence of the *lacO6* array and the *lacOA*-gRNA (which we previously showed to significantly slow down Cas9 diffusion), Cas9-HaloTag displays longer tracks.

With proteins displaying single diffusion behaviours, this observation bears little consequence. However, with proteins displaying several diffusion behaviours, as is the case with Cas9-HaloTag, this observation would over-estimate the fraction of fast molecules (freely diffusing) and conversely under-estimate the fraction of slow molecules (despite a correct absolute number of slow tracks). One way to circumvent this phenomenon and thus re-normalize our measurements would be to limit the number of labelled Cas9-HaloTag proteins to one per cell. However, this would prove challenging and technically difficult to achieve, due to protein expression variability. Moreover, it would then require significantly more images to reach statistically comparable sample sizes. With this in mind, we can now turn our attention to our quantifications of Cas9 various diffusion behaviours.

Protein behaviours

With the exception of Cas9-HaloTag freely diffusing in the absence of genomic DNA, absolute quantifications of Cas9-HaloTag D_{app} are not straightforward, for protein diffusion behaviours are not separate from each other. Rather, they form a mixture of DNA-interacting Cas9 proteins and freely diffusing ones. Real D_{app} quantification of each Cas9 behaviour would require fitting each distribution, with additional constraints: large sample sizes; presence of less than three different behaviours; which would need to be *really* different from each other. As an illustration, Rocha and colleagues[113] observed that with 5,000 tracks, conventional fitting methods could differentiate two behaviours, provided that their respective D_{app} were more than 2-fold different; fitting of more similar diffusions would completely miss their respective fractions (Figure 5.3). With three different behaviours, the task gets even more challenging. Never mind four behaviours.

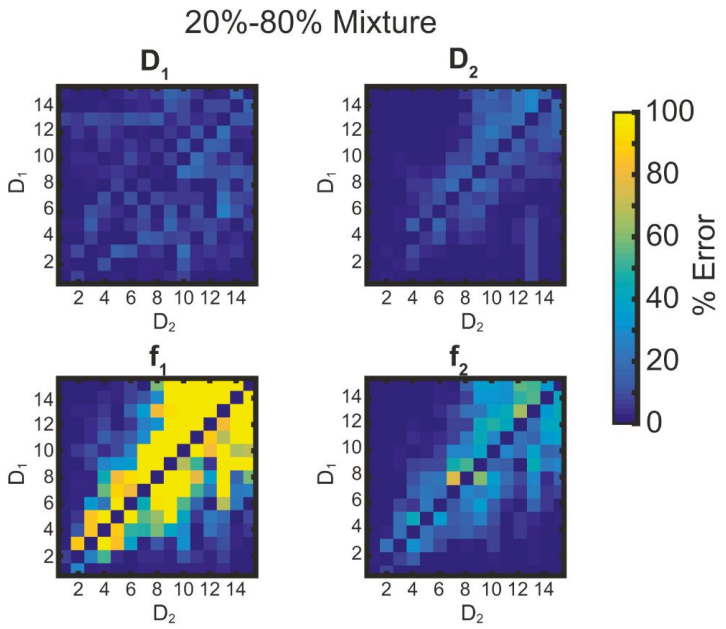

Figure 5.3. In a mixture of two diffusion behaviours, the quantification of D_{app} and their respective fractions is difficult. 3D simulations of confined Brownian motion generated 5,000 tracks with a mixture of two diffusion behaviours respectively characterizing 20% and 80% of their total amount, with D_{app} values varying between 1 µm²/s and 15 µm²/s. The comparison between the input (fixed) and the output (fitting) is plotted on a heat map, with levels of error going from blue to yellow. We see that regardless of D_{app} values, the fitting is accurate. However, the quantification of their respective fractions is very inaccurate, particularly with the smaller fraction (20%). Extracted from Rocha, J., Corbitt, J., Yan, T., Richardson, C. & Gahlmann, A. Resolving Cytosolic Diffusive States in Bacteria by Single-Molecule Tracking. *Biophys. J.* **116**, 1970–1983 (2019)[113].

Nevertheless, we suggested a simple model based on the measurement of slow tracks across all datasets, which gave us a first approximation of the quantification of the following diffusion fractions of the Cas9-HaloTag protein (Figure 4.26, Chapter 4);

- 64% freely diffusing,
- 8% with confined diffusion,
- 10% of Cas9-PAM interactions,
- 18% of ON-and-OFF-target interactions.

The most striking observation pertains to the freely-diffusing fraction of Cas9-HaloTag molecules which appears to account for more than half of the total. Accordingly, we hypothesized that Cas9 spent most of its time searching for its target, away from DNA. How does that compare with previous studies found in the literature?

Doudna and colleagues[81] made similar observations, stating that in mammalian cells, Cas9 searches predominantly through 3D diffusion. As a reminder, they used the catalytically-dead dCas9 mutant to target short interspersed nuclear elements (SINEs) of the B2 type present in the cells at about 350,000 copies. Their results showed that the protein search was dominated by 3D diffusion and that OFF-target events where short-lived, with an average residence time inferior to 1s. Quantifying the protein apparent diffusion coefficient, they observed diffusion shifts towards slower values as a function of gRNA association (Figure 5.4B).

Additionally, in the absence of gRNA, they measured a diffusion coefficient of the same order of magnitude as ours (between 1 and 10 $\mu m^2/s$). However, in the presence of a gRNA, they observed two distinct diffusion behaviours, separate from each other by as much as two orders of magnitude (10 and 0.1 $\mu m^2/s$). In contrast, not only did we not observe such a difference, but also in all conditions, our freely diffusing fraction was always predominant (Figure 5.4A).

Two reasons can explain those discrepancies. First, we used 22 targets, as opposed to their 300,000 SINEs targets. Normalizing by the size of the genome, their CRISPR complex still had 10-times more chances to find a target, thus accounting for an increased fraction of DNA-bound molecules. Second, they used the catalytically-dead version of the Cas9 protein (dCas9), which has been shown to stay bound to its target for several dozens of minutes[79], as opposed to the wild-type version which, even *in vitro*, was observed to stay bound to DNA for less than a minute[80]. Therefore, not only their dCas9-HaloTag proteins could find their targets quicker (10x), but they would also stay bound to them for much longer (10x), resulting in much larger fractions of slow particles (100x).

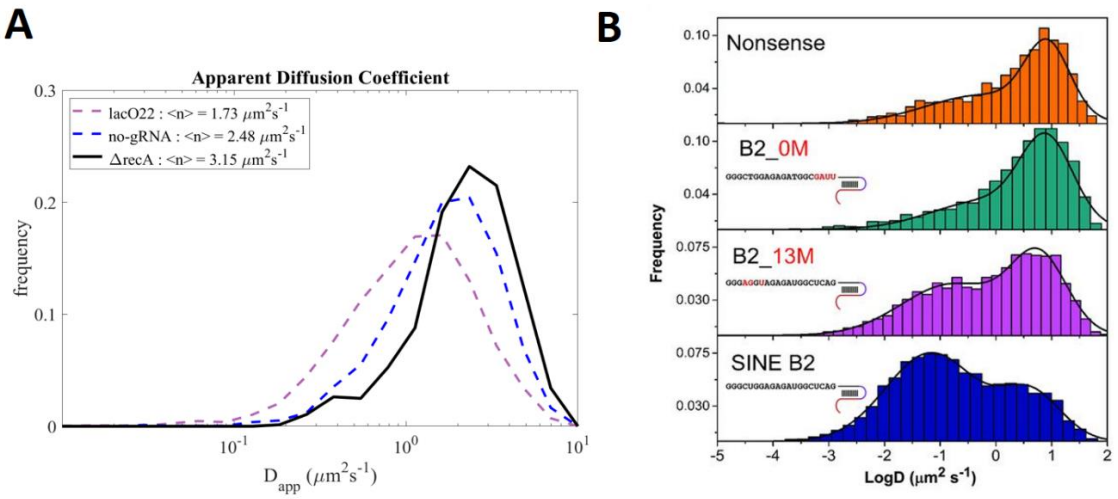

Figure 5.4. Comparison of our observation of Cas9-HaloTag diffusion with Doudna's dCas9-Halotag. (**A**) From this work are shown apparent diffusion coefficient distributions (D_{app}) of Cas9-HaloTag, without genomic content (black), with genomic content but without gRNA (blue) and with gRNA targeting 22 genomic targets (purple). Mean of distributions are indicated. (**B**) Extracted from Doudna's publication[81], distributions of dCas9-HaloTag diffusion coefficient, either alone (nonsense, orange), with a gRNA containing mismatches proximal to the PAM region (B2_0M, green), with a gRNA containing mismatches distal from PAM region (B2_13M, purple) and with a gRNA with the correct target sequence (SINE B2, blue).

