Stochastic Models for Replication Origin Spacings in Eukaryotic DNA Replication

Huw Day, N. C. Snaith

ARTICLE INFO

Keywords:
temporal point process
random matrix theory
Wigner’s surmise
Poisson point process
linear repulsion

ABSTRACT

We consider eukaryotic DNA replication and in particular the role of replication origins in this process. We focus on origins which are ‘active’ - that is, trigger themselves in the process before being read by the replication forks of other origins.

We initially consider the spacings of these active replication origins in comparison to certain probability distributions of spacings taken from random matrix theory. We see how the spacings between neighbouring eigenvalues from certain collections of random matrices have some potential for modelling the spacing between active origins. This suitability can be further augmented with the use of uniform thinning which acts as a continuous deformation between correlated eigenvalue spacings and exponential (Poissonian) spacings.

We model the process as a modified 2D Poisson process with an added exclusion rule to identify active points based on their position on the chromosome and trigger time relative to other origins. We see how this can be reduced to a stochastic geometry problem and show analytically that two active origins are unlikely to be close together, regardless of how many non-active points are between them. In particular, we see how these active origins repel linearly.

We then see how data from various DNA datasets match with simulations from our model. We see that whilst there is variety in the DNA data, comparing the data with the model provides insight into the replication origin distribution of various organisms.

1. Introduction

DNA is the genetic blueprint which allows for the creation of healthy and functioning cells. It contains all the information necessary for cells to perform various functions that are essential for life. DNA is stored as chromosomes. In eukaryotes (this is most types of organisms, including plants, fungi and animals), these chromosomes are linear in structure. In contrast, prokaryotes (bacteria and some single celled organisms) typically store DNA in circular chromosomes. DNA contains genes which provide the information for individual building blocks of the cell.

If we consider the cell as a factory (a producer of goods for the organism) and genes are the instructions about how to carry out various procedures, chromosomes are the instruction booklets. DNA replication is a crucial part of cell replication, which occurs in growth and repair of any organism. A key component in copying the genetic information is the structure of the chromosome.

Considering the linear structure of chromosomes in eukaryotes, we can compare replicating this information to how a photocopier copies a piece of paper. In a photocopier, the scanner reads the paper from top to bottom. On a linear chromosome, reading is done differently. Instead of going top to bottom, copying occurs at multiple points along the chromosome. These points are called replication origins. In each replication event, only some of these replication origins are licensed to trigger but some are not. When replication begins, the licensed origins will each trigger after some point in time. When an origin triggers, it sends out a replication fork, spreading out in both directions along the chromosome. This replication fork is analogous to the scanner in a photocopier, it reads and duplicates the genetic information it passes over. If a replication fork passes over a licensed replication origin before it triggers, we say the origin will become inactive (or passive) and will not trigger in this cycle of replication. We call licensed origins which trigger before they are made inactive, ‘active’ origins. Our interest will solely be on how these active origins are distributed, as seen in Figure 1.

An issue with replication forks however, is that they are unreliable. This means there is a chance that either end of the fork may stall irreversibly in place. If one replication fork stalls, this is not a major problem, since a replication fork approaching it from the other direction will still read the rest of the genetic information between them. The issue comes when two replication forks approaching each other both stall before colliding, leaving a section of unread genetic information in between them. We refer to these as double stall events.

If a double stall occurs when reading a chromosome, then the replicated DNA will be incomplete [1]. This can lead to mutations and in some cases genetic diseases. There are quality checkpoints and repair mechanisms in place to mitigate the damage from double stall events. It is also important to note that double stall events are not the only impediment to successful replication.

Due to natural selection, the distribution of replication origins on chromosomes is in some ways as efficient as it could be given external constraints. It is an interesting mathematical problem to study the statistics of these replication origins in various organisms. Additionally, better understanding these statistics could provide a better understanding for creating new treatments to treat genetic diseases [2].
Stochastic Models for Replication Origin Spacings in Eukaryotic DNA Replication

2. Random Matrix Theory, Exponential Distributions and Uniform Thinning

2.1. Point Processes

Although we include the precise definitions below, for this work it is sufficient to think of a point process as a collection of points scattered on a line, in a rectangle, or a more general space, according to some rule that controls their density and positioning.

Of the commonly studied point processes in probability theory, the archetypal one is the Poisson point process. The following definitions are taken from [3] pages 301 and 303 respectively.

Definition 1. (Point Process)

Let \( N \) be a function defined on a state space \( E \). Let \( \{X_n : n > 0\} \) be a series of random points in \( E \). For a reasonable subset \( A \subset E \), \( N_A \) is a random variable that counts the number of points \( \{X_n : n > 0\} \) in \( A \). More precisely:

\[
N_A = \sum_{n>0} I_{\{X_n \in A\}}
\]

where \( I_{\{\text{Event}\}} \) denotes the indicator function, which is equal to 1 if the event occurs and 0 otherwise.

Definition 2. (1D Poisson point process)

Let \( N \) be a point process on a state space \( E \subset \mathbb{R} \). \( N \) is a Poisson point process of intensity \( \lambda > 0 \) if:

1. For \( A \subset E \) we have that \( N_A \sim \text{Poisson}(|A|\lambda) \). More precisely, the probability that an interval (or collection of intervals) \( A \) of total length \( |A| \) contains \( k \) points is:

\[
P[N_A = k] = \begin{cases} 
    e^{-|A|\lambda} |A|^k k! & k = 0, 1, 2, \ldots \\
    0 & \text{otherwise}
\end{cases}
\]

2. If \( A_1, \ldots, A_k \) are disjoint subsets of \( E \), then \( N(A_1), \ldots, N(A_k) \) are independent random variables.

The Poisson point process produces points which have independent, uncorrelated spacings (the distances between neighbouring points). A point’s position is not influenced in any way by its neighbours - as if it didn’t ‘know’ where its neighbours are. In the case of a 1D Poisson point process, these spacings are exponential random variables with probability density function \( \lambda e^{-\lambda s} \) for \( s > 0 \) and 0 otherwise, where \( \lambda > 0 \) is the same parameter seen in Definition 2.

We can model replication origins on chromosomes as points on a line and then use some ideas from point processes to study them. In particular, we can generate a histogram of the spacings between adjacent points which we can then compare to probability density functions.

Considering Figure 2 (which is Figure 3A from [1]), we note that the spacings between \( S. \text{cerevisiae} \) origins appear to exhibit some sort of local repulsion (two origins are unlikely to be close together, which is made evident by the blue histogram being lower close to an inter-origin distance of 0) as well as a global confinement (two origins are also unlikely to be very far apart, which is made evident by the decrease of the histogram as the inter-origin distance gets larger, past 100). This is an example of correlation between points (the points notice one another), so we see that perhaps a Poisson process/exponential spacings (represented by the red line) is not the best model to use.

