Supplementary Model

1 Summary

In this supplementary note, we build a stochastic model of RecBCD's action on a DNA double strand break (DSB). As explained in the main text, the goal is to use this model to estimate the key biophysical parameters describing the mode of action of RecBCD *in vivo*. Given a set of parameters, the model can be used to assign a likelihood to our experimental data. Whichever set of parameters maximises this likelihood will provide the estimate we seek.

The model is of a hybrid "mechanistico-genomic" nature. On one hand, it draws from traditional stochastic modeling (discrete-time Markov chains) to represent the progression of the RecBCD complex on DNA and the various stages of DNA resection after a DSB; on the other hand, it incorporates precise genomic information to fix the position of Chi sites which are the master triggers of this process. It is this somewhat unusual combination of mechanistic and genomic information which allows us to exploit the data quantitatively and use it to investigate some of its underpinning biophysics.

This note is organised as follows. First, we review the extant knowledge and detail the simplifications we make to obtain the structure of our model. As is generally the case, the exercise of setting up the model is an excellent way to integrate the current biological understanding of the process. With the basic modeling choices in place, we analyse the mathematical structure of the model. We find that the model is simple enough that one can derive a closed formula for the resected single-stranded DNA segments produced by our idealised stochastic process. Then, we use this formula to compute the likelihood of the actual data according to various choices of parameters, and narrow down on a most likely set thereof. Finally, we expand on the discussion of our results presented in the Main Text.

2 Model

2.1 The mechanism of action of RecBCD

Extensive biochemical characterisations reviewed in Ref. [3] and in Ref. [9] demonstrate that the RecBCD complex loads on a DSB and translocates along DNA until it recognises a Chi site. Chi recognition is not certain, and RecBCD may read through several Chi sites before recognising one. Before recognition, the RecB and RecD motors are both engaged. As RecB is slower than RecD, a single strand loop accumulates ahead of RecB. Upon recognition, RecB becomes the lead motor and RecBCD's activity is modified so that the 5' strand is degraded, while RecA gets loaded on the 3' overhang. The loop formed prior to Chi recognition contributes to the 3' resected end that starts at the recognised Chi. Fig. 1 summarises the two stages of the resection process. Eventually RecBCD stops loading RecA and dissociates from DNA. This model is equally compatible with biochemical data of RecBCD activities obtained when the concentration of magnesium exceeds that of ATP or when the concentration of ATP exceeds that of magnesium.

2.2 Modeling choices

We translate this molecular knowledge in a series of modeling decisions and simplifying assumptions which we detail below. First, we model the recognition of a Chi site as a *stochastic* event. This seems natural as it is well observed that Chi recognition is not deterministic, and indeed only a stochastic model will allow us to get quantitative estimates on this important aspect of the process. Specifically, we assume that Chi sites are recognised by RecC with a probability p_χ which does not depend on the distance from the DSB, nor does it depend on the number of Chi sites previously encountered by RecBCD.

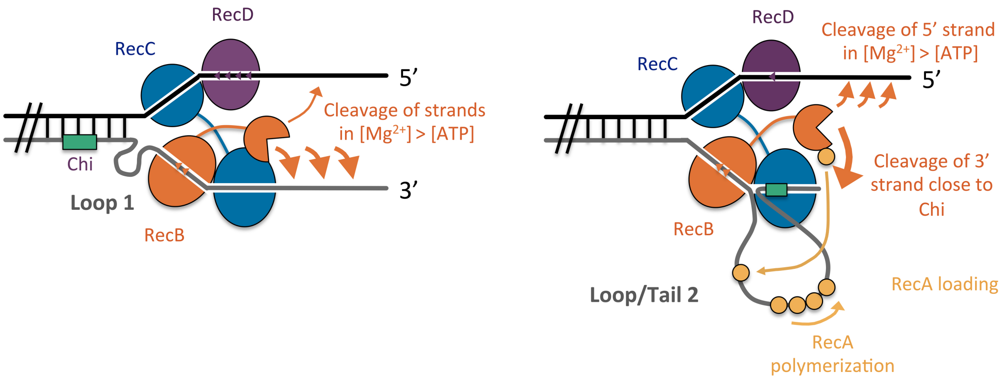

Figure 1. Sketch of DNA resection by RecBCD. Left panel - before Chi recognition: the RecB and RecD motors move along DNA and the RecB motor lags behind the RecD one; a loop forms ahead of RecB. Right panel - after Chi recognition: the entire RecBCD complex undergoes a conformational change which directs RecB's nuclease activity to the 5' strand, and induces the loading of RecA on the 3' one. In this schematic representation, the Chi site is shown held in its recognition site. However, the Chi site will be released either by disassembly of the RecBCD complex or at some point prior to this and the second single-stranded region will be converted from a loop to a tail.

We also model RecBCD's translocation in a stochastic way. The specific translocation mode depends on whether a Chi site has already been recognised or not. *Before* recognition, to take into account the different speeds v_B, and v_D of RecB and RecD, we distinguish two types of steps:
- one where the RecB and RecD motors move in unison with probability p_-;
- one where only the faster one, RecD, moves with complement probability $p_+ = 1 - p_-$.

In this mode, the mean ratio of the distances covered by RecD and RecB after any number of steps is given by $1/p_-$ (see §2.4), hence the speed ratio of the two motors is given by $v_B/v_D = p_- \leq 1$. This means that, consistently with Ref. [12], the model assumes that $v_B \leq v_D$.

Together with p_χ, estimating the speed ratio p_- is a key objective of the model.

After Chi recognition, the 3' strand is extended further and RecA is loaded on this strand. In this second mode, we assume that RecA is loaded uniformly on the single strand and we suppose that there is a constant probability p_{stop} for RecBCD to stop loading RecA (or to fall off) at each step.

