# Winning the cellular lottery: how proteins reach and recognize targets in DNA

**Sy Redding**

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

**COLUMBIA UNIVERSITY**

2015

# ABSTRACT

## Winning the cellular lottery: how proteins reach and recognize targets in DNA


## Sy Redding



Many aspects of biology depend on the ability of DNA-binding proteins to locate specific binding sites within the genome. This type of search process is required at the beginning of all site-specific protein-DNA interactions, and has the potential to act as the first stage of biological regulation. Given the difficulty of pinpointing a small region of DNA, within even simple genomes, it is expected that proteins are adapted to use specialized mechanisms, collectively referred to as facilitated diffusion [Berg *et al.*, 1981], to effectively reduce the dimensionality of their searches, and rapidly find their targets. Here, we use a combination of nanofabricated microfluidic devices and single-molecule microscopy to determine whether facilitated diffusion contributes to all DNA target searches. We investigate promoter binding by *E. coli* RNA polymerase, foreign DNA recognition by CRISPR-Cas complexes, and Rad51's homology search during recombination. In each example, we observe that the target searches proceed without extensive use of facilitated diffusion; rather, consideration of these non-facilitated target searches reveals an alternative search strategy. We show that instead of reducing the dimensionality of their searches, these proteins, reduce search complexity by minimizing unproductive interactions with DNA, thereby increase the probability of locating a specific DNA target.

# Table of Contents

# List of Figures

# Acknowledgments

I have been incredibly lucky over my graduate career to work with many fantastic scientists, several of which made significant contributions to this work. These contributions are also highlighted in the chapter headings. First, the study of RNA polymerase was a collaboration with Dr. Feng Wang. Second, experiments involving CRISPR-Cas complexes were a joint effort with Dr. Samuel Sternberg, with contributions from Prof. Jennifer Doudna, Prof. Blake Wiedenheft, Prof. Martin Jinek, Myles Marshall and Dr. Bryan Gibb. Third, the work presented on the mechanism of homologous recombination was in large part the efforts of Dr. Zhi Qi, in collaboration with Dr. Jayil Lee. Finally, I would like to thank Prof. Eric Greene, who taught me how to do science.

for Theodora

# Preface

DNA constitutes the genetic basis of all organismal life on our planet. However, because the information encoded in DNA follows a simple four-letter code, the amount of DNA necessary to build and operate an organism is enormous. For example, to manufacture and maintain a human being for their entire life requires the synthesis and repeated interpretation of roughly one light year of DNA. This amount of molecular information seems too immense to be looked through, even once. And yet, there is no end to examples in biology where not only is all of the information in DNA processed by the machinery of life, but it is done so accurately and efficiently.

This laborious business of DNA processing occurs in the cell. In humans, each cell contains a personal operating manual, written out in roughly one meter of DNA, packed in the space of a few thousand cubic micrometers. This DNA is maintained, read and interpreted, and replicated by droves of molecules called proteins. Through some fantastic mechanism(s), proteins in the cell are able to zero in on single pieces of DNA-based information, out of billions, at precisely the right time to play a specific role in the life of the organism.

To capture just how magnificent this feat is, it is helpful to translate the job of a protein to our own scale. An analogous task, for you or I, would be something like being set adrift in the ocean, and charged with locating a particular square meter of seawater. It is improbable that either of us would accomplish this goal before we expired; yet, our search would still be a bit easier than the protein's, because you and I, can (presumably) swim. Proteins rely on random fluctuations, effectively riding the waves of a cellular ocean until by happenstance they bump into the right place, whereas, people are active agents in determining their motion. A fairer analogy would require us to passively ride natural ocean currents, not that this condition makes the idea of finding a small square of ocean any more intractable.

The point is, the task set out for proteins in the cell, seems, at the outset, to be impossible, and, at the same time, central to the workings of all living organisms. Determining what is missing in the above picture of how proteins perform their cellular functions, and what tools the cell has at its disposal to drive proteins into achieving the impossible, is the subject of this thesis.

# Chapter 1

# DNA, DNA-binding proteins, and the biological target search

## 1.1 Introduction

The purpose of this first chapter is to orient the reader to how we think about DNA and DNA-binding proteins. We begin by describing lactose metabolism in *Escherichia coli* to highlight the role protein-DNA interactions play in biology and to introduce key concepts. Then we discuss the nature of DNA and DNA-binding proteins and how their makeup defines what happens when they come close to one another. Next, we discuss mechanisms proteins use to select for specific sequences of DNA. Then we introduce the single molecule experimental assay we use to probe protein-DNA interactions. Finally, we present two pairs of examples of single protein-DNA interactions. These examples confirm certain aspects of our interpretation of how DNA-binding proteins must search for target sites, but challenge other features, suggesting that the current interpretation is somewhat incomplete.

## 1.2 The *Lac* Operon

A classic example of how biology can turn on a single protein-DNA interaction is lactose metabolism in *E. coli* (Fig. 1.1). In general, *E. coli* prefer to eat glucose, but when times get tough, they will resort to other sugars, such as lactose [Moses and Prevost, 1966]. One reason for this preference is

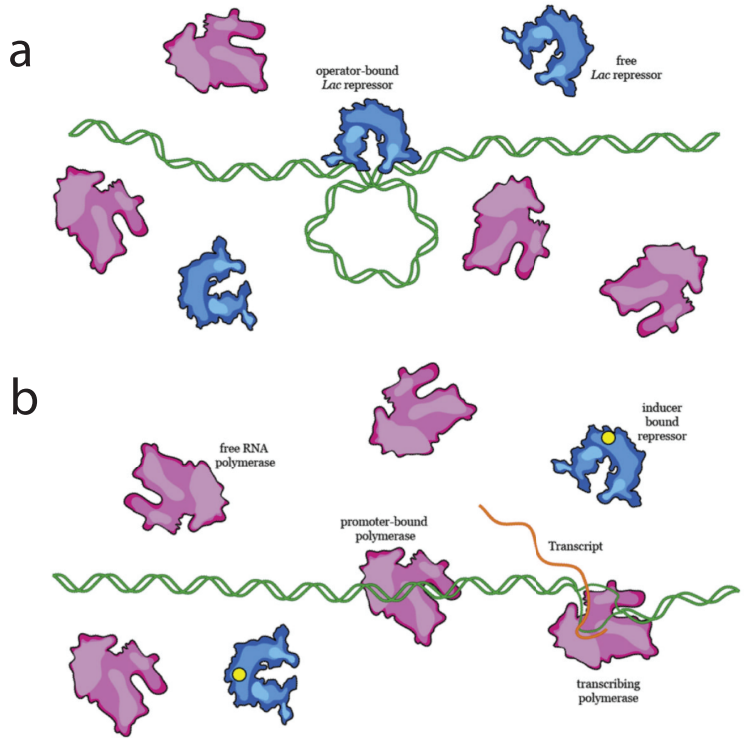


Figure 1.1: Regulation of the lac operon. (a) In the absence of lactose, the lac repressor protein binds to an operator sequence and prevents RNA polymerase from accessing its promoter. (b) When lactose is present, lac repressor dissociates from the DNA, allowing RNA polymerase to bind to its promoter and transcribe the genes necessary for lactose metabolism.

that, in order to eat lactose, the cell needs to produce additional enzymes. Within the bacterial chromosome, a set of genes called the *lac* operon, code for these proteins: *lac*Y encodes lactose permease, which cells need to acquire lactose from the environment; the *lac*Z gene encodes $\beta$-galactosidase, which breaks lactose down into the preferred glucose and galactose, which the cell can either further digest into glucose or discard; and the *lac*I gene encodes the *lac* repressor, which can shut down the cell's ability to express the *lac* operon encoded proteins whenever lactose is scarce [Lewis, 2005; Lewis, 2011; Beckwith, 2011]. Expression of the *lac*Y, *lac*Z, and *lac*I genes requires yet another protein, RNA polymerase. RNA polymerase binds to the DNA at the start of the *lac* operon, called the promoter, and copies the DNA into messenger RNA (mRNA) that will serve as the blueprint for the construction of *lac* proteins by ribosomes [Saecker *et al.*, 2011].

When glucose is available, individual cells only have a handful of *lac* proteins. This small supply of proteins is insufficient to meet the cell's energy needs in the event of a drastic change in carbon source; for example, if the bacteria's host decides to drink a glass of milk. To respond, the cell

needs lactose permease to pump the newfound lactose into the cell, it needs $\beta$-galactosidase to cleave lactose into simple sugars, and it needs RNA polymerase to immediately start transcribing the *lac* genes. Before any of this can happen, the *lac* repressor must switch the operon from off to on.

To turn the operon off, the *lac* repressor binds to the *E. coli* genome upstream of the *lac* genes, outcompeting RNA polymerase, and twists up the DNA, preventing RNA polymerase from gaining access to the promoter [Lewis, 2011; Lewis, 2005; Becker *et al.*, 2013] (Fig. 1.1a). However, in response to environmental changes, i.e., a milk bath, the operon needs to turn on. Importantly, when the *lac* repressor binds to lactose, it changes conformation, resulting in a rapid release of operator DNA. Now, with the repressor out of the way, the polymerase can bind to the promoter and express the operon [Lewis, 2011; Lewis, 2005] (Fig. 1.1b).

This transition is critical: if the *lac* repressor fails to dissociate from the operator site in the presence of lactose, then the cell incurs a competitive disadvantage relative to neighboring cells that are otherwise capable of faithfully regulating the *lac* operon. Conversely, if the *lac* repressor is slow to find the operon, or poor at fighting RNAP for purchase on the promoter, the operon will remain on and the cell will waste valuable energy producing *lac* proteins even in the absence of lactose.

The details of the *lac* system as outlined above are unique to the *lac* pathway, but, incredibly, the general structure of the regulation pathway is a common one in biology. Specifically, a great deal of biological regulation relies on one or more site-specific DNA-binding proteins both locating and recognizing specific locations of the genome at precisely the right time, so that they can turn on or off a particular gene, or set of genes. Thus, understanding how processes like the *lac* system function on the molecular level, and how that action translates into the reality of the cell, can give insight into life at all scales. The efficient design of the lactose metabolic pathway is remarkable because it is carried out through what are seemingly random events; namely, both the binding of the *lac* repressor and RNAP to the operon. This raises an important question: how do proteins like the *lac* repressor locate a specific region of a large molecule of DNA?

## 1.3 What is DNA binding?

To understand how the *lac* repressor works to achieve regulation, we first consider the nature of DNA and of protein-DNA interactions. DNA is comprised of two sugar-phosphate chains that wrap around one another to give DNA its right-handed double-helix shape and its overall negative charge [Watson and Crick, 1953; Wang *et al.*, 1982; Larsen *et al.*, 1991]. Connected to this scaffold are the DNA bases, adenine (A), guanine (G), cytosine (C), and thymine (T). As the negatively charged backbone wraps around the bases, it allows access to the base-pairing interior exclusively through two grooves that run along the molecule: a wide and shallow major groove and a narrow and deep minor groove [Rohs *et al.*, 2010; Shakked and Rabinovich, 1986]. This DNA double helix, is biology's preferred method of storage for its most important information: how to build a life. However, once stored in this structure, how does the cell access this information?

Consider how the DNA appears to a protein. The regularly spaced phosphates along the DNA molecule all carry a negative charge and can interact with positively charged amino acids in proteins, drawing the protein to the DNA [Jones *et al.*, 2003]. The energetic size of these interactions is generally large enough to pay for the energy lost in capturing a protein for a short amount of time [Lohman *et al.*, 1980]. Interactions like this, which rely almost exclusively on binding to the DNA backbone, are called non-specific [von Hippel and Berg, 1986]. This is because, while the chemical nature of DNA along its length is unique, a great deal of the electrostatic potential is not [1].

Non-specific binding does not read out the information in DNA. In order to do that, the protein must reach into the grooves of DNA and make direct contact with the DNA bases. It can then read the DNA by testing specific amino acids in the protein against the bases of DNA for their ability to form hydrogen bonds or hydrophobic contacts [Seeman *et al.*, 1976; von Hippel and Berg, 1986]. This type of binding is referred to as specific binding. For the *lac* repressor to bind specifically requires $\approx 36$ amino acids scattered across four separate *lac* repressors within a tetrameric complex, which recognize a pair of operator sequences 21-23 nucleotides in length [Kalodimos *et al.*, 2002]. But every protein has its own subset of amino acids that it uses as a cipher to interpret the

[1] This statement is overly simplified, and as a result conflates features of DNA recognition by proteins; e.g., shape readout of the major and minor grooves, which is a mechanism of specific binding (see below), but occurs through electrostatic interactions with the phosphate backbone. For an excellent discussion of recognition mechanisms, see R. Rohs, et al., 2010.

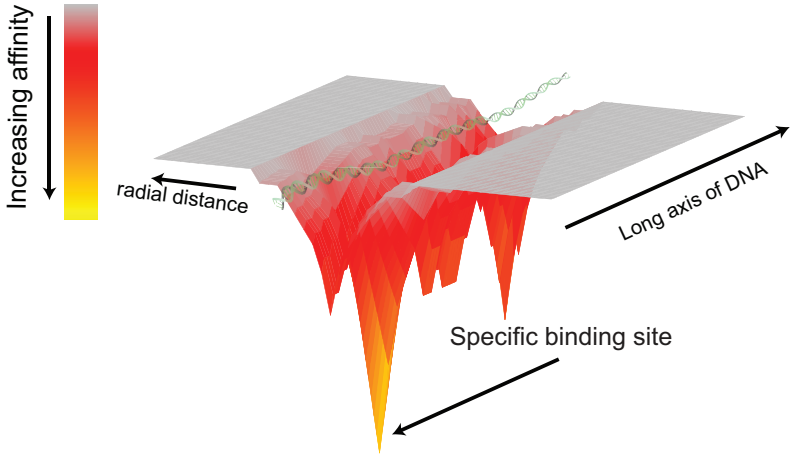


Figure 1.2: Specific and non-specific DNA binding. Schematic representation of a hypothetical three-dimensional surface plot showing the radial and longitudinal dependence of the binding energy for a hypothetical protein and a DNA molecule (in green).The minimum in the energy landscape would correspond to a specific binding site

information coded in DNA. On average, 24 protein residues participate to read out 12 nucleotides [Rohs *et al.*, 2010], and by altering the amino acids involved in DNA recognition, the cell can define how the DNA looks from the perspective of an individual protein. The issue of guiding the protein to a particular site then becomes a matter of ensuring that when a protein binds to DNA, the contacts between the protein and DNA are such that change in free energy is substantially lower at certain sites relative to other locations (Fig. 1.2) [Berg and von Hippel, 1987]. In general, this is axiom is true; proteins that recognize specific DNA sequences have a markedly lower energy at their specific site than all other sites [Kao-Huang *et al.*, 1977]. Figure 1.2 shows a cartoon of a hypothetical energetic landscape of a protein DNA interaction; that is, it shows what the DNA looks like to a particular protein.

It is important to realize that the chemistry that gives rise to protein recognition of specific sites affects how proteins interact with non-specific sites, as well. Consider again the contacts made between the *lac* repressor and its operator. If we reach into that interface and disrupt one of these interactions, the protein will retain substantial affinity for the operator sequence due to the remaining interactions; however, the free energy for this new complex will be greater [Betz *et al.*, 1986; Frank *et al.*, 1997]. If we then go into the complex and disturb yet another interaction, the free energy will likely rise again [Frank *et al.*, 1997], a fact borne out of the wealth of experimental investigations of target site mutations on binding energies. Allowing that

all protein-DNA complexes can be realized by a continuous perturbation of this toy model leads to the conclusion that a great deal of the affinity of a protein for any sequence of DNA arises from its ability to interact specifically with its specific target site (Fig. 1.2).

As a further corollary of the above, it is worth noting that, while it is possible to have a protein that binds DNA non-specifically but does not have a highly preferred specific target, the converse is not true: all site-specific DNA-binding proteins inherently possess some non-negligible affinity for non-specific DNA. The physical arguments above notwithstanding, for site-specific DNA-binding proteins, the absolute distinction between specific and nonspecific binding, being bound in the right place versus the wrong place, is really a question of biological relevance: proteins perform biological functions while bound to a specific sites.

## 1.4 The biological target search

Now that we have an idea of how the DNA and proteins see each other, we can address the question of how a protein searches across the genome for a specific site. You might imagine, from the myriad functions carried out by the billions of DNA-binding proteins across all kingdoms of life, that there would be innumerable mechanisms for locating specific target sites. However, despite the diversity of biology and biological processes to which they belong, many aspects of target searches are generally the same.

Each protein begins at the ribosome. Once assembled, it bounces around in the cell, riding thermal gradients, until it randomly encounters DNA. When the protein approaches the DNA, it is pulled close by the negatively charged backbone, while residues on the proteins surface help to orient the protein so that it can interrogate the captured sequence. The bound protein, with its critical residues extended into the DNA helix, then makes a calculation. It adds up the energy across the protein-DNA interface, some contacts paying energetic dividends while others yield losses. The best case for the protein is that, by pure chance, it has landed directly on its target site. However, the more likely scenario is that the protein has landed in the wrong place. In this case, the protein spends an amount of time, proportional to the strength of the interaction, bound to the DNA before falling off. At this point, the protein finds itself back in the cellular milieu, carried along by the random fluctuations of the cell, hopeful that it should reunite with the DNA, ideally, in the right

place.

For a site-specific DNA-binding protein, finding the target in this manner is daunting due to the sheer excess of non-target DNA. For example, there are three *lac* operators in the entire *E. coli* genome $\left(\approx 6 \cdot 10^6 \text{ base pairs in length}\right)$ [Lewis, 2005], and, finding one of these operators by chance, is a one-in-a-million occurrence. The initial journey to the DNA 99.9999% of the time leaves the protein bound to a random, non-specific sequence of DNA unrelated to the protein's specific biological function. One might well ask, how on earth biology works like this. How is anything alive if even the simplest of biological systems is working against those kinds of odds?

Remarkably, experiments monitoring the *lac* repressor as it searches for its operator sequence are inconsistent with the search model described above, where the protein relies solely on random collisions to find the target, referred to as a three dimensional (3D) search [Riggs *et al.*, 1970]. This finding initially appeared in an influential paper by Riggs, Bourgeois and Cohn, wherein the authors measured the kinetics of the *lac* repressor and *lac* operator [Riggs *et al.*, 1970]. Their measurement of the association rate (the rate at which the repressor finds the target from solution) was conspicuously large; the protein was getting to the operator too fast. In fact, the rate measured by Riggs, et al. was one and half orders of magnitude too fast. This result indicated that biology had in fact not left the regulation of the *lac* operon up to chance, because in this case, too fast means faster than would be expected from considering the protein and DNA operator association as a random collision. In discussing potential origins of this measured behavior, the authors offered up the following hypothesis.

> *It is therefore worth considering an extreme model of oriented diffusion. This model is that the lac repressor searches for the operator not by performing a three-dimensional random walk, but rather by binding to DNA and rolling or hopping along it, thus reducing the search for operator to only two dimensions.* [Riggs *et al.*, 1970]

The authors actually offer this explanation as a straw man and go on to explain how this proposed mechanism is unlikely to be the culprit in the enhanced rate [Riggs *et al.*, 1970]. Nevertheless, the result sparked a great deal of interest in the *lac* repressor's association kinetics and the kinetics of all protein DNA interactions [Hammar *et al.*, 2012; Elf *et al.*, 2007; Austin *et al.*, 1983; Gorman *et al.*, 2012], and within the decade, a mathematical treatment had been applied to the

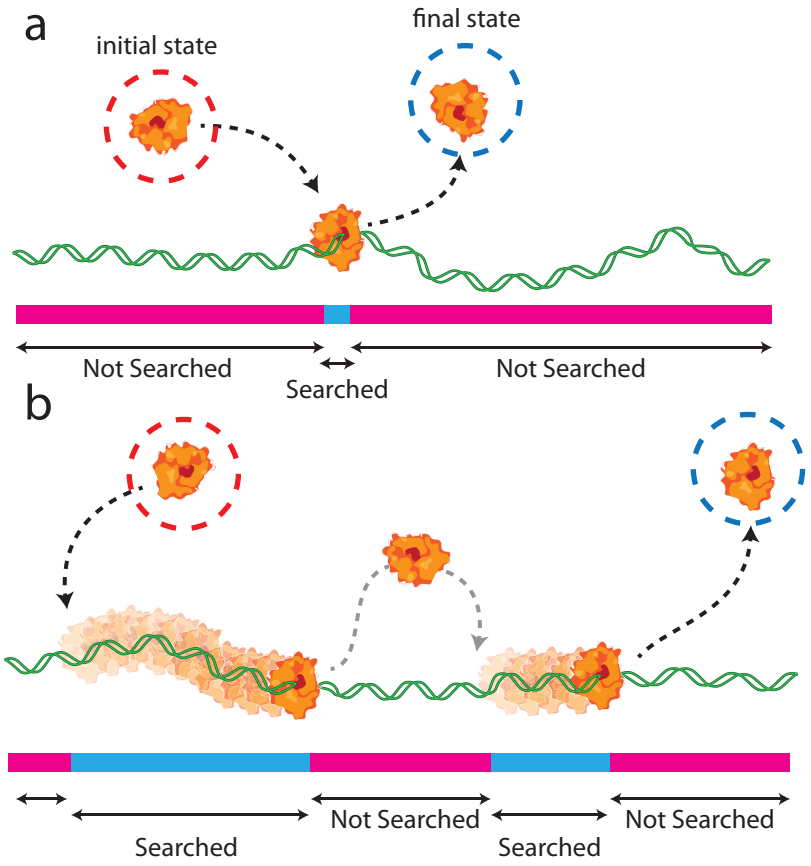


Figure 1.3: The facilitated diffusion model. (a) A purely 3D search, the protein binds to the DNA from solution, and interrogates only the DNA at the collision, before dissociating back into solution. The initial state (red circle) and the final state (blue circle) are both equilibrated states. (b) The facilitated search; the protein starts in solution (red circle), then binds to the DNA and engages in 1D sliding before dissociating from the DNA. Importantly, the protein rebinds to the DNA before equilibrating in solution, and engages in 1D sliding again before finally dissociating into and equilibrium solution state (blue circle).

straw man of Riggs et al. This theory, named the facilitated diffusion model, was in good agreement with experimental observations of repressor behavior, and seemed to provide a satisfactory answer to the question of how proteins find their targets [Berg *et al.*, 1981; Berg and Blomberg, 1976; Slutsky and Mirny, 2004; Halford and Marko, 2004].

### 1.4.1 Facilitated diffusion

The model of facilitated diffusion comes about from considering the role that non-target sites play along the path to target sites for a particular protein. Consider again the 3D search from above; the protein begins in solution before randomly encountering the DNA, at which point,

if the protein has found its target, the search is over. On the other hand, if the protein is in the wrong place, it falls off, returning to solution (Fig. 1.3a). The facilitated diffusion model posits that, while the protein is close and in contact with the DNA but in the wrong place, it should use that opportunity to take a look around at the neighboring DNA. It does this by either diffusing randomly while maintaining contact with the DNA or by disengaging momentarily before rebinding to the DNA at a nearby location (Fig. 1.3b). The key to the facilitated diffusion model is that the time between the initially unbound state and the final dissociated state for either search mechanism is the same. Importantly, in the 3D search model, the protein interrogates only one potential site in DNA for the target; however, in the case of a facilitated search, the protein has the opportunity to interrogate multiple sites during the same global binding event (Fig. 1.3). It then becomes clear how the facilitated diffusion model can explain the enhanced association rate of the *lac* repressor, suggesting that each time the *lac* repressor binds to DNA it scans multiple sites along the dimension of the DNA in search of the *lac* operator. While the facilitated diffusion model can be framed in a more mathematically sophisticated [Berg *et al.*, 1981; Berg and Blomberg, 1976; Slutsky and Mirny, 2004; Halford and Marko, 2004], or unnecessarily complicated way (Fig. A.1), the main ideas of the theory are captured in the above discussion. Ultimately, to verify such a model requires well-defined experiments[Winter *et al.*, 1981], or direct visualization of a protein as it searches for its target site [Wang *et al.*, 2013a; Wang *et al.*, 2006]. Accordingly, several studies, *in vivo*, *in vitro*, and *in silico*, have shown that the facilitated diffusion model is an accurate explanation of how the *lac* repressor, and some other site-specific DNA-binding proteins, finds their targets [Hammar *et al.*, 2012; Elf *et al.*, 2007; Wang *et al.*, 2006; Austin *et al.*, 1983; Wang *et al.*, 2013a; Koslover *et al.*, 2011; Schonhoft and Stivers, 2012]. What it is not clear is whether this model can be generalized to all site-specific DNA-binding proteins[Halford, 2009].

## 1.5 DNA curtains

To answer questions of how proteins interact with DNA, we use a technique called DNA curtains (Fig. 1.4a) [Greene *et al.*, 2010; Finkelstein and Greene, 2011]. This method allows us to arrange hundreds of individual molecules of DNA on the surface of a microscope slide (Fig. 1.4a). We then use fluorescence microscopy to watch individual proteins as they interact with the DNA in

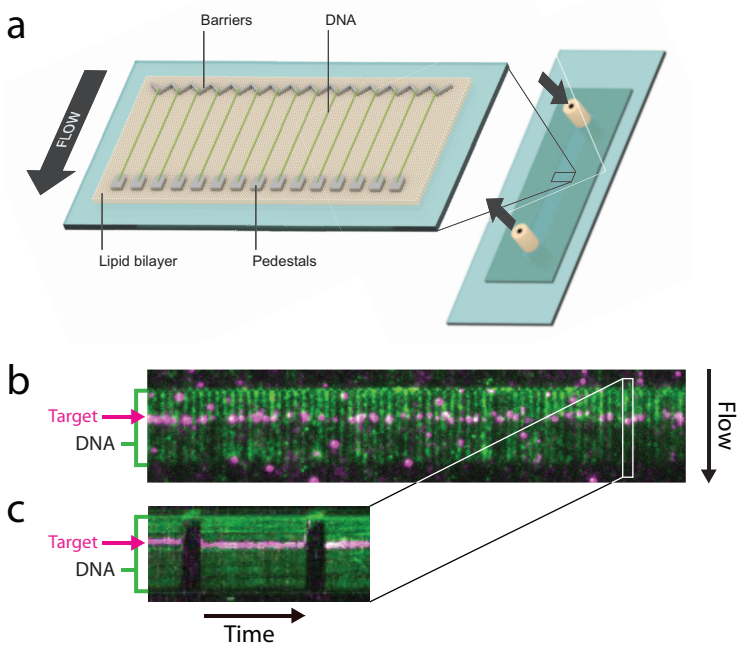


Figure 1.4: The DNA curtain assay. (a) schematic of flowcell. (b) Still image from a DNA curtain experiment showing several DNA molecules (green) bound by a site-specific DNA-binding protein (magenta). The binding site is indicated by the magenta arraw. The white box indicates the region of the corresponing movie chosen to project the kymogram in (c). (c) kymogram showing the the response of single tethered DNA to force

the curtain (Fig. 1.4b,c) [Greene *et al.*, 2010; Finkelstein and Greene, 2011].

The DNA curtain assay begins with manufacture of microscope slides with specialized features to capture and arrange DNA molecules in the experiment. These features, called barriers and pedestals (Fig 1.4a), are drawn onto the slide surface using a combination of electron-beam lithography and metal evaporation. Then by attaching a coverslip to these slides, we create a small reaction chamber where an experiment can be carried out (Fig. 1.4a).

Once the flowcell is assembled, the slide surface is coated with a lipid bilayer, where the lipid molecules in the bilayer freely diffuse in two dimensions. Importantly, the lipid molecules do not create a continuous bilayer, but instead are discontinuous across the nanofabricated barriers and pedestals. Among the lipids comprising the bilayer, a small number are biotinylated. We use these lipids to attach DNA to the bilayer; the DNA is also biotinylated, such that a single streptavidin is able to link the two. By applying flow through the flowcell, the DNA molecules are subjected to a drag force, which in turn causes the lipid they are attached to migrate through the bilayer with the DNA acting as a sail. The migrating lipid stops at the barrier due to the discontinuity in the bilayer,

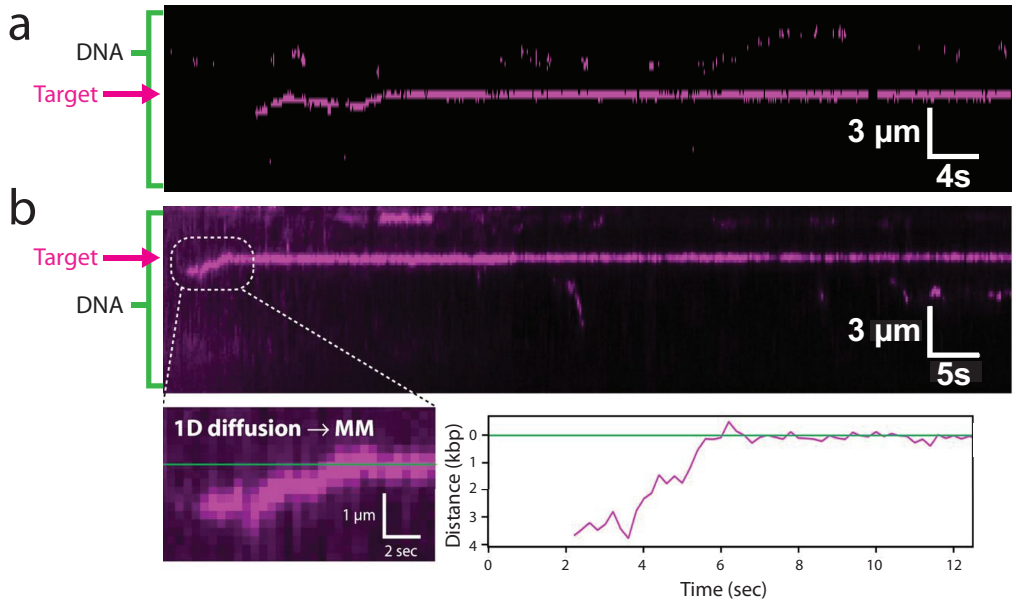


Figure 1.5: Single molecule examples of facilitated diffusion.

and all of the force on the DNA molecule is then transferred to extending the DNA in the direction of the flow (Fig. 1.4a). To illuminate fluorescently labeled DNA or proteins bound to the DNA, we use total internal reflection fluorescence microscopy (TIRFM). Breifly, a laser beam is reflected off the interface between the glass slide and the buffer in the flowcell, creating an exponentially diminishing wave that penetrates a shallow distance into the buffer, illuminating only fluorescent molecules near the surface, where the DNA curtains have been established [Greene *et al.*, 2010; Finkelstein and Greene, 2011] (Fig. 1.4b).

The data collected in a DNA curtain experiment comes in the form a movie, so we typically use a projection of the movie to represent the data; these projections are called kymograms (Fig. 1.4c). Kymograms are created by selecting the region of a movie encompassing a single DNA molecule and laying out each frame of the movie side by side. Therefore, the y-axis of a kymogram measures the position along a DNA molecule, and the x-axis measures the time (Fig. 1.4c).

## 1.6 Two cases that work

If the facilitated diffusion model answers the question of how proteins find specific sites in DNA, then we should be able to see the hallmarks of that activity in a DNA curtain experiment. As a first

example, we return to the *lac* repressor. This experiment uses DNA from bacteriophage $\lambda$, which has been altered to contain a single *lac* operator. The operator represents less than one fiftieth of one percent of the DNA length when stretched out in a DNA curtain. Then, by introducing fluorescently labeled *lac* repressor, we can visualize all of the interactions between the repressor and the DNA, binding both to its operator and to non-specific sites.

Figure 1.5a shows an example of the repressor, labeled in magenta, searching for and finding its operator sequence. As expected from the facilitated diffusion model, the *lac* repressor first binds to the DNA at a non-operator site. Then, held by its non-specific binding energy to the DNA, it begins to diffuse in one dimension along the length of the DNA until it encounters the operator sequence, indicated by the magenta arrow. Upon arrival at the target, the *lac* repressor then locks into a more stable binding form, and persists in that location for the duration of the experiment.

We introduced the *lac* repressor as an example of a protein with a overly tedious job to perform, finding the *lac* operator. But, it not only accomplished this goal, but did so much faster than expected. To explain this finding, we introduced a model of how DNA is surveyed by proteins, and here we show direct evidence that the arguments in the facilitated diffusion model are supported by reality. It is worth noting that the above constitutes the first direct visualization of the *lac* repressor searching for and finding its target [Wang *et al.*, 2013a], though previously it had been shown that the repressor could exhibit one-dimensional diffusion on DNA [Elf *et al.*, 2007].

Consider as a second example the mismatch repair proteins Msh2 and Msh6, which form a single complex (Msh2-6) in budding yeast [Gorman *et al.*, 2010]. The job of Msh2-6 in the cell is to follow around replication machinery and proofread newly synthesized DNA [Kunkel and Erie, 2005]. When replication mistakes are made, Msh2-6 binds these mismatches and recruits several proteins which, together, repair the DNA [Kunkel and Erie, 2005]. To visualize Msh2-6 searching for mispaired DNA, we generated a $\lambda$-DNA bearing three mismatched bases separated by 39 base pairs [Gorman *et al.*, 2012]. The kymogram in figure 1.5b shows Msh2-6 locating one of these mispaired bases (magenta arrow) by first binding distal to the target, and then engaging in one-dimensional diffusion to slide into alignment [Gorman *et al.*, 2012].

Here we have two proteins from two different kingdoms of life employing the same mechanisms to find their specific targets, suggesting that the qualities of the facilitated diffusion model common to all site-specific DNA-binding proteins [Halford and Marko, 2004].

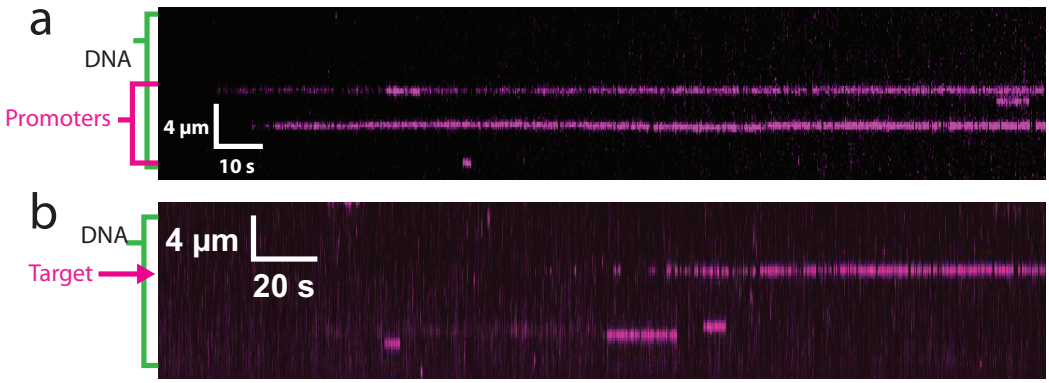


Figure 1.6: Single molecule examples of 3D searches.

## 1.7 Two cases that do not work

Given the apparent difficulty of protein target searches, it is attractive to assume that all DNA-binding proteins use facilitating mechanisms to speed up their respective search processes. However, figure 1.6 shows examples of two proteins which are, in the case of facilitated diffusion, rule breakers.

The first exception to the rule is *E. coli* RNA polymerase (RNAP) (Fig. 1.6a), which is the protein responsible for synthesis of all RNA in the bacterial cell [Ebright, 2000]. Using our DNA curtain assay, we visualized RNAP binding to promoter sequences native to the $\lambda$ phage genome. These promoters are scattered across one half of the DNA molecule (magenta bracket), and the kymogram (Fig. 1.6a) shows fluorescently labeled RNAP localizing to promoter sequences. Curiously, there is no indication in these experiments that RNAP is using facilitated diffusion; instead, the data suggest that RNAP directly binds to promoter sequences from solution by using a purely 3D search. The details of RNAP's interactions with DNA and the consequences of those interactions on protein target searches will be examined in Chapter 2.

