Alternative end-joining of DNA breaks

Schendel, Robin van

Alternative end-joining of DNA breaks

Robin van Schendel

Cover design & Layout: Robin van Schendel

Printing: Off Page, www.offpage.nl

ISBN: 978-94-6182-741-8

Alternative end-joining of DNA breaks

Proefschrift

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,

volgens besluit van het College voor Promoties

te verdedigen op donderdag 15 december 2016

klokke 13.45

door

Robin van Schendel

geboren te Rijswijk
in 1983

Promotiecommissie

Promotoren:	Prof. dr. M. Tijsterman	
	Prof. dr. J. Brouwer	
Leden Promotiecommissie:	Prof. J. den Dunnen	
	Prof. dr. H. te Riele	(NKI, Amsterdam)
	Dr. P. Knipscheer	(Hubrecht Institute, Utrecht)

TABLE OF CONTENTS

GENERAL INTRODUCTION AND AIM

In 1869 Friedrich Miescher was the first to discover DNA (deoxyribonucleic acid), at that time termed 'nuclein'[1]. 60 years later, in 1929, Phoebus Levene identified nucleotides as the building blocks of DNA, but it took until 1952 for scientists to realize that not proteins, but DNA is the carrier of genetic information[2]. This heritable information is vital to a cell's survival as it contains all the instructions to create life. DNA is composed of nucleotides that can contain four different bases: guanine, cytosine, adenine and thymine. The double helix structure of DNA consists of two complementary strands that are held together by hydrogen-bonds between base pairs that are exclusively formed by adenine – thymine and cytosine – guanine.

DNA is constantly threatened by endogenous as well as exogenous sources that can damage the DNA molecule and, if left unrepaired, these lesions can interfere with important cellular functions such as replication and transcription and will invariably lead to the loss of genetic information. It has been estimated that each of the ~10^{13} cells in the human body receives tens of thousands of DNA lesions per day[3]. Spontaneous hydrolysis of nucleotides is responsible for the bulk of base loss and results in the formation of abasic sites. Duplication of genetic information by DNA replication, which is essential for a cell to divide, poses another threat to the integrity of DNA as incorrect nucleotides may be incorporated or slippage of the replication machinery can occur, thereby inserting or deleting DNA. In addition to endogenous threats to genome stability, cells have to deal with various external causes of DNA damage such as ultraviolet (UV) light and ionizing radiation (IR). UV causes two adjacent pyrimidines (i.e. thymine and/or cytosine) to covalently bond and form a so-called intrastrand crosslink. IR is responsible for a plethora of lesions, including oxidative damage of bases, single-strand breaks and one of the most toxic lesions: double-strand breaks (DSBs). In addition, various genotoxic chemicals exist that can cause bulky adducts or interstrand crosslinks (ICLs). Cisplatin, a common anti-cancer drug, is able to physically connect both complementary DNA helices (i.e. an ICL), which will interfere with important cellular functions as the two DNA strands can no longer be separated.

It is therefore no surprise that cells have developed numerous DNA repair mechanisms to preserve the integrity and stability of DNA. Failure to properly repair DNA damage leads to the accumulation of mutations and can ultimately lead to malignant transformation. The main topic of this thesis is the repair of double-strand breaks, which I studied in the model organism *C. elegans*, a small nematode species of approximately 1 mm long. The simple fact that many of the DNA repair mechanisms found are conserved between humans and such a small organism as *C. elegans* is already an indication of their importance. In the remainder of this chapter I will introduce the DNA repair systems that exist to deal with DNA damage, followed by a brief introduction of next-generation sequencing. Its rapid development in the last decade has meant a game-changer for many scientists and in fact many of the discoveries presented in this thesis would not have been possible without it. Then I will introduce the model organism *C. elegans*, which has been extensively studied over the last 40 years. Finally, I will briefly outline the experimental chapters of this thesis.

DNA repair systems

In order to maintain genomic integrity cells have developed a broad range of protective mechanisms to cope with DNA damage. The pathways responsible for sensing, signalling and promoting DNA repair are collectively referred to as the DNA Damage Response (DDR). This multifaceted response to DNA damage together is responsible for the cell's outcome to genomic infliction: survival,

senescence (lost the capability to divide) or apoptosis (programmed cell death).

Base Excision Repair (BER)

Base excision repair (BER) is an important pathway primarily responsible for the repair of non-helix-distorting lesions. These include alkylated, oxidized and deaminated bases, the most common types of DNA damage. BER can be subdivided into two pathways: short- and long-patch BER, the main difference being that while long-patch BER results in a newly synthesized stretch of a few nucleotides, short-patch BER only inserts a single nucleotide. The activity of BER can be roughly divided into four steps: First, recognition of a damaged base and its subsequent removal by a glycosylase. Next, cleavage of the sugar backbone by an AP endonuclease, leaving a single nucleotide gap. Then, a polymerase is recruited to fill the gap and finally a DNA ligase will seal the gap by reconnecting the DNA backbone (Figure 1). Enzymes of BER are also responsible for restoring DNA single-strand breaks (SSBs)[4].

The importance of this pathway is illustrated by the high degree of conservation of BER between *E. coli* and mammals. Furthermore, deleterious mutations in BER genes have been shown to result in a higher mutation rate and an increased chance of developing cancer[5,6].

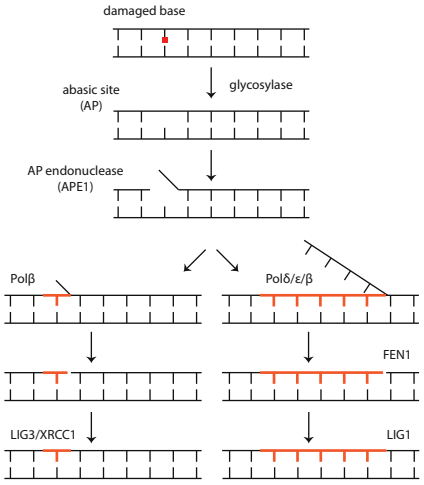


Figure 1. Base Excision Repair (BER). See text for details.

Nucleotide Excision Repair (NER)

Nucleotide excision repair (NER) is primarily responsible for the removal of helix-distorting lesions. A variety of DNA damage, such as UV-light and the anti-cancer drug cisplatin, can result in helix-distorting lesions. When such lesions arise in the transcribed strand and block an RNA polymerase they are repaired by transcription-coupled NER (TC-NER), while when present in the non-transcribed strand or in non-transcribed regions they are recognized by global genome NER (GG-NER). The primary difference between TC-NER and GG-NER is in damage recognition and signalling whereas the downstream repair steps are shared[7]. In GG-NER recognition takes place by protein complexes consisting of XPC and XPE and in TC-NER the stalled RNA polymerase recruits CSA and CSB. In both cases the next step is opening up the DNA via the multifunctional TFIIH complex. The lesion is then excised via the endonucleases XPF and XPG[8]. A DNA polymerase is

then brought in to fill the gap and finally a DNA ligase seals the break.

Defects in any of the xeroderma pigmentosum (XP) proteins, which are generally involved in NER, lead to the inability to repair damage caused by UV light. Patients with xeroderma pigmentosum thus have a greatly increased risk of developing skin cancer and have to minimize exposure to the sun throughout life.

Mismatch Repair (MMR)

Faithful duplication of genomic information is essential for survival and to improve the fidelity of DNA replication the cell is equipped with a highly efficient postreplicative DNA repair system called mismatch repair (MMR). Errors corrected by MMR include base-base mispairs, but also small insertion/deletion loops. The MMR pathway can discriminate between the templated and newly synthesized strand and scans the latter for errors. Upon recognition of a mismatch by the MutS-homologs (MSH2, MSH6 and MSH3 in mammals) the newly synthesized strand is nicked by MutL (MLH1 and PMS2 in mammals) and partly removed by the exonuclease EXO1[9]. The gap (approximately 150 bps) is then filled in by the replicative polymerases δ or ε. Final ligation is performed by LIGI (Figure 2). MMR reduces the rate of replication-associated errors by about 100-fold to 1 in 10^9 nucleotides[10].

Defects in MMR can lead to Lynch syndrome or hereditary nonpolyposis colon cancer (HNPCC). Patients that suffer from Lynch syndrome develop colon cancer at an early age. Microsatellite instability is another hallmark seen in Lynch syndrome patients and is caused by small insertions/deletions in regions of repetitive DNA, such as mono-, di- or tri-tracts[11].

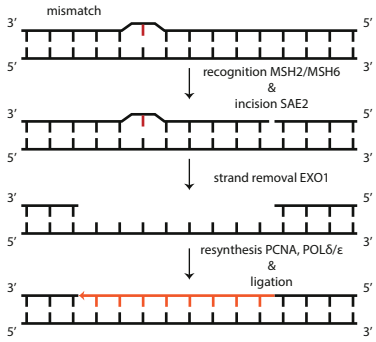


Figure 2. Mismatch Repair (MMR). See text for details.

Trans-Lesion Synthesis (TLS)

The replicative polymerases δ and ε have pivotal roles in DNA replication as they are responsible for lagging and leading strand synthesis respectively. Owing to their proof-reading capability these high fidelity polymerases have an error-rate of about 1 in 10^7 nucleotides[12]. A consequence of this high fidelity is their inability to incorporate a nucleotide opposite a damaged base thereby blocking replication. When this occurs the cell can switch to DNA damage tolerance pathways and one of the most studied pathways is trans-lesion synthesis (TLS)[13]. Upon replication fork stalling, specialized DNA polymerases (i.e. pol eta, kappa, rev1 and iota) are recruited to bypass the damage. Although these specialized TLS polymerases can efficiently bypass DNA damage,

they often do so by incorporation of an incorrect nucleotide opposite a damaged base[14]. Strictly speaking, TLS is not a DNA repair system as it does not repair DNA, but rather allows replication to continue past a damaged site to prevent replication fork collapse. The short-term benefit of continued replication outweighs the disadvantage of introducing point mutations as we also noted in Chapter 3 of this thesis.

The xeroderma pigmentosum variant (XPV) gene encodes for polymerase eta and this TLS polymerase is involved in the bypass of UV-damage. The absence of XPV leads to sensitivity to sunlight and patients develop malignant skin neoplasia at young age[15]. At a molecular level it has been shown that in the absence of (part of) TLS replication forks collapse, which leads to double-strand breaks and possible extensive loss of genetic information[16].

Interstrand Crosslink (ICL) Repair

Interstrand crosslink (ICL) repair is arguably the most complex DNA repair system as multiple repair pathways are involved in the removal and bypass of a single lesion. ICLs are extremely toxic to cells as both DNA strands are covalently linked, which inhibits strand separation and forms a physical block to both replication and transcription. Cells have developed a sophisticated repair system known as the Fanconi Anemia (FA) pathway to deal with ICLs. FA-deficient cells are extremely sensitive to crosslinking agents such as cisplatin and psoralen and up till now 19 different Fanconi genes are described (A, B, C, D1, D2, E, F, G, I, J, L, M, N, P, R, S, T, RAD51C and XPF). The current model for replication-associated ICL repair is as follows: as replication encounters and blocks at an ICL the FA-pathway responds by incision of the DNA at both sides of the crosslink. This process separates both strands and results in a double-strand break at the incised strand and in an unhooked nucleotide that is still crosslinked to the other (intact) strand. Replication then continues past the damage, likely via TLS. The incised strand is then repaired in an error-free manner via homologous recombination (HR) to restore genetic information at the break site (discussed below). As a final step the unhooked crosslink is removed by NER (Figure 3)[17].

Defects in any of the Fanconi genes lead to Fanconi Anemia, which is characterized by early development of blood cancer and bone marrow failure. About 60 percent of FA patients have congenital defects that include: short stature, abnormalities of the skin, head and arm[18]. How these congenital defects relate to the inability to repair ICLs is currently unknown.

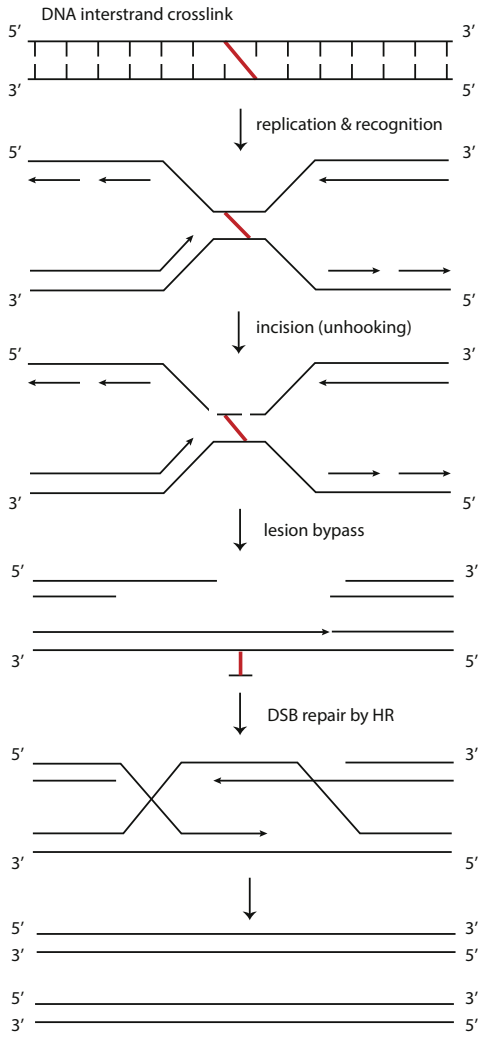


Figure 3. Interstrand Crosslink Repair. See text for details.

Homologous Recombination (HR)

A double-strand break (DSB) occurs when both strands of the DNA are broken and the DNA molecule is separated into two pieces. DSBs are the most dangerous lesion for a cell because chromosomes are physically broken. DSBs can be formed either directly, by for example ionizing radiation, or indirectly, by for example replication of single strand breaks (e.g. induced by topoisomerase inhibitors such as camptothecin) or by lesions induced by UV light and oxidation.

Cells can use homologous recombination (HR) to repair DSBs in a largely error-free manner by making use of the sister chromatid, which is present after replication, or the homologous chromosome as these contain homologous sequence. The central reaction to HR is homology search and DNA strand invasion by RAD51-coated ssDNA. A complex network of proteins is required to facilitate

invasion. First, recognition of the DSB takes place, which halts the cell cycle to allow for repair in an ATM-dependent manner[19]. Then, a complex consisting of MRE11, RAD50 and NBS1 (MRN complex) is recruited to resect the DSB ends, creating short 3′ overhangs[20]. Long-range resection is performed by EXO1 and DNA2 to expose the 3′ ssDNA overhangs, which are coated by RPA to prevent damage to the single-strand DNA (ssDNA) and prevent secondary structure formation. RPA is subsequently displaced from ssDNA by RAD51 in a BRCA2-dependent manner. The RAD51 filaments facilitate strand invasion by yet incompletely understood mechanisms. The invaded ssDNA subsequently serves as a primer from which extension takes place by a polymerase, mainly carried out by pol δ[21]. The elongated invaded strand is subsequently displaced and reannealed to the other side of the DSB, followed by a ligation step to finalize the reaction (Figure 4). When strand invasion is initiated from one broken DNA end and strand dissolution takes place this is termed synthesis-dependent strand annealing (SDSA). Alternatively, strand invasion is initiated from the other 3′ ssDNA end of the DSB as well, which leads to entangled DNA molecules, called a double holliday junction (dHJ). The dHJ can be resolved either by helicase and topoisomerase-mediated dissolution to give non-cross overs (NCOs) or cleaved by HJ resolvases, which results in both crossovers (COs) and NCOs[22].

The importance of HR for human health is underlined by the number of cancer predisposition syndromes that are associated by defects in HR genes such as ataxia telangiectasia (caused by mutations in ATM), Bloom's syndrome (caused by a mutation in BLM, a dHJ resolvase) and hereditary breast and ovarian cancer syndrome (HBOC) (caused by mutations in BRCA1 and BRCA2). Additionally, many homozygous mutations in HR genes in mice are lethal (e.g.. Brca1, Brca2, Rad51, Mre11, Rad50, NBS1), illustrating the vital importance of this repair system in mammals.

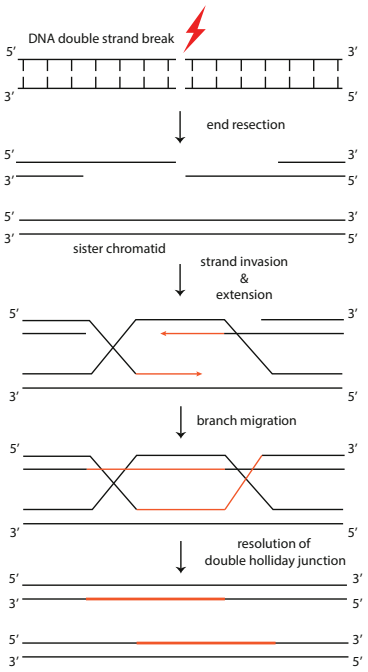


Figure 4. Homologous Recombination (HR). See text for details.

Non-homologous End Joining (NHEJ)

In addition to HR, cells are equipped with another DSB repair pathway called non-homologous end joining (NHEJ). In contrast to HR, NHEJ does not make use of a homologous template, but instead re-ligates the broken ends, which possibly leads to the loss of genetic information. It is therefore considered to be an error-prone pathway. NHEJ is the dominant repair pathway in G1 and early S phase when the sister chromosome is not available as a homologous template. Next to its pivotal role in repairing spontaneous DSBs it has another role in the repair of programmed DSBs that occur during V(D)J recombination, which allows for antibody diversification.

To repair a DSB, the ends are recognized and bound by the KU70/KU80 heterodimer, which has a high affinity for DNA ends. Then, DNA-PKcs is brought in to tether both ends and the ends are ligated by a protein complex consisting of Lig4 and XRCC4 (Figure 5). Some breaks seem to require end-processing prior to re-ligation and this can be carried out by the structure specific endonuclease Artemis or small gaps can be filled by polymerases mu and lambda[23]. Intriguingly, lower eukaryotes such as yeast and *C. elegans* lack DNA-PKcs and Artemis, but are NHEJ proficient[24].

Inactivation of XRCC4 and LIG4 in mice is lethal, indicating an absolute requirement for these proteins[25,26]. Mutations in KU70, KU80 or DNA-PKcs lead to viable mice, although they show severe phenotypes including: severe combined immunodeficiency (SCID, caused by the inability to perform V(D)J-recombination), sensitivity to radiation, early aging and neuronal apoptosis[27,28].

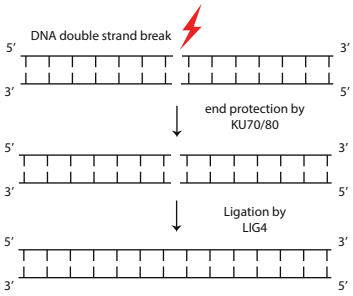


Figure 5. Non-Homologous End Joining (NHEJ). See text for details.

Alternative End Joining (Alt-EJ)

About two decades ago it became clear that next to HR and NHEJ, there was an alternative to repair DSBs: in the absence of Ku70, DSBs were still repaired and the repair footprints displayed small genomic deletions and the use of 3 – 16 nucleotides of (micro)homology for repair[29]. This pathway is currently known as alternative end joining (Alt-EJ) and there is now evidence that Alt-EJ can be divided in at least two sub-pathways. In the absence of LIG4 or XRCC4, which are involved in the final ligation step in NHEJ, all deletion footprints displayed microhomology. In contrast, KU70-deficient cells displayed two types of footprints where only one relies on microhomology. That suggests that binding of the KU70/80 complex to DSB-ends inhibits one of the Alt-EJ pathways[30]. Microhomology-mediated end joining (MMEJ) seems to depend on LIG3, although LIG1 has been shown to be able to partially substitute[31,32]. Repair by MMEJ as well as the second Alt-EJ pathway requires resection of the DNA to partially expose the DNA ends and this is thought to be performed by the MRN complex. MMEJ does not require any polymerase activity *per se*

as the homologous sequences will anneal and repair can be finalized by LIG3, possibly requiring an endonuclease to remove the DNA flaps. The second Alt-EJ pathway does require polymerase activity as the DNA requires extension. In *Drosophila* the A-family polymerase POLQ was shown to be involved in the alternative repair of DSBs[33]. A large part of this thesis concerns the role and mechanism by which POLQ repairs DSBs in *C. elegans*. By making use of various techniques including next-generation sequencing of genomic DNA, we identify POLQ as a major contributor to genome stability.

Next-Generation Sequencing

Prior to explaining the term next-generation sequencing I will first focus on the history of nucleic acid sequencing, which is simply determining the exact order of nucleotides in a given DNA or RNA molecule. As early as 1964 Robert Holley was able to sequence the 77 ribonucleotides of alanine tRNA, the tRNA that incorporates alanine into protein[34]. But it took until 1977 for Frederick Sanger and Walter Gilbert to independently develop sequencing methods for DNA by chain-termination and this technique remained the golden standard for over two decades[35,36]. In 1990 the initiative was taken to whole-genome sequence the complete human DNA, which consists of about 3.2 Gb (3,200,000,000 bases). The human genome project ended in 2003, two years ahead of time thanks to the increased speed and reduced cost of sequencing[37].

Since the completion of the first human genome the demand for cheaper and faster sequencing increased greatly. To allow for faster and cheaper sequencing, new methods were developed to replace the automated Sanger method, which is considered to be 'first-generation' sequencing. The new methods became known as next-generation sequencing or NGS. The combination of NGS-methods combined with massive parallel sequencing has made it possible for NGS platforms to nowadays sequence up to 600 Gb per run (i.e. 200 times the size of the human genome). Although each NGS platform employs different methods of sequencing, I will not discuss the differences here, but generally introduce the procedure to go from sample to analysing genomic data (see [38] for an excellent review on NGS methods).

First, the sample (DNA/RNA) has to be prepared. The sample is sheared into smaller fragments: typically ~500 bp in size, but this can vary depending on the application. Barcodes and adapters are ligated to the DNA-fragments. The adapters makes sure that all fragments have known primers at both ends from which sequencing can initiate. The barcodes allow for sequencing of several samples together as for example the *C. elegans* genome is only 100 Mb (32 times smaller than a human genome) and multiple samples can fit together in a sequencing lane. Once the library is constructed it is generally clonally amplified prior to sequencing. The actual sequencing is performed by synthesis. Each library fragment acts as a template onto which a new sequence is created by a polymerase. Sequencing occurs through cycles of washing and flooding the sequencing chamber with a known nucleotide to be incorporated. When incorporation of a nucleotide takes place this is detected (e.g. by a fluorescent or electrical signal) and digitally recorded. Fragments can be sequenced from one or both sides, depending on the NGS platform and the application.

NGS can be used for a wide range of applications, such as molecular diagnosis of inherited diseases, gene expression studies (RNA-Seq) to identify differential expressed genes, chromatin immunoprecipitation sequencing to identify binding locations of certain proteins (ChIP-seq), ribosome profiling to determine actively translated mRNAs (Ribo-Seq), Bisulphite sequencing to

determine methylation patterns, etc. I will focus here only on variant discovery in genomic DNA as that was the main purpose of the sequencing experiments that are described in this thesis.

After initial quality checks and filtering of erroneous reads the next step is to map all the reads to a reference genome (i.e. a representative example of a digital nucleic acid sequence) (Figure 6). The subsequent step is to identify variants, which are discrepancies between the reference genome and the sequenced sample. The most easily detectable variation is a single-nucleotide variant (SNV), which is a single base difference between the reference genome and the sample at a certain location. Some NGS-platforms deliver sequence information from both ends of a sheared DNA fragment, called paired-end reads. Paired-end reads are particularly useful to discover more complicated structural variants (i.e. deletions, insertions, inversions and translocations) as the two reads originate from a ~500bp fragment and therefore were very close together in the original sample. If for instance one read maps to one chromosome and the other to another chromosome it could indicate an interchromosomal translocation. Likewise, deletions can be detected as paired-end reads that map further apart in the reference genome than expected.

Variant discovery is intrinsically difficult and many software packages have been developed to tackle this problem. The split-read algorithm is a frequently used approach which makes use of the paired-end reads (e.g. Pindel[39] and Delly[40] implemented this approach). The algorithm is based on the assumption that if only one end of the pair can be mapped, the second cannot be mapped because it crosses a structural variation in the sample, which is not present in the reference genome (Figure 6). The unmapped read is then split into two parts and an attempt is made to re-map both split reads in the vicinity of the mapped read. The split can be done at various positions within one read and mapped at many positions and it is therefore computationally expensive to perform. The likelihood of being a true structural variation increases if multiple split-reads support a variation. To obtain sufficient confidence in the variant discovery it is common practice to have a genome coverage of at least 10-20 times (i.e. each nucleotide is seen at least 10-20 times on average) and to sequence multiple related samples to detect *de novo* structural variations.

One of the current milestones of NGS is to be able to sequence the entire human genome in <$1,000 (with an average coverage of ~30 times), although that goal has not been reached yet. A decade of NGS has produced an overwhelming amount of data and while more applications are being developed and existing ones improved, the amount of data will only expand. The next major challenge will be to efficiently utilize these data to increase our understanding of biology.

We used next-generation sequencing of genomic DNA of *C. elegans* to assay genomic changes in an unbiased way in several DNA repair-deficient backgrounds.

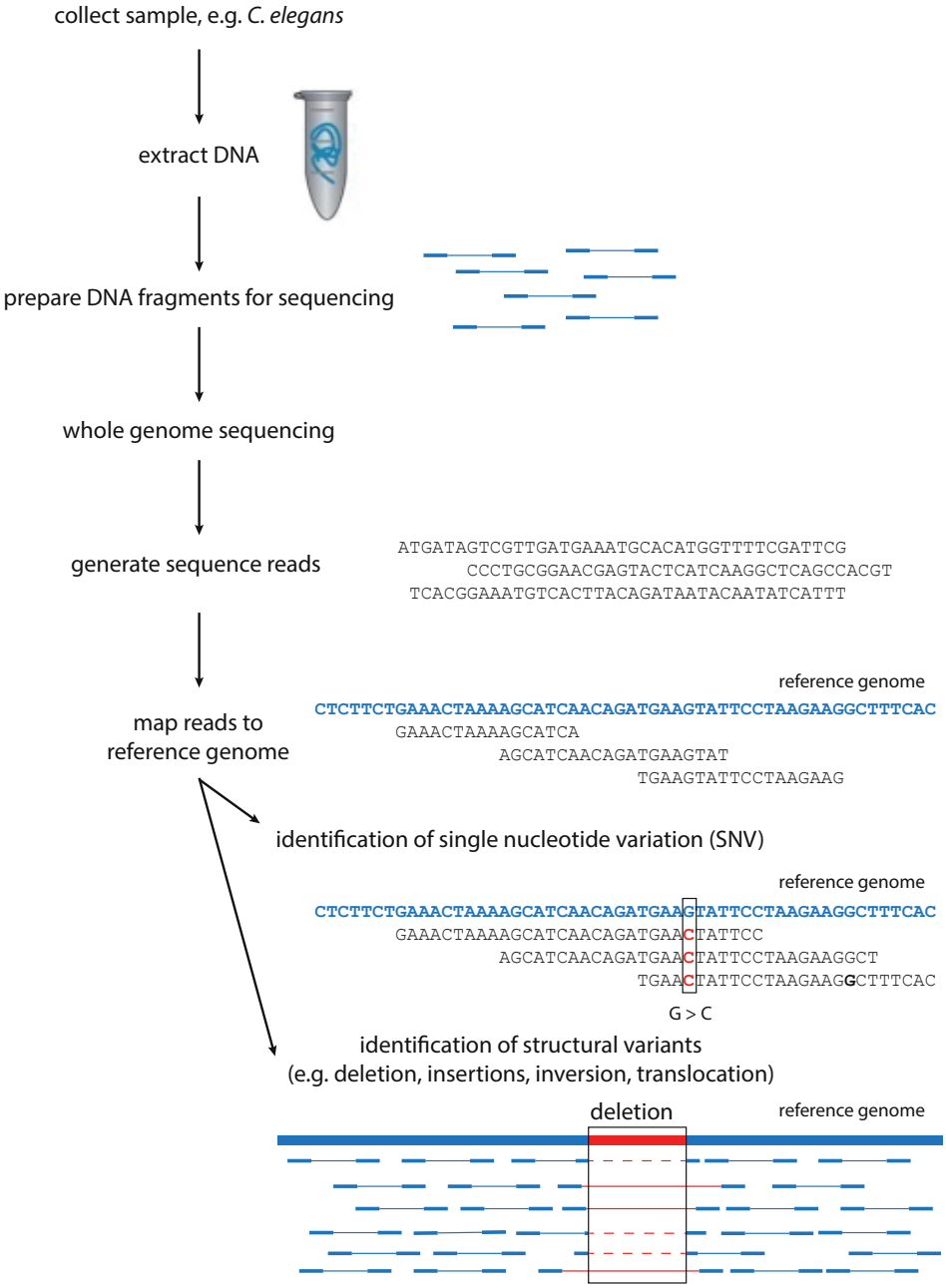


Figure 6. Next Generation Sequencing (NGS). An illustration of a typical NGS workflow as performed for sequencing of genomic DNA of *C. elegans*. See text for further details.

Caenorhabditis elegans

C. elegans was proposed as a model organism in 1974 by Sydney Brenner[41]. At the time *Drosophila* was already used, but Sydney Brenner deemed it too complex to study the nervous system. *C. elegans* is a 1 mm long transparent organism that feeds on bacteria and has a life-cycle of about 3.5 days in which it hatches and passes through four larval (L1 – L4) stages to become an adult. It is a hermaphroditic species making it a powerful genetic tool as progeny will carry (almost) the identical genetic information. Males (X0) are also occasionally born from a XX hermaphrodite, but are essentially the result of missegregation of the X chromosome during development of gametes. The presence of males, however, allows us to combine different mutations by simply crossing them. In 1998 *C. elegans* was the first multicellular organism to have its genome sequenced and published[42].

DNA repair mechanisms are highly conserved among eukaryotes and *C. elegans* is no exception. For many of the known DNA repair genes functional homologs have been identified and for many of the non-lethal genes loss-of-function alleles exist that can be requested from the Caenorhabditis Genetics Center (CGC). The recent development of CRISPR\Cas9 technology, which allows us to edit the genome of *C. elegans* in a way that could have never been done before (e.g. by endogenously tagging proteins by a fluorescent label, or to change specific amino acids in a gene) will inspire new and exciting research in this established model organism[43,44].

Aim and outline of this thesis

As loss of even a single DNA repair system can greatly increase the risk of cancer it is of critical importance to understand these cellular processes. The aim of this thesis is to further our understanding of the molecular details of DNA repair mechanisms, in particular DSB repair. Fundamental insight into these repair pathways will contribute to our understanding of biology and have the potential to assist in the development of anti-cancer drugs, by identifying new druggable targets. By using comparative genomics and whole-genome sequencing of propagated mutant as well as wild-type animals, we investigated the impact of various DNA repair systems on genome stability. This approach combined with specific assays to read out genome stability unexpectedly led to the discovery of a previously unknown DSB repair mechanism that depends on POLQ, which was found to be responsible for the majority of heritable genomic changes seen in *C. elegans*.

In **Chapter 2** we analyse the evolution of introns between several species of *C. elegans* and *Drosophila*. While many introns are conserved, some were lost during evolution. We perform an *in silico* analysis to compare lost and retained introns and identify microhomology between intron-exon junctions to be a determinant for increased intron loss.

In **Chapter 3** we make use of whole-genome sequencing to compare genomic alterations in *C. elegans* animals in wild-type, pol eta and pol kappa-deficient animals grown for many generations. In the absence of TLS we observe a distinct class of deletions occurring, which are between 50-300 bp. We find that these genomic scars are generated by a previously unknown DSB-repair pathway mediated by the A-family polymerase Theta (POLQ).

In **Chapter 4** we investigate the repair of DSBs in cells that give rise to the following generation (i.e. germ cells). To this end, we set up an assay to read out error-prone repair of DSBs generated by transposon jumps. As an independent readout we make use of the recently discovered CRISPR\Cas-9 system to induce DSBs in germ cells. In both assays we find the repair of breaks to be dependent on the activity of POLQ. Finally, by small-scale evolution experiments we identify

POLQ to be a key player in shaping the genome of *C. elegans* during evolution.

In **Chapter 5** we attempt to unveil the *in vivo* mechanism by which Polymerase Theta-mediated end-joining repairs DSBs. We show that most, if not all, EMS and UV/TMP-induced deletions are the result of POLQ-mediated repair. This finding allows for an in-depth analysis of ~10,000 deletion alleles that were generated in the last four decades of *C. elegans* research.

In **Chapter 6** I will summarize the main conclusions of this thesis and I will discuss some of the future perspectives that have emerged.

REFERENCES

1 R. Dahm Discovering DNA: Friedrich Miescher and the early years of nucleic acid research *Hum. Genet. 122*(6), 565 (2008).
2 A. D. HERSHEY and M. CHASE Independent functions of viral protein and nucleic acid in growth of bacteriophage *J. Gen. Physiol 36*(1), 39 (1952).
3 T. Lindahl and D. E. Barnes Repair of endogenous DNA damage *Cold Spring Harb. Symp. Quant. Biol. 65*, 127 (2000).
4 K. W. Caldecott Single-strand break repair and genetic disease *Nat. Rev. Genet. 9*(8), 619 (2008).
5 S. M. Farrington, *et al.* Germline susceptibility to colorectal cancer due to base-excision repair gene defects *Am. J. Hum. Genet. 77*(1), 112 (2005).
6 D. Starcevic, S. Dalal, and J. B. Sweasy Is there a link between DNA polymerase beta and cancer? *Cell Cycle 3*(8), 998 (2004).
7 J. A. Marteijn, *et al.* Understanding nucleotide excision repair and its roles in cancer and ageing *Nat. Rev. Mol. Cell Biol. 15*(7), 465 (2014).
8 E. C. Friedberg How nucleotide excision repair protects against cancer *Nat. Rev. Cancer 1*(1), 22 (2001).
9 A. B. Buermeyer, *et al.* Mammalian DNA mismatch repair *Annu. Rev. Genet. 33*, 533 (1999).
10 J. Pena-Diaz and J. Jiricny Mammalian mismatch repair: error-free or error-prone? *Trends Biochem. Sci. 37*(5), 206 (2012).
11 L. J. Rasmussen, *et al.* Pathological assessment of mismatch repair gene variants in Lynch syndrome: past, present, and future *Hum. Mutat. 33*(12), 1617 (2012).
12 T. A. Kunkel DNA replication fidelity *J. Biol. Chem. 279*(17), 16895 (2004).
13 P. L. Andersen, F. Xu, and W. Xiao Eukaryotic DNA damage tolerance and translesion synthesis through covalent modifications of PCNA *Cell Res. 18*(1), 162 (2008).
14 I. Saugar, M. A. Ortiz-Bazan, and J. A. Tercero Tolerating DNA damage during eukaryotic chromosome replication *Exp. Cell Res. 329*(1), 170 (2014).
15 J. E. Cleaver, *et al.* A summary of mutations in the UV-sensitive disorders: xeroderma pigmentosum, Cockayne syndrome, and trichothiodystrophy *Hum. Mutat. 14*(1), 9 (1999).
16 S. S. Lange, K. Takata, and R. D. Wood DNA polymerases and cancer *Nat. Rev. Cancer 11*(2), 96 (2011).
17 J. Zhang and J. C. Walter Mechanism and regulation of incisions during DNA interstrand cross-link repair *DNA Repair (Amst) 19*, 135 (2014).
18 J. Lanneaux, *et al.* [Fanconi anemia in 2012: diagnosis, pediatric follow-up and treatment] *Arch. Pediatr. 19*(10), 1100 (2012).
19 C. H. McGowan and P. Russell The DNA damage response: sensing and signaling *Curr. Opin. Cell Biol. 16*(6), 629 (2004).
20 C. Wyman and R. Kanaar DNA double-strand break repair: all's well that ends well *Annu. Rev. Genet. 40*, 363 (2006).
21 L. Maloisel, F. Fabre, and S. Gangloff DNA polymerase delta is preferentially recruited during homologous recombination to promote heteroduplex DNA extension *Mol. Cell Biol. 28*(4), 1373 (2008).
22 Y. Liu and S. C. West Happy Hollidays: 40th anniversary of the Holliday junction *Nat. Rev. Mol. Cell Biol. 5*(11), 937 (2004).
23 M. R. Lieber, *et al.* Flexibility in the order of action and in the enzymology of the nuclease, polymerases, and ligase of vertebrate non-homologous DNA end joining: relevance to cancer, aging, and the immune system *Cell Res. 18*(1), 125 (2008).
24 M. Shrivastav, L. P. De Haro, and J. A. Nickoloff Regulation of DNA double-strand break repair pathway choice *Cell Res. 18*(1), 134 (2008).
25 Y. Gao, *et al.* A critical role for DNA end-joining proteins in both lymphogenesis and neurogenesis *Cell 95*(7), 891 (1998).
26 D. E. Barnes, *et al.* Targeted disruption of the gene encoding DNA ligase IV leads to lethality in embryonic mice *Curr. Biol. 8*(25), 1395 (1998).
27 Y. Gu, *et al.* Growth retardation and leaky SCID phenotype of Ku70-deficient mice *Immunity. 7*(5), 653 (1997).
28 H. Li, *et al.* Deletion of Ku70, Ku80, or both causes early aging without substantially increased cancer *Mol. Cell Biol. 27*(23), 8205 (2007).
29 S. J. Boulton and S. P. Jackson Saccharomyces cerevisiae Ku70 potentiates illegitimate DNA double-strand break repair and serves as a barrier to error-prone DNA repair pathways *EMBO J. 15*(18), 5093 (1996).
30 C. Boboila, *et al.* Alternative end-joining catalyzes class switch recombination in the absence of both Ku70 and DNA ligase 4 *J. Exp. Med. 207*(2), 417 (2010).
31 C. Boboila, *et al.* Robust chromosomal DNA repair via alternative end-joining in the absence of X-ray repair cross-complementing protein 1 (XRCC1) *Proc. Natl. Acad. Sci. U. S. A 109*(7), 2473 (2012).
32 D. Simsek, *et al.* DNA ligase III promotes alternative nonhomologous end-joining during chromosomal translocation formation *PLoS. Genet. 7*(6), e1002080 (2011).
33 S. H. Chan, A. M. Yu, and M. McVey Dual roles for DNA polymerase theta in alternative end-joining repair of double-strand breaks in Drosophila *PLoS. Genet. 6*(7), e1001005 (2010).
34 R. W. HOLLEY, *et al.* STRUCTURE OF A RIBONUCLEIC ACID *Science 147*(3664), 1462 (1965).
35 A. M. Maxam and W. Gilbert A new method for sequencing DNA *Proc. Natl. Acad. Sci. U. S. A*

74(2), 560 (1977).
36 F. Sanger, S. Nicklen, and A. R. Coulson DNA sequencing with chain-terminating inhibitors *Proc. Natl. Acad. Sci. U. S. A 74*(12), 5463 (1977).
37 J. C. Venter, *et al.* The sequence of the human genome *Science 291*(5507), 1304 (2001).
38 M. L. Metzker Sequencing technologies - the next generation *Nat. Rev. Genet. 11*(1), 31 (2010).
39 K. Ye, *et al.* Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads *Bioinformatics. 25*(21), 2865 (2009).
40 T. Rausch, *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis *Bioinformatics. 28*(18), i333-i339 (2012).
41 S. Brenner The genetics of Caenorhabditis elegans *Genetics 77*(1), 71 (1974).
42 Genome sequence of the nematode C. elegans: a platform for investigating biology *Science 282*(5396), 2012 (1998).
43 S. Waaijers, *et al.* CRISPR/Cas9-targeted mutagenesis in Caenorhabditis elegans *Genetics 195*(3), 1187 (2013).
44 D. J. Dickinson, *et al.* Engineering the Caenorhabditis elegans genome using Cas9-triggered homologous recombination *Nat. Methods 10*(10), 1028 (2013).