We write τ_1 for the length of the single stranded loop (on the 3' strand ahead of RecB) at the time a Chi site is recognised. The mean value of τ_1 depends linearly on the distance of the said Chi site from the original DSB - the further the Chi site, the longer the loop. Similarly, we write τ_2 for the length of the single strand extension after Chi recognition and until RecBCD stops loading RecA (and possibly dissociates from DNA). Differently from τ_1, the value of τ_2 does not depend on which Chi site is recognised.

We also assume that the RecB subunit starts loading RecA only after Chi recognition (on the 3' resected strand). This means that the resected segment will begin at whichever Chi site is recognised and will have a total length of $\tau_1 + \tau_2$. And finally, we assume that whenever RecBCD falls off DNA *before* having recognised a Chi, the obtained single strand is not observable in the experiment as no RecA has been loaded.

Putting our choices together, we obtain a stochastic model (a discrete-time Markov chain) which generates the 3' resected segment onto which RecA is loaded. The model uses a restricted set of parameters $\mathcal{P}$ which consists of p_χ, p_-, and p_{stop}. Its overall structure is described in Fig. 2.

The model also includes the spatial configuration of Chi sites on the DNA. Let $I = \{1, \ldots, c\}$ be the set indexing the Chi sites in order of appearance after the DSB, we write λ_i for the distance of the i^{th} Chi site from the DSB with $\lambda_1 < \ldots < \lambda_c$. Because we know the genome sequence of the strain of interest, and the sequence of the Chi sites (5'-GCTGGTGG-3'), there is no need to make these sites explicit parameters of the model. It has been suggested that other sites can act as Chi-like motifs [2], but these are weaker and we do not take them into account.

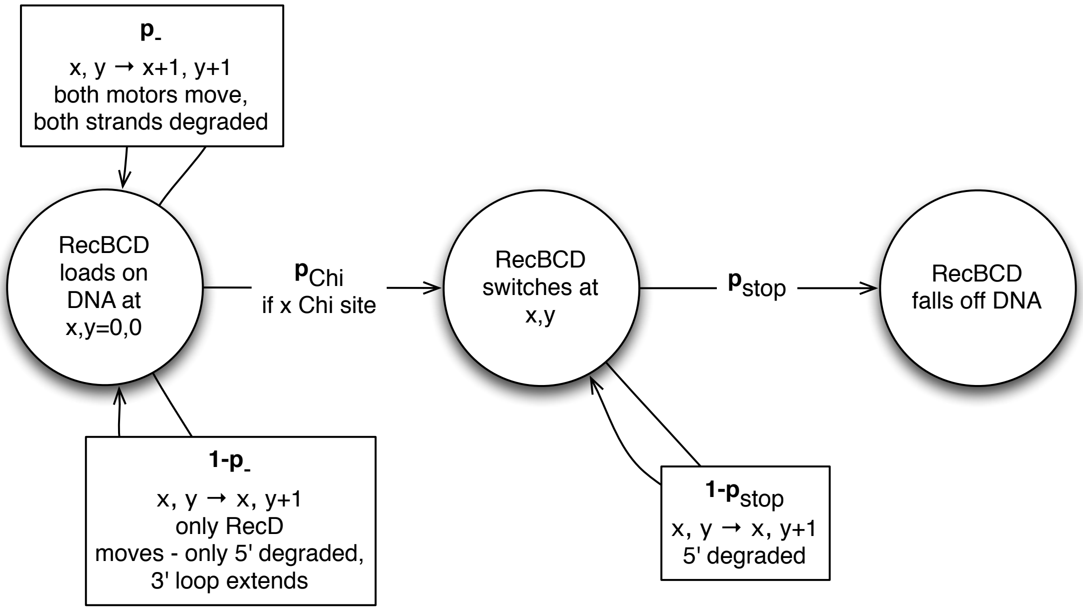

Figure 2. Decision tree for the model of DSB resection by RecBCD: x, y represent the respective DNA positions currently read by RecB and RecD; when x is a Chi site, with probability p_χ RecBCD switches to the mode where only 5' is degraded, else both motors continue to translocate along the dsDNA as before.

2.3 Variants

There are several other modeling options we could have considered. Let us mention two. A natural way to enrich the model would be to allow for reversible translocation of the motors, following the lines of the toy bimotor model developed in Ref. [10]. This would result in a smoother behaviour and potentially describe better the finer details of the biophysics of the motors. Another natural elaboration is to assume stochasticity in the parameters v_B, v_D governing the speed of the motors on DNA. Indeed, it has been shown recently that, *in vitro*, the pre-recognition translocation speed of RecBCD is itself fixed for an entire run by initial stochastic molecular events [7]. We discuss later whether incorporating this particular observation could result in a useful refinement of our model. With the simple model which we employ first, there is no need to predict the kinetics of the operation of RecBCD, and therefore no need to calibrate the time unit implicitly in our discrete-time modeling.

2.4 Derivation of the single strand distribution

There are three sources of randomness which jointly determine the segment produced by RecBCD:
- Y the (index of the) Chi site recognised by RecC,
- τ_1 the length of the single strand loop at the time a Chi site is recognised, and
- τ_2 the additional distance travelled by RecBCD after having recognised a Chi site.

In the following, we derive a simple formula for the distribution of these segments and their total length $\tau = \tau_1 + \tau_2$, and for the probability $Pr(x|\mathcal{P})$ that a nucleotide x is part of a segment.

Our first step is to calculate τ_1, assuming the Chi site recognised is at distance λ from the DSB. The value of τ_1 is given by the number of steps where RecB has not moved, and which have therefore resulted in extending the loop ahead of RecB, by the time RecB reaches λ.

Let $\mathcal{B}^-(X = k|n,p) = \binom{n+k-1}{k} p^n (1-p)^k$ denote the probability that a random variable X, distributed according to a *Negative Binomial* with parameters $n > 0$ and $p > 0$, takes a non-negative integer value k. The values of X track the number of failures needed to obtain n successes, each trial being independent, and p being the common probability of success. This translates directly to our setting, with n being the number of moves of RecB prior to Chi recognition, and p being p_- the probability of RecB moving.

