5

A single unresolved G4 quadruplex structure spawns multiple genomic rearrangements during animal development

Lemmens BB, van Schendel R, Tijsterman M.

Abstract

Faithfull DNA replication is crucial to prevent genetic heterogeneity and malignant transformation. How cells deal with endogenous DNA lesions and DNA secondary structures that hamper DNA replication is poorly understood, especially in biological contexts in which the lesions are insufficient in amount to cause cell cycle arrest. Here, we studied the genetic consequences of persistent DNA secondary structures during *C. elegans* development and found that a single unresolved G4 quadruplex can survive mitotic divisions, causing numerous deletions in the descending cells. We demonstrate that DOG-1/FANCJ is the principle helicase to resolve G4 quadruplexes in *C. elegans* and that in its absence these endogenous DNA secondary structures persist *in vivo* to create substrates for polymerase Theta-Mediated End Joining, providing further mechanistic insight in the mode of mutagenesis of G4 structures. Our data indicate that low frequency replication barriers escape detection and active processing, allowing them to cause multiple genomic rearrangements in proliferating tissues

Introduction

Every time a cell divides it has to replicate its entire genome to provide both daughter cells with equal amounts of genetic material and it has to do so in an accurate manner to ensure genome stability. This is especially important for multicellular organisms where the single-cell embryo has to replicate its genome several times to form all adult tissues, risking the accumulation of genetic alterations that may result in cellular dysfunction and/or malignant transformation.

Replication fidelity is maintained by the use of highly accurate replicative polymerases, which are double-checked by efficient mismatch and post-replication repair mechanisms (Loeb and Monnat 2008; Jiricny 2013). Additionally, cells need to deal with damaged DNA bases and DNA secondary structures that can obstruct the replicative polymerase and thus may hamper the cell to replicate its genome successfully (Budzowska and Kanaar 2009). Replication errors are a major source of spontaneous mutagenesis and promote cancer progression, yet how cells deal with replication impediments and how this impacts the mutational landscape is still unclear (Preston *et al.* 2010). Especially, how low frequency replication barriers affect genome stability during development is largely unknown.

We have recently shown that endogenous DNA damage as well as DNA secondary structures like G4 quadruplexes can cause genomic rearrangements during *C. elegans* development and evolution, and that deficiencies in specialized DNA polymerases or DNA helicases that deal with replication barriers can result in high levels of genomic instability in this animal (Koole *et al.* 2014; Roerink *et al.* 2014). In fact, animals lacking the FANCJ helicase DOG-1 show a thousand fold increase in deletion formation at G4 sites, illustrating the mutagenic potential of these endogenous DNA sequences and revealing the importance of such helicases to ensure genome stability (De and Michor 2011; Koole *et al.* 2014). We recently also have uncovered a novel error-prone DSB repair pathway that limits the consequences of such replication blocks, which we named Theta-mediated end joining (TMEJ), as it requires the well-conserved DNA polymerase Theta/POLQ-1 (Koole *et al.* 2014). Genetic deficiencies either leading to inefficient bypass of damaged DNA bases or the inability to resolve G4 quadruplexes typically resulted in 50-300 base pair (bp) deletions throughout the genome, which all were the product of TMEJ (Kruisselbrink *et al.* 2008; Koole *et al.* 2014; Roerink *et al.* 2014). It has been proposed that these deletions arise because unresolved replication blocks lead to DNA double-strand breaks (DSBs) that are subsequently repaired by TMEJ. However, at present, nothing is known about the fate of the initial blocking lesion, nor how these lesions induce DSBs that fuel TMEJ-dependent deletions.

G4 quadruplexes have been shown to be potent replication blocks *in vitro (Howell et al. 1996; Han et al. 1999; Edwards et al. 2014)*, yet little is known about their impact on replication progression *in vivo*. Recent insights on UV-induced DNA damage suggest that

local impairments of polymerase activity are unlikely to stall overall replication progression, as most DNA will be replicated via the convergence of opposing replication forks or via re-priming mechanisms downstream of the lesion (Blow *et al.* 2011; Elvers *et al.* 2011; Helleday 2013). Both scenarios promote gross DNA replication but are predicted to result in small single-strand (ss) DNA gaps opposite to the blocking lesions. How these ssDNA gaps are detected and resolved *in vivo* is still an open question, but the typical small 50-300bp deletions directly flanking replication-stalling G4 sites could imply that such gapped structures are converted into TMEJ substrates (Koole *et al.* 2014; Roerink *et al.* 2014).

Two fundamentally different models might link polymerase blockage to deletion formation: i) a direct conversion model, in which the blocking lesion and the resultant gapped DNA structure are directly converted into a DSB, either because ssDNA is intrinsically unstable or via active processing by nucleases, or ii) a persistent lesion model, in which the blocking lesion escapes detection/processing and the opposing ssDNA gap persists, eventually resulting in a DSB during the next S-phase as the ssDNA gap is replicated (Figure 1A).

Both scenarios can explain the occurrence of DSBs that could fuel TMEJ-dependent deletions, yet their consequences for animal development are fundamentally different. If unresolved replication barriers are converted directly into DSBs, they are lost during the repair process and pose a threat to genome stability only once. However, if unresolved replication barriers escape detection and persist during the life of the individual, they can continuously fuel genomic instability among dividing cells.

Here we provide evidence for a persistent lesion model, in which low-frequency replication blocks result in pre-mutagenic lesions (e.g. 50-300bp ssDNA gaps) that are not detected nor processed, yet result in TMEJ-dependent deletions in subsequent replication rounds. By combining in-depth deletion footprint analysis and transgenic reporter assays we found that in the absence of DOG-1 helicase activity, G4 quadruplexes persist *in vivo* and create pre-mutagenic lesions that in turn spawn multiple deletions throughout animal development. This work provides a framework to investigate the consequences of low frequency replication barriers for animal development and reveals that a single persistent DNA structure can lead to multiple genomic rearrangements in proliferating tissues – a feature that may significantly contribute to tumor heterogeneity and malignant transformation.

Results

MUS-81 or XPF-1 nucleases are not required for G4-induced deletion formation

Recent studies suggest that hard-to-replicate loci known as “fragile sites” can be processed by specialized topoisomerase complexes as well as structure-specific nucleases (*i.e.* Sgs1/Top3/Rmi1 and Mus81 respectively), which either resolve stalled replication intermediates directly or convert them into DSBs that are repaired later on (Manthei and Keck 2013; Minocherhomji

and Hickson 2014). Although the exact nature of these nuclease/topoisomerase substrates is not known, we hypothesized that similar mechanisms may promote DSB induction at hard to replicate G4 sites, creating a substrate for TMEJ.