Another example of a target search proceeding exclusively through a 3D search is the search for foreign DNA by the CRISPR protein Cas9 [Sternberg *et al.*, 2014]. Many bacteria utilize an RNA-based immune system referred to as the CRISPR-Cas system to hunt down and destroy foreign DNA [Wiedenheft *et al.*, 2012]. In *Streptocacus pyogenes*, more commonly known as flesh eating bacteria, the protein responsible for finding and degrading invading DNA (i.e., targets) is Cas9 [Jinek *et al.*, 2012]. To do this, Cas9 carries around a short piece of RNA that identifies foreign DNA sequences. When it locates a foreign DNA, it then uses two nuclease domains to cut the DNA

in half [Jinek *et al.*, 2012]. Using an RNA that targeted Cas9 to $\lambda$-DNA, we monitored Cas9 as it bound to DNA in a curtain experiment. The kymogram in figure 1.6b reveals that Cas9 also finds its targets (magenta arrow) directly from solution, without any evidence of facilitated diffusion. The mechanism of searching for and recognizing viral DNA in CRISPR immune systems will be explored further in Chapter 3.

Tied up in these two pieces of data is the core question this work seeks to address. We have two proteins, both of which play difficult and important roles in the cell, and yet, they do not use what is widely considered the fastest route to target binding. RNAP, arguably the most important protein in the cell, must find and transcribe thousands of individual genes over the course of the cell's life, and failure to do so results in death. Likewise, Cas9's job is to sniff out foreign invaders before they have a chance to take over the cell and murder their host in an effort to propagate their genes. If Cas9 fails, the cell will likely also die. Yet, in both of these cases, it seems the cell is fine with leaving the success or failure of these target searches up to chance.

On the other hand, the *lac* repressor and Msh2-6 both showed evidence of facilitated search processes. But, while both proteins are important to the cell, neither is essential for life. If either fails to do their job, the cell is disadvantaged, but, will by no means, die. The trend should be the opposite: the less essential a target search is for life, the less likely it should be that that search would employ facilitating mechanisms. Possibly, the premise I have presented is hollow; perhaps there is very little pressure on the cell to optimize searches because waiting for a one-in-a-million binding event is sufficient to sustain life. Alternatively, our understanding of target searches might be incomplete, and the cell may have found other ways to optimize target searches.

In the following three chapters, we will examine three cases where facilitated diffusion fails to explain how proteins locate and recognize targets in DNA. In Chapter 2, we will consider *E. coli* RNAP's search for promoter sequences. Chapter 3 will look at two pylogenetically distinct CRISPR immune systems, and how they manage the task of viral DNA recognition. Finally, Chapter 4 examines the mechanism behind target searches during homologous recombination. Surprisingly, facilitated diffusion is rebuffed in each of these cases, yet together they reveal a novel mechanism by which cells stack the odds of random collisions in its favor.

# Chapter 2

# Transcription initiation in *Escherichia coli*

---

This work was was originally published as: "The promoter search mechanism of *E. coli* RNA polymerase is dominated by three-dimensional diffusion", Feng Wang*, Sy Redding*, Ilya J. Finkelstein, Jason Gorman, David R. Reichman, and Eric C. Greene, Nat Struct Mol Biol. 2013 Feb; 20(2): 174181.

Author contributions: F.W. collected the RNAP experimental data, and F.W. and S.R analyzed the data. S.R. developed the theoretical analysis, conducted theoretical calculations, assisted with the RNAP data collection, and collected the data for lac repressor. I.J.F. assisted in establishing the single-molecule assays for QD-RNAP and QD-lac repressor. J.G. developed the substrate for the dig-QD measurements and collected the corresponding data. E.C.G. supervised the project and all authors co-wrote the paper.

---

## 2.1 Introduction

Transcription is the process of transferring the information encoded in DNA to RNA, and, to differing degrees, regulates gene expression across all kingdoms of life. The protein machinery responsible for transcription in *E. coli* is RNA polymerase (RNAP) [Haugen *et al.*, 2008; Browning and Busby, 2004; Saecker *et al.*, 2011; Nudler, 2009; Mendoza-Vargas *et al.*, 2009; Cho *et al.*, 2009]. Transcription can be broken up into three processes: initiation, locating the beginning of genes and

preparing to make RNA; elongation, where RNAP is actively generating RNA; and termination, or ending transcription [Haugen *et al.*, 2008; Browning and Busby, 2004; Saecker *et al.*, 2011; Nudler, 2009]. The most common ways cells regulate transcription is by (*i*) affecting how fast RNAP gets to a promoter, (*ii*) how well RNAP recognizes the promoter once it gets there, (*iii*) the ability of the polymerase to start writing RNA, or (*iv*) how well it continues writing RNA once it gets going [Haugen *et al.*, 2008; Browning and Busby, 2004; Saecker *et al.*, 2011; Nudler, 2009]. This chapter focuses on understanding how RNAP finds promoters in an effort to understand the role the target search for promoters plays in the regulation of genes.

In Chapter 1, it was discussed that, under certain conditions, the *lac* repressor can bind to its target faster than expected from pure 3D diffusion, and these accelerated association rates can be explained through a combination of 3D collisions and lower dimensional interactions with the DNA (Fig. A.1) [Riggs *et al.*, 1970]. This result, coupled with the unfavorable probability of locating a target by random chance, is often used to argue that facilitated diffusion contributes to all target searches [Halford and Marko, 2004; Halford, 2009]. There is, however, little evidence to support this generalization, and it is unclear whether the *lac* repressor is an appropriate model for the typical behavior of a DNA-binding protein and/or for target-search mechanisms.

Yet, recent studies seem to have confirmed this presumption and reported RNAP moving long distances along DNA by one-dimensional (1D) sliding [Singer and Wu, 1987; Ricchetti *et al.*, 1988; Kabata *et al.*, 1993; Guthold *et al.*, 1999; Harada *et al.*, 1999], and now, it is widely assumed that RNAP locates promoters through a process involving a 1D search [J. *et al.*, 2007]. Despite this, no promoter-association rate exceeding the 3D-diffusion limit has ever been reported[Roe *et al.*, 1984; Friedman and Gelles, 2012], and the potential contribution of facilitated diffusion to the promoter search process has been challenged in the literature [deHaseth *et al.*, 1998].

To help resolve the mechanism of the promoter search, we directly watched single molecules of *E. coli* RNAP as they searched for native promoters within the viral genome of bacteriophage $\lambda$. These observations revealed several transcriptional intermediates: nonspecific binding of RNAP to DNA, promoter-bound RNAP in the closed complex and open complex conformation, and actively transcribing RNAP. In an effort to interpret these data, we developed a theoretical framework to assess differing contributions to the overall association rate, which in turn dictated a redesign of our DNA curtain assay. Results from these experiments argue that facilitated diffusion likely does not

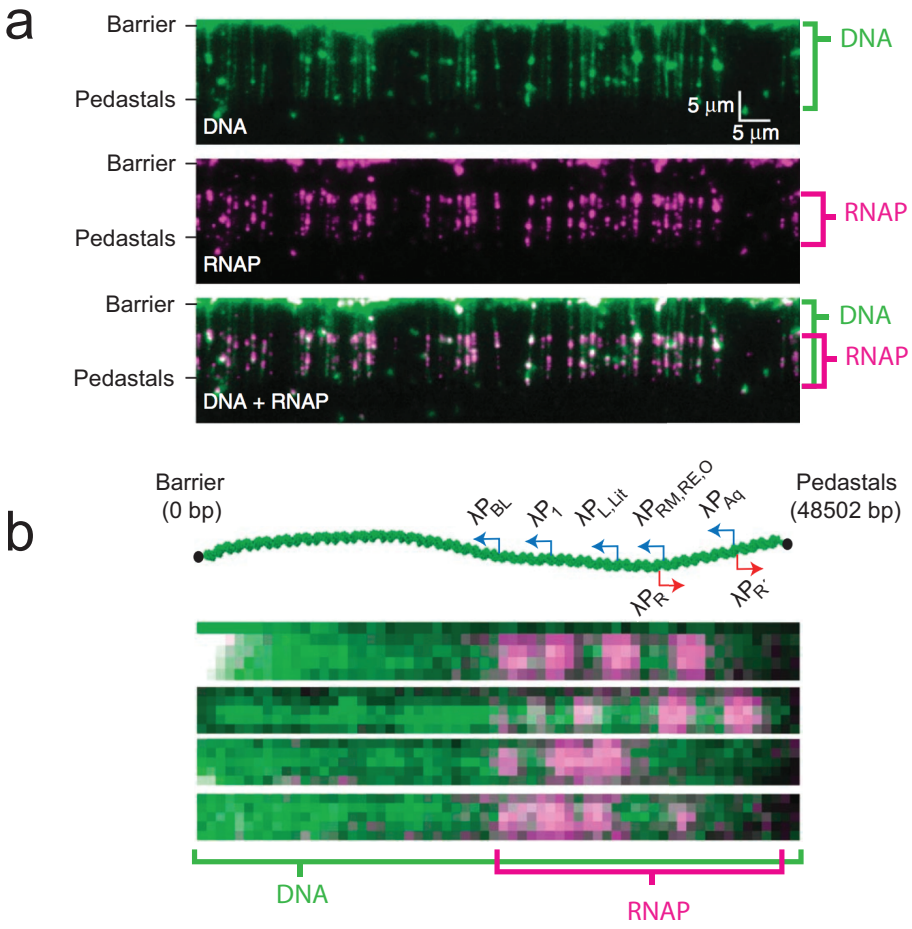


Figure 2.1: Single-molecule DNA-curtain assay for promoter-specific binding by RNA polymerase. (a) Two-color images of YOYO-1stained DNA (green) bound by QD-RNAP (magenta). (b) Schematic of the phage genome (48.5 kb), including relative locations and orientations of promoters aligned with images of QD-RNAP on single DNA molecules. Most RNAP is shown bound to the promoters, and the left half of the DNA that lacks promoters is essentially devoid of bound proteins.

contribute to the promoter search of RNAP at physiologically relevant protein concentrations, and highlight how protein concentration trumps the potential rate-accelerating benefits of facilitated diffusion. Finally, from this interpretation, we are able to determine the size of the targets that RNAP searches for, that is, we measure how big a promoter looks from the perspective RNAP.

## 2.2 Visualizing the promoter search by *E. coli* RNAP on DNA curtains

*E. coli* RNAP is among the best-characterized enzymes at the single molecule level, yet no study has conclusively established how RNAP locates promoters [Herbert *et al.*, 2008]. To distinguish among potential search mechanisms (Fig. 1.3, Fig. A.1) we used DNA curtains to visualize quantum dot-tagged RNAP (QD-RNAP) as they located and bound to native promoters in the genome of $\lambda$ phage (Fig. 2.1a,b). In earlier work from our laboratory, it was demonstrated that RNAP could bind promoters within the context of a DNA curtain, and that this binding was dependent on the full RNAP holoenzyeme [Finkelstein *et al.*, 2010]. Further, it was established that the QD-RNAP were active for transcription [Finkelstein *et al.*, 2010].

### 2.2.1 Promoter-association assays reveal known intermediates

Direct visualization of individual molecules of RNAP as they actively searched for promoters, revealed four potential binding intermediates, for brevity referred to as $\tau_0$, $\tau_1$, $\tau_2$, and $\tau_3$ events (Fig. 2.2a). $\tau_0$ events were short-lived ($\tau_0 = 5.5ms$) and were observed with either QDs alone or in the absence of DNA. We ascribed these events to random diffusion through the detection volume and they were not considered further. $\tau_1$ events were RNAP and DNA dependent, displayed short lifetimes ($\tau_1 = 30ms$), and occurred randomly along the DNA (Fig 2.2c). RNAP binding events corresponding to $\tau_2$ dissociated more slowly from the DNA ($\tau_2 = 3.5s$) and were strongly correlated with the location of known promoter sites in lambda DNA (Fig. 2.2c). Finally, $\tau_3$ binding events exhibited even slower dissociation ($\tau_3 \approx 6 \cdot 10^3 s$), coincided with known promoters (Fig. 2.2c), were resistant to challenge with heparin (a hallmark of open-complex formation) and could initiate transcription (Fig. 2.2b). These results are consistent with a reaction scheme where $\tau_1$ corresponds to nonspecifically bound RNAP, $\tau_2$ to closed complexes and $\tau_3$ to open complexes (Fig. 2.2d).

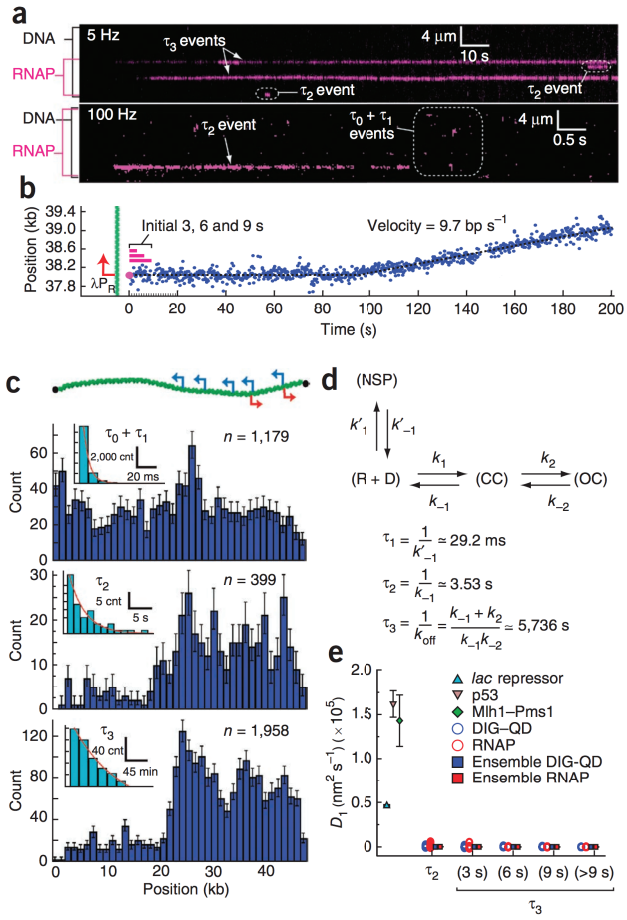


Figure 2.2: Visualizing single molecules of RNA polymerase as they search for and engage promoters. (a) Kymograms of RNAP binding to λ DNA, showing kinetically distinct intermediates. DNA is unlabeled, and RNAP is magenta. (b) Representative example of RNAP binding and initiating transcription from the $\lambda_{PR}$ promoter; for this assay, RNAP was premixed with all four NTPs immediately before injection into the sample chamber. Initial binding ($t = 0s$) is indicated as a magenta dot, and magenta bars highlight the first $39s$ of the reaction trajectory. (c) Binding distributions of kinetically distinct intermediates and corresponding lifetime measurements. A schematic showing the relative promoter location is included. Error bars indicate 70% confidence intervals obtained through bootstrap analysis. (d) Kinetic scheme reflecting observed intermediates. NSP, CC and OC refer to nonspecifically bound, closed complex and open complex, respectively; CC could also represent another intermediate preceding the open complex [Saecker *et al.*, 2011]. Kinetic parameters are not segregated for individual promoters; rather, they are considered collectively, and therefore reported values should be considered an average of all λ promoters. (e) Upper bound of observed diffusion coefficients for promoter-bound RNAP compared to immobilized DIG-QDs and other proteins known to undergo 1D diffusion [Elf *et al.*, 2007; Tafvizi *et al.*, 2011; Gorman *et al.*, 2010]. Diffusion coefficients are gamma distributed; therefore, we report the magnitude of the square root of the variance as error bars ($n \geq 50$ for all data sets)

Recently, the nonspecific lifetime of RNAP, $\tau_1$, was inferred from an *in vivo* study and was found to be identical to our determination of $\tau_1$ [Bakshi *et al.*, 2013]. Furthermore, the value we obtained for $\tau_3$ (that is, $k_{off}^{-1}$) is consistent with bulk biochemical data for the lifetime of the open complex [Dayton *et al.*, 1984; Brunner and Bujard, 1987; Hawley and McClure, 1980]. There are no reported measurements of $\tau_2$ in the literature, however our measurement of the ratio $\tau_2/\tau_3$ is in good agreement with bulk measurement of closed complex escape [Dayton *et al.*, 1984; Hawley and McClure, 1980]. But perhaps most illuminating, the constant $K_1 = k_1/k_{-1} = k_1\tau_2$, which is the equilibrium between RNAP in solution and the earliest stages of promoter capture, has been measured extensively *in vitro* [Dayton *et al.*, 1984; Brunner and Bujard, 1987; Hawley and McClure, 1980; Simons *et al.*, 1983; McClure, 1980]. Our measured value for $\tau_2$ is consistent with these measurements of $K_1$ if we assume that $k_1$ is equal to the 3D diffusion limited rate [von Hippel and Berg, 1989]. To be clear, *only* if we assume that there are no facilitating effects in RNAP's search for its promoters, do our direct measurements match independent biochemical studies of RNAP kinetics. We conclude that the intermediates observed in our assay reflect properties consistent with the literature and that the DNA-curtain assay can be used to probe the early stages of RNAP association that precede transcription.

### 2.2.2 No microscopically detectable 1D diffusion before promoter binding

Our results have demonstrated that QD-tagged RNAP is targeted to promoters in the DNA-curtain assay and that the experimental observables obtained from these assays recapitulated known reaction schemes and kinetic parameters for promoter association and dissociation. Unexpectedly, real-time observations of RNAP revealed no evidence for microscopic 1D diffusion by RNAP (described below). To ensure our experiment was not merely sampling buffer conditions biased against 1D mechanisms, we conducted our experiments over a range of ionic strengths (0 - 200 mM KCl, 0 - 10 mM $MgCl_2$), including all buffer conditions under which RNAP sliding had been previously reported. Under no conditions did we find evidence of microscopically detectable 1D diffusion of RNAP along the $\lambda$ DNA before promoter engagement.

This absence of 1D diffusion by RNAP is not because the DNA curtain assay somehow prevents the ability of DNA binding proteins to track along the DNA backbone. Recall the kymograms in figure 1.5, which show both the *lac* repressor and Msh2-6, searching for, and finding, their targets

via 1D diffusion. Furthermore, control experiments with RNAP from T7 phage were capable of extensive 1D diffusion in our assays [Wang *et al.*, 2013a]. Finally, we readily observed 1D movement for large (1.0 $\mu m$) beads coated with RNAP, suggesting that multivalent aggregates of RNAP may have confounded previous measurements [Wang *et al.*, 2013a].

In summary, we found no direct experimental evidence supporting an extensive contribution of 1D diffusion during the promoter search, which suggests that the promoter search by QD-tagged RNAP within the context of our DNA-curtain assays was dominated by 3D diffusion. As a further test of this interpretation, we next sought to determine the upper bounds for the observed 1D diffusion coefficients ($D_{1,obs}$) for QD-RNAP for the different reaction species. Unfortunately, given their transient nature ($\tau_1 \leq 30ms$), we could not determine $D_{1,obs}$ values for the nonspecifically bound RNAP (described below); however, we did calculate $D_{1,obs}$ for intermediates categorized as either closed ($\tau_2$) or open complexes ($\tau_3$ events) as well as for the first 3 to 9s after initial DNA binding for molecules of RNAP that subsequently initiated transcription in the presence of rNTPs (Fig. 2.2b,e). We also collected data sets for QDs covalently linked to the DNA through an internal DIG tag (DIG-QD). These data were collected identically to experiments concerning QD-RNAP (e.g. matching frame rate, truncated data size, etc.) to provide an indication of the extent to which DNA fluctuations contribute to measurements of $D_{1,obs}$. We then compared the resulting $D_{1,obs}$ values for DIG-QD and QD-RNAP to published values for several well characterized proteins known to undergo 1D diffusion (Fig. 2.2e). The $D_{1,obs}$ values for RNAP were all several orders of magnitude lower than values reported for the *lac* repressor [Elf *et al.*, 2007], p53 [Tafvizi *et al.*, 2011] and Mlh1Pms1 [Gorman *et al.*, 2010], which further argued against extensive 1D diffusion contributing to the promoter search. More convincing, the $D_{1,obs}$ values for RNAP ($\sim$ 15-100 $nm^2s^1$) were indistinguishable from values obtained for stationary DIG-QDs (Fig. 2.2e). It is important to recognize that the small $D_{1,obs}$ values obtained for RNAP cannot be interpreted as protein movement along the DNA but rather arise from the underlying diffusive fluctuations of the DNA itself. From these data, we concluded that promoter binding by *E. coli* RNAP is not preceded by microscopically detectable 1D diffusion. This conclusion leads to two possible interpretations: either promoter binding by RNAP is not preceded by 1D diffusion on any scale, or it occurs, but is faster than our image acquisition rate and/or over smaller distances than our spatial resolution.

## 2.3 What are we missing?

Concerned about what activities of the polymerase we might be missing out on in our experiments, we developed both a theoretical framework and a new DNA curtain experiment to allow access to information below the resolution of our microscope. Further details of the theory are presented in Appendix A; for brevity, we highlight key features and results. We began by recognizing that the flux of RNAP onto promoters is the result of two components: (i) direct binding to promoters from a fully equilibrated solution (that is, 3D diffusion) and (ii) rapid promoter binding from solution after dissociation from another region of DNA or after undergoing 1D diffusion along the DNA (that is, through facilitating mechanisms). The most important of these terms with respect to the promoter search by RNAP was direct binding from solution, which occurs at a rate of:

$$k_{\alpha}^{\psi}(t) = \frac{8}{\pi} D_3 C_0 \psi \int_0^{\infty} e^{-D_3 u^2 t} \left[ u \left( J_0^2(u\rho) + Y_0^2(u\rho) \right) \right]^{-1} du \tag{2.1}$$

, where $C_0$ is initial protein concentration, $D_3$ is the 3D-diffusion coefficient of QD-RNAP, $\psi$ is the effective target size, $\rho$ is the reaction radius, and $J_0$ and $Y_0$ are Bessel functions of the first and second kind, respectively.

The effective target size, $\psi$, is discussed further in Appendix A, but it is sufficient to consider the two terms, $\psi$ and $\rho$, as the height and radius of a cylinder which encapsulates the DNA that is sampled by RNAP's binding surface during a single encounter with the DNA. The target size should not be confused with promoter length; rather, it describes the range over which a bound protein can be out of register yet still recognize its target (Fig. 2.4a and Appendix A).

An important prediction arising from this formalism is that target-association rates for any protein become dominated by $k_{\alpha}^{\psi}$ as $C_0$ increases, implying that increased protein abundance can obviate any potentially accelerating contributions from facilitated diffusion, regardless of whether the protein in question is capable of hopping and/or sliding along DNA. In simple terms, facilitated diffusion to the target requires the protein to pass through non-promoter bound intermediates, which at low concentrations helps localize the protein and funnel it toward its target. But, the protein must spend some amount of time interrogating the DNA while engaged at non-specific sites in order to identify them as such. It is this wasted time that gives rise to the concentration dependence of facilitating mechanisms, because during the unproductive interrogation of non-target

sites, unbound proteins can proceed unimpeded to the target, provided a large enough pool of free protein exists. Notably, the physical behavior of individual proteins with respect to the search process is not changed, regardless of whether the concentration is high or low; the only thing that changes is the probability that the target is first locates through a direct collision (P3D) versus the probability of first engaging the target after undergoing facilitated diffusion along the DNA (PFD).

The question then becomes: at what concentration does 3D diffusion begin to dominate the search? This threshold depends on the strength of the facilitated pathway,which is determined by $D_1$ and the non-specific lifetime, $k'_{-1}$. We will refer to the concentration at which 3D target binding becomes favored as the facilitation threshold ($C_{thr}$): 3D target binding will be favored when the protein concentration equals or exceeds $C_{thr}$, whereas facilitated diffusion will be favored when the concentration is below $C_{thr}$. In addition, once effects on the association rate from facilitated diffusion are removed through increased protein abundance, $k_\alpha^\psi$ can be used to recover the effective target size, $\psi$ (Appendix A).

These parameters reflect dynamic physical properties of highly transient encounter complexes and cannot be accessed through traditional biochemical analysis of stable or metastable reaction intermediates (closed complexes, open complexes and so forth), nor can they be revealed through structural studies of static proteinnucleic acid complexes. To our knowledge, neither, $\psi$ nor $C_{thr}$ have been experimentally determined for any proteinnucleic acid interaction.

## 2.4 Single-molecule promoter-search kinetics

By definition, it is not possible to directly visualize submicroscopic events that contribute to target searches. However, we can obtain promoter-association rates from real-time single-molecule measurements (described below). These experimental values can then be compared to theoretical calculations, allowing us to extract critical features of search mechanisms that are otherwise obscured at the microscopic scale. This provides an independent assessment of the search process that is unhindered by existing spatial or temporal instrument resolution limits. Simply, we calculate $k_\alpha^\psi$ and measure the association rate, $k_a$; if the experimentally observed association rates exceed $k_\alpha^\psi$, then submicroscopic facilitated diffusion must be contributing to the search mechanism. In contrast, if the experimentally observed association rates are equal to $k_\alpha^\psi$, then the search mechanism

can be attributed to 3D collisions with no underlying contribution of submicroscopic facilitated diffusion (Fig 2.4).

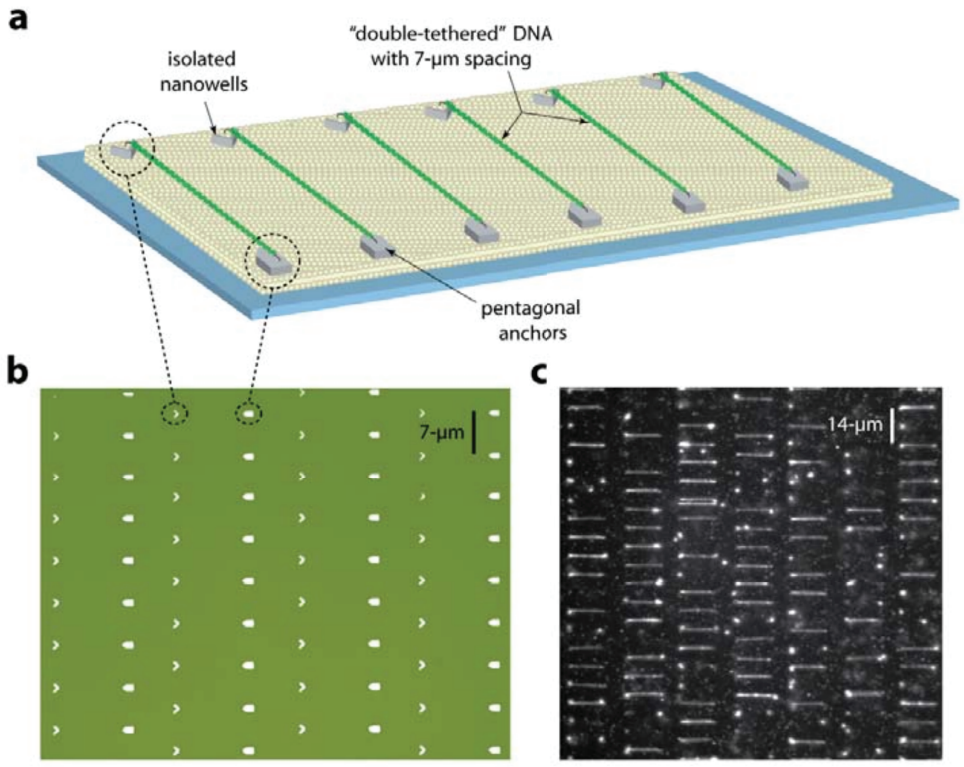


Figure 2.3: Parallel Array of Double-tethered Isolated (PARDI) Molecules. (a.) Schematic diagram of the new PARDI DNA curtain design used for the promoter association rate measurements. (b.) Optical image highlighting nanofabricated PARDI pattern design. (c.) Image of a typical PARDI field-of-view, showing the double-tethered, YOYO1-stained DNA molecules.

To avoid over interpreting our data or arbitrarily assigning variables (i.e. $C_{thr}, \psi, D_1$ etc.), we measured promoter association rates over a range of RNAP concentrations and used the above-mentioned trend in the association rate to evaluate RNAP's target search dynamics. Notably, with this assay, we did not measure closed- or open-complex formation; rather, we measured the instantaneous time at which single molecules of RNAP were initially detected at a promoter, conditioned upon their subsequent conversion to closed and then open complexes. These measurements were conducted at 100 $ms$ temporal resolution, which is appropriate, given that the slow downstream isomerization steps involved in promoter binding (for example, closed- and open-complex formation) occur on the order of seconds or minutes.

This assay utilized a new DNA curtain design (Fig 2.3), which was necessary to achieve two key

effects. First, it allowed for normalization of the DNA concentration both within an experiment and between multiple experiments. Association rates are bimolecular, and therefore the DNA concentration contributes to the association rate. By normalizing the DNA, we are able to determine $k_a$ as a pseudo first order rate. Second, this experimental format allowed us to precisely define all of the boundary conditions and parameters involved in calculating the predicted promoter association rate, $k_\alpha^\psi$ (for example, DNA geometry, DNA length, DNA density, number of accessible promoters, protein concentration, solution viscosity, temperature, ionic strength and so forth).

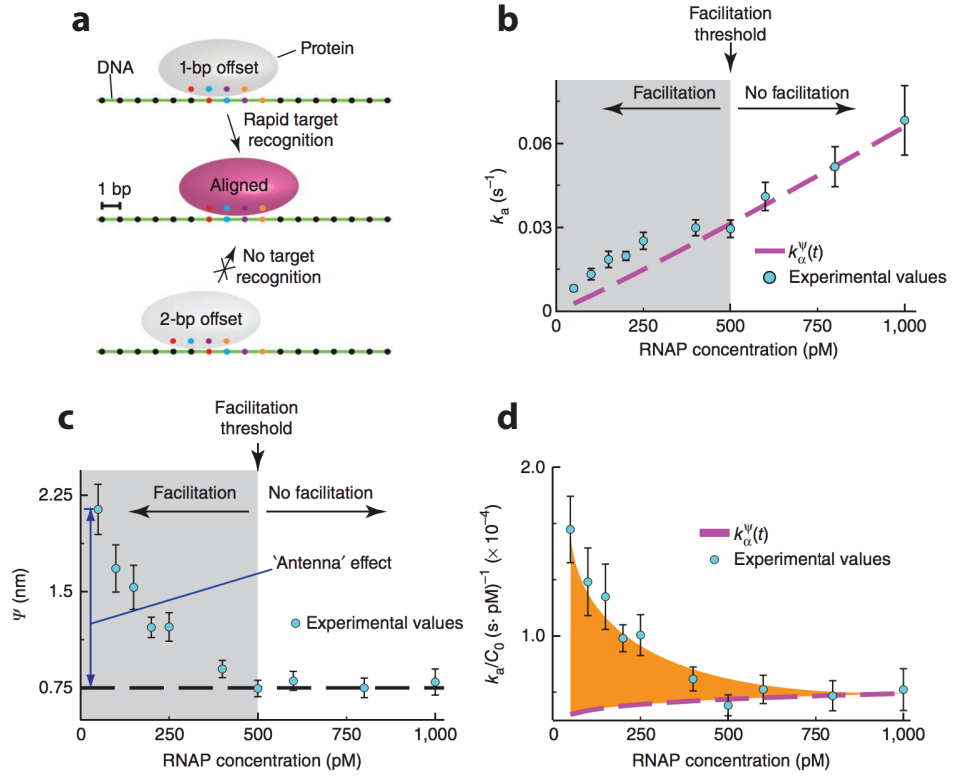


Figure 2.4: Single-molecule kinetics reveal that the promoter search is dominated by 3D diffusion. (a) Illustration of linear target size (a). For example, where $\psi = 2bp$, a 1-bp offset (in either direction) results in target recognition, but a 2-bp offset does not result in target recognition. (c) Observed promoter association rates ($k_a$). Dashed magenta line corresponds to $k_\alpha^\psi$ in the absence of facilitated diffusion (for $\psi = 0.75nm$), and experimental values above this line reflect rate enhancement due to facilitated diffusion. The boundary between the shaded and unshaded regions of the graph represents the facilitation threshold ($C_{thr}$; as indicated). (e) Effective target size ($\psi$) versus RNAP concentration. The dashed black line highlights the limiting value of $\psi$. (f) Rate acceleration ($k_a/C_o$) versus RNAP concentration. The difference between the experimental values and $k_\alpha^\psi$ reflects facilitated diffusion, and the orange shaded region represents the maximum possible acceleration due to 1D sliding and/or hopping.

Notably, the association rate of RNAP to promoters exceeded $k_\alpha^\psi$ below 500 $pM$ QD-RNAP, revealing that submicroscopic facilitated diffusion accelerated the promoter search by a factor of three at 25 $pM$ RNAP (Fig. 2.4b,d). However, at $\geq$ 500 $pM$ RNAP, $k_a$ converged to $k_\alpha^\psi$, which indicated that submicroscopic facilitated diffusion was short-circuited at higher concentrations, as expected (Fig. 2.4b,d). Although our results showed that QD-RNAP no longer benefits from facilitated diffusion at concentrations $\geq$ 500 $pM$, it must be recognized that $C_{thr}$ will vary for different proteins and/or different reaction conditions. For example, unlabeled RNAP (hydrodynamic radius $r = 7.4\ nm$) diffuses more rapidly through solution than QD-RNAP ($r \approx 13.4\ nm$), so we anticipate that promoter association with unlabeled proteins should converge to $k_\alpha^\psi$ at an even lower protein concentration, which would be reflected as a reduction in $C_{thr}$.

Furthermore, once the measured association rate collapsed to $k_\alpha^\psi$, we were able to determine the effective target size $\psi$ as $\sim$ 0.75 $nm$, corresponding to $\sim$ 3 $bp$ (Fig. 2.4c), which indicated that promoters would not be recognized if RNAP is more than $\pm$1.5 $bp$ out of register (Fig. 2.4a, Appendix A). The apparent increase in $\psi$ at low RNAP concentration reflected what is historically referred to as the antenna effect. At 25 $pM$ RNAP ($\psi = 2.23\ nm$), the antenna was just $\sim$ 1.48 $nm$ (corresponding to $\sim$ 6 $bp$ in our system); the very small size of the antenna indicated the limited contribution that facilitated diffusion (sliding and/or hopping) made to the promoter search even at the lowest RNAP concentrations tested (Fig. 2.4c,d).

An in vivo protein concentration of 1 $nM$ corresponds to just 1 protein molecule in a volume the size of an *E. coli* cell; therefore, an in vivo concentration of 50 $pM$ would be equivalent to an average of just $1/20^{\text{th}}$ of a molecule of RNAP per bacterium, which does not seem physiologically relevant. Taken together, our results demonstrate that although submicroscopic facilitated diffusion can moderately accelerate the promoter search, this acceleration only occurs at exceedingly low RNAP concentrations, whereas at physiologically relevant protein concentrations, the overall promoter search process should be dominated by 3D diffusion.