Hence, when RecB arrives at position λ, RecD is ahead at position $\lambda + \tau_1$, with the distance between the two, namely τ_1, being distributed as:

$$Pr(\tau_1 = k) = \mathcal{B}^-(\tau_1 = k|\lambda, p_-) \tag{1}$$

From this, we can write an explicit formula for the mean length of the loop as a function of λ - the position where recognition happens (measured as a distance from the DSB):

$$\tau_1 = \lambda(1 - p_-)/p_- \tag{2}$$

This formula is useful to evaluate the impact of p_- on the length of the loop. We can see from this that the mean ratio of the distances covered by the two motors is the mean of $(\lambda + \tau_1)/\lambda$. As a negative binomial has mean $n(1-p)/p$, we find that the mean speed ratio is $1 + (1-p_-)/p_- = 1/p_- = v_D/v_B$. In other words, p_- is none other than the v_B/v_D speed ratio.

Our second step is to evaluate the additional distance τ_2 travelled by RecBCD (with only RecB engaged, and the 5' strand being degraded) after recognition of the Chi site and until RecA loading stops. The 3' strand is extended until RecBCD stops loading RecA and/or dissociates, hence τ_2 follows a geometric distribution with parameter p_{stop}. We will write $\mathcal{G}(X = k|p) = \mathcal{B}^-(X = k|1, p) = (1-p)^k p$ for the geometric distribution of parameter p where the random variable X tracks the number $k \geq 0$ of failures.

Taking into account the fact that τ_1 and τ_2 are independent variables, we get the following expression for the distribution of the total length $\tau = \tau_1 + \tau_2$ of the segment produced by RecBCD, conditioned on the Chi site at λ being recognised:

$$Pr(\tau = z|\lambda) = \sum_{k=0}^{z} \mathcal{B}^-(\tau_1 = k|\lambda, p_-)\mathcal{G}(\tau_2 = z - k|p_{stop}) \tag{3}$$

The next step is to obtain the joint distribution of the 2D-random variable (τ, λ) where τ is the length of the segment, and λ is the distance from the DSB where the segment starts. As in our simple model, the DNA is always degraded up to the recognised Chi, λ takes values in the set of distances of Chi sites from the DSB, namely $(\lambda_i; \in I)$. The index Y of the Chi site eventually recognised is distributed as $\mathcal{G}(Y = i+1|p_\chi)$ for $0 \leq i < |I|$. (The offset by 1 comes from the fact that we start numbering Chi sites at 1).

Putting our calculations together we get the joint distribution:

$$Pr(\tau = z, \lambda = \lambda_i) = \mathcal{G}(Y = i+1|p_\chi) \sum_{k=0}^{z} \mathcal{B}^-(\tau_1 = k|\lambda_i, p_-)\mathcal{G}(\tau_2 = z - k|p_{stop}) \tag{4}$$

From this one can compute the *hit probability* of a nucleotide x, that is to say the probability of x to be included in the segment defined by τ and λ:

$$Pr(x|\mathcal{P}) = \sum_{\{i \in I|\lambda_i \leq x\}} \sum_{\{z \geq x - \lambda_i\}} Pr(\tau = z, \lambda = \lambda_i) \tag{5}$$

Note that the hit probability at x is zero unless one of the Chi sites before x is recognised. In particular, a 'runaway' RecBCD which fails to recognise *any* Chi, generates no segment and induces no RecA loading. Note also that the sum $\sum_x Pr(x|\mathcal{P})$ is not 1, as many x's receive hits simultaneously. In fact, $\sum_x Pr(x|\mathcal{P})$ is the mean number of hits, that is to say the mean length of the resected segment.

The hit probability depends strongly on the particular set of parameters $\mathcal{P}$ and we will exploit this dependency to estimate our three parameters: p_χ, p_-, and p_{stop}. By sampling the set of parameters, we can compute for each set how likely the data are according to this set -a quantity defined as the *likelihood* of the parameter set (see below for a precise definition). Provided we can do this sampling efficiently, we can obtain a precise 'heat map' of the parameter space, whose peaks will denote the maximally likely values of the parameters.

2.4.1 An approximation

In order to sample efficiently our parameter space, we use an approximation of $Pr(x|\mathcal{P})$ and replace τ_1 by its mean $\lambda_i(1-p_-)/p_-$. This is equivalent to supposing that the speed ratio v_B/v_D is constant.