2

MICROHOMOLOGY-MEDIATED INTRON LOSS (MMIL) DURING METAZOAN EVOLUTION

Robin van Schendel and Marcel Tijsterman

Department of Toxicogenetics, Leiden University Medical Center, The Netherlands

Published in Molecular Evolution & Biology 2013 May 26; 5 (6): 1212-1219

Abstract

How introns are lost from eukaryotic genomes during evolution remains an enigmatic question in biology. By comparative genome analysis of five *Caenorhabditis* and eight *Drosophila* species, we found that the likelihood of intron loss is highly influenced by the degree of sequence homology at exon-intron junctions: a significant elevated degree of microhomology was observed for sequences immediately flanking those introns that were eliminated from the genome of one or more sub-species. This determinant was significant even at individual nucleotides. We propose that microhomology-mediated DNA repair underlies this phenomenon which we termed microhomology-mediated intron loss (MMIL). This hypothesis is further supported by the observations that in both species i) smaller introns are preferentially lost over longer ones and ii) genes that are highly transcribed in germ cells, and are thus more prone to DNA double strand breaks, display elevated frequencies of intron loss. Our data also testify against a prominent role for reverse transcriptase-mediated intron loss (RTMIL) in metazoans.

Introduction

Introns are non-coding DNA sequences of ambiguous function that in eukaryotes interrupt exons and are removed from pre-mRNA by the splice machinery prior to translation. A question that has puzzled biologists already for over 30 years is how introns are introduced, maintained and lost from the genomes of eukaryotes. The "intron early theory" proposes that most introns were already present before eukaryotes and prokaryotes diverged, in the genome of their common ancestor. Subsequently, prokaryotes lost their introns and eukaryotes retained (at least some of) their introns. In an alternative model, known as the "intron late theory", introns were proposed to have emerged solely within the eukaryote lineage and accumulated in genomes over evolutionary time, especially in species that do not experience selection pressure for small genome size. The most early ancestral eukaryotic progenitor is assumed to contain already many introns, prior to initial divergence, based on the existence of introns in homologous genes across early diverged species[1-3].

While genomes of some vertebrate species contain >100,000 introns, others have extremely few: the genome of the parasite *Giardia lamblia*, as an example, contains only two introns[4], which may be explained by extensive intron loss in time. The increased availability of sequenced genomes has revealed, however, that rates of intron gain and loss can differ greatly between groups of species[2,4-12].

In numerous species a clear tendency can be observed towards introns being lost[2,5-7,10] and various intron-loss mechanisms have been proposed. Reverse transcription of mRNA and subsequent recombinational integration of the produced cDNA into the genome, also known as reverse transcriptase-mediated intron loss, has been suggested to explain cases where introns are lost while the surrounding exonic sequence remained perfectly intact13. A prediction from a model where reverse transcriptase starts at the 3' ends of mRNA is a bias of intron loss towards the 3' side (as cDNA synthesis would not always reach the 5' end of the mRNA, is expected). A trend towards more frequent loss of 3'-positioned introns was observed in *Drosophila*[14] and *Arabidopsis*[7]. More recently, modified versions of RTMIL were proposed, e.g. where the 3' end of an mRNA folds back on itself to serve as a primer for reverse transcription[15,16]. These models predict that adjacent introns will be more frequently lost than dispersed ones. For example in fungi numerous cases of intron loss could now be explained by this model[17]. No evidence was found in favor of this hypothesis in the nematode *C. elegans*[18].

We wondered whether another previously hypothesized mechanism of intron loss, i.e. error-prone DNA repair, could be responsible for the precise loss of introns from genomes. This thought was triggered when we anecdotally observed substantial sequence homology at the exon-intron junction of an intron in the *pcn-1* locus that was lost in *C. elegans*, but was still present in several other nematode species. In such cases, loss of the intronic sequence could be the result of DNA double-strand break (DSB) repair, guided by sequence homology near the break sites, as we previously have witnessed homology-driven DSB repair leading to intron-size deletions in *C. elegans* cells[19]. The likelihood of a small deletion leading to the exact removal of an intron is very low, but may be enhanced in cases where flanking sequences are homologous. We thus hypothesized that homologous sequences at the intron-exon junctions may direct repair of sporadic intronic DSBs leading to precise excision of the intron, a notion supported by glimpses of sequence homology surrounding introns that are uniquely present in the nematode *C. briggsae*[20], as if these sequences facilitated intron removal from the *C. elegans* genome.

Here, we have constructed datasets of conserved introns using either five *Caenorhabditis* or eight *Drosophila* species to uncover the mechanisms that are responsible for intron loss during evolution. Our large dataset allowed us to look in-depth into the current models of intron loss during evolution, even up to chromosome resolution, which was not possible until recently.

Results

Intron loss and gain in Caenorhabditis and Drosophila

We retrieved alignments of all protein sequences from *C. elegans*, *C. briggsae*, *C. remanei*, *C. brenneri* and *C. japonica* and re-inserted intron positions based on genome annotations. We restricted our analysis to regions of genes that were highly conserved: introns were only included if 15 amino acids on both sides of the intron were at least 50% identical across all species. Next, we identified all cases where an intron was lost at least once in four species; the evolutionary most distinct species *C. japonica* was used as an outgroup. Within 11,343 highly conserved loci we found 27,488 conserved introns. By further analyzing the conserved intron set, we found 2,753 cases of intron loss and 778 cases of potential intron gain; 19,444 introns had remained perfectly stable. 2,351 intron losses and 596 gains were found within a single species and 402 losses and 182 gains were located at ancestral nodes (Fig. 1A). Dollo parsimony was used to discriminate intron loss from intron gain. Independent parallel loss of the same intron was favored as an explanation over parallel gain of an intron in different species. If both loss and gain could explain an intron event, it was discarded from our analysis. The same analysis was performed for eight *Drosophila* species (Fig. 1B).

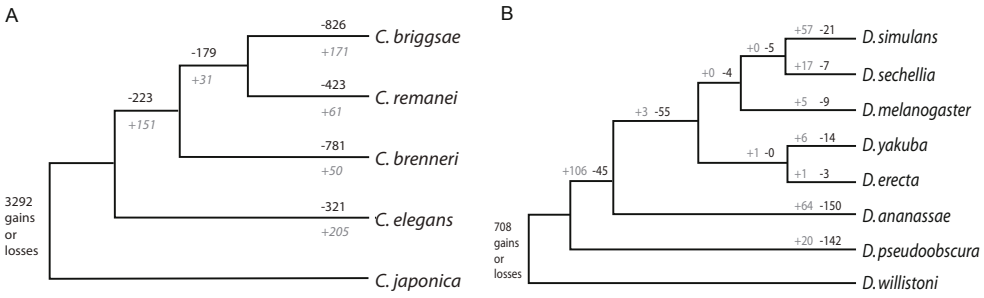


Figure 1. Intron dynamics in *Caenorhabditis* and *Drosophila* subspecies (A) Phylogenetic tree of *Caenorhabditis* species with number of introns lost (black) and gained (grey). (B) as in (A), but now for the *Drosophila* species. Genetic distances are not drawn to scale.

No reverse transcriptase-mediated intron loss in C. elegans and D. melanogaster

While reverse transcriptase-mediated intron loss (RTMIL) has been proposed to explain cases of precise intron loss in *Drosophila*[14,21] and other species[13], no evidence was found previously for this mechanism in *C. elegans*[18]. To further test this conclusion, we investigated our larger dataset, which also include additional nematode and fly species for two RTMIL predictions: preferential loss of 3′ over 5′ introns and preferential loss of adjacent introns over ones located more dispersed. While we observed a slight non-random distribution of intron loss, where the 3′ end of a locus is more susceptible than the 5′ end (Fig. S1A and S1B), we noticed that this bias is fully explained by a single peak of retained introns at the utmost 5′ side. We argue that this phenomenon can be

best explained by the notion that sequence elements regulating gene expression are frequently located in the first intron in *C. elegans*[22] and *Drosophila*[23] genes (Fig. S1C and S1D). Deletion of these introns may thus be under negative selection pressure[22,24]. We also failed to find support for the other projection of RTMIL. which is that pairs of adjacent introns are more frequently lost than dispersed pairs. Using the method published in[18], including Bonferroni correction for multiple testing, we found no difference in the number of expected and observed lost pairs of adjacent introns in *C. elegans* and *C. brenneri*. A small, but statistical difference was found in *C. briggsae* and *C. remanei* ($p < 0.01$, Fig S1E). The same analysis for *Drosophila* led to a surprising conclusion: we found a statistical difference only for *D. pseudoobscura* ($p < 0.05$). In the other six *Drosophila* species the number of cases of adjacent intron pair loss were not different from random chance (Fig S1F). Because *D. pseudoobscura* has been used to argue a role for RTMIL in flies[21], we wished to nuance that conclusion. Our data indicate that there is no support for a profound role of RTMIL in intron evolution in nematodes and flies, despite the notion of few atypical cases in flies where RTMIL seems the most logical explanation[14].

Microhomology is a determinant for intron loss

We next addressed the hypothesis of microhomology-mediated DNA repair underlying the disappearance of introns. We predicted that introns that were lost during evolution were more frequently surrounded by microhomologous sequences at their exon-intron borders, than those that were retained. In other words: is microhomology a determinant of intron loss? We restricted our analysis to the consensus splice donor (GT) and acceptor (AG) sequences and the immediately flanking two nucleotides of exonic sequences. Other intronic nucleotides as well as the wobble base (defined here as the nucleotide occupying the third position in a codon) of coding triplets were excluded. The rationale for eliminating the wobble position is as follows: as soon as an intron is lost, wobble bases surrounding the intron-exon junction lose their potential function in splicing. As a consequence, selection pressure on such non-coding nucleotides, if present, is likely lost together with the intron. The nature of the base at the time of analysis is therefore not informative as to the nature of the base at the time of intron loss. Thus, while the wobble bases may have contributed to the degree of microhomology at the time of intron loss, we eliminated them from our analysis. We subsequently determined the degree of homology by comparing the consensus splice donor nucleotides GT to the 2 outermost 5'-nucleotides of the 3' exon, and the consensus acceptor nucleotides AG to the 2 outermost 3'-nucleotides of the 5' exon. Identical nucleotides scored 1, non-identical scored 0. Non-coding wobble bases were omitted, hence the score window is maximized to 3. Figure 2B strikingly demonstrates that introns have indeed been more susceptible to being lost from genomes when they were flanked with homologous exon/intron junctions. While the group of retained introns in *Caenorhabditis* had a homology score of 1.37, lost introns scored 1.59 (with a scale from 0 to 3, ranging from no to perfect homology). Moreover, introns that were lost multiple times independently, scored even higher: 1.78 and 1.90 for 2 and 3 times being lost, respectively ($p < 0.001$ for each lost group compared to the retained group, $\chi 2$ test, $df = 3$). Phase one introns were excluded in this graph because they have a maximum score of 2 upon wobble base removal (Fig. S2). Figure 2D shows that sequence homology at each individual position of the junction contributed to the higher rates of intron loss in *Caenorhabditis*. To investigate the generality of this phenomenon, we performed a similar analysis on eight sequenced *Drosophila* species, resulting in a similar outcome: introns were more frequently lost when they had matching intron-exon junctions (Fig. 2C, 2E and S3). In *Drosophila* the group of

retained introns has a homology ranking of 1.37, lost introns score 1.69 ($p < 0.001$, $\chi 2$, df = 3).

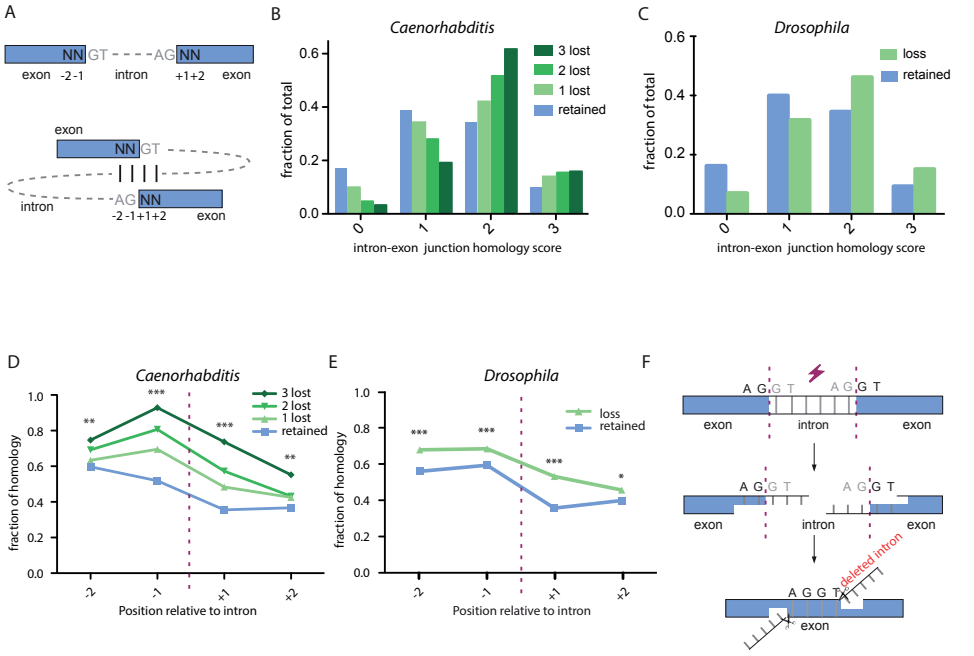


Figure 2. Microhomology-mediated intron loss (MMIL). (A) Schematic representation of the intron-exon junction alignment. For all intronic positions, the degree of homology was determined by comparing the consensus splice donor nucleotides GT to the 2 outermost 5′-nucleotides of the 3′ exon and the consensus acceptor nucleotides AG to the 2 outermost 3′-nucleotides of the 5′ exon. Identical nucleotides scored 1, non-identical scored 0. Non-coding wobble bases were omitted, hence the score window is maximized to 3. (B) The degree of intron-exon junction homology for intronic positions that suffered from 0, 1, 2 or 3 cases of intron loss. $\chi 2$ test (df = 3) was used to compare zero-lost group (n = 73,853) with the groups containing one loss (n = 1,832): $p < 0.001$, two losses (n = 528): $p < 0.001$ and three losses (n = 120): $p < 0.001$. (C) The degree of intron/exon junction homology for *Drosophila* intronic positions that suffered from zero (n = 99,864) or one or more (n = 1,385) losses ($\chi 2$ test, df = 3, $p < 0.001$). Homology scores for individual nucleotide positions as depicted in Fig. 3A for (D) *Caenorhabditis* and (E) *Drosophila*. * indicates $p < 0.05$, ** indicates $p < 0.01$ and *** $p < 0.001$. (F) A microhomology-mediated end-joining mechanism for intron loss.

Increased likelihood of loss for small introns

Sequence homology adjacent to DSBs is used in at least two error-prone DNA repair pathways, i.e. single-strand annealing and microhomology-mediated end-joining, the latter of which requires just a few identical bases on either side of the break[19,25]. Such pathways preferably use homologous sequence in close proximity to the DSB[26], and if DSB repair underlies the precise loss of introns, we expect shorter introns to be more prone to being lost. Because we earlier reasoned that the first introns in nematodes and flies possibly contain regulatory sequences and thus generally have greater length, we excluded all 5′ introns from our results. Our prediction was indeed met: we found smaller introns disappear at higher rates, both in *Caenorhabditis* (Fig. 3A) and in *Drosophila* (Fig. 3B). In *Caenorhabditis* the median intron size is 51 bp for introns that have been lost versus 57 bp for introns that have been retained ($p < 0.001$, Mann-Whitney U test). For *Drosophila* we found a median of 62 and 66 bp for lost and retained introns, respectively ($p < 0.001$, Mann-Whitney U test).

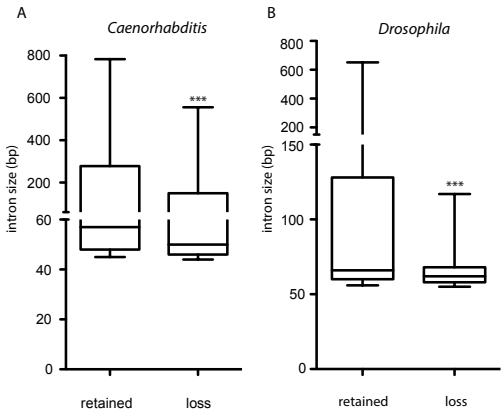


Figure 3. Preferential loss of small introns. A boxplot of the sizes of introns that were either 100% retained or found to be lost in at least one (A) *Caenorhabditis* or (B) *Drosophila* species. For the lost introns, we plotted the size of the introns that were retained at identical positions in neighboring species, excluding initial introns that possibly contain indispensable regulatory elements in the often larger introns. The median of introns that are lost was significantly smaller than that of retained introns for all *Caenorhabditis* ($p < 0.001$ (***)) and *Drosophila* species ($p < 0.001$ (***), Mann-Whitney U test). For *C. elegans*: n = 97,220 for retained introns; n = 10,465 for lost intron. For *Drosophila*: n = 142,967 for retained introns; n = 3,274 lost introns.

Germline expressed genes experience increased intron loss

We next questioned whether each gene is equally susceptible to losing one or more of its introns. One feature of a gene is its transcriptional status. Using a published dataset of germline expressed genes in *C. elegans*[27], we asked whether expression of a gene within the cells that pass on the genetic information to the next generation is of relevance. We found that ~47 % of genes that suffered from the loss of an intron are transcribed in germ cells (Fig. 4A). This is a significantly higher percentage than was found for genes that did not suffer from intron loss, which was ~38% (lost: 211 out of 450 genes versus retained: 2,555 out of 6,916 genes; $p<0.001$, $\chi 2$). A similar analysis was performed for *Drosophila* using a dataset retrieved from FlyAtlas[28]. This set contains all genes that are moderately expressed in both the ovary and the testis of the adult fly (6,141 out of 13,558). Also here, we found that germline gene expression increases the probability of intron loss (Fig. 4B), augmenting earlier work reporting elevated rates of intron loss for *Drosophila*[14] and mammals[5] for germline expressed genes. These observations are in perfect agreement with a DSB repair model of intron loss, as the more open chromatin structure of transcribed genes, as well as the activity of the transcription factories, are known to induce higher levels of DSBs in active genes[29-31].

X-chromosome germline expressed genes are less prone to intron loss

The *C. elegans* as well as the *D. melanogaster* genomes have been assembled into complete chromosomes. The constructed genomes allow us to plot the distribution of conserved and lost introns over the individual chromosomes. Using the reconstructed chromosomes, we asked whether the transcriptional status of genes influences the likelihood of losing an intron on each chromosome in a similar fashion. If intron loss were to be independent of their genomic location, a comparable distribution of lost and retained germline-expressed introns would be expected on each chromosome, and thus a ratio higher than one for lost/retained introns for all chromosomes. However, this is not what we observe: although this ratio is >1 for all autosomes, we found a clear decreased ratio (<1) on the X-chromosome in both *C. elegans* and *D. melanogaster* (Fig. 4C and 4D).

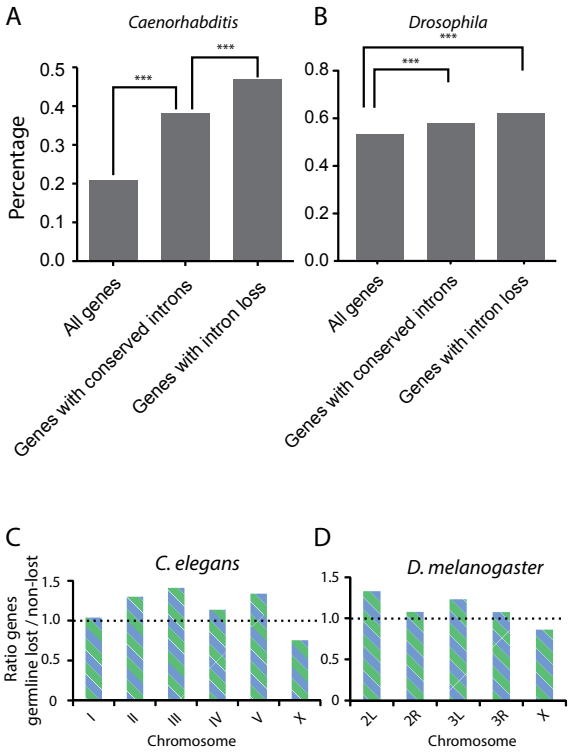


Figure 4. Increased likelihood of intron loss in germline-expressed genes in (A) *C. elegans* and (B) *D. melanogaster*. Our criteria for conserved introns, selecting on highly conserved surrounding exons, enriches for germline-expressed genes ($p < 0.001$, $\chi 2$ test). Germline expression was highly overrepresented in the class of genes with associated intron loss ($p < 0.001$, $\chi 2$ test). *** indicates $p < 0.001$. (C) Distribution of germline-expressed genes across the autosomes and the X-chromosome in *C. elegans*. For each chromosome the ratio between germline-expressing genes that have lost at least one intron and genes that contain only retained introns is plotted. (D) as in (C), but now for *D. melanogaster*. We find the same outcome as for *C. elegans*: introns located in germline-expressing genes on X are less prone to be lost compared to introns located on the autosomes.

Discussion

Recent studies have suggested DSB repair as being responsible for intron gains[4], leading to the suggestion that similar mechanisms might work for intron loss[7,32]. Using a comparative analysis of five *Caenorhabditis* and eight *Drosophila* species, we now show that the degree of microhomology at the exon-intron junction dictates the rate of intron loss in nematodes and flies, which supports a prominent role for error-prone DSB repair in changing the intron landscape. We call this phenomenon Microhomology-Mediated Intron Loss (MMIL).

Previously, non-homologous end-joining (NHEJ) has been suggested as a possible DNA repair mechanism for intron loss[7,14,32]. Although NHEJ can make use of a few nucleotides of microhomology to repair breaks[33], we disfavor this pathway to account for MMIL, mostly because this pathway plays little or no role in *C. elegans* germ cells[34]. Alternative error-prone DNA repair pathways, which have been shown to contribute to inheritable genome alteration in *C. elegans*[35], are known to be independent of the canonical NHEJ proteins CKU-70 and CKU-80[26,36]. The DSB repair mechanisms microhomology-mediated end-joining and single-stranded annealing use patches of (micro-) homology at either side of the break site to anneal in order to repair the DNA. Microhomology-mediated end-joining, although still rather ill defined, has been described as the pathway that uses only a few homologous nucleotides to establish contact between the two ends of the break. In our study we have restricted the analysis to only four positions because, apart from the splice donor and acceptor site, intronic sequences experience little selection pressure and can freely mutate without apparent consequences. The degree of microhomology at the exon/intron

border may thus very well have been more pronounced at the time the intron was lost. On an evolutionary time scale, DNA that is not under selective pressure will greatly vary between species that have relatively rapid turnover; it is estimated that each neutral base has been mutated 2-3 times since the divergence of *C. elegans* and *C. briggsae*[37]. We thus also restricted our analysis to regions of genes that were highly conserved: introns were included in our dataset only if 15 amino acids on both sides of the intron were at least 50% identical across all species. We also performed a more restrictive analysis using 100% identity in 6 amino acids on both sides, giving similar outcomes (data not included). For the same reason we omitted all wobble bases from our analysis, as also these are likely under less selective pressure after intron loss has occurred. It is thus more plausible that these bases in the current genome are different than at the moment the intron was lost. While this filter sharpens the analysis and outcomes, it is not essential, as without it, an earlier notion of elevated homology at the exon-intron border was previously spotted for *Drosophila*[21].

We found MMIL to better fit the presented data than RTMIL, which has been suggested to account for precise intron loss in other species, such as mammals and flies[5,14]. We did observe a slight bias for preferential intron retention at the 5' side of a locus, however, we consider it more likely that this effect is attributed to the retention of the first intron due to selection pressure on regulatory elements which are frequently located in the most 5' intron[38]. Indeed, the 5' conservation is no longer significant upon exclusion of the first intron (Fig. S1C-D). While the presence of microhomology is the quintessential feature to propose a MMIL model, two other observations are also in favor. Firstly, the projection that homologous sequences are preferably used when they are in close proximity to a break can explain why smaller introns are more frequently found to be lost than larger introns, in accordance with previous findings in *Drosophila*[14]. Interesting in this respect is that *C. elegans* genes that are expressed at higher levels tend to have shorter introns, which can increase the rate of intron loss if an intronic DSB occurs. We cannot, however, exclude other reasons for why smaller introns are more frequently lost over larger ones. Secondly, we found that genes that are germline-expressed are more susceptible to intron loss than those which are silent. This relationship could be explained by the notion that gene expression itself is a known inducer of DNA DSBs, which may ultimately lead to intron loss. The notion of enhanced intron loss in germline-expressed genes is in fact supportive of both the MMIL model as well as the RTMIL model. A difference between both models, however, is that RTMIL fully depends on transcriptional activity of the host gene in germ cells, whereas this dependency is far less strict for MMIL. RTMIL can thus not easily explain loss of introns in genes that are exclusively expressed in somatic tissue.

Surprisingly, we found that the preferential loss of introns from germline-expressed genes, while observed for all autosomes, is not seen for genes located on the X-chromosome. This is observed for both worms and flies. The *C. elegans* X-chromosome is silenced in early meiotic prophase in oogenic germ cells, and oocyte-enriched genes on the X-chromosomes are, on average, expressed at levels significantly lower than oocyte-enriched genes on autosomes[39]. In fact, transcription of several X-linked oocyte genes was found to be restricted to very late meiotic prophase I, a stage where DSBs are exclusively repaired via homologous recombination. This error-free repair pathway may thus protect X-linked genes from (intron) deletions at transcription-induced DSBs. While mechanisms of sex-chromosome inactivation have been observed for nematodes, flies and mammals[40,41], it is currently unknown whether they protect the sex chromosomes from mutations such as deletion of intronic sequences.

In summary, we here provide evidence that the presence of microhomology at the intron-exon junction is predictive for introns to be lost given enough time. We propose that the underlying mechanism for this MMIL phenomenon is microhomology-driven DNA double-strand break repair as this process is known to generate intron-size deletions, it explains why smaller introns are preferentially lost over larger ones, and it is in line with the observation that intron loss is more frequently found in actively transcribed genes, which are more susceptible to DNA damage. DNA repair may thus provide biological systems with the possibility to insert potential regulatory elements within encoding sequences as well as the means to remove them (Fig. 3D), even in a very precise manner, from genes that are under strong evolutionary pressure.

Materials and Methods

Protein alignments

Using the Ensembl Perl application program interface, alignments of protein sequences of *C. elegans*, *C. briggsae*, *C. remanei*, *C. brenneri* and *C. japonica* were retrieved (version 59[42]). Intron positions were re-inserted into the protein sequences and subsequent analysis was performed using custom Perl scripts. For *Drosophila*, the same analysis was performed for *D. simulans*, *D. sechellia*, *D. melanogaster*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura* and *D. willistoni* (version 59[42]).

Inferring intron loss

We restricted our analysis to regions of genes that were highly conserved: introns were included only if 15 amino acids on both sides of the intron were at least 50% identical across all species. Next, we identified all cases where an intron was lost at least once in four species; the evolutionary most distinct species *C. japonica* was used as an outgroup. The principle of Dollo parsimony was applied to the set of introns to distinguish parallel intron losses from intron gains. *C. japonica* and *D. willistoni* were used as outgroups in the *Caenorhabditis* and *Drosophila* analysis respectively.

Acknowledgements

This work was funded by European Research Council Starting Grant (203379, 'DSBrepair') to MT. We thank Jane van Heteren and Evelina Papaioannou for critically reading of the manuscript and members of the Tijsterman Lab for discussions.

Author Contributions

MT and RS wrote the paper. MT designed the study. RS wrote the Perl scripts and analyzed the data with MT.

Conflict of Interest

The authors declare that they have no conflict of interest.

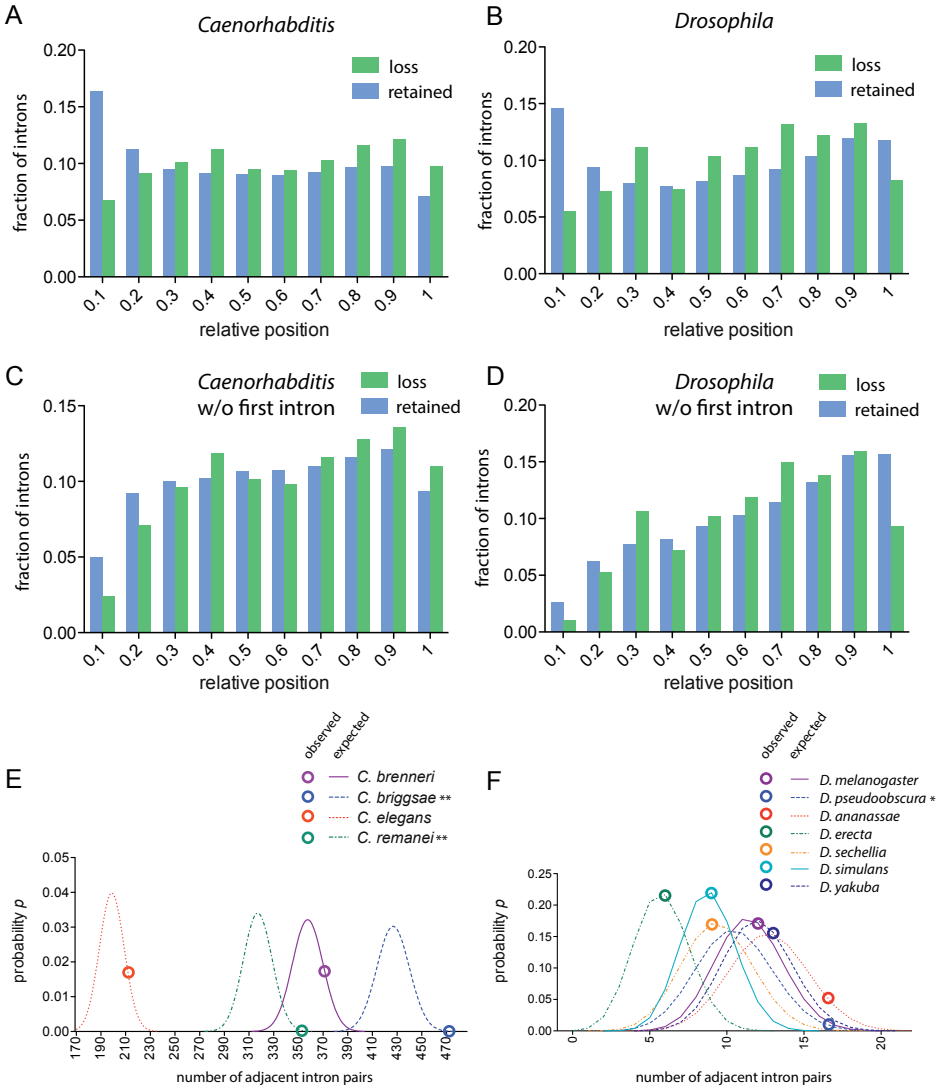


Figure S1. No evidence for RTMIL in *Caenorhabditis* and *Drosophila* subspecies. (A) Relative distribution of lost and retained introns for nematode genes. The position of the intron is determined by the number of bases upstream of an intron, divided by the number of bases in the coding region (including introns, excluding 3' and 5' UTRs). (B) as in (A), but now for Drosophila. (C and D) as in (A) and (B), but now all first introns were removed from the lost and non-lost dataset for both species. (E and F) The probability distribution for the total number of lost pairs of adjacent introns (see Formula 1 in (18)) for each analyzed Caenorhabditis species compared to C. japonica (C) or each Drosophila species compared to D. willistoni (D). Circles represent the absolute number of observed lost pairs (see Table S1 and S2), whereas the lines represent the distribution plot based on chance. For C. brenneri and C. elegans, the number of expected and observed pairs of adjacent intron loss was not significantly different. For C. briggsae and C. remanei, a small but significant difference ($p < 0.01$) was observed. For Drosophila we found no statistical difference in six out of seven subspecies between the number of observed and expected pairs of adjacent intron loss. A statistical difference was only observed for D. pseudoobscura ($p < 0.05$, including Bonferroni correction for multiple testing).

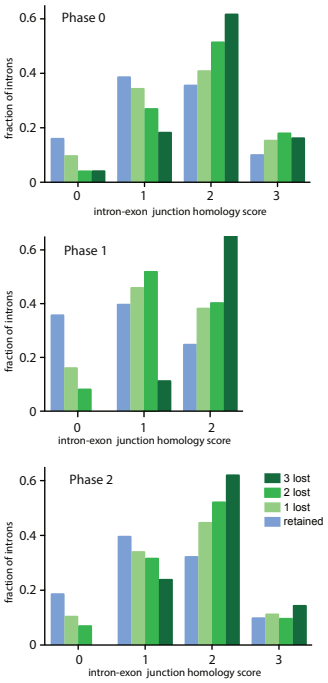


Figure S2. MMIL is evident in *Caenorhabditis* in all three different intron phases: a phase 0 intron is positioned between two codons, while a phase 1 intron disrupts a codon after the first position and a phase 2 intron after the second position. The degree of intron-exon junction homology is depicted for intronic positions that suffered from 0, 1, 2 or 3 cases of intron loss for phase 0 (A), phase 1 (B) and phase 2 (C). The absolute numbers for lost/total number of introns were: 1,683/47,962 for phase 0 introns (one wobble base), 658/24,025 for phase 1 introns (two wobble bases), and 799/28,373 for phase 2 introns (one wobble base).

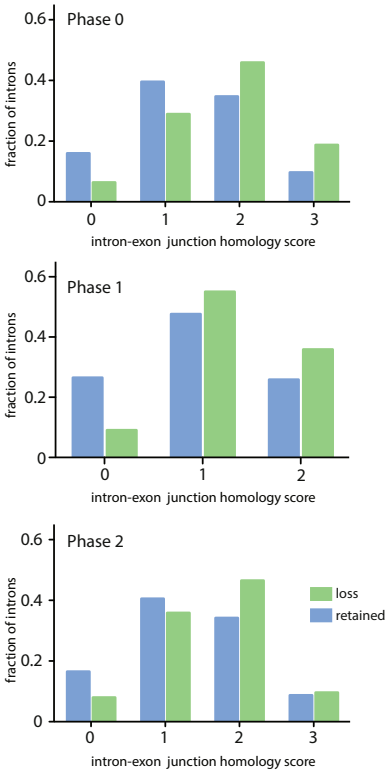


Figure S3. MMIL is evident in *Drosophila* in all three different intron phases. The degree of intron-exon junction homology is depicted for intronic positions that suffered from 0, 1, 2 or 3 cases of intron loss for phase 0 (A), phase 1 (B) and phase 2 (C). The absolute numbers for lost/total number of introns were: 453/60,200 for phase 0 introns (one wobble base), 276/43,104 for phase 1 introns (two wobble bases), and 301/39,664 for phase 2 introns (one wobble base).