In conclusion, we can hypothesize that at equal proportion of targets, we should have been able to observe a 1-log difference between the fast and slow diffusion behaviours. Regardless, we can extrapolate that with a single target, Cas9 would indeed spend the vast majority of its time freely diffusing.

This naturally brings us to a second study; that of Elf and colleagues[79] who measured a searching time (for dCas9) of about six hours, in *E.coli*. This measure is in agreement with our previous hypothesis of Cas9 spending most of its time searching for a target. However, our interpretations differ. We suggest that Cas9 does so freely diffusing, away from DNA, whereas they proposed that Cas9 is all this time interacting with DNA PAM sequences.

They based their interpretation on the fact that by taking images of dCas9-YPet with various exposure times, detecting a fluorescent spot would indicate that a protein stayed immobile for at least the length of the exposure. Without gRNA, this resulted in an averaged exposure time of 30 ms, which they interpreted as the time dCas9 takes to interrogate a PAM sequence. They went further by calculating that if Cas9 were to spend all its 6h-searching-time interacting with PAM sequences (which amount to 10^6 in the genome), it would indeed result in around 20 ms spent per PAM sequence, which was of the same order of magnitude as their previous measurement of 30 ms probing time. They hence concluded that Cas9 does spend all its searching time probing PAM sequences.

In response to this, we have two comments to make. First, despite the discrepancy of their 30 ms-probing time with what was previously found in the literature (750 ms in eukaryotic cells[81] and 1s *in vitro*[114]), they assumed that a similar order of magnitude between their measurement and their searching time estimation was sufficient to justify their interpretation. Yet, it is not because dCas9 spends as little as 30 ms to probe a PAM sequence that it necessarily spends all its searching time doing so. Second, a closer look at their measurements shows that around 30 ms, they only detected 0.2 fluorescent spots/cell, with a level of dCas9 expression amounting to five proteins/cell. Put together, this would mean that only 4% of their dCas9 proteins would actually be interacting with DNA for 30 ms (Figure 5.5). Considering that below 5 ms, they stated reaching the threshold of freely-diffusing particles, we are bound to wonder about their actual fraction of dCas9 proteins interacting with PAM sequences.

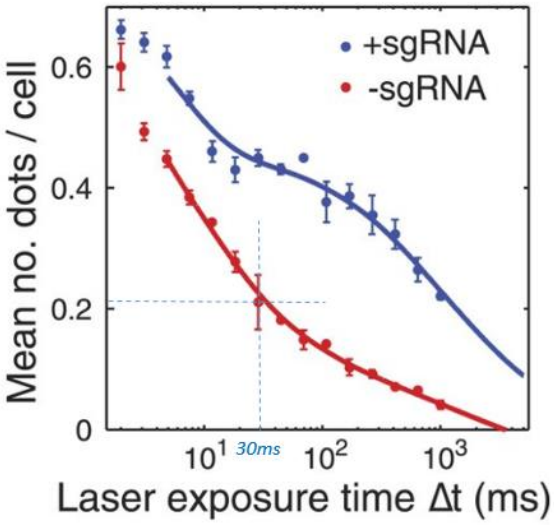

Figure 5.5. Distribution of dCas9-YPet detected spots/cell with varying exposure times. With varying exposure times (from 1 ms to 1 s), dCas9-YPet fluorescent spots were counted, either in the presence of a gRNA (blue), or in its absence (red). Highlighted are the coordinates of the point corresponding to 30 ms, without gRNA. Adapted from Jones, D. L. *et al.* Kinetics of dCas9 target search in Escherichia coli. *Science (80-.).* **9**, 1420–1424 (2017)[79].

Finally, we can compare both assays. Their technique is based on a collection of snapshots leading to the detection of immobile particles and further inference of protein behaviour. Ours relies on the direct observation of the complete range of behaviours the protein can adopt, in real time and in four different conditions. In other words, they infer whereas we visualize.

In conclusion, although we accept that dCas9 may take 30 ms to probe a PAM sequence, we do not believe there is sufficient evidence to support the theory following which Cas9 searching time is predominantly spent probing PAM sequences.

ON and OFF-targets observations

A further observation we make from our model relates to the fractions of ON and OFF-target interactions the Cas9 protein displayed. As explained in the previous chapter, we were unable to differentiate between the two types of interactions, due to technical limitations. However, when Cas9 is associated to a gRNA, yet with no target present on the genome, there can be no ON-target events. As a result, the corresponding fraction of 'ON and OFF-target' events actually limits itself to only OFF-target ones, which thus amount to 11% of the total number of tracks (Figure 5.6).

In contrast, the literature reported levels of OFF-target interactions ranging from 10^5-time less than ON-target ones (with gRNAs presenting more than 2 mismatches with the target) to just 10-time less (in the worst case, with only single mismatches)[47]. Comparing with our results indicating that with 22 targets, there are 18% of 'ON and OFF-target' events, we can extrapolate and state that even in the eventuality of those 18% being entirely the result of ON-target interactions, it would barely be twice more than the previously-measured fraction of OFF-target ones (11%). Despite the fact that we are comparing two separate datasets ('no-target' and '22 targets'), and should therefore remain prudent with any further quantification, we can still qualitatively remark that we are far from the '10-time less OFF-target than ON-target' report of the literature.

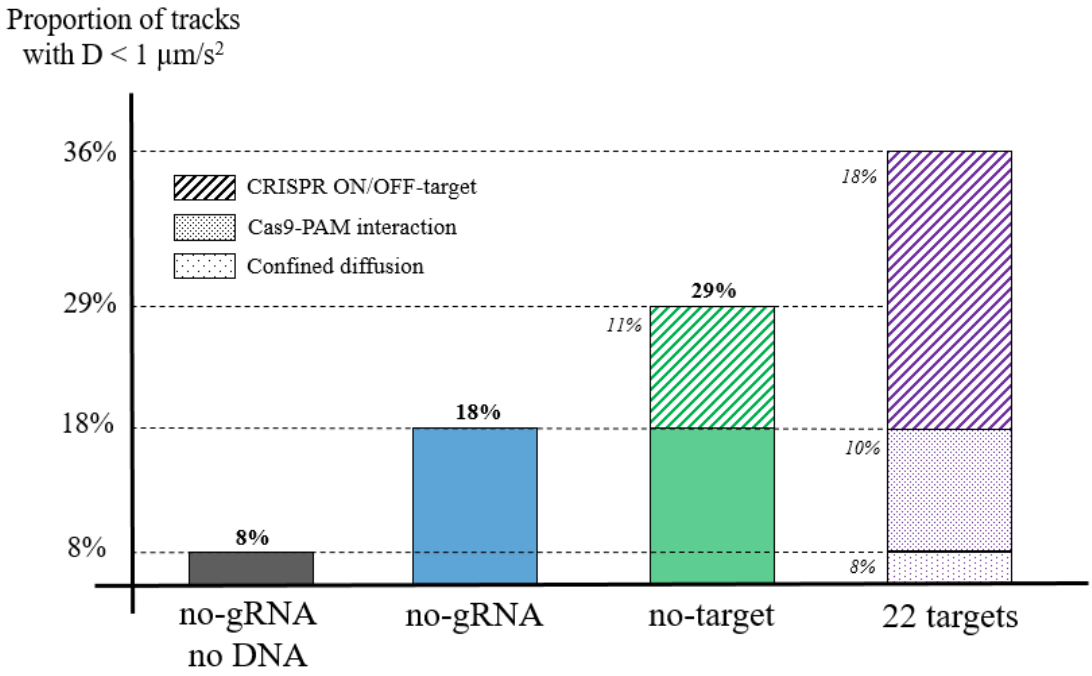

Figure 5.6. Quantification of the fractions of slow tracks in different datasets. Fractions of slow tracks (with D_{app} < 1 µm²/s) detected from cells containing Cas9-HaloTag with either no genomic DNA and no gRNA (black), genomic DNA but no gRNA (blue), a gRNA but no genomic target (green) or a gRNA and 22 genomic target (purple). Comparing the datasets together, we can quantify three main diffusion behaviours, indicated in the legend. Note that without target, there can be no ON-target interaction.

So what are the reasons of this new discrepancy, making us observe much more OFF-target interactions that reported in the literature? Let us remember that previous studies were relying on DNA modifications (indels) to detect such events, and as we hypothesized earlier in this work, those modifications may only represent a fraction of all OFF-target events. Indeed, some of those interactions may remain 'silent', with the CRISPR complex unwinding the DNA without reaching full homology, and thus preventing DNA cleavage. It is thus reasonable to assume that those 'silent' interactions may be much more frequent than those leading to DNA double-stranded breaks. Indeed, we observed a minimum of 1.6-fold less (or 40% less) OFF-target events, with 22 targets. Normalized to a single target, this number may drop even further (~2% less). Finally, we are here estimating the minimum ON/OFF difference; it may be the opposite, with OFF-target events actually much more frequent than their ON-target counterpart.