2.2. Wigner’s Surmise

Of particular interest to us is the idea of using the statistics of eigenvalues of certain randomly sampled matrices to model this system, as eigenvalues of random matrices tend to display local repulsion and a deficit of very large spacings compared to a Poisson point process. Random matrix statistics are more traditionally associated with modelling systems in various branches of physics which are outlined in great detail in [4], [5] and [6]. Random matrices also have applications in material science [7], signal detecting [8], [9], [10] and [11], wireless communication [12] and finance [13]. Many of these systems have certain matrices and eigenvalues associated with them, making it relatively intuitive to attempt to model them with random matrices. However, there is precedent for random matrix statistics being used to model physical systems where there is no such inherent matrix present. Perhaps one of the better known models is the Buses of Cuernavaca system [14], where the
times between passing buses were shown to display the same statistical behaviour as spaces between eigenvalues of random matrices. In that case there is not only statistical evidence, but also analytical calculations on a stochastic model that simulates the process in question.

An ensemble of random matrices that is a plausible model for replication origins is the Circular Orthogonal Ensemble of symmetric unitary matrices (COE) [5]. The eigenvalues of these matrices are points on a line (more precisely, on a circle that can be cut and uncurled to form a line segment) and the histogram of the spacings between these consecutive eigenvalues, when averaged over many COE matrices, is approximated by a function known as Wigner’s surmise.

We reach the probability density function for the spacing between eigenvalues. We re-scale our density so that the mean spacing is 1. Re-scaling is a common technique in random matrix theory in order to compare statistics between different datasets. In this paper, whenever we re-scale a dataset we will be re-scaling it so that the mean of the dataset is 1. Thus Wigner’s surmise is an approximation of the probability density function for the spacing between re-scaled eigenvalues of COE matrices for $s > 0$:

$$p(s) = \frac{\pi s}{2} \exp \left( -\frac{\pi s^2}{4} \right). \quad (1)$$

We see from this formula (as illustrated in Figure 3) that for small spacings ($s \sim 0$), the probability density function is small and so we would expect to not see two eigenvalues being close together since $p(s) \to 0$ as $s \to 0$. In particular, we see that $p(s)$ behaves like a linear function close to 0, an effect we refer to as ‘(local) linear repulsion’. Additionally we see that as $s \to \infty$, $p(s) \to 0$, faster than an exponential. We see that a point process with spacings from Wigner’s surmise would have points which are characterised by local repulsion but global confinement. This is in line with what we see in Figure 2.

For the DNA data, the datasets have several chromosomes per organism, so for each chromosome we calculate the replication origin spacings separately before collating all the spacings into one sample. Replication origins are not single points, they have length, but for our purposes, as is common in the literature ([1], [15]), we take the midpoint of the start and end of each origin (measured in base pairs) as their ‘position’. (We will explore in Section 4.2 the drawbacks of this assumption).

Equipped with a list of spacings between replication origins from the DNA data, we re-scale and normalise. We re-scale as described above whilst discussing Wigner’s surmise, making the mean spacing of the data equal to 1. We also normalise the spacing data.

Specifically, we say that we normalise a histogram/frequency diagram when we make the area under the plot equal to 1: we divide by the sum of all the heights of the bars.
the bin width. This has the effect of turning a histogram into something comparable to a probability density function plot. This gives us a common reference frame to compare all of our frequency plots.

We can produce cumulative frequency plots with mean spacing $1$ and ‘total’ $1$ so that our plot is comparable to a cumulative density function. We then plot this against the cumulative form of Wigner’s surmise, as seen in Figure 3.

In Figures 4, 5 and 6 we consider a re-scaled Wigner’s surmise, an exponential distribution of parameter $1$ (which will have mean value $1$) and spacings between replication origins on chromosomes of certain organisms, re-scaled so that the mean spacing is $1$. This allows for a comparison of the shape of the distributions; in Figures 4, 5 and 6, we see the replication origin spacings in blue, the spacings from a 1D Poisson process in green and Wigner’s surmise spacings in orange. We use the cumulative frequency because it gives a smoother curve to work with when the datasets we are working with are small.

K. lactis replication origin data is taken from [16], L. waltii replication origin data is taken from [17] and S. cerevisiae replication origin data is taken from [18].

We see that Wigner’s surmise has some utility in modelling K. lactis and L. waltii but due to the smaller sample size, it is difficult to draw any meaningful comparisons between the two DNA datasets. It is plausible that both could be modelled with a Wigner’s surmise.

Neither Wigner’s surmise nor the exponential appear to fit S. cerevisiae particularly well (they have RMS errors of 0.05307 and 0.0375 for Wigner and exponential respectively; for more detail on RMS error, see the next section). However, some sort of interpolation or continuous deformation between Wigner’s surmise and the exponential curve might be suitable. The next section leads us to one method of doing this.

2.3. Uniform thinning of COE eigenvalues

Recall that to compare to Wigner’s surmise, we generate COE random matrices and then calculate their eigenvalues, and then the corresponding spacings between consecutive eigenvalues.

The process of thinning occurs when we have our list of eigenvalues, say $\{\lambda_1, \lambda_2, \ldots, \lambda_{200}\}$. We define some thinning parameter $p \in (0, 1]$ and then remove each eigenvalue with probability $1 - p$. We can think of this process as having a biased coin toss with probability of heads $p$. We toss the coin for each eigenvalue $\lambda_i$. If the coin lands heads, the eigenvalue stays in our sample. If the coin lands tails, we remove that eigenvalue from our sample.

For a given value of $p$, we thin our eigenvalue sample. This has an interesting effect on the statistics.

Clearly if our thinning parameter $p = 1$ then we will not eliminate any of our sample and end up with our original...
Wigner’s surmise. As \( p \) gets smaller, we begin eliminating elements of sample with an increasing likelihood.

Two adjacent eigenvalues (say \( \lambda_j \) and \( \lambda_{j+1} \)) from our random matrix sample are correlated - we know they repel linearly but are also likely to be somewhat confined. If we eliminate an eigenvalue, then the two eigenvalues either side are now less correlated than they were before (if we eliminate \( \lambda_{j+1} \) then \( \lambda_j \) and \( \lambda_{j+2} \) become neighbours and their correlation will be weaker than if they were adjacent in the original unthinned sample).

As \( p \) approaches 0, the likelihood of any given eigenvalue being eliminated increases. In practical terms, our thinned sample starts to have increasingly large gaps between eigenvalues. Our thinned sample might look like: \{\lambda_1, \lambda_30, \lambda_64, \ldots\}. The bigger the gap between adjacent eigenvalues, the smaller the correlation. If we set \( p = 0 \) then we eliminate our entire sample and get no statistics. However, in the limit as \( p \) approaches 0, the spacing of our thinned sample are exponential/Poissonian statistics. This is exactly because the exponential points are not correlated.

Computationally, this is quite expensive. For every 200 x 200 random matrix, we have 200 eigenvalues. We then have to toss a coin with some bias \( p \) for each of those eigenvalues. That’s for just a single matrix in the sample. Data from millions of matrices are combined to generate a smooth approximation to thinned COE eigenvalue spacings.

To circumvent this computationally expensive method, we relied on Bournemann’s work from [19] and [20] involving Fredholm determinants.

Bournemann’s code was particularly useful to allow us to generate the curve of the probability density function for the spacings between thinned COE eigenvalues, which can be expressed in terms of Fredholm determinants.