Table S1

Adjacent intron pairs lost in *Caenorhabditis*

Species	Observed adjacent intron pairs lost	Expected adjacent intron pairs lost
C. brenneri	372	360
*C. briggsae***	465	427
C. elegans	213	199
*C. remanei***	351	317

** denotes $p < 0.01$

Table S2

Adjacent intron pairs lost in *Drosophila*

Species	Observed adjacent intron pairs lost	Expected adjacent intron pairs lost
D. melanogaster	12	11
*D. pseudoobscura**	17	10
D. ananassae	17	13
D. erecta	6	5
D. sechellia	9	9
D. simulans	9	9
D. yakuba	13	12

* denotes $p < 0.05$

REFERENCES

1 J. E. Stajich, F. S. Dietrich, and S. W. Roy Comparative genomic analysis of fungal genomes reveals intron-rich ancestors Genome Biol. 8(10), R223 (2007).

2 I. B. Rogozin, et al. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution Curr. Biol. 13(17), 1512 (2003).

3 A. Fedorov, A. F. Merican, and W. Gilbert Large-scale comparison of intron positions among animal, plant, and fungal genes Proc. Natl. Acad. Sci. U. S. A 99(25), 16128 (2002).

4 W. Li, et al. Extensive, recent intron gains in Daphnia populations Science 326(5957), 1260 (2009).

5 J. Coulombe-Huntington and J. Majewski Characterization of intron loss events in mammals Genome Res. 17(1), 23 (2007).

6 L. Y. Zhang, Y. F. Yang, and D. K. Niu Evaluation of models of the mechanisms underlying intron loss and gain in Aspergillus fungi J. Mol. Evol. 71(5-6), 364 (2010).

7 J. A. Fawcett, P. Rouze, and Y. Van de Peer Higher Intron Loss Rate in Arabidopsis thaliana Than A. lyrata Is Consistent with Stronger Selection for a Smaller Genome Mol. Biol. Evol. (2011).

8 A. Farlow, et al. Nonsense-mediated decay enables intron gain in Drosophila PLoS. Genet. 6(1), e1000819 (2010).

9 A. Coghlan and K. H. Wolfe Origins of recently gained introns in Caenorhabditis Proc. Natl. Acad. Sci. U. S. A 101(31), 11362 (2004).

10 C. B. Nielsen, et al. Patterns of intron gain and loss in fungi PLoS. Biol. 2(12), e422 (2004).

11 J. K. Colbourne, et al. The ecoresponsive genome of Daphnia pulex Science 331(6017), 555 (2011).

12 S. W. Roy and D. L. Hartl Very little intron loss/gain in Plasmodium: intron loss/gain mutation rates and intron number Genome Res. 16(6), 750 (2006).

13 S. W. Roy Intron-rich ancestors Trends Genet. 22(9), 468 (2006).

14 P. Yenerall, B. Krupa, and L. Zhou Mechanisms of intron gain and loss in Drosophila BMC. Evol. Biol. 11, 364 (2011).

15 A. L. Feiber, J. Rangarajan, and J. C. Vaughn The evolution of single-copy Drosophila nuclear 4f-rnp genes: spliceosomal intron losses create polymorphic alleles J. Mol. Evol. 55(4), 401 (2002).

16 D. K. Niu, W. R. Hou, and S. W. Li mRNA-mediated intron losses: evidence from extraordinarily large exons Mol. Biol. Evol. 22(6), 1475 (2005).

17 D. Croll and B. A. McDonald Intron gains and losses in the evolution of Fusarium and Cryptococcus fungi Genome Biol. Evol. 4(11), 1148 (2012).

18 S. W. Roy and W. Gilbert The pattern of intron loss Proc. Natl. Acad. Sci. U. S. A 102(3), 713 (2005).

19 D. B. Pontier and M. Tijsterman A robust network of double-strand break repair pathways governs genome integrity during C. elegans development Curr. Biol. 19(16), 1384 (2009).

20 W. J. Kent and A. M. Zahler Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment Genome Res. 10(8), 1115 (2000).

21 J. Coulombe-Huntington and J. Majewski Intron loss and gain in Drosophila Mol. Biol. Evol. 24(12), 2842 (2007).

22 K. R. Bradnam and I. Korf Longer first introns are a general property of eukaryotic gene structure PLoS. One. 3(8), e3093 (2008).

23 P. R. Haddrill, et al. Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content Genome Biol. 6(8), R67 (2005).

24 S. H. Ho, G. M. So, and K. L. Chow Postembryonic expression of Caenorhabditis elegans mab-21 and its requirement in sensory ray differentiation Dev. Dyn. 221(4), 422 (2001).

25 A. Decottignies Microhomology-mediated end joining in fission yeast is repressed by pku70 and relies on genes involved in homologous recombination Genetics 176(3), 1403 (2007).

26 M. McVey and S. E. Lee MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings Trends Genet. 24(11), 529 (2008).

27 X. Wang, et al. Identification of genes expressed in the hermaphrodite germ line of C. elegans using SAGE BMC. Genomics 10, 213 (2009).

28 V. R. Chintapalli, J. Wang, and J. A. Dow Using FlyAtlas to identify better Drosophila melanogaster models of human disease Nat. Genet. 39(6), 715 (2007).

29 M. C. Haffner, et al. Transcription-induced DNA double strand breaks: both oncogenic force and potential therapeutic target? Clin. Cancer Res. 17(12), 3858 (2011).

30 B. G. Ju, et al. A topoisomerase IIbeta-mediated dsDNA break required for regulated transcription Science 312(5781), 1798 (2006).

31 C. Lin, et al. Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer Cell 139(6), 1069 (2009).

32 A. Farlow, E. Meduri, and C. Schlotterer DNA double-strand break repair and the evolution of intron density Trends Genet. 27(1), 1 (2011).

33 M. R. Lieber, et al. Nonhomologous DNA end joining (NHEJ) and chromosomal translocations in humans Subcell. Biochem. 50, 279 (2010).

34 I. Clejan, J. Boerckel, and S. Ahmed Developmental modulation of nonhomologous end joining in Caenorhabditis elegans Genetics 173(3), 1301 (2006).

35 V. Robert and J. L. Bessereau Targeted engineering of the Caenorhabditis elegans genome following Mos1-triggered chromosomal breaks EMBO J. 26(1), 170 (2007).

36 J. E. Haber Alternative endings Proc. Natl. Acad. Sci. U. S. A 105(2), 405 (2008).
37 L. D. Stein, et al. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics PLoS. Biol. 1(2), E45 (2003).
38 M. Lynch and A. Kewalramani Messenger RNA surveillance and the evolutionary proliferation of introns Mol. Biol. Evol. 20(4), 563 (2003).
39 W. G. Kelly, et al. X-chromosome silencing in the germline of C. elegans Development 129(2), 479 (2002).
40 C. D. Meiklejohn, et al. Sex chromosome-specific regulation in the Drosophila male germline but little evidence for chromosomal dosage compensation or meiotic inactivation PLoS. Biol. 9(8), e1001126 (2011).
41 S. H. Namekawa and J. T. Lee XY and ZW: is meiotic sex chromosome inactivation the rule in evolution? PLoS. Genet. 5(5), e1000493 (2009).
42 P. J. Kersey, et al. Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species Nucleic Acids Res. 40(Database issue), D91-D97 (2012).

3

POLYMERASE THETA-MEDIATED END JOINING OF REPLICATION-ASSOCIATED DNA BREAKS IN *C. ELEGANS*

Sophie Roerink*, Robin van Schendel* and Marcel Tijsterman

*Co-first authors

Department of Toxicogenetics, Leiden University Medical Center, The Netherlands

Published in Genome Research 2014 March: 2014. 24: 954-962

ABSTRACT

DNA lesions that block replication fork progression are drivers of cancer-associated genome alterations, but the error-prone DNA repair mechanisms acting on collapsed replication are incompletely understood, and their contribution to genome evolution largely unexplored. Here, by whole genome sequencing of animal populations that were clonally propagated for over 50 generations, we identify a distinct class of deletions that spontaneously accumulate in *C. elegans* strains lacking translesion synthesis (TLS) polymerases. Emerging DNA double-strand breaks are repaired via an error-prone mechanism in which the outermost nucleotide of one end serves to prime DNA synthesis on the other end. This pathway critically depends on the A-family polymerase theta, which protects the genome against gross chromosomal rearrangements. By comparing the genomes of isolates of *C. elegans* from different geographical regions, we found that in fact most spontaneously evolving structural variations match the signature of polymerase Theta-Mediated End Joining (TMEJ), arguing that this pathway is an important source of genetic diversification.

INTRODUCTION

Identifying the mechanisms that fuel genome change is crucial for understanding evolution and carcinogenesis. Spontaneous mutagenesis is caused predominantly by misinsertions or slippage events of replicative polymerases that are missed by their proofreading domains, and not corrected by mismatch repair (Lynch 2008). Less frequently, but with a potentially much more detrimental effect, mutations can arise when DNA damage obstructs progression of DNA replication; and stalled replication forks eventually collapse, resulting in highly mutagenic double stranded breaks (DSBs). While error-free homologous repair, where the sister chromatid is used as a template, restores the original sequence, infrequent but highly mutagenic error-prone end joining processes can give rise to spontaneous deletions and tumor-promoting translocations (Mitelman et al. 2007).

To circumvent fork collapse at DNA damage, cells employ various alternative polymerases that are capable of incorporating nucleotides across DNA lesions, and are hence called translesion synthesis (TLS) polymerases. TLS acts on a wide variety of DNA lesions that can result from endogenous as well as exogenous genotoxic sources: DNA lesions that result from UV-light exposure, for instance, are efficiently bypassed by the well-conserved TLS polymerase eta (pol eta), inactivation of which in humans leads to the variant form of the skin cancer predisposition syndrome Xeroderma Pigmentosum (Johnson et al. 2007; Masutani et al. 1999b). Abundant *in vitro* studies demonstrate the involvement of TLS polymerases pol eta and pol kappa in bypass of lesions that are produced by endogenous reactive compounds, arguing that these polymerases are also essential for protection of the genome under unchallenged conditions (Fischhaber et al. 2002; Kusumoto et al. 2002; Haracska et al. 2000).

Although error-prone while replicating, and thus potentially causing misinsertions, TLS polymerases are thought to protect cells against the more mutagenic effects of replication fork collapse (Knobel and Marti 2011). Here, we investigate the contribution of TLS polymerases on the maintentance of genome stability and the mechanisms acting on stalled DNA replication, by characterizing *C. elegans* strains that are defective for the Y-family polymerases pol eta and pol kappa. Unexpectedly, we found that DSBs resulting from replication blocking endogenous lesions are not repaired via canonical DSB repair pathways but through an error-prone repair mechanism that critically depend on the A-family DNA polymerase theta (pol theta).

RESULTS

TLS polymerases protect genomes against spontaneous deletions

In previous work, we established the role of the *C. elegans* homologs of TLS polymerases pol eta (POLH-1) and pol kappa (POLK-1) in protection against a wide range of exogenous DNA damaging agents (Roerink et al. 2012). In these studies, we also sporadically observed readily recognizable mutant phenotypes during normal culturing of *polh-1polk-1* double mutant animals, which prompted us to suspect a prominent role for these Y-family of TLS polymerases in the prevention of spontaneous mutations (Figure S1). To address the nature of this increased mutagenesis in an unbiased way, we cultured populations of animals with specific defects in TLS for 60 generations, thus allowing spontaneous mutations to accumulate, and then sequenced their genomes (Figure 1A, Table S1 and Supplemental data file). Mutation accumulation (MA) lines of a wild-type strain (Bristol N2) and of the mismatch repair deficient strain *msh-6* - for which an ~100-fold higher mutation frequency has been reported (Tijsterman et al. 2002) - were sequenced as references. All

genomes have been sequenced with a minimal 12 fold base coverage (Table S1).

Although pol eta and pol kappa have reduced accuracy while replicating from undamaged as well as damaged DNA templates (Matsuda et al. 2000; Ohashi et al. 2000), we found that these proteins hardly contribute to base substitution processes or microsatellite instability under normal growth conditions: no significant elevation in the substitution or microsatellite mutation rates were found in *polh-1polk-1* worms as compared to wild-type controls (Figure 1B), which argues that another class of genetic changes must be responsible for the observed mutator phenotype. To detect other structural variations, we employed Pindel software, developed to identify deletions and/or insertions in whole-genome sequencing data (Ye et al. 2009). Strikingly, a unique class of deletions emerged in *polh-1* and *polh-1polk-1* mutants, which were not associated with repetitive loci, with sequences able to adopt stable secondary structure (*e.g.* G4 DNA), or with any other obvious genomic trait, and occurred at seemingly random locations throughout the genome (Figures 1C and S2). The vast majority of deletions ranged between 50 and 200 bp in size, with just a few exceptions being larger or smaller (Figure 1D). The median size, of 107 bp, was similar for deletions derived from either *polh-1* or *polh-1polk-1*-mutant animals (Figure 1D). Control wild-type and *msh-6* samples did not display any mutations from this class. Deletions occurred in *polh-1* single mutants with a rate of ~0.4 per animal generation, which translates to an average of ~0.03 deletion per genome per cell division. *polk-1* single mutants hardly suffered from deletions; however, *polh-1polk-1* double mutants had 5-fold increased rates of deletion induction as compared to *polh-1* single mutant animals, implying that *C. elegans* pol eta and pol kappa function redundantly on a subset of endogenous lesions.

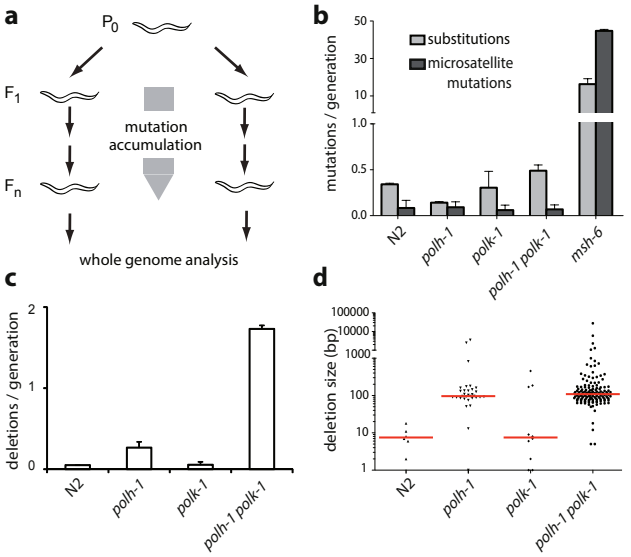


Figure 1. Spontaneous mutagenesis in TLS deficient strains. (A) Generation of mutation accumulation (MA) lines. For each genotype multiple populations were started by cloning out single worms from a single hermaphrodite P0. Cultures were propagated by transferring animals to new plates each generation. At generation Fn, a single animal was grown to a full population of which genomic DNA was isolated and subjected to whole genome sequencing on an Illumina HiSeq. (B) Substitution and microsatellite mutation rates for the indicated genotypes. Mutation rates are expressed as the number of mutations per generation divided by the number of nucleotides analysed. (C) Rates of structural variations for the indicated genotypes. (D) Size distribution of deletions in the different mutant backgrounds. The median sizes are indicated in red.

DSB induction in TLS deficient mutants

To further investigate the origin of the high number of deletions in *polh-1polk-1* deficient strains, we looked for manifestations of genomic instability in germ cells of these animals. We observed a mild but statistically significant increase in the number of foci of the DSB marker RAD-51 in proliferating germ cells of *polh-1polk-1* mutant animals (Figure S3A-B). Elevated levels of DSBs, are also suggested by the spontaneous emergence of dominant *him* mutants in *polh-1polk-1* mutant populations (Figure S1). This phenotype, which is defined by dominant inheritance of an increased number of males (XO) in predominantly hermaphroditic (XX) populations, has previously been found upon exposure to γ-irradiation and in mutants with enhanced telomere shortening, where it proved to result from X/autosome translocations (Herman et al. 1982; Meier et al. 2009). Despite these manifestations of enhanced replication stress in *polh-1polk-1* mutants, the levels of DSBs were insufficient to activate the two DNA damage checkpoints that operate in the *C. elegans* germline: cell cycle arrest and apoptosis (Gartner et al. 2000). We found neither a reduction in germ cell proliferation nor an increase of apoptotic bodies in *polh-1polk-1* mutant germlines, suggesting that TLS compromised germ cells proliferate in the presence of elevated levels of DSBs, with genomic deletions as a consequence (Figure S3C-E).

Footprints of error-prone DSB repair

To obtain mechanistic insight on the biology of deletion formation, we performed a detailed analysis on the sequence context of 141 *polh-1polk-1*-derived deletions (Supplemental data file). While the majority had simple deletion junctions (without inserts), about 25 percent of the footprints showed insertions of short sequence stretches (Figure 2A). Cases with inserts sufficiently long to faithfully trace their origin revealed that the inserted stretch, or part of it, is identical to sequences flanking the deletion (Figure 2B-C). This finding strongly suggests that DNA close to the break site was used as a template for *de novo* synthesis before both DNA ends were joined.

A DSB repair mechanism involving DNA synthesis is also suggested by the notion of a 'priming' nucleotide in more than 80 percent of all deletions: 83 of the 102 deletions without insert contain at the junction at least one nucleotide could have originated from either flank; in 51 cases this is restricted to a single nucleotide. To systematically assess the significance of this observation, we constructed deletion junction heat maps, which reflect the level of (micro)homology between 5' and 3' junctions (Figure 2D-F). We scored the degree of sequence identity in an 8 nt window, encompassing the 4 outermost nucleotides of the flanking sequence and the 4 nucleotides of the adjacent, but deleted, sequence. Indeed, compared to a randomly distributed simulated set, we found a very high similarity score for the nucleotide at the -1 position of the deletions and the +1 position of the opposing flanks ($p=7.3x10^{-17}$). Importantly, this profound degree of microhomology is restricted to only a single, the terminal nucleotide.

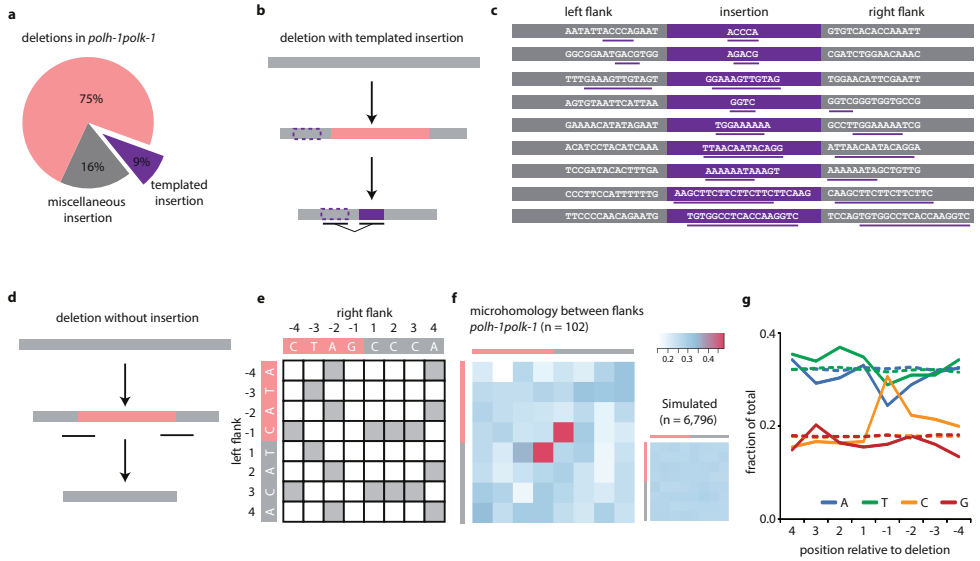


Figure 2. Deletion footprints in TLS mutants indicate a priming-based end joining mechanism. (A) Distribution of deletion footprints in *polh-1polk-1* mutants. (B) Schematic illustration of a deletion associated with a templated insertion. Deleted sequence in pink; newly inserted sequence in purple and its template boxed; non-altered DNA in grey. (C) Sequence context of deletions with templated insertions derived from *polh-1polk-1* animals. Matching sequences are underlined. (D) Schematic illustration of a deletion not accompanied by insertions. Deleted sequence in pink; non-altered DNA in grey. The eight nucleotide window -capturing neighbouring flanking and deleted sequences- that is used for the generation of the heat maps is underlined. (E) The strategy to score junction homology: for each simple deletion, matching bases between the 5′ and 3′ junction were scored 1, non-matching bases were scored 0, thus creating one map per deletion. (F) A heat map representing the sum of all individual deletion maps derived from *polh-1polk-1* animals. (n=102). A heat map for a simulated set of deletions (n=6796) with random distribution is displayed on the right. (G) Base composition at the 5′ and 3′ junctions. The flanking sequences have positive numbers, the deleted sequence have negative; -1 being the first nucleotide within the deletion. Dotted lines indicate the relative abundance of a particular base for a simulated set of deletions (n=6796).

Replication blocking endogenous damage resides at guanines

We next investigated whether the deletion specifics would reveal the nature of the spontaneous damage underlying fork stalling and break formation in TLS compromised animals using the following rationale: deletions in TLS deficient animals are likely brought about because of an inability to incorporate a base across endogenous lesions. If the nascent strand, blocked at the site of base damage, defines one end of the deletion junction, then the -1 position of the corresponding junction will represent the nucleotide complementary to the damaged base: it is the first base not to be incorporated. To test this hypothesis, we plotted the base distribution for each position of the junction and indeed found it not to be random at the -1 position, but rather being dominated by cytosines (Figure 2G). This result strongly argues that spontaneous base damage that requires pol eta and pol kappa to avoid DSB induction resides at guanines, which may point towards N2-dG and/or 8-oxo-dG adducted sites as a primary source of spontaneous mutagenesis.

Deletion formation is dependent on pol theta.

The frequent occurrence of templated insertions at the deletion junctions suggests the involvement

of a DNA polymerase to repair DSBs that are induced at replication-blocking dG bases. One candidate is the A-family DNA polymerase theta, which was previously implicated in repair of interstrand crosslinks in various models and in repair of transposition-induced DSBs in Drosophila (Muzzini et al. 2008; Shima et al. 2004; Yousefzadeh and Wood 2012). We recently identified a role for pol theta in preventing genomic instability at endogenous sequences that are able to fold into potentially replication blocking G-quadruplex structures (Koole et al. 2014). To test a possible role for this protein in deletion formation at spontaneous damage, we generated animals defective for *polh-1polk-1* and the *C. elegans* pol theta homolog *polq-1*. Strikingly, these animals are severely compromised in normal growth: while *polq-1* and *polh-1polk-1* animals had nearly wild-type growth characteristics, *polh-1polk-1polq-1* triple mutant animals had very much reduced fertility, albeit in a stochastic fashion, ranging from complete sterility to brood sizes of 30 percent of wild-type levels (Figure 3A). Associated with these fertility defects, we observed a profound increase in the number of RAD-51 foci in the proliferative zone of the germline as well as activation of the DNA damage checkpoint suggesting increased DNA end-resection and DSB signaling (Figure 3B-C, Figure S3E). From this we conclude that when damage cannot be bypassed, pol theta action safeguards animal fertility by preventing undesired HR-related processing of replication-associated breaks, which trigger checkpoint activation and prohibit proliferation.

Because the notion of endogenous damage blocking the replication fork can only be inferred indirectly from our data, we tested whether a similar detrimental effect of knocking out pol theta is also observed on bona-fide fork-stalling lesions, such as UV-induced photoproducts. Indeed, mutating pol theta strongly sensitizes *polh-1* mutant animals, but not otherwise wild type animals to UV exposure (Figure S3F), further strengthening the conclusion that pol theta action minimizes the toxic effects of persistent replication blocking DNA lesions, that result from either endogenous or exogenous source.

To study the role of pol theta in deletion formation on a molecular level, we assessed mutagenesis using an endogenous *unc-22* reporter gene (Figure 3D). We isolated spontaneous *unc-22* mutants from *polh-1polk-1* and *polh-1polk-1polq-1* populations and determined their molecular nature using PCR and Sanger sequencing. In perfect agreement with our whole-genome sequencing data, all *unc-22* mutations derived from *polh-1polk-1* animals were 50-200 bp deletions characterized by single nucleotide homology and templated insertions (Figure 3D, Table S2). In sharp contrast, *unc-22* mutants derived from *polh-1polk-1polq-1* triple mutants, while being induced at comparable rates, were of a completely different size category. Here, deletions were typically larger than 5 kb, with some spanning over 30 kb of genomic sequence, thus amplifying the deleterious impact of replication stalling lesions more than 100-fold (Figure 3D, Tables S2 and S3). We conclude that a pol theta-mediated end joining mechanism is responsible for the generation of small-sized deletions induced by replication fork stalling endogenous lesions. In its absence, large stretches of DNA surrounding DSBs are resected, resulting in abundant RAD-51 filament formation, mitotic checkpoint activation and excessive loss of DNA.

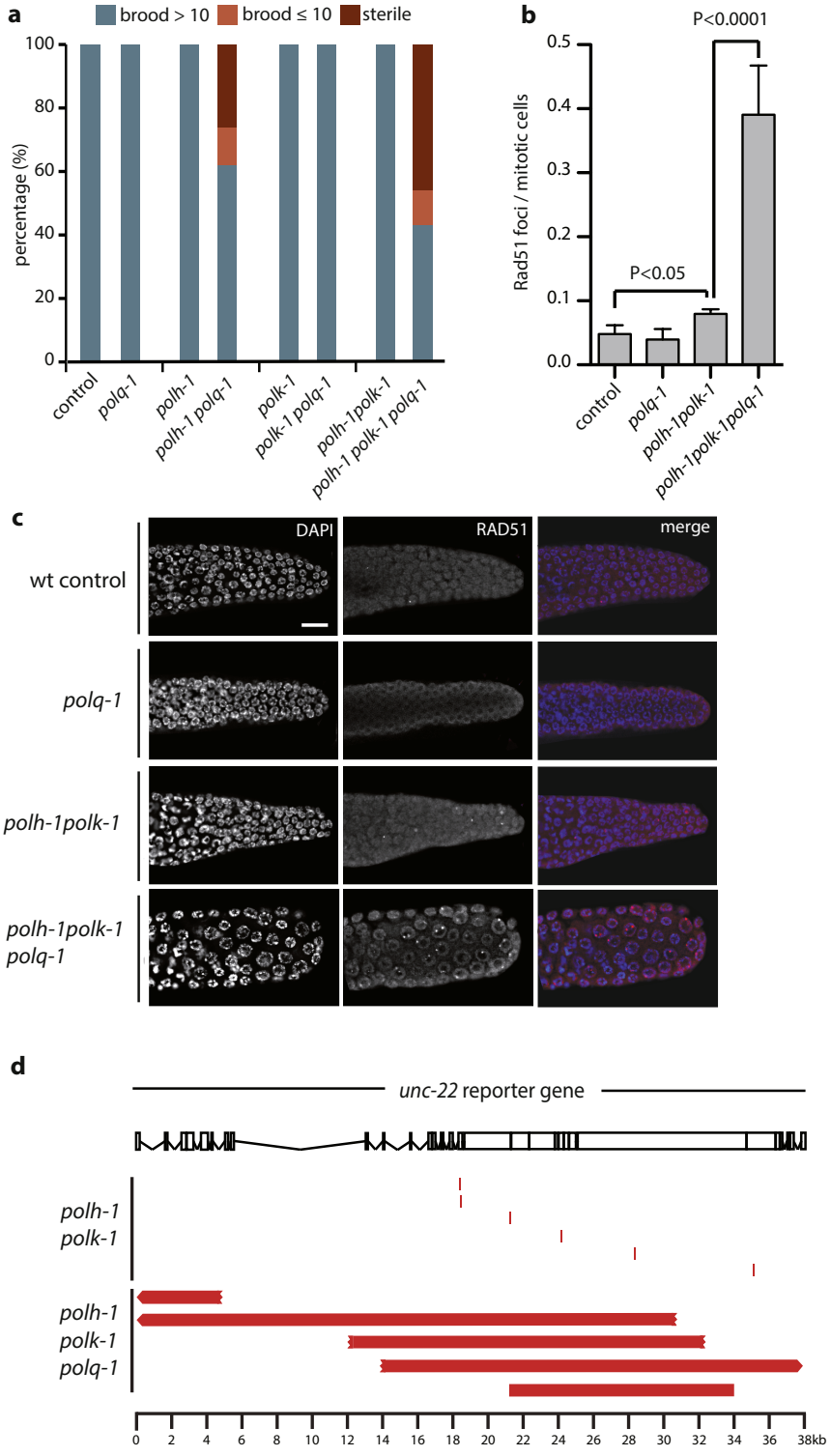

a
brood > 10
brood ≤ 10
sterile
percentage (%)
100
80
60
40
20
0
control
polq-1
polh-1
polh-1 polq-1
polk-1
polk-1 polq-1
polh-1polk-1
polh-1 polk-1 polq-1
b
P<0.0001
P<0.05
Rad51 foci / mitotic cells
0.5
0.4
0.3
0.2
0.1
0.0
control
polq-1
polh-1polk-1
polh-1polk-1polq-1
c
DAPI
RAD51
merge
wt control
polq-1
polh-1polk-1
polh-1polk-1
polq-1
d
unc-22 reporter gene
polh-1
polk-1
polh-1
polk-1
polq-1
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38kb

Figure 3. Pol theta mediates end joining of breaks in pol eta and pol kappa deficient animals. (A) Fecundity of single, double and triple knockout mutants of pol theta and TLS Polymerases pol eta and pol kappa. (B) Quantification and (C) representative pictures of RAD-51 immunostainings on germlines of the indicated genotype. Scale bar, 10 µm (D) Schematic representation of the *unc-22* reporter gene and spontaneous deletions (in red) isolated from either *polh-1polk-1* or *polh-1polk-1polq-1* mutant animals. Three out of five deletions extended beyond the borders of the *unc-22* locus.

Pol theta in wild-type C. elegans strains

The notion that we have uncovered a role for pol theta in genome protection under TLS deficient conditions raises the question: does pol theta-mediated repair also act when TLS is functional? Or in other words, how relevant is this error-prone repair pathway for animal fitness? We hypothesized that the action of an error-prone repair mechanism with such a clear and distinct signature, *i.e.* a distinct size class, single nucleotide homology and templated insertions, may leave its fingerprint in evolving genomes. For this reason, we compared the genomes of different natural isolates of *C. elegans*, to identify structural variations and defined their characteristics (Figure 4). The majority of deletions are of small size - 60 percent being smaller than 10 bp - while the number of deletions decreases with increasing size in an exponential manner. However, we found deletions in the size range 50-200 bp much more abundantly present than expected from this exponentially declining trend (Figure 4B). Moreover, deletions in this size range bear the pol theta signature: templated insertions and a strong overrepresentation (over 80 percent) of having at least one nucleotide homology (Figure 4C), which supports a role for pol theta in genome change during non-challenged growth. Unexpectedly, we observed templated insertions (2%) also in the small size range of deletions, and found also this class to be dominated by ≥1 nucleotide homology at the junction (Figure 4C-D), hinting to a much broader involvement of pol theta in genomic change, not being restricted to the creation of 50-200 bp deletions.

To further investigate the potential contribution of pol theta in spontaneous mutation induction under non-challenged growth conditions we used a forward mutagenesis assay that is based on the uncoordinated movement of animals carrying a dominant mutation (e1500) in the UNC-93 protein that affects muscle contraction (De Stasio et al. 1997; Greenwald and Horvitz 2003). This phenotype is suppressed by complete loss of function of *unc-93,* or by loss of one of several extragenic suppressor genes (*e.g. sup-9*, *sup-10*). We propagated populations of wild-type and *polq-1* mutant animals out of which we isolated and molecularly characterized revertants animals that had normal movement. Strikingly, the total number of revertants was increased fourfold in *polq-1* mutants (Figure 4E, Tables S4 and S5), demonstrating that pol theta action prevents mutation induction also in wild type animals during normal growth. The increased mutagenesis in *polq-1* is mainly attributed to a selective increase in large chromosomal deletions, similar to those previously identified in *unc-22* in *polh-1polk-1polq-1* deficient strains (Figure S5). Interestingly, we observed that one mutation class, *i.e.* small deletions of a size ≥3 bp, was completely absent in animals *polq-1* (3/28 in wild type vs 0/111 in *polq-1* mutants), arguing, together with the identification of pol theta signature carrying small-sized deletions in the genomes of natural isolates, that pol theta protect cells against arrest and the genome against large chromosomal DNA loss, but at the price of small deletions.

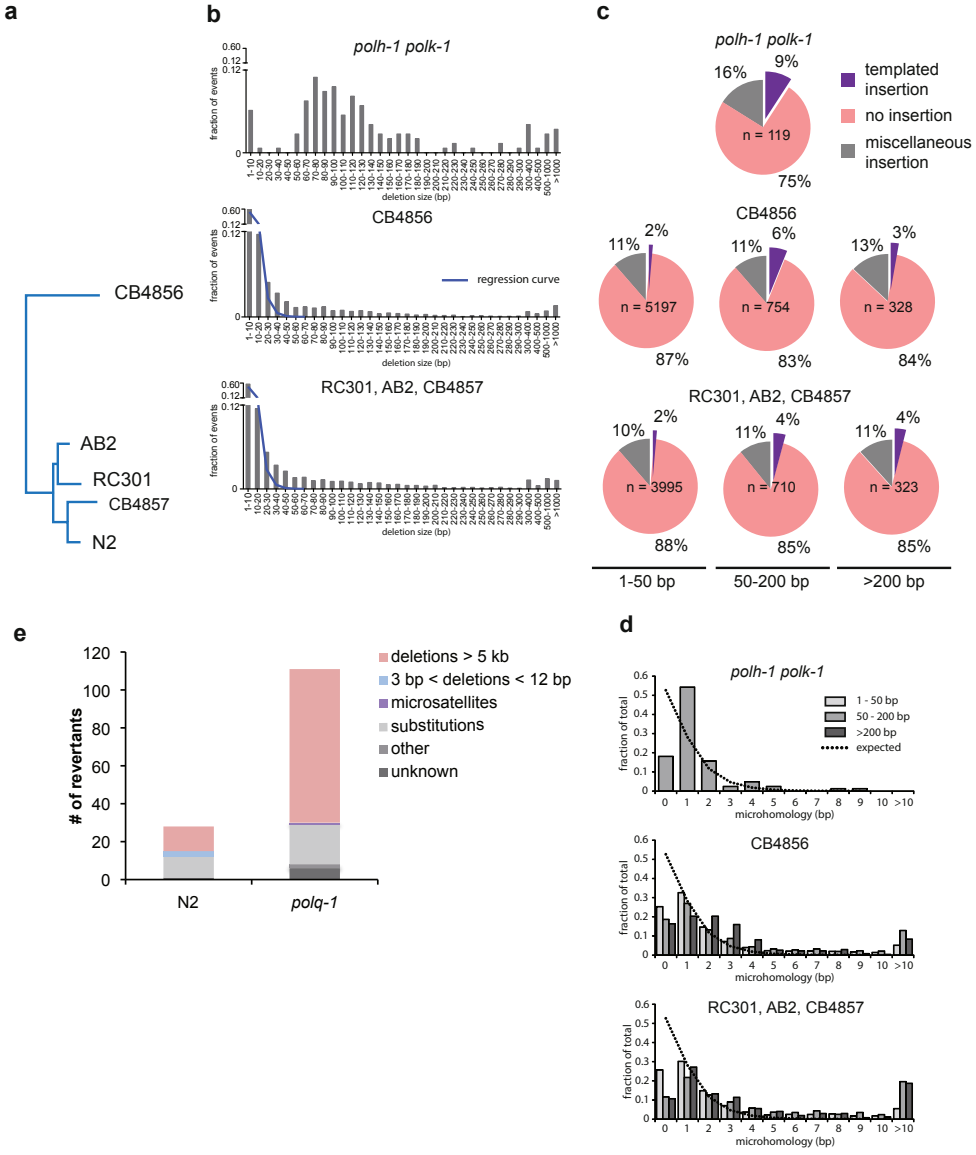


Figure 4. Signature of pol theta-mediated end joining in natural isolates of *C. elegans*. (A) Phylogenetic tree diagram of the different isolates of *C. elegans* used in this study. (B) Size distribution of deletions of evolutionary distinct *C. elegans* species compared to size distribution of *polh-1polk-1* derived deletions. An exponential regression curves describes the size distribution of deletions in both natural isolates up to 20 bp; deletions up to 200 bp are overrepresented. (C) Deletions in natural isolates, especially in size class 50-200 bp show templated insertions analogously to deletion footprints in *polh-1polk-1* animals. (D) Microhomology for deletions in natural isolates as compared to deletions in *polh-1polk-1* animals. (*E*) *unc-93* mutagenesis in *polq-1* worms and wild-type controls.

Discussion

TLS polymerases eta and kappa operate on endogenous lesions in an error free manner

Our data present the first evaluation of the contribution of two main members from the Y family polymerases eta and kappa on the stability of an animal's entire genome under unchallenged conditions. We show that these TLS polymerases prevent the induction of spontaneous deletions. Although *in vitro* studies demonstrated reduced accuracy of pol eta and pol kappa while replicating from undamaged and damaged DNA templates (Johnson et al. 1999; Fischhaber et al. 2002; Masutani et al. 1999a; Matsuda et al. 2000; Kusumoto et al. 2002; Ohashi et al. 2000; Haracska et al. 2000), our *in vivo* data show that the biologically desired bypass action of pol eta and pol kappa is largely error-free: their joined action prevents ~2 deletions per animal generation without significantly affecting the overall substitution rate (Figure 1B).

Deletions were found in animals deficient for pol eta, but not in pol kappa mutant strains. Pol kappa nevertheless can act on spontaneous damage as a greatly increased number of deletions result from the combined absence of both pol eta and pol kappa. This outcome argues that the two Y-family members function redundantly on a subset of endogenous lesions, a conclusion that is further supported by a similar genetic interaction for sensitivity towards the guanine alkylator MMS. Also for this exogenous source of DNA damage, animals deficient for both pol eta and pol kappa are profoundly more sensitive than animals deficient for only pol eta, while pol kappa disruption by itself only very mildly increases the sensitivity of wild-type worms. (Roerink et al. 2012). Under non-challenged conditions, we found deletion junctions to preferentially result from replication fork stalling at dG residues (Figure 2G), which may point towards N2-dG and/or 8-oxo-dG adducted sites as a primary source of spontaneous mutagenesis, as bypass activities of pol eta and pol kappa have been reported for these lesions (Avkin et al. 2004; Haracska et al. 2000).