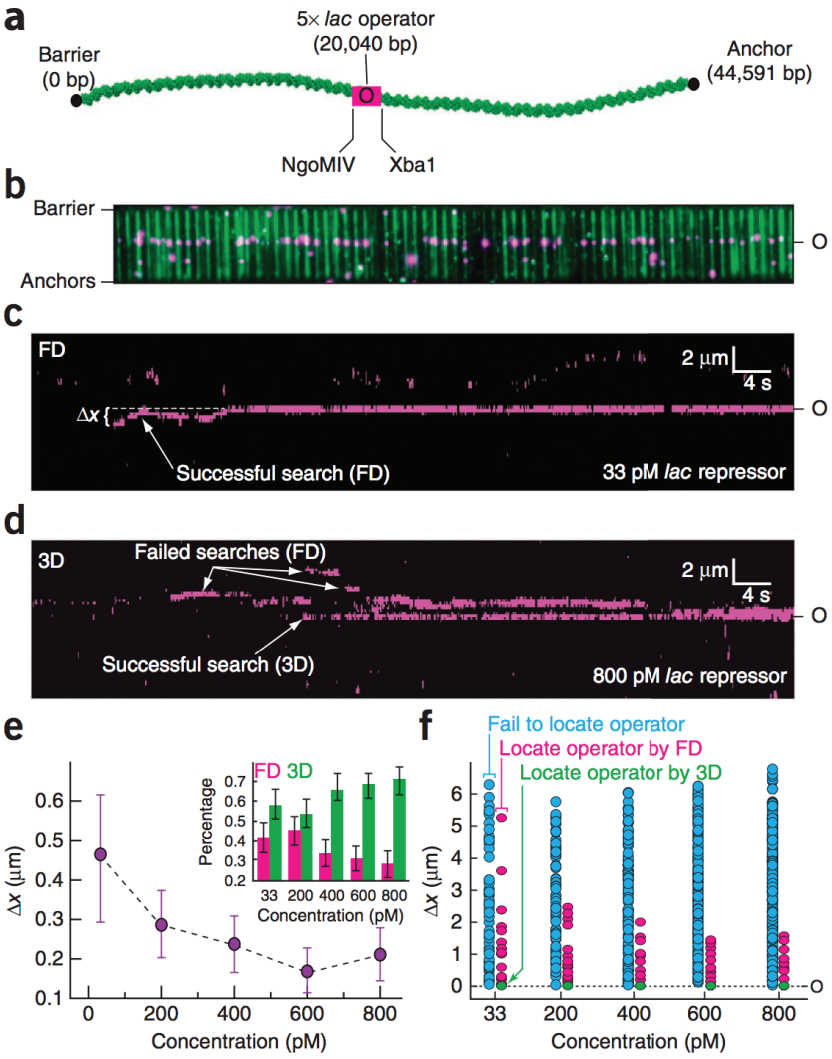


Figure 2.5: Protein concentration exerts a dominant influence on target searches, even for proteins capable of sliding on DNA. (a) DNA schematic showing the location of the 5 *lac* operator (O). (b) Two-color image of YOYO-1stained DNA (green) bound by QD-lac repressor (magenta). (c) Kymogram showing an example of lac repressor binding to nonspecific DNA and then diffusing in 1D to the operator; data were collected at $33pM$ *lac* repressor. The distance between the initial binding site and the operator is indicated as $\Delta x$. (d) Kymogram showing an example of direct operator binding in the absence of any detectable 1D sliding; data were collected at $800pM$ *lac* repressor. The successful search through 3D binding is highlighted, as are examples of molecules that searched through facilitated diffusion (FD) but failed to locate the operator. (e) Graph showing the mean value of $\Delta x$ as a function of protein concentration for proteins that successfully engage the operator. Inset, percentage of total operator-binding events that are attributable to facilitated diffusion (magenta) and 3D (green) at each protein concentration. Error bars, s.d. ($n \geq 54$ for each data point). (f) Graph of $\Delta x$ for all observed proteins. Blue data points correspond to proteins that fail to bind the operator, magenta data points are proteins that bind the operator after undergoing facilitated diffusion, and green data points correspond to 3D binding to the operator. All green data points within each column overlap at zero, but their fractional contribution to operator binding is shown as green bars in the inset of panel (e).

## 2.5 Increased protein abundance disfavors facilitated searches

One conclusion arising from our treatment of target searches is that increased protein abundance will diminish the contribution of facilitated diffusion. This concept is not unique to *E. coli* RNAP and will apply even to proteins that can diffuse long distances along DNA, because the probability of direct collisions (P3D) with the target always increases with increasing protein abundance and will eventual exceed the probability of target engagement through facilitated diffusion (PFD). As a simple illustration of this point, we used the DNA-curtain assay and a $\lambda$ DNA bearing 5-tandem 21-bp ideal *lac* operators [Finkelstein *et al.*, 2010] to qualitatively assess target binding by QD-tagged *lac* repressor (Fig. 2.5). These experiments were intentionally conducted at low ionic strength, such that nonspecific binding and 1D diffusion were greatly favored.

At low concentrations, many proteins initially bound to random, nonspecific sites and then diffused thousands of base pairs along the DNA before eventually binding the target; these events were categorized as having occurred through facilitated diffusion (Fig. 2.5c). Operator binding in the absence of microscopically detectable 1D diffusion was also observed; these events were categorized as 3D (Fig. 2.5d). For the proteins that successfully engaged the operator, the contribution of facilitated diffusion to the search process is reflected in the distance between the initial binding site and the operator ($\Delta x$) and the change in the ratio of facilitated diffusion to 3D events (Fig. 2.5e,f). As protein concentration increased, the mean value of $\Delta x$ decreased for the proteins that bound to the operator (Fig. 2.5e,f), and there was a corresponding increase in the fraction of events categorized as 3D (Fig. 2.5e, inset). At the highest concentration of *lac* repressor tested ($800pM$) $\sim 71\%$ of the total operator-binding events were attributed to 3D diffusion (Fig. 2.5e, inset). Technical limitations prevented titration to higher protein concentrations, but we anticipate that if the concentration were raised further, eventually all of the operator-binding events would occur through 3D diffusion. This conclusion will even extend into the submicroscopic regime as it does with RNAP. An in-depth analysis of the facilitation threshold and effective target size for the *lac* repressor (as provided above for RNAP) was beyond the scope of this work; however, the trend in these data clearly illustrate that the contribution of facilitated diffusion diminishes with increased protein abundance, even though the *lac* repressor is capable of sliding great distances on DNA under low-ionic-strength conditions.

Notably, at all concentrations tested, many molecules of *lac* repressor bound to random, non-

specific sites all along the length of the $\lambda$ DNA, and these proteins still exhibited 1D diffusion even when the concentration was raised (Fig. 2.5d,f); however, as concentration increased, this 1D diffusion should be considered nonproductive with respect to target association because most of the proteins that bound the operator first did so through 3D collisions (Fig. 2.5e, inset and Fig. 2.5d,f).

## 2.6 Discussion

Our results argue against facilitated diffusion at either the microscopic or submicroscopic scales being a significant contributing component of the *E. coli* RNAP promoter search. We also show that in general any potential contributions of facilitated diffusion can be overcome through increased protein abundance, even for proteins that can slide long distances on DNA. Facilitated diffusion and 3D collisions can be conceptually considered as two distinct competing pathways, either of which has the potential to result in target binding. 3D diffusion will always be favored at protein concentrations equal to or exceeding the facilitation threshold simply because the relative increase in protein abundance increases the probability of a direct collision with the target site (Fig. 2.6). In other words, just because a protein is physically capable of scanning long stretches of DNA during a non-specific binding event does not mean that these processes will accelerate target binding, because protein concentration can still dominate the overall search process.

A broader implication of this conclusion is that proteins present at low concentrations in living cells (for example, the *lac* repressor, with fewer than ten molecules per cell) may be more apt to locate targets through facilitated diffusion, whereas those present at higher concentrations (for example, RNAP, $\sim 3,000$ molecules per cell) may be more likely to engage their target sites through 3D diffusion.

### 2.6.1 Promoter searches in physiological settings

It is useful to consider how our promoter search results might translate in the promoter search by RNAP in a physiological setting. Our experimental setting differs substantially from much more complex physiological environments where the promoter search might be influenced by the presence of factors that can assist in the recruitment of RNAP to promoters, i.e. local DNA folding, higher

order chromatin architecture and macromolecular crowding (Fig. 2.6). Although we cannot yet quantitatively assess the influence of these parameters, we can consider how they might qualitatively affect the promoter search.

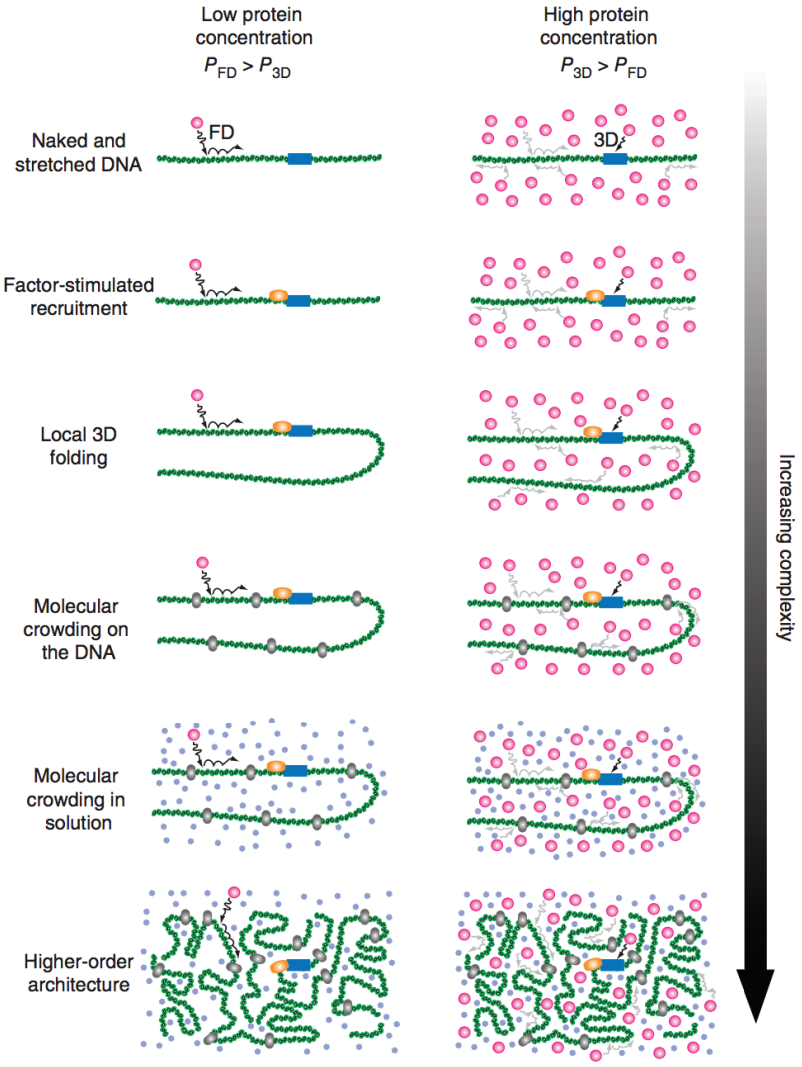


Figure 2.6: Increasingly complex environments encountered during in vivo searches. Facilitated diffusion (FD) will be favored at concentrations below the facilitation threshold because the initial encounter with the DNA will most often occur at nonspecific sites, so the probability (P) of target engagement through FD exceeds the probability of engagement through 3D (PFD>P3D). Concentrations equal to or exceeding the facilitation threshold will favor 3D because the relative increase in protein abundance increases the probability of a direct collision with the target site (PFD>P3D). FD-related processes such as sliding/hopping can still occur at high protein concentrations, but those proteins undergoing FD are less likely to reach the target site before those that collide directly with the target. Although the facilitation threshold will vary for different proteins and different conditions, higher protein concentrations will still favor 3D collisions irrespective of the local environment (e.g. the presence of recruitment factors, DNAbound obstacles, macromolecular crowding, local DNA folding) or global DNA architecture.

Transcriptional activators, such as catabolite activator protein (CAP), are commonly involved in the regulation of gene expression and can exert their effects either by facilitating recruitment of RNAP or by stimulating steps after recruitment (for example, open-complex formation, promoter escape and so forth) [Browning and Busby, 2004]. In scenarios involving factor-assisted recruitment, additional protein-protein contacts stabilize interactions between RNAP and the promoter. However, the presence of a transcriptional activator near a promoter should not fundamentally alter the search process by causing RNAP to start sliding and/or hopping along the DNA while executing its search; rather, it would just make the target appear larger to RNAP (that is, promoter plus factor, instead of just the promoter, resulting in an corresponding increase in $\psi$), which in turn reduces the facilitation threshold.

Factors that stimulate steps after recruitment would not influence the search because they exert their effects only after the promoter search is complete. Higher-order organization of DNA *in vivo* has the potential to promote 3D collisions and local concentration, but is not expected to favor 1D sliding and/or hopping, both of which can be considered as local events that are not influenced by global DNA architecture [Hu *et al.*, 2006]. In contrast, naked DNA stretched out at low dilution presents the most favorable possible conditions for 1D sliding and/or hopping [Halford, 2009; Bauer and Metzler, 2012]. The fact that we do not detect facilitated diffusion contributing to the promoter search by RNAP under conditions that should otherwise greatly favor hopping and/or sliding suggests these processes are unlikely to occur *in vivo*, simply owing to the more complex 3D DNA environment.

Molecular crowding, either in solution or on the DNA, is a nontrivial issue that can have both positive and negative impacts on DNA binding. Increased nonspecific binding can arise from macromolecular crowding in solution, owing to excluded volume effects [Minton, 2001], and any increase in nonspecific binding has the potential to promote facilitated diffusion. Although in the case of *E. coli* RNAP, increased nonspecific binding brought about through use of low-ionic-strength conditions still does not lead to microscopically detectable 1D diffusion, which suggests that any increased nonspecific affinity caused by excluded volume effects is unlikely to cause RNAP to start rapidly diffusing along DNA. The effects of macromolecular crowding on DNA arise from the presence of other nonspecific DNA-binding proteins, which can reduce nonspecific DNA-binding affinities through competitive inhibition [Graham *et al.*, 2011; Li *et al.*, 2009] and can also impede

1D diffusion along DNA through steric hindrance [Gorman *et al.*, 2010; Li *et al.*, 2009]. The net result of the seemingly opposed influences of macromolecular crowding in solution versus molecular crowding on the DNA has yet to be quantitatively explored, although one might anticipate that highly abundant proteins such as Fis and HU (each of which can be present at concentrations of up to $\sim$ 30-50 $\mu M$ in *E. coli*) would disfavor facilitated searches by restricting access to nonspecific sites [Li *et al.*, 2009].

In summary, there are at least four reasons why promoter searches in *E. coli* would not benefit from facilitated diffusion. First, the large in vivo concentration of RNAP 2-3$\mu M$ is much higher than $C_{thr}$ [Ishihama, 2000]. On the basis of our findings, if even a small fraction of the total RNAP present in a cell were free in solution, it is likely RNAP still locates promoters through 3D collisions rather than facilitated diffusion. Estimates have suggested that there are on the order of $\sim$ 550 molecules ($\sim 0.5\mu M$) of free $\sigma$70-containing RNAP holoenzyme in living bacteria [Ishihama, 2000]; if these estimates are correct, then the facilitation threshold would have to somehow increase by roughly three orders of magnitude in order for hopping and/or sliding to accelerate the promoter search *in vivo*. In contrast to RNAP, the *lac* repressor, which is thought to employ facilitated diffusion in vivo during its target search [Elf *et al.*, 2007; Hammar *et al.*, 2012], may need to do so to compensate for its much lower intracellular abundance (fewer than ten molecules per cell) and the corresponding scarcity of its targets (three lac operators per genome).

Second, long nonspecific lifetimes leads to slower searches due to time spent interrogating non-target DNA, and RNAP appears to be optimized to avoid wasting time by scanning nonspecific DNA [Halford, 2009; Berg and Blomberg, 1976; von Hippel and Berg, 1989]. Third, other proteins (for example, Fis, HU, IHF, HNS and so forth) may obstruct 1D diffusion, but such obstacles could be avoided through 3D searches [Li *et al.*, 2009]. Fourth, other steps are rate limiting during gene expression (for example, promoter accessibility, promoter escape, elongation and so forth) [deHaseth *et al.*, 1998; McClure, 1985; Reppas *et al.*, 2006; So *et al.*, 2011], which suggests that there is simply no pressure for RNAP to locate promoters faster than the 3D-diffusion limit. Finally, despite the much more complicated environments present in physiological settings, our general conclusion regarding the effects of protein abundance on target searches should remain qualitatively true because higher protein concentrations will increase the probability of direct target binding through 3D collisions.

For the chapters that follow, it is worth reiterating a few key features of RNAP's target search. Shown here, RNAP binds to DNA through series of short and weak interactions, comprising $\sim$ 3 $bp$ over $\sim$ 30 $ms$. During these interactions, if RNAP is able to sense promoter-like sequences, it can then prevent transition into longer-lived downstream states at obviously incorrect sites. Importantly, the target size measured here is likely too small to confer adequate information to completely specify genuine promoter sites *c.f.*[Cho *et al.*, 2009; Saecker *et al.*, 2011; Nudler, 2009] , but may bestow enough kinetic preference to drive a rapid 3D search toward promoters.

# Chapter 3

# Mechanisms of CRISPR interference

---

The Cas9 portion of the following work was was originally published as: "DNA interrogation by the CRISPR RNA-guided endonuclease Cas9", Samuel H. Sternberg*, Sy Redding*, Martin Jinek, Eric C. Greene, and Jennifer A. Doudna, Nature. 2014 Mar 6; 507(7490): 6267.

Author contributions: S.H.S. generated RNAs, conducted biochemical and single-molecule experiments, and assisted with single-molecule data analysis. S.R. conducted single-molecule experiments and data analysis, and assisted with the design and analysis of biochemical assays. M.J. cloned and purified 3x-FLAG-Cas9, and assisted with the design and interpretation of initial single-molecule experiments. S.H.S., S.R., M.J., E.C.G., and J.A.D. discussed the data and wrote the manuscript.

The Cascade portion has been recently submitted under the title: "DNA recognition and processing by the *E. coli* CRISPR/Cas system", Sy Redding, Samuel H. Sternberg, Myles Marshall, Bryan Gibb, Prashant Bhat, Chantal Guegler, Blake Wiedenheft, Jennifer A. Doudna, and Eric C. Greene.

Author contributions: S.R. conducted single-molecule experiments and data analysis and designed all biochemical assays. S.H.S. purified proteins and assisted with bulk biochemical experiments. M.M. and B.G. conducted bulk biochemical assays and purified proteins. C.G. P.B., and B.W. performed initial characterization of Cascade and Cas3. S.R., J.A.D., and E.C.G discussed the data and co-wrote the manuscript.

---

## 3.1 Introduction

Many bacteria and most archaea utilize an RNA-mediated adaptive immune system to provide protection from invading viruses and plasmids [Wiedenheft *et al.*, 2012]. This immunity relies on a DNA locus of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) and CRISPR-associated (Cas) proteins that function together to illicit an immune response (Fig 3.1). Bacteria harboring CRISPR-Cas loci respond to viral and plasmid challenge by integrating short fragments of foreign nucleic acids (protospacers) into the host chromosome bounded by short palindromic stretches of DNA (Fig. 3.1a). Transcription of the CRISPR array followed by enzymatic processing yields short CRISPR RNAs (crRNAs) that direct Cas protein-mediated cleavage of complementary target sequences within invading viral or plasmid DNA (Fig. 3.1b-f) [Brouns *et al.*, 2008; Garneau *et al.*, 2010; Barrangou *et al.*, 2007].

Close inspection of the CRISPR immune pathway exposes a serious problem. Foreign DNA is identified for destruction by directly base-pairing the crRNA to the DNA of the invader [Brouns *et al.*, 2008; Garneau *et al.*, 2010; Barrangou *et al.*, 2007]. Yet, this DNA sequence occurs in two places: once in the invading DNA and once again in the CRISPR locus, the site of crRNA synthesis (Fig 3.1a). So how is it that CRISPR-Cas complexes distinguish between DNA that belongs to the cell and foreign DNA?

The only indication of how this decision is made comes from inspecting the DNA context in which protospacers appear. Within the CRISPR locus, the protospacer is flanked on either side by short palindromic DNA repeats (Fig. 3.2a). Whereas in the foreign context, the protospacer lacks these repeats, but instead is abutted to a small conserved DNA motif, called a protospacer adjacent motif, or PAM (Fig. 3.2b). In the reference frame of the DNA, this is the only information available to CRISPR-Cas complexes for determining the difference between self and non-self.

CRISPR-Cas systems come in three types, Type I, II, and III, which can be further subdivided into subtypes [Wiedenheft *et al.*, 2012]. This chapter will focus on two of these systems: Type I-E, which is utilized in *Escherichia coli* [Brouns *et al.*, 2008], and Type II-A, the CRISPR classification of the system used in *Streptococcus pyogenes* [Jinek *et al.*, 2012]. In Type I-E, a ribonucleoprotein complex, Cascade, functions to detect the presence of invading DNA, using a single crRNA to recognize a specific 32 *bp* stretch of the invading DNA [Brouns *et al.*, 2008]. Once Cascade locates a target, it then recruits a trans-acting nuclease-helicase, Cas3, to digest the foreign DNA [Sinkunas

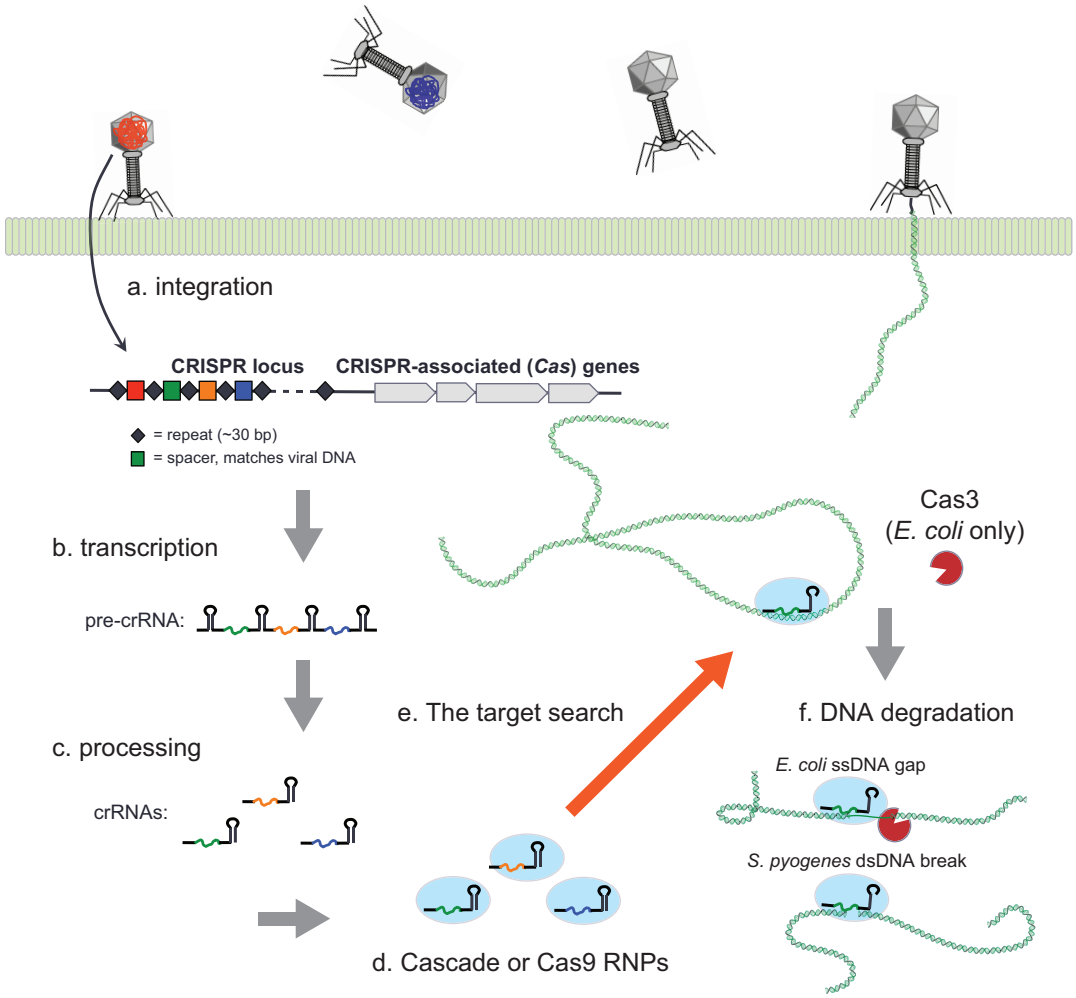


Figure 3.1: The CRISPR immune system. a, New protospacers are integrated into the CRISPR locus. b, Transcription of the CRISPR locus results in pre-crRNA that is then, c, processed into individual crRNAs. d, Cas complexes form around the crRNA, and use the RNA sequence to search for foriegn DNA, e. Once an invasive DNA sequence is found, f, Cas proteins then degrade the foreign DNA, thereby providing immunity.

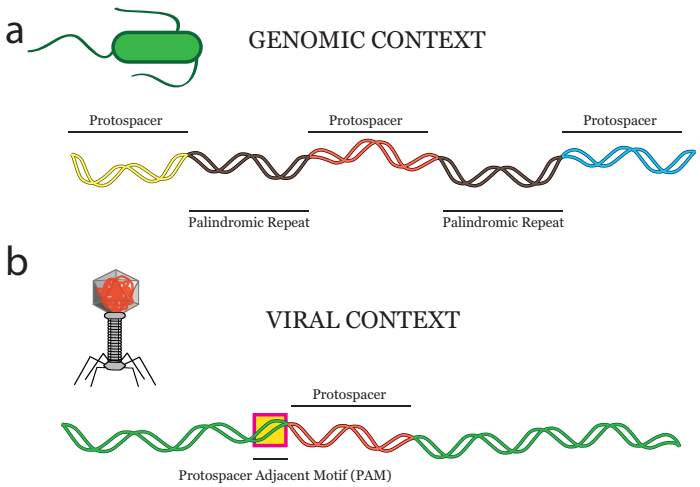


Figure 3.2: DNA context of protospacers. a, Within the host genome, protospacers occur in the CRISPR locus, where each protospacer is flanked by short palindromic DNA. b, Only in the virus, and adjacent to the protospacer, there is small DNA sequence which is highly conserved, called a PAM

*et al.*, 2011].

The Type II-A system employs a single protein, Cas9, as an RNA-guided endonuclease that uses a dual-guide RNA consisting of crRNA and a *trans*-activating crRNA (tracrRNA), which specify a 20 *bp* region of foreign DNA for target recognition and cleavage by a mechanism involving two nuclease active sites in Cas9 that together generate double-stranded DNA breaks (DSBs) [Jinek *et al.*, 2012].

RNA-programmed CRISPR-Cas complexes have proven to be a versatile tool for genome engineering in multiple cell types and organisms [Jinek *et al.*, 2013; Cong *et al.*, 2013; Mali *et al.*, 2013a; Hwang *et al.*, 2013; Wang *et al.*, 2013b; Bassett *et al.*, 2013; Gratz *et al.*, 2013]. Genomic engineering applications typically use Cas9, which makes site-specific DSBs that are repaired either by non-homologous end joining or homologous recombination, providing a facile means of modifying genomic information. In addition, catalytically inactive Cas9 or Cascade, alone or fused to transcriptional activator or repressor domains, have been used to alter transcription levels at sites targeted by guide RNAs [Qi *et al.*, 2013; Bikard *et al.*, 2013; Gilbert *et al.*, 2013; Maeder *et al.*, 2013; Perez-Pinera *et al.*, 2013; Mali *et al.*, 2013b]. Despite the remarkable ease in applying this technology, the fundamental mechanism that enables Cas9 or Cascade to locate specific targets within the vast sequence space of bacterial and eukaryotic genomes remains unknown.

## 3.2 Single-molecule visualization of CRISPR-Cas complexes

To determine how CRISPR-Cas complexes locate targets, we used our DNA curtain assay to visualize single Cas9 or Cascade molecules interacting with λ-DNA substrates (Fig. 3.3c,e). We purified both *S. pyogenes* Cas9 fused to C-terminal 3x-FLAG tag and *E. coli* Cascade, where the CasE subunit was fused to an N-terminal 3x-FLAG tag that enabled fluorescent labeling using anti-FLAG antibody-coated quantum dots (QDs) (Fig. 3.3a,b). We then generated guide RNAs (dual crRNA:tracrRNA) for Cas9 bearing complementarity to six different sites within λ-DNA and crRNAs for Cascade targeting the ribonucleotide-protein complex to three of the same locations (Fig. 3.3a,b). Control experiments confirmed that neither the 3x-FLAG tag nor QD inhibited target location for Cascade or Cas9 (not shown), and did not interfere with DNA cleavage by Cas9 [Sternberg *et al.*, 2014]. Additionally, we determined that all guide RNAs used in this study were functional [Sternberg *et al.*, 2014].

### 3.2.1 Programmed binding of Cas9 and Cascade

Initial experiments showing programmable Cas9 targeting were conducted with a nuclease-inactive version of Cas9 (D10A/H840A), dCas9 [Jinek *et al.*, 2012], to prevent cleavage of the DNA curtains. QD-tagged dCas9 and Cascade localized almost exclusively to the expected target sites in the DNA curtain assay (Fig 3.3c,e). Furthermore, both dCas9 and Cascade could be directed to any desired region of the phage DNA by redesigning the RNA guide sequence (Fig 3.3d,f). These results demonstrate that DNA targeting by Cas9 or Cascade is faithfully carried out in the DNA curtain assays.

*E. coli* Cascade requires crRNA to stably assemble, and the full complex is readily purified from its multiple subunits [Wiedenheft *et al.*, 2011]; therefore, all Cascade binding visualized in our assay is RNA mediated. Cas9, however, is a single protein, purified in the absence of RNA. Accordingly, we conducted controls with the apo- version of the protein to verify that the binding observed in DNA curtain assays was due to Cas9 loaded with RNA and not apo-Cas9 lacking guide RNA. Interestingly, apo-Cas9 also bound DNA but exhibited no sequence specificity (Fig. 3.4a,b) Attempts to measure the dissociation rate of DNA-bound apo-Cas9 were hampered by their exceedingly long lifetimes, placing a lower limit of at least 45 min on the actual lifetime (Fig.

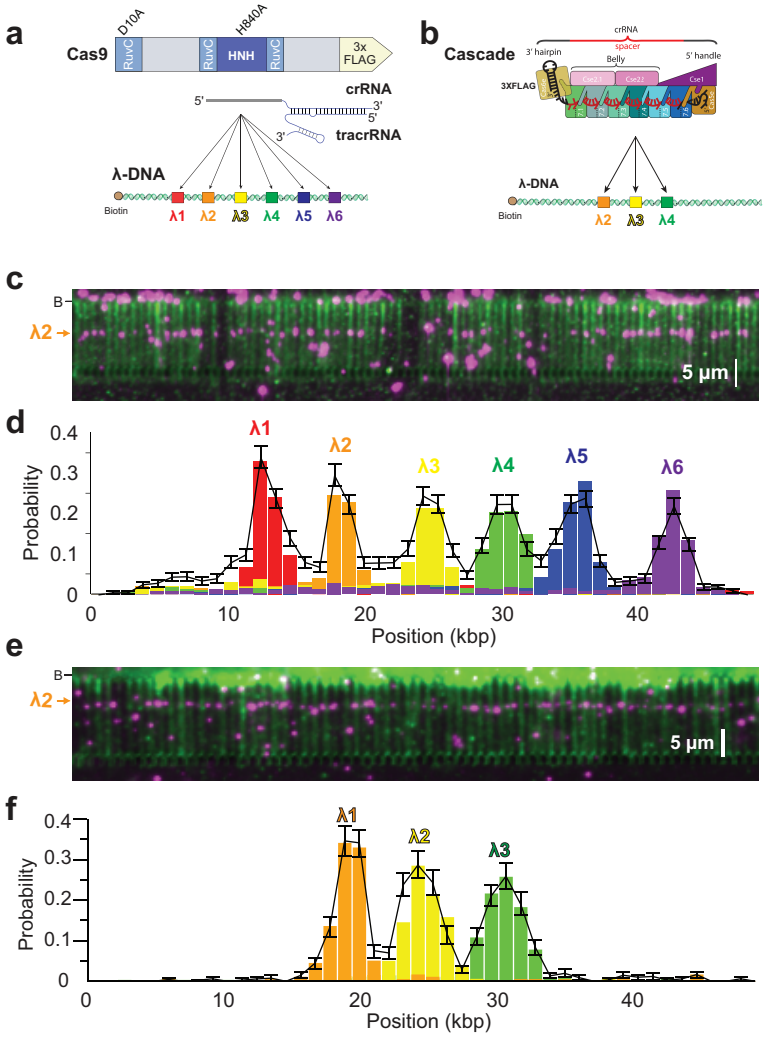


Figure 3.3: a, Wild-type Cas9 or dCas9 was programmed with crRNA:tracrRNA targeting one of six sites. b, Cascade was programmed with crRNA targeting three sites overlapping with Cas9 sites. c, YOYO1-stained DNA (green) bound by QD-tagged dCas9 (magenta) programmed with λ2 guide RNA. d, dCas9 binding distributions; error bars represent 95% confidence intervals obtained through bootstrap analysis. e, YOYO1-stained DNA (green) bound by QD-tagged Cascade (magenta) programmed with λ2 crRNA. f, Cascade binding distributions; error bars represent 95% confidence intervals obtained through bootstrap analysis.

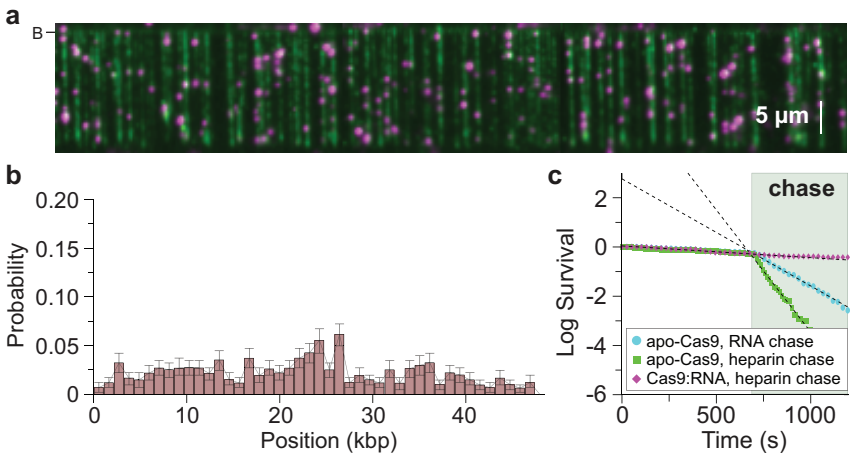


Figure 3.4: Apo-Cas9 binding activity. a, Image of apo-Cas9 bound to DNA curtains bound to apo-Cas9. b, Binding distribution of apo-Cas9; error bars represent 95% confidence intervals. c, Lifetimes of DNA-bound apo-Cas9 and Cas9:RNA after injection of λ2 crRNA:tracrRNA ($100nM$) or heparin ($10\mu gmL^{-1}$).

3.4c). Biochemical experiments revealed an upper limit of $\sim 25\ nM$ for the equilibrium dissociation constant ($K_d$) of this apo-Cas9:DNA complex, compared to $\sim 0.5\ nM$ for the Cas9:RNA complex bound to a *bona fide* target site.

We next asked whether DNA-bound apo-Cas9 could be distinguished from the Cas9:RNA complex based on a differential response to chases with free guide RNAs. To test this, we measured the lifetime of apo-Cas9 on DNA curtains before and after injection of crRNA:tracrRNA or heparin. Apo-Cas9 rapidly dissociated from the DNA in the presence of either competitor (Fig. 3.4c), and this result was verified with bulk biochemical assays [Sternberg *et al.*, 2014]. In contrast, target-bound Cas9:RNA was unaffected by heparin or excess crRNA:tracrRNA (Fig. 3.4c). These findings show that non-specifically bound apo-Cas9 has properties distinct from those of Cas9:RNA complexes bound to their cognate targets.