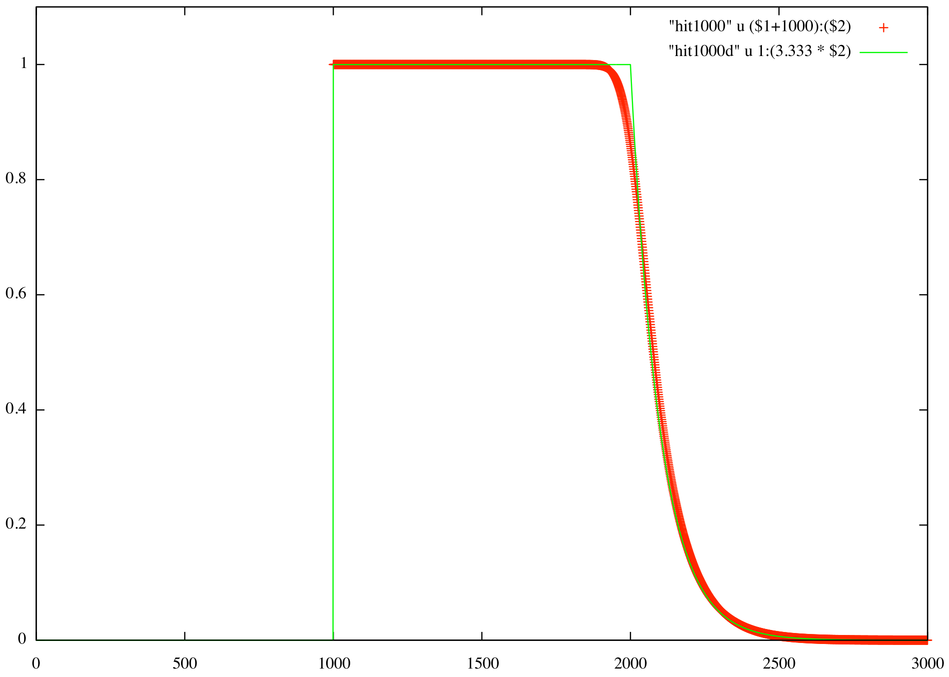

Figure 3. We use here for comparison a single transition site positioned at 1000 with $p_{stop} = 0.01$, $p_{-} = 0.5$, $p_{\chi} = 0.3$. Hence $\alpha = 2$ and the approximation (green) of $Pr(x|\lambda = 1000, \mathcal{P})$ is flat until position 2000. We see that it is quite close to the exact calculation (red).

With this approximation the expression for the hit probability, conditioned on recognition happening at λ_i, simplifies to:

$$Pr(x|\lambda = \lambda_i, \mathcal{P}) = \mathbf{1}_{\{\lambda_i \leq x \leq \alpha\lambda_i\}} + (1 - p_{stop})^{x - \alpha\lambda_i} \mathbf{1}_{\{x > \alpha\lambda_i\}} \tag{6}$$

with $\alpha = 1 + (1 - p_{-})/p_{-}$, and $\mathbf{1}_A$ the indicator function for A. This approximation incurs a negligible loss of precision as we see in Fig. 3 for a set of representative parameters. In general, the normalised error on the hit probability will be of the order of the coefficient of variation $1/\sqrt{\lambda_i(1 - p_{-})}$ which quickly becomes negligible as λ_i increases, as the closest Chi sites stand at $3kb$ from the DSB.

3 Data

The data consist of six data sets corresponding to the strains carrying 1 to 6 Chi sites at $3kb$ on the origin proximal side of the DSB. We focus on a $100kb$ region X on the origin proximal side of the DSB, as beyond this distance the signal reaches background noise level. This region does not contain any DSB-independent RecA binding loci (see Main Text) thus allowing us to apply the model described above on the entire region.

We use as input the $50bp$ reads mapped on the reference genome using novoalign version 2.0 (see Supplementary Methods). In order to compensate for any bias introduced by PCR amplification of DNA fragments before sequencing, multiple duplicate fragments (fragments starting and ending at the same positions) are replaced by a single 50bp read. The data are then processed by dividing the region in $250bp$ long non-overlapping bins and aggregating the reads that fall within each bin. The size of the bin is chosen as a trade-off between data robustness and resolution. As the bin size is much smaller than the expected size of a single strand coated by RecA (which is in the order of several kb) resolution should be minimally affected.

It remains to define and measure the *background* level of the RecA signal. To do this, we assume that there is no RecBCD-mediated loading of RecA before the Chi sites which stand closest to the break at about $3kb$. The RecA signal seen in this Chi-less region is treated as background, and we subtract its average level from the the binned data before comparison with the model.

3.1 Comparing model and data

In order to compare our model and the data, we rank sets of parameters $\mathcal{P}$ according to the probability they assign to the processed data within the region of observation X (see above). For adequate comparison the model results are aggregated in bins of $250bp$. One of the parameters which has a dimension, namely p_{stop} which is the inverse of a distance, is divided by the bin size 250.

The probability of detecting a nucleotide x in X (or the probability of observing a hit at x) can be written as:

$$F(x|\mathcal{P}) = \frac{Pr(x|\mathcal{P})}{\sum_{x\in X} Pr(x|\mathcal{P})} \tag{7}$$

This simple model assumes that the DNA fragments that are read are of the same length and located at identical positions as the initial single strand fragments onto which RecA is loaded. It also assumes that the DNA fragments are distributed uniformly and not biased by the sequence. This assumption is supported by the following arguments: (i) DNA fragments produced by sonication at the start of the ChIP process are unaffected by RecA binding. Hence fragments whether covered by RecA or not have the same probability of being sheared. (ii) As said above, PCR generated duplicates (identical fragments) are discarded, and then only the first $50bp$ (out of an average length of the fragment of $200bp$) of each remaining sequenced fragment is retained in the final hit count. (iii) The pileup data generated from the input samples (without RecA pulldown) show no sequence bias in the double strand break region (Fig. 4) (we note that the sequence GC content does not vary significantly in this region which does not contain horizontally transferred segments). (iv) While *E. coli* replication mechanism will lead to regions close to the origin showing a higher DNA copy number than regions close to the terminus of replication, the variability on the analysed region given a cell doubling time of 40 minutes is 0.25% and is therefore negligible.

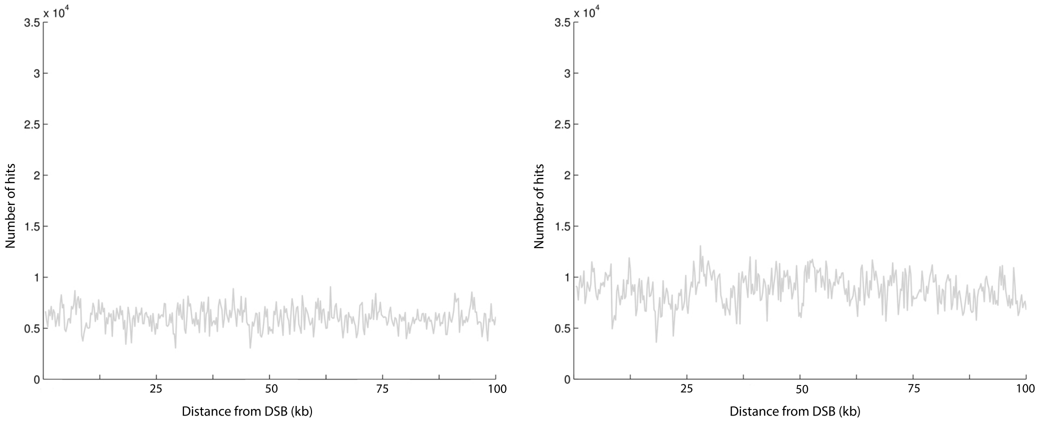

Figure 4. Two replicates of the hit counts per 1 kbp obtained from sequencing without RecA immuno-precipitation on the region of interest