In practical terms, this means that by varying our value of \( p \) we can continuously deform our Wigner’s surmise into an exponential distribution. We can use this deformation to best fit our thinned plot to our DNA data by finding the optimal value of \( p \).

To find the optimal value of \( p \), we sought to minimise the Root Mean Square error (RMS) between the data points and the respective model (e.g. COE eigenvalues with some uniform thinning applied). Fix the number of bins \( n \). For both the data and the COE eigenvalues with some thinning factor \( p \), we have their respective bin heights \( \{d_1, \ldots, d_n\} \) and \( \{w_1, \ldots, w_n\} \) respectively. We can calculate the RMS error between the two datasets by considering the squared difference between each data value and the model value, summing them, scaling this by the number of data points and then square rooting:

\[
\text{RMS Error} := \sqrt{\frac{1}{n} \sum_{i=1}^{n} (d_i - w_i)^2}.
\]

The RMS error will always be non-negative, and the closer the value is to 0, the smaller the error and the better the fit between our two distributions. In particular, the RMS error of a collection of points with itself will be 0.

For K. lactis in Figure 4, it is optimal to apply no thinning at all. It is interesting that COE eigenvalue spacings appear to be a significantly better fit for this dataset than the Poisson process spacings because it indicates a high level of correlation between the positions of the replication origins.

For L. waltii in Figure 7, we saw some slight systematic deviation from Wigner’s surmise (RMS from unthinned COE eigenvalue spacings is 0.0143) and found the optimal thinning factor is \( p = 0.93 \) (which yields a RMS error of 0.0111).

In the case of S. cerevisiae in Figure 8, there was quite significant deviation from Wigner’s surmise (RMS from unthinned COE eigenvalue spacings is 0.0530). The optimal thinning value is \( p = 0.28 \) (which yields a RMS error of 0.0093). This signifies quite a lot of deformation away from the original Wigner’s surmise. The spacings start to look more like Poissonian spacings.

We see that with the correct thinning factor, we can fit a curve to some of the datasets. However when we get to human DNA as seen in Figure 9 (data is taken from [21]), we see that uniform thinning is not sufficient. This motivated us to consider a different model.

3. 2D Poisson Point Model

3.1. Non-Uniform Exclusion Process

In this section we consider a stochastic model that seeks to be a direct analogue of the process we are modelling. We consider each ‘point’ (analogous to a replication origin) as having a position on the chromosome as well as a random length of time before it triggers and starts replication. Using
Stochastic Models for Replication Origin Spacings in Eukaryotic DNA Replication

Figure 8: Cumulative histogram plots comparing Poissonian spacings/exponential distribution of mean spacing 1, thinned \((p = 0.28)\) COE eigenvalues (labelled thinned COE) and re-scaled and normalised replication origin spacings from S. cerevisiae. RMS error between S. cerevisiae and thinned COE eigenvalue spacings 0.0093. RMS error between S. cerevisiae and exponential 0.0375. Spacing data taken from [18].

Figure 9: Cumulative histogram plots comparing Poissonian spacings/exponential distribution of mean spacing 1, an un-thinned \((p = 1)\) COE eigenvalues (labelled thinned COE) and re-scaled and normalised replication origin spacings from Human MCF7. RMS error between Human MCF7 and unthinned COE eigenvalue spacings 0.2588. RMS error between Human MCF7 and exponential 0.1957. Spacing data taken from [21].

these two pieces of information for each point, we perform some non-uniform exclusion process to select so-called ‘active’ points. Unlike Newman et al.’s work [1] we do not consider the role that stalling has on our model.

Begin by fixing the length of our chromosome to be some parameter \(L > 0\). The origins will all be located in \([0, L]\). We include a second axis to consider the trigger times of the origins. Whilst there are clearly limitations on these trigger times (the lifetime of a cell, for example), for our calculations, we can say that the trigger times could theoretically be anywhere in \([0, \infty)\).

We generate a 2D Poisson point process of parameter 1 on this space \([0, L] \times [0, \infty)\). This means that given a suitably nice set \(A \subset [0, L] \times [0, \infty)\) with area \(|A|\), we understand that:

\[ N_A := \text{Number of Poisson points in } A \sim \text{Poisson}(|A|). \]

(We use the notation \(\sim\) to say that two random variables behave with the same probability distribution).

Explicitly the probability of there being \(k\) points in the box \(A\) with area \(|A|\) is described as:

\[ \text{Prob}(N_A = k) = \frac{|A|^k e^{-|A|}}{k!} \text{ for } k = 0, 1 \ldots \] (2)

Once we have a number of points in \(A\), these points are scattered randomly but uniformly - they are no more likely to be in one part of \(A\) than any another. Furthermore, we have that given two disjoint subsets of our domain \(A\) and \(B\) (that is, two areas that do not overlap), the random variables \(N_A\) and \(N_B\) are independent. This uniquely defines our Poisson point process.

We end up with a collection of points in our domain. Now we seek to identify the so-called ‘active’ points and eliminate the non-active/passive points that get replicated by a replication fork emanating from a neighbouring origin. Denote a given Poisson point \((x_i, t_i)\) where the first entry indicates its position on the chromosome and the second indicates its trigger time.

We assume the speed of the replication fork emerging from each replication origin is constant and the same for all replication origins. (This is not an unreasonable assumption, see [22] for more on this). We will take this value to be 1. (Whilst more general calculations can be performed with general speed \(v\), we get all the information we need assuming this parameter to be 1). This means that the blue lines in Figures 10 and 11 represent the progress of a replication fork along the chromosome over time.

The following definition used in [23] but based on vocabulary from special relativity will prove useful in describing aspects of our model:

**Definition 3.** (Forward and Backward Light Cones)

- The forward light cone of a point \((x, t)\) is the set of points \(\{(x', t') : |x - x'| \leq t' - t\}\). The forward light cones track the progress of replication forks through space and time.

- The backward light cone of a point \((x, t)\) is the set of points \(\{(x', t') : |x - x'| \leq t - t'\}\).
Figure 10: A basic realisation of the 2D Poisson process with 3 points (marked by purple dots). Backward light cones are marked in red, forward light cones are marked in blue. In this instance we see that the highest point \((x_1, t_1)\) contains the other 2 points in its backwards cone, which means it would be made passive by \((x_2, t_2)\) without the presence of \((x_3, t_3)\). Similarly, \((x_2, t_2)\) is made passive by the lowest point. We can say that the lowest point \((x_3, t_3)\) is active and in this case makes the other 2 points passive.

If we consider this problem geometrically, we see that the isosceles triangle underneath each origin \((x_i, t_i)\) formed by two lines emitting from the point, of slope \(+1\) and \(-1\) respectively is its backward facing light cone as described in [23]). This triangle will have vertices \((x_i - t_i, 0)\), \((x_i + t_i, 0)\) and \((x_i, t_i)\). Upon closer inspection, it becomes apparent that any origin located in this triangle will make the point at the top of the triangle passive. The point \((x_j, t_j)\) is active if and only if its backward facing light cone is empty of points.

Additionally, we see that if you draw the upward triangle from each point (a natural upward extension of the vertices passing through \((x_i, t_i)\)), we obtain its forward facing light cone. The forward light cone marks the progress of the replication forks through space and time. We see that any point in this forward cone is made passive by our original point.