### 3.2.2 Catalytic activity of Cas9 is functional at the single molecule level

Initial experiments used catalytically inactive dCas9 to avoid DNA cleavage. Surprisingly, experiments performed with wild-type Cas9 also failed to reveal DNA cleavage. Rather, Cas9 molecules remained bound to their target sites; yielding identical results to those obtained using dCas9 (Fig. 3.5a). Controls confirmed that the imaging conditions did not inhibit Cas9 cleavage activity [Sternberg *et al.*, 2014]. These results suggested that Cas9 might cleave DNA but remain tightly bound to

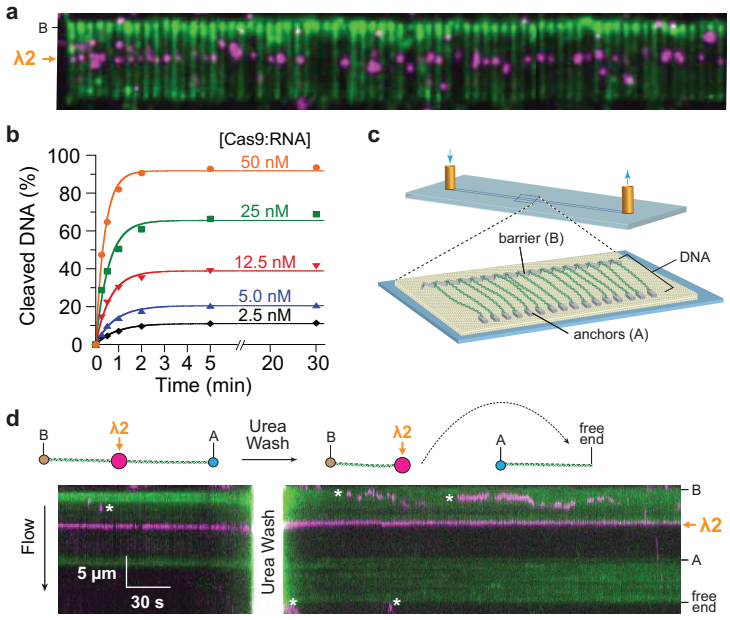


Figure 3.5: Cas9 remains bound to cleaved products. a, Wild-type Cas9:RNA bound to DNA curtains. b, Cleavage yield of 25$nM$ plasmid DNA is proportional to [Cas9:RNA]. c, Schematic of a double-tethered DNA curtain. d, Liberation of the cleaved DNA with 7 M urea; asterisks denote QDs that are attached to the lipid bilayer but not bound to the DNA.

both cleavage products, a hypothesis that was confirmed with biochemical gel shift assays [Sternberg *et al.*, 2014]. To determine whether stable product binding would prevent Cas9 from performing multiple turnover cleavage, we conducted plasmid DNA cleavage assays at varying molar ratios of Cas9 and target DNA, and measured the rate and yield of product formation (Fig. 3.5b). Surprisingly, the amount of product rapidly plateaued at a level proportional to the molar ratio of Cas9 to DNA, indicating that Cas9 does not follow Michaelis-Menten kinetics (Fig. 3.5b). Control experiments indicated that turnover also does not occur with short duplex DNA substrates and is not stimulated by either elevated temperature or an excess of free crRNA:tracrRNA [Sternberg *et al.*, 2014].

We next used double-tethered DNA curtains (Fig. 3.5a) to confirm that Cas9 catalyzed DNA cleavage in the single-molecule assays. Remarkably, when bound to target sites on λ-DNA, Cas9 failed to dissociate from the DNA even in the presence of heparin (10 $\mu g\ ml^{-1}$) (Fig. 3.4c) or up to 0.5$M$ NaCl. However, injection of 7$M$ urea caused Cas9 to release the DNA, and confirmed that the DNA was cleaved at the expected target site (Fig. 3.5d). These findings show that Cas9 remains tightly bound to both ends of the cleaved DNA, acting as a single-turnover enzyme.

## 3.3 Visualizing the target search of CRISPR-Cas complexes

### 3.3.1 Cas9 locates targets by 3D diffusion

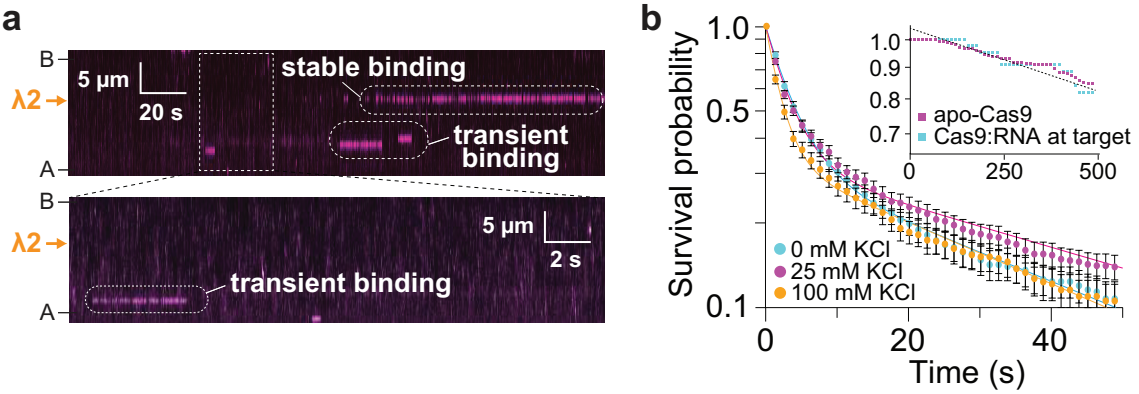


Figure 3.6: Experimentally observed Cas9 binding populations. a, Kymographs illustrating distinct binding events. b, Survival probabilities for non-target binding events; solid lines represent double-exponential fits. Inset: survival probabilities of DNA-bound apo-Cas9 and target DNA-bound Cas9:RNA.

To determine how Cas9 locates DNA targets, we visualized the target search process using double-tethered DNA curtains. For these assays, Cas9 programmed with $\lambda 2$ guide RNA was injected into the sample chamber, buffer flow was terminated, and reactions were visualized in real-time. These experiments revealed expected, long-lived, binding events at the target site and transient binding events at other sites on the DNA (Fig 3.6a). We saw no evidence of Cas9 associating with target sites through mechanisms involving facilitated diffusion; instead, all target association appeared to occur directly from solution through 3D collisions, reminiscent of *E. coli* RNAP.

The shorter-lived, non-specific binding events exhibited complex dissociation kinetics, and the simplest model describing these data was double-exponential decay with lifetimes of $\sim$ 3.3 and $\sim$ 58 seconds (at 25 $mM$ KCl) (Fig 3.6b). These lifetimes were readily distinguished from the long lifetimes of either target bound Cas9 or apo-Cas9 (Fig 3.6b inset). Furthermore, the experiments were conducted in the presence of a saturating (10-fold) molar excess of crRNA:tracrRNA to exclude contamination from apo-Cas9. This result indicates that at least two and possibly more binding intermediates exist on the pathway towards full target recognition. As discussed in Chapter 1, non-specific DNA binding typically involves electrostatic interactions with the sugar-

phosphate backbone, and therefore non-specific lifetimes tend to decrease rapidly with increasing ionic strength. Interestingly, the lifetimes of Cas9 bound at non-specific DNA sites were not appreciably affected by salt concentration (Fig 3.6b). One remarkable implication of this finding is that the Cas9 non-target binding events have characteristics more commonly attributed to site-specific association.

### 3.3.2 Cascade locates targets by 3D diffusion *and* facilitating mechanisms

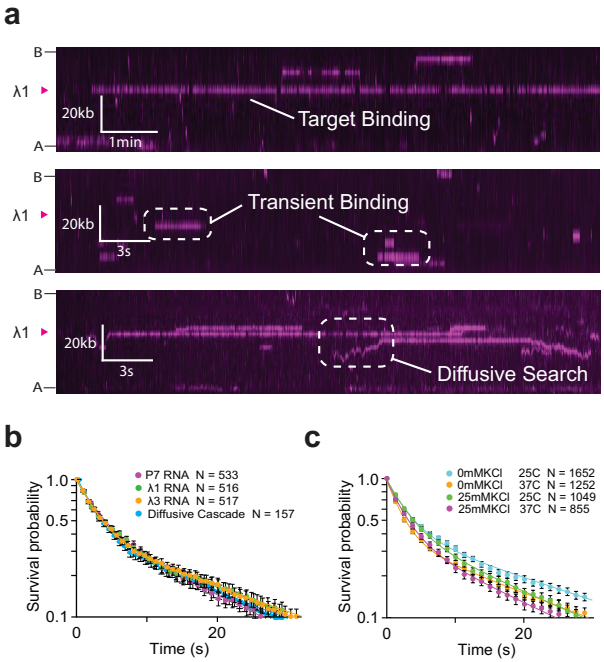


Figure 3.7: Experimentally observed Cas9 binding populations. a, Kymographs illustrating distinct binding events. b,c Survival probabilities for non-target binding events; solid lines represent double-exponential fits. b, Effect of different RNAs compared, as well as, the diffusive population of Cascade. c, Effect of experimental conditions on Cascade non-specific binding

Experiments directed at understanding Cascade's target search were performed as above with Cas9, with Cascade also programmed with a $\lambda 2$ crRNA. Comparable to the dynamics of Cas9, Cascade also exhibited long-lived binding events at the target site and transient binding events at other sites on the DNA (Fig 3.7a). However, in the case of Cascade, we were able to also visualize Cascade sliding along the DNA (Fig 3.7a). Curiously, this 1D motion was discontinuous; while moving along the DNA, Cascade often pauses, and, in fact, the most of time Cascade is bound to DNA it is in a non-diffusive state with its diffusive excursions constituting only a minority of

the total time in contact with DNA (Fig 3.7a). Overall, only $\sim$ 25% of Cascade binding events showed any microscopic diffusive motion, and these events typically only diffused for short stretches ($\sim$ 1-2 *kbp*) before transitioning into a non-diffusive complex. Furthermore, though rare, it was also possible to visualize Cascade locating and stably associating with target sites via this 1D pathway.

Importantly, both the lifetime of Cascade showing detectible motion and those events where 1D motion was absent, were identical (Fig 3.7b), indicating that they likely belong to the same population, and we only capture motion in cases where Cascade dwelled in the diffusive state long enough to escape the resolution of our experimental setup. The role of this diffusive state will be discussed further below.

Finally, Cascade non-target binding kinetics where phenomenologically identical to those of Cas9, i.e. two exponentially decaying states with lifetimes of $\sim$ 2.8 and $\sim$ 24 seconds (at $25mM$ KCl), regardless of the crRNA used to target the DNA (Fig. 3.7b). Additionally, as was the case for Cas9, Cascades non-specific DNA binding was not appreciably affected by salt concentration (Fig. 3.7c).

### 3.3.3 Cas9 and Cascade concern themselves with opposing ends of $\lambda$ DNA

To gain further insight into the nature of CRISPR-Cas search mechanisms, we measured the locations of all binding events for both Cas9 and Cascade (Fig. 3.8a,c). For the analysis, we included only Cascade binding events lacking 1D sliding to avoid ambiguity in determining position. However, if we separate Cascade binding trajectories showing motion into stationary and diffusive components, the position distribution of the stationary component is identical to that of Cascade events lacking 1D motion, giving further credence to the conclusion that we are sampling a single population (data not shown).

The lifetime of binding events for either Cas9 or Cascade did not vary substantially at different regions of the DNA (Fig 3.8a,c shown in color), which is to say that the heterogeneity of non-target bound lifetimes was not the result of long-lived binding of CRISPR-Cas complexes at a few specific locations in $\lambda$-DNA and shorter-lived binding elsewhere. However, the number of observed binding events was not uniformly distributed along the substrate, revealing an unequal recruitment to different regions of $\lambda$-DNA (Fig. 3.8a,c), suggesting that some underlying feature of $\lambda$-DNA might be influencing the target search. Interestingly, the distribution of Cascade binding events was anti-

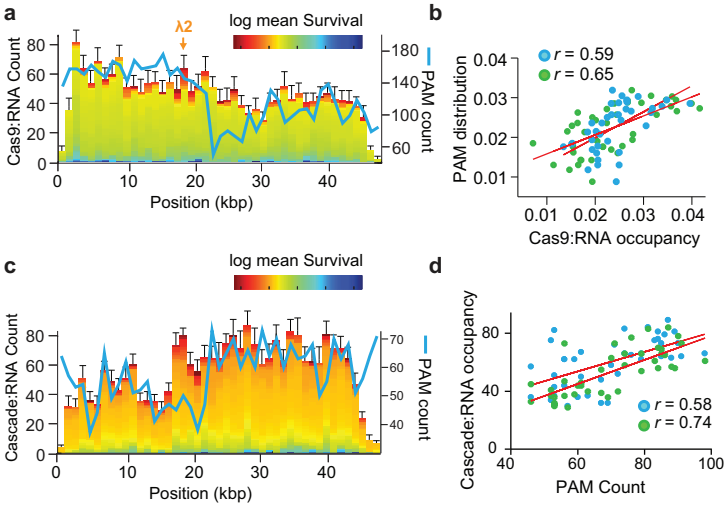


Figure 3.8: Cas9 and Cascade localize to PAM-rich regions during the target search. a, Distribution of Cas9 binding events ($N = 2,330$) and PAM density. Colour-coding reflects the binding dwell time relative to the mean dwell time. b, Correlation of PAM distribution and non-target Cas9 binding for $\lambda 2$ (blue) and spacer 2 (green) guide RNAs. c, Distribution of Cascade binding events ($N = 2,515$) and PAM density. Colour-coding is the same. d, Correlation of PAM distribution and non-target Cascade binding for $\lambda 2$ (blue) and P7 (green) guide RNAs.

correlated with Cas9's distribution (Fig 3.8a,c). This is an odd result for two reasons. First, both Cas9 and Cascade are loaded with crRNAs that target the same DNA sequence and it is reasonable to suspect that the binding of these Cas-complexes to DNA should reflect the required interactions between the crRNA and surveyed DNA. However, that Cascade and Cas9 are attracted to different regions suggests that these two complexes are more sensitive to elements of $\lambda$-DNA independent of the crRNA sequence. Second, an alternative hypothesis for the non-uniform binding to $\lambda$-DNA would be that these ribonucleoprotein complexes are reporting on the local melting temperature ($Tm$) of $\lambda$-DNA, because both proteins must melt the DNA in order to base pair their crRNA. Yet, while Cascade does bind more frequently to lower $Tm$ regions of $\lambda$-DNA, Cas9 does not. Instead, Cas9 seems to prefer binding to regions of higher $Tm$'s.

Instead, the most likely candidate for this signal is the PAM, the small conserved DNA sequence which always exists next to the protospacer only within the viral context (Fig 3.2). The PAM is in fact different for Cas9 and Cascade. Cas9 recognizes a $3'$-$NGG$-$5'$ PAM at the $3'$ end, and on the opposite DNA strand, of the protospacer. Cascade on the other hand, recognizes a $5'$-$AWG$-$3'$ PAM at the $5'$ end, and same DNA strand, of the protospacer. The $\lambda$ phage genome contains a total of $5,677$ $NGG$ PAM sites ( 1 PAM per 8.5 $bp$), and $3,151$ $AWG$ PAM sites ( 1 PAM

per 15.4 *bp*). Importantly, λ-DNA also has an unusual polar distribution of A/T- and G/C-rich sequences, which leads to an asymmetric distribution of PAMs, and importantly an opposing distribution of *NGG* PAMs to *AWG* PAMs (Fig. 3.8a,c). Both the Cas9 and Cascade binding site distributions were positively correlated with their respective PAM distributions (Fig. 3.8b,d), and we repeated this experiment using crRNAs having no complementary target sites within λ-DNA, and found no change in the observed binding lifetimes and even stronger correlations with the individual PAM distributions (Fig. 3.8b,d). These results, together with the insensitivity of short-lived binding events to ionic strength, suggest that the target search of both Cas9 and Cascade might be mediated by PAM sequences.

### 3.3.4 Cas9 binds exclusively at PAMs

To test the hypothesis that Cas9 uses PAM recognition as an obligate precursor to interrogation of flanking DNA for potential guide-RNA complementarity, we used competition assays to monitor the rate of Cas9-mediated DNA cleavage (Fig. 3.9a,b). In this assay, a good competitor would be sampled frequently by Cas9 and slow down the proteins search for and eventual cleavage of the labeled target. Alternatively, a bad competitor would have little to no effect on the proteins path to target. By measuring the change in the cleavage rate due to the presence of competitor DNA, we can extract the average amount of time that Cas9 spends sampling competitor DNA prior to locating and cleaving a radiolabeled substrate.

First, we used an unlabeled competitor DNA lacking *NGG* PAMs and bearing no sequence relationship to the crRNA. In these experiments the reaction proceeded at a rate almost indistinguishable from reactions lacking any competitor (Fig. 3.9b,c). Alternatively, using a competitor containing both a PAM and a fully complementary target sequence completely inhibited the reaction (Fig 3.9b). Next, a series of competitors were tested that bore no complementarity to the crRNA guide sequence but contained increasing numbers of PAMs (Fig. 3.9c,d). There was a direct correspondence between the number of PAMs, both within the competitor DNA and overall concentration, and the ability of a DNA competitor to interfere with target cleavage, indicating that the lifetime of Cas9 on competitor DNA increased with PAM density (Fig. 3.9c,d). These results demonstrate that the residence time of Cas9 on non-target DNA lacking PAMs is negligible, and support the hypothesis that the transient, non-target DNA binding events observed on the

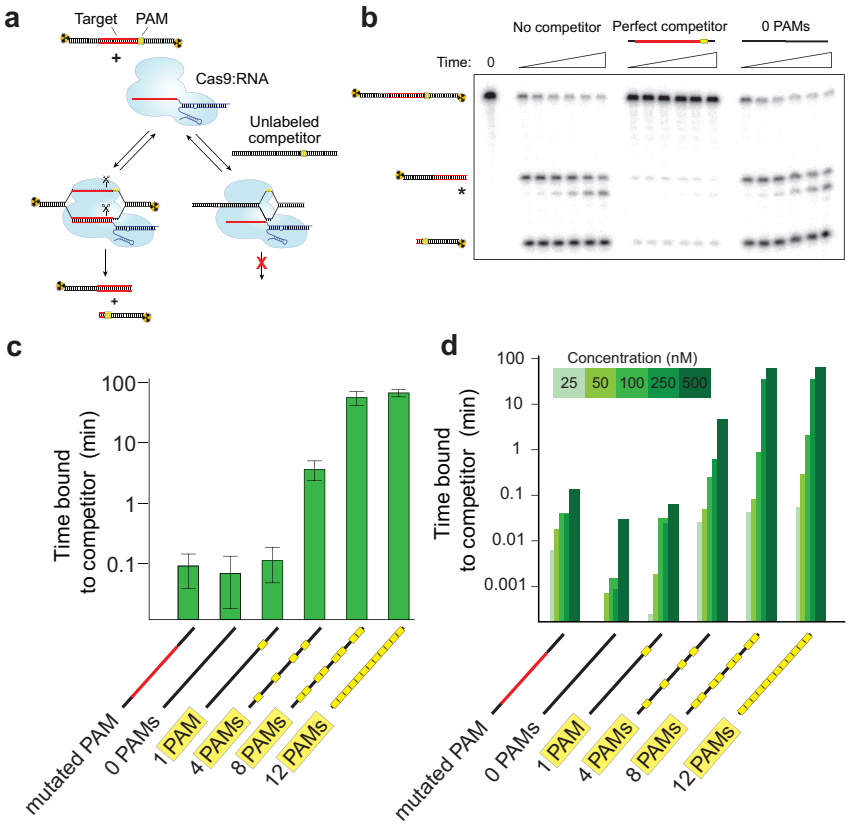


Figure 3.9: Cas9 searches for PAMs. a, Schematic of the competition cleavage assay. b, Cleavage assay with and without competitor DNAs. c, Quantitation of competition data (mean ± s.d.). Competitor cartoon representations show PAMs (yellow) and regions complementary to the crRNA (red). d, Competition data as in with c, but shown for all five concentrations of competitor tested.

DNA curtains likely occurred exclusively at PAM sequences. While Cas9 complexes undoubtedly sample DNA lacking PAMs, these rapid binding events are neither detectable in single-molecule assays or bulk binding experiments nor do they appreciably influence overall reaction kinetics in bulk biochemical assays.

To conclusively determine if Cas9's target search proceeds exclusively through PAM binding, we repeated the competition assay with a competitor bearing perfect complementarity to the crRNA, but with a single point mutation in the adjacent PAM (5′-*TCG*-3′) (Fig. 3.9c,d). As a competitor, this substrate failed to inhibit cleavage of the PAM bearing target DNA by Cas9 and behaved comparably to the non-target competitor DNA lacking PAMs, despite the fact that it contained perfect complementarity to the crRNA (Fig. 3.9c,d). Together, these results demonstrate that PAM recognition is the obligate first step during target recognition by Cas9.

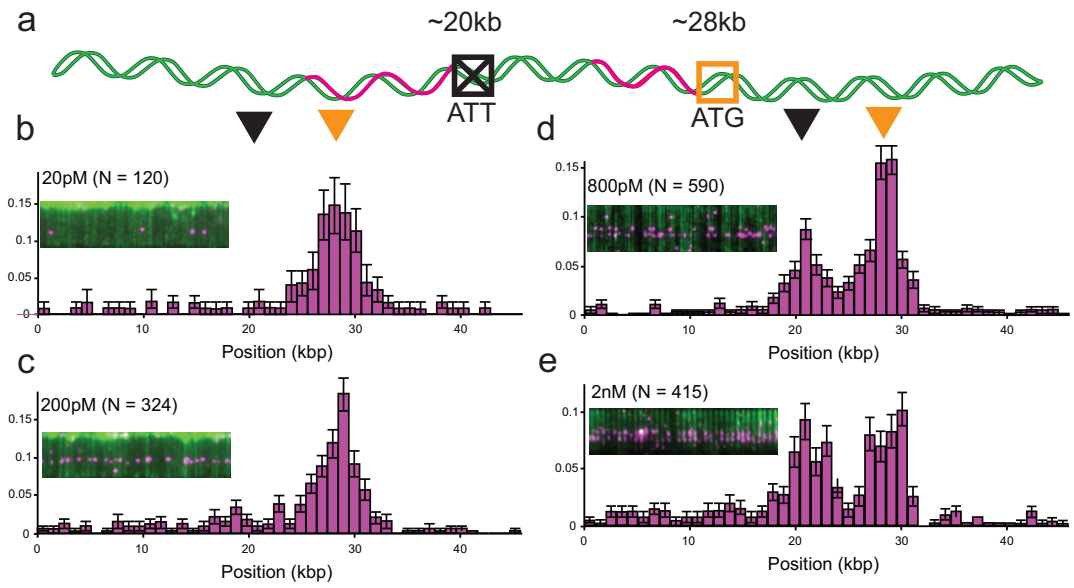


Figure 3.10: Cascade uses PAMs to rapidly locate targets. a, Cartoon of ePAM-λ-DNA, the wt-target is shown in orange, and the escape target in black. b-e, Binding distributions for four concentrations of Cascade, the wt-target is indicated by orange arrows and the escape target by black arrows. Errorbars represent 70% confidence intervals

### 3.3.5 PAM directs Cascade association

The biggest difference between *E. coli* Cascade and *S. pyogenes* Cas9, is that Cascade exhibits substantial nonspecific (i.e., non-PAM) binding, and because of this, competition experiments, like those above were not feasible. But, it is important to note, the mere existence of a diffusive state indicates that Cascade does not only interact with DNA in a PAM-dependent fashion. Yet, the binding we observed in our single molecule assay revealed a correlation between Cascade binding and *AWG* PAM sites. Given these results, and the behavior of Cas9, we arrived at the hypothesis that Cascade has the ability to bind non-specifically for short periods of time but still prefers PAM sites, and is likely captured rapidly at PAM sites due to their density in the λ-DNA substrate.

To test this idea, we developed an assay to assess whether PAMs act as a recruitment signal for Cascade binding. In these experiments we employed a modified DNA substrate, ePAM-λ-DNA, which contained two identical protospacer targets (Fig 3.10a). One target flanked by a true *ATG* PAM, and another target where the PAM contained a single base mutation, *ATT*. The protospacer with the *ATT* flanking sequence escapes CRISPR immunity *in vivo* [Datsenko *et al.*, 2012], and is referred to as an escape target; likewise the flanking *ATT* sequence is called an escape PAM. Single molecule experiments using ePAM-λ-DNA showed that targets adjacent to a true PAM

readily recruited Cascade at low concentrations (20-200 $pM$), whereas Cascade failed to occupy the escape target at the same concentrations (Fig 3.10b,d). However, at higher concentrations, Cascade was able to occupy both targets (Fig. 3.10 c,e). This result shows that there is an apparent $K_D$ defect for Cascade binding at escape sites, but the ability of Cascade to bind to DNA non-specifically allows the protein to remain local to the DNA long enough to either capture, or potentially induce, local melting of the DNA, and eventually capture the escape target. Cascades ability to locate targets bearing mutations, both in the PAM and protospacer sequences, plays an important role in adaptation of CRISPR immunity [Datsenko *et al.*, 2012; Blosser *et al.*, 2015; Fineran *et al.*, 2014]. Finally, we conclude that, while PAM binding is not a necessary precursor to target location and recognition by Cascade, it can direct the search process by using PAMs as signals of potential target sites, evident by the fact that the PAM-dependent search is able to locate PAM-flanked targets at concentrations an order of magnitude lower than for non-PAM bearing targets (Fig. 3.10 b-e).

## 3.4 Mechanism of RNA:DNA heteroduplex formation

After PAM recognition, CRISPR complexes must destabilize the adjacent duplex and initiate strand separation to enable base-pairing between the target DNA strand and the crRNA guide sequence. Because neither Cas9 nor Cascade has energy-dependent helicase activity, the mechanism of local DNA unwinding must rely upon thermally available energy. One possibility is that PAM binding could induce a general destabilization of the duplex along the length of the entire target sequence, leading to random nucleation of the RNA:DNA heteroduplex (Fig. 3.11a, top). Alternatively, PAM binding may cause only local melting of the duplex, with the RNA:DNA heteroduplex nucleating at the 3' end of the target sequence next to the PAM and proceeding sequentially towards the distal 5' end of the target sequence (Fig. 3.11a, bottom).

To distinguish between these two models for Cas9, we returned to our cleavage assay and designed a panel of DNA competitors in which the length and position of complementarity to the guide RNA was systematically varied. These competitors were intended to distinguish between the random nucleation and sequential unwinding models for heteroduplex formation based upon the predicted patterns of cleavage rate inhibition for each model (Fig. 3.11a). The ability of a

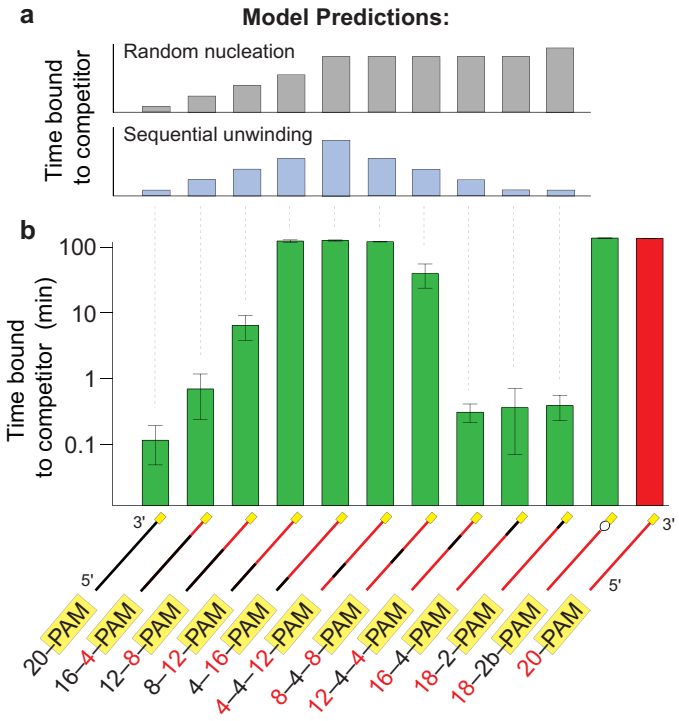


Figure 3.11: Cas9 unwinds dsDNA in a directional manner. a, Predicted data trends for the random nucleation or sequential unwinding models aligned with the corresponding data in b. b, Competition assays using substrates with variable degrees of crRNA complementarity, shown as in FIg. 3.9c,d. Numeric descriptions of the competitor DNAs indicate the regions of complementarity (red) or mismatches (black) to the crRNA sequence.

competitor DNA to inhibit substrate cleavage by Cas9 increased as the extent of complementarity originating at the 3′ end of the target sequence adjacent to the PAM increased (Fig. 3.11b). Inhibition increased dramatically when 12 or more base pairs were complementary to the crRNA guide sequence, which agrees with a previously reported requirement of an 8-12 nucleotide seed sequence for the Cas9-DNA cleavage reaction [Jinek *et al.*, 2012; Jiang *et al.*, 2013]. Strikingly, although competitors containing mismatches to the crRNA at the 5′ end of the target sequence competed effectively for Cas9 binding, competitors containing mismatches to the crRNA at the extreme 3' end immediately adjacent to the PAM were completely inert to binding (Fig. 3.11b). This was true even with a 2 *bp* mismatch followed by 18 *bp* of contiguous sequence complementarity to the crRNA. Therefore, when mismatches to the crRNA are encountered within the first two nucleotides of the target sequence, Cas9 loses the ability to interrogate and recognize the remainder of the DNA.

The pattern of inhibition observed with the different competitor DNAs indicates that sequence

homology adjacent to the PAM is necessary to initiate target duplex unwinding until the reaction has proceeded sufficiently far ( 12 *bp*, approximately one turn of an A-form RNA:DNA helix), such that the energy necessary for further propagation of the RNA:DNA heteroduplex falls below the energy needed for the reverse reaction. These findings suggest that formation of the RNA:DNA heteroduplex initiates at the PAM and proceeds through the target sequence by a sequential, stepwise unwinding mechanism consistent with a Brownian ratchet [Abbondanzieri *et al.*, 2005].

As a further test of this model, we used a DNA competitor that contained mismatches to the crRNA at positions 1-2 but was, itself, mismatched at the same two positions, forming a small bubble in the duplex. Despite the absence of sequence complementarity to the crRNA within the DNA bubble, this substrate was a robust competitor and bound Cas9 with an affinity nearly indistinguishable from that of an ideal substrate (Fig. 3.11b). Remarkably, this DNA could also be cleaved with near wild-type rates [Sternberg *et al.*, 2014]. We speculate that, with the energy of DNA melting already paid, Cas9 bypasses the mismatches and initiates nucleation of the RNA:DNA heteroduplex downstream of the bubble, thereby propagating strand separation through the remainder of the target.

Recent studies using fluorescence resonance energy transfer and short fluorescently labeled oligos determined that Cascade utilizes the same thermally ratcheted mechanism for heteroduplex formation [Rutkauskas *et al.*, 2015], suggesting that this mechanism is common to all CRISPR surveillance complexes.

## 3.5 The role of PAM in DNA degradation activity

### 3.5.1 The PAM triggers Cas9 nuclease activity

The results above indicate that PAM binding plays a central role in target recognition, and that for Cas9, introduction of a small bubble in the DNA target eliminates the need for RNA heteroduplex formation immediately adjacent to the PAM. One might expect PAM recognition to be dispensable for Cas9-mediated recognition and cleavage of a single-stranded DNA (ssDNA) target. Surprisingly, however, ssDNA substrates were cleaved more than two orders of magnitude slower than a double-stranded DNA (dsDNA) substrate (Fig. 3.12a,b), despite the fact that dCas9 bound both the dsDNA and ssDNA substrates with similar affinities (Fig. 3.12b). Importantly, Cas9 recognizes

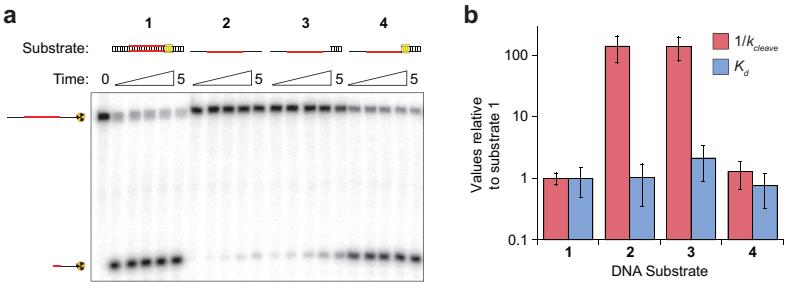


Figure 3.12: PAM recognition regulates Cas9 nuclease activity. a, Cleavage assay with single-stranded, double-stranded and partially double-stranded substrates. b, Relative affinities and cleavage rates (mean±s.d.).

the 5′-*NGG*-3′ PAM on the non-target DNA strand [Jinek *et al.*, 2012], so ssDNA substrates do not contain a PAM but rather the complement to the PAM sequence. We hypothesized that the absence of the PAM on the ssDNA might explain why an otherwise fully complementary target is resistant to cleavage. To test this possibility, we prepared hybrid substrates with varying lengths of dsDNA at the 3′ flanking sequence (Fig. 3.12a). Cleavage assays revealed that the ssDNA target strand could be activated for cleavage in the presence of flanking dsDNA that extended across the PAM sequence, but that this activating effect was lost when the dsDNA was truncated immediately before the PAM (Fig. 3.12a,b). Binding experiments confirmed these results were not a consequence of discrimination at the level of binding (Fig. 3.12b). Rather, the presence of the 5′-*NGG*-3′ PAM on the non-target strand was critical for some step of the reaction that occurred after binding. These data suggest that the PAM acts as an allosteric regulator of Cas9 nuclease activity.

### 3.5.2 The PAM is required to recruit Cas3 and license nuclease activity

In *E. coli*, Cascade's major role is as a surveillance complex, serving to locate foreign DNA and signal its presence to the cell. This signal is received by another Cas protein, Cas3, that is responsible for target degradation [Sinkunas *et al.*, 2011; Hochstrasser *et al.*, 2014; Mulepati and Bailey, 2013] (Fig. 3.1). Cas3 is a single stranded DNA nuclease, a 3′ to 5′ helicase, and digests Cascade bound DNA both *in vivo* and *in vitro* [Sinkunas *et al.*, 2011; Mulepati and Bailey, 2013]. In the *S. pyogenes* CRISPR pathway, Cas9 does the work of both Cascade and Cas3, and therefore it is crucial to the cell to minimize off target binding by Cas9, because there is the opportunity to accidentally cut the wrong DNA. For Cascade, minimization of off target binding is less stringent, presumably because

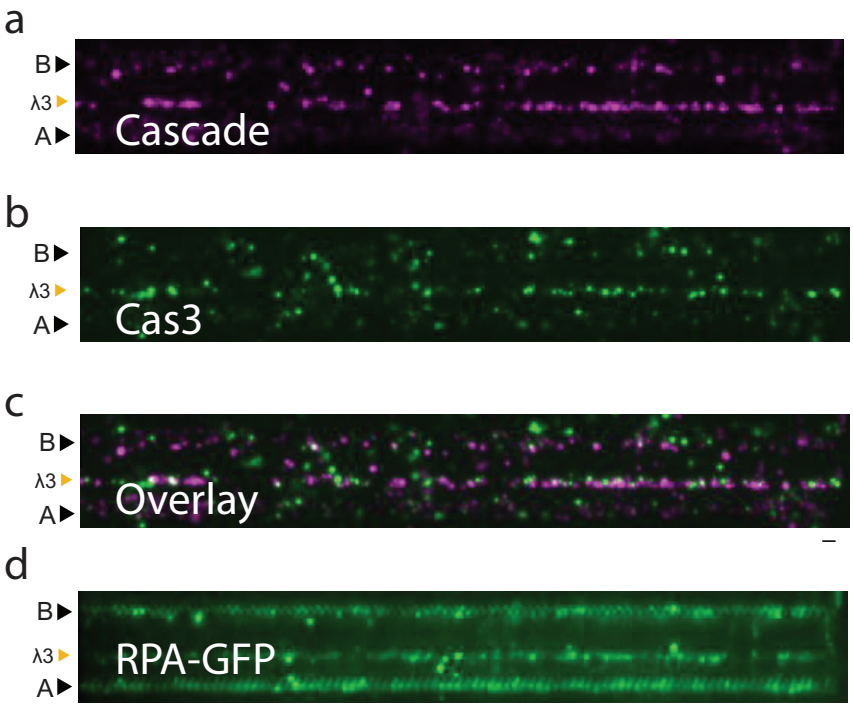


Figure 3.13: Cas3 binds to Cascade bound DNA. Widefield images of a, Cascade bound to the λ3 target, b, Cas3 colocalized to the λ3 target, c, overlay of a and b, and d, ATP, Cascade, and Cas3-dependent, RPA-GFP colocalization to the λ3 target.