An error-prone pol theta-mediated mechanism for repair of replication-associated DSBs

The footprints of the deletions that are suppressed by TLS polymerases fit best with a model in which one end of a DSB, induced at replication-blocking lesions, is extended using the other end as a template, with just a single base-paired nucleotide as a primer (explaining both single nucleotide homology and templated insertions). In this model, templated inserts can be explained as the result of iterative rounds of annealing and extension (Figure 5). The close proximity of insertions to their template also suggests that the extendable end of the DSB is not subjected to extensive trimming and suggests that DNA close to the break site was used as a template for *de novo* synthesis before both DNA ends were joined. A 'priming' nucleotide in more than 80 percent of all deletions further strengthened our hypothesized model of a DSB repair mechanism involving DNA synthesis. Further support is provided by the identification of a polymerase, the A-family polymerase pol theta, which we found to be essential for the formation of small-sized deletions. The molecular function of this protein in previously identified phenotypes, such as sensitivity towards crosslinking agents and radiation, as well as spontaneous genome instability in mice was largely unknown (Muzzini et al. 2008; Shima et al. 2004; Yousefzadeh and Wood 2012). We now show that a pol theta-dependent repair route provides cells with the ability to repair replication-associated breaks; we propose to refer to this pathway as TMEJ, for Pol Theta-Mediated End Joining, to set it apart from canonical NHEJ. We hypothesize that TMEJ may be specifically important in cases where the sister chromatid cannot be used as a DSB repair template for *e.g.* homologous recombination because that template still contains the original replication-blocking lesion (Figure

5). In favor of a role of TMEJ in preventing futile HR, we observed abundant RAD-51 filament formation, mitotic checkpoint activation and excessive loss of DNA in the absence of pol theta. When damage cannot be bypassed, pol theta action safeguards animal fertility by preventing undesired HR-related processing of replication-associated breaks, which would trigger checkpoint activation and prohibit proliferation.

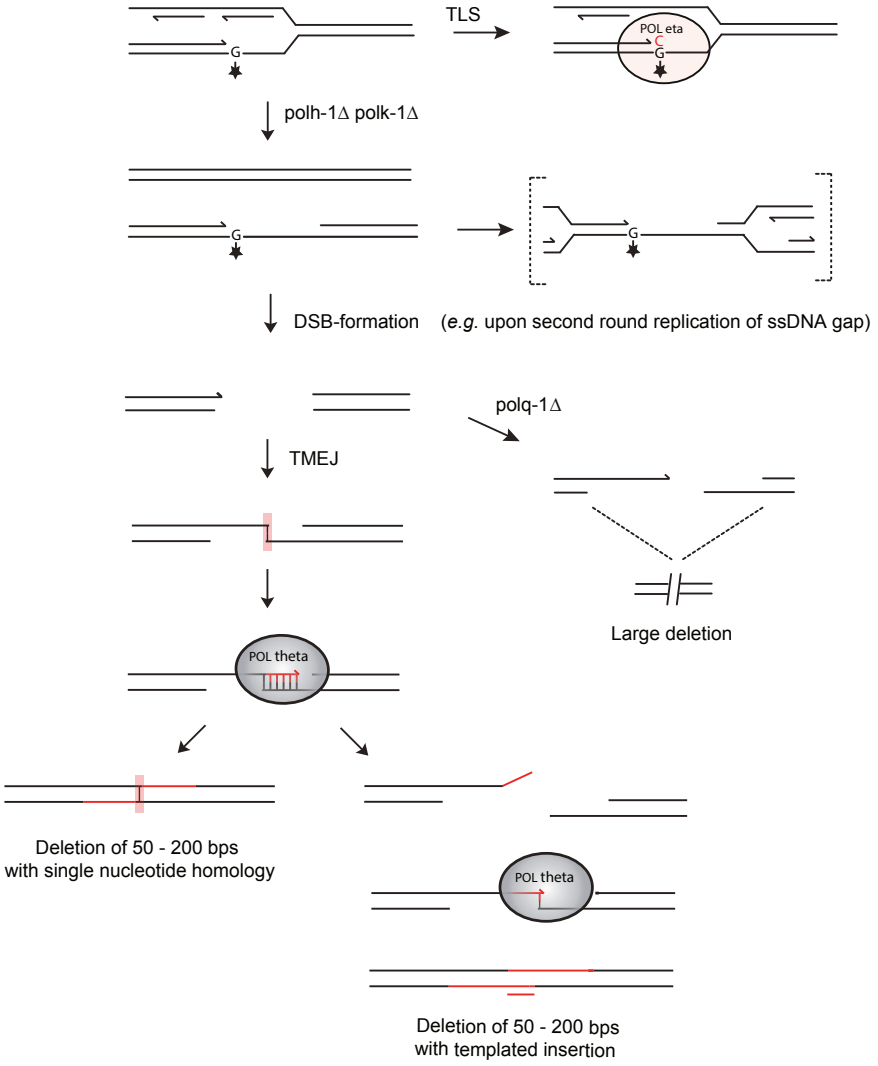


Figure 5. A tentative model for TMEJ of breaks induced at replication fork barriers. DNA lesions from endogenous sources - with increased frequency in the absence of functional TLS - causes replication fork blocks, leading to double stranded breaks. The broken ends are repaired by pol Theta-Mediated End Joining (TMEJ), which is stimulated by minimal priming of 1 base pair, explaining deletions with single nucleotide homology (left). Iterative cycles of priming, extending and dissociation will result in deletions with templated insertions (right). In pol theta deficient cells, DNA breaks resulting from replication fork stalling are differently processed, eventually leading to deletions of larger size.

Our model for TMEJ is conceptually different from the models that have previously been proposed to explain copy number variations and complex rearrangements in tumors and congenital disorders: microhomology-mediated break induced replication (MMBIR), and Fork stalling and Template switching (FoSTeS) (Hastings et al. 2009; Lee et al. 2007; Zhang et al. 2001). The genome rearrangements explained by these models are also characterized by the presence of limited sequence homology at the rearranged DNA junctions, however, both these models invoke the invasion of a 3' single strand end, either resulting from DNA breaks (MMBIR), or stalled forks (FoSTes) into the sister molecule or into another replication fork that is in 3D physical proximity, to reassure ongoing DNA replication. Our data on deletion junctions that result from blocked replication either at endogenous lesions (this manuscript) or secondary structures such as G4 DNA (Koole et al. 2014) favor an end-joining mechanism based on the presence of two-ended double strand breaks - which may be the result from replication of gapped DNA intermediates that form at persistent replication fork blocks (Figure 5) - as opposed to restarting replication of a one-ended break. The observation that Mus308, the Drosophila ortholog of pol theta, can act on dsDNA breaks resulting from P-element excision, is also in concert with an end-joining mechanism.

Another difference between TMEJ and MMBIR/FoSTeS relates to size; whereas TMEJ deletions are typically 50-200, the rearrangements that are explained by MMBIR/FoSTeS models span kilobases. Nevertheless, all models evoke the presence of flexible primer-template intermediates that can be extended in recurrent cycles, and imply DNA polymerase action. Important in that respect is the recent observation that MMBIR-type rearrangement in mammalian cells can be induced by replication stress and depend at least in part on the Pol delta subunit PolD4 (Costantino et al. 2014).

Of interest, while the vast majority of genomic rearrangements that we observed in TLS compromised animals are 50-200 deletions, we nevertheless found a very small number of more complex rearrangements (Supplemental data file). These events may, because of their complexity and size, be more resembling the complex rearrangement found in mammalian cells, however, their number was too limited to allow systematic analyses, and none were found in any of our other less sensitive phenotype-based assays, thus precluding genetic analysis at this stage.

TMEJ footprints in evolving genomes

The observation that pol theta also suppresses mutagenesis in wild-type animals, together with the notion that the signature of TMEJ is apparent in the genomes of natural isolates of *C. elegans* argues for a prominent role of this error-prone pathway to protect genomes against large chromosomal rearrangements. This role seems not to be restricted to replication fork stalls. While the class of 50-200 bp deletions that is seen in TLS deficient animals, is found overrepresented in genomes of natural isolates, the predominant fraction of deletions are smaller in size. Still, these smaller-sized deletions bear a TMEJ signature, in that they are characterized by single nucleotide homology and frequently are associated with templated insertions. A broader role for TMEJ, thus being responsible for many types of structural variations, is also supported by the *unc-93* forward mutagenesis assay, where small deletions (3 to 12 bp) were exclusively found in pol theta-proficient strains.

While mutagenic processes are drivers of evolution, they also fuel malignant transformation of cells. It is a current challenge to recognize specific classes of mutations in cancer genomes and attribute these either to underlying sources of DNA damage or to error-prone repair mechanisms. Identifying mutational signatures typifying specific repair processes is pivotal to this ambition.

Templated insertions and the use of minimal homology - two characteristics of TMEJ - have frequently been observed in higher order eukaryotes and in cancer tissues (Chen et al. 2010; Nik-Zainal et al. 2012; Carvalho et al. 2013), and have been ascribed to either classical non-homologous end joining or the molecularly ill-defined process of microhomology-mediated end joining (Honma et al. 2007; Kloosterman et al. 2012). Here, we describe a mechanistic alternative for repair of DSBs induced at stalled forks, which leaves a distinct and well-defined footprint in evolving genomes.

Methods

C. elegans genetics

All strains were cultured according to standard methods (Brenner 1974). Wild-type N2 (Bristol) worms were used in all control experiments. Alleles used in this study are: *polh-1 (lf31)*; *polh-1 (ok3317)*; *polk-1 (lf29)*; *polq-1(tm2026); msh-6(pk2504)*; *bcIs39[P(lim-7)ced-1::GFP + lin-15(+)]); unc-93(e1500)*. All mutant strains were backcrossed six times before performing experiments.

Whole genome sequencing of MA lines

Mutation accumulation (MA) lines were generated by cloning out F1 animals from one hermaphrodite. Each generation about five worms were transferred to new plates. MA lines were maintained for 60 generations or until severe growth defects developed. Single animals were then cloned out and propagated to obtain full plates for DNA isolation. Worms were washed off with M9 and incubated for one hour at room temperature while shaking, to remove bacteria from the animal's intestine. After two washes, worm pellets were lysed for two hours at 65°C with SDS containing lysis buffer. Genomic DNA was purified by using a DNeasy kit (Qiagen). Paired end (PE) libraries for whole genome sequencing (HiSeq2000 Illumina) were constructed from genomic DNA according to manufacturers' protocols with some adaptations. Shortly, 5 g DNA was sheared using a Covaris S220 ultrasonicator, followed by DNA end-repair, formation of 3'A overhangs using Klenow and ligation to Illumina PE adapters. Adapter-ligated products were purified on QIAquick spin columns (Qiagen) and PCR-amplified using Phusion DNA polymerase and barcoded Illumina PE primers for 10 cycles. PCR products of the 300 - 400 bp size range were selected on a 2% ultrapure agarose gel and purified on Qiaquick spin columns. DNA quality was assessed and quantified using an Agilent DNA 1000 assay. Four to five barcoded libraries were pooled in one lane for sequencing on a HiSeq.

Bioinformatic analysis

Image analysis, basecalling and error calibration was performed using standard Illumina software. For the analysis of the natural isolates paired-end whole genome sequence data was downloaded from the NCBI Sequence Read Archive (SRP011413) (Grishkevich et al. 2012), and sequence reads were mapped to the *C. elegans* reference genome (Wormbase release 225) by BWA. SAMtools was used for SNP and indel calling, with BAQ calculation turned off (Li et al. 2009). All non-unique SNPs and indels are considered to be pre-existing and were filtered out using custom Perl scripts. To identify microsatellite mutations and deletions we used Pindel, developed by Ye et al (Ye et al. 2009). A more detailed description of the bioinformatic procedures is enclosed in the supplemental information.

Microscopy

To study RAD-51 foci formation, germlines were dissected, freeze cracked and subsequently washed with 1% Triton and methanol (-20°C). RAD-51 was visualized by using an anti-RAD-51 rabbit monoclonal antibody and an Alexa488-labelled goat-anti-rabbit secondary antibody (Molecular Probes Inc), combined with 10 μg/mL DAPI. Dissected worms and eggs were mounted using Vectashield. Apoptosis was monitored using a *lim-7* driven *CED-1*::GFP fusion, which visualises sheath cells surrounding apoptotic germ cells. All microscopy was performed with a Leica DM6000 microscope.

UV sensitivity assay

To assess the sensitivity to germ cells to UV-exposure, young adults were exposed to various doses of UV light, and subsequently allowed to lay eggs for 48 hrs. 24 hrs later, the number of non-hatched eggs and surviving progeny were determined.

unc-22 mutagenesis assay

To identify spontaneous mutations in the *unc-22* muscle gene we started 50 populations by transferring a single animal to 9 cm plates seeded with OP50. In the case of the synthetically sick *polh-1polk-1polq-1* mutant, we started 200 populations with 5 worms per plate. Animals were washed off with 2 mM levamisole and transferred to 6-well plates to facilitate scoring of *unc-22* mutants, which are insensitive to the hypercontracting effects of the drug levamisole. Independent *unc-22* mutant animals were isolated. Genomic DNA was isolated from homozygous animals for subsequent PCR and sequence analysis.

unc-93 (e1500) mutagenesis assay

To generate a complete spectrum of spontaneous mutations we used a mutagenesis assay based on reversion of the socalled 'rubber band' phenotype, caused by a dominant mutation in the muscle gene *unc-93* (De Stasio et al. 1997; Greenwald and Horvitz 2003). Reversion of the *unc-93(e1500)* phenotype is caused by homozygous loss of *unc-93* or one of the suppressor genes *sup-9*, *sup-10*, *sup-11* and *sup-18*. *polq-1(tm2026) unc-93(e1500)* and *unc-93(e1500)* animals were singled to 2 x 400 6 cm plates. These plates were grown till starvation and equal fractions (chunks of 2 x 2 cm) were then transferred to 9 cm plates. Before these plates were fully grown, they were inspected for wild-type moving animals. From each starting culture only one revertant animal was isolated to ensure independent events.

Large chromosomal deletions in *unc-93*, *sup-9* and *sup-10* were identified by PCR amplification of exonic regions and two regions 5 kb upstream and downstream of the respective genes. Smaller genetic changes and substitutions were first classified into events in either the *unc-93* gene or in one of the suppressor genes by their ability to complement a known *unc-93* deletion allele. All *unc-93* exons were sequenced in revertant animals that failed to complement *unc-93*, whereas all exons of *sup-9* and *sup-10* were sequenced in revertants that complemented *unc-93*. *sup-11* or *sup-18* could not be subjected to molecular analysis due to lack of sequence data. Revertants that complemented *unc-93* but had not detectable mutation of *sup-9* or *sup-10*, were classified as 'unknown'.

Data access

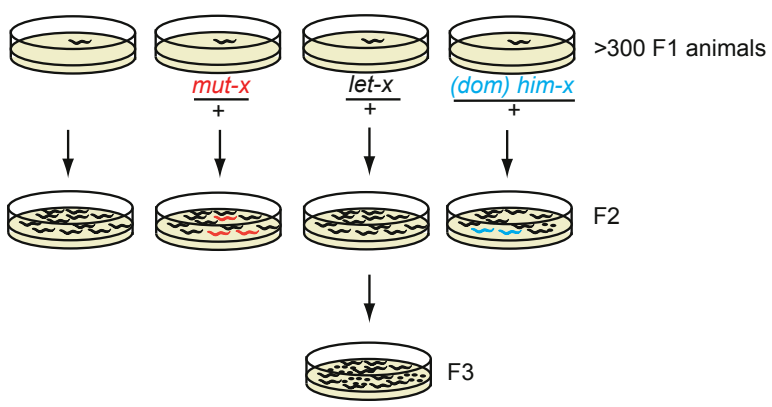
The sequencing data have been submitted to the NCBI Sequence Read Archive under accession number SRP020555. For the analysis of the natural isolates, paired-end whole genome sequence

data was downloaded from the NCBI Sequence Read Archive (SRP011413) (Grishkevich et al. 2012).

Acknowledgements

We thank the *C. elegans* Knockout Consortium, Shohei Mitani and the C. elegans Genetics Center for providing strains. We thank Wouter Koole and Jane van Heteren for critical reading of the manuscript and Bennie Lemmens and Harry Vrieling for discussions. MT is supported by grants from the European Research Council (203379, DSBrepair) and ZonMW/NGI-Horizon, Zenith.

a



b

Genotype	# analyzed plates	Mutants found
N2	340	0
polh-1(ok3317)	340	0
polh-1(lf31)	340	0
polk-1(lf29)	340	0
polh-1(ok3317);polk-1(lf29)	740	*dpy(3); ste(1); let(15); him(6)*
polh-1(lf31);polk-1(lf29)	340	*dpy(3); let(5); him(3)*
msh-6	300	20 visible mutants

Figure S1. Occurrence of spontaneous visible mutants in TLS defective strains. a, Experimental set-up to determine spontaneous mutagenesis: the F1 brood of non-mutant segregating hermaphrodites (P0) were singled to establish individual populations. These were inspected for mendelian segregation of abnormal phenotypes indicating the occurrence of a recessive mutations in the gametes of the P0. Mutants affecting body morphology (e.g. dumpy/dpy) or movement (i.e. uncoordinated/unc) can be scored in the F2 progeny. Mutations in essential genes (i.e. lethal/let) give rise to islands of dead eggs when populations are allowed to clear the food supply. Elevated numbers of males in the F2 progeny indicate a high incidence of males (him) phenotype, arguing for a dominant him mutation in the F1. b, Quantification of visible mutant phenotypes. The data for *msh-6* mutants have been published previously.

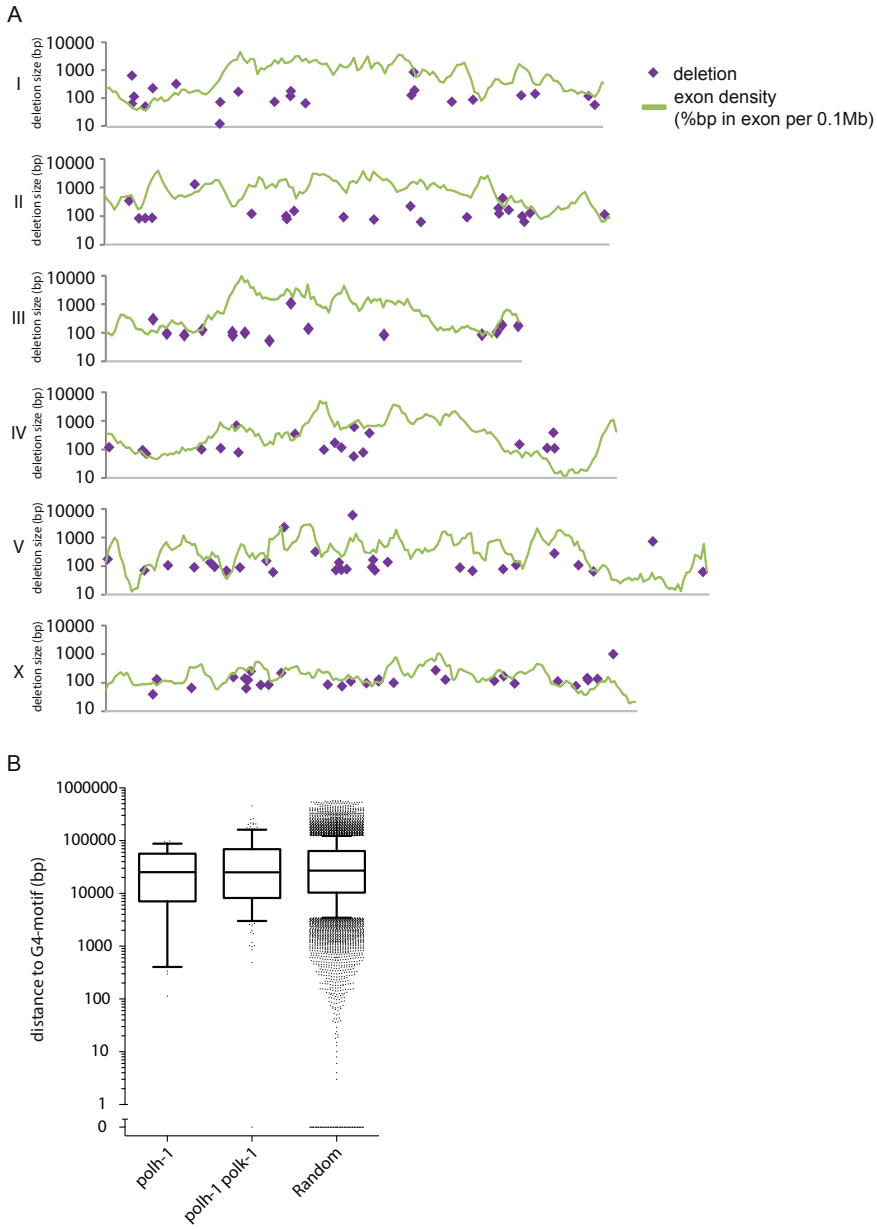


Figure S2. Genomic distribution of deletions in *polh-1polk-1* mutant animals. (A) Individual deletions (purple) were plotted onto a physical map of the C. elegans genome. The y-axis shows the size of the deletion on a logarithmic scale. The exon density is displayed in green (y-axis not shown). The length of the graph shows the size of the indicated chromosome relative to each other. (B) For each individual deletion the distance to the closest G4-motif G3-5N1-5G3-5N1-5G3-5N1-5G3-5 (1680 G4-motifs are present in the *C. elegans* genome) was determined. A random set of ~13,000 deletions with a size distribution similar to those observed in *polh-1 polk-1* mutants was plotted as a comparison. No statistical difference was found betweenthis random set and the set obtained from *polh-1* or *polh-1 polk-1* double mutant animals. Whiskers are drawn down to the 10th percentile and up to the 90th percentile. A distance of zero means that the nearest G4 motif is within the deletion.

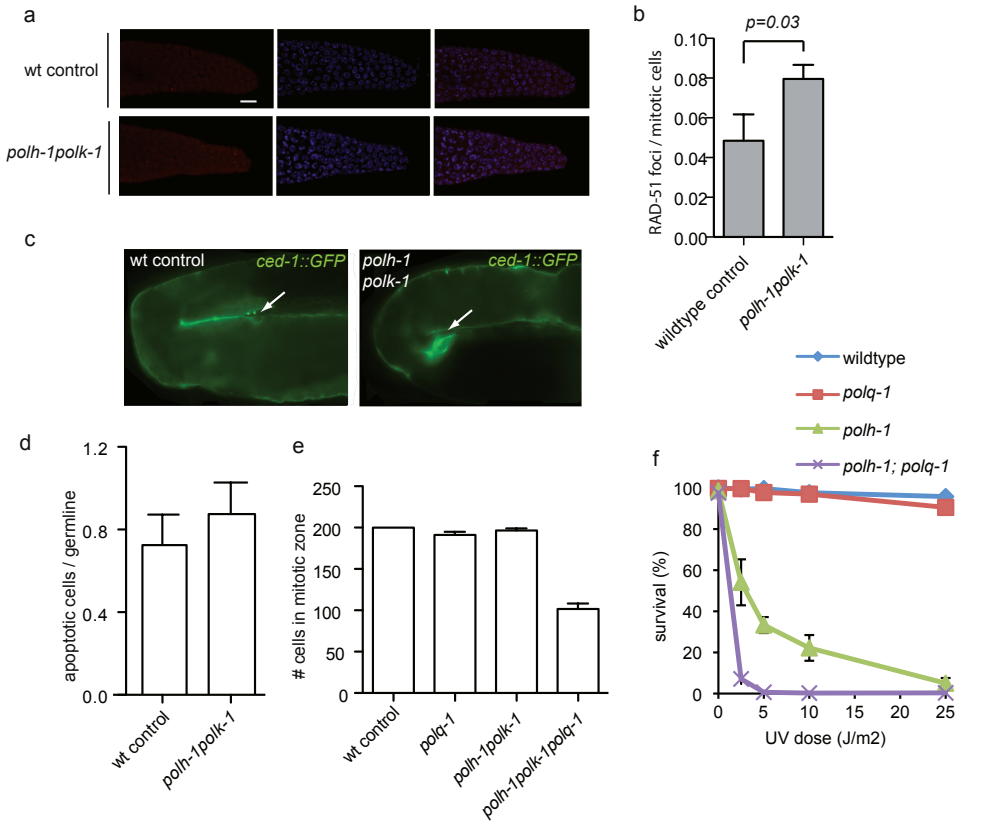


Figure S3. Analysis of DNA damage induction and apoptosis in single, double and triple mutants of *polh-1*, *polk-1* and *polq-1*. a. Representative images and b. quantification of RAD-51 foci for theindicated genotypes in nuclei present in the proliferative compartment of the *C. elegans* reproductivesystem. DAPI stainings in blue, RAD-51 in red. Scale bar, 10 µm c. Representative images of the bend ofthe gonad arm of animals transgenic for the apoptotic marker ced-1::GFP; cells in the process of apoptotic engulfment are indicated with arrows. Scale bar, 10 µm d. Quantification of apoptotic cells in *polh-1polk-1* mutant animals and wild-type controls. e. Quantification of the number of nuclei in themitotic region of the germline. A reduction in the number of cells in this region is an establishedoutcome of checkpoint activation. f. Sensitivity of *polh-1* and *polq-1* single and double mutants for exposure to UV, plotted as the fraction of surviving progeny after germline exposure of young adult worms.

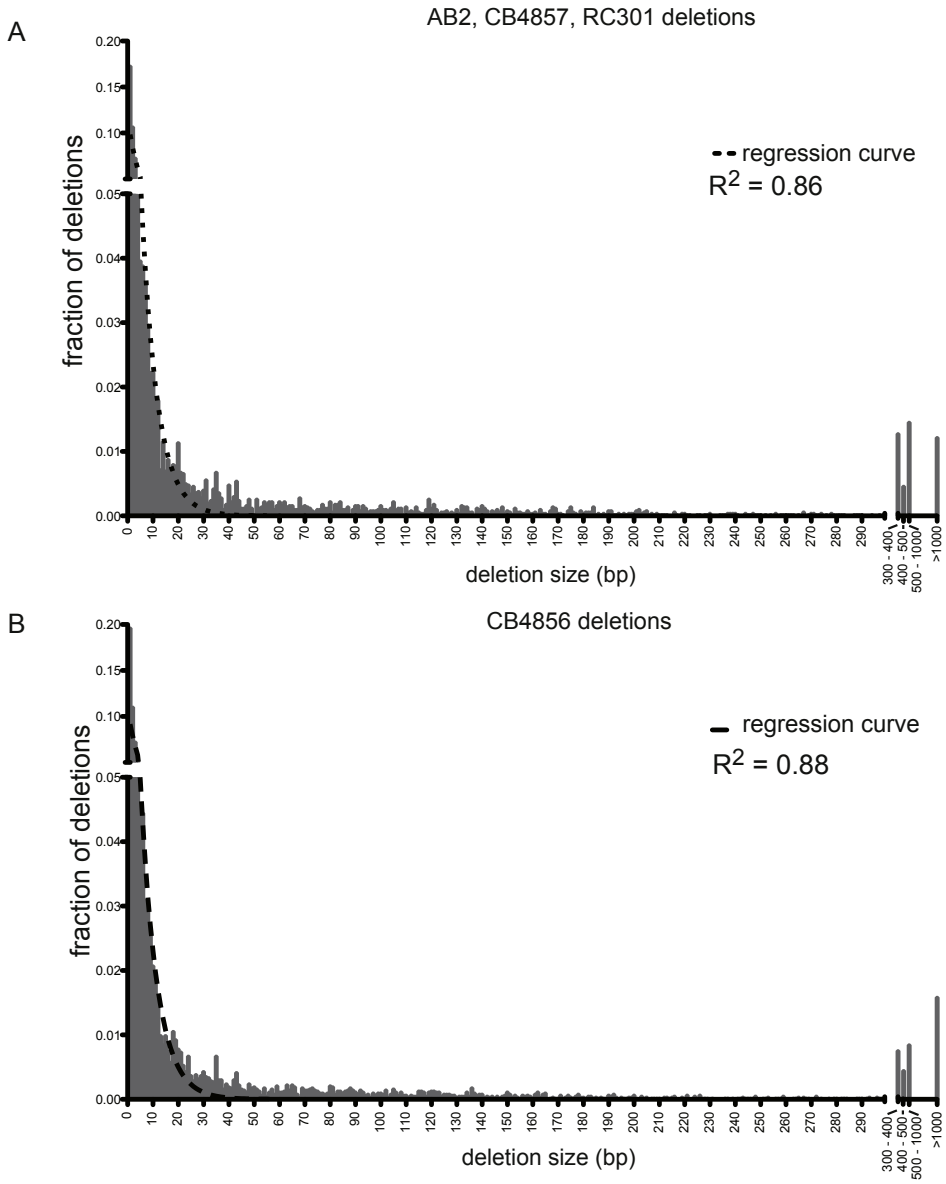
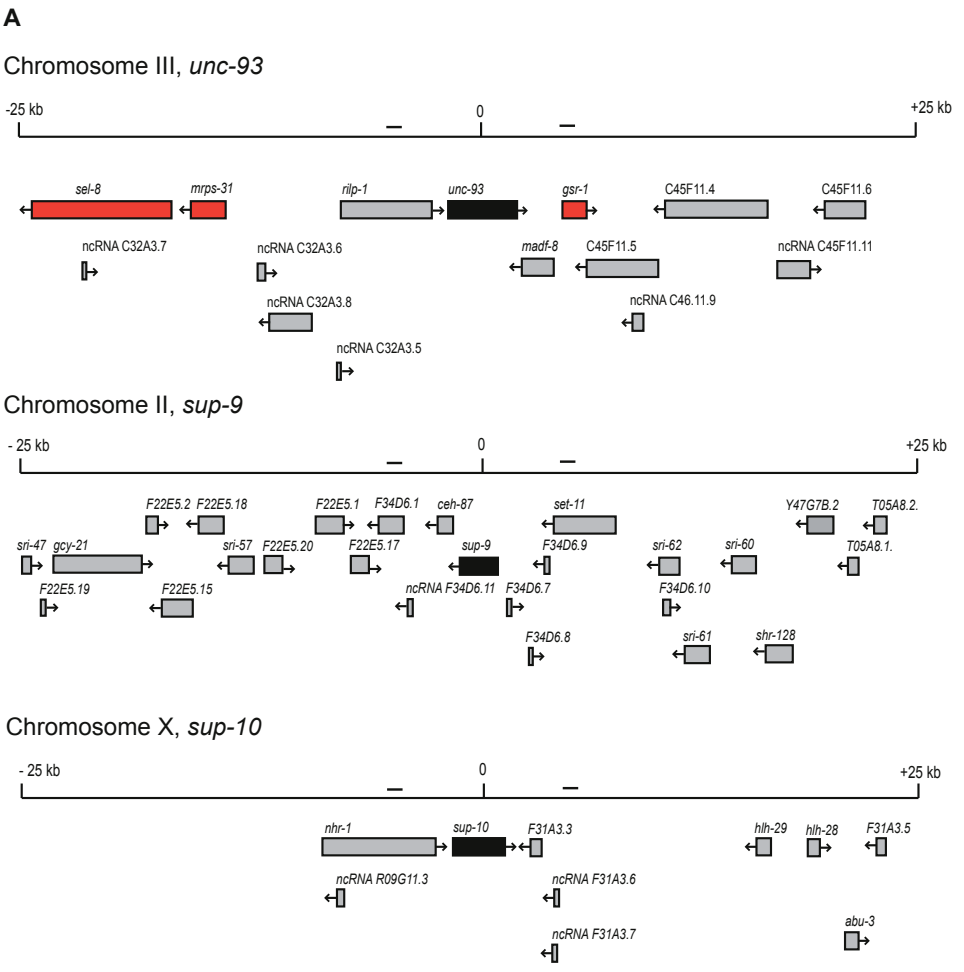


Figure S4. Histogram of size distributions is plotted of the various *C. elegans* natural isolates that were analyzed. Regression analysis showed that an exponential fit for deletion sizes up to 20bp approaches the actual distribution best. (A) the grouped distribution for AB2, CB4857 and RC301. (B) as in (A), but now for CB4856.

A



B

	unc-93	*sup-9*	*sup-10*
wildtype	0	2	11
polq-1	6	20	55

Figure S5. Selective occurence of large chromosomal deletions in regions that are devoid of essential genes in the *unc-93* mutagenesis assay. (A) Schematic representation of 50 kb regions surrounding the *unc-93*, *sup-9* and *sup-10* genes. Known essential genes are depicted in red. While *unc-93* is flanked by two essential genes, no essential genes are known in the 50 kb intervals around *sup-9* and *sup-10*. To estimate deletion sizes, amplification of PCR products at -5kb and +5kb positions has been tested. (B) Number of deletions larger than 5kb in *unc-93*, *sup-9* and *sup-10*.

Table S1. Whole genome sequencing statistics.

genotype	sample	# generations	# reads	average coverage	# bp >= 4x covered
N2	N2	60	45,258,326	28x	100,140,732
	N4	60	23,693,826	16x	99,675,920
polh-1(lf31)	H7	60	46,203,688	39x	100,229,062
	H8	60	44,982,616	37x	100,238,324
polk-1(lf29)	K1	60	41,517,548	21x	99,970,233
	K4	60	39,275,458	30x	100,235,635
	K9	60	40,037,564	24x	100,120,773
polh-1(lf31);polk-1(lf29)	D4	32	46,284,780	21x	99,911,564
	D13	25	38,712,292	29x	100,224,845
	D14	25	59,163,976	27x	100,202,641
msh-6 (pk2504)	M13	10	48,338,722	19x	99,236,278
	M15	10	44,129,942	12x	99,799,729

Table S2. unc-22 deletions in *polh-1polk-1* and *polh-1polk-1polq-1*.

	size	left flank	deletion left	deletion right	right flank	insertion
polh-1polk-1						
A	83 bp	GTACCTACTCA	CGTCCAAATG	TTATCGAAAA	GAACGTGTGC	–
B	74 bp	AATCCAGAAGT	CGATGACACC	CTTGGTTAGT	TATTTTTTGG	–
C	153 bp	ACAAGGCTGGG	CCTGGACAAC	TAAAGGCTGG	AGCCACTGTT	–
D	119 bp	GACTATCAAGG	CTGGTCAATC	TGATAACCCA	GAATACCAAT	AATCTGACTATCAAAGGAAATCTCAA-GAATCTGACTATCAAAG
E	93 bp	CTTGCAAAGGA	TCCATTTGGA	CACGTGACAA	CGGTGGATCA	–
F	71 bp	TGTGAAGCCTT	ACGGAACTGA	ACCACCAGTT	GTTACTTGGC	–
G		not identified				
polh-1polk-1 polq-1						
A	>4.7 kb					
B	>30.5 kb					
C	19 - 20.6 kb					
D	12660 bp	AAATGAGCACA	CTATTCTGTG	GAACAGGAGC	ATTTGGAGTT	
E	> 23.7 kb					
F		not identified				

Table S3. Frequency of *unc-22* mutations in *polq-1*, *polh-1polk-1* and *polh-1polk-1polq-1*.

Strain		total # plates scored	# plates containing one or more twitchers	estimated mutation rate
N2	wild-type	40	0	0.00E+00
XF152	*polq-1*	40	0	0.00E+00
XF507	*polh-1polk-1*	46	7	8.00E-06
XF840	*polh-1polq-1polk-1*	39	6	8.00E-06

Table S4. Sequence analysis of reversion mutants for *unc-93(e1500)*.

wild-type			
unc-93			
deletions > 5kb	0		
substitutions	6	cagttt(g>a)tctggc; C>Y	
		gacacg(t>a)cacagt; V>D	
		tgtctg(g>c)aatact; G>A	
		aaatat(c>t)gatttt; R>L	
		ggaatc(g>a)cggctt; T>A	
		tgttag(g>t)taatgg; splice	
other	1	gaatat(tcga>deleted)aaaactt	3bp > deletion > 12 bp
sup-9			
deletions > 5kb	2		
substitutions	1	ccattg(g>a)gactta; G>stop	
other	2	ccaata(gtga>deleted)cgtcat	3bp > deletion > 12 bp
		tctgta(ccgggtgggga>deleted)ggtctg	3bp > deletion > 12 bp
sup-10			
deletions > 5kb	11		
substitutions	3	cagttc(t>a)cttgta; L>H	
		tggaat(a>g)tggtcgg; M>V*	
		agccag(g>t)tttgta;; splice site mutation	
unknown	2		

*also tctttt(t>c)caacca in intron 150 bp upstream

Table S5. Sequence analysis of reversion mutants for *polq-1; unc-93(e1500)*.

polq-1			
unc-93			
deletions > 5kb	6		
substitutions	12	tgcgga(c>a)aagtcg; Q>K	
		cgttga(c>a)gattttc; T>K	
		gatctc(g>a)gatctg; G>R	
		ttccat(c>t)atttat; S>L	
		tttcta(c>a)ctcatg; T>N	
		tttcat(g>t)attgta; M>I	
		ggggag(c>a)caaatg; A>D	
		aagtcg(t>a)cggaaa; V>D	
		tccttt(c>t)gagaca; R>stop	
		tctata(c>a)attgtc; Y>stop	
		aatata(t>a)ttgctg; Y>stop	
		tgttag(g>a)taatgg; splice site mutation	
other	2	acgtca(ca>deleted)gttgaa	other
		ttttac(t>deleted)ttttag	microsatellite
sup-9			
deletions > 5kb	20		
substitutions	7	tcttcg(g>a)gctcac; G>E	
		gggtac(c>a)agtgga; Q>K	
		gtggag(c>a)atttta; A>E	
		ccattg(g>a)gactta; G>stop	
		aggcta(c>a)ggtcat; Y>stop	
		tccctg(c>t)aaactc; Q>stop	
		caagta(c>a)aacatg; Y>stop	
sup-10			
deletions > 5kb	55		
substitutions	1	atgtta(a>t)tataag; N>I	
other	1	gtgatg(a>deleted)catcaa	hairpin
unknown	7		

REFERENCES

Avkin S, Goldsmith M, Velasco-Miguel S, Geacintov N, Friedberg EC, Livneh Z. 2004. Quantitative analysis of translesion DNA synthesis across a benzo[a]pyrene-guanine adduct in mammalian cells: the role of DNA polymerase kappa. *J Biol Chem 279*: 53298–53305.