Put together, our results stress the need for a more comprehensive study of OFF-target events, for in both contexts of gene editing and antimicrobial alternatives, accurately knowing CRISPR-Cas9 entire range of DNA-interactions is paramount. Indeed, previous studies reported OFF-target interactions to last for up to 1s *in vitro*[114] which, coupled to other reports of Cas9's ability to affect gene regulation[35], could result in unwanted adverse effects; disturbing an accurate gene editing, or protecting a pathogenic bacteria to eliminate.

Fortunately, a number of techniques can limit the levels of unwanted OFF-target events. As a brief reminder, High-fidelity Cas9 mutants[46], anti-CRISPR proteins[49] and double-nickases techniques[47] are being developed. In parallel, improving our OFF-target detection methods is also extremely important, with ChIP-seq experiments[50], deep sequencing protocols[51] and *in-silico* predictions[115].

Alternative analysis methods

As shown above, the current analysis methods used in this study, namely the extraction of Mean Square Displacement (MSD) and apparent Diffusion Coefficient (D_{app}) present the main advantage of having been extensively described and used in the literature, thus facilitating comparisons with our results. However, those methods also present caveats such as the averaging of data (through the computation of MSD) and consequent loss of potentially significant information.

As an alternative, other analytical methods can be found such as the use of Radius of Gyration (described in Materials and Methods chapter) or the extraction of Single-Displacement. The latter is a particularly interesting example as it does not require any averaging and instead relies on the displacements undertaken by the protein, one frame after the next. A distribution of those displacements can then be obtained, with significantly more data points than for MSDs.

To illustrate this example, we computed the Single-Displacement distributions of our Cas9-HaloTag proteins in Figure 5.7, which offers a striking view of the diffusion shift operated by the protein in its four different conditions. In addition, the number of displacements resulting in the lacO22 distribution reaches 194,000 datapoints, compared with the 3,200 D_{app} points computed from the same sample.

Coupled to a simulation of Brownian motion, this amount of data could both be supportive (allowing the testing of additional hypotheses) and informative (through further data extraction, distribution fitting and inference).

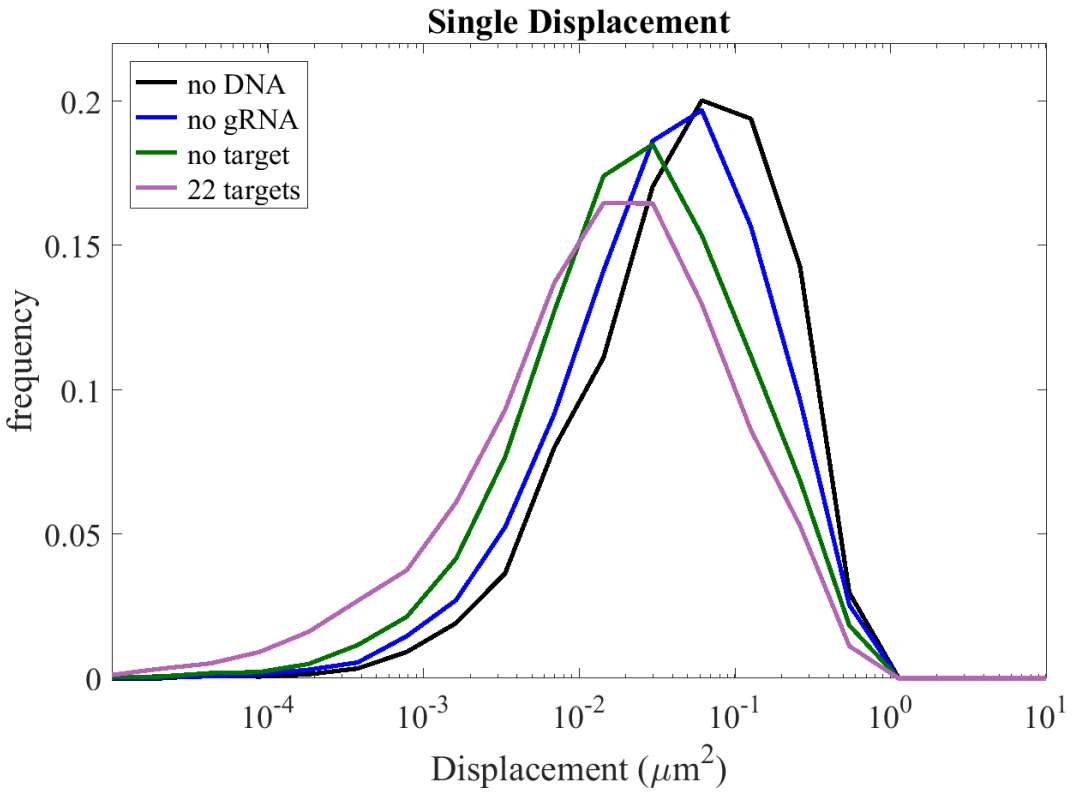

Figure 5.7. Distribution of Single-Displacement extracted from Cas9-HaloTag proteins. Movies describing cells containing the Cas9-HaloTag protein and either of the following constructs were observed: the *lacO22*-sgRNA and the *lacO22* array on the genome (22 targets, purple); the *lacO22*-sgRNA without target on the genome (no target, green); the protein without any sgRNA (no gRNA, blue); and finally, no sgRNA and no genomic DNA at all (no DNA, black).

Future work

In our lab, we now aim to build on our new expertise to expand our set of observations. Indeed, a good place to start would be to build a third array of *lacO* repeats, in order to validate our theory following which three different number of targets would reveal their respective fractions of ON and OFF-target events, on the assumption that their ON-target fraction linearly increases with the number of targets. This time, we would then be able to differentiate between the two fractions and quantify their respective presence in the population of tracks. We would finally know how much of each Cas9 displays, and which one is more frequent.

Additionally, further analyses would prove particularly useful to extract more information from the data we already obtained. As such, we could implement the fitting methods already illustrated by Uphoff and colleagues[71,112] applied to DNA repair proteins, which allowed the quantification of various behaviours. Along the same line, they also showed how such methods could be used to sort protein trajectories according to their diffusion behaviour, and subsequently identify their most frequent locations in an *E.coli* cell. Another direction would be to compare our Diffusion Coefficient analysis method with different ones, relying more on particle trajectories and less on statistical averaging, as illustrated with the Single-Displacement method.

Also, we greatly wish to apply our experiments to the catalytically-dead Cas9 mutant (dCas9), which was extensively studied in the literature. Given dCas9's ability to bind DNA for longer than its wild-type counterpart, we should be able to increase our fraction of slow tracks, thus increasing our detection sensitivity. Accordingly, the potential discrepancy between the two would be a direct measure of both proteins residence times (on the DNA). Additionally, such an endeavour would not only confront our expertise with that of other groups, but would also shed more light on the differences between mammalian and bacterial protein diffusion, in the context of chromatin structures. Finally, comparing both our Cas9 and dCas9 results would address new questions relating to the phenotypic activity of the protein, more specifically in the context of DNA repair and subsequent SOS-response activation.

Conclusion

In this work, we showed how fantastic a tool CRISPR-Cas9 was, and how it revolutionized both the way we do research and the way we apply it. Indeed, countless laboratories across the world now routinely use its genetic capabilities, which have pervaded our society in an attempt to improve every biological process; crop yields, biofuels, therapeutics, anti-mosquitoes, antimicrobials. I have to say, the future is exciting.

However, like every formidable tool, it has its limitations. Beyond potential ethical questionings, there are current technical shortcomings. We presented how CRISPR-Cas9 ability to cleave DNA is as powerful as it is accurate, and how a lack of control could be detrimental – if not dramatic. We described the range of behaviours CRISPR-Cas9 can adopt, and how despite its specific ON-target interactions with DNA, there were also non-specific OFF-target events. Those non-specific events are difficult to predict, difficult to detect and thus easy to underestimate.

In this study, we aimed to go beyond the current detection methods of OFF-target events which have focused on DNA modification. Instead, we decided to directly look at the protein and infer from its motion the type of behaviour it would adopt, specific or not. To do so, we developed a single-molecule microscopy assay designed to track Cas9 displacement in time and in various conditions; with-and-without genomic DNA, with-and-without gRNA and with-and-without target recognised by the gRNA. Put together, we successfully showed that we could track Cas9 but also build a first approximation of its fractions of respective behaviours. This resulted in the finding that the protein was predominantly diffusing in the cytoplasm, searching for its target. Additionally, we speculated that the fraction of OFF-target interaction was comparable to that of ON-target ones, if not larger. Finally, we showed how both fractions were highly dependent on the type of gRNA, characterized by its target sequence and thus genomic homologies.

