# DEVELOPMENT OF A HIGH-THROUGHPUT ASSAY TO MEASURE DNA MISMATCH REPAIR EFFICIENCY IN VIVO

by
GURCAN TUNC KAYIKCIOGLU

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
14 February 2021

# Abstract

Whether due to mutagens or DNA replication errors, mismatches arise spontaneously *in vivo*. If left unrepaired, accumulation of mutations at a high rate can be detrimental to the survival of the organism. Cells recognize the mismatches and repair them via a dedicated mismatch repair system. Although its efficiency has been shown to depend on the type and the sequence context of the mismatch, only a small subset of possible mismatched sequences could be examined thus far. In this work, I describe a high-throughput sequencing based approach that can assess the repair efficiency of many different mismatches in parallel, enabling a systematic analysis of the sequence effect on mismatch repair. In this scheme, an *in vitro* synthesized plasmid containing a single mismatch is introduced to an *E. coli* cell. If the mismatch is repaired prior to replication, all of the descendants will share the same sequence. If, however, replication precedes mismatch repair, the descendants have a mixture of two different sequences, and therefore the event frequencies of these two types provide information on the repair efficiency. Use of DNA barcodes enables obtaining single-molecule level information regarding the fate of each mismatch carrying molecules, through which the repair of 4434 different mismatches and 1300 insertion loops were monitored *in vivo* under various conditions. The results showed that CC mismatches are always poorly repaired whereas local sequence context is a strong determinant of the highly heterogeneous repair efficiency of TT, AG and CT mismatches. In contrast, most of the insertion loops were repaired with a high efficiency without an appreciable sequence context dependence. The measurement of the repair efficiency in mutant cell strains of different mismatch repair pathway mutants also showed a residual repair capability, potentially an indication of side-processes that lead to an apparent repair of mismatch bearing plasmids.

**Primary Reader and Advisor:** Taekjip Ha

**Secondary Reader:** Bin Wu

# Preface

In the majority of this work, I tried to find a new high-throughput method to measure DNA repair efficiency. The narrative that you are about to read will hence postulate the new method, will try to convince you that it can quantify the repair response with a reasonably high accuracy and precision, and will finally discuss the biological implications of our findings.

The First Chapter discusses the very first experimental approach we have attempted after a brief introduction to the topic. While this convinced us that the usage of barcoded vectors is a promising approach to pursue, the reader might quickly notice some shortcomings. While the experiment is cost-wise more affordable to perform by generating mismatch libraries via annealing standard purchasable oligos, this leads to information loss as the sequence of the DNA looks fully identical and hence the information about the original mismatch bearing plasmid might be lost. The DNA sequences we chose to sample in this part might also appear as poor random choices. That is mostly because we did not really choose them, but rather that we tried this barcode-based tracking idea of mine on the leftover material we had available from previous projects. Only later we attempted the experiments with more optimized sequences that at least sample the sequence space in a more balanced fashion using this same strategy.

In the Second Chapter, I tried to address the former and more significant shortcoming by introducing a second barcode on the mismatch libraries themselves. While this avoids the information loss problem I described above and decreases the cost to obtain DNA libraries, the quantity of material that can be obtained with contemporary array synthesis methods are rather prohibitively low. It took us a great effort to develop and optimize, to the best of our ability, a method to obtain mismatch libraries of sufficiently high quality and quantity with this approach. This allowed us to obtain a much clearer picture of what is actually happening in our dataset, yet still left some biological questions unanswered. While the locomotive of my dissertation is a novel high-throughput method we developed that can measure the repair efficiency of unprecedented number of mismatches with minimal effort, the number of possible sequence combinations increase exponentially with the length of DNA motifs being investigated. Yet, as the narrative will argue in more detail that the sequence contexts of concern might not be limited to 3 nucleotides or so, I

therefore will conclude this chapter by constructing an optimized experiment that can illuminate longer range effects by subsampling a certain subset of mismatches.

One common cause of mismatches in DNA is actually the imperfect fidelity of semi-conservative replication. As the two strands of the parent DNA unwind and DNA polymerases try to accurately synthesize a new strand based on the base sequence of the template strand, occasional misincorporations take place. But actually, DNA polymerases also misincorporate additional nucleotides or sometimes skip a nucleotide leading to insertion/deletion loops if uncorrected by proofreading. The same mismatch repair machinery is also capable of detecting these latter lesions, the sequence dependence of which I discuss in the Third Chapter. After a brief visit to the primitive method, I discuss our full characterization of this response against one or two base insertion/deletions analogous to mispairs again using the double barcoding strategy we developed.

The Fourth -and the Last- Chapter can be considered an appendix, in which I will actually totally deviate from mismatched DNA molecules or *E. coli* and rather introduce a new algorithm that can facilitate the analysis of single molecule movies to deduce FRET traces. The algorithm is unsupervised in that it does not require any user input to select point pairs to be able to deduce the transformation matrix between the two images. I later argue that for samples that contain a reasonably high density of molecules that emit signal in both channels, this could obviate the need to perform a calibration experiment, typically done with fluorescent beads in contemporary single-molecule research groups like ours.

The subject of the above sentences are oftentimes “we”. That is because none of this work would have been possible without the contribution of other people whom I happened to run across in my journey. I thank my advisor, Prof. Taekjip Ha, for his intellectual support throughout my graduate school experience. I also owe thanks to my former colleague Chang-Ting Lin, who helped me realizing the idea I conceptualized, especially at the early stages. Lengthy discussions with Prof. Kasper Hansen helped us interpret and move forward with the project, for which I am likewise deeply grateful. I similarly thank Prof. Richard Fishel for providing his critical view on our results and Prof. Winston Timp for generously allowing the usage of his Illumina MiSeq instrument. My past few years have taught me that the science involves convincing others, and

mine was no exception to this rule. In order to address the point of my friends, that the new data analysis approach might be only applicable to my images acquired on Prof. Ha's microscope, I ended up replicating the results on data acquired by Dr. Sonisilpa Mohapatra on Prof. Sarah Woodson's microscope, both of whom I am also indebted to.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# In vivo measurement of DNA repair efficiency

## 1.1 Abstract

Whether due to mutagens or DNA replication errors, mismatches arise spontaneously *in vivo*. If left unrepaired, accumulation of mutations at a high rate can be detrimental to the survival of the organism. Cells recognize the mismatches and repair them via a dedicated mismatch repair system. Although the efficiency of this response has been shown to depend on the type and the sequence context of the mismatch, only a small subset of possible mismatched sequences have been examined thus far. This is mostly due to the cost and labor intensive nature of the techniques that were available by the time these studies were published.

Here, I will introduce a novel high-throughput sequencing based approach that we have developed to address this limitation. In this scheme, a plasmid containing a single mismatch is introduced to an *E. coli* cell. If the mismatch is repaired prior to replication, all of the descendants will share the same sequence. If, however, the mismatch evades repair, the descendant cells will harbor a mixture of two different sequences after multiple rounds of replication. Therefore, the frequencies of these two types of events can be used to quantify the repair efficiency. By a PCR-based scheme, we generated a plasmid library tagged with a random DNA sequence that is essentially unique to each molecule. As the descendants of each plasmid will inherit this same DNA barcode, the replication products of the ancestor plasmid can be traced back without a necessity for any spatial confinement. The combination of these two propositions enables parallel assessment of a multitude of different mismatches.

## 1.2 Introduction

In its standard double stranded form, nucleotides on the two complementary strands of DNA form a Watson-Crick base pair. However, mismatches frequently occur due to external mutagens, spontaneous chemical changes or errors by the DNA replication machinery [1]. Such mismatches need to be corrected prior to the passage of the next replication fork to keep the mutation rates at biologically tolerable levels [2]. As the DNA is replicated in the cell, its two constituent strands are separated by replicative helicases and both single stranded DNAs (ssDNA) generated serve

as templates for semi-conservative DNA synthesis. The initial nucleotide incorporation attempt introduces, on average, about 1 incorrect nucleotide per $10^5$ in *Escherichia coli (E. coli)*, which is improved by two orders of magnitude by the 3' →5' proof-reading capability of the polymerase ( [3] and reviewed in [4]). While the combination of the selective nucleotide incorporation and proof-reading activities of the DNA polymerase means a residual misincorporation rate of 1 wrong nucleotide per about every 10 million nucleotides inserted, this final misincorporation rate of 100 mistakes per billion is still significant. Corresponding to generation of 1.38 mutations per hour in the 4.6Mbp genome of *Escherichia coli K-12* as it completes one full replication every 20 minutes [5], the handling of mispaired bases after the departure of the replication fork necessitates a dedicated mismatch repair (MMR) pathway (reviewed in [6–8]).

For the detection and correction of the continuously generated mismatches, a dedicated mismatch repair pathway (MMR) has evolved in *E. coli*, while homologous pathways also exist in eukaryotes (reviewed in [6, 7]). In *E. coli*, the MMR system starts with the MutS protein that recognizes a mismatch-bearing DNA with a high affinity, which then recruits MutL to the scar site. This MutS-MutL complex engages the nickase MutH and starting from the site of the introduced nick, the two strands are unwound by the helicase UvrD and a segment of the nicked strand is removed by exonucleases, generating a single-stranded region of typically a few hundred nucleotides in length. The resulting ssDNA gap is finally filled with a polymerase and the nick at the termini of the repaired region is closed by DNA ligase hence completing the repair response. The MMR system has a tendency to preserve the information inherited from the template strand as MutH introduces a nick preferentially on the nascent strand rather than a random choice. This discrimination is possible due to a temporary asymmetry in the methylation pattern immediately after DNA replication, as the nascent strand is devoid of methylation, whereas the template strand typically carries an N6-Methyladenosine at the GATC sequence motifs, the former of which is a better substrate for MutH.

*In vivo*, especially in organisms with a short doubling time, the MMR competes with the replication machinery and hence is not necessarily fully efficient. One can assess the MMR efficiency at different sequence positions by comparing the mutation accumulation rates in MMR-capable and

deficient cell strains. However, this indirect approach is confounded by the fact that a particular replicative mutation can arise via two different mismatch intermediates [3,9]. For example, both CT and AG mismatches, assuming the original base pair at that locus is CG, can lead to conversion of the C into an A. In addition, mutational outcomes would also depend on the sequence-dependent DNA synthesis error rate of replicative DNA polymerases. The misincorporation rate of DNA polymerases sharply varies for the nucleotide on the template strand to be replicated, the local sequence context on the template strand as well as the identity of the misincorporated nucleotide [10, 11].

An alternative approach is based on *in vitro* synthesis of DNA containing a mismatch of known identity and monitoring the daughter cells for potential heterogeneities following transformation into the organism of interest. If the mismatch carrying genetic element is designed in such a way that only one of the two strands codes for a reporter gene with an observable effect on the phenotype, either all or none of the offsprings of this repaired DNA would display this observable trait. In contrast, if the first replication fork reaches the mismatch before the detection by MMR or the mismatch detected by the MMR machinery is left unrepaired, the offsprings consist of a mixture coding for two different phenotypes. The constructs used in such studies are typically designed in such a way that one of these sequences coded by the two mispaired strands leads to a defective or truncated peptide. The reporter property of choice could be an endonuclease cut site [12], turbidity introduced by bacteriophages [13], a fluorescent protein such as GFP [14] or $\beta$-galactosidase gene that confers the cells blue color in response to 5-bromo-4-chloro-3-indolyl-$\beta$-D-galactopyranoside (a.k.a. x-gal) [15].

Using such strategies, it has been reported that the MMR response has a different efficiency depending on the identity of the mispaired nucleotides and the local sequence context of the mispair [6,7,13,15,16] (Table 1.2). However, such colony screening assays based on reporter genes are laborious and therefore generating sufficiently large data sets for data mining is impractical. Nevertheless, the number of combinations that need to be tested is very high, given that there are 4 different bases that can occupy each sequence position (A, C, G, T) and 8 possible canonical mismatch types (AA, AC, AG, CC, CT, GG, GT, TT). As an example, 1X coverage of all base

combinations up to the next-nearest neighboring basepairs requires conducting about $4^2 \cdot 8 \cdot 4^2 = 2048$ separate experiments. Yet, this pattern length of five actually is much shorter than the footprint of MutS. For comparison, the footprint of MutS on DNA was previously reported to range from 8 to 20 nucleotides, depending on the sequence of the mismatch under study and the assay method chosen [17, 18]. The product coded by one of the mispaired strands also needs to lead to an observable phenotypic change in practice, typically by causing a truncated translation or introducing an inactivating mutation. Certain codon changes code for either the same residue due to redundancy in the genetic code, or might still go silent as incorporation of a chemically similar residue might not have significant implications for the function of the protein. Therefore this requirement makes some measurements in the exactly desired sequence context challenging, if not impossible.

Instead, the approach I will describe here uses next-generation sequencing (NGS) to directly obtain the base sequence of the repair products and make quantitative comparisons to deduce the sequence context preference of MMR *in vivo* as it competes with the replication process. Similar to the phenotypic assays in the literature, the MMR process forces the information coded on the two strands to agree with each other, whereas the offsprings inherit conflicting information if the plasmid evades repair prior to replication. That is because of the fact that if a mismatch is repaired before the first replication occurs, one of the two strands is converted into the proper complementary of the other strand and subsequent replications will produce only one type of product. On the contrary, an ancestor plasmid evading repair gives rise to a mixture of two products that differ in base identity at the position of the mismatch. By comparing the relative abundance of repaired and unrepaired molecules, we can directly quantify the relative repair efficiency of a mismatch. However, such a conclusion can be reached if and only if single molecule level information regarding the introduced molecules can be obtained experimentally. A standard bulk sequencing experiment starting with multiple identical replicates of a mismatch carrying plasmid would otherwise yield a roughly 1:1 mixture of the two conflicting strands, instead, since each cell independently decides which DNA strand to keep during the repair process.

Not only *E. coli* thriving in a well-shaken LB culture would rapidly make multiple copies of each

of the transformed mismatch-bearing plasmids, but also a typical workflow for sample preparation for sequencing involves DNA amplification. To confer such a standard bulk system single-molecule experimental setup properties, we made use of random DNA sequences, which I will refer to as DNA barcodes in the sequel. DNA barcodes have been used as unique molecular identifiers both to improve the fidelity of DNA sequencing by tracking *in vitro* amplification products [19] and as a signature of cell identity in single-cell genome and transcriptome studies [20–24]. DNA barcoding was also used to accurately quantify *in vitro* polymerase error rates, which are otherwise obscured by the higher error rate of base calling of contemporary sequencing systems [11, 25].

Our approach uses next-generation sequencing to quantitatively compare the sequence context preference of MMR *in vivo*. We generated a barcoded vector library using PCR primers with a random tail, and transformed this barcoded vector library into *E. coli* after ligating to a mispaired DNA fragment library. As the barcodes are shared only among the descendants that originate from the same single molecule of ancestral plasmid, clustering based analysis of the replication products at single-molecule level reveals whether a repair event took place before the arrival of the replication fork to the mismatch-containing region at the single molecule level. Similar to the phenotypic assays in the literature, the MMR process forces the information coded on the two strands to agree with each other, whereas the offsprings inherit conflicting information if the plasmid evades repair prior to replication. By comparing the relative abundance of repaired and unrepaired molecules, we directly quantified the relative repair efficiency of many different mismatches in multiple sequence contexts in parallel.

## 1.3 Results

### 1.3.1 Estimation of the required data amount

Each individual mismatch carrying DNA has a binary fate, as the molecule is either repaired or left unrepaired before replication takes place. The major aim of this work is to quantify the repair efficiency of a particular mismatch as a function of the identity of the mispaired nucleotides (i.e. the mismatch type) and the base composition of the neighboring nucleotides (i.e. the sequence

context). Here, we define the repair efficiency as the relative frequency at which a mismatch experimentally introduced to a cell is outputted in the repaired state rather than the unrepaired state.

Being a frequency, the measurement of the repair efficiency requires independent observations of multiple plasmids carrying the identical mismatch type located within exactly the same sequence context. If we assume that a typical mismatch is repaired about p=90% of the occurrences, quantification of the repair efficiency such that the standard error of the reported measurements is within ± 0.05, N=36 different copies of the molecules carrying the identical mismatch need to be observed ($SE = \sqrt{Np(1-p)}/N$). Drawing inferences regarding the repair response requires observation of different mismatches in different sequence contexts. The assumption that repair has an inherent strand specificity and that the insertion orientation of a mismatch on a larger molecule is of importance will bring the total number of different sequence combinations for a full nearest neighbor sampling of all 8 mismatches to $4^1 \cdot 8 \cdot 4^1 = 128$, hence suggesting 4608 as the targeted number of plasmids to be independently observed in an ideal experiment that investigates nearest-neighbor effects. Extension of this dataset to contain all next-nearest neighbors increases the number of sequence combinations to $4^2 \cdot 8 \cdot 4^2 = 2048$ and the target number of plasmids to observe to 73 728. We hence sought a high-throughput method that can cope with this rapidly increasing number of independent observations and therefore resorted to DNA barcodes.

### 1.3.2 Assessment of the requirements on the barcode

In the workflow that I describe here, neither the transformed cells nor the extracted DNA are physically confined to a certain subcompartment of the growth medium, but they can diffuse freely, instead. Rather our ability to discern the kinship among the individual sequencing reads solely relies on the fact that the DNA barcodes act as reliable unique molecular identifiers. For this to happen, the set of all possible barcodes should be sufficiently diverse to ensure that each molecule is tagged with a distinguishably different barcode than other molecules in the same sample. As the experimental procedures introduce sequencing errors, the barcodes sampled should be separated from all neighbors by more than one mutation, so that the errors can be tolerated.

Let this distance be $\epsilon$, i.e. for two molecules of plasmid x and y carrying barcodes i and j, we conclude that x and y are the descendants of the same ancestor plasmid transformed to the cells if $d(i,j) \leq \epsilon$. In an ideal experiment with no replication or sequences errors, $d(i,j) = 0$ would be the parameter of choice, as barcodes on two sibling plasmids would be exactly identical.

Let $Prob_\epsilon$ denote the probability that in a data set comprising M clans, there exists at least two clans accidentally carrying the same barcode, despite originating from different ancestor molecules, if two barcodes different at most $\epsilon$ positions are considered identical. The $\epsilon = 0$ choice leads to the intuitive case that two barcodes are identical only if they are exactly identical. In this case, when barcodes of length L are used,

$$Prob_0(\exists i,j \ s.t. \ b_i = b_j) = 1 - \frac{\binom{4^L}{M} M!}{(4^L)^M} = 1 - \prod_{i=0}^{M-1} \frac{4^L - i}{4^L} \tag{1.1}$$

where the second term is the probability that each ancestor plasmid is assigned a distinct barcode. In a similar fashion, we can implement an upper-bound on the probability for $\epsilon > 0$. If barcodes $b_1, b_2, ..., b_k$ were already assigned up to the k'th clan, the disallowed barcodes when choosing $b_{k+1}$ for the (k+1)st clan are not only limited to $b_1, b_2, ..., b_k$, but all of their neighbors within a ball of radius $\epsilon$ should also be avoided. Then, for k+1'th clan, the allowed set of barcodes is,

$$A_{k+1} \equiv \{A, C, G, T\}^L \setminus \bigcup_{i=1}^{k} B(b_i, \epsilon) \tag{1.2}$$

With that definition, the probability of having at least one duplicate barcode assignment is,

$$Prob_\epsilon(\exists i \neq j \ s.t. \ b_i = b_j) = 1 - \prod_{k=1}^{M-1} m(A_k)/4^L \tag{1.3}$$

$$m\left(\bigcup_{i=1}^{k} B(b_i, \epsilon)\right) \leq \sum_{i=1}^{k} m(B(b_i, \epsilon)) \tag{1.4}$$

by union bound, and combined with Equation 1.2 it implies,

$$m(A_{k+1}) = 4^L - m\left(\bigcup_{i=1}^{k} B(b_i, \epsilon)\right) \geq 4^L - \sum_{i=1}^{k} m(B(b_i, \epsilon)) \tag{1.5}$$

and hence it follows that,

$$Prob_\epsilon(\exists i \neq j \ s.t. \ b_i = b_j) \leq 1 - \prod_{i=0}^{M-1} \frac{4^L - i - i\binom{L}{\epsilon}(4^\epsilon - 1)}{4^L} \tag{1.6}$$

where $m(B(b_i, \epsilon)) = \binom{L}{\epsilon}(4^\epsilon - 1)$ is the number of distinct $\epsilon$-neighbors of barcode $b_i$. The upper bound in Equation 1.6 is exact for $\epsilon = 0$ and reduces to Equation 1.1.

The probabilities for L = 25 bases are plotted in Figure 1.2a. For reference, 5MLx libraries have 94 thousand clans on average and $\epsilon = 2$ strategy was adopted for data analysis. At this regime, the probability that there exists in the whole data set at least a clan pair mis-tagged with the same barcode is about 1%. While such barcode clashes can lead to occasional mis-classifications of two C and V-type clans as U, this pessimistic upper bound is on the probability of having at least one misclassification in the entire data set, but not the probability that any two chosen clan to be perceived as one. This latter quantity shown in Figure 1.2b depends on the relative volume of $B(b_i, \epsilon)$, i.e.

$$Prob_\epsilon(Given \ i \neq j \ s.t. \ b_i = b_j) = \frac{m(B(b_i, \epsilon))}{m(\{A, C, G, T\}^L)} = \frac{\binom{L}{\epsilon}4^\epsilon}{4^L} \tag{1.7}$$

Both of these parameters suggest that a safe choice of L=25 bases for the barcode length to be sufficient to ensure that the probability of tagging two different clans with the same barcode by random choice is negligibly small.

### 1.3.3 Experimental generation of barcoded vectors

To be able to trace the replication products of each individual molecule of plasmid carrying a mismatch, we generated a vector library that is uniquely labeled with DNA barcodes, but is similar to the standard pUC19 plasmid otherwise [26]. We achieved this by a modified PCR in

which one of the primers carries a train of 25 randomly incorporated nucleotides on its 5' tail (Figure 1.1). While we did not need to make use of this feature, we also separated each block of 5 random nucleotides by a deterministically incorporated A, which could facilitate detection of potential shifts in the barcode during sequencing or replication at a low computational cost. We call these barcodes the 'tracing barcodes' because all plasmids sharing them must have descended from the same single plasmid DNA that carried the barcode (see 1.3.2), hence making it possible to trace the genealogy in a mixed cell culture.

We introduced two non self-complementary restriction sites at the two termini of the barcoded vector library via the 5' tails of the primers in addition to the barcodes. Our particular choice of the pair of restriction enzymes to be used here was mainly informed by the reports that the DNA ligases used in typical *in vitro* experiments are capable of abberrently ligating substrates even if the sticky ends are non-complementary or contain a gap, but that this effect becomes less significant, if the two overhangs are more dissimilar [27, 28]. The SacI and XhoI cut sites we incorporated lead to 3' and 5' overhangs upon a restriction double digest, respectively, hence reducing improper binding.

### 1.3.4 Mismatch library generation

To be able to observe the mismatch type and local sequence context dependence of the cellular repair response, we experimented with dsDNA libraries that carry one and only one mismatch per molecule by experimental design. To form such a library in a simple and cost-affordable way, we annealed fully pre-synthesized commercially purchased oligos to each other. In this scheme, each library shares one of the two constituent strands of the dsDNA, which we name as the 'common strand' of the library. The library generation protocol starts with the determination of the base sequence of the common strand, which ideally should not have a very strong secondary structure rendering the procession of the DNA polymerase or *in vitro* annealing attempts difficult, given that the sequence of the common strand represents the average oligo in the mixture. For the choice of the common strand, we rely on qualitative judgements based on the output of DNA folding softwares such as Mfold, whose output should contain only unstable conformations with a high

melting temperature [29].

The other strand of the library is almost identical to the reverse complementary sequence of the common strand, except that only one position contains a degenerate base containing all three bases other than the proper reverse-complementary. This means that if the common strand contains an A at a particular position, the variable strand can have V = { A, C, G } but not T. That is because incorporation of a T would yield proper dsDNA obeying Watson-Crick base pairing rule that is not a substrate of MMR. In a similar manner, to probe the mismatches at locations where the common strand contains a C, G, or T the variable strand would have H = { A, C, T }, D = { A, G, T }, and B = { C, G, T }, respectively. Each oligo representing a variable strand can carry a degenerate base B, D, H or V at one and only one position, but would otherwise exactly follow the reverse complementary of the sequence of the common strand. The inclusion of degenerate bases reduces the number of variable strands to be purchased by a factor of three, as each oligo can sample all three possible mismatches possible within that sequence context, forth possibility being forbidden as it leads to a dsDNA properly matched throughout the construct.

To be able to form circular plasmids by a DNA ligase, we formed compatible sticky ends at the termini of the mismatch libraries as on the termini of the barcoded vectors we generated (SacI and XhoI). For this, we opted to make the common strands longer by 4 nucleotides than the variable strand at both termini, where the unpaired sequence tetramers are complementary to the overhangs generated by the restriction enzymes ( 5' AGCT 3' and 3' AGCT 5'). To experimentally probe the repair properties of an n-nucleotide long sequence, it suffices to procure one n+8 nucleotid long common strand and n many n nucleotide long variable strands. For simplicity, we made an equimolar mixture of these n many variable ssDNA strands and annealed with the common strand by slow cooling from 98°C to room temperature on a thermal cycler in about an hour. We ligated thus formed mismatch library with the barcoded vector library to form circularized plasmids each carrying one unique random tag and one mismatch about 10 to 100 nucleotides downstream of the barcode.

### 1.3.5 Optimization of the choice of common strand sequence

Our protocol to generate mismatch-carrying plasmids involves a ligation step, hence requiring the presence of sticky ends at both termini. But this leaves the question regarding the choice of the consensus sequence to be used between these two restriction sites open. We conducted the initial attempts on DNA sequences available in our laboratory at the time of conception of the idea, and hence the sequence of choice does not carry any special mathematical or biological property. At later stages, we wanted to optimize the consensus sequence to provide a balanced coverage of mismatches and sequence contexts in the most labor- and cost-efficient way.

As an initial attempt, we targeted to measure the repair efficiency of all 8 types of mismatches within all possible nearest-neighbor sequence contexts. To achieve this, we sought a DNA sequence that contains each sequence trimer at least once to serve as the consensus sequence of the library. Since this sequence contains all 64 trimers possible with the 4-letter alphabet of DNA, it means all nearest neighbor combinations are contained within the library and that it suffices to anneal this common strand with variable strands that contain each possible base substitution at every position to represent 8 possible mismatches. We then aimed to obtain the shortest possible DNA sequence that represents all trimers. A simplistic approach could be to concatenate all trimers possible, through which one would obtain a $3 \cdot 4^3 = 192$-base long sequence. However, by considering the overlaps between the trimers, it is possible to obtain a shorter sequence that displays the same property. As an example, AAATA contains AAA, AAT and ATA trimers, but it is much shorter than the naive concatenation product AAAAATATA.

To be able to obtain the most compressed form of this sequence, we chose the consensus sequences of our libraries as the shortest sequence containing all sequence k-mers. Commonly attributed to De Bruijn, the most compressed sequence on an alphabet $\Sigma$ of size $||\Sigma|| = n$ containing all k-mers, B(n,k) is a cyclic sequence of length $n^k$ characters or $n^k + k - 1$ characters-long, if the former circular sequence is linearized. For our specific case sampling DNA base motifs, $\Sigma = \{A, C, G, T\}$, $n = 4$ and $k = 3$. We generated such De Bruijn sequences by concatenating Lyndon words in lexicographic order [30, 31]. Briefly, we generated a list of all necklaces of length n that are non-periodic, and concatenated the individual words into a super-string following Algorithm 1.

Here the Cyclic_DEB_Sequence is a cyclic sequence containing all sequence k-mers, Word denotes the constituting Lyndon words and incrementation is defined assuming base order A, C, G, T for DNA bases. To obtain a linear sequence still representing all k-mers (Linear_DEB_Sequence), we suffixed this circular sequence with the initial k-1 bases so that the k-mer at the break point of the cycle is still represented in the sequence.

**Algorithm 1:** Summary of the De Bruijn sequence generation procedure.

Initialize Cyclic_DEB_Sequence = ""
Initialize $Word$ = "A"
**while** $||Word|| > 0$ **do**
    Append $Word$ to Cyclic_DEB_Sequence
    $w \leftarrow Word$
    $x_i \leftarrow w_{i \ mod(||w||)}, \ \forall i = 1, 2, ..., n$
    $j \leftarrow min(k)$, such that $x_i = T, \ \forall i \geq k$
    $s_i \leftarrow x_i, \ \forall i = 1, 2, ..., j-2$
    $s_{j-1} \leftarrow x_{j-1} + 1$
    $Word \leftarrow s$
**end**
Linear_DEB_Sequence ← concatenate(Cyclic_DEB_Sequence[1,2,...,$4^k$], Cyclic_DEB_Sequence[1,2,...,k-1])
**Return** Linear_DEB_Sequence

Using this strategy, we obtained a 64 base long circular De Bruijn sequence shown in Figure 1.18a, which would be 66 base long in its linearized form and is possible to physically obtain using the contemporary technologies. However, we opted to cover this cyclic De Bruijn sequence by two different 60-base long chemically synthesized oligos that serve as common strands of two different mismatch libraries, as this arrangement was more cost-effective due to administrative reasons. The library DEB3L covers the first 60 bases out of 66 (TTT...CCG), whereas DEB3R covers the region omitted by DEB3L as its start position along the cycle is shifted (GCT...TCG). The two libraries overlap for most of their extent and hence provide a means to probe the repair efficiency in two different positions but within the same local sequence context. To avoid potential ligation related problems, and also to quantify stray event detections due to replication errors, the first and last 3 positions do not contain any mismatches in either library.

For each of these two sub-libraries, this approach requires purchase of 55 oligos including 1 common strand and 54 variable strands which contain one degenerate base to sample form a

mismatch upon annealing with the common strand. We mixed these 54 oligos and annealed to the respective common strand in two separate tubes, but pooled DEB3L and DEB3R immediately afterwards and performed the downstream steps on this mixed library. This implies that the two libraries encounter exactly the same conditions during the ligation, transformation, and sequencing library preparation, hence ruling out the effect of potential batch-to-batch variations on the final results. We deconvolved the sequencing data set during data analysis according to whether the obtained read resembles DEB3L or DEB3R consensus sequence more closely in Hamming distance and reported the results separately.

### 1.3.6 Algorithmic clan detection

After generating sticky ends using the two restriction sites also introduced by the primer pair, we ligated this linear barcoded vector to the short mismatch library generated by annealing chemically synthesized oligos. We transformed this plasmid library into electrocompetent K-12 *E. coli*, incubated the transformant bacteria overnight and extracted the plasmid library using a standard miniprep protocol. Out of this sample, we then amplified by PCR the region of interest, which comprises the tracing barcode, and the entirety of the sequence closely resembling the common strand, i.e. the inserted mismatch library, and sequenced using Illumina MiSeq platform. A typical such experimental attempt leads to $O(10^4)$ transformants providing $O(10^5)$ paired-end sequencing reads.

Using a custom-made C++ program, we imported all reads and organized them into groups that descended from the same ancestor plasmid by performing a density-based clustering (DBSCAN) on all obtained reads based on their tracing barcodes [32]. A pseudocode of DBSCAN is provided in Algorithm 2 and starts by looping over all unassigned reads one at a time and finds its immediate neighbors, which are defined as the barcodes in the obtained data set that are at most $\epsilon$ units away according to the provided measurement metric. If there are at least N many other data points in its $\epsilon$ neighborhood, the element is considered a core point and is used to seed a new cluster. Each neighbor node is added to the newly formed cluster center, and if any of the neighboring nodes are also core points themselves, their neighbors are added to the cluster. These recursive iterations

continue to recruit more nodes, till no other unprocessed neighbor nodes qualified to become a core point remains. With the inclusion of all core points and their neighbors, the current cluster is fully determined. The procedure continues by seeding other clusters at the remaining unprocessed core points till all core points have been processed, at which point the routine exits.

In practice, finding neighbors of a node is the most resource demanding step of this entire clustering procedure, which has a quadratic time complexity with respect to the number of nodes to process, as the barcode to barcode distance needs to be evaluated for each barcode pair in the dataset. Given the typical size of a data set (about 50 million reads per sample), repeated computation of the distances is CPU-intensive whereas the naive storage of a full pairwise distance matrix is similarly infeasible due to its memory intensive nature. To address these issues, we instead constructed a simple data structure which contains only a unique subset of all barcodes, and contains pointers to the duplicate barcodes which we detect by a pre-processing step involving a red-black tree [33]. It suffices to compute and store the distance matrix elements between these subset of unique barcodes, as the list of neighbors can easily be recovered by the union of all subset elements that are within the $\epsilon$ neighborhood and the duplicates they point to. As a second measure to reduce to computation time, we only compute the distance between elements that have sufficiently similar AT contents as the strong distance condition can hold only if this latter weaker condition holds. More precisely, for a given maximum distance threshold $\epsilon$, we perform an explicit pairwise distance computation only if the total number of A or T bases between the two barcodes differ by $\epsilon$ or less, or directly assign $d_{ij} = \infty$ otherwise. The neighbor finding procedure is outlined as a pseudocode in Algorithm 2.

The DBSCAN approach summarized above requires the computation of the $\epsilon$ neighborhood of barcodes. For this distance calculation between DNA barcodes, multiple metrics are possible. A rigorous approach could be to evaluate the number of total substitution or insertion/deletion mutations that needs to be made to reach from one barcode the other, which I will be referring as the edit distance. This necessitates the calculation of a global alignment by dynamic programming and is computationally much more costly. In the best case, a comparison between barcodes of length M=25 with Needleman-Wunsch algorithm would require populating a table that is M+1 by

**Algorithm 2:** Pseudo-code summarizing the barcode clustering algorithm employed.

```
Function DBSCAN(barcodeList, ε, N):
    neighborList ← FIND_NEIGHBORS(barcodeList, ε)
    corePts = {x ∈ barcodeList | ||neighborList_x|| > N}
    Initialize clanIndex to 1
    foreach x ∈ CorePts do
        if x has not been assigned to any clan then
            ADD_NEIGHBORS(x, neighborList, clanIndex, corePts)
            Increment clanIndex
        end
    end
return

Function ADD_NEIGHBORS(x, neighborList, clanIndex, corePts):
    Assign x to clan #clanIndex
    if x ∈ corePts then
        foreach z ∈ neighborList_x do
            ADD_NEIGHBORS(z, neighborList, clanIndex, corePts)
        end
    end
return

Function FIND_NEIGHBORS(barcodeList, ε):
    ATcontents ← Calculate A/T count of each barcodeList member
    Initialize neighborsList
    foreach x ∈ barcodeList do
        potentialNeighbors ← {z ∈ barcodeList| |ATcontents_z − ATcontents_x| < ε}
        foreach z ∈ potentialNeighbors do
            if distance(x, z) < ε then
                Append z to neighborsList_x
            end
        end
    end
return
```

M+1 that takes about 5 operations to populate each, taking about 3000 operations to compare two barcodes. This is because at each step, three possibilities need to be comparatively evaluated to find the locally optimal solution: a match/mismatch between current bases, insertion or deletion. A pseudocode of this approach is presented in Algorithm 3.

**Algorithm 3:** Determination of edit distance by dynamic programming

```
Function edit_distance(x, y):
    Initialize a (length(x)+1) by (length(y)+1) matrix table to 0
    Set table_{i0} = i
    Set table_{0i} = i
    foreach i ∈ {1, 2, 3, ..., length(x)} do
        foreach j ∈ {1, 2, 3, ..., length(y)} do
            if x_i = y_j then
                matchScore = table_{i-1,j-1}
            else
                matchScore = table_{i-1,j-1} + 1
            end
            insertionScore = table_{i-1,j} + 1
            deletionScore = table_{i,j-1} + 1
            table_{i,j} = min(matchScore, insertionScore, deletionScore)
        end
    end
    Report table_{length(x)+1,length(y)+1}
return
```

On the contrary, Hamming distance is computationally much more feasible, as only up to M nucleotide comparisons are to be made in the worst case scenario which occurs if the two barcdoes are identical up to the (M-1)st base (Algorithm 4). This approach does not tolerate any insertion/deletion (in/del) errors and rather considers all of the bases following an in/del as mismatched, which would in practice mean that some clans, in which certain members had undergone an insertion/deletion event, will be mistakenly considered as two separate clans potentially impacting the repair efficiency, more likely causing an increase than a decrease. While this error pattern has been reported to cause significant artifacts for certain experiments in the literature [34], we did not observe this to be a significant effect in practice for our case. In Figure 1.3, we analyzed the same experimental output corresponding to DEB3L and DEB3R libraries, and upon comparison of the results obtained by using the Hamming or edit distance metrics, we observed the results

to be very similar. We hence used in the remainder of this work exclusively Hamming distance for barcode clustering for computational efficiency, while detection of the adaptor, barcode and mismatch library positions still involved semi-global alignment as time cost of this operation grows only linearly with the size of the experimental data set.

**Algorithm 4:** Determination of Hamming distance

```
Function Hamming_distance(x, y):
    if length(x) ≠ length(y) then
        Return ∞
    end
    Initialize distance to 0
    foreach pos ∈ {1, 2, 3, ..., length(x)} do
        if x_pos ≠ y_pos then
            Increment distance
        end
    end
    Report distance
return
```

Because our choice of 25bp long tracing barcodes correspond to $4^{25} \approx 10^{15}$ different barcode possibilities, a group of plasmids sharing the same or highly similar barcodes are much more likely to be descendants of the same ancestor plasmid than sharing the random barcode by chance (Section 1.3.2). Hence each such cluster center that we detect using DBSCAN is essentially an equivalent of a colony on an LB-agar plate, which simply forms by the repeated expansion of a single bacterium at that spot during incubation. In fact, the vast diversity of the barcodes also provides a separation of barcodes from each other in the sequence space by a distance longer than a few mutations, thus enabling incorporation of tolerance to sequencing and replication errors. We achieve this by considering all tracing barcodes that are separated by at most $\epsilon = 3$ mismatches to belong to the same cluster, while we consider a cluster above noise level if there are at least N=10 different members in the cluster. In the sequel, we will refer to each such valid cluster center representing a group of reads corresponding to the same initial mismatch-bearing plasmid molecule a "clan", in analogy to its sociological counterpart that refers to "a group of people tracing descent from a common ancestor" according to the Merriam-Webster dictionary. Sequence composition of each clan tells us whether repair occurred or not by the time the first replication happens at a

single-molecule level detail.

### 1.3.7 Monitoring repair events by clan classification

In a cell, mismatches can result in mutations, as one of the two DNA polymerases will encounter an incorrect base sequence on its respective template strand during DNA replication. As such, the MMR and replicative mechanisms can be viewed as two competing processes. We define the repair efficiency of a mismatch as the fraction of all detected molecules that are repaired within the time window between the introduction of the plasmid into the cell and the arrival of the first replication fork to the mismatch. Under this definition where the first replication time is used as a stopwatch, the reported repair efficiencies do not necessarily reflect the absolute efficacy of a cell's repair response. Yet we believe this quantity is informative because the most mismatches will be generated during chromosome replication, and these mistakes need to be corrected before the next DNA replication event to avoid mutagenesis in the daughter cells. Although the clocks might tick at different rates between the chromosomal DNA and the extra-chromosomal plasmid, or between different loci on the same DNA based on its distance to the nearest replication origin, the clock will tick at about the same rate throughout the mismatch libraries we investigate, given that their end-to-end size is limited to 80 bases. Therefore, we consider this competition between repair and replicative processes a useful indicator of relative repair efficiencies of different mismatch sequences.

To decide if an individual clan is derived from a repaired or unrepaired ancestor plasmid, we determined a quantity we named "substitution prevalence" (s) which is the fraction of reads in a particular clan that carries the sequence of the variable strand rather than that of the common strand of the mismatch library, i.e.

$$s \equiv \frac{\#VariableStrands}{\#VariableStrands + \#CommonStrands} \tag{1.8}$$

If the mismatch remains unrepaired until the first replication, one of the two daughter plasmids inherits the consensus sequence via the common strand, whereas the other daughter plasmid receives the variable strand sequence. Subsequent rounds of plasmid replication would increase

the number of both plasmid variants, giving rise to a heterogeneous clan. Ideally, the strands of the plasmid will be independently replicated about equal number of times from there on, generating an equimolar final mixture of the common and variable strand variants of the plasmid ($\#VariableStrands = \#CommonStrands$). But while ideally half of the sequencing reads constituting the clan should carry the variable strand (s=0.5), in practice we expect a broader peak around this mean value due to binomial sampling and/or loss of information related to stochastic early stage cell death (Figure 1.4a, U-type clan). In contrast, plasmids that are repaired by keeping the common strand will carry no variable strand contribution at all, leading to a peak around s=0 (C-type clan). Similarly, a repair taking place by preserving the variable strand would result in a clan that carries exclusively variable strands, generating a peak centered around s=1 (V-type clan). While in the latter two cases, the repair should completely eradicate one of the two strands from the genealogy of that clan leading to peaks in the Dirac delta function ($\delta_{ij} = \{1, if\ i = j;\ 0,\ otherwise\}$, accidental mutagenesis during *in vivo* or *in vitro* DNA propagation as well as sequencing errors could still introduce a certain spread in practice, leading to peaks more resembling a beta function. Still, by a curve fitting routine for the three peaks centered around s=0, s=0.5, and s=1, we can quantify the relative frequency of repair and no-repair events in principle.

Figure 1.4 shows the substitution prevalence histogram (s-histogram) of clans with respect to the observed prevalence of the substitutions in a typical experiment, more specifically for a mismatch library that I will refer to as Single Mismatch Library (SML) from now on. The expected trimodal distribution of the s-histogram (Figure 1.4a) was observed from the experimental histogram of substitution prevalence values of all mismatches. Peculiarly, however, the peak corresponding to the V-type clans, i.e. those that were repaired by retaining the variable strand, was not centered at s=1, as we will discuss further later. Performing the experiment in $\Delta mutS$ cells [35], which are deficient in MMR response, caused a large increase the U type clans, i.e. those with intermediate s values (blue curve in Figure 2.2). In the sequel, we arbitrarily assume that the clans with a substitution rate between 0.1 and 0.9 to be unrepaired, or repaired otherwise (blue shading). I will improve on this assignment scheme in Chapter 2.

As a control experiment, annealing of two DNA oligos that are exact reverse-complementaries of each other forms a properly-paired dsDNA, ligation of which produces a circular plasmid that does not contain a DNA mismatch to trigger any repair (NPL). Regardless of the presence of a functional MMR, setting aside any replication or experimental errors, this experiment is expected to provide clans that represent the consensus sequence only, appearing as if there was a mismatch repaired by retaining the common strand. As expected, both wt and $\Delta mutS$ cells produced out of this non-mismatched DNA "library" s-histograms that only peaked at s=0. Overall, these analyses show that most of the repair or no-repair events we score indeed predominantly result from the MMR response of the cell against the intentionally introduced mismatches.

**Fine structures on s-histograms**

Despite the relatively high number of clans included in each bin ($\sim 100$ reads/clan), the substitution prevalence histograms contain well-distinguished peaks and troughs (Figure 1.4). These features are not due to random experimental fluctuations or artifacts, but are rather broadly applicable to other DNA libraries and are highly reproducible across experiments, as also shown for DEB3L in Figure 1.5. This observation can be partially accounted for by the statistics of integer division rather than pure measurement noise, as substitution prevalence is a ratio of two integers and the number of reads constituting a clan is typically on the order of 100. A second observation is related to the observed spread of the U-peaks, which in this case is broad enough to intermix with the C- and V-type peaks.

An accurate prediction of the expected fine features of the s-histograms is beyond the scope of this work and we did not pursue it experimentally any further, as it requires extensive characterization of the selection and sampling biases along the multi-step experimental procedure. However, as a simple approach, we can simulate two extreme scenarios related to where the most restrictive sampling occurs. If, following the transformation, plasmids are replicated with minimal bias without loss of any significant diversity and the major selective event is the choice of the subset of molecules to be sequenced at the very last step of the protocol (Figure 1.6e), the selection of the reads from the pool to make a U-type clan will resemble a binomial sampling procedure with

equal probability of choosing a common or variable strand and will result in a relatively sharper peak centered around s=0.5 (Figure 1.6b). If, however, there is another prior diversity-limiting sampling step in the workflow after which there is a form of unbiased amplification (Figure 1.6f), the peak of the U-type clans can become arbitrarily wide because the amplification of a drifted population should make a wider range of s-values accessible (Figure 1.6d). The s-histogram can be simulated using both binomial and uniform sampling approaches using the known clan size distribution of the sample and both approaches are capable of capturing the positions of the crests and troughs, even though not their magnitude (Figures 1.6c). In summary, both the fine features of the s-histograms and the wide spread of the peak are plausible outcomes of the experimental system. The s-histograms in the remainder of this text will not display these fine features for visual clarity (e.g. Figure 1.18b).

### Noise due to extra-cellular DNA

Our experiment generally records a high basal level of repair efficiency, source of which we could not decipher by this time. We observed this also to be the case for $\Delta mutS$ cells, in which MMR is not expected to be operational, suggesting that it can be a real reading reflecting the contribution of other non-canonical pathways. On the other hand, an erroneously high repair efficiency can be recorded out of unrepaired clans due to a failure to sequence variable or common strand components of a U-type clan containing both. One way that such a high background repair level emerges could be due to cell-free plasmids in the culture tube that are either released from the bacteria upon cell death or could be untransformed plasmids that are leftovers of electroporation. If they escape digestion, such stray plasmids could inflate the measured counts of C- and V-type clans, as the number density of these stray plasmids will very likely be lower than those that are maintained intracellularly at high copy numbers per cell and in a high number of daughter cells.

We therefore sought to comparatively check the relative quantity of plasmids within the cells vs. the available quantities in the growth medium that can participate in the NGS library preparations by a simple PCR-based test (Figure 1.7a). Following the recovery step of electroporation, we took aliquots from the growth medium at 10min, 1h, 3h or 16h timepoints. We centrifuged the aliquots

to separate the cell pellets from the growth medium, which we then re-suspended up to equal volume. We then used these samples as templates for a PCR and compared the quantity of the product generated by the band intensity after agarose gel electrophoresis. Our results suggest that while amplifiable plasmids can be detected in both fractions at any given time point if amplified thoroughly (40 cycles PCR in Figure 1.7b), the cell-free plasmids minimally contribute to the NGS libraries at lower thermal cycling that we normally use (20 cycles PCR). This suggests that our repair efficiency measurements are not significantly deflected by such stray molecules of low abundance.

### s-histograms of MMR mutants

We investigated the repair response of mutant cell strains where the only chromosomal copy of the one of the MMR pathway elements is replaced by a kanamycin cassette and asked if there is any variability between the clan substitution histograms of different MMR pathway mutants indicative of a difference in the overall residual repair response. Figure 1.8a displays the observed histograms obtained using our simplified library construction scheme by oligo annealing for various mutants. Whereas a deficiency in MutS or MutL lead to the emergence of a significant clan population that arise from unrepaired ancestral plasmids (intermediate s-values), the absence of MutH or UvrD, which operate downstream to MutS and MutL in the MMR pathway, lead to a less populated middle peak indicating a higher residual repair activity.

While the misclassification of U clans as C or V due to binomial sampling error is a possible reason that could increase the apparent repair efficiency of a mismatch, the extent of such an effect is expected to become more pronounced if the number of reads per clan (i.e. clan size) is lower. However, we did not observe such a parallel trend in the clan size distributions across these mutants that could explain the gradual disappearance of the impaired repair capability as the deleted MMR member is farther downstream (Figure 1.8b). Thus, we think that the different level of repair impediment posed by different mutations we investigated might be related to the degree to which these proteins are substitutable by alternative proteins in the cell. If true, this is an indication that the detection of the mismatch by MutS and MutL is much more essential

than the execution of the repair response itself, potentially because of the existence of alternative pathways that can repair the DNA once the mismatch is recognized, potentially delaying the initiation and/or progression of the plasmid replication.

Using the same SML library, we also compared the repair efficiency of wt cells with $\Delta uvrB$ cells, a deletion which is expected to impair the nucleotide excision repair pathway (reviewed in [36]) and observed that the repair efficiency is largely unaffected by the obstruction of this pathway (Figure 2.20c), suggesting that the repair response we measure is not dominated by the nucleotide excision repair response.

### 1.3.8 Calculation of individual repair efficiencies

Although our assay does not keep track of individual plasmids or their host cells by physical location or confinement, the unique DNA barcodes introduced to the system provides a means to distinguish the replication products originating from different ancestral plasmids. This property also enables the simultaneous assessment of a diverse set of mismatches, as the original mismatch type can be traced and inferred from the sequence composition of the corresponding clan obtained by clustering. In order to deduce the mismatch type and position on the corresponding ancestor plasmid among the starting material, we check the ensemble averaged sequence of the clan and compare with that of the consensus sequence of the mismatch library. The average sequence of each clan closely follows that of the sequence prototype of the DNA library used. While the error rates of *in vivo* and *in vitro* DNA amplification and sequencing will introduce deviations from the sequence of the common strand, they will have arbitrary base identities and their location in each clan member will also vary and hence averaging over multiple molecules will diminish their contribution to the overall substitution prevalence of the clan. On the contrary, the substitutions observed due to a mismatch on the original molecule will not cancel and lead to a significant substitution prevalence at the position of the by-then mismatch.

If the mismatch repair system removes the variable strand before the first replication event takes place, the ensemble averaged sequence of the clan will be indistinguishable from the consensus sequence. Therefore, while the C-clans can be easily detected by their low s parameter, regardless

of the identity of the mispaired bases or their position, the average sequence of the clan will closely follow the consensus sequence of the provided mismatch library, hence rendering the inference of the original single plasmid molecule giving rise to the detected C-clan impossible. In contrast, the emergence of deviations from the base sequence by an amount above the noise level suggests the presence of a mismatch at that sequence position in the ancestor molecule. We can hence guess the position of the mismatch by computing the difference matrix (d) between the actual base composition histogram of the clan and the expected base composition histogram (c) based on the known sequence of the consensus sequence, and report the position with the highest value. That is,

$$d_{ij} = \sum_{\forall k \in clan} \left( \delta_{i,k_j} - \delta_{i,c_j} \right) \tag{1.9}$$

and

$$p \equiv \underset{j}{argmax}\ d_{ij} \tag{1.10}$$

where $d_{ij}$ represents the difference matrix element for nucleotide position j=1,2,3,...N and base type i={A, C, G, T}; $c_j$ is the base identity of the consensus sequence at position j, and the summation is over all sequencing reads k that have been assigned to the clan by clustering. The mispaired bases constituting the mismatch can be deduced in an analogous way, as the shared sequence of the consensus strand is known *a priori* and the variable strand should carry the complementary of the observed substitution as seen by the common strand's frame or *vice versa*. The most likely base identity of the variable strand would be the most common substitution observed at mismatch position p, i.e.

$$b \equiv \underset{i}{argmax}\ d_{ip} \tag{1.11}$$

where the mismatch would form between b and the base complementary to $c_p$ ($c'_p$). As an example, if the most commonly substituted position of the clan dominantly has A's in the reads, whereas the consensus sequence contains a C, this means the common strand likely had a G at that particular position, whereas the variable strand had an A, and the mismatch was of AG type.

By the design principle of the mismatch library, at most one position and mismatch type per clan can deviate from the consensus sequence above significance level, i.e. the pair (p,b) is unique if it exists. We discard the clans that contain a scrambled sequence violating this assumption due to experimental errors.

Figure 1.9 exemplifies the expected event types and their interpretation using the scheme described in this section for the DEB3L library. For all the cases, the sequence indicated along the x-axis is that of the common strand that is identical for all clans by the design underlying the mismatch library. The variable strands deviate from the reverse complementary of the common sequence at one arbitrary position, but obey the Watson-Crick base pairing otherwise. The y-axis denotes the base identity of the variable strand that was observed at the indicated deviation position, which we obtained by subtracting the expected base composition histogram that consists of 0's and 1's based on the consensus sequence of the library from the actual base frequency distribution at each position calculated over the clan members. Each such histogram represents 1 clan, and the red tones indicate the detection of a base that is different from the consensus sequence of the variable strands whereas the blue tones correspond to reduced prevalence of a base that would have been expected at a position based on the consensus sequence. In the rest of the text, I will be presenting our single-mismatch level data in this matrix format.

In Figure 1.9a, we do not observe any significant negative or positive peaks in the difference histogram, which can happen if the clan members closely follow the consensus sequence. The original mismatch leading to this C-type clan cannot be deduced, hence this clan will have to be omitted. The blue pixel in the T-row of 1.9b is at a position where the common strand carries an A, based on which one normally would expect to observe a T. Instead, it was commonly substituted by a C as indicated by the red pixel, likely an indication of an AC mismatch. As the substitution was observed in about half of the clan members based on the color saturation levels, this clan was probably unrepaired, i.e. a U-type clan. Similarly, 1.9c indicates an unrepaired AG mismatch whereas 1.9d and 1.9e indicate TT and AC mismatches that were repaired by retention of the variable strand, i.e. V-type clans. Due to wrong base calls during sequencing or DNA amplification errors, it is possible to obtain clans that do not obey any of these event prototypes

(e.g. 1.9f and 1.9g), which I will be considering as noise and excluding from further analyses.

We assert whether a repair event has taken place on the $b - c'_p$ mismatch at position p, based on the prevalence of the most commonly observed substitution in the clan in accordance with the principle explained in detail in Section 1.3.7. If a repair event happens via retention of the variable strand (V-type clan), the substitution is observable in above threshold fraction of the clan members ($d_{bp} \geq t_{high} = 90\%$) and we record a repair event by incrementing the matrix element $V_{bp}$ counting the V-type clans for this position and mismatch type by one. Similarly, an unrepaired mismatch would have an intermediate substitution prevalence ($10\% = t_{low} < d_{bp} < t_{high} = 90\%$) and we would increment the corresponding matrix element $U_{bp}$ that counts the U-type clans. With this approach, clans displaying substitution among 10% or less of its members also represent a repair event, but using the common strand as the correct information resource rather than the variable strand. However, as I argued before, it is not possible to deduce the mismatch type and position in the ancestor plasmid solely based on this information and hence we disregard all such clans for further analyses. To compensate for this systematic loss of the repaired clans, we assume that the strand choice during repair is fully random and hence that the number of observed C and V-type clans should be about the same for each mismatch. Under this assumption, we define a proxy to the repair efficiency ($\eta$) as,

$$\eta'_{ij} \equiv \frac{2 \cdot Vij}{2 \cdot V_{ij} + U_{ij}} \tag{1.12}$$

## 1.4 15 600 repair efficiency measurements

Figures 1.10, 1.11, 1.12, 1.13 and 1.14 tabulate, in a heat map, the measured repaired efficiencies ($\eta'$) of mismatches as part of different consensus sequences tested under various conditions, defined as in Equation 1.12. In each plot, the base along the x-axis indicates the sequence of the common strand while the four rows show the base the variable strand contains at that position forming a mismatch, i.e. the rows from top to bottom represent the cases where the variable strand contains an A, T, C, or G, respectively. As an example, the red colored pixel on the FML output matrix (topmost) at the 5th column and G-row refers to a low repair efficiency for an AG mismatch that

is formed by an A on the common strand and a G on the variable strand and is located 5 bases after the last base of the mapping barcode. While four different bases can be incorporated at each position, only 3 out of 4 will lead to a mismatch, and the remaining fourth entry corresponding to a non-mismatch is marked by gray hatching as it does not correspond to a measurable quantity. A visual judgement of the figures suggests that while the majority of the mismatches were repaired at a very high efficiency ($\eta' \approx 1$, indicated by white and yellow tones), some mismatches were repaired with a noticeably lower efficiency ($\eta' \approx 0.3$, crimson tones).

The confidence level of the measurements will depend on the the total number of clans detected with the same mismatch at that particular position ($U_{ij}$+$V_{ij}$), whether it was repaired or unrepaired and hence is provided along with the repair efficiency matrices in the figures, where observing higher values is a likely indication of narrower error margins. To see this, we can propagate the errors in the measurement in Equation 1.12 as follows,

$$\delta\eta'_{ij} = \frac{-2V_{ij}}{(U_{ij}+2V_{ij})^2}\delta U_{ij} + \frac{2U_{ij}}{(U_{ij}+2V_{ij})^2}\delta V_{ij} \tag{1.13}$$

Or assuming independence of the counting statistics of the U and V-type clans, we can get

$$\Delta\eta'_{ij} = \frac{2\sqrt{V_{ij}^2(\Delta U_{ij})^2 + U_{ij}^2(\Delta V_{ij})^2}}{(U_{ij}+2V_{ij})^2} \tag{1.14}$$

Approximating the uncertainty in the measurement by the standard deviation and assuming that the clan counting statistics roughly follow Poissonian statistics, we can then assume that $\delta x = \sigma_x = \sqrt{x}$ and it follows that

$$\sigma_{\eta'_{ij}} = \frac{2\sqrt{U_{ij}V_{ij}(V_{ij}+U_{ij})}}{(U_{ij}+2V_{ij})^2} \tag{1.15}$$

We can verify that these results reflect our expectations on the error margins. Firstly, the accuracy in the measurement increases in parallel with the number of clans detected, since $\lim_{U\to\infty} \sigma_{\eta'} = \lim_{V\to\infty} \sigma_{\eta'} = 0$. Secondly, as $U_{ij}$ and $V_{ij}$ are all non-negative, an experimental error leading to an aberrant over-abundance of unrepaired clans ($\delta U_{ij} > 0$) or under-representation of

repaired clans ($\delta V_{ij} < 0$) result in under-estimation of the repair efficiency ($\delta \eta'_{ij} < 0$).

Depending on the sequencing depth and the transformation efficiency of the particular sample, we typically detected about 100 clans per each mismatch in a typical experiment. As a control measuring the frequency of events erroneously detected due to experimental errors, a few of the terminal positions do not carry any mismatches by design and hence no clan detection is expected corresponding to those entries on the matrices. Although some clans were still detected due to the experimental errors, we observed the frequency of such stray events to be much lower than the real signal level. These observations corroborate our opinion above that the error rate of our system is sufficiently low and the clans detected indeed mostly originate from the mismatches that we introduced.

As a global trend shared among most of the data sets, we note that a wild type (wt) cell is capable of repairing the majority of the mismatches with an average apparent efficiency of 77% and this repair capability was reduced down to 43% in ΔmutS and 46% ΔmutL strains. However, we observe the rate of repair to be relatively closer to wt, in the absence of either the nicking enzyme MutH (68%) or the helicase UvrD (72%), both of which operate downstream of MutS and MutL (Figure 1.15). As a further control, we also observed repair efficiency levels comparable to those in wt in a ΔuvrB strain, which is a part of the nucleotide excision repair pathway (reviewed in [37]) and observed the average repair efficiency to be 79%, deviation of which is within the error margin.

### 1.4.1 Location dependence of the repair efficiency

As a general trend in our datasets, we systematically observed the highest repair efficiencies at the mismatches closer to the ligation site adjacent to the barcoded region on our libraries (barcode-proximal side) compared to those farther away from this region (barcode-distal side). This trend especially reveals itself in Figure 1.14, as pixels on the left side harbor more yellow tones whereas the right side is dominated by redder tones. To facilitate the observation of this global trend, I replotted our experimentally obtained $\eta'$ values in Figures 1.16 and 1.17 as a function of the mismatch position with respect to the barcode-proximal ligation site. This representation disregards the

type of the mismatch and represents the repair efficiencies along the y-axis, rather than the color scale of the matrix representation I employed before, while traversing the x-axis from left to right means that the mismatch is positioned farther and farther away from the barcode as before. The monotonically decreasing global trend can be visually verified by the help of the gradual decrease in solid lines, which are the best-fitting 4th degree polynomials representing the respective data points.

As of the time of the submission of this dissertation, a clear mechanistic explanation of this gradual trend is still elusive. However, we observe that this global trend exists to a certain extent in all data sets that we have generated to varying extents. Firstly, a downwards trend is observable in both wt cells as well as various MMR pathway mutants that we have investigated. Secondly, we observe this phenomenon to be a sequence dependent feature as two libraries that experienced the same reaction conditions displayed drastically different trends. More specifically, both in wt and $\Delta mutS$ cells, DEB3R library has a significantly sharper drop in the apparent repair efficiencies than DEB3L, despite the fact that the two libraries were mixed in equal stoichiometry at the beginning of the experiment and were handled from that point on as a single mixed specimen. Strikingly, the sharpest drop happens within a roughly 10bp long middle segment that is rich in GC content. These suggest that the observed trend is not a phenomenon related to the response of MMR *per se*, but rather either it is a systematic measurement error, or it is related to other directional cellular response mechanisms against DNA such as nucleases that digest processively. Both of these two appear to play a role, which is what I briefly discuss next.

**The presence of nicks affect the apparent DNA repair efficiency**

In practice, nicks or gaps can arise on the introduced plasmids due to incomplete ligation between the barcoded vector and the mismatch library, as well as due to oxidative DNA damage during sample handling [38, 39]. The former effect can be experimentally exacerbated by treating the vector with an alkaline phosphatase before ligation, which removes the phosphate groups at the 5' termini and hence hinders the formation of two out of four required phosphodiester bonds at the ligation step as both T4 and T7 ligase employed in the study require 5' phosphate groups on their

substrates [40]. Our results led us to suspect that at least part of the above position dependence could be related to the potential DNA nicks on the transformed plasmids, as the decline in the repair efficiency became steeper when the nick formation was forced by the phosphatase treatment, regardless of the presence of a functional MMR (Figures 1.16a vs 1.16b).

For the other extreme, we also tried to exclude the contribution of the nicked or unligated molecules to the measurements, by including an exonuclease digestion step in the sample preparation workflow. If, following the ligation, an exonuclease is introduced to the sample, the molecules that are covalently closed will be protected from digestion as there is no exonuclease entry site, whereas some exonucleases can start digestion at the nicks. However, we continued to observe a similar gradually decreasing trend upon treatment with T5 exonuclease or simultaneous treatment with exonucleases III and VII (Figures 1.17a, 1.17b, 1.17c, 1.17d).

### Insertion orientation

We considered that the observed position dependence effects could be a real phenomenon caused by the proximity of certain sequence features on the vector itself. The *E. coli* MutH generates the nick at GATC sequence motifs, whose uneven distribution on the plasmid could have an impact on the apparent repair efficiencies. Figure 2.7b shows the position of the GATC distributions along the tUNC19 plasmid as DpnI cut sites, according to which the nearest GATC site is 550bp upstream to the barcode-proximal end of the library.

We hence tried flipping the insertion orientation of the mismatch library by switching the restriction sites on the barcoded vector, which can be achieved by changing the primer pair used for the production of the barcoded vector. This rearrangement brings the barcode distal side of the mismatch library with below average repair efficiencies closer towards the barcode where the measured repair efficiencies were previously higher. After this inversion we still observed a similar decay pattern in the original direction from the new barcode proximal side towards the barcode distal side, again the transition in DEB3R being sharper in the vicinity of the GC-rich center of the library as before (Figures 1.17a and 1.17b). Observation of such similar global trends in both insertion orientations might suggest that this effect is unlikely to be a coincidence due to a

difference in the consensus sequence of the all mismatch libraries of concern, but rather that the relative positioning with respect to the plasmid is of importance.

**Position dependent changes in the s-histograms**

Above, we noted a systematic difference in the repair efficiency of mismatches at the barcode-proximal and barcode-distal sides. As the repair efficiency is calculated by making use of the counts of C and V type clans, we asked if there is a position-dependent difference in the properties of the repaired or unrepaired clans and indeed, we observed a clear position dependence on both the characteristics of the sub-populations, as well as their level of occupancy. In the histograms, we expect 3 sub-populations (C,U,V). While the unrepaired clans have intermediate substitution values forming a widespread bell curve centered around s=0.5, the clans in which repair occurred by retaining the common strand (C) form a peak around s=0. We expected and observed this to hold for all mismatches, though the occupancy levels depended on the mismatch of concern. While we similarly expect the clans retaining the variable strand (V) to have a peak around s=1 for all mismatches, we observed this not to be the case. Instead, we found the V-peak to be broader than the C-peak, and that its center position to shift from around s=1 towards lower values monotonically as the mismatch position is varied from the barcode-proximal to the barcode-distal end of the inserted library.

DEB3 libraries differed in their drop and indeed, their substitution prevalence histograms presented in Figure 1.18c also differed. While DEB3L displayed a bimodal distribution with a well-defined C-type clan peak centered around s=0.1, the V-peak of the DEB3R was slightly shifted towards lower substitution prevalences than that of DEB3L indicating presence of two V-type clan subpopulations. DEB3R clans that are not in the C-peak (s >0.3) also have a notably more negative skewness (1.67 vs 0.84, p-value $<10^{-5}$ by one-tailed z-test). These variations are much more pronounced than the experimental variations, as the two curves representing two experimental duplicates lie close to each other in both cases. We observed this different behavior of DEB3L and DEB3R to be mostly attributable to the barcode-distal side of these libraries, as the equivalent histogram exclusively for the barcode-proximal 20 positions do not display this discrepancy, but

the barcode-distal 20 positions have well-separated V-peaks (Figures 1.18b and 1.18d). This observation corroborates our hypothesis above that this phenomenon is a highly sequence dependent feature as DEB3L and DEB3R libraries that experienced same reaction conditions still lead to different sub-populations.

As such effects could be contributed by cellular enzymatic machinery that attacks the introduced foreign plasmids via the nicks, we compared the s-histograms obtained by adding either T5 exonuclease alone or the combination of exonucleases III and VII. Starting from the nick, T5 exonuclease digests in the 5'→3' direction only, whereas the latter mixture is expected to lead to a more complete elimination of all nicked molecules as exonuclease III can digest 3'→5' and exonuclease VII is bidirectional and is able to attack circular single stranded DNA [41, 42]. The two different enzymatic procedures to remove nicked molecules did not lead to observable changes in the V-peaks of the barcode-proximal mismatches for DEB3L or DEB3R. However, exoIII/VII coctail caused a positive shift in the C-peak as well as increasing its spread. For the mismatches close to the barcode-distal end, the difference between two enzymatic treatment scenarios was minimal for DEB3L, whereas the reported strand choice bias of DEB3R significantly leaned towards the retention of the C-strand if T5 exonuclease was included, a tendency which got inversed by the exoIII/VII coctail.

While these observations suggest that the presence of DNA defects such as nicks might be an important confounding factor against the absolute quantification of repair efficiency, the potential experimental workarounds we have attempted to incorporate failed to address the observed features in our data set to a satisfactory level. Neither of the two experimental scenarios eliminated the asymmetry between the C- and V-peaks at the barcode-distal end, the contrast between the two mismatch libraries also remained. These results might mean that nick-initiated repair is only partially contributing to $\eta'$ or that none of the options we attempted were effective enough to fully eliminate nicked ligation products. In the remainder of the text, we hence opted not to include any exonucleases in the protocol. I will provide more experimental evidence to address this position dependence more in Chapter 2, where I will also attempt to remove this confounding factor at the data analysis stage. For the rest of this chapter, my focus will instead be on convincing ourselves

that our assay system reproducibly detects the mismatches that are repaired with an efficiency above or below average.

**Deduction of mismatch position is accurate**

Above, I claimed that it is possible to measure the repair efficiency of a particular mismatch in parallel out of a mixture containing a multitude of different mismatches in different sequence contexts. This system lead to some unexpected observations such as a reproducible global decline in repair efficiency from the barcode-proximal towards the barcode-distal end of the mismatch libraries stemming from a mechanistically not explicable behavior of the V-type clan peak in the s-histograms. While these effects might as well be real, we asked if it can be due to a potential data interpretation issue caused by the misinterpretation of the clans out of a complex mixture, either at the experimental or data analysis stage.

To verify if the co-measurement of different mismatches introduces such systematic biases, we designed a control experiment that makes use of 6 different thermally annealed mismatch-bearing specimen, each of which harbors only one kind of mismatch at an *a priori* known location. Since all six samples contain only a particular mismatch, it is possible to accurately infer the mismatch on the ancestor molecule of C-clans. As an example of a mismatch repaired with a low efficiency (low=L), we chose a CC mismatch out of the SML library at the 29th position from the barcode-proximal end, whereas as a mismatch repaired with high efficiency (high=H), we picked a CT mismatch that is at the 13th position. To sample the long scale effect of mismatch location with respect to the functional elements on a plasmid, we applied circular permutations on the SML sequence in such a way that these two mismatches are placed at around position 10 (proximal=P), 30 (mesial=M) or 60 (distal=D). This arrangement keeps the local sequence context similar across the P, M or D cases, while spanning the spectrum of different mismatches that we typically studied (Figure 1.20a).

For all 6 of such control specimens and using the same analysis technique we have used for the mismatch library cases that is based on the maximally substituted nucleotide in the clan's base frequency histogram, we could accurately detect the position of the mismatch we have incorporated

in (Figure 1.20b), as well as deducing the identity of the bases constituting the mismatch (Figure 1.20d). This suggests that without the *a priori* knowledge of the expected mismatch out of a large pool, it is possible to reliably infer them solely based on the base composition histogram. For each of the two mismatches, we compared their s-histograms at the three locations, and observed the repair efficiency of the CC mismatch to be lower around the barcode distal end (LD in Figure 1.20c) in comparison to the vicinity of the barcode-proximal end (LP). The repair efficiency of the CT mismatch was also lower around the barcode-distal end (HD in Figure 1.20e), but in contrast to the CC mismatch, the characteristics of the V and C-peaks were also position dependent: HM and HP contained bimodal V-peaks and both peaks shifted gradually away from the extreme limits of s. These observations suggest that the gradual trends we reported in Sections 1.4.1 and 1.4.1 are not a simple by-products of mismatch-type calling out of a large pool.

### 1.4.2 Detection of mismatches repaired with low efficiency

#### Description of our ternary classification method

As was discussed in Section 1.4.1, for reasons that are not fully clear by the time I am submitting this dissertation, we observed a high-position dependence in our clan histograms, which in turn leads to a potential position dependence of $\eta'$. Due to the presence of such global trends, a direct quantitative comparison of repair efficiencies of mismatches located at two well-separated neighborhoods might not be very reliable. The absolute measurements should also depend on the amplification biases and the number of distinct replication products sequenced per each original plasmid. The apparent repair efficiency that would be obtained from a data set with insufficient sequencing depth is expected to overestimate $\eta'$, as some of the U-type clans will be misclassified as V-type clans due to the statistical likelihood of losing the common sequences as part of random binomial sampling. While sampling error cannot cause misdetection of any V-type clans as U-type clans, this kind of measurement error can be introduced by sequencing errors. Any factor that causes a change in the strand choice bias (i.e. C to V type or *vice versa*) will similarly confound the measurements as C-clans are discarded during data analysis (Section 1.3.8). The presence of a global trend in $\eta'$ also means direct pairwise comparisons of $\eta'$ will yield an unrealistically high

estimate of the correlation coefficient.

While such confounding factors make the absolute quantitative comparisons unreliable and pose difficulties in their interpretation, they typically do not cause an obstacle against determination of the subset of outlier mismatches that are repaired much more or much less efficiently than the others. To circumvent these shortcomings, we implemented a ternary classification scheme, instead. To achieve this, we first determined the general trend in the repair efficiencies as a function of the base position with respect to the barcode-proximal end of the library, which we estimated by a least-squares fit to a 4th degree polynomial (black curves in Figure 1.21). Next, we assigned the mismatches significantly above or below the general trend to the High and Low efficiency classes, respectively. We considered the points that roughly follow this average level as members of the Medium repair efficiency class. For the aforementioned classifications, we computed the mean deviation of all points above or below this global trend curve and chose half of this average deviation as the significance cutoff. An experiment with infinite resolving power can very accurately and consistently classify the mismatches, and the classes assigned by two independent experimental replicates to the same mismatch should be identical, hence constituting a measure of correlation and/or reproducibility (Figure 1.21).

**Pairwise comparison of assigned classes**

To assess the strength of the correlation between two experimental data sets, we looked for the consistency between the assigned ranks to the same mismatches in their respective 3x3 rank ordering matrices in the form of a confusion matrix. In this representation, the comparison of two fully correlated datasets is ideally expected to produce a diagonal matrix, whereas two uncorrelated measurement series would randomly populate all nine entries to about the same level generating a roughly flat matrix. In our measurements, the assigned ranks are expected to correlate with each other if the repair efficiency is influenced by the physical characteristics of the local neighborhood, confounded by measurement noise. To quantify the significance of the repeatability of the rank ordering against the stochastic noise, we resort to the mean square contingency coefficient (a.k.a.

$\phi$ coefficient) [43] defined as

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{0\bullet}n_{1\bullet}n_{\bullet 0}n_{\bullet 1}}} \tag{1.16}$$

where $n_{11}$ and $n_{00}$ represent the number of true positive and true negative event counts detected, $n_{10}$ and $n_{01}$ are the false positive and false negative event counts. $\bullet \in \{0, 1\}$ implies summation over the two states, e.g. $n_{\bullet 1} = n_{11} + n_{01}$. If the algorithm throws an independent random output at each iteration, $n_{00} = n_{01}, n_{10} = n_{11}$ and hence $\phi = 0$, whereas a perfect agreement of the two data sets should yield exactly $\phi = 1$, since $n_{01} = n_{10} = 0$. In the case of an ideal anti-correlation, $n_{00} = n_{11} = 0$, and therefore $\phi = -1$. Experimental measurements could be anti-correlated, uncorrelated or positively correlated and thus $\phi \in [-1, +1]$, depending on the relative noise level. We extended this two-state statistical measure to assess the capability of reproducing the class assignments, where $\phi_k$ measures the ability to accurately detect the set of mismatches with rank=k against the other two classes and given by,

$$\begin{aligned}
\phi_1 &= \frac{n_{11}(n_{22} + n_{23} + n_{32} + n_{33}) - (n_{21} + n_{31})(n_{12} + n_{13})}{\sqrt{n_{\bullet 1}n_{1\bullet}(n_{2\bullet} + n_{3\bullet})(n_{\bullet 2} + n_{\bullet 3})}} \\
\phi_2 &= \frac{n_{22}(n_{11} + n_{13} + n_{31} + n_{33}) - (n_{12} + n_{32})(n_{21} + n_{23})}{\sqrt{n_{\bullet 2}n_{2\bullet}(n_{1\bullet} + n_{3\bullet})(n_{\bullet 1} + n_{\bullet 3})}} \\
\phi_3 &= \frac{n_{33}(n_{11} + n_{12} + n_{21} + n_{22}) - (n_{13} + n_{23})(n_{31} + n_{32})}{\sqrt{n_{\bullet 3}n_{3\bullet}(n_{1\bullet} + n_{2\bullet})(n_{\bullet 1} + n_{\bullet 2})}}
\end{aligned} \tag{1.17}$$

where $\bullet \in \{1, 2, 3\}$ implies summation as before. Figure 1.21 exemplifies the usage of this criterion as applied to our system. To evaluate the level of significance, a $\chi^2$ test was employed for 1 degree of freedom, since, $\phi = \sqrt{\frac{\chi^2}{N}}$, where N denotes the total number of elements included in the set, or equivalently, thrice the number of positions included. One can observe this relation by an analogous argument to [44]:

$$\begin{aligned}
\chi^2 &= \sum_{ij} \frac{(n_{ij} - n_{\bullet j}n_{i\bullet}/N)^2}{n_{\bullet j}n_{i\bullet}/N} \\
&= \frac{(Nn_{00} - n_{\bullet 0}n_{0\bullet})^2}{Nn_{\bullet 0}n_{0\bullet}} + \frac{(Nn_{11} - n_{\bullet 1}n_{1\bullet})^2}{Nn_{\bullet 1}n_{1\bullet}} + \frac{(Nn_{01} - n_{\bullet 0}n_{1\bullet})^2}{Nn_{\bullet 0}n_{1\bullet}} + \frac{(Nn_{10} - n_{\bullet 1}n_{0\bullet})^2}{Nn_{\bullet 1}n_{0\bullet}}
\end{aligned} \tag{1.18}$$

We can rearrange this equation as,

$$
\begin{aligned}
&= \frac{(Nn_{00} - n_{\bullet 0}n_{0\bullet})^2}{Nn_{\bullet 0}n_{0\bullet}} + \frac{(N(N - n_{\bullet 0} - n_{\bullet 0} - n_{00}) - (N - n_{\bullet 0})(N - n_{0\bullet}))^2}{Nn_{\bullet 1}n_{1\bullet}} + \\
&\frac{(N(n_{0\bullet} - n_{00}) - n_{\bullet 0}(N - n_{0\bullet}))^2}{Nn_{\bullet 0}n_{1\bullet}} + \frac{(N(n_{\bullet 0} - n_{00}) - (N - n_{\bullet 0})n_{0\bullet})^2}{Nn_{\bullet 1}n_{0\bullet}} \\
&= \frac{(Nn_{00} - n_{\bullet 0}n_{0\bullet})^2}{N}\left[\frac{1}{n_{\bullet 0}n_{0\bullet}} + \frac{1}{n_{\bullet 1}n_{1\bullet}} + \frac{1}{n_{\bullet 0}n_{1\bullet}} + \frac{1}{n_{\bullet 1}n_{0\bullet}}\right] \\
&= \frac{(n_{00}n_{11} - n_{10}n_{01})^2}{N}\frac{N^2}{n_{\bullet 0}n_{0\bullet}n_{\bullet 1}n_{1\bullet}} = N\phi^2
\end{aligned}
\tag{1.19}
$$

The minimum required level of $\phi$ to reach statistical significance is provided in Table 1.3.

**Assessment of experimental reproducibility via comparison of assigned classes**

Despite random fluctuations and the site-to-site variations in the absolute quantities of the measured repair efficiencies, our system has a reproducible statistical power to distinguish outlier mismatches that are easy or difficult to repair as measured by high positive values of the respective $\phi$-coefficients.

First, we used this comparative measure to assess the effect of various genetic backgrounds on apparent repair efficiency $\eta'$. Even if a system does not repair the introduced mismatches, some of the mismatches might be detected as inefficiently repaired mismatches due to random fluctuations caused by sampling errors, as such we expect the experimental class assignments to become a random process if an introduced mutation leads to a total loss of the MMR capabilities. Upon comparison of datasets obtained using SML, we indeed observed a very low correlation between outliers detected in wt cells and $\Delta mutS$ or $\Delta mutL$, suggesting an impairment of the repair process (Figure 1.22a). Deletion of *uvrB*, reported to be a member of nucleotide excision pathway but not MMR pathway, provided a similar subset of inefficiently repaired mismatches as wt, suggesting that the mismatch repair response was largely unaltered. Surprisingly, however, the outliers detected in $\Delta mutH$ and $\Delta uvrD$ *E. coli* strains were similar to wt cells' as well as to each other. This latter finding indicates that the repair capability was not as heavily impacted as was the case in $\Delta mutS$ or $\Delta mutL$ cells. While unexpected, this is consistent with our observation that the average repair efficiency in $\Delta mutH$ or $\Delta uvrD$ are much higher than in $\Delta mutS$ or $\Delta mutL$, and that MutH and

UvrD might not be equally essential for the repair of transformed plasmids.

Next, we asked whether biologically relevant modifications on the plasmids play a role in the sequence preference of the repair response. Similarly, the relative ease of repair of mismatches in plasmids harboring or devoid of single strand nicks were correlated in wt cells, but not in $\Delta mutS$ and $\Delta mutL$ (Figure 1.22b). Although we observed a global decrease in the measured DNA repair efficiency upon methylation in wt cells (Figure 1.8a), we observed that the detected outlier mismatches for wt cells were largely similar with or without methylation. A similarly high correlation was also observed for $\Delta mutH$ cells, but was weaker for $\Delta mutS$ cells, suggesting that the sequence properties rendering a mismatch inefficiently repaired are not significantly dependent on nicks or DNA methylation (Figure 1.22c).

In the prokaryotic cytoplasm, the plasmids are maintained in the fully methylated form, whereas the nascent strand during replication is methylated only on one strand for a brief period of time enabling strand-selective repair. As the DNA replication is an important source of mismatches, the MMR might have evolved especially to target such hemi-methylated DNA. To observe the potential effect of this epigenetic asymmetry on the subset of inefficiently repaired outlier mismatches, we performed a primer extension on a fully methylated vector library, thereby replacing one of the methylated strands with a newly synthesized unmethylated strand. Transformation of mismatch libraries ligated to such hemi-methylated plasmids indicated that such a hemi-methylation do not significantly alter the identity of poorly repaired outliers in a wt cell, either (Figure 1.23a), while leading to a systematic underestimation of the observed efficiencies as the symmetry assumption justifying Equation 1.12 does not hold. As a control against possible changes due to the extra primer extension step added to our vector generation workflow, we repeated this analysis procedure on a control library that went through the same vector preparation, except that the methylation step was omitted. The expected product of this new protocol is again a fully unmethylated mismatch-bearing plasmid, whose inefficiently repaired outliers in a wt cell were similar to that obtained using the unmethylated plasmid without the primer extension step (Figure 1.23b). As before, the reproducibility of this outlier set was much lower for $\Delta mutS$ cells due to impairment of MMR (Figure 1.23c). These observations suggest that neither nicks, nor methylation state has a

very important impact on the detected subset of relatively inefficiently repaired mismatches. Our results would agree with a model in which the critical decision to repair a particular mismatch has been mostly made during the detection step, but the execution by the downstream elements work more deterministically. Such chemical cues might be more important on "how" the repair will be performed, rather than "if".

Finally, we verified if the changes in our experimental design would influence the outlier mismatches. Of particular interest is the reduction of the barcode length from 25 to 20 bases, which reduces the barcode diversity by $4^5 = 1024$ fold, but does not have a significant effect on the detected outliers (Figure 1.24a). This observation corroborates our proposition in Section 1.3.2 that the chosen barcoding scheme labels virtually all plasmids uniquely and that the wrong clan calls due to tag redundancy is negligible, if not non-existent. On a similar note, we also verified that by the end of the incubation period, the cell culture has reached an equilibrium and that the inefficiently repaired mismatches detected are not significantly altered by an increase in the incubation time. While virtually all transformed DNA mismatches should be disbanded by the plasmid replication machinery early on during the incubation, excessive replication might lead to accumulation of replication error artifacts, hence leading to aberrant registration of certain replication error-prone DNA motifs as poorly repaired mismatches. To test this possibility, we transformed $\Delta uvrD$ cells with SML and harvested half of the bacterial culture the day after, whereas the other half was used to re-inoculate fresh medium and harvested after 1 more day of incubation. We compared the independently assigned classes to these two samples and observed a high correlation (Figure 1.24b). This argues in favor of the assumption that the detected outliers mismatches are stable by the time of the termination of the overnight incubation period.

### 1.4.3 Two closeby mismatches on the same molecule

As a further question, we investigated the combined effect of DNA molecules harboring two and only two mismatches that are located within 50bp from each other by design. Similar to the arguments above, the clans that are replication products following the successful repair of the ancestor plasmid are expected to exclusively consist of one type of product (C- or V-type clans), whereas

strands not repaired before the replication should provide a roughly equimolar mixture of the two strands (U-type clan). However, six different classes of clans need to be sought, unlike three in the single mismatch case (Figure 1.25). That both mismatches had been repaired using the common strand as the correct information source would produce clans that are identical to the consensus sequence (?+?). Hence, the progeny of double-mismatch containing library in which at least one of the two mismatches were repaired in this way do not reveal any information about the identity or the position of the mismatch on the ancestor plasmid. A similar loss of information is also the case, if one of the two mismatches were to be repaired based on the common strand, regardless of the other one remaining unrepaired (U+?) or repaired using the opposite strand (R+?). The type and position of the mismatch can be deduced by the entry in the difference histogram that most significantly deviates from the consensus sequence if both of the two mismatches were repaired using the variable strand (R+R), only one of them repaired keeping this strand whereas the second mismatch was left unrepaired (U+R) or neither of them could be repaired in time before the first plasmid replication happens (U+U).