3.2. Nearest Neighbour Spacing Calculation

The nearest neighbour spacing is the probability density function of the difference between adjacent points. Given a sample of ordered points (for example, eigenvalues) \(\{\lambda_1, \ldots, \lambda_n\}\) (with \(\lambda_j \leq \lambda_{j+1}, \forall j\)), we define the function \(p(s)\) for \(s > 0\) as the probability density of the nearest neighbour spacing of two eigenvalues. More precisely \(p(s)\) gives the probability density of the event that there is a gap of exactly \(s\) between any two points in our sample, say \(\lambda_{k+1} - \lambda_k\), with no points lying between them. The nearest neighbour spacing probability density says nothing about which pair of points the gap is between or where the gap occurs: for example it cannot tell us whether a large gap between \(\lambda_{k+1}\) and \(\lambda_k\) is followed by a small gap between \(\lambda_{k+2}\) and \(\lambda_{k+1}\).

For eigenvalues from a COE random matrix sample, Wigner’s surmise (eq. (1)), is a good approximation of the nearest neighbour spacing. It is of interest to perform a calculation for the nearest neighbour spacing between adjacent active points in our model.

When considering our stochastic model and which points are active and which are passive, it suffices to understand when certain areas of certain triangles are empty.

If we consider a fixed number of active points, then we can start understanding the distribution function of their positions and trigger times.

Consider two active points, denote them \((x_1, t_1)\) and \((x_2, t_2)\). For the purposes of our calculation, they will be ordered in terms of positions with \(x_1 \leq x_2\). There will be no ordering restriction on their trigger times. We ask what restrictions must be on the rest of the Poisson points for these two to be the only 2 active points.

In Figure 11, we see this setup. In order for the two chosen points to be active, their backward cones (red triangles) must be empty. Additionally, if another point was in the space between their backward and forward cones (either in the quadrilateral in the centre of Figure 11 or the shapes boxed in by the cones and the axis either side of our points), then such a point would also be active. We want to fix exactly 2 active points.

Consider placing a point somewhere in the plane in Figure 11. Now examine the lines immediately around this point:

- If all the lines are the red (from backward facing light cones) and black (the axes) then the point makes one of our chosen active points passive. This is not permitted.
• If all the lines are blue (from upward facing light cones) and black then the point is made passive by one of the active points. This is a permitted point.

• Finally, if the lines are a combination of red, blue and black then the point does not affect the fixed active points, but it is active itself, which means there are now 3 active points in the system. For our purposes, this is not permitted.

For these to be the only 2 active points, we need this mountain range shape (shaded in grey) area to be empty of points. We know from the definition of a Poisson process and eq. (2), that if we have a reasonable shape $A$ of area $|A|$ then the probability that that area has no points in it is just:

$$
\text{Prob}(A \text{ is empty}) = \text{Prob}(N_A = 0) = e^{-|A|}.
$$

So this just becomes a task of calculating the area of such a shape. We use an inclusion exclusion argument which will allow us to account for and deal with the case that the backward cones of our two active points intersect.

We calculate the area of the 3 big, blue-peaked triangles:

- The left most right angle triangle with vertices: $(0, x_1 + t_1), (x_1 + t_1, 0), (0, 0)$ which has area: $(x_1 + t_1)^2$.

- The blue-peaked triangle with vertices: $(x_1 - t_1, 0), \frac{1}{2}(t_2 + x_2^ - x_1 + t_1), (x_2 + t_2, 0)$ which has area $\frac{1}{2}(x_2 + t_2 - x_1 + t_1)^2$.

- The right most right angle triangle with vertices: $(x_2 - t_2, 0), (L, 0), (0, L + t_2 - x_2)$ which has area: $(L - x_2 + t_2)^2$.

We then want to take away the area of the two smaller red triangles under our active points because they have been counted twice. These have areas $t_1^2$ and $t_2^2$ respectively.

Note that we have avoided any problem with intersections of the smaller, red triangles with this inclusion exclusion method, counting the entire required area exactly once by adding the overlap area three times and subtracting it twice.

So we see that the total area we want empty is:

$$(x_1 + t_1)^2 + (L - x_2 + t_2)^2 + \frac{1}{4}(x_2 + t_2 - x_1 + t_1)^2 - t_1^2 - t_2^2.$$

Our goal is to integrate out both time variables in a way that keeps the 2 points active; it helps to write $s := x_2 - x_1$ for the distance between the active points. We can see that the middle triangle has the term $x_2 - x_1$ in its structure, but the edge triangles do not. We can circumvent this problem by considering our rectangular domain instead as cylindrical, setting $L = 2\pi$. We then just say that any cone that reaches a boundary on the left or the right, continues on the opposite boundary at the same height in the same direction.

From here it becomes reasonably simple to generalise our empty area, but noting that the new empty area will have the structure of two of our blue-peaked triangles.

In the cylindrical setting with two active points, $L = 2\pi$ and $s := x_2 - x_1$, the area we want empty is:

$$
\text{Area} = \frac{1}{4}(s + t_2 + t_1)^2 + \frac{1}{4}(2\pi - s + t_2 + t_1)^2 - t_1^2 - t_2^2.
$$

We write:

$$
p_2(s, t_1, t_2) := \exp(-\text{Area}).
$$

This density function represents the probability of having exactly two active points in our system, with trigger times $t_1$ and $t_2$ respectively and a distance $s$ between them.

We want to integrate out both time variables, but we need to be careful to ensure both of our points remain active. The height of the small, red triangle under our active points because they have been counted twice. These have areas $t_1^2$ and $t_2^2$ respectively.

We also want this to be true for the wraparound probability (replacing $s \rightarrow 2\pi - s$). Namely: $|t_1 - t_2| < 2\pi - s$. Otherwise, the trigger times can be infinitely big, provided they remain less than a distance $s$ from each other and positive. These bounds change for $n = 2$ precisely at $s = \pi$:

$$
-s > 2\pi \iff s < 2\pi - s \iff s < \pi.
$$

The final restriction is that the trigger times $t_1, t_2 \geq 0$. This causes some interaction with the lower bound. We see that with $s < \pi$ and $t_1 < s \iff t_1 - s < 0$, we have 0 as the dominant lower bound. We have two cases of bounds while $s < \pi$:

$$0 < t_2 < t_1 + s \text{ and } 0 < t_1 < s \text{ or } t_1 - s < t_2 < t_1 + s \text{ and } s < t_1 < \infty.
$$

We denote the nearest neighbour spacing function $P(s)$.