To investigate the genetic requirements for G4-induced deletion formation, we first studied the role of two highly conserved structure-specific nucleases, MUS81 and XPF. These nucleases have been implicated in DNA break formation at (yet-undefined) late replication intermediates and animals deficient for *xpf-1* and *mus-81* show elevated levels on genome instability in mitotic cells (Agostinho *et al.* 2013; O'Neil *et al.* 2013; Saito *et al.* 2013; Minocherhomji and Hickson 2014). We analyzed stochastic deletion formation in *dog-1* animals lacking either *mus-81* or *xpf-1* using a PCR-based assay that amplifies deletion products at endogenous G4 sites (e.g. Qua213 on chromosome I). Depletion of *mus-81* or *xpf-1* did not prevent G4 deletion formation in *dog-1* animals (Figure 1B). MUS-81 and XPF-1 are known to have (partially) redundant functions, which is exemplified by the fact that *mus-81; xpf-1* double mutants are synthetic lethal (Agostinho *et al.* 2013; O'Neil *et al.* 2013; Saito *et al.* 2013). However, the sensitivity of the PCR-based assay and the mutagenicity of such G4 sites allowed us to test G4 deletion formation in the few survivors of *dog-1 mus-81; xpf-1* triple mutants and found them still proficient in G4-induced deletion formation (Figure 1B). We therefore conclude that MUS-81 and XPF-1 are not required for G4-induced deletion formation.

Similarly, previous studies have found that neither the nuclease scaffold protein SLX-4 nor the Sgs1 homolog HIM-6 are needed for G4-induced deletion formation, further supporting the notion that unresolved G4 structures may not require direct processing to become mutagenic (Youds *et al.* 2006; Saito *et al.* 2009).

G4 deletion frequency reveals the presence of a persistent pre-mutagenic substrate

We next investigated a scenario where G4 quadruplexes are not processed in the absence of the DOG-1 helicase and therefore will persist during subsequent cellular divisions. In that case, the G4 will impose a physical block for the replicative polymerase during every S-phase, resulting in small ssDNA gaps in the nascent strand opposite of the stable G4. If these gaps are left unfilled, they are expected to collapse into DSBs during the next replication round, providing a substrate for TMEJ (Figure 1A). While TMEJ can repair the DSBs arising from the ssDNA gaps, it does not take away the replication-stalling lesion that caused the gap: the G4 quadruplex. This model therefore predicts that unresolved G4 structures will persist independent of the repair process and can cause new mutagenic events every time the cell divides.

A fundamental difference between the direct conversion model and the persistent lesion model is the fate of the G4 bearing DNA strand. In the first model the G4 bearing DNA strand is processed to form a deletion product, while in the latter scenario the G4 persists

and the gapped nascent strand provides the substrate for TMEJ (Figure 1A). Next to the obvious cellular implications, this difference in G4 fate will also affect the deletion frequency distribution within the animal population. While the direct conversion model predicts that an animal showing a deletion has lost the G4 structure, the persistent lesion model predicts that an animal showing a deletion still harbors a mutagenic G4, substantially increasing the chance of the occurrence of another deletion.

We set out to distinguish between these models by determining the stochastic deletion frequency of endogenous G4 sites per animal and correlate that to the frequency of animals bearing two unique deletions at the same genomic locus. The persistent lesion model predicts that the probability (and thus frequency) of a second deletion would be much higher than expected based on random chance. To be able to measure stochastic deletion frequencies among *dog-1* animals and at the same time capture multiple independent G4 deletion events within these animals, we devised a PCR-based assay on minimal dilutions of genomic DNA lysates (Figure 1C). In short, we extracted genomic DNA from individual animals (each consisting of ~900 somatic cells) and performed two independent nested PCRs, each on 1% of lysate, amplifying G4 sites on genetic material corresponding to ~9 somatic cells. To obtain stochastic deletions frequencies we extracted DNA from >190 *dog-1* proficient and >330 *dog-1* mutant animals and analyzed two endogenous G4 loci (*i.e.* Qua375 on chromosome I and Qua1277 on chromosome IV). For each locus, two independent PCR reactions were analyzed in parallel and unique deletion products were discriminated based on size by gel-electrophoreses (Figure 1C).

Using this approach we frequently detected two differently sized deletions per animals, suggesting that G4 deletions often co-occur during animal development (Figure 1D, highlighted in red). The identification of two different deletions per animal also indicated that the small sample size (corresponding to ~9 somatic cells) allowed us to amplify unique molecules and exclude potentially abundant/favored deletion products. While we found 4% and 6% of the *dog-1* deficient samples to have deletions at Qua375 and Qua1277, respectively, we detected no deletions in the *dog-1* proficient samples. Moreover, the deletions were randomly distributed over the animal population and no significant correlation was found between deletions at the two G4 loci (Figure S1). These data confirm DOG-1 as a potent suppressor of G4 deletions and indicate that we detected stochastic *in vivo* rearrangements.

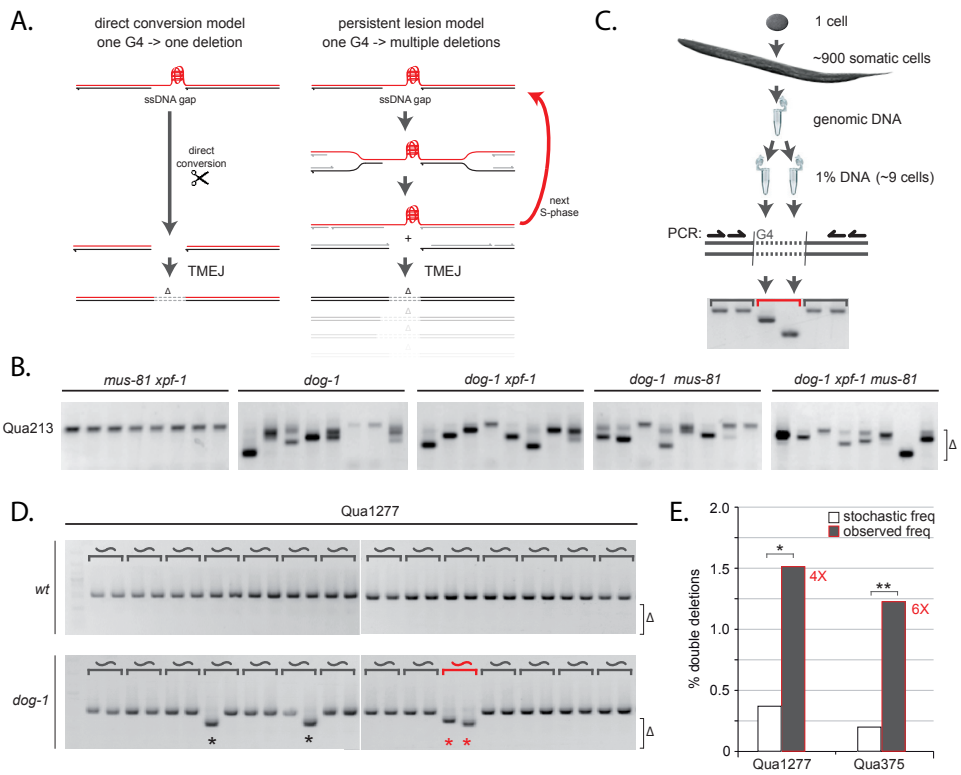

Figure 1. Overrepresentation of co-occurring G4 deletions in single animals

A. Two putative models for G4-induced deletions formation. In the "direct conversion model" the mutagenic G4 structure is lost during deletion formation, unlike the "persistent lesion model" where the G4 structure is maintained and can spawn multiple deletions in descending cells **B.** PCR analysis of G4 instability at endogenous G4 site Qua213; each well represents an independent PCR reaction on 10% single worm lysate; size-range of PCR-amplified deletions products is indicated by Δ. **C.** Schematic overview of PCR-based setup to identify multiple G4 deletions in single animals using parallel nested PCR reactions on 1% single worm lysates **D.** PCR analysis of G4 instability at endogenous G4 site Qua1277 in *dog-1* proficient (upper panel) and deficient animals (lower panel); Representative gel images are depicted used to determine the stochastic deletion frequencies among individual animals (asterisks) and directly extract the frequency of double deletions (red); each well represents an independent PCR reaction on 1% single worm lysate; size-range of PCR-amplified deletions products is indicated by Δ. ***E.*** Histogram depicts double deletion frequencies in 1% single animal lysates, as determined using the PCR-based assay depicted in Figure 1C. White bars represent expected double deletion frequencies based on the obtained stochastic deletion frequencies within the tested animal population (See methods section for details). Black bars represent double deletion frequencies that were identified directly (as highlighted in red in figure 1D). Asterisks indicate highly significant over-representation of the observed double deletion events within the tested population (*n=352 **n=576) as determined by hypergeometric testing (*p<0.003 **p<0.001).