The mismatch pairs that we investigated in our Double Mismatch Library (DML) were at most 50 bases apart from each other, and Figure 1.26 suggests that such closeby mismatches tend to have a common fate as clans containing different outcomes for the mismatches (U+?, R+? or U+R) are much less frequent than those in which either, both or none of the mismatches were repaired (R+R, U+U, ?+?). In wt cells, we observed the frequency of fully repaired (R+R or ?+?) clans to dominate over fully unrepaired clans (U+U). In contrast, ΔmutS or ΔmutL cell cultures were dominated by fully unrepaired clans (U+U). However, regardless of the presence of fully functional MMR system, observation of a clan in whose genealogy only one of the two mismatches repaired (U+R or U+?) was much less frequent than either of the three all-or-none type of clans. The event frequency of two nearby mismatches having been repaired using information from two opposite strands was also likewise very low in all five datasets (R+?). These observations were also the case if the construct used for transformation was nicked by design, albeit causing a measurable change in the apparent frequencies of these six classes.

Due to the sheer number of mismatch combinations possible, obtaining a dataset with a high

level of coverage is much more challenging, if not totally infeasible, than covering single mismatches. Following the chemical synthesis approach we described in this work, covering all possible mismatch combinations along the 53 nucleotide long sequence that we probed would require purchasing C(53,2)=1378 separate oligos. On similar grounds, the required sequence data would be much higher as analysis of the data sets would mean populating 4D tensors (4x53x4x53), instead of incrementing elements of 2D matrices (4x53). We hence only sampled a subset of all possible combinations in our experiments (594 out of 12402 possible) and limited our results to this position and mismatch type averaged format. However, our data suggest a significant variability between different mismatch combinations. Notably, the base spacing between the two mismatches positively correlates with the frequency of clans in which only one mismatch was repaired (R+U) accompanied by a gradual decrease in the frequency of fully repaired (R+R) and fully unrepaired classes (U+U in Figure 1.27). Such an increase in the co-repair probability of nearby mismatches would intuitively be expected to be high, because the MMR executes its function by removing a long stretch of DNA around the mismatch and it will be more likely for two closeby mismatches to coincidentally fall under the same removed ssDNA fragment and get repaired concurrently. However, this trend is more pronounced in the presence of nicks, and observed with or without functional MMR pathway, suggesting that it is not directly a result of the mismatch detection pathway only, but is likely to be contributed by a similar process that has a similar effect on the single mismatch case. As was discussed above, and despite availability of more experimental information in Chapter 2, the exact nature of this gradual trend in the repair efficiency is still elusive.

## 1.5 Conclusion

In summary, I introduced in this chapter a simple high-throughput method based on stochastic DNA barcoding and next generation sequencing that can be utilized to quantify relative repair efficiency of DNA mismatches. To demonstrate the validity of our approach, we performed measurements in wt as well as MMR deficient *E. coli* strains and observed a clear difference between the overall repair efficiencies. As a further verification, we detected none or only a few clans in the absence of intentionally introduced mismatches. With our method, we sought to determine

mismatches that tend to be less efficiently determined than the general trend and found them to be reproducible between experimental replicates and certain conditions in agreement with our expectations. I then suggested to generalize our approach to multiple mismatches on the same DNA strand and observed a clear connection between their fates in terms of repair, which is in agreement with our expectations due to the <50bp spacing in-between as opposed to the long stretches of strands removed during the MMR response that can reach to kilobases.

As a surprising outcome, we observed the repair efficiency to be highly dependent on its relative positioning on the ligated plasmid, to an extent depending on the consensus sequence of the mismatch library of concern. Although the repair efficiency has been reported to display genome-wide trends attributable to the relative location with respect to the replication origins [45], the sigmoidal transition we observe can be as sharp as within 20 nucleotides only and is not explainable by this proposition. While the repair efficiency difference between nicked and covalently closed molecules is in agreement with the literature (ex. [15]), it is also elusive why it makes this position dependence effect much more significant. Having said so, this global trend did not disappear after treatment with exonucleases capable of initiating digestion at the nicks, suggesting that this outcome might be a result of other sequence features on the vector. We tried to address this possibility by inverting the insertion orientation but did not observe a reversal of the trend. These observations do not prove, but suggest that it is caused by the sequence elements on the plasmid backbone, which could be the relative position with respect to the MutH binding site GATC, nearest being located about 0.5kb upstream of the tracing barcode.

I believe an accurate assessment of the difficult to repair sequence motifs is of importance, as an inability to correct inevitable replication mistakes is one of the reasons leading to accumulation of mutations. However, expansion of the method to deduce the repair efficiency of larger DNA motifs require more measurements. Currently, the bottleneck in the procedure is the generation of mismatch containing DNA libraries. While it provides an easy way, formation of mismatch libraries by annealing chemically synthesized oligos is prohibitively expensive, making generation of larger datasets infeasible. In fact, most of the data that is presented here has been possible thanks to the DNA libraries that we had at our disposal from previous studies. Furthermore, the

current workflow is also unable to distinguish between the repair products, if the substituted strand is removed. The compensation of this systematic data omission relies on the bold assumption that there is no strand selection bias in the absence of methylation cues or at least it is same for all mismatches sampled. In the next chapter, I will introduce an improved experimental approach that can address these shortcomings.

## 1.6 Materials and Methods

### 1.6.1 Preparation of trackable vector library (tUNC19)

We prepared chemically competent cells using NEBTurbo strain (NEB, C2984I) with E. coli Mix&Go Transformation kit with Zymobroth following manufacturers instructions (Zymo, T3001), hereafter referred to as Home-Made Competent Cells (HMCC). We transformed 50pg pUC19 (NEB, N3041S; Genbank L09137) into a 100 $\mu l$ HMCC aliquot by gentle agitation, plated on an LB-agar plate with $100\mu g/ml$ ampicillin, and extracted pUC19 plasmids using a standard Miniprep kit after overnight incubation in LB medium (Omega, E.Z.N.A. Plasmid Mini Kit I, D6942) inoculated by a single colony of these transformants. We used this product as template in a PCR reaction where one of the primers has 25 random bases (primer P2, denoted by N) in blocks of five N's separated by T's. A $800\mu l$ batch of the reaction contains 50ng pUC19 as template, 240pmol of each primer (P1 and P2) and Phusion 2X mastermix with HF buffer (NEB, M0531S). The PCR mix was split into $100\mu l$ aliquots and was subjected to 35 cycles of 10s 98°C denaturation, 30s 71°C primer annealing 60s at 72°C elongation phases preceded by additional initial denaturation at 98°C for 30s and followed by 72°C final extension for 2 min. The product was purified via QIAquick PCR purification kit (Qiagen, 28104) with 4ml buffer PB supplemented with $20\mu l$ 3M NaAc, otherwise according to manufacturers instructions. This procedure has a typical yield above $10\mu g$, on whose product we generate sticky ends and remove remaining pUC19 templates by a triple digest in 1X Cutsmart buffer with $8\mu l$ SacI-HF (NEB, R3156S), $8\mu l$ XhoI (NEB, R0146S) and $8\mu l$ DpnI (NEB, R0176S) in $400\mu l$ final volume. The reaction mixture was incubated for 1 hour at 37°C followed by a 30 min heat inactivation at 65°C. The barcoded vector to be ligated was purified with QIAquick

PCR purification kit out of this digestion mix.

### 1.6.2 Library preparation by oligo annealing

We dissolved all variable strand oligos to 200$\mu$M final concentration in T10 buffer and prepared an equimolar mixture of all variables strands (Please refer to 5.2 for a complete list). We thermally annealed this mix to the respective complementary strand of common sequence (C-FML, C-SML, C-DEB3L, C-DEB3R) by mixing 400fmol of each in 20$\mu$l T50 buffer and slowly cooling from 98°C down to room temperature in about 1 hour. To introduce 5' terminal phosphate groups, we incubated 60fmol annealed dsDNA library supplemented with 1$\mu l$ T4 polynucleotide kinase (NEB, M0201S) with 1mM ATP in 1X PNK buffer for 1 hour at 37°C followed by 30min inactivation at 65°C.

We used this product as the insert for ligation reaction. We ligated each library to our vector library by incubating at room temperature for 30 minutes to 2 hours as a mixture containing 600ng vector and and 3 to 6X molar excess of the respective insert supplemented with 20$\mu l$ quick ligase (NEB, M2200S) in 400$\mu l$ 1X quick ligation reaction buffer. The excess reagents and salt was removed by QIAquick PCR purification kit; where we followed the manufacturers recommendations except that the final product was eluted with 35$\mu l$ nuclease free water rather than provided elution buffer (EB).

### 1.6.3 Enzymatic modifications

To generate nicks in the final ligated construct, we optionally included a dephosplorylation step of the barcoded vector before ligation with 8 $\mu l$ shrimp alkaline phosphatase (NEB, M0371S) for 1 hour followed by an inactivation step at 65°C for 30min. As the ligation reaction is not 100% efficient, the products contain a mixture of ∼2.5kb plasmid carrying nicks on one, both or none of the strands as well as the unreacted linear DNA. When indicated, to remove these unsealed products, following the ligation, we supplemented the ligation mixture with 10U T5 exonuclease (NEB, M0363S) or simultaneously with 10U exonuclease VII (NEB, M0379S) and 100U exonuclease III (NEB, M0206S) and incubated for 1 hour at 37°C.

For the case that the vector backbone was indicated to have been methylated, we introduced methylation on adenines within GATC sequence contexts of the vector backbone by a deoxyadenosine methyltransferase (Dam, NEB, M0222S) treatment. We incubated $\approx 4\mu$g of vector DNA with 2 $\mu$l Dam (NEB, M0222S) and 1$\mu$l 0.8 $\mu$M S-adenosine methionine (SAM) in 40$\mu$l 1X Dam reaction buffer for 1 hour at 37°C before the double-digestion step. We purified the product with QIAquick PCR purification kit and verified the presence of methylation by DpnI or MboI cleavage [46, 47].

To generate hemi-methylated DNA, we treated the PCR product for the barcoded vector with deoxyadenosine methyltransferase as above, which puts methylation marks on both strands. Afterwards, we performed primer extension with 10X molar excess of primer P1 to displace one of the methylated strands in 1X Phusion mastermix prepared as above and incubated at 72°C for 3 min following a denaturation step at 98°C for 1 min. We purified this product with a PCR purification column and subjected to the same double-digest protocol described above to generate sticky ends.

### 1.6.4 Transformation

We generated electrocompetent cells following [48] using cell strains that we procured from *E. coli* stock center at Yale University (5.1). We used 1ml confluent culture to inoculate 1L autoclaved LB medium (Fisher Scientific, BP9723) supplemented with $50\mu g/ml$ kanamycin if required. We incubated a 4L flask at 250rpm constant shaking at 37°C till $OD_{600}$ reaches 0.3, which typically takes around 4 hours. We collected the cells by centrifugation at 2000g for 10 min, and re-suspended in equal volume of ice cold 10% glycerol in water, which had been sterilized by passing through 220nm filter. We repeated this wash step twice, after which re-suspend the cells in 5ml glycerol. We re-suspended the cells in 3ml glycerol and flash-froze $100\mu l$ aliquots in liquid nitrogen, which we stored at -80°C till usage.

We thawed an aliquot of electrocompetent cells on ice, to which we added $5\mu l$ of purified ligation product in water. We load this suspension on an ice-cold electroporation cuvette with 10mm gap width (Sigma-Aldrich, Z706078). We targeted 1700V with an Eppendorf Eporator, which usually results in an applied voltage of 1650V within 5ms. We recovered the cells from the cuvettes by a

wash with 500$\mu l$ SOC medium (NEB, B9020S) twice, pre-warmed to 37°C, a procedure which we repeated twice to improve the retrieval efficiency. We incubated this 1ml broth in a 50ml conical bottom centrifuge tube at 37°C for 1h, after which we added 9ml LB with final concentration of 100$\mu g/ml$ ampicillin. We plated 10$\mu l$ of this culture on an ampicillin containing agar plate for quality control purposes, whereas the rest was further incubated overnight in preparation for plasmid extraction.

### 1.6.5 High-throughput sequencing

We extracted the plasmid library from 10ml overnight cell culture using a standard Miniprep kit following the manufacturer prescribed protocol (Omega E.Z.N.A. Plasmid Mini Ki I, D6942). We used 5ng of this elute as PCR template to amplify out the region of interest using 0.5$\mu M$ of S1 and S2 primers with Phusion 2X mastermix. We employed 20 cycles of 10s 98°C denaturation, 20s 63°C primer annealing 10s at 72°C elongation phases preceded by additional initial denaturation at 98°C for 30s and followed by 72°C final extension for 2 min. To cleanup the PCR product, we incubated the product mixture with 20$\mu l$ Ampure XP beads (Beckman Coulter, A63880) for 5 min at room temperature. We retained the bead-bound material after keeping for 2 minutes on a magnetic rack (GE, 1201Q46). We washed the beads twice with 80% ethanol and eluted the material in 53$\mu l$ 10mM Tris pH8.5 by incubation for 2 min. We collected about 50$\mu l$ bead-free liquid 2 min after placing the material on a magnetic rack.

We performed 8 additional cycles of PCR with Nextera 24-Index kit for indexing before sample pooling (Illumina, FC-121-1011), for which we used 7.5$\mu l$ of elute as template, 7.5$\mu l$ suitable i5 and i7 primers with 38 $\mu l$ Phusion 2X mastermix. We followed manufacturers recommended thermal cycling protocol (95°C 3min, 98°C 30s, 55°C 30s, 72°C 30s, 72°C 5min). We also bead-purified 56$\mu l$ of this final product with 56$\mu l$ Ampure-XP and eluted with 28$\mu l$ 10mM Tris, pH8.5 buffer. We pooled the final products based on their Nanodrop and/or Qubit reading to desired relative contribution to the final pool to be sequenced. We found a mixture of 480 $\mu l$ Hbf buffer, 480$\mu l$ pooled 20pM library and 15$\mu l$ 20pM PhiX control library (Illumina, FC-110-3001) to provide a reasonable spot density. We typically used 300 cycles MiSeq v2 micro reagent kit (Illumina,

MS-102-1002) to perform a paired-end sequencing in-house for about 135-140 cycles each.

### 1.6.6 Data analysis

We retrieved raw *.fastq output from the MiSeq system and parsed with a home-made program implemented in C++ compiled with GCC v 9.3 and GNU Octave v5.2. The source code of the analysis toolkit can be accessed through the GitLab page, https://gitlab.com/tuncK/public/tree/master/fixseq-codes. The major steps taken in the course of analysis are as follows:

First, all the data is imported DNA by DNA, while the reads are parsed to locate the constant adapter segments using a Needleman-Wunsch algorithm with gap and mismatch penalties of -1 and match gain of +1. We take the reverse complement of the paired end reads and add to the dataset to enhance the SNR by reducing the effect of sequencing errors.

Secondly, we generate a unique set of all detected barcodes on a red-black tree, during which exact duplicates are detected and recorded. To account for sequencing errors that could artificially diversify barcodes from the same clan, we introduced error tolerance by implementing a density based scanning on the set of all extracted barcodes, where the minimum density threshold is N=10 within $\epsilon$=3 mutation distance. To reduce the memory usage, only a list of neighbors is stored rather than an explicit matrix listing pairwise distances. After all the barcodes are processed, barcodes failing the density criterion (noise) are discarded, and the core-points together with all neighbors are reported as clans. Interested readers are referred to [32] for the description of this DBSCAN algorithm.

Third, statistics in each clan is evaluated. Of particular use is the clan-wide ensemble-averaged base frequency, which is outputted into a *.hist text file following the .mat file format as a 4 x libraryLength matrix per each clan. For the purposes of downstream analyses, each such matrix is considered as one data point.

Finally, we import this *.hist file to count the number of repaired and unrepaired clans per each position and possible mismatch type. This deterministic decision is made with respect to a system-wide fixed threshold, typically set to [0.1, 0.9] for no-repair events, repaired otherwise. In the absence of a custom designed library with a secondary barcode that can keep track of mismatch

type in each clan, the clans that do not deviate from the common strand cannot be used to deduce the original mismatch type, and hence have to be discarded as ambiguity. We report the ratio of the repaired clan counts among the all detected clans with that same mismatch type, i.e.

$$\eta_{ij} = \frac{2N_{ij}^{repaired}}{2N_{ij}^{repaired} + N_{ij}^{unrepaired}} \tag{1.20}$$

A similar rational applies to insertion library scans (Chapter 3), but now we detect the frequency of shifts in each clan rather than single base substitutions. Processing a typical dataset containing a few hundred thousand reads with this procedure takes around 5 minutes on a standard quadcore desktop computer.

**Algorithm 5:** Summary of the data analysis procedure

$forwardReads \leftarrow$ Import all forward reads from file "sample#_R1.fastq"
$reverseReads \leftarrow$ Import all reverse reads from file "sample#_R2.fastq"
**foreach** $read \in reverseReads$ **do**
    $read \leftarrow$ reverse-complement($read$);
**end**
Fix $n =$ length of the barcode + mismatch library inserted
Initialize $acceptedReads = \varnothing$
**foreach** $full\_read$ *in Union*($forwardReads$,$reverseReads$) **do**
    $AdaptorEndPos \leftarrow$ Search 5' adaptor position
    **if** $full\_read$ *is shorter than n OR Adaptor not found OR Adaptor is severely shifted* **then**
        Ignore the $read$
    **end**
    $SeqOfInterest \leftarrow$ Extract the n-base long sub-sequence of full_read following the adaptor
    **if** $HammingDistance(sequenceOfInterest, expectedLibraryPrototype) > 5$ **then**
        $SeqOfInterest \leftarrow$ Shift the $SeqOfInterest$ by dynamic programming
    **end**
    **if** $HammingDistance(SeqeOfInterest, expectedLibraryPrototype) > 5$ **then**
        Extract $(barcode, probe)$ from $SeqOfInterest$ using $expectedLibraryPrototype$
        Add ($barcode$,$probe$) to $acceptedReads$
    **else**
        Ignore the $sequence$
    **end**
**end**

$barcodeClusters \leftarrow$ DBSCAN $(acceptedReads.barcodes, \epsilon = 3, N = 10)$
$clans \leftarrow$ Group $acceptedReads.probes$ w.r.t. $barcodeClusters$
Fix $PrototypeHistogram \leftarrow$ Build a normalized base composition histogram of the consensus sequence
Initialize 4 x length($SeqOfInterest$) matrices $repairedEvents$ and $unrepairedEvents$ to 0
**foreach** $clan \in clans$ **do**
    $ClanHistogram \leftarrow$ Build a normalized histogram of base composition of $clan$
    $DifferenceHistogram \leftarrow ClanHistogram - PrototypeHistogram$
    $(maxDeviationQty, maxDeviationPos, maxDeviationBaseType) \leftarrow findPeak(DifferenceHistogram)$
    **if** $maxDeviationQty < 3 \cdot secondHighestDeviationQty$ **then**
        Assert "low quality"
        Ignore the $clan$
    **end**
    **if** $maxDeviationQty < 0.1$ **then**
        Assert "repaired"
        Assert "MM type not inferrable"
    **else if** $maxDeviationQty > 0.9$ **then**
        Assert "repaired"
        Increment $repairedEvents(maxDeviationPos, maxDeviationBaseType)$
    **else**
        Assert "unrepaired"
        Increment $unrepairedEvents(maxDeviationPos, maxDeviationBaseType)$
    **end**
**end**
Report $totalClanCount \leftarrow repairedEvents + unrepairedEvents$
Report $repairEfficiency \leftarrow 2 \cdot repairedEvents/(2 \cdot repairedEvents + unrepairedEvents)$

## 1.7 Figures and tables

**Table 1.1:** List of acronyms used to describe different DNA libraries and required sequencing primers used in the setup.

| | | | |
|---|---|---|---|
| DEB3L | De Bruijn library Left | contains all trimers, IDT | S1 and S2_5 |
| DEB3R | De Bruijn library Right | contains all trimers, IDT | S1 and S2_5 |
| NPL | No Problem Library | No mismatch, IDT | S1 and S2_2 |
| SML | Single Mismatch library | IDT | S1 and S2_2 |
| FML | First Mismatch library | IDT | S1 and S2_1 |
| DML | Double mismatch library | IDT | S1 and S2_2 |

**Table 1.2:** Some past reports of repair efficiencies of different mismatch types, sorted in descending order. $\Delta$ refers to insertion loops

| Senior author | Ref | Organism | Mismatches (descending) | Notes |
|---|---|---|---|---|
| Kunkel | [15] | HeLa | AC >GT >CT | At position 87 |
| | | HeLa | GT >AC >CT | At position 89 |
| | | E. coli | GT >AC >CT | At position 87 |
| | | E. coli | GT = AC >CT | At position 89 |
| Modrich | [6] | E. coli | AC, GT >AA, GG, TT >AG, CC, CT | Review |
| Kunkel | [7] | S. cerevisiae | GT >CC >TT | Review |
| Radman | [13] | E. coli | GT >GG >TT >AC >AA >CC >AG >CT | |
| Fox | [16] | E. coli | AG >AC >AA,TT,CT >GG >GT >$\Delta T$ | $K_d$ of MutS |

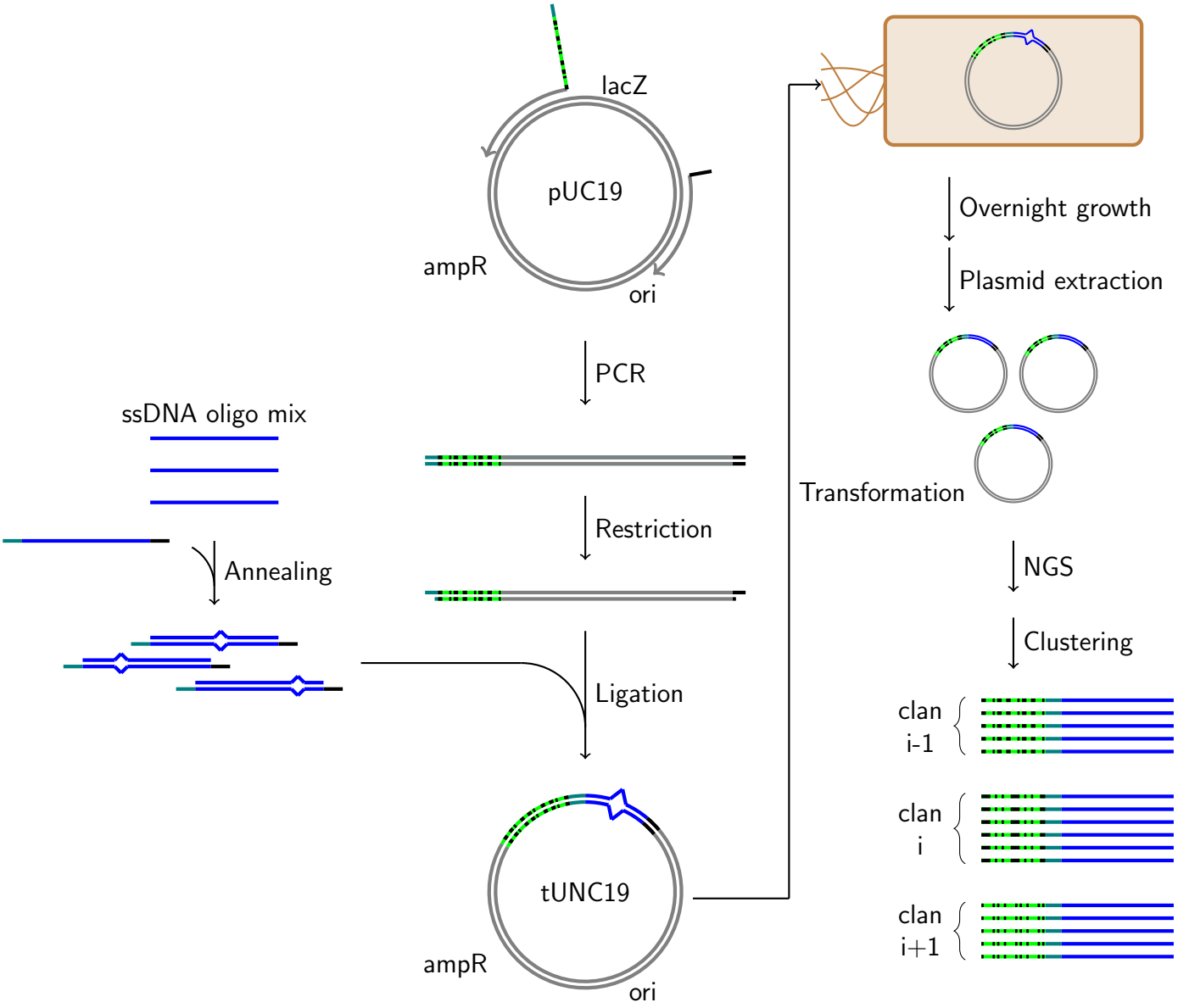



**Figure 1.1:** Schematic summary of the experimental design. A set of individually chemically synthesized oligos was purchased to serve as the variable strand and the library was formed by annealing their mixture to the common strand oligo to yield a heteroduplex carrying a mismatch. DNA barcodes were introduced to the pUC19 plasmid via a PCR reaction with primers containing a random base tail that uniquely labels each individual plasmid. The resulting linear PCR product was ligated to the mismatch library after generating sticky ends with a restriction double digest reaction. The plasmid library with mismatches was transformed into *E. coli* and after multiple rounds of replication, the extracted plasmid library was sequenced. Clustering of the reads with respect to the tracing barcodes segregate DNA sequences originating from the same ancestral plasmid, while the homogeneity or heterogeneity of the clans provides a means to detect whether the plasmids were repaired or unrepaired.

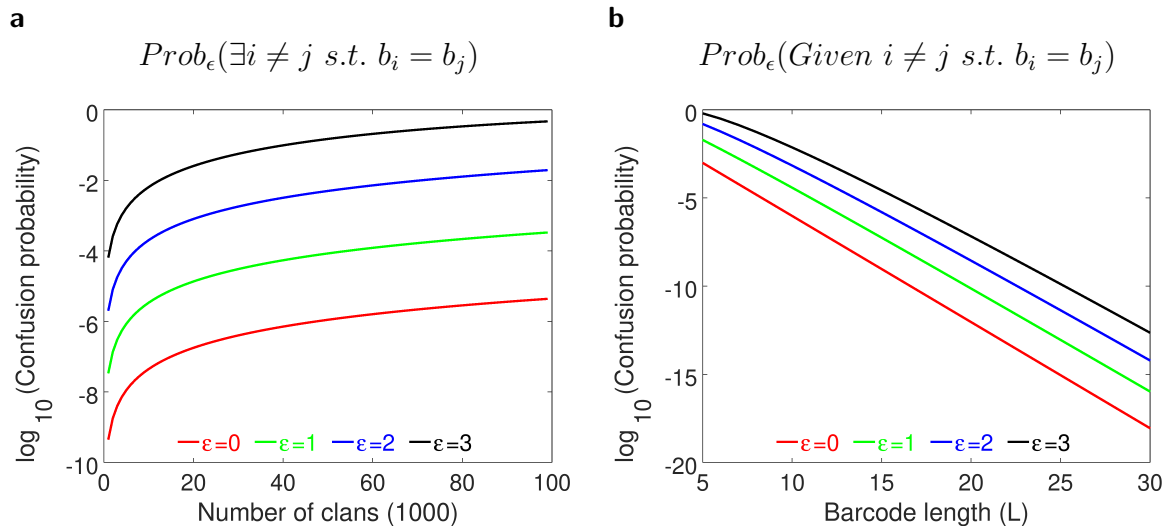



**Figure 1.2:** The probability of mis-tagging two plasmids in the starting material with the same barcode as an increasing function of the total number of clans in the sample. (a) The probability of observing at least two clashing clans in the full data set becomes more significant as the error-tolerance in data analysis is increased, which can lead to mis-classification of two repaired clans as unrepaired. (b) The probability that two randomly chosen molecules receive identical barcodes by chance, as a function of the barcode length. The 4 curves in each plot depicts the case with different error-tolerance during data analysis, where $\epsilon = 0$ indicates that clans contain barcodes that are exactly identical, whereas $\epsilon$ =1, 2, 3 indicate the cases that treat barcodes at 1, 2 or 3 substitutions from each other as identical.

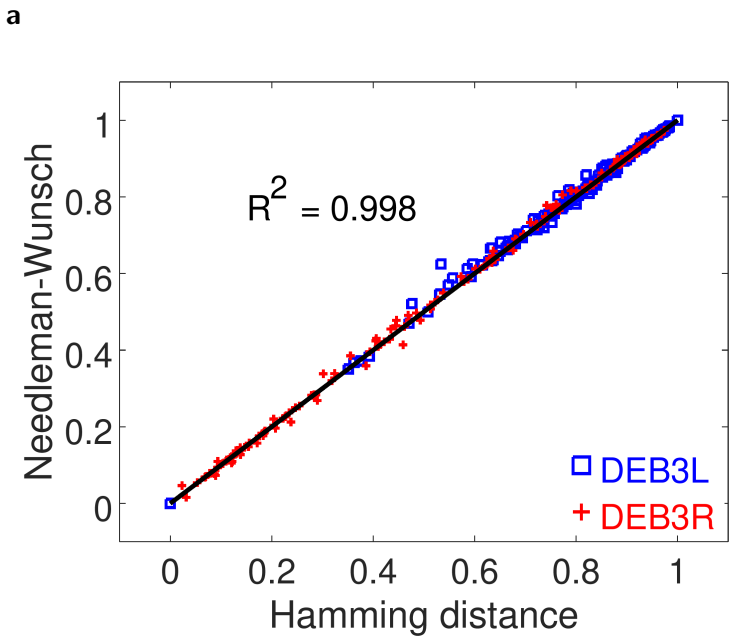



**Figure 1.3:** The choice of Hamming distance over edit distance for computational efficiency does not have a significant effect on the measurements, demonstrated by the comparison on DEB3 library results in wt cells with primer extension. Both coordinates of each data point represent the repair efficiency ($\eta'$) of a single mismatch, obtained by the analysis of the cumulative output from two experimental replicates using the respective distance metric in DBSCAN.

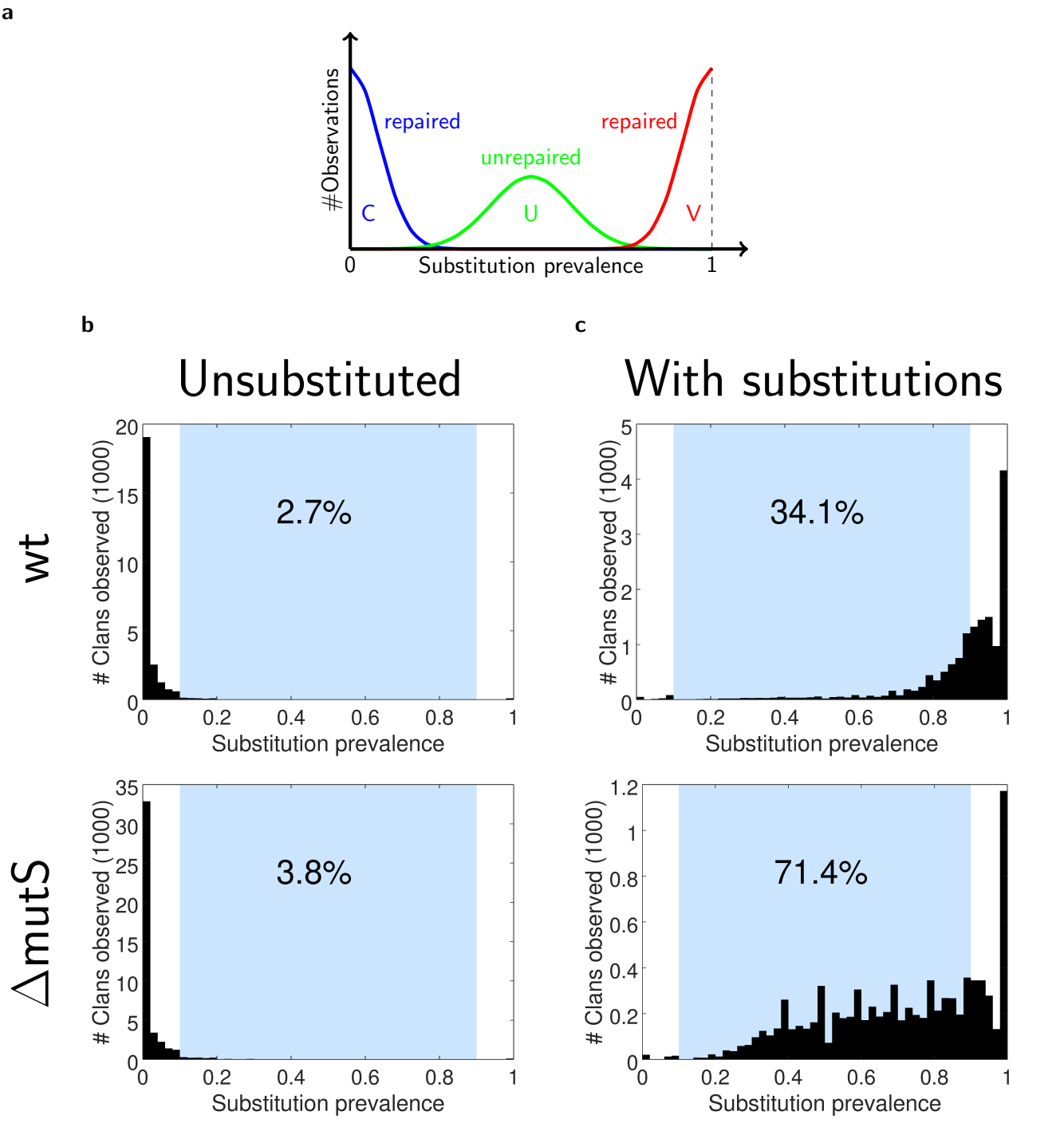



**Figure 1.4:** s-histograms with or without mismatches. (a) The prevalence of the substitutions against the consensus sequence of a DNA library provides a means to deduce if pre-replicative repair has taken place. Regardless of the activity of MMR, the samples constructed as a proper dsDNA without any substitutions predominantly yield clans with below-treshold substitution prevalences, if any (NPL, b). The treatment of wt cells with a library containing mismatches (SML) is dominated by repaired clans, whereas cells deficient in MMR yield a high amount of unrepaired clans (c). Impediment on MMR causes many more clans with unrepaired phenotype compared to the wt cells. Histograms were constructed from cumulated mismatches within all positions and mismatch types. Blue shaded region (substitution observable in 10 to 90% of all clan members) defines the zone we assumed to be the unrepaired clans for this data set.

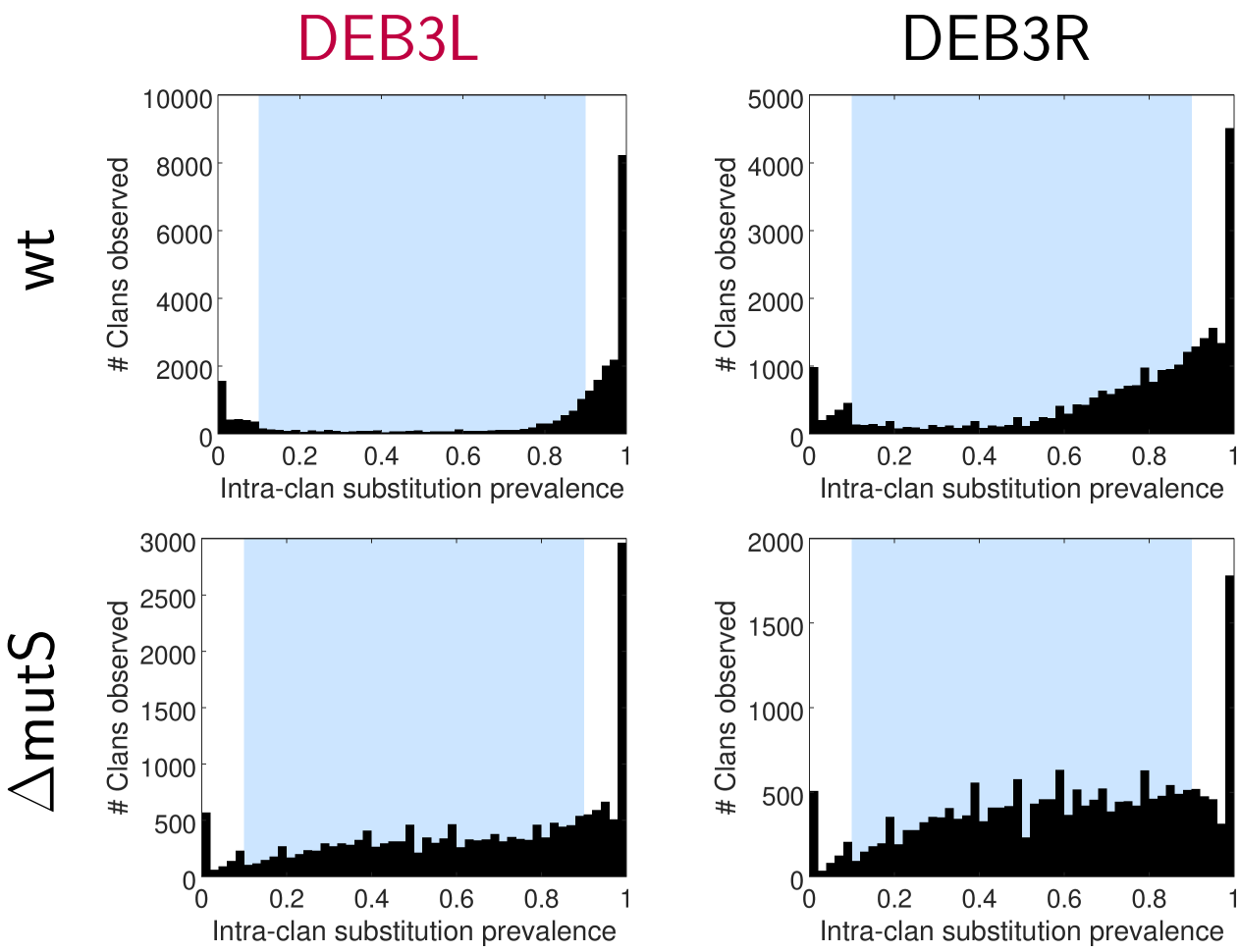



**Figure 1.5:** Same as in 1.4, but applied to the DEB3 library.

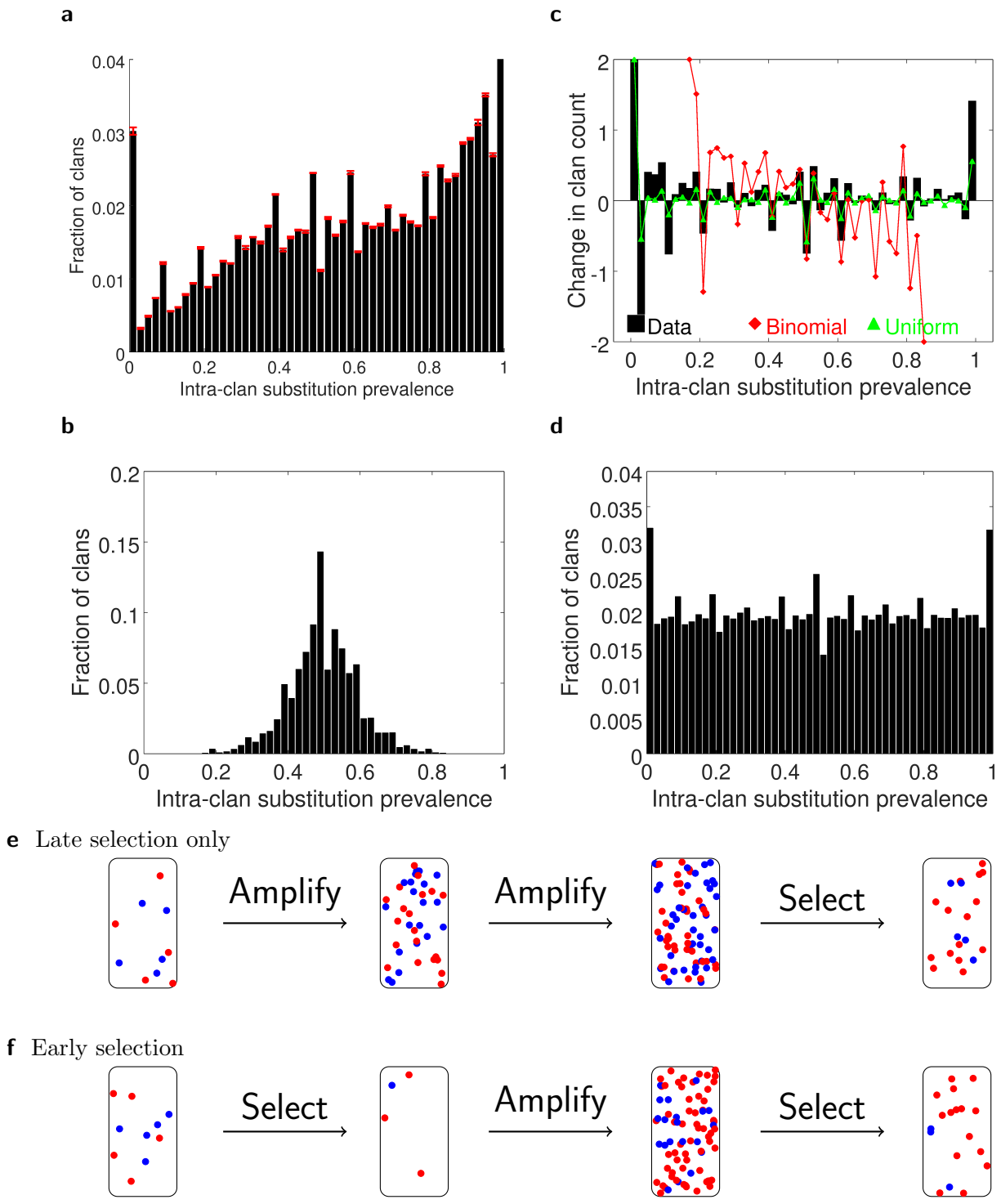
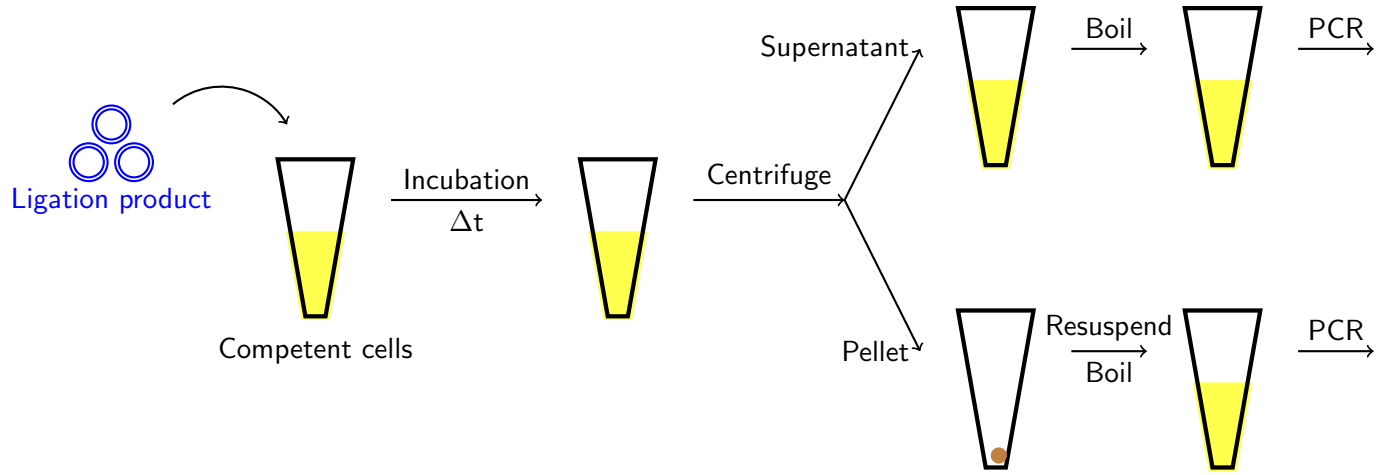



**Figure 1.6:** (a) The reproducible fluctuations observed in the substitution histograms can be partially explained by the integer counting statistics of DNA sequences that constitute the clans. Data are **mean** ± S.E.M. of four experimental DEB3L replicates in $\Delta mutS$ cells. (e) If the selection step leading to drift is late in the protocol, the U-peak will be dominated by binomial sampling process out of a very large DNA pool with a narrow spread (b). In an experiment with multiple DNA amplification and subsampling steps (f), the U-type clan distribution might deviate from binomial distribution and can attain more extreme values (d). The sign of 38 out of 49 jumps can be predicted correctly by the uniform model that disregards combinatorial likelihood based on binomial coefficients ($p=7 \cdot 10^{-5}$), whereas a full binomial model can predict 29 out of 49 (p=0.13, c).

**a**


Ligation product
Competent cells
Incubation
Δt
Centrifuge
Supernatant
Pellet
Boil
PCR
Resuspend
Boil
PCR

**b**

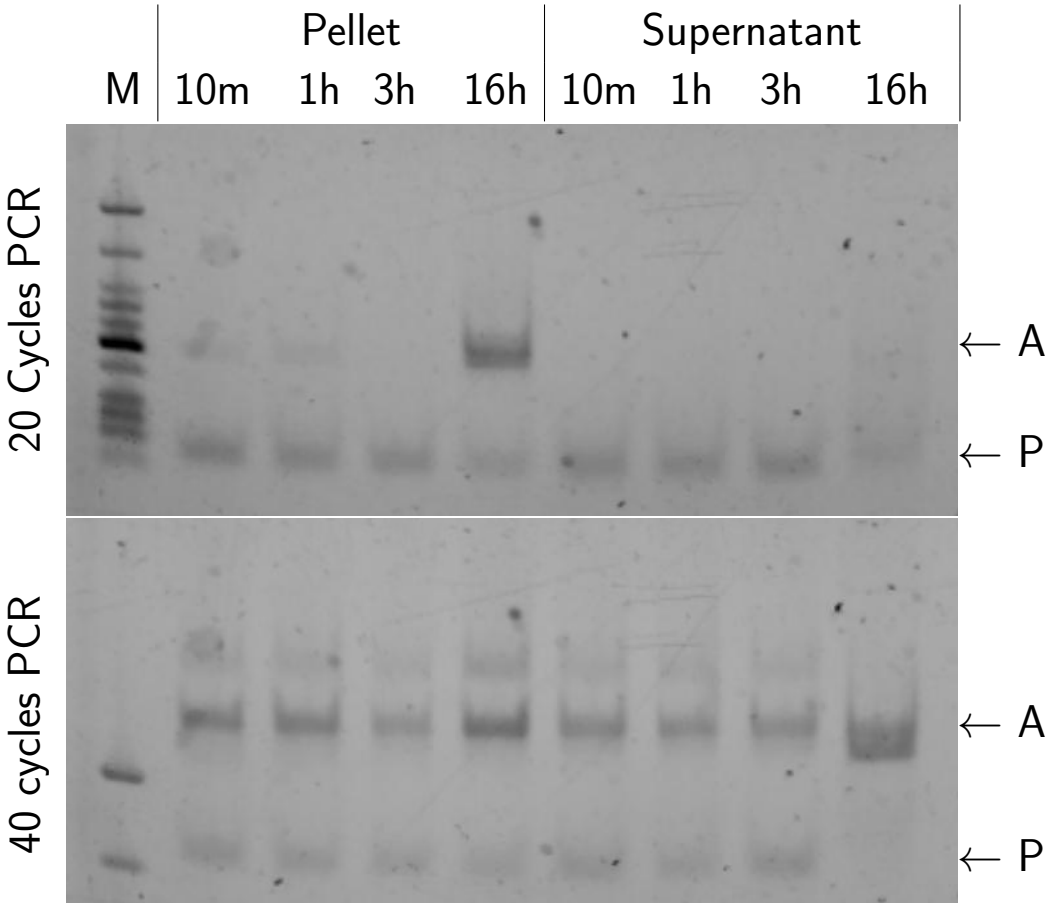

Pellet
Supernatant
M
10m
1h
3h
16h
10m
1h
3h
16h
20 Cycles PCR
40 cycles PCR
← A
← P
← A
← P

**Figure 1.7:** Comparison of plasmid quantities within the cells and the growth medium. Plasmids sequenced are mostly from within the cytoplasm, as the amount of plasmids within the cells well exceeds the free-floating plasmids in the growth medium that are leftovers of the transformation or have been released later upon cell lysis. (a) A simple experimental design to compare the amount of amplifiable plasmid available in the cells vs. the growth medium after the overnight recovery of the electroporated cells. (b) Image of a 2% agarose gel loaded with the PCR samples that were cycled for 20 or 40 cycles. Aliquots of the bacterial culture were collected after 10min, 1h, 3h or 16h of recovery following electroporation, and fractionated into pellet and supernatant. The pellet was re-suspended to the original volume in T10 buffer. 1 $\mu l$ of each fraction was used as PCR templates and the relative quantity was visually compared on an agarose gel after amplification. While some quantity of plasmid is always present in the growth medium and in the cell pellet and generates a visible band after 40 cycles of amplification, the quantity was too low to be observed after 20 cycles except in the cell pellet obtained after 16 hours. A: PCR amplicon; P:unused primers.

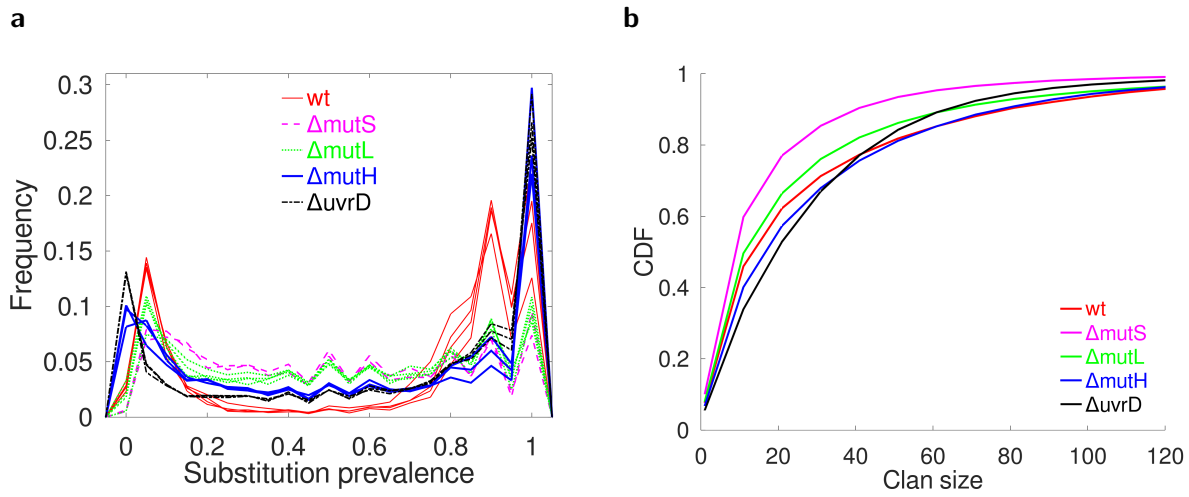



**Figure 1.8:** (a) Substitution prevalence histograms of various MMR pathway mutants. While all mutants have fewer mixed clans than wt cells (—), $\Delta mutS$(- - -) and $\Delta mutL$ ($\cdots$) cells' apparent repair was much less pronounced than in $\Delta mutH$ (—) or $\Delta uvrD$ cells (- · -). Each condition is represented by multiple curves representing independent experimental replicates. (b) Clan size distribution of the MMR pathway mutants. Each mutant is represented by a single curve depicting the CDF for the cumulated data from all relevant experimental replicates.

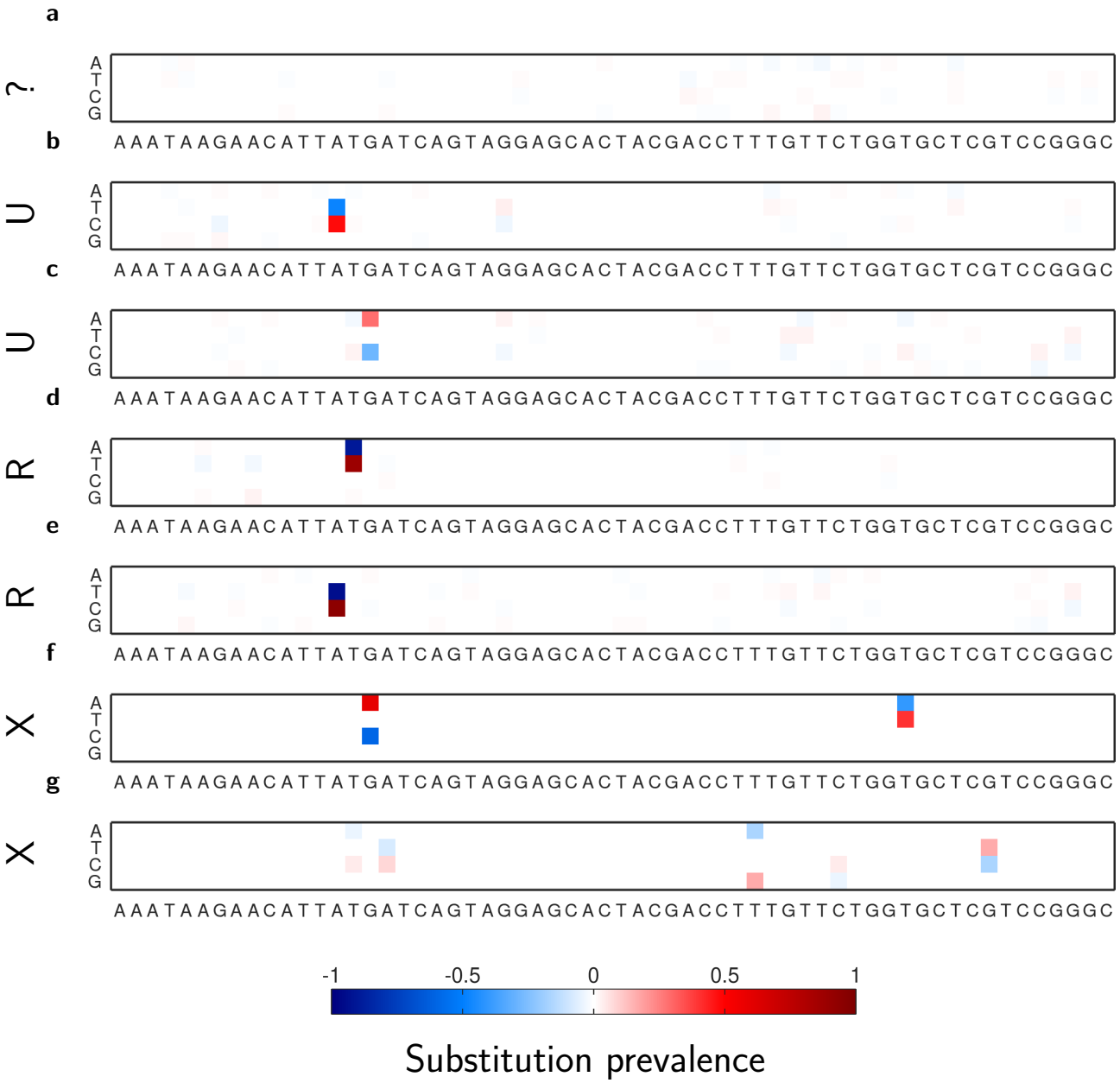



**Figure 1.9:** Outcome space expected out of a mismatch library (here DEB3L in $\Delta mutS$ cells), as viewed in terms of the difference between the observed base frequency distribution of the clan of interest and the expected base distribution based on the consensus sequence. A clan might get repaired making use of the common strand information, yielding a difference matrix of zeros and hence constitutes an ambiguity in the assignment of the original mismatch (?). Replication without repair (U) leads to a difference histogram with a single substitution observed in a subset of the clan members, whereas a successfull repair (R) yields a single substitution shared among virtually all clan members. Random replication, DNA amplification and sequencing errors may give rise to other patterns discarded in the workflow (X).

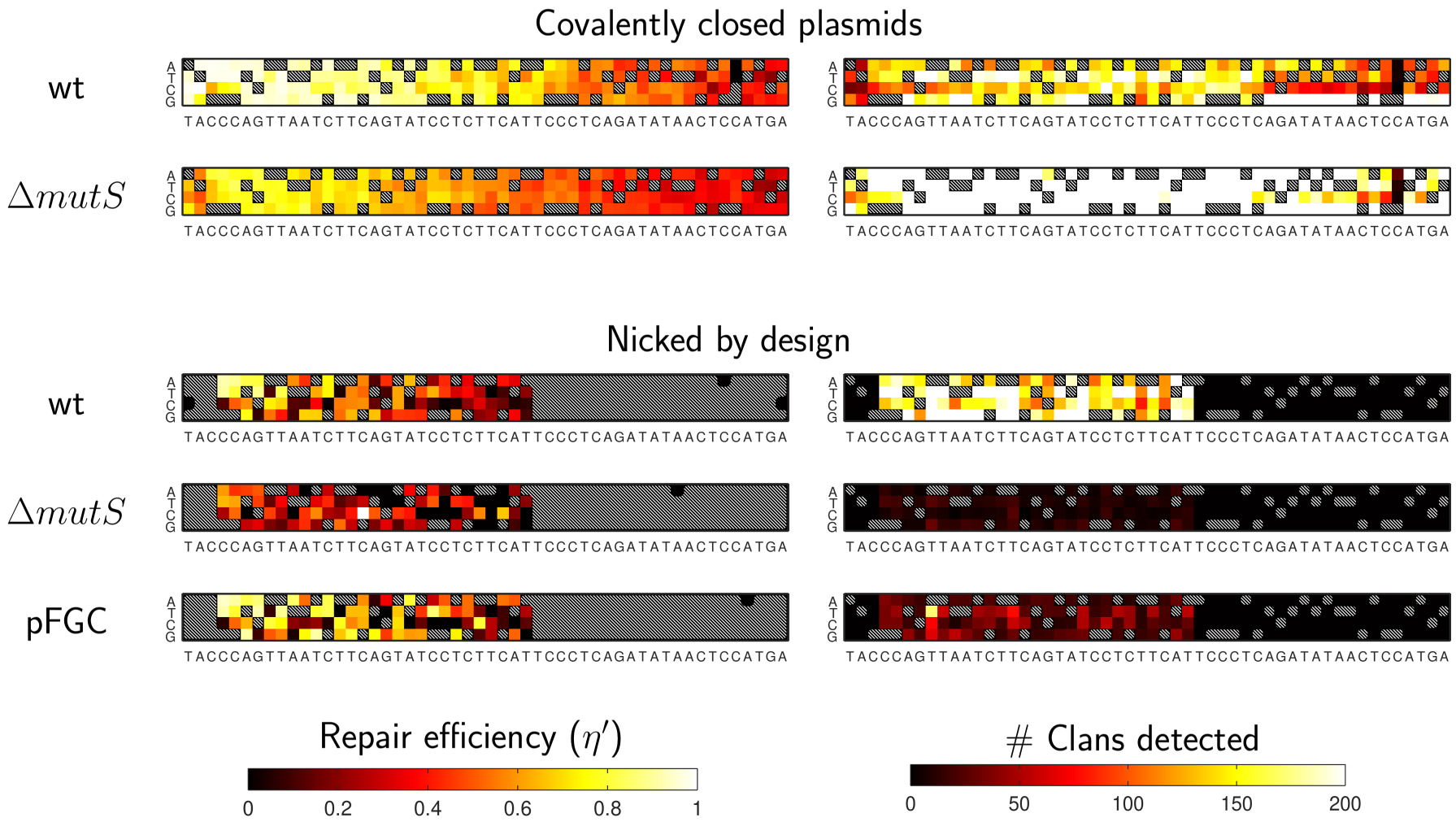



**Figure 1.10:** Repair efficiencies measured in FML. The measured repair efficiencies (left) are indicated with a colorscale, bright yellow indicating mismatches that are almost always repaired before any replication occurs whereas darker shades of red are mismatches that commonly slip repair machinery. The confidence in the measurement is correlated with the total number of clans detected that are attributable to the particular mismatch, whether repaired or unrepaired (right). Black-white hatching pattern ( ) shows positions that were not measured either because they were not contained in any clan (0/0 indeterminacy) or they obey Watson-Crick pairing rule and do not lead to a mismatch requiring correction. Mismatches were only introduced to positions [4,30] in the sample carrying nicks. pFGC refers to an attempt using a cell strain that contains the high copy pFGC5941 plasmid [49] to test the potential effect of the total amount of DNA in the cell that can potentially slow down replication and/or repair machineries.

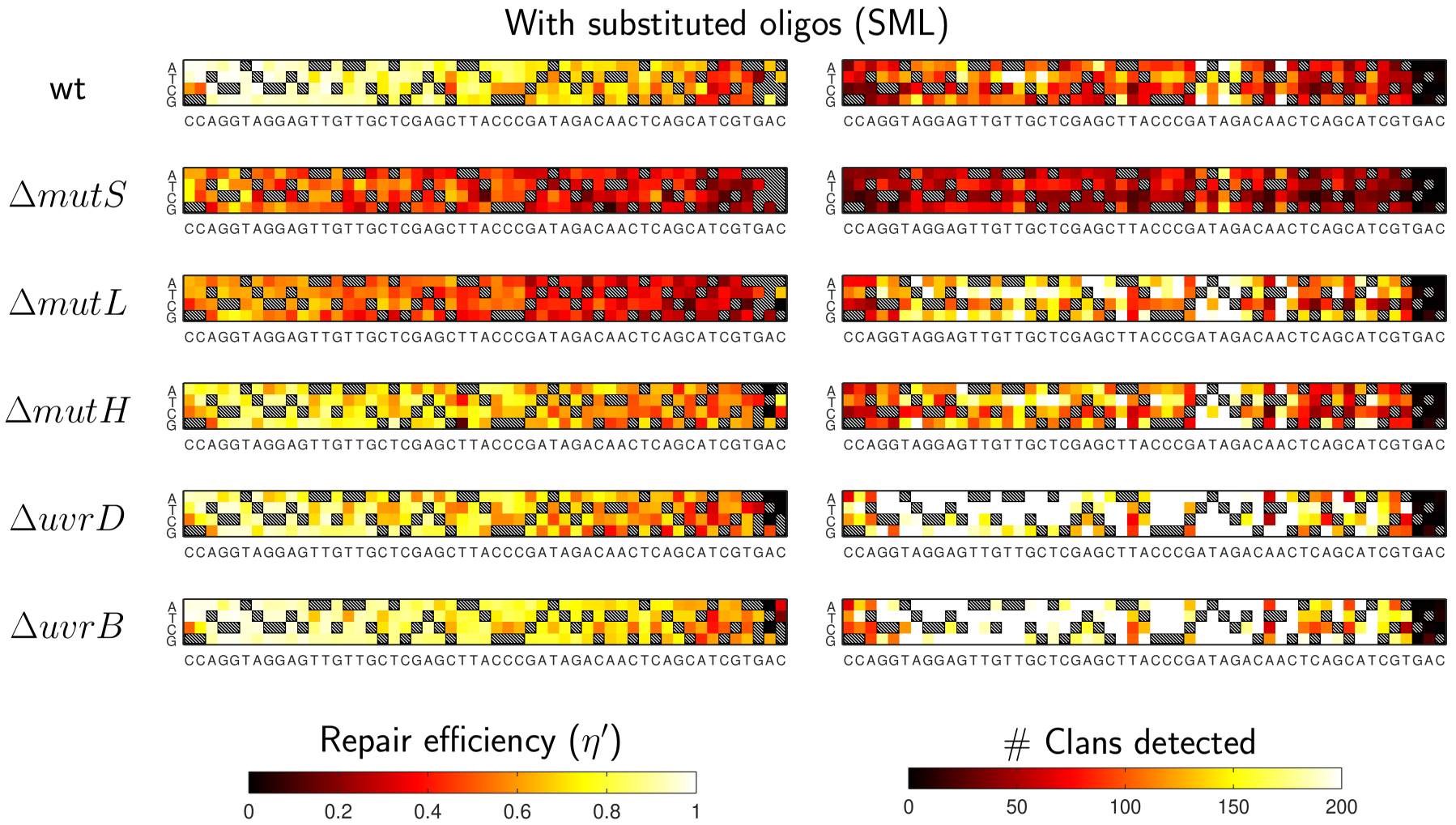



**Figure 1.11:** Same representation as in Figure 1.10, applied to SML consensus sequence. Last three positions do not carry any mismatches by design and hence no clan detection is expected.

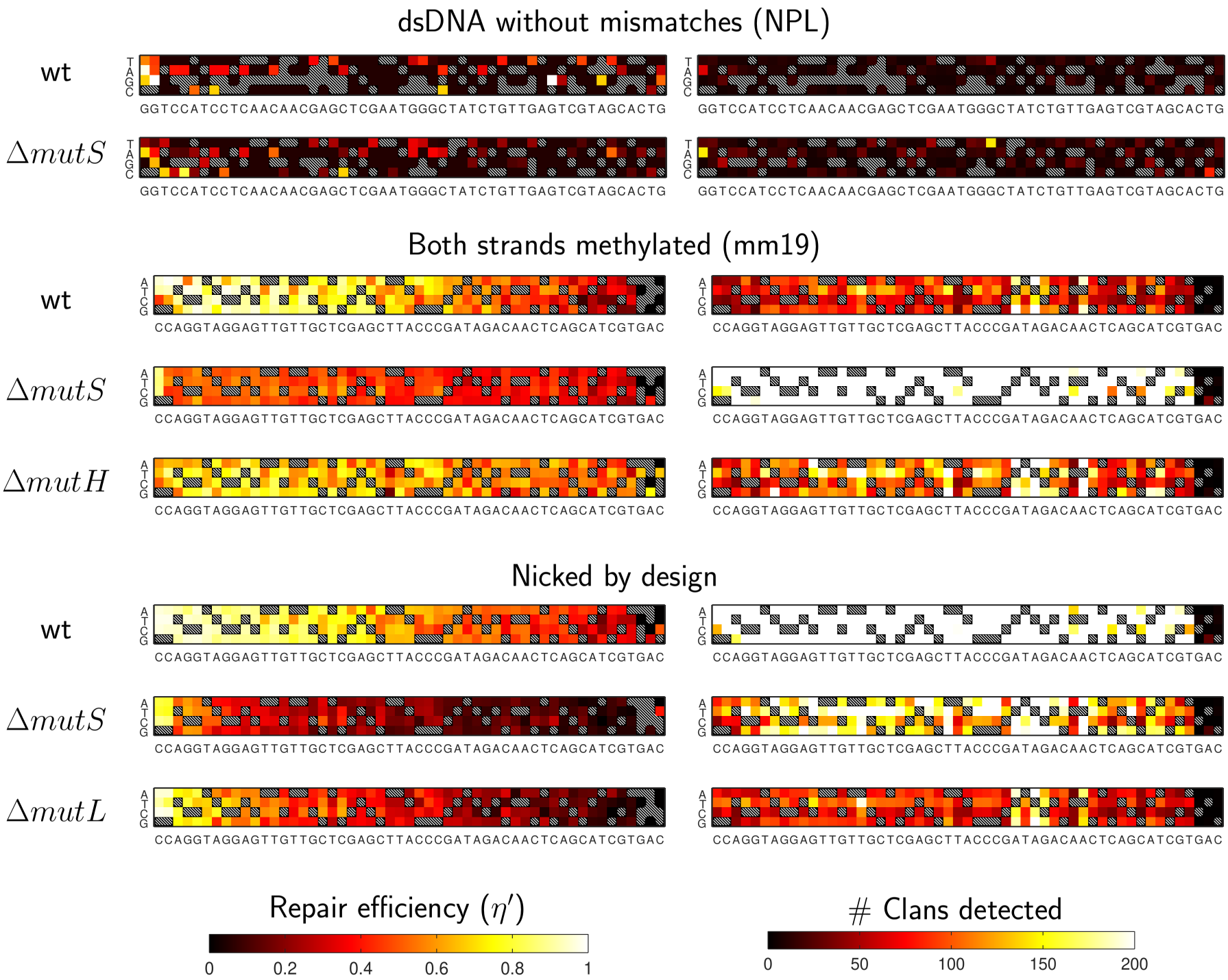



**Figure 1.12:** Same representation as in Figure 1.11. Last three positions do not carry any mismatches by design and hence no clan detection is expected, except for NPL, which does not have any subsitutions anywhere along the inserted DNA sequence.

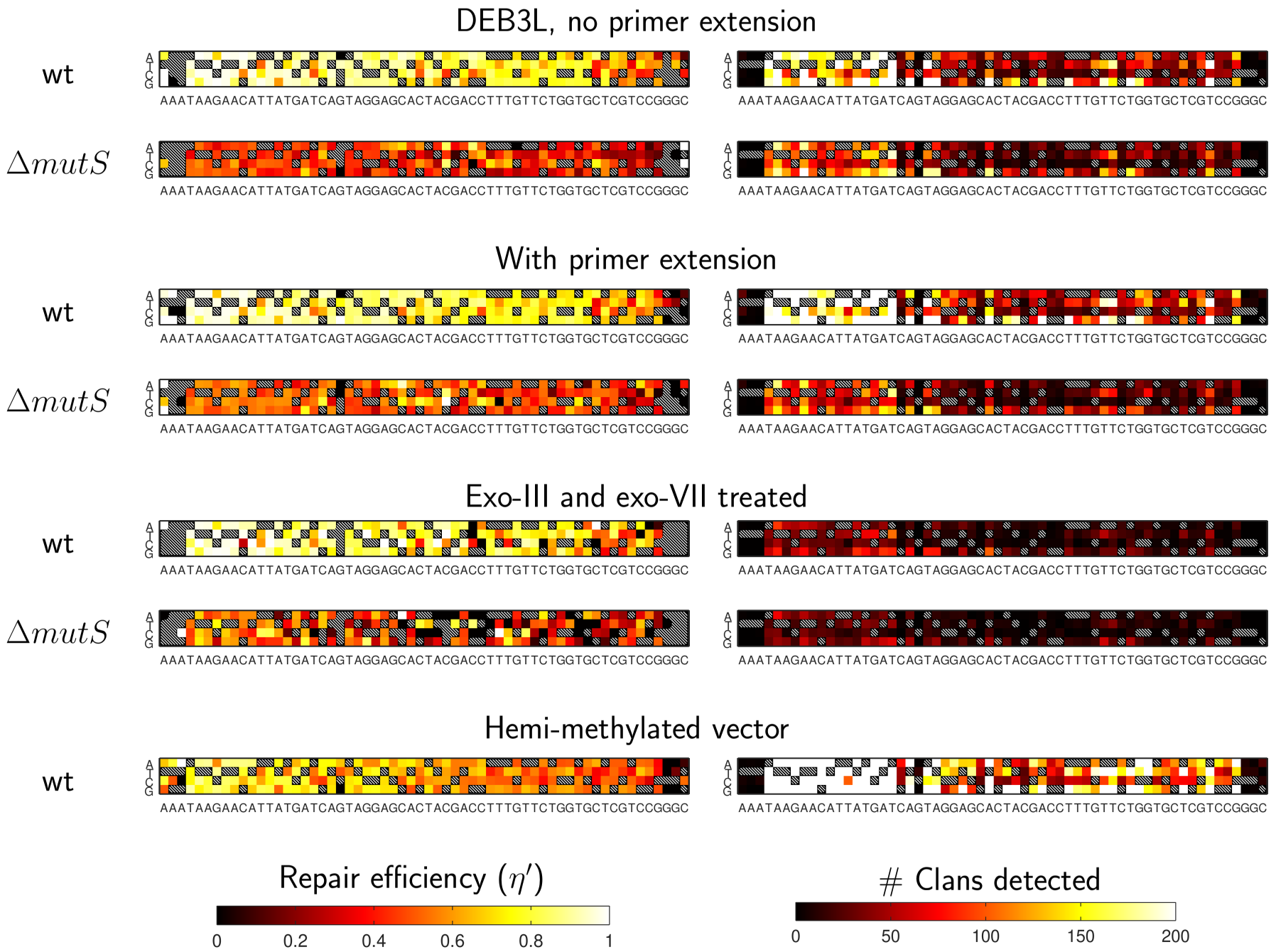



**Figure 1.13:** Same representation as in Figure 1.10, applied to DEB3L consensus sequence. When indicated, "primer extension" refers to annealing and extension of one primer on the barcoded vector library to resolve potential heteroduplexes at the tracing barcode site. All samples were treated with T5-exonuclease, except when indicated otherwise. First and last three positions do not carry any mismatches by design and hence no clan detection is expected.

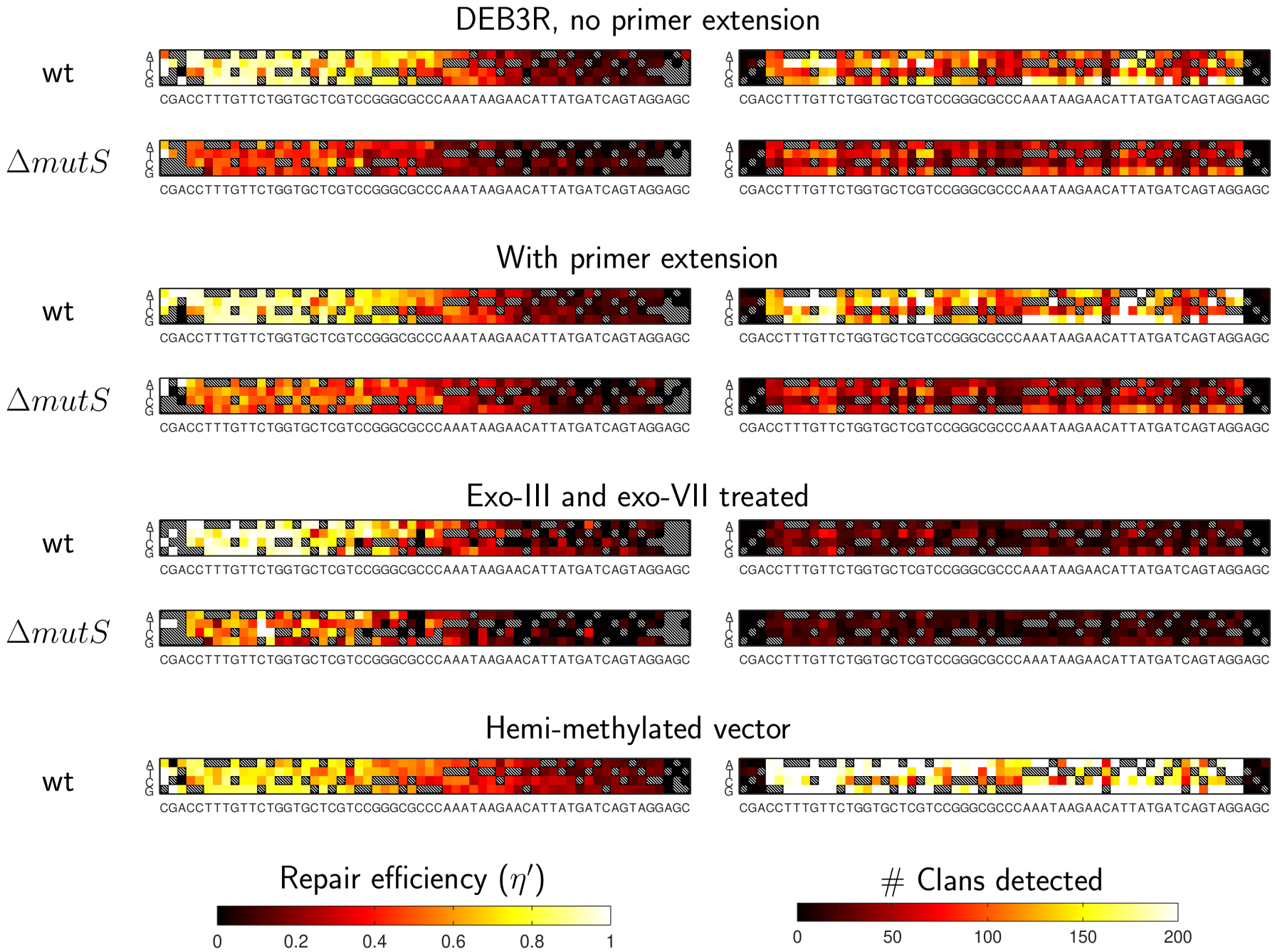



**Figure 1.14:** Same representation as in Figure 1.13, applied to DEB3R consensus sequence.

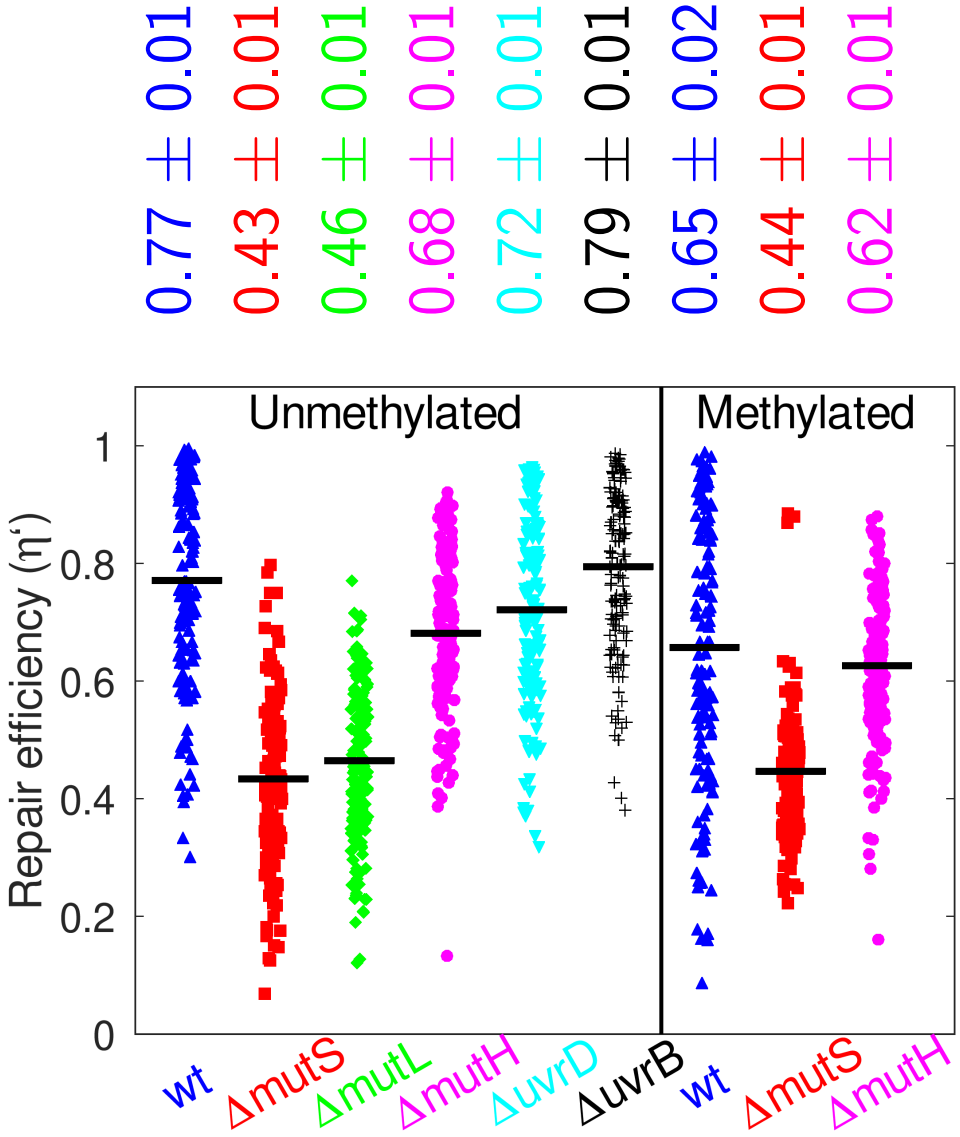



**Figure 1.15:** Comparison of repair efficiencies obtained with different mutant cell strains. Based on one-tailed z-test, the deviation of the means (indicated by horizontal black bars) of unmethylated samples from wt unmethylated sample are statistically highly significantly (p-values ranging from $4 \cdot 10^{-3}$ to $10^{-87}$), except $\Delta uvrB$ (p=0.10). The repair levels of methylated and unmethylated DNA in $\Delta mutS$ are similar (p = 0.18). Fully methylated DNA is repaired by wt or $\Delta mutH$ cells at a statistically highly significantly lower level ($p = 4 \cdot 10^{-7}$ and $p = 2 \cdot 10^{-4}$, respectively). The statistics reported above each group represent mean ± S.E.M. of 150 mismatches that are part of SML. Individual datapoints are cumulative of two to four experimental data sets, thick horizontal bars represent the mean of the respective group.

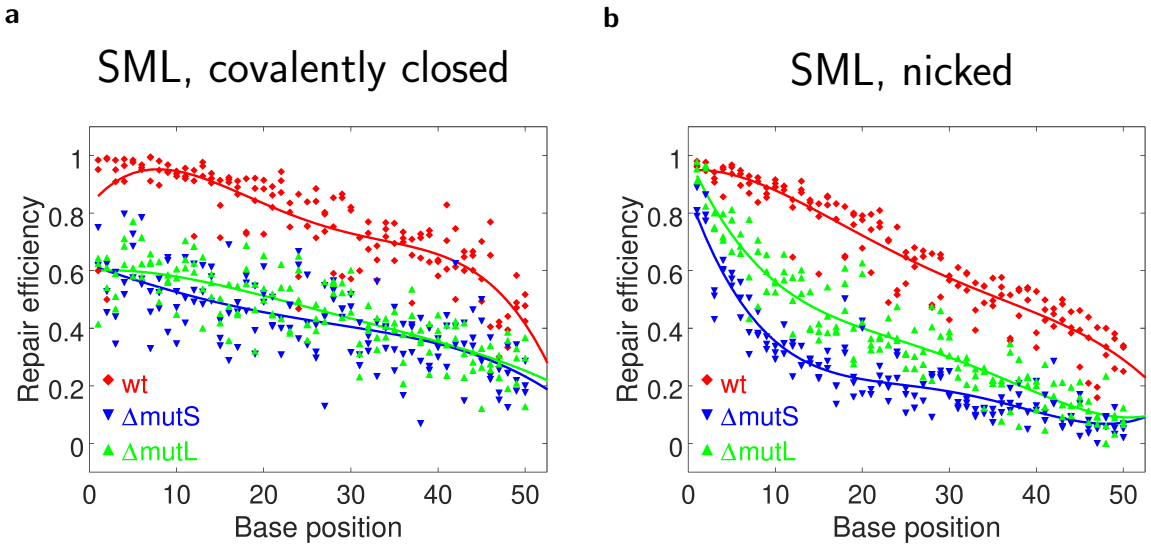



**Figure 1.16:** The observed repair efficiency is highly position dependent. The repair efficiency is higher close to the barcode-proximal side (base 1) and gradually decreases for mismatches that are farther away from the barcode (base 50), the extent of the trend being dependent on the DNA sequence, cell line and the presence of nicks. Straight lines are best-fitting fourth degree polynomials indicating the global trend in the respective dataset, each obtained from cumulative data out of independent experimental triplicates.

