PLOS ONE

RESEARCH ARTICLE

Highly mismatch-tolerant homology testing by RecA could explain how homology length affects recombination

Mara Prentiss[1]*, Dianzhuo Wang[1], Jonathan Fu[1], Chantal Prévost[2], Veronica Godoy-Carter[3], Nancy Kleckner[4], Claudia Danilowicz[1]

1 Department of Physics, Harvard University, Cambridge, Massachusetts, United States of America, **2** Laboratoire de Biochimie Théorique, Institut de Biologie Physico-Chimique, Paris, France, **3** Department of Biology, Northeastern University, Boston, Massachusetts, United States of America, **4** Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, United States of America

* prentiss@fas.harvard.edu

Citation: Prentiss M, Wang D, Fu J, Prévost C, Godoy-Carter V, Kleckner N, et al. (2023) Highly mismatch-tolerant homology testing by RecA could explain how homology length affects recombination. PLoS ONE 18(7): e0288611. https://doi.org/10.1371/journal.pone.0288611

Editor: Arthur J. Lustig, Tulane University Health Sciences Center, UNITED STATES

Received: March 16, 2023

Accepted: July 2, 2023

Published: July 13, 2023

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: MP: Chu Family Foundation and Harvard University; No CP: 'Initiative d'Excellence' program of the French State [DYNAMO, ANR-11-LABX0011]; No VGC: Northeastern University Skills for Capacity and Inclusion Program by the Howard Hughes Medical Institute through the Science Education Program; No NK: National Institute of Health, R35-GM-136322; No The funders had no

Abstract

In *E. coli*, double strand breaks (DSBs) are resected and loaded with RecA protein. The genome is then rapidly searched for a sequence that is homologous to the DNA flanking the DSB. Mismatches in homologous partners are rare, suggesting that RecA should rapidly reject mismatched recombination products; however, this is not the case. Decades of work have shown that long lasting recombination products can include many mismatches. In this work, we show that *in vitro* RecA forms readily observable recombination products when 16% of the bases in the product are mismatched. We also consider various theoretical models of mismatch-tolerant homology testing. The models test homology by comparing the sequences of L_{test} bases in two single-stranded DNAs (ssDNA) from the same genome. If the two sequences pass the homology test, the pairing between the two ssDNA becomes permanent. Stringency is the fraction of permanent pairings that join ssDNA from the same positions in the genome. We applied the models to both randomly generated genomes and bacterial genomes. For both randomly generated genomes and bacterial genomes, the models show that if no mismatches are accepted stringency is $\sim$ 99% when L_{test} = 14 bp. For randomly generated genomes, stringency decreases with increasing mismatch tolerance, and stringency improves with increasing L_{test}. In contrast, in bacterial genomes when $L_{test} \sim$ 75 bp, stringency is $\sim$ 99% for both mismatch-intolerant and mismatch-tolerant homology testing. Furthermore, increasing L_{test} does not improve stringency because most incorrect pairings join different copies of repeats. In sum, for bacterial genomes highly mismatch tolerant homology testing of 75 bp provides the same stringency as homology testing that rejects all mismatches and testing more than $\sim$75 base pairs is not useful. Interestingly, *in vivo* commitment to recombination typically requires homology testing of $\sim$ 75 bp, consistent with highly mismatch intolerant testing.

role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

RecA-family proteins occur in all domains of life [1]. RecA-family proteins promote repair of double strand breaks (DSB) in genomes [1–3]. Incorrect repair of double-strand breaks can cause cancer [4] and birth defects [4]. Thus, genome integrity should be maintained; however, genomic rearrangements [5, 6] can be highly advantageous. An example is bacterial acquisition of antibiotic resistances [7]. As a result, it is critical to understand how RecA-family proteins repair double-strand breaks and how that repair affects the balance between genome integrity and introducing rare and perhaps highly advantageous genomic rearrangements.

After a double strand break, resection of the broken double-stranded DNA (dsDNA) creates 3′-ssDNA tails [2, 8]. Those ssDNA tails are often referred to as invading strands. After an invading strand is covered by RecA protein, each ssDNA-RecA filament probes an intact copy of the broken chromosome. The sequence of the unbroken chromosome is probed by attempting to establish Watson-Crick pairing between the invading strand and one of the strands in the target thus forming a heteroduplex product and leaving the other strand in the target unpaired [2, 8]. The resulting structure is called a D-loop.

In vivo studies probed the homology dependence of recombination using exogenous sequences containing L contiguous base pairs that were homologous to the target DNA [9–11]. The homologous bases were flanked by extensive heterology. Those *in vivo* studies found that when $L < 20$ bp, recombination was not detectable. A steep exponential increase in recombination with length was observed for $20 \lesssim L \lesssim 75$ bp [9, 10], and a slow linear increase in recombination was observed for $L \gtrsim 75$ bp [10]. Thus, *in vivo* studies suggest that the behavior of recombination products as a function of L divides into three regions. The boundaries between the regimes occur at $\sim$ 20 bp and $\sim$ 75 bp. Furthermore, the *in vivo* studies suggest that testing more than 75 bp may not be very useful.

In vivo experiments have also considered how periodically spaced mismatches in otherwise homologous invading strands influence homologous recombination in yeast [12]. That work found that when the invading strand included 1 mismatch/6 bases repair was $\sim$ 5% efficient, whereas more frequent mismatch spacings did not produce observable repair. Furthermore, *in vitro* studies have found that RecA mediated homologous recombination is very tolerant of mismatches [13–17].

Though the detailed mechanisms underlying RecA mediated homologous recombination are not fully understood, it is believed that homology testing by RecA starts with the initial 8 bp homology test [16, 18–21] that accepts one mismatch [18]. That initial test is followed by homology testing that groups base pairs into triplets. Each triplet is tested for homology separately. For each triplet, the outcome of the homology test depends on the number of mismatches within the triplet [22]. Such a division of homology testing into tests of separate base pair triplets is consistent with the known structure of heteroduplex products. The crystal structure indicates that heteroduplex products are divided into nearly B-form base pair triplets that are separated by large rises [23]. The triplets are stabilized by protein residues that intercalate in the rises [18, 23].

It has long been known that RecA mediated homologous recombination can occur without ATP hydrolysis [24]. Some features of homology recognition are independent of hydrolysis. For example, the initial 8 bp test occurs both with [16] and without hydrolysis [18–21]; however, it is likely that some RecA features depend on hydrolysis. Importantly, without ATP hydrolysis, heteroduplex products longer than $\sim$ 20 bp are effectively irreversible [24], whereas with ATP hydrolysis 20 bp products remain highly reversible *in vitro* [25] and *in vivo* [9–11]. Thus, the reversal of heteroduplex products longer than 20 bp is clearly strongly influenced by ATP hydrolysis.

In this work we probe RecA strand exchange *in vitro* with ATP hydrolysis. Consistent with previous work in yeast [12], we show that if mismatches between the invading and complementary strands are distributed periodically, 1 mismatch/3 bases blocks formation of stable strand exchange products, whereas 1 mismatch/6 bases forms readily observable heteroduplex products. Given that homology-testing *in vitro* suggests that RecA family protein mediated strand exchange is very tolerant of mismatches, we use simple theoretical models to probe highly mismatch-tolerant homology testing. In all the theoretical models, two sequences containing L_{test} base pairs are tested for homology. If the two sequences pass the homology test, the pairing between the two sequences becomes stable. That stable pairing creates a heteroduplex product. We then consider the fraction of those heteroduplex products that correctly pair corresponding sequence regions in two chromosomes. When we apply homology testing to a "genome" consisting of randomly chosen bases, the fraction of incorrect pairings that passes the homology test always decreases as L_{test} increases. In contrast, applying the same homology tests to bacterial genomes indicates that when $L_{test} > 75$ bp, the fraction of incorrect pairings that passes the homology test does not decrease with increasing L_{test} because most incorrect products join different copies of sequences that are exactly repeated in the genome. Importantly, even if our models do not accurately reflect all features of RecA strand exchange *in vivo*, insights from robust features of the modeling results may greatly enhance our understanding of RecA mediated homology recognition.

Materials and methods

Sample preparation

The 180 bp rhodamine and fluorescein labelled dsDNA was prepared by initially annealing an internal rhodamine 90-nt ssDNA (O1) and a 5′end phosphorylated oligonucleotide (82 bases) (O2). Similarly, an 82-nt 5′-end phosphorylated oligonucleotide (O3) was annealed with a 98-nt oligonucleotide containing an internal fluorescein label (O4) to obtain another labelled dsDNA fragment. These two dsDNA labelled fragments were annealed and ligated overnight at 16˚C with T4 DNA ligase in ligase reaction buffer (50 mM Tris, 10 mM $MgCl_2$, 1 mM ATP, and 10 mM dithiothreitol, pH 7.5 (New England Biolabs (NEB)). The 180 bp construct was purified3 by gel on 3% agarose in TBE (Tris/Borate/EDTA) buffer for 2 hours (6 V/cm). The 180 bp band was visualized with a midrange UV trans-illuminator. Finally, the 180 bp dsDNA band was cut out and was extracted from agarose using a Nucleospin kit (Machery and Nagel, Bethlehem, PA). The product was finally concentrated using 100 kDa Amicon filters (Millipore). The oligonucleotide sequences (Integrated DNA Technologies (Coralville, IA)) used to prepare the dsDNA construct are listed in S1 Table.

Strand exchange reactions

DNA strand exchange *in vitro* was achieved by mixing 0.06 μM ssDNA-RecA filaments with 0.06 μM labeled dsDNA. The sequences of the oligonucleotides used for these filaments are listed in S1 Table. The ssDNA-RecA filaments were prepared by mixing the ssDNA (final concentration 6 μM in bases) with 2 μM RecA (NEB) at 3:1 ratio (bases: protein), 1 mM ATP, a regeneration system containing 10 U/ml of pyruvate kinase and 3 mM phosphoenolpyruvate, and 0.2 μM *E.coli* single-stranded binding protein, SSB (Abcam) in RecA buffer (70 mM Tris-HCl, 10 mM $MgCl_2$, and 5 mM dithiothreitol, pH 7.6). Filament formation proceeded at 37˚C for 10 minutes.

Fluorescence measurements

To measure the fluorescent signal of these strand exchange reactions, the ssDNA/RecA filament and dsDNA mixture was placed in a quartz cuvette. Initially, fluorescein emission was quenched by rhodamine because the outgoing and complementary strands were paired in the dsDNA. As strand exchange progresses, the outgoing and complementary strands separate and fluorescein emission increases. Using FluorEssence spectroscopy software and the automated FluoroMax spectrofluorometer (Horiba, Edison, NJ), the emission of the fluorescein label was read using 493 nm excitation wavelength and a 2 nm slit. The emission of the fluorescein fluorophore in counts per second (cps) was detected at 518 nm with a 2 nm slit. The reaction was run for 30 minutes, and emission measurements were collected every 1 second with an integration time of 0.5 s. Temperature was kept constant at 37˚C.