To obtain the expected frequency of animals carrying two independent stochastic deletions we considered the two genomic fractions as independent tests and multiplied the deletion frequency in the first sample to the frequency of unique deletions in the second sample (Figure 1E and see methods section for details). This data set also allowed us to extract the double deletion frequency directly, as we readily detected single animals with two differently sized deletions (Figure 1D, highlighted in red). Strikingly, the actual frequency of double deletions was four to six fold higher than predicted based on random chance ($p<0.003$), indicating that an animal carrying a deletion has a much higher chance to acquire another deletion than one would expect based on the deletion frequency in the population (Figure 1E). The observation that animals bearing a deletion are predisposed to have an additional deletion at the same genomic locus strongly argues for the existence of a pre-mutagenic lesion. The continued mutagenicity of the pre-mutagenic lesion implies that the source of genomic instability is not resolved, but instead persists during animal development, allowing it to cause multiple DSBs that are processed into deletions through TMEJ.

Unresolved G4 structures cause multiple site-specific deletions in single animals

If in the absence of DOG-1 helicase activity G4 structures persist and remain mutagenic, one would predict that the G4 structures that arise early in development would cause many different deletions within one animal. To see if we could identify more than two G4 deletions in individual *dog-1* animals we subjected the single animal lysates to multiple deletion tests (Figure 2A). As predicted by the persistent lesion model, lysates positive for a Qua375 deletion (indicative of a persistent G4) often showed extra deletions at this very same locus (5/14= 36%). In contrast, lysates previously found negative for Qua375 deletions rarely showed additional deletions (1/16= 6%). Similarly, we subjected lysates having two different Qua213 deletions (indicative of potential early events) to multiple testing rounds and identified several animals that had three or even five differently sized deletions among eleven test samples (5/18= 27%), while lysates lacking Qua213 deletions in the first two tests displayed no additional deletions (0/5= 0%) (Figure 2C). These observations are in line with the presence of a pre-mutagenic substrate only in a subset of animals and suggest that G4-induced genomic rearrangements can be very abundant in certain individuals, likely when the G4 structure manifested early in the cell lineage.

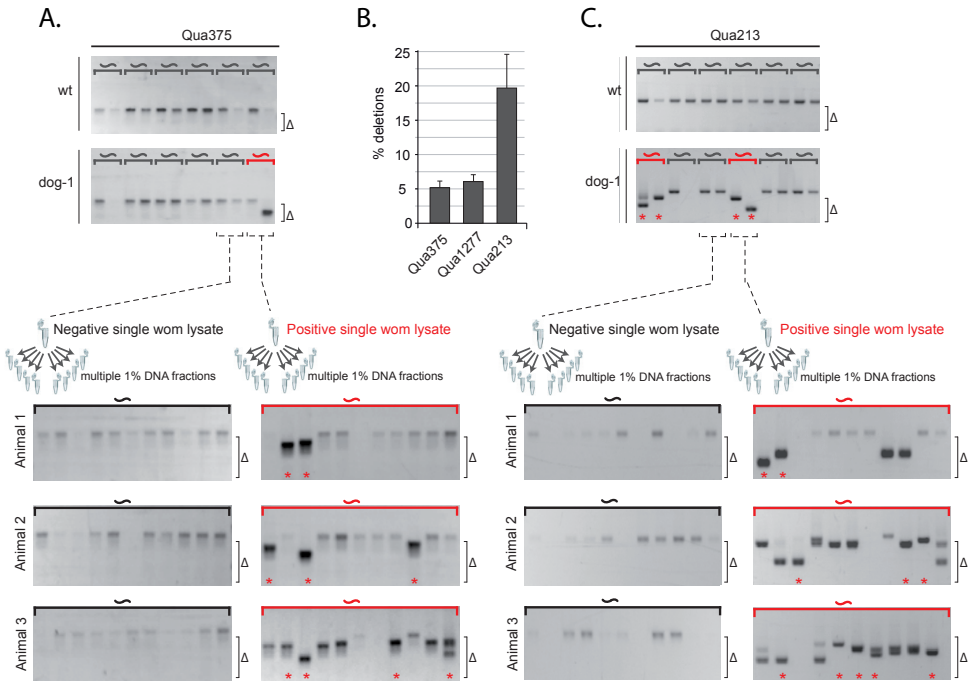

Figure 2. Co-occurrence of multiple locus-specific deletions in single animals
A. PCR analysis of G4 instability at endogenous G4 site Qua375 in wild-type and *dog-1* animals, upper and lower panel, respectively; each well represents an independent PCR reaction on 1% single worm lysate; size-range of PCR-amplified deletions products is indicated by Δ. Single worm lysates were first categorized based on the presence or absence of G4 deletions in two PCR test round and subsequently tested eleven times to test for additional G4 deletions. Three representative gel images for each category are depicted. Asterisk indicates unique deletion events based on product size. **B.** Histogram depicts double deletion frequencies in 1% single animal lysates, as determined using the PCR-based assay as depicted in Figure 1C. **C.** PCR analysis of G4 instability at endogenous G4 site Qua213 in wild-type and *dog-1* animals, upper and lower panel, respectively; each well represents an independent PCR reaction on 1% single worm lysate; size-range of PCR-amplified deletions products is indicated by Δ. Single worm lysates were first categorized based on the presence or absence of two unique G4 deletions in two PCR test round and subsequently tested eleven times to extract additional G4 deletions. Three representative gel images for each category are depicted. Asterisk indicates unique deletion events based on product size.

Recurrent 3′ deletion junctions in single animals argue for persistent replication blocks during development

Our observations thus far indicated that animals that cannot resolve G4 structures accumulate multiple site-specific deletions, suggestive of a local persistent entity that provokes deletion formation. To determine if the multiple deletion events within one animal were derived from a single persistent G4 structure, we needed to find unique characteristics of the causing lesion. Current technology does not allow direct monitoring of a single replication-blocking lesion

within a developing animal, a problem that has challenged research on endogenous (low frequency) replication barriers for years.

We nevertheless reasoned that each replication block could leave a unique genetic “scar” that would unveil its presence and exact genomic location. In fact, the exact position of G4 deletion junctions can serve as a distinctive mark, as we previously found that deletions at G4 sites are typically unidirectional with their 3′ junction close to the start of the G4 motif (Koole *et al.* 2014). The typical position of the 3′ deletion junctions, right at the start of the G4 motif, plausibly reflects the collision of the replicative polymerase with the stable G4 quadruplex (Figure 3A).