Cascade is not able to digest DNA; the control has been surrendered to Cas3 where it is now critical that Cas3 is recruited solely to *bona-fide* targets. Here, we have purified an N-terminal biotinylated Cas3 from *E. coli* and directly visualized QD-labelled Cas3 on DNA curtains. These experiments reveal that Cas3 is exclusively recruited to Cascade bound DNA, lacks detectible interactions with non-Cascade bound DNA, and digests DNA identified by Cascade (Fig. 3.13, Fig. 3.14a).

Because Cas3 is a single stranded nuclease, it leaves in its wake short ($\sim$ 200 *bp*) stretches of ssDNA (Fig 3.13d), and we are able to monitor Cas3-mediated digestion by labeling these ssDNA gaps with a single stranded binding protein from *Saccharomyces cerevisiae* fused to green fluorescent protein, (RPA-GFP) [Gibb *et al.*, 2014] (Fig. 3.13d and Fig. 3.14b). We show that Cas3 preferentially binds to DNA, only where Cascade is bound (Fig. 3.3f, Fig. 3.14a), and that in the presence of ATP, RPA-GFP also colocalizes to the λ-3 target site (Fig 3.13d, Fig. 3.14b). However, Cascade bound at escape sites, (*c.f.* Fig. 3.10e) failed to recruit Cas3 (Fig. 3.14c). This means that Cas3 either itself recognizes PAM sequences or recognizes Cascade only in the PAM bound state. Furthermore, recognition of Cascade at wild type target sites is required to license

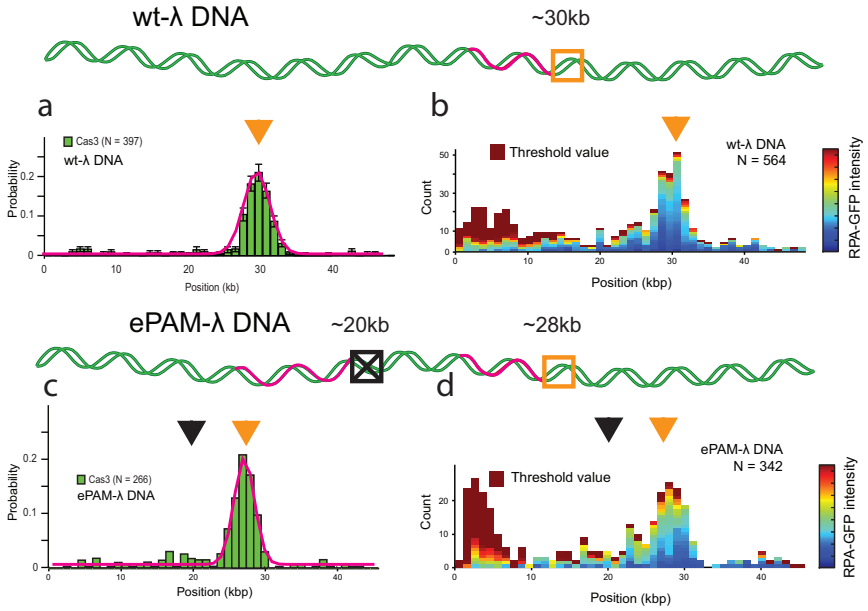


Figure 3.14: Cas3 preferentially binds and digests PAM-bearing targets. a, Cas3 binding distribution on wt-λ-DNA. b, Position distribution of RPA-GFP puncta following Cas3 digestion on wt-λ-DNA. The relative size of the ssDNA gap is related to the integrated intensity of the local GFP signal, shown in color. c, Cas3 binding distribution on ePAM-λ-DNA. d, RPA-GFP signal on ePAM-λ-DNA, data shown as in b.

Cas3 for DNA degradation (Fig. 3.14d), revealing that PAMs are as central to the regulation of CRISPR immunity in *E. coli* as they are in *S. pyogenes*.

## 3.6 Discussion

Our results suggest a general model for target binding and cleavage by CRISPR-Cas complexes hailing from evolutionarily distant organisms, which involves an unanticipated level of importance for PAM sequences at each stage of the reaction (Fig. 3.15, 3.16). Although minor details may differ, as in the case of the two systems presented here, we hypothesize that PAM interactions may play a similar role in all CRISPR RNA-guided immune systems.

The Cas9 target search begins with random collisions with DNA. However, rather than sampling all DNA equivalently, Cas9 accelerates the search by rapidly dissociating from non-PAM sites, thereby reducing the amount of time spent at off-targets. Only upon binding to a PAM site does Cas9 interrogate the flanking DNA for guide RNA complementarity (Fig. 3.15). The requirement for initial PAM recognition by Cas9 also eliminates the potential for suicidal self-targeting, since

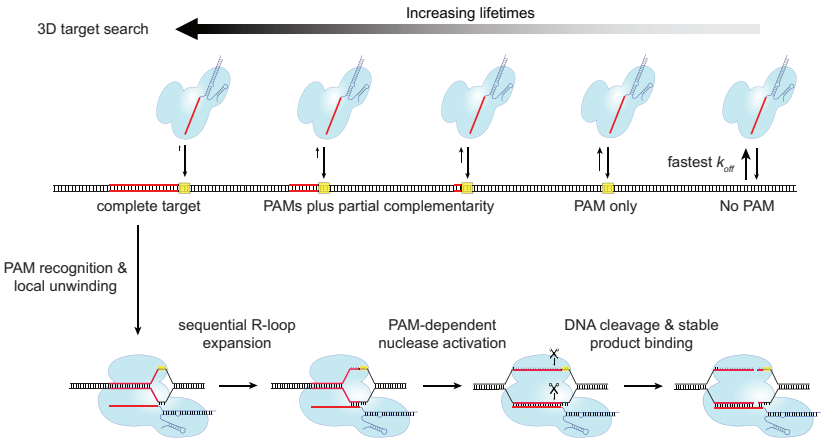


Figure 3.15: Model for target search, recognition and cleavage by Cas9. The search initiates through random three-dimensional collisions. Cas9 rapidly dissociates from non-PAM DNA, but binds PAMs for longer times and samples adjacent DNA for guide RNA complementarity, giving rise to a heterogeneous population of intermediates. At correct targets, Cas9 initiates formation of an RNA:DNA heteroduplex, and R-loop expansion propagates via sequential unwinding. The DNA is cleaved, and Cas9 remains bound to the cleaved products.

perfectly matching targets within the bacterial CRISPR locus are not flanked by PAMs.

Likewise, Cascade's search also uses PAMs to rapidly funnel Cascade into target sites, but also has a secondary and parallel pathway allowing for location and recognition of mutated target sites (Fig 3.16). This secondary pathway enables the *E. coli* CRISPR system to adapt to foreign DNA that has acquired an escape mutation; but, this added flexibility in Cascade's target search is likely permitted due to the strict requirement of PAMs at sites of DNA degradation by Cas3 (Fig 3.16).

Our results suggest that PAM recognition coincides with initial destabilization of the adjacent sequence, as evidenced from experiments using a bubble-containing DNA substrate, followed by sequential extension of the RNA:DNA heteroduplex. This mechanism explains the emergence of seed sequences, because mismatches encountered early in a directional melting-in process would prematurely abort target interrogation. Moreover, the complex dissociation kinetics observed on non-target $\lambda$-DNA would arise from heterogeneity in potential target sites as Cas9 or Cascade probe sequences adjacent to PAMs for guide RNA complementarity. Hybridization to a correct target then leads to activation of Cas9's nuclease domains or Cascade specific recruitment and licensing of Cas3. These steps also require PAM recognition, providing a surprising level of PAM-dependent regulation that ensures protection against self-cleavage of the CRISPR locus or near cognate sites scattered throughout the host genome.

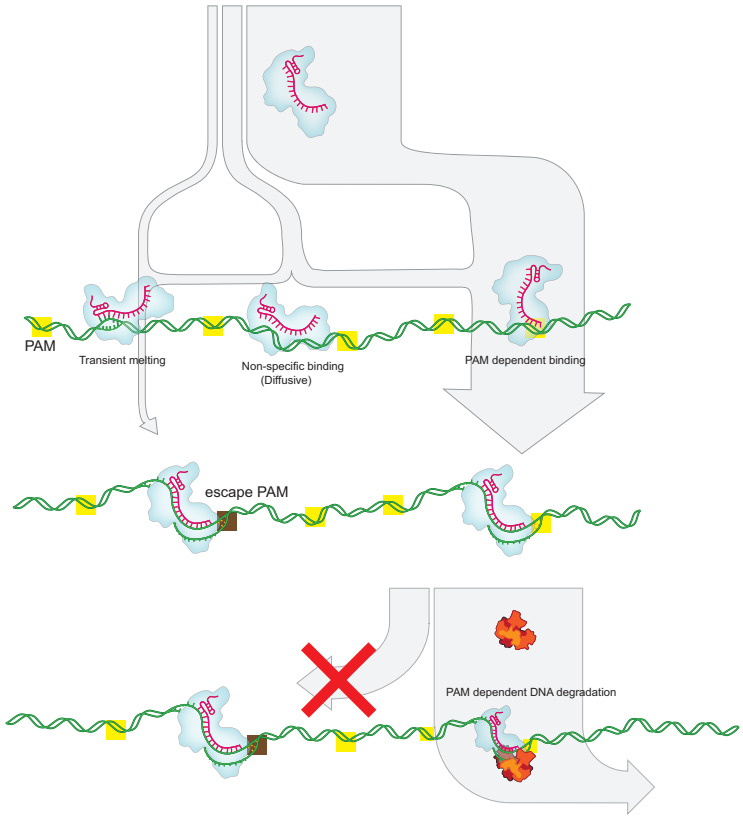


Figure 3.16: Model for target search and recognition by Cascade, and cleavage by Cas3. Cascade binding to DNA is either PAM-mediated (wide pathway) of non-specific (narrow path). Following dsDNA melting, Cascade only at PAM-bearing targets recruits Cas3 to digest foriegn DNA

Incredibly, each of the target searches we have considered thus far have remarkably similar qualities. Both RNAP and CRISPR interference complexes search for targets in DNA that lead to melting of dsDNA bubbles. These downstream reaction species can only be accessed through a substantial energetic barrier, which can lead to the protein or protein complex being trapped in local energy minimas. However, both the polymerase and the Cas complexes initiate their searches by identifying small signals in DNA, which have a higher probability of being the "right" place, thereby reducing the probability of the reaction ending up in a dead end. In the next chapter, we will discuss another target search that also requires DNA melting. This final target search highlights the magnitude of the advantage inherent to this search strategy.

# Chapter 4

# Mechanism of DNA sequence alignment during homologous recombination

---

This work was was originally published as: "DNA Sequence Alignment by Microhomology Sampling during Homologous Recombination", Zhi Qi, Sy Redding, Ja Yil Lee, Bryan Gibb, YoungHo Kwon, Hengyao Niu, William A. Gaines, Patrick Sung, Eric C. Greene, Cell. 2015 Feb 26;160(5):856-69.

Author contributions: Z.Q. designed and conducted the single-molecule experiments and data analysis. S.R. conducted all theoretical calculations and assisted in data analysis and experimental design. J.Y.L. assisted with single-molecule experiments, data analysis, and experimental design. B.G. expressed and purified human and yeast RPA and assisted with Rad51 characterization. Y.K., H.N., and W.G. purified yeast and human Rad51 and yeast Dmc1. E.C.G. supervised the project and wrote the manuscript with input from all co-authors.

---

## 4.1 Introduction

Homologous recombination (HR) is ubiquitous among all three kingdoms of life and serves as a driving force in evolution. HR is a major pathway for repairing DNA double-strand breaks (DSBs)

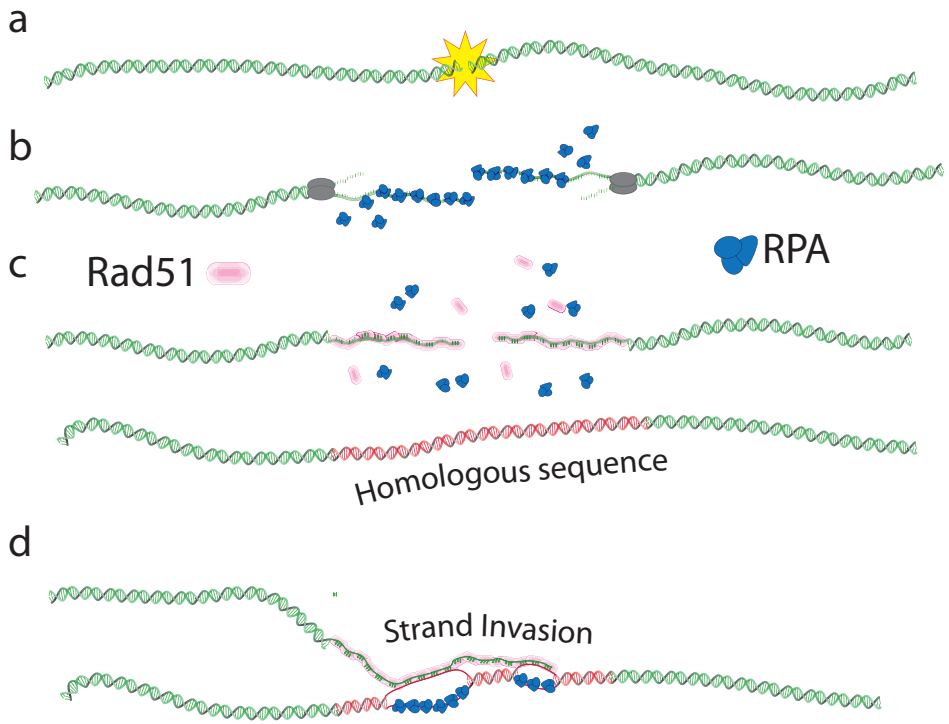


Figure 4.1: The homologous recombination pathway. a, DSB produced in the cell are, b, resected and coated with RPA. c, Rad51, exchanges with RPA to form PCs. d, The homology search; PCs search the genome for complementary DNA through strand invasion events

and single-strand DNA (ssDNA) gaps, and plays essential roles in repairing stalled or collapsed replication forks [Heyer *et al.*, 2010; Filippo *et al.*, 2008]. HR provides an alternative pathway for telomere maintenance [Eckert-Boulet and Lisby, 2010], can lead to the duplication of long regions of chromosomes [Smith *et al.*, 2007], and some organisms utilize HR as the sole means of initiating DNA replication [Hawkins *et al.*, 2013]. HR also generates genetic diversity and ensures proper chromosome segregation during meiosis [Neale and Keeney, 2006], and is a major source of phenotypic variation in many organisms [Fraser *et al.*, 2007; Hastings *et al.*, 2009]. In humans, aberrant HR underlies chromosomal rearrangements often associated with cancers, cancer prone syndromes, and numerous genetic diseases [Heyer *et al.*, 2010; Filippo *et al.*, 2008].

DSB repair in *Saccharomyces cerevisiae* has long served as paradigm for studying HR (Fig. 4.1) [Heyer *et al.*, 2010; Filippo *et al.*, 2008]. The DNA ends present at DSBs are first processed by $5'$ to $3'$ strand resection, yielding 3 ssDNA overhangs whose production coincides with the binding of replication protein A (RPA) (Fig. 4.1a,b). RPA is then replaced by either Rad51 or the meiosis-

specific recombinase Dmc1 (Fig. 4.1c), which is thought to have arisen by a gene duplication event early in the evolutionary history of eukaryotes [Lin *et al.*, 2006]. Rad51 and Dmc1 are both closely related to the *E. coli* recombinase protein RecA. These proteins are all DNA-dependent ATPases that form right-handed helical filaments on ssDNA, and the resulting presynaptic complexes (PCs) display a striking degree of conservation from bacteriophage to humans [Eggleston and Kowalczykowski, 1991].

Structural studies have revealed that the presynaptic ssDNA-protein filament is organized into base triplets that are maintained in near B-form conformation, but there is a 7.8 $\AA$ rise between adjacent triplets causing an overall extension of the ssDNA relative to B-form DNA [Chen *et al.*, 2008]. Once assembled on the ssDNA, Rad51/RecA recombinases must align their substrate with a homologous duplex elsewhere in the genome (Fig. 4.1d). This process is referred to as the homology search and it is conceptually similar to the target searches conducted by all other site-specific DNA-binding proteins [Barzel and Kupiec, 2008; Renkawitz *et al.*, 2014; Vonhippel and Berg, 1989]. The principles that govern sequence alignment during HR remain poorly understood because the corresponding intermediates are transient and asynchronous [Barzel and Kupiec, 2008; Renkawitz *et al.*, 2014].

What features are the recombinases searching for within dsDNA? How do they distinguish between nonhomologous and homologous sequences? Over what length scales do they test for homology? What distinguishes search intermediates from the commitment to strand exchange? These questions all pertain to the overarching issue of how homology is efficiently located given the vast sequence space encoded by the genome [Neale and Keeney, 2006]. We sought to address these questions by visualizing the homology search at the single-molecule level. Our results lead to a model in which 8-nt microhomology motifs serve as the fundamental units of molecular recognition by *S. cerevisiae* Rad51, and this initial event is distinct from subsequent strand invasion. We show that the physical principles underlying the ability of Rad51 to search for and align homologous DNA sequences are broadly conserved among the Rad51/RecA family members. This remarkable mechanism can drastically reduce the amount time necessary to align homologous dsDNA sequences.

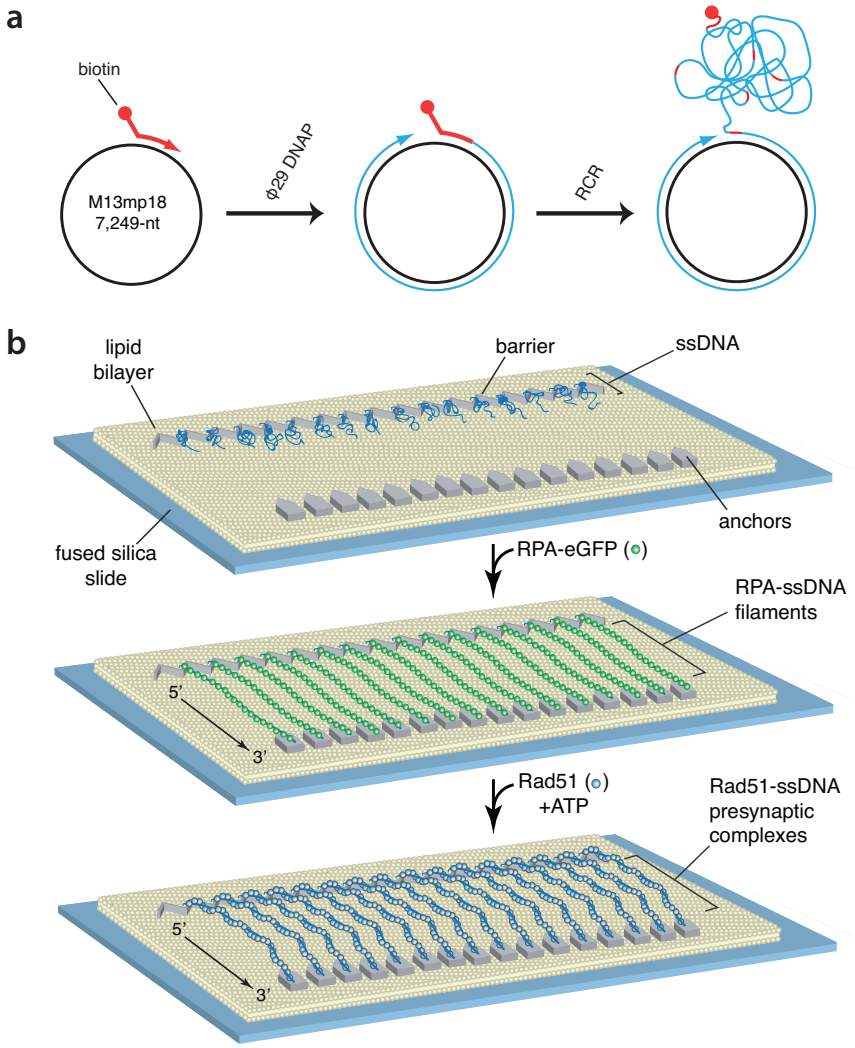


Figure 4.2: Single-stranded DNA curtains and presynaptic complex assembly. a, Outline of procedure for preparation of 5 biotinylated ssDNA substrate by rolling circle replication of a circular M13mp18 ssDNA template. b, Schematic illustrating the procedure for making double-tethered ssDNA curtains bound by RPA-eGFP, followed by the assembly of presynaptic complexes comprised of wild-type Rad51.

## 4.2 Assembly of Rad51 presynaptic complexes

We used ssDNA curtains and total internal refection fluorescence microscopy to visualize Rad51 PCs [Gibb *et al.*, 2014]. The ssDNA was generated using M13mp18 (7,249-nt) as a template for rolling circle replication (Fig 4.2a), and then anchored to the lipid bilayer in our flowcells through a biotin-streptavidin linkage, just like the anchoring of dsDNA in previous chapters (Fig. 4.2b). The ssDNA unravels when incubated with RPA-eGFP, and the downstream ends of the RPA-ssDNA are able to non-specifically anchor to the nanofabricated pedestals (Fig 4.2b). Addition of wild-type *S. cerevisiae* Rad51 led to efficient, ATP-dependent PC assembly (Fig. 4.2b).

## 4.3 Nonhomologous dsDNA capture by Rad51

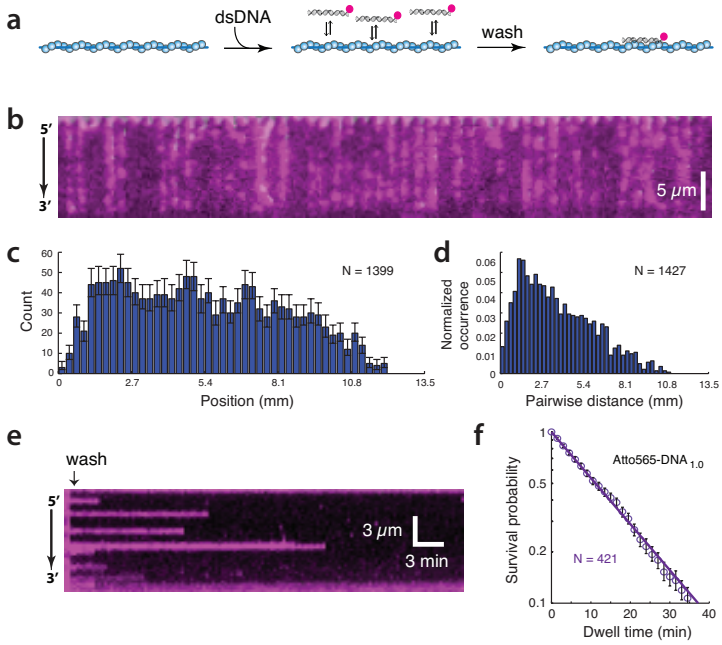


Figure 4.3: Visualizing dsDNA capture by Rad51. a, Strategy for detecting binding of Atto565-labeled dsDNA to the PCs. b, Wide-field image of Rad51 PCs bound to Atto565-DNA$_{1.0}$. c and d, Binding site distribution, d, and pair-wise distance distribution, d, of Atto565-DNA$_{1.0}$. e, Kymograph showing dissociation Atto565-DNA$_{1.0}$ from a single Rad51 PC; 100-msec frames were collected at 20-s intervals. f, Dissociation kinetics of Atto565-DNA$_{1.0}$. Unless otherwise stated, error bars for all binding site distributions and survival probability plots represent 70% confidence intervals obtained through bootstrap analysis.

Rad51/RecA recombinases must interrogate nonhomologous dsDNA while attempting to locate and align homologous sequences. We mimicked this process by testing the ability of the Rad51

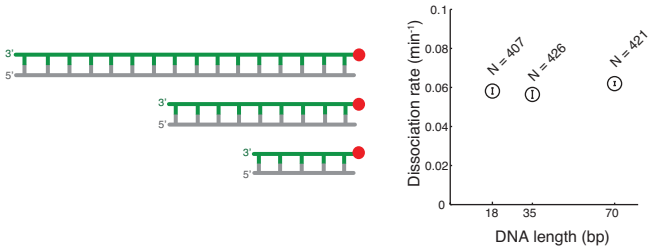


Figure 4.4: Influence of dsDNA fragment length on binding to the Rad51-ssDNA presynaptic complexes. a, Schematic of different duplex DNA substrates; the 35-bp and 18-bp substrates are truncations of the 70-bp sequence Atto565-$DNA_{1.0}$. b, Dissociation rates for the Rad51-ssDNA presynaptic complex for each of the different length dsDNA substrates.

PCs to interact with nonhomologous 70-base pair dsDNA oligonucleotides (Fig 4.3a). To visualize dsDNA binding, we injected Atto565 labeled dsDNA into the sample chamber; for brevity we designated this substrate $DNA_{1.0}$. Following a brief incubation, unbound dsDNA was flushed away and the remaining molecules were visualized. These experiments revealed $DNA_{1.0}$ bound to the PCs with no evident site preference within our resolution limits (Fig 4.3b,c,d), and most of the bound dsDNA (78.4%) exhibited single-step photo-bleaching [Qi *et al.*, 2015]. Controls with RPA-ssDNA (minus Rad51) confirmed that dsDNA capture was Rad51-dependent [Qi *et al.*, 2015]. In addition, the PCs rapidly disassembled when ATP was replaced with ADP, and the bound dsDNA was also quickly released when reactions were chased with ADP, indicating that dsDNA retention required the continued presence of Rad51 [Qi *et al.*, 2015]. Kinetic measurements yielded a dissociation rate ($k_{off}$) of $0.062 \pm 0.001 min^{-1}$ for $DNA_{1.0}$, corresponding to a lifetime of $\sim$ 16 minutes (Fig. 4.2e,f). This was an extraordinarily stable interaction for a seemingly nonhomologous dsDNA, and such long-lived intermediates would appear incompatible with an efficient search mechanism. We next sought to understand the physical basis for these long lifetimes.

### 4.3.1 Substrate length does not impact dsDNA retention

If nonhomologous dsDNA capture primarily involved nonspecific electrostatic contacts with the phosphate backbone, then the lifetime of the bound intermediates should vary with dsDNA length. We tested this possibility with 35-bp and 18-bp dsDNA substrates. Surprisingly, the truncated substrates bound tightly to the PCs, although more substrate and longer incubation times were required for initial engagement [Qi *et al.*, 2015]. We conclude that substrate length had a modest impact on initial association with the PC, but did not affect retention of the captured dsDNA,

suggesting that the observed intermediates were not maintained primarily through nonspecific contacts with dsDNA phosphate backbone.

### 4.3.2 Microhomology contributes to dsDNA capture

We next asked whether sequence microhomology might contribute to dsDNA capture. Analysis of $DNA_{1.0}$ revealed many short tracts of microhomology complementary to sequences scattered throughout the M13mp18 ssDNA, including 12 regions with $\geq$ 8-nts of microhomology (Fig. 4.5a,b).

Previous reports suggested that *E. coli* RecA can pair DNA substrates perhaps as short as 8-nt in length [De Vlaminck *et al.*, 2012; Hsieh *et al.*, 1992; Xiao *et al.*, 2006]. Based on this knowledge, we designed a new substrate ($DNA_{2.0}$), which retained global sequence composition to $DNA_{1.0}$, but lacked any microhomology $\geq$ 8-nt in length (Fig. 4.5d,e). While we readily detected capture of $DNA_{1.0}$ (Fig. 4.5c), we were unable to detect stable capture of $DNA_{2.0}$ under identical conditions (Fig. 4.5f), despite the fact that this substrate contains numerous tracts of microhomology $\geq$ 7-nt in length (Fig. 4.5d).

## 4.4 Stable dsDNA capture requires 8-nt tracts of microhomology

Our results imply that dsDNA capture involves 8-nt or longer tracts of microhomology. This hypothesis predicts that a single 8-nt tract of microhomology added to an otherwise nonhomologous dsDNA should confer stable association with the PC. We tested this prediction with a series of substrates bearing precisely 8-nt of microhomology (Fig. 4.6a). Remarkably, addition of a single 8-nt tract of microhomology was sufficient to confer stable binding of a nonhomologous dsDNA to the PC, and similar results were obtained for 8-nt microhomology motifs at different locations (Fig. 4.6b-e).

The binding site distributions and the pairwise distance distributions of $DNA_{2.1}$ (which contains a single 8-nt microhomology) revealed a $2.6 \pm 0.2$ $\mu m$ periodicity, consistent with the expectation that the dsDNA was captured at a single position on M13mp18.

The requirement for microhomology suggested that captured intermediates were retained through Watson-Crick pairing. This hypothesis predicts that the binding lifetime should scale with melting temperature ($T_m$), which was confirmed using substrates bearing 8-nt tracts of microhomology of

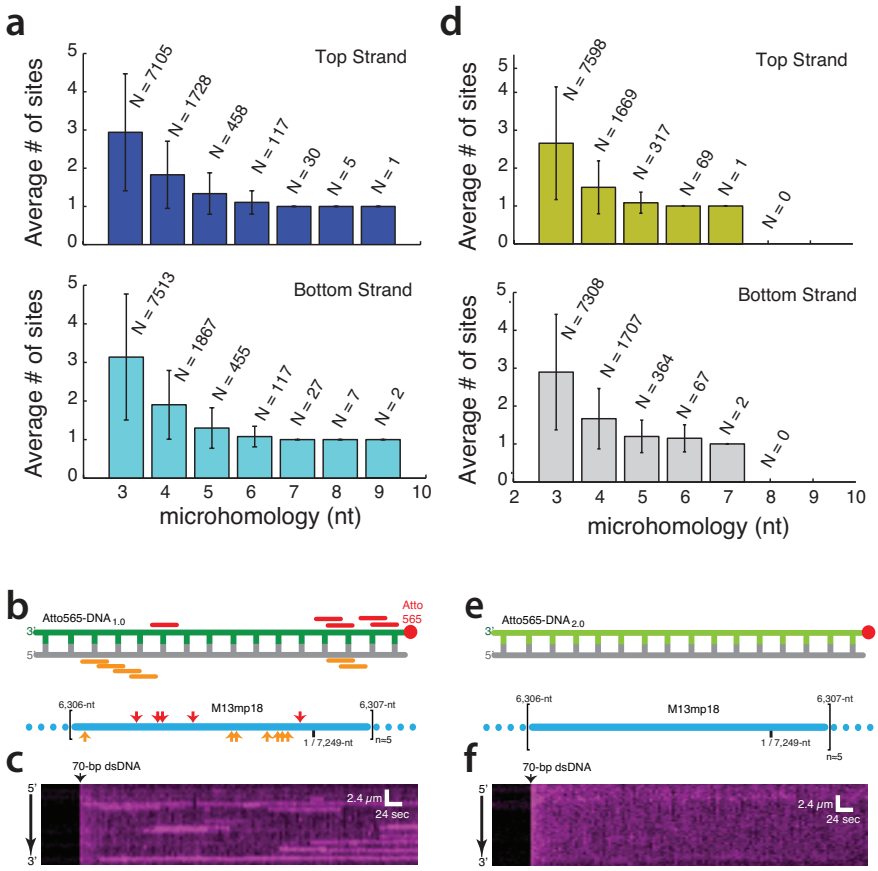


Figure 4.5: Stable capture of nonhomologous dsDNA. a, Analysis showing the total number (N) and the average number (±SD) for the given length of microhomology within each occupied 70-nt window along M13mp18. b, Positions of microhomology (≥ 8-nt) within Atto565-DNA$_{1.0}$ (color-coded bars indicate relative positions of microhomology within the dsDNA) and the schematic illustration showing the corresponding locations (indicated with color-coded arrowheads) of the tracts of microhomology along a single unit length M13mp18 ssDNA substrate (lower panel). Illustrations are not to scale. c, Kymograph showing binding of Atto565-DNA$_{1.0}$ to a single Rad51 PC; 100-ms frames were collected at 5-s intervals. d and e, Analysis (d) and schematic (e) of a re-designed 70-bp dsDNA (Atto565-DNA$_{2.0}$) lacking 8-nt tracts of microhomology. Error bars represent SD. f, Kymograph showing Atto565-DNA$_{2.0}$ incubated with a single PC; data were collected as in c.

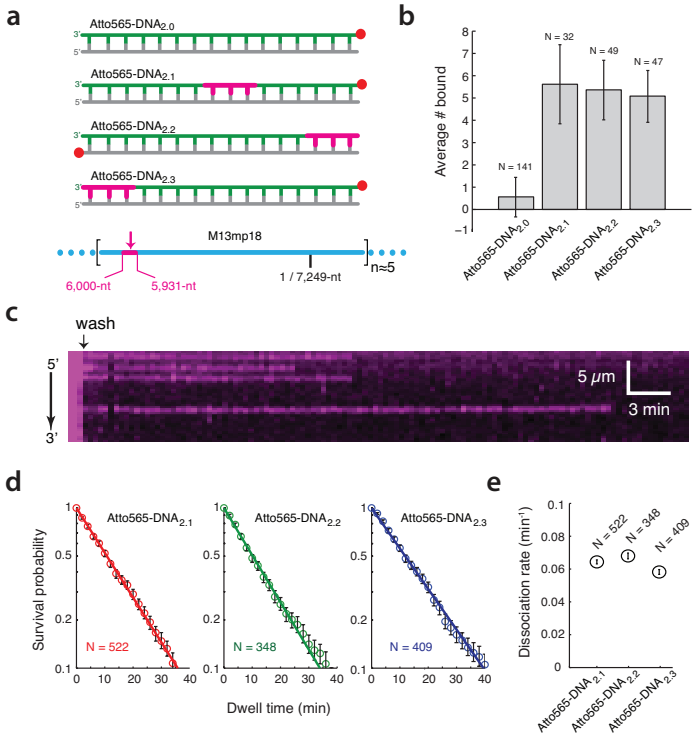


Figure 4.6: 8-nt Tracts of microhomology are sufficient for dsDNA capture. a, Substrates bearing a single 8-nt tract of microhomology (highlighted in magenta) at different positions within the 70-bp dsDNA. b, Average number of Atto565-dsDNA bound per PC. N corresponds to the number of PCs counted. Error bars represent SD. c, Kymograph showing an example of Atto565-DNA$_{2.1}$ dissociating from a PC. d and e, Survival probability plots (d) and dissociation rates (e) for each substrate.

varying AT-content (Fig. 4.7c). Moreover, the change in free energy ($\Delta\Delta G^{\ddagger}$) scaled with hydrogen bonding potential, with each hydrogen bond contributing $\sim 0.14\ k_bT$ to the binding of the 8-nt motif (Fig.). The modest contribution to overall stability for each hydrogen bond was consistent with the requirement that the homology search be driven by thermal fluctuations, and supports the notion that stretch-induced disruption of base stacking markedly destabilizes the Watson-Crick base pairs relative to B-DNA [Chen *et al.*, 2008].