Because of the symmetry in the setup, we can understand intuitively that the behaviour of $P(s)$ as $s$ approaches 0 from above should be the same as when $s$ approaches $2\pi$ from below. For $s < \pi$:

$$
P(s) := \int_{-s}^{0} \int_{t_1 - s}^{t_1 + s} p_2(s, t_1, t_2) dt_2 dt_1 + \int_{0}^{s} \int_{0}^{t_1 + s} p_2(s, t_1, t_2) dt_2 dt_1.
$$

For $s > \pi$, we simply replace $s \rightarrow 2\pi - s$, noting that $p_2(s, t_1, t_2)$ is unchanged by this transformation:

$$
P(s) := \int_{2\pi - s}^{\infty} \int_{t_1 - (2\pi - s)}^{t_1 + (2\pi - s)} p_2(s, t_1, t_2) dt_2 dt_1
$$
functions respectively:

\[ \text{erf} (z) := \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt. \]

\[ \text{erfi} (z) := \frac{2}{\sqrt{\pi}} \int_0^z e^{t^2} dt. \]

We will use the following facts without proof. Below we use the standard notation of \( i := \sqrt{-1} \) for the imaginary unit:

- \( \text{erf} (-z) = -\text{erf} (z). \)
- \( \text{erfi} (z) = -i \text{erf} (iz). \)
- \( \text{erf} (0) = 0, \text{erf} (+\infty) = 1. \)

We proceed with the first integral in the \( s < \pi \) case. We do this in steps. The first integral looks like:

\[
\int_{t_1-s}^{t_1+s} p_2(s,t_1,t_2) dt_2 = \frac{\sqrt{\pi}}{2} \exp \left( -\frac{(t_1 + s)^2}{4} - \frac{(t_1 + 2\pi - s)^2}{4} + t_1^2 - \frac{(t_1 + \pi)^2}{2} \right) \\
\times \text{erfi} \left( \frac{\sqrt{2}}{2} (t_2 + t_1 + \pi) \right) _{t_2 = t_1 + s}^{t_2 = t_1 - s} \\
= \frac{\sqrt{\pi}}{2} \exp \left( -\frac{(t_1 + s)^2}{4} - \frac{(t_1 + 2\pi - s)^2}{4} + t_1^2 - \frac{(t_1 + \pi)^2}{2} \right) \\
\times \text{erfi} \left( \frac{\sqrt{2}}{2} (s + \pi) \right) - \text{erfi} \left( \frac{\sqrt{2}}{2} (-s + \pi) \right). \\
\]

Next we integrate with respect to \( t_1 \) between \( t_1 = s \) and \( t_1 = \infty \) to get:

\[
\frac{\sqrt{2}}{4\sqrt{\pi}} \exp \left( -\frac{1}{2} s^2 + \pi s - \frac{3}{2} \pi^2 \right) e^{-2\pi t_1} \\
\times \left[ \text{erfi} \left( \frac{\pi - s}{\sqrt{2}} \right) - \text{erfi} \left( \frac{s + \pi}{\sqrt{2}} \right) \right] _{t_1 = \infty}^{t_1 = \infty} \\
= \frac{\sqrt{2}}{4\sqrt{\pi}} \exp \left( -\frac{1}{2} s^2 + \pi s - \frac{3}{2} \pi^2 \right) \\
\times \left[ \text{erfi} \left( \frac{\pi - s}{\sqrt{2}} \right) - \text{erfi} \left( \frac{s + \pi}{\sqrt{2}} \right) \right] _{t_1 = s}^{t_1 = \infty} \\
= \frac{\sqrt{2}}{4\sqrt{\pi}} \exp \left( -\frac{1}{2} s^2 + \pi s - \frac{3}{2} \pi^2 \right) \\
\times \left( \text{erfi} \left( \frac{s + \pi}{\sqrt{2}} \right) - \text{erfi} \left( \frac{\pi - s}{\sqrt{2}} \right) \right). \\
\]

Considering the second integral in the \( s < \pi \) case we first have:

\[
\int_0^{t_1+s} p_2(s,t_1,t_2) dt_2 \\
= \sqrt{\frac{\pi}{2}} \exp \left( -\frac{1}{2} s^2 - \frac{3}{2} \pi^2 + \pi s - 2t_1 \pi \right) \\
\times \left[ \text{erfi} \left( \frac{t_1 + \pi}{\sqrt{2}} \right) - \text{erfi} \left( \frac{\pi - s}{\sqrt{2}} \right) \right]. \\
\]

Next integrating with respect to \( t_1 \) between 0 and \( s \) we get:

\[
\frac{\sqrt{2}}{4\sqrt{\pi}} \exp \left( -\frac{1}{2} s^2 + \pi s - \frac{3}{2} \pi^2 \right) \\
\times \left[ 2e^{2\pi s} \text{erfi} \left( \frac{\pi}{\sqrt{2}} \right) - 2e^{2\pi s} \text{erfi} \left( \frac{\pi - s}{\sqrt{2}} \right) \right] \\
+ \text{erfi} \left( \frac{s + \pi}{\sqrt{2}} \right) - \text{erfi} \left( \frac{\pi - s}{\sqrt{2}} \right). \\
\]

Now we have to add these two cases together. Note that the two error functions from the first case cancel from the second case, so we are left with:

\[
P(s) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} s^2 + 3\pi s - \frac{3}{2} \pi^2 \right) \\
\times \left[ \text{erfi} \left( \frac{s + \pi}{\sqrt{2}} \right) - \text{erfi} \left( \frac{\pi - s}{\sqrt{2}} \right) \right]. \\
\]

When we complete the square for the polynomial in the exponential we get:

\[-\frac{1}{2} s^2 + 3\pi s - \frac{3}{2} \pi^2 = -\frac{1}{2} (s - \pi)^2 - \pi^2\]

which gives us the symmetrical structure around \( s = \pi \) which we wanted.

Expanding the error function in \( s \) around \( \frac{\pi}{\sqrt{2}} \):

\[
\text{erfi} \left( \frac{\pi - s}{\sqrt{2}} \right) = \\
\text{erfi} \left( \frac{s + \pi}{\sqrt{2}} \right) + O(s^2) \\
= \frac{\pi}{\sqrt{2}} \exp \left( -\frac{\pi^2}{2} \right) s + O(s^2). \\
\]
So we see the error function term in the bracket reduces to:

$$\text{erfi}\left(\frac{s}{\sqrt{2}}\right) - \text{erfi}\left(\frac{s - N}{\sqrt{2}}\right) = \exp\left(\frac{s^2}{2}\right) \frac{s}{\sqrt{2}} + O(s^2).$$

And so for small $s$ (so that the exponential factor in terms of $s$ goes to 1) we see linear repulsion:

$$P(s) = \frac{e^{s^2/2}}{2\sqrt{\pi}} s + O(s^2).$$

As noted above, we can see that because of the symmetrical nature of $P(s)$, we have exactly that $P(s) = P(2\pi - s)$. Intuitively, this is a consequence of the cylindrical setup we have, as two points approaching each other on one side (e.g. the front of the cylinder as you look at it) means that their distance is approaching $2\pi$ on the other side (e.g. around the back of the cylinder). More thoroughly, we see that $p_0(s, t_1, t_2) = p_0(2\pi - s, t_1, t_2)$ and that the arguments about the sizes of $t_1$ and $t_2$ are identical in terms of $s$ and $2\pi - s$.

4. Spacing Simulations

4.1. Comparing DNA Data with Simulations

We see from the previous section, calculating anything like a nearest neighbour spacing or a joint probability density function of the active points is still very intricate, even with just $N = 2$ active points. The calculation quickly becomes intractable for $N > 2$. However, when two points are sufficiently close together, they are unlikely to interact as strongly with any other active points. This means our calculation for small $s$ when $N = 2$ has utility when predicting the same behaviour even when $N > 2$.