To more precisely mark the start of a G4 deletion and correlate that to the potential configuration of a stable quadruplex fold, we determined the spectra of 3′ junctions of G4 deletions at two endogenous G4 sites that can adopt only one possible three-stacked quadruplex configuration, Qua1465 and Qua1466, and indeed found a very sharp distribution of 3′ junctions immediately flanking the 3′ outermost G of the G4 motif (Figure 3B). This narrow distribution of 3′ deletion junctions suggests that the position of the replication block at minimal G4 sites is relatively fixed.

The genome, however, also harbors many mutagenic G4 sites like Qua915 or Qua1277 that can adopt many different quadruplex configurations that satisfy the G4 consensus G_{3-5}-N_{1-3}-G_{3-5}-N_{1-3}-G_{3-5}-N_{1-3}-G_{3-5} (Kruisselbrink *et al.* 2008). Such G4 sites may inflict differently positioned replication barriers within the population, which would provide a window to discriminate between individual DNA secondary structures that arise during animal development. In line with that notion, 3′ deletion junctions at Qua915 and Qua1277 show a much broader spectrum than seen at minimal G4 sites and are found not only at the 3′ flank of the G4 motif but also within the G4 sequence itself, reaching up to the 3′ border of the outermost minimal G4 motif (Figure 3B). The observation that most deletion junctions at Qua915 and 1277 are not positioned immediately upstream of the G4 sequence (like with minimal G4 loci), but instead are located within the G4 sequence yet upstream of the minimal G4 consensus, strongly suggests that the pre-mutagenic lesion is in fact a DNA secondary structure.

If an unresolved G4 structure will not be processed and will persist during the animal’s lifetime, it is predicted to block the replicative DNA polymerase at the very same genomic position in subsequent replication cycles. Importantly, the strong correlation between the anticipated G4 structure and the 3′ deletion junctions allows us to test this hypothesis: If the position of G4 quadruplexes varies among individuals, but such a DNA structure persists within an individual, one expects to find variable 3′ deletion junctions within the population but recurrent 3′ junctions within single animals (Figure 3A). Our PCR-based assay allows us to detect multiple G4 deletions in individual animals and analyze the deletion footprints at the nucleotide level. We first analyzed the 3′ deletion junctions of five animals in which we found two different Qua1277 deletions and plotted their relative position to the G4 motif

(Figure 3C). While the position of the 3′ deletion junctions varied significantly between the different animals, the deletions within the animals had a very similar position of the 3′ junction, suggesting that the G4 structure indeed persisted during animal development and spawned different deletions.

Although the distributions of 3′ deletion junctions are narrow at minimal G4 motifs, they still spread a few nucleotides, implying variable polymerase progression upon G4 collision and/or minimal DNA processing. This intrinsic variability complicates the identification of persistent lesions as it obscures the translation of fixed G4 structures to deletion footprints. To facilitate the discrimination between stochastic G4 folding events and persistent G4 structures we searched for endogenous G4 sites that showed a bimodal distribution of 3′ deletion junctions. Bimodal junction spectra imply two distinct G4 folding possibilities that are spatially separated enough to identify two differently positioned replication blocks. The endogenous G4 site Qua375 could fulfill such criteria since it consists of a minimal G4 sequence (G_{15}) and an upstream guanine triplet (G_3) that are separated by 4 nucleotides.

We analyzed Qua375 deletions in >20 *dog-1* animals and determined the 3′ deletion junctions and found that deletions at Qua375 indeed can occur right next to the G_{15} tract, but primarily start in front of the extra guanine triplet, thus >7 nucleotides away from the minimal G4 motif (Figure 3A). The resultant bimodal distribution of 3′ deletion junctions indicated that quadruplexes at Qua375 adopt two distinct folding types: one consisting merely of the minimal G_{15} tract and another type that includes both the G-tract and the guanine triplet (Figure 3A, right panel). After characterizing stochastic Qua375 deletions among individual *dog-1* animals, we set out to determine deletion junctions within single animals. Like for Qua1277, we extracted two Qua375 deletion products per animals using our PCR assay on 1% DNA fractions and determined the 3′ deletion junctions (Figure 3B). Strikingly, all deletion pairs from individual animals (12/12) were from the same type, *i.e.* the position of the first deletion had perfect predictive value for the position of the second deletion (p=0.004). This strong correlation between 3′ deletion junctions within individuals was not limited to two G4 deletions, as also the cases having three or four different Qua375 deletions had matching 3′ junctions (p=0.003) (Figure 3C). The highly significant correlation between 3′ deletion junctions within single animals and the typical position next to G4 motifs strongly suggest that these genomic deletions have a common ancestor and that their position is determined by the anticipated polymerase block - the 3′ leg of the quadruplex fold. These observations together with the overrepresentation of double deletions within animals (Figure 1E) provide further evidence that G4 structures can arise during animal development and argue strongly for a scenario in which the DOG-1/FANCJ helicase is essential to resolve DNA secondary structures that would otherwise persist throughout the animal's life and cause multiple deletions.