Probabilities of incorrect pairings for a homology search following a DSB in a random genome: Analytical approach

For a random genome, the probability that one randomly chosen base will match another randomly chosen base is p = ¼. The probability of finding m or more correctly paired bases in a randomly chosen sample of n base pairs is then:

$$P(m,\ n)\ =\ \sum_{k=m}^{n}\binom{n}{k}p^{k}(1-p)^{n-k} \tag{1}$$

For the 8 bp test, n = 8 and for the triplet tests n = 3. In either case m is the difference between n and the number of mismatches.

Thus, for an 8 bp test with one mismatch n = 8 and m = 7, and for a triplet test that accepts two mismatches n = 3 and m = 1.

The probability of passing a series of tests is just the product of the probabilities for all of the tests. Thus, for a given L_{test}, the probability of passing an 8 bp test followed by a series of triplet tests is the product of the probability of passing the first 8 bp test times the probability of passing (L_{test}-8)/3 triplet tests, where $N_{mismatch}$ and $N_{mismatch8}$ are the number of mismatches accepted per triplet and 8 bp test, respectively.

$$\mathrm{Pass}\left(L_{test}, N_{mismatch,},\ N_{mismatch8}\right)\ =\ P(8-N_{mismatch8}, 8)\ P(3-N_{mismatch},\ 3)^{(Ltest-8)/3} \tag{2}$$

The formula implies that the results for a series of triplet tests that accept only one mismatch per triplet are much more stringent than the results of a test that accepts 1/3 mismatches in L_{test}-8 bases.

For a genome that includes L_{genome} randomly chosen bases, and a randomly chosen sequence of length L_{test}, in the genome there are on average Pass(L_{test}, $N_{mismatch}$, $N_{mismatch8}$) L_{genome} matching sequences of length L_{test}.

In DSB repair for a random genome, we assume that the sequence of the searcher is always included in the target; therefore, there is also always one correct pairing between the searcher and the target, regardless of the probability that a randomly chosen searching sequence would find a match. Thus, for a random genome during the homology search that follows a DSB, the

average number of pairings that would pass the RecA homology test is:

$$\begin{aligned} &\mathrm{Number}_{\mathrm{pass}}\left(\mathrm{N}_{\mathrm{mismatch}},\ \mathrm{N}_{\mathrm{mismatch}},\ \mathrm{L}_{\mathrm{test}},\ \mathrm{L}_{\mathrm{genome}}\right) \\ &= 1\ +\ \mathrm{Pass}(\mathrm{L}_{\mathrm{teste}},\ \mathrm{N}_{\mathrm{mismatch}}\ ,\ \mathrm{N}_{\mathrm{mismatch8}})\ \mathrm{L}_{\mathrm{genome}} \end{aligned} \tag{3}$$

Of those that pass, probability of being correct is then 1 out of the total number or

$$\begin{aligned} &\mathrm{Prob}_{\mathrm{pass}}\left(\mathrm{N}_{\mathrm{mismatck,}}\ \mathrm{N}_{\mathrm{mismatch8,}}\ \mathrm{L}_{\mathrm{test}},\ \mathrm{L}_{\mathrm{genome}}\right) \\ &=\ 1/\left(1\ +\ \mathrm{Pass}(\mathrm{L}_{\mathrm{test}},\ \mathrm{N}_{\mathrm{mismatch}},\ \mathrm{N}_{\mathrm{mismatch8}})\mathrm{L}_{\mathrm{genome}}\right) \end{aligned} \tag{4}$$

The probability that a pairing passes and is incorrect is then:

$$\begin{aligned} &\mathrm{Prob}_{\mathrm{incorrect}}\left(\mathrm{N}_{\mathrm{mismatch}},\ \mathrm{N}_{\mathrm{mismatch8,}}\ \mathrm{L}_{\mathrm{test}},\ \mathrm{L}_{\mathrm{genome}}\right) \\ &=\ 1/\left(1\ +\ \mathrm{Pass}(\mathrm{L}_{\mathrm{test}},\ \mathrm{N}_{\mathrm{mismatch}},\ \mathrm{N}_{\mathrm{mismatch8}})\mathrm{L}_{\mathrm{genome}}\right) \end{aligned} \tag{5}$$

We note that all predictions of the analytical treatments are independent of the following: 1. The directionality of strand exchange; 2. whether strand exchange is unidirectional or bidirectional; 3. whether the bases are tested iteratively or all at once.

Probabilities of incorrect pairings for a bacterial genome using the simplified model that is applied sparsely to the sequences of bacterial genomes or a random genome: Simulation approach

A position in the given strand of genome was randomly chosen as the position of the 5′ end of the invading strand. A second position in the given strand of the genome was randomly chosen to represent the testing position in an unbroken chromosome that pairs with the 5′ end of the invading strand. The sequences of the 8 bp starting at the 5′ end and extending in the 3′ direction were then compared. If the number of mismatches was $> \mathrm{N}_{\mathrm{mismatch8}}$, homology testing at that position terminated and a new testing position was chosen. If the number of mismatches was $\leq \mathrm{N}_{\mathrm{mismatch8}}$, then the next 3 bases on the 3′ side were tested for homology. Homology testing terminates, and a new testing position is chosen whenever the number of mismatches in a triplet $> \mathrm{N}_{\mathrm{mismatch}}$. If the length that has passed the homology tests reaches $\mathrm{L}_{\mathrm{test}}$, then the pairing is considered irreversible. If the test position and the search position are the same, the irreversible pairing is correct. If the test position and search position are different but there are no mismatches, then the irreversible pairing joins two copies of a repeat with length $\geq \mathrm{L}_{\mathrm{test}}$. If the irreversible pairing contains mismatches, the pairing is an error. Typically, 100–100000 different invading strand sequences were chosen, and the results represent the averages of the results for the chosen invading strand sequences out of 4.6 Mbp possible invading strand sequences.

The computer simulations only included $\sim$ 100 to 10000 simulated DSB positions, so the genome was sparsely sampled; however, we increased the number of simulated breaks until results were insensitive to the number of simulated breaks. In addition, when no mismatches are accepted it is possible to sample the entire genome, and results for sampling the entire genome are similar to results for sparse sampling. We note that results for sampling the entire genome and rejecting all mismatches are independent of the directionality of strand exchange and are not affected by whether strand exchange is unidirectional or bidirectional. Finally, we also applied the computer simulation to random genomes, and the results of the computer

simulations are in good agreement with the analytical results for sequences whose bases are chosen at random (random genomes).

We note that the results of these simulations are independent of whether the testing occurs simultaneously or iteratively. The model does assume that the triplet tests occur on the 3′ side of the 8 bp test, but results are insensitive to the positioning of the 8 bp test with respect to the triplet tests. Thus, bidirectional homology testing or 3′ to 5′ homology testing would give very similar results. Finally, exact homology testing that rejects all mismatches is independent of strand exchange directionality, and exact homology testing predicts the same stringency vs. L_{test} as homology testing in which the triplets are positioned to the 3′ side of the initial 8 bp test.

Exact probabilities of incorrect pairings when $N_{mismatch} = 0$ for entire bacterial genomes

Each possible 8 bp sequence was assigned a unique mapping number. The bacterial genome was divided into 8 bp sequences, each of which was assigned the corresponding mapping number. The 8 bp sequences were then sorted according to their mapping number, which grouped the entire genome into distinct 8 bp repeats. The total number of incorrect pairing locations for each member of the group is equal to the number of locations in the group -1. We repeated the same procedure for 11 bp sequences. To calculate exact repeats for 14, 17, 22, 33, 44, 55, and 99 bp we created a list of maps encoding a series of sequences with lengths $\leq$ 11 bp. For example, for $L_{test} = 14$ we used an 11 bp map and a 3 bp map and then grouped starting locations with the same values for both maps. For the *E.coli* MG1655 genome most 99 bp repeats occurred only twice, but some repeats occurred more than twice. The most frequent 99 bp repeat occurred 9 times. We obtained histograms of the long repeats by sorting the 99 bp sequences according to the starting position and counting the number of sequential starting positions.

Given the number of unique repeats with length L_{test} and the frequency of each repeat ($frequency_{repeat}$), we calculated the probability that a DSB in a bacterial genome would lead to an incorrect final pairing. For each repeat, the number of incorrect targets in the genome is given by ($frequency_{repeat}$ -1), whereas the number of correct targets is 1. Thus, the probability that a sequence matched pairing of the repeat is incorrect is ($frequency_{repeat}$ -1)/ ($frequency_{repeat}$). The probability that a DSB creates an invading strand terminating in this repeat is ($frequency_{repeat}$)/(genome length). Thus, the total probability that an invading strand consisting of the particular repeat leads to an incorrect pairing is ($frequency_{repeat}$ -1)/(genome length). Summing over the result for all unique repeats gives the total probability that a DSB will lead to a sequence matched incorrect pairing with length L_{test}.

We note that all predictions of the exact probabilities for bacterial genomes are independent of the following: 1. The directionality of strand exchange. 2. Whether strand exchange is unidirectional or bidirectional. 3. Whether the bases are tested iteratively or all at once.

Calculation of the number of distinct long repeat pairs in bacterial genome

To determine the prevalence of longer repeats, each possible 12 bp sequence was assigned a unique mapping number. The bacterial genome was divided into 12 bp sequences, each of which was assigned the corresponding mapping number. The 12 bp sequences were then sorted according to their mapping number, which grouped the entire genome into distinct 12 bp repeats. Longer repeats were probed by extending the 12 bp sequences within each mapping group and counting only those pairs that had no mismatches over the length L_{test}. If the number is non-zero, that implies that there must be some repeats that have lengths $\geq L_{test}$. If most

repeats only occur twice, the ratio of the number of distinct pairs to the length of the genome gives the probability that a DSB will lead to an incorrect pairing of that length.

Probabilities of incorrect pairings for a bacterial genome using the simplified model with Chi sites that sparsely samples bacterial genomes

A DSB position in the genome was randomly chosen. The nearest Chi site on the 5′ side of the DSB was found. The 3′ end of the invading strand was positioned at the 3′ end of that Chi site. The homology test considers the L_{test} bases on the 5′ side of the Chi site.