3.2 Parameter estimation

The amount of DNA obtained before PCR amplification (of which the only role is to produce enough material for sequencing) is small enough that with a very high probability, no two reads come from the same individual DSB event. This means that the hits recorded from each read are approximately statistically independent. Hence we can conceptualize the experiment as drawing repeatedly and independently from a pool of nucleotides (the total amount of DNA collected in a given experiment), some of them being marked (included in a resection segment), and some not. Then again, because the sample taken from the pool is extremely 'thin', we can assume that the drawing is with replacement, and follows therefore a multinomial distribution. The likelihood of the data, can then be written to a very good approximation in this simple way:

$$\mathcal{L}(\mathbf{n}|\mathcal{P}) = C(\mathbf{n}) \prod_{x\in X} F(x|\mathcal{P})^{n_x} \tag{8}$$

where X is our $100kb$ region of interest, x ranges in the observation region X, $\mathbf{n}$ is the sequence n_x of x's hit counts in the data, and $C(\mathbf{n})$ is a multinomial coefficient. Taking a logarithm of the above, and forgetting $C(\mathbf{n})$ which does not depend on $\mathcal{P}$, and therefore plays no role in the maximisation of $\mathcal{L}$, we arrive at the following objective:

$$\sum_{x\in X} n_x \log F(x|\mathcal{P}) \tag{9}$$

that is to say we wish to find the value of $\mathcal{P}$ which maximises the above expression.

To estimate this best set of parameters, we follow a simple strategy and sample the $[0,1]^3$ interval as follows:
- $[0.5, 1]$ with step size 10^{-2} for $p_- = v_B/v_D$,
- $[0.1, 0.7]$ with step size 10^{-2} for p_χ,
- $[0.810^{-4}, 1.2 \times 10^{-4}]$ with step size 4×10^{-6} for p_{stop}.

The sampling was implemented using a Matlab script (available upon request) to compute the log-likelihood and locate the global maximum. The obtained optimal values are shown in Fig. 5. The box plots describe the likelihood of each parameter in the neighbourhood of these optimal values (see below).

3.3 Discussion

As one can see in Fig. 3 (Main Text), our parsimonious model reiterates the data rather well for the optimal parameters. This suggests that the model has indeed captured some of the salient aspects of the mechanisms at play in the real system.

To gauge the local log-likelihood distribution at higher resolution than our initial grid sampling, we ran a Metropolis-Hastings algorithm starting at the previously identified global maximum. The jump sizes are taken to be uniformly distributed within $\pm\epsilon$, where ϵ is the resolution of the mesh used in the grid sampling (see right above). The associated random walk samples the immediate neighbourhood of our best estimate (10000 steps for each data set). The results shown as box plots in Fig 5 confirm the presence of strong local maxima.

With the obtained parameters, the size of the loop, namely τ_1 in our notations, will be of the order of $0.05/0.95 \times 10^2 kb \sim 5kb$ at the far $100kb$ end of the Chi site range in X (the Chi sites most distant from the DSB), while τ_2, the other component of the length of the single strand is independent of the site of recognition and of the order of $1/p_{stop} \sim 10kb$. The resected segment will take a range of values which is bounded below by τ_1. As the efficiency of the search for an homologous sequence for repair depends on the length of the segment, the τ_1-"loop" might have a determinant role to play. In addition, the loop also plays a role in the loading of RecA, and therefore efficient loading might also depend on τ_1.

The values predicted for p_{stop} and $v_B/v_D = p_-$ are stable across the 6 different data sets, as they should, as these values are meant to capture mechanistic parameters that are independent of the conditions of the experiments. On the other hand, the value of the recognition probability varies from one data set to the next: there is a decrease in the predicted p_χ as the number of Chi sites in the initial array increases. The Chi sites in the array are separated by only $10bp$. The trend which we observe in p_χ is likely due to them being placed too close in the array for the Chi recognition subunit, RecC, to work independently on each site. This means in turn that the most robust estimate of p_χ is likely to be found in the case of the array containing 1 Chi site. We will focus on this data set below.

3.4 Model comparison

Estimates of v_B/v_D using our initial model (referred to below as the basic model) are close to 1 and substantially higher than reported in the literature [9]. This raises the question as to whether