We also tested how microhomology length influenced dsDNA capture (Fig. 4.7d). We were unable to detect any stable binding intermediates in these assays when the 8-nt tract of microhomology was decreased to 7-nt (DNA$_{2.6}$), in agreement with the conclusion that 8-nts of microhomology was necessary for stable dsDNA capture (Fig. 4.7d, see below). In contrast, increasing the 8-nt tract of microhomology to 9-nt reduced the dissociation rate, and additional length increases resulted in step-wise reductions in the dissociation rates in precise 3-nt increments (Fig. 4.7d, see below).

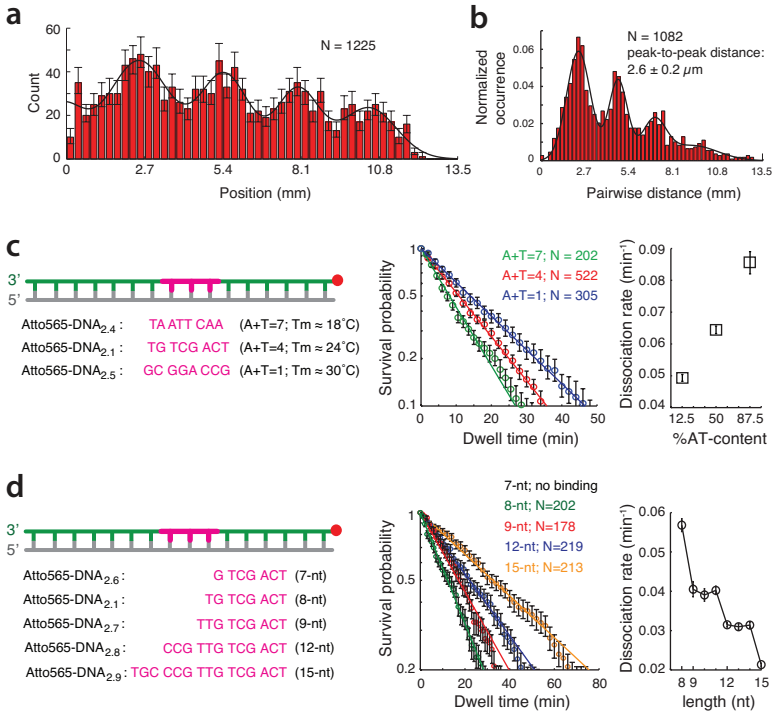


Figure 4.7: 8-nt Tracts of microhomology are sufficient for dsDNA capture. a and b, Binding distribution (a) and pairwise distance distribution (b) for Atto565-DNA$_{2.1}$. c, Design, survival probability plots, and dissociation rates for DNA substrates bearing a single 8-nt tract of microhomology with varying AT-content. d, Design, survival probability plots, and dissociation rates for substrates bearing 8- to 15-nts of microhomology; sequences and survival probability curves for the 10-nt, 11-nt, 13-nt, and 14-nt substrates are omitted for clarity. There was no detectable binding activity for Atto565-DNA$_{2.6}$ in these assays.

The microhomology requirement, the periodic binding patterns, and the influence of AT-content and microhomology length all suggested that the bound intermediates were maintained through Watson-Crick interactions.

## 4.5 Transient dsDNA sampling by Rad51

Rad51 did not stably capture dsDNA lacking 8-nt tracts of microhomology, but it must be sampling these molecules for homology.

Even microhomology-bearing dsDNA must in most instances be transiently sampled, because the vast majority of bimolecular encounters will occur at nonhomologous sites. Therefore, the 70-bp substrates used in our assays offered the unique potential for exploring how Rad51 coated ssDNA samples and rejects dsDNA while searching for homology. We detected these transient

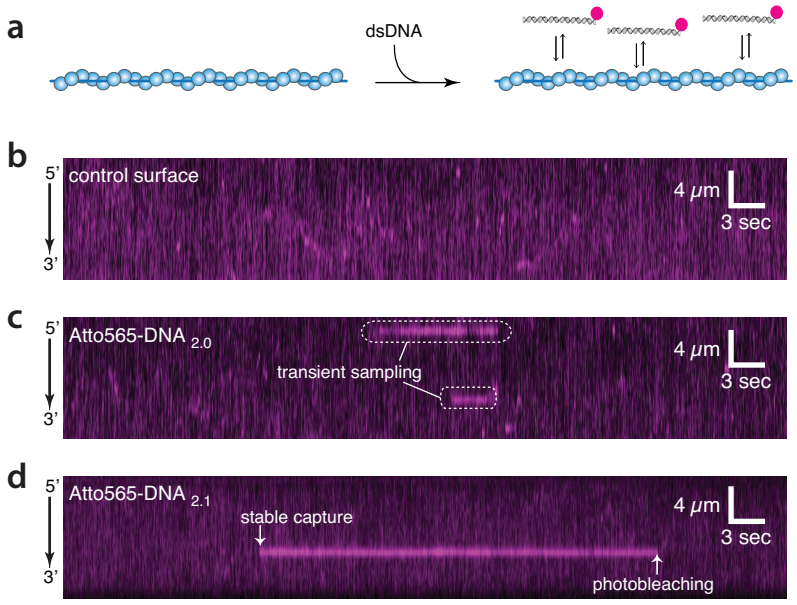


Figure 4.8: Transient sampling dsDNA lacking microhomology. a, Strategy for visualizing dsDNA sampling at 60-ms resolution. b,c and d, Kymographs showing (b) Atto565-DNA$_{1.0}$ in the absence of the PC (control surface), and Rad51 PCs sampling (c) Atto565-DNA$_{2.0}$ or (d) Atto565-DNA$_{2.1}$.

intermediates by visualizing reactions in real-time at 60-millisecond ($msec$) resolution (Fig. 4.8a-d). Remarkably, the survival probabilities of substrates lacking $\geq$ 8-nt of microhomology (DNA$_{2.0}$) did not decay exponentially, but rather scaled as a power-law, with 50% of the molecules dissociating within 0.54 seconds (Fig. 4.9a), even though this substrate harbors numerous $\geq$ 7-nt tracts of microhomology (Fig. 4.5). Power-law dependence was also observed over short time regimes for a substrate bearing a single 8-nt tract of microhomology (DNA$_{2.1}$), whereas the lifetimes were limited by photo-bleaching at longer time scales, as expected (Fig. 4.9a).

We next conducted real-time measurements with DNA$_{2.6}$, which differs from DNA$_{2.1}$ by just a single nucleotide (Fig. 4.7d); as indicated above, this single nucleotide change reduces the 8-nt tract of microhomology to 7-nts, and abolishes stable capture of this substrate by Rad51. Instead, DNA$_{2.6}$ exhibits power-law distributed dissociation kinetics with 50% of the molecules dissociating within 0.82 seconds (Fig. 4.9a). These findings indicate that all the dsDNA substrates were initially sampled through the same pathway, as revealed by its characteristic power-law dependence, but only substrates bearing 8-nts of microhomology transitioned into the long-lived state.

A crucial implication of this power-law behavior is that the transient sampling events cannot be ascribed to a single conformational state, but rather reflect the existence of a highly diverse

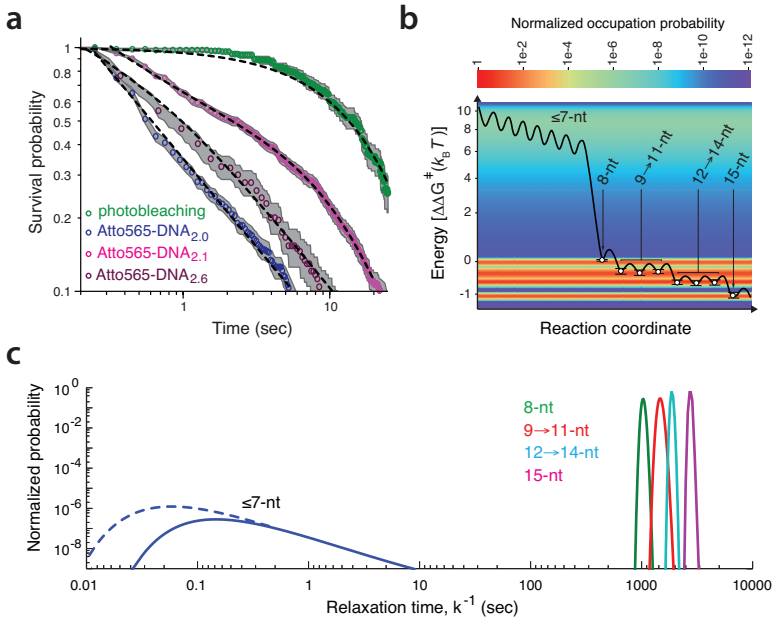


Figure 4.9: Analysis of transient dsDNA sampling. a, Log-log plot revealing the power-law dependence of the transient search intermediates. Dashed lines represent a single exponential fit to the photo-bleaching data, power-law fits for Atto565-DNA$_{2.0}$ and Atto565-DNA$_{2.6}$, and combination of a power-law and single exponential fit for Atto565-DNA$_{2.1}$. b, Energy landscape describing dsDNA sampling and strand invasion by Rad51. The heat map and open circles ($\pm$SD) represent calculated values for normalized occupation probability and $\Delta\Delta G^{\ddagger}$ values based on experimental data, respectively. The black line is a representation of the landscape and the heights of the energy barriers between states is for illustrative purposes only. Additional details are presented in the main text and Appendix B. c, Distribution of kinetic rates for dsDNA sampling and capture by Rad51. Solid lines represent experimental data and the dashed line reflects intermediates that are sampled too rapidly to be detected (Appendix B).

ensemble of different states [Austin *et al.*, 1975; Frauenfelder *et al.*, 1991]. The physical basis for this power-law dependence is readily understood given the vast number of potential intermediates. If one assumes a recognition model involves a 1-nt target size (Appendix A), then a 70-bp dsDNA can be misaligned with a total of $453,652$ distinct sites on M13mp18, each of which can give rise to energetically distinct states based on differences in sequence composition (Chpt 1, Appendix B). These considerations highlight the tremendous challenge faced during the homology search, even within our simplified experimental system.

### 4.5.1 Energy landscape for dsDNA sampling and strand invasion by Rad51

Our data provide a free energy landscape describing dsDNA sampling and strand invasion by Rad51 (Fig. 4.9b and Appendix B). The initial search process is characterized by transient sampling intermediates that encompass a broad distribution of energetic states, which could reflect thousands of distinct complexes as Rad51 interrogates different regions of dsDNA for homology (Fig. 4.9b,c). Recognition of an 8-nt tract of microhomology results in a $8.2\ k_bT$ drop in the energy barrier ($\Delta\Delta G^{\ddagger}$), and gives rise to a $\geq 4$ order-of-magnitude decrease in dissociation kinetics, providing a robust length-based mechanism for kinetically discriminating against sequences that are unlikely to be fully homologous (Fig. 4.9b,c). This length-based microhomology recognition event is the single largest change in the energy landscape, and most likely reflects a conformational transition within the Rad51-ssDNA-dsDNA ternary complex the exact nature of which remains to be explored.

The finding that recognition of an 8-nt tract (as opposed to either 6- or 9-nts) coincided with the largest drop in free energy was not anticipated given that the ssDNA within the PC is organized into base triplets [Chen *et al.*, 2008]. Following microhomology capture, Rad51 can probe the flanking the DNA for additional homology while attempting strand invasion. Pairing with a $9^{\text{th}}$nt results in an additional $\sim 0.4\ k_bT$ reduction in free energy, revealing that incorporation of the $9^{\text{th}}$nt enabled more stable engagement of the $3^{\text{rd}}$ base triplet (Fig. 4.9b,c). All subsequent reductions in free energy occurred in precise 3-nt increments, suggesting that the ssDNA bound by Rad51 was organized into base triplets (Fig. 4.9b,c), as observed for *E. coli* RecA [Chen *et al.*, 2008], and that the quantized reductions in binding energy were the functional consequence of this triplet organization. Together, these findings also indicate that capture of the first 8-nt tract of microhomology is energetically and mechanistically distinct from the subsequent reactions involved in strand invasion, suggesting

that recognition of the $9^{\text{th}}$nt demarks the beginning of actual strand exchange, allowing subsequent reactions to take place in 3-nt steps (Fig.).

## 4.6 Extensive sliding or intersegmental transfer do not contribute to microhomology capture

Prior smFRET measurements suggested that 1-dimensional (1D) sliding might contribute to DNA alignment by RecA over short distances [Ragunathan *et al.*, 2013]. However, in agreement with prior biochemical studies [Adzuma, 1998], our data revealed no detectable evidence of 1D sliding of the dsDNA, although we do not rule out the possibility that sliding might take place over short distances ($\leq 270\ nm$ i.e. submicroscopic). Furthermore, the kinetic preference discrimination at 8nt sites also argues against an appreciable role for a 1D search, which would greatly limit the lifetime of the dsDNA at sites $\leq$ 7-nt in length.

Other studies have shown that sequence alignment by RecA involves intersegmental transfer (Appendix A) [Forget and Kowalczykowski, 2012]. We found no evidence that the 70-bp dsDNA molecules moved by intersegmental transfer [Qi *et al.*, 2015]; however, we emphasize that these results do not argue against intersegmental transfer as a crucial component of the Rad51 homology search (see below), rather, our findings are as anticipated for a search entity engaging a single unit-length binding element.

## 4.7 Facilitated exchange promotes turnover of dsDNA bound to the presynaptic complex

Stand invasion in *S. cerevisiae* can be detected within approximately $\sim$ 10-60 minutes of DSB formation, so the search for homology must be completed within this time window. However, 8-nts is insufficient to define a sequence as statistically unique within the *S. cerevisiae* genome, and it is difficult to envision how recombination could be executed on a relevant time scale if the PC became kinetically trapped every time it encountered a $\geq$ 8-nt tract of microhomology. This implies the existence of unknown mechanisms for disrupting these intermediates.

One possibility is that specific enzymes might disrupt intermediates involving short micro-

homology motifs; there are numerous helicases/translocases with the potential to fulfill such a role (e.g. Mph1, Srs2, Sgs1, Rdh54 and/or Rad54) [Heyer *et al.*, 2010; Renkawitz *et al.*, 2014; Filippo *et al.*, 2008]. We do not exclude the possibility that these or other proteins may contribute to the homology search, perhaps by promoting the turnover of Rad51 bound to incorrect 8-nt tracts of microhomology. However, Rad51, like many other Rad51/RecA family members, can catalyze strand exchange *in vitro* with no need for these accessory factors despite the potential for sequence misalignment at any of the hundreds of 8-nt microhomology motifs present in the plasmids typically used for these assays, underscoring that the ability to search for homology is an intrinsic property of Rad51/RecA proteins.

Therefore we asked whether a more fundamental mechanism(s) might promote dissolution of microhomology-bound intermediates. It has recently been recognized that facilitated exchange can contribute to disruption of protein-nucleic acid interactions [Gibb *et al.*, 2014; Graham *et al.*, 2011; Sing *et al.*, 2014], and may be a general but underappreciated phenomenon that influences macromolecular interactions under crowded physiological settings. Facilitated exchange reflects the existence of microscopically dissociated intermediates, which only undergo macroscopic dissociation when competing interactions arise from other molecules in the local environment. These concepts are readily extended to reactions involving the PC.

We considered the possibility that dissolution of intermediates arising from captured microhomology might be promoted by facilitated exchange with other dsDNA molecules. The hypothesis that DNA might disrupt search intermediates is intriguing given the high concentration of DNA within the nucleus and the potential ubiquity of such a mechanism. To test this hypothesis we asked whether dsDNA bound to the PCs was released more rapidly into free solution when challenged with free competitor dsDNA. For this, fluorescent $DNA_{1.0}$ was pre-bound to the PCs, and the reactions were chased with unlabeled $DNA_{1.0}$ (Fig. 4.10). Remarkably, the competitor chase was able to accelerate the macroscopic dissociation rate up to 3-fold (Fig. 4.10b-d). We conclude that free dsDNA can accelerate turnover of dsDNA bound to the PCs consistent with a mechanism involving facilitated exchange.

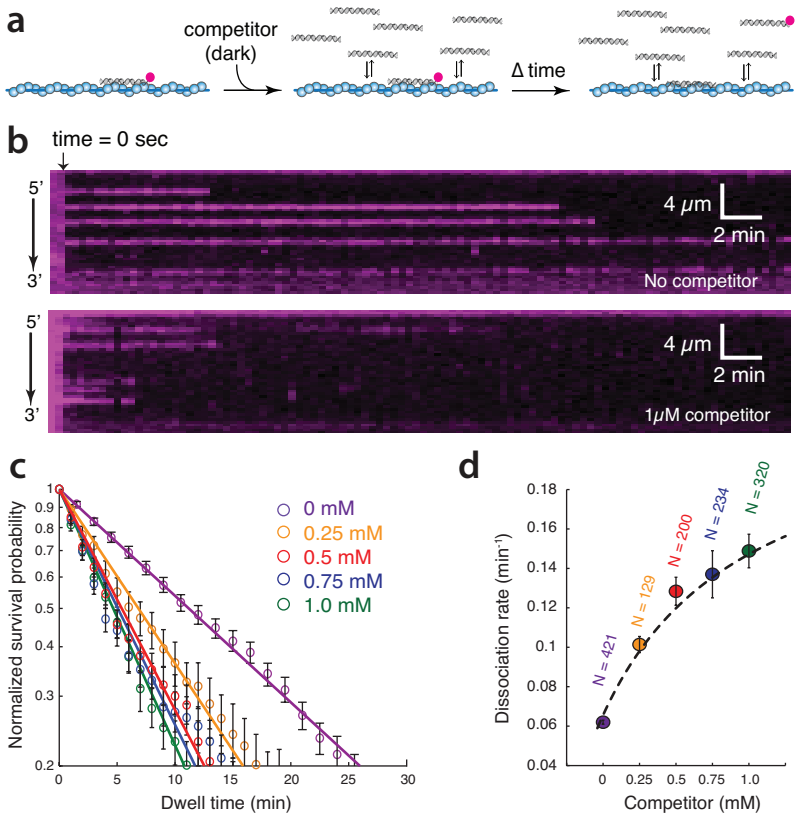


Figure 4.10: Facilitated exchange of captured intermediates. a, Strategy for quantifying dsDNA dissociation after injection of unlabeled competing DNA. b, Kymographs showing the dissociation of Atto565-$DNA_{1.0}$ from the Rad51 PC in the absence (upper panel) and presence (lower panel) of unlabeled competitor $DNA_{1.0}$. c and d, Dwell time analysis of dissociation kinetics (c) and dissociation rates (d) for Atto565-$DNA_{1.0}$ when chased with varying concentrations of dark $DNA_{1.0}$. The dissociation rates as a function of dark competitor are fit to a Hill-type curve with an intercept conveying the reaction in the absence of competitor. N corresponds to the number of Atto565-DNA molecules measured. Error bars represent SD.

### 4.7.1 Sequence and length requirements for facilitated exchange

PCs capture dsDNA through 8-nt tracts of microhomology, implying that facilitated exchange might involve overlapping tracts of microhomology. If correct, then facilitated exchange should only occur with competitor substrates bearing identical 8-nt tracts of microhomology. Indeed, reactions with two different Atto565-labeled substrates and series of competitors confirmed that facilitated exchange required overlapping tracts of microhomology (Fig. 4.11a-d), and exchange was abolished if the competing microhomology was shifted by even a single nucleotide in either direction (not shown).

We next tested how facilitated exchange was influenced by microhomology length. The in-

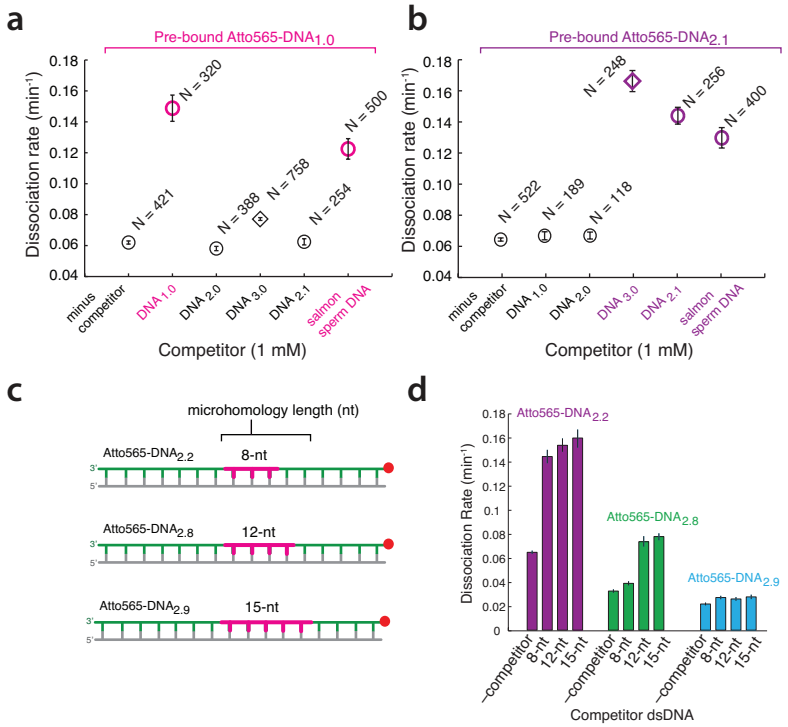


Figure 4.11: (Length and overlap requirements for facilitated exchange. a and b, Dissociation rates for Atto565-DNA$_{1.0}$ (a) and Atto565-DNADNA$_{2.1}$ (b) when challenged with different competitor substrates ($1\mu M$ each), as indicated; like colors correspond to competitors bearing overlapping tracts of microhomology, competitors lacking overlapping microhomology are shown in black. N corresponds to the number of Atto565-DNA molecules measured. Error bars represent SD. c and d, Schematic (c) and corresponding (d) data for substrates used to test the influence of microhomology length and alignment on facilitated exchange.

creased stability of substrates bearing longer tracts of microhomology (see Fig. 4.7d) was reflected in the finding that shorter tracts of microhomology were more readily exchanged with longer tracts, whereas longer tracts of microhomology were more resistant to exchange with shorter tracts (Fig. 4.11d). Moreover, a 15-nt tract of microhomology was sufficient to render a bound substrate completely resistant to facilitated exchange. Together, these results demonstrate that facilitated exchange requires overlapping microhomology, indicate that once the PC has engaged a particular dsDNA it ignores substrates lacking overlapping microhomology, and suggest that facilitated exchange can lead to preferential association with longer microhomology motifs. These results also imply the existence of a length-based threshold of 15-nts as perhaps demarking the commitment to strand exchange; reversibility at this stage of the reaction would likely require the action of accessory proteins dedicated to dissolution of aberrant strand exchange intermediates [Heyer *et al.*, 2010;

Filippo *et al.*, 2008].

In addition to facilitated exchange, Atto565-labeled substrates bearing an 8-nt microhomology motif were also displaced from the PC when challenged with a fully homologous 70-bp substrate ($DNA_{3.0}$), but only if the fully homologous substrate overlapped in sequence with the bound dsDNA (Fig. 4.11a,b). This finding implies that the initiation of strand exchange with a fully homologous substrate anywhere along the PC would be sufficient to drive disruption of captured 8-nt tracts of microhomology located at adjacent positions along the PC, ensuring that stand invasion could progress unimpeded once homology was correctly identified.

## 4.8 Joint molecules made with fully homologous dsDNA resist disruption

The results presented above lead to four predictions for reactions involving homologous substrates: (*i*) initial sampling of the homologous substrate should exhibit power-law dependence over short time regimes; (*ii*) a fully homologous substrate should bind to all locations bearing $\geq$ 8-nt of microhomology; (*iii*) a captured homologous substrate should exhibit two categories of lifetimes corresponding to those molecules bound to microhomology motifs and those that are bound to the full region of homology; and (*iv*) the captured intermediates should be differentially affected when chased with competitor dsDNA.

We tested these predictions using a homologous 70-bp substrate ($DNA_{3.0}$); analysis of this substrate revealed $\geq$ 8-nt tracts of microhomology at 19 distinct sites on M13mp18 ssDNA (Fig. 4.12a). As anticipated, the initial sampling intermediates exhibited characteristic power-law behavior, reflecting the existence of a diverse ensemble of transient encounter complexes (Fig. 4.12b). Once captured, lifetime analysis of the bound dsDNA revealed the existence of two spatially distinct populations: shorter-lived intermediates, and longer-lived intermediates that displayed a periodic binding distribution as expected for the unique 70-nt region of homology (Fig. 4.12c,d). As predicted, only the shorter-lived intermediates were disrupted when challenged with competing dsDNA, whereas the longer-lived complexes were resistant to facilitated exchange (Fig. 4.12e,f). We conclude that Rad51 utilizes a length-based microhomology recognition mechanism even when presented with a fully homologous substrate and that reaction products generated through strand

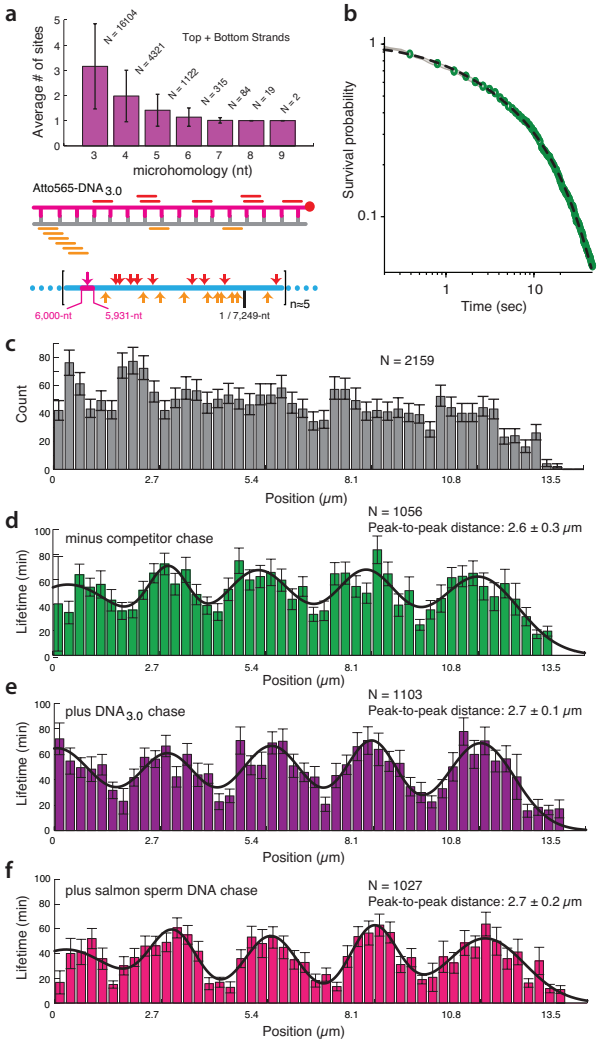


Figure 4.12: Sampling and Capture of a Fully Homologous Substrate. a, Microhomology analysis and schematic of the 70-bp homologous dsDNA substrate (Atto565-$DNA_{3.0}$) highlighting the 8-nt tracts of microhomology complementary to the M13mp18 ssDNA substrate. b, Power-law dependence of search intermediates observed with $DNA_{3.0}$. The dashed black line shows combination of a power-law and single exponential fit (to account for photo-bleaching) to the data. c, Observed binding distribution of Atto565-$DNA_{3.0}$ at time zero. d, Lifetime distribution of Atto565-$DNA_{3.0}$ in the absence of competitor dsDNA challenge. e and f, Lifetime distribution of Atto565-$DNA_{3.0}$ when challenged with either (e) $1\mu M$ $DNA_{3.0}$ or (f) $1\mu M$ salmon sperm DNA.

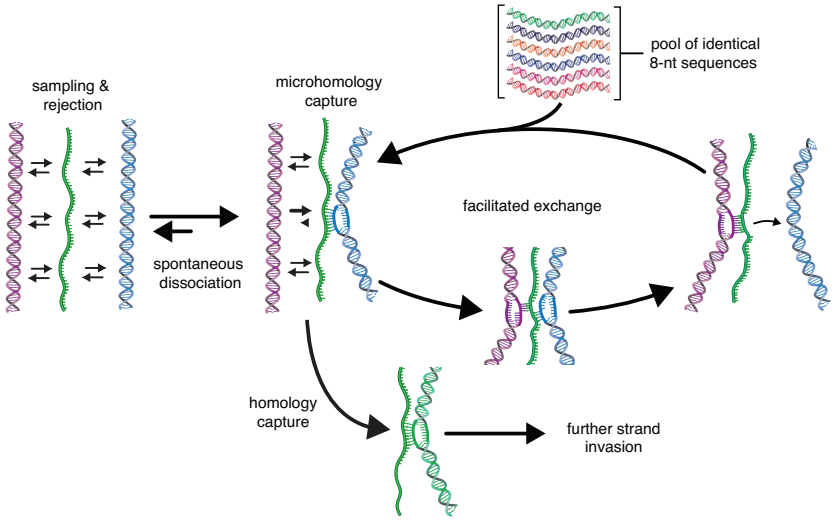


Figure 4.13: Sequence alignment model. Model depicting a homology search mechanism involving rapid sampling and rejection of DNA lacking microhomology, followed by eventual capture of an 8-nt tract of microhomology and facilitated exchange allowing for an iterative search through sequence space. Additional details are presented in the main text.

invasion of the fully homologous 70-bp substrate were highly stable.

## 4.9 Model for DNA sequence alignment during HR

Our results are unified in a model for how Rad51 aligns DNA sequences during HR (Fig. 4.13). For clarity, Figure 4.13 depicts a single interacting unit; we anticipate multiple unit-length interactions will occur throughout the PC, as expected for intersegmental transfer [Forget and Kowalczykowski, 2012]. We propose that Rad51-ssDNA filaments sample dsDNA in 8-nt increments and quickly reject any sequences lacking 8-nt tracts of contiguous microhomology. This stage of the reaction is characterized by a complex energetic landscape as Rad51 quickly explores a vast amount of sequence space. The presence of an 8-nt tract of microhomology allows dsDNA to be captured through Watson-Crick pairing, enabling Rad51 to probe the flanking duplex for additional complementarity while attempting more extensive strand exchange. If pairing with a $9^{\text{th}}$ nt is successful, then the resulting intermediates are rendered more stable by virtue of more extensive Watson-Crick pairing in precise 3-nt increments, eventually crossing a threshold ($\sim$ 15-nts) beyond which they are much less susceptible to either spontaneous dissociation or facilitated exchange. In contrast, if further strand invasion fails, then any search intermediates bound to incorrect 8-nt tracts of microhomology can be disrupted by either spontaneous dissociation or facilitated exchange. Alternatively, successful

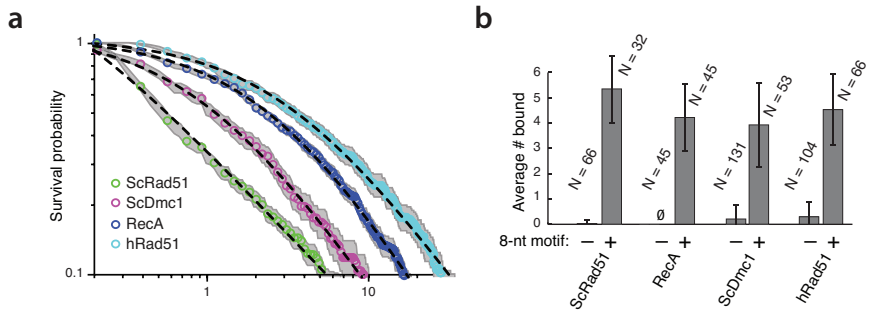


Figure 4.14: Transient sampling of PCs across distant relatives of the RecA gene family. a and b, Plots showing power-law behavior during dsDNA sampling (a) and microhomology-dependent binding (b) for E. coli RecA, hRad51, and *S. cerevisiae* Dmc1, Data presented for ScRad51 are reproduced from Figures 4.9 and 4.7 for comparison. The plus and minus 8-nt motif designations in (b) correspond to Atto565-DNA$_{2.1}$ and Atto565-DNA$_{2.0}$, respectively

capture of full homology anywhere along the length of the PC will also disrupt any existing search intermediates allowing unimpeded strand exchange.

This model also hints at a deeper understanding for how *E. coli* RecA might search for homology RecA can capture as little as 8-nt of homology [Hsieh *et al.*, 1992], and re-evaluation of the $1,762$-nt ssDNA and $48,502$-bp dsDNA sequences used to substantiate the RecA intersegmental transfer mechanism reveals a total of $2,089$ tracts of microhomology 8-nt in length [Forget and Kowalczykowski, 2012]. We suggest that RecA PCs may establish numerous points of contact with dsDNA through these short tracts of microhomology.

## 4.10 A conserved search mechanism for the Rad51/RecA recombinases

The salient feature of our model for the homology search is that it kinetically minimizes nonproductive interactions with short ($\leq$ 7-nt) dsDNA sequences that have little chance of being the fully homologous target. This assertion is based upon two key features of *S. cerevisiae* Rad51: (*i*) rapid sampling and rejection of dsDNA lacking microhomology motifs through a mechanism characterized by its distinctive power-law dependence; and (*ii*) length-specific kinetic selection of microhomology tracts (Fig. 4.13). We next asked whether human Rad51 (hRad51), *S. cerevisiae* Dmc1, and *E. coli* RecA behaved similarly.

Remarkably, all three proteins displayed power-law behavior while transiently sampling dsDNA

that lacked 8-nt microhomology motifs, with 50% of the sampling events dissociating before 3.5, 1.1, and 2.5 seconds for hRad51, ScDmc1, and RecA, respectively (Fig. 4.14a); and all three proteins preferentially captured substrates harboring 8-nts of microhomology (Fig. 4.14b). These results revealed that recognition of an 8-nt microhomology motif coincided with $\sim$ 6.1, $\sim$ 6.5, and $\sim$ 6.2 $k_bT$ ($\Delta\Delta G^{\ddagger}$) reductions in the free energy landscapes for hRad51, ScDmc1, and RecA, respectively, reflecting the drastic differences in affinity for dsDNA with and without an 8-nt tract of microhomology. These findings suggest that the ability to interrogate dsDNA through a mechanism involving length-specific microhomology recognition emerged early in the evolutionary history of the RAD51/recA gene family.

## 4.11 Discussion

The genetic transactions that take place during HR are governed by the physicochemical properties of the macromolecules that promote these reactions, and a full appreciation for the elegance of DNA recombination requires a detailed understanding of the underlying mechanistic principles. Our work suggests that length-specific kinetic selection of 8-nt microhomology motifs underlies the intrinsic ability of the Rad51/RecA recombinases to efficiently align homologous sequences, and mechanistically distinguishes this process from the 3-nt steps that take place during strand exchange. The use of microhomology motifs as recognition elements has crucial implications for understanding how DNA sequences are aligned during HR.

### 4.11.1 Microhomology recognition minimizes search complexity

The advantages of a length-based microhomology recognition can be illustrated by considering its influence on the amount of sequence space that must be interrogated during the homology search. The information that must be processed in order to align two homologous sequences can be quantitatively described as search complexity, which reflects the number of sites a searching entity must visit within the genome while attempting to locate a unique sequence (Fig. 4.13). A full treatment of search complexity is presented in Appendix B; here we highlight key concepts and their relevance to HR. In brief, search complexity can be defined as:

$$complexity(bp \cdot genome^{-1}) = \frac{2n}{l}(o-n-1)(l-n-1)4^{-n} \tag{4.1}$$

; where $n$ is the length of microhomology used during the search, $l$ is the length of the genome, and $o$ is PC length. Any value for search complexity $\geq 1.0\ bp{\cdot}genome^{-1}$ indicates that the PC needs to, on average, sample more than a genome equivalents worth of sites before locating homology; e.g., for an organism with a genome of one million base pairs, a search complexity of $1bp \cdot genome^{-1}$ indicates that the PC would, on average, need to sample and amount of DNA equal to 100% of the genome (i.e., one million base pairs) before locating homology. Values $> 1.0\ bp \cdot genome^{-1}$ reflect a search that is accelerated relative to genome size; e.g., a search complexity of $0.1\ bp \cdot genome^{-1}$ indicates that only one tenth of the genome would need to be sampled to locate homology. The above considers the target search as purely three dimensional, and does not include potential reductions in complexity due to facilitating mechanisms, because our data does not suggest a critical role for these mechanisms in the homology search.