Strikingly, those observations were in contradiction with previous studies, which claimed that (i) Cas9 spent most of its time interacting with DNA (PAM sequences) and (ii) with far less OFF-target events. We confronted those interpretations with (i) our direct observations of freely diffusing Cas9, and suggested that (ii) most of the OFF-target interactions detected in previous studies were only a fraction of a larger ensemble, invisible to past detection techniques and largely underestimated.

In conclusion, we end this work by encouraging further studies on the quantification, as well as on the identification of the potential roles of those 'silent' OFF-target events, which could well be interfering with DNA maintenance mechanisms (replication, regulation, repair).

Finally, thank you for the time you have spared reading this material, and I hope you enjoyed reading it as much as I enjoyed collecting, discussing and writing it.

Appendix

Table 1. Plasmids.

Plasmid	Genotype	Source
pXZ02	pOSIP-CT, *araC pBAD-Cas9*	
pXZ03	*pBAD33, araC pBAD-Cas9*	Gift from SJR
pXZ05	pUA66, sgRNA-*prtA*	Gift from SJR
pXZ06	pOSIP-KH, *PsfiA::gfp*	Gift from SJR
pXZ08	pKD46, pE-FLP	Addgene[83]
pXZ10	pOSIP-KT	Addgene[83]
pXZ11	pOSIP-CT	Addgene[83]
pXZ13	pOSIP-KL, leaky pTet-Cas9	Gift from DB
pXZ15	pOSIP-KL, tight pTet-Cas9	Gift from DB
pXZ17	pOSIP-KL, leaky pTet-Cas9-HaloTag	
pXZ19	pOSIP-KL, tight pTet-Cas9-HaloTag	
pXZ20	*pOSIP-CT, araC pBAD-Cas9-HaloTag*	
pXZ23	pUA66: gRNA(empty)	Gift from SJR
pXZ24	pTOF(Δ*recA*)	Gift from VA
pXZ32	crRNA (*lacZ1*)	Gift from DB
pXZ33	crRNA (*lacZ2*)	Gift from DB
pXZ36	pUC, *lacO6*	Built with EGF
pXZ37	pUC, *lacO14*	Built with EGF
pXZ38	pUC, *lacO22*	Built with EGF
pXZ39	pUC, *lacO30*	Built with EGF
pXZ40	pOSIP-KO, *lacO6*	
pXZ42	pOSIP-KO, *lacO22*	
pXZ44	pUA66, sgRNA-*lacOA* (array)	
pXZ45	pUA66, sgRNA-*lacO* (genome)	
pXZ46	pBAD-HaloTag	Gift from HT

SJR: Sebastian Jaramillo-Riveri
DB: David Bikard
VA: Vincent Amarh
EGF: Edinburgh Genome Foundry
HT: Hannah Taylor

Plasmid Construction.

For each plasmid, the backbone was either extracted with restriction enzymes (digestion), or amplified by PCR. They were then associated via Gibson Assembly with one or several fragments, themselves either coming from gBlocks or from PCR amplicons. In the table below are indicated the plasmids, their contents, their backbones, the templates from which the PCR were realised, with their respective amplification primers. Note that for pXZ40 and pXZ42, simple restriction-ligation was performed. For pXZ44 and pXZ45, whole-plasmid amplification was used instead, with the sgRNA target sequence introduced on the amplification primers.

Table 2. Description of templates for plasmid construction.

Plasmid	Content	Backbone (digest)	Fragments	Amplification Primers	Checking Primers
pXZ02	*pOSIP-CT, araC pBAD-Cas9*	pXZ11	pXZ03 (PCR)	**pOSIP-CT**: oXZ14, oXZ24 **pXZ03**: oXZ22, oXZ23	**oXZ25, oXZ26** **oXZ27, oXZ28**
pXZ17	*pOSIP-KL, pTet-Cas9-HaloTag (leaky)*	pXZ13 (SacI, NarI)	gBlock gXZ01	-	oXZ15, oXZ16
pXZ19	*pOSIP-KL, pTet-Cas9-HaloTag (tight)*	pXZ15 (SacI, NarI)	gBlock gXZ01	-	oXZ15, oXZ16
pXZ20	*pOSIP-CT, araC pBAD-Cas9-HaloTag*	pXZ02 (PmlI, NarI)	gBlock gXZ01, pXZ02	**pXZ02**: oXZ29, oXZ30	**oXZ27, oXZ28**
pXZ40	*pOSIP-KO, lacO6*	pOSIP-KO (BamHI, NarI)	pXZ36 (BamHI, NarI)	**-**	**oXZ136, oXZ137**
pXZ42	*pOSIP-KO, lacO22*	pOSIP-KO (BamHI, NarI)	pXZ38 (BamHI, NarI)	**-**	**oXZ136, oXZ137**
pXZ44	pUA66, sgRNA-*lacOA* (array)	pXZ05	-	**oXZ150, oXZ151**	**oXZ149, oXZ152**
pXZ45	pUA66, sgRNA-*lacO* (genome)	pXZ05	-	**oXZ154, oXZ155**	**oXZ149, oXZ152**

Table 3. *E.coli* strains.

Strain	Genotype	Plasmid	Source/Comment
MG1655	F- lambda- *ilvG- rfb*-50 *rph*-1		*E. coli* K-12
BW27783	*Δ(araD-araB)567, ΔlacZ4787(::rrnB-3), λ⁻, Δ(araH-araF)570(::FRT), ΔaraEp-532::FRT, φP$_{cp8}$araE535, rph-1, Δ(rhaD-rhaB)568, hsdR514*		[105]
XZ12	BW27783	pXZ03	Gift from SJR
XZ13	BW27783	pXZ05	Gift from SJR
XZ14	BW27783	pXZ03, pXZ05	
XZ17	DH5a pir	pXZ06	Gift from SJR
XZ19	E811	pXZ08	[83]
XZ20	BW27783, HK022::PsfiA-mGFP		
XZ30	DH5a pir	pXZ02	
XZ31	DH5a pir		Gift from MEK
XZ33	E811	pXZ11	
XZ35	DH5a pir	pXZ13	Gift from DB
XZ37	DH5a pir	pXZ15	Gift from DB
XZ50	MG1655, HK022::PsfiA-mGFP, Lambda::(leaky)pTet-Cas9		
XZ54	MG1655, HK022::PsfiA-mGFP, Lambda::(tight)pTet-Cas9		
XZ58	DH5a pir	pXZ17	
XZ62	DH5a pir	pXZ19	
XZ74	MG1655, Lambda::(leaky)pTet-Cas9		
XZ78	MG1655, Lambda::(tight)pTet-Cas9		
XZ82	MG1655, HK022::PsfiA-mGFP, Lambda::(leaky)pTet-Cas9-HaloTag		
XZ86	MG1655, HK022::PsfiA-mGFP, Lambda::(tight)pTet-Cas9-HaloTag		
XZ90	MG1655, Lambda::(leaky)pTet-Cas9-HaloTag		
XZ94	MG1655, Lambda::(tight)pTet-Cas9-HaloTag		
XZ99	MG1655, HK022::PsfiA-mGFP, Lambda::(tight)pTet-Cas9	pXZ05	gRNA(ptrA)
XZ106	MG1655, Lambda::(tight)pTet-Cas9	pXZ05	gRNA(ptrA)
XZ110	MG1655, HK022::PsfiA-mGFP, Lambda::(tight)pTet-Cas9-HaloTag	pXZ05	gRNA(ptrA)
XZ114	MG1655, Lambda::(tight)pTet-Cas9-HaloTag	pXZ05	gRNA(ptrA)
XZ117	BW27783, P21::pBAD-Cas9		
XZ119	MG1655, P21::pBAD-Cas9, HK022::PsfiA-mGFP		
XZ121	BW27783, P21::pBAD-Cas9	pXZ05	gRNA(ptrA)
XZ122	MG1655, P21::pBAD-Cas9, HK022::PsfiA-mGFP	pXZ05	gRNA(ptrA)
XZ124	DH5a pir	pXZ20	
XZ125	DH5a pir	pXZ20, pXZ05	gRNA(ptrA)
XZ128	BW27783, P21::pBAD-Cas9-HaloTag		
XZ130	MG1655, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP		
XZ136	BW27783, P21::pBAD-Cas9-HaloTag	pXZ05	gRNA(ptrA)
XZ137	MG1655, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP	pXZ05	gRNA(ptrA)