Simplified probabilistic model of homology testing in triplets

A random number generator creates a number between 0 and 1. $Tripletpass_M$ describes the probability that a triplet with M mismatches will pass the triplet homology test. A triplet passes the homology test if $Tripletpass_M$ is greater than the random number. We first considered homology tests with $Tripletpass_M = 1$ when M = 0, which means that the triplet is completely sequence matched. We considered two fundamentally different homology testing models. In one model, $Tripletpass_M$ is the same for all $M > 0$. This represents the case in which collective interactions within the triplet allow a single mismatch to destabilize the triplet. For this model in which all triplets containing mismatches have the same probability of passing. We considered two different passing probabilities: $Tripletpass_M = 0.25$ and $Tripletpass_M = 0.5$.

We also considered models in which each mismatch decreases the probability that a triplet will pass the homology test. In that model, a random number generator creates a number between 0 and 4. We ran a model in which a homology test of a triplet passes the triplet if the number of mismatches is less than or equal to the value determined by the random number generator. Since 0 mismatches is less than or equal to all those values, a triplet with 0 mismatches has a 100% chance of passing the homology test. Similarly, 1 mismatch, 2 mismatches, or 3 mismatches have a 75%, 50%, or 25% chance of passing the homology test, respectively. We ran a second more promiscuous homology test in which 1 mismatch, 2 mismatches, or 3 mismatches have a 75%, 50%, or 50% chance of passing the homology test. Thus, the second test is the same as the first except for the probability that a completely mismatched triplet will pass the test.

Results

In vitro experiments probing RecA strand exchange in the presence of several periodically spaced single mismatches

Previous work has suggested that with ATP hydrolysis strand exchange can accept a single mismatch [17]. To gain some insight into the mismatch tolerance of triplet-based homology testing with ATP hydrolysis, we considered invading strands with periodically spaced single mismatches. How the invading strand is divided into triplet homology tests depends on the position at which strand exchange starts; however, periodically spaced single mismatches with 1 mismatch per 3 bases will always have exactly 1 mismatch/triplet. Similarly, periodically spaced single mismatches with 1 mismatch per 6 bases will have exactly 1 mismatch in every other triplet. The mismatch-containing triplets will be separated by single sequence-matched triplets.

We observed strand exchange using FRET due to fluorophores positioned in the homoduplex (Fig 1A). When the homoduplex is base paired, emission of the fluorescein is quenched by the nearby rhodamine fluorophore. When base pairing is disrupted and both fluorophores are separated, fluorescein emission increases. In particular, strand exchange is probed by measuring the interaction between unlabeled 98-nt ssDNA/RecA filaments and labelled 180 bp

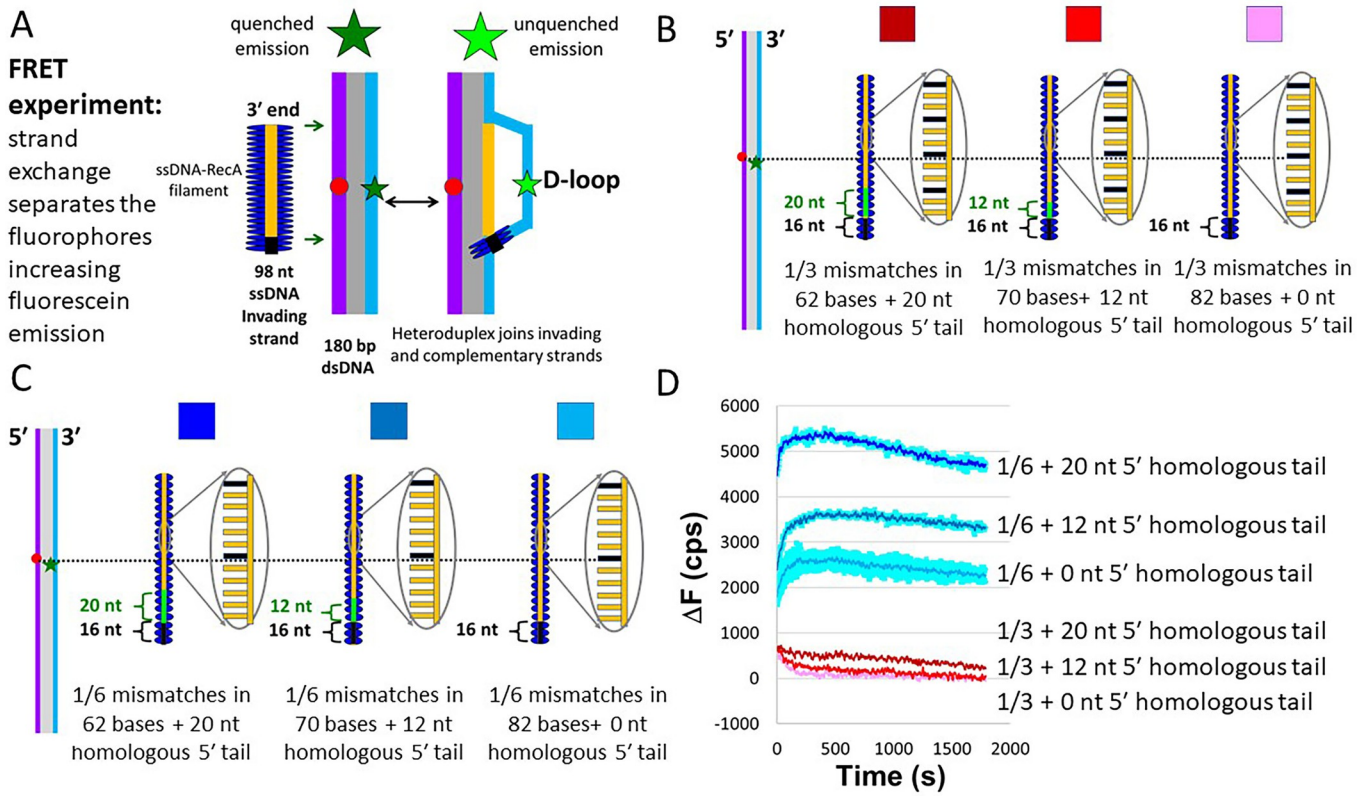

Fig 1. Strand exchange across single mismatches monitored using FRET. (A). Schematics of interactions of 180 bp dsDNA with 98-nt filaments containing different degrees of homology. The purple and light blue lines represent the complementary and outgoing strands, respectively. The gray rectangle indicates base paired regions. The dark blue ovals represent RecA. The red circle (rhodamine) and green star (fluorescein) show the locations of the fluorophores along the 180 bp dsDNA. All the invading strands contain 98 nt. The 16 heterologous bases nearest the 5′-end of the invading strand are shown in black. The green brackets indicate the region of the dsDNA that is partially homologous to the invading strand. The orange regions of the invading strand are partially homologous to the dsDNA. **(B).** Schematics of the experiments with mismatches periodically spaced 1 mismatch/3 bases. The orange regions of the invading strand include the periodically spaced mismatches, whereas the black and green regions are completely heterologous or completely homologous, respectively. The completely heterologous and homologous regions are also indicated by the black and green brackets. **(C).** Same as B but for invading strands with 1 mismatch every 6 bp. **(D).** Change in fluorescein emission (△F) vs. time curves. The dark red, bright red, and pink curves indicate results for 1/3 periodically spaced mismatches with a 20, 12, or 0 nt homologous tail on the 5′-side. The dark blue, medium blue, and light blue curves indicate results for 1/6 periodically spaced mismatches with a 20, 12, or 0 nt homologous tail on the 5′-side. Error bars represent root mean square deviation from two independent runs.

https://doi.org/10.1371/journal.pone.0288611.g001

dsDNA. The same dsDNA was used in all the experiments. The mismatches between the invading and complementary strands were controlled by the sequence of the invading strand (S1 Table).

The dsDNA had fluorophores positioned 90 bp (fluorescein, green star) and 89 bp (rhodamine, red circle) (Fig 1A) from the end of the dsDNA that is homologous to the 3′ end of the invading strand. The unlabeled ssDNA used for the fully homologous filament contained a 16-nt heterologous tail in order to have a total length of ∼ 100 bases (Fig 1A). The ∼ 100 nt length is long enough for ssDNA-RecA filaments to be stable with ATP hydrolysis. The ssDNA-RecA filament was mixed with the 180 bp dsDNA, and the resulting emission of the fluorescein label in the dsDNA was measured over time. To monitor strand exchange, we measure △F, the change in fluorescein emission, as a function of time.

In one set of experiments, the periodic mismatches were distributed over the entire invading strand (Fig 1B and 1C); therefore, strand exchange products with 1 mismatch/3 bases would be rejected by the initial 8 bp test [18], regardless of the tolerance of the triplet tests. To gain insight into the triplet tests we performed additional experiments in which the invading

strands included homologous regions at the 5′ end that can pass the 8 bp test, as well as regions with periodically spaced mismatches that would be subject to triplet testing if the initial 8 bp test is passed (Fig 1B and 1C).

We considered interactions with 1 mismatch/3 bases (Fig 1B) and 1 mismatch/6 bases (Fig 1C). In the absence of a homologous 5′ tail, 1 mismatch/3 bases would always be rejected by the initial 8 bp test, and indeed we do not observe any significant fluorescence in that case (pink curve, Fig 1D). When a 12 or 20 nt homologous tail is present, 8 bp tests within the homologous tail can be passed. After that test is passed, strand exchange can progress through the remainder of the invading strand. Having a 12 nt homologous tail is insufficient to produce an increase in emission (Fig 1D, bright red curve) whereas a 20 nt tail might result in a very small increase in emission, but the increase is not statistically significant (dark red curve, Fig 1D). Thus, the results of invading strands with 1 mismatch/3 bases suggest that the triplet testing does not always pass triplets containing one mismatch.

Results of experiments with 1 mismatch/6 bases (Fig 1C) show that invading strands with 1 mismatch/6 bases can pass the initial 8 bp test (Fig 1D, blue curves). Thus, even without a homologous tail, invading strands with 1 mismatch/6 bases can provide information on the probability that a triplet with a single mismatch will be incorporated in a strand exchange product. Consistent with the results in yeast [12], interactions with 1 mismatch/6 bases show an increase in fluorescence due to formation of strand exchange products. The emission for 1 mismatch/6 bases was $\sim$ 25% of the emission for a perfectly matched invading strand. Thus, like Rad51 [12], RecA sometimes yields heteroduplex products from invading strands containing 1/6 = 16% mismatches. Unsurprisingly, the emission increases as the length of the homologous region at the 5′ end increases.

In these bulk experiments, we cannot determine the probability that a triplet containing a single mismatch will pass a homology test and be incorporated in a heteroduplex product;

Table 1. List of terms used in the homology testing models.