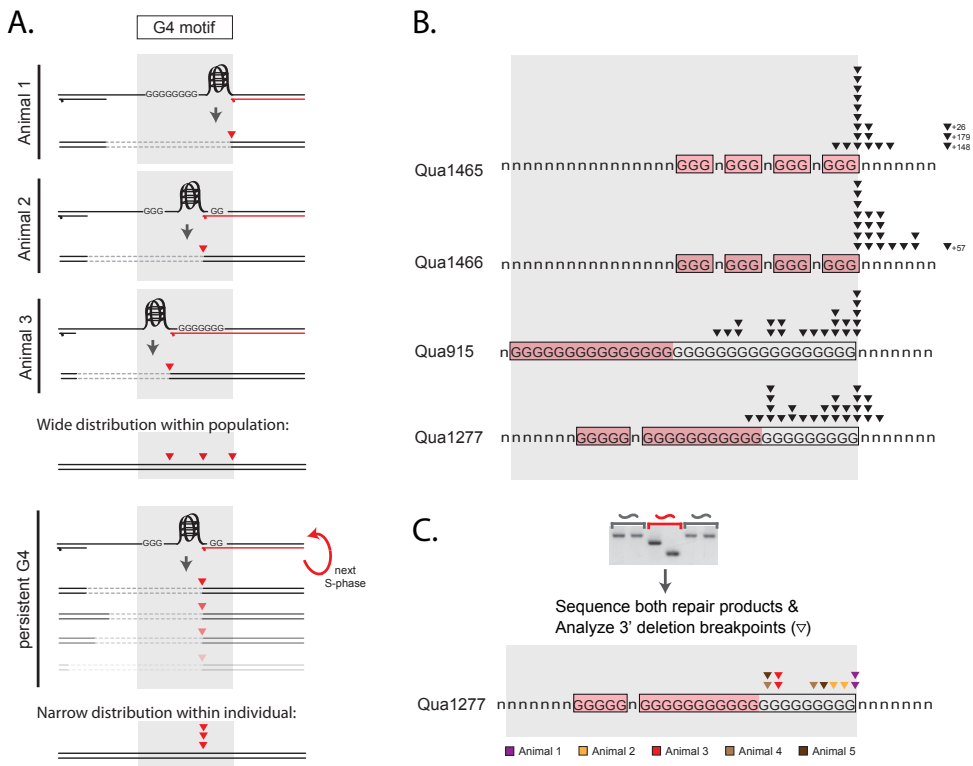

Figure 3. Spectra of 3′ deletion junctions correlate with G4 quadruplex configurations A. Schematic representation of possible G4 quadruplex positions in G-rich motifs and its strong association with 3′ deletion junction position (red triangle). While stochastic G4 folding events within the population are predicted to occur randomly along the possible folding positions and thus produce a wide spectrum of 3′ deletion junctions, persistent G4 structures are predicted to stay at one position and thus cause a very narrow spectrum of 3′ deletion junctions within individuals. **B.** Spectra of 3′ deletion junctions at indicated G4 loci. Each black triangle represents a 3′ deletion junction identified in a unique individual. **C.** Experimental setup and results of 3′ deletion junction analysis at Qua1277 in single *dog-1* deficient animals. The 3′ deletion junctions identified in 1% lysate fractions of the same individual are color-coded as indicated by the legends.

Although 3′ junctions of G4 deletions within an animal are very similar, no significant correlation was found for deletion size (Figure 4D). This observation suggests that the blocking lesion persists, but the resultant pre-mutagenic substrates (e.g. ssDNA gaps) can vary, leading to stochastic 5′ junctions. Similarly, we did not find any predictive value for the presence of flank insertions among double deletions, suggesting that subsequent repair events also occur on uniquely formed substrates. All together, these data imply that a single G4 structure can persist and give rise to various lesions that are processed independently by TMEJ.

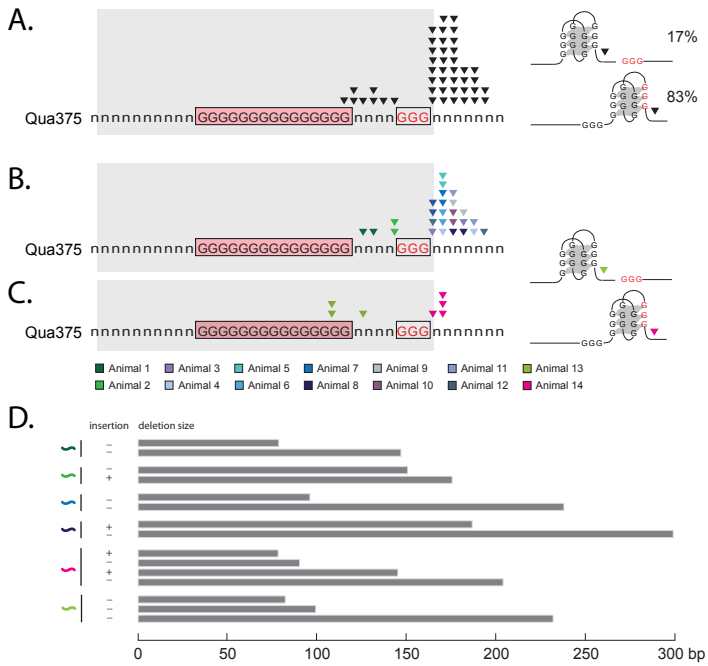

Figure 4. Recurrent 3′ deletion junctions at endogenous G4 site in single animals
A. Bimodal spectrum of 3′ deletion junctions at Qua375. Each black triangle represents a 3′ deletion junction identified in a unique individual. Illustrations on the right portray expected G4 quadruplex configurations at Qua375 and their relative frequencies **B and C.** Results of 3′ deletion junction analysis at Qua375 in single *dog-1* deficient animals. The 3′ deletion junctions identified in 1% lysate fractions of the same individual are color-coded as indicated. Illustrations on the right portray expected G4 quadruplex configurations at Qua375 in animal 13 and 14 in which several recurrent 3′ deletion junctions were identified. **D.** Histogram depicts sizes of G4 deletions in the individual animals of which the 3′ deletion junctions are depicted above. Individuals are color-coded as indicated in Figure 4C. Presence of flank insertions is indicated (+/-).

Visualization of multiple G4 deletion events during organ development

To substantiate our PCR-based findings we set out to visualize multiple G4 deletion events directly in single animals using a LacZ-based reporter system. We constructed transgenic animals that express LacZ only when a G4-induced deletion brings the reporter ORF in frame with the upstream ATG start codon (Figure 5A). The G4 reporter construct is driven by a *myo-2* promoter, which results in specific expression in pharyngeal muscle cells. The *C. elegans* pharynx is a well-characterized organ composed of cells with different embryonic origins and its muscle cells undergo approximately ten mitotic divisions after fertilization of the zygote. We hypothesized that also these cells could be subject to persistent lesions, resulting in multiple deletions in different cells of the pharynx.

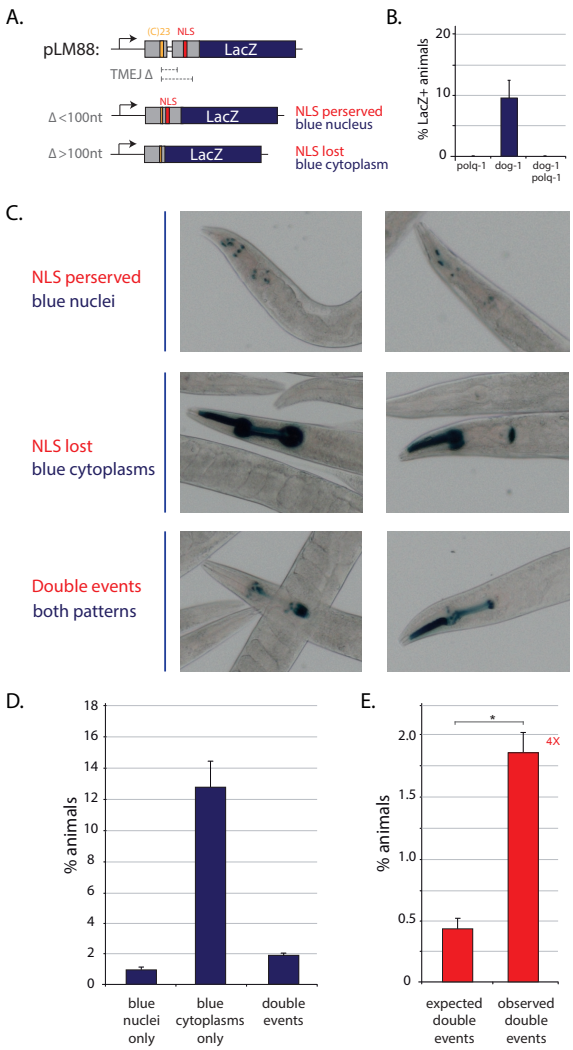

Figure 5. Transgenic G4 instability reporter reveals co-occurring deletions of different sizes during pharynx development

A. Schematic diagram of G4 instability reporter pLM88 and repair outcomes via TMEJ. While relative small deletions will keep the downstream nuclear localization signal (NLS) intact, larger deletions will exclude the NLS from the restored LacZ ORF. **B.** Histogram shows quantification of stochastic pLM88 ORF correction in three asynchronous populations measured by the percentage of LacZ positive animals of the indicated genotype. Average percentage of LacZ-positive animals of three independent experiments is depicted and error bars represent S.E.M. **C.** Representative pictures of stochastic LacZ expression patterns of *dog-1* deficient pLM88 animals. **D.** Histogram depicts quantification of stochastic LacZ patterns in synchronized populations of *dog-1* deficient pLM88 animals (as illustrated in figure 5C). Average percentage of LacZ-positive animals of three independent experiments is depicted and error bars represent S.E.M. **E.** Comparison of the observed frequency of double events and the expected frequency of double staining patterns based on the frequency of the individual nuclear/cytoplasmic LacZ patterns. Asterisk indicates highly significant over-representation of the observed double events within the population (n=1933) as determined by hypergeometric testing ($p<0.001$).

