Inevitability and containment of replication errors for eukaryotic genome lengths spanning Megabase to Gigabase

Al Mamun, Mohammed; Albergante, Luca; Moreno, Alberto; Carrington, Jamie T.; Blow, John; Newman, Timothy J.

Published in:
Proceedings of the National Academy of Sciences

DOI:
10.1073/pnas.1603241113

Publication date:
2016

Document Version
Peer reviewed version

Inevitability and containment of replication errors for eukaryotic genome lengths spanning Megabase to Gigabase

Mohammed Al Mamun[*], Luca Albergante[*], Alberto Moreno, Jamie T. Carrington, J. Julian Blow, Timothy J. Newman[+].

School of Life Sciences, University of Dundee, Dundee DD1 5EH, UK

[*] These authors contributed equally
[+] Correspondence to t.newman@dundee.ac.uk

Short Title: Containment of replication errors in eukaryotes

Significance Statement

Errors in DNA replication can never be completely avoided. By combining a minimal model that takes into account the positions of replication origins (the regions on the DNA where replication initiates) with experimental evidence, we show that genome size strongly influences the frequency of replicative errors. Our work reveals: i) simple eukaryotes are able to achieve a very low probability of replicative errors by having a moderate number of origins placed at regular intervals, ii) this strategy is ineffective in eukaryotes with larger genomes, such as human, for which replicative errors are inevitable, and iii) in these organisms even moderate numbers of origins can provide containment of replication errors to very low levels, which can be repaired subsequently.

Abstract

The replication of DNA is initiated at particular sites on the genome called replication origins (ROs). Understanding the constraints that regulate the distribution of ROs across different organisms is fundamental for quantifying the degree of replication errors and their downstream consequences. Using a simple probabilistic model we generate a set of predictions on the extreme sensitivity of error rates to the distribution of ROs, and how this distribution must therefore be tuned for genomes of vastly different sizes. As genome size changes from Megabases to Gigabases we predict that regularity of RO spacing is lost, that large gaps between ROs dominate error rates but are heavily constrained by the mean stalling distance of replication forks, and that for genomes spanning ~100 Megabases to ~10 Gigabases errors become increasingly inevitable but their number remains very small (three or less). Our theory predicts that the number of errors becomes significantly higher for genome sizes greater than ~10 Gigabases. We test these predictions against datasets in yeast, *Arabidopsis*, *Drosophila* and human, and also through direct experimentation on two different human cell lines. Agreement of theoretical predictions with experiment and datasets is found in all cases, resulting in a picture of great simplicity, whereby the density and positioning of ROs explain the replication error rates for the entire range of eukaryotes for which data is available. The theory highlights three domains of error rates: negligible (yeast), tolerable (metazoan) and high (some plants), with the human genome at the extreme end of the middle domain.

\body

Introduction

The proper maintenance of genetic information is of fundamental importance to the survival of all organisms and many molecular mechanisms exist to ensure that the genetic sequence encoded by DNA is maintained unaltered generation after generation (1–3). To preserve the integrity of genetic information and to avoid aberrant ploidy it is crucial that the entire DNA is copied exactly once: replicating only part of the DNA results in potential corruption of genes and replicating certain parts of the DNA more than once would perturb chromosome structure and strongly affect gene dosage (4–6). Not surprisingly, regions of under- and over-replicated DNA are common in cancer (7, 8).

DNA replication is a particularly complex process in eukaryotic organisms with large genomes distributed across multiple chromosomes. Multiple checkpoints exist to ensure that once replication starts the whole DNA is faithfully replicated before the chromosomes are segregated. Under- and over-replication of DNA are prevented by using predefined points of replication initiation called Replication Origins (ROs) (3, 9).

During late mitosis and the G1 phase of the cell division cycle, each potential RO is 'licensed' for a single initiation event by being loaded with MCM2-7 double hexamers. In order to prevent re-replication of DNA segments, the ability to license new origins ceases before cells enter S phase. During this phase, hundreds to thousands of licensed ROs are activated throughout the genome (10). Bidirectional replication forks are established at active ROs, each driven by a single MCM2-7 hexamer, allowing DNA polymerases to copy the DNA (Figure 1a). Despite being highly reliable molecular machines, replication forks can on rare occasions irreversibly stall (11). The activation of additional ROs can overcome the problem of irreversibly stalled replication forks, as a new fork will eventually *meet* the stalled one hence replicating all of the intervening DNA. However, if adjacent right-moving and left-moving forks stall and no additional ROs are available between them, the DNA in-between the two forks will remain unreplicated (Figure 1b). This phenomenon constitutes a major replication error for the cell, which is commonly called a double-fork stall (DFS) (Figure 1). The occurrence of DFSs is therefore a key obstacle for cells to either avoid or overcome in order to maintain replication fidelity. The molecular processes underlying the management of DFSs are an active field of study and insults to these processes have been associated with different pathologies (11–13).

In our previous work, we introduced a simple probabilistic theory to determine the probability of replication failure arising from DFSs for a given set of ROs in a genome (14). The theory depends on two key assumptions, i.e. that the cell has no time constraint in completing the process (i.e. that all licensed ROs are allowed to be activated as necessary), and that there is a constant small probability per nucleotide for each individual replication fork to irreversibly stall. Mathematical analysis of the theory showed that in organisms with a genome length comparable to yeasts (~10 Mbp), evenly distributing the ROs throughout the genome optimally reduces the replication errors due to irreversible fork-stalling to levels observed in experiments. In accordance with the theoretical prediction, a strong bias towards evenly distributed

ROs was observed in biological data derived from different yeast species (14). The theory relies on a single unknown parameter, the median stall distance (denoted by N_s), which describes the typical stalling distance of replication forks (RF) in eukaryotes (14). Our theory was used to obtain an estimate N_s from the probability of DFS and the RO distribution. The value obtained (N_s~12 Mb) is remarkably close to direct experimental measurements.

In this article we extend our theory to study much larger genomes (100 Mbp – 10 Gbp), which are typically found in metazoa and plants. Our theory requires as input the positions of ROs along the genome and yields a number of clear predictions concerning the rates of DFSs, using both mathematical and computational approaches. These predictions were tested on available datasets describing RO distribution in one plant (*Arabidopsis*) (15), one invertebrate (*Drosophila*) (16) and two independent human datasets (reporting different human cell-lines) (17, 18) (SI Table 1). Note that the two human datasets have been derived using different approaches to RO detection and hence the number and positions of ROs vary between them. The two datasets are largely compatible with reported 70% overlap in genomic sites containing ROs in both datasets (18) (See also SI Figure 1), and therefore can be used to test the robustness of our theory to experimental and biological variation.