Strain	Genotype	Plasmid	Source/Comment
XZ145	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP		
XZ148	BW27783, P21::pBAD-Cas9, HK022::PsfiA-mGFP	pXZ05	gRNA(ptrA)
XZ149	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP	pXZ05	gRNA(ptrA)
XZ151	MG1655	pXZ32	crRNA(lacZ1)
XZ152	MG1655	pXZ33	crRNA(lacZ2)
XZ157	BW27783	pXZ23	gRNA(empty)
XZ158	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP	pXZ23	gRNA(empty)
XZ159	XL1-blue	pXZ24	Gift from VA
XZ161	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP, ΔrecA		
XZ162	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP, ΔrecA	pXZ23	gRNA(empty)
XZ163	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP, ΔrecA	pXZ05	gRNA(ptrA)
XZ165	BW27783	pXZ46	pBAD-HaloTag
XZ179	BW27783, P21::pBAD-Cas9, HK022::PsfiA-mGFP	pXZ32	crRNA(lacZ1)
XZ180	BW27783, P21::pBAD-Cas9, HK022::PsfiA-mGFP ; *unstable*	pXZ33	crRNA(lacZ2)
XZ187	BW27783, P21::pBAD-Cas9, HK022::PsfiA-mGFP	pXZ23	gRNA(empty)
XZ189	DH5a pir	pXZ36	
XZ190	DH5a pir	pXZ37	
XZ191	DH5a pir	pXZ38	
XZ192	DH5a pir	pXZ39	
XZ195	DH5a pir	pXZ40	
XZ197	DH5a pir	pXZ42	
XZ199	DH5a pir	pXZ45	
XZ200	DH5a pir	pXZ44	
XZ210	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP, 186::lacO6		
XZ211	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP, 186::lacO22		
XZ219	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP, 186::lacO6	pXZ23	gRNA(empty)
XZ220	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP, 186::lacO6	pXZ05	gRNA(ptrA)
XZ221	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP, 186::lacO6	pXZ45	gRNA(lacO)
XZ222	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP, 186::lacO6	pXZ44	gRNA(lacOA)
XZ226	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP, 186::lacO22	pXZ23	gRNA(empty)
XZ227	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP, 186::lacO22	pXZ05	gRNA(ptrA)
XZ228	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP, 186::lacO22	pXZ44	gRNA(lacOA)
XZ235	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP	pXZ44	gRNA(lacOA)
XZ236	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP	pXZ38, pXZ44	gRNA(lacOA)
XZ237	BW27783, P21::pBAD-Cas9-HaloTag, HK022::PsfiA-mGFP, 186::lacO22	pXZ38, pXZ44	gRNA(lacOA)

SJR: Sebastian Jaramillo-Riveri
MEK: Meriem El Karoui
DB: David Bikard
VA: Vincent Amarh
HT: Hannah Taylor

Table 4. Description of strains used in experiments.

Experiment	Strains (XZ#)	Thesis paragraph
Qualitative Cas9 activity	99, 110, 148, 149	3.2.1
Quantitative Cas9 activity	148, 149, 158, 187	3.2.2
Cas9-HaloTag signal specificity	158, 187	3.3.3
Protein counting	149, 158, 187	3.2.4
HiLo signal specificity	149, 158, 187	3.3.1
Tracking: HaloTag	165 and BW27783	3.3.2
Tracking: Cas9-HaloTag	149, 158, 187	3.3.3
Qualitative CRISPR-*lacO* activity	149, 158, 210, 219, 220, 221, 222, 226, 227, 228, 235	4.2.1, 4.2.2
Quantitative CRISPR-*lacO* activity	226, 228	4.2.3
Tracking: CRISPR-*lacO6*	187, 219, 222	4.3.1
Tracking: CRISPR-*lacO22*	187, 226, 228	4.3.2
Tracking: no target	187, 226, 235	4.3.3
Tracking: plasmid targets	187, 226, 235, 236, 237	4.3.4
Tracking: no genomic DNA	158, 162, 187	4.3.4

Table 5. Primers.

For primers containing overhangs, the annealing part is shown in bold. oXZ150 and oXZ154 contain SacI restriction sites, to facilitate downstream screening of plasmid construction. For oXZ151-155, red base-pairs (bps) indicate the gRNA target sequences, blue bps the SacI added restriction site and the underlined bps the sequence overlapping their paired primer.

Primer	Sequence (5' → 3')	Comment
oXZ14	*GTCAGCTAGGAGGTGACTAA**CTGCAGGC ATGCCTCGAGAT***	Building pXZ02, anneals pOSIP-CT
oXZ15	TATTCAGTCCTAGTGGTTGC	Checking pXZ17, pXZ19, forw.
oXZ16	TGAGGTTGAAGGTAATTCCA	Checking pXZ17, pXZ19, rev.
oXZ22	GGATCCTCTAGAGGGCTTACTAAAAGCC AGATAACAGTAT**TTATGACAACTTGACG GCTA**	Building pXZ02, anneals pXZ03
oXZ23	GGCGCCATGCATCTCGAGGCATGCCTGC AG**TTAGTCACCTCCTAGCTGACTC**	Building pXZ02, anneals pXZ03
oXZ24	TAGCCGTCAAGTTGTCATAA**ATACTGTT ATCTGGCTTTTA**	Building pXZ02, anneals pOSIP-CT
oXZ25	GCAGACAAGCCCGTCAGGGCGCGT	Checking pXZ02, forw.1
oXZ26	TGTTACTGCGGCGCATGGAAGCGA	Checking pXZ02, rev.1
oXZ27	AGTCCAGAAGATAACGAACAAAAA	Checking pXZ02, pXZ20, forw.2
oXZ28	TTGTATTGCCAATATACCTTTCCGG	Checking pXZ02, pXZ20, rev.2
oXZ29	GGGCGGCATAAGCCAGAAAATATCG	Building pXZ20, anneals pXZ02
oXZ30	AAGCAGCGGGAGCTCCAAGATTCGT	Building pXZ20, anneals pXZ02

oXZ136	GAATTCGAGCTCGGTACCCG	Checking and Seq. pXZ40, pXZ42, forw.
oXZ137	AATCGCCAGAGAAATCTACG	Checking and Seq. pXZ40, pXZ42, forw.
oXZ138	CTCATACAATCGGAAACACAAT	Sequencing pXZ42
oXZ139	CCTGTAGTACAGCGCTTATG	Sequencing pXZ42
oXZ141	CGCGCCGCTAGATTCAACC	Sequencing pXZ40, pXZ42
oXZ144	GAGCGGATAACAATTCTTGAT	Sequencing pXZ42
oXZ145	TCTCCCGCTTGTCATGCC	Sequencing pXZ42
oXZ148	ACTTCGGACGTCTAAGAAAC	Sequencing pXZ40, pXZ42
oXZ149	CTGCAGTTCATTCAGGGCAC	Checking and Seq. pXZ44, pXZ45, forw.
oXZ150	GAACTAGTATTATACCTAGGACTGAGCT AGCTGTCAA*GAG**CTC***GAGGTGAAGACGA AAGG	Building pXZ44, anneals pXZ05
oXZ151	TAATACTAGTTCCGCTCACAATTCCACA TGGTTTAA**GAGCTATGCTGGAAACAGCA**	Building pXZ44, anneals pXZ05
oXZ152	AAAGGCCCAGTCTTTCGACT	Checking and Seq. pXZ44, pXZ45, rev.
oXZ154	ATACTAGTATTATACCTAGGACTGAGCT AGCTGTCAA*GAG**CTC***GAGGTGAAGACGA AAGG	Building pXZ45, anneals pXZ05
oXZ155	TAATACTAGTATGTTGTGTGGAATTGTG AGGTTTAA**GAGCTATGCTGGAAACAGCA**	Building pXZ45, anneals pXZ05

Table 6. Additional sequences.