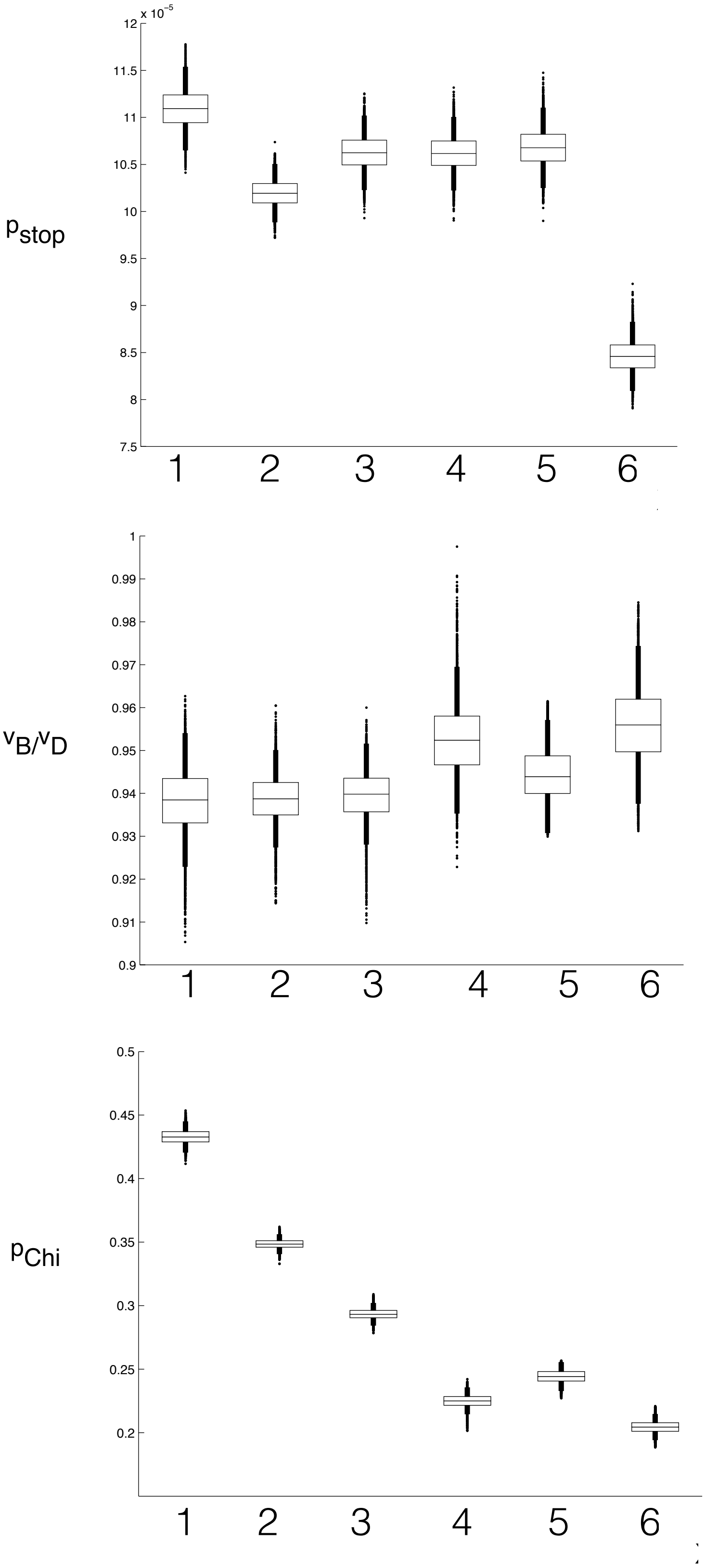

Figure 5. Boxplots describing the likelihood of the three model parameters p_{stop} (top), $p_{-} = v_B/v_D$ (middle), and p_{χ} (bottom) in the neighbourhood of the optimal values for each of the six strains (enumerated from 1 to 6 on the x-axis). The local distribution is obtained by a Markov Chain Monte Carlo exploration of the neighbourhood starting at the optimum. The boxplots centre lines show the medians; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots.

the signal in the data is strong enough to allow a correct estimation of v_B/v_D or whether the model would fit the data equally well if v_B/v_D was simply fixed to 1. In that case, the actual v_B/v_D may still be different from 1 as observed *in vitro*, but this would suggest that the data do not allow its correct estimation. To compare the performance of the basic model when fixing v_B/v_D to 1 or estimating it from the data, we used a BIC (Bayesian Information Criterion [8], [6]) score. BIC takes into account the log-likelihood (L) of the data but penalises models with higher complexity (i. e. a larger number of independent parameters (q)) to a greater extent than conventional log-likelihood ratio tests. The BIC score is defined as follows:

$$BIC = -2L + 2q \log n \tag{10}$$

where n stands for the total number of observations $n = \sum_x n_x$. Note that we use the objective function defined in Eq.9 instead of the true log-likelihood L to calculate the BIC score in Eq.10. The objective function differs from the true log-likelihood by a function dependent on the data only ($\log C(\mathbf{n})$ in Eq.8) but not on the parameters of the model. This is a convenient strategy since the part of the log-likelihood function independent of the parameters is not necessary to compare the models. Table 1 shows the BIC scores of the models where v_B/v_D is either fixed to 1 or estimated. In all cases except the data set with an array of 6 Chi sites, the model where v_B/v_D is estimated from the data is strongly preferred, indicating that v_B/v_D is important to explain the data. In the case of the 6 Chi array, most of the signal is concentrated at the array and there is little signal away from the DSB. The estimation of v_B/v_D is directly dependent on τ_1, and τ_1 is linearly dependent on the distance from the DSB and is better estimated if there is enough signal away from the DSB. It is therefore not surprising that in that dataset v_B/v_D cannot be estimated reliably.

Table 1. BIC scores computed for both models under consideration and all 6 strains with different number of Chi sites

N Chi	BIC($v_B/v_D = 1$)	BIC($v_B/v_D < 1$, estimated from data)	Best Model
1	92145	92087	v_B/v_D estimated (very strong)
2	188849	188701	v_B/v_D estimated (very strong)
3	111294	111197	v_B/v_D estimated (very strong)
4	90378	90319	v_B/v_D estimated (very strong)
5	80722	80652	v_B/v_D estimated (very strong)
6	86471	86481	v_B/v_D fixed (strong)

3.5 Mixture model

So far, our model is assuming a constant immutable ratio between the velocities of the two motors in RecBCD. But, recent *in vitro* experiments [7] demonstrate that, in fact, RecBCD operates (at the single molecule level) with a bimodal distribution of velocities. It is tempting to investigate whether a mixture between two modes described by different sets of parameters would explain the data better. The new model can be described as follows:

$$Pr'(x \mid r, \mathcal{P}_1, \mathcal{P}_2) = r \cdot Pr(x \mid \mathcal{P}_1) + (1 - r) \cdot Pr(x \mid \mathcal{P}_2) \tag{11}$$

where r is the probability of choosing the first set of parameters $\mathcal{P}_1$ and $1 - r$ is the probability of choosing $\mathcal{P}_2$ accordingly.