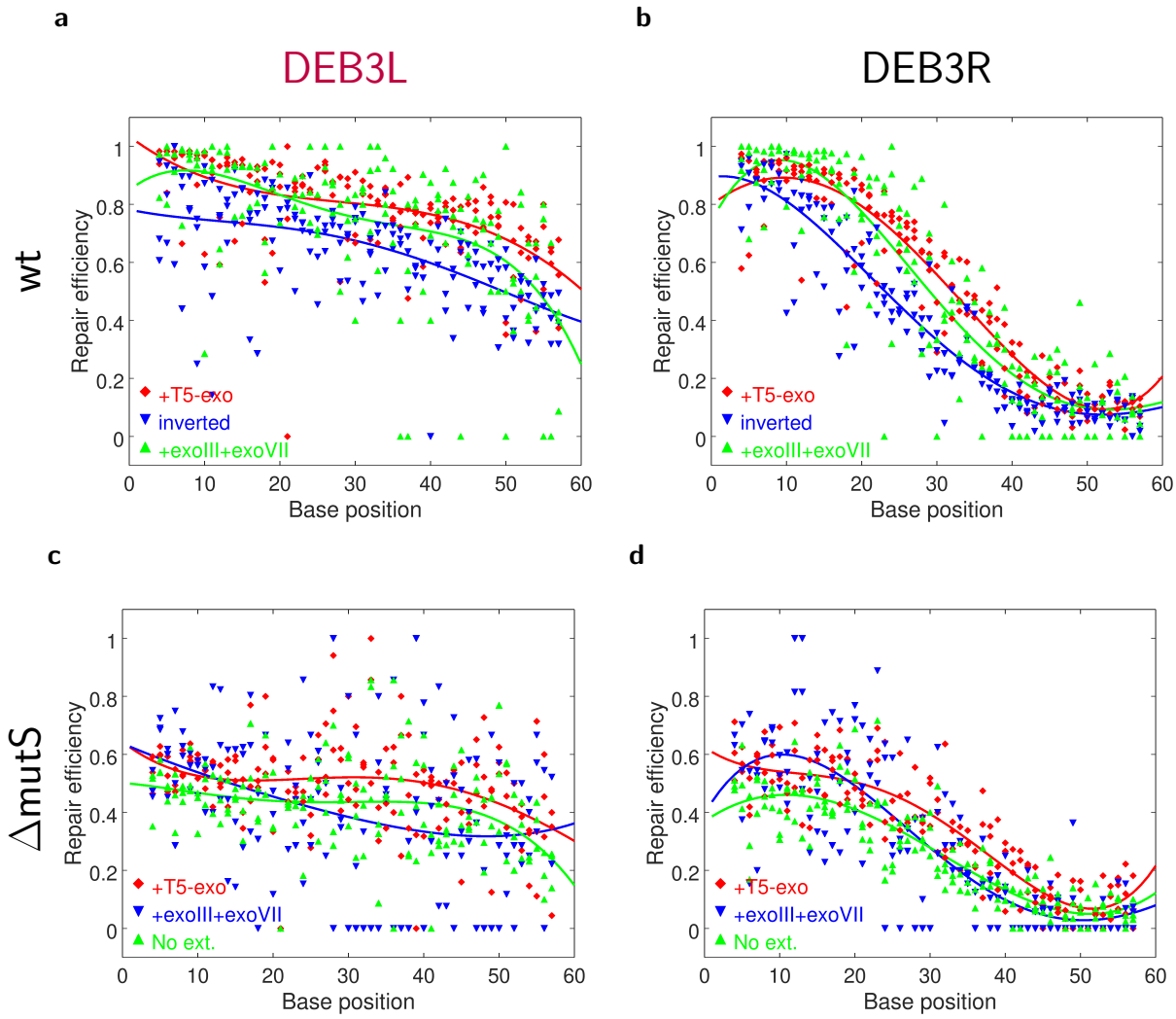



**Figure 1.17:** The position dependence observed in DEB3 libraries, demonstrated in an analogous way to Figure 1.16. Qualitatively same trend is observable if the library insertion direction is flipped ("inverted"). The omission of bubble mitigation procedure by primer extension also does not have a significant effect ("No ext"). Exonuclease treatment attempt to eliminate nicked molecules do not abolish the observed trends ("+T5-exo" or "+exoIII+exoVII"). Solid curves are best-fitting fourth degree polynomials indicating the global trend in the respective data set, each obtained from cumulative of two to four experimental replicates.

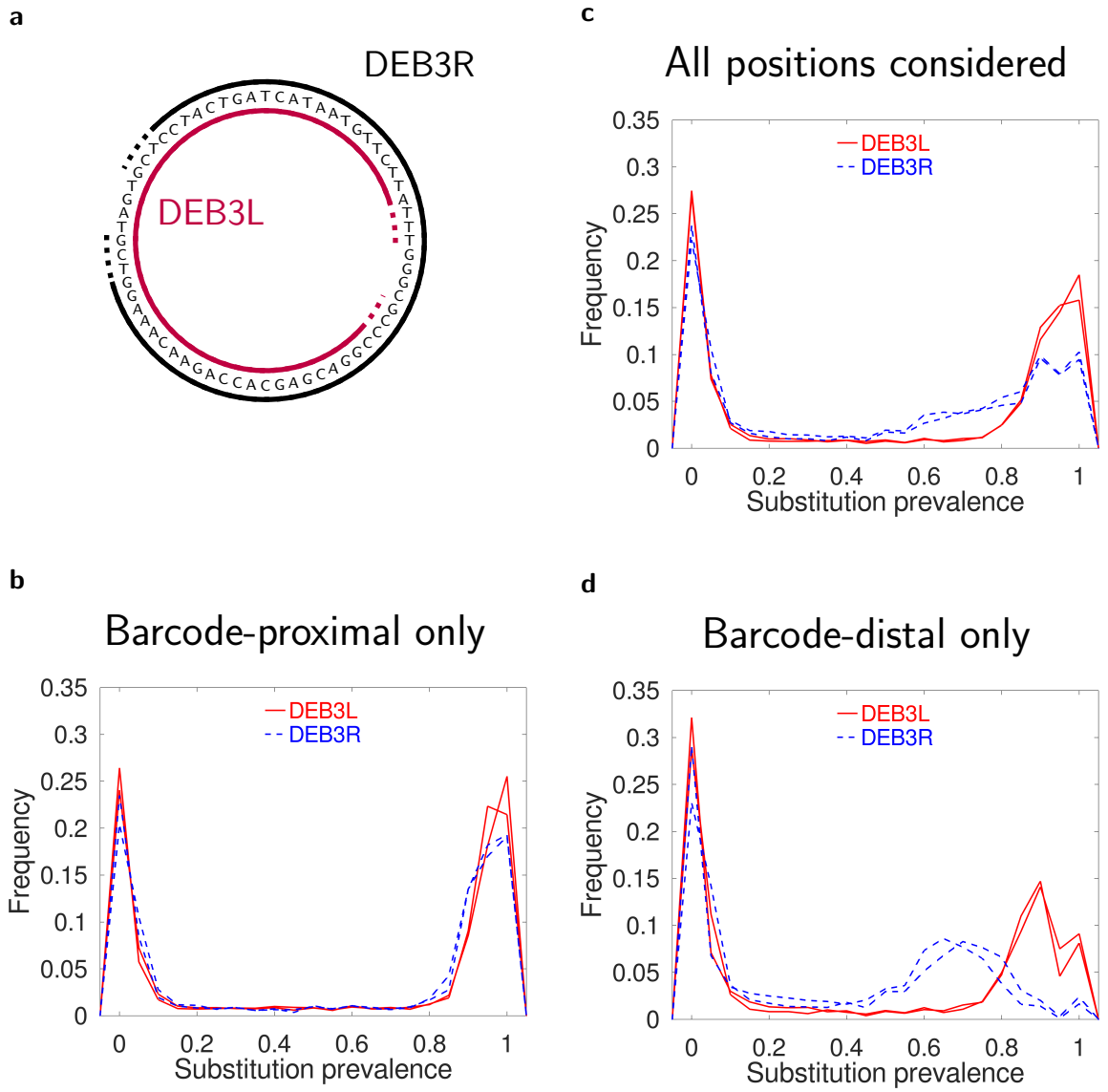



**Figure 1.18:** (a) A de Bruijn sequence containing all sequence trimers and the DEB3L and DEB3R mismatch libraries sampling this sequence. Position dependent shifts on the clan prevalence histograms are sequence dependent. (c) Substitution prevalence histograms differ for DEB3L and DEB3R libraries. This difference largely originates from the discrepancy observed in the barcode-distal end of the library (histogram of last 20 positions only, d), while the barcode proximal end behave largely similarly in the two cases (histograms of the first 20 positions only, b). Each condition is represented by two separate curves of identical color, representing two independent experimental replicates.

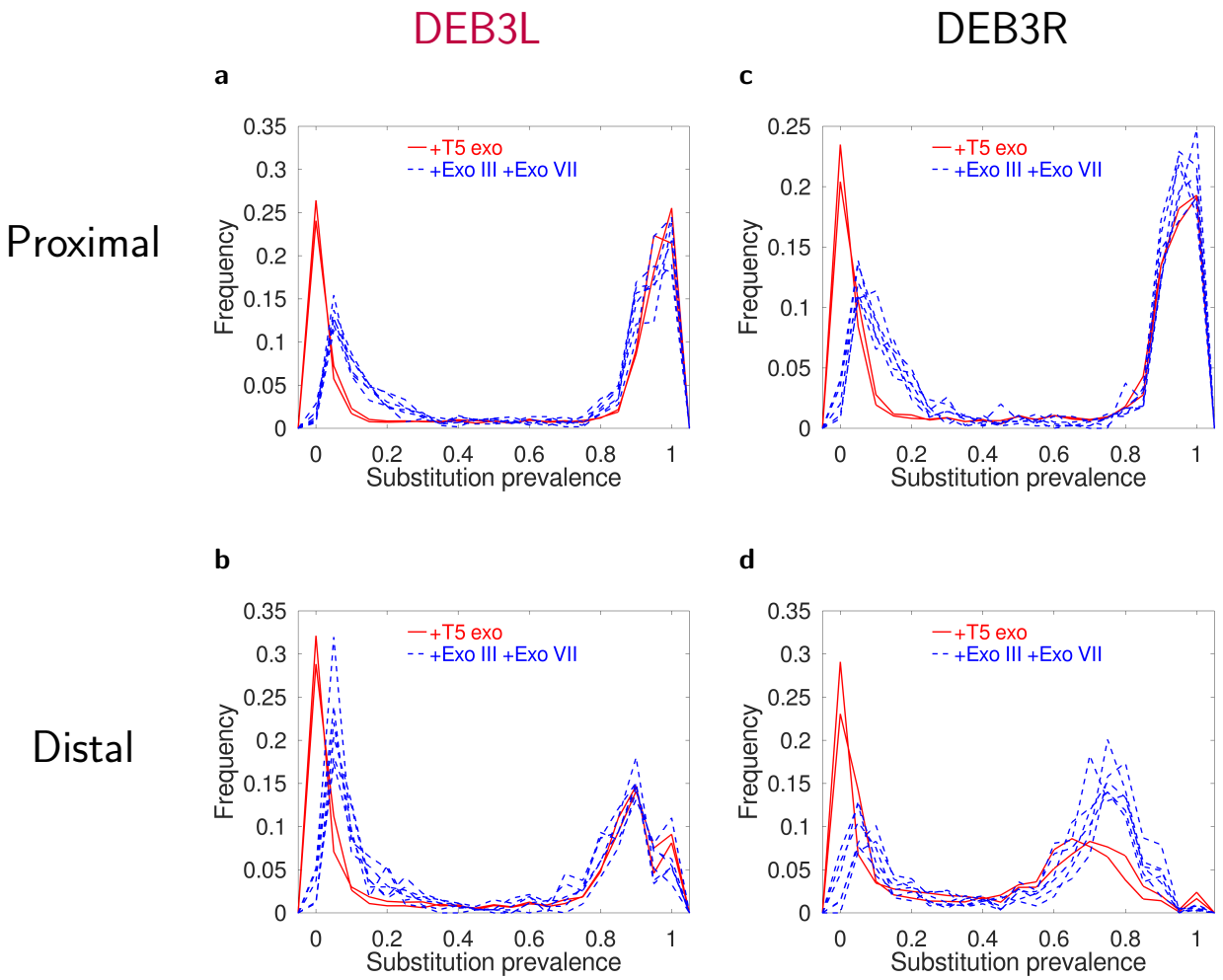



**Figure 1.19:** The nicks on the DNA partially explains the position dependence of clan prevalence distribution. The histogram for the barcode-proximal 20 positions is less affected by the exonuclease treatment to eliminate the nicked molecules for DEB3L (a) or DEB3R (c). The position of the V-peaks shift for both libraries with both exonuclease treatment protocols, but the extent is more limited for DEB3L (b) than DEB3R (d). A switch to a 3'→5' exonuclease system shifts the retained strand preference from common strand to the variable strand in DEB3R. Experiments were performed on an equimolar mixture of DEB3L and DEB3R in wt cells. T5 exonuclease was tested on 2 replicates shown in two solid red lines (—), 6 experimental replicates with exonucleases III & VII treatment of the plasmid are indicated in blue dashed line (- - -).

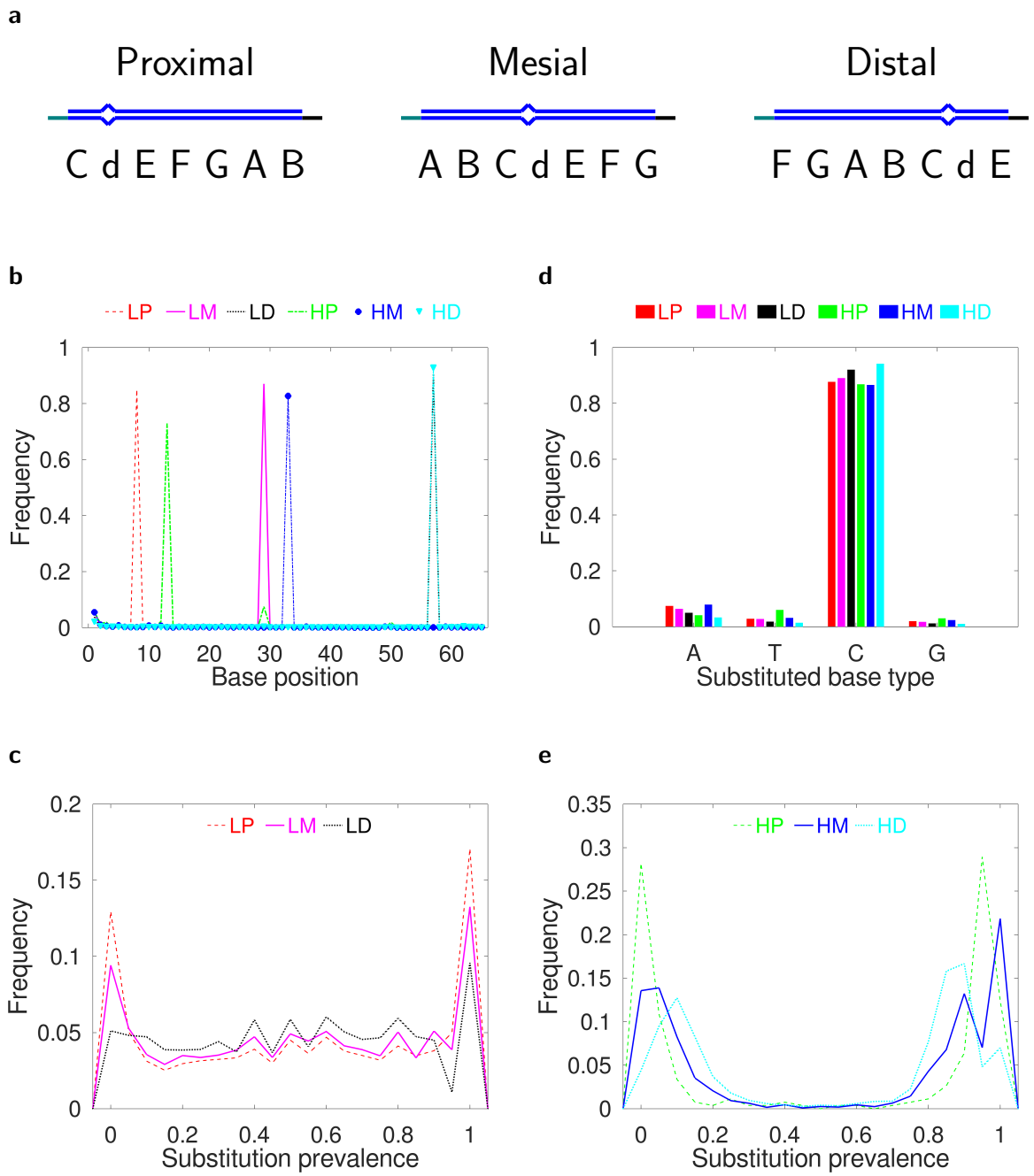



**Figure 1.20:** A control experiment using 6 mismatch libraries each containing only one type of mismatch. (a) SML sequence is cyclically permuted to bring the mismatches to the barcode-proximal, mesial and distal positions while keeping the immediate sequence context constant. (b) The position of the mismatch deduced by the maximally substituted nucleotide position of each clan. (d) Deduced mismatch type using the identity of the most commonly substituted base at the deduced position of the mismatch. Substitution prevalence histograms for the three control libraries that exclusively contain a CC (c) or CT mismatch (e). All data are calculated from the combined output of two independent experimental replicates.

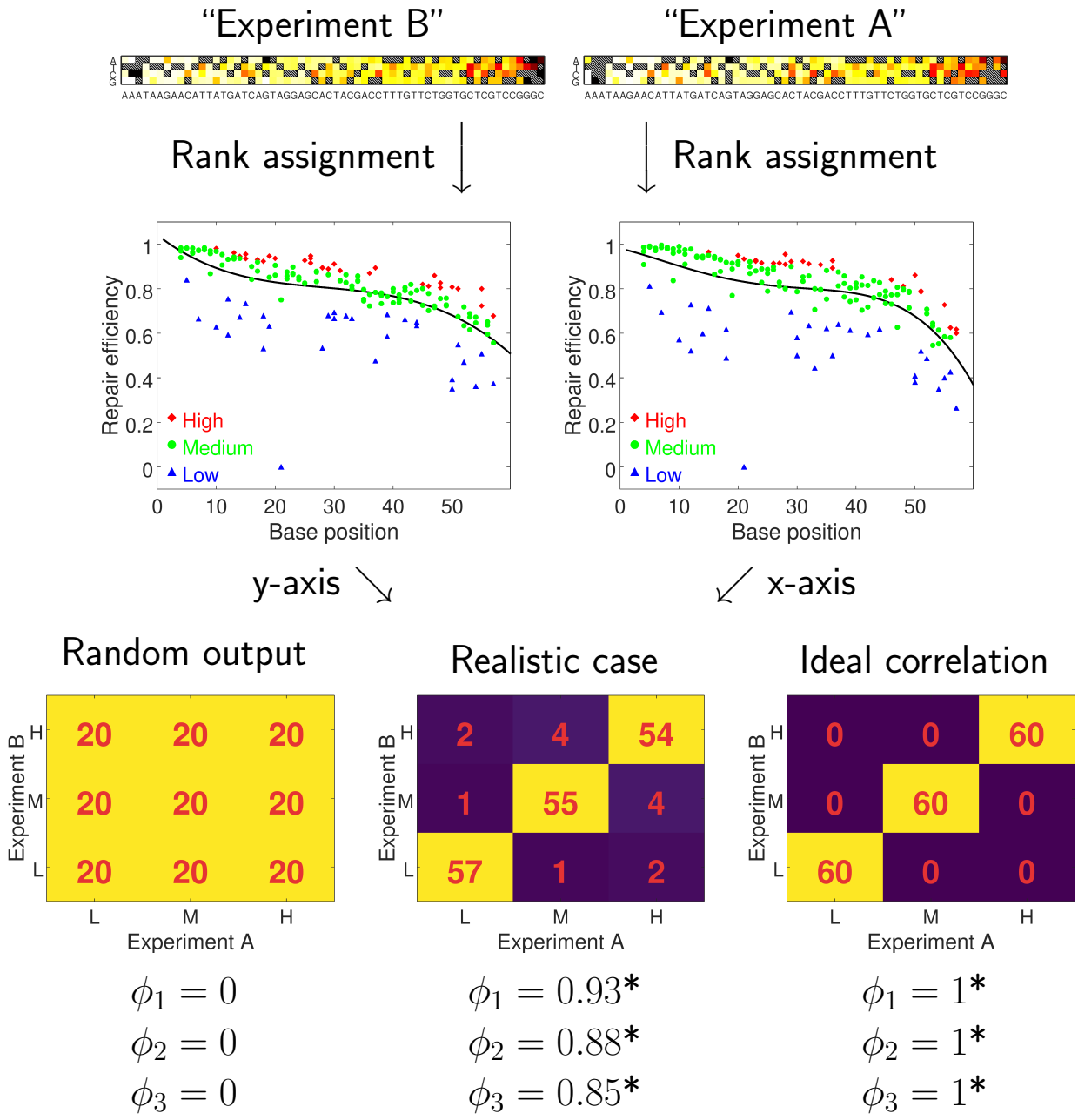



**Figure 1.21:** Ternary classification of mismatches based on relative ranking. For each experimental dataset, the global trend in the repair efficiency is estimated by fitting with a 4th degree polynomial (black line). The mismatches are assigned to High, Medium and Low repair efficiency classes if their repair efficiency exceeds, is within 0.5 mean deviation of or falls below the general trend, respectively. Different data sets are compared by means of the correlation between the ranks assigned to the same mismatch, evaluated by the mean square contingency coefficient ($\phi$). $\phi$ attains a value of 0 for totally random output and 1 for an idealized fully reproducible system and other values for realistic experimental cases. Based on Table 1.3, statistically highly significant $\phi$ values (p-value <0.01) are indicated with an asterisk (*).

**Table 1.3:** Minimum value of $\phi$ required for statistical significance ($\phi_{p<0.05}$) and statistical high significance $\phi_{p<0.01}$, as a function of the total number of positions included in the data set (N).

| N | $\lvert\phi_{p<0.05}\rvert$ | $\lvert\phi_{p<0.01}\rvert$ | N | $\lvert\phi_{p<0.05}\rvert$ | $\lvert\phi_{p<0.01}\rvert$ | N | $\lvert\phi_{p<0.05}\rvert$ | $\lvert\phi_{p<0.01}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.96 | 2.58 | 21 | 0.43 | 0.56 | 120 | 0.18 | 0.24 |
| 2 | 1.39 | 1.82 | 22 | 0.42 | 0.55 | 140 | 0.18 | 0.24 |
| 3 | 1.13 | 1.49 | 23 | 0.41 | 0.54 | 160 | 0.17 | 0.22 |
| 4 | 0.98 | 1.29 | 24 | 0.40 | 0.53 | 180 | 0.15 | 0.20 |
| 5 | 0.88 | 1.15 | 25 | 0.39 | 0.52 | 200 | 0.15 | 0.19 |
| 6 | 0.80 | 1.05 | 26 | 0.38 | 0.51 | 300 | 0.14 | 0.18 |
| 7 | 0.74 | 0.97 | 27 | 0.38 | 0.50 | 400 | 0.11 | 0.15 |
| 8 | 0.69 | 0.91 | 28 | 0.37 | 0.49 | 500 | 0.10 | 0.13 |
| 9 | 0.65 | 0.86 | 29 | 0.36 | 0.48 | 600 | 0.09 | 0.12 |
| 10 | 0.62 | 0.81 | 30 | 0.36 | 0.47 | 700 | 0.08 | 0.11 |
| 11 | 0.59 | 0.78 | 35 | 0.33 | 0.44 | 800 | 0.07 | 0.10 |
| 12 | 0.57 | 0.74 | 40 | 0.31 | 0.41 | 900 | 0.07 | 0.09 |
| 13 | 0.54 | 0.71 | 45 | 0.29 | 0.38 | 1000 | 0.07 | 0.09 |
| 14 | 0.52 | 0.69 | 50 | 0.28 | 0.36 | 2000 | 0.06 | 0.08 |
| 15 | 0.51 | 0.67 | 55 | 0.26 | 0.35 | 3000 | 0.04 | 0.06 |
| 16 | 0.49 | 0.64 | 60 | 0.25 | 0.33 | 4000 | 0.04 | 0.05 |
| 17 | 0.48 | 0.62 | 70 | 0.23 | 0.31 | 5000 | 0.03 | 0.04 |
| 18 | 0.46 | 0.61 | 80 | 0.22 | 0.29 | 6000 | 0.03 | 0.04 |
| 19 | 0.45 | 0.59 | 90 | 0.21 | 0.27 | 7000 | 0.03 | 0.03 |
| 20 | 0.44 | 0.58 | 100 | 0.20 | 0.26 | 8000 | 0.02 | 0.03 |

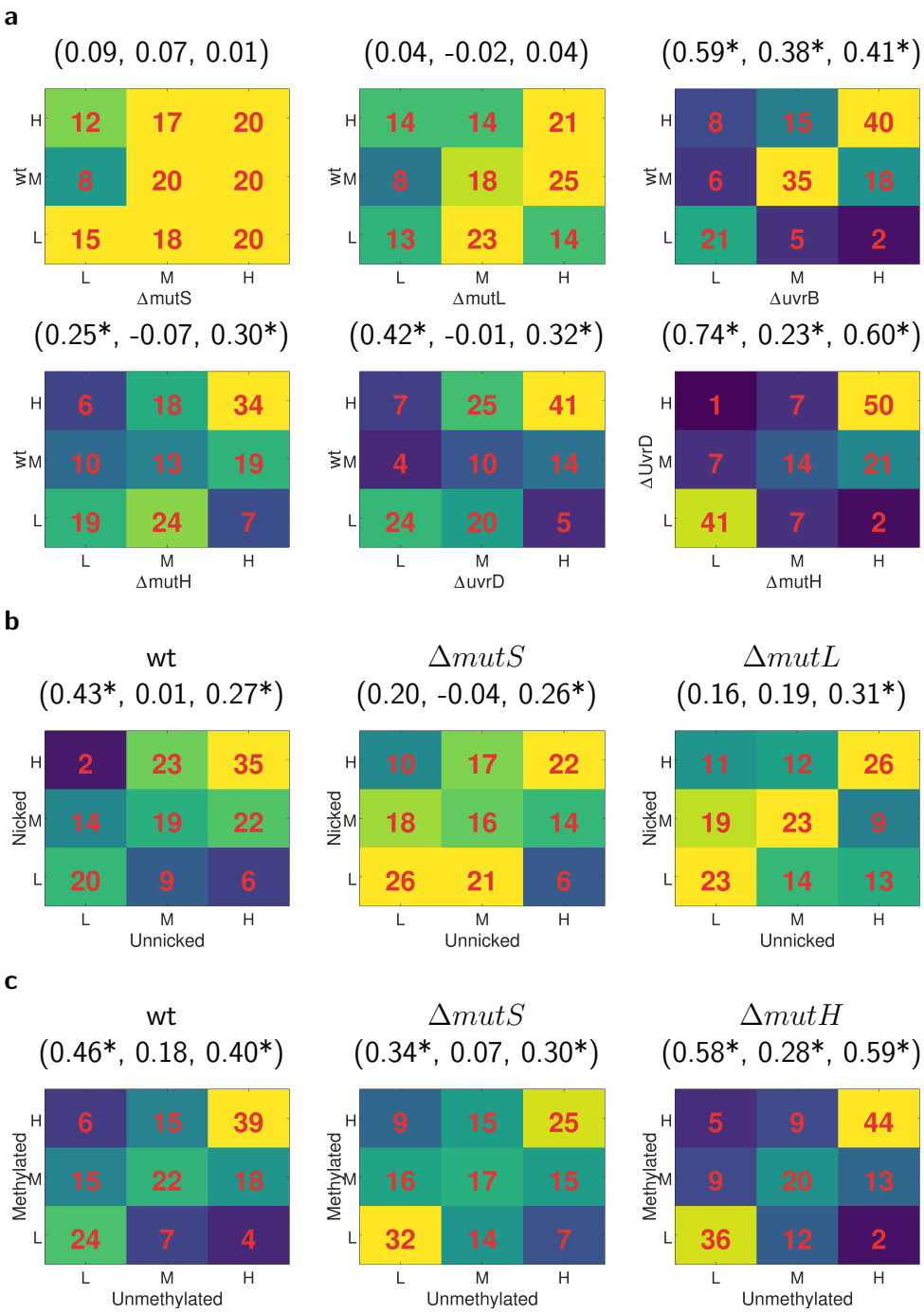



**Figure 1.22:** Ternary class comparison matrices for SML. (a) The correlation between datasets obtained with a functional MMR pathway correlate more significantly than MMR deficient cells. The relative efficiency of mismatches obtained with and without methylation largely correlates except in $\Delta mutS$ cells (c), introduction of nicks similarly provides similar relative efficiency in wt but not in $\Delta mutS$ or $\Delta mutL$ cells (b). The triplets indicated above individual matrices indicate the mean square contingency coefficients $(\phi_1, \phi_2, \phi_3)$, respectively. Based on Table 1.3, statistically highly significant $\phi$ values (p-value <0.01) are indicated with an asterisk (*). All samples are cumulative of two to four experimental replicates, samples were neither methylated nor nicked unless indicated otherwise.

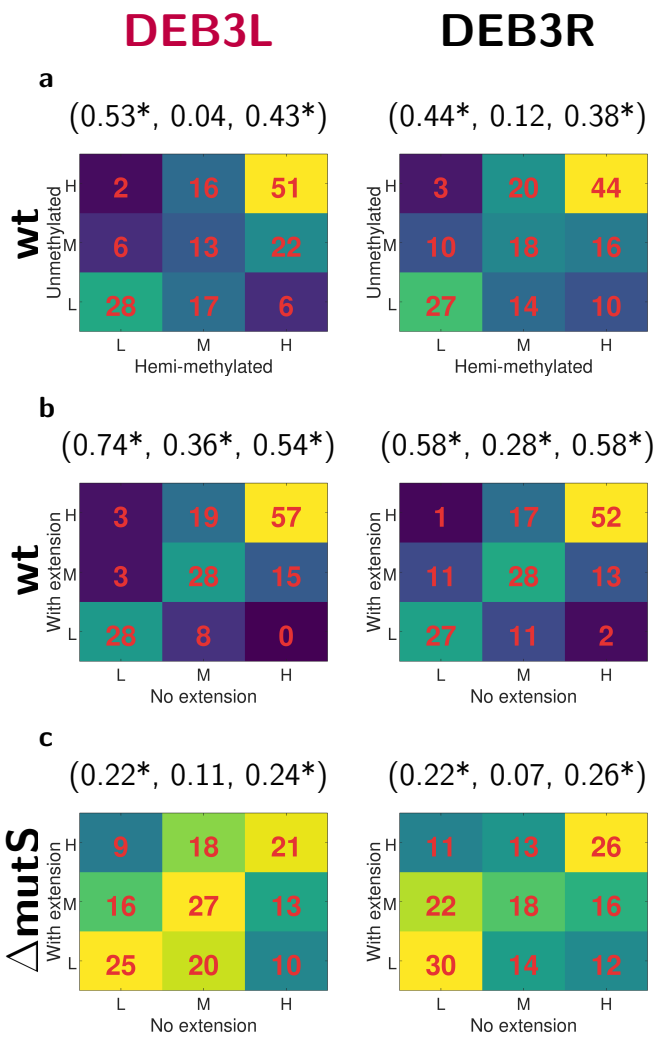



**Figure 1.23:** Ternary class comparison matrices analogous to Figure 1.22, but for DEB3 library, considering the assigned classes of all 162 mismatches per dataset. Relative repair efficiency of mismatches is largely independent of the presence of asymmetric methylation (a). Removal of potential heteroduplexes by primer extension mostly conserves the set of difficult to repair outliers in wt cells (b), but the correlation is weaker in $\Delta mutS$ cells (c).

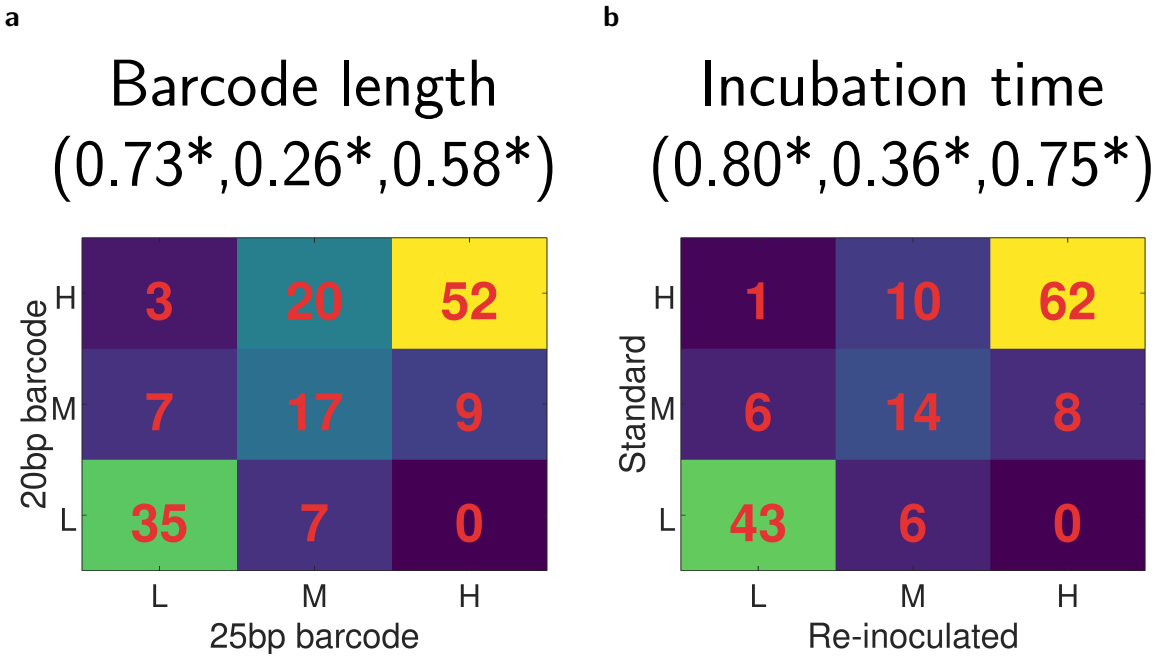



**Figure 1.24:** Chosen barcode length and incubation time are adequate and the proposed system tracks clans with an acceptable accuracy. (a) The chosen barcode diversity is adequate and reduction of barcode length from 25 to 20bp does not cause a significant change in the detected outliers observed in wt cells against methylated plasmids. (b) The assay is insensitive to excessive cell porpagation as the assigned ranks are mostly unchanged if the overnight grown culture of ΔuvrD cells is used to re-inoculate a fresh culture and sequenced after one more overnight incubation. All data sets for the comparisons were drawn from SML.

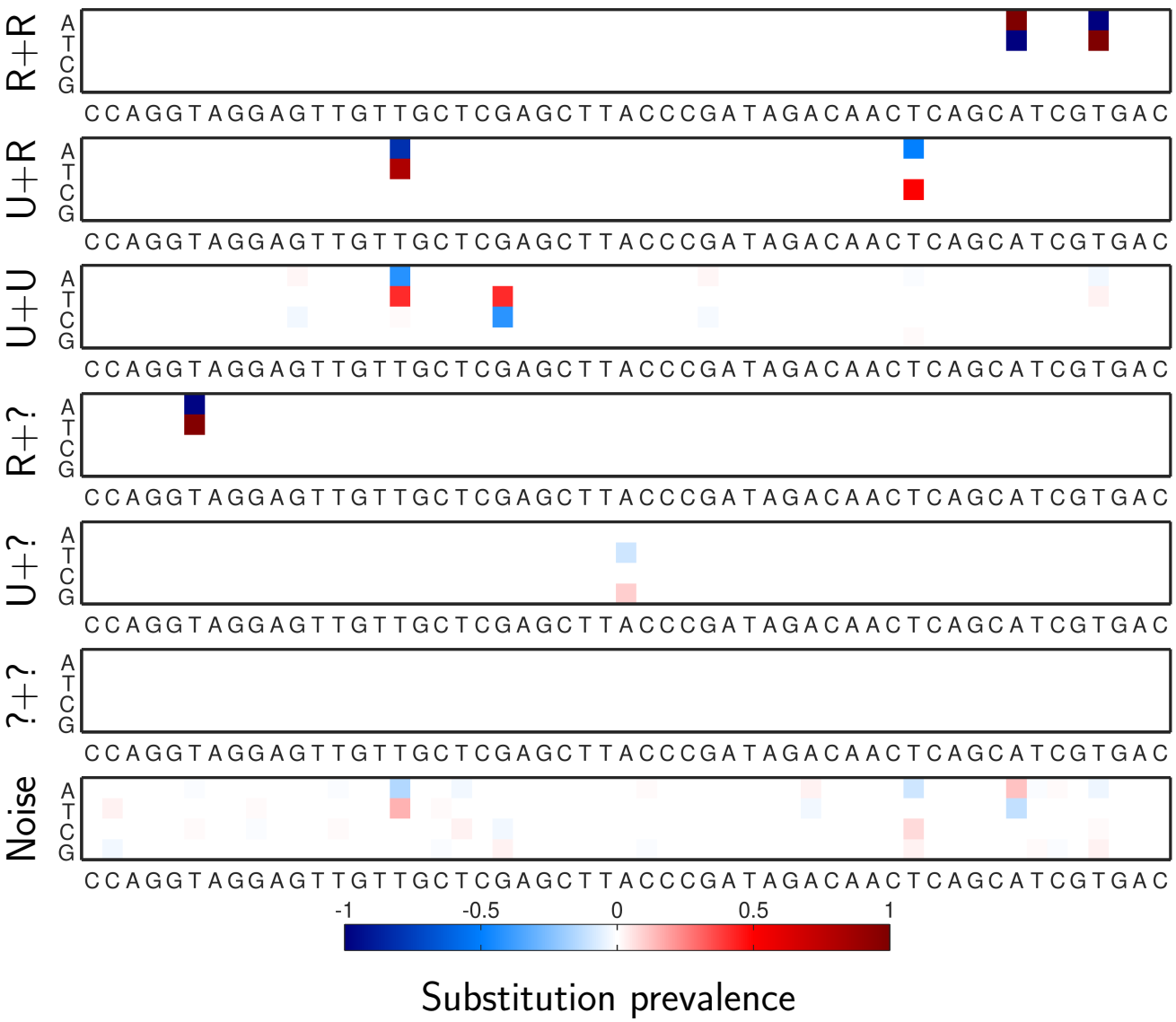



**Figure 1.25:** Outcome space expected from the double mismatch library. Example single-clan difference histograms that represent the detectable outcome space from a starting sample containing two and only two mismatches by design, otherwise analogous to Figure 1.9. Each mismatch can be repaired by keeping the variable strand (R) or by retaining the common strand (?) or undergo replication directly without a preceding repair event (U). Of the 7 types of events that can be detected, only three (R+R, R+U, U+U) convey information regarding the type and position of the mismatch on the ancestor molecule.

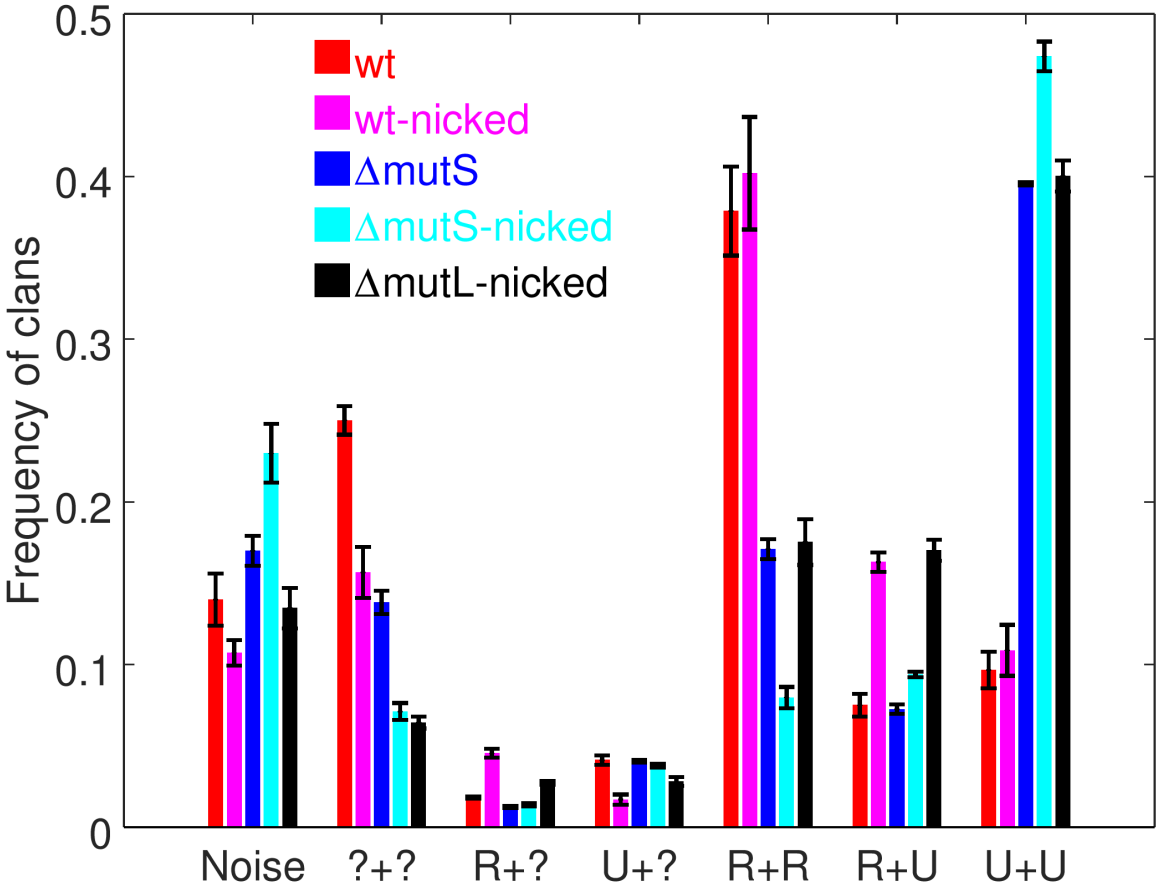



**Figure 1.26:** Detected frequency of 7 expected classes of events, averaged over the position and the type of the mismatches. Results are reported as mean ± S.E.M of three experimental replicates.

**a**

5'-ggTCCatCCTcaACAacGAGctCGAatGGGctATCtgTTGagTCGtaGcaCTG-3'

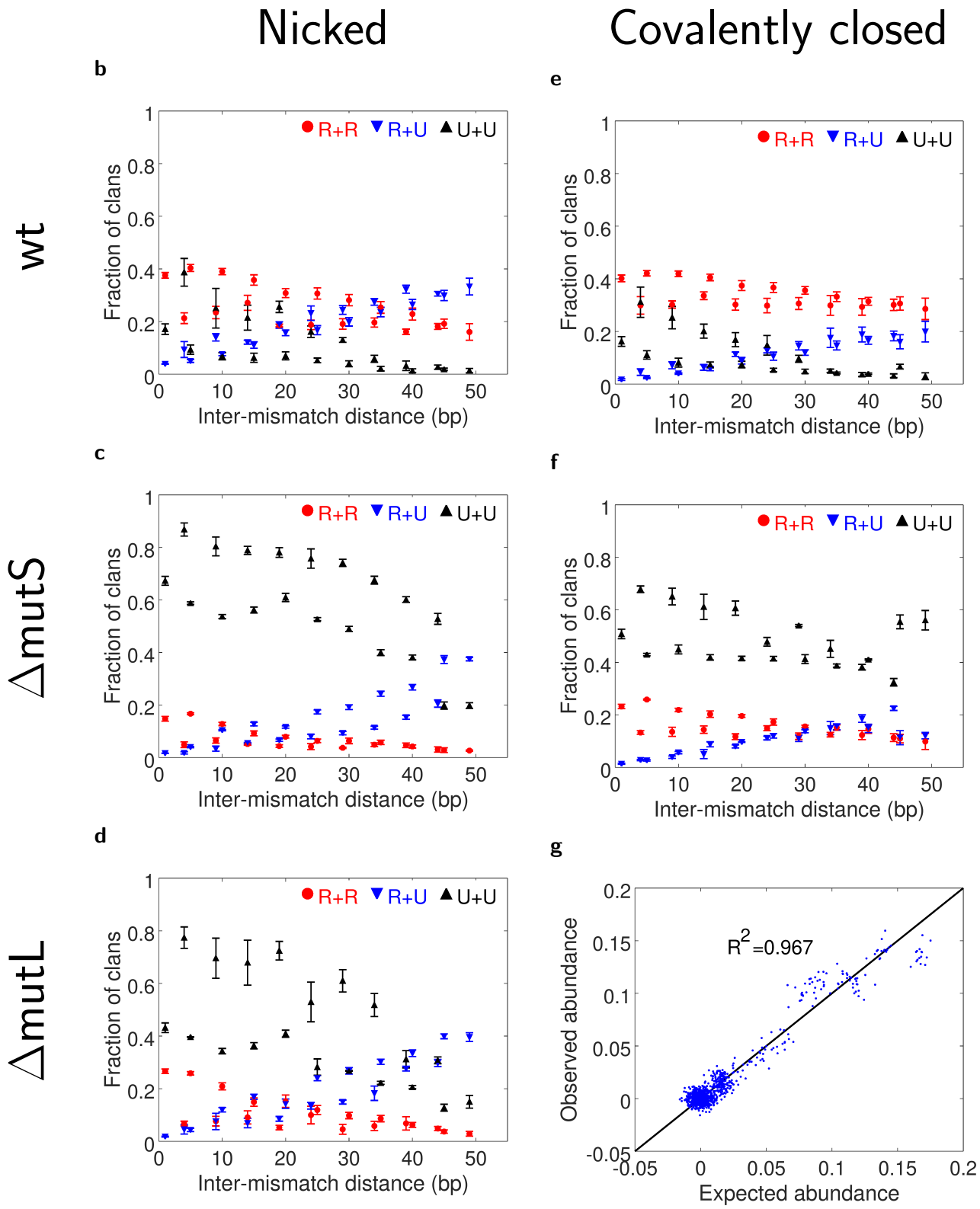


**Figure 1.27:** The relative frequency of three unambiguious clan classes depends on the base spacing between the two mismatches. The substituted positions within the sequence studied to form mismatches are shown in lower case in (a), the other positions are not sampled as part of any double mismatch combination. Results are reported as mean ± S.E.M of three experimental replicates, the distance combinations that are not represented in the library have been omitted for clarity (0/0 indeterminacy). Note that the alternating oscillation behavior is an expected artifact of the library design, which samples the mismatch pair possibilities non-uniformly (g, cumulative of all five data sets).

**Table 1.4:** The list of all experimental conditions tested. All samples were sequenced with Illumina MiSeq.

| Exp No | cell strain | vector backbone | library | Date sequenced | i7 index | i5 index |
|---|---|---|---|---|---|---|
| 1 | ΔmutS | tUNC19 | DEB3L+DEB3R | 02012019 | N701 | S517 |
| 2 | ΔmutS | tUNC19 | DEB3L+DEB3R | 02012019 | N702 | S517 |
| 3 | ΔmutS | tUNC19-x | DEB3L+DEB3R | 02012019 | N703 | S517 |
| 4 | ΔmutS | tUNC19-x | DEB3L+DEB3R | 02012019 | N704 | S517 |
| 5 | wt | tUNC19-x | DEB3L+DEB3R | 02012019 | N705 | S517 |
| | | | | 06022019 | | |
| 6 | wt | tUNC19-x | DEB3L+DEB3R | 02012019 | N706 | S517 |
| | | | | 06022019 | | |
| 7 | wt | tUNC19 | DEB3L+DEB3R | 06022019 | N701 | S503 |
| 8 | wt | tUNC19 | DEB3L+DEB3R | 06022019 | N702 | S503 |
| 9 | wt | hm19 | DEB3L+DEB3R | 19022019 | N703 | S504 |
| 10 | wt | hm19 | DEB3L+DEB3R | 19022019 | N704 | S504 |
| 11 | wt | 91CNUt | DEB3R | 12032019 | N701 | S502 |
| 12 | wt | 91CNUt | DEB3L+DEB3R | 12032019 | N702 | S502 |
| 13 | wt | 91CNUt | DEB3L | 12032019 | N703 | S502 |
| 14 | wt | 91CNUt | DEB3L | 12032019 | N704 | S502 |
| 15 | wt | tUNC19 + exoIII/VII | DEB3L+DEB3R | 27032019 | N701 | S503 |
| 16 | wt | tUNC19 + exoIII/VII | DEB3L+DEB3R | 27032019 | N702 | S503 |
| 17 | wt | tUNC19 + exoIII/VII | DEB3L+DEB3R | 27032019 | N703 | S503 |
| 18 | wt | tUNC19 + exoIII/VII | DEB3L+DEB3R | 27032019 | N701 | S504 |
| 19 | wt | tUNC19 + exoIII/VII | DEB3L+DEB3R | 27032019 | N702 | S504 |
| 20 | wt | tUNC19 + exoIII/VII | DEB3L+DEB3R | 27032019 | N703 | S504 |
| 21 | ΔmutS | tUNC19 + exoIII/VII | DEB3L+DEB3R | 27032019 | N704 | S503 |
| 22 | ΔmutS | tUNC19 + exoIII/VII | DEB3L+DEB3R | 27032019 | N705 | S503 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 23 | ΔmutS | tUNC19 + exoIII/VII | DEB3L+DEB3R | 27032019 | N706 | S503 |
| 24 | wt | tUNC19 | NPL | 21122017 | N704 | S503 |
| 25 | wt | tUNC19 | NPL | 21122017 | N705 | S503 |
| 26 | wt | tUNC19 | NPL | 21122017 | N706 | S503 |
| 27 | ΔmutS | tUNC19 | NPL | 21122017 | N701 | S504 |
| 28 | ΔmutS | tUNC19 | NPL | 21122017 | N702 | S504 |
| 29 | ΔmutS | tUNC19 | NPL | 21122017 | N703 | S504 |
| 30 | wt | tUNC19 | DML | 07052018 | N705 | S503 |
| 31 | wt | tUNC19 | DML | 07052018 | N705 | S504 |
| 32 | wt | tUNC19 | DML | 07052018 | N705 | S517 |
| 33 | wt | tUNC19, nicked | DML | 12072017<br>19072017 | N704 | S517 |
| 34 | wt | tUNC19, nicked | DML | 12072017<br>19072017 | N705 | S517 |
| 35 | wt | tUNC19, nicked | DML | 12072017<br>19072017 | N706 | S517 |
| 36 | ΔmutS | tUNC19 | DML | 07052018 | N706 | S503 |
| 37 | ΔmutS | tUNC19 | DML | 07052018 | N706 | S504 |
| 38 | ΔmutS | tUNC19 | DML | 07052018 | N706 | S517 |
| 39 | ΔmutS | tUNC19, nicked | DML | 12072017<br>19072017 | N704 | S504 |
| 40 | ΔmutS | tUNC19, nicked | DML | 12072017<br>19072017 | N705 | S504 |
| 41 | ΔmutS | tUNC19, nicked | DML | 12072017<br>19072017 | N706 | S504 |
| 42 | ΔmutL | tUNC19, nicked | DML | 12072017<br>19072017 | N704 | S503 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 43 | ΔmutL | tUNC19, nicked | DML | 12072017<br>19072017 | N705 | S503 |
| 44 | ΔmutL | tUNC19, nicked | DML | 12072017<br>19072017 | N706 | S503 |
| 45 | wt | tUNC19, nicked | FML | 26052017 | N701 | S503 |
| 46 | wt | tUNC19, nicked | FML | 26052017 | N702 | S503 |
| 47 | wt | tUNC19, nicked | FML | 26052017 | N703 | S503 |
| 48 | wt | tUNC19, nicked | FML | 26052017 | N704 | S503 |
| 49 | wt | tUNC19 | FML | 02112017 | N701 | S517 |
| 50 | wt | tUNC19 | FML | 02112017 | N702 | S517 |
| 51 | wt | tUNC19 | FML | 02112017 | N703 | S517 |
| 52 | ΔmutS | tUNC19 | FML | 02112017 | N704 | S517 |
| 53 | ΔmutS | tUNC19 | FML | 02112017 | N705 | S517 |
| 54 | ΔmutS | tUNC19 | FML | 02112017 | N706 | S517 |
| 55 | ΔmutS | tUNC19, nicked | FML | 26052017 | N705 | S503 |
| 56 | ΔmutS | tUNC19, nicked | FML | 26052017 | N706 | S503 |
| 57 | pFGC | tUNC19, nicked | FML | 26052017 | N701 | S502 |
| 58 | pFGC | tUNC19, nicked | FML | 26052017 | N702 | S502 |
| 59 | wt | mm19, 20bp barcode | SML | 09102017 | N702 | S502 |
| 60 | wt | mm19, 20bp barcode | SML | 09102017 | N704 | S504 |
| 61 | wt | mm19, 20bp barcode | SML | 09102017 | N705 | S502 |
| 62 | wt | mm19 | SML | 21122017 | N701 | S503 |
| 63 | wt | mm19 | SML | 21122017 | N702 | S503 |
| 64 | wt | mm19 | SML | 21122017 | N703 | S503 |
| 65 | wt | tUNC19, nicked | SML | 12072017<br>19072017 | N705 | S502 |
| 66 | wt | tUNC19, nicked | SML | 12072017<br>19072017 | N702 | S517 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 67 | wt | tUNC19, nicked | SML | 12072017<br>19072017 | N703 | S517 |
| 68 | wt | tUNC19 | SML | 21082017 | N703 | S517 |
| 69 | wt | tUNC19 | SML | 21082017 | N704 | S517 |
| 70 | wt | tUNC19 | SML | 21082017 | N705 | S517 |
| 71 | wt | tUNC19 | SML | 21082017 | N702 | S517 |
| 72 | ΔmutH | mm19 | SML | 03042018 | N705 | S502 |
| 73 | ΔmutH | mm19 | SML | 03042018 | N706 | S502 |
| 74 | ΔmutH | mm19 | SML | 03042018 | N701 | S502 |
| 75 | ΔmutH | tUNC19 | SML | 03042018 | N702 | S502 |
| 76 | ΔmutH | tUNC19 | SML | 03042018 | N703 | S502 |
| 77 | ΔmutH | tUNC19 | SML | 03042018 | N704 | S502 |
| 78 | ΔmutS | mm19 | SML | 09102017 | N706 | S517 |
| 79 | ΔmutS | mm19 | SML | 09102017 | N702 | S517 |
| 80 | ΔmutS | mm19 | SML | 09102017 | N705 | S517 |
| 81 | ΔmutS | tUNC19, nicked | SML | 12072017<br>19072017 | N701 | S504 |
| 82 | ΔmutS | tUNC19, nicked | SML | 12072017<br>19072017 | N702 | S504 |
| 83 | ΔmutS | tUNC19, nicked | SML | 12072017<br>19072017 | N703 | S504 |
| 84 | ΔmutS | tUNC19 | SML | 21082017 | N704 | S504 |
| 85 | ΔmutS | tUNC19 | SML | 21082017 | N705 | S504 |
| 86 | ΔmutL | tUNC19, nicked | SML | 12072017<br>19072017 | N706 | S502 |
| 87 | ΔmutL | tUNC19, nicked | SML | 12072017<br>19072017 | N702 | S503 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 88 | ΔmutL | tUNC19, nicked | SML | 12072017<br>19072017 | N703 | S503 |
| 89 | ΔmutL | tUNC19 | SML | 21082017 | N703 | S503 |
| 90 | ΔmutL | tUNC19 | SML | 21082017 | N704 | S503 |
| 91 | ΔmutL | tUNC19 | SML | 21082017 | N705 | S503 |
| 92 | ΔmutL | tUNC19 | SML | 21082017 | N703 | S504 |
| 93 | ΔuvrD | tUNC19 | SML | 14092017 | N705 | S503 |
| 94 | ΔuvrD | tUNC19 | SML | 14092017 | N702 | S503 |
| 95 | ΔuvrD | tUNC19 | SML | 14092017 | N703 | S503 |
| 96 | ΔuvrD | tUNC19, re-inoculated | SML | 14092017 | N704 | S503 |
| 97 | ΔuvrD | tUNC19, re-inoculated | SML | 14092017 | N705 | S517 |
| 98 | ΔuvrD | tUNC19, re-inoculated | SML | 14092017 | N706 | S517 |
| 99 | ΔuvrB | tUNC19 | SML | 25102017 | N704 | S503 |
| 100 | ΔuvrB | tUNC19 | SML | 25102017 | N705 | S503 |
| 101 | ΔuvrB | tUNC19 | SML | 25102017 | N706 | S503 |
| 102 | wt | tUNC19 | LP | 07102019 | N704 | S504 |
| 103 | wt | tUNC19 | LP | 07102019 | N704 | S504 |
| 104 | wt | tUNC19 | LM | 07102019 | N704 | S504 |
| 105 | wt | tUNC19 | LM | 07102019 | N704 | S504 |
| 106 | wt | tUNC19 | LD | 07102019 | N704 | S504 |
| 107 | wt | tUNC19 | LD | 07102019 | N704 | S504 |
| 108 | wt | tUNC19 | HP | 07102019 | N704 | S504 |
| 109 | wt | tUNC19 | HP | 07102019 | N704 | S504 |
| 110 | wt | tUNC19 | HM | 07102019 | N704 | S504 |
| 111 | wt | tUNC19 | HM | 07102019 | N704 | S504 |
| 112 | wt | tUNC19 | HD | 07102019 | N704 | S504 |
| 113 | wt | tUNC19 | HD | 07102019 | N704 | S504 |

# Chapter 2

# In vivo tracking of DNA strand choice bias

## 2.1 Abstract

For all discussions in Chapter 1, it was assumed that the mispaired DNA is repaired by making a random choice between the two constituent strands of the mismatch bearing DNA with conflicting information. That means a cell keeps about half of the time one of the two strands, whereas this strand is eliminated before any replication takes place other half of the time. This assumption was made mostly out of necessity than that it reflects the reality. Because *E. coli* cells are actually capable of differentiating between the two DNA strands that are products of semi-conservative DNA replication via their epigenetic marks. Here we will relax this assumption and independently account for the two types of repair products. For this, we made the "common strand" of the library distinguishable by means of a second set of barcodes in addition to the previously introduced tracing barcodes to obtain single-molecule level information.

Equipped with this improvement, we measured the repair efficiency of all 8 types of mismatches in all nearest and next-nearest sequence contexts to obtain the largest repair efficiency data set that was ever reported to my knowledge. Our results show that CC mismatches are always poorly repaired, whereas the local sequence context is a strong determinant of the highly heterogeneous repair efficiency of TT, AG and CT mismatches. We also repeated the nearest neighbor context measurements on MMR mutants and observed the absence of MutS to be the most detrimental to the observed repair efficiencies. The chapter concludes by a brief comparison of our data with the literature to explore the structural and biochemical reasons behind the differential repair response.

## 2.2 Introduction

In the previous chapter, I introduced our novel idea of making use of DNA barcodes to recuperate single-molecule level information out of a bulk bacterial culture. In this way, although the cells were grown without any physical restraints against mixing, the replication products of the same ancestor plasmid could be deduced. This approach allowed us to monitor the elimination of one of the two conflicting strands of a mismatch-bearing DNA and hence to deduce a measure of repair efficiency.

In this scheme, the workflow that I described uniquely labels the vector only, to generate a barcoded derivative of pUC19 that I loosely referred to as tUNC19. We inserted into this new barcoded vector a mismatch library by ligation, which we formed by annealing chemically synthesized oligos that we had procured. These two ssDNA oligos are reverse-complementary to each other, except at one position, hence generating upon annealing one and only one mismatch per molecule by design. This strategy provides a very simple means to obtain a mismatch library in a cost- and labor-efficient way: to scan mismatches in a 60 base long region, it suffices to procure 61 oligos that are of about that length.

To make the construction of these mismatch libraries feasible, we had kept the sequence of one of the strands constant (the so called common strand) whereas the position and type of the mismatch was varied by making modifications on the base sequence of the other strand (the variable strand). As all mismatches exactly shared the same common strand, it did not convey any information regarding the identity of the mismatch on the ancestor plasmid. This rendered some of the data systematically uninterpretable, because if the molecule was repaired based on the information coded by the common strand and the removal of the variable strand, all replication products of this plasmid will have the sequence of the common strand of the library and hence it is impossible to accurately attribute and record this information to the mismatch on the ancestor molecule.

Because of this shortcoming, we had to make use of the repair efficiency definition of Equation 1.12, where we first ignore the real counts of C-type clans and then compensate for this systematic ignorance by assuming that the number of V- and C-type clans should be about the same. This assumption was made mostly out of necessity more than it being an accurate reflection of the biological reality. The reports in the literature indicate on the contrary that an *E. coli* oftentimes differentiates between the two DNA strands [50]. This strand choice serves as a mechanism to keep the mutations at bay by the retention of the strand which is more likely to contain the correct information. For a brief time window, the products of the semi-conservative DNA replication will differ in their epigenetic marks. Since DNA polymerases synthesize the complementary chains using unmodified dNTPs in the cytoplasm, the nascent strand is completely devoid of such epigenetic

marks. On the contrary, the parental strand that has served as the template strand is not re-synthesized during this replication procedure and hence conserves its methylated state.

This epigenetic asymmetry forms the basis of the strand discrimination of *E. coli* MMR [6]. The repair response starts with the detection of the mismatch carrying DNA by MutS, which then recruits MutL to the site and this recognition complex recruits the DNA nicking enzyme MutH. The strand that is cut by MutH will be removed by the unwinding activity of the helicase UvrD and destined to degradation by exonucleases thus generating a single-stranded region a few hundred nucleotides in length. The resulting ssDNA gap is finally filled with DNA polymerase and the nicks are closed by DNA ligase, a procedure which copies the sequence information encoded on the unnicked strand. The MMR system has a tendency to preserve the information inherited from the template strand as MutH introduces a nick preferentially on the nascent strand. This strand discrimination capability of MutH is possible due to a transient asymmetry in the methylation pattern immediately after DNA replication, because the nascent strand is initially devoid of methylation, whereas the template strand typically carries an N6-methyladenosine within the GATC tetrameric sequence motifs. This asymmetry will eventually disappear as the methylation marks are established on both strands by DNA methylase Deoxyadenosinemethylase (Dam).

Another mechanism that can potentially introduce a strand choice is transcription coupled repair (reviewed in [51]), which is the enhancement of the repair efficiency of the transcribed strand of dsDNA in comparison to the coding strand. If during transcription, a DNA lesion is encountered, RNA polymerase stalls in the vicinity and repair associated proteins are recruited to the lesion site, such as UvrA, UvrB and Mfd, while Mfd has recently been reported to be non-essential [52]. The initiation of the repair procedure releases the stalled RNA polymerase and a single-stranded gap is generated by double incision on the template strand by UvrC and unwinding activity by UvrD. As a final step, this gap is filled with DNA polymerase and sealed by DNA ligase. While transcription coupled repair is primarily implicated for nucleotide excision pathway mainly targeting UV-damage adducts such as cyclobutane pyrimidine dimers, a potential cross-talk between the DNA repair pathways might be at play as MMR elements MutS and MutL have been reported to be required for the transcription coupled repair [53].

These phenomena imply that in real biological systems the random strand choice assumption that we introduced in Chapter 1 is not on solid grounds. In Chapter 1, we abolished the asymmetric methylation driven strand preference by constructing vectors out of PCR products that are not methylated by design on either strand or are methylated on both strands *in vitro* in a symmetric manner. Here we will relax this assumption altogether instead and independently account for the two types of repair products. For this, we made the "common strand" of different mismatch library members distinguishable from each other by means of a second set of barcodes. This is in addition to the previously introduced "tracing barcodes" that keep track of the replication products. The sequence of this second barcode is physically linked to the mismatch library, enabling a systematic analysis of the sequence effect on mismatch repair. As a drawback, this approach renders the experimental workflow much more involved, as it is not possible to form mismatches by annealing oligo libraries containing highly diverse barcode regions, but rather copying the barcode information across the strands is necessary.

## 2.3 Results

### 2.3.1 Overview of the experimental design

Similar to the methodology adopted in Chapter 1, we generated a barcoded vector library by a PCR-based scheme where one of the primers carries 25 random nucleotides (Figure 2.1). Into this barcoded vector library, we ligated a mismatch library of short (∼100bp) dsDNA fragments, which contain one and only one mismatch per molecule by design. This ligation product was transformed into *E. coli*, in which the plasmid is allowed to replicate overnight multiple rounds, and we monitored if a repair occurred before the arrival of the first replication fork. We achieved this in a high-throughput manner by generating an amplicon library out of the extracted plasmid mixture and clustering the reads obtained from a next-generation sequencer with respect to the barcode sequences. Due to the high diversity of the barcodes, it is virtually impossible for two plasmids in these datasets to carry the same barcode sequence, unless they are replication products of the same original molecule, which both of them inherited the barcode from. This implies that

each of the clusters we obtained are the replication products of the same single molecule of ancestral plasmid. We then check each cluster to see if they contain contributions from both strands of the mismatched DNA, indication of an unrepaired molecule. Clusters containing exclusively variable or common strand sequence correspond to repaired molecules.

In this chapter, we will follow the same design principle, and the same experimental methodologies, except that the mismatch library that is investigated as part of this barcoded plasmid library contains a short second barcode of itself. Rather than directly annealing commercially synthesized ssDNA oligos, we made use of oligo pools that can contain as many as 100 000 different elements. Each oligo pool member codes for one different version of the variable strand and carries a short barcode of known sequence, and a known relationship exists between the sequences of the barcode and the associated variable strand. We copied this barcode information to the common strand by primer extension after partial annealing of the common and variable strands. The presence of this second barcode enables inference of original mismatch type and position that gave rise to a plasmid that was repaired based on the common strand sequence. Below, I will start by describing this new approach we used to generate the mismatch library and the associated data analysis workflow. For other parts of the workflow, please refer to Chapter 1.

### 2.3.2 Double barcoding strategy for single molecule tracking of mismatch repair

Given that DNA has 4 types of nucleotides (A,C,G,T), and that each of them have 1 proper Watson-Crick complement, 8 types of mismatches exist (AA,CC,GG,TT,AC,AG,CT,GT). We aimed to design an experiment that samples each of these mismatches in a balanced way such that each mismatch is sampled in each sequence context at least once. To ensure this, we sought to obtain an optimal DNA sequence that represents all sequence k-mers to serve as the consensus sequence of all the variable strand elements. We achieved this by generating a De Bruijn sequence on alphabet $\Sigma = \{A, C, G, T\}$ following a similar reasoning as in Section 1.3.5. A de Bruijn sequence is the theoretically shortest possible sequence that contains every sequence k-mers and can be obtained via Algorithm 1.

For experimental simplicity, we started by generating a mismatch library that contains all mismatches within all possible nearest-neighbor contexts, which can be achieved by computing the De Bruijn sequences containing all trimers (k=3). The shortest such sequence is 64 bases long but is cyclic, which can be linearized into a 66 base long sequence. We named one such sequence 3ML1 and used as the consensus sequence of our library (Figure 2.9a), i.e. the ensemble average of all variable strands would be this 3ML1 sequence. To span all possible mismatches, we obtained a DNA library that contains individual DNA sequences each of which deviate from this 3ML1 sequence by one nucleotide only, and hence serving as the variable strands of the mismatch library. The quantity of ssDNA obtainable from a contemporary oligo pool is limited to fmol ranges. To scale up the amount of variable strands, we amplified the purchased ssDNA library using a modified primer pair where the strand to be eliminated has a 5' phosphate group making them prone to digestion by $\lambda$ exonuclease, whereas the strand to be retained is protected by five consecutive phosphothioate bonds and a 5' terminal biotin (Figure 2.1) [54–56]. The treatment of the PCR product hence yields an ssDNA library that contains the same sequences as the oligo pool that serves as the starting material. We separately procured a chemically synthesized sequence that is reverse-complementary of the consensus sequence as before to serve as the common strand of the mismatch library.

By annealing the common and variable strands, we can assemble the desired mismatch library that contains all possible mismatches along the 3ML1 sequence. However, MMR can take place by keeping either of the two strands and the retention of the common strand would lead to a loss of information regarding the mismatch type and the position on the ancestor plasmid, because being the reverse complementary of the 3ML1 consensus strand, common strand is identical among all mismatch library members. Such systematic loss of information would make it impossible to correctly account for about half of the repaired clans. To overcome this challenge, we introduced a second set of barcodes 6 to 8 base pairs in length to the 5' end of the variable strands that are uniquely linked to the identity of the mismatch and its position, which I will be referring to as "mapping barcodes" from now on. To avoid introducing a systematic bias, each mismatch is redundantly represented by multiple different mapping barcodes, but each mapping barcode

is used only once and hence represents a mismatch unambiguously. To be more precise, the relationship between the mapping barcode sequences and variable strand sequences is surjective but non-injective.

The presence of barcodes on the elements of the variable strand library pool poses an experimental complication, because it is not possible to thermally anneal this mixture with the common strand directly. As each variable strand is different at the mapping barcode site, a forced annealing with the common strand of single kind would lead to formation of bubbles in the heteroduplex. The bias that could be imparted on our measurements by such bubbles is uncertain: while multi-base insertion/deletion loops are known to be poor substrates of MMR [57], the exact length of the bubbles and the response of the MMR against them is not well-characterized. We therefore circumvented this ambiguity by annealing the common and variable strands to form a partial duplex only, so that the mapping barcode segment is left unpaired. We then used oligo extension via Klenow fragment DNA polymerase to complete the heteroduplex into dsDNA and copying the mapping barcode onto the common strand in the process.

We had introduced two non self-complementary restriction sites (SacI and XhoI) at the two termini of the barcoded vector library via the 5' tails of the primers in addition to the barcodes. We also included these same restriction sites by the design of the mismatch library sequences such that after generating sticky ends with a double restriction digestion reaction, we can ligate this mismatch library with the barcoded vector library. For this ligation step, we opted for T7 ligase for its higher specificity for proper sticky-end ligation in comparison to other typical ligases such as T4 [40]. We transformed this plasmid library into electrocompetent K-12 *E. coli* (BW25113), incubated the transformant bacteria overnight and sequenced the extracted plasmids using Illumina MiSeq or HiSeq platforms. We used a clustering approach on the tracing barcodes obtained as part of this NGS output to deduce the relative frequency of repair events, which I will describe next.

### 2.3.3 Repair efficiency can be deduced by classifying clans

To organize the obtained reads into groups that descended from the same ancestor plasmid, we performed density-based clustering (DBSCAN) on all obtained reads based on their tracing barcodes [32]. Because our choice of 25bp long tracing barcodes correspond to $4^{25} \approx 10^{15}$ different possibilities, a group of plasmids sharing the same or highly similar barcodes are much more likely to be descendants of the same ancestor plasmid than sharing the random barcode by chance (Section 1.3.2 and Figure 1.2). Sequence composition of such a clan therefore tells us whether repair has already taken place or not by the time the first replication occurs.

As was discussed in Section 1.3.7, to decide if an individual clan derives from a repaired or unrepaired ancestor plasmid, we determined the "substitution prevalence" (s) which is the fraction of reads in a particular clan that carries the sequence of the variable strand, viz.

$$s \equiv \frac{\#VariableStrands}{\#VariableStrands + \#CommonStrands} \tag{2.1}$$

An unrepaired clan has about equal numbers of reads representing the variable and common strands of the library ($\#VariableStrands = \#CommonStrands$ and s=0.5), because multiple rounds of semi-conservative replication produces daughter plasmids independently using both strands as the template. A repair event preceding the first replication will enforce the conformity of the two strands to the Watson-Crick base pairing rule, and hence the strand that was removed during repair will never be used as template for replication. The removal of common strands will lead to clans in which 100% of the reads will represent the variable strand ($\#CommonStrands = 0$ and s=1), whereas on the other extreme, the removal of the variable strand will lead to clans exclusively consisting of common strands ($\#VariableStrands = 0$ and s=0).

Following the same nomenclature as in Chapter 1, I will refer to these three types of clans as U, V and C-type, with the exception that C-types are this time distinguishable via the mapping barcodes (Figure 2.2a). Due to an interplay of experimental errors and finite sampling of a binary pool, we expect the three clan types to form relatively broad peaks rather than sharp spikes. In Figure 2.2b, the expected trimodal distribution was observed from the histogram of substitution prevalence values of all mismatches. Peculiarly, however, the peak corresponding to the V-type

clans, i.e. those that were repaired by retaining the variable strand, was not centered at s=1 as we will discuss further below (red curve). Performing the same experiment in $\Delta mutS$ cells, which are deficient in MMR response, caused a large increase the U type clans, i.e. those with intermediate s values (blue curve in Figure 2.2). As a further control, performing the experiments in wild type cells using fully base paired plasmids of same sequence as the common strand (3CL1) showed a histogram peaked at s=0 (black curve in Figure 2.2). Overall, these analyses show that most of the repair or no-repair events we score indeed predominantly result from the MMR response of the cell against the intentionally introduced mismatches.

### 2.3.4 The mismatch on the ancestor plasmid can be deduced using the mapping barcodes

The approach described so far groups the data in clans and can subject them to ternary classification based on if and how the repair has happened for each single plasmid (C, U, V clans). Using the concept of mapping barcodes that I introduced before, we can also deduce the type and the position of the mismatch on the ancestor plasmid that gave rise to this clan easily. To do this, we process the dataset clan by clan and for each read constituting the clan, we extract the segment immediately following the SacI restriction site (GAGCTC) as the location of the mapping barcodes is well-determined by design. We then computed an ensemble-averaged mapping barcode of each clan by calculating the base frequency histogram and a majority voting scheme. Using the sequence of this average barcode, we refer to the pre-determined look-up table that stores the relationship between the mapping barcode sequence and the linked variable strand sequence. *A posteriori* knowledge of the variable strand sequence on the ancestor plasmid is equivalent to knowing the mismatch type and position because the other strand of the plasmid is of known shared sequence among all plasmids in the pool and the violation of the base pairing rule can be found by checking these two sequences one base at a time in linear time.

By design, the mapping barcodes assigned to each variable strand is separated from each other by at least 2 mutations. This means any experimental error that causes a mutation in the mapping barcode will convert the legitimate barcode to a barcode that is not part of the lookup table and

hence flagging the clan for deletion. Confusion of the mismatch identity is possible if there are 2 or more mutations in the barcode sequence, which is less likely to happen. By reading these mapping barcodes, we indeed could deduce the type and sequence of the mismatch present in the original single plasmid with a very high accuracy, unbiased by whether a repair has taken place or which strand was retained. To reach this conclusion, we built a confusion matrix comparing the type or position of the mismatch that we deduced by the help of the mapping barcodes (along the y-axis) with the mismatch type or position we deduced by building the base frequency histogram of the clan and checking where the most deviation from the consensus sequence of the library occurs (along the x-axis) as was done in Chapter 1. For the latter quantity, we only included the U and V clans, as it is not possible to accurately deduce the mismatch identity without mapping barcodes with this strategy. If the strategy I proposed above works perfectly, the mismatch identity that we obtain with the two approaches should agree and the confusion matrices should be diagonal (i.e. an identity matrix), or we should get a randomized matrix of roughly equal matrix elements if the relationship does not hold at all due to high experimental noise. The matrices in Figure 2.3, which closely follow the former case, suggest that it is possible to deduce the mismatch type very accurately out of a diverse mixture both using MMR-capable and MMR-deficient cells.

### 2.3.5 Position dependent s-histograms

The mapping barcodes enable an accurate inference of the mismatch identity on the ancestor plasmid that leads to a particular clan, unbiased by the repair outcome. Using this approach, we can obtain the s-histograms of individual mismatches, and as Figure 2.16 exemplifies, the histograms we obtain for each mismatch qualitatively resemble the tri-modal distribution we expected (Figure 2.2a). However, the peak corresponding to V-type clans shifted monotonically from around s=1 to s=0.6 as the position of the mismatch moved further away from the tracing barcode (barcode-proximal to barcode-distal). In contrast, the C-type clan peak stayed precisely at s=0 for all mismatches we examined.

This could potentially be an artifact of the particular sequence we chose for this set of experiments (3ML1), as some DNA translocating proteins involved in DNA repair have been reported to

sharply switch behavior at certain DNA sequence motifs, as exemplified by the switching behavior of RecBCD at Chi sites [58]. But our results show that this gradual shift in the V-peak position is evident for all DNA libraries we have experimented with, albeit the steepness of the transition as well as the position of the inflection vary in a manner that depends on the consensus sequence of the library. Figure 2.7a shows the center position (p in Equation 2.2) of the deduced V-peaks 5MLx libraries, where the color scheme represents the overall GC content of the sublibrary and higher AT content often lead to more significant changes in the V-peak position. That a similar trend was observed in Chapter 1 using oligo annealing based mismatch libraries also suggests that this is not an artifact caused by the PCR-based mismatch library production scheme we developed.

To visualize these position dependent global trends in a more systematic way, we represented the occupancy level of each such histogram in gray value and represented the s-histogram of each consecutive position as a stack of rows in Figure 2.5. At each base position from the barcode-proximal to the barcode-distal terminus of the library, it is possible to form 3 different mismatches, which can have very different s-histograms. For simplicity, each row represents the average of the three s-histograms per position, and darker colors show a higher occupancy level of the indicated state. Figures 2.5a and 2.5b show the compact representation of s-histograms in wt and $\Delta mutS$ cells, respectively. In both cases, we observe a sharp peak at s=0, representing C-type clans, while the intermediate s-values are much more populated in $\Delta mutS$ cells, an indication of a higher abundance of U-type clans and lower repair capacity. In both cell strains, but much more obviously in wt cells, we observe a third population around high s-values, representing the V-type clans.

As I will describe the methodological details in Section 3.4.2, we then classified each clan in a position dependent manner, and calculated the relative abundance of the three subpopulations (Figures 2.5c and 2.5d). In wt cells, the U-type clans constituted about 10% of the entire population, corresponding to a global repair efficiency of 100%-10%=90%, whereas this went down only to 50% in $\Delta mutS$ cells. Surprisingly, the center of the V-type clan peak at s=1 (red colored) gradually and monotonically shifted towards more intermediate s values. Compared to the C peak, the V peak was also broader, accompanied by a side-peak at around s=1 for all mismatches.

We do not have a mechanistic explanation of this effect, but various control experiments suggest

that this behavior is rather related to the directionality of the barcoded vector. First and foremost, we observed a qualitatively similar shift in ΔmutS cells: while the V-peak is not as distinct, the observed upper limit of the s-values change in a similar way as in wt cells. The relatively smaller V-type peak exactly at s=1 also gets less and less populated towards the barcode distal side, suggesting this systematic effect is not fully caused by the MMR, but rather is a result of other responses that preferentially keep the variable strand sequence.

As a second observation, we considered that this gradual trend might perhaps be related to the design of the consensus sequence we have chosen, as the GC content of the barcode-distal half of the library was much higher than the barcode-proximal half that can potentially affect the results (33% vs 64%). To check if such a bias can result from the library design, we flipped the insertion orientation of the 3ML1 mismatch library on the barcoded vector by switching the SacI and XhoI restriction sites on the original barcoded vector (via PCR primers invP1 and invP2 to obtain cNUT91 out of pUC19). This scheme brings the mismatches that were previously farther away from the tracing barcode to its close proximity, hence should result in a change in the directionality of the trends, if they are related to the mismatch library itself. But on the contrary, we observed the V-peak position to gradually drift to lower values towards the barcode-distal end, similar in direction as before (Figure 2.6). As another possibility, this could be an artifact due to the directionality of our library synthesis procedure that involves 3'→5' oligo array synthesis, multiple rounds of PCR involving 5'→3' DNA elongation and strand-selective 5'→3' digestion. However, this experiment and that the trend observed using oligo annealing based mismatch libraries produce similar trends argue also against this hypothesis.

And finally, we noticed that, despite using a paired-end sequencing approach, our NGS scheme had an inherent directionality that could have introduced a directional bias. That is because the forward reads always start from the barcode-proximal end of the amplicon and proceed towards the barcode-distal end, and the reverse reads occur from the barcode-distal end towards the barcode-proximal end of the amplicon. By the manufacturer's design, the forward reads always precede the reverse reads and the overall sequence chemistry leads to a slightly different error profile for the two reads [59]. We hence considered changing the sequencing order by inverting the sequencing

adapters on the final amplicons to be sequenced by switching the overhangs on the PCR primers that were used as part of our library preparation protocol (Primers revS1 and revS2_4). This re-arrangement makes the temporally first reads to proceed from the barcode-distal towards the barcode-proximal end of the library and the following read to start at the barcode-proximal end, instead. Comparison of Figures 2.6c and 2.6d suggest that a similar trend exists in both sequencing scenarios and hence the position dependent trend cannot be accounted for by a preferred orientation of our sequencing protocol or insertion orientation of the mismatch library.

### 2.3.6 Computation of the repair efficiency using clan counts

As in Chapter 1, we monitored the DNA repair response in competition with the replication of the plasmid. Or more rigorously, we defined the repair efficiency of a mismatch as the fraction of molecules that are repaired within the time window between the introduction of the plasmid into the cell and the arrival of the first replication fork to the mismatch, the latter of which physically eliminates the mismatch. We defined the repair efficiency ($\eta$) of a particular mismatch as the fraction of clans that were repaired, i.e. C- or V-type clans rather than U.

Above, I showed example s-histograms in which the characteristics of the C, U, V sub-populations were observed to be position dependent. Most notably, both the spread and the center of the V-peak gradually drifted from its expected position s=1 towards lower values as the mismatch position approached the barcode-distal end. To avoid any systematic bias introduced by this phenomenon, we implemented a position-dependent boundary between U and V populations, obtained by a fixed offset from the V-type clan peak position by applying a moving threshold to decide between U and V outcome cases. To find the high-cutoff $c_{high}$, we used curve fitting on the substitution histogram of each individual mismatch using the below function and a built-in Levenberg-Macquardt minimizer using the function prototype:

$$f(s) = k\delta(s) + le^{\frac{-(s-p)^2}{q}} \tag{2.2}$$

where k, l, p, q are the four degrees of freedom that vary to optimize the fit of the model function to the observed histogram. We performed an interpolation by fitting all $p_i$ with a 6th

degree polynomial to find an interpolation for the peak position along s, $\widetilde{p_i}$, that depends on substitution position i, but not on the substituted base type j. *Ad hoc*, we considered all clans with a substitution frequency above $c_{high} = \widetilde{p_i} - 0.15$ as repaired by keeping the variable strand of the library. For the C vs. U decision problem, we implemented a system-wide *ad hoc* fixed threshold of $c_{low}(i) = 0.1$, giving rise to:

$$assertion : \begin{cases} C, & 0 \leq s \leq c_{low} \\ U, & c_{low} < s < c_{high} \\ V, & c_{high} \leq s \leq 1 \end{cases} \tag{2.3}$$

Finally, for a mismatch at position i and formed by the substituted base type j, both the repair efficiency ($\eta_{ij}$) and strand choice bias ($\beta_{ij}$) can be computed, using the number of unrepaired clans ($U_{ij}$), number of repaired clans using the common ($C_{ij}$) or the variable strand ($V_{ij}$) as the correct information source by the cell.

$$\eta_{ij} = \frac{C_{ij} + V_{ij}}{C_{ij} + V_{ij} + U_{ij}} \tag{2.4}$$

$$\beta_{ij} = \frac{C_{ij} - V_{ij}}{C_{ij} + V_{ij}} \tag{2.5}$$

and the total number of data points used for the calculation of each mismatch necessitates $T_{ij} = C_{ij} + V_{ij} + U_{ij}$ to be as high as possible for precision (Section 2.5.6).