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

Carvalho CMB, Pehlivan D, Ramocki MB, Fang P, Alleva B, Franco LM, Belmont JW, Hastings PJ, Lupski JR. 2013. Replicative mechanisms for CNV formation are error prone. *Nat Genet*.

Chen J-M, Cooper DN, Férec C, Kehrer-Sawatzki H, Patrinos GP. 2010. Genomic rearrangements in inherited disease and cancer. *Semin Cancer Biol 20*: 222–233.

Costantino L, Sotiriou SK, Rantala JK, Magin S, Mladenov E, Helleday T, Haber JE, Iliakis G, Kallioniemi OP, Halazonetis TD. 2014. Break-induced replication repair of damaged forks induces genomic duplications in human cells. *Science 343*: 88–91.

De Stasio E, Lephoto C, Azuma L, Holst C, Stanislaus D, Uttam J. 1997. Characterization of revertants of unc-93(e1500) in Caenorhabditis elegans induced by N-ethyl-N-nitrosourea. *Genetics 147*: 597–608.

Fischhaber PL, Gerlach VL, Feaver WJ, Hatahet Z, Wallace SS, Friedberg EC. 2002. Human DNA polymerase kappa bypasses and extends beyond thymine glycols during translesion synthesis in vitro, preferentially incorporating correct nucleotides. *J Biol Chem 277*: 37604–37611.

Gartner A, Milstein S, Ahmed S, Hodgkin J, Hengartner MO. 2000. A conserved checkpoint pathway mediates DNA damage--induced apoptosis and cell cycle arrest in C. elegans. *Mol Cell 5*: 435–443.

Greenwald IS, Horvitz HR. 2003. unc-93(e1500): A behavioral mutant of Caenorhabditis elegans that defines a gene with a wild-type null phenotype. *Genetics 96*: 147–164.

Grishkevich V, Ben-Elazar S, Hashimshony T, Schott DH, Hunter CP, Yanai I. 2012. A genomic bias for genotype-environment interactions in C. elegans. *Mol Syst Biol 8*: 587.

Haracska L, Yu SL, Johnson RE, Prakash L, Prakash S. 2000. Efficient and accurate replication in the presence of 7,8-dihydro-8-oxoguanine by DNA polymerase eta. *Nat Genet 25*: 458–461.

Hastings PJ, Ira G, Lupski JR. 2009. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet 5*: e1000327.

Herman RK, Kari CK, Hartman PS. 1982. Dominant X-chromosome nondisjunction mutants of Caenorhabditis elegans. *Genetics 102*: 379–400.

Honma M, Sakuraba M, Koizumi T, Takashima Y, Sakamoto H, Hayashi M. 2007. Non-homologous end-joining for repairing I-SceI-induced DNA double strand breaks in human cells. *DNA Repair 6*: 781–788.

Johnson RE, Prakash S, Prakash L. 1999. Efficient bypass of a thymine-thymine dimer by yeast DNA polymerase, Poleta. *Science 283*: 1001–1004.

Johnson RE, Yu S-L, Prakash S, Prakash L. 2007. A role for yeast and human translesion synthesis DNA polymerases in promoting replication through 3-methyl adenine. *Mol Cell Biol 27*: 7198–7205.

Kloosterman WP, Tavakoli-Yaraki M, van Roosmalen MJ, van Binsbergen E, Renkens I, Duran K, Ballarati L, Vergult S, Giardino D, Hansson K, et al. 2012. Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. *CellReports 1*: 648–655.

Knobel PA, Marti TM. 2011. Translesion DNA synthesis in the context of cancer research. *Cancer Cell Int 11*: 39.

Koole W, van Schendel R, Karambelas AE, van Heteren JT, Okihara KL, Tijsterman M. 2014. A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. *Nat Commun 5*.

Kusumoto R, Masutani C, Iwai S, Hanaoka F. 2002. Translesion synthesis by human DNA polymerase eta across thymine glycol lesions. *Biochemistry 41*: 6090–6099.

Lee JA, Carvalho CMB, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell 131*: 1235–1247.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics 25*: 2078–2079.

Lynch M. 2008. The cellular, developmental and population-genetic determinants of mutation-rate evolution. *Genetics 180*: 933–943.

Masutani C, Araki M, Yamada A, Kusumoto R, Nogimori T, Maekawa T, Iwai S, Hanaoka F. 1999a. Xeroderma pigmentosum variant (XP-V) correcting protein from HeLa cells has a thymine dimer bypass DNA polymerase activity. *EMBO J 18*: 3491–3501.

Masutani C, Kusumoto R, Yamada A, Dohmae N, Yokoi M, Yuasa M, Araki M, Iwai S, Takio K, Hanaoka F. 1999b. The XPV (xeroderma pigmentosum variant) gene encodes human DNA polymerase

eta. *Nature 399*: 700–704.
Matsuda T, Bebenek K, Masutani C, Hanaoka F, Kunkel TA. 2000. Low fidelity DNA synthesis by human DNA polymerase-eta. *Nature 404*: 1011–1013.
Meier B, Barber LJ, Liu Y, Shtessel L, Boulton SJ, Gartner A, Ahmed S. 2009. The MRT-1 nuclease is required for DNA crosslink repair and telomerase activity in vivo in Caenorhabditis elegans. *EMBO J 28*: 3549–3563.
Mitelman F, Johansson B, Mertens F. 2007. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer 7*: 233–245.
Muzzini DM, Plevani P, Boulton SJ, Cassata G, Marini F. 2008. Caenorhabditis elegans POLQ-1 and HEL-308 function in two distinct DNA interstrand cross-link repair pathways. *DNA Repair 7*: 941–950.
Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, et al. 2012. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell 149*: 979–993.
Ohashi E, Bebenek K, Matsuda T, Feaver WJ, Gerlach VL, Friedberg EC, Ohmori H, Kunkel TA. 2000. Fidelity and processivity of DNA synthesis by DNA polymerase kappa, the product of the human DINB1 gene. *J Biol Chem 275*: 39678–39684.
Roerink SF, Koole W, Stapel LC, Romeijn RJ, Tijsterman M. 2012. A Broad Requirement for TLS Polymerases η and κ, and Interacting Sumoylation and Nuclear Pore Proteins, in Lesion Bypass during C. elegans Embryogenesis. *PLoS Genet 8*: e1002800.
Shima N, Munroe RJ, Schimenti JC. 2004. The mouse genomic instability mutation chaos1 is an allele of Polq that exhibits genetic interaction with Atm. *Mol Cell Biol 24*: 10381–10389.
Tijsterman M, Pothof J, Plasterk RHA. 2002. Frequent germline mutations and somatic repeat instability in DNA mismatch-repair-deficient Caenorhabditis elegans. *Genetics 161*: 651–660.
Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics 25*: 2865–2871.
Yousefzadeh MJ, Wood RD. 2012. DNA polymerase POLQ and cellular defense against DNA damage. *DNA Repair*.
Zhang Y, Wu X, Yuan F, Xie Z, Wang Z. 2001. Highly frequent frameshift DNA synthesis by human DNA polymerase mu. *Mol Cell Biol 21*: 7995–8006.

4

POLYMERASE θ IS A KEY DRIVER OF GENOME EVOLUTION AND OF CRISPR/CAS9-MEDIATED MUTAGENESIS

Robin van Schendel, Sophie Roerink, Vincent Portegijs, Sander van den Heuvel and Marcel Tijsterman

Department of Human Genetics, Leiden University Medical Center, The Netherlands

Published in Genome Research 2015 June: 2015: 7394

Abstract

Cells are protected from toxic DNA double-strand breaks by a number of DNA repair mechanisms, including some that are intrinsically error-prone, thus resulting in mutations. To what extent these mechanisms contribute to evolutionary diversification remains unknown. Here, we demonstrate that the A-family polymerase theta (POLQ) is a major driver of inheritable genomic alterations in *C. elegans*. Unlike somatic cells, which employ non-homologous end joining (NHEJ) to repair DNA transposon-induced DSBs, germ cells use polymerase theta-mediated end joining, a conceptually simple repair mechanism requiring only one nucleotide as a template for repair. Also CRISPR/Cas9-induced genomic changes are exclusively generated through polymerase theta-mediated end joining, refuting a previously assumed requirement for NHEJ in their formation. Finally, through whole genome sequencing of propagated populations, we show that only POLQ proficient animals accumulate genomic scars that are abundantly present in genomes of wild *C. elegans*, pointing towards POLQ as a major driver of genome diversification.

Introduction

Identifying the mechanisms that drive heritable genome alterations is important for our understanding of carcinogenesis, inborn disease and evolution. Several repair mechanisms exist to avoid the potentially detrimental effects of DNA breaks: homologous recombination (HR) repairs DSBs in an error-free manner, but only when an undamaged template is available; non-homologous end joining (NHEJ) joins the ends of a DNA break without the use of a repair template, frequently resulting in sequence alterations[1]. In addition to these two well-established repair modes, other genetically less-defined mechanisms operate, mostly under circumstances that are more rare and incompletely understood. An alternative end joining (alt-EJ) pathway was described which generally manifests only when NHEJ is compromised[2-4]. The A-family Polymerase theta (POLQ) was recently identified to play a major role in alt-EJ of DSBs in *Drosophila*, *C. elegans*, mice and humans[5-10]. Several other functions have been suggested for POLQ, besides operating in alt-EJ, which includes bypassing DNA lesions[11-13] and influencing the timing of DNA replication origin firing[13, 14]. Mice lacking functional POLQ show a very mild enhanced chromosome instability phenotype, which is exacerbated in combination with a deficiency in ATM, a kinase involved in the repair of DSBs[13, 15]. The recent discovery that HR-deficient tumours are dependent on repair by POLQ also argues that HR and alt-EJ can act on similar substrates, and importantly identifies POLQ as a druggable candidate target for cancer therapy[5]. The physiologically relevant contexts for when alt-EJ is the repair route of choice are, however, largely unknown. Recent work in *C. elegans* suggested that POLQ is important in repairing replication-associated DSBs in cells that fail to bypass endogenous DNA lesions DSBs[9], or unwind thermodynamically stable DNA structures[6]. Other observations point to the predominance of alt-EJ in germ cells: de novo genome deletions and chromotripsis-like chromosome rearrangements underlying congenital disease are frequently characterized by microhomology at their junctions[16], a feature that has thus far been characteristic for alt-EJ[17]. Such a scenario would also be compatible with the observed lack of expression of key NHEJ proteins during specific (DSB-repair proficient) stages of gametogenesis in vertebrates[18, 19]. To identify the contribution of DSB repair pathways to inheritable genome change, we studied error-prone repair of DSBs in germ cells of *C. elegans,* and surprisingly found this to be entirely dependent on POLQ-mediated alternative end joining. Moreover, we found POLQ-1 action to be solely responsible for the vast majority of insertion/deletions that occur during natural evolution of *C. elegans*.

Results

Transposon breaks are repaired by POLQ-mediated end joining

In *C. elegans* DNA transposons of the Mariner family are a natural source of genome change: upon hopping into a new location, transposons leave behind a DSB that in somatic cells is repaired by NHEJ[20], but in germ cells is either repaired error-free by HR[21] or error-prone by an EJ mechanism that is currently unknown[20, 22]. We first inspected the genomes of 45 sequenced natural isolates of *C. elegans*[23, 24] for genomic scars associated with DNA transposition. Although we found 93 unique transposon insertions in 23 isolates, too few deletions were identified at known transposon sites (<10) for a systematic analysis of deletion junctions (see Supplementary Fig. 1, Supplementary Data 1-2). The high insert versus deletion ratio is in line with previous data arguing that transposon-induced DSBs are predominantly repaired in an error-free manner[21]. To study error-prone repair

we next stimulated DNA transposition under laboratory conditions (by genetically inactivating transposon silencing[25]) and phenotypically monitored DSB repair in germ cells. To this end, animals were used that carry a frame-disrupting Tc1 element in the endogenous *unc-22* gene, which makes them move uncoordinatedly. Tc1 excision followed by imprecise repair of the resulting break can lead to ORF restoration, and the frequency of wild type-moving animals in populations of uncoordinated animals thus reflects the frequency of error-prone repair of transposon-induced DSBs in germ cells (Fig. 1a-b). In line with previous findings[22], we found that NHEJ deficiency did not affect the frequency (2.6E-4 and 2.3E-4, for wild type and *lig-4* mutant animals, respectively) or pattern of Tc1-induced genomic alterations: in both genetic backgrounds the spectrum is highly variant, showing 26 distinct deletion products in 103 isolated wild type animals and 16 distinct footprints in 36 isolated *lig-4* mutant animals (Fig. 1c, Supplementary Data 3). We next found that deficiencies in genes in other DSB repair pathways i.e. homologous recombination (*brc-1*, the worm homolog of mammalian breast cancer gene BRCA1) or single strand annealing (*xpf-1*/*ercc-1*) also did not affect the mutation spectrum of insertions/deletions (indels) at Tc1-induced breaks (Fig. 1c, Supplementary Fig. 2), nor did defects in mismatch repair or translesion synthesis (see Supplementary Fig. 2, Supplementary Data 4). However, in depth analysis of >100 deletion footprints derived from wild type populations provided a strong clue about the identity of the repair process that is responsible for their generation: ~79% of all deletions that were simple (that lost only the Tc1 element and some flanking nucleotides, *n*=43) displayed single nucleotide homology, a feature that was recently attributed to the action of an alternative form of end joining that critically depends on the A-family polymerase theta (POLQ)[6, 9]. In addition, another described feature of polymerase theta-mediated end joining (TMEJ) stood out in this collection of repair products: 24% of all deletions contained, in addition to the loss of the Tc1 element and a few flanking nucleotides, DNA inserts of which the sequence was identical to sequences in close proximity to the DSB, so-called templated inserts[26, 27]. Indeed we found that inactivation of *polq-1*, the gene encoding POLQ, dramatically affected the outcome of transposon-induced DSB repair: a profound reduction (>20 fold) in the number of deletion products was observed and also the spectrum of the remaining products greatly changed (Fig. 1d). No templated inserts were found, and one class of footprints, which is devoid of single-nucleotide homology and may have been the result of blunt ligation of limitedly processed ends, dominated the spectrum (32 out of 39 repair products). We conclude from these data that TMEJ is responsible for >95% of error-prone repair of transposon-induced breaks in germ cells of *C. elegans*. Reconstructing how individual templated inserts came about (see Supplementary Fig. 3) allows us to construct a detailed mechanistic model for TMEJ on DSBs, in which minute base pairing interactions of two 3' ssDNA tails at either side of the break are sufficient to prime DNA synthesis by POLQ-1, leading to a DNA complementarity-driven stabilization of the broken ends.

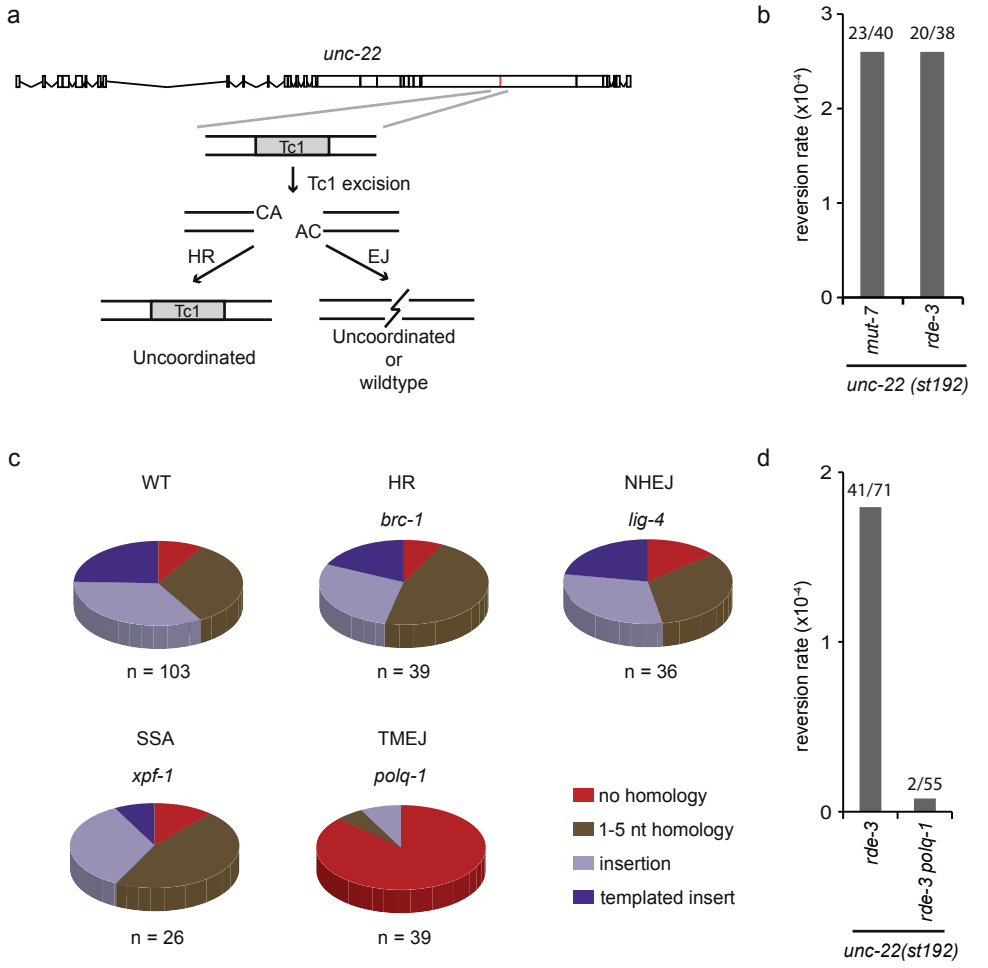


FIGURE 1. Error-prone repair of transposon-induced DSBs requires POLQ-1. A. Schematic representation of the experimental system to monitor repair of Tc1-induced DSBs. Tc1-encoded transposases can excise a frame-disrupting Tc1 element (*unc-22::st192*) from the endogenous *unc-22* gene, thus resulting in a DSB within the *unc-22* ORF with non-complementary 3' overhangs of two nucleotides. In case of repair through homologous recombination (HR), the original (Tc1-containing) sequence will be restored without affecting the phenotype of progeny cells. Error-prone end joining (EJ) can lead to *unc-22* ORF correction, which, when occurring in germ cells, will result in wild type-moving progeny born out of uncoordinatedly moving *unc-22* mutant animals. B. Reversion frequencies of Tc1 for two different genetic backgrounds (*rde-3* and *mut-7*) that de-repress transposon silencing[53, 54]. For each mutant background about 20 populations were scored for the presence of revertants and experiments were performed in duplicate. The total number of populations that were assayed and the number of populations that contained at least one revertant animal is indicated. Populations contained, on average, 2000 animals C. Distribution of footprints in *unc-22(st192)* for the indicated genomic backgrounds; all strains were also *rde-3* deficient. The number of independently derived reversion alleles is depicted underneath. Distinct footprints (26 in repair-proficient animals) were classified into 4 separate categories: i) simple deletions without homology at the deletion junction (red), ii) simple deletions with 1-5 bp of sequence homology at the deletion junction (brown), iii) deletions that also contained insertions (light blue), and iv) deletions with associated insertions that were identical to sequences immediate flanking the break (blue). D. Quantification of the *unc-22*(*st192*) reversion frequency in *rde-3* and *polq-1; rde-3* mutant backgrounds. The number of populations that were assayed and the number of populations that contained at least one revertant animal is indicated. Populations contained, on average, 2400 animals.

POLQ-mediated repair of CRISPR/Cas9-induced breaks

To further substantiate this finding and also to look at substrate specificity, we next studied DSB breaks that were brought about by the clustered, regularly interspersed, short palindromic repeats (CRISPR) RNA-guided Cas9 nuclease[28]. CRISPR/Cas9 technology is used to create mutants in a broad spectrum of biological systems, including worms, flies, fish, plants and mice[29-32]. The basic principle is to generate a DSB by introducing a guide RNA, which forms a RNA:DNA duplex at a target site, which is then recognized and cut by Cas9. It has been suggested that CRISPR/Cas9-induced breaks are repaired by NHEJ in these systems. However, we here show that CRISPR/Cas9-mediated germline transformation in *C. elegans* is entirely mediated by TMEJ, and not by NHEJ. We created mutant animals by microinjecting CRISPR plasmids targeting three sites at two distinct loci into the gonadal syncytium of hermaphroditic *C. elegans (*Fig. 2a). Deletion alleles were generated with ~10% efficiency per progeny that has been successfully transformed (Fig. 2b-c, Supplementary Table 2). Most of the obtained alleles had a small deletion, with a median size of approximately 13 bp for each target (Fig. 2d, Supplementary Data 5). This outcome is in agreement with all currently available worm data on CRISPR alleles, arguing little effect of the target's sequence context or genomic environment on the outcome of repair. We found that inactivation of NHEJ, by disrupting either *lig-4* or *cku-80* (*C. elegans* Ku80) (Fig. 2d, Supplementary Fig. 4), did not change the frequency or the type of genomic alterations, thus ruling out a role for canonical NHEJ in CRISPR/Cas9-mediated germ cell transformation. In contrast, the efficiency of successful CRISPR/Cas9 targeting dropped at least 6 fold for all targets in *polq-1*-deficient animals (Fig. 2c). Moreover, the mutants that were obtained in this background had deletions that were ~1000 fold larger, ~10-15 kb on average (Fig. 2d). We thus conclude that TMEJ is responsible for repair of blunt CRISPR/Cas9-induced DSBs in germ cells giving rise to inheritable alleles. Here, as in the processing of transposon-induced breaks, TMEJ action results in a typical signature: 7% of CRISPR/Cas9 breaks are characterized by templated inserts and 80% of simple junctions have single nucleotide homology (see Supplementary Fig. 5). Break-ends that are processed by POLQ also appear to be quite stable, as many deletions have their junction exactly at the position where the blunt-end DSB is made and have lost only few base pairs at one of either ends (see Supplementary Fig. 4). The demonstration that POLQ acts dominantly in end joining of CRISPR/Cas9-mediated DSBs raises the question whether it also acts to suppress HR-mediated homologous repair of CRISPR/Cas9 breaks. We found, however, with two different target-repair template combinations that homologous targeting is not more efficient in *polq-1* animals (see Supplementary Fig. 6).

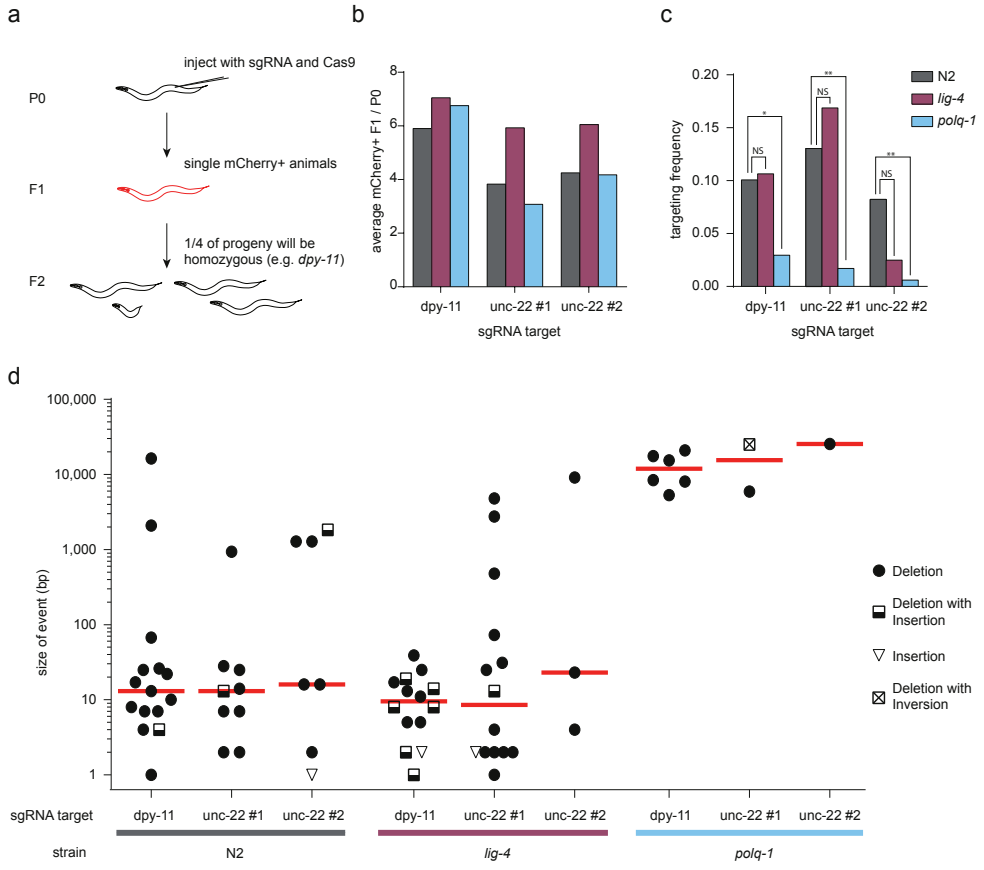


FIGURE 2. CRISPR/Cas9-induced mutations are generated through TMEJ. A. Schematic illustration of the strategy to generate mutants via CRISPR/Cas9 technology in *C. elegans*. Hermaphroditic animals (P0) are microinjected with plasmids that provide germline expression of Cas9 and of guide RNAs that target genes of interest (*dpy-11* and *unc-22*). A marker plasmid that results in somatic mCherry expression was co-injected. Only mCherry-positive progeny animals (F1) were clonally grown because these have, when compared to non-expressing progeny animals, a higher chance of carrying a (heterozygous) mutation in the targeted gene. Homozygous mutant animals will manifest in a Mendelian manner in the brood (F2) of transformed F1's because of hermaphroditism. B. A quantification of the efficiency of transgenesis in animals of different genotype. The average number of mCherry-expressing animals per injected P0 animal is indicated for each sgRNA target. More than 20 animals were injected per experimental condition. C. A quantification of the efficiency of CRISPR/Cas9-induced gene targeting per sgRNA target in animals of different genotype. The frequency is defined as the number of mutant alleles divided by the number of successfully transformed F1 progeny animals. A Fisher's exact test was used to determine statistical significance. (NS - non-significant, * $p < 0.05$, ** $p < 0.01$) D. A size representation of CRISPR/Cas9-induced mutants that were obtained in wild type, *lig-4* and *polq-1* mutant animals. Three different sgRNAs, targeting two genes were used. The median is indicated in red.

POLQ-mediated repair drives genome evolution

Our data reveal a critical role for POLQ in the repair of DSBs in germ cells of *C. elegans*, but does not address the question how relevant TMEJ is for genome change under unperturbed growth. What is the contribution of error-prone DSB repair to genome evolution? We previously found a TMEJ fingerprint in the genomes of *C. elegans* strains that were isolated from different parts of the globe, however, very little could be concluded as to the scale of the involvement, the source of the instability, or the possible presence of redundant pathways that may have similar outcomes[9]. Using two complementary approaches we now provide evidence that TMEJ plays a previously unrecognized major role in genome diversification. First, we sequenced two of the most diverged *C. elegans* strains known, and used these, together with recently sequenced natural isolates of *C. elegans*[23, 24], to reconstruct the nature of ~17,000 unique insertions/deletions (indels). Single nucleotide variants and indels at microsatellite repeats were excluded from the analysis, as these are likely the product of replication errors and not of error-prone DSB repair. We found the indels in the natural strains to be highly similar to those accumulating in the standard laboratory strain Bristol N2 when grown under laboratory conditions (Fig. 3a). Small deletions (< 500bp), which comprise the vast majority of the indels, had a very similar size distribution in all samples and were characterized by a high degree of single nucleotide homology at the deletion junctions. Particularly the latter feature is characteristic for TMEJ of DSBs[6, 9]. Then, to test whether POLQ is indeed required for the generation of spontaneous indels, we clonally grew wild type and *polq-1* mutant animals for over 50 generations and then sequenced their genomes (Fig. 3b, Supplementary Table 3). While the induction rate of single nucleotide variations (0.25 SNV per generation, see Supplementary Fig. 7, Supplementary Data 6) was identical in wild type and *polq-1* mutants, the induction rate for deletions was strikingly different: we detected small-sized deletions (median size of 7 bp) only in wild type animals. This class of mutations was completely absent in the genomes of *polq-1* animals (Fig. 3c, Supplementary Table 4-5). Instead, extensive deletions (median size of ~13,500 bp) were found, which *vice versa* were not detected in POLQ-proficient animals, suggesting that in the absence of POLQ, the substrates that would induce small deletions are processed differently, leading to massive deletions, which are easily lost from populations because of negative selection. Together, these data argue that the vast majority of indels that are accumulating during nematode evolution is the direct result of POLQ action.

Discussion

Our data show an unprecedented importance for alternative end joining, which depends on POLQ, in repairing DSBs in the germ cells of *C. elegans*. Previous work has led to the realisation that DSBs in *C. elegans* germ cells are either repaired in an error-free manner, through HR, or via an end-joining pathway that is different from classical NHEJ[21, 22, 33]. We here show that DSBs resulting from transposon mobilisation or through the action of the Cas9 endonuclease are repaired via POLQ-mediated end joining, a mechanism that uses single nucleotide homology and leads to small sized deletions (of about ~7-13 bp), occasionally accompanied by templated insertions. The reason why NHEJ does not act on these breaks is not known, but it is not because NHEJ is absent from germ cells: we previously demonstrated NHEJ activity on meiotic breaks in animals that were mutated in the worm ortholog of the end-resection factor CtiP[34]. Also, the Fanconi Anaemia pathway has been shown to restrict NHEJ activity in germ cells[35].

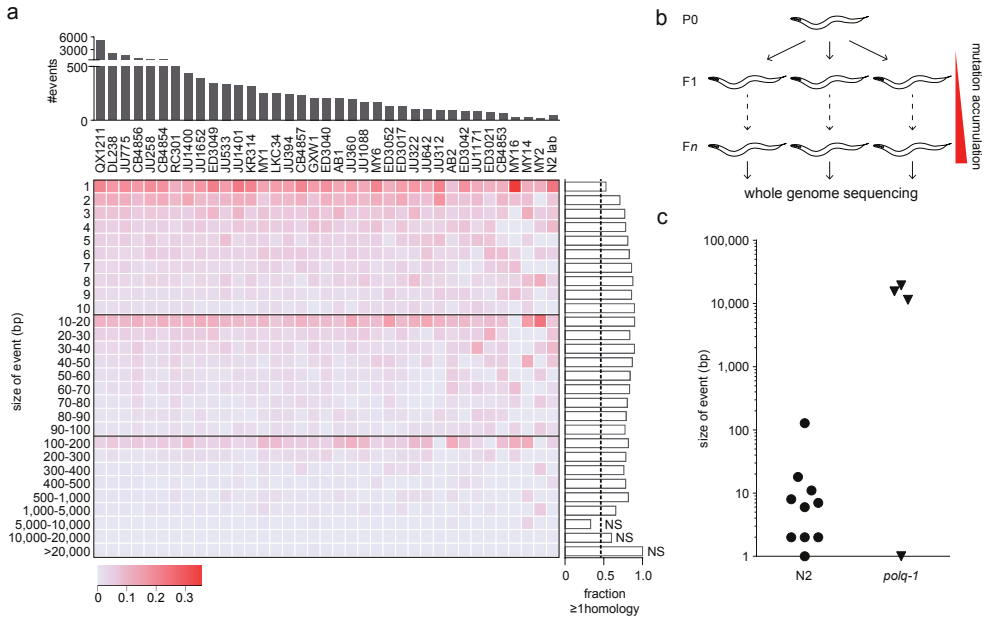


FIGURE 3. TMEJ is a driver of genomic diversification in *C. elegans*. A. A heat map representation of all genomic deletions events that were uniquely present in natural isolates of *C. elegans*, in which deletions are binned to size. The intensity of the colour reflects the percentage of deletions in each bin; the number of deletions for each strain is plotted above the heat map. The lane "N2 lab" represents deletions that accumulated in the Bristol N2 strains upon culturing in three different laboratories. For each size bin the fraction of microhomology ≥1 is plotted to the right of the heat map. The calculated ratio, as well as an empirically determined ratio, for the presence of microhomology ≥1 is 0.47 for a randomly distributed set of deletions in the *C. elegans* genome[9], which is represented by a dashed line. All size bins display a statistically elevated level of microhomology ($p < 0.001$, binomial test), except for deletions >5000, which were rare (n=19): NS indicates no statistically significant difference to the expected ratio of 0.47. B. Schematic illustration of the experimental setup reflecting small-scale evolution. Progeny animals (F1) from a single hermaphrodite (P0) are picked to separate plates to establish independent populations that were thus isogenic at the start of culturing. To establish bottlenecks and to carefully keep tract of the number of generations (n), a small number of progeny animals were transferred to new plates each generation. DNA was isolated from the progeny of a single animal (Fn) and sequenced by Next-Generation Sequencing technology with a base coverage of ~30 for each sample. C. A dot plot representing all unique deletion events that were found in the genomes of wild type (N2) and *polq-1* mutant animals.

An alternative explanation for the inability of NHEJ to process DSBs may be that (restricted) end-resection is very efficient in cycling germ cells – early embryonic cell cycles are devoid of recognisable G1 and G2 cell cycle stages – thus leading to 3' ssDNA overhangs onto which KU70/KU80 complexes do not nucleate a NHEJ reaction. The recent demonstration that POLQ can extend the 3' hydroxyl end of a 3' ssDNA tail when minimally paired with another DNA molecule with a 3' overhang supports the idea that transposon- or Cas9-induced breaks in germ cells are processed to have 3' overhanging ends[36]. In this scenario, POLQ-mediated end joining repairs DSBs that are processed to feed into HR, but which do not necessarily have an error-free template available, for instance because the break is introduced prior to DNA replication, or because both

sister chromatids sustain a break. This notion is supported by the recent demonstration that POLQ-mediated repair is very prominent in cases where replication-associated DSBs have unavailable sister chromatids[6], or in HR compromised genetic backgrounds[5, 27].

We found that POLQ functionality is causally involved in the generation of small indels that are abundantly present in the genomes of wild isolates of *C. elegans*. It argues that physiological DSBs in germ cells are repaired through TMEJ, generating inheritable genome alterations. At present, surprisingly little is known about which mechanisms shape the genome of an animal by generating the mutations onto which natural selection can act. Part of this lack of knowledge is because it is extremely difficult to prove experimentally, even for classes of mutations for which a very likely mechanism has been put forward, such as monotract expansions and contractions through polymerase slippage. Evidence for causality is ideally obtained by witnessing a reduction in mutagenesis upon inactivation of a candidate mechanism. The very low frequency of spontaneous mutagenesis in unperturbed conditions is complicating this issue even further. We mimicked evolution by growing animals for over 50 generation (under laboratory conditions) and then sequenced their entire genome to obtain sufficient data points to address questions concerning spontaneous mutagenesis. We surprisingly found that POLQ is causally involved in the generation of the vast majority of small indels in wild type animals. This class of indels are also abundantly present in the genomes of wild isolates of *C. elegans*, and our data thus strongly suggest that a mutagenic activity of POLQ is responsible for a major class of genome change during evolution. It is impossible to prove that these indels result from processing of physiological DSBs, however, we consider this very likely because the outcome of POLQ action on programmed DSB is grosso modo identical in nature to the indels that accumulate during evolution, with respect to size, use of single nucleotide homology and the occasional presence of templated inserts. In the absence of POLQ, the mutagenic outcomes are far worse, *i.e.* deletions are ~1000 fold larger in size. POLQ thus acts to protect cells but with a small price which manifest as small-sized genomic scars. Which DNA repair pathway is responsible for generating the sizable deletions manifesting in POLQ deficient genetic backgrounds will be the subject of further investigation – the deletion junctions are not characterized by extensive use of homology, which disfavours single strand annealing (SSA) acting as a redundant and mutagenic mechanism to process DSBs

Surprisingly, on an organismal level only mild phenotypes result from the absence of POLQ: mice develop normally and are fertile, with a slightly elevated level of genome instability and a subtle, but distinct, reduction in antibody diversification[5, 15]. Whether POLQ is also a natural driver of genome variation in human germ cells or (cancerous) somatic cells sustaining cell viability at the expense of mutation induction is yet unknown but the presence of microhomology and the occasional presence of template inserts at junctions of copy number variations, deletions and translocations as well as in junctions observed in chromotripsis[16, 37, 38] supports such a scenario. Therefore, inhibiting POLQ may, apart from sensitizing cells towards replication stress[9], restrict the adaptive response of oncogenically transformed cells and thus impair cancer maturation[13, 39].