Term	Definition
Invading strand	ssDNA formed from one side of a DSB. RecA binds to the ssDNA, forming an ssDNA-RecA filament
Homoduplex DNA	dsDNA in the unbroken chromosome. It is composed of the complementary and outgoing strands.
Complementary strand	One of the DNA strands in the homoduplex. The ssDNA-RecA filament tests homology by trying to establish Watson-Crick pairing between the invading and complementary strands.
L_{test}	Number of base pairs that are tested for homology to determine whether an attempted pairing between L_{test} bases in the invading and complementary strands should become permanent.
Heteroduplex product	dsDNA formed by Watson-Crick pairing of L_{test} bases in the invading and complementary strands after the homology test is passed. In the models, homology testing stops after the L_{test} bp heteroduplex product forms.
stringency	The fraction of interactions that forms an L_{test} bp heteroduplex product that correctly pairs a sequence region in the invading strand with the corresponding sequence region in the complementary strand.
α	When L_{test} bases are tested for homology as a group, the homology test is passed if the total number of mismatches in the L_{test} bases is $\leq \langle L_{test}$.
$N_{mismatch}$	When deterministic homology testing is done in triplets, a triplet passes the homology test if the number of mismatches in the triplet is $\leq N_{mismatch}$.
$N_{mismatch8}$	When deterministic homology testing includes an 8 bp test, the 8 bp test is passed if the number of mismatches in the 8 bp is $\leq N_{mismatch8}$.
$Tripletpass_M$	When triplet homology testing is probabilistic, $Tripletpass_M$ is the probability that a triplet will pass a homology test if the triplet contains M mismatches, where M = 0, 1, 2, or 3.

https://doi.org/10.1371/journal.pone.0288611.t001

however, the probability that a triplet containing a single mismatch will pass a homology test is greater than 0, but less than 1.

Analytical treatment of the deterministic homology testing L_{test} base pairs in a random genome

The *in vitro* experiments presented above suggest that RecA can form readily observable heteroduplex products that contain 16% mismatches. This poor homology stringency might indicate that RecA mediated homologous recombination alone could not lead to accurate DSB repair; however, statistical considerations show that highly mismatch tolerant homology testing of a large number of base pairs can provide extremely accurate homology recognition. Thus, we studied whether there are L_{test} values that could allow highly mismatch tolerant homology testing to almost always lead to correct DSB repair.

The *in vitro* results in this paper indicate the probability that a triplet containing a single mismatch will pass a homology test is greater than 0, but less than 1; however, we first consider simple deterministic homology testing models in which the probability of passing a test is always 1 for interactions that meet a homology threshold and always 0 for interactions that fail to meet that threshold. Such models can provide insight into how the accuracy of mismatch tolerant homology testing is influenced by L_{test}, even if the models do not accurately capture the behavior of RecA mediated homologous recombination. Table 1 summarizes the terms used in the homology testing models.

We begin by applying the deterministic models to "genomes" consisting of randomly chosen bases. We refer to these sequences as "random genomes". We begin with random genomes for two reasons: 1. Simple analytical formulas describe the results of applying deterministic models to random genomes (Materials and Methods); and 2. Comparing results for random genomes to results for actual bacterial genomes allows us to highlight how the non-random nature of bacterial genome sequences could affect homologous recombination.

To highlight the influence of testing in triplets [22] and using an initial 8 bp test [16, 18–21], we consider three simple deterministic models (Fig 2). The first model considers the total number of mismatches in all L_{test} base pairs (Figs 2Bi, S1Ai, S1Bi and S1Ci). If the number of mismatches in L_{test} base pairs is $> \alpha L_{test}$, then the homology test always rejects the attempted base pairing and no stable heteroduplex product is formed. Otherwise, the homology test is passed, and a stable heteroduplex product is formed. Mismatch tolerance increases with α. Importantly, the result of the homology test is completely insensitive to the distribution of the mismatches within the L_{test} base pairs. We present this simple and very incorrect model to highlight how homology recognition is improved by dividing homology testing into base pair triplets and incorporating an initial 8-bp test.

The light blue lines with circular symbols in Fig 3 show the result for a homology test that does not accept any mismatches ($\alpha = 0$). In the random genome, good stringency can be achieved by testing < 20 bp, which is why early work suggested that it was reasonable that RecA rejects homologous products that extend over less than 20 bp [11].

In Fig 3A, the curves indicated by the red triangles, gray diamonds, and black squares represent results for $\alpha = 1/3$, 2/3, and 3/3, respectively. Fig 3A shows that homology testing of ~ 50 bp provides good stringency for $\alpha = 1/3$, but for $\alpha = 2/3$, stringency is poor even if 120 bp are tested.

The second model divides the L_{test} base pairs into triplets, and then applies a homology test to each of the triplets (Figs 2Bii, S1Aii, S1Bii and S1Cii). Passing a homology test of L_{test} base pairs requires that no triplet includes more than $N_{mismatch}$ mismatches. Otherwise, the homology test is always failed. The third model applies an 8 bp test that accepts one mismatch [18], with the remainder of the L_{test} -8 bases tested using the triplet test (S1Aiii, S1Biii and S1Ciii Figs). In both

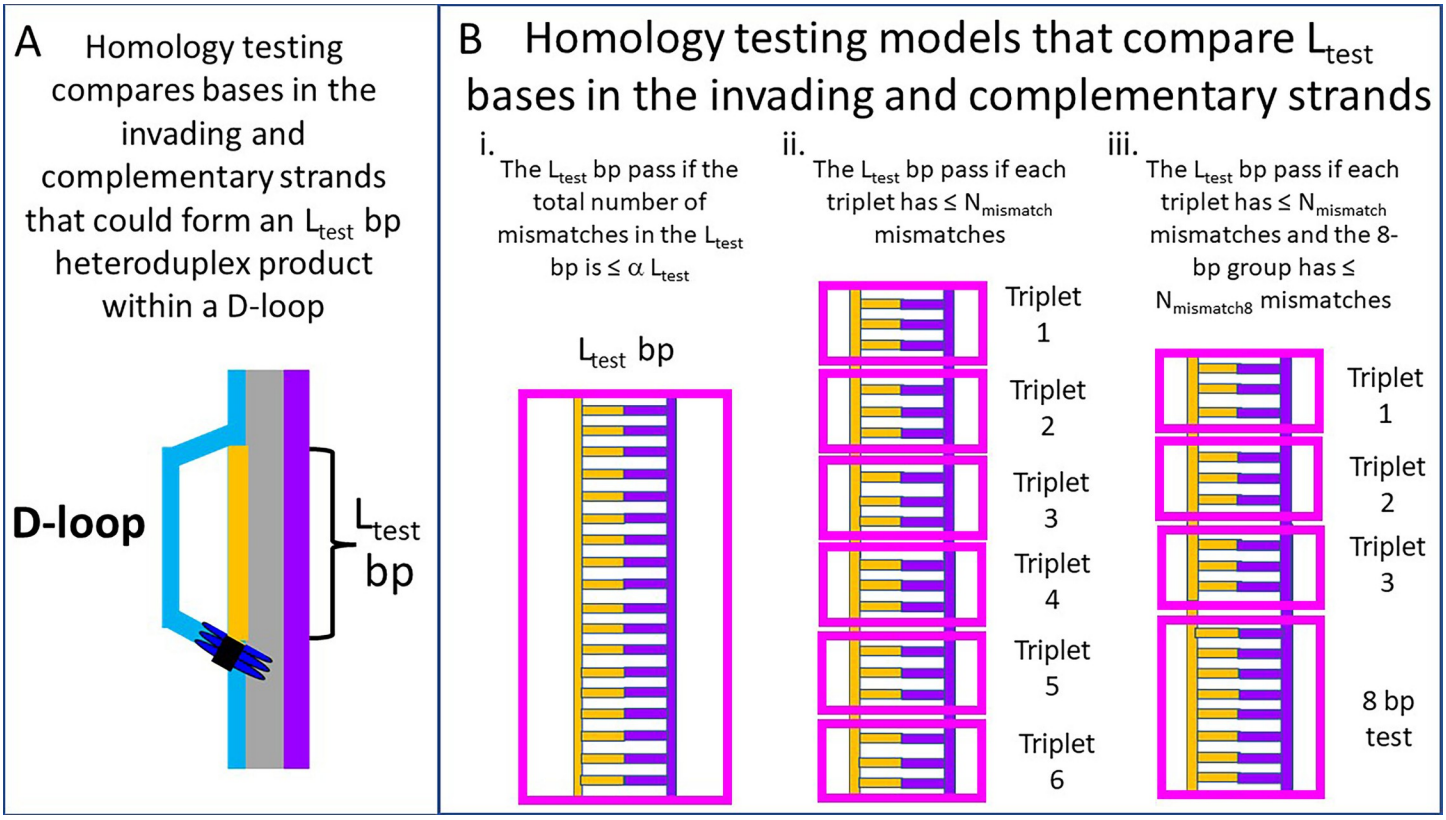

Fig 2. Illustrations of simple homology testing models that compare L_{test} base pairs in the invading and complementary strands. (A). RecA mediated homologous recombination forms a D-loop in which the invading (orange) and complementary (purple) strands bind displacing the outgoing strand (blue). The gray regions indicate base pairing. **(B)**. Different homology testing models: i. the test is passed if the total number of mismatches in L_{test} bp is ≤ α L_{test}. The strictness of the test depends on the value of α If α= 0, no mismatches are accepted. ii. The L_{test} base pairs are divided into triplets. Six separate triplets are shown. Each triplet is tested separately. An individual triplet passes the homology test if it includes ≤ $N_{mismatch}$ mismatches. The L_{test} bp pass the homology test if every individual triplet passes the triplet test. If $N_{mismatch}$ = 0, no mismatches are accepted. iii. Triplet testing is the same as in ii, but testing also includes an 8 bp test. The 8 bp test is passed if the 8 bp include ≤ $N_{mismatch8}$ mismatches. The L_{test} bp pass the homology test if the 8 bp test is passed and every individual triplet passes the triplet test.

https://doi.org/10.1371/journal.pone.0288611.g002

the second and third models, homology testing is sensitive to the distribution of mismatches within the L_{test} base pairs.

Fig 3B shows results for the model divides the L_{test} base pairs into triplets and then applies a homology test to each of the triplets. Again, there is a simple analytical formula that gives the stringency as a function of L_{test}. In Fig 3B, the curves with the red triangles, gray diamonds, and black squares correspond to results for triplet tests with $N_{mismatch}$ = 1, 2, and 3. Thus, the total number of allowed mismatches in Fig 3B is ($N_{mismatch}$/3) L_{test}. Importantly, for a given α, the total number of mismatches allowed in Fig 3A is α L_{test}, and the symbolism for the curve in Fig 3A corresponding to α is the same as the symbolism in Fig 3B corresponding to $N_{mismatch}$/3 = α. Thus, one can determine whether triplet testing improves stringency by comparing a curve in Fig 3B to the curve in Fig 3A that is represented by the same symbols.