In order to better visualise this, we can simulate the process in Python. This means we can compare our Poisson process model with the DNA data to see how well they match.

For our point process, we generate 2D Poisson points, identify the active points and put their x-coordinates into a list. Then we calculate the nearest neighbour spacings of these positions. We do this for several realisations of the same process to get a larger dataset. We re-scale this data so that the mean spacing is 1 (for comparison purposes). We then make this into a cumulative frequency plot, normalised so the ‘total’ frequency is 1. This becomes an approximation of a cumulative density function.

We see in Figure 12 and Figure 13 where we plotted the re-scaled and normalised cumulative frequency plot of the replication origin spacings from S. cerevisiae and Drosophila KC (data is taken from [25]) respectively that there is evidence that our Poisson model is a better model for the replication origins in some organisms than Wigner’s surmise or an exponential distribution. RMS errors for S. cerevisiae with Wigner’s surmise and our Poisson model are 0.0349 and 0.0155 respectively.

In Figure 14, we compare our Poisson model to a dataset of human replication origins. The shape of the data is very different to yeast or fruit flies, characterised by a large number of smaller spacings and then a small (but still significant) number of extremely large outliers.

We calculate the lower and upper quartiles of the data, $Q_1$ and $Q_3$ which mark the end of the bottom and the start of the top 25% of the data respectively. Outliers are defined as data points which are more than 1.5 times the interquartile range ($Q_3 - Q_1$) above the upper quartile or below the lower quartile. (The whiskers extend to the last data point below $Q_3 + 1.5(Q_3 - Q_1)$ and above the first data point above $Q_1 - 1.5(Q_3 - Q_1)$, hence the apparent asymmetry in the boxplots).

This is characterised in Figures 15 and 16, the boxplots for the re-scaled spacing data in the S. cerevisiae and the Human MCF7 datasets. These infrequent, large spacings could indicate a clustering effect, where we have a clump of origins followed by a large gap. Alternatively, they could
Stochastic Models for Replication Origin Spacings in Eukaryotic DNA Replication

Figure 13: The re-scaled cumulative frequency plot of the replication origin spacings from Drosophila KC (a fruit fly strain) against the Poisson model. The second plot also includes spacings from Wigner’s surmise and an exponential distribution. RMS error between Drosophila KC and Poisson Model 0.0155. RMS error between Drosophila KC and Wigner’s surmise 0.0349. Spacing data taken from [25].

Figure 14: The re-scaled cumulative frequency plot of the replication origin spacings from a dataset marked Human MCF7 against the Poisson model. The second plot also includes spacings from Wigner’s surmise and an exponential distribution. The fit to the data is not good for any of the proposed models, particularly for smaller spacings. RMS error between Human MCF7 and Poisson Model 0.2299. Spacing data taken from [21].

indicate areas of the chromosome with a smaller density than the majority of points.

In Figure 17 we see a plot of the Human MCF7 without outliers and we note that the fit is closer to the Poisson model, but still not, by any means, good. We can see that the initial growth of the DNA data curve is less prominent as the mean has now been skewed less to the right.

4.2. Applying Linear Regression to the Data

The two main methods of simulating replication origin spacings on chromosomes which we have used are thinning the eigenvalues of random COE matrices and our Poisson model. In both the COE and the Poisson model, local linear repulsion is present - i.e. close to 0, the probability density functions of spacings between neighbouring points behave like a straight line with positive slope \( p(s) \sim As \) as \( s \) tends towards 0 (for some constant \( A > 0 \)). Both of these models are reasonable approximations for some of the replication spacings taken from particular organisms. For completeness, we now check that the replication origin spacing data taken from organisms also shows some sort of linear relationship close to 0.

Linear regression is a well-documented method of testing for the linear relationship between variables. A good citation for the general methods discussed in this subsection is [26].

We perform linear regression on the data starting from the smallest possible spacings and increasing the subset of the data we analyse to see how long the linear relation persists.

For each collection of data points, we attempt to match a line of best fit and calculate Pearson’s Product Moment Correlation Coefficient (PMCC) for each fit as well as a p-value.

The correlation coefficient takes a value between -1 and +1, with -1 indicating a perfect negative linear relationship (the slope is down and to the right) and +1 indicating a perfect positive linear relationship (i.e. the slope is up and to the right). Here, by perfect ‘linear relationship’, we mean that the RMS error between the straight line of best fit...
Stochastic Models for Replication Origin Spacings in Eukaryotic DNA Replication

Figure 15: Box plot of the spacings of the S. cerevisiae dataset as seen in Figure 12. The y-axis represents the spacings with re-scaling but no normalisation (i.e. the mean spacing is 1 but the frequencies are the same as the data, instead of being normalised to act like a probability density). The black box indicates the interquartile range of the data, with the orange line representing the median. The threshold for outliers is 1.5 times the interquartile range above or below the upper and lower quartiles, which is marked by the black ‘whiskers’. Values from the dataset that are outliers are marked as dots. Spacing data taken from [18].

Figure 16: Box plot of the spacings of the Human MCF7 dataset as seen in Figure 14. The y-axis represents the spacings with re-scaling but no normalisation (i.e. the mean spacing is 1 but the frequencies are the same as the data, instead of being normalised to act like a probability density). Outliers (1.5 times the interquartile range above or below the upper and lower quartiles) are marked as circles. The entire interquartile range here is visualised as a single line, indicating how extreme and prominent the outliers are in this dataset. Spacing data taken from [21].

Figure 17: The same plot as Figure 14, except the Human MCF7 replication origins have had all the outliers (pictured as black dots in Figure 16) removed from the dataset before the mean spacing was re-scaled (the remaining spacings have mean value 1). RMS error between Human MCF7 (without outliers) and Poisson Model 0.0897. Spacing data taken from [21].

through the data and the data itself is identically 0, i.e. the straight line perfectly goes through every single data point. A correlation coefficient close to 0 is indicative of little observable correlation.

If we were to apply this process to 2 data points, then 3, then 4 and pick the collection of points which gave the highest PMCC, we would always take the linear regression performed on just 2 point, as this would yield a PMCC of +1. Instead we perform PMCC in increments. For example, start from 0 and then perform linear regression on the first 5 bins. Then calculate the linear regression on the first 10 bins. starting from 0 and then working through histogram bins. For each collection of data points, we attempt to match a line of best fit and calculate Pearson’s Product Moment Correlation Coefficient (PMCC) for each fit as well as a p-value.

We used this method instead of just arbitrarily choosing the first, say tenth, of the data in order to get a better idea of how long the linear relationship lasts in each dataset.

The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a PMCC at least as extreme as the one computed from these datasets. A lower p-value means that it is more likely that our observations are not a chance finding.

We optimised for the largest PMCC so that we could see the best (i.e. most positive) linear relationship. We plotted the p-values as well to get a better understanding of how significant these linear relationships were.