Methods

C. elegans genetics

Nematodes were cultured on standard NGM plates at 20 degrees[40]. The following alleles were used in this study: *rde-3 (ne298); mut-7 (pk204); unc-22 (st192::Tc1); lig-4 (ok716); xpf-1 (e1487); ercc-1 (tm2073); brc-1 (tm1145); exo-1(tm1842); mlh-1(gl516); polh-1(lf31); polq-1 (tm2026); cku-*

80 (rb964).

Reversion assay to identify mutations by Tc1 transposition

Animals carrying *unc-22 (st192::Tc1)*, *rde-3(ne298)* or *mut-7(pk204)*, and wild type or mutant alleles of DNA repair genes were cultured, keeping track of the presence of the transposon in *unc-22* by selecting for worms that are Unc and by PCR analysis diagnostic for *unc-22::Tc1*. To assay error-prone repair of a DSB at the endogenous *unc-22* locus, single animals were transferred to 6 cm agar plates seeded with OP50 and propagated until starvation. Each experiment typically contained 30-50 plates per genotype. Plates were inspected for the presence or absence of non-Unc wild-type moving revertants. The reversion frequency is calculated by assuming a Poisson distribution for reversion[41]: Reversion frequency = $-\ln(P_0) / 2n$, where P_0 is the fraction of plates that did not yield revertants, and n is the number of animals that were screened per plate. From plates containing revertant animals, one non-Unc animal was transferred to a new plate and the molecular nature of the events that restored UNC-22 function were determined by PCR analysis and Sanger sequencing on DNA isolated from their brood.

CRISPR/Cas-9-induced mutations and homologous recombination

Plasmids were injected using standard C. elegans microinjection procedures. Briefly, one day prior to injection, L4 animals were transferred to new plates and cultured at 15 degrees. Gonads of young adults were injected with a solution containing: 20ng μl^{-1} pDD162 (Peft-3::Cas9) (Addgene 47549)[42], 20ng μl^{-1} pMB70 (u6::sgRNA with appropriate target (Supplementary Table 1)), 60ng μl^{-1} pBluescript, 10ng μl^{-1} pGH8, 2.5ng μl^{-1} pCFJ90, 5 ng μl^{-1} pCFJ104. Progeny animals that express mCherry were picked to new plates 3-4 days post injection. The progeny of these animals was inspected for Mendelian segregation of the corresponding phenotype. For gene targeting through homologous recombination the following injection mix was used: 30ng μl^{-1} Peft-3::Cas9 (Addgene 46168)[43], 100ng μl^{-1} pMB70 (u6::sgRNA with appropriate target for HDR #1 GFP or HDR #2 SNP), 30 ng μl^{-1} HDR template (pVP042 or pVP048), 10ng μl^{-1} pGH8, 2.5ng μl^{-1} pCFJ90, 5 ng μl^{-1} pCFJ104. PCRs with primers diagnostic for HR products at the endogenous locus were performed on F2 populations, where one primer resided in the repair template and the other just outside the homology arm (pVP042 GFP Fw: GAGAGAGGCGTGAAACACAAAG, Rv: TTTGGGAAGGTACGTCCGTC 1796 bp product or pVP048 Fw: GGCGCATGCACATAATCTTTCA, Rv: CCAGTGAGCTGCTCTTGAAGA 1610bp product). See Supplementary Data 5 and Supplementary Table 1-2 for more details.

Plasmid construction

pVP042 was generated to insert sequences encoding an N-terminal protein tag (FKBP-eGFP) into the endogenous *gpr-1* locus. DNA fragments were inserted into the pBSK vector using Gibson Assembly (New England Biolabs). Homologous arms of 1650 bp upstream and 1573 bp downstream of the *gpr-1* cleavage site were amplified from genomic DNA using KOD Polymerase (Novagen). Codon-optimized FKBP was synthesized (Integrated DNA technologies) and codon-optimized eGFP was amplified from pMA-eGFP (a kind gift of Anthony Hyman) and inserted directly 5' of the ATG of *gpr-1*. Five mismatches were introduced in the sgRNA target site to prevent cleavage of knockin alleles. pVP048 was generated to alter a single codon in the endogenous *lin-5* coding sequences. DNA fragments were inserted into the pBSK vector using Gibson Assembly (New England Biolabs). Homologous arms of 1568 bp upstream and 1557 bp downstream of

the *lin-5* cleavage site were amplified from cosmid C03G3 using KOD Polymerase (Novagen), a linker containing the altered cleavage site was synthesized (Integrated DNA technologies). Seven mismatches were introduced in the sgRNA target site to prevent cleavage of knockin alleles.

DAPI staining

L4 worms were picked and allowed to age 20-24 hrs. Gonad dissection was carried out in 1 x EBT (25mM HEPES-Cl pH 7.4, 118 mM NaCl, 48 mM KCl, 2mM CaCl, 2mM MgCL, 0.1% Tween 20 and 20 mM sodium azide). An equal volume of 4% formaldehyde in EBS was added (final concentration is 2% formaldehyde) and allowed to incubate for 5 min. The dissected worms were freeze-cracked in liquid nitrogen for 10 min, incubated in methanol at -20°C for 10 min, transferred to PBS/0.1% Tween (PBST), washed 3x10 min in PBS/1% Triton-X and stained 10 min in 0.5 µg ml^{-1} DAPI/PBST. Finally samples were de-stained in PBST for 1 h and mounted with Vectashield. Gonads were analysed using Leica DM6000 microscope.

Small-scale evolution and bioinformatic analysis

Mutation accumulation lines were generated by cloning out F1 animals from one hermaphrodite. Each generation, about three worms, were transferred to new plates. MA lines were maintained for 50-60 generations. Single animals were then cloned out and propagated to obtain full plates for DNA isolation. Worms were washed off with M9 and incubated for 2 h while shaking to remove bacteria from the intestines. Genomic DNA was isolated using a Blood and Tissue Culture Kit (Qiagen). DNA was sequenced on a Illumina HiSeq2000 machine according to manufacturers' protocol. Image analysis, base calling and error calibration was performed using standard Illumina software. Raw reads were mapped to the C. elegans reference genome (Wormbase release 235) by BWA[44]. SAMtools[45] was used for SNV and small indel calling, with BAQ calculation turned off. To identify larger indels and microsatellites, GATK[46] and Pindel[47] was used. In cases that only one of the software identified the structural variation, visual inspection was carried out using IGV[48]. Variations were marked as true if covered by both forward and reverse reads, and at least five times covered, while no reads were found that supported the reference genome while all other samples of the identical genotype supported the reference genome. For the analysis of natural isolates the same criteria were used, but the output was restricted to Pindel and only unique calls were included. In addition, deletions were only included when showing a >3-fold coverage drop of the deleted sequence, but normal coverage in at least 5 other natural isolates. All sequencing data, including the natural isolates DL238 and QX1211, have been submitted to the NCBI Sequence Read Archive (SRA) with accession ID (SRP046600). Two sequenced N2 strains can be found at accession ID (SRP020555). Genome sequences of other C. elegans natural isolates were obtained from[23, 24]; the genome sequence of PX174 is identical to RC301[49] and was excluded from the analysis. The genome of different cultures of N2 were derived from the National Institute of Genetics Japan (NCBI SRA: DRP001005), from the 50 Helminth Genome Initiative (submitted by the Sanger Center, NCBI SRA: ERX278110) and our own data (SRP020555, SRP046600).

Transposon Evolution

RetroSeq[50] was used to find genomic positions of transposons that are not present in the C. elegans reference genome (WB235). Retroseq discovery was run in align mode, using a transposon reference file containing all known Tc/mariner-like transposons. A custom script was written to identify those locations that showed hallmarks of a transposon insertion, which is duplication of a

flanking TA or TCA sequence, interrupted by a novel DNA sequence (indicative of an insertion). Once a position was identified in one natural isolate, all other natural isolates were analysed. Occasionally, RetroSeq was unable to identify the specific type of transposon. In those cases, more than one possible transposon was assigned to that location. To identify potential transposon deletions Pindel was used in which ≥ 8 supporting reads was set as a threshold and 0 reads should support the reference genome. The majority of the deletions were present in multiple natural isolates and were excluded from the analysis as these likely represent transposon insertions in the lineage that include the reference genome.

Phylogenetic Tree

The phylogenetic tree was created using high-quality SNV calls (SNV quality score ≥100) throughout all natural isolates with ≥5 reads (and more than 80% of the reads supporting the SNV) and supported by both forward and reverse reads. These criteria applied to the genomes of 44 natural isolates and N2 and resulted in 565,662 SNVs. PLINK[51] was used for pruning pairs with $r^2 > 0.3$ in a sliding 50-marker window at 5-marker steps and minor allele frequency SNPs were filtered out (< 0.05), leaving 22,487 informative SNPs. SNPhylo[52] was subsequently used to create the phylogenetic tree. Bootstrap analysis was performed 1,000 times to determine the reliability of each branch in the tree.

Acknowledgments

We thank the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440), for proving strains, Mike Boxem for plasmids, Jane van Heteren for comments to the manuscript, Harry Vrieling for discussions and Karin Brouwer for initial experiments on Tc1-induced break repair. MT is supported by grants from the European Research Council (203379, DSBrepair), the European Commission (DDResponse), and ZonMW/NGI-Horizon.

Author contributions

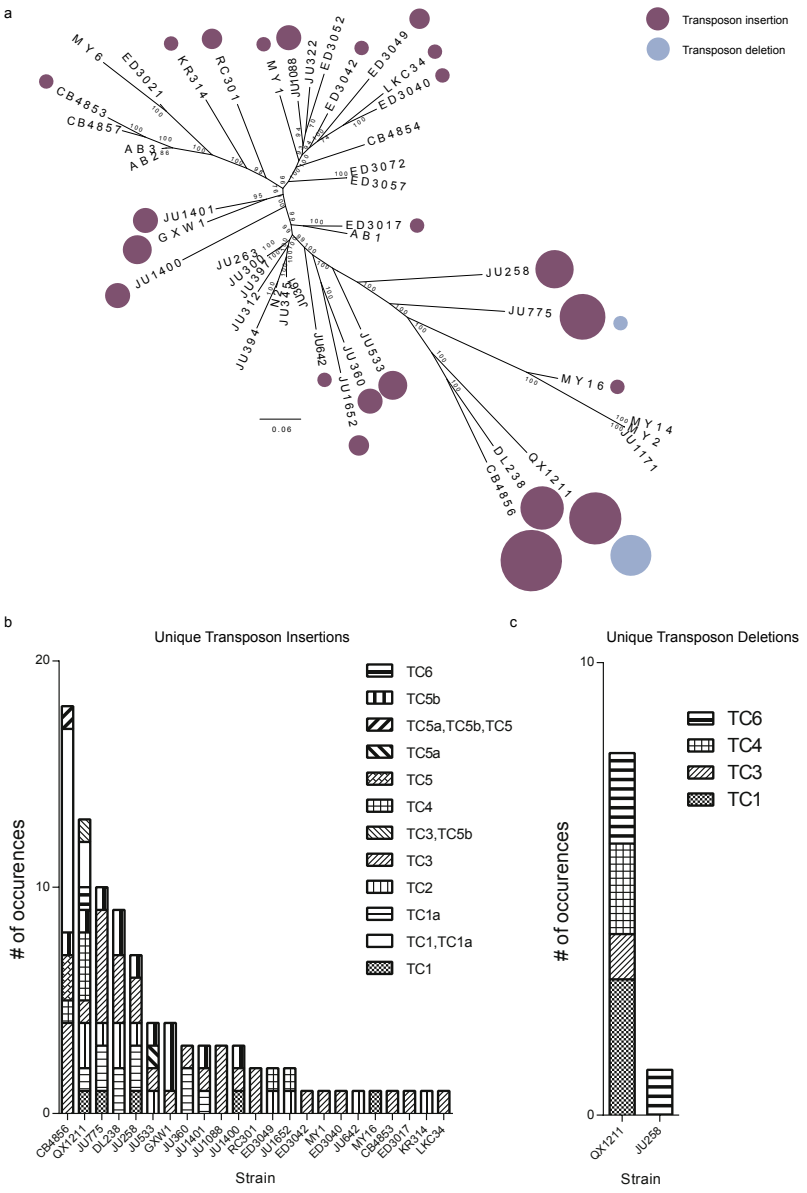
R.v.S. and M.T. conceived and designed the study. R.v.S. and S.R. performed the experiments. R.v.S. performed the bioinformatics analysis. V.P and S.v.d.H. generated reagents and advised on CRISPR/Cas9-related experimental procedures. All authors interpreted the experimental data. R.v.S. and M.T. wrote the manuscript.

Conflict of interest

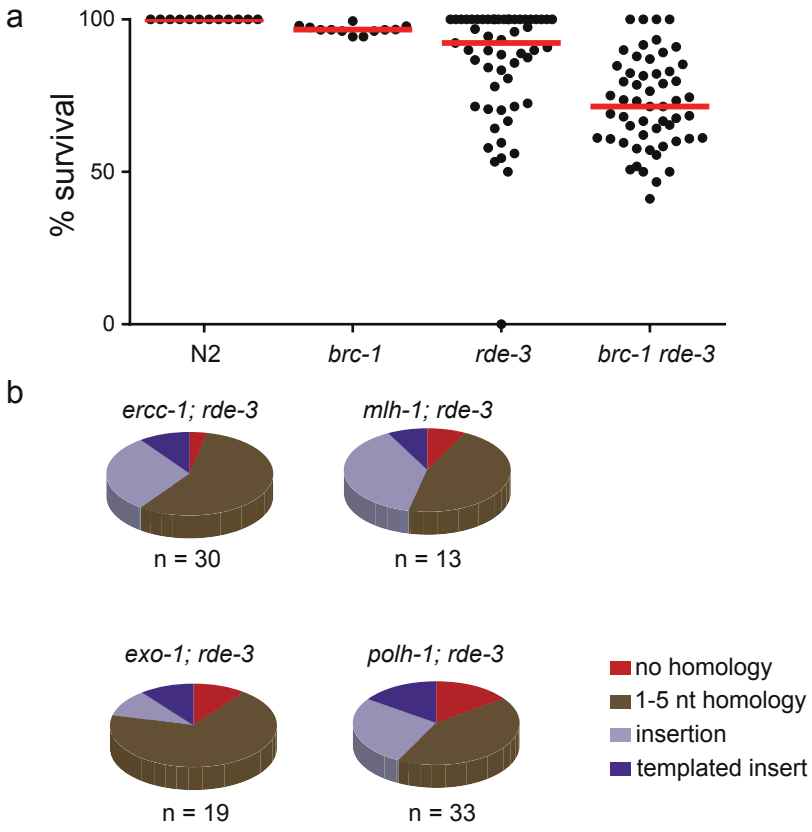
Authors declare no conflict of interest

Accession codes

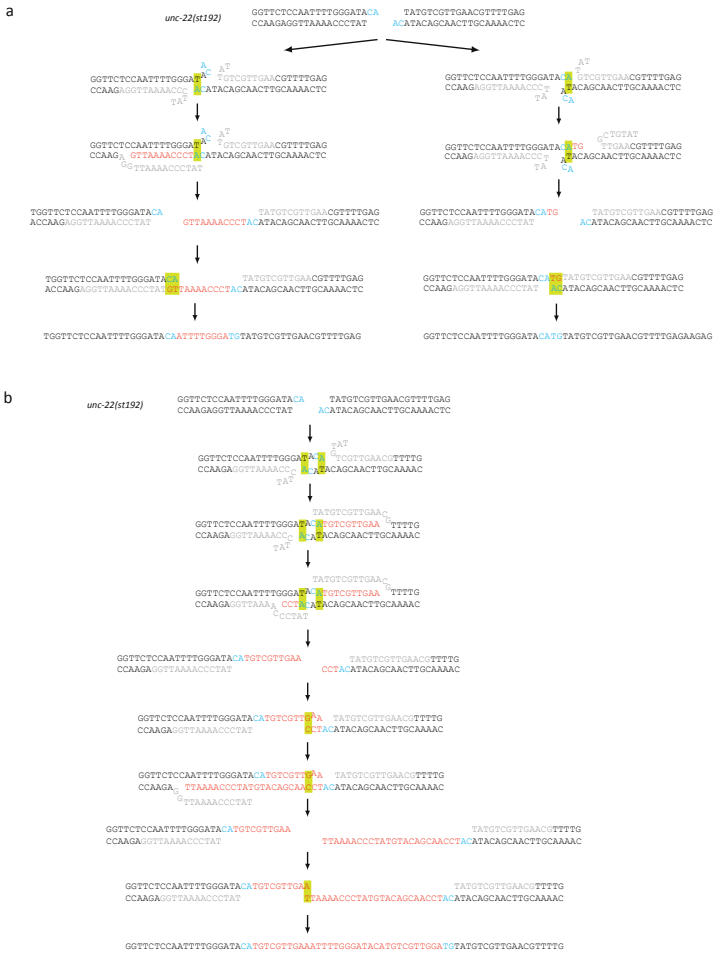
Raw sequences have been made publicly available at NCBI SRA (Accession code SRP046600).



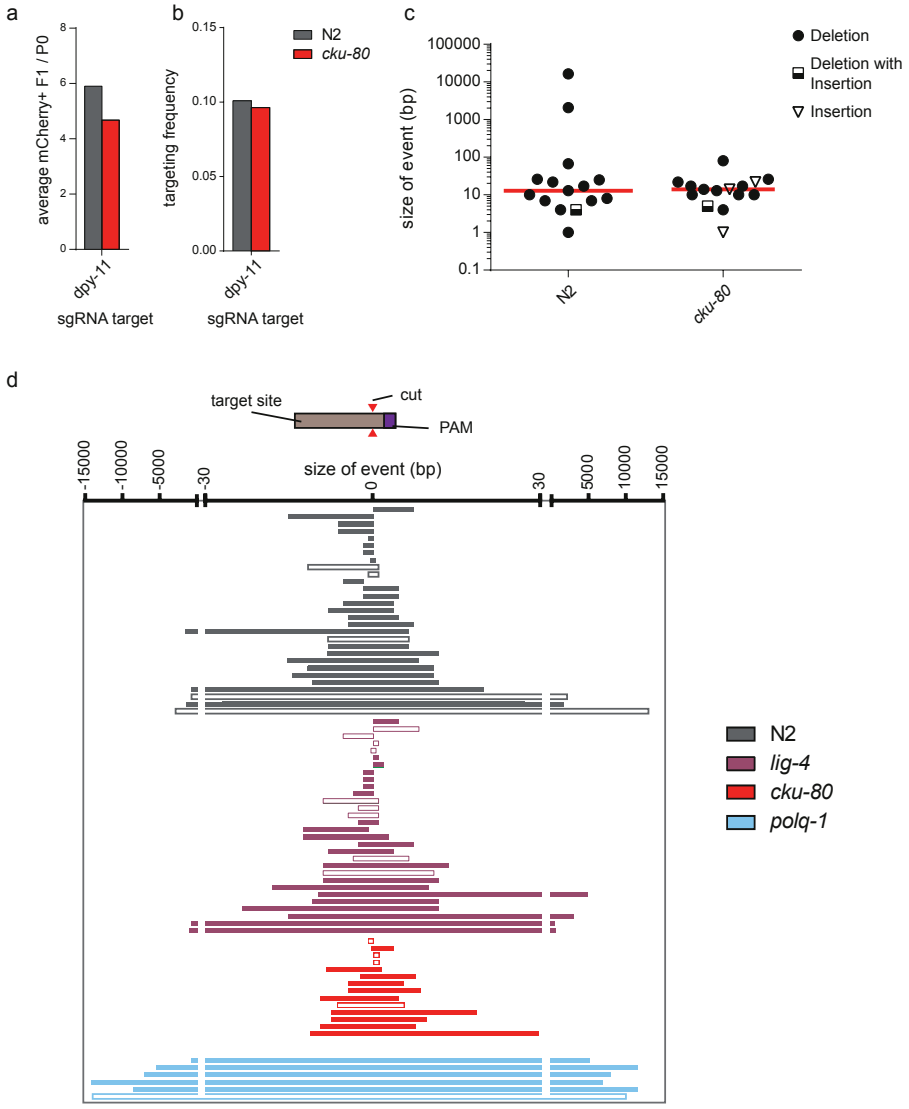
Supplementary Figure 1. DNA transposition in natural isolates of *C. elegans*. A. A phylogenetic tree was constructed from ~22,000 informative SNPs (See Material & Method section for details) present in 45 natural isolates. The outcomes of standard bootstrap analysis (1000 times) are plotted for each branch point. Transposon insertions were identified by RetroSeq, which was specifically designed to find such events in paired-end sequence data. The surface area of the plotted circle reflects the number of insertions (purple) and potential deletions (blue) that are unique to each strain: because the tree is unrooted, it cannot be concluded from this analysis whether the events that are marked as deletions (blue) in QX1211, one of the most diverged strains, are not in fact de novo insertions in a parent-of-origin that spawned all isolates after a split with QX1211. B. The number of unique insertions per type of transposon in each strain is plotted. C. The number of copies per type of transposon that was uniquely absent in each strain is plotted.



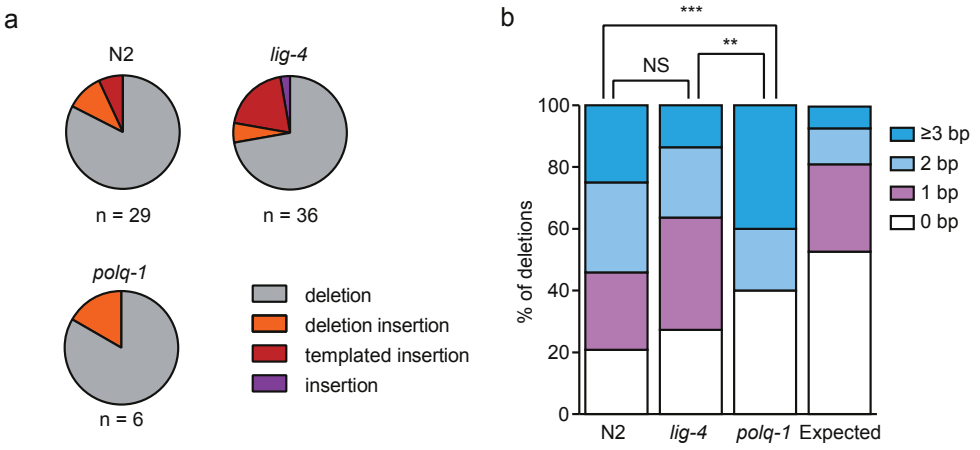
Supplementary Figure 2. Genetic analysis of error-prone repair of transposon-induced breaks in *C. elegans* germ cells. A. Transposon breaks-induced embryonic death. Bee swarm plot in which embryo survival is plotted for strains in which transposition is silenced (N2) or de-repressed in germ cells (*rde-3*) and are either proficient or deficient for the homologous recombination gene *brc-1*. Each dot represents the offspring of one animal; the percentage is calculated as the number of hatched larvae divided by the number of total eggs laid. For both N2 and *brc-1*-deficient animals the survival of at least 10 P0 animals was scored, while for the *rde-3*-deficient strains at least 50 P0 animals was scored. The red line represents the median survival for each strain. B. Distribution of footprints in *unc-22(st192)* for the indicated genomic backgrounds. The number of independently derived reversion alleles is depicted underneath. Distinct footprints were classified into 4 separate categories: i) simple deletions without homology at the deletion junction (red), ii) simple deletions with 1-5 bp of sequence homology at the deletion junction (brown), iii) deletions that also contained insertions (light blue), and iv) deletions with associated insertions that were identical to sequences immediate flanking the break (blue).



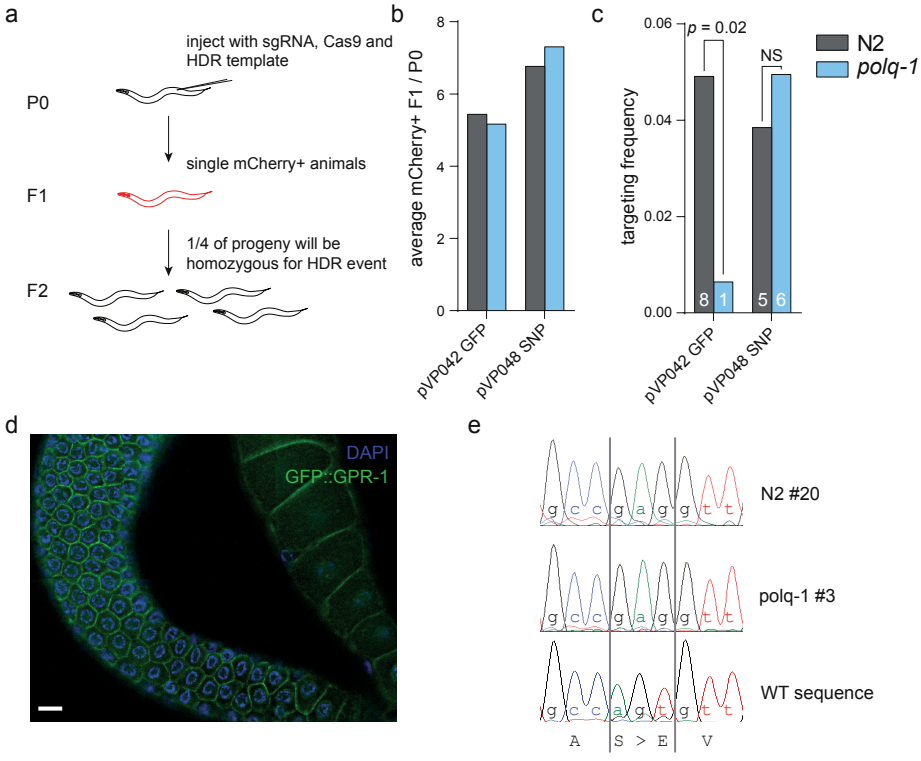
Supplementary Figure 3. Molecular model for TMEJ-generated templated inserts. A. Schematic illustration of the consecutive steps of TMEJ of a Tc1-induced DSB leading to the most commonly found templated inserts (12/103 for the outcome displayed on the left; 24/103 for the outcome displayed on the right). The sequence context of *unc-22(st192)* upon excision of Tc1 is displayed, with the 3'CA overhangs in blue. In the first round of the cycle the outermost 3' base (A) has served as a primer for POLQ action by base-pairing (boxed in yellow) to the first available T of the opposite flank; the left flank in the left panel and the right flank in the right panel. Newly synthesized DNA, through the action of POLQ, is displayed in red. Nucleotides that are either displaced by POLQ action or absent because of DSB processing prior to POLQ action are depicted in grey. The formation of the resulting intermediate, that is presumably energetically more stable because of the extended base-pairing of the newly synthesized DNA to its template, is apparently not always driving the process into the generation of simple deletions (without insertions, but with single nucleotide homology). Instead, for thus far unknown reasons, further extension is abrogated, and subsequently the outmost 3' base will search for a new match to re-anneal and again serve as a primer in a second attempt to join both ends. It is noteworthy that the most prominent templated inserts (left and right panel) are conceptually identical: in the first cycle DNA synthesis is continued up to the point where the two outermost 3' nucleotide of newly synthesized DNA can base-pair with the outermost 2 nucleotides of the template strand. B. Re-iteration of the steps displayed in A can explain even the most complex inserts. In the illustrated case, both flanks served as template for DNA synthesis; the left flank 3 times and the right flank 2 times, and all DNA synthesis events were primed with 1 nt base pairing.



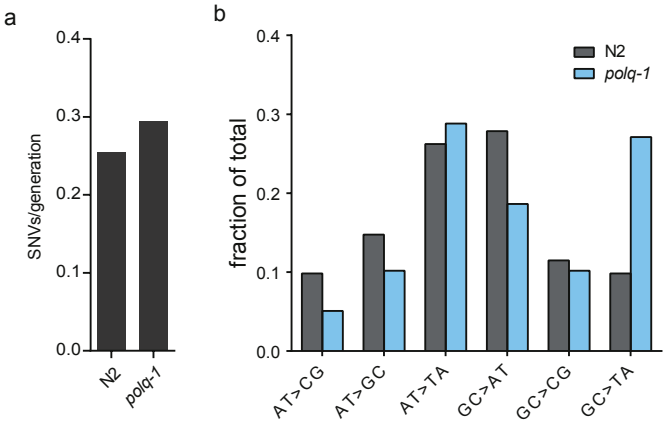
Supplementary Figure 4. Genetic and molecular analsyis of CRISPR/Cas9-induced genome rearrangements. A-C. Error-prone repair of CRISPR/Cas9-induced DSBs is independent of NHEJ protein CKU-80. A. A quantification of the efficiency of transgenesis in wild type (N2) and *cku-80*-deficient animals. The average number of mCherry-expressing animals per injected P0 animal is indicated. At least 20 animals were injected per strain. B. A quantification of the efficiency of CRISPR/Cas9-induced gene targeting of the *dpy-11* locus in wild type (N2) animals and *cku-80* deficient animals. The frequency is the number of mutant alleles divided by the number of successfully transformed F1 progeny animals. C. A size representation of CRISPR/Cas9-induced *dpy-11* mutants that were obtained in wild type and in *cku-80* mutant animals. The median is indicated in red. D. A visual representation of the CRISPR/Cas9-induced *dpy-11*, *unc-22* (target 1) and *unc-22* (target 2) alleles that were obtained in the strains of indicated genotype. 0 (bp) defines the cut-site of the sgRNA/Cas9 complex, and the orientation of the target and PAM site relative to 0 is depicted. Bars represent the DNA sequence that is lost in each allele. Closed bars represent simple deletions; open bars represent insertions, deletion with insertions and deletions with inversions.



Supplementary Figure 5. Types and homology distribution of CRISPR/Cas9 induced mutations. A. The distribution of mutational classes in strains of indicated genotype. B. Quantification of the extent of microhomology for the simple deletions obtained in strains of indicated genotype. The distribution that is expected if deletions were randomly distributed is also indicated. The distribution in wild type (N2) and *lig-4* mutant animals is statistically not significantly different (NS), however, both are different from *polq-1*. (** $p < 0.01$, *** $p < 0.001$, T-test)



Supplementary Figure 6. No increase in Homology Directed Repair (HDR) in animals defective for POLQ-1
A. A schematic illustration of the strategy to generate CRISPR/Cas9-induced alleles via HDR in *C. elegans*. Hermaphroditic animals (P0) are microinjected with a plasmid that provides germline expression of Cas9, a guide RNA that targets a gene of interest, and a plasmid that has a template for HDR. A marker plasmid that results in somatic mCherry expression is also co-injected. Only mCherry-positive progeny animals (F1) were clonally grown because these have, when compared to non-expressing progeny animals, a higher chance of carrying a (heterozygous) mutation in the targeted gene. Homozygous mutant animals will manifest in a Mendelian manner in the brood (F2) of transformed F1's because of hermaphroditism. B. A quantification of the efficiency of transgenesis in animals of different genotype. The average number of mCherry-expressing animals per injected P0 animal is indicated for each target locus. At least 20 animals were injected per target per strain. C. A quantification of the efficiency of CRISPR/Cas9-induced gene targeting in wild type and *polq-1* mutant animals for the indicated locus. The frequency is the number of mutant alleles divided by the number of successfully transformed F1 progeny animals. A Fisher's exact test was used to determine statistical significance. D. Representative image of successful HDR-mediated targeting of GFP to the endogenous *gpr-1* locus using CRISPR/Cas9. GPR-1::GFP (green) expression is visible in the cortex of germ cells in the distal (left) and proximal (right) area of the gonadal syncytium. DAPI staining in blue marks nuclei. Scale bar = 5 µm. E. Sanger sequences of two SNP alleles that were generated via CRISPR/Cas9-induced HDR. The wild type and mutant DNA as well as the amino acid sequence is indicated.



Supplementary Figure 7. Comparable SNV induction rates and distributions in wild type (N2) and *polq-1* mutant animals. A. Quantification of the SNV induction rate (SNVs per generation) for the indicated genetic background. The data were obtained from sequencing 4 times 60 generations of wild type growth and 4 times 50 generations of polq-1 growth. B. The base composition of the SNVs that accumulated in wild type and polq-1 mutant animals. Both distributions are comparable, apart from an elevated rate of GC>TA mutations in POLQ-1 deficient animals.