The curves represented by the blue circles are identical in Fig 3A and 3B. Thus, if all mismatches are rejected, dividing homology testing into triplets offers no advantage. In contrast, if some mismatches are accepted, testing in triplets can bring a significant advantage. The advantage of triplet testing can be seen by comparing the curves represented by the red triangles and the gray diamonds. In Fig 3A, the curve with the gray diamonds indicates that α= 2/3

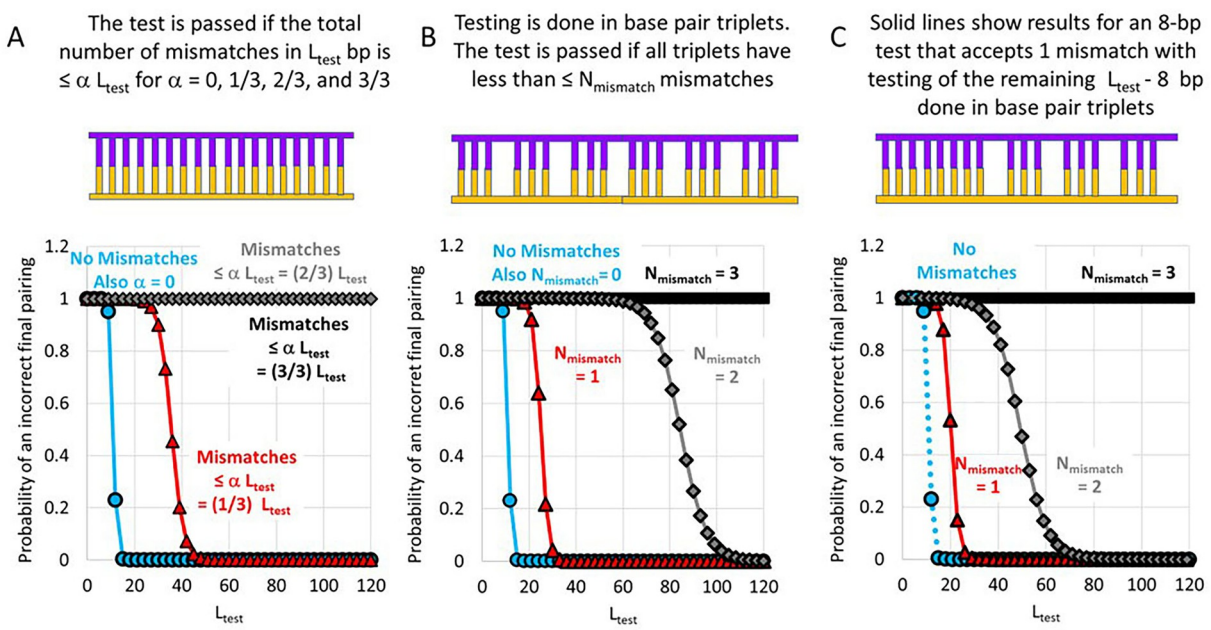

Fig 3. Probability of an incorrect final pairing vs. L_{test} for a random genome. In all panels, the light blue (circles) curve shows the result if no mismatches are accepted. **(A)**. The red (triangles), gray (diamonds), and black (squares) curves show the result if the number of accepted mismatches is αL_{test} = (m/3) L_{test} where m = 1, 2, or 3, respectively. The results for m = 2 and m = 3 are indistinguishable. (**B**). The curves show the results in which the L_{test} base pairs are grouped into triplets. Each of the L_{test}/3 triplets is tested for homology. The homology test of an individual triplet is passed if the number of mismatches in the triplet is less than $N_{mismatch}$. All L_{test} base pairs pass the homology test if all L_{test}/3 triplet homology tests are passed. The red, gray, and black curves with black outlined triangles, diamonds, and squares show the result for $N_{mismatch}$ = 1, 2, or 3, respectively. (**C**). The solid curves show results for homology testing that includes an 8-bp homology test that accepts one mismatch. ($N_{mismatch8}$ = 1) and homology testing of (L_{test}-8)/3 base pair triplets. The triangle, diamond, and square symbols indicate results for triplet tests that accept $N_{mismatch}$ bp for $N_{mismatch}$ = 1, 2, and 3, respectively.

https://doi.org/10.1371/journal.pone.0288611.g003

provides poor stringency even after testing 120 bp; however, Fig 3B indicates that testing in triplets allows $N_{mismatch}$ = 2 to provide good stringency by testing only ∼ 50 bp.

Finally, we considered the third testing model based on previous work suggesting that in an initial 8-bp test [16, 18, 20, 21] is followed by testing in base triplets [22] (Fig 2Biii and S1Aiii, S1Biii and S1Ciii Fig). Again, there is an analytical formula for the stringency, and that formula is independent of the order in which the tests are performed. Thus, the results do not depend on the directionality of strand exchange or the order in which the 8-bp test and the triplet tests are performed.

The results for testing that includes an 8-bp test are shown by the solid lines in Fig 3C. Comparison between the solid curves with the same symbols in Fig 3B and 3C demonstrates that for both for $N_{mismatch}$ = 1 (red triangles) and for $N_{mismatch}$ = 2 (gray diamonds), the initial 8-bp test decreases the L_{test} required to achieve a given stringency. The required L_{test} decreases because the 8 bp test is stricter than either triplet test. Interestingly, with an 8 bp test, even triplet tests that accept 2 mismatches can provide ∼99% stringency by testing ∼ 75 bp. For $N_{mismatch}$ = 3 (black squares) the 8 bp test offers a negligible improvement because the triplet test never rejects any pairing. In sum, Fig 3 shows that dividing L_{test} into groups that are tested for homology separately can greatly improve the stringency of mismatch-tolerant homology testing; however, the stringency of mismatch intolerant testing is not affected by dividing L_{test} base pairs into groups. We note that all the results in Fig 3 are independent on whether homology testing occurs iteratively or simultaneously (Fig 3); however, iterative testing vastly

improves searching speed. Furthermore, in iterative testing, searching speed improves if the 8 bp test occurs before the triplet tests.

Importantly, though different curves in Fig 3 represent various homology testing models, all the curves show that stringency as a function of L_{test} divides into two regimes. At low values of L_{test}, most sequences can pass a homology test at several positions in the genome; therefore, stringency is very poor and insensitive to L_{test}. At higher values of L_{test}, most sequences can only pass the homology test at the corresponding position in the genome, and stringency increases exponentially with L_{test}. Of course, the L_{test} value that divides the regime increases as homology testing becomes less strict.

Deterministic homology testing in bacterial genomes

Bacterial genomes do not consist of randomly chosen bases. Instead, they contain long repeats. S2A Fig shows histograms of the distribution of repeats in *E.coli* MG1655 that are longer than 99 bp. Importantly, there are 900 positions in a 1000-bp repeat at which a DSB could create an invading strand that includes 100 bp that occur at least one other position in the genome. Thus, S2B Fig shows a graph of the number of times that unique 100-bp sequences appear in the genome.

To determine the stringency for a homology test that considers L_{test} base pairs in the *E. coli* MG1655 genome and rejects all attempts to form a heteroduplex product that includes mismatches, we counted the number of 100-bp sequences that occur more than once in the given strand (Materials and Methods). Even a homology test that rejects all mismatches could not reject sequences that join different copies of these 100-bp repeats. S2C Fig indicates that there are more than 13,000 unique 100-bp sequences that occur exactly twice in the given strand of the *E. coli* MG1655, and one 100-bp sequence has 9 copies in that given strand.

For this work, we make the simplifying assumption that DSB are uniformly distributed across the genome. Furthermore, we assume that homology testing is also uniformly distributed across the genome. Given those assumptions, there is a >1% probability that a DSB at a random position in the *E. coli* MG1655 genome would lead to a sequence matched heteroduplex product that joins two different copies of a 100 bp sequence. The dark blue diamond in Fig 4B indicates the results of this calculation.

We now compare the results for the random genomes to the results for *E. coli* MG1655 (Fig 4A and 4B). The light blue lines in Fig 4B indicate that when $L_{test} = 21$, if no mismatches are accepted, the stringency for random genomes is $\sim$ 1-1x10^{-6}, which is much better than 99%. The stringency in the random genome is high because in a random genome the probability that two bases accidentally match is ¼ (S3 Fig). Thus, the probability that two 21-bp sequences in a random genome match is $1/4^{21} \sim 2x10^{-13}$. This low probability implies that random genomes do not contain long repeats. Additionally, the probability that a 21-bp sequence has a match in a 5x10^6 nt sequence is 2x10^{-13} x 5x10^6 $\sim$ 10^{-6}, consistent with the result shown by the dotted light blue line in Fig 4B. Therefore, Fig 4B indicates that if no mismatches are accepted the stringency achieved in a random genome when $L_{test} = 21$ is much better than the stringency achieved in a bacterial genome when $L_{test} = 100$ bp because when $L_{test} > 20$ bp most incorrect pairings in bacterial genomes join different copies of repeats. In sum, the long repeats in bacterial genomes limit stringency to $\sim$ 99% unless L_{test} is larger than 100 bp.