The benefits of microhomology recognition can now be explored by considering the impact on search complexity (Fig. 4.15a-d). The most important revelation from this analysis is that search complexity decreases exponentially with the minimal length of microhomology necessary for dsDNA recognition. The source of this exponential dependence is evident given that for any genome short sequences will always have many exact matches, while longer sequences will always have fewer exact matches. For example, any defined 3-nt motif occurs on average once every 639-bp, and there would be $\sim 377,229$ such sequences in the *S. cerevisiae* genome (Fig. 4.15d). In contrast, 8-nt motifs will on average occur just once every $65,536$-bp, and there would only be $\sim 762$ identical 8-nt motifs in the yeast genome (corresponding to an *in vivo* concentration of $\sim 0.3\ \mu M$ for any given 8-mer). As a consequence, a search utilizing a single 8-nt motif would only need to interrogate just $\sim 0.01\%$ of the genome to locate the homologous target, and the vast majority of the genome could be kinetically ignored. Indeed, a homology search involving length-specific recognition of 8-nt motifs, while kinetically minimizing interactions with shorter sequence motifs, would effectively eliminate $\geq 99.9\%$ of the genome for species ranging from *E. coli* to humans.

Genetic and physical measures of the ssDNA overhangs generated during DSB repair suggest that *S. cerevisiae* PCs are $\sim$ 100-4,000 nts in length, and it is especially informative to consider how search complexity varies within this length regime. For a search utilizing 8-nt tracts of microhomology, a 100-nt PC would only need to process information content corresponding to one hundredth of the genome (Fig. 4.15c, inset), a $4,000$-nt PC would only need to sample half of the

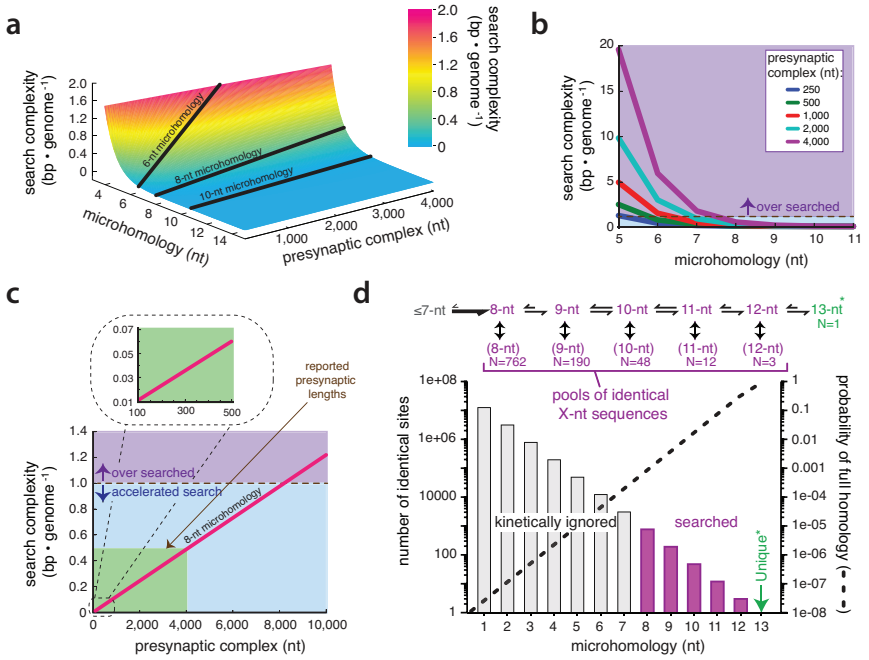


Figure 4.15: Calculations of search complexity. a, Surface plot showing how search complexity varies with PC length and the length of microhomology necessary for dsDNA interrogation. b, Variation in search complexity for search models employing different lengths of microhomology, as indicated. c, Relationship between search complexity and PC length for recognition involving 8-nt of microhomology. The green shaded region encompasses length estimates for S. cerevisiae PCs. d, Fraction of the *S. cerevisiae* genome that can be kinetically ignored when employing a length-dependent search mechanism based on recognition of 8-nt motifs.

genome (Fig. 4.15c), and search complexity would not enter the over searched regime until PC length exceeded $\sim 8,000$-nt (Fig. 4.15c). In contrast, if one assumes a model without microhomology recognition (i.e. $n = 1$), then PCs ranging from 100-4,000-nt in length might have to process information equivalent to 2,500-100,000% of the genome. These considerations illustrate how simply subdividing the search into length-based microhomology recognition elements can drastically reduce the time necessary to align homologous sequences. We will return to this idea below, and in the final chapter.

### 4.11.2 Physiological implications for HR and DSB repair

The reductionist treatment of search complexity presented above excludes any potential effects of accessory factors, chromatin structural proteins, chromosome organization, etc. Interpretation of our results within the context of these physiological realities leads to several important insights and predictions. First, end resection, PC assembly, and the homology search are often presented as distinct stages of DSB repair. However, there is no reason to believe that these reactions are completely uncoupled, and the relative timing of these events dictates how much information must be processed during the homology search. Our results predict a substantial benefit to beginning the homology search as soon as possible after initiating DSB resection (Fig. 4.15a-c).

Second, for mechanisms involving length-dependent microhomology recognition, the fractional reduction in search complexity is the same regardless of genome size. Although longer recognition motifs offer the potential for further reductions in search complexity, this would compromise reversibility because of the greater enthalpic penalty incurred for disruption of a larger binding surface, which could ultimately lead to misalignment of DNA sequences trapped in local minima. Moreover, assuming a randomized nucleotide distribution, the length required to statistically define a given sequence as unique does not vary drastically across species. For instance, average lengths of just 12, 13, and 17 nucleotides are sufficient to uniquely define most sequences within the *E. coli*, *S. cerevisiae*, and human genomes, respectively (Appendix B). These considerations imply that there may be little or no evolutionary pressure to utilize longer tracts of microhomology to compensate for variations in genome size. Notably, real genomes contain repetitive sequences and other regions of low sequence complexity (e.g. rDNA and tRNA genes, transposons, centromeres, telomeres, etc.), and such regions would require longer sequences to define uniqueness, or else may

suffer from a greater potential for misalignment during HR. Interestingly, recombination within these regions is often suppressed and/or otherwise tightly regulated [Eckert-Boulet and Lisby, 2010; Eckert-Boulet and Lisby, 2009; Pan *et al.*, 2011; Sasaki *et al.*, 2010], perhaps reflecting in part the unique challenges faced by the recombination machinery in these regions of low sequence complexity.

Third, PC organization affects the amount of information that must be processed during the homology search. The preceding discussion assumes a contiguous PC consisting of all possible overlapping 8-nt units (Appendix B). However, search complexity declines by an entire order of magnitude if the PC is segregated into non-overlapping 8-nt sections, and intermediate subdivisions are similarly beneficial (Appendix B). It is not known whether PCs *in vivo* are comprised of uninterrupted Rad51/RecA filaments, or whether they contain protein-free gaps and/or other physical discontinuities (e.g. other HR proteins). Our results suggest some proteins could promote HR by segregating Rad51/RecA filaments into non-overlapping functional units.

Fourth, once the PC has engaged a particular 8-nt tract of microhomology it can undergo exchange with other regions of dsDNA bearing the same microhomology, but resists exchange with unrelated sequences. Moreover, shorter tracts of microhomology are more readily exchanged with longer tracts, reflecting the higher stability of intermediates held together by longer tracts of Watson-Crick pairing. Preferential exchange with longer tracts of microhomology may yield a hierarchy of increasingly stable intermediates, which might in turn funnel the PC through progressively smaller pools of sequences leading to the homologous target (Fig. 4.13).

Fifth, compartmentalization of the search through either spatial organization or steric occlusion will decrease search complexity linearly with respect to the amount of sequence accessible for interrogation. Benefits are readily envisaged if homologous chromosomes are physically juxtaposed, as anticipated for sister chromatids immediately following DNA replication, and accumulating evidence suggests that homologous sequences also have a greater probability of being juxtaposed at other points in the cell cycle [Barzel and Kupiec, 2008; Gladyshev and Kleckner, 2014; Weiner and Kleckner, 1994]. Similarly, restricting search intermediates to the linker DNA between nucleosomes could reduce search complexity by $\sim$ 75% based on nucleosome occupancy of the *S. cerevisiae* genome.

# Chapter 5

# Conclusion

## 5.1 Summary

Over the course of this work, we have looked at the target search mechanisms of three biologically diverse proteins or protein complexes, hailing from organisms spanning all of life, from bacteria to humans. We began by pointing out the unenviable odds of target searches that rely on random occurrences, and wondered how it is that life persists, when such a ubiquitous element of DNA-based reactions is so statistically improbable. Driven by the realization that cells are continuously playing, and apparently winning, this cellular lottery, we sought to deepen our understanding of how proteins find their targets in DNA.

In Chapter 1, we considered the implications of the physical nature of both DNA and DNA-binding proteins, leading to a model that could explain how cells might solve the target search problem. This model, facilitated diffusion, was first described by Berg and von Hippel, and contends that the search is accelerated by the addition of local effects on concentration due to non-specific binding [Berg and Blomberg, 1976; **?**]. Importantly, all protein-DNA association processes are facilitated to some extent, because it is unphysical for a site-specific DNA-binding protein to lack non-specific binding activity, and it is this latter activity that provides the basis of facilitation. Accordingly, we presented two examples, the *lac* repressor and Msh2-6, where the mechanics of facilitation was a reality for biological macromolecules. However, we also examined several examples of target searches, which were devoid of extensive use of facilitating mechanisms. We then set out to answer how the cell solves these non-facilitated searches, and to learn what options the cell has

at its disposal to drive proteins toward their targets when facilitating effects no longer hold sway over the global target search.

The first non-facilitated search we looked at was the promoter search of *E. coli* RNA polymerase. From this example, we revealed the effect of concentration on searching molecules, and questioned whether or not the cellular pressure we surmised as the motivation for optimizing target searches, was an actuality for abundant proteins. At least for RNAP, it seemed, that any pressure on the target search may have been alleviated by higher protein numbers or didn't exist in the first place, resulting in a target search deficient in facilitating mechanisms; A fact we observed as a modest acceleration of the target search from facilitating effects only present at vanishingly small protein numbers.

From these studies on RNAP, we were able to determine, for the first time, the effective target size of a DNA-binding protein. Given the results presented in latter chapters, it is compelling to consider again how RNAP might be locating promoters. RNAP's job is to locate a promoter, and then open a bubble in DNA. This second DNA melting step requires the transduction of a great deal of thermal energy into the system, and it is hard to imagine that the cell would spend this energy whimsically. Our data suggests a mechanism for RNAP to avoid wasteful interaction with non-promoter DNA. We show that RNAP first looks for a small target, comprising an effective surface area comparable to $\sim$ 3 bases, after which, it transitions into the closed complex, and finally, the open complex. So it seems that instead of looking for full promoter sequences straight away, where RNAP can be trapped by irreversible binding at incorrect locations, RNAP first looks for short promoter-like sequences. Then, only once RNAP finds a short DNA sequence it likes, does it commit to melting the DNA.

The parallels between this revised interpretation of RNAP's promoter search and how CRISPR-Cas proteins locate protospacers in foreign DNA are apparent. In Chapter 2, we considered the target search of Cas9 from *S. pyogenes* and Cascade from *E. coli*, and arrived at a similar model. Specifically, Both Cas9 and Cascade utilized a small sequence element, called a PAM, to direct their respective target searches. We showed that Cas9 preferentially binds to PAM sequences and rapidly rejects non-PAM DNA, and only after binding to a PAM, does Cas9 commit to melting the DNA. In *E. coli*, we saw that Cascade also used PAMs to help it efficiently find protospacers. But, interestingly, Cascade relinquished some PAM binding specificity for the ability to recognize

mutated targets, and adapt to an ever-changing viral environment. Generally, for both systems, the target search proceeded through an initial, weakly bound, state, where discrimination between PAM and non-PAM DNA could be maximized, before transitioning into a more stable interaction, affording the CRISPR-Cas complexes a rapid path to target binding.

In the last chapter, this strategy of separating a search process into two pieces was employed by Rad51-ssDNA filaments to search for homologous dsDNA. The major difference between the previous examples and Rad51s search is the use of a fixed sequence motif; e.g., Cas9 looks for a fixed 3-nt motif, whereas Rad51 looks for variable 8-nt motifs. The rest of the Rad51 story is a familiar one; Rad51 searches for homology by first looking for a small portion of its target before testing the flanking DNA for further base matching. The difference in stability for substrates bearing $\leq$ 7-nt versus 8-nt of microhomology minimizes off-target interactions, ensuring that Rad51 spends most of the search interrogating sequences that already have a high probability of being a homologous target.

## 5.2 Final remarks

Given the striking similarities between each of the target searches presented here, it is attractive to speculate a general search strategy that may be common to many DNA binding proteins. This simple model, which we have named a reduced complexity (RC) search, makes two assertions. First, initial binding occurs through limited interactions, constituting only a portion of the complete binding interaction. Second, the efficiency of the RC search rests on the strength of a kinetic disparity that selects for a subset of these interactions for further interrogation.

Presented this way, the second assertion seems like a restatement of specific and non-specific binding. In Chapter 1, we discussed how the characteristics of specific binding give rise to an energetic landscape where lower energies coincide with the locations of specific binding sites; that is, a kinetic disparity between specific and non-specific binding. It is helpful to think about the initial binding in our model as specific binding because it has the same physical origins and characteristics. The key difference, however, is in the extent of the interaction, and therefore we offer a third assertion: the initial encounter with the DNA in an RC search should be a weak interaction.

RC searches direct proteins to their targets by minimizing off target interactions. In the toy

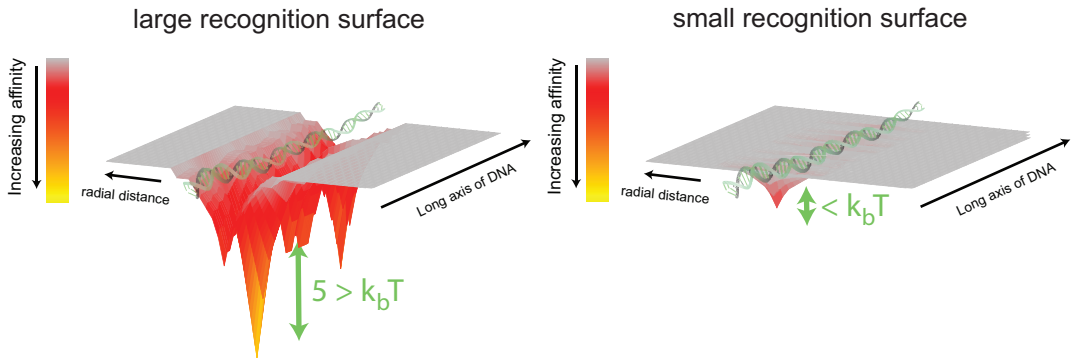


Figure 5.1: Hypothetical protein-DNA interaction landscapes as a function of binding surface. a, The hypothetical surface for a large binding interface, i.e., the *lac* repressor. Green arrow represents the standard deviation of the energetic landscape. b, Surface as in a, but for a small binding interface.

model from Chapter 1, we reduced protein-DNA binding to a set of individual interactions, which the protein tests against the DNA at each site to define the binding energy. Consider how that landscape varies as we change the number of interactions in the protein-DNA interface (Fig. 5.1). When several interactions are involved in binding, specific sites are bound very tightly, even with a modest number of interactions, and the cumulative binding energy is sufficient to define uniqueness [Slutsky and Mirny, 2004]. However, this specificity comes at the price of a very rough energetic terrain, comprised of several local minima (Fig. 5.1a). Alternatively, when the number of interactions in the protein-DNA interface is small, binding at specific sites cannot be terribly strong, and as a consequence there are several pseudo-targets that have to be investigated. But, as the specificity diminishes, so does binding at off-target sites (Fig. 5.1b).

For example, recall the *lac* repressor and its 16 interactions with its operator sequence. If a single one of these interactions is incorrect, the energetic change is very small ($\Delta G < 0.4\ k_bT$) relative to the total binding energy ($K_D \approx 5E^{-10}M$) [Kalodimos *et al.*, 2002]. Alternatively, PAM binding comprises a handful of interactions [Anders *et al.*, 2014]; yet, If one of these interactions is disturbed, it constitutes a greater fraction of the overall binding energy, and, as we have shown, fully disrupts the protein-DNA interaction (Fig. 3.9).

So then how does the cell get the best of both cases? How can it minimize off-target interactions *and* conserve strong binding at target sites? Notably, this problem has been considered previously, but within the framework of the facilitated diffusion model [Slutsky and Mirny, 2004]. Sliding along the DNA is a major part of why facilitated diffusion works; yet, sliding also requires the binding

energy between the DNA and protein to be a fairly smooth function relative to temperature. However, the large interactions necessary for specific binding give rise to an energetic landscape that is restrictive to sliding [Slutsky and Mirny, 2004] (Fig. 5.1b). Interestingly, to resolve this inconsistency, the diffusive search was split into two pieces [Slutsky and Mirny, 2004]. Specifically, two inter-convertible populations: one scanning state, which slides freely on a smooth landscape; and, a recognition state, which engages intermittently and allows for specific binding [Slutsky and Mirny, 2004]. Here, we offer an alternative splitting of the search process. The first half of our search is characterized by weak binding intermediates, most of which can be rapidly escaped. However, a few of these interactions are just strong enough to foster specific binding.

The facilitated diffusion model suggests that cellular pressures on the target search should result in optimization of non-specific interactions. Alternatively, from the perspective of the RC search, the response should be minimization of non-specific interactions, with the cell instead electing to optimize the 3D search.

The reality for most proteins is likely somewhere between. As cells adapt to environmental pressures, some systems in the cell may adopt to minimize search complexity, while other systems may invest facilitated mechanisms. What stands out, is that cells do not seem to have left target searches up to chance. Each example presented here, described a target search, that, at the outset, sounded statistically unlikely at best; and yet, by tuning the interactions between the protein and DNA, the cell managed to stack the odds of the target search in its favor.

# Bibliography

[Abbondanzieri *et al.*, 2005] E. A. Abbondanzieri, W. J. Greenleaf, J. W. Shaevitz, R. Landick, and S. M. Block. Direct observation of base-pair stepping by rna polymerase. *Nature*, 438(7067):460–5, 2005.

[Adzuma, 1998] K. Adzuma. No sliding during homology search by reca protein. *Journal of Biological Chemistry*, 273(47):31565–31573, 1998.

[Anders *et al.*, 2014] C. Anders, O. Niewoehner, A. Duerst, and M. Jinek. Structural basis of pam-dependent target dna recognition by the cas9 endonuclease. *Nature*, 513(7519):569–+, 2014.

[Austin *et al.*, 1975] R. H. Austin, K. W. Beeson, L. Eisenstein, H. Frauenfelder, and I. C. Gunsalus. Dynamics of ligand-binding to myoglobin. *Biochemistry*, 14(24):5355–5373, 1975.

[Austin *et al.*, 1983] R. H. Austin, J. Karohl, and T. M. Jovin. Rotational diffusion of escherichia coli rna polymerase free and bound to deoxyribonucleic acid in nonspecific complexes. *Biochemistry*, 22(13):3082–90, 1983.

[Bakshi *et al.*, 2013] S. Bakshi, R. M. Dalrymple, W. Li, H. Choi, and J. C. Weisshaar. Partitioning of rna polymerase activity in live escherichia coli from analysis of single-molecule diffusive trajectories. *Biophys J*, 105(12):2676–86, 2013.

[Barrangou *et al.*, 2007] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero, and P. Horvath. Crispr provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819):1709–12, 2007.

[Barzel and Kupiec, 2008] A. Barzel and M. Kupiec. Finding a match: how do homologous sequences get together for recombination? *Nature Reviews Genetics*, 9(1):27–37, 2008.

[Bassett *et al.*, 2013] A. R. Bassett, C. Tibbit, C. P. Ponting, and J. L. Liu. Highly efficient targeted mutagenesis of drosophila with the crispr/cas9 system. *Cell Rep*, 4(1):220–8, 2013.

[Bauer and Metzler, 2012] M. Bauer and R. Metzler. Generalized facilitated diffusion model for dna-binding proteins with search and recognition states. *Biophys J*, 102(10):2321–30, 2012.

[Becker *et al.*, 2013] N. A. Becker, J. P. Peters, 3rd Maher, L. J., and T. A. Lionberger. Mechanism of promoter repression by lac repressor-dna loops. *Nucleic Acids Res*, 41(1):156–66, 2013.

[Beckwith, 2011] J. Beckwith. The operon as paradigm: normal science and the beginning of biological complexity. *J Mol Biol*, 409(1):7–13, 2011.

[Berg and Blomberg, 1976] O. G. Berg and C. Blomberg. Association kinetics with coupled diffusional flows. special application to the lac repressor–operator system. *Biophys Chem*, 4(4):367–81, 1976.

[Berg and von Hippel, 1987] O. G. Berg and P. H. von Hippel. Selection of dna binding sites by regulatory proteins. statistical-mechanical theory and application to operators and promoters. *J Mol Biol*, 193(4):723–50, 1987.

[Berg *et al.*, 1981] O. G. Berg, R. B. Winter, and P. H. von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. models and theory. *Biochemistry*, 20(24):6929–48, 1981.

[Betz *et al.*, 1986] J. L. Betz, H. M. Sasmor, F. Buck, M. Y. Insley, and M. H. Caruthers. Base substitution mutants of the lac operator: in vivo and in vitro affinities for lac repressor. *Gene*, 50(1-3):123–32, 1986.

[Bikard *et al.*, 2013] D. Bikard, W. Jiang, P. Samai, A. Hochschild, F. Zhang, and L. A. Marraffini. Programmable repression and activation of bacterial gene expression using an engineered crispr-cas system. *Nucleic Acids Res*, 41(15):7429–37, 2013.

[Blosser *et al.*, 2015] T. R. Blosser, L. Loeff, E. R. Westra, M. Vlot, T. Kunne, M. Sobota, C. Dekker, S. J. Brouns, and C. Joo. Two distinct dna binding modes guide dual roles of a crispr-cas protein complex. *Mol Cell*, 58(1):60–70, 2015.

[Brouns *et al.*, 2008] S. J. Brouns, M. M. Jore, M. Lundgren, E. R. Westra, R. J. Slijkhuis, A. P. Snijders, M. J. Dickman, K. S. Makarova, E. V. Koonin, and J. van der Oost. Small crispr rnas guide antiviral defense in prokaryotes. *Science*, 321(5891):960–4, 2008.

[Browning and Busby, 2004] D. F. Browning and S. J. Busby. The regulation of bacterial transcription initiation. *Nat Rev Microbiol*, 2(1):57–65, 2004.

[Brunner and Bujard, 1987] M. Brunner and H. Bujard. Promoter recognition and promoter strength in the escherichia coli system. *EMBO J*, 6(10):3139–44, 1987.

[Chen *et al.*, 2008] Z. C. Chen, H. J. Yang, and N. P. Pavletich. Mechanism of homologous recombination from the reca-ssdna/dsdna structures. *Nature*, 453(7194):489–U3, 2008.

[Cho *et al.*, 2009] B. K. Cho, K. Zengler, Y. Qiu, Y. S. Park, E. M. Knight, C. L. Barrett, Y. Gao, and B. O. Palsson. The transcription unit architecture of the escherichia coli genome. *Nat Biotechnol*, 27(11):1043–9, 2009.

[Cong *et al.*, 2013] L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang. Multiplex genome engineering using crispr/cas systems. *Science*, 339(6121):819–23, 2013.

[Datsenko *et al.*, 2012] K. A. Datsenko, K. Pougach, A. Tikhonov, B. L. Wanner, K. Severinov, and E. Semenova. Molecular memory of prior infections activates the crispr/cas adaptive bacterial immunity system. *Nat Commun*, 3:945, 2012.

[Dayton *et al.*, 1984] C. J. Dayton, D. E. Prosen, K. L. Parker, and C. L. Cech. Kinetic measurements of escherichia coli rna polymerase association with bacteriophage t7 early promoters. *J Biol Chem*, 259(3):1616–21, 1984.

[De Vlaminck *et al.*, 2012] I. De Vlaminck, M. T. J. van Loenhout, L. Zweifel, J. den Blanken, K. Hooning, S. Hage, J. Kerssemakers, and C. Dekker. Mechanism of homology recognition in dna recombination from dual-molecule experiments. *Molecular Cell*, 46(5):616–624, 2012.

[deHaseth *et al.*, 1998] P. L. deHaseth, M. L. Zupancic, and Jr. Record, M. T. Rna polymerase-promoter interactions: the comings and goings of rna polymerase. *J Bacteriol*, 180(12):3019–25, 1998.

[Ebright, 2000] R. H. Ebright. Rna polymerase: structural similarities between bacterial rna polymerase and eukaryotic rna polymerase ii. *J Mol Biol*, 304(5):687–98, 2000.

[Eckert-Boulet and Lisby, 2009] N. Eckert-Boulet and M. Lisby. Regulation of rdna stability by sumoylation. *DNA Repair*, 8(4):507–516, 2009.

[Eckert-Boulet and Lisby, 2010] N. Eckert-Boulet and M. Lisby. Regulation of homologous recombination at telomeres in budding yeast. *Febs Letters*, 584(17):3696–3702, 2010.

[Eggleston and Kowalczykowski, 1991] A. K. Eggleston and S. C. Kowalczykowski. An overview of homologous pairing and dna strand exchange proteins. *Biochimie*, 73(2-3):163–176, 1991.

[Elf *et al.*, 2007] J. Elf, G. W. Li, and X. S. Xie. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*, 316(5828):1191–4, 2007.

[Filippo *et al.*, 2008] J. S. Filippo, P. Sung, and H. Klein. Mechanism of eukaryotic homologous recombination. *Annual Review of Biochemistry*, 77:229–257, 2008.

[Fineran *et al.*, 2014] P. C. Fineran, M. J. Gerritzen, M. Suarez-Diez, T. Kunne, J. Boekhorst, S. A. van Hijum, R. H. Staals, and S. J. Brouns. Degenerate target sites mediate rapid primed crispr adaptation. *Proc Natl Acad Sci U S A*, 111(16):E1629–38, 2014.

[Finkelstein and Greene, 2011] I. J. Finkelstein and E. C. Greene. Supported lipid bilayers and dna curtains for high-throughput single-molecule studies. *Methods Mol Biol*, 745:447–61, 2011.

[Finkelstein *et al.*, 2010] I. J. Finkelstein, M. L. Visnapuu, and E. C. Greene. Single-molecule imaging reveals mechanisms of protein disruption by a dna translocase. *Nature*, 468(7326):983–7, 2010.

[Forget and Kowalczykowski, 2012] A. L. Forget and S. C. Kowalczykowski. Single-molecule imaging of dna pairing by reca reveals a three-dimensional homology search. *Nature*, 482(7385):423–U178, 2012.

[Frank *et al.*, 1997] D. E. Frank, R. M. Saecker, J. P. Bond, M. W. Capp, O. V. Tsodikov, S. E. Melcher, M. M. Levandoski, and Jr. Record, M. T. Thermodynamics of the interactions of lac repressor with variants of the symmetric lac operator: effects of converting a consensus site to a non-specific site. *J Mol Biol*, 267(5):1186–206, 1997.

[Fraser *et al.*, 2007] C. Fraser, W. P. Hanage, and B. G. Spratt. Recombination and the nature of bacterial speciation. *Science*, 315(5811):476–480, 2007.

[Frauenfelder *et al.*, 1991] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

[Friedman and Gelles, 2012] L. J. Friedman and J. Gelles. Mechanism of transcription initiation at an activator-dependent promoter defined by single-molecule observation. *Cell*, 148(4):679–89, 2012.

[Garneau *et al.*, 2010] J. E. Garneau, M. E. Dupuis, M. Villion, D. A. Romero, R. Barrangou, P. Boyaval, C. Fremaux, P. Horvath, A. H. Magadan, and S. Moineau. The crispr/cas bacterial immune system cleaves bacteriophage and plasmid dna. *Nature*, 468(7320):67–71, 2010.

[Gibb *et al.*, 2014] B. Gibb, L. F. Ye, S. C. Gergoudis, Y. Kwon, H. Niu, P. Sung, and E. C. Greene. Concentration-dependent exchange of replication protein a on single-stranded dna revealed by single-molecule imaging. *PLoS One*, 9(2):e87922, 2014.

[Gilbert *et al.*, 2013] L. A. Gilbert, M. H. Larson, L. Morsut, Z. Liu, G. A. Brar, S. E. Torres, N. Stern-Ginossar, O. Brandman, E. H. Whitehead, J. A. Doudna, W. A. Lim, J. S. Weissman, and L. S. Qi. Crispr-mediated modular rna-guided regulation of transcription in eukaryotes. *Cell*, 154(2):442–51, 2013.

[Gladyshev and Kleckner, 2014] E. Gladyshev and N. Kleckner. Direct recognition of homology between double helices of dna in neurospora crassa. *Nature Communications*, 5, 2014.

[Gorman *et al.*, 2010] J. Gorman, A. J. Plys, M. L. Visnapuu, E. Alani, and E. C. Greene. Visualizing one-dimensional diffusion of eukaryotic dna repair factors along a chromatin lattice. *Nat Struct Mol Biol*, 17(8):932–8, 2010.

[Gorman *et al.*, 2012] J. Gorman, F. Wang, S. Redding, A. J. Plys, T. Fazio, S. Wind, E. E. Alani, and E. C. Greene. Single-molecule imaging reveals target-search mechanisms during dna mismatch repair. *Proc Natl Acad Sci U S A*, 109(45):E3074–83, 2012.

[Graham *et al.*, 2011] J. S. Graham, R. C. Johnson, and J. F. Marko. Concentration-dependent exchange accelerates turnover of proteins bound to double-stranded dna. *Nucleic Acids Res*, 39(6):2249–59, 2011.

[Gratz *et al.*, 2013] S. J. Gratz, A. M. Cummings, J. N. Nguyen, D. C. Hamm, L. K. Donohue, M. M. Harrison, J. Wildonger, and K. M. O'Connor-Giles. Genome engineering of drosophila with the crispr rna-guided cas9 nuclease. *Genetics*, 194(4):1029–35, 2013.

[Greene *et al.*, 2010] E. C. Greene, S. Wind, T. Fazio, J. Gorman, and M. L. Visnapuu. Dna curtains for high-throughput single-molecule optical imaging. *Methods Enzymol*, 472:293–315, 2010.

[Guthold *et al.*, 1999] M. Guthold, X. Zhu, C. Rivetti, G. Yang, N. H. Thomson, S. Kasas, H. G. Hansma, B. Smith, P. K. Hansma, and C. Bustamante. Direct observation of one-dimensional diffusion and transcription by escherichia coli rna polymerase. *Biophys J*, 77(4):2284–94, 1999.

[Halford and Marko, 2004] S. E. Halford and J. F. Marko. How do site-specific dna-binding proteins find their targets? *Nucleic Acids Res*, 32(10):3040–52, 2004.

[Halford, 2009] S. E. Halford. An end to 40 years of mistakes in dna-protein association kinetics? *Biochem Soc Trans*, 37(Pt 2):343–8, 2009.

[Hammar *et al.*, 2012] P. Hammar, P. Leroy, A. Mahmutovic, E. G. Marklund, O. G. Berg, and J. Elf. The lac repressor displays facilitated diffusion in living cells. *Science*, 336(6088):1595–8, 2012.

[Harada *et al.*, 1999] Y. Harada, T. Funatsu, K. Murakami, Y. Nonoyama, A. Ishihama, and T. Yanagida. Single-molecule imaging of rna polymerase-dna interactions in real time. *Biophys J*, 76(2):709–15, 1999.

[Hastings *et al.*, 2009] P. J. Hastings, J. R. Lupski, S. M. Rosenberg, and G. Ira. Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8):551–564, 2009.

[Haugen *et al.*, 2008] S. P. Haugen, W. Ross, and R. L. Gourse. Advances in bacterial promoter recognition and its control by factors that do not bind dna. *Nat Rev Microbiol*, 6(7):507–19, 2008.

[Hawkins *et al.*, 2013] M. Hawkins, S. Malla, M. J. Blythe, C. A. Nieduszynski, and T. Allers. Accelerated growth in the absence of dna replication origins. *Nature*, 503(7477):544–+, 2013.

[Hawley and McClure, 1980] D. K. Hawley and W. R. McClure. In vitro comparison of initiation properties of bacteriophage lambda wild-type pr and x3 mutant promoters. *Proc Natl Acad Sci U S A*, 77(11):6381–5, 1980.

[Herbert *et al.*, 2008] K. M. Herbert, W. J. Greenleaf, and S. M. Block. Single-molecule studies of rna polymerase: motoring along. *Annu Rev Biochem*, 77:149–76, 2008.

[Heyer *et al.*, 2010] W. D. Heyer, K. T. Ehmsen, and J. Liu. Regulation of homologous recombination in eukaryotes. *Annual Review of Genetics, Vol 44*, 44:113–139, 2010.

[Hochstrasser *et al.*, 2014] M. L. Hochstrasser, D. W. Taylor, P. Bhat, C. K. Guegler, S. H. Sternberg, E. Nogales, and J. A. Doudna. Casa mediates cas3-catalyzed target degradation during crispr rna-guided interference. *Proc Natl Acad Sci U S A*, 111(18):6618–23, 2014.

[Horatio Scott Carslaw, 1986] John Conrad Jaeger Horatio Scott Carslaw. *Conduction of Heat in Solids*. Clarendon Press, Cambridge, UK, New York, 1986.

[Hsieh *et al.*, 1992] P. Hsieh, C. S. Cameriniotero, and R. D. Cameriniotero. The synapsis event in the homologous pairing of dnas - reca recognizes and pairs less than one helical repeat of dna. *Proceedings of the National Academy of Sciences of the United States of America*, 89(14):6492–6496, 1992.

[Hu *et al.*, 2006] T. Hu, A. Y. Grosberg, and B. I. Shklovskii. How proteins search for their specific sites on dna: the role of dna conformation. *Biophys J*, 90(8):2731–44, 2006.

[Hwang *et al.*, 2013] W. Y. Hwang, Y. Fu, D. Reyon, M. L. Maeder, S. Q. Tsai, J. D. Sander, R. T. Peterson, J. R. Yeh, and J. K. Joung. Efficient genome editing in zebrafish using a crispr-cas system. *Nat Biotechnol*, 31(3):227–9, 2013.

[Ishihama, 2000] A. Ishihama. Functional modulation of escherichia coli rna polymerase. *Annu Rev Microbiol*, 54:499–518, 2000.

[J. *et al.*, 2007] Berg J., Tymoczko J., and Stryer L. *Biochemistry*. W.H. Freeman and Company; New York, 2007.

[Jiang *et al.*, 2013] W. Jiang, D. Bikard, D. Cox, F. Zhang, and L. A. Marraffini. Rna-guided editing of bacterial genomes using crispr-cas systems. *Nat Biotechnol*, 31(3):233–9, 2013.

[Jinek *et al.*, 2012] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. A programmable dual-rna-guided dna endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–21, 2012.