### Measured repair efficiency values

Using the strategy outlined above, we individually measured the repair efficiency of 192 mismatches as part of 3ML1 (Figure 2.9a). As a global characteristic shared among the most of the data sets, I note that a wild type (wt) cell is capable of repairing the majority of mismatches with a high efficiency ($\eta = 0.89 \pm 0.15 = \mu \pm \sigma$, Figure 2.8a) and the apparent repair efficiency is significantly lower in ΔmutS cells ($\eta = 0.43 \pm 0.06$, n=198, $p < 10^{-5}$ by one tailed z-test, Figure 2.8b). Because MMR should not occur in ΔmutS cells, we expect $\eta = 0$ to uniformly hold, and all

observed apparent repair is attributable to side pathways that result in loss of one plasmid strand, experimental artifacts and sampling errors. Therefore, unless otherwise indicated, we will report the scaled repaired efficiency ($\eta_s$) obtained by re-scaling the measured range, from $\eta_{min} = 0.37$ to $\eta_{max} = 1$, to the scale range from 0 to 1, i.e.

$$\eta_s \equiv \frac{\eta - \eta_{min}}{\eta_{max} - \eta_{min}} \tag{2.6}$$

The repair efficiency quantified as $\eta_s$ was highly reproducible across independent experimental replicates for the libraries we studied (Figures 2.8e, 2.8f, 2.8g).

Figure 2.10 tabulates, as a heat map, $\eta_s$ of individual mismatches as part of different consensus sequences tested in wt cells. In each plot, the base along the x-axis indicates the sequence of the common strand while the four rows show the base the variable strand contains at that position forming a mismatch, i.e. the rows from top to bottom represent the cases where the variable strand contains an A, T, C, or G, respectively. As an example, the red colored pixel on the 3ML1 output matrix (topmost) at the 5th column and G-row refers to a low repair efficiency for an AG mismatch that is formed by an A on the common strand and a G on the variable strand and is located 5 nucleotides after the last nucleotide of the mapping barcode. While four different bases can be incorporated at each position, only 3 out of 4 will lead to a mismatch, and the remaining fourth entry corresponding to a non-mismatch case is marked by gray hatching as it is not a repair machinery substrate. While the majority of the mismatches were repaired with a very high efficiency ($\eta_s \approx 1$, indicated by white and yellow tones), some mismatches were repaired with a noticeably lower efficiency ($\eta_s \approx 0.2$, crimson tones). The diversity of the pixel colors suggest that, even when the C-type clans are taken into account via this double barcoding approach, the repair efficiency of mismatches vary significantly but reproducibly.

### 2.3.7 AG, CT, TT mismatches are difficult to repair in a context dependent manner

Next, we expanded our dataset to sample a larger sequence subspace by including the next-nearest neighbor nucleotides surrounding the mismatches. As the shortest sequence containing all sequence pentamers is 1028 bases in length (Figure 2.9b), too long for oligonucleotide synthesis as a single piece, we divided this sequence into 13 sub-sequences (5ML1, 5ML2, ..., 5ML13) and performed measurements in 13 independent sets of experiments. Collectively referred to as 5MLx (x=1,2,...,13), they together represent each mismatch type in all sequence contexts out to the next-nearest neighbors. While the exact precision is dependent on both the number of reads per clan and the total number of clans for the same mismatch, $\eta_s$ values were reproducible between experimental replicates. The reproducibility in 3ML1 library was much higher than in 5MLx libraries likely because of the much higher sequencing depth (111 vs 27 median reads per clan, respectively; Figure 2.8 and Figure 2.22). While the 3ML1 and 5MLx libraries, as well as individual sub-libraries comprising 5MLx significantly differed in number of clans attributable to each individual mismatch, we observed the correlation between this quantity and scaled repair efficiency to be weak both at the library level (Figure 2.22c) and at the single mismatch level (Figure 2.22f). In particular, 3ML1 had an average clan count more than an order of magnitude higher than 5MLx except 5ML8 (4294 vs 425 clans/mismatch), while the average scaled repair efficiency only varied by less than 1% (0.832 vs 0.838). For comparison, the root mean square deviation between the two 3ML1 replicate $\eta_s$ measurements in wt cells was 0.018. We hence consider the effect of non-uniform sampling in our assay to be present, but limited in extent.

Our large data set containing more than 10 000 repair efficiency measurements showed clearly that some mispaired bases are repaired much better than others. If all mismatches were repaired with about the same repair efficiency confounded by some sampling error, one would expect all entries in Figure 2.10 to be randomly distributed both to the columns and rows of the matrices. Yet, an opposite qualitative judgment can be formed by observing that the number of dark-red pixels, indicative of a low repair efficiency, are highly represented in the C row, which indicates that poorly repaired mismatches often contain a C on the variable strand. In addition, our experimental

design attempts to compress the library sequences, which tends to cluster similar nucleotide motifs together at certain zones along the prototype sequence as opposed to a random distribution of 4 constituent bases. The tendency of the pixels with low repair efficiency to accumulate more densely at certain zones along the common strand (x-axis) again suggests that for those mismatches showing less than near-complete efficiency, the sequence context might be a determinant of repair efficiency. We therefore asked if there are any generalizable properties that makes a mismatch more likely to evade repair.

We first checked the effect of the base identity of the nucleotides that are mispaired on the observed repair efficiency. Accounting for the symmetry between the two DNA strands, a mismatch can form between AA, AC, AG, CC, CT, GG, GT and TT nucleotides. Disregarding the sequence context of these 8 mismatch types, we observed that essentially all CC mismatches were repaired with a low efficiency ($\mu \pm \sigma = 0.35 \pm 0.13$), whereas virtually all AA, AC, GG, and GT mismatches were efficiently repaired irrespective of the sequence context ($0.92 \pm 0.07, 0.93 \pm 0.04, 0.94 \pm 0.04, 0.93 \pm 0.04$, respectively (Figure 2.11a). On the other hand, we observed a wider range of repair efficiency for AG, CT and TT mismatches ($0.81 \pm 0.17, 0.82 \pm 0.19, 0.75 \pm 0.22$, respectively), suggesting that the sequence context is an important determinant of how efficiently these three mismatch types can be repaired.

### 2.3.8 Nearest-neighbors cannot fully describe the repair efficiency variations

Figures 2.11d - 2.11f show three sequence contexts that are identical up to the nearest neighbors, as the GT (unboxed), CT (pink boxes) and TT mismatches (green boxes) have a T on both the 5' and 3' sides in all three cases. Still, the repair efficiencies of the mismatches differed significantly from each other, suggesting that the nearest-neighbors are not sufficient to explain the observed variability, but rather that the repair efficiency is influenced by the neighbors that are 2 nucleotides or more away from the position of the mismatch.

Among the three highly context dependent mismatches, CT and TT are mispairing between two pyrimidines (YY), whereas AG is between two purines (RR). To test whether the identity

of the mispaired nucleotides influences the sequence context dependence of repair efficiency, we checked in the 5MLx library for each occurrence of these three mismatch types and separately considered the base identity of the nearest neighbor bases and whether the next-nearest neighbors are purine (R) or pyrimidine (Y) (Figure 2.12e-2.12d). In this scheme, each distinct occurrence of the mismatch in the 5MLx is represented by a bead whose color indicates $\eta_s$. The particular cell of the table that each colored bead is positioned in reflects the identity of the bases neighboring the mismatch. The 10 columns along the x-axis indicate the base identity of the 5' and 3' immediate neighbors that are labeled as P and Q, respectively. In a similar manner, the placement along the y-axis is according to whether the second nearest bases are occupied by an R or Y.

For AG mismatches, the lowest repair activity occurs if the mismatch is flanked by CC or GA as the 5' and 3' immediate neighbors of A, that is occupying positions labeled as P and Q, respectively. Yet, the repair still happens efficiently if the 5' and 3' next-nearest neighbors are Y and R, respectively. In contrast, placement of the mismatch between a 5' C and 3' A or G, alternatively a T before and an A after the mismatch led to more difficult to repair outliers of CT mismatches, and the presence of a Y at the 3' next-nearest neighbor position exacerbated this deficiency. Investigated with a similar approach, TT mismatches were poorly repaired more often if the immediate neighbors are CG or TA, but the next nearest neighbors had a less pronounced effect. Hence, we concluded that there is no unifying 5' - 3' nearest neighboring base combination that generates all difficult to repair mismatch outliers, but rather that there are certain pentamer sequence contexts specific to a mismatch type that influences their repair efficiency.

### 2.3.9 Sequence context effect extends beyond next-nearest neighbors

Based on the results I presented so far, we had previously concluded that the repair efficiency is heavily dependent on the mismatch type: while CC mismatches are poorly repaired, the efficiency of the response against AG, CT and TT mismatches was sequence context dependent (Figures 2.11a). We also had found examples of mismatches such that changing the next-nearest neighbors lead to a change in the observed repair efficiency, even though the mismatch type and the nearest neighbors were the same (Figures 2.11d - 2.11f). This suggested that the next-nearest neighbors

have a measurable influence on the repair efficiency, and this led us to ask to what extent our sequence motif descriptors containing up to the 2nd nearest neighbor nucleotides capture the variation in the repair efficiency.

To assess this, we designed a mismatch library that measures DNA mismatches within different heptamer contexts. If a description of a mismatch with its 2nd degree neighbors is an adequate model, then changing the base composition of the 3rd or higher order neighboring nucleotides should not lead to an appreciable change in $\eta_s$. In contrast to the 3ML1 and 5MLx libraries representing each and every mismatch in all trimer and pentamer contexts, it is experimentally infeasible to sample all sequence heptamers, because the shortest sequence containing all heptamers (i.e. the De Bruijn sequence for k=7) is impractically long ($4^7 + 7 - 1 = 16390$ bp). Rather than sampling the full sequence space, we instead opted to subsample the sequence subspace that represents the three arbitrarily-chosen pentamer motifs that previously showed variability in mismatch repair efficiency (CAAAA, AAAAT and GAAAT). We therefore sampled the repair efficiency of these three pentamers while varying the immediately surrounding bases, or equivalently, our desired subset contains all heptamers that contain either of these three pentamers at their core.

While algorithmically finding the theoretically shortest DNA sequence is possible by brute force search or dynamic programming, finding the exact solution of this problem is highly resource demanding as it is equivalent to solving a traveling salesman problem that is of complexity class NP-hard [60]. We instead resorted to a greedy algorithm, which at each iteration of the loop, connects the nodes that are separated by the shortest edge and iterates until no nodes are left that are unconnected. The routine quits by reporting a compressed concatenated sequence by traveling through the path formed by the joined edges, which has been shortened in comparison to the naively concatenated sequence, but is not necessarily the theoretically shortest sequence.

For this purpose, we made use of a custom metric as follows: if $x_i$ and $x_j$ are two sequence k-mers, we define the distance between them by,

$$d(x_i, x_j) = ||x_i x_j|| - ||x_i|| \tag{2.7}$$

where the first term represents the length of the concatenated sequence and the latter term is the

length of the sequence before the current iteration. The "concatenation" we perform here accounts for the possible compression by making use of the overlaps between the 3' end of $x_i$ and 5' end of $x_j$, hence $0 \leq d(x_i, x_j) \leq ||x_j||$. We then aim to (approximately) minimize for the total distance traversed by optimizing for the sequence in which the nodes should be added, i.e.

$$\underset{\{a_k\}}{argmin} \sum_k d(x_{a_k}, x_{a_{k+1}}) \tag{2.8}$$

As an example, Figure 2.13a shows a directed graph with four nodes. The numbers along the edges describe the distance between the nodes, if the nodes were included in the direction indicated by the arrow: the node pointed by the arrow head is to follow the node indicated by the tail of the arrow. Following the definition above, d(AAAAA,AAAAT) = 1, as the hexamer AAAAAT contains both of the pentamers in the prescribed order by an increase in length by only 6-5=1. However, d(AAAAT,AAAAA) = 5, as the decamer AAAATAAAAA is the shortest possible sequence that contains both pentamers in the given order. A potential pseudo-optimal path would join the nodes AAAAA→AAAAT at the first iteration and obtain AAAAAT, then CAAAA→AAAAAT and get CAAAAAT, then TAAAC→CAAAAAT to report the final sequence TAAACAAAAAT that contains all four pentamers.

In our specific case, as there will be 16 heptamers per each pentamer core investigated, the length of the sequence obtained by naive concatenation of all the required heptamers would be $16 \cdot 7 \cdot 3 = 336$ bases long, whereas the pseudo-optimized sequence with the above approach above is 244 base long, hence providing 27% compression (Figure 2.13b). Due to the limit on the maximum length of solid-phase synthesized oligos, we still had to split this sequence into 4 sub-libraries (SSL1, SSL2, SSL3, SSL4) similar to the 13 sublibraries constituting the 5MLx library and measured the repair efficiency of all mismatches possible at any position along these four libraries independently.

As before, we observed the apparent raw repair efficiency of wt cells to be significantly higher than $\Delta mutS$ cells ($\mu \pm \sigma = 0.81 \pm 0.12$ vs $0.62 \pm 0.10$, $p < 10^{-10}$ with one tailed z-test). The scaled repair efficiencies measured for the libraries are shown in Figure 2.14a, displaying a high $\eta_s$ for most mismatches, apart from certain outliers, hence producing matrices dominated by bright tones with the usual crimson patterns indicating inefficiently repaired mismatches similar to the

previous outputs. Also in agreement with our past observations, we observed AA, AC, GG, GT to be efficiently repaired and CC mismatches to be poorly repaired in general, whereas the repair efficiency of AG, CT and TT mismatches showed the strongest context dependence, a trend common in all libraries we studied (Figures 2.11a, 2.11c and 2.11b). However, we observed $\eta_s$ to be globally lower for SSLx than 5MLx for all mismatch types ($\mu \pm SE = 0.69 \pm 0.01$ vs $0.83 \pm 0.00$, $p < 10^{-10}$ by one tailed z-test). While SSLx has a significantly higher overall AT content (78% vs 51%) that could have rendered the local melting temperature lower, we found only a weak correlation between $\eta_s$ and the AT content of the surrounding nonamer sequence context in the two libraries, but a more pronounced positive correlation between the sublibrary-wide average repair efficiency and the overall GC content of the sublibrary (Figures 2.21a and 2.21b).

A 4-sublibrary overlay of the pairwise comparison of the three experimental replicates in wt cells suggests that the measured $\eta_s$ are reproducible (Figure 2.14b). The correlation between experimental replicates was weaker for SSLx than 3ML1, which we again attribute to the higher sequencing depth of the latter (median number of reads per clan= 27 vs 111; Figures 2.22a and 2.14c). We then asked if neighbors that are more than 3 nucleotides away from the position of the mismatch still have a detectable influence on the repair efficiency. If heptamer motifs are sufficient descriptors of the repair efficiency, the base composition of the nucleotides before and after this motif should have a negligible influence on the observed repair efficiency and $\eta_s$ should be similar even if these neighbors are changed. A qualitative analysis of the heptamer motifs in different higher order sequence contexts suggests that $\eta_s$ depends on these distant neighbors. In particular, the common-sequence heptamer CAAAAAT is represented within 6 different sequence contexts in our libraries and we observe $\eta_s$ to vary in a sequence context dependent manner (Figure 2.14d). If the bases immediately preceding CAAAAAT at the barcode proximal and distal sides are AT respectively, we observed the repair efficiency within the heptamer window to be high (top, $\langle \eta_s \rangle = 0.83$), intermediate for GG (middle, $\langle \eta_s \rangle = 0.68$) and low for the AA case (bottom, $\langle \eta_s \rangle = 0.48$). These observations suggest that a full characterization of the repairability of different DNA motifs should consider at least up to the 4th degree neighbors on both sides.

### Cosine similarity

In contrast to the immediate neighbors of a mismatch, one intuitively suspects that the nucleotides farther away can potentially have a less significant impact on the structure and dynamics of the vicinity of the mismatch. If this is the case, it is possible that the sequence context effect of the distant neighbors can be simplified and perhaps influence the repair by changing the affinity of DNA repair machinery to the substrate independent of the mismatch type itself. To determine if the distant sequence context changes $\eta_s$ by a constant scale factor for all mismatches at the center of these heptamer windows, we compared the similarity of the patterns in all of these excerpts containing CAAAAAT by means of cosine similarity. For this, we represented the repair efficiency of the AA, AC, AG mismatches at the center of the heptamer as a vector in a 3-dimensional repair efficiency space, and calculated the angle ($\theta$) between these 3-dimensional vectors in a pairwise fashion, where higher angles indicate less correlated changes in the repair efficiency of the three central mismatches at that same position (Figure 2.15a). This angle can be calculated by the scalar product of the vectors, that is,

$$\theta(\mathbf{x}, \mathbf{y}) = arccos\left(\frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}\right) \tag{2.9}$$

where $\mathbf{x}$ and $\mathbf{y}$ are vectors each representing one such motif. We shifted the position of the origin to the average repair efficiency position by subtracting the average repair efficiency of all motifs for each coordinate, viz. $x_i = \eta_i - \langle \eta_{ij} \rangle_j$.

Figure 2.15b displays the pairwise similarity angle between these chosen motifs, where the bases along the rows and columns show the 4th degree neighbor nucleotide on the common strand at the 3' and 5' sides, respectively. Apart from the self-comparisons along the diagonal that are perfectly co-linear, we observed the mismatch type preference to be the most similar for the cases where the 4th degree neighbors are AT and CC ($\theta = 55^o$). The angles in general appeared to be high, indicating that changes in the repair efficiency of the three mismatches we investigated were not linearly correlated but rather varying somewhat independently. Strikingly, we noticed two cases where both of the 4th neighbors were C's, but 5th neighbors differed and the mismatch type dependence of the repair efficiency was still more different the former two examples ($112^o$)

indicative of a weak positive correlation. In comparison, $\theta = 80.1^o \pm 1.8^o$ for 72 similar windows randomly chosen out of SSL3.

I will conclude this section by attempting to estimate the length of the 3D vector due to experimental errors using the experimental standard error ($\sigma$) in determination of the three independent Cartesian coordinates, viz.

$$r = \sqrt{\eta_{AA}^2 + \eta_{AC}^2 + \eta_{AG}^2} \tag{2.10}$$

Using which, we can estimate the propagation of errors on the measurements by,

$$\delta r = \frac{x\delta x + y\delta y + z\delta z}{r} \tag{2.11}$$

$$\sigma_r = \frac{\sqrt{x^2\sigma_x^2 + y^2\sigma_y^2 + z^2\sigma_z^2}}{r} \tag{2.12}$$

Table 2.1 estimates the error on the measurements using this formula, according to which only the nonamer whose first and last bases are T (labeled TT) does not deviate from the mean more than the error margin of the experiment. Both based on the variability of the repair efficiency between the mismatches that are within the same heptamer sequence context, as well as the weak positive correlation we observed by means of cosine similarity, we concluded that while short DNA motifs are accurate predictors of repairability, $\eta_s$ cannot be fully predicted by a simplified model considering only sequence trimers, pentamers or even nonamers but a fully predictive model should rather contain at least dodecamer-level detail.

**Table 2.1:** The propagation of experimental error margin to the length of calculated deviation vectors.

| $4^o$ neighbors | $\sigma(\eta_{AA})$ | $\sigma(\eta_{AC})$ | $\sigma(\eta_{AG})$ | $\delta r$ | $\|\|A\|\|$ |
|---|---|---|---|---|---|
| AT | 0.058 | 0.038 | 0.092 | 0.071 | 0.334 |
| CC | 0.016 | 0.023 | 0.090 | 0.027 | 0.054 |
| AA | 0.148 | 0.014 | 0.037 | 0.044 | 0.409 |
| GG | 0.033 | 0.015 | 0.079 | 0.032 | 0.085 |
| TT | 0.087 | 0.097 | 0.178 | 0.114 | 0.058 |
| CC | 0.016 | 0.023 | 0.090 | 0.075 | 0.088 |

### 2.3.10 Prediction of mutational signatures

In the course of the life of the organism or the species, countless genome damage, maintenance and cell division events occur, during which mutations accumulate. The rate of accumulation of different types of mutations are generally different, and these tendencies are sometimes a sign of the biochemical state of the cell. Referred to as "mutational signatures", such patterns can serve as useful tools to understand the affected cellular machinery leading to the diseased state, such as deciphering the common etiology underlying different cancer cases [61].

One common mechanism mutations in the genome can originate through is via mismatched intermediates arising due to DNA replication errors. The frequency at which different types of mismatches accumulate over time highly depends on the sequence context, as well as on the biological processes available in the cell that are involved in the generation or correction of mispairs. While certain nucleotide permutations might give rise to DNA motifs in which the likelihood of a replication error is much higher, the repair of the mistakes within certain sequence motifs might be more difficult, hence also leading to a mutation hot spot. Due to alterations in the DNA sequence or larger scale chromosomal re-arrangements, a mutant DNA polymerase might operate with a lower fidelity, alternatively the proof-reading or MMR capabilities might have been reduced. In more quantitative terms, the relative frequency of observing a mutation will be directly proportional to the number of occurrences of the motif in the genome, number of replication errors made; and inversely proportional to the repair capability.

To numerically estimate the relative abundance of mutations that would be expected within a particular sequence context, we computed the product $f = ae(1 - \eta_s)$ where:

**f, the relative frequency with which a particular mutation should be observed in the genome**

The total number of occurrences of a point mutation can be directly determined by counting deviations from the reference sequence, providing a means to quantify this frequency. As a comparison set obtained with such a strategy, we obtained the human mutation signatures for single base substitutions from Sanger Institute. As an example, the signature SBS1 is given in Figure 2.16a. Here, the six differentially colored groups represent the substitution point mutations from one base to another, whereas the 16 labels along the x-axis per each group denote the identity of the bases surrounding the mutation. The bar heights represent the relative frequency of each type of 96 possible substitutions, the distribution of which constitutes the signature characterizing the metabolic state of the cell.

In this signature model we also adopted, only the type of the substitution and the base identity of nearest neighbor nucleotides are considered, which -accounting for symmetry operations- generates a 96 dimensional vector. Because of the symmetry imposed on the two strands by Watson-Crick base pairing rule, of the $C(4, 2) = 12$ possible base changes, only 6 are independent. As an example, a G>A transition is equivalent to C>T on the opposite strand and hence does not need to be separately accounted for and SBS1 is highly rich in this mutation type. Yet, the immediate nucleotides surrounding this mutation site had a large impact on the prevalence of C>T transition in this signature: for most mutations, the original C was followed by a a G on the 3' side. The aim of this section is to predict mutational signatures that are similar to SBS1.

**a, the number of occurrences of that sequence context in the genome**

If a certain DNA motif is more commonly observed in the genome, it is more likely to observe mutations within it *ceteris paribus*. To account for the difference in the abundance of different DNA motifs in the human genome, we obtained the human reference genome sequence version

38 (GRCh38.p13) and only considered the primary assemblies corresponding to the nuclear DNA. Similarly, for the speculated signatures based on *E. coli* genome, we obtained the assembly version ASM584v2 for K-12 MG1655 and counted the number of occurrences of each trimer and pentamer in the genome, during which overlaps between consecutive k-mers were allowed.

A list of the five most and the least abundant trimer or pentamers can be found in Tables 2.2 and 2.3, inspecting which one can realize that the number of occurrences of sequence motifs might differ by more than two orders of magnitude. In the human genome, AT rich motifs were more common, whereas the most frequent pentamers in *E. coli* genome were relatively higher in GC content. An observation of the relative frequency of the trimers (Figure 2.16b) and pentamers (Figure 2.16c) sorted by rank also show that the disparity between the k-mers is much higher in the human genome. For the computations, we considered both strands of the genome, i.e. the relative abundance of a particular motif is proportional to the average of the incidence of the motif itself and its reverse-complement in the reference sequence. As an example, this means TTTTT is present in the human genome with relative frequency of $(1.00+0.99)/2 = 0.995$.

**Table 2.2:** The frequency of the five most and least abundant DNA trimers and pentamers in the human genome.

| Rank | Trimer | Relative freq. | Pentamer | Relative freq. |
|---|---|---|---|---|
| 1 | TTT | 1.00 | TTTTT | 1.00 |
| 2 | AAA | 0.99 | AAAAA | 0.99 |
| 3 | ATT | 0.95 | ATTTT | 0.64 |
| 4 | AAT | 0.94 | AAAAT | 0.64 |
| 5 | AGA | 0.91 | TATTT | 0.58 |
| 5 | ACG | 0.41 | TCGCG | 0.07 |
| 4 | GCG | 0.37 | CGCGT | 0.06 |
| 3 | CGC | 0.37 | ACGCG | 0.06 |
| 2 | TCG | 0.24 | CGTCG | 0.06 |
| 1 | CGA | 0.23 | CGACG | 0.06 |

**Table 2.3:** The frequency of the five most and least abundant DNA trimers and pentamers in the *E. coli* genome.

| Rank | Motif | Relative #motifs | Motif | Relative #motifs |
|---|---|---|---|---|
| 1 | CGC | 1.00 | CCAGC | 1.00 |
| 2 | GCG | 0.99 | CGCCA | 1.00 |
| 3 | TTT | 0.95 | GCCAG | 0.97 |
| 4 | AAA | 0.94 | GCTGG | 0.97 |
| 5 | CAG | 0.91 | CTGGC | 0.96 |
| 5 | GGG | 0.41 | ACTAG | 0.02 |
| 4 | CTC | 0.37 | TCTAG | 0.01 |
| 3 | GAG | 0.37 | CTAGA | 0.01 |
| 2 | TAG | 0.24 | CCTAG | 0.01 |
| 1 | CTA | 0.23 | CTAGG | 0.01 |

**e, the probability that a replication error occurs in that sequence context**

The probability that DNA polymerase makes a misincorporation error is not the same genome-wide, but rather would depend on the type of the misincorporation made and its sequence context. For this purpose, we re-analyzed the Magnifi assay results, which aims to estimate the relative frequency of DNA polymerase mistakes as a function of both the identity of the misincorporated nucleotide and the local sequence context [10]. In brief, the assay boosts the replication error rates by providing an imbalanced dNTP mixture during oligo extension, where the correct nucleotide to be incorporated is much less abundant than the incorrect nucleotides and hence significantly reduces the sequencing depth needed to detect otherwise scarce misincorporation events. As an example, if the so named Error Enriched Site (EES) contains a T on the template strand, the correct elongation would require a dATP, whose quantity is greatly overwhelmed by dCTP, dGTP and dTTP in the reaction buffer.

We obtained the fastq files using SRA Toolkit v.2.10. and re-calculated the misinscorporation rates using a simple in-house simple C++ program compiled with GCC v9.3. The code imports the reads from a fastq file to the memory and filters for sequences that carry an exactly matching

sequence to the CS2 adapter sequence (AGACCAAGTCTCTGCTACCGTA) and the 3' end of the extension primer (ATCACCAGGTGT). We then determined the expected base identity at the error enriched site (EES) by checking the 100 nucleotide long template sequence surrounding the EES against the 4 possible construct designs reported in the study:

- `GGGCTAGTCGTCTGTATAGG TCTCCGCTTCTTTCCTTTCCTT BBBNBBB`
  `TTCTTCTCTCTCCTTTTTGTTTGCTGTTTGGTGTTGTTTTGGGTGTTCTCC`
- `GGGCTAGTCGTCTGTATAGG ACACCGCAACAAACCAAACCAA VVVNVVV`
  `AACAACACACAGGAAAAAGAAAGGAGAAAGGAGAAGAAAAGGGAGAACACC`
- `GGGCTAGTCGTCTGTATAGG TGTGGGGTTGTTTGGTTTGGTT DDDNDDD`
  `TTGGTGGATATGAGAAAGGTTGGATGTTAGTTGTTGAGTTAGGAGTTGTTG`
- `GGGCTAGTCGTCTGTATAGG TCTTCCCTTCTTTCCTTTCCTT HHHNHHH`
  `TTCCTCCATATCACACACCTTCCATCTTACTTCTTCACTTACCACTTCTTC`

We discarded the reads that violate all these four prototypes and quantified the number of reads that carry an A, C, G or T at the EES (position N). We then computed the relative mutation propensity by only considering the two bases immediately before and after the EES, thereby averaging over the higher order local sequence context. The design principle of the assay ensures that the nucleotide to be inserted at EES and up to the 3rd degree neighbors are different, and therefore this approach cannot quantify mutation events that occur if the EES and the surroundings share the same sequence. As an example, the authors could not make measurements within pentamers AAAAA, ACACC, or CCAAC because the EES (the central base) carries an A, and none of the immediate neighbors can contain an A, as the T required for DNA synthesis across it is scarcely available in the reaction. CCACC, CTATC or GGACG would be examples of pentamers within which measurements were possible. Because of this, I will ignore the contributions of these possible scenarios to the mutational signature in the rest of this section.

In eukaryotic cells, the major contribution to replication of the leading and lagging strands is performed by DNA polymerases $\epsilon$ and $\delta$, respectively [62]. Unfortunately, the error profiles of

such major replicative DNA polymerases were not available to the required level of detail to my knowledge, whether *in vitro* or *in vivo*. We instead had to further assume that the error profile of the replicative DNA polymerases are similar to the *in vitro* error rate of common polymerases used in lab settings. As a proxy for a polymerase that has proofreading activity, I will make use of the Magnifi data acquired using Phi29 polymerase, which resembles DNA polymerases $\epsilon$ and $\delta$ as it has proofreading activity. To mimic the aberrantly mutagenic DNA replication, we used the error propensity measurements using DNA polymerase IV, instead. For both enzymes, we made use of the data acquired using the $1 : 10^8$ correct/incorrect nucleotide abundance ratio.

### $\eta_s$, repair efficiency of the mismatch intermediate(s)

A mismatch that is repaired before any replication attempt is made does not contribute to the mutational load, whereas an unrepaired mismatch will lead to a mismatch about half of the time, as one of the two conflicting strands will be the template strand identical to the reference sequence. For the purposes of this exercise, I will be assuming that the epigenetic marks are made into efficient use by the cells to ensure a 100% strand bias favoring the retention of the template strand. This assumption implies that repair always happens correctly, if it happens, by reversing the DNA polymerase misincorporation error. The repair efficiency was reported to depend on the chromatin packaging in eukaryotes, and the nucleosomes are positioned along the chromosomal DNA in a sequence dependent manner, but I will be ignoring this chromatin-packaging mediated sequence bias for simplicity.

Using the simple approach described above, Figures 2.17a and 2.17c display the estimated mutational signatures if the sequence composition resembles that of *E. coli*, depending on whether the replication occurs with or without proof-reading. I would like to emphasize that the mutational signatures have been normalized to sum up to 1, and hence the intuitively expected increase in the overall mutation rate in the latter case is not expected to be directly observable on the plots. Similarly, Figures 2.17b and 2.17d show the mutation signatures obtained by using the human genome.

Finally, we asked if these human-genome adjusted mutational signatures are similar to the

reported signatures reported for human cancer cells [61]. Treating each signature as a vector in a 96 dimensional space, we hypothesized that similar signatures should be more close to being co-linear, as opposed to two different signatures and hence made use of the cosine similarity concept, in an analogous way to Equation 2.9. We observed that none of the two signatures we hence speculated about resemble any of the 72 cancer signatures very closely (Figures 2.18a and 2.18b). The most pairwise angles ($\theta$) were about $80^o$, indicating that these speculated signatures are almost orthogonal to the COSMIC signatures. Since the repair efficiency data we have made use of here was obtained in wt cells, under the additional assumption that the evolutionary well-conserved nature of the MMR proteins make the universal physical properties of DNA the major determinant of the differential repair response across taxonomic kingdoms, we expect the cancer signatures whose etiologies are linked with well-characterized defects in MMR (namely signatures 6, 15, 21, 26, 44) not to have exceptionally higher cosine similarity (i.e. lower $\theta$) than the others. We indeed observed that the mean value of the angles we estimated (blue bars) did not significantly deviate from the overall mean neither with ($79.0^o \pm 6.5^o$ vs $78.9^o \pm 5.7^o$, p=0.50 by one tailed z-test) or without proofreading ($83.0^o \pm 3.1^o$ vs $80.9^o \pm 4.6^o$, p=0.40). We excluded signatures 14 and 20 from this analysis as their etiologies indicate defects in MMR capabilities as well as concurrent mutations in DNA polymerase $\epsilon$ and $\delta$, respectively.

Having said the above, the conclusions of this section are much more speculative than it should be. While the inadequacies of the available data was evident, the main aim of this section was to suggest an approach that the repair efficiency data could be made useful. Further experiments are essential to characterize the replication error rate in full detail. A more descriptive model that takes the chromatin packaging in human nuclei into account would also very likely improve the accuracy of the postulated signatures.

### 2.3.11 Apparent repair efficiency in mutant cells

In the above sequel, I reported the repair efficiency measurements in wt *E. coli* cells, where MMR is functional and most mismatches were repaired efficiently. In contrast, experiments in $\Delta$mutS cells lead to a significant reduction in the observed repair efficiencies, hence suggesting that the

observed response is mostly an output of the MMR pathway (Figure 2.8b). While we observed a statistically highly significant decrease in the repair efficiency of all mismatches in these MMR deficient cells compared to wt cells, this failure to repair was far from being complete, but rather $\Delta mutS$ cells could repair the introduced mismatches about half of the time.

As we deleted farther downstream elements of MMR, we observed the repair machinery to be less and less impaired by the mutation. While the wt cells repaired mismatches on 3ML1 with $0.89 \pm 0.15$ ($\eta = \mu \pm \sigma$), the $\Delta mutS$ cells only could repair with an apparent efficiency of $0.43 \pm 0.06$. The repair efficiency observed in $\Delta mutL$ ($0.49 \pm 0.10$), $\Delta mutH$ ($0.54 \pm 0.12$) and $\Delta uvrD$ ($0.64 \pm 0.20$) were progressively higher but remained between those of wt and $\Delta mutS$ cells. That is, the availability of a functional MutS is more essential then MutL or MutH whereas absence of cytoplasmic UvrD impedes repair but is the most tolerable for repair to succeed. In contrast, the overall appearance of the substitution prevalence histograms remained similar for all mutants 2.19b. On the other hand, the spread of the $\eta$ distribution in each mutant strain also gradually increased as the mutation shifted towards the farther downstream elements of the MMR pathway. This latter effect can be observed qualitatively from the extent of pixel-to-pixel variation observable in $\eta$ matrices for each mutant strain (Figure 2.19a) or by the gradual increase in the spread of the $\eta$ distributions mentioned above. We observed these phenomenon not to be specific response to the 3ML1 library only (Section 1.3.7).

To obtain a quantitative measure of this differential level of sequence dependent repair in different mutant strains, we compared the raw apparent repair efficiency ($\eta$) of well-repaired and poorly-repaired mismatches across strains. In all mismatch libraries we reported so far, all CC mismatches were poorly repaired, whereas GG mismatches were well-repaired by a wt cell regardless of the sequence context (Figures 2.11a, 2.11b and 2.11c). While both CC and GG mismatch types were poorly repaired by $\Delta mutS$ cells ($\mu_\eta \pm \sigma_\eta = 0.31 \pm 0.06$ vs. $0.28 \pm 0.03$ for CC and GG, respectively), we observed the difference between the mean repair efficiency of these two groups to widen stepwise rather than in a all-or-none manner (Figure 2.19c). While a repair efficiency gap between these two groups emerges in $\Delta mutL$ ($0.43 \pm 0.06$ vs $0.50 \pm 0.06$) and $\Delta mutH$ cells ($0.43 \pm 0.06$ vs $0.62 \pm 0.05$), the extent of the gap in $\Delta uvrD$ ($0.38 \pm 0.04$ vs $0.82 \pm 0.05$) approaches

to that observed in wt cells ($0.56 \pm 0.08$ vs $0.98 \pm 0.00$). These results mean that while the loss of function of MutS disrupts the sequence-dependent DNA repair, absence of downstream proteins do not have an equally significant impact. Our data therefore make us think that the major determinant of DNA repair efficiency is at the detection step by MutS and to a lesser extent by MutL, but not significantly influenced by the DNA cleavage or unwinding steps.

As the tUNC19 plasmids are derived from pUC19 plasmids carrying high-copy number pBR322 ori [63], at a given time there will be many copies of the replicated plasmids in each cell, hence casting doubts about whether homologous recombination is a potential contributor to the obtained readings. To assess this hypothesis, we also repeated the repair efficiency measurements using the same 3ML1 mismatch library, but in ΔrecA cells that are defective in homologous recombination and observed a high correlation with our measurements in wt cells (Figure 2.20b). This observation suggests that the repair response we observe is not dominantly contributed by potential recombination events between plasmids carrying the variable or common variants of plasmids in the cytoplasm to a significant degree. On a similar note, we had measured the repaired efficiency of SML library in $\Delta UvrB$ cells and compared with the results obtained in wt cells (Figure 2.20c). The poorly repaired mismatches in both cell types were the same, suggesting that the involvement of the nucleotide excision pathway is not very significant, either.

### 2.3.12 The influence of AT content on repair efficiency

As the percentage of GC content (% of G or C nucleotides on the DNA of interest) is a determinant of dsDNA stability as well as DNA-protein interactions, we next checked if there is any correlation between the repair efficiency we measured and the GC content of the underlying DNA sequence. For this, we first checked the local GC content surrounding the mismatch, which we defined as the 9-nucleotide window centered at the position of the mismatch. Figure 2.21a shows a scatter plot of $\eta_s$ with respect to the local GC content, where each red dot represents one mismatch in 3ML1, 5MLx or SSLx library. We observed that the repair efficiency moderately depends on the local GC content.

As a second measure of GC content, we considered the consensus sequence of the entire mis-

match library and compared with the ensemble averaged repair efficiency of all mismatches sampled as part of this library ($\langle \eta_s \rangle$). We observed that the repair efficiency positively correlated with the total GC content of the library (Figure 2.21b). The 3ML1 library contains all trimers exactly once and hence provides a DNA sequence that is very close to 50% in GC content (purple). The individual sublibraries of 5MLx vary in GC content (blue), and the observed mean repair efficiency varied in a way that depends on the GC content. SSLx is highly enriched in AT content, as it is constructed to contain heptamer motifs all harboring central TTT's and the repair on all 4 constituent sublibraries was less efficient (red). These trends then collectively argue that the repair response is positively correlated with the GC content.

### 2.3.13 The influence of the data quantity

In this new experimental scheme that I propose here to measure the repair efficiency of DNA mismatches, the accuracy of the measurements depends on two different statistical factors. First and foremost, the ternary classification of clans as C, U and V should be performed accurately. Secondly, the C, U, V clans should be detected with the correct frequency, so that the repair efficiency ratio can be evaluated accurately.

The former accuracy criterion depends on the sequencing depth of the experiment in the form of number of sequencing reads per clan. If the sequencing depth is too low, the misclassification probability of U clans as C or V-clans will increase, since the probability of subsampling a clan exclusively consisting of common or variable strands out of an equimolar mixture of the two will be higher if the number of draws is lower. On the contrary, the likelihood of a sequencing error causing misclassification of a C or V-type clan as a U-type clan will be higher for a smaller clan. As the mapping barcodes help eliminating the second type of errors, the overall expectation is that a shallower sequencing depth would lead to a higher apparent repair efficiency level.

Figures 2.22a and 2.22d compare the median clan size in 3ML1 and 5MLx libraries, respectively, according to which the sequencing depth of 3ML1 is higher by about 5 fold. Figures 2.22b and 2.22e show the s-histograms for the clans obtained from these two libraries, either for the entire dataset (black solid lines) or those whose sizes are below median (green dashed) or above median (purple

dotted). Apart from the wider C-peak distributions for the below median clans, no significant qualitative difference is observable, an indication of sufficient sequencing coverage.

On the other hand, the latter accuracy criterion is related to the total number of clans detected per each mismatch, as the greater number of single plasmid molecules with the identical mismatch will improve the confidence in the deduced relative frequency of repaired clans. As the typical value of repair efficiency we measured is around 90%, a U-type clan is detected with only about 10% probability. If the total number of clans per mismatch is insufficiently low, some mismatches will be detected as if repaired with 100% efficiency as the accidental likelihood of not detecting any U-type clan will be higher. The scatter plots in Figures 2.22c and 2.22f show a mild negative correlation between $\eta_s$ and the number of clans detected per mismatch at the sublibrary level and single mismatch level, respectively.

### 2.3.14 Modifications on the vector design

Being a pUC19 derivative, the barcoded vector we used in the majority of the experiments that I have been referring to as tUNC19 carries a lac promoter immediately following the cloning site of the mismatch library into the construct (Figure 2.7b). This brings the barcode-distal end of the investigated mismatch libraries to the close proximity of a promoter, which can confound our measurements due to competition between transcription and repair machineries to bind to the same DNA substrate or via an equivalent of transcription coupled repair mechanism [51]. As our approach relies on the plasmid replication as the time keeper, a potential delay of the replication initiation or progression due to binding of transcription factors or elongation of an R-loop might also affect the repair efficiencies we report [64]. To assess the potential effect of this phenomenon on our system, we repeated our assay on a shortened version of tUNC19 that is generated by a PCR on pUC19 using primers P2 and P4. We hence obtained a new vector short19 that omits the lac promoter but has the same DNA sequence as tUNC19 otherwise. We observed the results obtained with this truncated plasmids to highly correlate with those obtained with the same 3ML1 library as part of the full tUNC19, hence suggesting that transcription coupled repair has a non-dominant effect on our results, if any (Figures 2.23a and 2.23b). This observation also suggests

that the asymmetric behavior of C- and V-type clans in the s-histograms is likely not due to the strand asymmetry between the coding and non-coding strands of the plasmid.

As a second concern, the barcoding PCR in our workflow can generate vectors that are improperly annealed under certain circumstances. At any stage during the thermal cycling procedure, the recently synthesized strands by the annealing of barcode-bearing primer (Primer P2) to an already barcoded DNA strand that was synthesized in the previous cycles with a different barcode will result in a hetroduplex that carries two different barcodes on the two annealed strands. As the amplification procedure proceeds, the quantity of the primers will also gradually decrease, hence increasing the likelihood that the full length ssDNA carrying different barcodes aberrantly annealing to each other as the temperature is decreased for the primers to anneal. Both of these scenarios can lead to molecules that contain a bubble at the location of the tracing barcode, but are properly annealed dsDNA along the rest of the 2.5kb long amplicon. To investigate the effect of such putative bubbles on the mismatch carrying plasmids, we attempted a control experiment where one of the allegedly improperly annealed strands in the PCR products is displaced by an additional primer extension step. To achieve this, we added 10 fold molar excess of primer P1 to the tUNC19 PCR product in the presence of Phusion polymerase and subjected to one cycle primer extension. We then used this product I call tUNC19-X for the same downstream steps: we ligated this product with the same 3ML1 library as before and measured the repair efficiency of individual mismatches. We did not observe a systematic discrepancy between the results obtained with and without this additional primer extension step, suggesting that either such aberrant products do not exist to a very high extent among our plasmids, or that their presence does not have a significant effect on the reported repair efficiencies (Figures 2.23c and 2.23d). We hence generated the rest of our dataset without this treatment.

### 2.3.15 Information entropy

We then sought a systematic approach to assess the data quality we obtain by quantifying the variability in the observed base frequencies. In contrast to the oligo annealing based approach to generate mismatch libraries, oligo pool synthesis and multiple PCR amplifications are part

of the current workflow. This multi-step process is high fidelity but is still error-prone, and potential errors in the DNA sequences would accumulate leading to detection of stray events. Among the experimentally introduced artifacts, potential PCR errors before the transformation will only change the identity of the plasmid library member, as they take place before the sample is seen by the cells and replicated, so they are likely to introduce a second mismatch upon partial annealing with the common strand. Such misincorporations will hence change the response of the cell potentially against all mismatches as the detection of a secondary mismatch will trigger the re-synthesis of the vicinity thereby causing the repair of the original mismatch. On the other hand, errors introduced during the intra-cellular plasmid replication and the sequencing library preparation following the plasmid extraction will not change the properties of the mismatch bearing plasmid that the cell responds to, but rather will confound the interpretation of the data. However, any deviation from the consensus sequence will be disregarded unless it is at the same position as the original mismatch, because the mismatch position and identity is encoded by the mapping barcode and deviations from the designed sequence are detectable. For this reason, this latter type of errors are likely less detrimental.

As a measure of experimental errors arising from the accumulation of mutations, we resort to the information entropy concept, which is proportional to the degree of randomness observed and is defined by [65]:

$$S = -\sum_{i} p(x_i) log(p(x_i)) \tag{2.13}$$

which as applied to a standard DNA sequence consisting of 4 standard basis would be:

$$S_i = -\sum_{N=\{A,C,G,T\}} Prob_i(N)\ log_2\left(Prob_i(N)\right) \tag{2.14}$$

where $S_i$ denotes the entropy of the DNA read at the i'th nucleotide position, and $Prob_i(N)$ is the experimentally deduced probability of observing base the N at the i'th position. For a fully randomized DNA, the information entropy as defined above reaches its maximum value of 2, since $\forall i, N, Prob_i(N) = 0.25$ and thence $S_i = -4 \cdot 0.25 \cdot log_2(0.25) = 2$. On the other extreme, for a

perfect dsDNA without any mismatches or substitutions, we expect a pure product representing the consensus sequence ($\{C_i\}$) only, as was the case with experiments performed using library NPL or 3CL1. In this case $\forall i, Prob_i(N) = \delta_{NC_i}$ and therefore $S_i = -\sum_N \delta_{NC_i} \cdot log_2(\delta_{NC_i}) = -log_2(\delta_{C_iC_i}) = 0$, i.e. the base sequence is fully determined by the consensus sequence alone and no extra information is available in any read. Entropy obtained from an experimental output will be between these two limits in practice.

For our mismatch libraries, $C_i$ is the pre-defined consensus sequence of the mismatch library and hence indicates the most likely base to be observed at position i. In particular, an ideal mismatch library carrying a random single mismatch at a random position with equal probabilities, we have

$$Prob_i(N) = \begin{cases} 1 - 3 \cdot \frac{b}{3L} & N = C_i \\ \frac{b}{3L} & N \neq C_i \end{cases} \tag{2.15}$$

because for a mismatch library of length L, 3L different mismatches are possible, where each member is ideally observed with about the same frequency. Using this we can reach,

$$\begin{aligned} S_i &= -3 \cdot \frac{b}{3L} \cdot log_2\left(\frac{b}{3L}\right) - 1 \cdot \left(1 - 3 \cdot \frac{b}{3L}\right) \cdot log_2\left(1 - 3 \cdot \frac{b}{3L}\right) \\ S_i &= -\frac{b}{L} \cdot log_2\left(\frac{b}{3L}\right) + \left(\frac{b}{L} - 1\right) \cdot log_2\left(1 - \frac{b}{L}\right) \end{aligned} \tag{2.16}$$

where $b = (\beta + 1)/2$ is the frequency of observing the substituted strand rather than the consensus strand. This latter parameter is a measure of the strand choice bias, which will be assumed to be 0.5 in the absence of selectively deposited epigenetic marks promoting preferential retention of the common or variable strand. The relationship between the library length and the expected information entropy for an idealized library is plotted in Figure 2.24a as a function of b. As would be intuitively expected, a higher b parameter leads to a higher level of entropy, as it means a higher prevalence of the variable strands rather than common strands in the mixture. Under these assumptions, we reach that $S \sim 0.1$ for an idealized library uniformly sampling mismatches along a library length L$\sim$50bp, but should remain as small as possible if no mismatches are introduced to the system by design.

Out of an experimental data set, this entropy parameter can be extracted by obtaining the

base frequencies for all sequences that are part of all clans detected. The stark contrast between the entropy distributions of mismatch devoid (NPL) and mismatch containing (SML) libraries is displayed in Figure 2.24b for the oligo annealing based libraries of Chapter 1. Our theoretical model developed in Equation 2.16 is indicated by the blue line, which correctly captures the order of magnitude of S. Both in the presence and absence of an active MMR system in the cell, the information entropy in the absence of deliberately introduced substitutions (NPL) is much lower than when they are introduced intentionally (SML), while the consensus sequences of the two were kept identical. I would also like to note that the substitution mutations during sample preparation or inaccurate base calling will also contribute to the measured entropy. In fact, for the 3ML1 library, whose preparation involves an oligo array synthesis and PCR, we observe the background entropy level for the analogous control library without mismatches (3CL1) to be much higher than NPL (Figure 2.24c). But despite the potential DNA polymerase errors, the entropy levels in fully-paired DNA were still less than the corresponding entropy level for the mismatch-bearing cases with or without active MMR (Figure 2.24d). To be more precise, for 64 out of the 66 positions available in the mismatch library, the entropy of 3CL1 in wt cells was the lowest of the three samples, whereas the relative entropy ranks of 3ML1 in wt and $\Delta mutS$ cells were less determined. The relatively higher apparent entropy of wt cells can potentially be due to inherent strand bias during repair favoring the retention of the substituted strand. All in all, the general trend in the information entropies we measured were in line with our expectations.

### 2.3.16 The effect of DNA methylation on the observed repair response

Having measured the repair efficiency on naked *in vitro* synthesized DNA, we asked if DNA modifications have an observable effect. In particular, *E. coli* MMR is known to be methylation sensitive, as the nascent strand during DNA replication is unmethylated for a brief period of time, while the template strand is methylated hence forming a physical basis for a strand-selective repair procedure. As such, the preferential removal of the unmethylated strand can diminish the mutational load caused by misincorporation errors. In the form presented in the text so far, this system has been malfunctional as both strands of the vectors were intentionally devoid of methylation, being

PCR products using standard unmodified dNTPs. We asked if and how methylation effects our repair efficiency results.

First, we attempted to introduce pre-methylated plasmids to the cells. To achieve this, we included an *in vitro* deoxyadenosine methylase (Dam) treatment step of tUNC19 preparation workflow following the PCR step. This process introduces methyl groups on adenines located in GATC sequence tetramer motifs symmetrically on both strands (vector mm19), hence we expect about equal population of C- and V-type clans as was observed in wt cells with unmethylated plasmids. We similarly tried to measure the repair efficiency of hemi-methylated DNA. To best of our knowledge, a simple *in vitro* system to produce hemi-methylated vectors is not available. Instead, we produced fully methylated DNA as above and then replaced one of the methylated strands by primer extension throughout the vector backbone by supplying only one of the primers used in the previous PCR (vector hm19). As this newly synthesized strand is a standard PCR product polymerized using unmethylated dNTPs, the strand attached to the V strand of the mismatch library in the ligated plasmids is unmethylated, while the C-strand remains methylated. This arrangement is expected to increase the retention of C-strands during repair, while triggering preferential removal of V-strands (Figure 2.25a).

A biologically more realistic description of the repair response would involve the presence of a bias favoring presence of one of the two strands in more often than 50% of the reads that can significantly alter the observed entropy levels. For a system that preferentially retains the common strand, the expected entropy will also be lower (lower values of b in Figure 2.24a) whereas a tendency to report more reads originating from the variable strand should increase the measured entropy (higher b). A strand selection bias in the former form can be experimentally induced by using the hemi-methylated vector backbone described above (hm19). We experimentally verified this proposition on the DEB3L/DEB3R libraries based on oligo annealing (Figure 2.25b). Compared to its fully unmethylated counterpart (tUNC19), we observed that this hemi-methylated sample (hm19) had systematically lower entropy. A similar experiment with the PCR-based 3ML1 library lead to similar conclusions, and further revealed that a strand selection bias is much more significant in wt cells compared to MMR deficient cell strains $\Delta mutS$ or $\Delta mutH$. Surprisingly,

we did not observe a significant entropic difference between the two methylation states for $\Delta dam$ cells, which might be attributable to the replication-related properties of a hemi-methylated origin of replication (Figures 2.25c and 2.25d).

At single-molecule level, the asymmetric methylation marks also changed the population of C-U-V clans beyond the replicate-to-replicate variation favoring the retention of the C-strand over V-strand in comparison to the fully unmethylated case (Figures 2.26a and 2.26c). The RMS discrepancy between the strand retention bias ($\beta = (\#C - \#V)/(\#C + \#V)$) measured between two experimental replicates was 0.09 for the hemi-methylated case and 0.04 for the unmethylated case, whereas the mean bias of hemi-methylated plasmids was higher than the unmethylated case by 0.20. Yet, these changes remained limited in extent. In comparison, under certain conditions, MMR has been reported to provide near-complete shift in the strand bias favoring the retention of the methylated strand on a hemi-methylated $\lambda$ DNA ($\beta = 0.96$ or $\beta = -0.92$), as opposed to the same constructs symmetrically methylated on both strands ($\beta = 0.03$, p-value $<10^{-5}$ by one-tailed z-test for both cases) [50].

The limited extent of the change in the strand choice bias might be caused by very fast methylation of the unmethylated plasmids immediately following transformation before either repair or replication takes place. As the wt strain used in the experiments endogenously expresses deoxyadenosine methyl transferase (Dam) and hence can potentially fully methylate the transformed plasmids in the cytoplasm before replication or repair takes place, we next repeated our experiments in $\Delta dam$ cells [46]. When we performed the measurements with unmethylated vectors ligated to 3ML1 library, we observed the results of $\Delta dam$ cells to highly correlate with the wt cells' response (Figure 2.20a). This observation suggests that a potential methylation state introduced intra-cellularly following the transformation of unmethylated plasmids does not have an important effect on the repair efficiency of different mismatches relative to each other. In contrast, the introduction of hemi-methylated plasmids into $\Delta dam$ cells significantly increased $\eta$ for all poorly repaired mismatches (Figure 2.26b). The increased $\eta$ in the absence of Dam might be explained by the reported inhibition of the *in vivo* replication of plasmids carrying a hemi-methylated pBR322 replication origin [66]. While Dam is not part of the MMR pathway, such a

potential replication delay could grant a functionally intact repair machinery more time to respond against the mismatches, and hence increasing the apparent repair efficiency globally, because our assay uses the first passage time of the replication machinery as the timekeeper.

Next, we compared the strand choice bias during repair for unmethylated vs. hemi-methylated plasmids by referring to the substitution prevalence histograms as before (Figure 2.26d). In $\Delta dam$ cells, we observed the C-peak to be retained more frequently in the presence of hemi-methylation marks as opposed to the total absence of methylation marks on either strand. Yet, the magnitude of this bias was low in magnitude ($\langle\beta\rangle = -0.15$ vs $\langle\beta\rangle = -0.11$, respectively, p-value=0.09). That the strand selection bias is limited also in $\Delta dam$ cells argues against possible fast methylation of our introduced plasmids by Dam in wt cells, hence justifying our approach. This argument is further justified by the fact that, a similar preferential retention of C-strands in $\Delta mutH$ cells was not observed. All in all, while the well-characterized methylation-dependent response of the MMR pathway is evident in our results, the potential uncertainty regarding the methylation state of the repair templates is likely not an important factor confounding the repair efficiencies we reported.

### 2.3.17 The repair response of cells that are both MMR and methylation deficient

We previously had observed that the symmetrically or asymmetrically deposited methylation marks alter the repair response against the mismatch libraries we have introduced to an observable degree. In particular, the abolition of *in vivo* methylation reduces the relative repairability of the mismatches that are repaired with an intermediate repair efficiency, while the overall results closely correlate with that observed in wt cells (Figure 2.20a). Introduction of hemi-methylated DNA significantly increased the global apparent repair efficiency levels (Figure 2.26b), while at least a portion of this effect can potentially be caused by the delayed replication of hemi-methylated plasmids. While the deposition of such hemi-methylation marks lead to an enhanced preference to retain the common strand of the mismatch libraries (Figure 2.25b), the strength and the direction of this preference was less clear in ΔMMR or $\Delta dam$ cells (Figure 2.25d). These latter measurements were also confounded by the presence of active cytoplasmic Dam that can change the methylation

state of the transformed plasmids before repair or replication takes place.

We hence further investigated the effect of the methylation on the apparent repair response in cell strains in which both an MMR pathway element and the methylation enzyme Dam is unavailable due to the double mutant genotype ($\Delta$MMR $\Delta dam$). For each of the four double mutants, we independently measured the repair efficiency of mismatches that are represented in the 3ML1 mismatch library that was ligated into an unmethylated (tUNC19, Figure 2.27a), symmetrically methylated on both the common and variable strands (mm19, Figure 2.27b), or on the common strand only (hm19, Figure 2.27c). In general, the apparent repair efficiency of a fully methylated plasmid was the highest of the three cases, whereas the unmethylated plasmids were repaired with the poorest efficiency. For the unmethylated case, similar to our observations before, deletion of *mutS* ($\mu \pm \sigma = 0.42 \pm 0.07$) or *mutL* ($0.44 \pm 0.07$) had a significant negative impact on the repair capability than cell strains that lacked MutH ($0.61 \pm 0.14$) or UvrD ($0.63 \pm 0.15$). Albeit the absolute magnitude of the measured repair efficiency varied, this trend was the case for all three epigenetic states we have studied.

We next compared the substitution prevalence histograms of the double mutant cell lines with $\Delta dam$ cells and observed in all cases the three subpopulations that we typically observed before (Figure 2.28a). While the C- and V-type peaks in $\Delta mutH$ $\Delta dam$ and $\Delta uvrD$ $\Delta dam$ cells were much more pronounced, in all cases the V-type peaks were much broader than C-type peaks, similar to our past observations. A separate investigation of each cell type for the three different methylation states did not reveal a major systematic difference between the histograms, except a slightly higher population of U-peaks at the expense of C- or V-peaks when unmethylated plasmids were transformed, whereas the U-type clan population in fully methylated plasmids was the lowest of the three (Figures 2.28b-2.28f). On a similar note, the comparison of substitution prevalence histograms of $\Delta$MMR $\Delta dam$ double mutants with the corresponding $\Delta$MMR only cells that contain a functional *in vivo* methylation system suggests that the C-, U-, and V- type clans are about equally present in these two groups (Figures 2.29b-2.29e). This similarity was also the case between $\Delta mutH$ and $\Delta mutH$ $\Delta dam$ cells transformed with hemi-methylated plasmids (Figure 2.29f). On the contrary, wt cells had a significantly higher proportion of C- and V-type

clans compared to $\Delta dam$ cells (Figure 2.29a).

The high correlation between the repair efficiencies measured in unmethylated and hemi-methylated plasmids suggest that in all MMR-Dam double mutant cell lines, the sequence dependence of the repair response is very similar in the two sets (Figure 2.30a). This contrasts with the enhanced apparent repair in the cells with functional MMR, where the apparent repair efficiency of a hemi-methylated DNA was substantially higher. A similar comparison between the fully methylated and unmethylated plasmids transformed into MMR-Dam double mutants do not reveal a major difference in repair response against the individual mismatches, except that the mismatches on methylated plasmids were repaired with a slightly higher repair efficiency on average (Figure 2.30b). The sequence preference of repair was also mostly similar with or without intra-cellular methylation capability (Figure 2.30c). In contrast, an analogous approach to the repair efficiencies measured in ΔMMR only cells suggest that mismatches that were repaired with an intermediate efficiency in MMR mutants are oftentimes well-repaired in a wt cell (Figure 2.30d).

In $\Delta mutH$ $\Delta dam$ and $\Delta uvrD$ $\Delta dam$ cells, we observed a notable residual repair capability, whose extent was sequence dependent, as revealed by the presence of pixels with orange and red tones as opposed to the yellow overall trend in Figure 2.27. We finally asked whether the sequence dependent characteristics of this residual repair activity is similar to that of the MMR response in wt cells. For all three methylation cases we have investigated, the repair efficiency of $\Delta mutH$ $\Delta dam$ (Figure 2.31a) and $\Delta uvrD$ $\Delta dam$ (Figure 2.31b) cells correlated with that of wt cells, except that the mismatches with intermediate repair efficiency were repaired with a disproportionately poor efficiency in these mutant cells, revealing itself as a v-shaped distribution on the scatter plots. Of the three methylation states, the residual repair efficiency of fully methylated plasmids followed the wt repair response the most closely and the hemi-methylated plasmids displayed the most deviation. The residual repair efficiency in $\Delta mutH$ $\Delta dam$ and $\Delta uvrD$ $\Delta dam$ cells well-correlated with each other for all three methylation cases (Figure 2.31c). For the hemi-methylated case, the sequence preference of a $\Delta mutH$ cell was similar with or without the availability of Dam, similar to the observation as in Figure 2.30c (Figure 2.31d). In summary, the sequence preference of the residual mismatch repair efficiency is mostly similar to that of the wt cell's MMR response. If

MutS binding is the primary determinant of this sequence dependence, this similarity might be an indication that the mismatch detection is common between the two repair mechanisms even though the corrective action is taken by alternative effector proteins.

## 2.4 Conclusion

In this chapter I described the double-barcoding strategy we developed as an improved method to quantify the repair efficiency of mismatches using random DNA sequences as heritable unique molecular identifiers. Making use of the next-generation sequencing methods whose output has been increasing at diminishing costs [67], our method can scan large mismatch libraries in a non-labor intensive way with only about 3 days of hands-on processing. While the efficiency of the cellular response against the mismatch bearing DNA has been measured as early as in 1980s, an equivalent high-throughput method has not been reported in the literature to my knowledge. Using this method, we measured the *in vivo* repair efficiency of 4434 distinct mismatches in K-12 *E. coli*, thereby sampling all possible mismatches within all pentameric sequence contexts at least once. I believe these results constitute the most comprehensive repair efficiency compendium reported in the literature to date.

In agreement with the literature, our results indicate that most of the mismatches on our plasmids can be efficiently repaired whereas CC mismatches are often missed by the repair machinery [6]. We observed a strong sequence context dependence for AG, CT and TT mismatches and assessed the effect of dinucleotides immediately before and after these mismatches. The context dependence is not limited to the nearest-neighbors and extents to farther neighbors. Given that the reports regarding the footprint of MutS on dsDNA range from 8 to 20 nucleotides [17,18], one might suspect that the sequence contexts as large as 10th nearest neighbor might have an impact on the repair efficiency of the mismatch. In certain regions of the genome, it is likely that sequence contexts matter even on larger length scales as was shown for one specific sequence previously [68]. As a particular example, for an average TT mismatch our measurements yield $\langle \eta_s \rangle = 0.748$, whereas in a pentameric sequence context containing a TT mismatch neighbored by an A at the 5' side and followed by a C on the 3' end, our results indicate $\eta_s = 0.191$ hence

suggesting it to be a much more difficult to repair mismatch than usual. I would like to point that while this novel method provides an easy and feasible approach to scan the repairability of different mismatches within one or a few template sequences, it is not very suitable for the determination of such longer-range sequence effects as the length of the sequence template rapidly increases with the length of the sequence motifs of concern. As such, while our results can capture the repair efficiency difference between mismatches, we observe that the sequence context effect extends to the base composition of the loci at least as far as 4 nucleotides away from the position of the mismatch.

Our results indicate that CC mismatches are always poorly repaired, whereas AG, CT or TT mismatches are repaired in a sequence context dependent manner and these observations largely agree with the MutS sliding clamp formation propensity studied by surface plasmon resonance, where the authors observed the affinity of MutS to be low against CC, TT and AG mismatches [69]. We calculated $\eta_s$ values of the sequences used in the SPR study and plotted against the binding affinity as revealed by $K_d$. For all difficult to repair mismatches, i.e. those with a low $\eta_s$, MutS had a low binding affinity (Figure 2.32e). Likewise, comparing with the biochemical data on human MutS homologs, we found that sequences with low $\eta_s$ values have low $k_{cat}$ values [70] (Figure 2.32d).

Our expectation that the observed repair efficiencies should be highly influenced by the interactions of MutS with mismatched DNA is supported by the observation that the greatest reduction in the repair capability was imparted by the absence of MutS, and the effect was progressively lower for MutL, MutH and UvrD, respectively. This suggests that the primary determinant of repairability is the detection of the mismatches, which could potentially delay the replication initiation or progression, hence giving more opportunity for the repair to occur. In contrast, that the absence of MutH or UvrD leads to an incomplete reduction in the repair capacity can possibly suggest alternative side pathways that can lead to the (apparent) repair of a mismatch along multiple points of canonical MMR pathway, provided that the mismatch detection has successfully occurred.

While the MMR is an evolutionary well-conserved response across kingdoms, in eukaryotes,

both the efficiency of DNA repair and the incidence of DNA damage depend on the relative positioning with respect to nucleosome core particles [71, 72]. While the DNA that wrapped around nucleosomes are generally repaired with a lower efficiency in comparison to linker DNA, the wound DNA segments with minor groves facing towards the histones also have a measurably lower repair efficiency than those with minor groves facing outwards. Hence the repair efficiency of a eukaryotic cell against a genomic mismatch in its full biological context might harbor finer prints than *E. coli*. On the other hand, previous studies have indicated that both the mismatch type and the local sequence context have a measurable effect on the local physical properties of DNA that could influence its shape and rigidity [73–77]. As such, the differential response of the MMR to the mismatches can be related to the change in the affinity or binding mode of MutS or downstream members of the pathway rather than being a highly species-specific property of the MMR elements, suggesting that the sequence-dependent characteristics we reported here can potentially be generalized to other organisms. In fact, DNA mismatches have also been studied using all atom molecular dynamics (MD) simulations, revealing how the mismatch types and the flexibility of the flanking sequences influence DNA structure dynamics [78]. We found that mismatches with low $\eta_s$ values according to our data have above average helical twist (Figure 2.32b) and narrower than average minor grooves according to MD simulations (Figure 2.32c). Such poorly repaired mismatches also reduce the local dynamics of DNA, as reported by the extent of breathing observed in the course of the simulations (Figure 2.32a).

Having said so, our interpretations of the results are far from being complete. To start with, we observe a high baseline level of repair at about 35% for the inefficiently repaired mismatches or in MMR-inoperative cells. As our main focus was on the repairability of mismatches relative to each other, we opted to report the outcomes after eliminating this baseline effect in the majority of this work. This high baseline can be partially explained by the sampling bias that leads to a systematic misclassification of U-type clans as C or V-type, due to a failure to detect either the variable or the common strand components of the clan. Such a binomial sampling error would introduce a significant measurement bias much more if the clan size is small. In fact, we observed this high baseline to occur in measurements performed in ΔmutS cells, where MMR pathway is

inoperative, suggesting that it is not directly related to MMR. However, we cannot rule out side-pathways that lead to repair or apparent repair events, if any, which could imply that at least part of this baseline is actually biologically relevant. In fact, among our data on MMR pathway mutants, we also observed a stepwise disruption of MMR capability rather than its all-or-none involvement. Also taking into account that our assay quantifies the repair efficiency against the clock that is governed by the replication time of the particular plasmid and locus of choice, I believe a comparative usage of $\eta_s$ reported in this study is more accurate rather than reporting an absolute quantity.

Secondly, we observed a position-dependent trend in the outcome populations. At the level of substitution prevalence histograms, this effect revealed itself as shifted peak positions or unpredictable gradual changes in the spread of the subpopulations. At the repair efficiency level, this translated as a gradual monotonic decrease, whose direction was linked to the relative positioning with respect to the functional elements on the plasmid used. When reporting the results, we attempted to eliminate this effect during data processing, as observation of a similar trend in MMR deficient cells suggest that this likely is not a phenomenon directly related to the response of the MMR pathway. To date, a full understanding of this effect is still elusive.

A cell's ability to accurately propagate genetic information depends on the accuracy of DNA synthesis, as well as the efficiency of the mounted response against occasional mistakes that arise and I hope the outcomes of this study will be helpful to illuminate the properties of the latter factor. While this study was conducted on *E. coli* for simplicity, I believe the outcomes directly related to the physical characteristics of DNA will be paralleled in homologous systems as MMR is evolutionary well-conserved [79]. I also would like to note that the approach we demonstrated in this work can be extended to higher organisms such as yeast or human cell lines by choosing compatible genetic elements and hence can inform studies on cancer or even optimal design of a synthetic genomes in the future, given the redundancy in the genetic code.

## 2.5 Materials and methods

### 2.5.1 Library preparation using oligoarrays

For 3ML1, we purchased a 4007 member custom designed oligo library consisting of 132 nucleotide long oligos (Twist Biosciences, CA). Each oligo contains a primer binding site (Z6 and Z13) for amplifications, a XbaI and XhoI cut site followed by a 7 base long mapping barcode. We computationally generated a 66 base long consensus sequence representing all sequence triplets in the form of a third order De Bruijn sequence on a 4 letter alphabet. This property ensures that all nearest neighbor sequence contexts are represented once and only once, hence achieving the theoretically most efficient way of sampling in terms of sequencing requirement. Each individual member of the library differs from this consensus sequence by one base at an arbitrary position, and the mapping between the 7bp barcode (separated by at least 2 mutations from each other) and the expected mismatch position is known *a priori* as a lookup table exists by design. To ensure that the observed repair efficiency is not a pure artifact of mapping barcode difference between different library members, each individual mismatch is represented by multiple tracing barcodes ($\sim$3 distinct mapping barcodes per each mismatch for 5MLx, $\sim$20 per mismatch for 3ML1). Barcodes generating undesirable extra restriction sites were computationally discarded during the design stage to avoid truncated inserts.

For the pentamer library (5MLx), we applied the same protocol except that we purchased a DNA library comprising a total of 9828 different oligos, each 142 bases long (Genscript, NJ). The De Bruijn sequence sampling all pentamers is 1028 bases long, and it is not possible to procure a diverse DNA library consisting of oligos of this length with the current technology. We split the sequence down to 13 sub-sequences, sampling 84 out of 1028 bases with 4 base overlap at the termini. To encode the position and the type of the mismatch, we included 6-base long barcodes, each separated by at least 2 mutations in the Hamming space. By design, the library consists of 13 sub-libraries and each sub-library is flanked by a different adapter pair combination for selective amplification of the chosen sub-library. By choosing the proper primer pair during the PCR (Z13/Z14 and Z1/Z2/Z3/Z4/Z5/Z6/Z7), we can obtain the sub-library of choice and hence

each 5MLx sublibrary was prepared, transformed and analyzed separately in the workflow.

We bought the sub-sampling heptamer libraries (SSLx) from Agilent (Wilmington, DE, G7220A) and applied the same protocol, except that we could obtain a clean product without making use of emPCR, and therefore we omitted this step. We included all four sub-libraries in the same oligo pool and amplified by choosing the proper primer pair during PCR (Z13 and Z1/Z2/Z3/Z4). The full list of sequences that are part of all three oligo pools can be accessed via the supplementary materials on the associated Gitlab page: https://gitlab.com/tuncK/dissertation

### 2.5.2 Mismatch library generation

We amplified the oligo pool we received through two consecutive rounds of emulsion PCR (emPCR) following manufacturers instructions (ChimerX, catalog# 3600). Briefly, per each $50\mu l$ aqueous reaction volume and 300 $\mu l$ oil phase, we amplified 10 fmol of the original oligo library with unmodified 500nM primer Z6 and 500mM primer Z13 with Phusion polymerase (NEB, M0530L). The product was purified following manufacturers specifications, which comprise breaking up the emulsion in n-butanol, centrifugation for phase separation and DNA purification using a silicate column from the non-organic phase, and recovery of the amplified DNA with $50\mu l$ of the provided aqueous elution buffer. We re-amplified this purified product with 500nM Thio-Z13 and 500nM Phospho-Z6 primers with Phusion polymerase. While we again employed emulsion conditions for 3ML1 for this second step, we could re-amplify 5MLx or SSLx without emulsion conditions. While being highly dependent on the sequence template, a typical PCR reaction provided about 10 $\mu g$ of product per each ml of reaction volume after this second step. The thermal cycler protocol for both steps is 98°C 30s; 25x(98°C 10s; 63°C 20s; 72°C 10s); final elongation at 72°C for 2min; and the product is held indefinitely at 4°C afterwards.

We digested the phosphorylated strand by incubating each $\mu g$ of the above obtained PCR product with 2.5U $\lambda$ exonuclease (NEB, M0262L) in 100 $\mu l$ respective manufacturer-supplied 1X reaction buffer at 37°C for 1 hour followed by a 30min heat-inactivation step at 75°C. The ssDNA product was purified with a PCR purification column (Qiagen) and was thermally annealed to the respective common strand (e.g. 3ML1-C, 5ML1-C, ...; purchased from IDT) in T50 by slow cooling

on a thermal cycler from 98°C to 37°C in 1h, typically containing around $2\mu M$ of each strand. About 15pmol of this partial-duplex was extended with 1.25U Taq polymerase (NEB, M0273L) in standard Taq buffer supplemented with $200\mu M$ dNTP by incubation at 65°C for 20min. The product was purified using a PCR clean-up column (QIAGEN), and the elute was subjected to double digest by 40U SacI-HF (NEB, R3156L) and 40U XhoI (NEB, R0146L) in 1X Cutsmart buffer for 2h at 37°C. The product was purified with a PCR clean-up column and this elute is hereafter referred to as insert.

### 2.5.3 Electroporation

We ligated the above insert with a barcoded vector library (tUNC19 or its methylated derivatives) that were obtained above. A typical ligation includes about 1-2 $\mu g$ vector, 3:1 insert:vector ratio, 150kU T7 ligase (M0318L) in 1ml 1X T7 DNA ligase reaction buffer incubated for 30 min at room temperature. The product was cleaned with a PCR clean-up column and eluted with 50 $\mu l$ water. We mixed 5 $\mu l$ of this elute with 100 $\mu l$ ice-cold electrocompetent cells in a 1.5ml tube and immediately transferred into a pre-chilled electroporation cuvette with a 1mm gap width (Sigma-Aldrich, Z706078) and applied 1700V for about 5ms (Eppendorf Eporator, 4309000027). We quickly washed the cuvettes with 500 $\mu l$ SOC twice and recovered at 37°C for 1 hour in a 50ml conical tube (Corning, CLS430829). We then added 9 ml room-temperature LB, with a final concentration of 100 $\mu$g/ml ampicillin and incubated overnight with constant 250 rpm shaking at 37°C. As a quality control measure, we spread 100 $\mu l$ of this final culture on an LB-agar plate with 100 $\mu$g/ml ampicillin to confirm efficient ligation and transformation. A typical experiment yields around 50-100 colonies after overnight incubation at $37^o$C, each colony corresponding to 100 expected clans due to the dilution factor.

### 2.5.4 Next-generation sequencing

We extracted the plasmid library from a 10ml overnight LB culture using a standard Miniprep kit following the prescribed protocol (Omega E.Z.N.A. Plasmid Mini Ki I, D6942). We used 5ng of this elute as PCR template to amplify out the portion of interest using 0.5 $\mu$M of S1 and S2_4

primers with Phusion 2X mastermix. We employed 20 cycles of 10s 98°C denaturation, 20s 63°C primer annealing 10s at 72°C elongation phases preceded by additional initial denaturation at 98°C for 30s and followed by a 72°C final extension for 2 min. To cleanup the product, we incubated the product mixture with 20 $\mu$l Ampure XP beads (Beckman Coulter, A63880) for 5 min. We retained the bead-bound material after keeping for 2 minutes on a magnetic rack (GE, 1201Q46). We washed the beads twice with 200 $\mu$l 80% ethanol and eluted the material in 53 $\mu$l 10mM Tris pH8.5 by incubation for 2 min. We collected about 45-50 $\mu$l bead-free liquid 2 min after placing the material on magnetic rack.

We performed 8 additional cycles of PCR with Nextera 96-Index kit for indexing before sample pooling (Illumina, FC-131-2001), for which we used 7.5 $\mu l$ of the elute as template, 7.5$\mu$l of chosen i5 and i7 primers with 38 $\mu l$ Phusion 2X. We followed the manufacturers recommended thermal cycling protocol (95°C 3min, 98°C 30s, 55°C 30s, 72°C 30s, 72°C 5min). We also bead-purified 56$\mu l$ of this final product with 56 $\mu l$ Ampure-XP and eluted with 28 $\mu l$ 10mM Tris, pH8.5 buffer, following the same procedure described above otherwise. We pooled the final products based on their Nanodrop readings to the desired relative number density. We mostly used paired-end 2x150 cycles sequencing on Illumina HiSeq-2500 platform with 15% PhiX spiked in (Genewiz, NJ). For shallower sequencing, we used 300 cycles MiSeq v2 micro reagent kit (Illumina, MS-102-1002) to perform a paired-end sequencing for about 140-150 cycles each. While varying from sample to sample, we found a mixture of 480 $\mu l$ Hbf buffer, 100-120 $\mu l$ pooled-denatured 20 pM library and 10-20 $\mu$l 20pM PhiX control library (Illumina, FC-110-3001) to provide a reasonable spot density in general.

### 2.5.5 Calculation of repair efficiency

We retrieved the raw *.fastq output from the MiSeq/HiSeq system and parsed with a home-made program implemented in C++ and GNU Octave v5.2. The source code of the analysis toolkit can be accessed through the Gitlab page: https://gitlab.com/tuncK/public/tree/master/fixseq-codes. For each individual sample sequenced by HiSeq, we typically obtained around 50 million paired-end reads per sample, processing which requires up to 25 GB RAM and

10-20 hours of CPU wall time on a standard 24-core computation node at Maryland Advanced Research Computing Center (MARCC), parallelized via OpenMP, or 3-6 hours on a 96-core c5 elastic compute cloud (Amazon Web Services). The major steps in the course of analysis workflow are as follows:

First, all the data is imported into the memory read by read, while reads are parsed to locate the constant adapter segments using a Needleman-Wunsch algorithm with gap and mismatch penalties of -1 and match gain of +1. We take the reverse complement of the paired end reads and add to the dataset to enhance the SNR by reducing the effect of sequencing errors (Algorithm 6). Using the a priori known library prototype, we extract the segment corresponding to the clan barcode that immediately follows the end position of the adapter and the mismatch carrying library segment which follows the barcode after a 6bp gap due to the SacI restriction site (GAGCTC). If this extracted library deviates from the library prototype by more than 5 substitutions, we attempt to re-align the sequence by applying Needleman-Wunsch dynamic programming algorithm. Reads that still deviate from the prototype by more than 5 bases or do not carry a clearly identifiable adapter sequence are omitted. This step is typically performed locally on a standard personal computer after which we export the list of barcode-read pairs as a .csf file.

Second, we transfer this compressed file to a high performance computing cluster and generate a unique set of all detected barcodes on a red-black tree, during which exact duplicates are detected and recorded. To account for sequencing errors that could artificially diversify barcodes from the same clan, we introduced error tolerance by implementing a density based clustering on the set of all detected barcodes, where the minimum density threshold of N=10 should be reached within $\epsilon$=3 Hamming distance. To reduce the memory usage, only a list of neighbors is stored rather than an explicit matrix listing pairwise distances. After all barcodes are processed, barcodes failing the density criterion (noise) are discarded, and the core-points together with all neighbors are reported as clans. Interested readers are referred to [32] for the description of this DBSCAN algorithm.

Third, the base distribution frequency in each clan is evaluated individually, as a 4xlibrary_length matrix per clan. For the purposes of the further analyses, each such matrix is considered as one data point. The output for each sample is hence a 3D tensor, which is out-

putted into a *.hist obeying the *.mat file format.

Fourth, we import this *.hist file to GNU Octave v5.2 to interpret the results for each mismatch in the input library individually. To attribute a particular clan to the mismatch that its ancestor plasmid was carrying, we make use of the mapping barcodes. Out of the base frequency histogram, we use the initial 6 to 8 rows to extract the maximum voted base combination for the mapping barcode. Using these ensemble-averaged mapping barcodes of each clan, we check the pre-determined lookup table. We omit clans whose mapping barcodes do not have an exact hit, as the mapping barcodes are separated from each other by multiple mismatches making conversion of one barcode into another unlikely.

Finally, we process the list of 4xlength matrices clan by clan to count the number of unrepaired clans ($U_{ij}$) or repaired clans using the common ($C_{ij}$) or the variable strand ($V_{ij}$) as the correct information source by the cell. Knowing these three variables for each mismatch position i and substituted base type j, both the repair efficiency ($\eta_{ij}$) and strand choice bias ($\beta_{ij}$) can be calculated according to Equations 2.4 and 2.5.

### 2.5.6 Theoretical assessment of sampling uncertainty

Collection of a higher number of clans corresponding to the same mismatch will increase the precision and accuracy of the reported quantities. The uncertainty on the repair efficiency ($\delta\eta_{ij}$) that we report here will be inversely proportional to the absolute counts of the 3 clan types that could be detected. To be more precise,

$$\delta\eta_{ij} = \frac{-(C_{ij}+V_{ij})\delta U_{ij} + U_{ij}\delta C_{ij} + U_{ij}\delta V_{ij}}{(C_{ij}+U_{ij}+V_{ij})^2} \tag{2.17}$$

As expected, the uncertainty in $\eta$ declines as the number of measured number of clans of either type increases ($\lim_{C_{ij}\to\infty} \delta\eta_{ij} = \lim_{U_{ij}\to\infty} \delta\eta_{ij} = \lim_{V_{ij}\to\infty} \delta\eta_{ij} = 0$). I would like to note that the counts of these three types of clans are not independent due to the constraint on the total number of oligos obtained carrying the same mismatch in the starting pool. Let this starting material contain a total of $c_{ij}$ many mismatch carrying plasmids. Then, $C_{ij} + U_{ij} + V_{ij} = c_{ij}$, differentiating which one finds that $\delta C_{ij} + \delta U_{ij} + \delta V_{ij} = \delta c_{ij}$.