To discriminate unique deletion events in single animals we exploited the fact that deletions arising from a single replication block can be of different size; we cloned a nuclear localization signal (NLS) 90bp downstream of the G4 motif, allowing small deletions to keep the NLS intact and big deletions to exclude the NLS but still render the LacZ reporter gene functional. As expected, *dog-1* deficiency resulted in a stark increase in LacZ expressing cells and we detected both nuclear and cytoplasmic staining patterns (Figure 5B and 5C). In line with previous reports showing that >75% of G4 deletions are bigger than 100bp (Koole *et al.* 2014), we found the vast majority of LacZ expressing animals to display cytoplasmic LacZ patterns (~80%), indicating that often the NLS was lost (Figure 5D). Moreover, both LacZ patterns were *polq-1* dependent, indicating that TMEJ was indeed responsible for these deletion events (Figure 5B). We next determined the frequency of LacZ ORF correction events and found 15% of the animals to display cytoplasmic staining patterns and 3% to have nuclear LacZ staining patterns. Notably, animals harboring cells with nuclear LacZ often displayed additional cells with cytoplasmic staining, suggesting that they suffered (at least) two unique G4 deletion events (Figure 5D). Importantly, the frequency of such double events was significantly higher (>4 fold) than expected based on the stochastic frequencies of the individual patterns ($p<0.001$), confirming that G4 deletions co-occur during animal development (Figure 5E). These observations imply that somatic cells are prone to suffer from a G4 deletion when other cells in the lineage sustained a deletion at that very same locus, substantiating the notion that *dog-1* deficient animals harbor local pre-mutagenic substrates that promote the formation of multiple G4 deletions in descending cells. All together, these data argue that unresolved G4 structures can persist during animal development and present a potent source of genetic mosaicism in somatic tissues.

Discussion

Faithfull DNA replication is crucial to all life forms and is continuously challenged by lesions that can obstruct the replicative polymerase. Endogenous replication barriers are thought to be a major source of spontaneous mutagenesis, potentially promoting malignant transformation. However, little is known about the consequences of endogenous replication barriers for animal development, especially when present at low physiological levels. In recent years, several defense mechanisms are identified that promote DNA synthesis past replication barriers, including specialized translesion polymerases that can bypass damaged DNA bases (e.g. POLH/POLH-1) and helicases that can resolve DNA secondary structures (e.g. FANCJ/DOG-1). Deficiencies in these translesion polymerases or DNA helicases result in cancer prepositions syndromes in humans and increased spontaneous mutagenesis in *C. elegans* (Masutani *et al.* 1999; Levitus *et al.* 2005; Koole *et al.* 2014; Roerink *et al.* 2014). Recent studies imply a major role for TMEJ in repairing DSBs inflicted by unresolved replication

barriers, typically resulting in deletions of 50-300bp in size (Koole *et al.* 2014; Roerink *et al.* 2014). Yet how unresolved replication barriers promote the formation of genomic deletions on a mechanistic level remained elusive. Here, we provide evidence for a model in which low frequency replication barriers escape detection and direct processing, allowing them to spawn multiple TMEJ-mediated genomic rearrangements in descending cells.

We studied the mutagenic role of endogenous G4 quadruplexes: guanine-rich DNA secondary structures that are very stable under physiological conditions and potently block replication *in vitro* (Howell *et al.* 1996; Han *et al.* 1999; Huppert 2010). While the biochemical properties of G4 quadruplexes have been well characterized, their formation and putative function *in vivo* has remained elusive and a subject of debate ever since they were first described (Guschlbauer *et al.* 1990). While recent studies suggest that G4 quadruplexes may be functionally important to regulate transcription, DNA replication and/or telomere maintenance, the use of thermodynamically stable G4 quadruplexes as regulatory entities is controversial, mainly because such structures are intrinsically recombinogenic and would pose serious problems to many DNA metabolic processes (Tarsounas and Tijsterman 2013). Several DNA helicases are reported to resolve G4 quadruplexes *in vitro* (including *FANCJ, PIF1* and *BLM*), presenting several conceivable defense mechanisms to neutralize the malicious properties of G4 quadruplexes.

The data presented here indicate that G4 quadruplexes persist during *C. elegans* development in the absence of the DOG-1/FANCJ helicase, implying that effective redundant activities that can resolve these stable DNA structures *in vivo* are absent. Moreover, the persistent mutagenicity of G4 motifs suggests that G4 quadruplexes can be highly stable throughout the animal's lifetime and support the notion that these enigmatic DNA structures do occur *in vivo*.

The frequency and 3′ junction analysis of G4 deletions presented here argue that G4 quadruplexes arise stochastically within the animal population but remain stable in *dog-1* deficient individuals, resulting in several TMEJ-mediated deletions during animal development. These findings were substantiated by a LacZ-based reporter system able to detect different G4 deletions in single animals, showing that G4 deletions are prone to co-occur in somatic tissues. We propose a model where unresolved G4 quadruplexes block the replicative polymerase, leading to short stretches of unreplicated DNA opposite the persistent G4 structure (*i.e.* 50-300bp ssDNA gaps). These ssDNA gaps present pre-mutagenic lesions that might not stall mitotic division. Upon a second round of replication, the ssDNA gaps are converted into replication-born DSBs, which via TMEJ result into 50-300bp deletions. Importantly, the G4-bearing strand remains intact during the process of deletion formation, causing the replicative polymerase to run again into the persistent G4 quadruplex and create a new 50-300bp ssDNA gap. In agreement with this model we found the co-occurring deletions in single animals to have matching 3′ junctions (suggestive of a persistent polymerase block) but stochastic 5′

junctions (suggestive of independent ssDNA gap sizes/DSB substrates). Furthermore, the high co-occurrence of G4 deletions in single animals argues that the mutagenic lesions are not lost during deletion formation but instead promote G4 instability during animal development.

The persistence of unresolved G4 quadruplexes also implies that these DNA secondary structures are not processed directly by structure-specific nucleases to create DSBs and TMEJ-mediated deletions. The data presented here thus suggest that the biological consequences of low frequency replication barriers are fundamentally different to the replication stress induced via overall DNA polymerase inhibition (e.g. by aphidicolin) or nucleotide depletion (e.g. by hydroxyurea). The latter treatments impede efficient DNA replication, resulting in DNA breaks and gaps at hard to replicate chromosomal regions called “fragile sites”. Recent data indicate that break formation at fragile sites is an active process that requires structure-specific nucleases such as MUS81 and XPF (Minocherhomji and Hickson 2014). Here we found deletion formation at endogenous G4 sites to occur independent of these nucleases, suggestive of a different source of genome instability. We propose that inhibition of DNA synthesis by DNA polymerase inhibitors or nucleotide depletion may result in large sections of unreplicated DNA that are subject to nuclease cleavage, while low frequency replication barriers might inflict relatively small ssDNA gaps that escape detection and nuclease cleavage. This latter attribute allows unresolved replication barriers to persist during mitotic divisions and cause multiple ssDNA gaps that induce DSBs in subsequent replication rounds, ultimately resulting in numerous genomic rearrangements in proliferating tissues.