Our theoretical and computational analysis leads to a series of direct predictions, which are all found to be consistent with all datasets analysed, revealing a picture of great simplicity. The robustness of DNA replication in eukaryotes can be maintained so long as the largest replicon (inter-RO distance) is well below the median stall distance N_s. For organisms with larger genomes, such as typical vertebrates and plants, DFSs are highly likely even if the mean replicon length is small. These organisms therefore require mechanisms to deal with DFSs, and in related experimental work, we provide experimental evidence for one such post-replicative mechanism (19). For cells with such repair mechanisms the burden of equally spacing ROs is lifted; far more important is the distribution of larger replicons (relative to N_s) from which DFS events are most likely to arise. Our theory also indicates that the number of DFSs becomes unwieldy for genomes significantly greater than 10 Gb, and this additional challenge may play a central role in limiting the genome size of higher eukaryotes.

Results

The ‘central equation’ for determining replication errors

In our previous work (14) we derived a mathematical equation for the genome-wide probability of DFSs, based on the distribution of ROs and the median stalling distance N_s. The published equations depend on the largest replicon being significantly smaller than N_s. Although this limitation holds true for yeast genomes, it does not apply to replication origins that have been mapped in mammalian cells. As described in Materials and Methods, we use the same theoretical framework to derive more general equations that are applicable to genomes containing arbitrarily large replicons. In order to utilise our theoretical results we require detailed information on the location of ROs. A number of datasets have been published that provide the locations

of ROs in eukaryotes along with the total genome length (denoted by N_g in the following). In this work, we have used origin mapping data from *Saccharomyces cerevisiae* (20), *Schizosaccharomyces pombe* (21), *Arabidopsis thaliana* (15), *Drosophila melanogaster* (16) and from 5 human tissue culture cell lines (IMR-90, HeLa, hESC, iPSC and K562) from Besnard et al. (17) (denoted by 'B' in the following) and Picard et al. (18) (denoted by 'P'). Since the work of Picard et al. used more modern techniques (particularly in peak identification) it might be considered a more reliable dataset; comparison with the Besnard et al. is useful in assessing the experimental uncertainties in some of the data.

Because of the very low probability of a DFS in any given replicon, we can show that the statistics of DFSs are Poisson to a very high level of accuracy (see SI Text), and that the probability of no DFSs genome-wide has the form $\exp(-\lambda)$. Thus, a great deal of information concerning the probabilities of DFSs for given genome can be obtained from the single parameter λ. We remind the reader that for a Poisson distribution, λ also describes both the mean and the variance of the distribution. For a given genome with K ROs, we denote the replicons by the K-1 values N_i (with i=1, …, K-1). These data can then be used in the 'central equation' arising from our theory (Eq. 1):

$$\lambda = \log(2)\frac{N_g}{N_s} - \sum_{i=1}^{K} \log\left(1 + \log(2)\frac{N_i}{N_s}\right) \qquad \text{(Eq. 1)}$$

This expression for λ contains a single unknown parameter N_s – i.e. the number of replicated bases along the DNA beyond which 50% of replication forks irreversibly stall. This is inversely proportional to the very small probability of stalling per nucleotide (14).

On the right-hand-side of (Eq. 1), we can identify the two distinct contributions of the genome length (first term) and of the RO distribution (second term). Genome length determines a baseline probability of DFSs that can be lowered by increasing the number of ROs and/or changing their distribution along the genome: indeed, as we have shown previously (14), for a given number of ROs, equally distributing them across the genome is the optimal arrangement to minimize the probability of DFSs. This establishes a hierarchy of contributions to the probability of DFSs, with genome length being the most important factor, followed by RO number and then RO distribution (Figure 2).

In organisms with relatively small genomes, such as yeasts (~10 Mbp), an average density of 1 RO per ~20 Kbp allows the maintenance of very small probabilities of genome-wide DFSs. Application of (Eq. *M*11) to the yeast datasets gives values around 10^{-3} for the probability of one or more DFSs, consistent with our previous analysis. With the increase in genome size from around 10 Mbp (in yeasts) to around 10 Gbp (in human), (Eq. *M*11) shows that the probability of DFSs increases by approximately two orders of magnitude, to more than 0.5 for human genomes (Figure 3a). This huge increase in error rate occurs despite essentially no shift in the mean replicon size (Figure 3b). Therefore it is absolutely necessary for these organisms to have molecular machinery able to repair DFSs.

The bias towards uniformly spaced replication origins is progressively lost in larger genomes

The regularity of the RO distribution can be assessed by computing the coefficient of variation of the replicon lengths, denoted by *R*, defined as the ratio of their standard deviation to their mean. For a perfectly uniform distribution of equally spaced ROs, *R* is equal to 0. On the other hand, computational analysis indicates that when ROs are randomly distributed on the genome, the value of *R* is very close to 1 (14).

In the yeast genomes (diploid genome sizes ~20 Mbp), we previously showed that their RO distributions were strongly biased towards uniform spacing with values of *R* ranging from 0.72 to 0.77 (Figure 3c). The probability of DFSs is very small in yeasts due to they small genome size, and optimization of the RO positions by lowering *R* reduces this even further. However, as discussed above, organisms with larger genomes have a significantly higher probability of DFS events, which results in the need for additional molecular mechanisms to cope with the consequences (19) and the presence of such mechanisms means there is little to be gained in uniformly ordering ROs on the genomes. Thus, our expectation is that *R* should be significantly larger in organisms with larger genomes compared to the values found in yeast. Statistical analysis of the available data confirms this expectation (Figure 3c). *Arabidopsis* and *Drosophila* (diploid genome sizes ~250 Mbp) have values of *R* around unity (i.e. approximating a random distribution). Particularly striking is the fact that in human genomes (~6,000 Mbp), the values of *R* are *significantly larger* than unity, indicating that ROs are not spaced purely randomly and that both the number and size of large replicons is significantly greater than expected by chance. This unexpected distribution has important consequences that are discussed below.

The probability of a DFS in a given replicon increases with the replicon length according to (Eq. *M*5) (Materials and Methods) and is plotted in Figure 3d. The probability has a strongly non-linear form: increasing as the square of the replicon length for lengths much less than the stalling distance, and saturating to unity for lengths significantly greater than the stalling distance. Figure 3e provides a graphical representation that highlights the dramatic shift in variation of replicon lengths, or equivalently the per replicon rate of DFS, by plotting the predicted probability of DFSs across the largest chromosome of different organisms. It is apparent that the variation in probability of error increases by approximately one order of magnitude from yeast to *Drosophila*, and then again by approximately one order of magnitude from *Drosophila* to human.