It is worth noting that stringency for bacterial genomes stringency cannot be expressed by a simple analytical formula, and the search technique that we used to find the exact matches does not extend to inexact matches. Thus, we used computer simulations to study how mismatch tolerance influences stringency in bacterial genomes (Materials and Methods). The solid curves with circular symbols in Fig 4 show results for computer simulations of homology

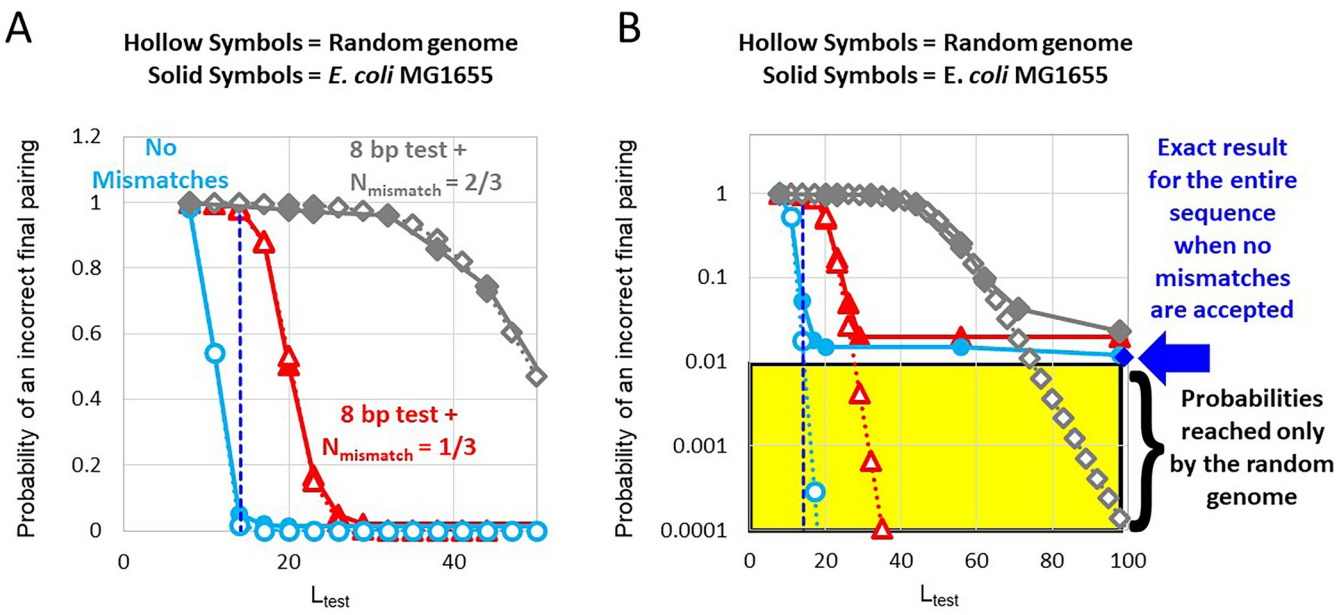

Fig 4. Probability that after a DSB RecA will create a final incorrect pairing as a function of L_{test} for the random genome (hollow symbols) and the *E. coli* MG1655 genome (solid symbols). **(A)**. The probability that a final pairing will be incorrect as a function of L_{test} and $N_{mismatch}$. The blue line shows results for sparse homology testing that requires complete sequence matching. The red and gray lines with triangle and diamond symbols indicate results for sparse homology testing when the 8 bp test that accepts one mismatch is followed by triplet tests that accept 1and 2 mismatches, respectively. The dark blue vertical dashed line indicates $L_{test} = 14$. **(B)**. Same as A with a logarithmic y-axis. The blue diamond highlighted by the blue arrow indicates the exact result for homology testing the entire genome when $L_{test} = 98$ and no mismatches are accepted. The yellow region indicates the stringencies that are achieved in the random genome but are not reached in the *E. coli* MG1655 genome.

https://doi.org/10.1371/journal.pone.0288611.g004

testing using an 8-bp test that accepts one mismatch and triplet testing of L_{test}-8 base pairs for an *E. coli* MG1655 genome. For comparison, results for this test in a random genome are shown by dotted curves with hollow-square symbols. In both cases, the blue curves show results for tests that reject all mismatches. The remaining curves show results for an 8-bp test that accepts one mismatch, where the remainder of the L_{test}-8 bases are then tested in base pair triplets. The red and gray curves show results for triplet tests that accept 1 and 2 mismatches, respectively. The results for the random genome and the bacterial genome seem nearly identical (Fig 4A). Fig 4B shows the same results as Fig 4A, except that the y axis is logarithmic. In Fig 4B, it is clear that the results for the random genome and the results for the bacterial genome diverge in the region where stringency $> 99\%$ because at those stringencies most incorrect heteroduplex products join different copies of long repeats.

For L_{test} values that give poor stringency in the random genome, the *E. coli* results agree well with the results for the random genome (Fig 4) because most groups of L_{test} bp that pass the homology test in the *E. coli* genome do not join different copies of long repeats. In contrast, for L_{test} values that provide $> 99\%$ stringency in a random genome (yellow highlighted region in Fig 4B), the *E. coli* results show much poorer stringency than the random genome. Importantly, once the *E. coli* stringency reaches $\sim 99\%$, increasing L_{test} does not improve stringency. This saturation of stringency occurs for both sparse sampling and complete sampling (S4 Fig) because most incorrect groups of L_{test} base pairs that pass the homology test join different copies of long repeats (S5 Fig). Other bacterial genomes also show a similar saturation of stringency with L_{test} because they too contain long repeats (S6–S8 Figs). In sum, independent of any feature of RecA-mediated homologous recombination, if the 3′ end of the invading strand is located at a random position in the genome, then even if homology testing rejects all

mismatches more than 1% of the 100 bp groups that pass the homology test would include complementary and invading strands from different copies of long repeats (Fig 4). Furthermore, if testing rejects all mismatches, then testing more than $\sim$ 14 bp does not significantly improve stringency.

After a DSB, if repair follows the RecBCD pathway the RecBCD protein interacts with the ends of the broken dsDNA [2, 26–28]. The function of RecBCD changes when it recognizes the 8-bp sequence GGCGGCGG, which is called the Chi site [28–30]. If there were no Chi sites in long repeats, then no heteroduplex product could join different copies of long repeats if the RecBCD pathway were followed; however, some Chi sites are positioned in long repeats [31]. Thus, to determine how terminating invading strands at or near Chi sites influences mismatch-tolerant homology recognition, we performed various simulations.

For Chi site simulations, the DSB occurs at a randomly chosen position in an *E. coli* MG1655 genome; however, the invading strand sequence used in the simulation terminates in the nearest Chi site on the 5′ side of the DSB. The homology test is then applied to the L_{test} base pairs on the 5′ side of the Chi site. We note that regardless of the direction of strand exchange, these are the L_{test} base pairs that are included in the heteroduplex that extends to the 3′ end of the invading strand.

There are $\sim$ 500 Chi sites in each strand of the 4.6 Mbp *E. coli* MG1655 genome [26]. Thus, the separation between Chi sites frequently extends over thousands of base pairs. As a result, thousands of different DSB positions will produce the same invading strand sequence. A detailed consideration of the distribution of Chi sites in genomes is required to determine whether the RecBCD pathway increases or decreases the probability that one side of a DSB will lead to a pairing between different copies of a long repeat. Interestingly, our simulations indicate that stringency as a function of L_{test} is not affected by whether or not the invading strand terminates in a Chi site (S8 Fig). These simulations only consider the invading strand formed by one side of the DSB. If the RecBCD pathway is followed, no single DSB could create two invading strands from the same long repeat [31].

Importantly, though different curves in Figs 4 and S6 represent various homology testing models, all the curves show that stringency as a function of L_{test} divides into three regimes. The first two regimes are shared with random genomes; however, for bacterial genomes an additional new regime begins when stringency as a function of L_{test} saturates. That asymptotic value occurs because pairings join different copies of long repeats. Finally, for bacterial genomes there is an eventual increase in stringency once $L_{test} \sim$ 1000 bp (S4 Fig), consistent with the histogram of repeat lengths shown in S2 Fig.

More realistic homology testing

The results of the deterministic homology recognition models provide insight into how L_{test} affects stringency. The results also highlight how long repeats in bacterial genomes limit the accuracy of homology testing, even if the testing does not accept any mismatches. Unfortunately, our *in vitro* results indicate that the deterministic models do not accurately capture homology testing by RecA; the *in vitro* results suggest that 1-bp mismatch is sometimes accepted and that 1-bp mismatch is sometimes rejected, whereas in the deterministic model the mismatch would either always be accepted or always be rejected. Therefore, we also considered models where the acceptance of mismatches is probabilistic. For example, in such models the probability of accepting a single mismatch in a triplet might be 75%, which implies that the probability that the mismatch would be rejected would be 25%.

Unfortunately, the bulk FRET measurements in Fig 1 do not allow us to accurately determine the probability that a triplet homology test will accept 1, 2, or 3 mismatches within that

triplet. Thus, we considered various probabilistic models. In the models, the probability of passing a triplet test is characterized by $tripletpass_M$, where M represents the number of mismatches in the triplet. It is known that RecA family proteins sometimes reject perfectly homologous sequences [32]. Thus, one might think that it would be useful to run simulations in which perfectly homologous triplets are sometimes rejected ($tripletpass_0 < 100\%$); however, if for all M tripletpassM is proportional to $tripletpass_0$, then the stringency vs L_{test} for a homology test with $tripletpass_0 < 100\%$ is same as the stringency vs L_{test} for $tripletpass_0 = 100\%$. Reducing $tripletpass_0$ below 100% increases searching times; therefore, for all the results shown in Fig 4, we chose $tripletpass_M = 100\%$ for completely matched triplets. This is perfectly compatible with many sequence-matched products reversing before they extend over L_{test} base pairs. Indeed, simulations in which perfectly matched sequences could reverse before reaching L_{test} showed the same stringency vs. L_{test} as simulations in which sequence-matched regions never reverse.

The various models we considered may not exactly capture the details of RecA-mediated homologous recombination *in vivo*, but they allow us to test whether stringency as a function of L_{test} is very sensitive to the details of the probabilistic homology testing. If the results of our models are insensitive to details of the models, then the modeling results may provide insight to mismatch tolerant homology testing by RecA.

Results of probabilistic homology testing are shown in Fig 5. The triangle and diamond symbols in Fig 5 repeat the curves shown in Fig 4 that indicate results of deterministic homology testing of *E. coli* MG1655 genomes using an initial 8-bp test followed by triplet tests that reject any triplet with more than $N_{mismatch}$ mismatches. The results of the deterministic tests can be compared with results of several different probabilistic testing models that include an initial 8-bp test that is followed by triplet tests that depend on the number of mismatches in each triplet.

Previous work has indicated that collective interactions between bases might allow a single mismatch to destabilize a triplet [22, 33, 34]. In that case, $tripletpass_M$ would be the same for all $M > 0$. In Fig 5A the orange and purple curves show results for $tripletpass_M = 25\%$ or 50%, respectively, for all $M > 0$. For the data shown in Fig 1, when there is a 20-nt homologous tail the invading strand with 1 mismatch per 6 bases includes 10 mismatched triplets. Thus, the probability of getting through all 10 mismatches would be $.25^{10} < 10^{-6}$ or $.5^{10} < 10^{-3}$. Of course, a FRET signal may not require progressing through all the mismatched triplets.

To further probe the robustness of the results, we considered probabilistic models in which each additional mismatch decreases the stability of the triplet. In this case $tripletpass_M$ would decrease with M (orange and purple curves in Fig 5B). The orange curve shows results for $tripletpass_M = 75\%$, 50%, and 25% for M = 1, 2, and 3 mismatches, respectively; therefore, for $tripletpass_M = 75\%$ the probability of getting through all 10 mismatched triplets would be $.75^{10} \sim 6\%$.

For bacterial genomes the predicted stringencies as a function of L_{test} fall in the range between the predictions of the simplistic deterministic model for $N_{mismatch} = 1$ (Fig 5, red curve) and $N_{mismatch} = 2$ (Fig 5, gray curve). We also considered other probability distributions that are compatible with the *in vitro* results shown in Fig 1, and those results also fall largely in the same range. In sum, for probabilistic models consistent with the *in vitro* results shown in Fig 1, the stringencies as a function of L_{test} are insensitive to the details of the models, including whether or not all mismatched triples have the same probability.