**Algorithm 6:** Extraction of clans from the raw NGS output, coded as a C++ program and typically requires high-performance computing.

```
forwardReads ← Import all forward reads from file "sample#_R1.fastq"
reverseReads ← Import all reverse reads from file "sample#_R2.fastq"
foreach read ∈ reverseReads do
    read ← reverse-complement(read);
end
Fix minLength = Total length of tracing_barcode, SacI site, mapping barcode and mismatch_library
Initialize acceptedReads = ∅
foreach full_read ∈ Union(forwardReads, reverseReads) do
    AdaptorEndPos ← Search the last base of the 5' adapter in full_read
    if full_read is shorter than minLength OR Adaptor not found OR AdaptorEndPos ∉[15,30] then
        Ignore the full_read
    end
    SeqOfInterest ← Extract the n-base long sub-sequence of full_read following AdaptorEndPos
    if HammingDistance(sequenceOfInterest, expectedLibraryPrototype) >5 then
        SeqOfInterest ← Shift SeqOfInterest by dynamic programming
    end
    if HammingDistance(SeqeOfInterest, expectedLibraryPrototype) ≤ 5 then
        Extract (tracing_barcode, mismatch_probe) from SeqOfInterest using the expected library
          prototype
        Append (tracing_barcode, mismatch_probe) to acceptedReads
    else
        Ignore the full_read
    end
    Export acceptedReads into sample#.csf file
end

Import acceptedReads from sample#.csf file
barcodeClusters ← DBSCAN(acceptedReads.tracing_barcode, ϵ = 3, N = 10)
all_clans ← Group acceptedReads.mismatch_probe w.r.t. barcodeClusters
clan_histograms ← #clans x ||SeqOfInterest|| x 4 matrix containing base composition histograms of
  all_clans
Return clan_histograms
```

**Algorithm 7:** Calculation of the repair efficiency from clans. Coded as an Octave script and executed on a standard personal computer.

```
Import clan_histograms
foreach clan ∈ all_clans do
    if clan.mapping_barcode is invalid then
        clan.MMTypeID = NULL
    else
        (MMbase, MMpos) ← Refer to the database for the clan.mapping_barcode
        clan.MMTypeID ← (MMbase, MMpos)
    end
end

foreach (MMbase, MMpos) do
    relevant_clans ← {clan ∈ all_clans | clan.MMTypeID == (MMbase, MMpos)}
    substitution_prevalences = ∅
    foreach clan ∈ relevant_clans do
        s ← Evaluate %reads ∈ clan containing MMbase at MMpos
        Append s to substitution_prevalences
    end
    Append substitution_prevalences to all_substitution_prevalences
    substitution_prevalence_hist ← Build a normalized histogram of substitution_prevalences
    (a, b, c, d) ← Curve fitting on substitution_prevalence_hist using Eq. 2.2
    if d < 0.05 then
        Append c to V_peak_positions(MMpos)
    end
end

Fitted_peak_pos ← Polynomial fitting on V_peak_positions
UV_boundary[MMpos] ← Fitted_peak_pos[MMpos] − 0.15
Initialize ||SeqOfInterest||x4 matrices Fcounts, Ucounts, Vcounts to 0
foreach (MMbase,MMpos) do
    relevant_clans ← {clan ∈ all_clans | clan.MMTypeID == (MMbase, MMpos)}
    relevant_prevalences ← all_substitution_prevalences(relevant_clans)
    Ccounts[MMbase, MMpos] = m({s ∈ relevant_prevalences ∩ [0, 0.1] })
    Ucounts[MMbase, MMpos] = m({s ∈ relevant_prevalences ∩ (0.1, UV_boundary[MMpos]) })
    Vcounts[MMbase, MMpos] = m({s ∈ relevant_prevalences ∩ [UV_boundary[MMpos], 1] })
end

Report totalClanCount_ij ← Ccounts_ij + Ucounts_ij + Vcounts_ij
Report η_ij ← (Ccounts_ij + Vcounts_ij)/(Ccounts_ij + Ucounts_ij + Vcounts_ij)
Report β_ij ← (Ccounts_ij − Vcounts_ij)/(Ccounts_ij + Vcounts_ij)
```

We then obtain for the uncertainty in $\eta_{ij}$,

$$\delta\eta_{ij} = \frac{(C_{ij} + U_{ij} + V_{ij})(\delta V_{ij} + \delta C_{ij}) - (C_{ij} + V_{ij})\delta c_{ij}}{(C_{ij} + U_{ij} + V_{ij})^2} \tag{2.18}$$

which, using Equation 2.17, we can rearrange into,

$$\delta\eta_{ij} = \frac{\delta C_{ij} + \delta V_{ij} - \eta_{ij}\delta c_{ij}}{C_{ij} + U_{ij} + V_{ij}} \tag{2.19}$$

If we further assume that the errors in these three parameters are independent, we reach,

$$\Delta\eta_{ij} = \frac{\sqrt{(\Delta C_{ij})^2 + (\Delta V_{ij})^2 + \eta_{ij}^2(\Delta c_{ij})^2}}{C_{ij} + U_{ij} + V_{ij}} \tag{2.20}$$

Under the additional assumption that the mismatch library elements are similar to each other, $\delta c_{ij} = \sigma_c$ holds for the frequency distribution of all elements of the starting library, and that the uncertainties in the counts of two repaired clan counts are Poisson distributed, it will then hold that $\delta x \sim \sigma_x = \sqrt{x}$ and we will get,

$$\sigma_{\eta_{ij}} \approx \frac{\sqrt{C_{ij} + V_{ij} + \eta_{ij}^2\sigma_c^2}}{C_{ij} + U_{ij} + V_{ij}} \tag{2.21}$$

### 2.5.7 Recombineering

To obtain both methylase and an MMR member deficient cells, we replaced the chromosomal copy of the *dam* gene with a chloramphenicol resistance (camR) cassette via $\lambda$ red mediated homologous recombination in each MMR-deficient cell strain that we had procured ($\Delta mutS$, $\Delta mutL$, $\Delta mutH$, or $\Delta uvrD$) [80].

We generated camR containing dsDNA fragments to serve as substrates of recombination by a PCR where the overhangs of R1p and R2p primers are homologous to the up- and downstream loci of the dam gene whereas 3' termini bind to the pGGAselect vector (NEB, N0309). A total of 400 $\mu l$ PCR mix containing Q5 hot start polymerase (NEB, M0494S), 5ng pGGA plasmid and 500nM R1p and R2p primers was subjected to the following thermal cycling protocol: 98°C 30s, 30x(98°C

10s, 68°C 20s, 72°C 25s), and a 2 minute final extension at 72°C. The product was purified with a PCR cleanup column and the carry-over template plasmid was digested by 20U DpnI in 1X Cutsmart buffer (NEB, B7204S) at 37°C for 1 hour, followed by a 30 min denaturation step at 65°C. The undigested DNA was again column purified and used for recombination.

To trigger recombination, we obtained pREDTAI (Addgene, #51627) which provides the three required components of the $\lambda$ red system, namely $\gamma$ (a RecBCD inhibitor), $\beta$ (an ssDNA binding protein) and *exo* (a 5'→3' dsDNA exonuclease). We purified the pREDTAI plasmid from the procured cell strain and transformed 2ng pREDTAI into 100 $\mu l$ electrocompetent ΔMMR cells. We recovered the transformants in 1ml SOC for 1 hour at 30°C, of which we plated 100 $\mu l$ on LB-agar plates containing 100 $\mu g/ml$ ampicillin. After an overnight incubation at 30°C, we selected single colonies to inoculate liquid LB cultures, which were grown overnight at 30°C.

Out of these ΔMMR and pREDTAI containing cells, we obtained electrocompetent cells by inoculating 200ml LB supplemented with 2% (w/v) L-arabinose for induction, 50$\mu g/ml$ ampicillin, 25 $\mu g/ml$ kanamycin and 2 ml overnight culture of the respective strain. In about 6-10 hours at 30°C, the culture reached OD600∼0.5, at which point the cells were harvested by centrifugation in 50ml conical tubes for 10 min at 2000g. The cells were washed thrice with 10% filter-sterilized glycerol solution, after which the final pellet was re-suspended in 2ml 10% glycerol, aliquoted, flash frozen and stored at -80°C until use.

We electroporated about 200 ng of the above-mentioned PCR product containing the camR cassette into $\lambda$ system expressing electrocompetent cells and recovered in 1 ml SOC for 2 hours at 30°C with constant agitation. After addition of 9ml LB supplemented with 35$\mu g/ml$ chloramphenicol and 50$\mu g/ml$ kanamycin, the resulting 10ml liquid culture was further incubated overnight at 30°C. This confluent overnight culture was streaked on an agar plate containing 17$\mu g/ml$ chloramphenicol and 25$\mu g/ml$ kanamycin. After an overnight growth at 37°C, single colonies were selected for liquid culture in LB with 35$\mu g/ml$ chloramphenicol and 50$\mu g/ml$ kanamycin.

After an overnight incubation at 37°C with 200 rpm constant agitation, the *dam* locus was PCR amplified using Q5 polymerase and 500nM of "Dam Verify Fwd" and "Dam Verify Rev" primers, via the following thermal cycling protocol: 98°C 60s, 30x(98°C 10s, 63°C 20s, 72°C 30s),

and a 2 minute final extension at 72°C. The success of the recombination was verified by Sanger sequencing of this amplicon separately again using the "Dam Verify" primers. To rule out possible contamination of the cell cultures during this multi-step procedure, the identity of the MMR mutation expected based on the starting cell strain was also independently verified by Sanger sequencing of the PCR amplicons of the *mutS, mutL, mutH* and *uvrD* loci, using the relevant forward/reverse primer pair and Q5 polymerase obeying a similar thermal cycling protocol. The glycerol stocks of these verified double-mutant cell strains were used to produce electrocompetent cells following the standard experimental protocol described above.

## 2.6 Figures and tables

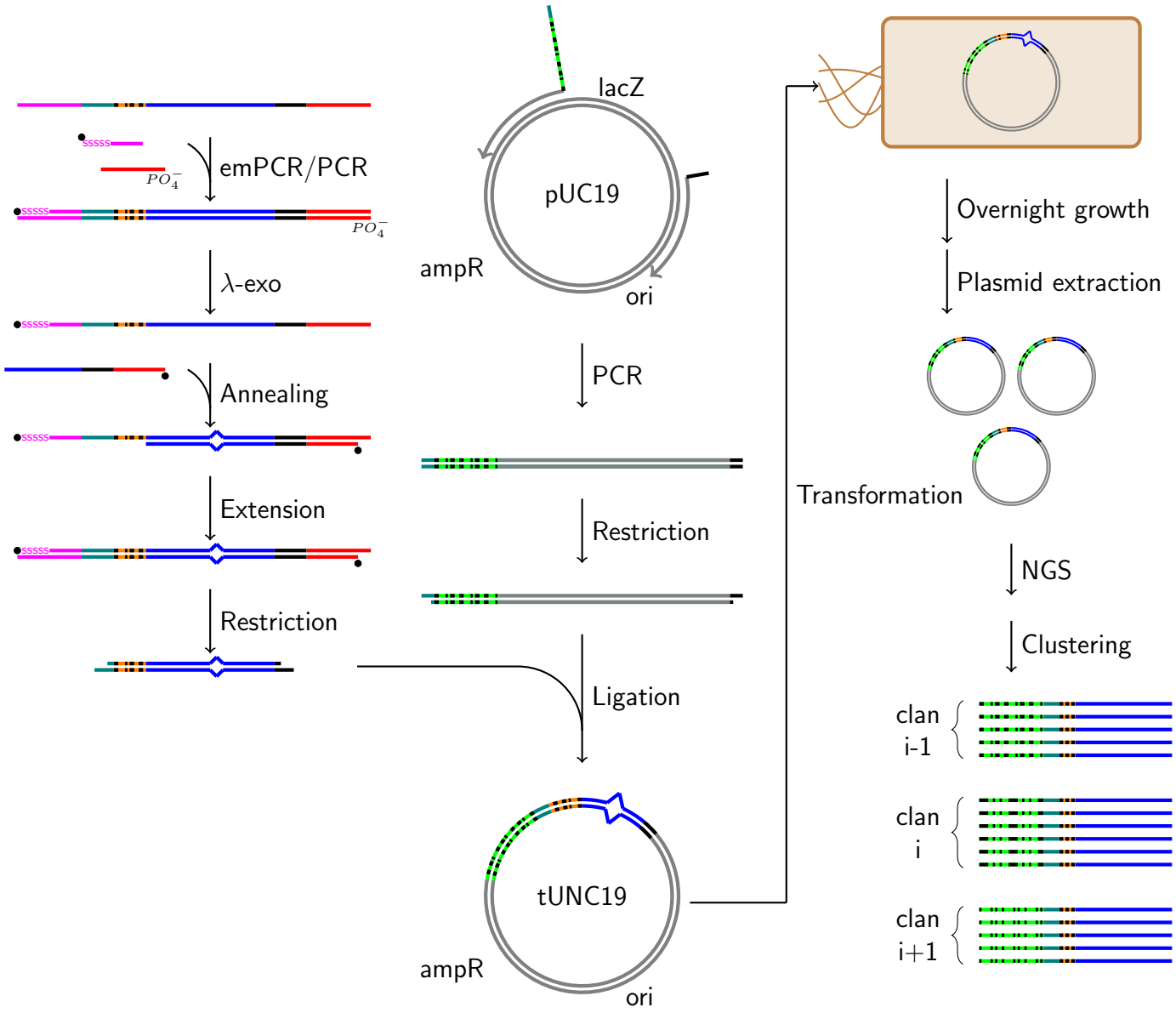



**Figure 2.1:** Schematic summary of the experimental design. An array synthesized oligo library was purchased and amplified with modified primers conferring an ability to selectively digest one of the strands. The hence obtained ssDNA library is annealed to an oligo of common sequence forming a mismatch library. The mapping barcode information is copied to the other strand by primer extension. DNA barcodes were introduced to pUC19 plasmid via a PCR reaction with primers containing a random base tail that uniquely labels each individual plasmid. The resulting linear PCR product was ligated to the mismatch library after generating sticky ends with a restriction double digest reaction. The plasmid library with mismatches was transformed into *E. coli*. After multiple rounds of replication, the extracted plasmid library is sequenced. Clustering of the reads with respect to the tracing barcodes segregate DNA sequences originating from the same original plasmid, whereas the mapping barcodes are used as a lookup table for the mismatch type. The homogeneity or heterogeneity of the clans provide a means to repaired and unrepaired plasmids, respectively.

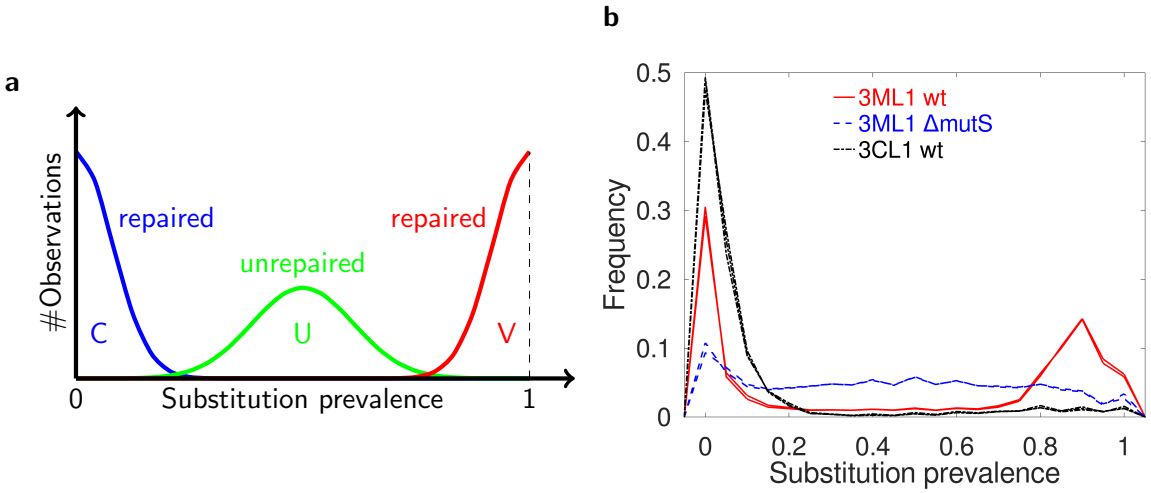



**Figure 2.2:** The observed substitution frequency histogram depends on the presence of mismatches in the input library and the avalability of the cellular MMR response. (a) We expect 3 distinct subpopulations in histograms for C-U-V type clans. (b) Transformation of a true mismatch library (3ML1, —) to wt cells produces all three C-U-V populations, whereas the same input library generates dominantly U-type clans in ΔmutS cells (- - -). In wt cells, 3CL1 control library consisting of two fully annealed strands without forming mismatches dominantly generates C-type clans as opposed to 3ML1 (···). Each dataset is represented by two curves corresponding to two independent biological replicates.

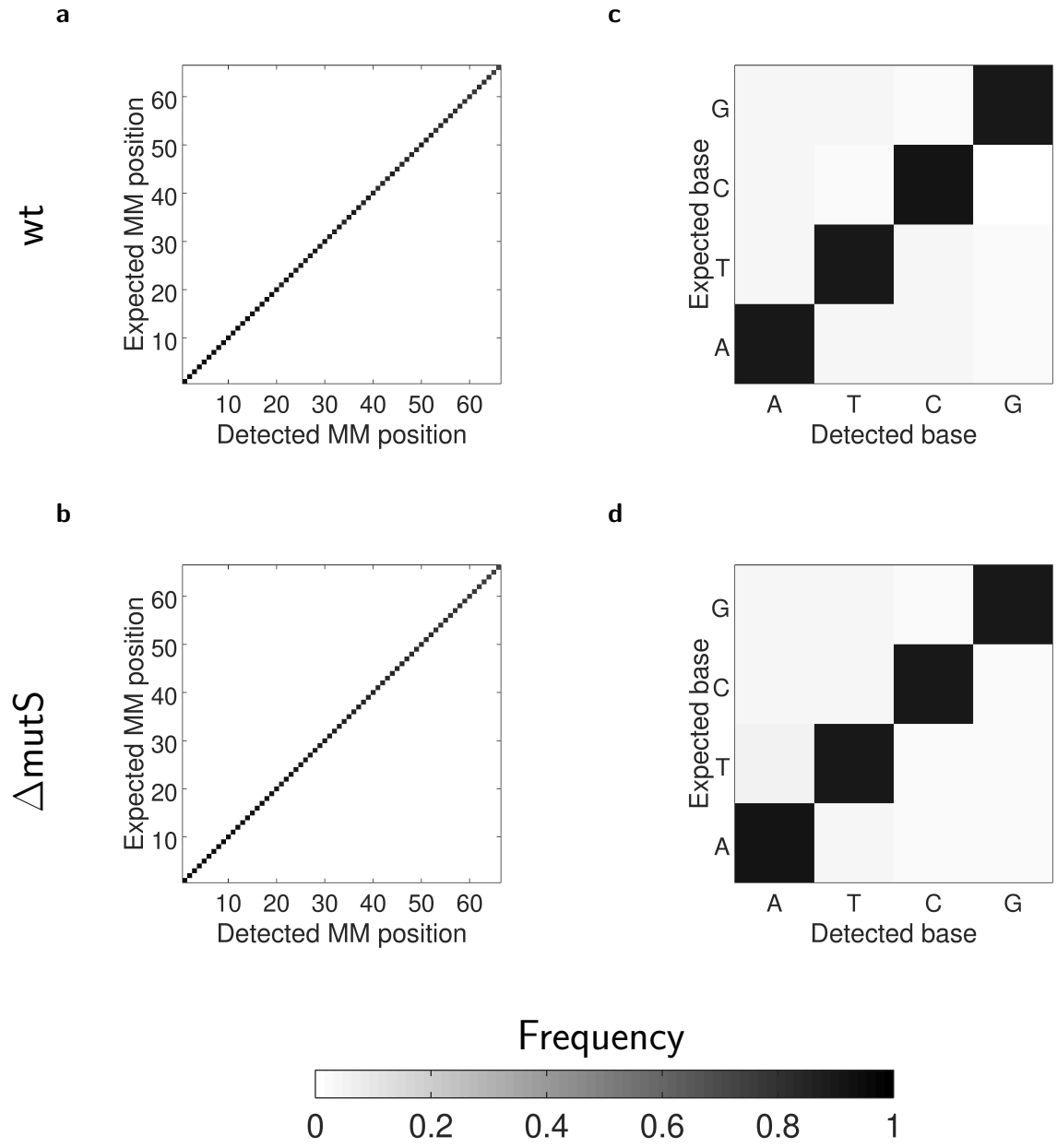



**Figure 2.3:** Confusion matrices comparing the mismatch position (a, b) and mismatch type (c, d) conclusions reached using the mapping barcodes of 3ML1 library (y-axis) with those deduced using the position and base identity of the maximum substituted base element within the clan (x-axis). For both wt (a, c) and $\Delta mutS$ (b, d) cells, clans whose substition prevalence were less than <0.15 were omitted due to the low confidence level of deducing the original mismatch identity by the histogram method. Each clan is recorded as one event for its respective entry, and each row of the matrices is normalised to 1, such that entries along a row yield a probability of observing the experimental outcome, given the input material. Higher gray values indicate higher frequency of observing a particular event, and an ideal system with negligible synthesis, amplification and sequencing errors is expected to yield a diagonal matrix.

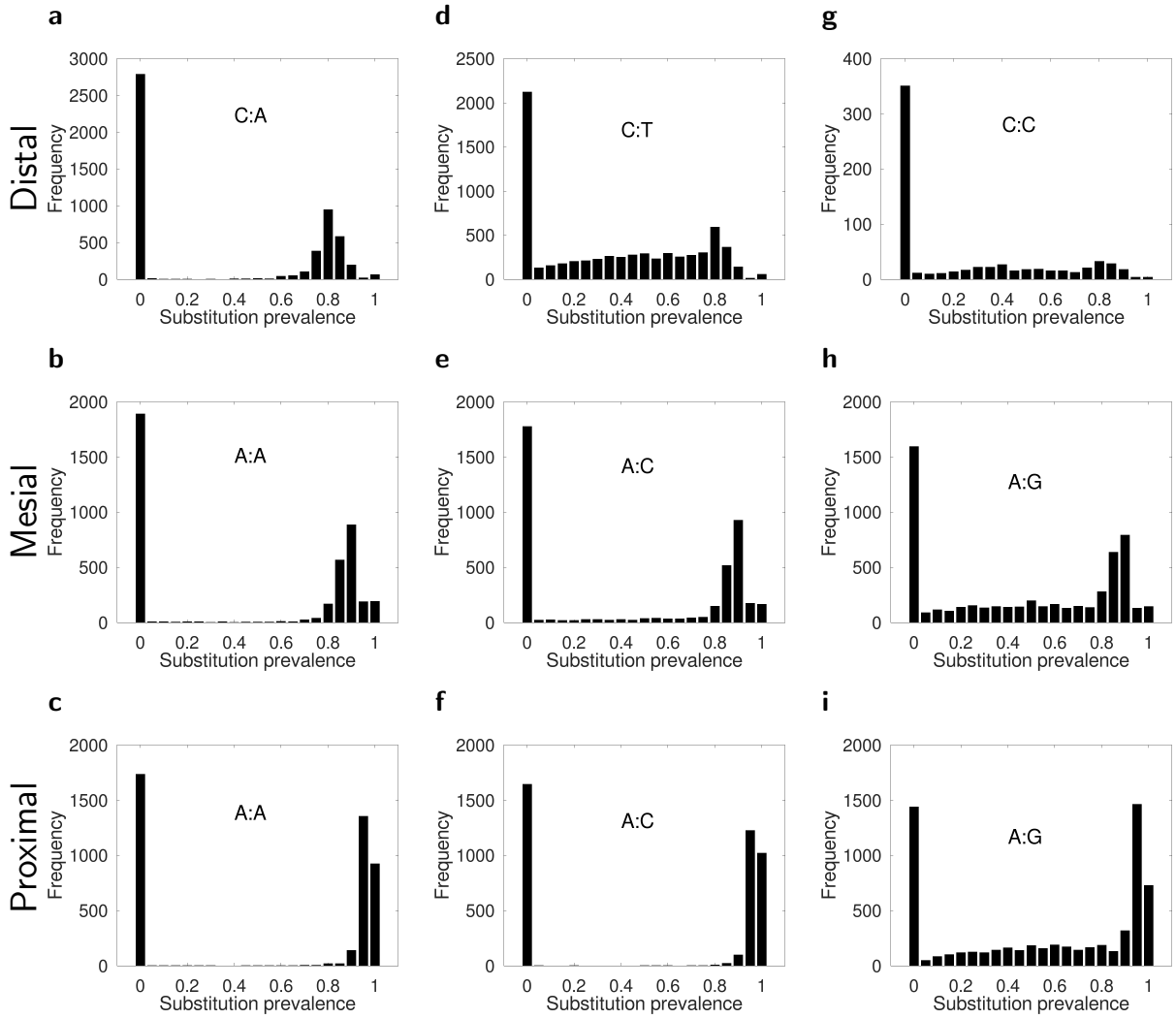



**Figure 2.4:** Example substitution prevalence histograms of mismatches sampled. Clan substitution prevalence histograms for 9 different mismatches at the proximal end of the mismatch library with respect to the tracing barcode (c,f,i, base position 5), around the middle of the library (b,e,h, base position 35) and at the distal end from the barcode (a,d,g, base position 62) measured in wt cells using 3ML1. The identity of the mispaired nucleotides is indicated in each plot.

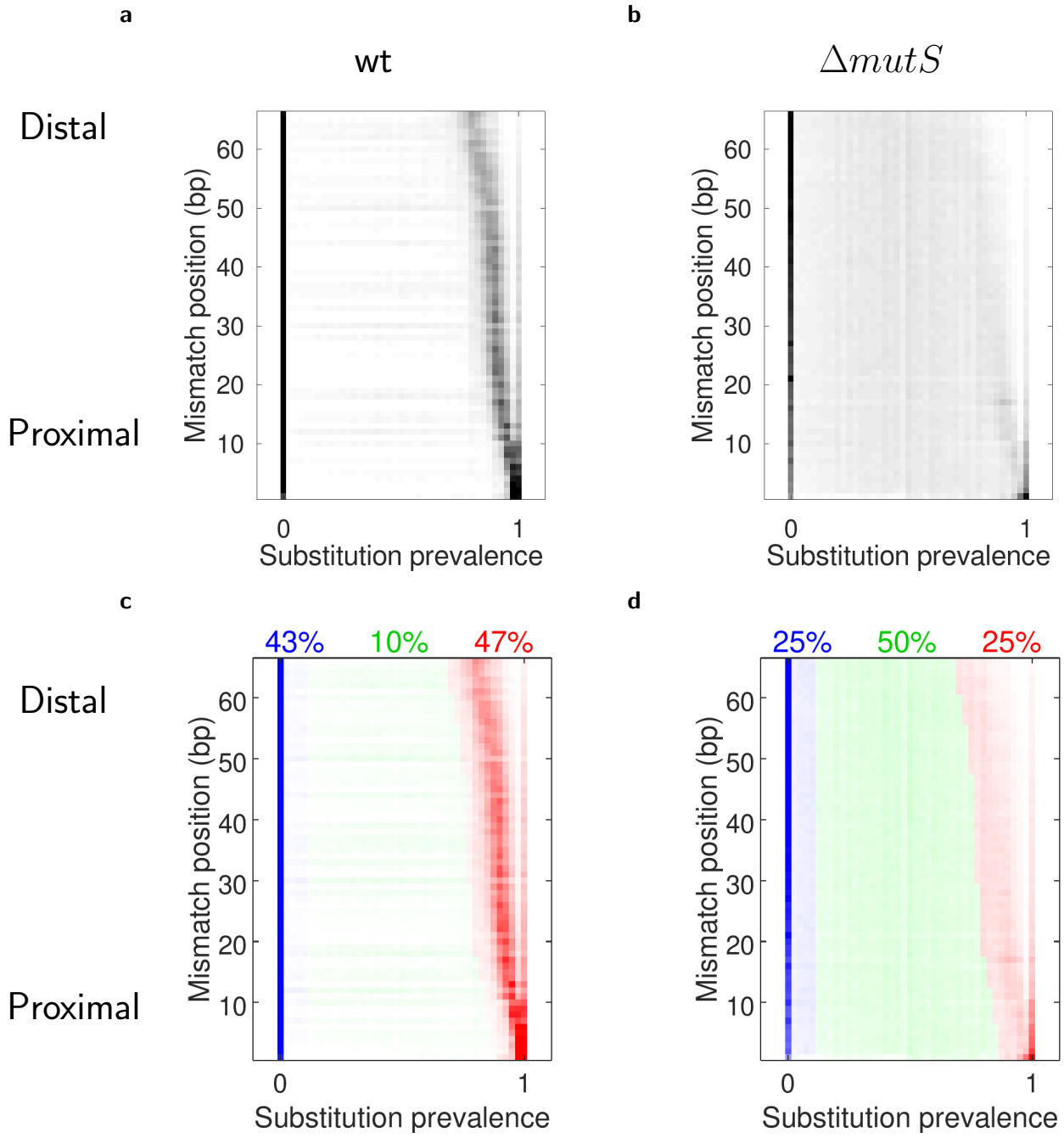



**Figure 2.5:** (a-b) Combined histograms of substitution prevalences of clans along 3ML1 library in wt and ΔmutS cells. Each row of the histograms represents one mismatch position where the counts for three mismatch possibilities at the same position were cumulated into a single row. (c-d) Same histograms, but after the clans are classified as C, U, V type as described in Section 3.4.2.

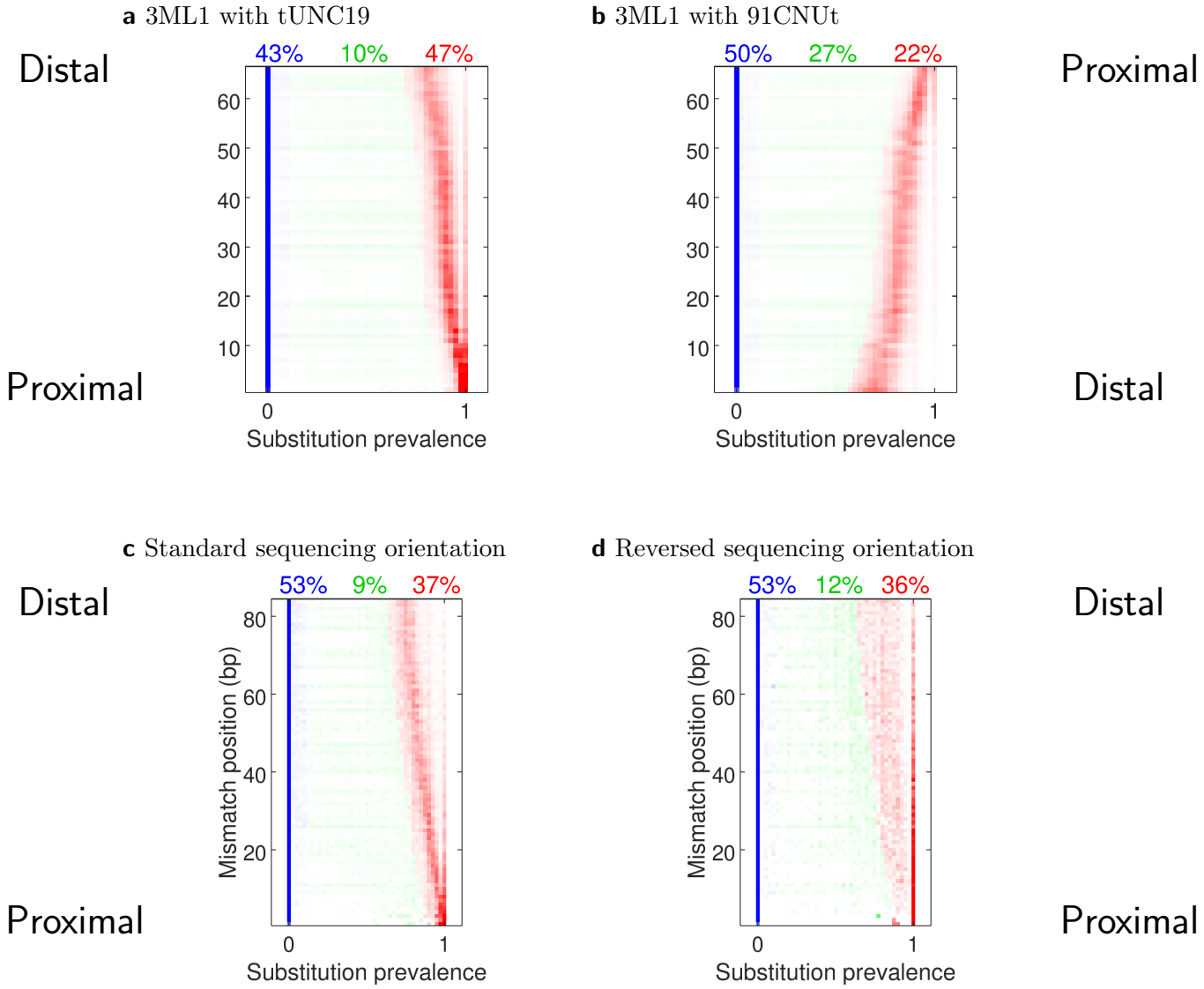



**Figure 2.6:** The effect of insertion or sequencing orientations on s-histograms. (a,b) Clan substitution prevalence histograms for the 3ML1 library, measured using the standard insertion orientation ("TTT-end" proximal to the barcode) vs. inverted insertion orientation to the plasmid ("TTT-end" distal to the barcode). The inversion of the insertion orientation leads to an inversion in the high-substitution peak shift in wt cells. (c,d) Clan substitution prevalence histograms for the 5ML1 library, same sample was sequenced either using the standard sequencing order (the variable strand is read first, then the common strand; c) or in the opposite order (the common strand is read first, then the variable strand; d). A similar peak shift is observed in both cases, suggesting that this is not a pure artifact of an gradual increase in sequencing errors in the course of the sequencing process. All data were obtained in wt cells.

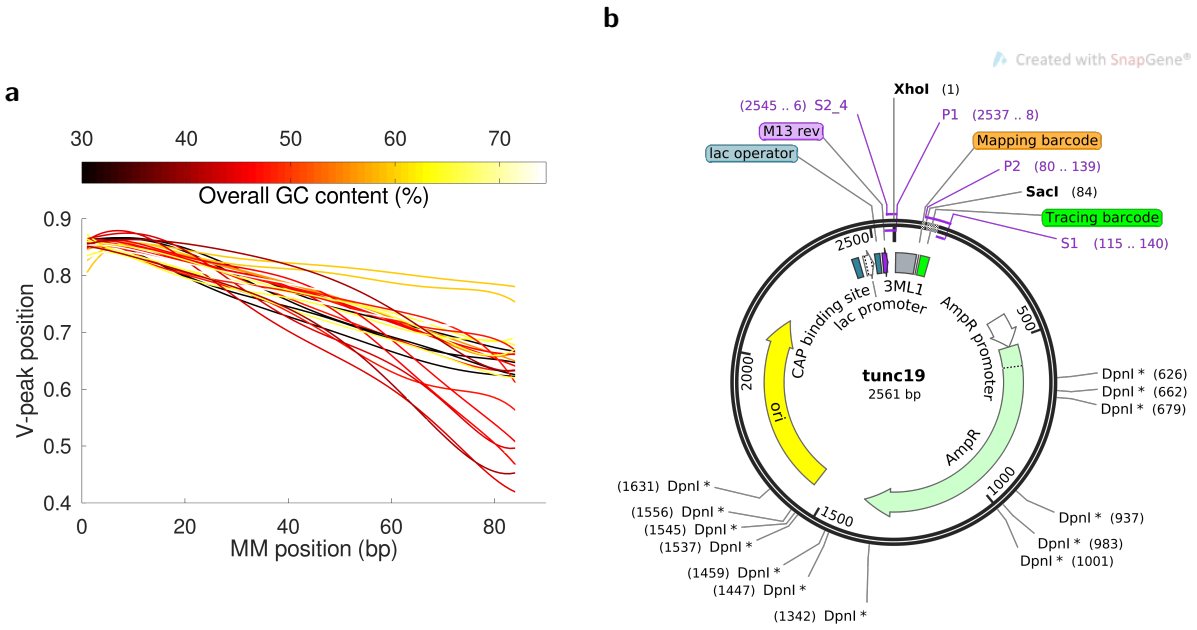



**Figure 2.7:** The effect of the GC content and the plasmid layout on the position dependence. (a) The position of the V-clan peak (i.e. the fitting parameter p in Equation 2.2) as a function of the position along the mismatch library and overall GC content of the 5MLx sublibrary. (b) Cartoon representation of the tUNC19 barcoded vector, here shown with 3ML1 library inserted. The positions of MutH attack sites (GATC) are the same as DpnI.

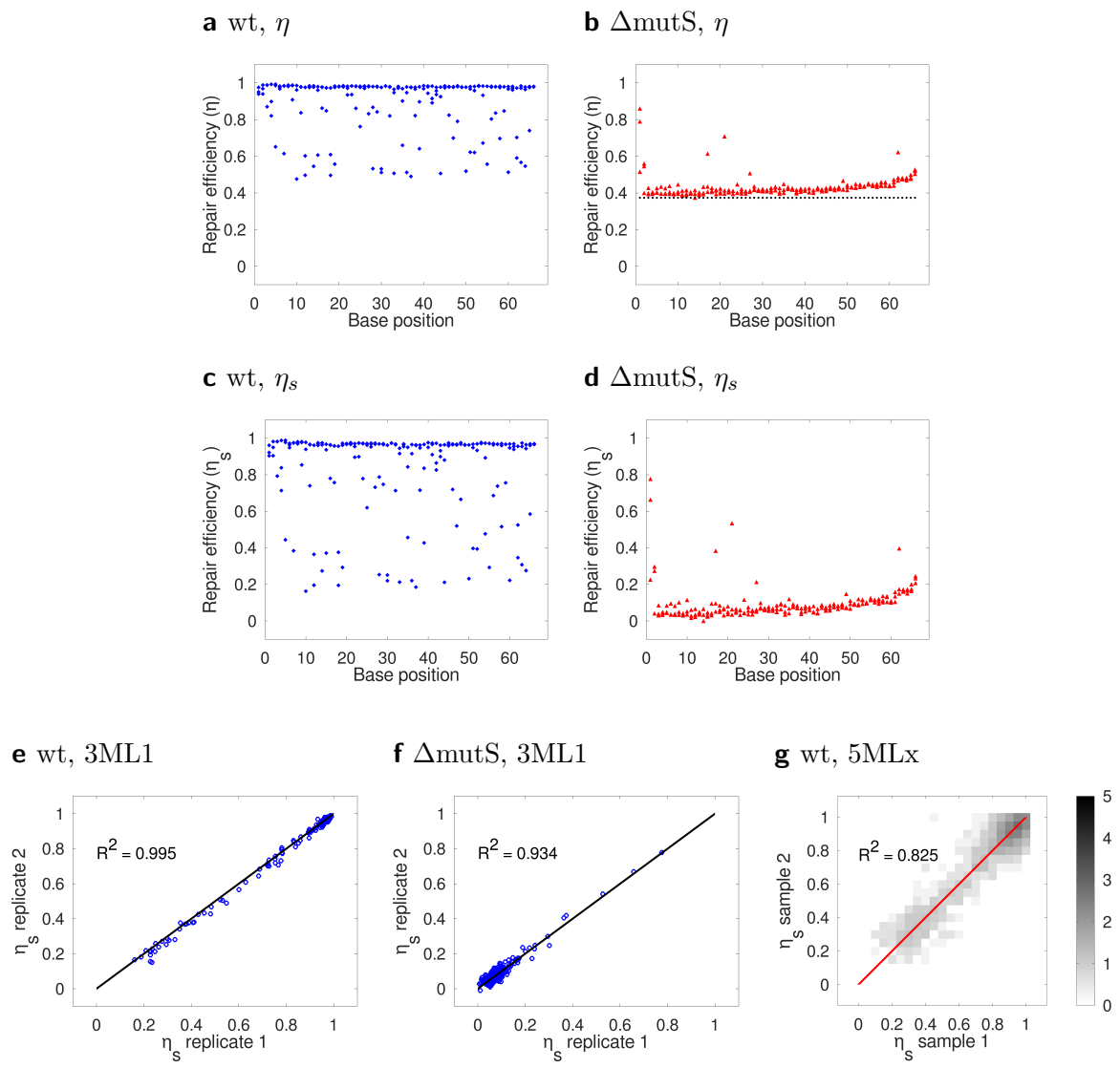



**Figure 2.8:** Scaled repair efficiency as a reproducible measure. (a-b) The raw repair efficiencies have a high background level attributable both to sampling errors and potential side pathways. (c-d) After scaling the raw repair efficiencies based on the lowest measurable value in ΔmutS cells (0.37, $\cdots$ in Figure b). (e-g) Comparison of the repair efficiencies ($\eta$) measured using unmethylated plasmids carrying 3ML1 mismatch library reported as the mean of two biological replicates, independently treated starting from the ligation step. Each blue marker (**o**) represents one mismatch sampled within the 3ML1 library, regardless of the position or type of the mismatch. The solid black line (—) indicates the x=y diagonal, on which the data points are expected to lie if the experiment is fully reproducible. $R^2(x,y) = 1 - \langle (x-y)^2 \rangle / \sigma_x \sigma_y$ indicates the coefficient of determination.

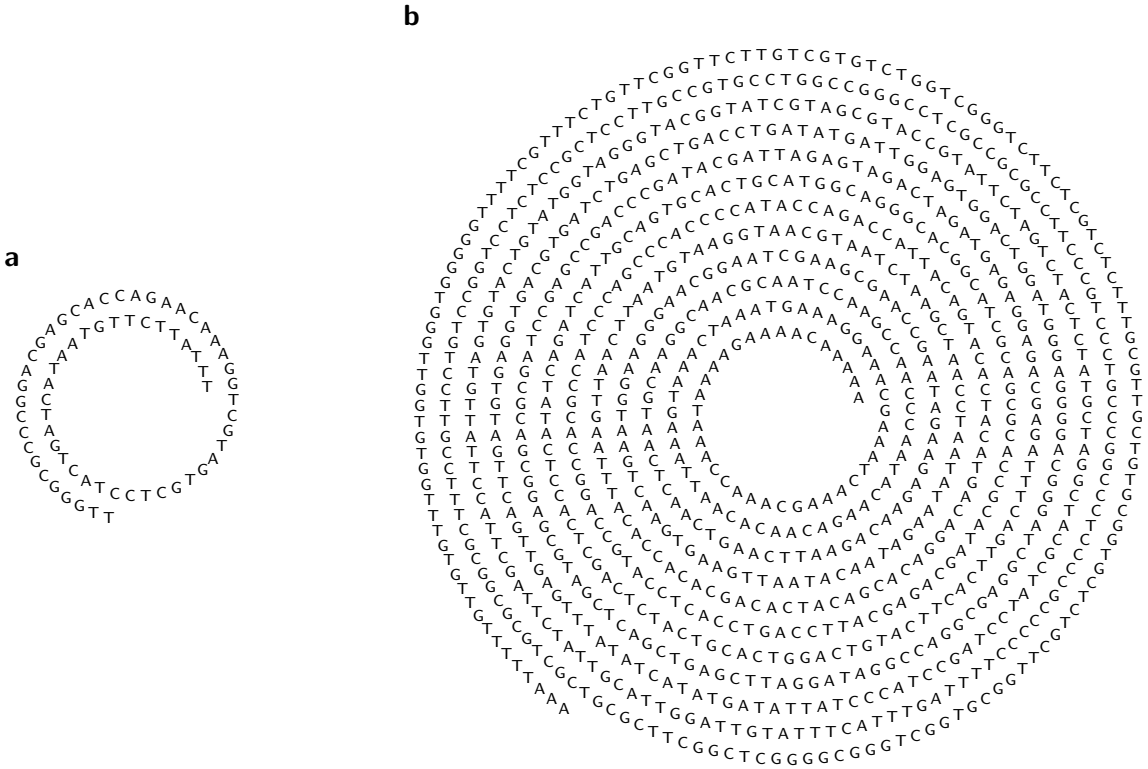



**Figure 2.9:** De Bruijn sequences containing all sequence trimers (k=3, a) and all sequence pentamers (k=5, b), brought into the linear form.

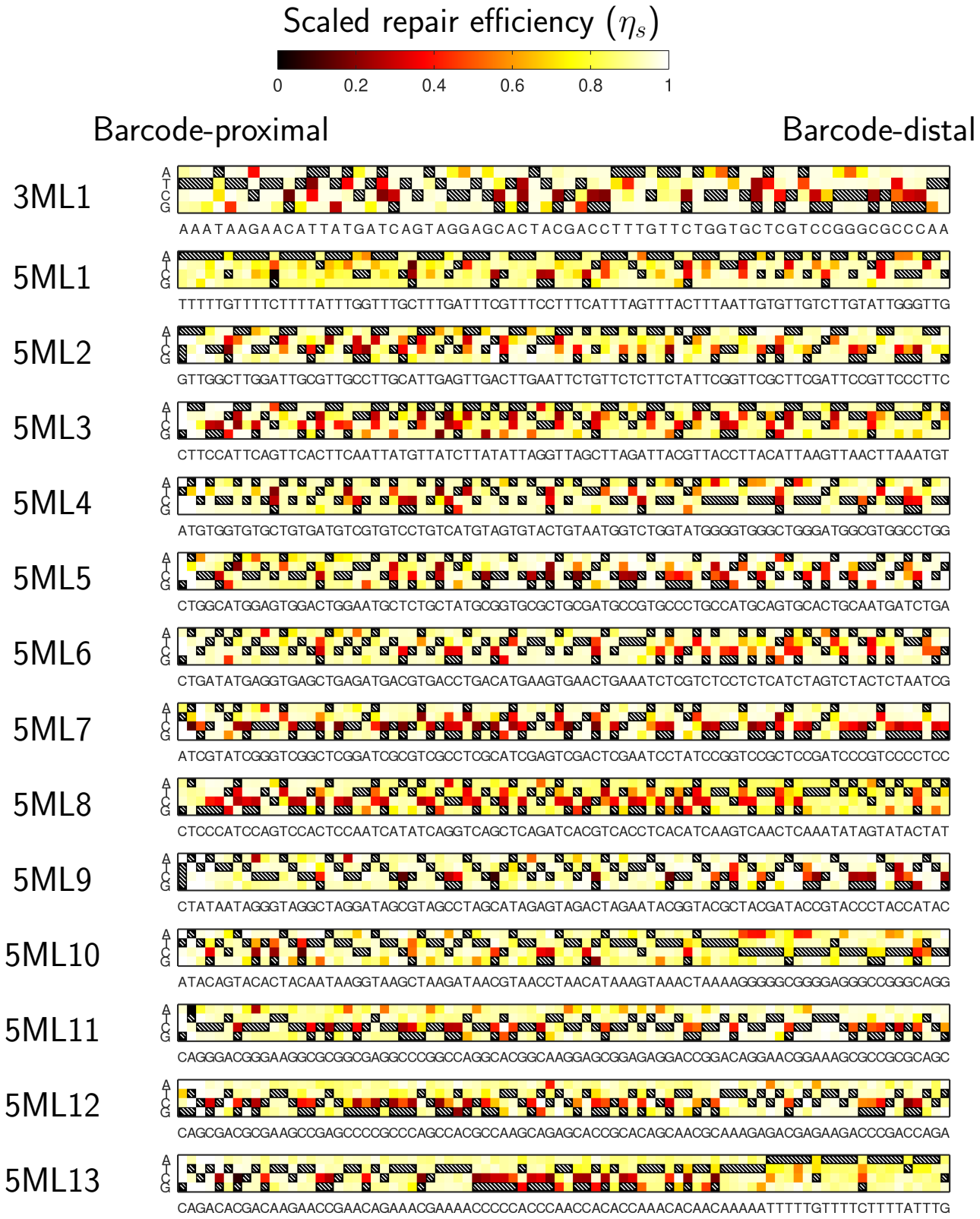

Scaled repair efficiency ($\eta_s$)
0
0.2
0.4
0.6
0.8
1
Barcode-proximal
Barcode-distal
3ML1
AAATAAGAACATTATGATCAGTAGGAGCACTACGACCTTTGTTCTGGTGCTCGTCCGGGCGCCCAA
5ML1
TTTTTGTTTTCTTTTATTTGGTTTGCTTTGATTTCGTTTCCTTTCATTTAGTTTACTTTAATTGTGTTGTCTTGTATTGGGTTG
5ML2
GTTGGCTTGGATTGCGTTGCCTTGCATTGAGTTGACTTGAATTCTGTTCTCTTCTATTCGGTTCGCTTCGATTCCGTTCCCTTC
5ML3
CTTCCATTCAGTTCACTTCAATTATGTTATCTTATATTAGGTTAGCTTAGATTACGTTACCTTACATTAAGTTAACTTAAATGT
5ML4
ATGTGGTGTGCTGTGATGTCGTGTCCTGTCATGTAGTGTACTGTAATGGTCTGGTATGGGGTGGGCTGGGATGGCGTGGCCTGG
5ML5
CTGGCATGGAGTGGACTGGAATGCTCTGCTATGCGGTGCGCTGCGATGCCGTGCCCTGCCATGCAGTGCACTGCAATGATCTGA
5ML6
CTGATATGAGGTGAGCTGAGATGACGTGACCTGACATGAAGTGAACTGAAATCTCGTCTCCTCTCATCTAGTCTACTCTAATCG
5ML7
ATCGTATCGGGTCGGCTCGGATCGCGTCGCCTCGCATCGAGTCGACTCGAATCCTATCCGGTCCGCTCCGATCCCGTCCCCTCC
5ML8
CTCCCATCCAGTCCACTCCAATCATATCAGGTCAGCTCAGATCACGTCACCTCACATCAAGTCAACTCAAATATAGTATACTAT
5ML9
CTATAATAGGGTAGGCTAGGATAGCGTAGCCTAGCATAGAGTAGACTAGAATACGGTACGCTACGATACCGTACCCTACCATAC
5ML10
ATACAGTACACTACAATAAGGTAAGCTAAGATAACGTAACCTAACATAAAGTAAACTAAAAGGGGGCGGGGAGGGCCGGGCAGG
5ML11
CAGGGACGGGAAGGCGCGGCGAGGCCCGGCCAGGCACGGCAAGGAGCGGAGAGGACCGGACAGGAACGGAAAGCGCCGCGCAGC
5ML12
CAGCGACGCGAAGCCGAGCCCCGCCCAGCCACGCCAAGCAGAGCACCGCACAGCAACGCAAAGAGACGAGAAGACCCGACCAGA
5ML13
CAGACACGACAAGAACCGAACAGAAACGAAAACCCCCACCCAACCACACCAAACACAACAAAAATTTTTGTTTTCTTTTATTTG

**Figure 2.10:** Scaled repair efficiencies ($\eta_s$) measured for each individual mismatch as part of a mismatch library containing all sequence triplets (3ML1), and 13 sublibraries collectively covering all pentamer sequence contexts (5MLx). The sequences mentioned along the x-axis show the strand of the common sequence shared among all library elements. For each matrix entry, the variable strand differs from the complementary of this sequence at the x-position of the element (A, C, G or T as indicated along the y-axis) and forces the formation of a mispair with the common strand. Black-white hatching patterns (▨) represent proper Watson-Crick base pairing not leading to a mismatch to be repaired. All data pertain to wt cells, extracted out of aggregated output of all relevant experimental replicates.

$$5' - SPNQR - 3'$$
$$3' - \bar{S}\bar{P}M\bar{Q}\bar{R} - 5'$$

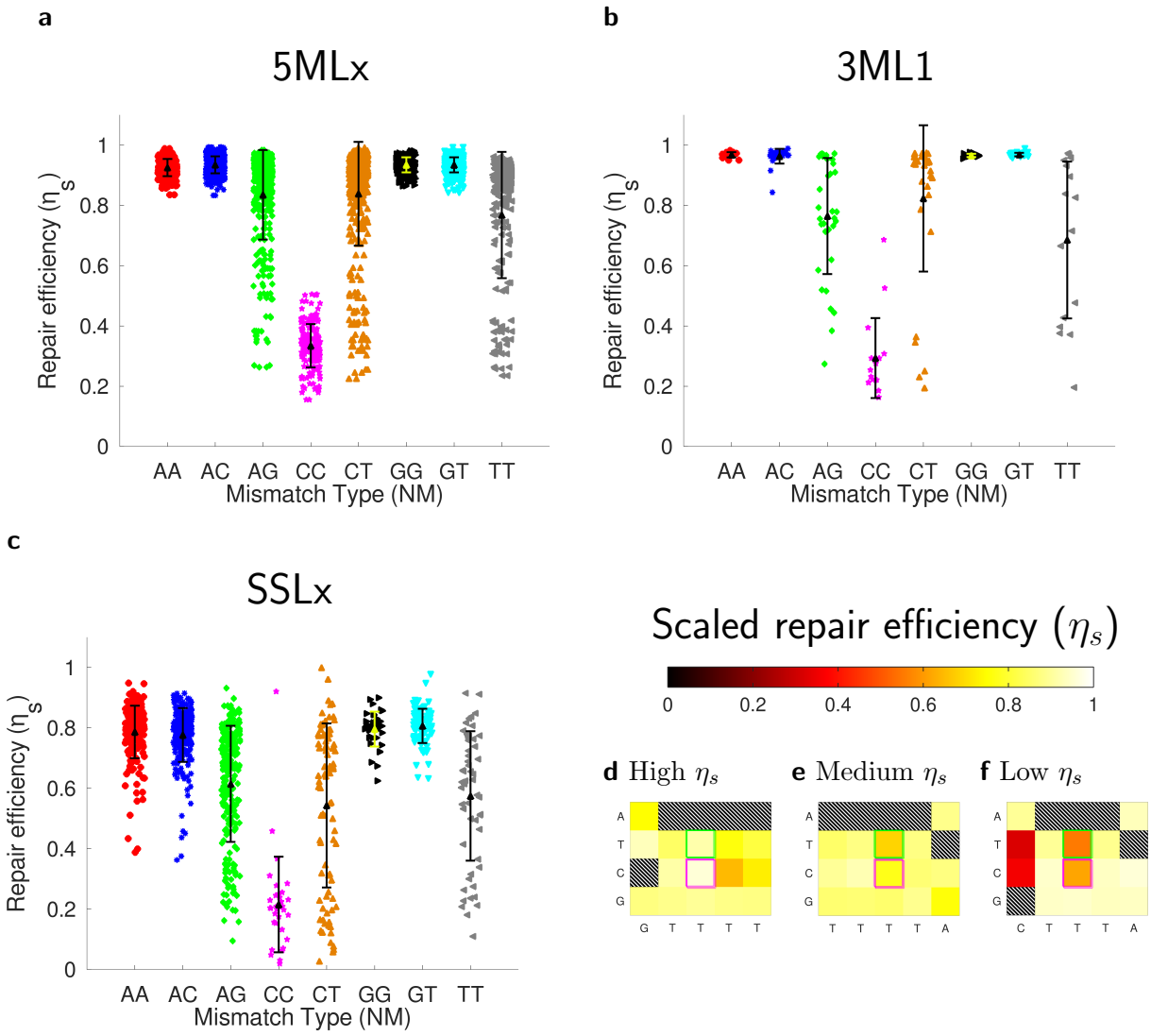
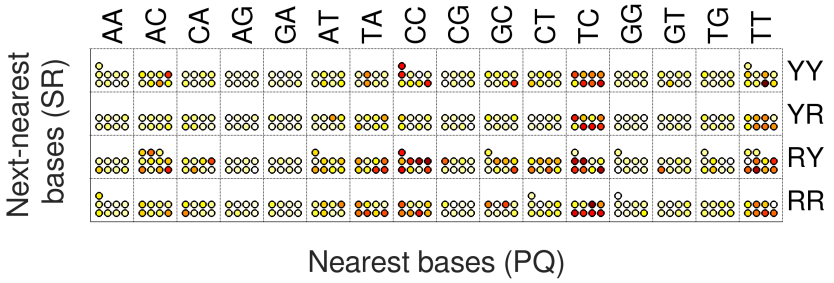



**Figure 2.11:** Comparison of $\eta_s$ measured for different mismatch types in 5MLx (a), 3ML1 (b) and SSLx (c) libraries largely agree that AG, CT and TT mismatches diplay the most sequence context dependence. Only mismatches measured within at least 2 different contexts diverging by a standard deviation less than 0.1 are included. (d - f) Example excerpts from the 5MLx shown in Figure 2.10 examplifying the sequence effect on repair efficiency, where the TT and CT mismatches with identical nearest neighbors but different next-nearest neighbors differ in $\eta_s$.

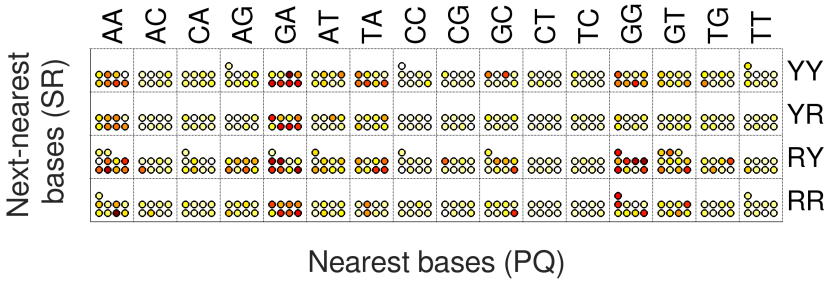
**a**

5’ - S P A Q R - 3’
3’ - S̄ P̄ G Q̄ R̄ - 5’



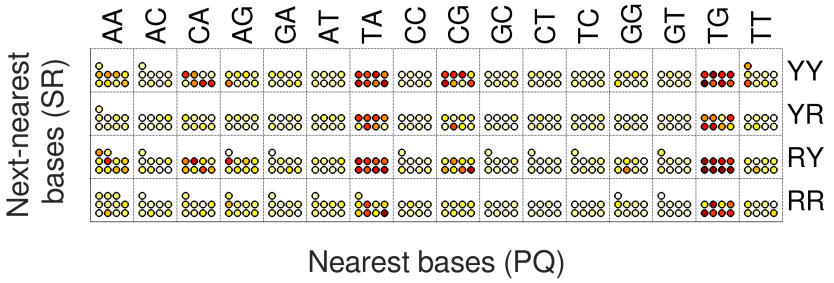

**b**

5’ - S P G Q R - 3’
3’ - S̄ P̄ A Q̄ R̄ - 5’



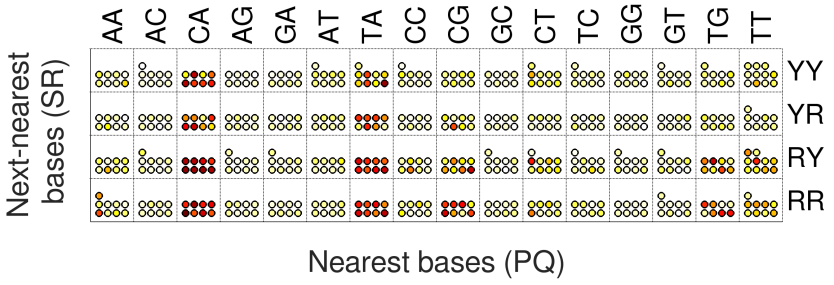

**c**

5’ - S P C Q R - 3’
3’ - S̄ P̄ T Q̄ R̄ - 5’




**d**

5’ - S P T Q R - 3’
3’ - S̄ P̄ C Q̄ R̄ - 5’




**e**

5’ - S P T Q R - 3’
3’ - S̄ P̄ T Q̄ R̄ - 5’

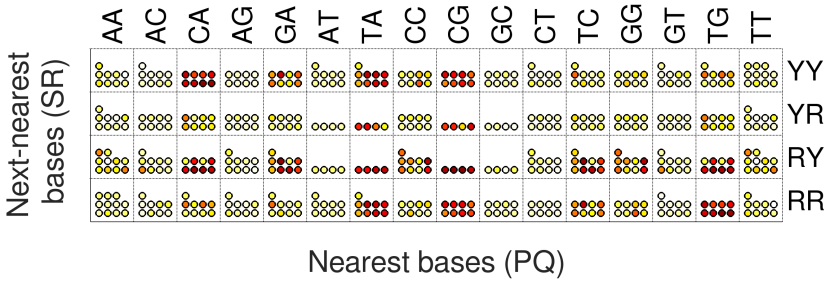


**Figure 2.12:** The effect of the nucleotides immediately adjacent to the mismatches (x-axis, positions P and Q) and the next-nearest nucleotides (y-axis, positions S and R) on the repair efficiency of each mismatch type. Each bead represents one independent measurement performed as part of 5MLx and the local sequence context obeys the sequence pattern corresponding to the cell the bead is located in. Coloring of the beads represents $\eta_s$ same as the matrices presented in Figure 2.10, obtained by cumulated clans from two experimental replicates. R: Purine base (A or G), Y: pyrimidine base (C or T). All neighbours are paired with their respective Watson-Crick pair, denoted by a bar ($\bar{N}$). Beads are unequally distributed due to the possibility of extra sampling of certain mismatches close to the termini of the mismatch libraries or symmetry operations.

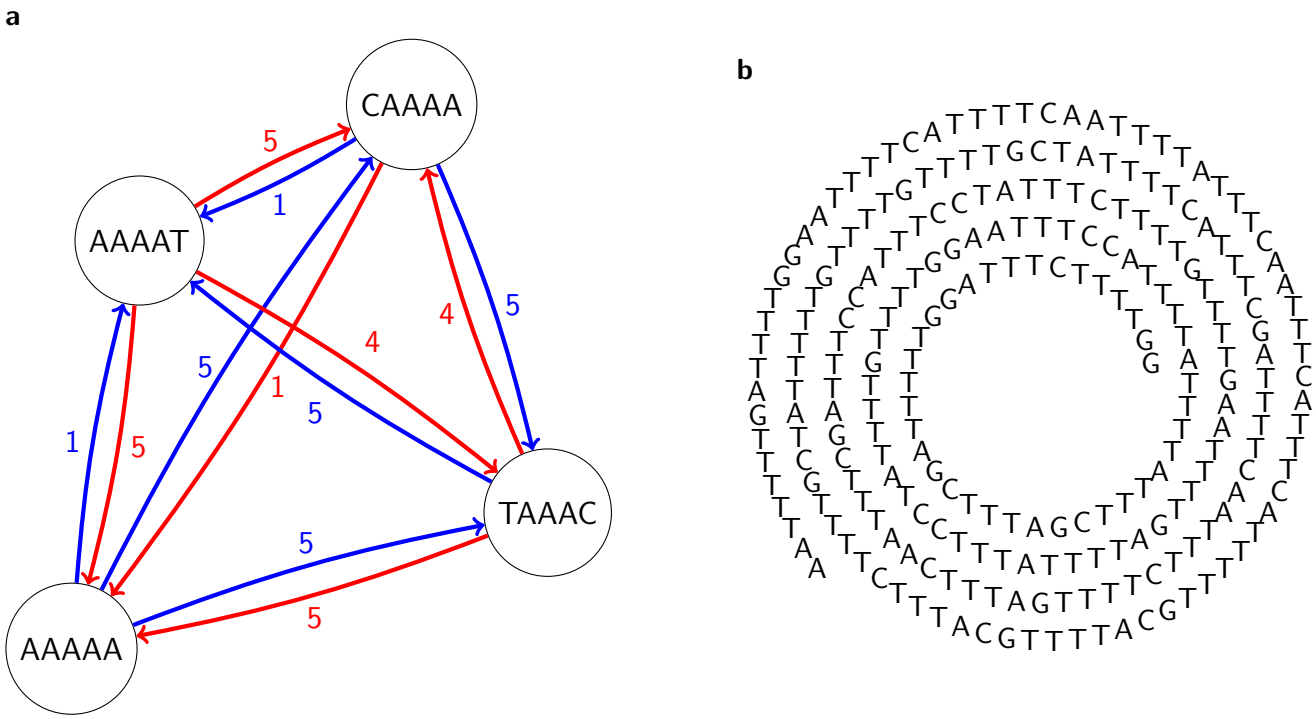



**Figure 2.13:** The construction of SSLx consensus sequence. (a) The distance measure implemented as part of the greedy algorithm to obtain the consensus sequence of SSLx. Distance is not commutative, and the order of the visited nodes is indicated by the arrowhead of the bi-directional graph (ex: d(CAAAA,AAAAA) = 1). (b) The pseudo-optimized consensus sequence of SSLx using the greedy algorithm.

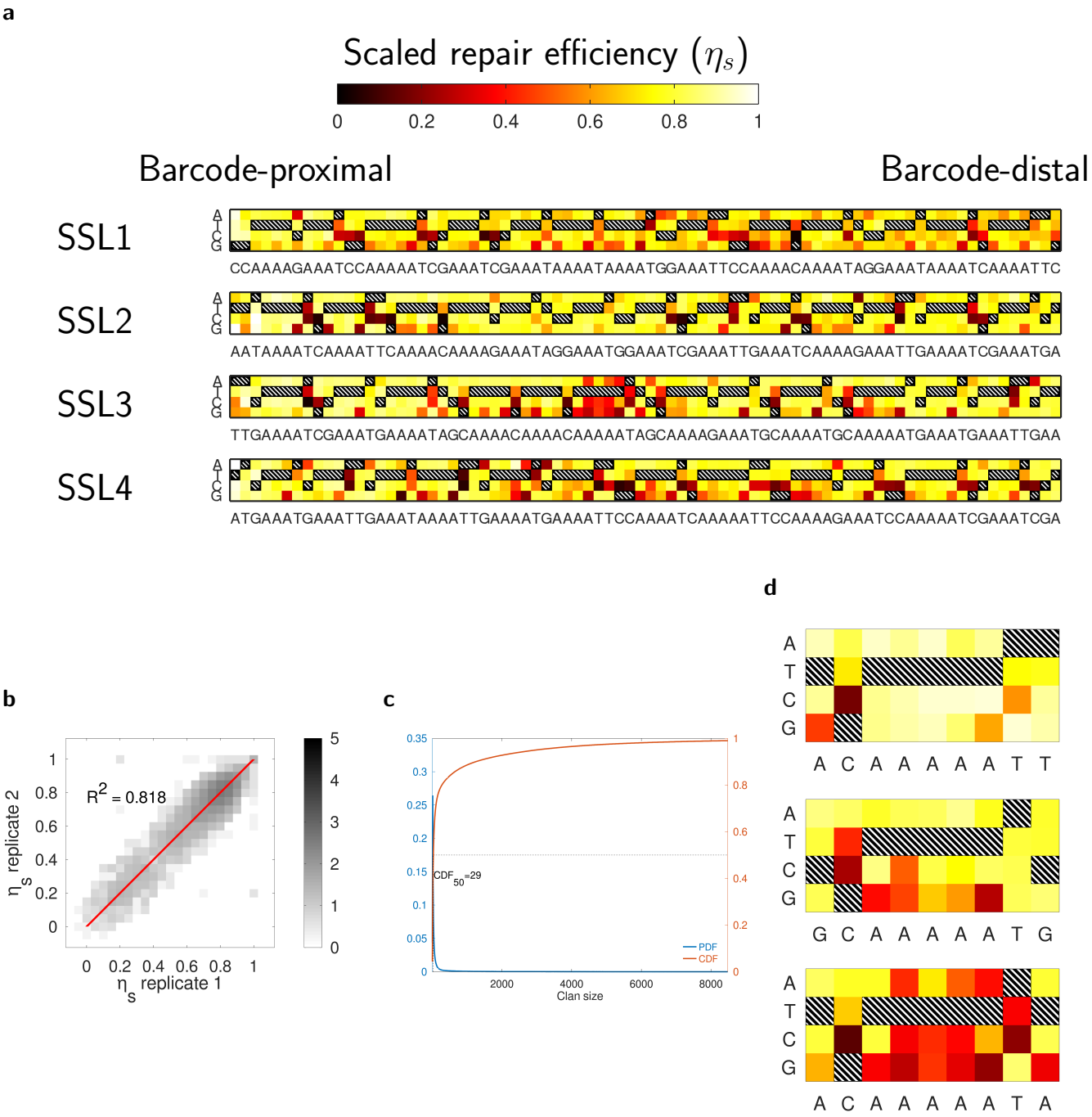



**Figure 2.14:** Results obtained using SSLx mismatch library. (a) Scaled repair efficiency matrices ($\eta_s$) of 4 sublibraries together subsampling the heptamer sequence contexts (SSLx). (b) Comparison of $\eta_s$ measured in triplicates, shown as a density map. Each of the four sublibraries are overlaid on the same graph, all possible replicate pairs were compared redundantly. The distribution of the number of reads per clan (i.e. clan size) for SSLx (c), analogous to Figure 2.22a. (d) Three example excerpts of identical sequence heptamers surrounded by different 4th degree neighbors show different $\eta_s$. From top to bottom, the windows are centered at position 62 of 5ML13, positions 63 and 36 of SSL3.

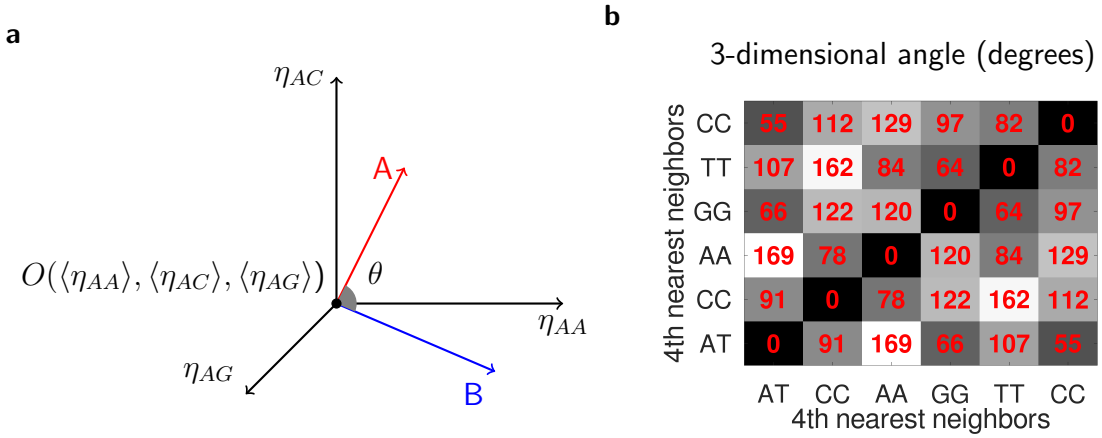



**Figure 2.15:** Comparison of the heptamer motifs via cosine similarity. (a) Definition of a 3D space spanned by the deviation of $\eta_s$ from its mean value of the three mismatches (AA, AC, AG) at the center of heptamer motifs on the common sequence. (b) Pairwise assessment of correlated changes in $\eta_s$ for three mismatches located at the middle of heptamer motifs on the common strand of the form shown in Figure 2.14d. The matrix entries indicate the angle evaluated by $\theta = arccos(\mathbf{A}\cdot\mathbf{B}/||\mathbf{A}||\cdot||\mathbf{B}||)$, lower angles are a signature of a higher correlation. The axes' labels show the base identity of the 4th degree neighbors before and after the shared CAAAAAT on the common strand, respectively.

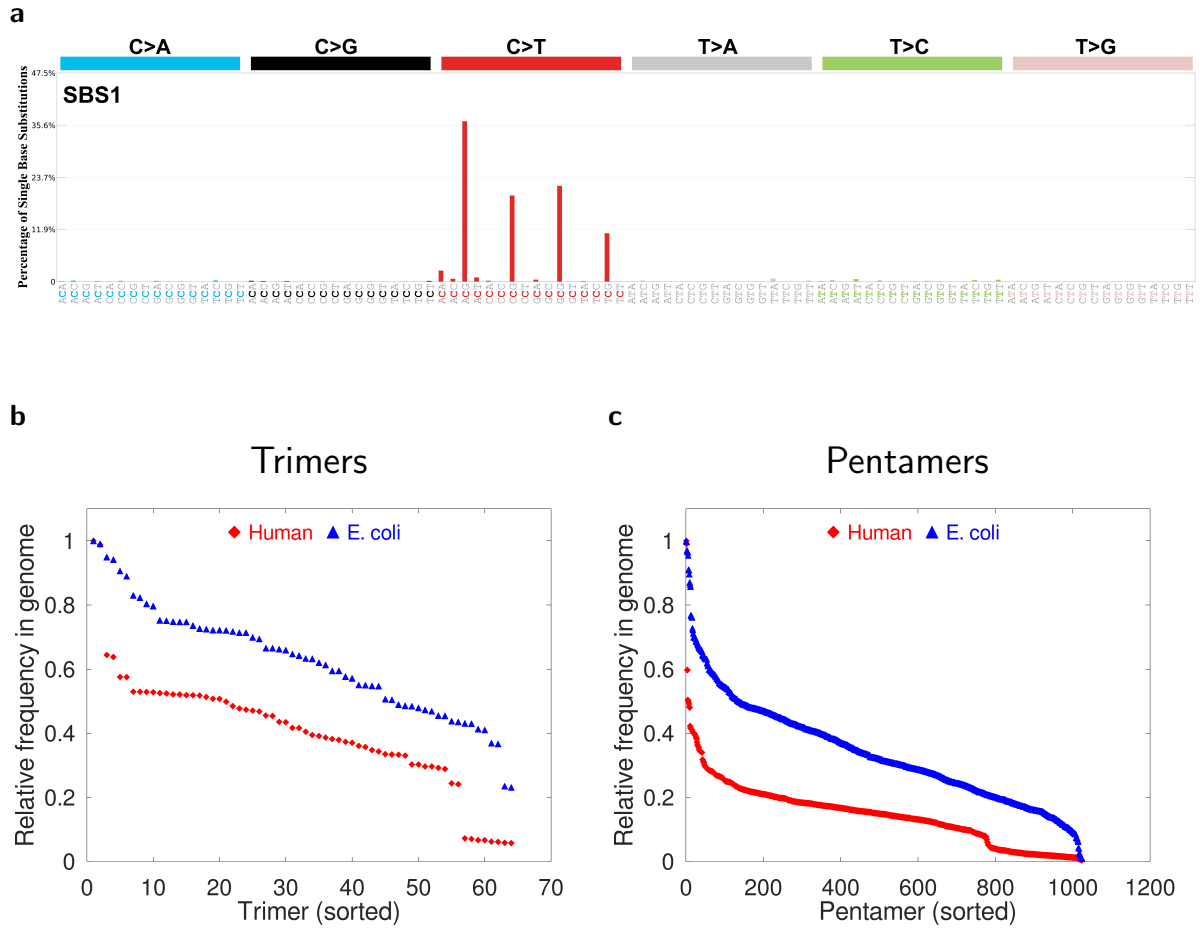



**Figure 2.16:** Mutational signatures depend on the k-mer frequencies in the genome. (a) SBS1 as an example mutational signature from the COSMIC compendium. Comparison of the frequency of DNA trimers (b) and pentamers (c) in the human genome and the *E. coli* genome. Plotted relative frequencies reflect the ratio of the abundance of the k-mer and the most abundant k-mer. Overlaps between motifs were allowed during counting.

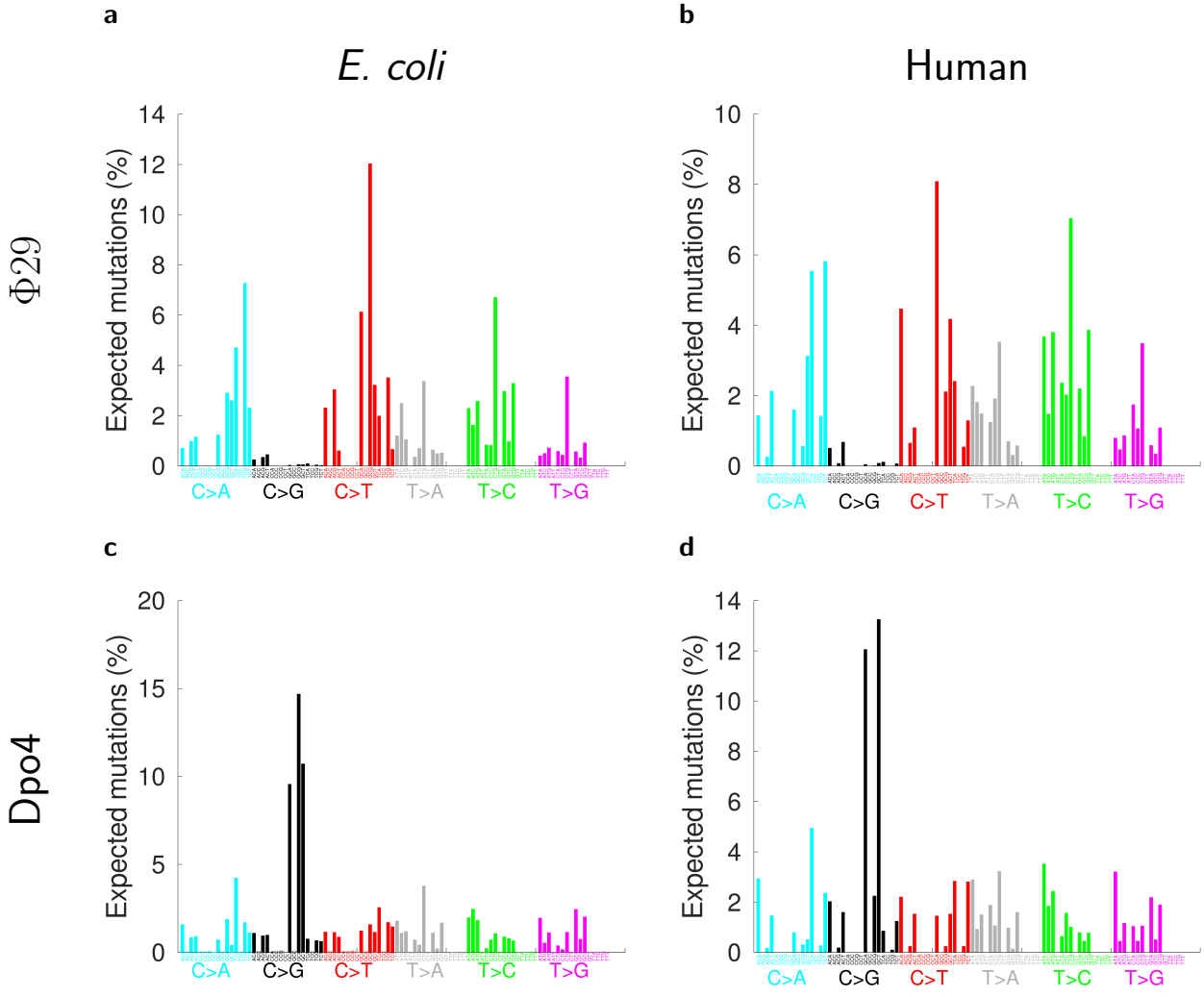



**Figure 2.17:** The speculated mutational signatures obtained using our simple model described in Section 2.3.10 using human (b, d) and *E. coli* (a, c) genomic pentamer frequencies, using the DNA replication error profiles estimated by the proofreading capable Φ29 (a, b) and incapable Dpo4 DNA polymerases (c, d). Each signature is normalized by the total expected mutational load across 96 dimensions.

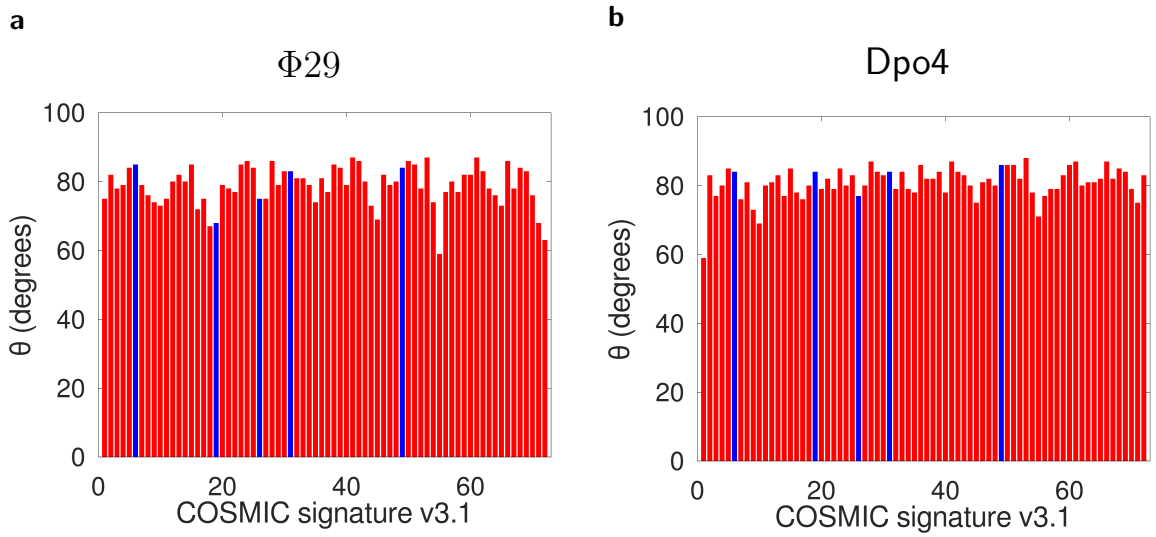



**Figure 2.18:** The cosine similarity between COSMIC and the speculated signatures for human cell lines using the Φ29 (a) and Dpo4 (b) replication error profiles. The blue bars represent the COSMIC signatures that are known to be associated with MMR defects (signatures 6, 15, 21, 26, 44), excluding those with concurrent mutations leading to hypermutant DNA replication (signatures 14 and 20).

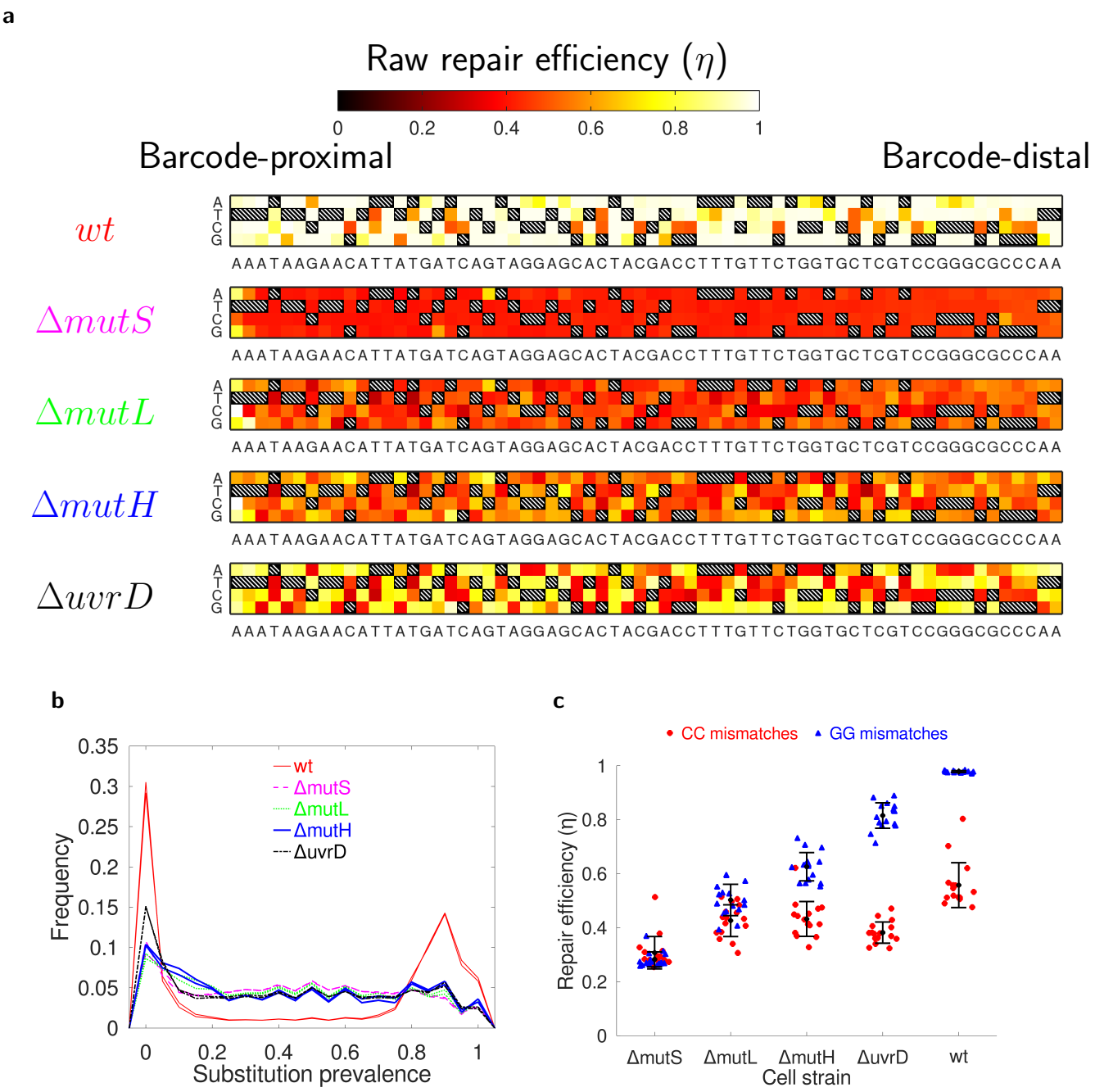



**Figure 2.19:** The effect of MMR mutations on the repair efficiency. (a) Repair efficiency measurements obtained within 3ML1 template sequence in various MMR pathway mutant strains. (b) Substitution prevalence histograms of MMR pathway mutants. Each condition is represented by two separate curves of identical color representing independent experimental duplicates. (c) Comparison of the repair efficiency of CC (red) and GG (blue) mismatches in MMR pathway mutant strains. Each individual data point represents one occurrence of a CC or GG mismatch located within a different sequence context as part of 3ML1. $\eta$ values indicated along the y-axis are the mean of two experimental replicates of the indicated condition, while the error bars show $\mu \pm \sigma$ of each group.