Large replicons in human genomes cause the most errors but are bounded by the stalling distance

Consistent with our analysis of the values of *R*, we would expect the largest replicons in the genome to be very significantly different in diploid genomes of size ~20 Mbp, ~250 Mbp and ~6 Gbp (represented by yeasts, *Drosophila*/*Arabidopsis* and human respectively), with significantly larger replicons appearing in those genomes with *R* larger than unity. As seen in Figure 4a, this is exactly what is observed, with the largest replicons being ~60 Kbp in yeasts (~120 Kbp expected for a random distribution), 151 Kbp in Drosophila (207 Kbp expected if random), 773 Kbp in

Arabidopsis (663 Kbp expected if random) and ~5 Mbp in human (~300 Kbp expected if random). This can also be seen by the significant increase in outliers in the box plots of replicon lengths for the different organisms considered (Figure 4b). As is clear from Figure 3d, the probability of a DFS in a given replicon increases dramatically as the length of the replicon approaches the median stalling distance N_s. To avoid almost inevitable errors arising from a single replicon, we would expect the length of the largest replicon in the entire genome to be bounded by N_s, and this is indeed what is observed in the data. In the B dataset, we find that the largest replicons in each human cell-line are 3.59 Mbp (IMR90), 3.71 Mbp (hESC), 3.71 Mbp (iPSC), and 4.29 Mbp (HeLa); while in P we find 5.65 Mbp (IMR90), 5.73 Mbp (HeLa), and 5.94 Mbp (K562). Interestingly, the largest replicons appear to be bounded by approximately one half of the stalling distance, which means that the largest replicon in each human cell line contributes a predicted error rate of approximately 5%. We note that all the datasets used for our analysis rely on genomic sequencing data. As such, large regions of repetitive DNA will not be sequenced accurately, and yet are likely to contain ROs. These false negatives imply that the largest replicons measured provide an upper bound rather than a definite value, though we do not expect large numbers of missed ROs (19). The future use of more advanced techniques, for example single cell sequencing, will shed more light on this aspect.

In the human genome, given that errors are very likely, we can determine the range of replicon lengths that are the main contributors to the DFS. We grouped the replicons into five cohorts: very small (XS; <1 Kbp), small (S; 1-10 Kbp), medium (M; 10-100 Kbp), large (L; 100 Kbp-1 Mbp) and very large (XL; >1 Mbp). The frequency of replicons in these five cohorts is shown for IMR90 from the B and P studies in Figure 5a and 5b. The most common range of replicons is small and medium respectively, the shift from ‘small’ to ‘medium’ being due to the coalescence of small replicons in the Picard et al. study. ‘Large’ and ‘very large’ replicons appear only at low frequency. Despite this, Figure 5c and 5d show that the cohort of ‘large’ replicons dominates as the source of error, which is due to the fact that the DFS probability increases non-linearly with the replicon length (Figure 3d). The error rate due to the small number of ‘very large’ replicons is significantly smaller compared to the ‘large’ replicons. An important consequence of this finding is that there will be a very limited impact on genome-wide error rates from false negatives, which primarily affect the distribution of ‘very large’ replicons.

Interestingly, in both datasets, for all cell lines a closer examination of the error rates in the vicinity of the ‘large’ cohort shows a surprisingly statistically uniform distribution of error rate, which is suggestive of ROs being placed so as to “spread the risk” of error across size scales. In Figures 5e and 5f, the probability of DFS in each 10 kbp interval in the range 10 - 300 kbp is shown for the Besnard et al. (Figure 5e) and Picard et al. (Figure 5f) datasets for primary IMR90 cells. These are the replicons that contribute the most to the DFS probability. The maxima are relatively broad, particularly for the B dataset, for which the probability of DFS in each 10kbp is approximately constant at 0.030-0.035 across replicons spanning from 40 kbp to 200 kbp. For replicons significantly smaller than the stalling distance, one can infer from the theory that ROs are placed in such a way to give a power law, with a frequency of DFSs that decreases as the inverse square of the replicon length thereby spreading the probability of a DFS equally amongst all size classes (described by (Eq. *M*17) in the SI Text). Figure 5g and 5h show that there is a remarkable concordance between the

theoretical frequency distribution (in blue) with the frequency distribution in the data for IMR90 cell-line in both datasets (in red). There is also excellent agreement with the theoretical distribution in all the other cell-lines in both datasets (SI Figures 2, 3 and 4). These results can be interpreted in terms of "spreading the damage" as widely as possible in the replicon size region of maximal DFS errors, as a power law is the most effective way to delocalize errors from any single cohort of replicon lengths.

Replication errors are common but low in number for higher eukaryotes

As discussed above, our theory predicts that the distribution of the number of DFSs in a given genome is Poisson-distributed to a very high degree of accuracy. We have applied our theory to the human cell lines datasets to test this prediction. As shown in Figure 6, for all cell lines, from both laboratories, the distribution of DFSs is indeed Poisson-distributed, regardless of being primary or tumoural cell lines. Statistical analysis confirms that the computationally derived probability distribution of DFSs is statistically indistinguishable from the fitted Poisson distribution. Interestingly, we find a very low probability (<10%) of encountering more than three DFSs in the replication of the entire diploid human DNA per cell cycle. Therefore, despite the high probability of the presence of DFSs (~80%), in ~90% of cells undergoing DNA replication the expected number of DFSs is predicted to be three or less, with one or two errors being the most likely occurrences. Indeed, we find that the parameter λ (i.e., the mean number of errors) that characterizes the distribution of DFSs ranges from 1.67 to 2.15 in Besnard et al. (17) and from 1.21 to 2.05 in Picard et al. (18).

Given that DFSs in human cell lines are almost inevitable, it is somewhat surprising to find that their number is quite sharply constrained to be essentially one, two or three. This might indicate that the mechanism that deals with such errors has a very low capacity. If, as suggested in Moreno et al. (19), the defects induced by DFSs can be resolved in the following cell cycle by segregating unreplicated DNA to daughter cells, DNA strand breaks could be generated at each DFS. Because the number of illegitimate ways that double strand breaks could be correctly rejoined increases as the factorial of the number of breaks, this might constrain the number of tolerated DFSs to about 3 or less. We provide a rationale for putative biological mechanisms in the discussion section, and our arguments lead us to consider two different "biomarkers" for double strand breaks which would arise from DFS errors: these are the presence of 53BP1 nuclear bodies in the G1 phase of the subsequent cell cycle, and the presence of ultrafine anaphase-bridges (UFBs) during mitosis. Our theory suggests that the number of both 53BP1 nuclear bodies and UFBs are distributed as a Poisson with a value of λ between one and two.