Content	Sequence (5' → 3')	Comment
gRNA(ptrA)	GGTGCGCATCATAAAGTAAG	Target sequence
gRNA(lacO)	ATGTTGTGTGGAATTGTGAG	Target sequence
gRNA(lacOA)	TCCGCTCACAATTCCACATG	Target sequence
crRNA(lacZ1)	TCACTGGCCGTCGTTTTACAACGTCGTGAC	Target sequence
crRNA(lacZ2)	CCATTACGGTCAATCCGCCGTTTGTTCCCA	Target sequence
lacO6 array	*GATCCTCTAGAGGGCTTACTAAAAGCCAGATAACAG TATGCGTATTTG*GCGCGCCGCTAGATTCAACCAAAG ACAATTGTTA**TCCGCTCACAATTCCACATG**TGGTTG GCTTGTGTCAAGGTGTGAATTGTTA**TCCGCTCACAA TTCCACATG**TGGACACTCTGCAGCCTCTCAGTAATT GTTA**TCCGCTCACAATTCCACATG**TGGCTTCGCATC GAAGTGCTAGAAATTGTTA**TCCGCTCACAATTCCAC ATG**TGGGATACAGGTCCCGTGGTAAGAATTGTTA**TC CGCTCACAATTCCACATG**TGGCCGTAGTAGTATTTA CGTCGAATTGTTA**TCCGCTCACAATTCCACATG**TGG TAGTCACCGTTCACCTCTACACTAGT*CTAGTAGCGG CCGCTGCAGGCATGCCTCGAGATGCATGG*	*6 lacO* targets in bold, cloning sequences in italics. Example of one repeat highlighted in yellow, with *lacO* operator region underlined, and PAM region in red.
lacO22 array	*GATCCTCTAGAGGGCTTACTAAAAGCCAGATAACAG TATGCGTATTTG*GCGCGCCGCTAGATTCAACCAAAG ACAATTGTTA**TCCGCTCACAATTCCACATG**TGGTTG GCTTGTGTCAAGGTGTGAATTGTTA**TCCGCTCACAA TTCCACATG**TGGACACTCTGCAGCCTCTCAGTAATT	*22 lacO* targets in bold, cloning sequences in italics.

	GTTA**TCCGCTCACAATTCCACATG**TGGCTTCGCATC GAAGTGCTAGAAATTGTTA**TCCGCTCACAATTCCAC** **ATG**TGGGATACAGGTCCCGTGGTAAGAATTGTTA**TC** **CGCTCACAATTCCACATG**TGGCCGTAGTAGTATTTA CGTCGAATTGTTA**TCCGCTCACAATTCCACATG**TGG TAGTCACCGTTCACCTCTACAATTGTTA**TCCGCTCA** **CAATTCCACATG**TGGGCTCATACAATCGGAAACACA ATTGTTA**TCCGCTCACAATTCCACATG**TGGCTAGTT CCATGGACCAGACAAATTGTTA**TCCGCTCACAATTC** **CACATG**TGGAGCTTCTGGCCTCGATCAAGAATTGTT A**TCCGCTCACAATTCCACATG**TGGATCTGACGAACT CTATTCGCAATTGTTA**TCCGCTCACAATTCCACATG** TGGCAACAGGGCAAGGGGAAGCTAATTGTTA**TCCGC** **TCACAATTCCACATG**TGGCCACTTGAATCATGCATC CAAATTGTTA**TCCGCTCACAATTCCACATG**TGGAGG TGAGAGGAGTTGGAGAGAATTGTTA**TCCGCTCACAA** **TTCCACATG**TGGCCTGTAGTACAGCGCTTATGAATT GTTA**TCCGCTCACAATTCCACATG**TGGGCATGACAA GCGGGAGACTAAATTGTTA**TCCGCTCACAATTCCAC** **ATG**TGGCCTTCCCATACATTCTCGAGAATTGTTA**TC** **CGCTCACAATTCCACATG**TGGCGATTCGTCATTGGG ATAACAATTGTTA**TCCGCTCACAATTCCACATG**TGG AGTTGGCATGCTAACCGAGAAATTGTTA**TCCGCTCA** **CAATTCCACATG**TGGATGTACCCTCGCAAGTTGAGA ATTGTTA**TCCGCTCACAATTCCACATG**TGGAGTAGA CCGTGCAGGGTAAGAATTGTTA**TCCGCTCACAATTC** **CACATG**TGGTCCTCCAACAGGCCTCTTTCAATTGTT A**TCCGCTCACAATTCCACATG**TGGGGAGTAGAGGAC GACTACACACTAGT*CTAGTAGCGGCCGCTGCAGGCA* *TGCCTCGAGATGCATGG*	Example of one repeat highlighted in yellow, with *lacO* operator region underlined, and PAM region in red.
gBlock gXZ1	AAATATTATTCATTTATTTACGTTGACGAATCTTGG AGCTCCCGCTGCTTTTAAATATTTTGATACAACAAT TGATCGTAAACGATATACGTCTACAAAAGAAGTTTT AGATGCCACTCTTATCCATCAATCCATCACTGGTCT TTATGAAACACGCATTGATTTGAGTCAGCTAGGAGG TGACAGCGGTGGCGGTGGCAGTAACGATGGATCCGA AATCGGTACTGGCTTTCCATTCGACCCCCATTATGT GGAAGTCCTGGGCGAGCGCATGCACTACGTCGATGT TGGTCCGCGCGATGGCACCCCTGTGCTGTTCCTGCA CGGTAACCCGACCTCCTCCTACGTGTGGCGCAACAT CATCCCGCATGTTGCACCGACCCATCGCTGCATTGC TCCAGACCTGATCGGTATGGGCAAATCCGACAAACC AGACCTGGGTTATTTCTTCGACGACCACGTCCGCTT CATGGATGCCTTCATCGAAGCCCTGGGTCTGGAAGA GGTCGTCCTGGTCATTCACGACTGGGGCTCCGCTCT GGGTTTCCACTGGGCCAAGCGCAATCCAGAGCGCGT CAAAGGTATTGCATTTATGGAGTTCATCCGCCCTAT CCCGACCTGGGACGAATGGCCAGAATTTGCCCGCGA GACCTTCCAGGCCTTCCGCACCACCGACGTCGGCCG CAAGCTGATCATCGATCAGAACGTTTTTATCGAGGG TACGCTGCCGATGGGTGTCGTCCGCCCGCTGACTGA AGTCGAGATGGACCATTACCGCGAGCCGTTCCTGAA TCCTGTTGACCGCGAGCCACTGTGGCGCTTCCCAAA CGAGCTGCCAATCGCCGGTGAGCCAGCGAACATCGT CGCGCTGGTCGAAGAATACATGGACTGGCTGCACCA GTCCCCTGTCCCGAAGCTGCTGTTCTGGGGCACCCC AGGCGTTCTGATCCCACCGGCCGAAGCCGCTCGCCT GGCCAAAAGCCTGCCTAACTGCAAGGCTGTGGACAT	Building pXZ17, pXZ19 and pXZ20

	CGGCCCGGGTCTGAATCTGCTGCAAGAAGACAACCC GGACCTGATCGGCAGCGAGATCGCGCGCTGGCTGTC TACTCTGGAGATTTCCGGTTAACTCGAGTAAGGATC TCCAGGCATTGCAGGCATGCCTCGAGATGCATGGCG CCTAACCTAAACTGACAGGCATCAAATTAAGCAGAA GG	

Table 7. Media composition.

Generic Media	**Imaging Media**	**HaloTag fixation**	
M9 1X	M9 1X	Formaldehyde 2.5%	
Glucose 0.2% w/v	Glucose 0.2% w/v	PBS 1X	
$MgSO_4$ 2mM	MgSO4 2mM		
$CaCl_2$ 0.1mM	$CaCl_2$ 0.1mM		
MEM Essential 1X	LB 10%		
MEM Non-Essential 1X			

Table 8. Antibiotic concentrations.

Antibiotics	**Concentration**
Kanamycin (Kn)	50 μg/ml
Chloramphenicol (Cl)	30 μg/ml
Ampicillin (Am)	100 μg/ml
Anhydrotetracycline	1μM

Table 9. Single-Molecule Tracking script description

0. Cell cropped manually from single-molecule movie
1. Segmentation of the cell
    a. Extracts information from each frame of the movie, using `read_stacks` and `SumIm` developed by A. Lepore;
    b. Computes the intensity threshold used to obtain the binary mask;
    c. Extracts information from the identified cell.
2. Tracking of the fluorescent molecules within the cell
    a. Runs the frames of the movie through `findpeak_timeTOtrack`, developed by A. Lepore, to identify the fluorescent peaks
        i. Applies a band-pass filter that suppresses pixel noise (`bpass`, from Crocker et al.)
        ii. Identifies peaks from their intensity (`pkfnd`, from E. R. Dufresne)
        iii. Identifies peaks' centre position by fitting the peaks to a 2D Gaussian (`centfind`, Eric R. Dufresne)
    b. Tracks the identified peaks frame by frame (`track`, from Crocker et al.)

c. Restricts them to those within the cell (`SelTracks`, from A. Lepore)

3. Tracking data visualisation
   a. Shows tracks overlapping the first frame of the cropped movie
   b. Shows distribution of molecules' occupancy, on a pixel heat map

4. Mean Square Displacement (MSD) and Diffusion Coefficient (D) computation
   a. MSD and Single Displacement respectively computed with `MSDcalculations` and `SDcalculations` (A. Lepore)
   b. Visualise MSD as a function of lag-time
   c. D computed and distribution visualised (log scale)

Table 10. Parameters used in the tracking script.