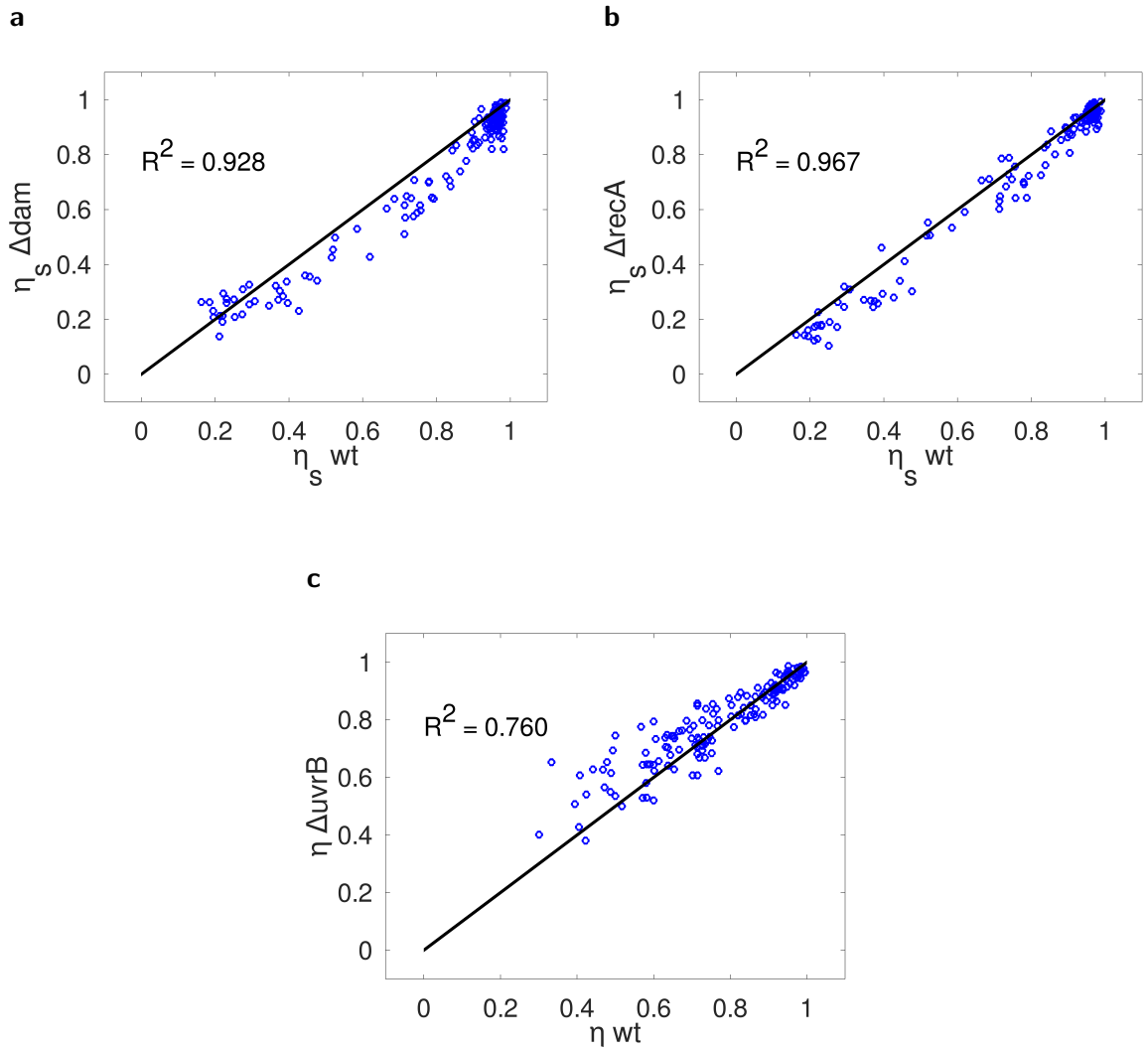



**Figure 2.20:** Comparison of $\eta_s$ of unmethylated 3ML1 library measured in $\Delta dam$ (a) and $\Delta recA$ cells (b) correlate with wt data. Comparison of $\eta'$ measures using SML in wt and $\Delta uvrB$ cells (c).

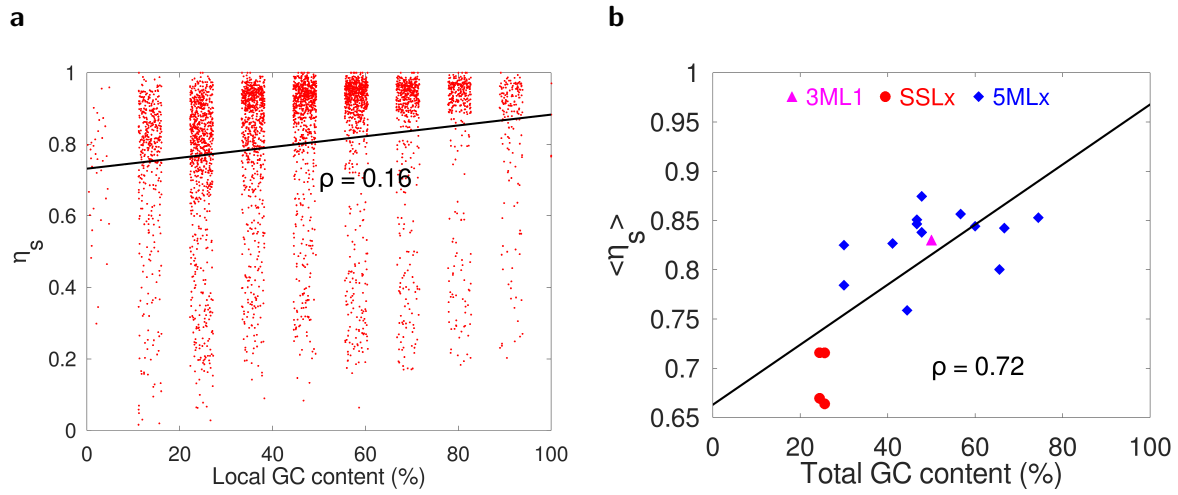



**Figure 2.21:** The effect of AT content on repair efficiency. (a) Correlation between $\eta_s$ and local AT content. Each individual mismatch that was sampled in 5MLx or SSLx is represented with a red dot, where the x-axis includes random noise component for data visualisation. Local AT content is the percentage of bases within the 9-base window centered at the position of the mismatch that are A or T, $\rho$ indicates Pearson's correlation coefficient, i.e. $\rho(x, y) = cov(x, y)/\sigma_x\sigma_y$. (b) The correlation between $\langle\eta_s\rangle$ across all mismatches that is part of the same sublibrary and the overall AT percentage. Each sublibrary is indicated by one data point.

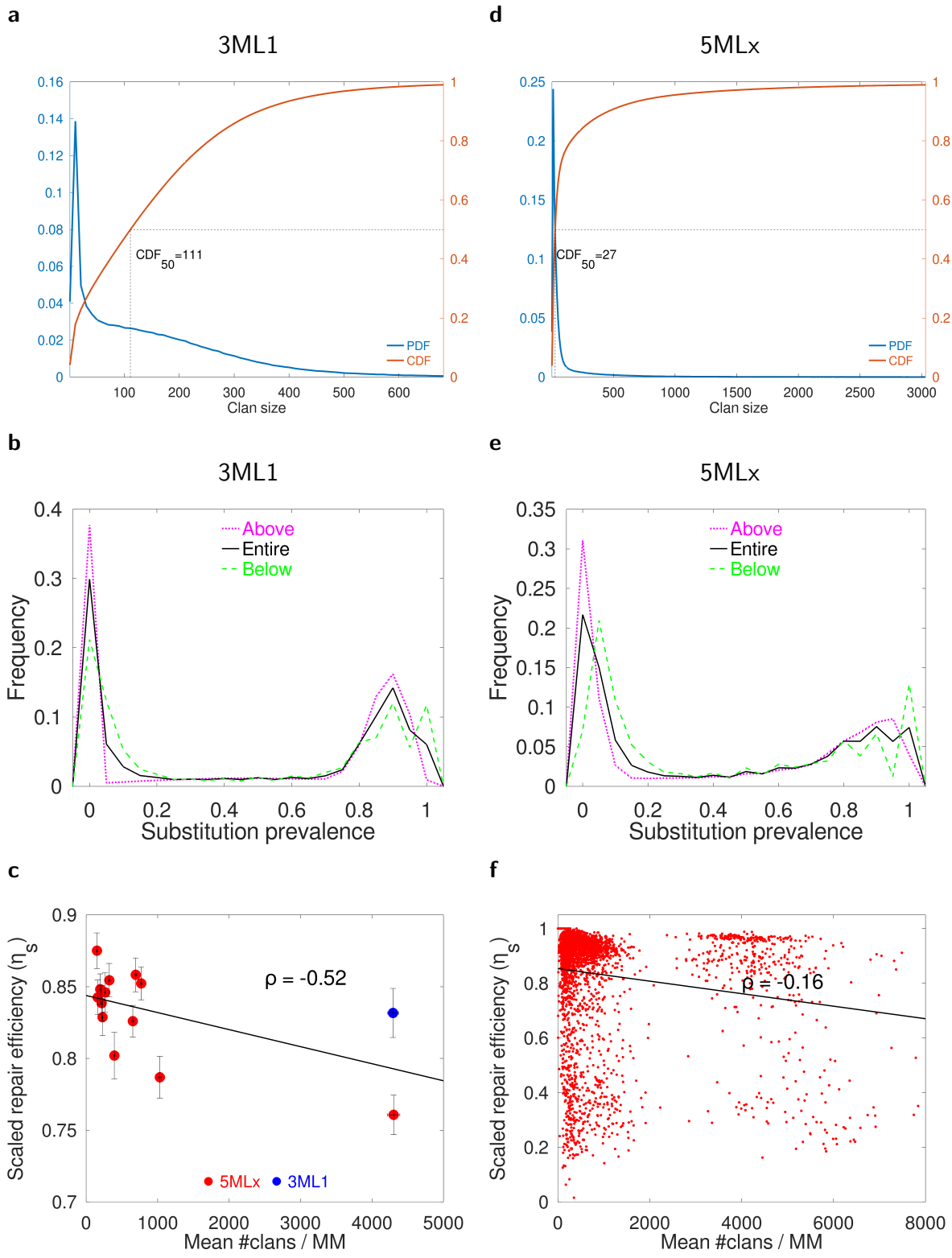

a
3ML1
CDF50=111
PDF
CDF
Clan size
d
5MLx
CDF50=27
PDF
CDF
Clan size
b
3ML1
Above
Entire
Below
Frequency
Substitution prevalence
e
5MLx
Above
Entire
Below
Frequency
Substitution prevalence
c
ρ = -0.52
5MLx
3ML1
Scaled repair efficiency (ηs)
Mean #clans / MM
f
ρ = -0.16
Scaled repair efficiency (ηs)
Mean #clans / MM

**Figure 2.22:** The effect of data quantity on the measurements. The distribution of number of reads per clan (i.e. clan size) for 3ML1 (a) and 5MLx (d). $CDF_{50}$ values represent the clan size at which the cumulative distribution function (CDF) reaches 0.5, i.e the median of the distribution. (b,e) The substitution prevalence distributions indicate the C-U-V populations detected in the below median (- - -) and above median ($\cdots$) sized clans, along with the distribution for the entire population (—). The reported median of 27 reads/clan for 5MLx libraries is based on the combined data for all 13 constituent sublibraries. The median values of individual sublibraries were 37, 39, 47, 80, 39, 127, 34, 20, 72, 50, 50, 56 and 40 reads/clan, respectively. (c) Correlation between the number of clans detected per mismatch in the library and the library-wide mean repair efficiency. Error bars indicate $\mu \pm \sigma$ for each sublibrary. (f) Correlation between the number of clans detected for each individual mismatch and the scaled repair efficiency. Each dot represents the mean measurements of experimental replicates for a a single mismatch that was sampled as part of either 3ML1 or 5MLx. $\rho$ indicates Pearson correlation coefficient, i.e. $\rho(x, y) = cov(x, y)/\sigma_x \sigma_y$.

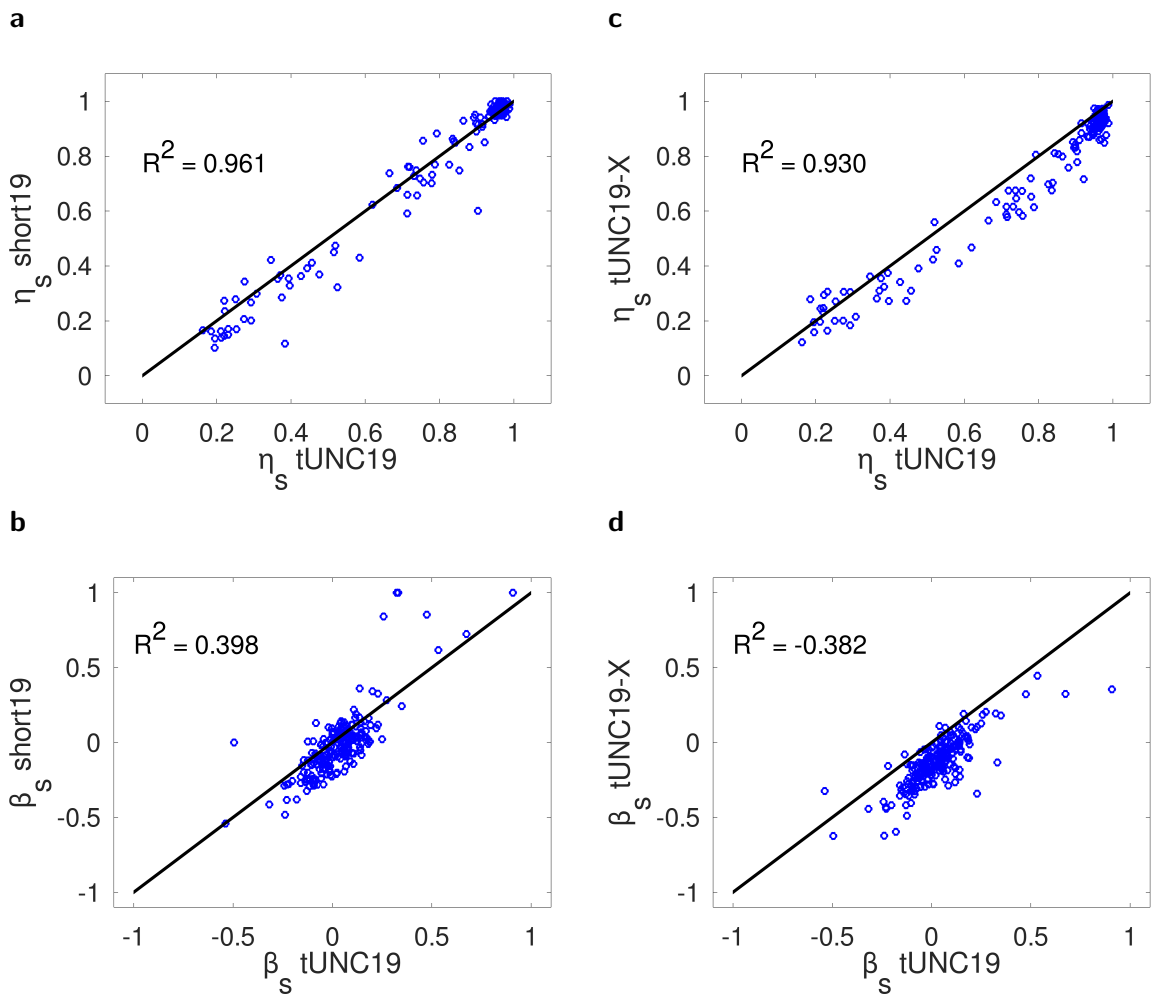



**Figure 2.23:** Changes on the vector backbone do not alter observations significantly. Comparison of the scaled repair efficiencies ($\eta_s$) and strand retention biases ($\beta$) measured in wt cells with or without a nearby lac promoter in the barcode-distal end of the mismatch library (a and b) and with or without the bubble-mitigating primer extension procedure (c and d). Introduction of a primer extension step to remove potential heteroduplex bubbles in the tracing barcode region of the vector does not lead to a significant change in $\eta_s$, nor the presence of the promoter. Each data point represents the repair efficiency of an individual mismatch on 3ML1 as deduced by the cumulated data from two experimental replicates, $R^2$ indicates the coefficient of determination.

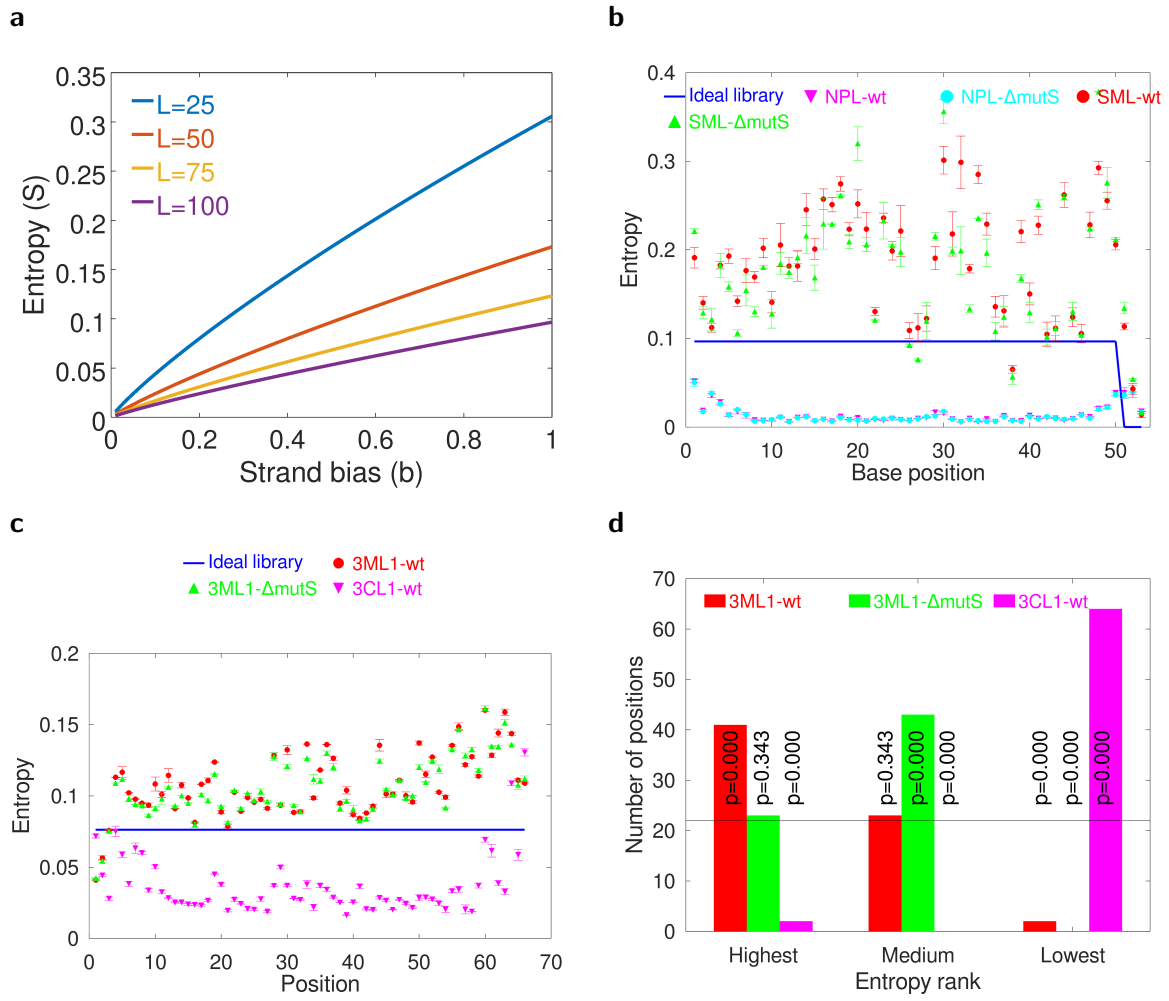



**Figure 2.24:** Information entropy as a measure of sequence heterogeneity. (a) Expected information entropy as a function of library length (L) and strand bias (b). (b) The observed entropy as a function of base position along the mismatch library in SML and its corresponding control sample without mismatches NPL. (c) The observed information entropy in 3ML1 and its corresponding control sample 3CL1. Information entropy of the DNA libraries expressed as mean ± S.E.M. of 2 to 4 experimental replicates. (—) corresponds to the expected entropy for a library with a uniform substitution probability in the absence of experimental errors. (d) Histogram of relative entropy rank of three 3ML1 samples. Each base position is recorded as 1 event and the p-values reflect the one-tailed probabilities explicitly calculated by the bionomial formula testing the observed deviation from the fully random rank distribution with 1/3 probability each (—).

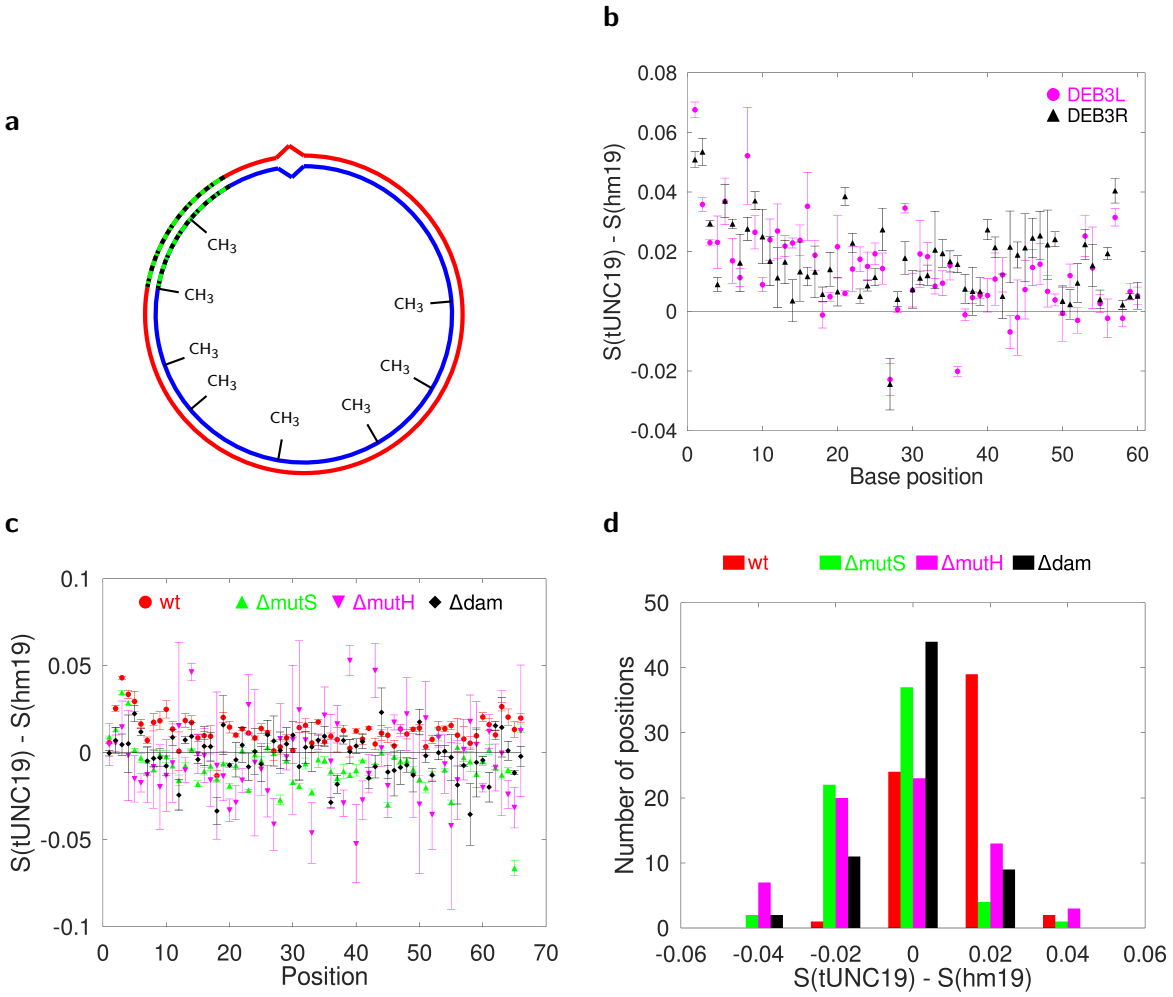



**Figure 2.25:** Hemi-methylation induces strand choice bias. (a) The hemi-methylation scheme places adenine methylations on the strand that is ligated to the common strand of the mismatch library, whereas the variable strands will be preferentially removed. The difference in entropy levels observed in the absence and presence of hemi-methylation for DEB3 (b) and 3ML1 (c) libraries. Information entropy of DNA libraries are expressed as mean ± S.E.M of 2 to 4 experimental replicates. The thin black lines (—) represent the absence of any entropic effect due to hemi-methylation (i.e. $\Delta S = 0$). (d) The effect of genetic background on the entropy change due to hemi-methylation. Each event count in the histogram corresponds to 1 out of 66 variable base positions in 3ML1.

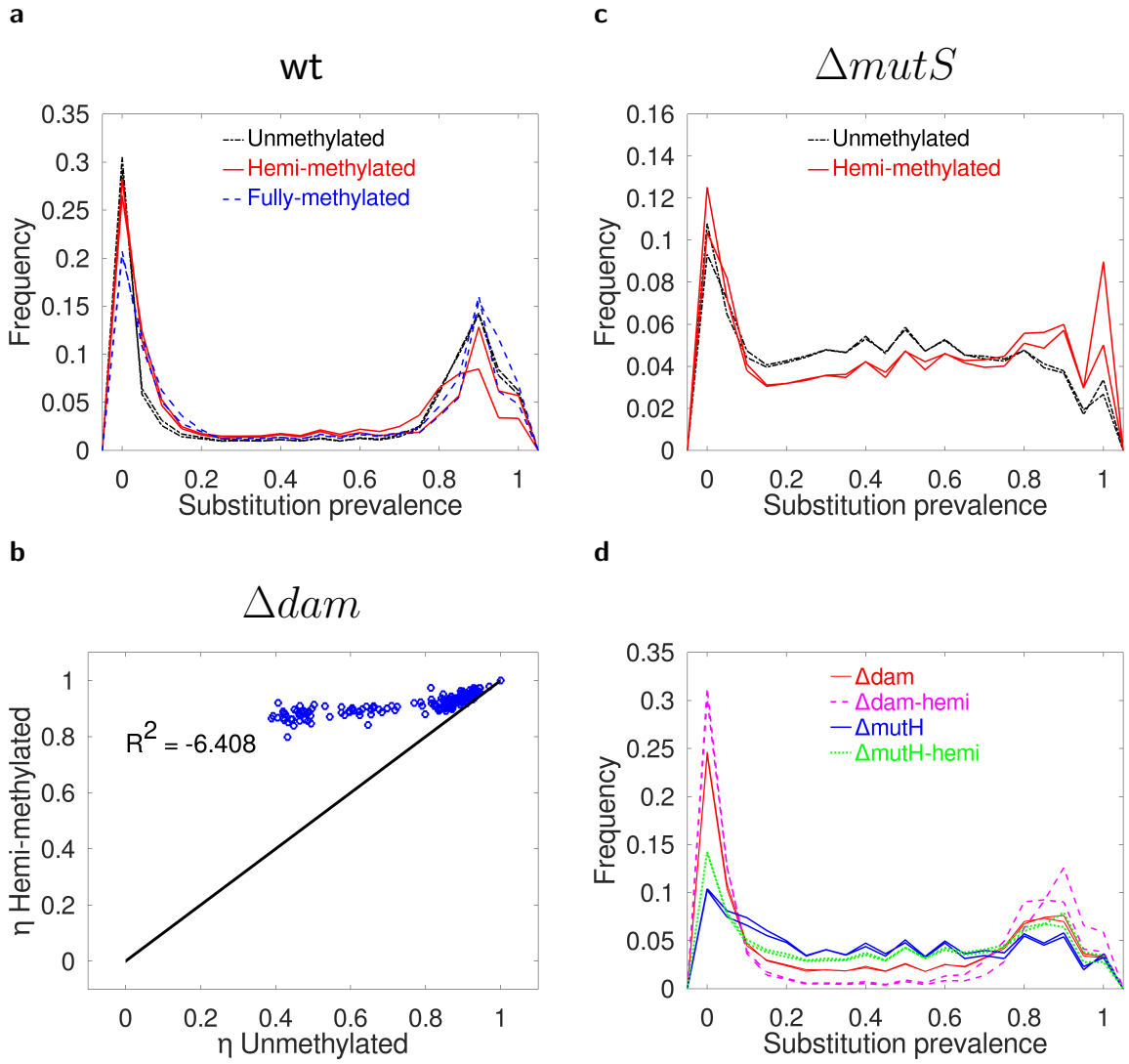



**Figure 2.26:** The effects of methylation on the observed repair response. Comparison of the substitution prevalence histograms under different methylation states in wt (a) and ΔmutS cells (c). (b) The comparison of raw repair efficiency ($\eta$) of 3ML1 mismatches measured in Δ*dam* cells, as part of unmethylated vs. hemi-methylated plasmids. Each data point represents one individual mismatch, whose repair efficiency is averaged over 2 experimental replicates. (d) Substitution prevalence histograms of unmethylated and hemi-methylated plasmids carrying the 3ML1 mismatch library, measured in Δ*dam* or Δ*mutH* cells. Each independent experimental duplicate is indicated by a separate curve of identical form and color.

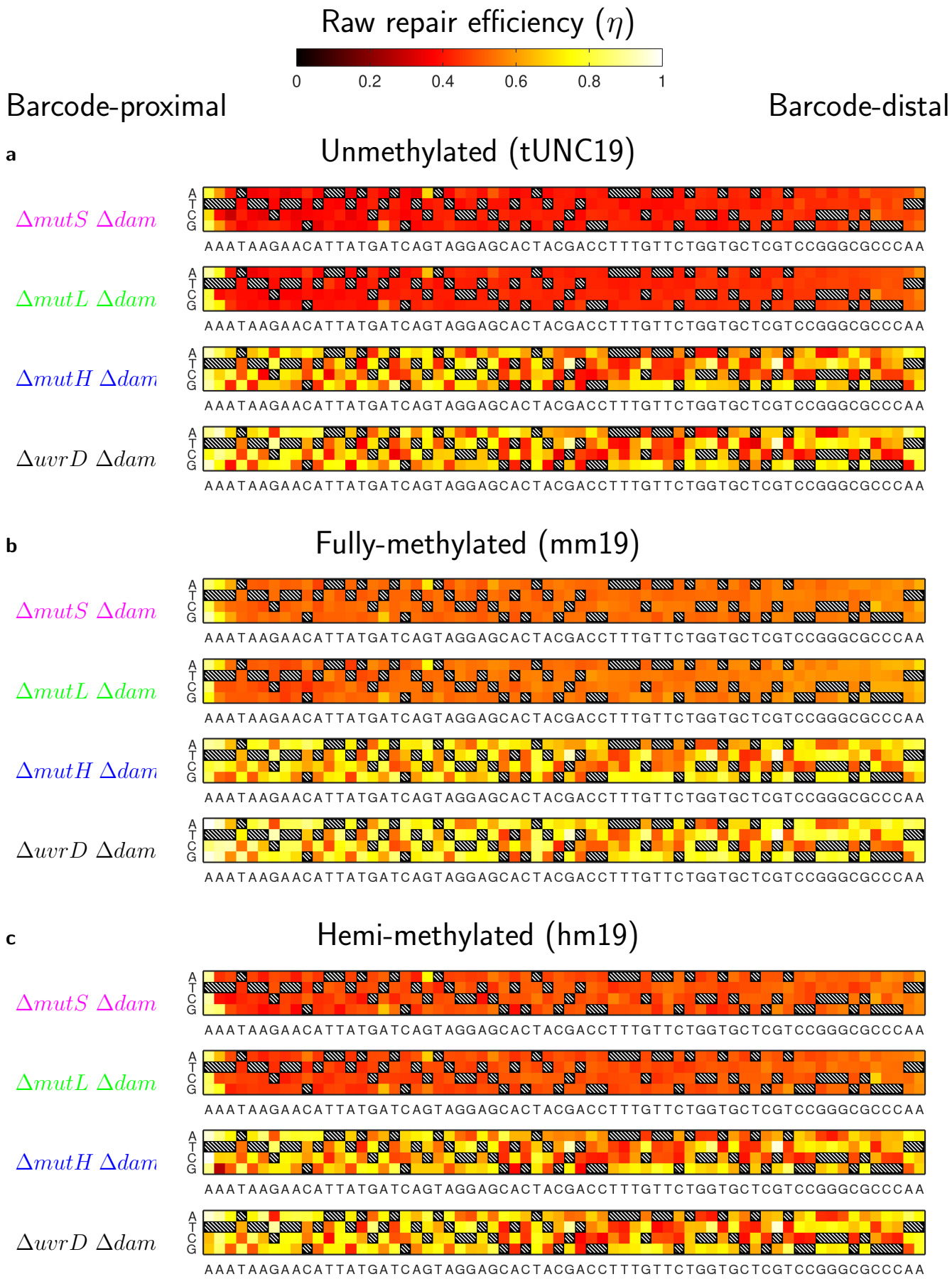

Raw repair efficiency ($\eta$)
0
0.2
0.4
0.6
0.8
1
Barcode-proximal
Barcode-distal
a
Unmethylated (tUNC19)
b
Fully-methylated (mm19)
c
Hemi-methylated (hm19)
$\Delta mutS\ \Delta dam$
$\Delta mutL\ \Delta dam$
$\Delta mutH\ \Delta dam$
$\Delta uvrD\ \Delta dam$
A
T
C
G
AAATAAGAACATTATGATCAGTAGGAGCACTACGACCTTTGTTCTGGTGCTCGTCCGGGCGCCCAA

**Figure 2.27:** The effect of concurrent deletion of MMR pathway elements and *dam* gene on the observed repair efficiency. Repair efficiency measurements were obtained within the 3ML1 template sequence ligated into barcoded vectors that were devoid of methylation at the time of transformation (tUNC19, a), symmetrically methylated on both strands (mm19, b) or methylated only on the strand carrying the common sequence (hm19, c). $\eta$ values reflect the repair efficiency calculated out of the combined data obtained from two experimental replicates of each indicated condition.

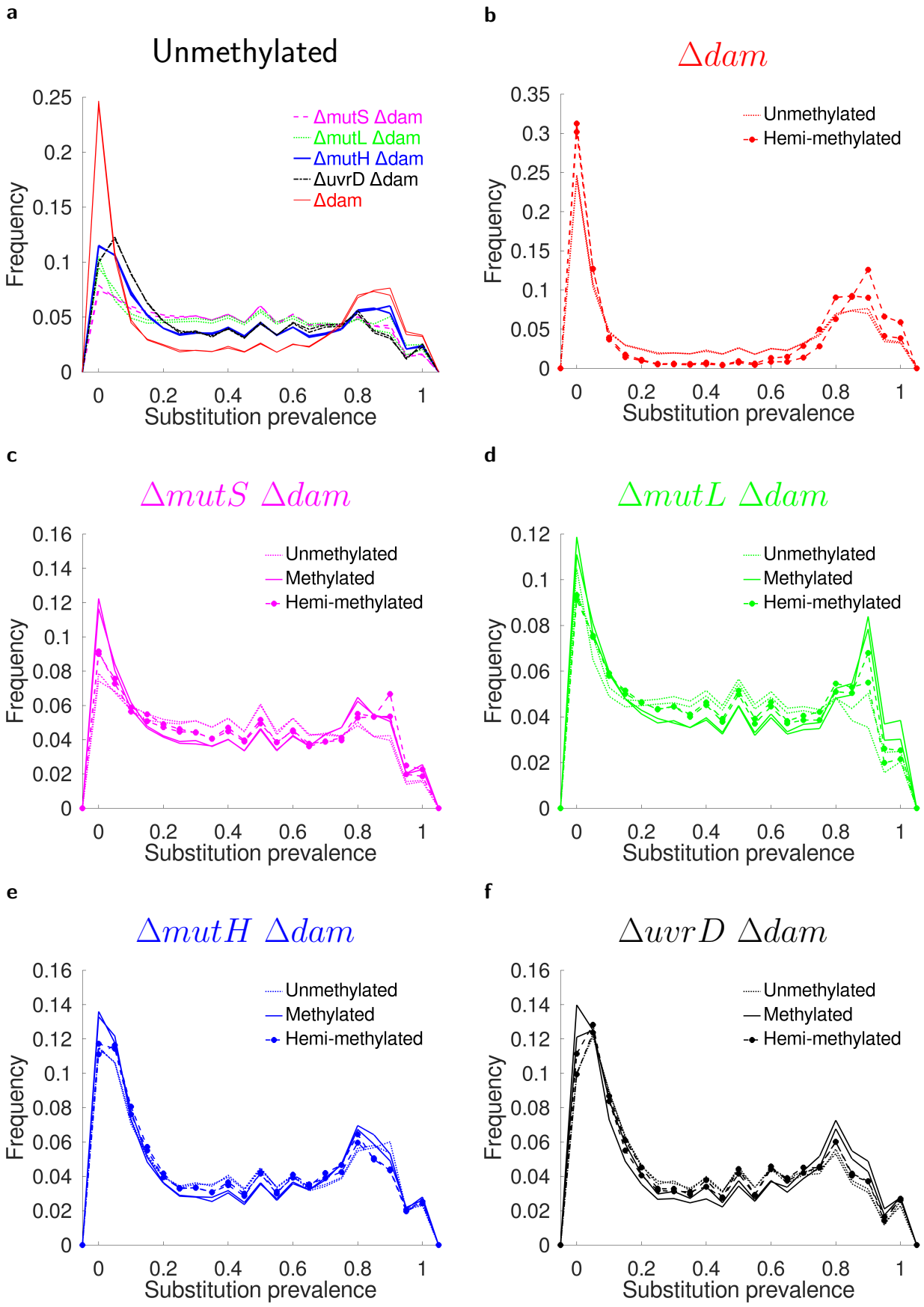

a
Unmethylated
ΔmutS Δdam
ΔmutL Δdam
ΔmutH Δdam
ΔuvrD Δdam
Δdam
Frequency
Substitution prevalence
b
Δdam
Unmethylated
Hemi-methylated
Frequency
Substitution prevalence
c
ΔmutS Δdam
Unmethylated
Methylated
Hemi-methylated
Frequency
Substitution prevalence
d
ΔmutL Δdam
Unmethylated
Methylated
Hemi-methylated
Frequency
Substitution prevalence
e
ΔmutH Δdam
Unmethylated
Methylated
Hemi-methylated
Frequency
Substitution prevalence
f
ΔuvrD Δdam
Unmethylated
Methylated
Hemi-methylated
Frequency
Substitution prevalence

**Figure 2.28:** Substitution prevalence histograms of dam-MMR pathway double mutants obtained using the 3ML1 consensus sequence. (a) Comparison of the s-histograms for various double mutants obtained with unmethylated input DNA. (b-f) S-histograms obtained in double mutant cell strains obtained using unmethylated (tUNC19), methylated (mm19) and hemi-methylated (hm19) barcoded vectors. Data obtained in each mutant cell strain is represented by two individual curves representing two independent experimental duplicates.

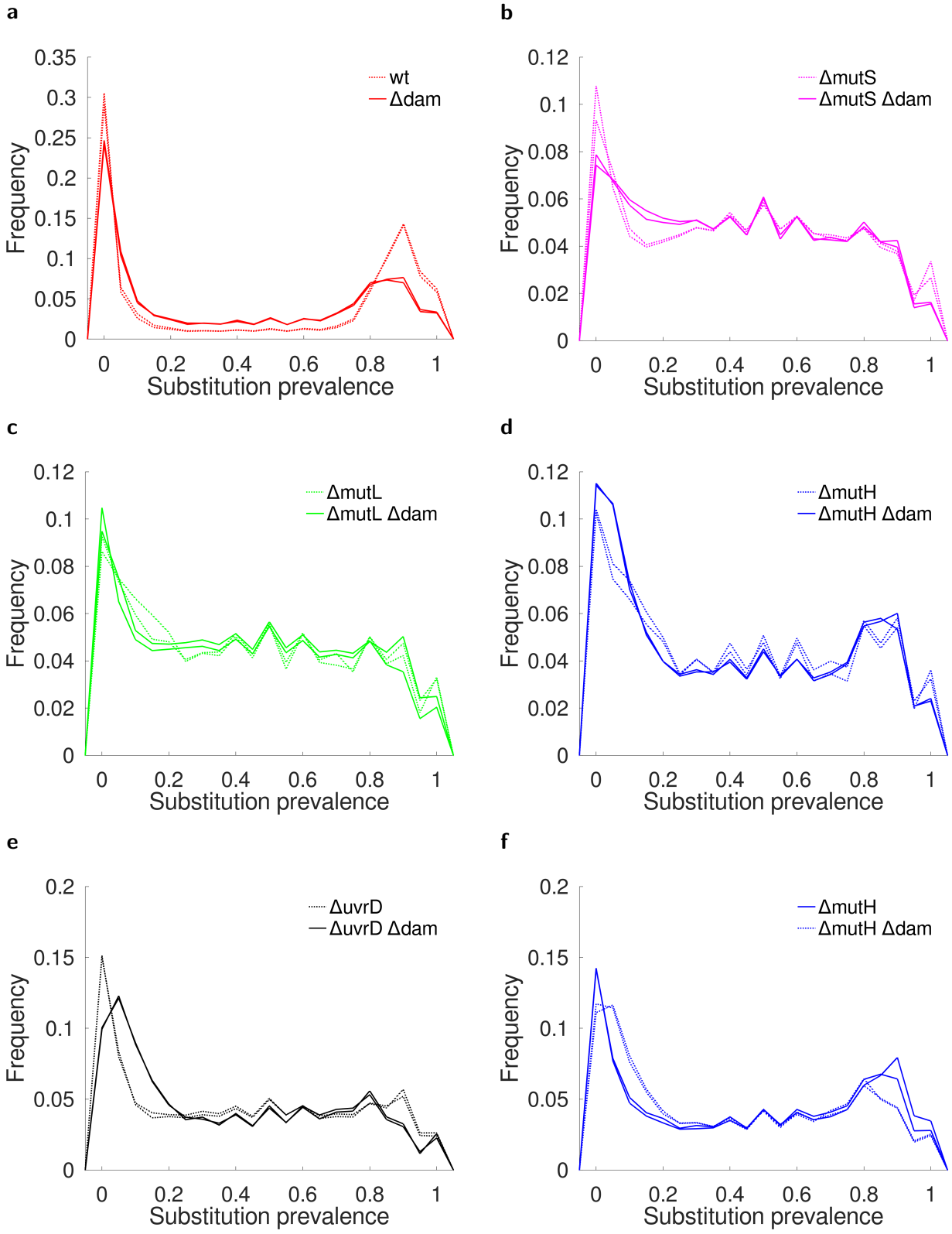

a
wt
Δdam
Frequency
Substitution prevalence
b
ΔmutS
ΔmutS Δdam
Frequency
Substitution prevalence
c
ΔmutL
ΔmutL Δdam
Frequency
Substitution prevalence
d
ΔmutH
ΔmutH Δdam
Frequency
Substitution prevalence
e
ΔuvrD
ΔuvrD Δdam
Frequency
Substitution prevalence
f
ΔmutH
ΔmutH Δdam
Frequency
Substitution prevalence

**Figure 2.29:** S-histogram comparisons between ΔMMR and ΔMMR Δ*dam* cells. Additional deletion of *dam* has a small but measurable effect on the substitution prevalence histograms of cells with intact MMR (a) or MMR mutant cells transformed with unmethylated plasmids (b-e). (f) S-histograms of MutH deficient cells transformed with hemi-methylated plasmids. Data obtained in each mutant cell strain was obtained using 3ML1 consensus sequence ligated into unmethylated input DNA and each condition is represented by two separate curves representing two independent experimental duplicates.

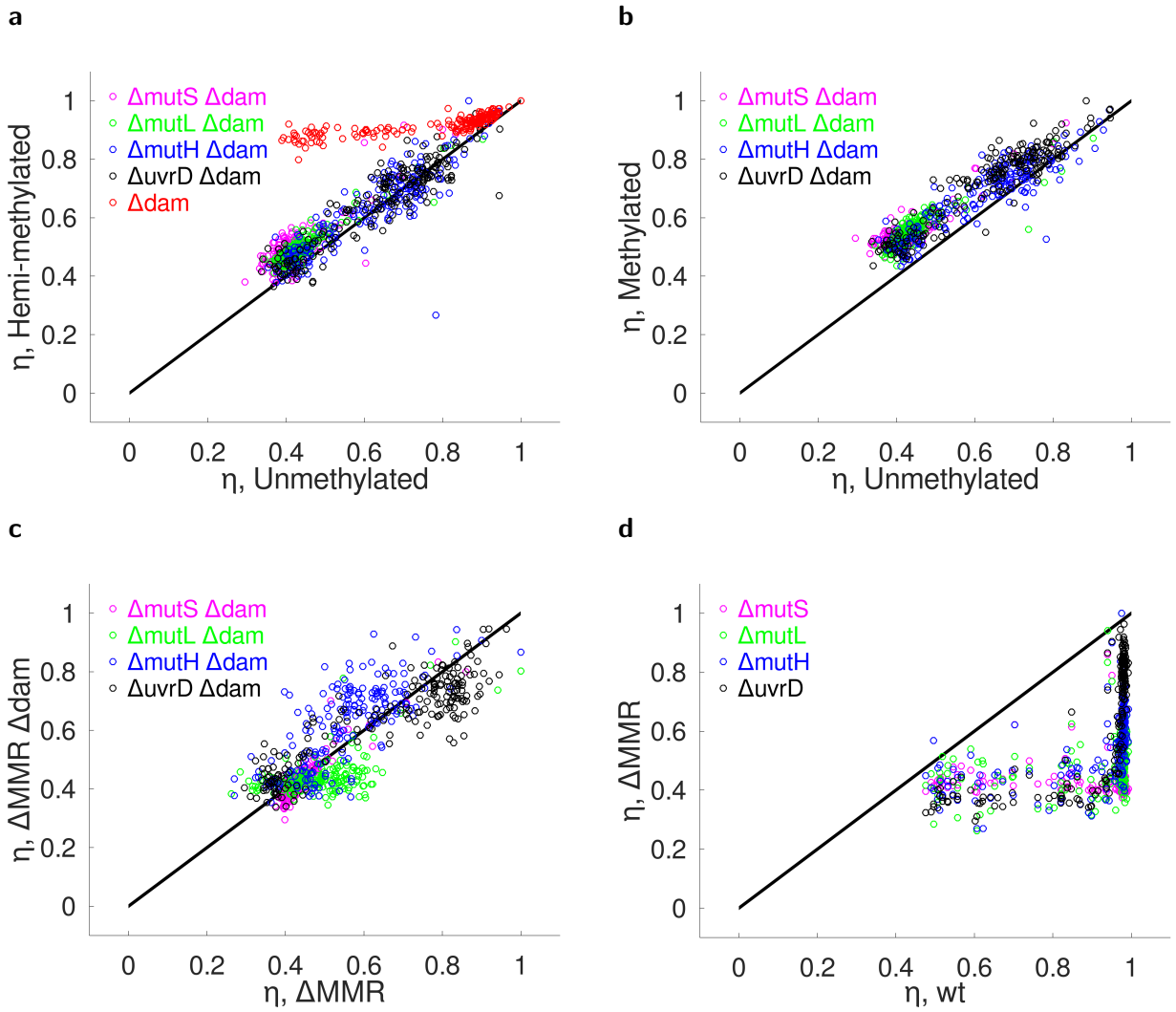



**Figure 2.30:** Comparison of repair efficiencies for ΔMMR Δ*dam* double mutants. (a) Hemi-methylation does not change the measured repair efficiencies relative to the unmethylated case in ΔMMR Δ*dam* cells, but boosts the apparent repair capability in MMR-capable cells. (b) Symmetric methylation of both strands increase the observed repair efficiencies to only a minor degree in all ΔMMR Δ*dam* cells. (c) In the 4 MMR-mutant strains, additional deletion of *dam* does not have a profound effect on the observed repair efficiencies of unmethylayted plasmids. (d) In cells with *in vivo* methylation capability, additional mutation of any MMR element causes a major deviation from the repair efficiencies observed in wt cells. In all panels, each data point represents the repair efficiency of an individual mismatch as part of 3ML1 consensus sequence and quantified out of the cumulated output of two experimental replicates for the indicated condition.

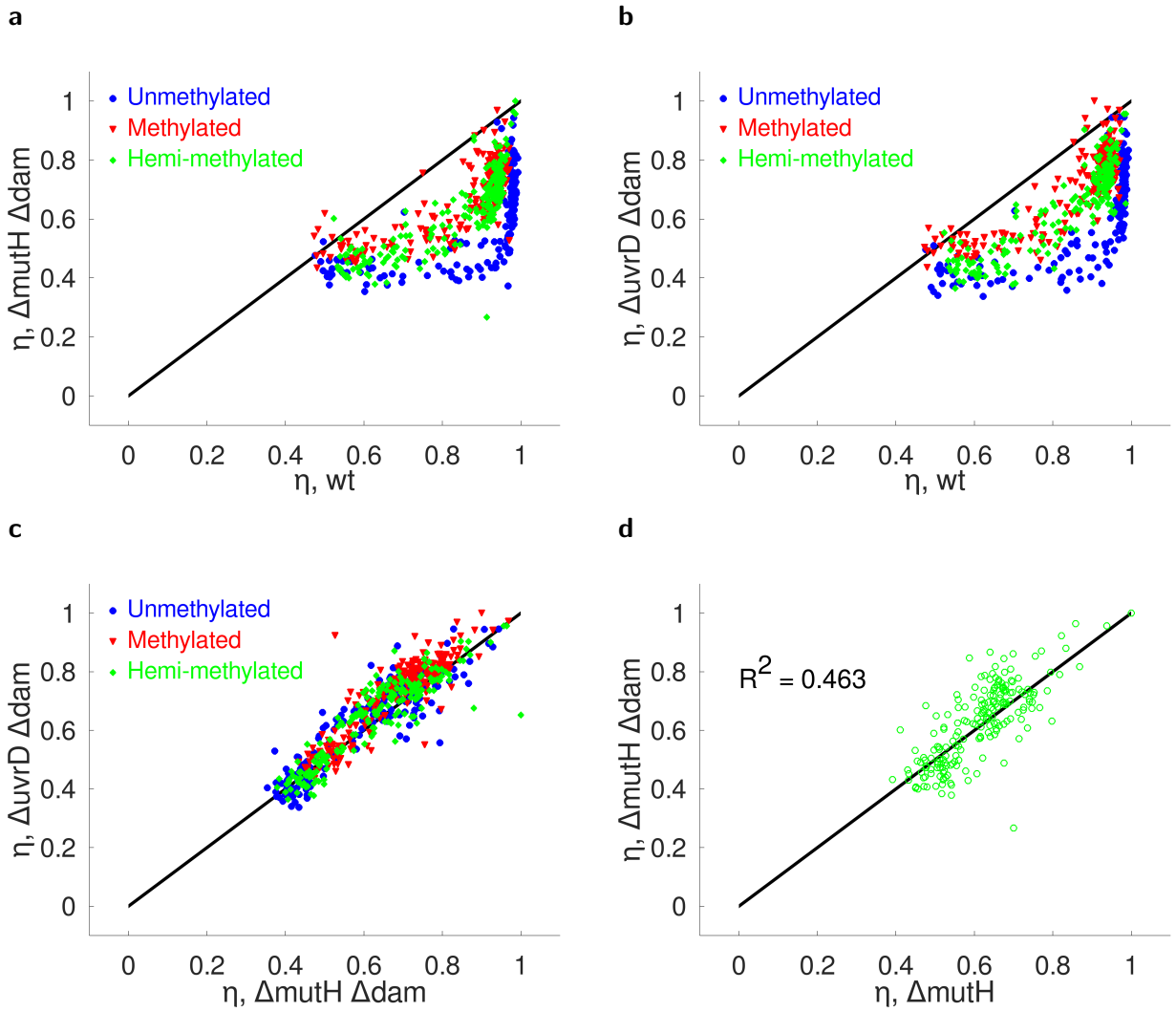



**Figure 2.31:** The residual repair efficiency in ΔMMR cells resembles MMR response in wt cell. While the mismatches with intermediate repair efficiencies are disproportionally negatively affected by Δ*mutH* Δ*dam* (a) or Δ*uvrD* Δ*dam* (b) mutations, the observed residual repair capability has a similar sequence preference as wt cells for unmethylated (tUNC19), methylated (mm19) or hemi-methylated (hm19) barcoded vectors. (c) These observed sequence dependent trends are similar between Δ*mutH* Δ*dam* and Δ*uvrD* Δ*dam* cells. (d) In Δ*mutH* cells, the repair efficiency of hemi-methylated plasmids are not globally impacted by the presence of cytoplasmic Dam. In all panels, each data point represents the repair efficiency of an individual mismatch as part of 3ML1 consensus sequence and quantified by using cumulated output of two experimental replicates for the indicated condition.

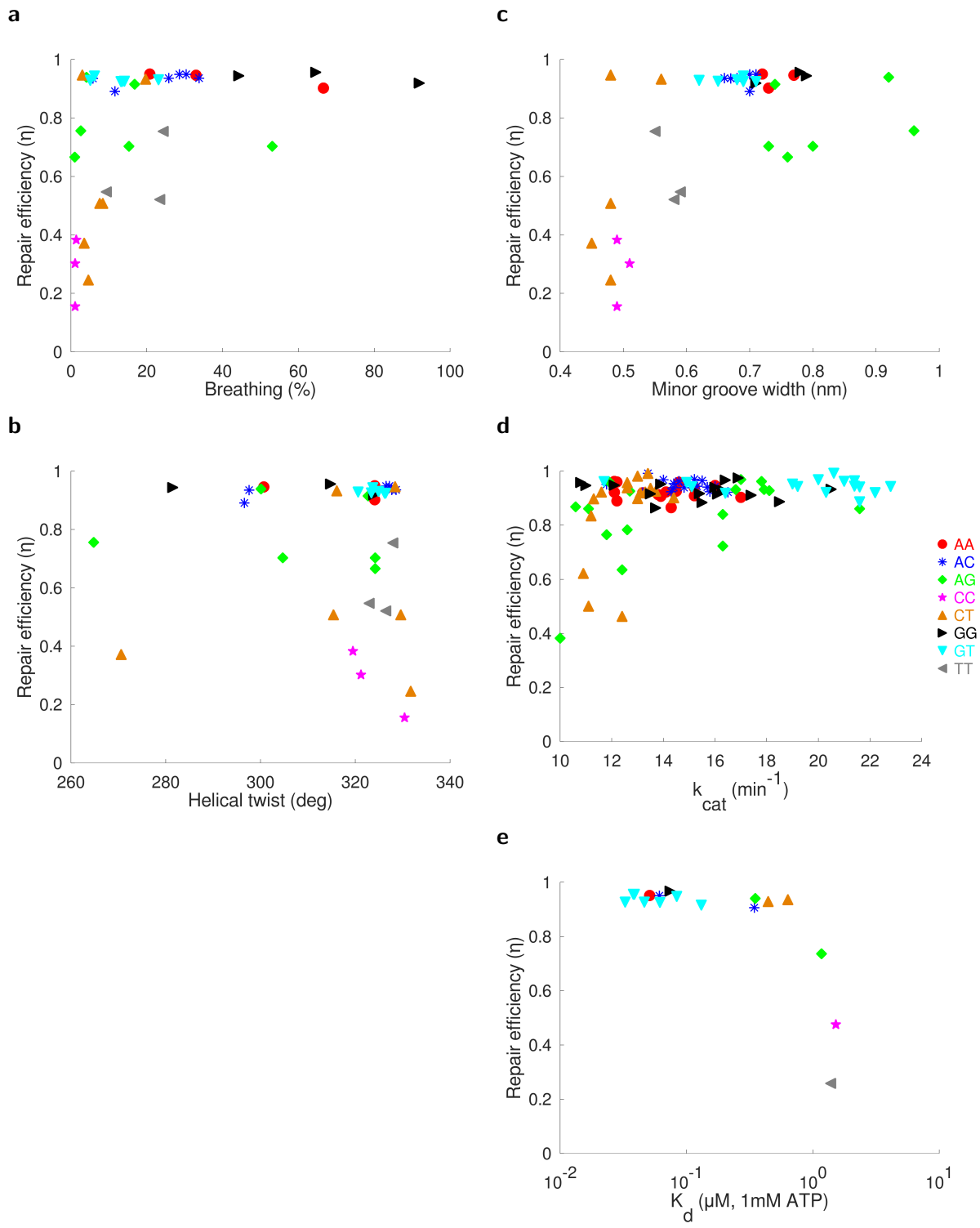



**Figure 2.32:** The comparison of repair results with biophysical properties of DNA and MutS. (a,b,c) The apparent repair efficiency compared with structural parameters obtained from MD simulations on DNA carrying mismatches [78]. Comparison of scaled repair efficiency ($\eta_s$) by considering the next-neighbour sequence context measured in 5MLx vs breathing, helical twist and minor groove width, respectively. (d,e) All inefficiently repaired mismatches have low $k_{cat}$, but a high $K_d$ for human MutS homologue [70].

**Table 2.4:** Full list of all experiments using the libraries that make use of the double barcoding strategy.

| Exp No | cell strain | vector backbone | library | Date sequenced | i7 index | i5 index |
|---|---|---|---|---|---|---|
| 1 | wt | tUNC19 | 3ML1 | MiSeq03102019 | N701 | S517 |
| | | | | HiSeq16102019 | | |
| 2 | wt | tUNC19 | 3ML1 | MiSeq03102019 | N702 | S517 |
| | | | | HiSeq16102019 | | |
| 3 | $\Delta mutS$ | tUNC19 | 3ML1 | MiSeq03102019 | N703 | S517 |
| | | | | HiSeq16102019 | | |
| 4 | $\Delta mutS$ | tUNC19 | 3ML1 | MiSeq03102019 | N704 | S517 |
| | | | | HiSeq16102019 | | |
| 5 | wt | hm19 | 3ML1 | HiSeq16102019 | N705 | S517 |
| 6 | wt | hm19 | 3ML1 | HiSeq16102019 | N705 | S517 |
| 7 | wt | 91CNUt | 3ML1 | HiSeq16102019 | N710 | S517 |
| 8 | wt | 91CNUt | 3ML1 | HiSeq16102019 | N711 | S517 |
| 9 | wt | short19 | 3ML1 | HiSeq09012020 | N710 | S504 |
| 10 | wt | short19 | 3ML1 | HiSeq09012020 | N711 | S504 |
| 11 | $\Delta mutS$ | hm19 | 3ML1 | HiSeq09012020 | N701 | S505 |
| 12 | $\Delta mutS$ | hm19 | 3ML1 | HiSeq09012020 | N702 | S505 |
| 13 | $\Delta recA$ | tUNC19 | 3ML1 | HiSeq19022020 | N701 | S517 |
| 14 | $\Delta recA$ | tUNC19 | 3ML1 | HiSeq19022020 | N702 | S517 |
| 15 | $\Delta dam$ | tUNC19 | 3ML1 | HiSeq19022020 | N703 | S517 |
| 16 | $\Delta dam$ | tUNC19 | 3ML1 | HiSeq19022020 | N714 | S517 |
| 17 | wt | mm19 | 3ML1 | HiSeq19022020 | N707 | S517 |
| 18 | wt | mm19 | 3ML1 | HiSeq19022020 | N710 | S517 |
| 19 | wt | tUNC19-X | 3ML1 | HiSeq19022020 | N711 | S517 |
| 20 | wt | tUNC19-X | 3ML1 | HiSeq19022020 | N712 | S517 |
| 21 | wt | tUNC19-X | 3ML1 | HiSeq19022020 | N704 | S517 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 22 | wt | tUNC19 | 3CL1 | HiSeq19022020 | N705 | S517 |
| 23 | wt | tUNC19 | 3CL1 | HiSeq19022020 | N706 | S517 |
| 24 | wt | tUNC19 | 3CL1 | HiSeq19022020 | N715 | S517 |
| 25 | $\Delta dam$ | hm19 | 3ML1 | HiSeq10092020 | N701 | S508 |
| 26 | $\Delta dam$ | hm19 | 3ML1 | HiSeq10092020 | N702 | S508 |
| 27 | $\Delta mutL$ | tUNC19 | 3ML1 | HiSeq10092020 | N703 | S508 |
| 28 | $\Delta mutL$ | tUNC19 | 3ML1 | HiSeq10092020 | N704 | S508 |
| 29 | $\Delta mutH$ | tUNC19 | 3ML1 | HiSeq10092020 | N705 | S508 |
| 30 | $\Delta mutH$ | tUNC19 | 3ML1 | HiSeq10092020 | N706 | S508 |
| 31 | $\Delta uvrD$ | tUNC19 | 3ML1 | HiSeq10092020 | N707 | S508 |
| 32 | $\Delta uvrD$ | tUNC19 | 3ML1 | HiSeq10092020 | N710 | S508 |
| 33 | $\Delta dam$ | tUNC19 | 3ML1 | HiSeq10092020 | N711 | S508 |
| 34 | $\Delta dam$ | tUNC19 | 3ML1 | HiSeq10092020 | N712 | S508 |
| 35 | $\Delta mutH$ | hm19 | 3ML1 | HiSeq10092020 | N714 | S507 |
| 36 | $\Delta mutH$ | hm19 | 3ML1 | HiSeq10092020 | N715 | S507 |
| 37 | $\Delta mutS\ \Delta dam$ | tUNC19 | 3ML1 | HiSeq05012021 | N701 | S502 |
| 38 | $\Delta mutS\ \Delta dam$ | tUNC19 | 3ML1 | HiSeq05012021 | N702 | S502 |
| 39 | $\Delta mutS\ \Delta dam$ | mm19 | 3ML1 | HiSeq05012021 | N703 | S502 |
| 40 | $\Delta mutS\ \Delta dam$ | mm19 | 3ML1 | HiSeq05012021 | N704 | S502 |
| 41 | $\Delta mutS\ \Delta dam$ | hm19 | 3ML1 | HiSeq05012021 | N705 | S502 |
| 42 | $\Delta mutS\ \Delta dam$ | hm19 | 3ML1 | HiSeq05012021 | N706 | S502 |
| 43 | $\Delta mutL\ \Delta dam$ | tUNC19 | 3ML1 | HiSeq05012021 | N707 | S502 |
| 44 | $\Delta mutL\ \Delta dam$ | tUNC19 | 3ML1 | HiSeq05012021 | N710 | S502 |
| 45 | $\Delta mutL\ \Delta dam$ | mm19 | 3ML1 | HiSeq05012021 | N711 | S502 |
| 46 | $\Delta mutL\ \Delta dam$ | mm19 | 3ML1 | HiSeq05012021 | N712 | S502 |
| 47 | $\Delta mutL\ \Delta dam$ | hm19 | 3ML1 | HiSeq05012021 | N714 | S502 |
| 48 | $\Delta mutL\ \Delta dam$ | hm19 | 3ML1 | HiSeq05012021 | N715 | S502 |
| 49 | $\Delta mutH\ \Delta dam$ | tUNC19 | 3ML1 | HiSeq11012021 | N701 | S503 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 50 | $\Delta mutH\ \Delta dam$ | tUNC19 | 3ML1 | HiSeq11012021 | N702 | S503 |
| 51 | $\Delta mutH\ \Delta dam$ | mm19 | 3ML1 | HiSeq11012021 | N703 | S503 |
| 52 | $\Delta mutH\ \Delta dam$ | mm19 | 3ML1 | HiSeq11012021 | N704 | S503 |
| 53 | $\Delta mutH\ \Delta dam$ | hm19 | 3ML1 | HiSeq11012021 | N705 | S503 |
| 54 | $\Delta mutH\ \Delta dam$ | hm19 | 3ML1 | HiSeq11012021 | N706 | S503 |
| 55 | $\Delta uvrD\ \Delta dam$ | tUNC19 | 3ML1 | HiSeq11012021 | N707 | S503 |
| 56 | $\Delta uvrD\ \Delta dam$ | tUNC19 | 3ML1 | HiSeq11012021 | N710 | S503 |
| 57 | $\Delta uvrD\ \Delta dam$ | mm19 | 3ML1 | HiSeq11012021 | N711 | S503 |
| 58 | $\Delta uvrD\ \Delta dam$ | mm19 | 3ML1 | HiSeq11012021 | N712 | S503 |
| 59 | $\Delta uvrD\ \Delta dam$ | hm19 | 3ML1 | HiSeq11012021 | N714 | S503 |
| 60 | $\Delta uvrD\ \Delta dam$ | hm19 | 3ML1 | HiSeq11012021 | N715 | S503 |
| 61 | wt + exoIII + exoVII | tUNC19 | 5ML1 | MiSeq27112019 | N703 | S504 |
| 62 | wt + exoIII + exoVII | tUNC19 | 5ML1 | N/A | N704 | S504 |
| 63 | Same sample rS1-rS2 | tUNC19 | 5ML1 | MiSeq27112019 | N705 | S504 |
| 64 | Same sample rS1-rS2 | tUNC19 | 5ML1 | N/A | N706 | S504 |
| 65 | wt | tUNC19 | 5ML1 | HiSeq09012020 | N701 | S504 |
| | | | | MiSeq27112019 | | |
| 66 | wt | tUNC19 | 5ML1 | HiSeq09012020 | N702 | S504 |
| 67 | wt | tUNC19 | 5ML2 | HiSeq09012020 | N701 | S503 |
| 68 | wt | tUNC19 | 5ML2 | HiSeq09012020 | N702 | S503 |
| 69 | wt | tUNC19 | 5ML3 | HiSeq09012020 | N701 | S502 |
| | | | | MiSeq27112019 | | |
| 70 | wt | tUNC19 | 5ML3 | HiSeq09012020 | N702 | S502 |
| 71 | wt | tUNC19 | 5ML4 | HiSeq09012020 | N707 | S502 |
| 72 | wt | tUNC19 | 5ML4 | HiSeq09012020 | N712 | S502 |
| 73 | wt | tUNC19 | 5ML5 | HiSeq09012020 | N710 | S502 |
| 74 | wt | tUNC19 | 5ML5 | HiSeq09012020 | N711 | S502 |
| 75 | wt | tUNC19 | 5ML6 | HiSeq09012020 | N703 | S503 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 76 | wt | tUNC19 | 5ML6 | HiSeq09012020 | N704 | S503 |
| 77 | wt | tUNC19 | 5ML7 | HiSeq09012020 | N714 | S502 |
| 78 | wt | tUNC19 | 5ML7 | HiSeq09012020 | N715 | S502 |
| 79 | wt | tUNC19 | 5ML8 | HiSeq09012020 | N705 | S503 |
| 80 | wt | tUNC19 | 5ML8 | HiSeq09012020 | N706 | S503 |
| 81 | wt | tUNC19 | 5ML9 | HiSeq09012020 | N707 | S503 |
| 82 | wt | tUNC19 | 5ML9 | HiSeq09012020 | N710 | S503 |
| 83 | wt | tUNC19 | 5ML10 | HiSeq09012020 | N703 | S502 |
| | | | | MiSeq27112019 | | |
| 84 | wt | tUNC19 | 5ML10 | HiSeq09012020 | N704 | S502 |
| 85 | wt | tUNC19 | 5ML11 | HiSeq09012020 | N705 | S502 |
| 86 | wt | tUNC19 | 5ML11 | HiSeq09012020 | N706 | S502 |
| 87 | wt | tUNC19 | 5ML12 | HiSeq09012020 | N711 | S503 |
| 88 | wt | tUNC19 | 5ML12 | HiSeq09012020 | N712 | S503 |
| 89 | wt | tUNC19 | 5ML13 | HiSeq09012020 | N714 | S503 |
| 90 | wt | tUNC19 | 5ML13 | HiSeq09012020 | N715 | S503 |
| 91 | wt | tUNC19 | SSL1 | HiSeq18082020 | N701 | S507 |
| 92 | wt | tUNC19 | SSL1 | HiSeq18082020 | N702 | S507 |
| 93 | wt | tUNC19 | SSL1 | HiSeq18082020 | N705 | S507 |
| 94 | $\Delta mutS$ | tUNC19 | SSL1 | HiSeq18082020 | N706 | S507 |
| 95 | wt | tUNC19 | SSL2 | HiSeq18082020 | N701 | S508 |
| 96 | wt | tUNC19 | SSL2 | HiSeq18082020 | N702 | S508 |
| 97 | wt | tUNC19 | SSL2 | HiSeq18082020 | N705 | S508 |
| 98 | $\Delta mutS$ | tUNC19 | SSL2 | HiSeq18082020 | N706 | S508 |
| 99 | wt | tUNC19 | SSL3 | HiSeq18082020 | N701 | S510 |
| 100 | wt | tUNC19 | SSL3 | HiSeq18082020 | N702 | S510 |
| 101 | wt | tUNC19 | SSL3 | HiSeq18082020 | N705 | S510 |
| 102 | $\Delta mutS$ | tUNC19 | SSL3 | HiSeq18082020 | N706 | S510 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 103 | wt | tUNC19 | SSL4 | HiSeq18082020 | N701 | S511 |
| 104 | wt | tUNC19 | SSL4 | HiSeq18082020 | N702 | S511 |
| 105 | wt | tUNC19 | SSL4 | HiSeq18082020 | N705 | S511 |
| 106 | $\Delta mutS$ | tUNC19 | SSL4 | HiSeq18082020 | N706 | S511 |

# Chapter 3

# In vivo tracking of insertion loop repair

## 3.1 Abstract

The focus of the previous two chapters was on DNA mismatches, which form by the improper pairing between the bases on opposite strands, hence violating the canonical Watson-Crick base pairing rule. As I have argued before, such problems might lead to mutations during semi-conservative DNA replication as the two strands carry conflicting information and such a conflict will be undetectable by the two DNA polymerases replicating the leading and lagging strands of DNA independently. The DNA replication machinery can also inadvertently add or skip nucleotides during strand elongation, generating insertion/deletion loops where the number of nucleotides on the two complementary strands will this time be unbalanced. Cellular systems are capable of repairing such DNA defects as well, characterization of which will be the topic of this chapter. Using a similar strategy to DNA mismatches, we studied in our setup all possible DNA insertions in all nearest neighbor sequence contexts and our results indicate that all insertion loops are repaired with a very high efficiency regardless of the sequence context. While repair happened exclusively by the retention of the shorter strand if there are insertion loops with two nucleotides, the strand choice was random if only one extra nucleotide was inserted at a time.

## 3.2 Results

Although generally observed at lower rates *in vivo* than mismatches, an insertion loop can also arise due to replication slippage [3]. DNA polymerases also generate deletion errors as well as insertion errors, albeit at about two orders of magnitude less frequently than mismatches [25, 81]. The same MMR pathway characterized in the previous chapters is also capable of triggering an efficient repair response against the insertion loops of one or more bases, although the affinity of MutS significantly drops for longer insertions and it has been reported that no repair is triggered by insertion loops made of 5 or more extra nucleotides [16, 57, 69]. We here asked if this repair efficiency of insertion/deletion loops depend on the base identity of the extra unpaired nucleotide or its local sequence context.

### 3.2.1 Insertion library construction by oligo annealing

As the simpler method we have available, we first resorted to generation of insertion libraries using an analogous strategy to Chapter 1 and experimented with insertion loop containing specimen obtained by annealing of two short DNA strands that are imperfect reverse-complementary of each other. To be able to observe the insertion type and the local sequence context dependence of the cellular repair response, we experimented with dsDNA libraries that carry one and only one insertion per molecule by experimental design. To form such an insertion loop library in a simple and cost-affordable way, we annealed fully pre-synthesized commercially purchased oligos to each other. We adopted the same consensus sequence as SML mentioned in Chapter 1 as the consensus sequence of this insertion library (IL), while the reverse-complementary of this consensus sequence served as the common strand of the library that was shared among all members of the insertion library. The variable strand of the library follows the base sequence of the consensus sequence, except that between two arbitrarily chosen consecutive bases, it contains an extra nucleotide.

We chose this extra nucleotide to be a degenerate base N={A, C, G, T}, each incorporated at this position at about equal probability during chemical synthesis. Each oligo representing a variable strand can carry the degenerate base N at one and only one position, but would otherwise exactly follow the reverse complementary of the sequence of the common strand, thereby leaving exactly one nucleotide of the variable strand unpaired upon annealing. This means each obtained oligo will sample all 4 insertion types that is possible at that location and hence reducing the number of variable strands to be purchased by a factor of 4. In total contrast to the mismatch library construction case, where the proper complementary base of the common strand was to be avoided, insertion of any base at any position along the consensus sequence will result in an improperly annealed DNA that is a substrate for the repair machinery.

To simplify the downstream ligation step, we again formed compatible sticky ends at the termini of the mismatch libraries as on the termini of the barcoded vectors we generated (SacI and XhoI). For this, we opted to make the common strands longer by 4 bases than the variable strand at both termini, where the unpaired sequence tetramers are complementary to the overhangs generated by the restriction enzymes (5'-AGCT-3' and 3'-AGCT-5'). To experimentally probe the repair

properties of an n-base long sequence, it suffices to procure 1 n+8 base long oligo to serve as the common strand of the library and n-1 many n+1 base long variable strands. We first made an equimolar mixture of these n-1 variable ssDNA strands and annealed with the common strand by slow cooling from 98°C to room temperature on a thermal cycler. We ligated thus formed mismatch library with the barcoded vector library to form circularized plasmids each carrying one unique random tag and one insertion at an arbitrary position in the proximity of the barcode (Figure 1.1).

### 3.2.2 Classification of clans

The presence of a unique tracing barcode on each and every single molecule of the extra inserted nucleotide carrying plasmid makes it possible to deduce the clans in a similar way to mismatches by density based clustering, which as an output provides clans as groups of reads that are descendants of the same ancestor molecule. By checking whether the sequence composition of these clans are heterogenous or homogeneous, the efficiency of the cellular response against various insertion loops can also be observed following a similar overall strategy as for the mismatches, albeit with some minor differences in the expected base distribution histograms of the clans that contain contributions from the variable strand.

Replication of a mismatch carrying DNA as well as an extra inserted base carrying plasmid can both produce broadly three types of sequences: C-, U- or V-type clans. A C-type clan represents a repair event that leads to the complete elimination of the variable strand and hence the ensemble averaged sequence of the clan would closely follow the consensus sequence of the mismatch library (Figure 3.1a). As such, while being accurately detectable, it is not possible to deduce any information about the insertion in the ancestor DNA molecule and we had to exclude C-type clans from the rest of the analysis in both cases.

V-type clans also represent repair events, but the repair has occurred by retaining the variable strand of the library and the ensemble-averaged sequence of the clan can be used to infer the position of the inserted base as well as its identity (Figures 3.1b and 3.1c). Out of a mismatch library, we would expect this averaged sequence to deviate from the consensus sequence at one and only one position, whereas the introduction of insertion loops would cause a frame shift with

respect to the common strand. All of the bases following the insertion position will be shifted by 1 base towards the barcode-distal side, starting at the position of the insertion in the ancestor molecule. The identity and position of this first shifted base provides a means to deduce the identity and the position of the extra base in the ancestor DNA. To determine this in practice, we subtract the expected base composition histogram based on the consensus sequence of the library from the actual base composition histogram that was experimentally observed. That is,

$$d_{ij} = \sum_{\forall k \in clan} \delta_{i,k_j} - \delta_{i,c_j} \tag{3.1}$$

and

$$p \equiv min\left(\{j | d_{ij} > t_{low} = 0.1\}\right); \tag{3.2}$$

$$b \equiv \underset{i}{argmax}\ d_{ip} \tag{3.3}$$

where the inserted b-base is after the p'th position of the consensus sequence based on the difference histogram matrix d, constructed according to the experimentally obtained sequencing reads k that have been assigned to the clan of interest during clustering. In a successfully repaired clan, either all or none of the members are expected to display such a shift, while U-type clans, in which a repair event has not taken place in time, are expected to contain a roughly equal ratio of the two possible types of replication products such that only half of the clan members will appear shifted (Figures 3.1d and 3.1e). To make an algorithmic binary classification decision between the U or V-type clans, we can use the insertion prevalence ratio ($d_{bp}$). To improve the accuracy of this step against sequence dependent frequency of sequencing errors, we performed a curve fitting on the base composition histograms and obtained a corrected value for the insertion prevalence by providing $s = d_{bp}$ as the initial guess, viz.

$$f_{ij}(s|clan) = \sum_{\forall k \in clan} \left(\delta_{i,k_j} - (1-s)\delta_{i,c_j} - s\delta_{i,v_j^p}\right) \tag{3.4}$$

where f is the objective function to be minimized and $v_j^p$ in the third term is the base identity

on the p'th variable strand expected at position j. If a repair event happens via retention of the variable strand (V-type clan), the substitution is observable in above threshold fraction of the clan members ($d_{bp} \geq t_{high} = 90\%$) and we record a repair event by incrementing the matrix element $V_{bp}$ counting the V-type clans for this position and mismatch type. Similarly, an unrepaired mismatch would have an intermediate insertion prevalence ($10\% = t_{low} > d_{bp} > t_{high} = 90\%$) and we would increment the corresponding matrix element $U_{bp}$ that counts the U-type clans.

Using this methodology, we can deduce the type of the insertion in the ancestor plasmid. As an example, Figures 3.1b and 3.1c correspond to an insertion of a T and G, respectively, that were repaired before replication. The A insertions in Figures 3.1d and 3.1e, based on the bimodal nature of the clans, were not repaired. Due to experimental errors, clans that cannot be interpreted under this ternary classification scheme can arise, which we excluded from further analyses. In particular, an insertion/deletion error in a significant subpopulation of the clan members can shift the a portion of the clan members, causing the s parameter to significantly differ throughout the region of interest (Figure 3.1f). Similarly, an insertion/deletion error can cause the emergence of a third population, which likewise renders the results uninterpretable (Figure 3.1g).

### 3.2.3 Insertion - deletions loops are repaired efficiently

Using the above methodology, we can hence easily quantify the "insertion prevalence" in each clan analogous to the "substitution prevalence" concept we made use of for mismatch libraries before. By consulting the s-histogram for IL (Figure 3.3), one can readily deduce that the wt cells are capable of repairing insertion errors with a very high accuracy as the peak for the U-type clans is very scarcely populated in comparison to the equivalent histogram for mismatches (Figure 1.8a). Due to the inability to correctly account for the origin of the detected C-type clans, it is not possible to correctly estimate the repair efficiency of individual insertion loops in an unbiased way without the inclusion of mapping barcodes. Yet in Chapter 1, we adopted the calculable parameter defined in Equation 1.12, a workaround which we similarly adopted here as a comparative tool between different insertions (Figure 3.2b). While wt cells could repair an average insertion loop with an efficiency of $\mu \pm SEM = 0.92 \pm 0.01$, $\Delta mutS$ cells repaired with a statistically highly significantly

lower level of $0.51 \pm 0.00$ (N=157, $p < 10^{-5}$ by one tailed z-test), albeit the contribution of these other non-MutS mediated pathways or experimental sampling errors was considerably higher for the insertion library case in comparison to its mismatch library counterpart ($0.43 \pm 0.01$, $p < 10^{-5}$).

When all possible mismatches along a common sequence are sampled, 1 out of 4 base possibilities per each position leads to proper Watson-Crick pairing. In contrast, when an extra base is inserted to mimic an indel error, an analogous set of invalid matrix entries does not exist. However, when an insertion library is designed by introduction of each all four bases between each consecutive base doublets of a consensus sequence, some of the library members obtained will be identical. If ATCG is the example consensus sequence, addition of a T after the first base A or second base T both lead to the same variable strand ATTCG and the exact physical position of the insertion loop in the plasmid is hence spatially uncertain and might potentially be dynamic. When the output is represented in an analogous matrix format, where the x-axis displays the sequence of the common strand and the y-axis the bases inserted between the bases at that position, some of the matrix entries are systematically redundant (Figure 3.2a). Due to this uncertainty, even though the figures in this work will report identical repair efficiency values redundantly at all such positions, it should be understood that no claim regarding a well-defined physical location for the inserted nucleotides is implied.

Having said so, these values should be interpreted cautiously, because in Chapter 1, we calculated the repair efficiency of mismatches by assuming that -both plasmid strands harboring or lacking methylation marks- the two strands of DNA are indistinguishable to the repair machinery and hence there is no significant strand choice bias. This assumption allowed the computation of a repair efficiency via Equation 1.12. While it might not provide sufficient information to guide the correct repair action, an insertion-deletion defect is inherently asymmetric. This means that the assumptions leading to 1.12 do not hold, a shortcoming which we will again address by using mapping barcodes.

### 3.2.4 Double barcoding strategy to track insertion loops

Using the same strategy as in Chapter 2 (Figure 3.4), we generated a barcoded plasmid library each member of which carries one and only one insertion by the design of the starting oligo library. In two different sets of experiments, we sampled either all single nucleotide insertions (SIL) or two consecutive nucleotide insertions (DIL) possible along the 66 bp De Bruijn sequence containing all sequence triplets. This sequence is identical to the consensus sequence of the 3ML1 library, hence representing all sequence triplets exactly once. Again making use of the mapping barcodes, the insertion position as well as the identity of the inserted extra nucleotide can be deduced in an unbiased way via the pre-defined relationship by the construction of the oligo pool we have designed. We evaluated the repair fate in a similar way as for mismatches, except that the clans that has preserved the variable strand appear as shifted as was described in Section 3.2.2, rather than displaying an apparent point substitution.

In agreement with our previous results on single-barcoded IL mismatch library, we observed that wt cells were able to repair virtually all mismatches with a high efficiency, as the zone on the histogram that corresponds to U-type clans is almost unpopulated (red curve in Figure 3.5a). The V-peak was only slightly higher than the C-peak, suggesting that the strand choice bias exists but it is far from being complete. $\Delta mutS$ cells, on the other hand (blue dashed curve) generated a noticeably higher proportion of U-type clans than wt cells, albeit it was still less than its mismatch library counterpart. As a control, we repeated the same analysis procedure on the 3CL1 control sample, whose base sequence is identical to the 3ML1, but does not carry any insertion loops nor any mismatches, and therefore is not subject to any repair. Just as expected, 3CL1 in wt cells gave rise to only C-peak elements with near-complete penetration, suggesting that the the erroneous detection of clans with insertions due to experimental errors is very low (black curve).

Insertion loops containing two consecutive nucleotides were also well-repaired in wt cells (red curve in Figure 3.5b), but in contrast to single nucleotide insertions, the repair displayed a strong strand choice bias favoring the retention of the common strand. In fact, this preference to keep shorter strand was complete within the experimental error margin, as the histogram obtained for DIL and 3CL1 essentially overlapped. Surprisingly, we observed this high repair with strong strand

choice bias towards the common strand to be the case even in $\Delta mutS$ cells, in which MMR is inoperative. This is in contrast with the SIL case, where the presence of a functional MMR made a distinguishable difference in the histograms, albeit the extent of the difference between the two cell strains was small.

### 3.2.5 Repair of insertion-deletion loops is sequence independent

Next, we investigated the potential insertion type and the sequence context dependence of the repair response mounted against inserted extra nucleotides in the SIL and DIL libraries. We hypothesized that the plasmids carrying an insertion that have not undergone a repair would give rise to a mixed clan containing both shifted and unshifted sequences, and hence we used this insertion prevalence parameter to decide on a clan-by-clan basis if an individual molecule was repaired. We first calculated an ensemble-averaged mapping barcode, which we searched in the look-up table to deduce the position of the inserted base and its identity, or equivalently the expected sequence of the variable strand. We observed that the accuracy of the apparent mismatch type assignment based on these pre-determined mapping barcodes is lower in the insertion case than the mismatch case (Figure 2.3 vs Figure 3.7). In particular, the mapping accuracy both for the type and position of the insertion loop was lower in wt cells than in $\Delta mutS$ cells for SIL, suggesting the possibility that this these apparent replication errors might be a real outcome of an error-prone repair process on the introduced DNA lesions, rather than being fully explainable by the batch to batch variation of oligo array synthesis or DNA amplification.