We have performed an experimental analysis of 53BP1 in IMR90 cells and both 53BP1 and UFBs in U2-OS cells, and measured the frequency of their occurrence during the cell cycle at a single cell level (19). In agreement with our predictions, the experimental distributions of both 53BP1 nuclear bodies and UFBs fit to a Poisson distribution (Figures 7a, 7b and 7c). Statistical analyses indicate that both a naïve fitting using the mean of the data and a more advanced approach that accounts for potential errors introduced by the experimental procedure of the immunofluorescence experiments (Figures 7a, 7b and 7c) produce distributions which are not statistically different from Poisson distributions for both 53BP1 nuclear bodies (*P* values between 0.61 and 1 for both IMR90 and U2-OS cells) and UFBs (*P* values between 0.53 and 1

for U2-OS cells). Additionally, the fitted λ values, 0.52 (naïve) and 0.54 (filtered) in IMR90 and 1.64 (naïve) and 1.89 (filtered) in U2-OS cells for 53BP1 nuclear bodies, and 1.27 (naïve) and 1.19 (filtered) for UFBs, are in line with the expectation of a limited number of DFSs. Moreno et al (2016) (19) show that the number of 53BP1 nuclear bodies and UFBs follows a Poisson distribution in the HeLa cell line with λ values of 0.94 (naïve) and 1.12 (filtered) for 53BP1 and 1.43 (naïve) and 1.19 (filtered) for UFBs (19). Taken together, these results provide good agreement of our theory with the available data and reinforce the connection between 53BP1 nuclear bodies and UFBs to DFSs. The analysis of UFBs in unperturbed IMR90 cells was not possible due to experimental difficulties related to the fact that this cell line is not immortalized.

As a more quantitative analysis, we compared the λ values obtained by direct calculation from the RO distribution of different human cell lines and the experimental λ values estimated from the distribution of 53BP1 and UFBs. Note that comprehensive RO distribution data are not available for the cell line used for the UFB experiments (U2-OS) and diversity has been observed in RO-distribution across different cell lines (17). Moreover, both 53BP1 and UFBs are likely to provide only an approximation of the number of DFSs as they appear also in the presence of non-DFS associated double strand breaks. Despite these limitations, a comparison of the λ values indicates that experimental measures are in excellent agreement with theoretical prediction (Figure 7d). Additional comparisons with the λ values obtained from HeLa reinforce our conclusions (Figure 7d). Interestingly, the range of variation observed in the experimental value of λ is matched by the range of variation of our model predictions, suggesting that our methodology is correctly capturing experimental variations.

In both IMR90 and HeLa cells the experimentally derived λ obtained from 53BP1 nuclear bodies data is approximately half of the theoretical estimate obtained from the RO mapping data. This is also true for UFBs in HeLa cells. So long as the density of ROs is small, it is straightforward to show that doubling the density of ROs halves the value of λ. Hence the factor of two difference in the experimental and theoretical values of λ could indicate that around half of the genomic ROs are missing in the current datasets (e.g. due to difficulties in detecting ROs that fire very rarely or ROs positioned in repetitive regions of the DNA). This line of reasoning is also consistent with a potential issue with the largest measured replicon being approximately 4 Mbp; the issue being that the replication time for such a gap would be significantly longer than typical S-phase (ca. 8 hours) (22). If the true RO density is twice that measured, one can show that the largest gap would be halved, giving a value of 2 Mbp which is in line with the estimate of 2 Mbp for the longest stretch of DNA that could be replicated in the duration of S-phase (assuming a fork speed of approximately 2 Kbp per minute (23), and remembering that a large replicon will be replicated almost symmetrically by forks travelling from either end).

Effect of variation of the stalling distance

In applying our theory to the RO position data for various human cell lines, we can vary the numerical value of the median stalling distance N_s and measure the effect on the expected number of DFSs. This allows us to gauge the extent to which our conclusions are robust to the variation of the only parameter in our analysis for which

we do not have strong experimental data. Both theoretical and biological estimates indicate that N_s is approximately 10 Mbp (14, 24). However, a precise estimate of this value is difficult to determine *in vivo*. The stalling distance is inversely proportional to the very small probability of an irreversible stalling event per nucleotide replicated, which because of the conservation of the basic replication machinery is likely to be relatively well conserved across eukaryotes.

First, we analyzed the overall probability of DFSs occurring as N_s is varied. In all the human cell lines considered we observe a characteristic transition around 5 Mbp: below this value the probability of observing DFSs saturates at one (Figure 8a). Therefore, DFSs are inevitable for smaller values of N_s as one might expect. Importantly, our analysis indicates diminishing returns when N_s is increased to much larger values: even for N_s around 30 Mbp, error rates are sufficiently high (1 in 5 cells would experience a DFS during S phase) that additional DFS repair mechanisms are still required. Therefore, in higher eukaryotes with large genomes the pressure to maintain genome stability is most easily resolved by additional safeguard mechanisms to deal with consequences of DFSs, rather than by stabilizing the replication machinery to give such a large N_s that DFSs can be avoided with the regular RO distribution found in eukaryotes with smaller genomes.

Our analysis stresses the inevitability of DFS errors during replication of the human genome and calls for a shift in our approach with respect to how the problem has been viewed in the past. On varying the median stalling distance in human cells, the probability of exactly one DFS genome-wide reaches a maximum between 10 and 15 Mbp, depending on the particular cell line and dataset used (Figures 8b and 8c). Furthermore, on varying the stalling distance, we find that the probability of exactly two or exactly three DFSs occurring also have peaks in the range 6-10 Mbp, again depending on the cell line and the dataset used (Figures 8b and 8c). To probe the likelihood of small number of errors occurring, we plotted the probability of observing *one, two or three* DFSs as stalling distance was varied (Figures 8d and 8e). These results show a very pronounced maximum for N_s around 10 Mbp in the B dataset, and around 8 Mbp in the P dataset. In summary, our analysis of the available RO distribution in a variety of human cell lines and in different datasets indicate that only for N_s in the vicinity of 10 Mbp the number of DFSs is constrained between zero and three.

Finally, we can measure the average number of DFSs when N_s is varied. This number is equal to the λ parameter of a Poisson distribution, and therefore allows a direct comparison to our experimental measures. As expected, the average number of DFSs decreases from a large value as N_s is increased (Figures 8f and 8g). As explained in the previous section, fitting the Poisson distribution to 53BP1 and UFB experimental data gives values of λ between 0.54 and 1.89 (the values are shown in Figure 8f and 8g as black, blue and red lines). The intersection of the decaying curve with these two lines provides another independent estimate of the stalling distance, which we find to be between 8 and 16 Mbp depending on the cell line and dataset used. Our analysis of the statistics of DFSs in human cell data on varying the stalling distance therefore provides very strong evidence for the robustness of this parameter with a value in the range 8-15 Mbp, consistent with previous estimates from our analysis of yeast RO distributions, and direct experimental estimates (14, 24).