Discussion

Many aspects of RecA homology testing are not fully understood. In agreement with previous published results [13–17], this work presented *in vitro* data suggesting that even with ATP hydrolysis, RecA homology testing is highly mismatch tolerant (Fig 1). We now consider

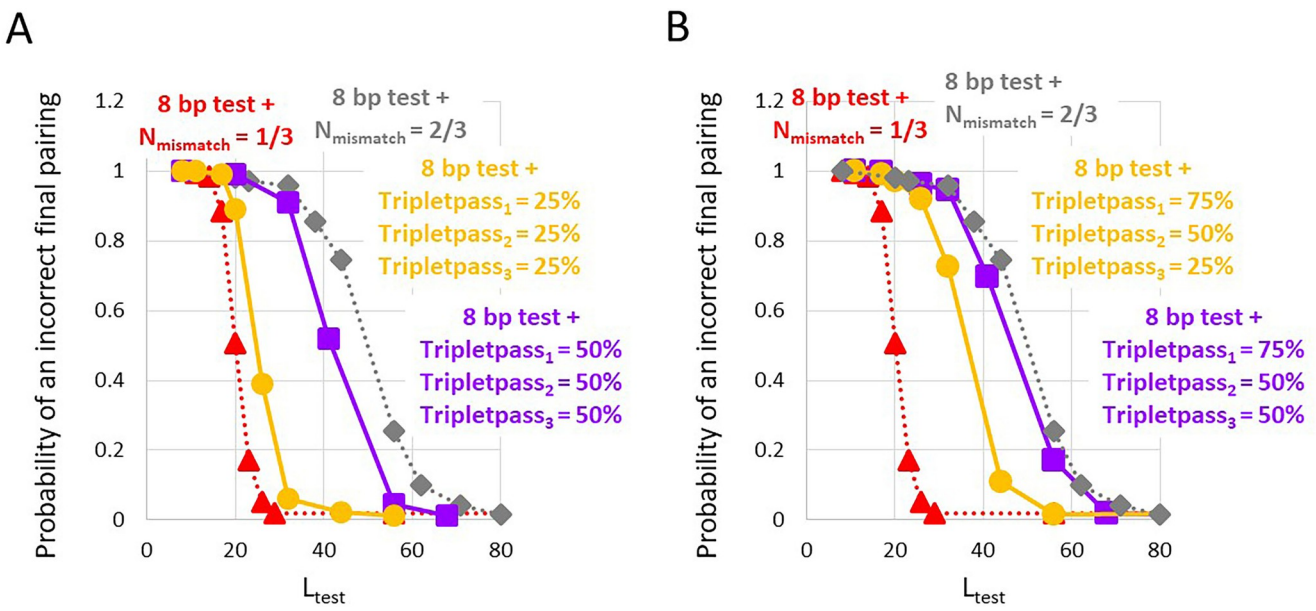

Fig 5. Probability that after a DSB RecA will create a final correct or incorrect pairing as a function of L_{test} for *E. coli* MG1655 genome for deterministic and probabilistic homology testing. **(A)**. Probability that a final pairing will be incorrect as a function of L_{test} and $N_{mismatch}$. The red and gray lines with triangle and diamond symbols indicate results for sparse homology testing when the 8-bp test that accepts one mismatch is followed by triplet tests that accept 1 or 2 mismatches, respectively. The orange and purple curves represent various probabilistic testing strategies. For both the orange and purple curves $tripletpass_M$ = 100% for completely matched triplets. The orange and purple curves with the circular and square symbols represent results for homology tests in which $tripletpass_M$ = 25% or 50%, respectively, for any triplet that contains at least one mismatch. **(B)** The curves with the triangular and diamond symbols are the same as in A. For both the orange and purple curves $tripletpass_M$ = 100% for completely matched triplets. The orange curve shows results for $tripletpass_M$ = 75%, 50%, and 25% for 1, 2, and 3 mismatches, respectively. The purple curve shows results when $tripletpass_M$ = 75%, 50%, and 50% for 1, 2, and 3 mismatches, respectively.

https://doi.org/10.1371/journal.pone.0288611.g005

origins and implications of the high mismatch tolerance of RecA mediated homologous recombination.

Searching speed is an important requirement for many sequence testing systems. We speculate that highly accurate sequence recognition requires rigidly held Watson-Crick partners that would slow homology testing; therefore, we propose that the mismatch tolerance of RecA is a consequence of structural features that speed DSB repair.

Importantly, the modeling results in this paper suggest that for bacterial genomes stringency vs. L_{test} for bacterial genomes divides into three regimes: 1. At small L_{test}, stringency is very poor and insensitive to L_{test} because almost every possible sequence of length L_{test} occurs many times in the genome. 2. At moderate L_{test}, stringency increases exponentially with L_{test} because most sequences of length L_{test} occur only once in the genome. 3. At large L_{test}, stringency saturates because testing more bases does not improve stringency since most incorrect products join different copies of long repeats.

We now review results of recombination frequency as a function of the homologous sequence length, L. Significant recombination requires a homologous segment that extends over more than 20 bp [9–11]. The frequency of recombination increases approximately exponentially as L increases from ~20 to ~75 bp; however, when L is larger than approximately 75 bp, recombination increases linearly with L [10]. Thus, the recombination frequency as a function of L divides into three distinct regimes: 1. For $L < 20$ bp recombination is negligible. 2. For $20 < L < 75$ bp recombination shows a steep exponential increase with L. 3. For $L > 75$ bp there is a very slow linear increase in recombination with L. In sum, both recombination as a function of L and predicted stringency as a function of L_{test} divide into three regimes.

We propose that highly mismatch tolerant homology testing underlies the following L dependent features of recombination: 1. Recombination is negligible when $L < 20$ because when $L_{test} < 20$ bp almost all products are wrong, whereas almost all 20 bp products are correct if testing rejects all mismatches. 2. Recombination increases strongly with L for $20 \leq L \leq 75$ bp because stringency increases strongly with L_{test} from $20 \leq L \leq 75$ bp. 3. Recombination increases slowly with L when $L > 75$ bp because testing more bases does not significantly improve stringency since most incorrect products join different copies of long repeats (Figs 3–5).

Finally, this work has shown that the stringency that can be achieved by RecA alone depends strongly on L_{test}, so it is important to consider what might govern L_{test} *in vivo*. With ATP hydrolysis even 150 bp products are highly unstable [25]. In contrast, *in vivo* incorporation of regions of accidental homology saturates at $\sim$ 75 bp [9, 10]; therefore, *in vivo* L_{test} is unlikely to be governed by the length-dependent stability of heteroduplex products. Thus, we speculate that irreversible alignment between the broken and unbroken chromosomes usually depends on significant polymerization by DNA polymerase Pol IV [35, 36], which requires that heteroduplex products extend over $\sim$ 50–75 bp [37].

Other proteins may also influence the stringency of DSB repair. Even after testing 98 bp, some products of highly mismatch tolerant testing could contain mismatches (S9 Fig). MutS, MutL, and UVrD could combine to reverse sufficiently mismatched heteroduplex products [38], consistent with recombination of closely related bacterial genomes being blocked by MutS [39].

Finally, the $\sim$1% of DSB repairs that join different copies of repeats would allow extensive Pol IV polymerization and would not be reversed by mismatch repair. Thus, if those pairings are not reversed by another protein, the pairings could lead to cell death or to genomic rearrangement.

In sum, we propose that the level of genomic alterations produced during recombinational repair reflects a critical balance between highly mismatch-tolerant RecA-mediated strand exchange and intervention by other cellular components. The balance is presumably tuned by evolutionary forces to meet the requirements of rapid repair, genetic stability, and genetic variation, which may vary according to the cellular environment.

Supporting information

S1 Fig. Illustration of different homology testing schemes.
(DOCX)

S2 Fig. Distributions of repeated sequences in the *E. coli* MG1655 genome.
(DOCX)

S3 Fig. Enumeration of all 16 possible base pair sequences.
(DOCX)

S4 Fig. Predicted incorrect DSB repairs as a function of L_{test} for $N_{mismatch} = 0$ for sparse sampling and complete sampling.
(DOCX)

S5 Fig. Probability that a DSB will result in an incorrect final pairing vs. L_{test} if all mismatches are rejected for different bacterial genomes.
(DOCX)

S6 Fig. Saturation in the decrease in incorrect pairings as L_{test} increases.
(DOCX)

S7 Fig. Ratio of the number of distinct pairs of starting locations in the given strands of bacterial genomes that share a repeat of length L to the genome length as a function of L.
(DOCX)

S8 Fig. Probability that a DSB will result in an incorrect final pairing vs. L_{test} if invading strands all terminate in Chi sites.
(DOCX)

S9 Fig. Predicted incorrect DSB repairs for L_{test} = 98.
(DOCX)

S1 Table. Sequences used for experiments in Fig 1.
(TIF)

Acknowledgments

We acknowledge useful interactions with Benjamin Tang, Adam Kaufman, and Sofia Roitman.

Author Contributions

Conceptualization: Mara Prentiss.

Formal analysis: Mara Prentiss, Chantal Prévost, Nancy Kleckner.

Funding acquisition: Mara Prentiss.

Investigation: Mara Prentiss, Dianzhuo Wang, Jonathan Fu, Claudia Danilowicz.

Methodology: Mara Prentiss.

Software: Mara Prentiss.

Writing – original draft: Mara Prentiss.

Writing – review & editing: Mara Prentiss, Chantal Prévost, Veronica Godoy-Carter, Nancy Kleckner, Claudia Danilowicz.