[Jinek *et al.*, 2013] M. Jinek, A. East, A. Cheng, S. Lin, E. Ma, and J. Doudna. Rna-programmed genome editing in human cells. *Elife*, 2:e00471, 2013.

[Jones *et al.*, 2003] S. Jones, H. P. Shanahan, H. M. Berman, and J. M. Thornton. Using electrostatic potentials to predict dna-binding sites on dna-binding proteins. *Nucleic Acids Res*, 31(24):7189–98, 2003.

[Kabata *et al.*, 1993] H. Kabata, O. Kurosawa, I. Arai, M. Washizu, S. A. Margarson, R. E. Glass, and N. Shimamoto. Visualization of single molecules of rna polymerase sliding along dna. *Science*, 262(5139):1561–3, 1993.

[Kalodimos *et al.*, 2002] C. G. Kalodimos, A. M. Bonvin, R. K. Salinas, R. Wechselberger, R. Boelens, and R. Kaptein. Plasticity in protein-dna recognition: lac repressor interacts with its natural operator 01 through alternative conformations of its dna-binding domain. *EMBO J*, 21(12):2866–76, 2002.

[Kao-Huang *et al.*, 1977] Y. Kao-Huang, A. Revzin, A. P. Butler, P. O'Conner, D. W. Noble, and P. H. von Hippel. Nonspecific dna binding of genome-regulating proteins as a biological control mechanism: measurement of dna-bound escherichia coli lac repressor in vivo. *Proc Natl Acad Sci U S A*, 74(10):4228–32, 1977.

[Koslover *et al.*, 2011] E. F. Koslover, M. A. Diaz de la Rosa, and A. J. Spakowitz. Theoretical and computational modeling of target-site search kinetics in vitro and in vivo. *Biophys J*, 101(4):856–65, 2011.

[Kunkel and Erie, 2005] T. A. Kunkel and D. A. Erie. Dna mismatch repair. *Annu Rev Biochem*, 74:681–710, 2005.

[Larsen *et al.*, 1991] T. A. Larsen, M. L. Kopka, and R. E. Dickerson. Crystal structure analysis of the b-dna dodecamer cgtgaattcacg. *Biochemistry*, 30(18):4443–9, 1991.

[Lewis, 2005] M. Lewis. The lac repressor. *C R Biol*, 328(6):521–48, 2005.

[Lewis, 2011] M. Lewis. A tale of two repressors. *J Mol Biol*, 409(1):14–27, 2011.

[Li *et al.*, 2009] G. W. Li, O. G. Berg, and J. Elf. Effects of macromolecular crowding and dna looping on gene regulation kinetics. *Nature Physics*, 5(4):294–297, 2009.

[Lin *et al.*, 2006] Z. G. Lin, H. Z. Kong, M. Nei, and H. Ma. Origins and evolution of the reca/rad51 gene family: Evidence for ancient gene duplication and endosymbiotic gene transfer. *Proceedings of the National Academy of Sciences of the United States of America*, 103(27):10328–10333, 2006.

[Lohman *et al.*, 1980] T. M. Lohman, P. L. deHaseth, and Jr. Record, M. T. Pentalysine-deoxyribonucleic acid interactions: a model for the general effects of ion concentrations on the interactions of proteins with nucleic acids. *Biochemistry*, 19(15):3522–30, 1980.

[Maeder *et al.*, 2013] M. L. Maeder, S. J. Linder, V. M. Cascio, Y. Fu, Q. H. Ho, and J. K. Joung. Crispr rna-guided activation of endogenous human genes. *Nat Methods*, 10(10):977–9, 2013.

[Mali *et al.*, 2013a] P. Mali, J. Aach, P. B. Stranges, K. M. Esvelt, M. Moosburner, S. Kosuri, L. Yang, and G. M. Church. Cas9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol*, 31(9):833–8, 2013.

[Mali *et al.*, 2013b] P. Mali, L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, and G. M. Church. Rna-guided human genome engineering via cas9. *Science*, 339(6121):823–6, 2013.

[McClure, 1980] W. R. McClure. Rate-limiting steps in rna chain initiation. *Proc Natl Acad Sci U S A*, 77(10):5634–8, 1980.

[McClure, 1985] W. R. McClure. Mechanism and control of transcription initiation in prokaryotes. *Annu Rev Biochem*, 54:171–204, 1985.

[Mendoza-Vargas *et al.*, 2009] A. Mendoza-Vargas, L. Olvera, M. Olvera, R. Grande, L. Vega-Alvarado, B. Taboada, V. Jimenez-Jacinto, H. Salgado, K. Juarez, B. Contreras-Moreira, A. M.

Huerta, J. Collado-Vides, and E. Morett. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in e. coli. *PLoS One*, 4(10):e7526, 2009.

[Minton, 2001] A. P. Minton. The influence of macromolecular crowding and macromolecular confinement on biochemical reactions in physiological media. *J Biol Chem*, 276(14):10577–80, 2001.

[Moses and Prevost, 1966] V. Moses and C. Prevost. Catabolite repression of beta-galactosidase synthesis in escherichia coli. *Biochem J*, 100(2):336–53, 1966.

[Mulepati and Bailey, 2013] S. Mulepati and S. Bailey. In vitro reconstitution of an escherichia coli rna-guided immune system reveals unidirectional, atp-dependent degradation of dna target. *J Biol Chem*, 288(31):22184–92, 2013.

[Neale and Keeney, 2006] M. J. Neale and S. Keeney. Clarifying the mechanics of dna strand exchange in meiotic recombination. *Nature*, 442(7099):153–158, 2006.

[Nudler, 2009] E. Nudler. Rna polymerase active center: the molecular engine of transcription. *Annu Rev Biochem*, 78:335–61, 2009.

[Pan *et al.*, 2011] J. Pan, M. Sasaki, R. Kniewel, H. Murakami, H. G. Blitzblau, S. E. Tischfield, X. Zhu, M. J. Neale, M. Jasin, N. D. Socci, A. Hochwagen, and S. Keeney. A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell*, 144(5):719–731, 2011.

[Perez-Pinera *et al.*, 2013] P. Perez-Pinera, D. D. Kocak, C. M. Vockley, A. F. Adler, A. M. Kabadi, L. R. Polstein, P. I. Thakore, K. A. Glass, D. G. Ousterout, K. W. Leong, F. Guilak, G. E. Crawford, T. E. Reddy, and C. A. Gersbach. Rna-guided gene activation by crispr-cas9-based transcription factors. *Nat Methods*, 10(10):973–6, 2013.

[Qi *et al.*, 2013] L. S. Qi, M. H. Larson, L. A. Gilbert, J. A. Doudna, J. S. Weissman, A. P. Arkin, and W. A. Lim. Repurposing crispr as an rna-guided platform for sequence-specific control of gene expression. *Cell*, 152(5):1173–83, 2013.

[Qi *et al.*, 2015] Z. Qi, S. Redding, J. Y. Lee, B. Gibb, Y. Kwon, H. Y. Niu, W. A. Gaines, P. Sung, and E. C. Greene. Dna sequence alignment by microhomology sampling during homologous recombination. *Cell*, 160(5), 2015.

[Ragunathan *et al.*, 2013] K. Ragunathan, C. Liu, and T. Ha. Reca filament sliding on dna facilitates homology search (vol 1, e00067, 2012). *Elife*, 2, 2013.

[Redner, 2001] Sidney Redner. *A guide to first-passage processes.* Cambridge University Press, Cambridge, UK, New York, 2001.

[Renkawitz *et al.*, 2014] J. Renkawitz, C. A. Lademann, and S. Jentsch. Dna damage mechanisms and principles of homology search during recombination. *Nature Reviews Molecular Cell Biology*, 15(6):369–383, 2014.

[Reppas *et al.*, 2006] N. B. Reppas, J. T. Wade, G. M. Church, and K. Struhl. The transition between transcriptional initiation and elongation in e. coli is highly variable and often rate limiting. *Mol Cell*, 24(5):747–57, 2006.

[Ricchetti *et al.*, 1988] M. Ricchetti, W. Metzger, and H. Heumann. One-dimensional diffusion of escherichia coli dna-dependent rna polymerase: a mechanism to facilitate promoter location. *Proc Natl Acad Sci U S A*, 85(13):4610–4, 1988.

[Riggs *et al.*, 1970] A. D. Riggs, S. Bourgeois, and M. Cohn. The lac repressor-operator interaction. 3. kinetic studies. *J Mol Biol*, 53(3):401–17, 1970.

[Roe *et al.*, 1984] J. H. Roe, R. R. Burgess, and Jr. Record, M. T. Kinetics and mechanism of the interaction of escherichia coli rna polymerase with the lambda pr promoter. *J Mol Biol*, 176(4):495–522, 1984.

[Rohs *et al.*, 2010] R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann. Origins of specificity in protein-dna recognition. *Annu Rev Biochem*, 79:233–69, 2010.

[Rutkauskas *et al.*, 2015] M. Rutkauskas, T. Sinkunas, I. Songailiene, M. S. Tikhomirova, V. Siksnys, and R. Seidel. Directional r-loop formation by the crispr-cas surveillance complex cascade provides efficient off-target site rejection. *Cell Rep*, 2015.

[Saecker *et al.*, 2011] R. M. Saecker, Jr. Record, M. T., and P. L. Dehaseth. Mechanism of bacterial transcription initiation: Rna polymerase - promoter binding, isomerization to initiation-competent open complexes, and initiation of rna synthesis. *J Mol Biol*, 412(5):754–71, 2011.

[Sasaki *et al.*, 2010] M. Sasaki, J. Lange, and S. Keeney. Genome destabilization by homologous recombination in the germ line. *Nature Reviews Molecular Cell Biology*, 11(3):182–195, 2010.

[Schonhoft and Stivers, 2012] J. D. Schonhoft and J. T. Stivers. Timing facilitated site transfer of an enzyme on dna. *Nat Chem Biol*, 8(2):205–10, 2012.

[Seeman *et al.*, 1976] N. C. Seeman, J. M. Rosenberg, and A. Rich. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci U S A*, 73(3):804–8, 1976.

[Shakked and Rabinovich, 1986] Z. Shakked and D. Rabinovich. The effect of the base sequence on the fine structure of the dna double helix. *Prog Biophys Mol Biol*, 47(3):159–95, 1986.

[Simons *et al.*, 1983] R. W. Simons, B. C. Hoopes, W. R. McClure, and N. Kleckner. Three promoters near the termini of is10: pin, pout, and piii. *Cell*, 34(2):673–82, 1983.

[Sing *et al.*, 2014] C. E. Sing, M. O. de la Cruz, and J. F. Marko. Multiple-binding-site mechanism explains concentration-dependent unbinding rates of dna-binding proteins. *Nucleic Acids Research*, 42(6):3783–3791, 2014.

[Singer and Wu, 1987] P. Singer and C. W. Wu. Promoter search by escherichia coli rna polymerase on a circular dna template. *J Biol Chem*, 262(29):14178–89, 1987.

[Sinkunas *et al.*, 2011] T. Sinkunas, G. Gasiunas, C. Fremaux, R. Barrangou, P. Horvath, and V. Siksnys. Cas3 is a single-stranded dna nuclease and atp-dependent helicase in the crispr/cas immune system. *EMBO J*, 30(7):1335–42, 2011.

[Slutsky and Mirny, 2004] M. Slutsky and L. A. Mirny. Kinetics of protein-dna interaction: Facilitated target location in sequence-dependent potential. *Biophysical Journal*, 87(6):4021–4035, 2004.

[Smith *et al.*, 2007] C. E. Smith, B. Llorente, and L. S. Symington. Template switching during break-induced replication. *Nature*, 447(7140):102–105, 2007.

[So *et al.*, 2011] L. H. So, A. Ghosh, C. Zong, L. A. Sepulveda, R. Segev, and I. Golding. General properties of transcriptional time series in escherichia coli. *Nat Genet*, 43(6):554–60, 2011.

[Sternberg *et al.*, 2014] S. H. Sternberg, S. Redding, M. Jinek, E. C. Greene, and J. A. Doudna. Dna interrogation by the crispr rna-guided endonuclease cas9. *Nature*, 507(7490):62–7, 2014.

[Tafvizi *et al.*, 2011] A. Tafvizi, F. Huang, A. R. Fersht, L. A. Mirny, and A. M. van Oijen. A single-molecule characterization of p53 search on dna. *Proc Natl Acad Sci U S A*, 108(2):563–8, 2011.

[von Hippel and Berg, 1986] P. H. von Hippel and O. G. Berg. On the specificity of dna-protein interactions. *Proc Natl Acad Sci U S A*, 83(6):1608–12, 1986.

[von Hippel and Berg, 1989] P. H. von Hippel and O. G. Berg. Facilitated target location in biological systems. *J Biol Chem*, 264(2):675–8, 1989.

[Vonhippel and Berg, 1989] P. H. Vonhippel and O. G. Berg. Facilitated target location in biological-systems. *Journal of Biological Chemistry*, 264(2):675–678, 1989.

[Wang *et al.*, 1982] A. H. Wang, S. Fujii, J. H. van Boom, and A. Rich. Molecular structure of the octamer d(g-g-c-c-g-g-c-c): modified a-dna. *Proc Natl Acad Sci U S A*, 79(13):3968–72, 1982.

[Wang *et al.*, 2006] Y. M. Wang, R. H. Austin, and E. C. Cox. Single molecule measurements of repressor protein 1d diffusion on dna. *Phys Rev Lett*, 97(4):048302, 2006.

[Wang *et al.*, 2013a] F. Wang, S. Redding, I. J. Finkelstein, J. Gorman, D. R. Reichman, and E. C. Greene. The promoter-search mechanism of escherichia coli rna polymerase is dominated by three-dimensional diffusion. *Nat Struct Mol Biol*, 20(2):174–81, 2013.

[Wang *et al.*, 2013b] H. Wang, H. Yang, C. S. Shivalila, M. M. Dawlaty, A. W. Cheng, F. Zhang, and R. Jaenisch. One-step generation of mice carrying mutations in multiple genes by crispr/cas-mediated genome engineering. *Cell*, 153(4):910–8, 2013.

[Watson and Crick, 1953] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, 1953.

[Weiner and Kleckner, 1994] B. M. Weiner and N. Kleckner. Chromosome pairing via multiple interstitial interactions before and during meiosis in yeast. *Cell*, 77(7):977–991, 1994.

[Wiedenheft *et al.*, 2011] B. Wiedenheft, G. C. Lander, K. Zhou, M. M. Jore, S. J. Brouns, J. van der Oost, J. A. Doudna, and E. Nogales. Structures of the rna-guided surveillance complex from a bacterial immune system. *Nature*, 477(7365):486–9, 2011.

[Wiedenheft *et al.*, 2012] B. Wiedenheft, S. H. Sternberg, and J. A. Doudna. Rna-guided genetic silencing systems in bacteria and archaea. *Nature*, 482(7385):331–8, 2012.

[Winter *et al.*, 1981] R. B. Winter, O. G. Berg, and P. H. von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. the escherichia coli lac repressor–operator interaction: kinetic measurements and conclusions. *Biochemistry*, 20(24):6961–77, 1981.

[Xiao *et al.*, 2006] J. Xiao, A. M. Lee, and S. F. Singleton. Direct evaluation of a kinetic model for reca-mediated dna-strand exchange: The importance of nucleic acid dynamics and entropy during homologous genetic recombination. *Chembiochem*, 7(8):1265–1278, 2006.

# Appendix A

# Facilitated diffusion

## A.1 Introduction

The simplest model of the association rate of DNA-binding proteins is one where both the protein and DNA target sequence are considered to be diffusing spheres [Halford and Marko, 2004]. Once they come close to one another, they react. This reaction happens at a rate, $k_a$:

$$k_a = 4\pi\rho\left(D_{DNA} + D_{protein}\right)\left(1 + \mathcal{O}(t^{-1/2})\right) \tag{A.1}$$

.

Here, $D$ is the respective diffusion coefficient, and $\rho$ is the reaction radius, which will be defined below. This relation is commonly cited as an upper limit on the speed at which a diffusion-controlled reaction may occur [Halford, 2009]. However, early measurements of the rate at which the *lac* repressor associates to its operon sequence exceeded this limit [Riggs *et al.*, 1970]. The paradox of faster-than-diffusion association was resolved for site-specific association of proteins by including mechanisms of lower dimensionality: hopping, sliding, jumping, and intersegmental transfer, in what has been termed the facilitated diffusion model (Fig A.1) [Berg *et al.*, 1981].

With facilitated diffusion, the association rate has the form $k_{FD} = k_a(1+\zeta)$ where $\zeta$ depends on the dissociation rate of the protein from non-specific DNA, the diffusion coefficient of the protein on the surface of DNA, and the protein concentration [Halford and Marko, 2004].

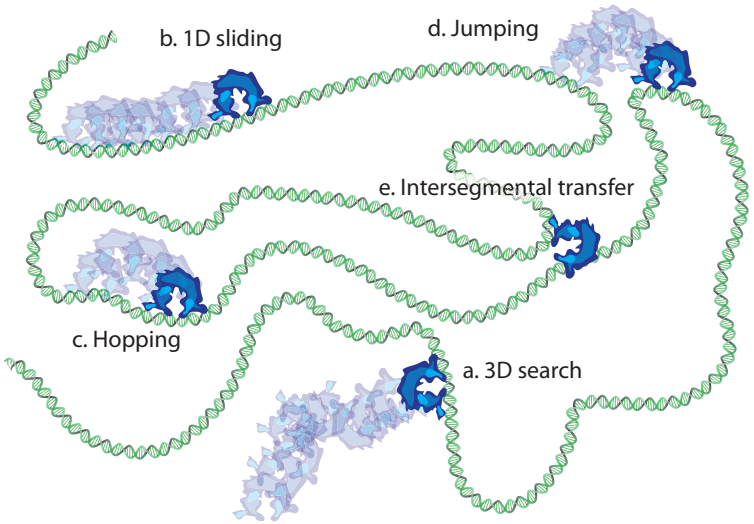


Figure A.1: Diffusion-based models for how proteins might search for binding targets: a. random collision through 3D diffusion; b. 1D sliding, wherein the protein moves without dissociating from the DNA; c. 1D hopping, involving a series of microscopic dissociation and rebinding events; d. Jumping, involving a microscopic dissociation and rebinding across a larger linear region of DNA due to the 3D organization of the chromosome; and e. intersegmental transfer, involving movement from one distal location to another via a looped intermediate.These mechanisms are not mutually exclusive, and the latter four are categorized as facilitated diffusion because by reducing dimensionality they allow target association rates exceeding limits imposed by 3D diffusion.

## A.2 The Reaction Radius

The theoretical framework of the facilitated diffusion model separates DNA-bound from free protein states at a distance, $\rho$, which is referred to as the reaction radius. The motion of the protein beyond this distance is expected to be free thermal diffusion in three dimensions. While, within $\rho$ the protein is expected to be Brownian as well, however, there is additional friction along the DNA axis resulting from the interactions between the protein and the DNA. $\rho$ is then dependent on the size of the protein and the DNA as well as the ionic strength of the solution, as it describes the point at which we are comfortable discarding the gradient of the radial portion of the electrostatic potential of DNA. Typically, $\rho$ is chosen to be the sum of the radii of the searching protein and the DNA plus the Debye screening length, $r_{db}$.

## A.3 The target size

Investigations of the sequence-specific binding of proteins commonly includes of foot-printing and/or mutation methods, revealing both the extent of the protein-DNA interface and minimal consensus sequences. However, a fundamental question remains. That is, over what range can the protein be out of register with the target sequence and still recognize it? If we move the protein 1bp to the left or right of perfectly centered on the target, is the protein heavily biased toward registered binding or does the protein act as if it is has been placed on a random sequence for which it has no preference? If the answer is the former, then this begs the question, exactly how far out of register is it necessary to move the protein until the latter is true. This is the concept of the target size.

## A.4 The Effective Target Size

While, above we considered the effect of lateral displacement on target recognition, we also need to consider the effect of protein orientation with respect to the target. The importance of orientation arises due to the fact that, usually, the entire surface of the protein does not carry out the function of sequence recognition. For example, consider a protein as a Janus particle, where half of the surface recognizes DNA sequences, and the other half carries out some enzymatic function. If we consider the search process to consist of only proteins colliding with the DNA from solution, half of the particles which collide with the target sequence will recognize it, as the other half would have encountered the DNA in an unproductive orientation (i.e. with the enzymatic surface). While, it is difficult to consider the motion of a protein about its own axis in calculation, the effect of orientation can be accounted for by altering the target size. For the example above, an effective target size equal to half of the size of the target size would account for the enzymatic surface, while allowing every encounter to be productive.

## A.5 Calculation of the association rate of RNAP

We consider the DNA to be initially void of bound protein and immersed in an isotropic distribution of RNAP molecules at concentration $C_0$. Then, the initial (first encounter) association rate of proteins to the DNA is identical to the flux of proteins across an absorbing cylinder of radius $\rho$ and

length $L$, where $L = 48,502$ bp. This flux can be found from the solution to the radial diffusion equation, subject to the following boundary conditions.

$$C(r,0) = C_0 \qquad \text{for } r > \rho$$
$$C(\rho,t) = 0$$
$$C(\infty,t) = C_0$$

The Laplace transformed solution satisfying these boundary conditions is given by [Redner, 2001]:

$$C(r,s) = \frac{C_0}{s}\left(1 - \frac{K_0\left(r\sqrt{s/D}\right)}{K_0\left(\rho\sqrt{s/D}\right)}\right) \tag{A.2}$$

Where $D$ is the 3D diffusion coefficient and $K_0$ is the modified Bessel function of the second kind. The solution to the above in the time domain can be written as [Horatio Scott Carslaw, 1986]:

$$C(r,t) = \frac{2C_0}{\pi}\int_0^\infty e^{-Du^2t}\frac{J_0(\rho u)Y_0(ru) - J_0(ru)Y_0(\rho u)}{u\left(J_0^2(\rho u) + Y_0^2(\rho u)\right)}du \tag{A.3}$$

To determine the rate of association per unit length, we find the flux of proteins across the boundary at $\rho$, and then integrate this flux over the entire surface of the DNA:

$$k_\alpha = 2\pi\rho D\frac{d}{dr}C(r,t)\bigg|_{r=\rho} = \frac{8}{\pi}DC_0\int_0^\infty e^{-Du^2t}\left[u\left(J_0^2(\rho u) + Y_0^2(\rho u)\right)\right]^{-1}du \tag{A.4}$$

From the above we also find $k_\alpha^\psi = k_\alpha\psi$, that is, the rate of association for a target size of $\psi$.

## A.6 Calculation of Effective Target Size

To estimate the limiting rate of association, we find the average number of binding events per unit length up to a time, $t$.

$$\langle n(t)\rangle = \frac{8}{\pi}C_0\int_0^\infty \frac{1 - e^{-Du^2t}}{u^3\left(J_0^2(\rho u) + Y_0^2(\rho u)\right)}du \tag{A.5}$$

Now, if there are $N$ promoter sites, each of an effective length $\psi$ of DNA, then the probability of randomly choosing one of these sites is $N\pi/L$. That is to say, on average, it takes $L/N\psi$ random collision events until a promoter is found. We then ask for the time such that $\langle n(t)\rangle = L/N\psi$. We set $\tau = tD/\rho^2$ and then for each value of $C_0$, $D$, $\rho$, $N$, and $\psi$ there is a $\tau$ such the following equality is true.

$$\psi^{-1} = \frac{8}{\pi} N\rho^2 C_0 \int_0^\infty \frac{1 - e^{-u^2 t}}{u^3 \left(J_0^2(u) + Y_0^2(u)\right)} du \tag{A.6}$$

When the concentration reaches a value such that $k_\alpha^\psi$ the predominant contributor to the association rate, the above calculation yields the effective target size. Furthermore, at any concentration higher than this value, the above continues to give the same result for $\psi$. However, at lower concentrations, this calculation will over estimate $\psi$, due to the combined influence of hopping and sliding. Traditionally, this would be referred to as the antenna effect [Riggs *et al.*, 1970].

# Appendix B

# Kinetic rate analysis and search complexity for Rad51's target search

## B.1 Free energy calculations

For dsDNA substrates harboring $\geq$ 8-nt of microhomology, the free energy barrier, $\Delta G^{\ddagger}$, for escape from a potential well can be related to the rate via the following:

$$k_d = Ae^{-\frac{\Delta G^{\ddagger}}{k_b T}} \tag{B.1}$$

where $A$ is a jump frequency, $k_b$ is the Boltzmann constant, and $T$ is the temperature. The difference in between the barrier heights between two different escape processes can be compared, leading to the following equation:

$$\Delta\Delta G^{\ddagger} = \Delta G_2^{\ddagger} - \Delta G_1^{\ddagger} = k_b T \ln\left(\frac{k_d^1}{k_d^2}\right) \tag{B.2}$$

All reported $\Delta\Delta G^{\ddagger}$ values were normalized such that $\Delta G^{\ddagger}$ for Atto565-DNA$_{2.1}$ (which contains a single 8-nt tract of microhomology) is zero, and the experimentally measured data used to calculate $\Delta\Delta G^{\ddagger}$ were the $k_d$ values for the different lengths of microhomology obtained from the survival probability data.

The dissociation kinetics for dsDNA substrates lacking $\geq$ 8-nt of microhomology exhibit a power law distribution, and therefore cannot be described by a single energetic well or a single dissociation

rate. Instead, these data for substrates are consistent with the existence of a large number of states spanning a large range of binding energies and cannot be defined by a single dissociation rate, but rather reflect a broad distribution of dissociation rates. In the case of a distribution of rates, or in other words a distribution of barrier heights, $\Delta G^{\ddagger}$, we find $\Delta\Delta G^{\ddagger}$ as above, but now we weight these values according to their probability density $g(k_d \propto \Delta G^{\ddagger})$. The distribution of dissociation rates, $g(k_d)$, leading to an observed power law survival can be determined as described:

$$g(k_d) = \frac{(\tau_0 k_d)^n e^{-\tau_0 k_d}}{RT\ \Gamma(n)} \tag{B.3}$$

where $\Gamma(n)$ is the gamma function, and $\tau_0$ and $n$ are estimated from the power law fit to the data. The data used to calculate $\Delta\Delta G^{\ddagger}$ for the substrate lacking $\geq$ 8-nt of microhomology were obtained from the experimentally measured survival probability data for Atto565-DNA$_{2.0}$. As above, the reported $\Delta\Delta G^{\ddagger}$ values were normalized such that $\Delta G^{\ddagger}$ for Atto565-DNA$_{2.1}$ (which contains a single 8-nt tract of microhomology) is zero.

## B.2 Search Complexity

During the homology search, the presynaptic complex must interrogate distinct sites within the genome to determine the degree of complementarity. Here, we outline how a microhomology-based search can accelerate this process. We first describe the magnitude of the reduction in search complexity that results from sampling for defined tracts of microhomology by considering the prevalence of particular nucleotide sequences.

The first aspect to consider is the probability of randomly generating a particular sequence of length $n$. For large sequences, the frequency of each base pair can be considered random and independent, therefore the probability of each base pairs occurrence is 0.25; note that for the sake of simplicity we do not account for biased G/C-content, and the potential for skewed sequence composition has little impact on our overall conclusions. To determine the likelihood of a particular sequence of length $n$, we simply find the product of the probability of a particular base at each site within the genome. For instance, the probability of a sequence one base in length being an A is 0.25, whereas the probability of a four base sequence of AAAA is $(0.25)(0.25)(0.25)(0.25) = 0.0039$. More generally, the probability of a particular sequence of length $n$ may be written as $p_n = 4^{-n}$.

We must next consider the probability that a particular sequence of length $n$ occurs within a larger sequence of length$l$. In a sequence of length $l$, there are $l - n + 1$ overlapping sequences of length $n$. The probability that any one of these sequences is not the particular sequence we are looking for is $1 - p_n$. Then the probability that none of the $l - n + 1$ sequences are the particular sequence is the product of all of the individual probabilities, which is written as $(1 - p_n)^{l-n+1}$. From this result we then find the probability of finding a particular sequence of length $n$ occurring within a larger sequence of length $l$, as one minus the probability that the sequence does not occur, which is given as:

$$p_n^l = 1 - \left(1 - 4^{-n}\right)^{l-n+1} \tag{B.4}$$

The above probabilities can be readily converted into a frequency. For instance, the complete *S. cerevisiae* diploid genome is $\sim 24$ Mbp in length, corresponding to a total of $\sim 48$-meganucleotides of ssDNA sequence. This means that each unique sequence, which occurs with a probability of $p_n$, has $\sim 48$ million chances to appear. For example, a sequence of eight base pairs has a probability of $p_8 = 1.5E^5$ to occur and is expected to appear $\sim 762$ times in a genome the size of S. cerevisiae. By contrast there are $\sim 2,947$ identical seven bp sequences or $\sim 184$ identical nine bp sequences. We can express the frequency of finding a particular sequence of length $n$ within a genome of length $l$ as:

$$f_n(l) = 2(l - n + 1)4^{-n} \tag{B.5}$$

It is usually assumed that in order to understand the complexity of a particular target search one needs only to consider the prevalence of potential target sequences within the genome. In the case of the homology search, however, this assumption grossly underestimates the number of non-targets, as the sequence and length of the ssDNA bound within the presynaptic complex itself contributes to the search complexity. An intuitive understanding of how presynaptic complex length influences search complexity can be established by considering that if the searching entity embedded within the presynaptic complex where comprised of a just a single nucleotide (e.g., A), then, on average, it would have to visit all other matching base pairs within the entire genome before locating the correct region of homology. Now, let the presynaptic complex contain two adjacent nucleotides (e.g., AA) each independently searching for a defined site within the genome. In this case each of

the two different adenine bases must independently probe each base in the genome, leading to a two-fold increase in search complexity.

We are therefore concerned with determining the probability that a sequence element of size $n$ (i.e. microhomology) that occurs within a particular sequence of length $o$ (i.e. the length of the presynaptic filament) also occurs in a larger sequence of length $l$ (i.e. the genome). As above, there are $o - n + 1$ overlapping sequences of length $n$ that occur in a sequence of length $o$. Then, using the same method as above, the probability of finding at least one continuous homologous region of length $n$ among the possible $n$-sized sequences available against the substrate of length $l$ is given by:

$$p_n(l|o) = 1 - \left(\left(1 - 4^{-n}\right)^{l-n+1}\right)^{o-n+1} \tag{B.6}$$

Moreover, the above may also be cast as a frequency, which is given by the sum of the frequencies of each individual occurring sequence:

$$f_n(l) = 2(o - n + 1)(l - n + 1)4^{-n} \tag{B.7}$$

Notably, the above considers all possible overlapping sequences of length $n$ that occur in the presynaptic complex (Fig. B.1 upper panel). It may be the case that searching elements within the presynaptic complex and are in fact not overlapping. In the extreme case of a presynaptic complex consisting of completely independent non-overlapping elements (Fig. B.1, lower panel), the number of $n$-sized sequences present is $o/n$, which would replace $o - n + 1$ in the equation above.

We can now calculate search complexity $\mathcal{C}$ by relating the above equations for frequency of occurrence $f_n(l|o)$ by accounting for the total number of potential target sites (i.e. sites within the genome bearing tracts of microhomology of length $n$) in relation to the total number of sites $n$ nucleotides in length (regardless of sequence composition) present within a genome of length $l$, which is written as:

$$\mathcal{C} = \frac{n}{l} f_n(l|o) \tag{B.8}$$

The results arising from this equation are presented in Figures 4.7d-g and B.1, and the further implications of these calculations in relationship to our experimental studies and the homology search

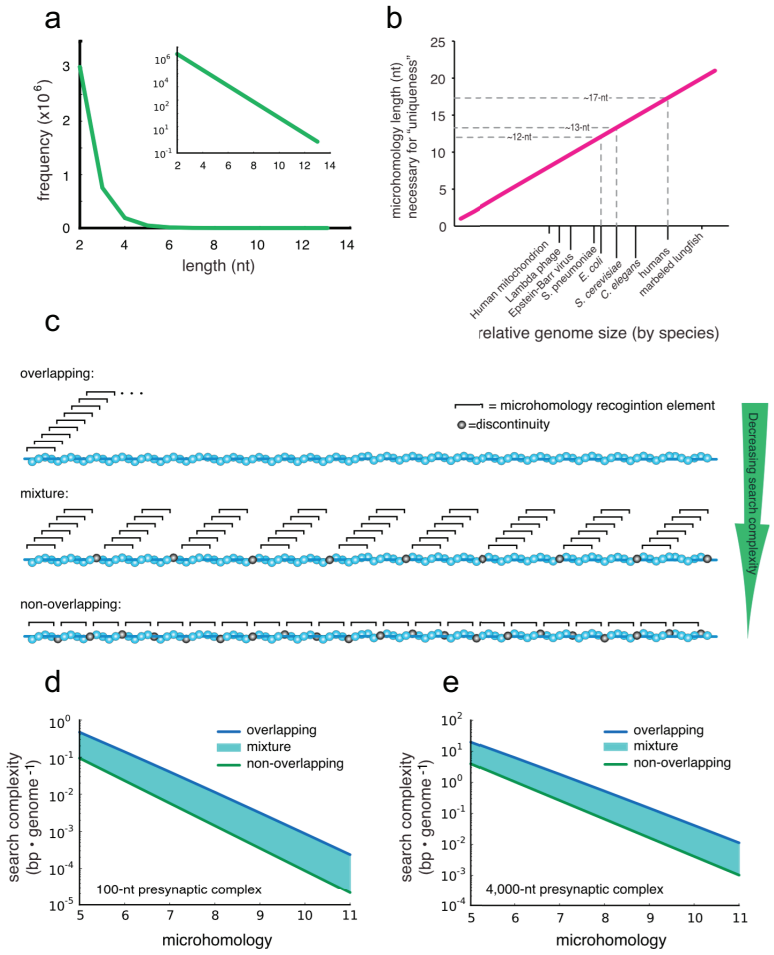


Figure B.1: a. Calculated frequency for any sequence of microhomology of a defined length for the diploid *S. cerevisiae* genome; the inset shows the same microhomology frequency data on a semi-log plot. b. Graph showing the average sequence length necessary to define a unique site with the genomes of various organisms; the x axis is organized based on relative size of each genome. The dashed lines highlight the average length necessary to define a given sequence as unique within the genomes of E. coli, S. cerevisiae, and humans. Note that these calculations assume a randomized genome of uniform A/T/G/C-content. c. Schematic illustrating overlapping, mixture and non-overlapping models describing the potential division of the presynaptic complex into functional units of a defined length. All calculations presented in the main text are based upon an overlapping model, which assumes any stretch of ssDNA of a fixed length can be used to interrogate duplex DNA. The non-overlapping model assumes that the recognition elements are functionally and physically separated from one another in precise increments equal to the length of microhomology that is interrogated during the homology search. The mixture model is a combination of the overlapping and non-overlapping organizations. d and e. Plots showing the influence of model selection on search complexity for presynaptic complexes that are either (D) 100-nt or (E) 4,000-nt in length.

in general are described Discussion section of the main text. The most important concept arising from this theoretical treatment of the homology search problem is the exponential dependence of search complexity on the length ($n$) of the fundamental elements within the presynaptic complex that are responsible for conducting the search. Our experimental work shows that the homology search involves recognition of 8-nt units of microhomology, which results is a substantial reduction on the overall complexity of the homology search.

It should be noted that we are interpreting search complexity as a proxy for the time it would take to complete a particular target search process (i.e. lower search complexity leads to more rapid searches, whereas higher search complexity will result in slower searches). This interpretation of search complexity makes the implicit assumption that there is a large kinetic difference between binding intermediates based on a particular length of sequence microhomology; we experimentally demonstrate that this interpretation of search complexity is valid for homologous recombination based upon the $> 4$ order-of-magnitude difference between binding intermediates with and without 8-nt tracts of microhomology.