In Table 1, we have values obtained using linear regression on several DNA datasets, as well as Wigner’s surmise and our Poisson model. Figures 18 and 19 are scatters plots

Huw Day and N. C. Snaith: Preprint submitted to Elsevier
Table 1
A table showing observed linear correlation in various DNA samples, optimised for largest PMCC.

| Sample             | PMCC  | P-Value |
|--------------------|-------|---------|
| K. lactis          | 0.8483| 3.848E-3|
| L. waltii          | 0.7566| 1.094E-3|
| S. cerevisiae      | 0.7872| 3.803E-5|
| Drosophila KC      | 0.9649| 6.825E-12|
| Poisson Model      | 0.9749| 3.372E-13|
| Wigner’s surmise   | 0.9999| 2.089E-35|
| Human MCF7         | 0.9541| 7.281E-11|

Figure 18: Linear regression used to look for local linear repulsion from the re-scaled (mean spacing 1) nearest neighbour spacings of Wigner’s surmise. The points represent the frequency (not the probability) of certain re-scaled spacings. This histogram has a bin width of 1/50. The line of best fit has been drawn to show the fit of the data. PMCC=0.9999 with a p-value of 2.089E-35.

Figure 19: Linear regression used to look for local linear repulsion from the re-scaled (mean spacing 1) nearest neighbour spacings of our Poisson model. The points represent the frequency (not the probability) of certain re-scaled spacings. This histogram has a bin width of 1/50. The line of best fit has been drawn to show the fit of the data. PMCC=0.9749 with a p-value of 3.372E-13.

Figure 20: Linear regression used to look for local linear repulsion from the re-scaled (mean spacing 1) nearest neighbour spacings of K. lactis, data taken from [16]. The points represent the frequency (not the probability) of certain re-scaled spacings. This histogram has a bin width of 1/10. The line of best fit has been drawn to show the fit of the data. PMCC=0.8483 with a p-value of 3.848E-3.

The flat stretch close to the origin seen in Figure 23 and 24 is noteworthy but not necessarily unexpected based on our method of analysis. Despite each replication origin being an area of chromosome with width, we modelled them as single points at just the midpoint of the two ends of the origin. This means that we would not expect interorigin distances to be smaller than a certain value, because the gap between the end of one origin and the start of another is always smaller than the distance between their midpoints. There is some literature [27] that considers so called interbubble distances (spacing from the far end of an origin to the near end of the next origin) instead of distances between midpoints. When we use the distance between midpoints...
and replication bubbles fuse together, the centre of the two fused bubbles is incorrectly assumed to be an origin and the location of the actual origin is missed (as noted in [28]). For our purposes it was easier to simulate replication if we considered each replication origin as a single point, so the interorigin distances were far more natural to use than the interbubble distances.

In Figures 25 and 26 we see a side-by-side comparison of the replication origin lengths and replication origin spacings (before any re-scaling of the data). Comparing these plots we see a disparity in the overlaps between the two. In the K lactis dataset, there are no origins of length within the interquartile range of the spacings. In contrast, 8.24% of the Drosophila KC origin lengths and 0.34% of the Human MCF7 origin lengths are within the interquartile range of their respective spacings. For the Drosophila KC dataset, the overlap provides a viable explanation for the lack of small interorigin spacings. This overlap is visible in Figure 25. For the Human MCF7 dataset, given that the plot in 24 has extremely narrow bin width, we see the same pattern close to 0. In other datasets where we see a similar overlap, we also find no very small interorigin spacings, whereas for datasets where the scale of the length of the spacings is very different from the scale of the origins we don’t see the same obvious deficit of small spacings, as in the case of K lactis.
Stochastic Models for Replication Origin Spacings in Eukaryotic DNA Replication

5. Discussion

Eukaryotic DNA replication relies on replication origins spaced along chromosomes to read genetic information with bidirectional replication forks. We can consider the distribution of replication origin spacings on chromosomes in various organisms.

In the literature, some attempts have been made to understand the spacing distribution of replication origins on chromosomes. In [1], we saw a detailed study on the effects of stalling, in particular double stall events, on replication origin spacings in various strains of yeast. Newman et al concluded that the effect these events had was significant, which served as inspiration for our implementation of uniform thinning in our first model. Their comparisons of replication origin spacings with simulated exponential distributions helped them to conclude that correlation between replication origins was evident and provided an early guide for many of the simulations we performed for this paper.

Not only were uncorrelated points unsuitable for modelling the spacings, but due to the prevalence of large spacings, the issue of the random completion problem emerges [27], [29]. The large spacings, in conjunction with the limitations on the replication cycle length meant a high likelihood of large gaps of unreplicated chromosome being present at the end of the replication cycle. In order to avoid this, a different spacing regime is required, where such large gaps are less likely to be present. We know that there is some correlation between replication origins and we have been able to specify that this correlation often appears to involve local linear repulsion.

We have proposed two different viable probabilistic models for modelling these replication origin spacing distributions. One model is more naïve and only considers the positions of replication origins and the spacings between them (thinned COE eigenvalues). The other model is more intricate and considers the replication trigger timings of each replication origin (Poisson model). We have promising analytical and numerical evidence of these two different stochastic processes being good models for various spacing data from replication origins on chromosomes. We have additionally shown that linear repulsion occurs in many different organisms with high significance and low p-values.

We have seen from analytical spacing calculations and linear regression that both of our probability models display local linear repulsion. This agrees with our linear regression analysis on some of the replication origin spacing data which also displays local linear repulsion. Even the human dataset displays repulsion, see Figure 24, although we do not have the evidence to claim the repulsion is linear.

With the human dataset, we have seen that the spacings are characterised by a few extremely large outlying spacings and many smaller spacings. When we filter out the large outliers, the distribution of the remaining spacings moves closer to our models, but they would still require some modification to be of use in analysing the human dataset.

One possible modification is considering a model where the likelihood of finding replication origins in a certain area
of chromosome is higher in some places than in others. In the language of point processes, we could say that the point process of replication origins on chromosomes could be inhomogeneous. We say a point process is homogeneous if its intensity is constant - that is, the average density of points in space does not depend on where you look. An inhomogeneous point process would be a process where the intensity is a function of the position in space. The average density of points in space does depend on where you look. More details can be found in [3]. Both of the models we have proposed in this paper have been homogeneous.

It is important to note that our Poisson model is only homogeneous if it is imposed on a cylinder and is not true if it imposed on a rectangle. If our Poisson model was set up in the rectangular setting, whilst the initial 2D Poisson points generated would be homogeneous by definition, the exclusion process would be less likely to exclude a point closer to the boundaries of the chromosome. To see this, consider a replication origin on the boundary. In effect, it only has to worry about a replication fork running it over from one direction when we consider the rectangular model. The closer a given origin gets to the boundary, the less likely it is to be affected by replication origins between the given origin and the boundary. This effect is eliminated when we consider a cylindrical model.

In practical terms, the inhomogeneous placement of replication origins means that certain parts of the chromosome are more or less likely to have replication origins present, independent of the proximity of neighbouring origins. This mechanism was not considered in our mathematical model, which generated a homogeneous 2D Poisson point process. How origins are distributed does have stochastic elements to it [1] but there is also evidence of location dependence [21].