Parameter	Function	Value	Note
`param.Xtime`	-	`150`	`objective magnification`
`param.pixel2micron`	-	`16/param.Xtime`	`converts px in µm`
`param.exp_time`		`0.010`	`exposure time (in seconds`
`param.nstacks_track`	`findpeak_timeTOtrack`	`500`	`number of frames used to track`
`param.nstacks_seg`	`read_stacks`	`200`	`number of frames used to segment`
`j0=1`		`1`	`first frame to be read`
`param.thrfpeak`	`pkfnd`	`12.5`	`threshold multiplicative factor`
`param.pnoise`	`bpass`	`1`	`noise`
`param.psize`	`pkfnd, bpass`	`4`	`size group of pixel`
`param.pgauss`	`centfind`	`7`	`size for gaussian fit`
`param.maxd`	`track`	`8`	`maximum dispersion`
`para.mem`	`track`	`0`	`memory`
`para.good`	`track`	`2`	`eliminates very short trajectories`
`para.quiet`	`track`	`0`	`to obtain text output`

The whole algorithm and additional functions are available on GitHub at the following address: https://github.com/xzaoui/PhD_scripts.

Bibliography

1. Waterhouse, P. M., Wang, M. B. & Lough, T. Gene silencing as an adaptive defence against viruses. *Nature* **411**, 834–842 (2001).

2. Sorek, R., Kunin, V. & Hugenholtz, P. CRISPR - A widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.* **6**, 181–186 (2008).

3. Marraffini, L. A. & Sontheimer, E. J. CRISPR Interference Limits Horizontal Targeting DNA. *Science (80-.).* **322**, 1843–1845 (2008).

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

5. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–63 (2005).

6. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607 (2011).

7. Hale, C. R. *et al.* RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. *Cell* **139**, 945–956 (2009).

8. Waters, L. S. & Storz, G. Regulatory RNAs in Bacteria. *Cell* **136**, 615–628 (2009).

9. Makarova, K. S. *et al.* Evolution and classification of the CRISPR-Cas systems. *Nature Reviews Microbiology* **9**, 467–477 (2011).

10. Daniel, H. H., Jeremy, S., Emmanuel, F. M. & Karen, E. N. A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes. *PLoS Comput. Biol.* **1**, (2005).

11. Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. A putative RNA-interference-based immune system in prokaryotes: Computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology Direct* **1**, (2006).

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

13. Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463**, 568–571 (2010).

14. Makarova, K. S., Aravind, L., Wolf, Y. I. & Koonin, E. V. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol. Direct* **6**, (2011).

15. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science (80-.).* **337**, 816–822 (2012).

16. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nat Methods* **10**, 957–963 (2013).

17. Jiang, F. & Doudna, J. A. CRISPR–Cas9 Structures and Mechanisms. *Annu. Rev. Biophys.* **46**, 505–529 (2017).

18. Wiedenheft, B. *et al.* RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc. Natl. Acad. Sci.* **108**, 10092–10097 (2011).

19. Jiang, F., Zhou, K., Ma, L., Gressel, S. & Doudna, J. A. A Cas9-guide RNA complex

preorganized for target DNA recognition. *Science (80-.).* **348**, 1477–1481 (2015).

20. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).

21. Liang, F., Han, M., Romanienko, P. J. & Jasin, M. Homology-directed repair is a major double-strand break repair pathway in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 5172–5177 (1998).

22. Li, X. & Heyer, W. D. Homologous recombination in DNA repair and DNA damage tolerance. *Cell Research* **18**, 99–113 (2008).

23. Lieber, M. R. The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. *Annu. Rev. Biochem.* **79**, 181–211 (2010).

24. Lamarche, B. J., Orazio, N. I. & Weitzman, M. D. The MRN complex in double-strand break repair and telomere maintenance. *FEBS Letters* **584**, 3682–3695 (2010).

25. Dillingham, M. S. & Kowalczykowski, S. C. RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiol. Mol. Biol. Rev.* **72**, 642–71 (2008).

26. Radman, M. SOS repair hypothesis: phenomenology of an inducible DNA repair which is accompanied by mutagenesis. *Basic life sciences* **5 A**, 355–367 (1975).

27. Courcelle, J., Khodursky, A., Peter, B., Brown, P. O. & Hanawalt, P. C. Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient Escherichia coli. *Genetics* **158**, 41–64 (2001).

28. Zhang, A. P. P., Pigli, Y. Z. & Rice, P. A. Structure of the LexA-DNA complex and implications for SOS box measurement. *Nature* **466**, 883–886 (2010).

29. Sassanfar, M. & Roberts, J. W. Nature of the SOS-inducing signal in Escherichia coli. The involvement of DNA replication. *J. Mol. Biol.* **212**, 79–96 (1990).

30. Michel, B. After 30 years of study, the bacterial SOS response still surprises us. *PLoS Biology* **3**, 1174–1176 (2005).

31. Moore, J. K. & Haber, J. E. Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **16**, 2164–73 (1996).

32. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).

33. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science (80-.).* **339**, 819–823 (2013).

34. Standage-Beier, K., Zhang, Q. & Wang, X. Targeted Large-Scale Deletion of Bacterial Genomes Using CRISPR-Nickases. *ACS Synth. Biol.* **4**, 1217–1225 (2015).

35. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-γuided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).

36. Bikard, D. *et al.* Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res.* **41**, 7429–7437 (2013).

37. Bikard, D. *et al.* Development of sequence-specific antimicrobials based on programmable CRISPR-Cas nucleases. *Nat. Biotechnol.* **32**, 1146–1150 (2014).

38. Cohen, J. CRISPR patent fight revived. *Science (80-.).* **365**, 15.2-16 (2019).

39. Kallimasioti-Pazi, E. M. *et al.* Heterochromatin delays CRISPR-Cas9 mutagenesis but does

not influence the outcome of mutagenic DNA repair. *PLOS Biol.* **16**, e2005595 (2018).

40. Verkuijl, S. A. & Rots, M. G. The influence of eukaryotic chromatin state on CRISPR–Cas9 editing efficiencies. *Current Opinion in Biotechnology* **55**, 68–73 (2019).

41. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).

42. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).

43. Wong, N., Liu, W. & Wang, X. WU-CRISPR: Characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biol.* **16**, 218 (2015).

44. Xiao-Jie, L., Hui-Ying, X., Zun-Ping, K., Jin-Lian, C. & Li-Juan, J. CRISPR-Cas9: A new and promising player in gene therapy. *J. Med. Genet.* **52**, 289–296 (2015).

45. Zhang, X. H., Tee, L. Y., Wang, X. G., Huang, Q. S. & Yang, S. H. Off-target effects in CRISPR/Cas9-mediated genome engineering. *Molecular Therapy - Nucleic Acids* **4**, e264 (2015).

46. Vakulskas, C. A. *et al.* A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. *Nat. Med.* **24**, 1216–1224 (2018).

47. Cho, S. W. *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* **24**, 132–141 (2014).

48. Fu Y, Sander JD, Reyon D, Cascio VM & Joung JK. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32**, 279–84 (2014).

49. Shin, J. *et al.* Disabling Cas9 by an anti-CRISPR DNA mimic. *Sci. Adv.* **3**, (2017).

50. Wu, X. *et al.* Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32**, 670–676 (2014).

51. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–198 (2015).

52. Kim, D. *et al.* Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat. Methods* **12**, 237–43, 1 p following 243 (2015).

53. Farasat, I. & Salis, H. M. A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation. *PLOS Comput. Biol.* **12**, e1004724 (2016).

54. Jiang, F. & Doudna, J. A. CRISPR-Cas9 Structures and Mechanisms. (2017). doi:10.1146/annurev-biophys

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

56. SHIMOMURA, O., JOHNSON, F. H. & SAIGA, Y. Extraction, purification and properties of aequorin, a bioluminescent. *J. Cell. Comp. Physiol.* **59**, 223–239 (1962).

57. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W. & Prasher, D. C. Green fluorescent protein as a marker for gene expression. *Science (80-.).* **263**, 802–805 (1994).

58. Dickson, R. M., Cubittt, A. B., Tsient, R. Y. & Moerner, W. E. On/off blinking and switching behaviour of single molecules of green fluorescent protein. *Nature* **388**, 355–358 (1997).

59. Betzig, E. *et al.* Imaging intracellular fluorescent proteins at nanometer resolution. *Science (80-.).* **313**, 1642–1645 (2006).