The persistent lesion model, as presented here, also provides an explanation why replication-born DSBs caused by unresolved replication barriers are not repaired via error-free mechanisms like homologous recombination (HR), but instead are repaired via mutagenic pathways such as TMEJ. If unresolved replication barriers remain present during the process of DSB formation, the replication-born DSBs (derived from the ssDNA gaps) cannot be repaired via HR using the sister-chromatid, because this template still contains the persistent lesion (Figure 1A). In contrast to HR, TMEJ does not require an undamaged repair template, allowing DSB repair in the presence of a lesion-bearing sister chromatid. Repair via TMEJ comes with a price, however, as it results in small deletions. The notion that a single unresolved replication barrier can cause multiple DSBs that are bound to be repaired via mutagenic means illustrates the potential hazard of such lesions and emphasizes the need for helicases such as DOG-1/FANCJ to safeguard genome stability in proliferating tissues.

The concepts mentioned above may also have important implications for cancer development, as the mutagenic burden of unresolved replication barriers is predicted to grow every time the cell divides, fueling genetic heterogeneity in fast proliferating tissues. We show here that a single unresolved G4 quadruplex can cause many deletions and promote genetic heterogeneity in somatic tissues, a feature often found in tumor tissues and known to correlate with poor prognosis (Burrell *et al.* 2013). Given the significant enrichment of G4 motifs at

structural genomic variations in cancer tissues (De and Michor 2011), unresolved replication barriers may present a potent source of genomic instability during cancer evolution.

Materials and Methods

Genetics

All strains were cultured according to standard *C. elegans* procedures (Brenner 1974). Alleles used in this study include: *LGI; dog-1(gk10), mus-81(tm1937), LGII; xpf-1(e1487), LGIII: polq-1(tm2026), LG unknown; lfIs77 [pLM88].*

PCR-based assays to identify G4 deletions at endogenous loci

Stochastic deletion formation at endogenous G4 DNA loci was assayed using a PCR-based approach. Genomic DNA was isolated either from single worms or pools of worms and subjected to nested rounds of PCRs with primers that flank a G4 motif; all amplicons are >1kb in size. PCR-based methods are highly sensitive and allows detection of low-frequency genomic rearrangements: G4 deletion products are preferentially amplified because they are smaller than the abundant wild-type products and lack the G4 motif that hampers DNA replication *in vitro* (Pontier *et al.* 2009).

To capture independent G4 deletion events in individuals, L4 stage animals were used (1 worm per 10ul lysis reaction) and ~0.1ul lysate (1%) was transferred into 15ul PCR reactions using a 384 pin replicator (Genetix X5050). Subsequently 0.2ul of PCR product was used for 15ul nested PCR reactions. PCR reactions were typically run for 35 cycles with 54°C primer annealing and 72°C extension for 120 seconds. x The following primers (5′-3′) were used: Qua213; ctcagccaaggctacaaac, gatacgtgtacatgaatagtc, ccggcaattacacatttgcc, caaaactgtcgcctgacctc,Qua1277;ggggagaagccgcatccaa,cacatggagacggagagaaac, cctgacaaacgcctactctc,gaatccctttaatttggcaatag,Qua375;ctagttcagggtatctggac, ccttctctcgaagcgcgacc, ggacggagagtcaataaaatc, cgaggtaaagtgcccgcaatc.

Deletion junctions were analyzed by Sanger sequencing.

To obtain stochastic deletions frequencies at Qua1277 and Qua375, we analyzed two independent PCR reactions on 1% lysate fractions of >190 *dog-1* proficient and >330 *dog-1* deficient L4 animals. Unique deletion products were discriminated based on size by gel-electrophoreses. To obtain the expected frequency of animals carrying two independent stochastic deletions we considered the two genomic lystate fractions as independent tests and multiplied the deletion frequency in the first sample to the frequency of unique deletions in the second sample. Dominant deletions products that resulted in identical deletions in both lysate fractions were assigned positive only in the first sample but not in the second, because these two deletions products did not represent independent stochastic events. In fact, subsequent testing of singe worm lysates that displayed two identical deletions resulted exclusively in

identical PCR products (5/5), indicating that such deletion products act dominantly in the PCR reaction and likely prevent the amplification of actual independent stochastic events, actually resulting in an underestimation of unique double deletion events. To determine whether there was significant over-representation of double deletion events, we used a hypergeometric test for overlap between the stochastic deletion frequency (based on the deletion frequency among independent individuals and size of the population) and the observed deletion frequency within the population sample carrying at least one unique deletion. This test returns the probability of a given number of sample successes, given the sample size, population successes and population size (hypergeometric p-value).

LacZ-based transgenic reporter assay to visualize G4 deletions

Transgenic strains were obtained by microinjection of reporter construct pLM88 [myo-2::C23::stops::NLS::LacZ] and mCherry-based co-expression markers pGH8 pCFJ104 to generate *lfIs77 (Frokjaer-Jensen et al. 2008)*. To visualize stochastic G4 deletions, clonal lines of *dog-1* deficient *lfIs77* animals were synchronized by bleaching and ~200 L1 animals were grown on OP50 plates and stained for LacZ expression three days later (Pothof *et al.* 2003). To obtain LacZ expression frequencies, >4 synchronized populations of three independent clonal *dog-1 lfIs77* lines were analyzed.

Supplementary information

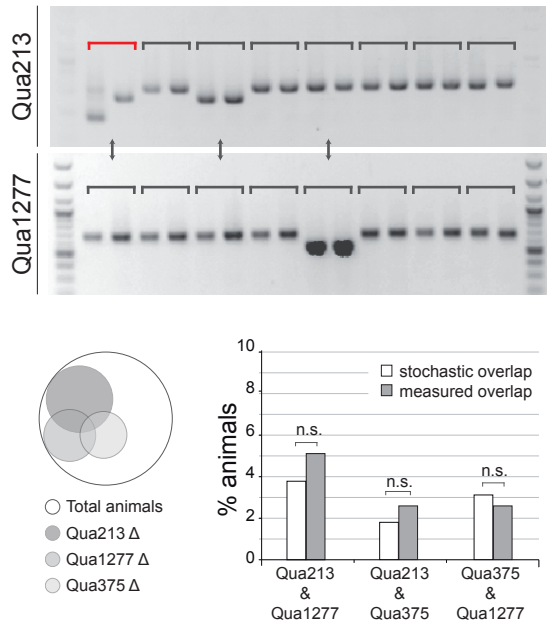

Figure S1. No significant correlation among G4 deletions detected at different genomic loci
PCR analysis of G4 instability at endogenous G4 loci Qua213, Qua1277 and Qua375. Representative gel images are depicted used to monitor stochastic deletions at Qua213 and Qua1277 in the same set of *dog-1* deficient animals (upper panels). Arrows illustrate no correlation between deletion events at the different loci in the same animal. Venn diagram shows distribution of G4 deletion events among 156 animals tested for all three loci. Histogram depicts expected frequencies of animals showing stochastic G4 deletions at both indicated G4 loci, assuming the deletion events are independent random events (white bars), as well as the observed frequencies of animals showing G4 deletions at both indicated G4 loci (grey bars). n.s. indicates a non-significant over-representation of the overlap between the indicated G4 deletion samples within the animal population (n=156) as determined by hypergeometric testing ($p > 0.20$), strongly suggesting that these are random independent events.