References

1. Bell JC, Kowalczykowski SC. RecA: Regulation and Mechanism of a Molecular Search Engine. Trends in Biochemical Sciences. 2016; 41(6):491–507. https://doi.org/10.1016/j.tibs.2016.04.002. PMID: 27156117
2. Symington LS. End resection at double-strand breaks: mechanism and regulation. Cold Spring Harb Perspect Biol. 2014; 6(8):a016436. https://doi.org/10.1101/cshperspect.a016436 PubMed PMID: PMC4107989. PMID: 25085909
3. Cox MM. Motoring along with the bacterial RecA protein. Nature Rev Mol Cell Biol. 2007; 8:127–38. https://doi.org/10.1038/nrm2099 PMID: 17228330
4. Alves I, Houle AA, Hussin JG, Awadalla P. The impact of recombination on human mutation load and disease. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2017; 372(1736). Epub 2017/11/08. https://doi.org/10.1098/rstb.2016.0465 PMID: 29109227; PubMed Central PMCID: PMC5698626.
5. Hughes D. Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. Genome Biology. 2000; 1(6):reviews0006.1. https://doi.org/10.1186/gb-2000-1-6-reviews0006 PMID: 11380986
6. Hughes D. Impact of homologous recombination on genome organization and stability. In: R.L. C, editor. Organization of the Prokaryotic Genome. Washington: ASM Press; 1999. p. 109–28.
7. Thi TD, López E, Rodríguez-Rojas A, Rodríguez-Beltrán J, Couce A, Guelfo JR, et al. Effect of recA inactivation on mutagenesis of Escherichia coli exposed to sublethal concentrations of antimicrobials. J Antimicrob Chemother. 2011; 66(3):531–8. https://doi.org/10.1093/jac/dkq496 PMID: 21212055

8. Kowalczykowski SC. An overview of the molecular mechanisms of recombinational DNA repair. Cold Spring Harb Perspect Biol. 2015; 7(11):a016410. https://doi.org/10.1101/cshperspect.a016410 PubMed PMID: WOS:000368535700002. PMID: 26525148

9. Lovett ST, Hurley RL, Sutera VA, Aubuchon RH, Lebedeva MA. Crossing over between regions of limited homology in *Escherichia coli*: RecA-dependent and RecA-independent pathways. Genetics. 2002; 160(3):851–9. PubMed PMID: WOS:000174790600004.

10. Watt VM, Ingles CJ, Urdea MS, Rutter WJ. Homology requirements for recombination in Escherichia coli. Proc Natl Acad Sci, USA 1985; 82(14):4768–72. https://doi.org/10.1073/pnas.82.14.4768 PubMed PMID: WOS:A1985ANF2200044. PMID: 3161076

11. Shen P, Huang HV. Homologous recombination in Escherichia coli—dependence on substrate length and homology. Genetics. 1986; 112: 441–57. https://doi.org/10.1093/genetics/112.3.441 PMID: 3007275

12. Anand R, Beach A, Li K, Haber J. Rad51-mediated double-strand break repair and mismatch correction of divergent substrates. Nature. 2017; 544:377. https://doi.org/10.1038/nature22046 https://www.nature.com/articles/nature22046#supplementary-information. PMID: 28405019

13. Bucka A, Stasiak A. RecA-mediated strand exchange traverses substitutional heterologies more easily than deletions or insertions. Nucleic Acids Res. 2001; 29(12):2464–70. https://doi.org/10.1093/nar/29.12.2464 PMID: 11410652

14. Morel P, Stasiak A, Ehrlich SD, Cassuto E. Effect of length and location of heterologous sequences on RecA-mediated strand exchange. J Biol Chem. 1994; 269(31):19830–5. https://doi.org/10.1016/S0021-9258(17)32095-1. PMID: 8051065

15. Rosselli W, Stasiak A. The ATPase activity of RecA is needed to push the DNA strand exchange through heterologous regions. EMBO J. 1991; 10(13):4391–6. https://doi.org/10.1002/j.1460-2075.1991.tb05017.x PMID: 1836761

16. Bazemore LR, Folta-Stogniew E, Takahashi M, Radding CM. RecA tests homology at both pairing and strand exchange. Proc Natl Acad Sci USA 1997; 94(22):11863–8. https://doi.org/10.1073/pnas.94.22.11863 PubMed PMID: WOS:A1997YD50600023. PMID: 9342328

17. Sagi D, Tlusty T, Stavans J. High fidelity of RecA-catalyzed recombination: a watchdog of genetic diversity. Nucleic Acids Res. 2006; 34(18):5021–31. https://doi.org/10.1093/nar/gkl586 PMID: 16990254

18. Danilowicz C, Yang D, Kelley C, Prévost C, Prentiss M. The poor homology stringency in the heteroduplex allows strand exchange to incorporate desirable mismatches without sacrificing recognition in vivo. Nucleic Acids Res. 2015; 43(13):6473–85. https://doi.org/10.1093/nar/gkv610 PubMed PMID: WOS:000359776500029. PMID: 26089391

19. Yang DR, Boyer B, Prévost C, Danilowicz C, Prentiss M. Integrating multi-scale data on homologous recombination into a new recognition mechanism based on simulations of the RecA-ssDNA/dsDNA structure. Nucleic Acids Res. 2015; 43(21):10251–63. https://doi.org/10.1093/nar/gkv883 PubMed PMID: WOS:000366410900025. PMID: 26384422

20. Qi Z, Redding S, Lee JY, Gibb B, Kwon Y, Niu H, et al. DNA Sequence alignment by microhomology sampling during homologous recombination. Cell 2015; 160: 856–69. https://doi.org/10.1016/j.cell.2015.01.029 PMID: 25684365

21. Hsieh P, Camerini-Otero CS, Camerini-Otero RD. The synapsis event in the homologous pairing of DNAs: RecA recognizes and pairs less than one helical repeat of DNA. Proc Natl Acad Sci USA. 1992; 89:6492–6. https://doi.org/10.1073/pnas.89.14.6492 PMID: 1631148

22. Lee JY, Terakawa T, Qi Z, Steinfeld JB, Redding S, Kwon Y, et al. DNA recombination base triplet stepping by the Rad51/RecA family of recombinases. Science. 2015; 349(6251):977–81. https://doi.org/10.1126/science.aab2666 PubMed PMID: WOS:000360646800054. PMID: 26315438

23. Chen Z, Yang H, Pavletich NP. Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures. Nature. 2008; 453:489–94. https://doi.org/10.1038/nature06971 PMID: 18497818

24. Menetski JP, Bear DG, Kowalczykowski SC. Stable DNA heteroduplex formation catalyzed by the Escherichia coli RecA protein in the absence of ATP hydrolysis. Proc Natl Acad Sci USA 1990; 87(1):21–5. https://doi.org/10.1073/pnas.87.1.21 PMID: 2404275

25. Danilowicz C, Vietorisz E, Godoy-Carter V, Prévost C, Prentiss M. Influences of ssDNA-RecA filament length on the fidelity of homologous recombination. J Mol Biol. 2021;433(18). https://doi.org/10.1016/j.jmb.2021.167143 PubMed PMID: WOS:000686349400018. PMID: 34242669

26. Smith GR. How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. Microbiol Mol Biol Rev. 2012; 76(2):217–28. https://doi.org/10.1128/MMBR.05026-11 PubMed PMID: WOS:000305508000003. PMID: 22688812

27. Spies M, C Kowalczykowski S. Homologous recombination by RecBCD and RecF pathways. Washington, D.C.: ASM Press; 2005.

28. Dillingham MS, Kowalczykowski SC. RecBCD enzyme and the repair of double-stranded DNA breaks. Microbiol Mol Biol Rev. 2008; 72(4):642–71. https://doi.org/10.1128/MMBR.00020-08 PubMed PMID: WOS:000261384900004. PMID: 19052323

29. Amundsen SK, Sharp JW, Smith GR. RecBCD enzyme "Chi recognition" mutants recognize Chi recombination hotspots in the right DNA context. Genetics. 2016; 204(1):139–52. Epub Epub 2016 Jul 8. https://doi.org/10.1534/genetics.116.191056 PubMed Central PMCID: PMCID: PMC5012381. PMID: 27401752

30. Handa N, Yang L, Dillingham MS, Kobayashi I, Wigley DB, Kowalczykowski SC. Molecular determinants responsible for recognition of the single-stranded DNA regulatory sequence, chi, by RecBCD enzyme. Proc Natl Acad Sci USA. 2012; 109(23):8901–6. https://doi.org/10.1073/pnas.1206076109 PubMed PMID: WOS:000304991100032. PMID: 22603794

31. Li C, Danilowicz C, Tashjian TF, Godoy VG, Prévost C, Prentiss M. The positioning of Chi sites allows the RecBCD pathway to suppress some genomic rearrangements. Nucleic Acids Res. 2019; 47 (4):1836–46. https://doi.org/10.1093/nar/gky1252 PubMed PMID: WOS:000467961200023. PMID: 30544167

32. Coic E, Martin J, Ryu T, Tay SY, Kondev J, Haber JE. Dynamics of homology searching during gene conversion in Saccharomyces cerevisiae revealed by donor competition. Genetics. 2011; 189(4):1225–33. https://doi.org/10.1534/genetics.111.132738 PubMed PMID: WOS:000298412100008. PMID: 21954161

33. Cisse II, Kim H, Ha T. A rule of seven in Watson-Crick base-pairing of mismatched sequences. Nature Structural & Molecular Biology 2012; 19: 623–7.

34. Peacock-Villada A, Coljee V, Danilowicz C, Prentiss M. ssDNA pairing accuracy increases when abasic sites divide nucleotides into small groups. Plos One. 2015; 10(6):e0130875. https://doi.org/10.1371/journal.pone.0130875 PubMed PMID: WOS:000358147500105. PMID: 26115175

35. Henrikus SS, Wood EA, McDonald JP, Cox MM, Woodgate R, Goodman MF, et al. DNA polymerase IV primarily operates outside of DNA replication forks in Escherichia coli. PLOS Genetics. 2018; 14(1): e1007161. https://doi.org/10.1371/journal.pgen.1007161 PMID: 29351274

36. Henrikus SS, van Oijen AM, Robinson A. Specialised DNA polymerases in Escherichia coli: roles within multiple pathways. Current Genetics. 2018; 64(6):1189–96. https://doi.org/10.1007/s00294-018-0840-x PMID: 29700578

37. Lu D, Danilowicz C, Tashjian TF, Prévost C, Godoy VG, Prentiss M. Slow extension of the invading DNA strand in a D-loop formed by RecA-mediated homologous recombination may enhance recognition of DNA homology. J Biol Chem. 2019; 294(21):8606–16. https://doi.org/10.1074/jbc.RA119.007554 PubMed PMID: WOS:000471108200027. PMID: 30975899

38. Tham KC, Hermans N, Winterwerp HH, Cox MM, Wyman C, Kanaar R, et al. Mismatch repair inhibits homeologous recombination via coordinated directional unwinding of trapped DNA structures. Mol Cell. 2013; 51(3):326–37. Epub 2013/08/13. https://doi.org/10.1016/j.molcel.2013.07.008 PMID: 23932715; PubMed Central PMCID: PMC3781583.

39. Zahrt TC, Maloy S. Barriers to recombination between closely related bacteria: MutS and RecBCD inhibit recombination between Salmonella typhimurium and Salmonella typhi. Proc Natl Acad Sci USA. 1997; 94(18):9786–91. Epub 1997/09/02. https://doi.org/10.1073/pnas.94.18.9786 PMID: 9275203; PubMed Central PMCID: PMC23269.