Effect of varying the number of licensed ROs

Interestingly, amongst the cell types we analysed, there was no major difference in the mean replicon length (Figure 3b). Figure 9 shows how decreasing mean replicon length would reduce the probability of DFSs in a *generic* organism. The black, light-blue, and blue lines illustrate the mean replicon length to achieve a fixed probability of DFSs under the optimal situation of equally spaced ROs. All the datasets analyzed in the article have a mean replicon length ranging between 10 and 100 Kbp (shaded pink in Figure 9). Because of the relatively small genome sizes of yeasts, so long as ROs are evenly spaced this mean replicon length can achieve a tolerable DFS probability of ~0.1%, similar to the chromosome mis-segregation rate (14). In order to maintain a low probability of DFSs as in yeasts, longer genomes would require a much lower mean replicon length or in other words, much higher density of ROs on the genome. Since the MCM2-7 double hexamer that licenses an RO has a footprint of ~60 bp (25, 26) this provides an absolute limit to the possible replicon length (dashed line in Figure 9). It is just about possible for organisms with ~6,000 Mbp genomes to achieve yeast-like DFS probabilities, but the genome would have to be almost completely packed with MCM2-7, which might leave the genome unable to perform its major function of providing the template for transcription. Since this is an implausible saturation for normal cells, additional post-replicative mechanisms must be in place to deal with the inevitable DFSs. For this reason, regularity in RO distribution is not an effective safeguard against DFSs in organisms with larger genomes.

Discussion

Faithful DNA replication is fundamental to preserve the genetic content of cells and to avoid the severe pathologies which arise when DNA is improperly replicated. The appropriate location and activation of Replication Origins (ROs) is fundamental to ensuring that replicative errors are minimized. Here we show that understanding the principles that govern distribution of ROs provides new quantitative insights into the way that different organisms maintain genetic integrity. By using a probability theory approach, based on a one-parameter model with simple yet plausible assumptions, we have developed a set of measures and predictions that further this understanding. The excellent agreement of our theoretical predictions with experimental data strongly supports the validity of our model assumptions. Moreover, it allows us to explore the rich system-level diversity of features and constraints associated with DNA replication.

Replicative errors are inevitable in larger genomes

Increased phenotypic complexity of organisms is generally associated with larger genome length and metazoans have much larger genomes compared to yeast: the diploid human genome is approximately 600 times larger than the haploid yeast genome. Despite this large difference in genome size, the replication machinery is essentially conserved (4). Over the past few decades, much effort has been devoted to understanding the molecular mechanisms involved in eukaryotic DNA replication and the associated damage-repair mechanisms. However, less is known about the system-level structures and processes that allow replication fidelity across the different scales of eukaryotic complexity, mirrored by genome lengths spanning over three orders of magnitude across yeast to human. We have used a theoretical approach, previously validated in yeasts (14), to predict the probability of DFSs for different organisms with widely different genome lengths, and for which detailed RO distribution data are available.

Our ‘central equation’ shows that there is a hierarchy of contributions to the probability of DFS, with genome length being the most important factor, followed by RO number and then RO distribution. This effectively creates different classes of probabilities of DFS errors ($\sim 10^{-3}$, $\sim 10^{-2}$, and ~1) for the respective classes of organisms according to their genome lengths (~20 Mbp, ~250 Mbp and ~6 Gbp). Interestingly, amongst the cell types we analysed, there was no major difference in the density of ROs i.e. mean replicon length. One possible explanation for this is that in order to make a significant effect on reducing DFSs, the RO density in organisms with genomes of 250 Mbp or more would lead to excessive clashes with the transcriptional machinery. The third component of our equation – the uniformity of replicon length, i.e. R – also reflects these classes (with values <1, ~1 and >1 respectively), indicating that as the probability of DFSs approaches 1 in larger genomes, the pressure towards a regular RO distribution is lifted.

Inevitability is mitigated by containment in longer genomes and beyond

DFSs are the primary cause of DNA double strand breaks during replication (27–29), and are likely to be major contributors for the development of cancer and other

pathologies, such as ones associated with aging (30, 31). The inevitability of DFSs in longer genomes requires the presence of cellular mechanisms, which are able to deal with such errors in an efficient manner. In related experimental work, we provide experimental evidence for one such post-replicative mechanism, involving the segregation of unreplicated DNA via UFBs and its protection by 53BP1 before being resolved in the next S phase (19). We have demonstrated very good agreement in the numbers and statistical distribution of experimental measurements of both 53BP1 and UFBs with the predictions of Poisson statistics from our theory, supporting the validity of our conclusions, and indicating that DFSs in the experimental systems are well approximated as independent events.

Analysis of the data available for human cell-lines within our theoretical framework shows that RO density and distribution constrain the number of DFSs per cell cycle to three or less for nearly all cells. This may partially be explained by the difficulty in properly recombining two strands of DNA when end-joining is used. For example, if four DFSs occur and need to be fixed, eight strands will be generated and only one of the 24 theoretically possible combinations is correct. From our experimental observations, cells with large numbers of 53BP1 nuclear bodies and UFBs showed increased blebbing and apoptosis. This suggests that large numbers of DFSs could compromise the working of the cell and the efficiency of the repair mechanism. Thus, our theory, in light of the experimental data, shows a contingent trade off between inevitability of DFS occurrence and the difficulty of its resolution (i.e. apparently requiring sophisticated molecular machinery for detection and repair). It is worth stressing that our central equation for λ, the mean number of DFSs, contains very large numerical values, i.e. N_g and N_s, as well as thousands of replicon lengths. Therefore, in principle, the formula could have produced values for λ of almost arbitrary magnitude, either much less than or much greater than unity. It is striking that our theoretical predictions from the central equation yield values for λ close to unity and in such strong agreement with experimental data.

Another important requirement for the containment of replicative errors in larger genomes is an upper limit in the length of large replicons. Longer replicons correspond to a higher probability of DFSs (Figure 3d). Our theory indicates that the largest tolerable replicons in human cell-lines are bounded by ~0.5 N_s, and interestingly the largest replicons found in experimental datasets are around 0.3 N_s. In addition, we have analysed human cell line data within our theoretical framework, and by varying N_s we are able to clearly show that the probability of observing a number of DFSs equal to one, two or three is maximized for N_s in the region of 10 Mbp. This value for N_s is in excellent agreement with previous experimental and theoretical estimates in human cell lines and yeasts (14, 24). Due to the universality of replication machinery across the eukaryotes and the necessity of error containment in larger genomes, we propose this N_s value to be robust and universal in eukaryotes. A further signature of the containment mechanisms associated with the inevitable errors in human genomes can be found in the distribution of the risk among replicons of different sizes: a relatively narrow range of replicons (of size ~40 to ~200 kbp) contributes the most to DFSs, with the different replicon sizes in this range contributing approximately equally to the risk.

As a final note, it is worth stressing that some organisms, particularly plants, have very large genomes, with N_g as large as ~100 Gbp (32). Our theory would predict in

such cases that the number of DFSs becomes much larger than three, and in the region of ten or more. Interestingly it has been observed that the cell cycle length in plants undergoes a dramatic lengthening as genome size exceeds about 25 Gb (32), potentially reflecting the significantly greater burden of DFS detection and correction in these organisms. We would predict similar effects for ploidy variants within the same species. We currently do not have genome-wide RO distribution data for these organisms to test this idea, but this would provide further opportunities for gaining new understanding of the system-level strategies that eukaryotes employ to minimize replication errors.