We then considered each read that is a member of the clan individually, and assigned the read to be a representation of the common or variable strand, depending to which the read's Hamming distance was smaller. During this procedure, we disregarded the reads that are farther away from both the common and variable strands by more than 2 substitutions. Processing of each clan member in this fashion provides a measure regarding the insertion prevalence, which is the fraction of variable strands detected with this method. For simplicity, we considered clans whose insertion prevalence is in [0, 0.1], (0.1, 0.9) and [0.9, 1] to be C, U and V-type clans, respectively. In this manner, we analyzed the entire dataset clan by clan, classifying each clan individually, and

incrementing the relevant matrix element that keeps the record of C, U or V type clans per each insertion type and position by considering the information provided through the mapping barcode. This procedure provides 3 matrices that count the C, U, or V-type clans, hence enabling the usage of the repair efficiency ($\eta$) defined by Equation 2.4, as before.

Applying this strategy, the apparent repair efficiencies of individual insertion loops that we obtained are shown in Figures 3.5c and 3.5d. In wt cells, we observed virtually all single and double insertion loops to be repaired with a very high efficiency ($\eta = 0.91 \pm 0.04$ and $\eta = 1.00 \pm 0.00$, respectively). In comparison to mismatches, we observed the contribution by the canonical MMR to the apparent repair efficiency to be less significant for single nucleotide insertion loops, as the repair efficiencies observed in $\Delta mutS$ cells were significantly lower than their wt counterparts ($0.64 \pm 0.11$, p-value $< 10^{-5}$ by one tailed z-test), but still higher than that observed for the repair efficiency of 3ML1 mismatch library ($0.43 \pm 0.06$, p-value $< 10^{-5}$). The contrast between the MMR capable and incapable cells was practically non-existent for the double nucleotide insertion case as the apparent repair efficiency in $\Delta mutS$ cells was also near-complete ($1.00 \pm 0.00$). While the overall position dependence of these raw repair efficiencies is more subtle in the insertion library than the mismatch case, we observe a very high repair efficiency at the barcode-proximal end for SIL, with or without a functional MMR (Figure 3.5c and Figure 3.6). In contrast, out of a starting material uniformly sampling insertions at all positions, we detected only clans whose ancestral plasmids contained a two-base insertion loop at the barcode-distal terminal of the library, but no clans with an insertion loop at the barcode-proximal side (Figure 3.5d and Figure 3.6). While we do not have unequivocal evidence supporting this hypothesis, the systematic under-representation of barcode-proximal insertions among the all detected clans and their above average higher efficiency might be explained by a potential incomplete extension and ligation during our library preparation procedure, rendering some of our constructs prone to nuclease attacks in the cytoplasm.

## 3.3 Conclusion

In this Chapter, I described the repair response we observed using barcoded plasmids carrying insertion/deletion loops. We observed that in wt cells, almost all insertion loops are repaired very

efficiently, and although the measured values slowly decrease from left to right as was observed in other libraries, variation between different insertion types is quite small. The very high repair efficiency of wt cells was not only limited to single nucleotide insertion loops, but also double nucleotide insertions that we studied via our double-barcoded oligo pool based mismatch library generation strategy. This high efficiency of repair could be a result of the high affinity of MutS to DNA carrying insertion-deletion loops [16, 82]. While MutS-deficient cells generated more unrepaired clans than wt cells, the apparent efficiency of the repair response against insertion loops was near-complete, suggesting a potential involvement of alternative cellular response mechanisms. Some of these observations might be related to nuclease attacks on the foreign constructs that we introduced into the cytoplasm. Alternatively, it could be explained by our failure to obtain the claimed constructs with insertion loops due to a potential shortcoming of our library preparation protocol, most likely failure of the Klenow polymerase to elongate the annealed constructs across the mapping barcodes due to the presence of insertion loops close by. If such a polymerase-dependent artifact exists, it is likely related to polymerase binding rather than DNA polymerase induced loop removal as this step of our workflow does not involve any DNA polymerases that are capable of performing 3' to 5' proofreading activity. Furthermore, the relative occupancy levels of C-U-V type clans were comparable with the single barcoded oligo annealing based protocol, which does not involve any polymerase activity for the synthesis of the insertion loop bearing constructs.

For single nucleotide insertions, the cells did not have a very strong preference as to which strand is to be kept during the repair and produced roughly equal proportions of C- and V-type clans despite the inherent asymmetry of an indel loop. Interestingly, however, the introduction of two consecutive extra bases as part of the variable strands lead to a strong strand bias that favored the retention of the shorter common strand, i.e. the C-type clans. While the repair of this double insertion library was near-complete, the repair appeared virtually exclusively through the retention of the common strand, whereas no significant quantity of U- or V-type clans could be detected.

I would like to point your attention to the fact that as seen by our experimental setup, the two DNA libraries with extra one or two nucleotides inserted on the variable strand effectively also

generates a data set that properly contains the data set which studies all possible deletions on the common strand. As such, these measurements can be viewed as the characterization of the repair efficiency of deletion loops. In the biological world, the propagation of such DNA polymerase extra-incorporation errors as indel mutations have an even greater potential to harm the fitness of the organism, as it might lead to the accumulation of truncated dysfunctional proteins. In that regard, it would be a beneficial trait for the cell to be able to detect and repair insertion loops with a very high efficiency. The strong strand choice bias of the cell upon encountering multiple inserted extra nucleotides might perhaps be explained in this evolutionary context. Given the high fidelity of the DNA replication machinery, the probability of generating multiple false insertions in a row would be very improbable. Yet, considering that the reported frequency of committing deletion errors is about two orders of magnitude lower than an insertion error, an encountered two-base imbalance between the two annealed strands is about 4 orders of magnitude more likely to be an insertion error than being caused by accidental omission of two consecutive bases during the replication of the other strand. Considering this, a cellular machinery that has evolved to degrade the strands with extra unpaired nucleotides might confer the cell an evolutionary advantage by keeping the mutation rates at bay.

## 3.4 Materials and methods

### 3.4.1 Generation of double barcoded mismatch libraries

Here, we applied the same approach and the experimental procedure as in Chapter 2 to the insertion libraries. However, rather than a point substitution, each variable strand of SIL contains 1 extra nucleotide with respect to the consensus sequence of 3ML1. The position and type (A, C, G or T) of the added base varies along the library, hence sampling all possible insertions. SIL and DIL were purchased as a single synthetic oligo pool containing a total of 11220 oligos (Genscript, NJ). SIL contains 3300 oligos, 126 bases long each, and containing 7-base long mapping barcodes. DIL contains 7920 oligos, 128 base long each and containing 8 base-long mapping barcodes. In both cases, each mapping barcode is separated from the other mapping barcodes in the respective

library by at least 2 subsitution mutations by design. We selectively amplified SIL or DIL by using the directed primer pair (Z13 or Z14 along with Z6) in the emPCR stage, followed by a large scale standard PCR following the same protocol that was described for mismatches before.

### 3.4.2 C-, U-, V-type clan assignment in insertion libraries

While we followed a similar analysis workflow as in the mismatch libraries, the detection of insertion loops requires a slightly different algorithmic approach as the potential DNA sequences that can be obtained from each molecule are not related by a single base substitution with respect to a reference sequence (reverse complement of the sequence of the common strand), but rather reads that originate from the variable strand are shifted starting from the position of the insertion and onwards. To decide the fraction of clan members that represent the variable strand information, we first calculated the mapping barcode representing the ensemble of clan members by majority voting. Next, we refer to the look-up table to deduce the variable strand forming the insertion loop that carries this particular mapping barcode. Then we loop over the clan members read by read and for each individual read, we check its Hamming distance to the common strand and the mapped variable strand. We compare the two distances and assert that the read represents the strand to which its distance is smaller. If the read is differs from the common or variable strand sequence by more than 2 mutations, we regard the read as noise and omit. After processing all members constituting the clan, we compute the insertion prevalence (i), which is the fraction of total clan members that were determined to be derived from the variable strand.

Due to the indeterminacy about the physical location of the insertion loop, as well as observing that both the peaks at s=0 and s=1 are well defined, we did not attempt a position-dependent correction unlike the case for the mismatches. Instead, we directly assigned all clans with $s \in [0,\ 0.1]$ as C-type, $s \in (0.1,\ 0.9)$ as U-type, and $s \in [0.9,\ 1]$ as V-type. We reported the raw repair efficiency ($\eta$) using the counts of C, U, V-type clans for each insertion type as before (Equation 2.4). A pseudo-code of this routine can be found in Algorithm 8.

**Algorithm 8:** Repair efficiency calculation out of double barcoded insertion libraries

```
forwardReads ← Import all forward reads from file "sample#_R1.fastq"
reverseReads ← Import all reverse reads from file "sample#_R2.fastq"
foreach read ∈ reverseReads do
    read ← reverse-complement(read);
end
Fix minLength = Total length of tracking_barcode, SacI site, mapping barcode and insertion_library
Initialize acceptedReads = ∅
foreach full_read ∈ Union(forwardReads, reverseReads) do
    AdaptorEndPos ← Search the last base of the 5' adapter in full_read
    if full_read is shorter than minLength OR Adaptor not found OR AdaptorEndPos ∉[15,30] then
        Ignore the full_read
    end
    SeqOfInterest ← Extract the n-base long sub-sequence of full_read following AdaptorEndPos
    if HammingDistance(sequenceOfInterest, expectedLibraryPrototype) >5 then
        SeqOfInterest ← Shift SeqOfInterest by dynamic programming
    end
    if HammingDistance(SeqeOfInterest, expectedLibraryPrototype) ≤ 5 then
        Extract (tracing_barcode, insertion_probe) from SeqOfInterest using the expected library prototype
        Append (tracing_barcode, insertion_probe) to acceptedReads
    else
        Ignore the full_read
    end
    Export acceptedReads into sample#.csf file
end

Import acceptedReads from sample#.csf file
barcodeClusters ← DBSCAN(acceptedReads.tracing_barcode, ϵ = 3, N = 10)
foreach clan ∈ all_clans do
    if clan.mapping_barcode is invalid then
        clan = NULL
    else
        expected_variable_strand ← Refer to the database for clan.mapping_barcode
        foreach read ∈ clan do
            distC ← Hamming_distance(read, common_strand)
            distV ← Hamming_distance(read, expected_variable_strand)
            if distV <distC ∧ distV <2 then
                Increment Vcount
            end
            if distC <distV ∧ distC <2 then
                Increment Ccount
            end
        end
        totalCount ← Vcount + Ccount
        if totalCount ≥ 10 then
            Print: (Vcount, Ccount, expected_variable_strand)
        end
    end
end
```

## 3.5 Figures and tables

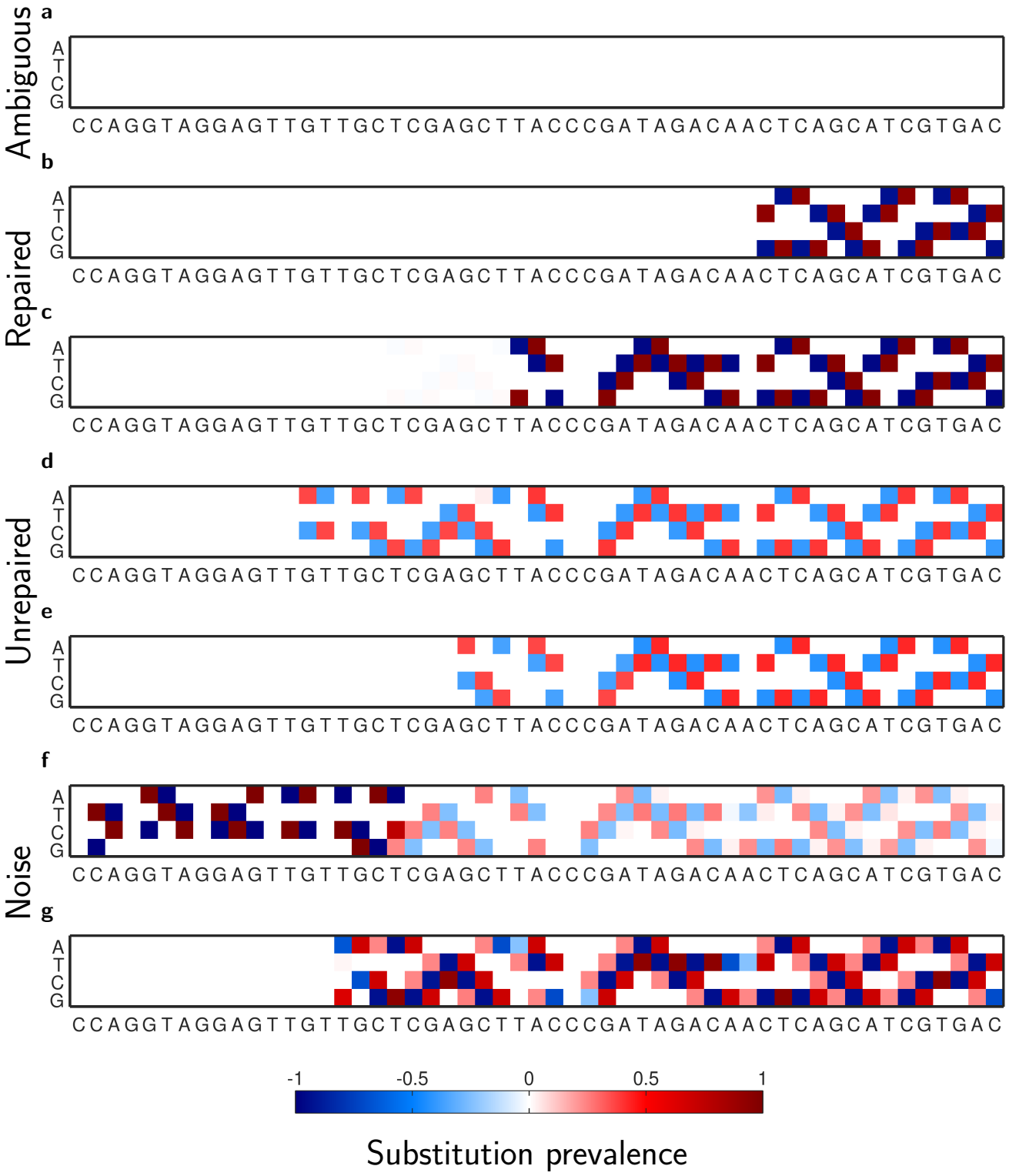



**Figure 3.1:** Detectable events expected from an insertion library, analogous to Figure 1.9. (a) An ambiguous clan in which the correction was made based on the common strand, due to which insertion type on the ancestor plasmid and the insertion position cannot be inferred. (b,c) Example clan histograms that were repaired according to the variable strand information. (d,e) About half of the reads constituting unrepaired clans are shifted from the consensus sequence, whereas repair event shifts all members. Starting position of the shift reveals the insertion position and type. (f,g) Uninterpretable clans due to experimental errors that are omitted.

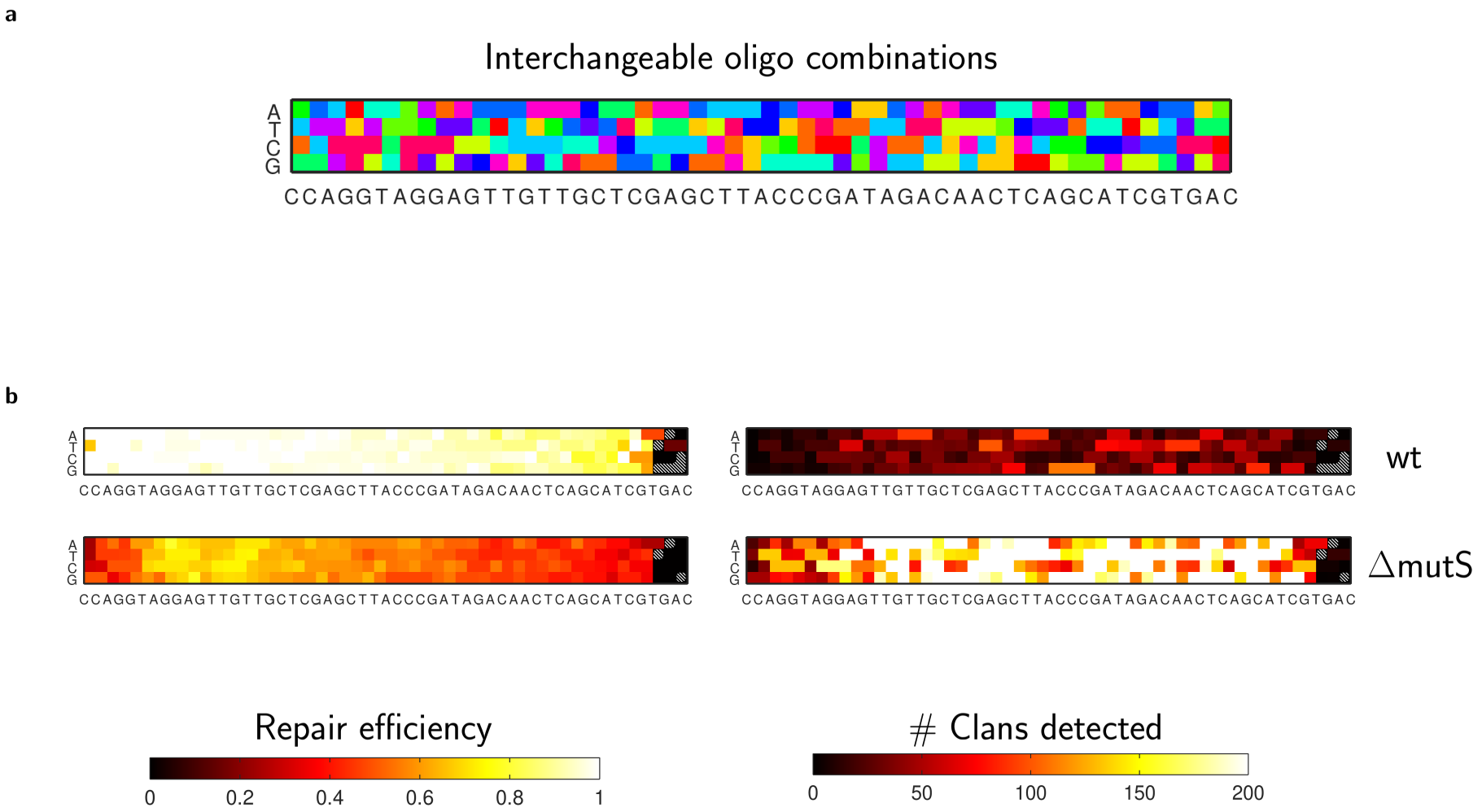



**Figure 3.2:** Repair efficiencies of insertion loop as part of IL following the same representation as in Figure 1.10. (a) Although all four base insertions lead to a physical loop to repair, some insertions can be de-localized to the neighboring positions, if indicated with the same color. (b) The measured repair efficiency of individual nucleotide insertions and the total number of clans detected for for an A,T,C,G insertion (y-axis) between two adjacent nucleotides along the 53 base long consensus sequence (x-axis). Data have been aggregated out of three experimental replicates.

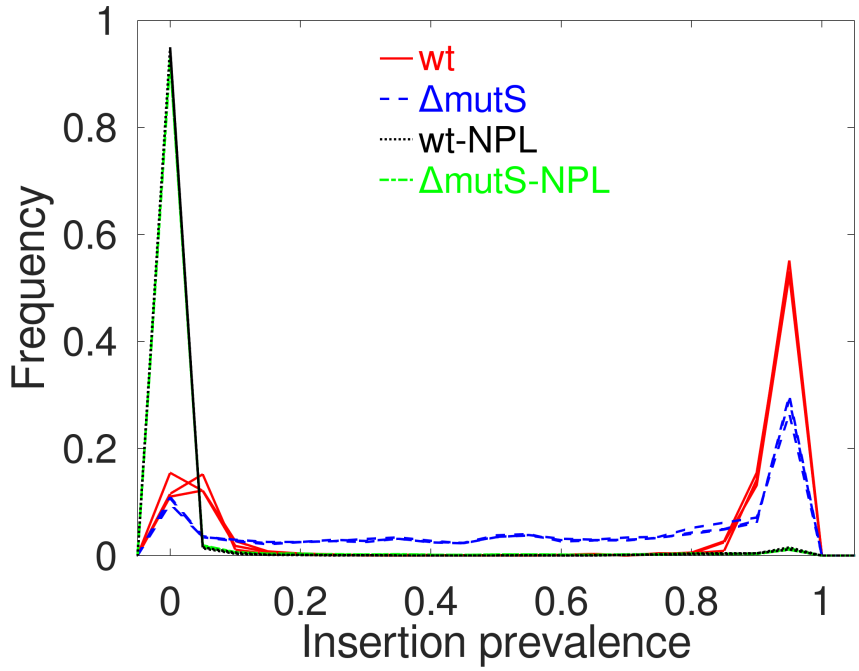



**Figure 3.3:** Insertion prevalence histogram of IL, wt (—) or $\Delta mutS$(- - -) cells. NPL measured in wt ($\cdots$) or $\Delta mutS$(- $\cdot$ -) cells are control samples without any insertion loops that have the same sequence as the common strand of IL. Each condition is represented by multiple curves representing independent experimental replicates.

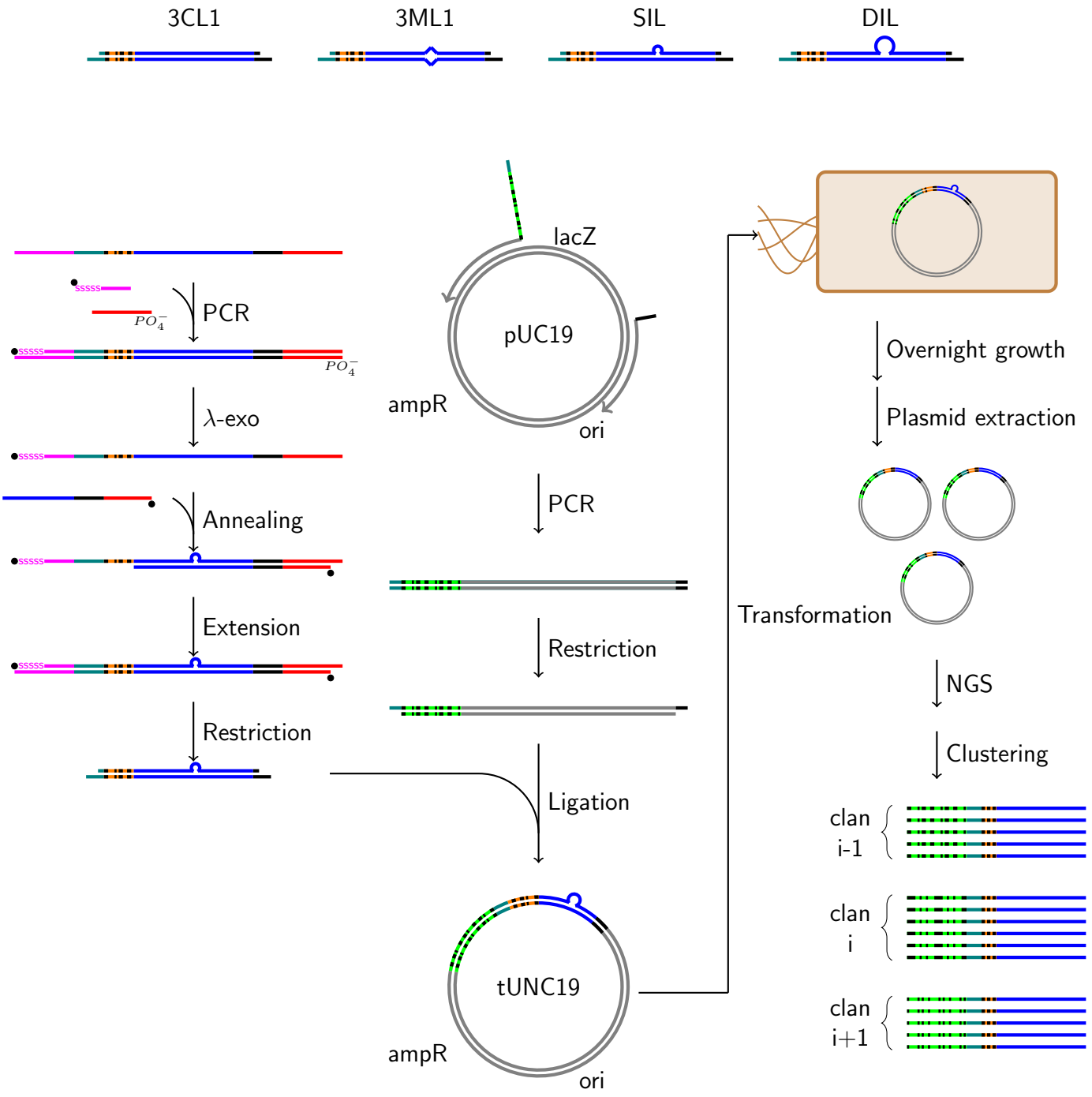



**Figure 3.4:** Illustrated summary of the experimental design, analogous to Figure 2.1. 3CL1 is a proper dsDNA that carries a De Bruijn sequence representing all sequence trimers, whereas 3ML1 contains one and only one mismatch along this sequence by design at an arbitrary location. SIL contains an extra nucleotide on the variable strand at an arbitrary location, hence generating an insertion loop upon annealing with the 3ML1 common strand, while DIL contains two extra adjacent nucleotides inserted between each two consecutive bases of the consensus sequence and leading to two nucleotide insertion loops.

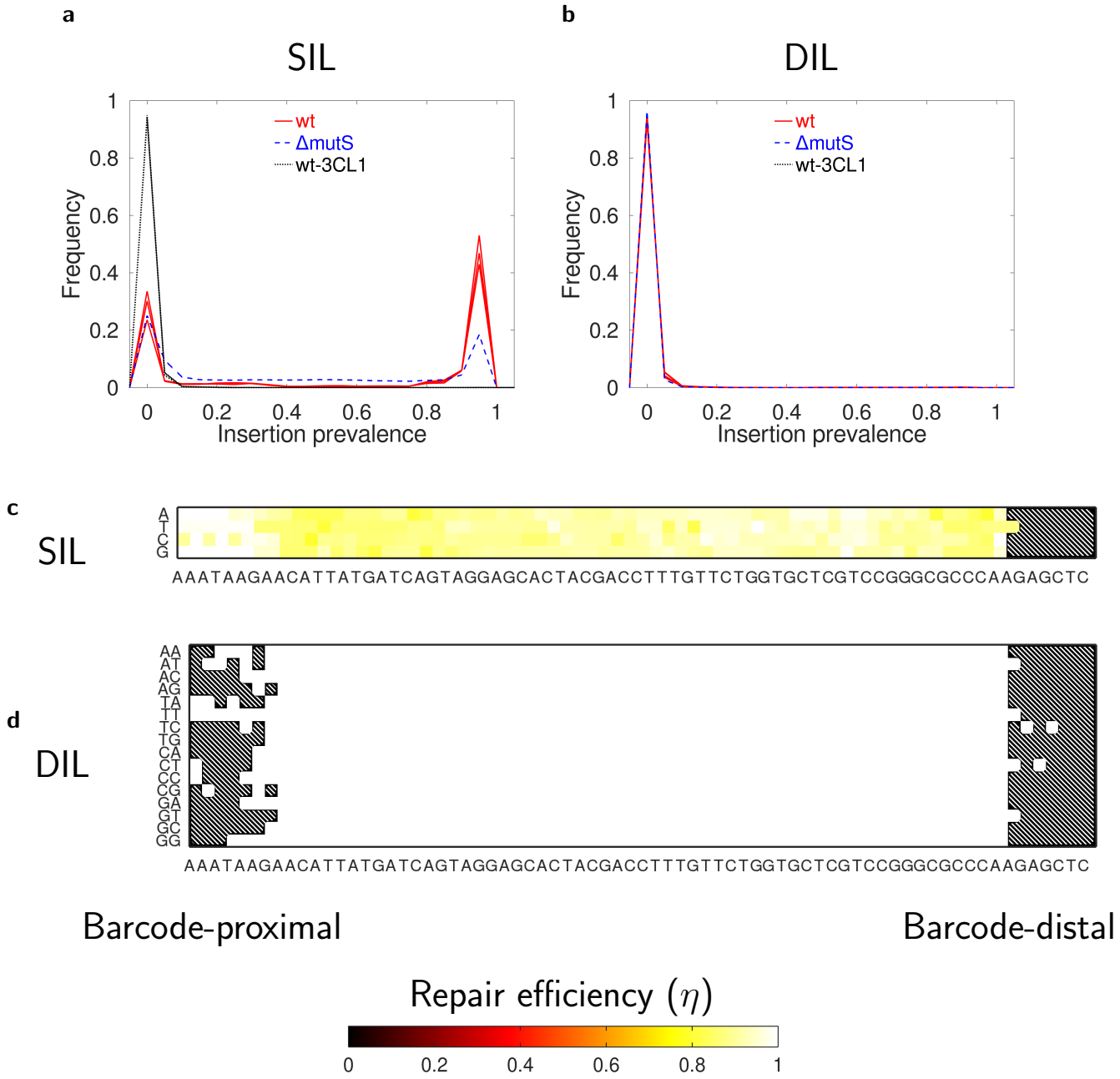



**Figure 3.5:** Insertion prevalence histograms of clans as a function of the position for single (SIL, a) or double base insertions (DIL, b). 3CL1 represents a control sample analyzed with the same workflow, carries the identical base sequence, but without any insertions (···). While Δ*mutS* cells (- - -) have more mixed clans compared to wt (—) for single base insertions, essentially all clans are concentrated at the two peaks for the double-base insertions. Each experimental replicate is indicated with a separate curve of same color. Raw repair efficiencies ($\eta$) were measured for individual single (c) and double (d) unpaired bases mimicking cellular response against replicative insertion/deletion errors acquired in wt cells. The plotted data is extracted from the clans extracted from 3 SIL and 7 DIL experimental replicates.

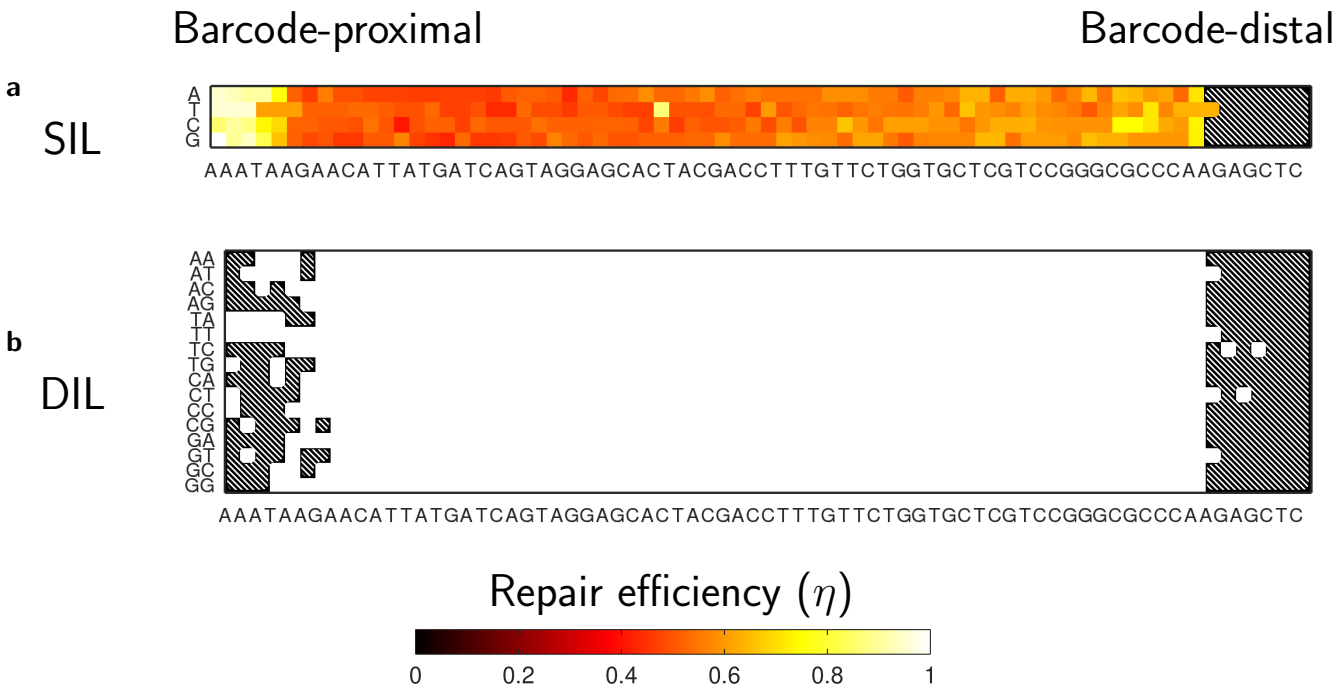



**Figure 3.6:** Repair efficiency of single (a) and double insertion loops (b) in ΔmutS cells. Black-white hatching patterns (▨) represent cases related to which no C, U or V clans were detected. No insertions were included in the starting library at the barcode-distal side, whereas the absence of clans with insertions in the barcode-proximal side might be due to inefficient ligation due to the presence of insertion close to the ligation site. Each panel represents the clans extracted from a single experimental replicate. The pixels that result in an identical variable strand are redundantly represented by the same value. The fraying of data points into the shaded region within which no bases were intentionally inserted result from this choice of representation.

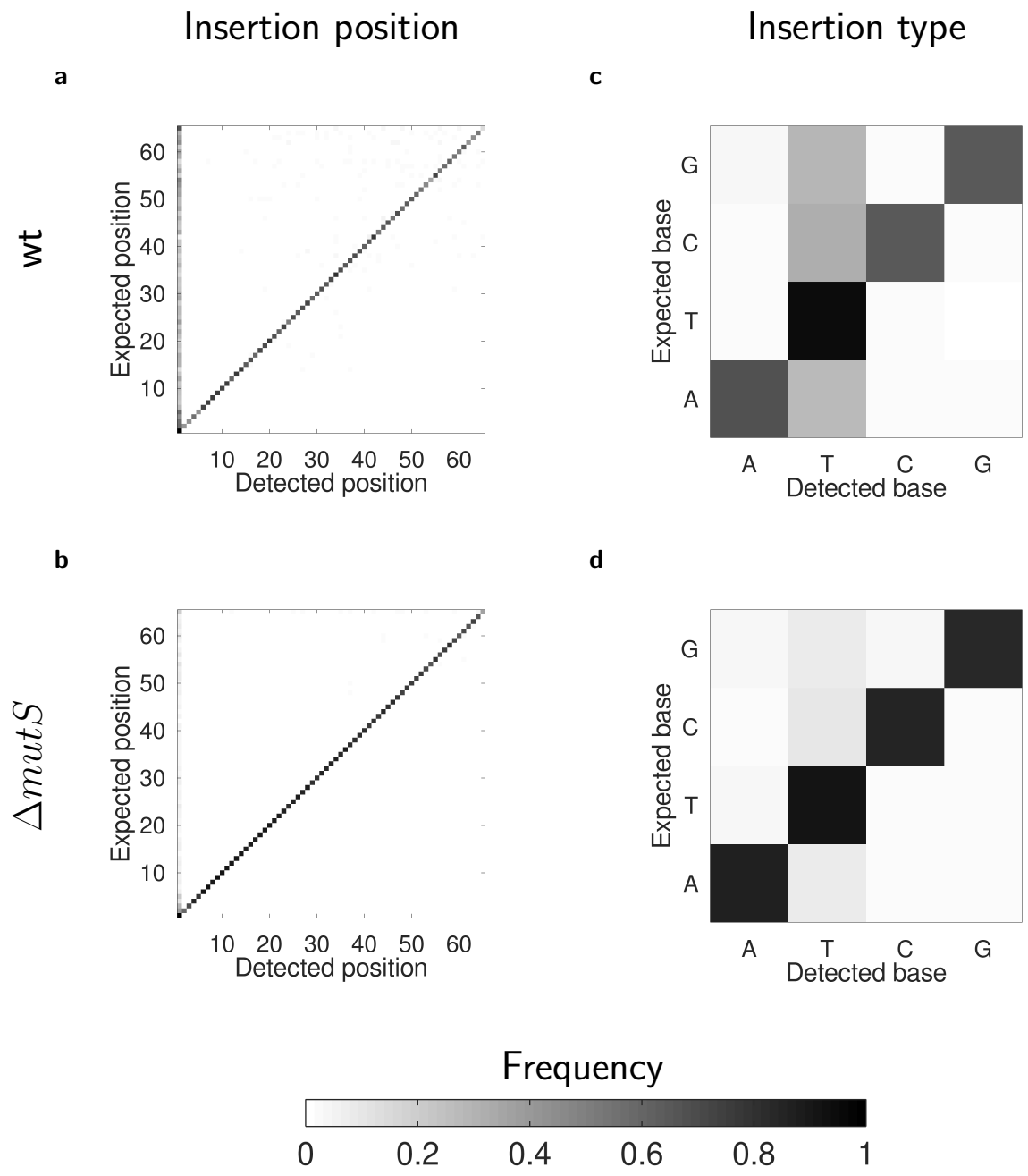



**Figure 3.7:** Confusion matrices comparing the position of the single inserted base and its type conclusions reached using the look-up barcodes (y-axis) with those deduced using the position and base identity of the most likely variant of the variable strand within the clan (x-axis). wt data was compiled using the combined output of three experimental replicates. Clans whose substitution prevalences are less than <0.2 were omitted due to the low confidence level of deducing the identity of the insertion in the starting material by the histogram method. Each clan was recorded as one event for its respective entry, and each row of the matrices was normalized to 1, such that entries along a row yield a probability of observing the experimental outcome, given the input material. Higher gray values indicate higher frequency of observing a particular event, and an ideal system with negligible synthesis, amplification and sequencing errors is expected to yield a diagonal matrix.

**Table 3.1:** A detailed list of all experimental conditions tested involving insertion-deletion loops.

| exp. no | cell strain | vector backbone | library | date sequenced | i7 index | i5 index |
|---|---|---|---|---|---|---|
| 1 | wt | tUNC19 | IL | MiSeq25102017 | N701 | S504 |
| 2 | wt | tUNC19 | IL | MiSeq25102017 | N702 | S504 |
| 3 | wt | tUNC19 | IL | MiSeq25102017 | N703 | S504 |
| 4 | $\Delta mutS$ | tUNC19 | IL | MiSeq25102017 | N704 | S504 |
| 5 | $\Delta mutS$ | tUNC19 | IL | MiSeq25102017 | N705 | S504 |
| 6 | $\Delta mutS$ | tUNC19 | IL | MiSeq25102017 | N706 | S504 |
| 7 | wt | tUNC19 | SIL | HiSeq22062020 | N701 | S503 |
| 8 | wt | tUNC19 | SIL | HiSeq22062020 | N702 | S503 |
| 9 | wt | tUNC19 | SIL | HiSeq22062020 | N703 | S503 |
| 10 | $\Delta mutS$ | tUNC19 | SIL | HiSeq22062020 | N704 | S503 |
| 11 | $\Delta mutS$ | tUNC19 | SIL | HiSeq22062020 | N705 | S503 |
| 12 | wt | tUNC19 | DIL | HiSeq22062020 | N706 | S503 |
| 13 | wt | tUNC19 | DIL | HiSeq22062020 | N707 | S503 |
| 14 | wt | tUNC19 | DIL | HiSeq22062020 | N710 | S503 |
| 15 | wt | tUNC19 | DIL | HiSeq22062020 | N711 | S503 |
| 16 | wt | tUNC19 | DIL | HiSeq22062020 | N712 | S503 |
| 17 | $\Delta mutS$ | tUNC19 | DIL | HiSeq22062020 | N714 | S503 |
| 18 | $\Delta mutS$ | tUNC19 | DIL | HiSeq22062020 | N715 | S503 |

# Chapter 4

# A novel unsupervised algorithm to process multi-channel single molecule images

## 4.1 Abstract

This fourth and last chapter is neither about the DNA itself nor *E. coli*, but more about improving our day to day microscopy workflow and hence both the system of concern and its language deviates from the previous chapters. While it did not lead to any biological discoveries of importance, this simple idea might prove practically useful, which is about aligning single molecule microscopy images. For a typical multi-wavelength detection based experiment, the different detection channels need to be aligned with respect to each other to deduce an accurate map between the pixel coordinates of each channel. A typical experimenter tackles this problem by imaging a test target that emits in all channels to a certain degree, typically consisting of sparsely distributed fluorescent beads in practice. The operator then manually locates peaks corresponding to the same bead in different channels and assigns this as a constraint that the applied transformation should satisfy. The coordinate transformation can be deduced after obtaining a sufficient number of constraints. Here I argue that the pipeline of multi-channel single-molecule image analysis can be simplified by the help of a simple automated algorithm, hence saving time to the microscope operator, as well as reducing the usage of experimental resources. The routines I describe here can be executed on a standard personal computer in almost real time. In contrast to more advanced computer vision tools, it has a highly intuitive nature and does not require any advanced mathematical knowledge. With this simple approach, the single molecule image stacks can be processed accurately without any human input.

## 4.2 Introduction

Single-molecule imaging microscopy experiments commonly make use of sparsely distributed samples on a slide surface. In its one form, the aim of the experiment is to observe the number of the surface-attached molecules and their co-localization if there are multiple components, thereby providing a quantitative measure of the equilibrium affinity between molecules [83]. This approach typically requires sequential imaging of the sample conjugated with different fluorophores and quantification of the number of spots detected that overlap between the frames when the

individual images are overlaid. In a slightly different form, the fluorophores can be chosen to have overlapping excitation and emission spectra, leading to the exchange of photon-induced excitation between molecules, a phenomenon commonly referred to as Foerster Resonance Energy Transfer (FRET). As this is an exclusively short-range process, it can illuminate the structural states a construct is at, as well as providing information regarding kinetics of transitions between them [84].

While a wide variety of optical components that can be used to accurately overlay image components on a detector surface exists, inherent chromatism of the optical components, as well as the human limit of precision in fine adjustments are still limited, leaving a residual discrepancy in the spatial positions of the objects observed with different imaging settings. Hence for a typical multi-wavelength detection based experiment, the different detection channels need to be computationally aligned with respect to each other to deduce an accurate map between the pixel coordinates of each channel. A typical experimenter tackles this problem in practice by imaging a test target that emits in all channels to a certain degree, typically consisting of sparsely distributed fluorescent beads such as TetraSpeck microspheres (Invitrogen). To understand the process, it would be helpful to observe the current process flow widely employed in our group at the time this work was published [85]:

1. The operator assembles a slide chamber typically using passivated glass or quartz slides and coverslips. The chamber is loaded with a solution containing the specimen of biological interest, which can be a single molecule or an assembly of multiple subunits that have a reasonable affinity to each other such that they will co-localize to the same physical location due to interactions between them. The sample is conjugated with two or more fluorescent probes.

2. Upon excitation, the fluorophores emit at slightly different wavelengths and this emitted signal can be split using a dichroic mirror in the emission path and reflected on two adjacent sub-compartments of a sensitive camera, typically emCCD or CMOS.

3. The operator is interested in a quantitative comparison of the spot counts or signal intensities in different channels that correspond to the same molecule. Either for signal quality reasons

or because of interest in the temporal changes, the operator records a short movie consisting of 10 to 10 000 frames. This data is stored on the hard drive essentially in the form of a stack of images. Each image usually has 512x512 to 2048x2048 pixels, harboring sparsely distributed random spots surrounded by pixels that carry only background-level signal intensity.

4. Although the operator can visually verify the rough alignment of the images with respect to each other and attempt a calibration by adjusting the beam alignments, a precise relation between the data channels is not known *a priori* due to the remaining imperfections in the beam alignment. To deduce this pixelwise bijective relation, the operator generates one more specimen that can be easily detected in both signal channels, typically consisting of fluorescent beads. The operator acquires a separate short movie with this bright and stable test target.

5. The operator processes the bead movie by randomly choosing three spots in one signal channel and visually deciding the corresponding peak position in the other image subcompartment(s). The existing algorithm solves an exactly determined linear algebra problem to search for a transformation matrix satisfying these three user-defined constraints. The algebraic system is self-consistent, unless the operator makes a mistake and accidentally relates the same point to multiple points.

6. A program routine uses this transformation matrix to project the signal channel to the other(s). Based on the overlap obtained between peaks, it provides statistics of overall quality of this routine. Based on his/her educated judgement, the operator either accepts the proposed transformation or repeats this manual mapping procedure till a transformation matrix is obtained that provides a reasonable overlap by making a different choice of peak triplets each time.

7. Using the transformation matrix obtained from this short procedure, the operator analyses the original movie(s). For this, first a projection image is generated out of the multi-frame image stack, either by maximum intensity projection or average of a few of the frames in the image stack. Peak positions in all signal channels are independently determined by Gaussian

fitting out of this projected image and the transformation determined above is applied to one channel of the image to find the correspondence between the peak pairs.

8. The intensity of each spot is integrated one by one for each frame of the movie. This yields a binary file that contains one time trace per each signal channel and each unique molecule reported in corresponding pairs.

9. As the mapping between the channels is very sensitive to small changes in the emission path of the instrumentation, the operator repeats this procedure at regular intervals, ideally before and/or after each set of experiments.

In the sequel, I will propose an alternative strategy that not only makes this routine fully-automated and hence faster, but also provides more accurate results due to the ability to benefit from solving an over-determined linear system. As no human input is needed, the results of this approach is also more deterministic and is subject to fewer human errors. The routine is tolerant to moderate experimental errors such as background level variations, shift, shear or rotation of the image channels with respect to each other. As long as a reasonable number of point pairs are still available, the approach is also highly tolerant to photobleaching or imperfect dye conjugation chemistry that leads to orphan spots in either or both of the detection channels.

Depending on the configuration of the setup, image at each timepoint either contains multiple images from multiple cameras or one image itself is split into sub-images. Without loss of generality, from now on I will describe how to map two emission channels of a single molecule fluorescence resonance energy transfer (smFRET) microscope. Although I will focus my attention mostly on image stacks split into two sub-compartments vertically, the methodology that I propose could also be applied to other similar systems. Due to the availability of the instrumentation to acquire data for a proof-of-principle experiment, the constructs are typically going to be labeled with Cy3 and Cy5, hence the green-red color theme. But the approach I derived are immediately applicable to any other fluorophore FRET pair or two arbitrary fluorophores on the same construct sequentially imaged without relying on any energy transfer interaction between them.

I first will start by portraying the mathematical question to be solved, and then describe a routine to detect the two different fluorescence channels out of a composite image. For the

sake of completeness, I then will show a simple and efficient algorithm to detect signal peaks out of a sparsely populated image. Next, I will describe some intuitive features to describe the local neighborhood of a peak, so that the peak pairs in the two signal channels can be detected. The chapter will conclude by demonstration of the proposed mapping algorithm on some real microscopy images.

## 4.3 Results

### 4.3.1 Mathematical statement of the problem

Each real molecule on the slide will have images that appear at different pixel coordinates in both emission channels. Let these peaks to be at $(x_1, y_1)$ and $(x'_1, y'_1)$, respectively. Assuming a linear transformation, the relationship between the two in the most general form will be given by:

$$\begin{bmatrix} x'_1 \\ y'_1 \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \begin{bmatrix} f \\ g \end{bmatrix} \tag{4.1}$$

where a, b, c, d, f and g are the unknowns to be determined. f and g refer to the translations whereas the matrix of a-d can accommodate rotations, scaling and skewnesses. I note that we can also express this equation in the below equivalent format:

$$\begin{bmatrix} x'_1 \\ y'_1 \end{bmatrix} = \begin{bmatrix} a & b & f \\ c & d & g \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} \tag{4.2}$$

Or after some further rearrangements,

$$\begin{bmatrix} x'_1 \\ y'_1 \end{bmatrix} = \begin{bmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_1 & y_1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ f \\ c \\ d \\ g \end{bmatrix} \tag{4.3}$$

a, b, c, d, f, g are the six unknowns of this equation, to be able to solve for which we need at least six equations. The necessity to manually find at least three point pairs in two channels stems from the fact that each pair on the 2D detector plane provides 2 independent constraints. As there are about N=300 peaks per channel, the system is usually over-determined and will look in the form:

$$\begin{bmatrix} x'_1 \\ y'_1 \\ x'_2 \\ y'_2 \\ \vdots \\ x'_N \\ y'_N \end{bmatrix} = \begin{bmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_1 & y_1 & 1 \\ x_2 & y_2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_2 & y_2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_N & y_N & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_N & y_N & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ f \\ c \\ d \\ g \end{bmatrix} \tag{4.4}$$

which is of the form $\vec{b} = A\vec{u}$ and an approximate solution can be easily obtained by the pseudo-inverse, i.e. $\vec{u} = (A^T A)^{-1} A^T \vec{b}$. Note that obtaining $\vec{u}$ addresses the ultimate aim of this discussion, however this requires the knowledge of a set of point pairs $\{(x, y, x', y')\}$ of at least 3 elements to define A and $\vec{b}$. In the current format, the microscopist can achieve this by visually selecting a subset $N \geq 3$ of available positions in the two channels to be able to obtain a determined linear system. Below I will argue that this can be achieved programmatically, rather than being performed manually.

### 4.3.2 Segmentation of image channels

In a compound image that is a juxtaposition of two signal channels side by side, we need to deduce the borderline between the two channels to be able to binary classify the detected peaks. To achieve this, we benefit from the fact that the image optics usually generates an image artifact at the transition zone with a below-average noise level in practice. The aim of this section is then to find the best line that separates the two zones by localizing this pattern.

We expect the boundary to cause a relatively sharper transition between the two zones than the rest of the image. To detect such locations, we can calculate the relative intensity of the image gradients ($||\nabla M||/M$) by a convolution. That is, the raw image being M, and K being a (2n+1)x(2n+1) Gaussian kernel defined by

$$K_{ij} \propto -(i-n)e^{-\frac{(i-n)^2+(j-n)^2}{2\sigma^2}} \tag{4.5}$$

Given a user-defined threshold parameter t, we then search for the set of pixels with above threshold level, i.e.

$$Q \equiv \left\{ (i,j) \mid \frac{1}{M_{ij}} \sqrt{(M*K)_{ij}^2 + (M*K^T)_{ij}^2} > t \right\} \tag{4.6}$$

where the two convolution terms $(A*B)$ in the square root compute the derivative along the x and y axes, respectively. Hence Q represents the pixels neighboring sharp transitions. In practice, these points should be accumulated around the edge separating the two channels, but random noise in the image causes the set Q to have a significant number of members elsewhere, as Figure 4.2c shows.

We now need to find the equation of the line that majority of the set Q describes despite the significant noise level. In 2D, a line would be represented as y=mx+b, but for steep lines (i.e. high m), this formula would cause numerical instability in obtaining the solution. Instead, here I use polar coordinates, according to which a line is defined by

$$\rho_o - \rho \; cos(\theta - \theta_o) = 0 \tag{4.7}$$

in which $\rho_o$ is the distance to the origin and $\theta_0$ is the angle that this normal makes with the x-axis (Figure 4.1a). Resorting to Hough Transform [86, 87], we can see that this equation can be considered a parametric equation where $\rho_o$ and $\theta_o$ are the unknowns. As such, through each single point $(p_x, p_y) = (p_r, p_\theta) \in Q$, a bundle of lines can pass (Figure 4.1b) and there will be a corresponding sinusoidal curve in the Hough space spanned by $\rho_o \in [0, \infty)$ and $\theta_0 \in [0, \pi)$ (Figure 4.1c).

As the main edge we seek is supported by many co-linear points (Figure 4.1d), we loop over all elements in Q to build a histogram and choose the most voted value as the true value of the pair $(r_0, \theta_0)$ (Figure 4.1e). The corresponding line in the image space (Figure 4.1f) separates the input image into three zones, the third being the unusable transition zone ($\rho_0 - \rho cos(\theta - \theta_0) > d$, $\rho_o - \rho\ cos(\theta - \theta_o) < -d$ and $|\rho_o - \rho\ cos(\theta - \theta_0) = 0| \leq d$).

The application of this approach to a real microscopy image is shown in Figure 4.2, where the starting image (Figure 4.2a) was first subjected to the differentiation described by Equation 4.6 (Figure 4.2b) and then thresholded to obtain a binary image displaying the pixels around which sharp transitions happen (Figure 4.2c). By choosing the highest voted parameter out of this histogram in the Hough space (Figure 4.2d), the edge was accurately located and the image was segmented into three ones (red for Cy5, green for Cy3 emission channel and black thick line as the boundary zone), regardless of the orientation of the input image (Figure 4.3).

### 4.3.3 Detection of the peaks

The proposed algorithm does not suggest any novelties in this regard, however an easier and more CPU-efficient approach compared to the commonly used analysis workflow will be described here for completeness. We start by generating an estimation on the background levels throughout the image. Due to the characteristics of total internal reflection (TIR) illumination mode, the background tends to be far from uniform but rather high around the beam center and relatively lower towards the edges. Surfaces that contain defects might also cause complications, as the accumulation of molecules at non-passivated sites may lead to localities of high background. To

capture such trends, we convolute the normalized image with the Gaussian kernel below:

$$K_{ij} \propto e^{-\frac{(i-n)^2+(j-n)^2}{2\sigma^2}} \tag{4.8}$$

Out of this smoothened image, we then obtain a binary image by simple thresholding:

$$B_{ij} = \begin{cases} 1, & M_{ij} - (M * K)_{ij} \geq 0.2 \\ 0, & otherwise \end{cases} \tag{4.9}$$

where 1's indicate the pixels that contain a high signal level, and hence are likely to contain a peak due to a fluorescent molecule. Pixels with 0 are treated as empty areas that contain about background level signal. We segment this binary image pixelwise by labeling each disconnected non-zero cluster with a different integer by following a connected component labeling algorithm [88].

To achieve this, one makes three passes through the binary image. In the first pass, we label each pixel with a cluster index starting with 1 and traversing the matrix unidirectionally row by row (Figure 4.4a). If the current pixel on the binary image contains a 0, representing noise, the associated entries in the label matrix are also labeled as noise (X). If a signal carrying entry is encountered, then it can be part of the same peak represented by the cluster indices at the left or top-left, top or top-right neighbors. Or else, if all these 4 neighbors are noise, then one seeds a new peak by assignment of the next unused cluster index (Figure 4.4b). The discovery of new peaks can continue in this fashion (Figure 4.4c), unless more than one of the previously processed neighbors contain a non-noise cluster label and their labels differ from each other (Figure 4.4d). In this case, the conflict is resolved by the realization that these differentially labeled clusters are actually parts of one larger cluster, and the usage of multiple different labels for this same cluster was actually a mistake. To save computational power, the equivalency of the labels can be recorded for later use (Figure 4.4e) and the image processing continues pixel by pixel in this fashion until all pixels are processed. If a label redundancy table is used, these labels should be replaced with a unified label at this time by making one more pass through the image.

At this point each connected component of the binary image B is labeled with a shared label that is unique to each component representing a peak (Figure 4.4f). We now will estimate the

pixel coordinates of each such peak by calculating the centroid position and the area of each peak, that is

$$\begin{aligned} \langle x_k \rangle &= \sum_{ij} i M_{ij} 1_k(i,j)/I_k \\ \langle y_k \rangle &= \sum_{ij} j M_{ij} 1_k(i,j)/I_k \end{aligned} \tag{4.10}$$

where k refers to the k'th peak and $1_k$(i,j) is the indicator function that attains 1 if pixel (i,j) belongs to k'th peak and 0 otherwise. The common normalization factor is the the total intensity of the peak:

$$I_k = \sum_{ij} M_{ij} 1_k(i,j) \tag{4.11}$$

While the accuracy provided by Equation 4.12 is sufficient for our purposes, a refinement is possible by Gaussian fitting of the intensity matrix M using these results as initial guess. If needed, an initial estimate of the peak width can also be obtained by making use of the formula $\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2$ and the former required term can be similarly evaluated by,

$$\begin{aligned} \langle x_k^2 \rangle &= \sum_{ij} i^2 M_{ij} 1_k(i,j)/I_k \\ \langle y_k^2 \rangle &= \sum_{ij} j^2 M_{ij} 1_k(i,j)/I_k \end{aligned} \tag{4.12}$$

### 4.3.4 Mapping of peaks

Up to this point, we have detected the two emission channels and determined the peak positions. Ideally, each spot in the green channel will have a corresponding spot in the red channel. I now will describe two intuitive features that can be used to accurately detect the pairs of spots based on their local neighborhood by "point feature matching".

#### Directional neighbor counts

A typical human observer judges the corresponding peak pairs by the help of patterns formed by the neighboring spots. If the patterns formed by the relative positioning of the neighboring spots are visually similar, the two picked spots are also likely to be the corresponding images of the same bead on the slide. An easy way to define a characteristic feature describing a spot is hence by

deducing a systematic description of the neighbor peak positions algorithmically.

As a simplistic approach, here I will consider the number of neighbors in each spatial predetermined sector surrounding the spot of interest (Figure 4.5a). For this, I split the local neighborhood of the spot into about 20 angular and 20 radial sectors and count how many neighbors fall into each sector, which defines a roughly 20x20 feature matrix for each spot (Figure 4.5b). Each individual spot in each image channel will have a feature matrix of its own, and therefore processing the image yields two stack of matrices, each containing about a few hundred features for each spot in the field of view (Figure 4.6a). If two spots from the green channel and the red channel are actually due to emissions from the same bead, then the positioning of the local neighbors will be very similar to each other and hence the two feature matrices will elementwise be very similar. After a pairwise comparison of these feature matrices, it is very likely that the two spots whose feature matrices are separated by the smallest distance are likely to be the corresponding spots, hence allowing pair assignment for each peak (Figure 4.6b).

While ideally feature matrices of the corresponding peak pairs are expected to be identical, some peaks might be orphan spots without a corresponding pair, requiring omission of false assignments by thresholding (i.e. reject if feature matrix to feature matrix distance >8). Also due to experimental errors on the peak position determination, the relative location of the neighbors might in reality vary, necessitating allowing imperfect matches between matrices as a match, as well as requiring a careful choice of the dimensions of the spatial sectors. For this, we can estimate the error margin on the radial coordinates by error propagation on $r^2 = x^2 + y^2$, from which we can reach,

$$\delta r = \frac{x\delta x + y\delta y}{r} \tag{4.13}$$

$$\Delta r \sim \frac{\sqrt{x^2(\Delta x)^2 + y^2(\Delta y)^2}}{r} \sim \Delta x \tag{4.14}$$

which only depends on the imaging precision, but is independent of the chosen range of the polar coordinates in the Cartesian plane. The imprecision in the angular coordinates have a

similar confounding effect on the determination of the local features, as a significant change in the azimuthal coordinates might cause some of the neighbor spots to shift into the next angular sector. As the matrix counts will change uncorrelatedly in the two image channels, this might lead to false pair assignments. The uncertainty in the angular coordinates can similarly be connected to the localization uncertainty starting from $tan\theta = y/x$ by,

$$\begin{aligned} sec^2\theta \ \delta\theta &= \frac{x\delta y - y\delta x}{x^2} \\ \delta\theta &= \frac{x\delta y - y\delta x}{x^2 + y^2} \end{aligned} \tag{4.15}$$

$$\Delta\theta \sim \frac{\sqrt{x^2(\Delta y)^2 + y^2(\Delta x)^2}}{x^2 + y^2} \sim \frac{\sqrt{(x^2 + y^2)(\Delta x)^2}}{r^2} = \frac{\Delta x}{r} \tag{4.16}$$

The least accuracy in determining the counts in each angular bin comes from the neighbors that are too close to the spot of interest, which we avoid in practice by the fact the number of neighbors grows quadratically with $r$. These two conclusions suggest that essentially entire neighborhood of the spot can be taken into consideration when constructing the feature, as the systematic omission of any particular region will likely not lead to a very significant improvement in the accuracy.

I would like to note that the feature I propose in this section is shift invariant, as only relative positions of the neighbors to the spot of concern are involved, but not the absolute pixel coordinates. The approach would be tolerant to a certain degree to the localization error of the experiment, while the extent will be limited by the error tolerance that can be allowed during feature comparisons. As Figures 4.5d and 4.5e demonstrate, the features are not rotation or scale invariant as a significant difference between the image channels has a potential to significantly alter the deduced feature matrices.

### Angles between neighbor pairs

As a second approach to construct reliable features based on local neighborhood, we can consider the angles that form between the neighbors. Or more precisely, we can consider the angles that form between the two rays that emanate from the spot of interest located at the vertex of the angle towards the neighbors that are located at a specific radial distance range from this vertex

(Figure 4.7a). To serve as a feature that can be used for comparisons, we then form a sorted list of all angles between all pairs of neighbors in the allowed band (red spots). If two spots in two image channels belong to the same signal emitter on the specimen, then the local neighbors will also be located at similar locations and hence a high number of similar angles that are present in both lists is a potential sign that the two spots should be assigned as pairs.

It is noteworthy that these features are automatically scale and shift invariant, as only angles are considered. As the rotation of the entire image would increment the azimuthal position of all neighbors by the same amount, the proposed feature that consists of a list of difference in angular positions is rotation invariant. The experimental determination of angular positions will, however, be error-prone due to the limit of localization precision. From the geometric outlay (Figure 4.8a), as well as assuming that the thickness of the ring is small and the localization precision is isotropic, we can estimate the uncertainty in the measured angles between the two points located at $(x_1, y_1)$ and $(x_2, y_2)$ as,

$$\begin{aligned} cos\theta &= \frac{x_1x_2 + y_1y_2}{r^2} \\ \delta\theta &= \frac{-1}{r^2 sin\theta}\left(x_1\delta x_2 + x_2\delta x_1 + y_1\delta y_2 + y_2\delta y_1\right) \\ \Delta\theta &\sim \frac{1}{r sin\theta}\Delta x \end{aligned} \tag{4.17}$$

As $\lim_{\theta\to 0}\Delta\theta = \lim_{r\to 0}\Delta\theta = \infty$, we deduce that to avoid angles with very high detection errors, we should exclude from the list below-threshold angles ($\theta \geq 0.2rad$). A choice of a belt with a too low radius will also increase misassigned pairings due to high error margins, thus $r \sim 50px$. The quadratic increase of number of angles between all neighbor pairs necessitates a limit on the total number of neighbors included. In practice, I achieved this by considering among neighbors only those that are closer than a certain cutoff distance ($r < 70px$). In addition, only the smallest 100 angles are included in the feature vector.

To take the imprecision in angular measurements into account, I imposed an error tolerance such that the angles from the two lists are considered a match if they deviate at most by this tolerance level ($\delta = 0.005rad$). While the choice of this tolerance level is arbitrary, if too high a tolerance is chosen, the significance of the matching angles found in two lists will be too low that the false matches will dominate the output (Figure 4.8b). The approach suggested above is

capable of providing the solution with sufficient accuracy. Table 4.1 lists the parameters obtained after 20 independent runs of the same bead image file. The accuracy of the outputted values are amply high that would satisfy a typical experimenter.

### 4.3.5 Finer mapping

So far in the procedure, we obtained two sets of point coordinates and an accurate transformation between these two sets is sought after to make them overlap as closely as possible. Mathematically well-established algorithms that address this point set registration problem exist, but they are often conceptually very complex that makes their usage by a typical microscopist difficult [89]. After detecting a few ($\geq 3$) correspondences with the above approach, we use this information to seed an iterative closest point approach [90], which in our case typically convergences after the first iteration. One can accomplish this by applying the deduced rough transformation on the peak coordinates from the green channel to the red channel and assigning the closest point in the red channel as the corresponding peak, if it is reasonably close ($\leq 5px$) and repeat until convergence. In case there are more corresponding point pairs found, only a random subset of 200 pairs is used to save computational power, as seeking a least squares approximation to the linear system in Equation 4.4 is quadratic in time complexity.

For this latter final finer map, I use a higher degree model with more parameters as the amount of available data is now not as limited as before:

$$\begin{bmatrix} x_1' \\ y_1' \end{bmatrix} = \begin{bmatrix} a & b & c & d & f \\ h & i & j & k & l \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ x_1^2 \\ y_1^2 \\ x_1 y_1 \end{bmatrix} + \begin{bmatrix} g \\ m \end{bmatrix} \tag{4.18}$$

which can similarly be cast into a linear equation of the form:

$$\begin{bmatrix} x'_1 \\ y'_1 \\ x'_2 \\ y'_2 \\ \vdots \\ x'_N \\ y'_N \end{bmatrix} = \begin{bmatrix} x_1 & y_1 & x_1^2 & y_1^2 & x_1y_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & x_1 & y_1 & x_1^2 & y_1^2 & x_1y_1 & 1 \\ x_2 & y_2 & x_2^2 & y_2^2 & x_2y_2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & x_2 & y_2 & x_2^2 & y_2^2 & x_2y_2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_N & y_N & x_N^2 & y_N^2 & x_Ny_N & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & x_N & y_N & x_N^2 & y_N^2 & x_Ny_N & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ f \\ g \\ h \\ i \\ j \\ k \\ l \\ m \end{bmatrix} \tag{4.19}$$

which can be solved with a similar strategy as Equation 4.4 was solved.

### 4.3.6 Artifacts due to labeling chemistry

The discourse above made use of a slide prepared with fluorescent beads, colloquially referred to as a "bead slide". The bead slide is prepared by diluting fluorescent beads, typically TetraSpeck beads (Invitrogen), ∼100nm in diameter, in a stable aqueous buffer such as T50 and has the advantage of providing a very bright and stable signal in both signal channels. This ensures that the patterns observed in the two sub-compartments are very similar and hence constitutes the optimal case for the analysis procedure.

To simplify the experimental workflow, it is desirable to remove the bead image but rather extract that information directly from the experimental movie, instead. Assuming that a sufficiently reliable centroid finding approach is adopted, in a real experimental case, the accuracy of position

detection is governed by the SNR of the image, which is bounded by the formula [91]:

$$\Delta x = \sqrt{\frac{s^2 + a^2/12}{N} + \frac{8\pi s^4 b^2}{a^2 N^2}} \tag{4.20}$$

where b is the standard deviation of the background signal intensity, N is the photon count, s is the standard deviation of the spot size and a is the physical size of each pixel on the detector. The reader can observe that a high position accuracy requires high photon counts, which is amply available with bead slides but not necessarily possible when using single organic fluorophores as they tend to photobleach before a high photon count can be reached (e.g. Cy3-Cy5 conjugated DNA). From Equations 4.16, 4.14 and 4.17, we can observe that the uncertainty in determining both proposed features linearly depends on the localization accuracy, regarding which the beads provide a distinct advantage than a single-molecule image obtained from typical fluorophores. In this respect, it is highly desirable that any proposed approach does not completely fail if the peak positions are inaccurate due to lower photon count that is practically achievable. Table 4.2 compares the fine mapping parameters obtained using movies acquired using a bead slide vs. cyanine dye labeled DNA. All 12 mapping parameters obtained with the two samples are very similar, supporting the idea that usage of a bead slide is redundant if a sufficiently well-labelled biological sample is available. The final iterations lead to accurate fine mapping parameters that indeed makes the signal spots emitted by Cy3 to overlap with Cy5 spots when the transformation is applied (Figure 4.9). Therefore, obtaining an accurate map with low photon counts that are typically attainable out of relatively dimmer organic fluorophores is still possible.

As a second concern, the molecules might not have been labeled with both of the fluorescent dyes, but rather only one fluorophore due to imperfect labeling chemistry, or alternatively one of the fluorophores might have been lost due to photobleaching, both of which are commonly observed when working with organic fluorophores but not bead slides. Such molecules will only be detected in one signal channel but not both and hence can increase the noise as perceived by the above described algorithm. To be applicable to non-bead slides, the approach should also tolerate a certain fraction of such orphan spots and Figure 4.10 displays the peak-to-peak correspondence deduced in an *in silico* experiment assessing this situation. Each panel makes use of the same

composite bead slide image, out of which the red and green emission channels have been detected followed by peak finding. To simulate the presence of orphan spots, I deliberately removed a certain percentage of the detected peaks from the list of detected peaks before mapping. Peaks can be randomly missing from the green channel (rows), the red channel (columns) or both, at a percentage of all detected spots in that channel as indicated along the respective axes. The algorithm is capable of correctly finding peak pairs corresponding to the same bead in both signal channels, even if 40% of the spots are orphan, as can be judged by the fact that there are corresponding pairs indicated with blue lines. In contrast, 40% random spots missing from both channels leads to a mapping failure as only 2 corresponding peaks could be detected whereas the theoretical minimum requirement is 3 pairs. I hence believe that the proposed automated procedure can tolerate down to 80% coupling efficiency, below which it might not be possible to deduce the mapping parameters, in which case the routine quits with a fatal error.

## 4.4 Conclusion

In this last Chapter, which can be regarded as a simple tutorial on single molecule microscopy image analysis, I introduced a novel unsupervised algorithm to process image stacks that contain multiple juxtaposed signal channels, as commonly is the case for smFRET measurements. While my proposed approach consists of multiple steps, each of which have more advanced alternatives, I believe the routines included here are intuitive and hence can easily be adapted for routine use.

Apart from a rough instrument specific calibration, none of the procedures described here require any user input. From the start to the end, the scripts could be executed on a standard personal computer in seconds with GNU Octave v5.2. While optimization of the computational efficiency is beyond the scope of this work, further improvements are possible via multi-threading or usage of more advanced data structures. In that regard, this proposal can substantially decrease the hands-on time of a microscope operator by abolishing the necessity to hand-pick corresponding spots, and associated efforts to ensure the mapping quality mainly via trial and error. Running independently without operator input, this idea might also reduce the human errors that are primarily associated with a poor choice of point pairs to set the constraints.

Another benefit of making use of an algorithmic approach is the relative ease of establishing an over-determined linear system to solve. While the amount of point pairs that a microscopist can manually pick within a reasonable time is limited, it is possible to computationally impose more than 100 coordinate relationships to be approximately satisfied, hence the accuracy of the deduced transformation is expected to be high. This ability to obtain more point pairs also confers a tolerance to the uncertainty in the peak center positions as well as the presence of orphan spots without corresponding peaks in the other channel, thanks to which a separate bead slide image acquisition purely for calibration purposes can be made redundant. Apart from leading to time and cost savings, the ability to deduce the absolutely indispensable mapping transformation parameters out of the real experimental image stacks also makes the experimental data self-contained, thereby reducing the risk of data loss due to the accidental loss of mapping data.

While various computer vision tools exist in the literature that try to address similar problems, a complete proposal to automate single-molecule data processing workflow, to my knowledge, has not been reported in the literature, nor has a simpler approach made its way to our own pipelines at the time of the submission of this dissertation. In a hope that this novel unsupervised algorithmic approach I describe is useful, the source code is available through Gitlab along with sample data to process: https://gitlab.com/tuncK/public/tree/master/usmap.

## 4.5 Figures and tables

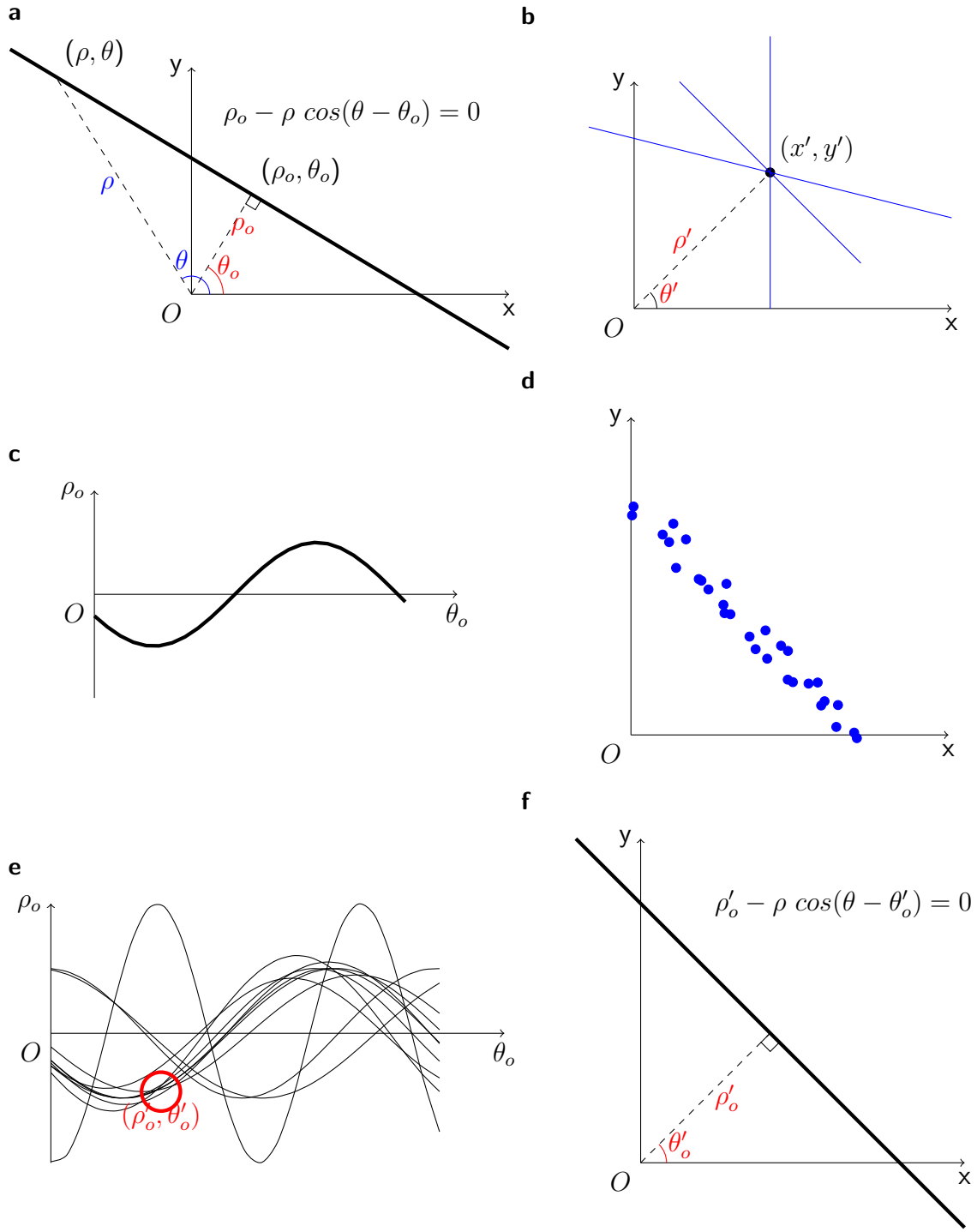



**Figure 4.1:** Pictorial summary of Hough transform. (a) For computational stability, each point that is part of an edge in the image is represented in polar coordinates. An infinite number of lines can pass through each point in the Cartesian plane (b). The coordinates of this point puts a constraint on the two polar parameters of a line, in such a way that each point in the real space transforms to a sinusoidal curve in the plane spanned by these two parameters (c). When a histogram is built by transforming each point in the image, the points along an edge will accumulate constructively (d), providing the line fitting parameters $(\rho'_o, \theta'_o)$ (e). Inverse Hough transformation of $(\rho'_o, \theta'_o)$ provides the sought-after line in the real space (f).

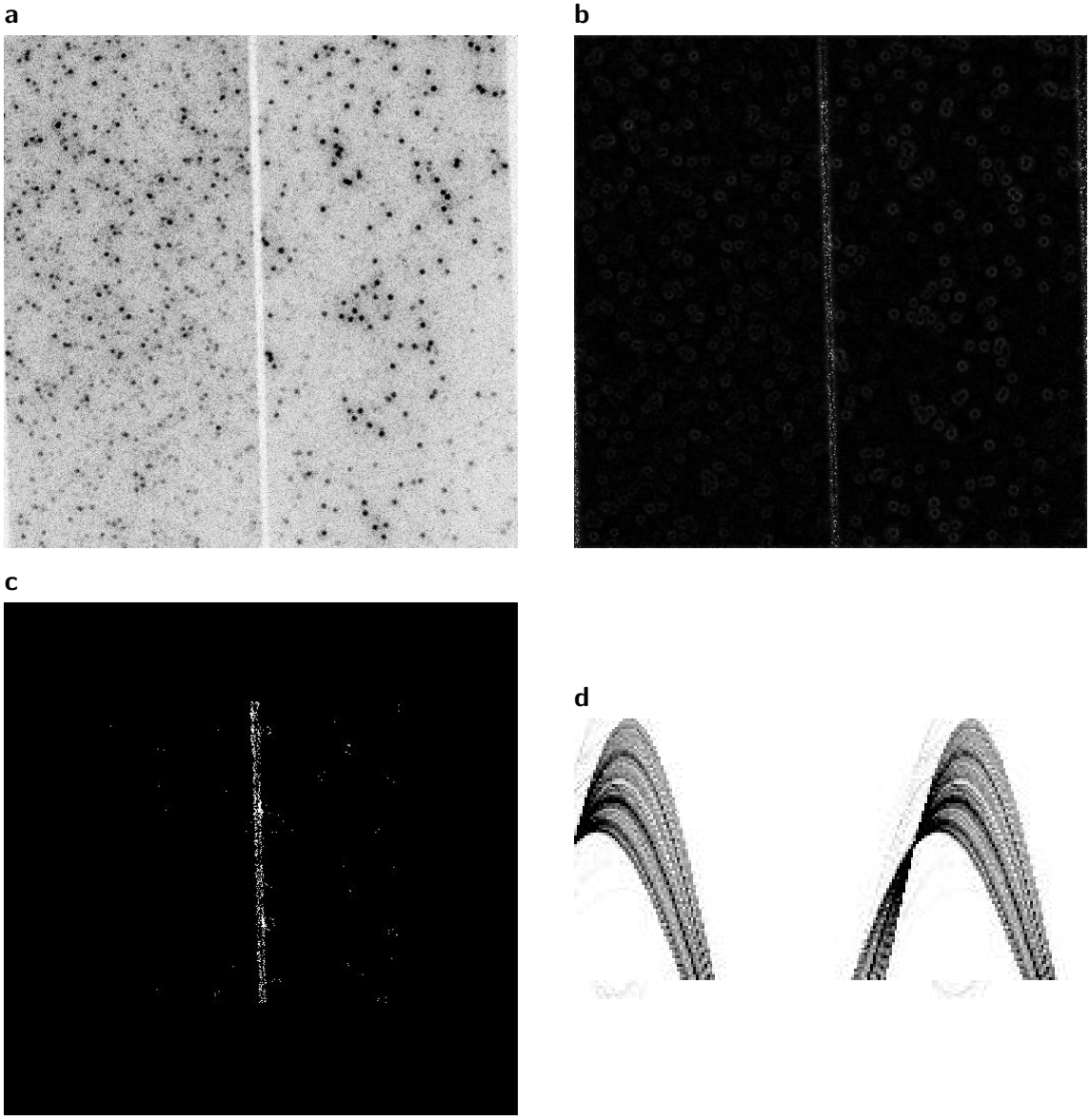



**Figure 4.2:** Boundary detection in practice. (a) A sample microscope image of Cy3/Cy5 double labeled DNA on a PEG slide, illuminated with 532nm laser under a smTIRF microscope, shown in negative and enhanced for contrast (1-3·image). (b) Gradient operation of Equation 4.6 applied to the input image. (c) Positions of edge points after thresholding b, i.e. the set Q in the text. (d) Histogram constructed by the overlay of sinusoidals resulting from the Hough transform of all above-threshold points, shown in negative.

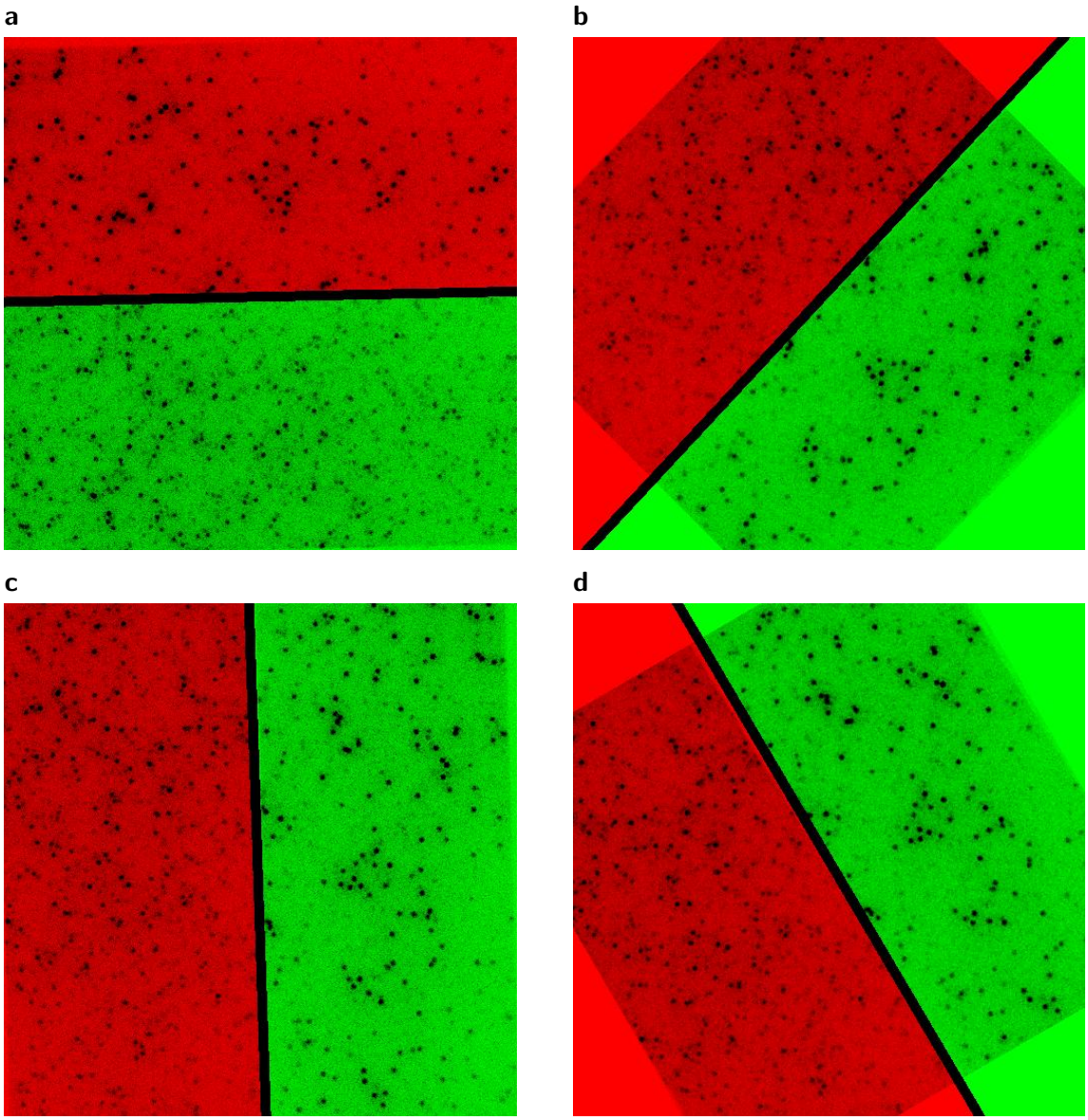



**Figure 4.3:** Boundary detection can be performed for arbitrary orientations of the separation line. The images are split by the black line into two false colored red or green zones representing the two signal channels, when the separation line makes about $0^o$ (a), $45^o$ (b), $90^o$ (c) or $120^o$ (d) with the x-axis.