Supplementary Table 1. CRISPR targets

Target	Sequence	Chromosome	Start	End
dpy-11	GCAAGGATCTTCAAAAAGCATGG	V	6512819	6512842
unc-22 #1	GACTGCTTGCGGAGAGAGCAAGG	IV	11985320	11985343
unc-22 #2	GAAAAGCAAGATGCTGCCACTGG	IV	11985349	11985372

Supplementary Table 2. CRISPR injections

Strain	Allele	Target	Injected P0s	mCherry$^+$ F1	Targeting succes
N2		dpy-11	20	118	12
RB873	*lig-4 (ok716)*	dpy-11	20	141	16
XF152	*polq-1 (tm2026)*	dpy-11	37	250	8
N2		unc-22 #1	18	69	9
RB873	*lig-4 (ok716)*	unc-22 #1	14	83	14
XF152	*polq-1 (tm2026)*	unc-22 #1	38	117	2
N2		unc-22 #2	20	85	7
RB873	*lig-4 (ok716)*	unc-22 #2	20	121	3
XF152	*polq-1 (tm2026)*	unc-22 #2	40	167	1
N2		HDR #1 GFP	30	163	8
XF152	*polq-1 (tm2026)*	HDR #1 GFP	30	155	1
N2		HDR #2 SNP	23	168	6
XF152	*polq-1 (tm2026)*	HDR #2 SNP	17	115	5
RB964	*cku-80 (ok861)*	dpy-11	40	187	18

Supplementary Table 3. Whole genome sequence information

Sample_ID	Allele	#Generations	Average Coverage
N2_2	N2	60	28x
N2_3	N2	60	63x
N2_4	N2	60	16x
N2_50	N2	60	27x
XF151_E50	*polq-1 (tm2026)*	50	31x
XF151_G50	*polq-1 (tm2026)*	50	18x
XF151_H50	*polq-1 (tm2026)*	50	38x
XF151_I50	*polq-1 (tm2026)*	50	44x

Supplementary Table 4. CNVs

SampleGroup	Sample	Chr	Size	Start	End
N2	N2_3	III	128	1407845	1407973
N2	N-4	I	18	4336437	4336455
N2	N-2	X	11	16256644	16256655
N2	N-4	I	8	7667695	7667695
N2	N-2	II	7	10938339	10938346
N2	N-2	V	6	18324033	18324039
N2	N2_50	V	2	12400541	12400543
N2	N2_3	II	2	13447058	13447060
N2	N-4	II	2	11519211	11519213
N2	N-2	I	1	4524154	4524155
polq-1	XF152_I50	IV	19397	8038373	8057770
polq-1	XF152_E50	V	15588	15466490	15482078
polq-1	XF152_E50	V	11424	3111176	3122600
polq-1	XF152_E50	II	1	1517738	1517738

Supplementary Table 5. Microsatellite changes

Sample	Sample ID	Chr	Size	Start	End	Change	TractType	TractString
N2	N2_3	IV	3	1377786	1377789	DEL	MONO	gggGGGGGGGGG
N2	N2_3	X	3	2042489	2042492	DEL	MONO	gggGGGGGGGGGGGGG
N2	N2_3	X	2	3470709	3470711	DEL	MONO	ggGGGGGGGGGGG
N2	N2_50	II	1	240972	240972	SINS	MONO	aAAAAAAAAA
N2	N2_3	II	1	3409898	3409899	DEL	MONO	tTTTTTTTT
N2	N2_50	IV	1	3679599	3679600	DEL	MONO	gGGGGGGGGG
N2	N2_3	III	1	4430962	4430963	DEL	MONO	aAAAAAA
N2	N2_3	I	1	6149746	6149747	DEL	MONO	tTTTT
N2	N2_3	X	1	8972343	8972343	SINS	MONO	tTTTTTTTTT
N2	N2_50	II	1	10591087	10591088	DEL	MONO	tTTTTTTTTT
N2	N2_3	III	1	11300284	11300285	DEL	MONO	tTTTTTTT
N2	N2_50	I	1	12862421	12862422	DEL	MONO	tTTTTTTTTTTT
N2	N2_50	II	1	13626403	13626404	DEL	MONO	aAAAAA
N2	N2_3	X	1	17032785	17032785	SINS	MONO	aAAAAAAAA
N2	N-2	I	1	3050224	3050225	DEL	MONO	aAAAAAAAAA
N2	N-2	II	1	1372830	1372831	DEL	MONO	tTTTTTTT
N2	N-2	II	1	6800228	6800229	DEL	MONO	aAAAAAAAAA
N2	N-4	II	1	13626403	13626404	DEL	MONO	aAAAAA
N2	N-4	III	1	10729931	10729932	DEL	MONO	tTTTTTTT
N2	N-4	III	1	11049705	11049706	DEL	MONO	tTTTTTTT
N2	N-2	III	1	13307901	13307902	DEL	MONO	tTTTTTTTT
N2	N-4	III	1	11764750	11764750	SINS	MONO	tTTTTTTT
N2	N-4	V	1	16781480	16781480	SINS	MONO	tTTTTT
N2	N-4	X	1	17032785	17032785	SINS	MONO	aAAAAAAA
N2	N2_3	II	1	8535067	8535067	SINS	MONO	aAAAAAAAAAA
N2	N2_50	X	1	8861877	8861878	DEL	MONO	tTTTTTTTTT
N2	N-4	I	1	1264431	1264432	DEL	MONO	aAAAAAAAAA
N2	N-4	II	1	5020332	5020333	DEL	MONO	aAAAAAAA
N2	N-4	II	1	7188545	7188546	DEL	MONO	tTTTTTTT
N2	N-4	III	1	13307140	13307141	DEL	MONO	gGGGGGGGGG
N2	N-4	IV	1	16981105	16981106	DEL	MONO	aAAAAAAAAA
N2	N2_3	II	1	4642303	4642304	DEL	MONO	aAAA
polq-1	XF152_H50	X	14	9225274	9225288	DEL	DI	agagagagagagagAGAGAGAGAGAGAGAGAGAGA-GAGAGAGAGAGAGAG
polq-1	XF152_I50	I	9	1546401	1546410	DEL	NINE	tcggcaaatTCGGCAAATTCGGCAAATTCGGCAAATTC-GGCAAAT
polq-1	XF152_G50	III	3	12566581	12566581	SINS	TRI	agaAGAAGAAGAAGA
polq-1	XF152_H50	I	3	14124788	14124788	SINS	MONO	aaaAAAAAAAAAAAA
polq-1	XF152_H50	V	2	11496752	11496754	DEL	MONO	ggGGGGGGGGGGG

polq-1	XF152_I50	III	2	5416080	5416082	DEL	MONO	ggGGGGGGGGGGGG
polq-1	XF152_H50	III	1	212206	212207	DEL	MONO	tTTTTTTTTT
polq-1	XF152_G50	X	1	2312399	2312399	SINS	MONO	cCCCCCCCCCC
polq-1	XF152_I50	V	1	3449534	3449535	DEL	MONO	tTTTT
polq-1	XF152_I50	IV	1	5199068	5199069	DEL	MONO	tTTTTTT
polq-1	XF152_G50	X	1	6580098	6580099	DEL	MONO	aAAAAAAA
polq-1	XF152_E50	V	1	8296059	8296060	DEL	MONO	aAAAAAAA
polq-1	XF152_H50	IV	1	8633324	8633325	DEL	MONO	tTTTTTT
polq-1	XF152_E50	I	1	10093794	10093795	DEL	MONO	aAAAAAAAAA
polq-1	XF152_E50	IV	1	13304597	13304598	DEL	MONO	aAAAAAA
polq-1	XF152_E50	X	1	16192386	16192387	DEL	MONO	aAAAAAAAAA
polq-1	XF152_H50	V	1	17052577	17052577	SINS	MONO	aAAAAAAAAA
polq-1	XF152_I50	V	1	17711073	17711074	DEL	MONO	aAAAAAAAAA
polq-1	XF152_H50	V	1	18421950	18421950	SINS	MONO	tTTTTTTTTTT
polq-1	XF152_E50	X	1	15785644	15785645	DEL	MONO	gGGGGGGGGG

REFERENCES

1. Hoeijmakers,J.H. Genome maintenance mechanisms for preventing cancer. *Nature* 411, 366-374 (2001).
2. Wang,H. *et al.* Biochemical evidence for Ku-independent backup pathways of NHEJ. *Nucleic Acids Res.* 31, 5377-5388 (2003).
3. Wilson,T.E., Grawunder,U., & Lieber,M.R. Yeast DNA ligase IV mediates non-homologous DNA end joining. *Nature* 388, 495-498 (1997).
4. Boulton,S.J. & Jackson,S.P. Identification of a Saccharomyces cerevisiae Ku80 homologue: roles in DNA double strand break rejoining and in telomeric maintenance. *Nucleic Acids Res.* 24, 4639-4648 (1996).
5. Ceccaldi,R. *et al.* Homologous-recombination-deficient tumours are dependent on Poltheta-mediated repair. *Nature* 518, 258-262 (2015).
6. Koole,W. *et al.* A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. *Nat. Commun.* 5, 3216 (2014).
7. Mateos-Gomez,P.A. *et al.* Mammalian polymerase theta promotes alternative NHEJ and suppresses recombination. *Nature* 518, 254-257 (2015).
8. McVey,M. & Lee,S.E. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet.* 24, 529-538 (2008).
9. Roerink,S.F., van,S.R., & Tijsterman,M. Polymerase theta-mediated end joining of replication-associated DNA breaks in C. elegans. *Genome Res.* 24, 954-962 (2014).
10. Yousefzadeh,M.J. *et al.* Mechanism of suppression of chromosomal instability by DNA polymerase POLQ. *PLoS. Genet.* 10, e1004654 (2014).
11. Yoon,J.H., Roy,C.J., Park,J., Prakash,S., & Prakash,L. A role for DNA polymerase theta in promoting replication through oxidative DNA lesion, thymine glycol, in human cells. *J. Biol. Chem.* 289, 13177-13185 (2014).
12. Seki,M. *et al.* High-efficiency bypass of DNA damage by human DNA polymerase Q. *EMBO J.* 23, 4484-4494 (2004).
13. Shima,N., Munroe,R.J., & Schimenti,J.C. The mouse genomic instability mutation chaos1 is an allele of Polq that exhibits genetic interaction with Atm. *Mol. Cell Biol.* 24, 10381-10389 (2004).
14. Fernandez-Vidal,A. *et al.* A role for DNA polymerase theta in the timing of DNA replication. *Nat. Commun.* 5, 4285 (2014).
15. Shima,N. *et al.* Phenotype-based identification of mouse chromosome instability mutants. *Genetics* 163, 1031-1040 (2003).
16. Kloosterman,W.P. *et al.* Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. *Cell Rep.* 1, 648-655 (2012).
17. Villarreal,D.D. *et al.* Microhomology directs diverse DNA break repair pathways and chromosomal translocations. *PLoS. Genet.* 8, e1003026 (2012).
18. Ashwood-Smith,M.J. & Edwards,R.G. DNA repair by oocytes. *Mol. Hum. Reprod.* 2, 46-51 (1996).
19. Hamer,G. *et al.* Function of DNA-protein kinase catalytic subunit during the early meiotic prophase without Ku70 and Ku86. *Biol. Reprod.* 68, 717-721 (2003).
20. Robert,V. & Bessereau,J.L. Targeted engineering of the Caenorhabditis elegans genome following Mos1-triggered chromosomal breaks. *EMBO J.* 26, 170-183 (2007).
21. Plasterk,R.H. The origin of footprints of the Tc1 transposon of Caenorhabditis elegans. *EMBO J.* 10, 1919-1925 (1991).
22. Robert,V.J., Davis,M.W., Jorgensen,E.M., & Bessereau,J.L. Gene conversion and end-joining-repair double-strand breaks in the Caenorhabditis elegans germline. *Genetics* 180, 673-679 (2008).
23. Grishkevich,V. *et al.* A genomic bias for genotype-environment interactions in C. elegans. *Mol. Syst. Biol.* 8, 587 (2012).
24. Thompson,O. *et al.* The million mutation project: a new approach to genetics in Caenorhabditis elegans. *Genome Res.* 23, 1749-1762 (2013).
25. Sijen,T. & Plasterk,R.H. Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. *Nature* 426, 310-314 (2003).
26. Yu,A.M. & McVey,M. Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions. *Nucleic Acids Res.* 38, 5706-5717 (2010).
27. Chan,S.H., Yu,A.M., & McVey,M. Dual roles for DNA polymerase theta in alternative end-joining repair of double-strand breaks in Drosophila. *PLoS. Genet.* 6, e1001005 (2010).
28. Jinek,M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816-821 (2012).
29. Waaijers,S. *et al.* CRISPR/Cas9-targeted mutagenesis in Caenorhabditis elegans. *Genetics* 195, 1187-1191 (2013).
30. Gratz,S.J. *et al.* Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease. *Genetics* 194, 1029-1035 (2013).
31. Hwang,W.Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.* 31, 227-229 (2013).
32. Mali,P. *et al.* RNA-guided human genome engineering via Cas9. *Science* 339, 823-826

(2013).
33. Clejan,I., Boerckel,J., & Ahmed,S. Developmental modulation of nonhomologous end joining in Caenorhabditis elegans. *Genetics* 173, 1301-1317 (2006).
34. Lemmens,B.B., Johnson,N.M., & Tijsterman,M. COM-1 promotes homologous recombination during Caenorhabditis elegans meiosis by antagonizing Ku-mediated non-homologous end joining. *PLoS. Genet.* 9, e1003276 (2013).
35. Adamo,A. *et al.* Preventing nonhomologous end joining suppresses DNA repair defects of Fanconi anemia. *Mol. Cell* 39, 25-35 (2010).
36. Kent,T., Chandramouly,G., McDevitt,S.M., Ozdemir,A.Y., & Pomerantz,R.T. Mechanism of microhomology-mediated end-joining promoted by human DNA polymerase theta. *Nat. Struct. Mol. Biol.*(2015).
37. Carvalho,C.M. *et al.* Replicative mechanisms for CNV formation are error prone. *Nat. Genet.* 45, 1319-1326 (2013).
38. Boeva,V. *et al.* Breakpoint features of genomic rearrangements in neuroblastoma with unbalanced translocations and chromothripsis. *PLoS. One.* 8, e72182 (2013).
39. Higgins,G.S. *et al.* A small interfering RNA screen of genes involved in DNA repair identifies tumor-specific radiosensitization by POLQ knockdown. *Cancer Res.* 70, 2984-2993 (2010).
40. Brenner,S. The genetics of Caenorhabditis elegans. *Genetics* 77, 71-94 (1974).
41. Mori,I., Moerman,D.G., & Waterston,R.H. Analysis of a mutator activity necessary for germline transposition and excision of Tc1 transposable elements in Caenorhabditis elegans. *Genetics* 120, 397-407 (1988).
42. Dickinson,D.J., Ward,J.D., Reiner,D.J., & Goldstein,B. Engineering the Caenorhabditis elegans genome using Cas9-triggered homologous recombination. *Nat. Methods* 10, 1028-1034 (2013).
43. Tzur,Y.B. *et al.* Heritable custom genomic modifications in Caenorhabditis elegans via a CRISPR-Cas9 system. *Genetics* 195, 1181-1185 (2013).
44. Li,H. & Durbin,R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 25, 1754-1760 (2009).
45. Li,H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 25, 2078-2079 (2009).
46. McKenna,A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297-1303 (2010).
47. Ye,K., Schulz,M.H., Long,Q., Apweiler,R., & Ning,Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics.* 25, 2865-2871 (2009).
48. Thorvaldsdottir,H., Robinson,J.T., & Mesirov,J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178-192 (2013).
49. Andersen,E.C. *et al.* Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. *Nat. Genet.* 44, 285-290 (2012).
50. Keane,T.M., Wong,K., & Adams,D.J. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics.* 29, 389-390 (2013).
51. Purcell,S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559-575 (2007).
52. Lee,T.H., Guo,H., Wang,X., Kim,C., & Paterson,A.H. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC. Genomics* 15, 162 (2014).
53. Zeng,F., Baldwin,D.A., & Schultz,R.M. Transcript profiling during preimplantation mouse development. *Dev. Biol.* 272, 483-496 (2004).
54. Ketting,R.F., Haverkamp,T.H., van Luenen,H.G., & Plasterk,R.H. Mut-7 of C. elegans, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. *Cell* 99, 133-141 (1999).

5

GENOMIC SCARS GENERATED BY POLYMERASE THETA REVEAL THE VERSATILE MECHANISM OF ALTERNATIVE END-JOINING

Robin van Schendel, Jane van Heteren, Richard Welten and Marcel Tijsterman

Department of Human Genetics, Leiden University Medical Center, The Netherlands

Published in PLOS Genetics 2016 October

ABSTRACT

For more than half a century, genotoxic agents have been used to induce mutations in the genome of model organisms to establish genotype-phenotype relationships. While inaccurate replication across damaged bases can explain the formation of single nucleotide variants, it remained unknown how DNA damage induces more severe genomic alterations. Here, we demonstrate for two of the most widely used mutagens, *i.e.* ethyl methanesulfonate (EMS) and photo-activated trimethylpsoralen (UV/TMP), that deletion mutagenesis is the result of polymerase Theta (POLQ)-mediated end joining (TMEJ) of double strand breaks (DSBs). This discovery allowed us to survey many thousands of available *C. elegans* deletion alleles to address the biology of this alternative end-joining repair mechanism. Analysis of ~7,000 deletion breakpoints and their cognate junctions reveals a distinct order of events. We found that nascent strands blocked at sites of DNA damage can engage in one or more cycles of primer extension using a more downstream located break end as a template. Resolution is accomplished when 3' overhangs have matching ends. Our study provides a step-wise and versatile model for the *in vivo* mechanism of POLQ action, which explains the molecular nature of mutagen-induced deletion alleles.

INTRODUCTION

DNA mutations fuel evolution of organisms giving rise to speciation, and of cells within an organisms giving rise to cancer. Two replication-associated mechanisms are responsible for most if not all single nucleotide variants (SNVs) as well as small insertions/deletions (indels) at repetitive sequences: i) copying errors made by the replicative polymerases delta and epsilon, which are mostly undone by DNA mismatch repair, and ii) replication of damaged DNA by specialized so-called translesion synthesis (TLS) polymerases. TLS polymerases, in contrast to the replicative polymerases, have the ability to extend nascent DNA strands across non- or poorly coding damaged bases, often leading to mutation. It is, however, less well understood which mechanisms are responsible for other types of genomic alterations, such as deletions that are larger than a few bases.

A recent study that involved whole genome analysis of *C. elegans* animals that were propagated for many generations revealed that vast majority of accumulating deletions larger than 1 bp required the activity of the A-family polymerase Theta (POLQ). Upon unperturbed growth, wild-type *C. elegans* genomes accumulate SNVs as well as deletions but the latter class was strikingly absent in strains that were defective for POLQ[1]. Instead, much more dramatic chromosomal rearrangements were noticed indicating that POLQ action protects the genome against deterioration but at the cost of a small genomic scar. A similar profile of mutagenesis was observed resulting from DNA double-strand break repair, which hinted towards DSBs as being a very prominent source of genome diversification during evolution, and towards error-prone DSB repair as the mechanism responsible for this type of genome alterations[1].

The first demonstration of POLQ acting on DSBs was made in *Drosophila*: *in vivo* processing of artificially-induced DSBs in POLQ-mutant flies deviated from that in wild-type flies[2]. POLQ deficiency did not increase sensitivity to ionizing radiation, yet it did greatly exacerbate hypersensitivity in flies impaired in homologous recombination. Apparently, a POLQ-dependent DSB-repair pathway can act as a backup in HR-compromised circumstances. Indeed, recent work on human POLQ revealed a strong synergistic relationship between the HR pathway and POLQ-mediated DSB repair[3,4]. The synthetic lethal nature of this genetic interaction may be of great clinical importance as it identifies POLQ as a druggable target for tumours carrying mutations in HR genes. Another indication that POLQ repairs DSBs in contexts where HR is compromised came from genetic studies performed in *C. elegans*. Here it was shown that POLQ-mediated repair is the only pathway (also in HR-proficient conditions) capable of repairing replication-associated DSBs that are induced when persistent DNA damage or stable secondary structures cause a permanent block to DNA replication[5,6]. It was subsequently shown that these DSBs result from inheritable ssDNA gaps opposite to the strand containing the damage, which could thus not serve as a template for HR[7].

Extensive analyses of repair products in both flies and worms provided a clear signature of POLQ-mediated DSB repair with two prominent features: i) the notion of microhomology at the repair junctions, a feature previously ascribed to non-canonical end-joining also called alternative end-joining[8,9], and ii) the occasional presence of so-called template inserts: deletions that contain, at the deletion junction, the inclusion of a DNA insert (hereafter called delins). These inserts are of variable length but their origin can be mapped to DNA regions that lie in very close proximity to the DSBs ends that produced the delins. Similar hallmarks can be found for POLQ-mediated DSB repair in human and mouse cells[4,10]. A recent *in vitro* study provided a molecular explanation for the prominent presence of microhomology at the DSB repair junctions: repair reactions with purified protein showed that two base pairs of complementarity is enough for human POLQ to pair

and extend 3' overhangs of partially double-stranded oligonucleotides[11].

Although it is now becoming increasingly clear that POLQ plays an evolutionarily conserved role in DSB repair, how POLQ acts *in vivo* to explain all the observed consequences remains to be elucidated. Over the last four decades, the *C. elegans* community has used EMS and UV/TMP to generate many thousands of deletion alleles, but the underlying mechanism has remained unknown. Here, we demonstrate that mutagen-induced replication breaks in *C. elegans* germ cells are exclusively repaired by POLQ. This publically available allele collection, reflecting ~7,000 *in vivo* POLQ-mediated end joining reactions, allows us to analyse and describe the POLQ-mediated repair mechanism in great detail.

RESULTS

POLQ-deficient animals are hypersensitive to EMS and UV/TMP

To investigate whether POLQ plays a general role in the processing of mutagen-induced DNA damage, we assayed embryonic survival in animals that were exposed to two of the most widely used mutagens in *C. elegans*: EMS, which causes alkylating damage, and TMP, which, upon exposure to UVA light, results in monoadducts and crosslinks. We found *polq-1*-deficient animals to produce more unviable embryos than wild-type animals when exposed to EMS (Fig 1A, S1 Fig), but not to the extent observed in animals that are defective for polymerase eta (*polh-1*), a translesion synthesis (TLS) polymerase that is involved in replicative bypass of DNA damage[12]. A similar mild hypersensitivity was observed when *polq-1*-mutant animals were incubated with TMP and subsequently exposed to UVA (Fig 1B, S1 Fig), in agreement with previously published work[13]. In addition to monitoring the survival of embryos, we monitored their ability to produce functional gametes. Complete or partial sterility of daughters from exposed mothers is another phenotype that is related to genotoxic stress, likely because germ cells, or their progenitors, are more susceptible to DNA damage-induced arrest, apoptosis, and mitotic catastrophe[14]. Indeed, at EMS or UV/TMP doses where the brood size of exposed mothers were only moderately affected in both wild-type and *polq-1*-mutant animals (Fig 1C-D) dramatic sterility was observed in *polq-1* but not in wild-type progeny animals (Fig 1E-F): 99% versus 16% median reduction, in brood for EMS-treated animals, and 65% versus 5% for UV/TMP-treated animals. These data establish a prominent role for POLQ in protecting germ cells against EMS and UV/TMP-induced toxicity.

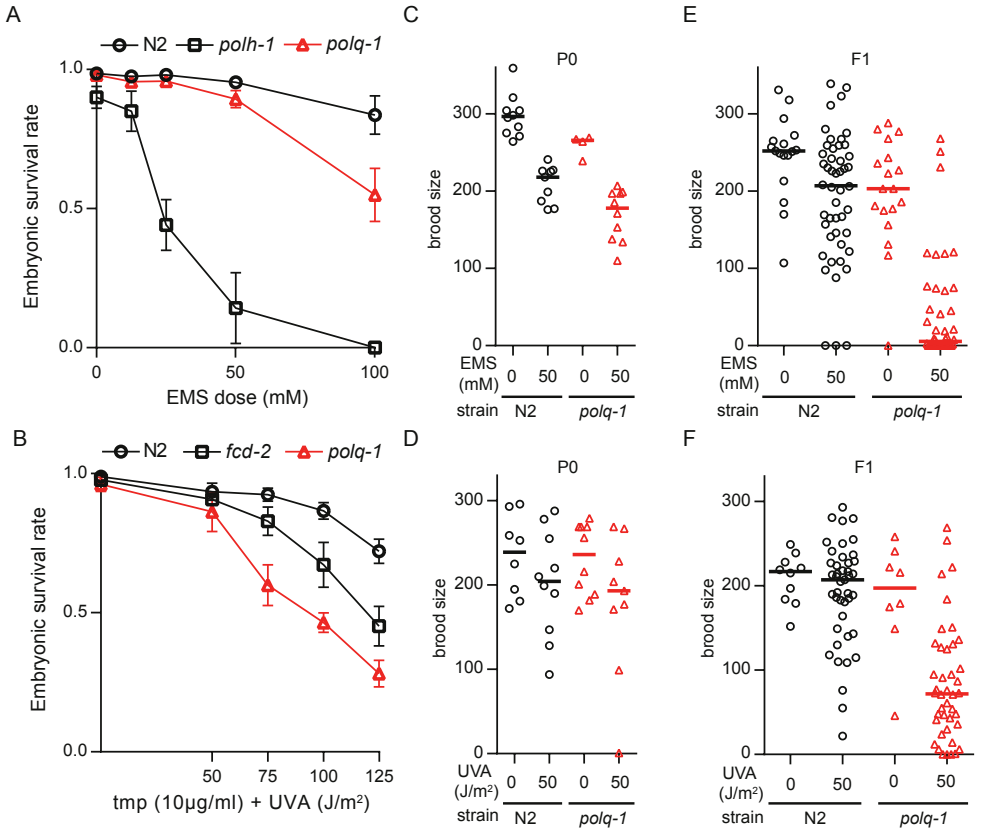


Fig 1. POLQ-deficient animals are hypersensitive to EMS and UV/TMP A. Sensitivity to EMS exposure. **B.** Sensitivity to UV/TMP treatment. L4 animals of the indicated genotype were exposed to DNA damaging treatments and survival was quantified by counting dead embryos versus living progeny in the next generation. **C-D.** The total brood (eggs + larvae) was determined for P0 animals of the indicated genotype that were mock treated or treated with EMS (C) or UV/TMP (D). Lines represent the median for each dataset. **E-F.** The total brood was determined for F1 animals that originated from P0 animals that were either mock treated or treated with EMS (E) or UV/TMP (F). Lines represent the median for each dataset.

EMS and UV/TMP-induced deletions are dependent on POLQ

EMS and UV/TMP are widely used mutagens in *C. elegans* to create loss-of-function alleles[15]. Given the sensitivity of *polq-1* animals towards these agents we wanted to investigate whether POLQ functionality is relevant for generating these alleles. EMS predominantly alkylates guanine which can be bypassed, leading predominantly to GC>AT transitions[15-17]. Deletions also result from EMS treatment through yet unknown biology[17]. UV/TMP treatment results in a different spectrum of mutations: for this mutagen, deletions dominate base pair substitutions[17,18], but also here, the underlying mechanism of deletion formation is unknown. To address the candidate role of POLQ in producing deletion alleles, we created libraries of mutagenized wild-type and *polq-1*-mutant animals and screened them for deletions. We used standard protocols that were previously used by numerous laboratories and consortia leading to the ~10,000 *C. elegans* deletion alleles that

are currently available[19-21]. The general concept of these protocols is to find by PCR a smaller than wild-type product for a target of interest in pooled broods of mutagenized animals; then use a sib-selection strategy to isolate the mutant allele (S2 Fig and Methods section). Because the progeny of mutagenized *polq-1*-animals have a reduced brood size (Fig 1E-F), we screened the F1 generation, and not the F2, which allowed us to inspect the same number of animals for *polq-1*-mutant and wild-type genotypes. We screened the libraries for deletions using eight different amplicons, all ~1 kb in size. Positive pools were chased by PCR of less-complex pools and individual library addresses (in duplicate) to exclude false positives (See Methods for details). This strategy proved to be robust and specific as deletion alleles were readily detected in wild-type animals exposed to either EMS or UV/TMP, but not in mock-treated animals (Fig 2A-B and S2B-C Fig). In contrast, we did not find a single deletion allele in libraries of either EMS- or UV/TMP-mutagenized *polq-1* animals (Fig 2A-B). From this data we conclude that EMS- and UV/TMP-induced deletion mutagenesis, in the size range of 50 bp up to ~1 kb, requires functional POLQ.

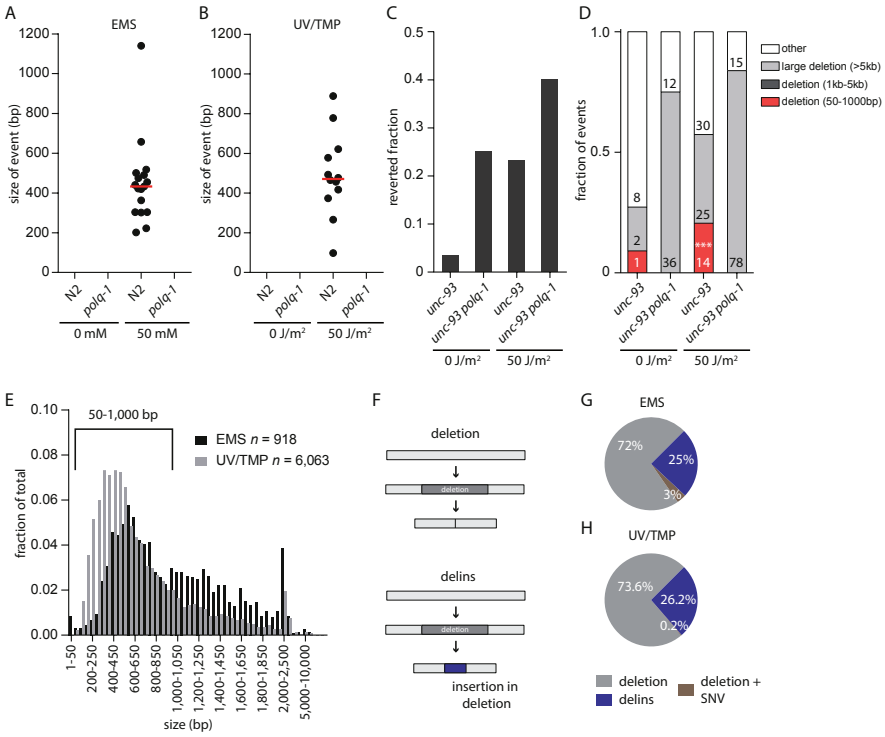


Fig 2. EMS and UV/TMP-induced deletion alleles are dependent on POLQ. A-B. Size distribution for all confirmed deletion events found in EMS (A) or UV/TMP (B) mutagenized libraries. Red bars represent the median deletion size. **C**. Fraction of populations that contained *unc-93(e1500)* revertant animals. At least 250 populations were assayed per experimental condition. **D**. Distribution of *unc-93* reversion-footprints for the indicated genotype and experimental condition. The class of 50-1000bp was found to be statistically different between treated *unc-93* and *unc-93 polq-1* animals. The category 'other' includes wild-type sized PCR products, which based on previous experiments mostly reflect base substitutions. ($p<0.001$, Fisher's exact test, indicated by ***) **E**. Size distribution of EMS- and UV/TMP-induced deletions generated by the *C. elegans* community. Only the deletions 50 – 1,000 bp (918 and 6,063 for EMS and UV/TMP-induced deletions, respectively) were used in subsequent analyses. **F**. Graphic representation of the two different types of deletions. The upper panel illustrates a simple deletion, in which only sequence is lost; the bottom panel reflects a delins, in which loss of

To further validate this conclusion we investigated UV/TMP-induced mutagenesis in a more unbiased fashion by catching loss-of-function mutations in an endogenous genomic target, *unc-93*. A dominant mutation in the transmembrane protein UNC-93, *unc-93(e1500)*, causes worms to move uncoordinatedly. Loss of UNC-93 expression, or of one of its cofactors SUP-9 and SUP-10 results in a reversion to wild-type movement, which provides an easy phenotypic manner to monitor loss of function mutagenesis. We exposed POLQ-proficient and -deficient animals, carrying the *unc-93(e1500)* allele to TMP with or without UVA irradiation to introduce crosslinks. Wild-type-moving animals were isolated from the brood of exposed animals and subsequently inspected for deletions in *unc-93, sup-9* and *sup-10*. The mutants that did not, by DNA gel electrophoresis, reveal a deletion in any of the three genes are likely the result of single nucleotide variations (SNVs) and were not further analysed. In treated wild-type animals, we observed an increase in two distinct categories of deletions (Fig 2C-D): one class, comprising of small, 50 bp to 1 kb, deletions with median size of ~100 bp (S2D Fig), and another class in which deletions are substantially larger, being >5 kb in size (Fig 2D). No deletions were found in the size range 1-5 kb. UV/TMP-treated *polq-1*-deficient animals were, however, devoid of small deletions, while the ratio of very large deletions further increased (Fig 2C-D). Based on these data and the PCR-based screenings of UV/TMP-treated mutant libraries, we conclude that the vast majority (if not all) of small deletions in the range of 50 bp up to at least 1 kb are the result of POLQ action. In its absence large deletions manifest, which, in agreement with our previous work, argue that POLQ prevents large genomic alterations at replication blocking DNA lesions at the expense of relatively small deletions[1,5,6].

Replication approaches to one nucleotide from the damage

Above, we demonstrate that deletion alleles isolated from libraries of EMS- and UV/TMP-treated populations are the result of POLQ action. This notion allows us to systematically analyse a uniquely rich collection of ~2,000 EMS- and ~8,000 UV/TMP-induced deletion alleles that were generated by the *C. elegans* community to elucidate the *in vivo* mechanism of POLQ action. Fig 2E displays the sizes for all ~10,000 alleles, for which the sequence information was retrieved from WormBase[22]. The majority of alleles are between 50 bp and 1kb and can be categorized into two groups: i) simple deletions, which make up the majority of events (~70-75%) in both the EMS and in the UV/TMP dataset, and ii) deletions that are accompanied by an insertion of a small segment (median: 5 bp for both sets) of novel DNA; we refer to this class (~25-30%) of alleles as delins (Fig 2F-H). We set out to characterize the ~5,000 deletions and ~1,800 delins, filtered to size (50-1,000 bp), into great detail.

First, we investigated the base composition of deletion junctions to further examine an earlier reported relationship in POLQ-mediated mutagenesis between the position of a deletion breakpoint and the position of a replicating blocking lesion: we previously found for deletions resulting from replication blocking G-quadruplexes that one of the breakpoints maps close to the replication impediment[6]. This led to a model where deletions result from processing the 3' hydroxyl ends of blocked nascent strands. DNA lesions induced by EMS and UV/TMP also have the potential to block replication, and we thus questioned whether cognate deletions close to their breakpoints carry the signature of EMS- or UV/TMP-inflicted base damage. More precisely, if one

sequence is accompanied with the insertion of *de novo* sequence. **G-H**. Pie chart representation of the fraction of deletions and delins that were isolated from EMS (G) and UV/TMP (H) mutagenized libraries. Deletions + SNV represent cases where a SNV is found in close proximity to a deletion.

of both breakpoints results from processing a stable but reactive nascent strand that was extended up to the damaged base, then the first nucleotide immediately downstream of the breakpoint (the -1 position) might reveal the nature of the replication impediment (see Fig 3A for a graphical illustration of this concept). Indeed, we found a clear non-random base composition at position -1: for EMS we found an overrepresentation of cytosine (Fig 3B and S3 Fig), which perfectly fits the damage spectrum of EMS predominantly ethylating guanines[16,17]. Blocked DNA synthesis, incapable of extending across a damaged guanine, would result in a 3′ hydroxyl end immediately upstream of a cytosine. Also for deletions induced by UV/TMP we found at the -1 position a clear mutagen-specific overrepresentation of a particular base, in this case an adenine (Fig 3C), which reflect TMPs reactivity towards thymines[23]. Strikingly, and in contrast to the EMS spectrum, we here also observed a non-random distribution at the +1 position, being a thymine. This outcome suggests that UV/TMP-induced deletions are preferentially induced at sites where replication is blocked by a thymine that is preceded by an adenine, a conclusion that is further supported by probing the datasets with pairs of nucleotides (S3 Fig). This prevalent signature is in perfect agreement with the preference of psoralens to intercalate into and react with 5′TA in duplexed DNA[24,25]. Without further genetic dissection, however, it is impossible to discriminate between interstrand crosslinks at 5′TA sites or monoadducts (or DNA-protein complexes) formed at sites of preferred intercalation, being responsible for POLQ-dependent deletion formation. Irrespective which lesion, our data indicates that replication can proceed right up to the base that is damaged by the psoralen moiety.

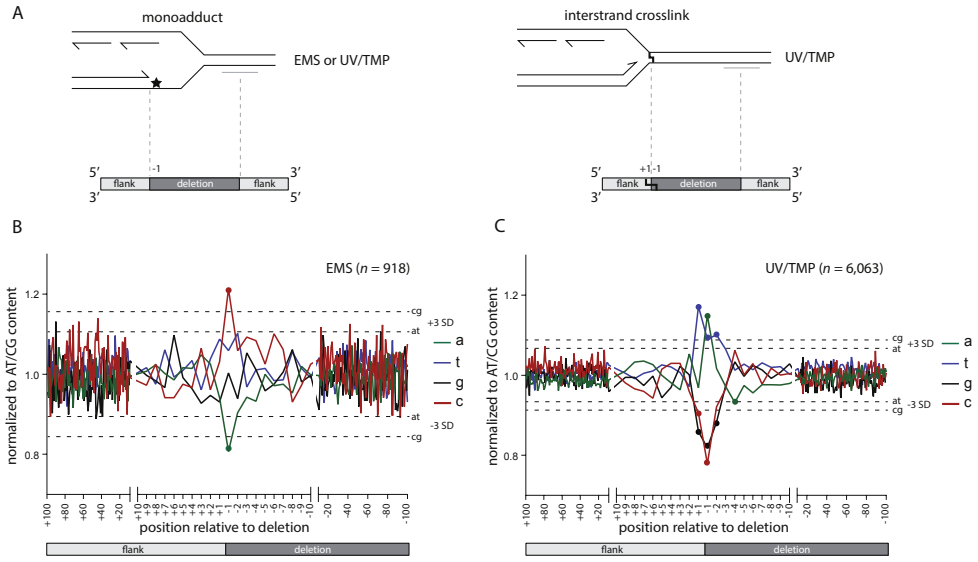


Fig 3. Replication approaches to one nucleotide from the damage. A. Schematic illustration of the concept that one junction of DNA-damage-induced deletions is defined by the nascent strand blocked at sites of DNA damage. In this hypothesis, the replication-blocking lesion may dictate position -1, being the outermost nucleotide of the lost sequence. **B-C**. The base composition of all breakpoints, normalized to the relative AT/CG content around the breakpoints (from +100 to -100). Position +100 to +1 reflects the sequence that is retained in the deletion alleles; position -1 to -100 reflects the sequence that is lost. Dashed lines represent three times the SD. Data points outside these boundaries are marked with a dot.

Our analysis of ~7,000 mutagen-induced deletion alleles reveals a clear lesion-specific signature in POLQ-mediated deletion formation. Importantly, a single replication fork block triggers such a deletion, as we observed a damage signature at only one of both breakpoints (S4 Fig). The position of the damage with respect to the deletion junction supports a mechanistic model where the nascent strand blocked at the site of base damage is not subjected to extensive trimming but instead is reactive towards a POLQ-mediated end-joining reaction that has small sized deletions as an end-product. The putative mechanism responsible for generating the other reactive end at a 50-1,000 bp distance will be discussed later, but we will provide evidence that, with respect to reactivity, it is indistinguishable from the blocked nascent strand.

Single nucleotide priming is sufficient to initiate repair by POLQ

We reveal above that the terminal nucleotide of the nascent strand, blocked at the site of base damage, is retained in the repair product, it is the base immediately flanking the deletion, but does it also guide repair? To address this question we compiled all simple deletions from the UV/TMP dataset that had the signature T_{+1},A_{-1} composition at one of both breakpoints, because only for this subclass (*n*=1,248) the identity of the terminal nucleotide of the nascent strand is known, i.e. a thymine. We then tested the following prediction: if this 3' thymine is guiding repair of the break, by providing a minimal primer for POLQ, a thymine should be overrepresented at the -1 position of the opposite flank (Fig 4A for a graphical illustration). This is indeed what we found: Fig 4B shows that the composition of the donor sequence opposite to the blocked nascent strand is completely random apart from position -1, which is dominated by a thymine. A similar conclusion results if we use an approach that is blind to the replication-obstructing base and does not restrict the analysis to a single nucleotide. For each of the ~5,000 alleles we established the degree of homology between both breakpoints by scoring the degree of sequence identity in a 16-nt window, encompassing the 8 outermost nucleotide of the flanking sequence and the 8 nucleotides of the adjacent but deleted sequence (see Fig 4C for a schematic illustration of the approach). These plots were subsequently compiled to generate heat maps for the different category of alleles. In both the UV/TMP-induced (*n*=4,461) and the EMS-induced deletions (*n*=662) crosstalk between both breakpoints is observed, but only for the nucleotide at the -1 position of the deletion and the +1 position of the opposing flank (Fig 4D). This outcome lends further support to the hypothesis that the terminal base of one end, upon minimal pairing with the opposing template, is guiding POLQ-mediated repair.