Materials and Methods

1) Experimental setup

For the 53BP1 and UFBs experiments, U2OS and IMR-90 cell lines from the American Type Culture Collection (ATCC) were maintained in Dulbeccos's Modified Eagle's medium (DMEM, Invitrogen), supplemented with 10% FBS (Invitrogen) and penicillin and streptomycin at 37˚C in 5% CO_2. Standard immunofluorescence protocols were used for the 53BP1 and UFBs staining. Briefly, cells were fixed with 4% formaldehyde, permeabilised with 0.1% Triton in PBS and blocked in 0.5% fish gelatin (Sigma, G-7765). Samples were incubated overnight with primary antibodies. To specify G1 phase cells they were incubated with 40 μM EdU (Invitrogen) for 30 minutes prior to fixation, and then incubated with Cyclin A (abCam, 1:300, ab16726). For the detection of 53BP1 cells were also stained with GFP (1:2000, abCam ab13970). To stain incorporated nucleotides the Click-iT-EdU kit was used as instructed by the manufacturers (Invitrogen, C10337). For staining UFBs, cells were incubated with BLM (1:200 Santa Cruz, sc-7790). Alexa secondary antibodies (Invitrogen) were used for 1 hour. Microscopy images were acquired using an Olympus IX70 deltavision deconvolution microscope and a CCD camera. Data from microscopy experiment were analysed using Volocity 3D analysis software (Perkin Elmer).

2) Datasets used and statistical analysis

Limited direct experimental evidence exists on ROs in plants and metazoa and most data focus on the genomic density, rather than localization, of ROs (33, 34). Therefore, the main results of our article are framed in the context of available datasets describing genome-wide RO-positions. Less high-quality datasets have been considered where appropriate to provide additional challenge to the theoretical predictions and their interpretation. *Saccharomyces cerevisiae* ROs were obtained from the highly curated OriDB (20) with selection criteria discussed in (14). To provide additional validation, we considered another yeast species in this article: *Schizosaccharomyces pombe* (21). RO distribution data were also obtained for the following multicellular organisms: *Arabidopsis thaliana* (15), *Drosophila melanogaster* (16) and human. Human data for the four cell lines IMR90, HeLa, hESC and iPSC were derived as discussed in (17) and different datasets for IMR90, HeLa and K562 cell lines were obtained from (18). The summary of the datasets is presented in SI Table 1.

When RO positions were defined by genomic ranges, the middle point of the range was used as the genomic location of the RO. Moreover, to limit the problems associated with technological limitations in sequencing the centromeric regions of chromosomes, the largest replicon of each chromosome (corresponding to the centromeric region) was excluded from the analysis in all the organisms considered.

Probabilities of DFSs were obtained from RO position data using the formulas detailed in the following mathematical derivations. To allowed standardized comparisons in computing the probability of DFS, all the organisms were considered as diploid. Poisson fits of the computationally derived distribution of DFSs were computed using the probability of no DFSs. Poisson fits of the experimental data were computed using the mean (naïve) or by minimizing the difference from the frequencies of DFS strictly larger than zero (filtered). Differences between distributions were computed using Chi-Squared tests.

3) Model derivation and mathematical details

Derivation of central equation

The baseline assumptions that have been used to construct the mathematical model have been described elsewhere (14) and will not be discussed here. In yeast the size of the largest replicon i.e. inter-RO distance is significantly smaller than N_s. This size difference allowed the introduction of approximations, which could be used to obtain simpler formulas in our previous work (14). This is not valid in human genomes, and therefore we could not rely on the approximations previously used. Hence, various quantities had to be re-derived to avoid previously introduced approximations, and we provide the more general derivations below.

Let *D* be the distance between two adjacent ROs located respectively at $n = 0$ and $n = N$, where N-1 is the number of nucleotides within *D*. As shown in (14), the probability of a double stall in *D* '(DSD)' is given by the following expression:

$$\mathrm{Prob(DSD)} = \sum_{n=0}^{N-1} (1-q)^n q[1-(1-q)^{N-n}] \qquad (M1)$$

Therefore

$$\mathrm{Prob(DSD)} = q\sum_{n=0}^{N-1}(1-q)^n - q\sum_{n=0}^{N-1}(1-q)^n(1-q)^{N-n}$$

$$= q\sum_{n=0}^{N-1}(1-q)^n - q\sum_{n=0}^{N-1}(1-q)^N$$

Evaluating the sums using the formula for a geometric series we have,

$$\mathrm{Prob(DSD)} = q\left(\frac{1-(1-q)^N}{1-(1-q)}\right) - Nq\,(1-q)^N$$

$$= 1-(1-q)^N - Nq\,(1-q)^N$$

Thus,

$$\mathrm{Prob(DSD)} = 1-(1+Nq)(1-q)^N \qquad (M2)$$

Expressing the product as the exponential of the sum of the logarithms gives

$$(1-q)^N = \exp(N\log(1-q)) \qquad (M3)$$

Since q is an extremely small number, $\log(1-q) \approx -q$, and hence

$$(1-q)^N = \exp(-Nq) \qquad (M4)$$

Combining Eq. (*M2*) with Eq. (*M4*), we obtain

$$\text{Prob(DSD)} = 1-(1+Nq)\cdot\exp(-Nq) \qquad (M5)$$

Let us define the distance between the adjacent $(k+1)^{\text{th}}$ and k^{th} ROs as N_k. The probability of double stall between this pair of ROs will be denoted as P_k. Thus,

$$P_k = 1-(1+N_kq)\exp(-N_kq) \qquad (M6)$$

The genome-wide probability of no double stall, which will be denoted as Prob(NDS), is given by the product of probability of no double stalls in each replicon, i.e.

$$\text{Prob(NDS)} = \prod_k (1-P_k) \qquad (M7)$$

Combining Eq. (*M6*) and Eq. (*M7*), we have

$$\text{Prob(NDS)} = \left\{\prod_k (1+N_kq)\right\}\cdot\left\{\prod_k (\exp(-N_kq))\right\} \qquad (M8)$$

Let N_g be the genome length, then

$$\sum_k N_k = N_g$$

Thus

$$\prod_k (\exp(-N_kq)) = \exp\left(-q\sum_k N_k\right) = \exp(-qN_g) \qquad (M9)$$

Similarly,

$$\prod_k (1+N_kq) = \prod_k \exp(\log(1+N_k\,q)) = \exp\left(\sum_k \log(1+N_k\,q)\right) \qquad (M10)$$

Therefore, combining (*M*8), (*M*9) and (*M*10) we have

$$\text{Prob(NDS)} = \exp(-q\,N_g)\cdot\exp\left(\sum_k \log(1+N_k\,q)\right)$$

Or

$$\text{Prob(NDS)} = \exp\left(-q\,N_g + \sum_k \log(1+N_k\,q)\right)$$

Let N_s be the median stalling distance, we have shown before (14) that $q = \log(2)/N_s$. Hence

$$\text{Prob(NDS)} = \exp\left(-\frac{\log(2)\,N_g}{N_s} + \sum_k \log\left(1+\frac{\log(2)\,N_k}{N_s}\right)\right) \qquad (M11)$$

As given by Eq. 1 in the main text, where the negative of the quantity in parentheses is denoted by λ. Further derivations are provided in the SI Text.