60. Ashkin, A. Design for an optical cw atom laser. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 12108–12113 (2004).

61. Simmons, R. M. & Finer, J. T. Glasperlenspiel II. *Curr. Biol.* **3**, 309–311 (1993).

62. Simmons, R. Molecular motors: Single-molecule mechanics. *Curr. Biol.* **6**, 392–394 (1996).

63. Pilizota, T. *et al.* A molecular brake, not a clutch, stops the Rhodobacter sphaeroides flagellar motor. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 11582–11587 (2009).

64. Hamdan, S. M., Loparo, J. J., Takahashi, M., Richardson, C. C. & Van Oijen, A. M. Dynamics of DNA replication loops reveal temporal control of lagging-strand synthesis. *Nature* **457**, (2008).

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

66. Subach, F. V. *et al.* Photoactivatable mCherry for high-resolution two-color fluorescence microscopy. *Nat. Methods* **6**, 153–159 (2009).

67. Urh, M. & Rosenberg, M. HaloTag, a Platform Technology for Protein Analysis. *Curr. Chem. Genomics* **6**, 72–8 (2012).

68. Axelrod, D. *Cell-Substrate Contacts Illuminated by Total Internal Reflection Fluorescence*. *THE JOURNAL OF CELL BIOLOGY* **89**, (1981).

69. Manley, S. *et al.* High-density mapping of single-molecule trajectories with photoactivated localization microscopy. (2008). doi:10.1038/NMETH.1176

70. Yao, Z. & Carballido-López, R. Fluorescence Imaging for Bacterial Cell Biology: From Localization to Dynamics, From Ensembles to Single Molecules. *Annu. Rev. Microbiol.* **68**, 459–476 (2014).

71. Kapanidis, A. N., Lepore, A. & El Karoui, M. Rediscovering Bacteria through Single-Molecule Imaging in Living Cells. *Biophysical Journal* **115**, 190–202 (2018).

72. Elf, J., Li, G.-W. & Xie, S. Probing Transcription Factor Dynamics at the Single-Molecule Level in a Living Cell. *Science (80-.).* **316**, 1191–1194 (2007).

73. Simson, R. *et al.* Structural mosaicism on the submicron scale in the plasma membrane. *Biophys. J.* **74**, 297–308 (1998).

74. Saxton, M. J. & Jacobson, K. Single-particle tracking: Applications to membrane dynamics. *Annual Review of Biophysics and Biomolecular Structure* **26**, 373–399 (1997).

75. Michalet, X. 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 (2010).

76. Mean squared displacement - Wikipedia. Available at: https://en.wikipedia.org/wiki/Mean_squared_displacement.

77. Einstein, A. & Cowper, A. D. Investigations on the theory of the Brownian movement. *Dover Publ.* (1956).

78. Billaudeau, C. *et al.* Contrasting mechanisms of growth in two model rod-shaped bacteria. *Nat. Commun.* **8**, (2017).

79. Jones, D. L. *et al.* Kinetics of dCas9 target search in Escherichia coli. *Science (80-.).* **9**, 1420–1424 (2017).

80. Shibata, M. *et al.* Real-space and real-Time dynamics of CRISPR-Cas9 visualized by high-

speed atomic force microscopy. *Nat. Commun.* **8**, (2017).

81. Knight, S. C. *et al.* Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science* **350**, 823–6 (2015).

82. Schleif, R. AraC protein, regulation of the L-arabinose operon in Escherichia coli, and the light switch mechanism of AraC action. *FEMS Microbiol. Rev.* **34**, 779–796 (2010).

83. St-Pierre, F. *et al.* One-step cloning and chromosomal integration of DNA. *ACS Synth. Biol.* **2**, 537–541 (2013).

84. Martin, K., Huo, L. & Schleif, R. F. The DNA loop model for ara repression: AraC protein occupies the proposed loop sites in vivo and repression-negative mutations lie in these same sites. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 3654–3658 (1986).

85. Schleif, R. AraC protein: A love-hate relationship. *BioEssays* (2003). doi:10.1002/bies.10237

86. Huisman, O. & D'Ari, R. Effect of suppressors of SOS-mediated filamentation on sfiA operon expression in Escherichia coli. *J. Bacteriol.* **153**, 169–75 (1983).

87. Los, G. V. *et al.* HaloTag: A novel protein labeling technology for cell imaging and protein analysis. *ACS Chem. Biol.* **3**, 373–382 (2008).

88. Liss, V., Barlag, B., Nietschke, M. & Hensel, M. Self-labelling enzymes as universal tags for fluorescence microscopy, super-resolution microscopy and electron microscopy. *Sci. Rep.* **5**, (2015).

89. Grimm, J. B. *et al.* A general method to improve fluorophores for live-cell and single-molecule microscopy. *Nat. Methods* **12**, 244–250 (2015).

90. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).

91. Murphy, K. C., Campellone, K. G. & Poteete, A. R. PCR-mediated gene replacement in Escherichia coli. *Gene* **246**, 321–330 (2000).

92. Chung, C. T., Niemela, S. L. & Miller, R. H. One-step preparation of competent Escherichia coli: Transformation and storage of bacterial cells in the same solution (recombinant DNA). *Pnas* **86**, 2172–2175 (1989).

93. Mix and Go! E.coli Transformation Kit. Available at: https://www.zymoresearch.com/products/mix-and-go-e-coli-transformation-kit.

94. Medical Uv Crosslinker Factory and Suppliers China - Customized Products Wholesale - Scientz. Available at: http://www.scientzbio.com/uv-fixer/medical-uv-crosslinker.html.

95. MetaMorph, Microscope Imaging, Microscopy Analysis Software | Molecular Devices. Available at: https://www.moleculardevices.com/products/cellular-imaging-systems/acquisition-and-analysis-software/metamorph-microscopy.

96. Liu, C., Liu, Y. L., Perillo, E. P., Dunn, A. K. & Yeh, H. C. Single-Molecule Tracking and Its Application in Biomolecular Binding Detection. *IEEE J. Sel. Top. Quantum Electron.* **22**, (2016).

97. Jaramillo-Riveri, S. I. Growth influences the single cell variability of the DNA damage response in Escherichia coli. *Edinburgh Res. Arch.* (2019).

98. Detect Cell Using Edge Detection and Morphology - MATLAB & Simulink Example - MathWorks United Kingdom. Available at: https://uk.mathworks.com/help/images/detecting-a-cell-using-image-segmentation.html;jsessionid=b00c683923786df87a8ba12fa355.

99. Lepore, A. *et al.* Quantification of very low-abundant proteins in bacteria using the HaloTag

and epi-fluorescence microscopy. *Sci. Rep.* **9**, (2019).

100. Okumus, B. *et al.* Mechanical slowing-down of cytoplasmic diffusion allows in vivo counting of proteins in individual cells. *Nat. Commun.* **7**, 11641 (2016).

101. Matlab Particle Tracking. Available at: http://site.physics.georgetown.edu/matlab/.

102. Crocker, J. & Grier, D. Methods of Digital Video Microscopy for Colloidal Studies. *J. Colloid Interface Sci.* **179**, 298–310 (1996).

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

104. Bell, J. C. & Kowalczykowski, S. C. RecA: Regulation and Mechanism of a Molecular Search Engine. *Trends Biochem. Sci.* **41**, 491–507 (2016).

105. Khlebnikov, A., Datsenko, K. A., Skaug, T., Wanner, B. L. & Keasling, J. D. Homogeneous expression of the PBAD promoter in Escherichia coli by constitutive expression of the low-affinity high-capacity araE transporter. *Microbiology* **147**, 3241–3247 (2001).

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

107. RecBCD coordinates repair of two ends at a DNA double-strand break, preventing aberrant chromosome amplification. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6061781/. (Accessed: 12th April 2020)

108. Cui, L. & Bikard, D. Consequences of Cas9 cleavage in the chromosome of Escherichia coli. *Nucleic Acids Res.* **44**, 4243–4251 (2016).

109. Cox, M. M. *et al.* The importance of repairing stalled replication forks. *Nature* **404**, 37–41 (2000).

110. Stracy, M., Uphoff, S., Garza De Leon, F. & Kapanidis, A. N. In vivo single-molecule imaging of bacterial DNA replication, transcription, and repair. *FEBS Letters* **588**, 3585–3594 (2014).

111. Barrangou, R. Cas9 targeting and the CRISPR revolution. *Science* **344**, 707–708 (2014).

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

113. Rocha, J., Corbitt, J., Yan, T., Richardson, C. & Gahlmann, A. Resolving Cytosolic Diffusive States in Bacteria by Single-Molecule Tracking. *Biophys. J.* **116**, 1970–1983 (2019).

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

115. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).