To estimate this best set of parameters, we follow a simple strategy and sample the $[0, 1]^6$ interval in two rounds as follows:

Step 1:

- $[0, 1]$ with step size 10^{-1} for $p^1_- = v^1_B/v^1_D$,
- $[0, 1]$ with step size 10^{-1} for p^1_χ,
- $[0, 1]$ with step size 10^{-1} for $p^2_- = v^2_B/v^2_D$,
- $[0, 1]$ with step size 10^{-1} for p^2_χ,
- $[0.8 \times 10^{-4}, 1.44 \times 10^{-4}]$ with step size 4×10^{-6} for p_{stop}.
- $[0.5, 1]$ with step size 10^{-1} for r

Step 2:

- $[0.8, 1]$ with step size 2×10^{-2} for $p^1_- = v^1_B/v^1_D$,
- $[0.2, 0.4]$ with step size 2×10^{-2} for p^1_χ,
- $[0.4, 0.6]$ with step size 2×10^{-2} for $p^2_- = v^2_B/v^2_D$,
- $[0.7, 1]$ with step size 2×10^{-2} for p^2_χ,
- $[0.8 \times 10^{-4}, 1.2 \times 10^{-4}]$ with step size 4×10^{-6} for p_{stop}.
- $[0.5, 0.6]$ with step size 2×10^{-2} for r

We find that the optimal mixture is driven by $r = 54\%$ for the first set of parameters: $p^1_\chi = 0.26$, $v^1_B/v^1_D = 0.86$, and $1 - r = 46\%$ for the second one: $p^2_\chi = 0.86$, $v^2_B/v^2_D = 0.58$ and p_{stop}=1.04×10^{-4}.

Fig. 6 shows the induced split in the space of parameters. The first thing to notice is that this is a 'real' mixture, in the sense that the two modes are very distinct, and their respective weights are similar. In particular, the recognition probabilities become very different in both modes, and different from the initial model and the *in vitro* estimates [4, 11, 12].

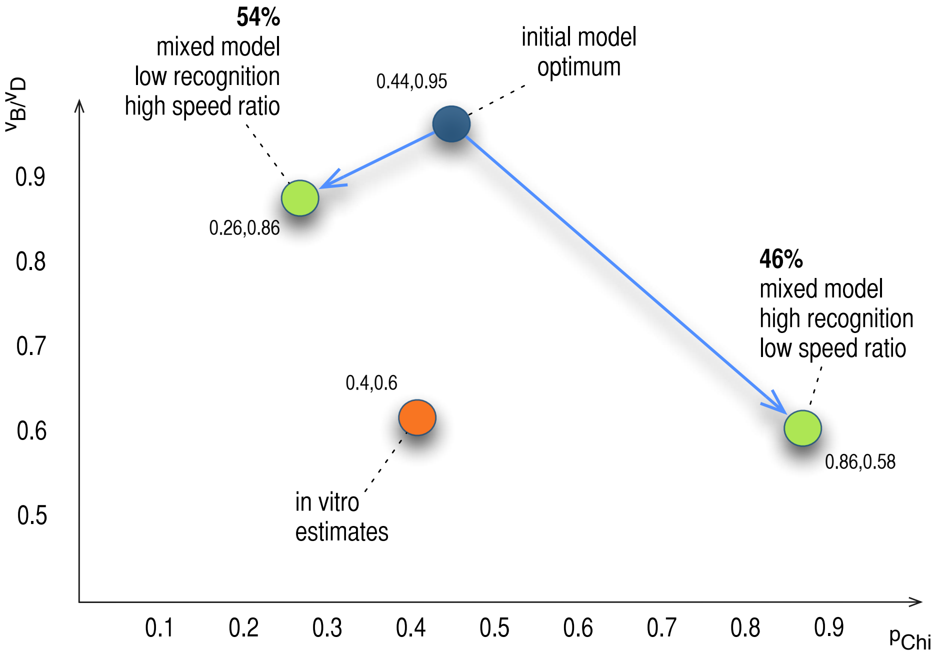

Figure 6. The two parameter sets in the mixture model compared to the optimal parameters of the initial model. Percentages indicate the probabilities of the low-recognition/high-ratio mode (46%), and of the high-recognition/low-ratio mode (54%).

Fig. 7 shows the marked improvement on fitting for the first peak (at the position of the first Chi). The improvement is noticeable both in the proximal region, where the high-recognition mode allows the model to fit better the initial peak (solid line), and compares well with the initial fit (dotted line); and, at the far end, where the low-recognition mode delineates the finer details of the data better as well (one sees the presence of the two Chi sites clearly in the prediction). This is confirmed by the BIC scores (see Table 2) that indicate a strong preference for the mixture model.

It is instructive to compare the predicted mean values of τ_1 for all four parameter sets (including the one coming from *in vitro* estimates [12]. If we compare these mean values at $60kb$ we get:

in vitro estimates: $60\frac{4}{6} \sim 40kb$

initial model: $60\frac{5}{95} \sim 3kb$

mixture model:
- *low recognition mode* $60\frac{14}{86} \sim 9kb$
- *high recognition mode* $60\frac{42}{58} \sim 43kb$