4) Software used

Data analysis was performed using R version 2.15 and RStudio version 0.98.978 (www.rstudio.com).

Author contributions

MAM and LA provided original concepts, performed mathematical calculations, designed and implemented the computational experiments, analysed the data, and wrote the paper. AM provided original concepts, performed some of the biological experiments, and wrote the paper. JTC performed some of the biological experiments. JJB provided original concepts and wrote the paper. TJN provided original concepts, developed the mathematical model, performed mathematical calculations and wrote the paper.

Acknowledgments

The authors are grateful to Dianbo Liu and Sam Palmer for helpful discussions. AM, JTC and JJB acknowledge support from Cancer Research UK (grant C303/A14301) and the Wellcome Trust (grant WT096598MA). MAM, LA and TJN acknowledge support from the Scottish Universities Life Science Alliance. TJN acknowledges support from the National Institutes of Health (Physical Sciences in Oncology Centers, U54 CA143682). The authors also acknowledge High Performance Computer resources partially supported by the Wellcome Trust (Strategic Grant 097945).

References

1. Nielsen O, Løbner-Olesen A (2008) Once in a lifetime: strategies for preventing re-replication in prokaryotic and eukaryotic cells. *EMBO Rep* 9(2):151–156.

2. Bebenek A (2008) [DNA replication fidelity]. *Postepy Biochem* 54(1):43–56.

3. Blow JJ, Ge XQ, Jackson DA (2011) How dormant origins promote complete genome replication. *Trends Biochem Sci* 36(8):405–414.

4. Sclafani RA, Holzen TM (2007) Cell Cycle Regulation of DNA Replication. *Annu Rev Genet* 41:237–280.

5. Diffley JFX (2011) Quality control in the initiation of eukaryotic DNA replication. *Philos Trans R Soc Lond B Biol Sci* 366(1584):3545–3553.

6. Arias EE, Walter JC (2007) Strength in numbers: preventing rereplication via multiple mechanisms in eukaryotic cells. *Genes Dev* 21(5):497–518.

7. Hastings P, Lupski JR, Rosenberg SM, Ira G (2009) Mechanisms of change in gene copy number. *Nat Rev Genet* 10(8):551–564.

8. Dayal JHS, Albergante L, Newman TJ, South AP (2015) Quantitation of multiclonality in control and drug-treated tumour populations using high-throughput analysis of karyotypic heterogeneity. *Converg Sci Phys Oncol* 1(2):025001.

9. Blow JJ, Dutta A (2005) Preventing re-replication of chromosomal DNA. *Nat Rev Mol Cell Biol* 6(6):476–486.

10. Alver RC, Chadha GS, Blow JJ (2014) The contribution of dormant origins to genome stability: From cell biology to human genetics. *DNA Repair* 19:182–189.

11. Cobb JA, et al. (2005) Replisome instability, fork collapse, and gross chromosomal rearrangements arise synergistically from Mec1 kinase and RecQ helicase mutations. *Genes Dev* 19(24):3055–3069.

12. Ghosal G, Chen J (2013) DNA damage tolerance: a double-edged sword guarding the genome. *Transl Cancer Res* 2(3):107–129.

13. Mazouzi A, Velimezi G, Loizou JI (2014) DNA replication stress: Causes, resolution and disease. *Exp Cell Res* 329(1):85–93.

14. Newman TJ, Mamun MA, Nieduszynski CA, Blow JJ (2013) Replisome stall events have shaped the distribution of replication origins in the genomes of yeasts. *Nucleic Acids Res* 41(21):9705–9718.

15. Costas C, et al. (2011) Genome-wide mapping of Arabidopsis thaliana origins of DNA replication and their associated epigenetic marks. *Nat Struct Mol Biol* 18(3):395–400.

16. Cayrou C, et al. (2011) Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome Res* 21(9):1438–1449.

17. Besnard E, et al. (2012) Unraveling cell type–specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat Struct Mol Biol* 19(8):837–844.

18. Picard F, et al. (2014) The Spatiotemporal Program of DNA Replication Is Associated with Specific Combinations of Chromatin Marks in Human Cells. *PLoS Genet* 10(5):e1004282.

19. Moreno A, et al. (2016) Unreplicated DNA remaining from unperturbed S phases passes through mitosis for resolution in daughter cells. *PNAS*: in press.

20. Siow CC, Nieduszynska SR, Müller CA, Nieduszynski CA (2012) OriDB, the DNA replication origin database updated and extended. *Nucleic Acids Res* 40(D1):D682–D686.

21. Hayashi M, et al. (2007) Genome-wide localization of pre-RC sites and identification of replication origins in fission yeast. *EMBO J* 26(5):1327–1339.

22. Cooper GM (2000) The Eukaryotic Cell Cycle. *The Cell: A Molecular Approach. 2nd Edition.* (Sunderland (MA): Sinauer Associates, Boston University). 2nd Ed. Available at: http://www.ncbi.nlm.nih.gov/books/NBK9876/.

23. Méchali M (2010) Eukaryotic DNA replication origins: many choices for appropriate answers. *Nat Rev Mol Cell Biol* 11(10):728–738.

24. Maya-Mendoza A, Petermann E, Gillespie DAF, Caldecott KW, Jackson DA (2007) Chk1 regulates the density of active replication origins during the vertebrate S phase. *EMBO J* 26(11):2719–2731.

25. Remus D, et al. (2009) Concerted loading of Mcm2-7 double hexamers around DNA during DNA replication origin licensing. *Cell* 139(4):719–730.

26. Evrin C, et al. (2009) A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. *Proc Natl Acad Sci* 106(48):20240–20245.

27. Unno J, et al. (2013) Artemis-dependent DNA double-strand break formation at stalled replication forks. *Cancer Sci* 104(6):703–710.

28. Allen C, Ashley AK, Hromas R, Nickoloff JA (2011) More forks on the road to replication stress recovery. *J Mol Cell Biol* 3(1):4–12.