By modelling replication origins on chromosomes with homogeneous point processes, we have implicitly assumed that there is no location dependence. Literature has shown this to be untrue in multiple cases such as in S. cerevisiae ([30], [31], [32], [33], [34], [35]) where the origin positions are well-known and mapped out, and in humans ([21], [36]) where there is strong evidence that at least some replication origins are location dependent. There are some attempts to address this in [37] by having each replication origin trigger time as an independent (but with different firing rate) exponential random variable. Later attempts in the same work vary that rate throughout the evolution of the chromosome.

The fact that our homogeneous models with purely stochastic and no location dependent origin distributions do not model humans well also suggests that the location factor is important. Our analysis also revealed the importance of the large outlying spacings in the human datasets. Such a large collection of outliers are not present in any of the other datasets we analysed. Future work might consider the positions of such large outliers.

From a mathematical modelling perspective, it makes sense to begin by assuming points are uncorrelated. This is done for spacings in [1] and also with the trigger timings in [15]. For our model, we started with uncorrelated Poisson points and then used an exclusion process which emulates the replication forks travelling along chromosomes in the replication cycle in order to produce correlated points. These correlated points show local linear repulsion, in line with statistical evidence from some of the data. There are other works which have proposed different timing regimes such as [38] which proposed using Gaussian distributed trigger times for origins on S. cerevisiae.

Our work approaches current problems in genetics from a mathematical standing. Whilst we made several assumptions and simplifications of the biological processes we modelled, many of these assumptions (replication origins having 0 width, replication fork moving at constant rates etc.) are taken in previous works from a more biological point of view and considered fairly reasonable. Using a more probabilistic approach and taking tools from different areas of mathematics (namely random matrix theory and the study of higher dimensional point processes), we found ourselves well equipped to approach the problem of understanding gaps between replication origins.

We focused here on the behaviour with small spacings and not on the occurrence of larger gaps. Because of this, we have not addressed whether either of our models present a viable solution to the random completion problem. It is worth noting, however, that the spacing distribution of eigenvalues of COE matrices, which models well the replication origin spacing data from K. lactis, and to a slightly lesser extent L. waltii, has a tail that decays faster than exponential and so has a reduced number of large spacings. In future work we could look at the likelihood of large spacings between neighbouring active replication origins.

The utility in our second model is particularly noteworthy. As we have built simulations to describe this process, we are able to put further restrictions on the model; for example in future work we could make the initial 2D Poisson process inhomogeneous, favouring certain areas of the chromosome (as discussed in [21]) or varying distributions/intensity of the trigger timings (as discussed in [15] and [38]). There is even scope to consider the effects that stalling has on our system, as originally considered in [1]. We could also use these simulations to capture information about replication time and compare with conclusions in other work (for example [39]). However, for the purposes of this paper, we focused on results on origin spacings.

Acknowledgements

We thank Beth Kent, Jack Simons, Lohini Sri Ram, Nor Farah Wahidah, Patrick Nairne, Sam Stockman, Zen Kok for their undergraduate projects on this topic. This gave us a head start which made approaching this problem considerably easier.
Thank you Ben Carter and Ellen Spackman for helping us understand the genetics content of this problem.

Thank you to Marton Balazs and Thomas Maddox for your insight and suggestions on how we could improve our stochastic model.

Funding
The first author was supported and funded by the EPSRC, which is part of UKRI.

Declarations of interest: none.

References
[1] Timothy J Newman, Mohammed A Mamun, Conrad A Nieduszynski, and J Julian Blow. Replication stall events have shared the distribution of replication origins in the genomes of yeasts. Nucleic Acids Research, 41(21):9705–9718, 2013.
[2] David HM Steffin, Emily M Hsieh, and Rayne H Rouce. Gene therapy: current applications and future possibilities. Advances in Pediatrics, 66:37–54, 2019.
[3] Sidney I Resnick. Adventures in Stochastic Processes. Springer Science & Business Media, 1992.
[4] Peter J Forrester. Log-gases and Random Matrices (LMS-34). Princeton University Press, 2010.
[5] Madan Lal Mehta. Random Matrices. Elsevier, 2004.
[6] Terence Tao. Topics in Random Matrix Theory, volume 132. American Mathematical Soc., 2012.
[7] Richard L Weaver. Spectral statistics in elastodynamics. Journal of the Acoustical Society of America, 85(3):1005–1013, 1989.
[8] Pascal Bianchi, Merouane Debahb, Mylène Maïda, and Jamal Najim. Performance of statistical tests for single-source detection using random matrix theory. IEEE Transactions on Information theory, 57(4):2400–2419, 2011.
[9] Shira Kritchman and Boaz Nadler. Determining the number of components in a factor model from limited noisy data. Chemometrics and Intelligent Laboratory Systems, 94(1):19–32, 2008.
[10] Boaz Nadler, Federico Penna, and Roberto Garello. Performance of eigenvalue-based signal detectors with known and unknown noise level. In 2011 IEEE International Conference on Communications (ICC), pages 1–5. IEEE, 2011.
[11] Alexei Onatski. Testing hypotheses about the number of factors in large factor models. Econometrica, 77(5):1447–1479, 2009.
[12] Romain Couillet and Merouane Debahb. Random Matrix Methods for Wireless Communications. Cambridge University Press, 2011.
[13] Jean-Philippe Bouchaud and Marc Potters. Financial applications of random matrix theory: a short review. Cambridge University Press, 2009.
[14] Nicholas Rhind. DNA replication timing: random thoughts about origin firing. Nature Cell Biology, 8(12):1313–1316, 2006.
[15] Ivan Liachko, Anand Bhaskar, Chanmi Lee, Shau Chee Claire Chung, Bik-Kwoon Tye, and Uri Keich. A comprehensive genome-wide map of autonomously replicating sequences in a naive genome. PLoS Genetics, 6(5):e1000946, 2010.
[16] Cheuk C Siow, Sian R Nieduszynska, Caroline A Müller, and Conrad A Nieduszynski. Oridb, the DNA replication origin database updated and extended. Nucleic Acids Research, 40(D1):D682–D686, 2012.
[17] Conrad A Nieduszynski, Shin-ichiro Hiraga, Prashanth Ak, Craig J Benham, and Anne D Donaldson. Oridb: a DNA replication origin database. Nucleic Acids Research, 35(suppl_1):D40–D46, 2007.
[18] Folkmar Bornemann, Peter J Forrester, and Anthony Mays. Finite size effects for spacing distributions in random matrix theory: circular ensembles and riemann zeros. Studies in Applied Mathematics, 138(4):401–437, 2017.
[38] Alessandro PS de Moura, Renata Retkute, Michelle Hawkins, and Conrad A Nieduszynski. Mathematical modelling of whole chromosome replication. *Nucleic Acids Research*, 38(17):5623–5633, 2010.

[39] Jean-Michel Arbona, Arach Goldar, Olivier Hyrien, Alain Arneodo, and Benjamin Audit. The eukaryotic bell-shaped temporal rate of DNA replication origin firing emanates from a balance between origin activation and passivation. *Elife*, 7:e35192, 2018.