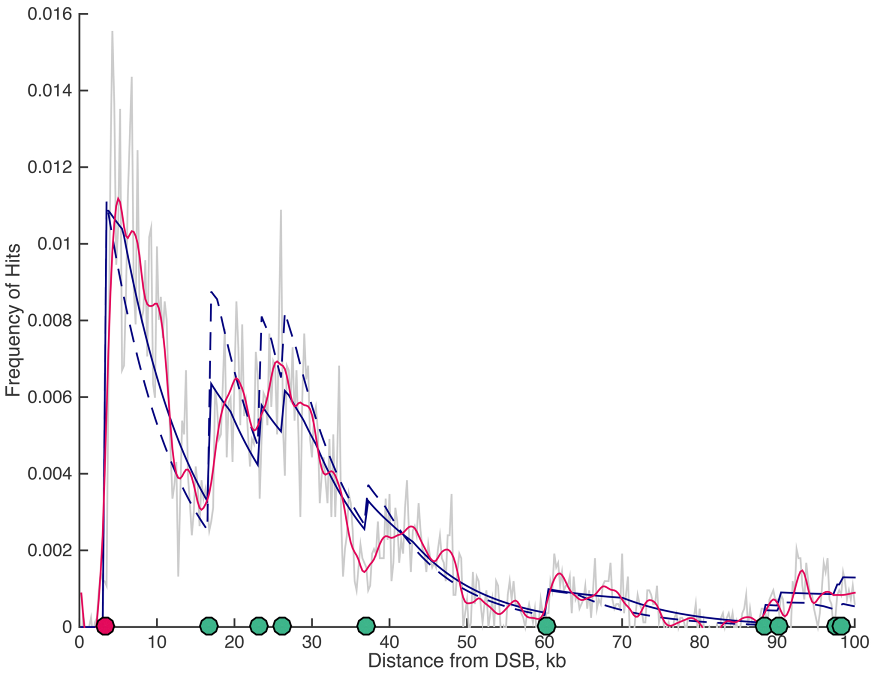

Figure 7. Comparison of the 1 Chi data set and the predictions of the optimal mixed model (solid line, $p_{stop} = 1.04 \times 10^{-4}, p^1_\chi = 0.26, v^1_B/v^1_D = 0.86, p^2_\chi = 0.86, v^2_B/v^2_D = 0.58, r = 54\%$), and the optimal basic one (dotted line, $p_{stop} = 1.12 \times 10^{-4}, p_\chi = 0.44, v^1_B/v_D = 0.95$). The Chi sites are depicted by green circles except for the position of the Chi array which is in red. The grey line shows the raw data binned into 250 bp bins. The red curve represent the smoothed data with a 'loess' filter (bandwith 5700, span 0.057).

At first, the prediction for the high recognition mode of the mixed model ($45kb$) seems improbably long. However, it is important to note that this model predicts a very high probability of Chi recognition. Given that Chi sites are present on average every $5kb$ on the chromosome [13], in this mode RecBCD would very rarely travel such a large distance before Chi recognition. One can calculate the mean τ_1 over all Chi sites, assuming a Chi site every $5kb$ and taking into account the probability of Chi recognition: both the high and low recognition modes of the mixed model have a mean τ_1 of the same order of magnitude ($5\frac{14}{86}\frac{100}{26} \sim 3.1kb$ and $5\frac{42}{58}\frac{100}{86} \sim 4.2kb$). This value is significantly higher than that predicted from the initial model ($5\frac{5}{95}\frac{100}{44} \sim 0.6kb$). It might be that a minimal loop size is important to ensure efficient loading of RecA at Chi which could be reflected in the predictions of the mixed model.

Table 2. Comparison of the mixture model and the basic model for the data set with 1 Chi site in the Chi-array. The BIC scores have been computed using Eq. 10

BIC basic model(v_B/v_D estimated)	BIC mixture model	Preferred model
92087	91735	mixture model (very strong)

As all molecular systems, the double-strand break repair system is faced with trade-offs. The density of Chi sites found on the chromosome together with the imperfect recognition thereof could be interpreted as a sign that recognition accuracy is traded off against some additional desirable properties. Such properties could be: speed of execution of the resection, optimisation of the length of the segment on which RecA will be loaded and the search for homology will be based [5], control of the variance of this length. Taking these new quantitative insights into account, and insofar as the model captures well the general features of the hit counts, and their dependency on the variations of the Chi distributions, one can use it as a quantitative tool in the investigation of the reasons for the genomic distribution of Chi sites [13]. Specifically, one can ask whether this distribution is judiciously adjusted to the generation of a resected single strand which optimises the performance of the RecA-based homology search and hence of the entire DSB repair process. The hypothetic single molecule *in vivo* bimodal behaviour, which our data-driven model suggests, would avail the cell with a larger palette of repair options, and thus should be integral to this investigation.

References

[1] Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[2] Keith C Cheng and Gerald R Smith. Cutting of Chi-like sequences by the RecBCD enzyme of Escherichia coli. *Journal of Molecular Biology*, 194(4):747–750, 1987.

[3] Mark S Dillingham and Stephen C Kowalczykowski. RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiology and Molecular Biology Reviews*, 72(4):642–671, 2008.

[4] Dan A Dixon and Stephen C Kowalczykowski. The recombination hotspot χ is a regulatory sequence that acts by attenuating the nuclease activity of the e. coli recbcd enzyme. *Cell*, 73(1):87–96, 1993.

[5] Anthony L Forget and Stephen C Kowalczykowski. Single-molecule imaging of dna pairing by reca reveals a three-dimensional homology search. *Nature*, 482(7385):423–427, 2012.

[6] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[7] Bian Liu, Ronald J Baskin, and Stephen C Kowalczykowski. DNA unwinding heterogeneity by recbcd results from static molecules able to equilibrate. *Nature*, 500(7463):482–485, 2013.

[8] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[9] Gerald R Smith. How recbcd enzyme and chi promote dna break repair and recombination: a molecular biologist's view. *Microbiology and Molecular Biology Reviews*, 76(2):217–228, 2012.

[10] Evgeny B Stukalin, Hubert Phillips III, and Anatoly B Kolomeisky. Coupling of two motor proteins: a new motor can move faster. *Physical review letters*, 94(23):238101, 2005.

[11] Andrew F Taylor and Gerald R Smith. Recbcd enzyme is altered upon cutting dna at a chi recombination hotspot. *Proceedings of the National Academy of Sciences*, 89(12):5226–5230, 1992.

[12] Andrew F Taylor and Gerald R Smith. RecBCD enzyme is a DNA helicase with fast and slow motors of opposite polarity. *Nature*, 423(6942):889–893, 2003.

[13] Fabrice Touzain, Marie-Agnès Petit, Sophie Schbath, and Meriem El Karoui. DNA motifs that sculpt the bacterial chromosome. *Nature Reviews Microbiology*, 9(1):15–26, 2010.