29. Jones RM, Kotsantis P, Stewart GS, Groth P, Petermann E (2014) BRCA2 and RAD51 Promote Double-Strand Break Formation and Cell Death in Response to Gemcitabine. *Mol Cancer Ther* 13(10):2412–2421.

30. Bohgaki T, Bohgaki M, Hakem R (2010) DNA double-strand break signaling and human disorders. *Genome Integr* 1(1):15.

31. Li H, Mitchell JR, Hasty P (2008) DNA double-strand breaks: a potential causative factor for mammalian aging? *Mech Ageing Dev* 129(7-8):416–424.

32. Francis D, Davies MS, Barlow PW (2008) A strong nucleotypic effect on the cell cycle regardless of ploidy level. *Ann Bot* 101(6):747–757.

33. Wong PG, et al. (2011) Cdc45 limits replicon usage from a low density of preRCs in mammalian cells. *PloS One* 6(3):e17533.

34. Mahbubani HM, Chong JPJ, Chevalier S, Thömmes P, Blow JJ (1997) Cell Cycle Regulation of the Replication Licensing System: Involvement of a Cdk-dependent Inhibitor. *J Cell Biol* 136(1):125–135.

Figure legends

Figure 1. Potential outcomes arising from ROs licensed on a DNA segment. DNA is denoted as a single black line. Prior to S phase entry, four origins (denoted by I, II, III and IV) are licensed by binding a double hexamer of Mcm2-7 proteins (blue). As an origin fires, both Mcm2-7 single hexamers are converted into an active CMG helicase (pink). (a) RO 'II' is dormant and passively replicated by the fork coming from RO 'I'; replication is complete. (b) Red crosses depict the fork stalling. Previously dormant RO 'II' is fired to complete the replication of DNA between stalled forks. However, as there is no RO licensed between RO 'III' and 'IV', the DNA between two stalled forks in this part remains unreplicated and complete replication is compromised. This figure is adapted from previous work (14).

Figure 2: Schematic of the 'central equation'. The genome length is the dominant contributor to the overall replication error due to fork stalling, followed by the number of licensed ROs and lastly by their distribution.

Figure 3: a) Predicted probability of one or more DFSs for various eukaryotic genomes using the 'central equation' from the model. b) Measured mean replicon length across the same genomes from the corresponding experimental datasets. c) Computed *R*-values from the same eukaryotic datasets; note, the dashed bars represent simulated *R*-values for virtual genomes of the same length and RO density, but assuming ROs to be randomly distributed. d) The probability of a DFS, denoted P(DFS), is plotted as a function of increasing replicon length. The estimated median fork-stalling distance, N_s (10 Mbp), is highlighted on the *x*-axis. P(DFS) starts to increase sharply as soon as the replicon size reaches approximately half the value of N_s; note that the *x*-axis has a log scale. e) The calculated probability of a DFS inside replicons plotted against normalized chromosomal lengths for the largest chromosomes in budding yeast, *Drosophila*, *Arabidopsis* and the IMR90 cell-line from two human datasets (B and P).

Figure 4: a) Measured lengths of the largest replicons are shown in each dataset alongside the dashed bars showing the value obtained for virtual genomes of the same length and RO density, but assuming ROs to be randomly distributed. b) The distribution of genome-wide replicon lengths plotted in boxplot format for budding yeast, *Drosophila*, *Arabidopsis* and the IMR90 cell-line from two human datasets (B and P).

Figure 5: Data in the left and right columns is from the IMR90 human datasets B and P respectively. a & b) Frequency of replicons in each cohort; defined according to the following size ranges, $<10^3$ bp = XS, 10^3–10^4 bp = S, 10^4–10^5 bp = M, 10^5–10^6 bp = L, $>10^6$ bp = XL. c & d) Probability of DFS in each cohort of the replicons. e & f) Higher resolution plot of probability of DFS at the transition from "medium (M)" to "large (L)" gap cohorts, contributing most towards the P(DFS); red bars show the bins

with maximum P(DFS) in respective datasets. g & h) Theoretical frequency distribution of replicons inferred from the plots e & f are presented in blue; grey shows the actual frequency distribution in those bins in the data and red highlights the red bins in e & f.

Figure 6: Theoretical prediction for the distribution of the number of DFSs based on the RO positions in each human cell-line datasets (using data from both B and P); also shown, as lines and dots, are best fits to a Poisson distribution.

Figure 7: a) Experimental distribution of three different replicates of 53BP1 nuclear bodies in the IMR90 cell-line fitted with a naïve Poisson (i.e. taking the mean of the data as λ) (gray) and a filtered Poisson (i.e. ignoring the frequncies of zero counts to account for potential error from immunofluorescence staining) (lightgray). The single fitting with the average of the three replicates (not statistically different) is shown. b) Experimental distribution of 53BP1 nuclear bodies in the U2-OS cell-line fitted with a naïve Poisson (i.e. taking the mean of the data as λ) (gray) and a filtered Poisson (i.e. ignoring the frequncies of zero counts to account for potential error from immunofluorescence staining) (lightgray). c) Experimental distribution of UFBs in the U2-OS cell-line fitted with a naïve Poisson (gray) and a filtered Poisson (lightgray). d) Values of the Possion parameter λ obtained from experimental fits of 53BP1 nuclear bodies in IMR90, U2-OS and HeLa; and UFBs in U2-OS and HeLa are compared with theoretical values obtained from different cell lines in Figure 6.

Figure 8: a) Based on the RO distributions in the various human datasets, theoretical predictions of the percentage of cells with DFSs is plotted as a function of the parameter N_s (median stalling distance); the percentage is essentially 100% when $N_s <$ 5 Mbp and this percentage is still non-trivially high even when $N_s >$ 20 Mbp. b & c) Theoretical predictions of the probability of one, two and three DFSs is shown as a function of N_s. d & e) Theoretical predictions of the probability of one, two, or three DFSs is shown as a function of N_s. f & g) Expected numbers of DFSs in different cell-lines are plotted against N_s; in black, blue and red are the experimentally obtained expected number of 53BP1 nuclear bodies in IMR90, U2-OS and HeLa; and UFBs in U2-OS and HeLa cell-lines respectively. Crossing points of the black, blue and red lines over the curves provide an independent estimate for the plausible range of N_s (vertical lines) by directly comparing experimental data with theoretical predictions.

Figure 9: Highlighting the issues faced to maintain small DFS error rates for genomes of increasing length: Theoretical prediction of the average replicon length as a function of increasing genome length, to maintain a fixed probability of DFS, for three different values of this probability; diamonds show the positions of yeast, *Arabidopsis*, *Drosophila* and human respectively, obtained from the datasets of RO positions. The pink shadow highlights the biologically relevant range for mean replicon lengths as per all eukaryotic datasets available. The dashed red line marks the footprint for the MCM2-7 double hexamer, below which any replicon length is biologically unrealistic.