Acknowledgements

The authors thank Shohei Mitani (National Bioresource Project, Japan) and the Caenorhabditis Genetics Center for strains; Jane van Heteren, Ron Romeijn and Maartje van Kregten for valuable technical support.

References

Agostinho, A., B. Meier, R. Sonneville, M. Jagut, A. Woglar *et al.*, 2013 Combinatorial regulation of meiotic holliday junction resolution in C. elegans by HIM-6 (BLM) helicase, SLX-4, and the SLX-1, MUS-81 and XPF-1 nucleases. PLoS Genet 9**:** e1003591.

Blow, J. J., X. Q. Ge and D. A. Jackson, 2011 How dormant origins promote complete genome replication. Trends Biochem Sci 36**:** 405-414.

Brenner, S., 1974 The genetics of Caenorhabditis elegans. Genetics 77**:** 71-94.

Budzowska, M., and R. Kanaar, 2009 Mechanisms of dealing with DNA damage-induced replication problems. Cell Biochem Biophys 53**:** 17-31.

Burrell, R. A., N. McGranahan, J. Bartek and C. Swanton, 2013 The causes and consequences of genetic heterogeneity in cancer evolution. Nature 501**:** 338-345.

De, S., and F. Michor, 2011 DNA secondary structures and epigenetic determinants of cancer genome evolution. Nat Struct Mol Biol 18**:** 950-955.

Edwards, D. N., A. Machwe, Z. Wang and D. K. Orren, 2014 Intramolecular telomeric G-quadruplexes dramatically inhibit DNA synthesis by replicative and translesion polymerases, revealing their potential to lead to genetic change. PLoS One 9**:** e80664.

Elvers, I., F. Johansson, P. Groth, K. Erixon and T. Helleday, 2011 UV stalled replication forks restart by re-priming in human fibroblasts. Nucleic Acids Res 39**:** 7049-7057.

Frokjaer-Jensen, C., M. W. Davis, C. E. Hopkins, B. J. Newman, J. M. Thummel *et al.*, 2008 Single-copy insertion of transgenes in Caenorhabditis elegans. Nat Genet 40**:** 1375-1383.

Guschlbauer, W., J. F. Chantot and D. Thiele, 1990 Four-stranded nucleic acid structures 25 years later: from guanosine gels to telomer DNA. J Biomol Struct Dyn 8**:** 491-511.

Han, H., L. H. Hurley and M. Salazar, 1999 A DNA polymerase stop assay for G-quadruplex-interactive compounds. Nucleic Acids Res 27**:** 537-542.

Helleday, T., 2013 PrimPol breaks replication barriers. Nat Struct Mol Biol 20**:** 1348-1350.

Howell, R. M., K. J. Woodford, M. N. Weitzmann and K. Usdin, 1996 The chicken beta-globin gene promoter forms a novel “cinched” tetrahelical structure. J Biol Chem 271**:** 5208-5214.

Huppert, J. L., 2010 Structure, location and interactions of G-quadruplexes. FEBS J 277**:** 3452-3458.

Jiricny, J., 2013 Postreplicative mismatch repair. Cold Spring Harb Perspect Biol 5**:** a012633.

Koole, W., R. van Schendel, A. E. Karambelas, J. T. van Heteren, K. L. Okihara *et al.*, 2014 A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. Nat Commun 5**:** 3216.

Kruisselbrink, E., V. Guryev, K. Brouwer, D. B. Pontier, E. Cuppen *et al.*, 2008 Mutagenic capacity of endogenous G4 DNA underlies genome instability in FANCJ-defective C. elegans. Curr Biol 18**:** 900-905.

Levitus, M., Q. Waisfisz, B. C. Godthelp, Y. de Vries, S. Hussain *et al.*, 2005 The DNA helicase BRIP1 is defective in Fanconi anemia complementation group J. Nat Genet 37**:** 934-935.

Loeb, L. A., and R. J. Monnat, Jr., 2008 DNA polymerases and human disease. Nat Rev Genet 9**:** 594-604.

Manthei, K. A., and J. L. Keck, 2013 The BLM dissolvasome in DNA replication and repair. Cell Mol Life Sci 70**:** 4067-4084.

Masutani, C., R. Kusumoto, A. Yamada, N. Dohmae, M. Yokoi *et al.*, 1999 The XPV (xeroderma pigmentosum variant) gene encodes human DNA polymerase eta. Nature 399**:** 700-704.

Minocherhomji, S., and I. D. Hickson, 2014 Structure-specific endonucleases: guardians of fragile site stability. Trends Cell Biol 24**:** 321-327.

O'Neil, N. J., J. S. Martin, J. L. Youds, J. D. Ward, M. I. Petalcorin *et al.*, 2013 Joint molecule resolution requires the redundant activities of MUS-81 and XPF-1 during Caenorhabditis elegans meiosis. PLoS Genet 9**:** e1003582.

Pontier, D. B., E. Kruisselbrink, V. Guryev and M. Tijsterman, 2009 Isolation of deletion alleles by G4 DNA-induced mutagenesis. Nat Methods 6**:** 655-657.

Pothof, J., G. van Haaften, K. Thijssen, R. S. Kamath, A. G. Fraser *et al.*, 2003 Identification of genes that protect the C. elegans genome against mutations by genome-wide RNAi. Genes Dev 17**:** 443-448.

Preston, B. D., T. M. Albertson and A. J. Herr, 2010 DNA replication fidelity and cancer. Semin Cancer Biol 20**:** 281-293.

Roerink, S. F., R. van Schendel and M. Tijsterman, 2014 Polymerase theta-mediated end joining of replication-associated DNA breaks in C. elegans. Genome Res 24**:** 954-962.

Saito, T. T., D. Y. Lui, H. M. Kim, K. Meyer and M. P. Colaiacovo, 2013 Interplay between structure-specific endonucleases for crossover control during Caenorhabditis elegans meiosis. PLoS Genet 9**:** e1003586.

Saito, T. T., J. L. Youds, S. J. Boulton and M. P. Colaiacovo, 2009 Caenorhabditis elegans HIM-18/SLX-4 interacts with SLX-1 and XPF-1 and maintains genomic integrity in the germline by processing recombination intermediates. PLoS Genet 5**:** e1000735.

Tarsounas, M., and M. Tijsterman, 2013 Genomes and G-quadruplexes: for better or for worse. J Mol Biol 425**:** 4782-4789.

Youds, J. L., N. J. O'Neil and A. M. Rose, 2006 Homologous recombination is required for genome stability in the absence of DOG-1 in Caenorhabditis elegans. Genetics 173**:** 697-708.