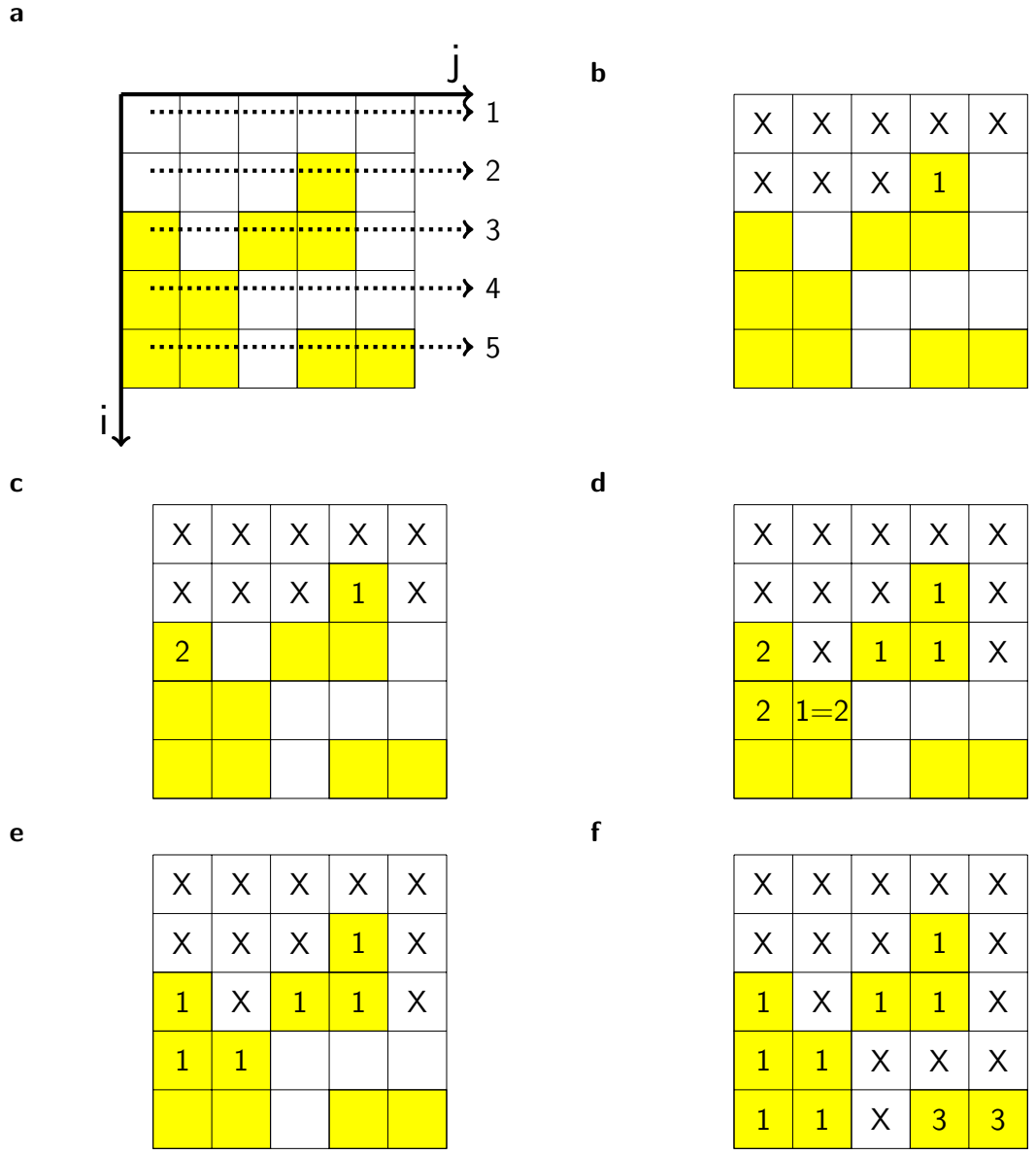
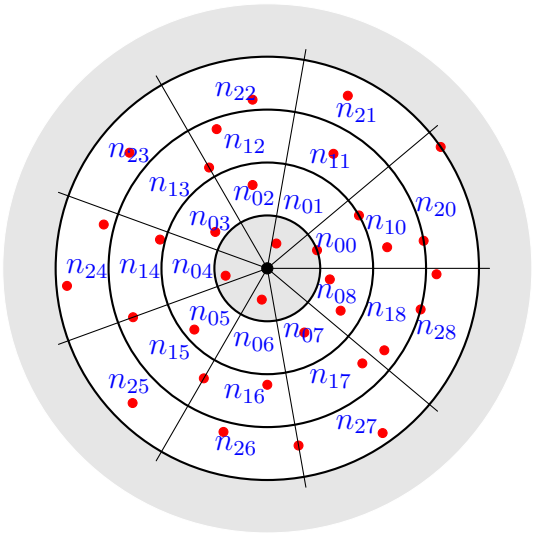



**Figure 4.4:** Peak detection by the class assignment procedure. (a) The binary image is traversed row by row in a unidirectional fashion. Zero-valued pixels (white cells) are skipped as noise, (X), whereas at the pixels with signal (yellow cells) a cluster is seeded (b,c). If a neighbor of an unprocessed pixel contains a cluster, same cluster is expanded. In case a non-noise pixel is encountered whose two neighbors had been assigned to different clusters (d), these clusters are merged by relabeling (e). The routine quits after all pixels in the binary image have been processed (f).

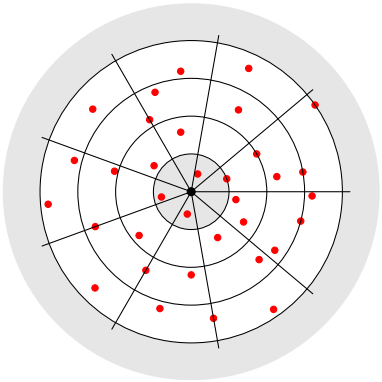
**a**



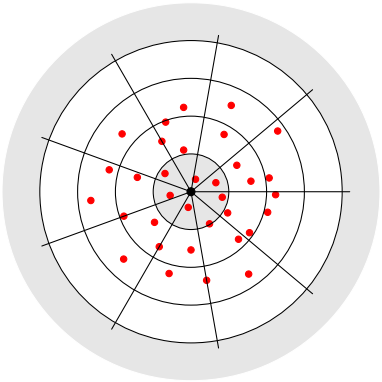

**b**

| $n_{ij}$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 2 |
| 1 | 2 | 1 | 2 | 0 | 1 | 1 | 2 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 |

**c**



**d**



**e**

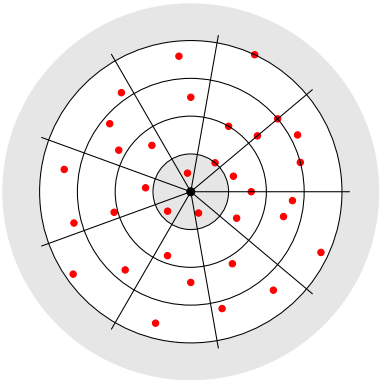


**Figure 4.5:** Neighbor count per sector as a feature. (a) Description of the local pattern by counting the neighbor points in polar coordinates. (b) The matrix describing the neighbor count as seen by the solid black point in part a. Each matrix entry shows the number of neighboring points in the band segment where each row is a different range in radial distance, and columns represent different angular coordinate sectors. This neighbor counting approach would fail due to re-scaling (d) or rotation (e) with respect to the reference channel (c) at an extent comparable to the distance and angle resolution of the user-defined sectors.

**a**

$$M_1^g = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 2 & 0 \\ 0 & 2 & 0 & 1 & 2 & 0 & 0 & 2 & 1 & 1 \\ 4 & 3 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

$$M_2^g = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 2 \\ 1 & 2 & 1 & 2 & 0 & 1 & 1 & 2 & 1 & 1 \\ 2 & 1 & 1 & 1 & 1 & 2 & 1 & 1 & 2 & 2 \end{bmatrix}$$

$$M_3^g = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 2 & 2 & 4 & 1 & 2 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 2 & 1 & 1 & 0 & 0 \end{bmatrix}$$

$$\vdots$$

$$M_1^r = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 2 \\ 1 & 1 & 1 & 1 & 2 & 0 & 0 & 0 & 2 & 1 \\ 0 & 1 & 1 & 0 & 0 & 2 & 1 & 1 & 2 & 3 \end{bmatrix}$$

$$M_2^r = \begin{bmatrix} 0 & 0 & 1 & 1 & 2 & 1 & 1 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 & 0 & 0 & 1 & 1 & 2 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 2 \end{bmatrix}$$

$$M_3^r = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 2 \\ 1 & 2 & 1 & 2 & 0 & 1 & 2 & 2 & 1 & 1 \\ 2 & 1 & 1 & 1 & 1 & 2 & 1 & 1 & 2 & 3 \end{bmatrix}$$

$$\vdots$$

**b**

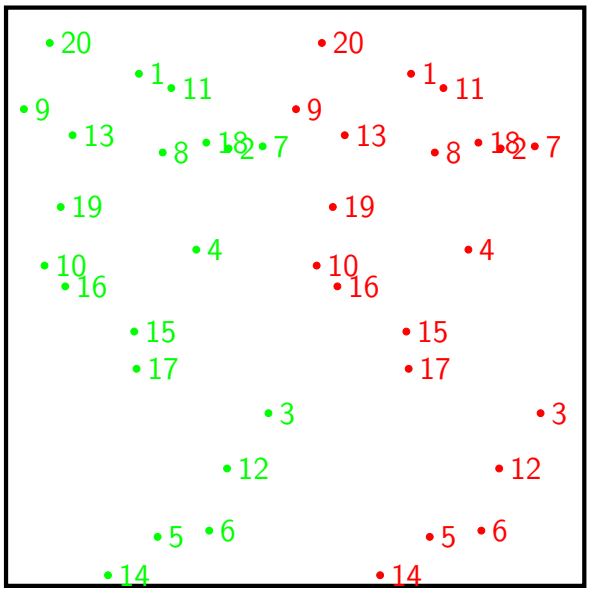



**Figure 4.6:** The correspondence between the peaks can be deduced by a comparison of the neighbor count matrices. (a) Each peak in each channel can be described by an independent neighbor count matrix, leading to a list of feature matrices that can be compared with each other elementwise to deduce the peak pairs corresponding to the same construct on the slide surface. This leads to a list of constraints on the coordinate pairs on the two signal channels (b).

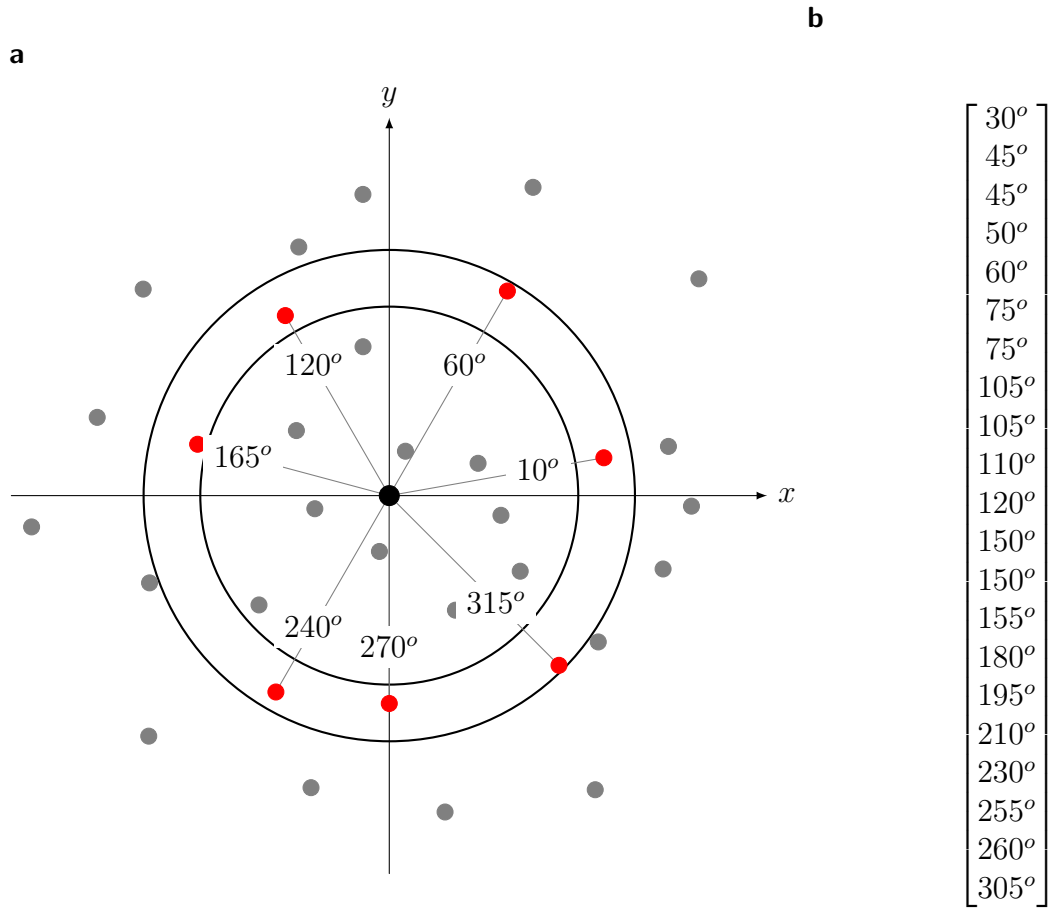



**Figure 4.7:** Angles between rays to the neighbor spots as a feature (a) Angles between each pair of neighbors that are located within a pre-defined radial distance range can be used as a computational feature to identify corresponding spots on a slide image. (b) The sorted list forming a vector representing the central spot using this scheme that can be used for point comparisons. Each entry is the angle between two neighbors within the acceptable band, as an example, both $120^o - 10^o = 110^o$ and $240^o - 120^o = 120^o$ are on the list as there are three neighbors in acceptable band that make $10^o$, $120^o$ *and* $240^o$ with the x-axis.

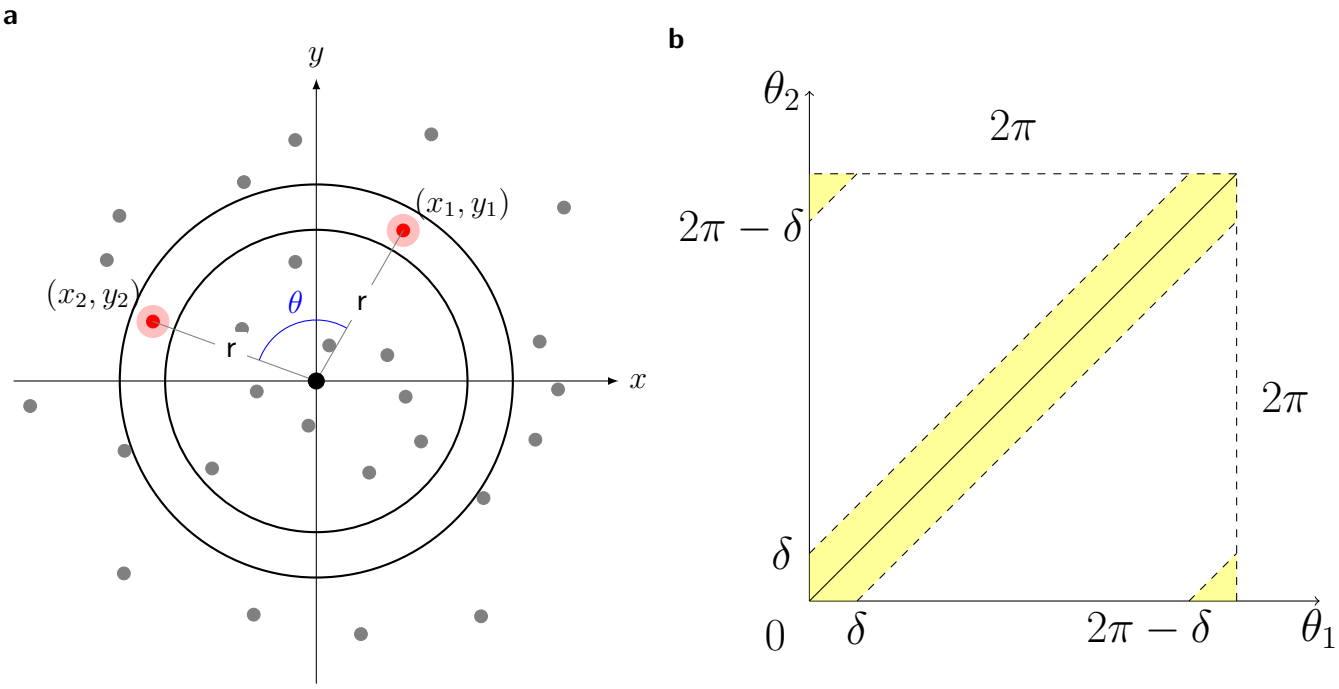



**Figure 4.8:** Experimental error tolerance of angle features. (a) The estimation of the error in the angle measurement between vertices extending toward two arbitrarily chosen neighbors for a given imprecision in the spot localization in a microscopy image. (b) Quantification of the false positive detection rate as a function of angular error tolerance during comparison of two angles out of the feature list. An arbitrarily chosen pair of polar angles $(\theta_1, \theta_2)$ can be anywhere in $[0, 2\pi)x[0, 2\pi)$, whereas consideration of two angles that are similar up to a threshold value ( $|\theta_1 - \theta_2| < \delta$ ) will lead to a zone roughly following $\theta_1 = \theta_2$ that is recorded as a hit. The probability that a false positive hit is recorded is given by the relative area of the yellow zone, i.e. $p-value = \frac{2\delta 2\pi}{(2\pi)^2} = \delta/\pi$.

**Table 4.1:** Mapping parameters obtained by comparing the list of pairwise angles out of a bead slide movie. Data is provided as mean and standard deviation. Two spots were considered to be corresponding if #angles matching within 0.005rad was at least 28.

| #Point pairs used | | a | b ($x10^{-4}$) | f | c ($x10^{-4}$) | d | g |
|---|---|---|---|---|---|---|---|
| 3 | $\mu$ | 0.9977 | -2.5697 | 254.66 | -8.0668 | 1.0028 | -0.14087 |
| | $\sigma$ | 0.0163 | 105.78 | 2.2998 | 78.639 | 0.0058 | 1.58211 |
| 5 | $\mu$ | 0.9991 | 9.2749 | 254.11 | 29.341 | 1.0041 | -0.89959 |
| | $\sigma$ | 0.0019 | 12.627 | 0.28 | 24.580 | 0.0014 | 0.35193 |
| 10 | $\mu$ | 0.9990 | 7.2867 | 254.10 | 15.644 | 1.0041 | -0.73294 |
| | $\sigma$ | 0.0017 | 6.7229 | 0.28 | 18.499 | 0.0006 | 0.18030 |
| 20 | $\mu$ | 0.9984 | 5.6190 | 254.21 | 19.264 | 1.0038 | -0.69567 |
| | $\sigma$ | 0.0010 | 3.6901 | 0.13 | 8.3494 | 0.0004 | 0.07742 |
| 100 | | 0.9987 | -4.6952 | 254.43 | 13.059 | 1.0032 | -0.48269 |
| Ideal | | 1 | 0 | 256 | 0 | 1 | 0 |

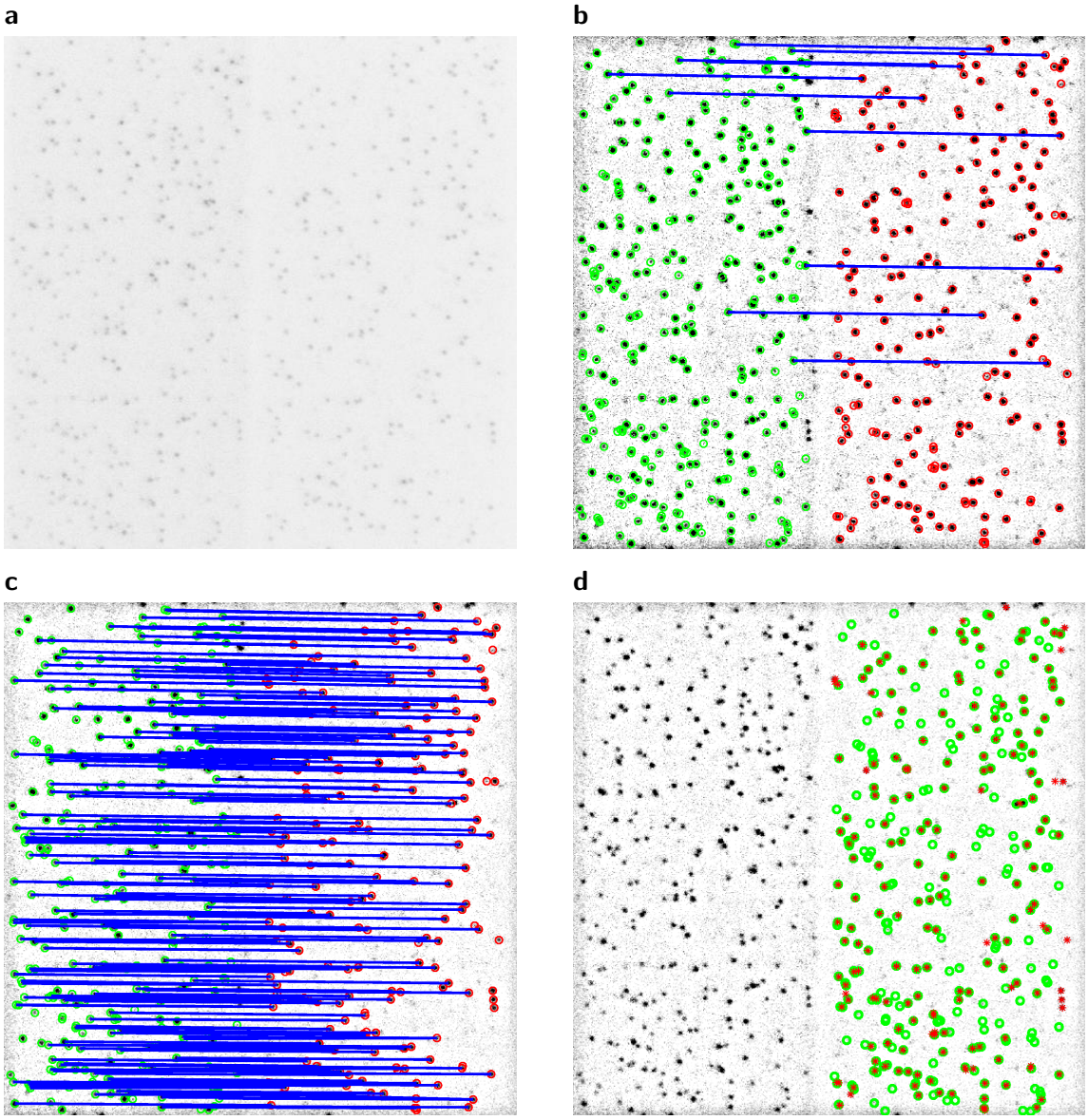



**Figure 4.9:** Deduction of mapping parameters out of an smFRET movie, obtained using PEG slide immobilized Cy3/Cy5 double conjugated DNA oligo, chemically synthesized by IDT. (a) Input image to be processed obtained by alternating excitation with 532nm and 638nm, processed by maximum intensity projection across the two modalities, shown in negative (b). Correspondence between spot pairs deduced by the initial attempt via angle between pairs of neighbors. (c). The correspondence obtained after one round of iterative closest point approach, using which a final map is evaluated. (d) Application of the deduced transformation on the left (Cy3) channel to the right (Cy5) channel.

**Table 4.2:** The 12 fine-mapping parameters obtained using the angles between neighbor pairs. The raw input image was either acquired on 100nm Tetraspeck fluorescent beads illuminated with 532nm laser or Cy3/C5 double conjugated DNA alternatively illuminated with 532nm and 638nm. The deviations indicate the standard error of the mean out of 16 different experimental replicate movies of dye-conjugated DNA and 5 replicate movies of fluorescent beads.

| | | Beads | | Cy3/Cy5 DNA | |
|---|---|---|---|---|---|
| Parameter | Ideal | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| a | 1 | 0.9938 | 0.0016 | 0.9938 | 0.0008 |
| b ($10^{-6}$) | 0 | -59.855 | 734.66 | -259.02 | 248.74 |
| c ($10^{-6}$) | 0 | 3.1952 | 4.6798 | 4.7771 | 3.7743 |
| d ($10^{-6}$) | 0 | 0.5236 | 1.5369 | -0.2628 | 0.6124 |
| e ($10^{-6}$) | 0 | -1.3402 | 0.9731 | 0.3256 | 1.7139 |
| f | 256 | 254.18 | 0.0412 | 254.42 | 0.0696 |
| g ($10^{-6}$) | 0 | -2903.7 | 1156.3 | -617.14 | 1065.2 |
| h | 1 | 0.9930 | 0.0011 | 0.9932 | 0.0003 |
| i ($10^{-6}$) | 0 | 8.8282 | 4.3115 | 2.7018 | 3.7737 |
| j ($10^{-6}$) | 0 | 0.1215 | 1.7204 | -0.2677 | 0.5409 |
| k ($10^{-6}$) | 0 | 1.6311 | 2.1214 | 0.5683 | 1.400 |
| l | 0 | 4.9404 | 0.1331 | 4.8104 | 0.0918 |

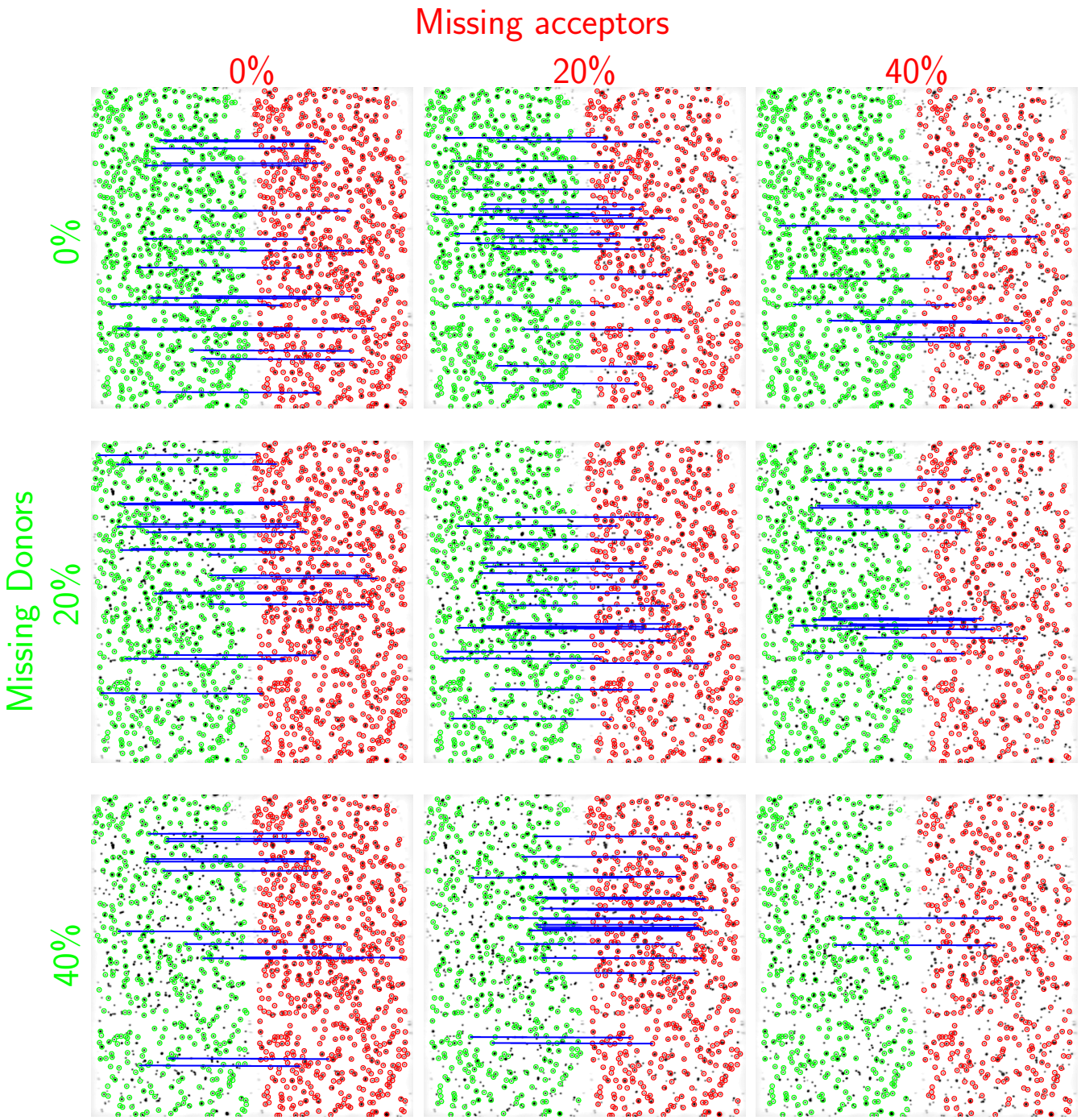



**Figure 4.10:** The proposed approach works despite incomplete dye conjugation and photobleaching. The unavailability of a donor or acceptor leads to the presence of orphan spots in the field of view without a corresponding spot in the other emission channel. Percentages in each row or column indicate the fraction of spots randomly chosen and removed (not circled) from the dataset after detection to simulate incomplete labeling. Random removal of 40% or more spots leads to spurious matches and mapping failure. Only up to 20 matches are displayed for visual clarity. hitThreshold = 25 and acceptanceThreshold = 0.005.

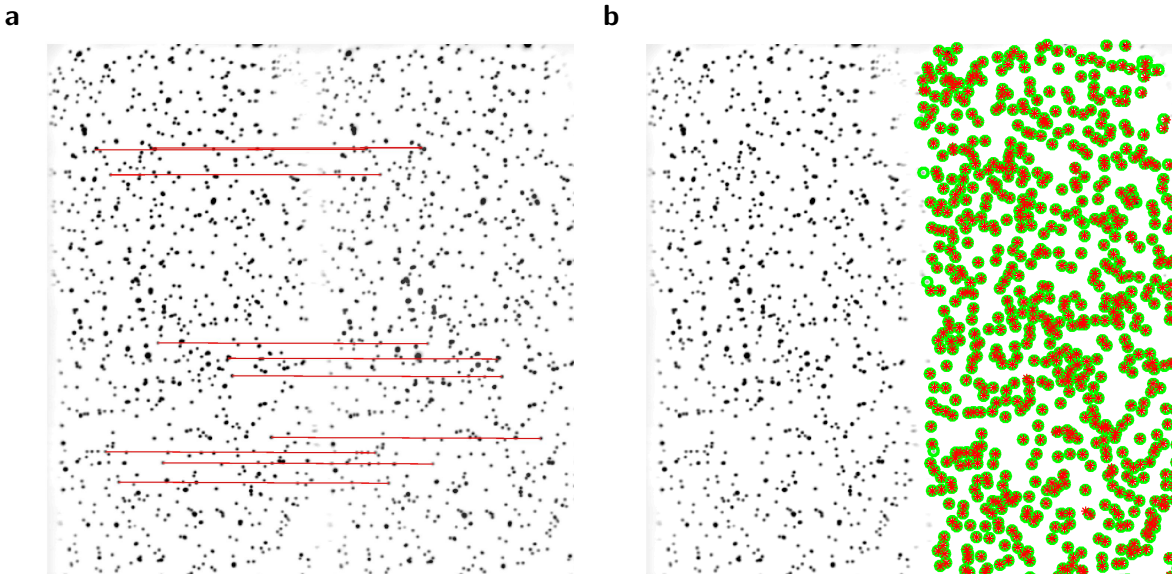



**Figure 4.11:** Angle between each pair of neighbors at a defined radial distance range leads to accurate mapping of emission channels. (a) n=10 pairs were chosen by searching for counterparts of randomly chosen spots on the right channel of a bead image. The detected spot pairs were connected for visualization by the red lines (—). (b) The obtained map from this procedure was used to determine the projection of all detected spots on the left channel on the right channel. The green circles (o) represent the position of spots as detected directly on the right channel, whereas the red stars (*) are the projected positions of the left channel. Virtually all points from the two sets overlap almost perfectly, an indication of successful mapping.

# Chapter 5

# Appendices

## 5.1 Detailed information on cell strains

We purchased K-12 collection of *E. coli* from the Coli Genetic Stock Center in Yale University [35]. The genetic background of the strains used in the studies are as follows:

| | | | |
|---|---|---|---|
| wild type | BW25113 | #7636 | $F^-, \Delta(araD-araB)567, \Delta lacZ4787(::rnB-3), \lambda^-, rph-1, \Delta(rhaD-rhaB)568, hsdR514$ |
| $\Delta mutS$ | JW2703-2 | #10126 | $F^-, \Delta(araD-araB)567, \Delta lacZ4787(::rnB-3), \lambda^-, rph-1, \Delta mutS738::kan, \Delta(rhaD-rhaB)568, hsdR514$ |
| $\Delta mutL$ | JW4128-1 | #10971 | $F^-, \Delta(araD-araB)567, \Delta lacZ4787(::rnB-3), \lambda^-, rph-1, \Delta mutL720::kan, \Delta(rhaD-rhaB)568, hsdR514$ |
| $\Delta mutH$ | JW2799-2 | #10186 | $F^-, \Delta(araD-araB)567, \Delta lacZ4787(::rnB-3), \lambda^-, rph-1, \Delta mutH756::kan, \Delta(rhaD-rhaB)568, hsdR514$ |
| $\Delta uvrD$ | JW3786-5 | #10752 | $F^-, \Delta(araD-araB)567, \Delta lacZ4787(::rnB-3), \lambda^-, rph-1, \Delta uvrD769::kan, \Delta(rhaD-rhaB)568, hsdR514$ |
| $\Delta uvrB$ | JW0762-2 | #8819 | $F^-, \Delta(araD-araB)567, \Delta lacZ4787(::rnB-3), \lambda^-, rph-1, \Delta uvrB751::kan, \Delta(rhaD-rhaB)568, hsdR514$ |
| $\Delta dam$ | JW3350-2 | #11675 | $F^-, \Delta(araD-araB)567, \Delta lacZ4787(::rnB-3), \lambda^-, rph-1, \Delta dam-722::kan, \Delta(rhaD-rhaB)568, hsdR514$ |
| $\Delta recA$ | BW26355 | #7651 | $F^-, \Delta(araD-araB)567, \Delta lacZ4787(::rnB-3), \lambda^-, rph-1, \Delta recA635::kan, \Delta(rhaD-rhaB)568, hsdR514$ |

Out of the above cell strains we procured, we constructed the new cell strains below via recombineering where the *dam* was replaced with a chloramphenicol resistance gene driven together with its own promoter cloned from the pGGAselect vector.

| | |
|---|---|
| $\Delta mutS\ \Delta dam$ | $F^-, \Delta(araD-araB)567, \Delta lacZ4787(::rnB-3), \lambda^-, rph-1, \Delta mutS738::kan, \Delta(rhaD-rhaB)568, hsdR514, \Delta dam-722::cam$ |

$\Delta mutL\ \Delta dam$ $F^-, \Delta(araD - araB)567, \Delta lacZ4787(::rnB - 3), \lambda^-, rph - 1, \Delta mutL720 :: kan, \Delta(rhaD - rhaB)568, hsdR514, \Delta dam - 722 :: cam$

$\Delta mutH\ \Delta dam$ $F^-, \Delta(araD - araB)567, \Delta lacZ4787(::rnB - 3), \lambda^-, rph - 1, \Delta mutH756 :: kan, \Delta(rhaD - rhaB)568, hsdR514, \Delta dam - 722 :: cam$

$\Delta uvrD\ \Delta dam$ $F^-, \Delta(araD - araB)567, \Delta lacZ4787(::rnB - 3), \lambda^-, rph - 1, \Delta uvrD769 :: kan, \Delta(rhaD - rhaB)568, hsdR514, \Delta dam - 722 :: cam$

## 5.2 Detailed list of all DNA sequences used

All oligos were purchased from IDT, standard desalted unless indicated otherwise.

### 5.2.1 Primers used for recombineering and verifications

The **bold** segments indicate non-binfing sequences homologous to the *dam* locus.

R1p

/5/ **GTCGGAGCTTTCTCCACAGCCGGAGAAGGTGTAATTAGTTAGTCAGC** GAAGATCCTTTGATCTTTTCTACGGGGT /3/

R2p

/5/ **ATACTGTTTCATCCGCTTCTCCTTGAGAATTATTTTTTCGCGGGT-GAAAC** TATCCGCTCATGAGTAGCACCA /3/

Dam verify Fwd

/5/ AAGCGGTATCTACATTGCCAGC /3/

Dam verify Rev

/5/ CAAGGATTTCAGCACCATTGGC /3/

MutS verify Fwd

/5/ CCATCACACCCCATTTAATATCAGGGA /3/

MutS verify Rev

/5/ CGATAGCAAAAGACTATCGGGAATTGTTATTACA /3/

MutL verify Fwd

/5/ GTACGGTGACGACGCCAGATC /3/

MutL verify Rev

/5/ TCGCCTTAGGCAGGCTCGC /3/

MutH verify Fwd

/5/ ACACTGCGAATATTCGGCACATAATTGC /3/

MutH verify Rev

/5/ CGGCAGGTCAAAGCGATGGCTA /3/

UvrD verify Fwd

/5/ TAAGGTGCGCAGCACCGCAT /3/

UvrD verify Rev

/5/ TTGCGCTTCTCCGCCCAACC /3/

### 5.2.2 Primers for the barcoded vector

Primer P1

/5/ TTTTTTT CTCGAG GCAAGCTTGGCGTAATCATGGTCAT /3/

Primer P2 - 25bp barcode

/5/ TTTTTTT GAGCTC NNNNNTNNNNNTNNNNNTNNNNNTNNNNN TGCGGTATTTCACACCGCATATGGT /3/

Primer P2 - 20bp barcode

/5/ TTTTTTT GAGCTC NNNNNTNNNNNTNNNNNTNNNNN TGCGGTATTTCACACCGCATATGGT /3/

Primer inv P1

/5/ TTTTTTT GAGCTC GCAAGCTTGGCGTAATCATGGTCAT /3/

Primer inv P2

/5/ TTTTTTT CTCGAG NNNNNTNNNNNTNNNNNTNNNNNTNNNNN TGCGGTATTTCACACCGCATATGGT /3/

Primer P4

/5/ TTTTTTT CTCGAG GGTGCCTAATGAGTGAGCTAACTCAC /3/

pUC19 Sanger sequencing primer

/5/ CACAGCTTGTCTGTAAGCGG /3/

### 5.2.3 Primers for sequencing library preparation

Primer S1

/5/ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CATATGCGGTGTGAAATACCGCA /3/

Primer S2

/5/ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG GCTAGTACCTCAATATAGACTCCCT /3/

Primer S2_2

/5/ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG AAGCTTGCCTCGACAGAATAGGAAC /3/

Primer S2_3

/5/ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG GCCAAGCTTGCCTCGAGCT /3/

Primer S2_4

/5/ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ATTACGCCAAGCTTGCCTCGAG /3/

Primer S2_5

/5/ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ATTACGCCAAGCTTGCCTCGA /3/

Primer S2_6

/5/ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ATTACGCCAAGCTTGCGAGCT /3/

Primer S2_7

/5/ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG TCACTCATTAGGCACC CTCGAG /3/

The orientation on the sequencing chip was inverted by using:

Reverse S1

/5/ TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ATTACGCCAAGCTTGCCTCGAG /3/

Reverse S2_4

/5/ GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG CATATGCGGTGTGAAATACCGCA /3/

Indexing primers, where [i5] and [i7] denote 8bp multiplexing barcodes (purchased from Illumina, FC-131-1001):

i7 primer

/5/ CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG /3/

i5 primer

/5/ AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTC /3/

### 5.2.4 Oligos to construct mismatch libraries

FML library prototype

/5/ ATGGGTCAATTAGAAGTCATAGGAGAAGTAAGGGAGTCTATATTGAGGTACTAGC /3/

C (FML)

/5/ TCGA GCTAGTACCTCAATATAGACTCCCT TACTTCTCCTATGACTTCTAATTGACCCAT AGCT /3/

SML/DML/IL/NPL library prototype

/5/ GGTCCATCCTCAACAACGAGCTCGAATGGGCTATCTGTTGAGTCGTAGCA CTGGTTCCTATTCTG

C2 (SML/DML/IL/NPL)

/5/ TCGA CAGAATAGGAACCAG TGCTACGACTCAACAGATAGCCCATTCGAGCTCGTTGTTGAGGATGGACC AGCT

DEB3L prototype (54 oligos that contain 1 substitution each)

/5/ TTTATTCTTGTAATACTAGTCATCCTCGTGATGCTGGAAACAAGACCACGAGCAGGCCCG /3/

C - DEB3L

/5/ TCGA CGGGCCTGCTCGTGGTCTTGTTTCCAGCATCACGAGGATGACTAGTATTACAAGAATAAA AGCT /3/

DEB3R prototype (54 oligos that contain 1 substitution each)

/5/ GCTGGAAACAAGACCACGAGCAGGCCCGCGGGTTTATTCTTGTAATACTAGTCATCCTCG /3/

C - DEB3L

/5/ TCGA CGAGGATGACTAGTATTACAAGAATAAACCCGCGGGCCTGCTCGTGGTCTTGTTTCCAGC AGCT /3/

### 5.2.5 Cyclically permuted oligos for single mismatch control experiment

The small case letters indicate the position of the mismatch upon annealing.

L: Low repair efficiency (forms a CC mismatch)

H: High repair efficiency (forms a CT mismatch)

P: Mismatch is at the barcode-proximal position

M: Mismatch is at the barcode-mesial position

D: Mismatch is at the barcode-distal position

Variable strand of L-P

/5/ TCGAATGcGCTATCTGTTGAGTCGTAGCACTGGTTCCTATTCTGGGTCCATCCT-CAACAACGAGC /3/

Common strand of L-P

/5/ TCGA GCTCGTTGTTGAGGATGGACCCAGAATAGGAACCAGTGCTACGACT-CAACAGATAGCcCATTCGA AGCT /3/

Variable strand of L-M (mismatch is at the same position as in SML)

/5/ GGTCCATCCTCAACAACGAGCTCGAATGcGCTATCTGTTGAGTCGTAGCACTG-GTTCCTATTCTG

Common strand of L-M (same as C2)

/5/ TCGA CAGAATAGGAACCAGTGCTACGACTCAACAGATAGCcCATTC-GAGCTCGTTGTTGAGGATGGACC AGCT /3/

Variable strand of L-D

/5/ TTGAGTCGTAGCACTGGTTCCTATTCTGGGTCCATCCTCAACAACGAGCTCGAAT-GcGCTATCTG /3/

Common strand of L-D

/5/ TCGA CAGATAGCcCATTCGAGCTCGTTGTTGAGGATGGACCCAGAATAG-GAACCAGTGCTACGACTCAA AGCT /3/

Variable strand of H-P (mismatch is at the same position as in SML)

/5/ GGTCCATCCTCAcCAACGAGCTCGAATGGGCTATCTGTTGAGTCGTAGCACTG-GTTCCTATTCTG /3/

Common strand of H-P (same as C2)

/5/ TCGA CAGAATAGGAACCAGTGCTACGACTCAACAGATAGCCCATTC-GAGCTCGTTGtTGAGGATGGACC AGCT /3/

Variable strand of H-M

/5/ TAGCACTGGTTCCTATTCTGGGTCCATCCTCAcCAACGAGCTCGAATGGGCTATCT-GTTGAGTCG /3/

Common strand of H-M

/5/ TCGA CGACTCAACAGATAGCCCATTCGAGCTCGTTGtTGAGGATGGACCCA-GAATAGGAACCAGTGCTA AGCT /3/

Variable strand of H-D

/5/ TCGAATGGGCTATCTGTTGAGTCGTAGCACTGGTTCCTATTCTGGGTCCATCCT-CAcCAACGAGC /3/

Common strand of H-D

/5/ TCGA GCTCGTTGtTGAGGATGGACCCAGAATAGGAACCAGTGCTACGACTCAACA-GATAGCCCATTCGA AGCT /3/

### 5.2.6 Primers used to generate double-barcoded mismatch libraries

Primer Phospho-Z1

/5/5Phos/ ATGGGACCGCATCGTAGCTT /3/

Primer Phospho-Z2

/5/5Phos/ CGGCTGAATGGTACCCGATA /3/

Primer Phospho-Z3

/5/5Phos/ GAGCGCAGCTGGTGTAGACA /3/

Primer Phospho-Z4

/5/5Phos/ CGACTTCGAATTCAGCACGT /3/

Primer Phospho-Z5

/5/5Phos/ ACACCCGCTCGATCCCTTAT /3/

Primer Phospho-Z6

/5/5Phos/ GGTTGAACAACCGCCGGTAT /3/

Primer Phospho-Z7

/5/5Phos/ AAATTGCCGTTGCGGATTTC /3/

Primer Thio-Z13, (*) indicates the position of a phosphothioate bond

/5/biotin/ C*G*C*T*C*CTTGTTGTACTCGCA /3/

Primer Thio-Z14, (*) indicates the position of a phosphothioate bond

/5/biotin/ T*G*A*C*G*GAGGATAGAAGGCCA /3/

### 5.2.7 Double barcoded mismatch library sequence prototypes

Adaptor sites are indicated in blue and red, while the restriction sites are shown in **bold font**. N's indicate the random bases that constitute the mapping barcodes, ideally containing ACGT bases with 25% probability each. Required primer pair for amplification are indicated in parentheses following the library name.

3ML1 prototype (Z13, Z6)

/5/ TTCTAGA CGCTCCTTGTTGTACTCGCA **GAGCTC** NNNNNNN TTTATTCTTG-TAATACTAGTCATCCTCGTGATGCTGGAAACAAGACCACGAGCAGGCCCGCGGGTT **CTCGAG** ATACCGGCGGTTGTTCAACC /3/

5ML1 prototype (Z13, Z1)

/5/ CGCTCCTTGTTGTACTCGCA **GAGCTC** NNNNNN AAAAACAAAA-GAAAATAAACCAAACGAAACTAAAGCAAAGGAAAGTAAATCAAATGAAAT-TAACACAACAGAACATAACCCAAC **CTCGAG** AAGCTACGATGCGGTCCCAT /3/

5ML2 prototype (Z13, Z2)

/5/ CGCTCCTTGTTGTACTCGCA **GAGCTC** NNNNNN CAACCGAACC-TAACGCAACGGAACGTAACTCAACTGAACTTAAGACAAGAGAAGATAAGC-CAAGCGAAGCTAAGGCAAGGGAAG **CTCGAG** TATCGGGTACCATTCAGCCG /3/

5ML3 prototype (Z13, Z3)

/5/ CGCTCCTTGTTGTACTCGCA **GAGCTC** NNNNNN GAAGGTAAGT-CAAGTGAAGTTAATACAATAGAATATAATCCAATCGAATCTAATGCAATG-GAATGTAATTCAATTGAATTTACA **CTCGAG** TGTCTACACCAGCTGCGCTC /3/

5ML4 prototype (Z13, Z4)

/5/ CGCTCCTTGTTGTACTCGCA **GAGCTC** NNNNNN TACACCACAC-GACACTACAGCACAGGACAGTACATCACATGACATTACCAGACCATACCC-

CACCCGACCCTACCGCACCGGACC **CTCGAG** ACGTGCTGAATTCGAAGTCG /3/

5ML5 prototype (Z13, Z5)

/5/ CGCTCCTTGTTGTACTCGCA **GAGCTC** NNNNNN GACCGTACCT-CACCTGACCTTACGAGACGATACGCCACGCGACGCTACGGCACGGGACG-GTACGTCACGTGACGTTACTAGACT **CTCGAG** ATAAGGGATCGAGCGGGTGT /3/

5ML6 prototype (Z13, Z6)

/5/ CGCTCCTTGTTGTACTCGCA **GAGCTC** NNNNNN GACTATACTC-CACTCGACTCTACTGCACTGGACTGTACTTCACTTGACTTTAGAGCAGAG-GAGAGTAGATCAGATGAGATTAGC **CTCGAG** ATACCGGCGGTTGTTCAACC /3/

5ML7 prototype (Z13, Z7)

/5/ CGCTCCTTGTTGTACTCGCA **GAGCTC** NNNNNN TAGCATAGC-CCAGCCGAGCCTAGCGCAGCGGAGCGTAGCTCAGCTGAGCTTAGGATAG-GCCAGGCGAGGCTAGGGCAGGGGAGG **CTCGAG** GAAATCCGCAACGGCAATTT /3/

5ML8 prototype (Z14, Z1)

/5/ TGACGGAGGATAGAAGGCCA **GAGCTC** NNNNNN GAGGGTAG-GTCAGGTGAGGTTAGTATAGTCCAGTCGAGTCTAGTGCAGTGGAGTG-TAGTTCAGTTGAGTTTATATCATATGATA **CTCGAG** AAGCTACGATGCGGTCCCAT /3/

5ML9 prototype (Z14, Z2)

/5/ TGACGGAGGATAGAAGGCCA **GAGCTC** NNNNNN GATATTATCC-CATCCGATCCTATCGCATCGGATCGTATCTCATCTGATCTTATGCCATGC-GATGCTATGGCATGGGATGGTATG **CTCGAG** TATCGGGTACCATTCAGCCG /3/

5ML10 prototype (Z14, Z3)

/5/ TGACGGAGGATAGAAGGCCA **GAGCTC** NNNNNN TATGTCATGTGATGTTATTCCATTCGATTCTATTGCATTGGATTGTATTTCATTTGATTTTCCCCCGCCCCTCCCGGCCCGTCC **CTCGAG** TGTCTACACCAGCTGCGCTC /3/

5ML11 prototype (Z14, Z4)

/5/ TGACGGAGGATAGAAGGCCA **GAGCTC** NNNNNN GTCCCTGCCCTTCCGCGCCGCTCCGGGCCGGTCCGTGCCGTTCCTCGCCTCTCCTGGCCTGTCCTTGCCTTTCGCGGCGCGTCG **CTCGAG** ACGTGCTGAATTCGAAGTCG /3/

5ML12 prototype (Z14, Z5)

/5/ TGACGGAGGATAGAAGGCCA **GAGCTC** NNNNNN GTCGCTGCGCTTCGGCTCGGGGCGGGTCGGTGCGGTTCGTCTCGTGGCGTGTCGTTGCGTTTCTCTGCTCTTCTGGGCTGGTCT **CTCGAG** ATAAGGGATCGAGCGGGTGT /3/

5ML13 prototype (Z14, Z6)

/5/ TGACGGAGGATAGAAGGCCA **GAGCTC** NNNNNN GTCTGTGCTGTTCTTGGCTTGTCTTTGCTTTTGGGGGTGGGTTGGTGTGGTTTGTGTTGTTTTTAAAAACAAAAGAAAATAAAC **CTCGAG** ATACCGGCGGTTGTTCAACC /3/

SSL1 prototype (Z13, Z1)

/5/ CGCTCCTTGTTGTACTCGCA **GAGCTC** NNNNNNN GGTTTTCTTTAGGTTTTTAGCTTTAGCTTTATTTTATTTTACCTTTAAGGTTTTGTTTTATCCTTTATTTTAGTTTTAAG **CTCGAG** AAGCTACGATGCGGTCCCAT /3/

SSL2 prototype (Z13, Z2)

/5/ CGCTCCTTGTTGTACTCGCA **GAGCTC** NNNNNNN TTATTT-
TAGTTTTAAGTTTTGTTTTCTTTATCCTTTACCTTTAGCTTTAACTT-
TAGTTTTCTTTAACTTTTAGCTTTACT **CTCGAG** TATCGGGTACCATTCAGCCG
/3/

SSL3 prototype (Z13, Z3)

/5/ CGCTCCTTGTTGTACTCGCA **GAGCTC** NNNNNNN AACTTT-
TAGCTTTACTTTTATCGTTTTGTTTTGTTTTTATCGTTTTCTTTACGTTT-
TACGTTTTTACTTTACTTTAACTT **CTCGAG** TGTCTACACCAGCTGCGCTC /3/

SSL4 prototype (Z13, Z4)

/5/ CGCTCCTTGTTGTACTCGCA **GAGCTC** NNNNNNN TACTTTACTTTAACTTTATTT-
TAACTTTTACTTTTAAGGTTTTAGTTTTTAAGGTTTTCTTTAGGTTTTTAGCTTTAGCT
**CTCGAG** ACGTGCTGAATTCGAAGTCG /3/

SIL prototype (Z13, Z6, 126 bases including one inserted extra base)

/5/ CGCTCCTTGTTGTACTCGCA GAGCTC NNNNNNN TTTATTCTTGTAATACTAGT-
CATCCTCGTGATGCTGGAAACAAGACCACGAGCAGGCCCGCGGGTT CTCGAG ATAC-
CGGCGGTTGTTCAACC /3/

DIL prototype (Z14, Z6, 128 bases, including inserted two extra bases)

/5/ TGACGGAGGATAGAAGGCCA GAGCTC NNNNNNNN TTTATTCTTGTAATAC-
TAGTCATCCTCGTGATGCTGGAAGCAAGACCACGAGCAGGCCCGCGGGTT CTCGAG
ATACCGGCGGTTGTTCAACC /3/

### 5.2.8 Oligos used as common strands of double-barcoded mismatch libraries

We annealed the ssDNA libraries we obtained to chemically synthesized oligos, 100 bases each, purchased with HPLC purification individually.

C-3ML1

/5/ CGGTAT **CTCGAG** AACCCGCGGGCCTGCTCGTGGTCTTGTTTCCAGCATCACGAGGATGACTAGTATTACAAGAATAAA /3/

C-5ML1

/5/ ATCGTAGCTT **CTCGAG** GTTGGGTTATGTTCTGTTGTGTTAATTTCATTTGATTTACTTTCCTTTGCTTTAGTTTCGTTTGGTTTATTTTCTTTTGTTTTT /3/

C-5ML2

/5/ GTACCCGATA **CTCGAG** CTTCCCTTGCCTTAGCTTCGCTTGGCTTATCTTCTCTTGTCTTAAGTTCAGTTGAGTTACGTTCCGTTGCGTTAGGTTCGGTTG /3/

C-5ML3

/5/ GGTGTAGACA **CTCGAG** TGTAAATTCAATTGAATTACATTCCATTGCATTAGATTCGATTGGATTATATTCTATTGTATTAACTTCACTTGACTTACCTTC /3/

C-5ML4

/5/ TTCAGCACGT **CTCGAG** GGTCCGGTGCGGTAGGGTCGGGTGGGGTATGGTCTGGTAATGTCATGTGATGTACTGTCCTGTGCTGTAGTGTCGTGTGGTGTA /3/

C-5ML5

/5/ GATCCCTTAT **CTCGAG** AGTCTAGTAACGTCACGTGACGTACCGTCCCGTGCCGTAGCGTCGCGTGGCGTATCGTCTCGTAAGGTCAGGTGAGGTACGGTC /3/

C-5ML6

/5/ CCGCCGGTAT **CTCGAG** GCTAATCTCATCTGATCTACTCTCCTCTGCTCTAAAGT-CAAGTGAAGTACAGTCCAGTGCAGTAGAGTCGAGTGGAGTATAGTC /3/

C-5ML7

/5/ TGCGGATTTC **CTCGAG** CCTCCCCTGCCCTAGCCTCGCCTGGCCTATCC-TAAGCTCAGCTGAGCTACGCTCCGCTGCGCTAGGCTCGGCTGGGCTATGCTA /3/

C-5ML8

/5/ ATCGTAGCTT **CTCGAG** TATCATATGATATAAACTCAACTGAACTACACTCCACT-GCACTAGACTCGACTGGACTATACTAACCTCACCTGACCTACCCTC /3/

C-5ML9

/5/ GTACCCGATA **CTCGAG** CATACCATCCCATGCCATAGCATCGCATGGCATAA-GATCAGATGAGATACGATCCGATGCGATAGGATCGGATGGGATAATATC /3/

C-5ML10

/5/ GGTGTAGACA **CTCGAG** GGACGGGCCGGGAGGGGCGGGGGAAAATCAAAT-GAAATACAATCCAATGCAATAGAATCGAATGGAATAACATCACATGACATA /3/

C-5ML11

/5/ TTCAGCACGT **CTCGAG** CGACGCGCCGCGAAAGGCAAGGACAGGCCAGGAGAG-GCGAGGAACGGCACGGACCGGCCCGGAGCGGCGCGGAAGGGCAGGGAC /3/

C-5ML12

/5/ GATCCCTTAT **CTCGAG** AGACCAGCCCAGAAGAGCAGAGAAACGCAACGA-CACGCCACGAGACGAACCGCACCGACCCGCCCCGAGCCGAAGCGCAGCGAC /3/

C-5ML13

/5/ CCGCCGGTAT **CTCGAG** GTTTATTTTCTTTTGTTTTTAAAAACAACACAAACCA-CACCAACCCACCCCCAAAAGCAAAGACAAGCCAAGAACAGCACAGAC /3/

C-SSL1

/5/ CCGCATCGTAGCTT **CTCGAG** CTTAAAACTAAAATAAAGGATAAAACAAAACCT-TAAAGGTAAAATAAAATAAAGCTAAAGCTAAAAACCTAAAGAAAACC /3/

C-SSL2

/5/ AATGGTACCCGATA **CTCGAG** AGTAAAGCTAAAAGTTAAAGAAAACTAAAGT-TAAAGCTAAAGGTAAAGGATAAAGAAAACAAAACTTAAAACTAAAATAA /3/

C-SSL3

/5/ AGCTGGTGTAGACA **CTCGAG** AAGTTAAAGTAAAGTAAAAACGTAAAACGTAAA-GAAAACGATAAAAACAAAACAAAACGATAAAAGTAAAGCTAAAAGTT /3/

C-SSL4

/5/ CGAATTCAGCACGT **CTCGAG** AGCTAAAGCTAAAAACCTAAAGAAAACCT-TAAAAACTAAAACCTTAAAAGTAAAAGTTAAAATAAAGTTAAAGTAAAGTA /3/

**Table 5.1:** Commonly used abbreviations describing different plasmid constructs used in this study.

| Vector | Primers for production | Primers for sequencing | Description |
|---|---|---|---|
| tUNC19 | P1 & P2 | S1 & S2_4 | Standard barcoded library, a derivative of pUC19, used by default unless indicated otherwise. |
| mm19 | P1 & P2 | S1 & S2_4 | Both strands of the plasmid are methylated. Obtained by Dam treatment of the tUNC19 PCR amplicon. |
| hm19 | P1 & P2 | S1 & S2_4 | Hemi-methylated vector. Obtained by performing primer extension on Dam treated amplicon to displace one of the two methylated strands. |
| tUNC19-X | P1 & P2 | S1 & S2_4 | Undergoes the same primer extension procedure as hm19, but was not subjected to Dam treatment leading to a fully unmethylated product. |
| short19 | P1 & P4 | S1 & S2_7 | Unmodified vector similar to tUNC19, but the lacZ promoter immediately upstream has been eliminated by the choice of primers. |
| 91CNUt | invP1 & invP2 | S1 & S2_6 | The positions of the restriction sites are flipped, resulting in inverted insertion orientation of the mismatch libraries compared to tUNC19. |

## 5.3 References


[1] Sawami Kobayashi, Michael R. Valentine, Phuong Pham, Mike O'Donnell, and Myron F. Goodman. Fidelity of escherichia coli dna polymerase iv: Preferential generation of small deletion mutations by dntp-stabilized misalignment. *Journal of Biological Chemistry*, 277(37):34198–34207, 2002.

[2] R Wagner and M Meselson. Repair tracts in mismatched dna heteroduplexes. *Proceedings of the National Academy of Sciences*, 73(11):4135–4139, 1976.

[3] R M Schaaper. Base selection, proofreading, and mismatch repair during dna replication in escherichia coli. *Journal of Biological Chemistry*, 268(32):23762–23765, 1993.

[4] Iwona J. Fijalkowska, Roel M. Schaaper, and Piotr Jonczyk. Dna replication fidelity in escherichia coli: a multi-dna polymerase affair. *FEMS Microbiology Reviews*, 36(6):1105–1121, 2012.

[5] Frederick R. Blattner, Guy Plunkett, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau, and Ying Shao. The complete genome sequence of escherichia coli k-12. *Science*, 277(5331):1453–1462, 1997.

[6] Paul Modrich. Dna mismatch correction. *Annual Review of Biochemistry*, 56(1):435–466, 1987. PMID: 3304141.

[7] Thomas A. Kunkel and Dorothy A. Erie. Eukaryotic mismatch repair in relation to dna replication. *Annual Review of Genetics*, 49(1):291–313, 2015. PMID: 26436461.

[8] Paul Modrich. Mechanisms in e. coli and human mismatch repair (nobel lecture). *Angewandte Chemie International Edition*, 55(30):8490–8501, 2016.

[9] R M Schaaper and R L Dunn. Spontaneous mutation in the escherichia coli laci gene. *Genetics*, 129(2):317–326, 1991.

[10] Alexandra M de Paz, Thaddeus R Cybulski, Adam H Marblestone, Bradley M Zamft, George M Church, Edward S Boyden, Konrad P Kording, and Keith E J Tyo. High-resolution mapping of DNA polymerase fidelity using nucleotide imbalances and next-generation sequencing. *Nucleic Acids Research*, 46(13):e78–e78, 04 2018.

[11] Justin Jee, Aviram Rasouly, Ilya Shamovsky, Yonatan Akivis, Susan R. Steinman, Bud Mishra, and Evgeny Nudler. Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. *Nature*, 534:693–696, 2016.

[12] D K Bishop and R D Kolodner. Repair of heteroduplex plasmid dna after transformation into saccharomyces cerevisiae. *Molecular and Cellular Biology*, 6(10):3401–3409, 1986.

[13] Christiane Dohet, Robert Wagner, and Miroslav Radman. Repair of defined single base-pair mismatches in escherichia coli. *Proceedings of the National Academy of Sciences*, 82(2):503–505, 1985.

[14] Bisheng Zhou, Changjiang Huang, Junhua Yang, Jianxin Lu, Qiaoxiang Dong, and Lu-Zhe Sun. Preparation of heteroduplex enhanced green fluorescent protein plasmid for in vivo mismatch repair activity assay. *Analytical Biochemistry*, 388(1):167 – 169, 2009.

[15] D C Thomas, J D Roberts, and T A Kunkel. Heteroduplex repair in extracts of human hela cells. *Journal of Biological Chemistry*, 266(6):3744–51, 1991.

[16] James Brown, Tom Brown, and Keith R. Fox. Affinity of mismatch-binding protein muts for heteroduplexes containing different mismatches. *Biochemical Journal*, 354(3):627–633, 2001.

[17] S S Su and P Modrich. Escherichia coli muts-encoded protein binds to mismatched dna base pairs. *Proceedings of the National Academy of Sciences*, 83(14):5057–5061, 1986.

[18] Tassadite Selmane, Mark J Schofield, Sunil Nayak, Chunwei Du, and Peggy Hsieh. Formation of a dna mismatch repair complex mediated by atp. *Journal of Molecular Biology*, 334(5):949 – 965, 2003.

[19] Isaac Kinde, Jian Wu, Nick Papadopoulos, Kenneth W. Kinzler, and Bert Vogelstein. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences*, 108(23):9530–9535, 2011.

[20] Freeman Lan, Benjamin Demaree, Noorsher Ahmed, and Adam R. Abate. Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nature Biotechnology*, 35:640 EP –, May 2017.

[21] Assaf Rotem, Oren Ram, Noam Shoresh, Ralph A. Sperling, Michael Schnall-Levin, Huidan Zhang, Anindita Basu, Bradley E. Bernstein, and David A. Weitz. High-throughput single-cell labeling (hi-scl) for rna-seq using drop-based microfluidics. *PLOS ONE*, 10(5):1–14, 05 2015.

[22] Todd M. Gierahn, Marc H. Wadsworth II, Travis K. Hughes, Bryan D. Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J. Christopher Love, and Alex K. Shalek. Seq-well: portable, low-cost rna sequencing of single cells at high throughput. *Nature Methods*, 14:395 EP –, Feb 2017.

[23] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202 – 1214, 2015.

[24] Naomi Habib, Inbal Avraham-Davidi, Anindita Basu, Tyler Burks, Karthik Shekhar, Matan Hofree, Sourav R. Choudhury, François Aguet, Ellen Gelfand, Kristin Ardlie, David A. Weitz, Orit Rozenblatt-Rosen, Feng Zhang, and Aviv Regev. Massively parallel single-nucleus rna-seq with dronc-seq. *Nature Methods*, 14:955 EP –, Aug 2017.

[25] David F. Lee, Jenny Lu, Seungwoo Chang, Joseph J. Loparo, and Xiaoliang S. Xie. Mapping DNA polymerase errors by single-molecule sequencing. *Nucleic Acids Research*, 44(13):e118–e118, 05 2016.

[26] Jan Norrander, Tomas Kempe, and Joachim Messing. Construction of improved m13 vectors using oligodeoxynucleotide-directed mutagenesis. *Gene*, 26(1):101 – 106, 1983.

[27] Dan Y. Wu and R.Bruce Wallace. Specificity of the nick-closing activity of bacteriophage t4 dna ligase. *Gene*, 76(2):245 – 254, 1989.

[28] Vladimir Potapov, Jennifer L. Ong, Rebecca B. Kucera, Bradley W. Langhorst, Katharina Bilotti, John M. Pryor, Eric J. Cantor, Barry Canton, Thomas F. Knight, Thomas C. Evans, and Gregory J. S. Lohman. Comprehensive profiling of four base overhang ligation fidelity by t4 dna ligase and application to dna assembly. *ACS Synthetic Biology*, 7(11):2665–2674, 2018. PMID: 30335370.

[29] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406–3415, 07 2003.

[30] Harold Fredricksen and James Maiorana. Necklaces of beads in k colors and k-ary de bruijn sequences. *Discrete Mathematics*, 23(3):207 – 210, 1978.

[31] Jean-Pierre Duval. Génération d'une section des classes de conjugaison et arbre des mots de lyndon de longueur bornée. *Theoretical Computer Science*, 60(3):255 – 283, 1988.

[32] Martin Ester, Hans-Peter Kriegel, Joerg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *AAAI*, KDD-96(037):226–231, 1996.

[33] Rudolf Bayer. Symmetric binary b-trees: Data structure and maintenance algorithms. *Acta Informatica*, 1(4):290–306, Dec 1972.

[34] Akshay Tambe and Lior Pachter. Barcode identification for single cell genomics. *BMC Bioinformatics*, 20(1):32, Jan 2019.

[35] Tomoya Baba, Takeshi Ara, Miki Hasegawa, Yuki Takai, Yoshiko Okumura, Miki Baba, Kirill A Datsenko, Masaru Tomita, Barry L Wanner, and Hirotada Mori. Construction of escherichia coli k-12 in-frame, single-gene knockout mutants: the keio collection. *Molecular Systems Biology*, 2(1):2006.0008, 2006.

[36] James J. Truglio, Deborah L. Croteau, Bennett Van Houten, and Caroline Kisker. Prokaryotic nucleotide excision repair: the uvrabc system. *Chemical Reviews*, 106(2):233–252, 2006. PMID: 16464004.

[37] Aziz Sancar. Dna excision repair. *Annual Review of Biochemistry*, 65(1):43–81, 1996. PMID: 8811174.

[38] Maura Costello, Trevor J. Pugh, Timothy J. Fennell, Chip Stewart, Lee Lichtenstein, James C. Meldrim, Jennifer L. Fostel, Dennis C. Friedrich, Danielle Perrin, Danielle Dionne, Sharon Kim, Stacey B. Gabriel, Eric S. Lander, Sheila Fisher, and Gad Getz. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research*, 41(6):e67–e67, 01 2013.

[39] Anthony P. Breen and John A. Murphy. Reactions of oxyl radicals with dna. *Free Radical Biology and Medicine*, 18(6):1033 – 1077, 1995.

[40] Aidan J. Doherty, Stephen R. Ashford, Hosahalli S. Subramanya, and Dale B. Wigley. Bacteriophage t7 dna ligase: Overexpression, purification, crystallization, and characterization. *Journal of Biological Chemistry*, 271(19):11083–11089, 1996.

[41] J R Sayers and F Eckstein. Properties of overexpressed phage t5 d15 exonuclease. similarities with escherichia coli dna polymerase i 5'-3' exonuclease. *Journal of Biological Chemistry*, 265(30):18311–18317, 1990.

[42] Lovett S. The dna exonucleases of escherichia coli. *EcoSal Plus*, 2011.

[43] K. Pearson. *Drapers' Company Research Memoirs*, chapter XIII: On the theory of contingency and its relation to association and normal correlation, pages 1–35. Cambridge University Press, 1904.

[44] J. P. Guilford. The phi coefficient and chi square as indices of item validity. *Psychometrika*, 6(1):11–19, Feb 1941.

[45] Scott A. Lujan, Anders R. Clausen, Alan B. Clark, Heather K. MacAlpine, David M. MacAlpine, Ewa P. Malc, Piotr A. Mieczkowski, Adam B. Burkholder, David C. Fargo, Dmitry A. Gordenin, and Thomas A. Kunkel. Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. *Genome Research*, 24(11):1751–1764, 2014.

[46] G E Geier and P Modrich. Recognition sequence of the dam methylase of escherichia coli k12 and mode of cleavage of dpn i endonuclease. *Journal of Biological Chemistry*, 254(4):1408–13, 1979.

[47] Richard E. Gelinas, Phyllis A. Myers, and Richard J. Roberts. Two sequence-specific endonucleases from moraxella bovis. *Journal of Molecular Biology*, 114(1):169 – 179, 1977.

[48] Nicholas Renzette. Generation of transformation competent e. coli. *Current Protocols in Microbiology*, 22(1):A.3L.1–A.3L.5, 2011.

[49] Arthur Kerschen, Carolyn A. Napoli, Richard A. Jorgensen, and Andreas E. Müller. Effectiveness of rna interference in transgenic plants. *FEBS Letters*, 566(1-3):223–228, 2004.

[50] Patricia J. Pukkila, Janet Peterson, Gail Herman, Paul Modrich, and Matthew Meselson. Effects of high levels of dna adenine methylation on methyl-directed mismatch repair in escherichia coli. *Genetics*, 104(4):571–582, 1983.

[51] Ann Ganesan, Graciela Spivak, and Philip C. Hanawalt. Chapter 2 - transcription-coupled dna repair in prokaryotes. In Paul W. Doetsch, editor, *Mechanisms of DNA Repair*, volume 110 of *Progress in Molecular Biology and Translational Science*, pages 25 – 40. Academic Press, 2012.

[52] Ogun Adebali, Yi-Ying Chiou, Jinchuan Hu, Aziz Sancar, and Christopher P. Selby. Genome-wide transcription-coupled repair in escherichia coli is mediated by the mfd translocase. *Proceedings of the National Academy of Sciences*, 114(11):E2116–E2125, 2017.

[53] I Mellon and G N Champe. Products of dna mismatch repair genes muts and mutl are required for transcription-coupled nucleotide-excision repair of the lactose operon in escherichia coli. *Proceedings of the National Academy of Sciences*, 93(3):1292–1297, 1996.

[54] T T Nikiforov, R B Rendle, M L Kotewicz, and Y H Rogers. The use of phosphorothioate primers and exonuclease hydrolysis for the preparation of single-stranded pcr products and their detection by solid-phase hybridization. *Genome Research*, 3(5):285–291, 1994.

[55] Xi-Peng Liu and Jian-Hua Liu. The terminal 5' phosphate and proximate phosphorothioate promote ligation-independent cloning. *Protein Science*, 19(5):967–973, 2010.

[56] Yusuf E. Murgha, Jean-Marie Rouillard, and Erdogan Gulari. Methods for the preparation of large quantities of complex single-stranded oligonucleotide libraries. *PLOS ONE*, 9(4):1–10, 04 2014.

[57] M Carraway and M G Marinus. Repair of heteroduplex dna molecules with multibase loops in escherichia coli. *Journal of Bacteriology*, 175(13):3972–3980, 1993.

[58] Susan K. Amundsen, Andrew F. Taylor, Manjula Reddy, and Gerald R. Smith. Intersubunit signaling in recbcd enzyme, a complex protein machine regulated by chi hot spots. *Genes & development*, 21(24):3296–3307, Dec 2007. 18079176[pmid].

[59] Manuel Allhoff, Alexander Schönhuth, Marcel Martin, Ivan G. Costa, Sven Rahmann, and Tobias Marschall. Discovering motifs that induce sequencing errors. *BMC Bioinformatics*, 14(5):S1, Apr 2013.

[60] Michael Held and Richard M. Karp. A dynamic programming approach to sequencing problems. *Journal of the Society for Industrial and Applied Mathematics*, 10(1):196–210, 1962.

[61] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 10 2018.

[62] Scott A. Lujan, Jessica S. Williams, and Thomas A. Kunkel. Dna polymerases divide the labor of genome replication. *Trends in Cell Biology*, 26(9):640–654, Sep 2016.

[63] Celeste Yanisch-Perron, Jeffrey Vieira, and Joachim Messing. Improved m13 phage cloning vectors and host strains: nucleotide sequences of the m13mpl8 and puc19 vectors. *Gene*, 33(1):103 – 119, 1985.

[64] S B Haase, S S Heinzel, and M P Calos. Transcription inhibits the replication of autonomously replicating plasmids in human cells. *Molecular and Cellular Biology*, 14(4):2516–2524, 1994.

[65] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[66] David W. Russell and Norton D. Zinder. Hemimethylation prevents dna replication in e. coli. *Cell*, 50(7):1071–1079, Sep 1987.

[67] Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, Jun 2016.

[68] Scott A. Lujan, Jessica S. Williams, Zachary F. Pursell, Amy A. Abdulovic-Cui, Alan B. Clark, Stephanie A. Nick McElhinny, and Thomas A. Kunkel. Mismatch repair balances leading and lagging strand dna replication fidelity. *PLoS genetics*, 8(10):e1003016–e1003016, 2012. 23071460[pmid].

[69] Flora S. Groothuizen, Alexander Fish, Maxim V. Petoukhov, Annet Reumer, Laura Manelyte, Herrie H. K. Winterwerp, Martin G. Marinus, Joyce H. G. Lebbink, Dmitri I. Svergun, Peter

Friedhoff, and Titia K. Sixma. Using stable MutS dimers and tetramers to quantitatively analyze DNA mismatch recognition and sliding clamp formation. *Nucleic Acids Research*, 41(17):8166–8181, 07 2013.

[70] Anthony Mazurek, Christopher N. Johnson, Markus W. Germann, and Richard Fishel. Sequence context effect for hmsh2-hmsh6 mismatch-dependent activation. *Proceedings of the National Academy of Sciences*, 106(11):4177–4182, 2009.

[71] Oriol Pich, Ferran Muiños, Radhakrishnan Sabarinathan, Iker Reyes-Salazar, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. Somatic and germline mutation periodicity follow the orientation of the dna minor groove around nucleosomes. *Cell*, 175(4):1074 – 1087.e18, 2018.

[72] Kadir C. Akdemir, Victoria T. Le, Justin M. Kim, Sarah Killcoyne, Devin A. King, Ya-Ping Lin, Yanyan Tian, Akira Inoue, Samirkumar B. Amin, Frederick S. Robinson, Manjunath Nimmakayalu, Rafael E. Herrera, Erica J. Lynn, Kin Chan, Sahil Seth, Leszek J. Klimczak, Moritz Gerstung, Dmitry A. Gordenin, John O'Brien, Lei Li, Yonathan Lissanu Deribe, Roel G. Verhaak, Peter J. Campbell, Rebecca Fitzgerald, Ashby J. Morrison, Jesse R. Dixon, and P. Andrew Futreal. Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nature Genetics*, 52(11):1178–1188, Nov 2020.

[73] Monika Sharma, Alexander V. Predeus, Shayantani Mukherjee, and Michael Feig. Dna bending propensity in the presence of base mismatches: Implications for dna repair. *The Journal of Physical Chemistry B*, 117(20):6194–6205, 2013. PMID: 23621762.

[74] Pablo D Dans, Alexandra Balaceanu, Marco Pasi, Alessandro S Patelli, Daiva Petkevičiūtė, Jürgen Walther, Adam Hospital, Genís Bayarri, Richard Lavery, John H Maddocks, and Modesto Orozco. The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules. *Nucleic Acids Research*, 47(21):11090–11102, 10 2019.

[75] Stephanie Geggier and Alexander Vologodskii. Sequence dependence of dna bending rigidity. *Proceedings of the National Academy of Sciences*, 107(35):15421–15426, 2010.

[76] Reza Vafabakhsh and Taekjip Ha. Extreme bendability of dna less than 100 base pairs long revealed by single-molecule cyclization. *Science*, 337(6098):1097–1101, 2012.

[77] Aakash Basu, Dmitriy G. Bobrovnikov, Zan Qureshi, Tunc Kayikcioglu, Thuy T. M. Ngo, Anand Ranjan, Sebastian Eustermann, Basilio Cieza, Michael T. Morgan, Miroslav Hejna, H. Tomas Rube, Karl-Peter Hopfner, Cynthia Wolberger, Jun S. Song, and Taekjip Ha. Measuring dna mechanics on the genome scale. *Nature*, Dec 2020.

[78] Giulia Rossetti, Pablo D. Dans, Irene Gomez-Pinto, Ivan Ivani, Carlos Gonzalez, and Modesto Orozco. The structural impact of DNA mismatches. *Nucleic Acids Research*, 43(8):4309–4321, 03 2015.

[79] Flora S. Groothuizen and Titia K. Sixma. The conserved molecular machinery in dna mismatch repair enzyme structures. *DNA Repair*, 38:14 – 23, 2016. DNA mismatch Repair.

[80] Daiguan Yu, Hilary M. Ellis, E-Chiang Lee, Nancy A. Jenkins, Neal G. Copeland, and Donald L. Court. An efficient recombination system for chromosome engineering in escherichia coli. *Proceedings of the National Academy of Sciences*, 97(11):5978–5983, 2000.

[81] Jessica S. Williams, Anders R. Clausen, Stephanie A. [Nick McElhinny], Brian E. Watts, Erik Johansson, and Thomas A. Kunkel. Proofreading of ribonucleotides inserted into dna by yeast dna polymerase $\epsilon$. *DNA Repair*, 11(8):649 – 656, 2012.

[82] Indranil Biswas and Peggy Hsieh. Identification and characterization of a thermostable muts homolog from thermus aquaticus. *Journal of Biological Chemistry*, 271(9):5040–5048, 1996.

[83] Ankur Jain, Ruijie Liu, Yang K. Xiang, and Taekjip Ha. Single-molecule pull-down for studying protein interactions. *Nature Protocols*, 7(3):445–452, Mar 2012.

[84] T Ha. Single-molecule fluorescence resonance energy transfer. *Methods (San Diego, Calif.)*, 25(1):78—86, September 2001.

[85] Rahul Roy, Sungchul Hohng, and Taekjip Ha. A practical guide to single-molecule fret. *Nature Methods*, 5(6):507–516, Jun 2008.

[86] P.V.C. Hough. Machine Analysis of Bubble Chamber Pictures. *Conf. Proc. C*, 590914:554–558, 1959.

[87] Richard O. Duda and Peter E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM*, 15(1):11–15, January 1972.

[88] Linda G. Shapiro. Connected component labeling and adjacency graph construction. In T. Yung Kong and Azriel Rosenfeld, editors, *Topological Algorithms for Digital Image Processing*, volume 19 of *Machine Intelligence and Pattern Recognition*, pages 1 – 30. North-Holland, 1996.

[89] Steven Gold, Anand Rangarajan, Chien-Ping Lu, Suguna Pappu, and Eric Mjolsness. New algorithms for 2d and 3d point matching: pose estimation and correspondence. *Pattern Recognition*, 31(8):1019 – 1031, 1998.

[90] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, Feb 1992.

[91] Russell E. Thompson, Daniel R. Larson, and Watt W. Webb. Precise nanometer localization analysis for individual fluorescent probes. *Biophysical Journal*, 82(5):2775 – 2783, 2002.

## 5.4 Curriculum vitae

### 5.4.1 Education

- Johns Hopkins University, Baltimore, MD, USA 2016 - 2021
  PhD, Biophysical Chemistry
- University of Illinois, Urbana, IL, USA 2014 - 2015
  MSc, Biophysics & quantitative biology
- Bogazici University, Istanbul, Turkey 2009 - 2014
  BSc, Molecular biology and genetics
  BSc, Physics
- Ankara Ataturk Anadolu Lisesi, Ankara, Turkey 2005 - 2009
  Mathematics and natural sciences

### 5.4.2 Research experience

- Research assistant, Johns Hopkins University, Baltimore, MD, USA 01/2016 - 02/2021
  High-throughput assay development to detect error-prone DNA sequence motifs
- Intern scientist, Revolve Biotechnologies Inc., Baltimore, MD, USA 08/2018 - 11/2018
  Custom-design antibody engineering for *in vitro* detection kits
- Research assistant, University of Illinois, Urbana, IL, USA 08/2014 - 01/2016
  Characterization of DNA flexibility via single molecule fluorescence microscopy
- Research assistant, Bogazici University, Istanbul, Turkey 06/2013 - 08/2014
  Photo-acoustic microscope development
  Computational modeling of tumor vasculature
- Research intern, Centre of Structural Biology, Aarhus, Denmark 01/2012 - 07/2012
  Synthesis and purification of engineered proteins for x-ray crystallography

### 5.4.3 Scientific publications & presentations

- **Tunc Kayikcioglu**, Chang-Ting Lin, Kasper Hansen, Taekjip Ha "High Throughput Measurement of DNA Repair Efficiency in vivo", in preparation.

- Aakash Basu, Dmitriy G. Bobrovnikov, Zan Qureshi, **Tunc Kayikcioglu**, Thuy T. M. Ngo, Anand Rajan, Sebastian Eustermann, Basilio Cieza, Mike Morgan, Miroslav Hejna, H. Tomas Rube, Karl-Peter Hopfner, Cynthia Wolberger, Jun Song and Taekjip Ha (2020) "Measuring DNA mechanics on the genome scale", Nature 10.1038/s41586-020-03052-3.

- Anustup Poddar, Muhammad S. Azam, **Tunc Kayikcioglu**, Maksym Bobrovskyy, Jichuan Zhang, Xiangqian Ma, Piyush Labhsetwar, Jingyi Fei, Digvijay Singh, Zaida Luthey-Schulten, Carin K. Vanderpool, Taekjip Ha (2021) "Effects of individual base-pairs on in vivo target search and destruction kinetics of small RNA", Nature Communications 10.1038/s41467-021-21144-0.

- Esra Aytac-Kipergil, Aytac Demirkiran, Nasire Uluc, Seydi Yavas, **Tunc Kayikcioglu**, Sarper Salman, Sohret Gorkem Karamuk, Fatih Ilday and Mehmet Unlu (2016) "Development of a Fiber Laser with Independently Adjustable Properties for Optical Resolution Photoacoustic Microscope" Scientific Reports 6:38674.

- **Tunc Kayikcioglu**, Chang-Ting Lin, Taekjip Ha (2018) "Massively parallel measurement of DNA mismatch repair efficiency in vivo", poster at Biophysical Society Annual Meeting, San Francisco, USA.

- **Tunc Kayikcioglu**, Chang-Ting Lin, Taekjip Ha (2017) "Massively parallel measurement of DNA mismatch repair efficiency", talk at CPLC Symposium, Urbana, USA.

- **Tunc Kayikcioglu**, Thuy Ngo, Aakash Basu, Sangwoo Park, Qiucen Zhang, Taekjip Ha (2016) "Single Molecule Toolbox: Measuring DNA Flexibility" Physics of Living Systems Annual Meeting, Boston, USA.

- **Tunc Kayikcioglu**, Taekjip Ha (2016) "DNA Flexibility Does Not Show Appreciable Tem-

perature Dependence", poster presentation at Biophysical Society Annual Meeting, Los Angeles, USA.