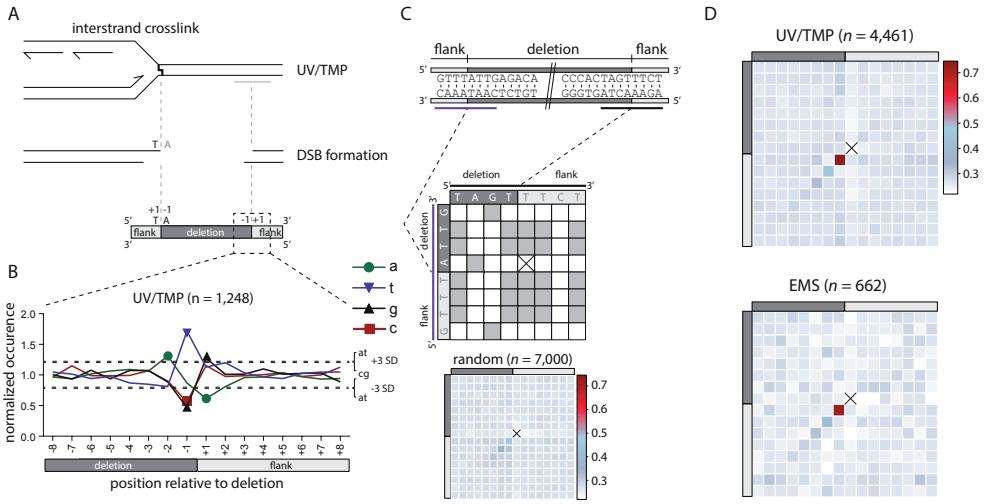


Fig 4. POLQ-mediated repair is characterized by single nucleotide homology. A. Schematic illustration of a replication fork blocked at an UV/TMP-induced crosslink that subsequently leads to a DSB, which is repaired by POLQ leading to a deletion of the intervening sequence. One reactive end of the DSB is determined by the nascent strand blocked by an UV/TMP-induced crosslink that predominantly links thymines in opposite strands when in a 5'TA configuration. **B**. Deletion alleles that contain a 5'TA at the (+1, -1) position of one of their breakpoints are analysed (*n*=1,248) for the base composition at the opposite breakpoint. Dashed lines represent three times the SD, which is determined by the base composition of the region between -100 and +100. **C**. Schematic illustration of how microhomology between breakpoints is determined in an unbiased manner. For each allele a table is constructed that allows for the scoring of homology between both breakpoints that give rise to a deletion. Each position of the upstream breakpoint (purple) is compared to each position of the downstream breakpoint (black). Identical nucleotides score 1, non-identical score 0. Subsequently, a heat map is constructed by summing all scores for all events at each position divided by the number of events. For reference purposes, a heat map was constructed for 7,000 deletions randomly created *in silico* throughout the genome. Of note, all alleles are annotated in keeping with maximal 5' conservation, which here dictates that the base at the -1 position at the 5' side is never identical to the +1 position at the 3' side: in such a case, that base will shift to the +1 position at the 5' side. As a consequence of this rule, the position marked by a cross will have no microhomology score, while the +1,-1 position is slightly elevated. The extent of this methodological skewing can be noticed in the analysis of the random set of deletions. **D**. Heat maps for UV/TMP- and EMS-induced deletions. Heat map contains 16 bases overlapping each breakpoint; 8 bases immediately flanking the deletion (light grey) and 8 bases immediately inside the deletion (dark grey).

Templated inserts and simple deletions have a common origin

Once priming has been established and extension has commenced there are two possible fates: i) continuation and further processing; in which case the outcome will be a deletion with single nucleotide identity at the junction, or ii) discontinuation. If, in the latter case, the extended end serves as a new nucleation site for yet another round of POLQ-mediated repair, templated inserts will result (Fig 5A). If so, delins are suspected to have some features identical to those described above for simple deletions. To address this, and to further dissect the *in vivo* mechanism of POLQ-dependent mutagenesis, we characterized the ~25-30% of mutagen-induced deletion alleles that are accompanied by small insertions in great detail. First we placed them, based on their size and suspected origin, in different categories (Fig 5B): ~47-50% are so small (<5 bp) that their origin is untraceable, and another 5-10% are larger in size but their sequence does not provide enough certainty as to their origin. However, ~40-45% of delins (~700) have inserts with sufficient

sequence information to reveal their source: apart from a small percentage (~3%) that comprise of sequences mapping to distant sites at the same chromosome or to other chromosomes (S5 Fig), the majority (~37-44%) maps very close to the deletion. These insertions are either completely or partially identical to parts of the flanking sequences and have been designated 'templated inserts' because of a presumed role for the flanking DNA to serve as a template for a repair reaction. Because the majority of templated inserts map a few bases away from the deletion junction (the template is located within the flank) a number of parameters can be investigated centred around the questions: i) what defines the start of POLQ-mediated DNA synthesis, ii) what defines the end, and iii) how accurate is it?

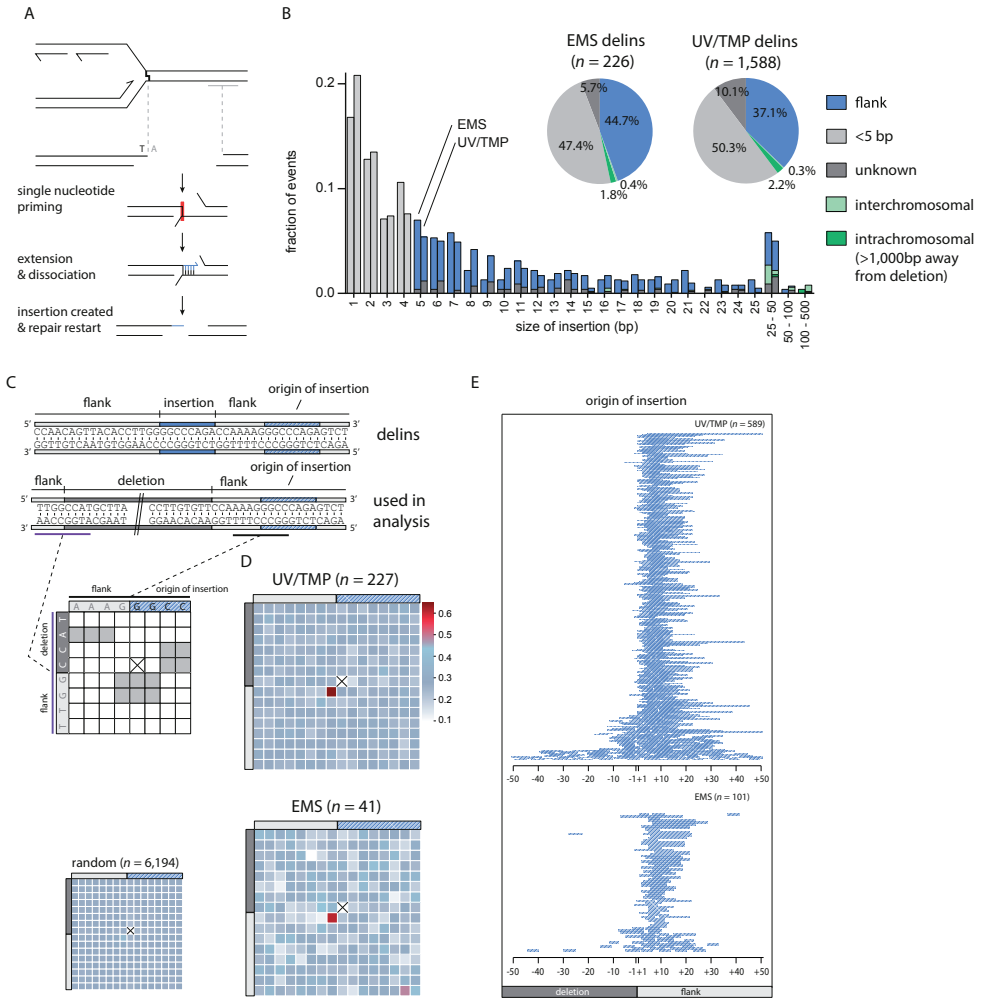


Fig 5. Hallmarks and genesis of delins A. Schematic illustration of the concept that templated insertions are generated by POLQ-mediated extension of one reactive 3' end (*e.g.* the nascent strand blocked at sites of base damage) using the other end as a template: single nucleotide priming and disrupted extension can lead to delins formation. **B.** Size distribution of insertions found in EMS- and UV/TMP-derived delins. For 47-50% of delins the insert size is too small (<5 bp) to uniquely identify their origin. 37-44% of delins can be mapped to within 20 bp flanking the breakpoint. Another 2-3% of delins are copied from inter- or intrachromosomal (>1000bp away

from deletion) locations. For 6-10% of delins no apparent source could be identified. **C**. Schematic illustration for how microhomology is determined between the sequence that was used as a template for the generation of an insertion (the template) and the opposite breakpoint (the primer). A typical delins is portrayed at the sequence level as an example in which both the insertion (in blue) as its identified origin (in striped blue) is indicated. Underneath is another representation of the same delins, now containing the deleted sequence. This configuration is used in the subsequent analysis, where for each delins a table is constructed in which the bases overlapping with the 5' side of the insertion origin (black) are compared to the bases that are overlapping the opposite breakpoint (purple). Identical nucleotides score 1, non-identical score 0. Subsequently, a heat map is constructed by summing all scores for all events divided by the number of events at each position. For reference purposes, a heat map was constructed for ~6,000 delins with perfect templated flank insertion randomly created *in silico* throughout the genome. Of note, at one position such a comparison cannot be done because the start and end nucleotide of an insertion is never identical to the deleted part of a delins and are thus always 0 (crossed out). As a result some other positions become slightly overrepresented as can be appreciated from the *in silico* generated delins. **D**. Heat map for UV/TMP- and EMS-induced delins for which the origin of the inserts are mapped. **E**. Visual representation of the origins of flank insertions for UV/TMP- and EMS-induced delins. A single line represents one mapped flank insertion and is drawn relative to its cognate breakpoint with '-' for deleted and '+' for retained sequences.

With respect to the start, we focused on templated inserts that are 100% identical to sequences in their flanks to avoid possible ambiguity in interpretation. For both UV/TMP and EMS-induced alleles (n=227 and 41, respectively) we found that templated inserts, similar to simple deletions, are primed by a single base pair. This priming becomes apparent when the base composition of one breakpoint is plotted to the base pairs that are neighbouring the sequence that served as a template for extension (Fig 5C-D). Overrepresentation of sequence identity is confined to one position, the +1 base of one breakpoint (the reactive end) and the base flanking the origin of the insert in the opposite breakpoint (the template), providing further confirmation that a single base pair is sufficient to drive POLQ-mediated repair. We found that ~85% of inserts originate from priming within 10 base pairs of the breakpoints (Fig 5E), which could point to homology search close to the end of the available sequence.

Templated inserts result from template switching and reiterated priming

The observed similarities in the initiation steps of deletions that are simple and those that include a templated insert means that the difference between both outcomes is the consequence of a downstream step, for instance, discontinuity of POLQ action. The determinants influencing discontinuity in the repair reaction are currently unknown but it is a remarkable frequent event as ~25% of all alleles have insertions. From plotting the size of all inserts (Fig 5B), we infer that templated inserts do not have a minimal length: although it is impossible to reliably map inserts of only one or a few bases to the flanking sequences, we observe that the percentage of inserts that can be mapped is constant, yet high, over the complete range of small insert size. This notion argues that also the very small, unmappable, insertions are flank-derived. Fig 5B also shows that while template inserts are overall rather small (<25 bp), they do not have a preferred size. Instead, a gradual decline in length is observed which may suggest that comprehensive extension prevents discontinuity. Still, we also found inserts where stretches of more than 20 consecutive bases have been templated, indicating that substantial base pairing can still be disrupted before the two opposite ends are irreversibly connected. Whether POLQ dissociates from the template in this process or whether POLQ facilitates template switching is an interesting question as the latter option could serve to broaden the resolving potential of POLQ-mediated repair. Some delins have complex combinatorial inserts with two or more mostly overlapping templated inserts, arguing for

reiterative steps of priming, extension and dissociation. In most of these cases (16 out of 17) only one flank provided the template, which hints towards directionality in POLQ-mediated resolution.

To complete repair of aborted reactions, it seems plausible that another round of priming and extension is required, analogous to the biology leading to simple deletions, only in this case, one end has been extended using the other end as a template. To test this hypothesis, we again created heat maps, but here compared the terminal bases of the origin of the template inserts as well as their flanking bases (as this constitutes the new reactive end), to the border of the same flank, which in this scenario is considered the opposing end (Fig 6A). We indeed found support for a single base pair priming reaction as also here a clear overrepresentation of single nucleotide identity is observed (Fig 6B-C). Our combined analysis thus supports a model, where simple deletions and template inserts result from the same chemistry, displaying the same features, the only difference being an aborted POLQ-mediated extension of a single base paired-primed intermediate.

Probing the entire collection of ~10,000 EMS- and UV/TMP-induced *C. elegans* deletion alleles for single nucleotide identity at break junctions and the presence of template inserts suggest that POLQ-mediated end joining is responsible for the majority of deletions in a 50-3,000bp range (S6 Fig).

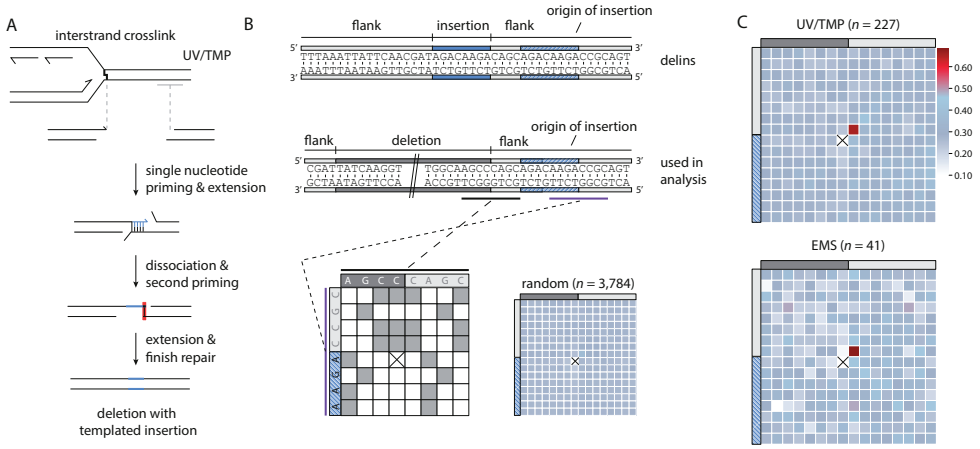


Fig 6. Primer-template switching results in delins formation. A. Schematic illustration of how primer template switching followed by POLQ-mediated extension and resolution results in a templated insertion. The requirement of single-nucleotide homology in POLQ-mediated end joining predicts that the nucleotide directly 3' of the templated insertion (blue line) is typically identical to the outermost nucleotide of the 'acceptor' breakpoint. This prediction is highlighted by the red box. **B**. As in Fig 5C, but here for the end of the origin of templated insert and the adjacent deletion junction. As an example a typical delins is portrayed at the sequence level in which both the insertion (in blue) as its identified origin (in striped blue) is indicated. Underneath is another representation of the same delins, now containing the deleted sequence. This configuration is used in the subsequent analysis. Of note, at one position such a comparison cannot be done because the start and end nucleotide of an insertion is never identical to the deleted part of a delins and are thus always 0 (crossed out). As a result some other positions become slightly overrepresented as can be appreciated from the *in silico* generated delins. **C**. Heat map for UV/TMP and EMS-induced delins where the insertion origin could be faithfully traced back to the immediate flank.

POLQ activity is error prone

At present it is unknown what underlies the discontinuity in POLQ-mediated repair that leads to delins instead of simple deletions. One possibility is polymerase errors. POLQ is a relatively error-prone polymerase generating single base errors at rates 10- to more than 100-fold higher than other polymerase A family members[26]. Mismatches resulting from wrongly incorporated nucleotides may reduce POLQ's processivity and promote dissociation and/or template switching. One observation provides strong support for such a scenario: the frequency of errors observed in templated inserts is extremely high as compared to mutations in the flanks of the simple deletions, while for both repair products the flank has served as a template for POLQ action. Although ~30% of all templated inserts are perfect, in the sense that they do not show mismatches, another 15% can be matched to the flank through a single run of consecutive bases if one mismatch or one slippage event is allowed (Fig 7A). It can thus be argued that at least 1 in 3 templated inserts suffers from a mutation which translates to an error rate of ~1 in 30 base pairs during templated extension (average insert size = ~10bp). In sharp contrast, we found only few mutations in the flanks of ~4,500 UV/TMP- induced simple deletions. Assuming that here POLQ is required to extend the reactive end with at least 10 bp, we calculate an error rate of <1 in 3,000 bp for simple deletions. To explain the >100 fold higher mutation frequency in extension leading to templated inserts, we propose that POLQ errors in fact provoke template switching, thus are causal to the formation of delins. A supporting observation is that mismatches are more frequently found closer to where the reaction is abrogated (Fig 7B).

POLQ replication errors could result from replicating non-damaged or damaged DNA. The *in vitro* demonstrated bypass activity of POLQ may help to extend past base damage or abasic sites. We mostly found incorrect incorporation of adenines opposite to any nucleotide other than a thymine (Fig 7C), making up for half of all mismatches, which fits with the preferential incorporation of adenine that has been observed for POLQ *in vitro*[27].

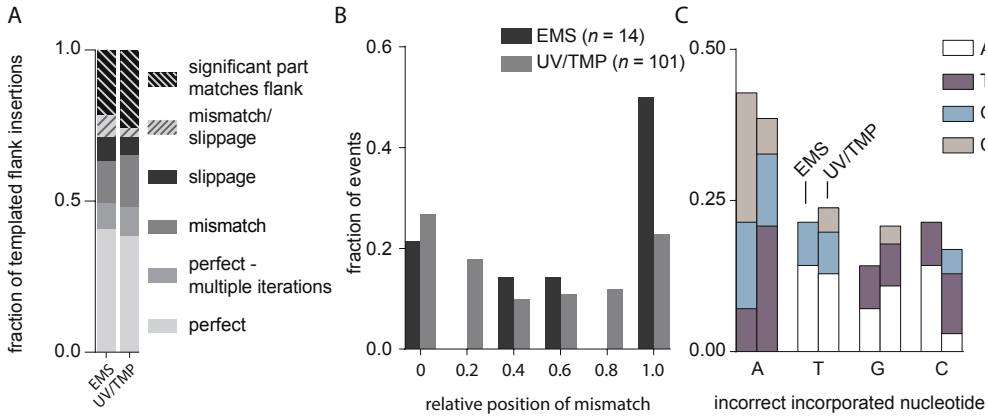


Fig 7. POLQ activity is error prone. A. The fraction of templated flank insertions derived from a single origin is greatly increased when we allow a SNV or a slippage-event in a microsatellite (≥4 bp). B. The relative position of mismatches in delins is plotted for each mutagen relative to the insertion. C. Fraction of incorrect incorporated nucleotides in EMS and UV/TMP deletions, grouped by nucleotide misincorporation.

Mutagen-induced deletions are the product of DSB repair

Finally, using this unique dataset of ~7,000 *in vivo* POLQ reactions we re-evaluated the assumption

that POLQ acts to protect against mutagen-induced damage by acting on replication-associated DSBs. Despite having demonstrated that POLQ-mediated end joining is a stand-alone DSB-repair pathway that is able to process bona fide DSBs[1], it remained difficult to formally prove that a DSB is an intermediate in a repair reaction that produces simple deletions and templated inserts that were previously also found to accumulate in mutants defective for TLS polymerases. Through combining the features that characterize POLQ-mediated deletions, a mutagen, *i.e.* UV/TMP, that leaves a signature in the final product, and the sheer size of the collection analysed here, we are now able to establish that replication-associated deletion mutagenesis results from the processing of two opposing 3' extendable ends, hence a DSB. Above, we have shown that a nascent strand blocked at a site of base damage can serve as a single nucleotide primer to be extended, using a donor sequence, located 50-1,000 bp away, as a template. In Fig 8, we show that there is an equal likelihood of finding the reciprocal event: that the sequence immediately upstream of the blocked fork has served as a template for a priming, reactive end that is located 50-1,000 bp more downstream. This argues that POLQ-mediated repair, as in repairing bona fide DSBs, here acts to connect two 3' reactive ends. It is currently unknown whether POLQ-mediated repair of replication-associated DSBs necessitates end-resection to create sizable 3' ssDNA regions (which then function as primer or as template). *In vitro*, human POLQ can extend ssDNA molecules intra-molecularly through a fold-back-stimulated templated reaction[28]. Here, by probing the delins for inserts that had a reverse-complement orientation with respect to their flanking matches we indeed found *in vivo* support for 3' extension in which both the primer and the template reside on the same DSB end (S7 Fig).

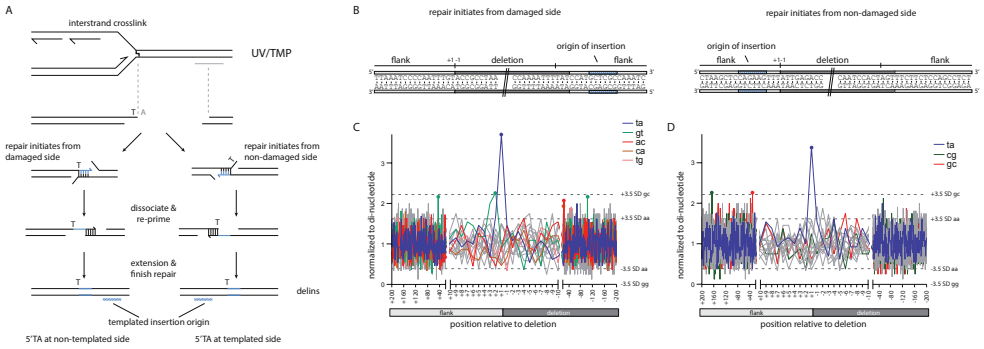


Fig 8. Mutagen-induced deletions are the product of DSB repair. A. Schematic illustration of a replication-blocking lesion that is converted to a DSB and finally results in a templated flank insertion. The 5'TA causing the deletion defines one end of the break, while the composition of the other end is unknown. By using the 5'TA together with the side of origin of templated insertions we can determine the reactivity of both 3' break ends: if the 5'TA is on the opposite side of the templated insertion origin, repair initiated from the damaged side. On the other hand if both are on the same side then repair is initiated from the non-damaged side. **B**. Examples of two delins, portrayed at the sequence level, where either the 5' side (left drawing) or the 3' side (right drawing) potentially served as a primer to initiate repair. **C**. Analysis that probes the (+1,-1) junction of the side opposite to the flank containing the insertion origin. Dashes lines represent 3.5 times the SD. Only the largest and smallest variations for individual dinucleotides are shown. Only dinucleotide sets containing at least one position (marked by dots) that is >3.5 times the SD are shown in color. **D**. As in B, but in this case the (+1,-1) junction of the side that contains the insertion origin is analysed.

DISCUSSION

In this study, we have shown that EMS and UV/TMP-induced DSBs are predominantly repaired via POLQ-mediated repair and in-depth analysis of ~7,000 unique deletion footprints allowed us to unveil important characteristics of the *in vivo* repair mechanism. We found that mutagen-induced deletions are the product of alternative DSB repair in which one end is produced by the replication machinery that approached the damage up to one nucleotide. Base pairing of the terminal nucleotide of the blocked nascent strand to single stranded DNA at the opposite break end primes POLQ to polymerize, resulting in DNA tracts that are templated by the sequence immediately flanking the DSB. Further processing of the ensuing stable joints produces simple deletions. However, in case DNA synthesis is interrupted, likely resulting from POLQ errors, a primer-template switch is induced in which the newly formed terminal nucleotides again pair in order for POLQ-mediated extension to continue. We find that one or more cycles of such templated DNA synthesis and primer-template switching can fully explain the composition of deletions that are associated with inserts.

From a conservative point of view, POLQ-mediated repair is a surprisingly elegant solution to the problem how to repair a DSB while keeping the loss of genetic information to an absolute minimum: the repair reaction does not depend on removal of nucleotides to create ligatable ends. It is thus an intriguing idea that nature, perhaps because of the polarity in DNA synthesis being in a 5′ to 3′ direction, has evolved DNA repair and recombination mechanisms that use or tolerate extensive 5′ but not 3′ end-resection; it is obvious that having both these activities prominently used inside nuclei would constitute a great threat to genomes. We have shown here for POLQ-mediated repair of DSBs that the 3′ end of a DNA molecule is very stable and acts as a nucleation site in the repair reaction. Using a specialized polymerase to extend and as such stabilize minimally paired 3′ ends, as opposed to trimming by exonucleases provides a simple yet powerful and versatile solution to a complex problem. One striking aspect of *C. elegans* POLQ is the notion of single nucleotide homology. The degree of microhomology in (POLQ-dependent and potentially POLQ-independent) alternative end-joining in a number of other biological systems, such as mouse, human and also plants appear to concern more bases, frequently 3 to 4 bp[4,10,29]. It is yet unclear whether this difference reflects species specific adaptation to the enzyme or differences in the context in which POLQ was studied: a recent *in vitro* study using purified human POLQ demonstrated pairing and extension of 3′ overhangs with just two nucleotides of homology[11]. Another perhaps more striking difference in POLQ-mediated repair between species is the composition of insertions that are found in between the break junctions. While insertions in *C. elegans* are mostly derived from a single proximal location, footprints in other species suggest that POLQ is more promiscuous, because inserts often originate from multiple locations, which is suggestive of iterative rounds of abortive repair[4,29]. It is currently unknown what is the cause of this apparent discrepancy between POLQ-mediated repair in different species, but it is of interest to note that mammalian POLQ has evolved to include three additional loop regions in the polymerase domain. One of these loops, loop2, was recently implicated in non-templated terminal transferase activity[28]. The ability to add random nucleotides to the 3′ end of a DSB-repair intermediate may help to generate more opportunity for microhomology-mediated templated resolution.

We have previously shown that POLQ is the primary pathway acting on DSBs that result from DNA replication blocking endogenous lesions[5-7]. An intriguing question concerns the size distribution of resulting deletions: as also shown here, one junction is defined by the replication fork impediment, but

what defines the other end? Genetic and molecular dissection of replication-obstructing G-quadruplex structures has led to the model where a replication-stalling DNA lesion results in a ssDNA gap downstream of the impediment[5-7]. More recently, we provided evidence supporting the idea that it is this gap that is responsible for a DSB (with ends 50 to a few hundred bps apart) when the gapped strand is replicated in the next S-phase[7]. POLQ-mediated alternative end-joining subsequently acts on these replication-associated DSBs, instead of HR, which cannot repair the break using the sister chromatid as the latter still contains the replication-blocking impediment (see [7] for details).

In this study, we demonstrate an identical genetic requirement for the repair of DSBs resulting from mutagen exposure; however, it is yet uncertain which replication-blocking lesions are causative. EMS induces a plethora of lesions[30] some of which have been shown to be potent blocks of the replicative polymerases[31], whereas UV/TMP treatment generates psoralen monoadducts on thymines and interstrand crosslinks with a great preference for thymines. Whether deletions induced by UV/TMP are the result of ICL or monoadducts is an outstanding question because the notion of preferential junction formation at 5'TA sites is not discriminatory. Although this outcome perfectly fits a scenario of replication up to the first damaged base of juxtaposed T-T ICLs, it also fits to replication blocking at monoadducts that are preferentially induced at 5'TA sites. The hypersensitivity of *C. elegans* POLQ mutant animals towards alkylating and crosslinking agents (as also observed for POLQ/Mus308 mutant *Drosophila*) may seem to contradict to an apparent lack of sensitivity in other systems, such as POLQ knockout mouse cells. We suspect this difference to primarily originate from the fact that *C. elegans* toxicity assays, especially those encompassing early embryonic cell divisions, are very sensitive to perturbations of DNA replication[12,32].

Exposure to mutagens, such as EMS and UV/TMP, is widely used to induce random mutations in a great variety of organisms other than *C. elegans*, such as *Drosophila*, *Zebrafish*, *Arabidopsis*, Tomato, and mouse. Although EMS-induced damage predominantly induces SNVs, in all these biological systems deletions have been observed ranging in size from a few base pairs to numerous kb[18,20,33-39], and it will be of great interest to investigate whether the causal involvement of POLQ-mediated repair is evolutionary conserved.

In this work, we have linked a specific type of mutations, i.e. deletions of small size, to carcinogenic mutagens that are used in clinical setting. It is becoming increasingly important to establish causal relationships between the exact type and nature of their DNA damaging agents and genome alterations, especially because of the growing interest in mutational signatures in cancer genomes. Recently, the altered genomes of cancer cells are not only inspected for potentially cancer promoting (driver) mutations but also for signatures that testify to the history of the tumour, with respect to genetic makeup and/or environmental exposure[40]. Currently, the majority of these signatures are based on single base substitutions and their surrounding DNA context, but cancer genomes are loaded with copy number variations, deletions and insertions, and also gross chromosomal rearrangements that are likely resulting from mutagenic DNA repair processes[41,42]. It will be interesting to inspect cancer genomes, especially those evolving in cancer cells that are characterized by a defect in homologous recombination for genomic scars that carry the signature of POLQ-mediated end joining, to also determine the contribution of this mutagenic pathway to tumorigenesis.

METHODS

C. elegans genetics

Standard methods and conditions for culturing *C. elegans* were used[15]. The alleles used in this study were: *polh-1*(lf31); *polq-1* (tm2026); *fcd-2* (tm1298). Bristol N2 was used as wild type in all experiments.

Nematode mutagenesis

Mutagenesis with EMS was performed at 12.5mM, 25mM, 50mM or 100mM according to standard protocols[15]. In brief, populations were synchronized by alkaline hypochlorite treatment and eggs were allowed to hatch o/n. L1 worms were plated out on 9cm NGM agar plates seeded with *E. coli* (OP50) and grown at 20 degrees. Two days later L4 worms were washed off the plates and treated for 4 hours with EMS dissolved in M9.

A similar staging protocol was used for UV/TMP mutagenesis. Subsequently, animals of the L4 stage were treated for one hour with 10µg/ml TMP (Sigma, T6137, stock: 100mg dissolved in 40ml acetone) dissolved in M9. Animals were distributed onto non-seeded NGM plates and exposed to UVA irradiation (366nm; CAMAG 29200 Universal UV LAMP) at a dose rate of 160µW/cm^2 (Blak-Ray UV-meter model no. J221), after which the animals were transferred to standard OP50/NGM plates.

Sensitivity assays

Staged animals were exposed to either EMS or UV/TMP at the L4 larval stage and per experimental condition four plates each containing three worms were started. After a 24-36-hour period of egg laying the mothers were removed. The number of (dead) eggs and hatched progeny (after 24 hours) was determined. All experiments were performed in triplicate. We determined the brood size for animals by collecting eggs from individual hermaphrodites in sequential periods of 24 hours. For each period the number of (dead) eggs and hatched progeny (after 24 hours) was determined and then added.

Deletion library PCR assay

For each deletion library ~80,000 animals were used for synchronization by hypochlorite treatment (0.5M NaOH, 2% hypochlorite) and overnight starvation. Animals of the L4 stage were treated with EMS (50mM), UV/TMP (50 J/m^2) or mock-treated. P0 animals were removed by hypochlorite treatment 24 hours post-UV/TMP-treatment, and after o/n hatching ~100,000 F1 animals were transferred to 10 9 cm plates and were grown for two days at 20 degrees. Then, animals were collected by rinsing the plates with M9 and distributed over 10 96-well plates such that each well contained ~100 worms in a 5 µl volume. To this 10 µl of lysis buffer was added and animals were subsequently subjected to a standard lysis protocol to liberate the DNA. All 10 plates were pooled into 1 master plate (using 10 µl original DNA mixture), which was used for another round of pooling by combining 10µl from each of the eight wells in a column, finally yielding one row of 12 wells for library. Prior to performing nested PCRs for eight different genomic targets (see Supplementary Table 1), the DNA was digested with the thermostable restriction enzyme PspGI. Upon detection of a smaller-than-wild-type product in the pools, PCRs were repeated on the master plate and then on individual plates. The PCR products of the samples that remained positive during this deconvolution exercise (in duplicate) were sequenced. We considered a result a false positive if

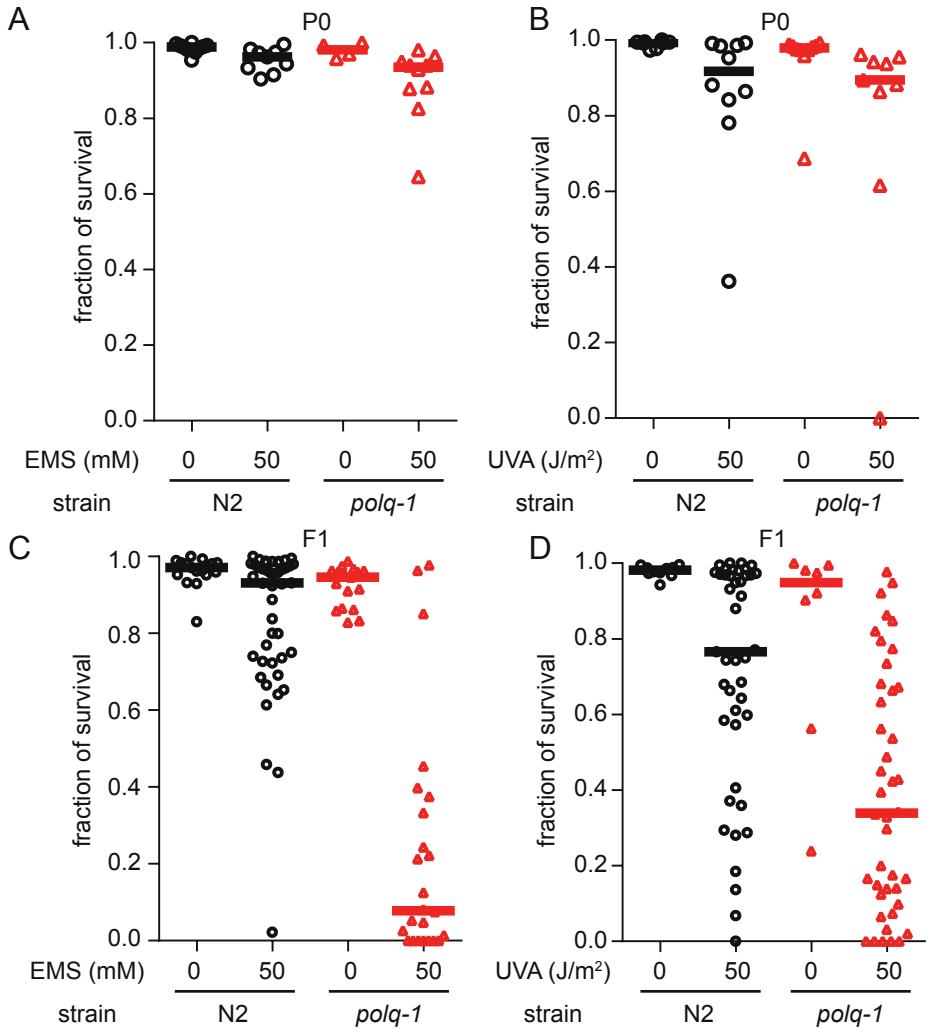
the samples of lower complexity failed to reproduce the PCR product.

Bioinformatic analyses

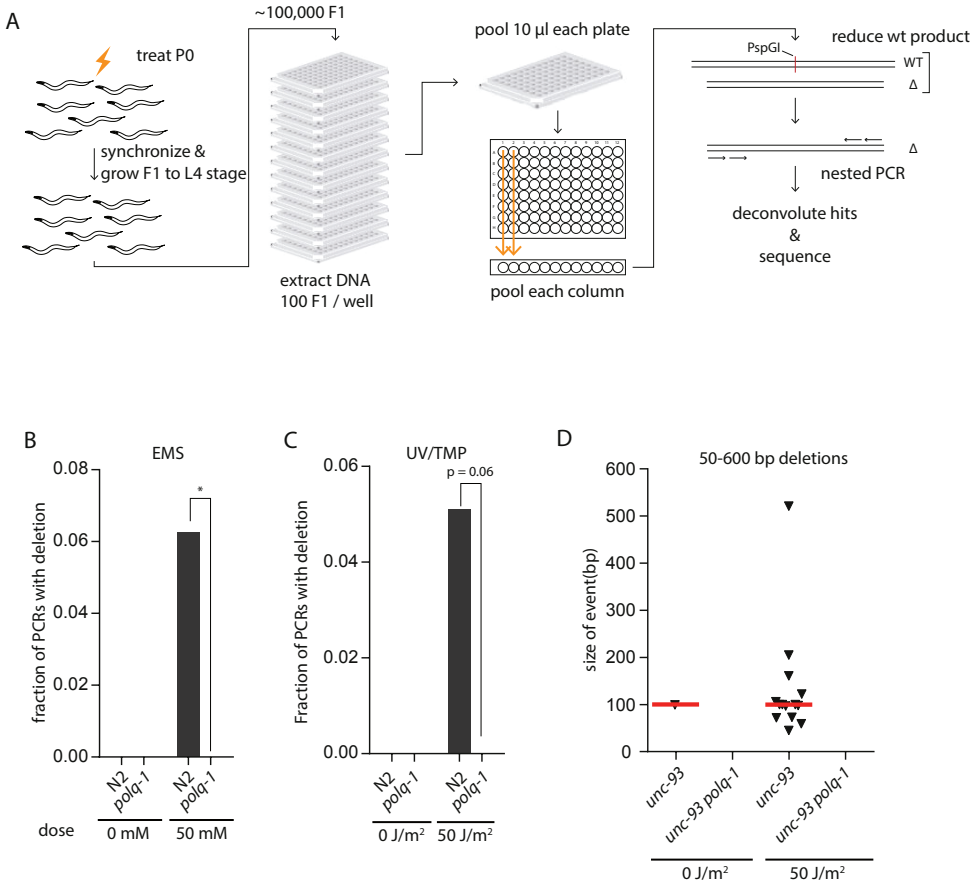
The sequence information for publically available deletion alleles was retrieved from WormBase (WS243). A custom Java program was written to analyse and annotate the WormBase alleles (available upon request). We included a number of additional stringency criteria: 1) the coordinates of the allele should match the information about the allele's left and right border sequence, 2) insertions within deletions should be as minimal as possible, 3) insertions that contained one or more Ns were discarded. In addition, for cases where sequence homology at the junctions allowed for more than one possible mapping position we placed the homology at the retained flank of the 5' side. To identify the origin of the insertions in the delins alleles we i) performed BLAST for insertions ≥15 nt, and ii) used a custom-made algorithm aimed to find the longest common substring, i.e. the longest possible match between a stretch of the insertion (≥5 nt) and the sequence that is in close proximity of the junctions (≤50 nt of each flank and 50 nt within either side of the deletion). All deletion alleles used in our analyses can be found in S3 Table.

ACKNOWLEDGEMENTS

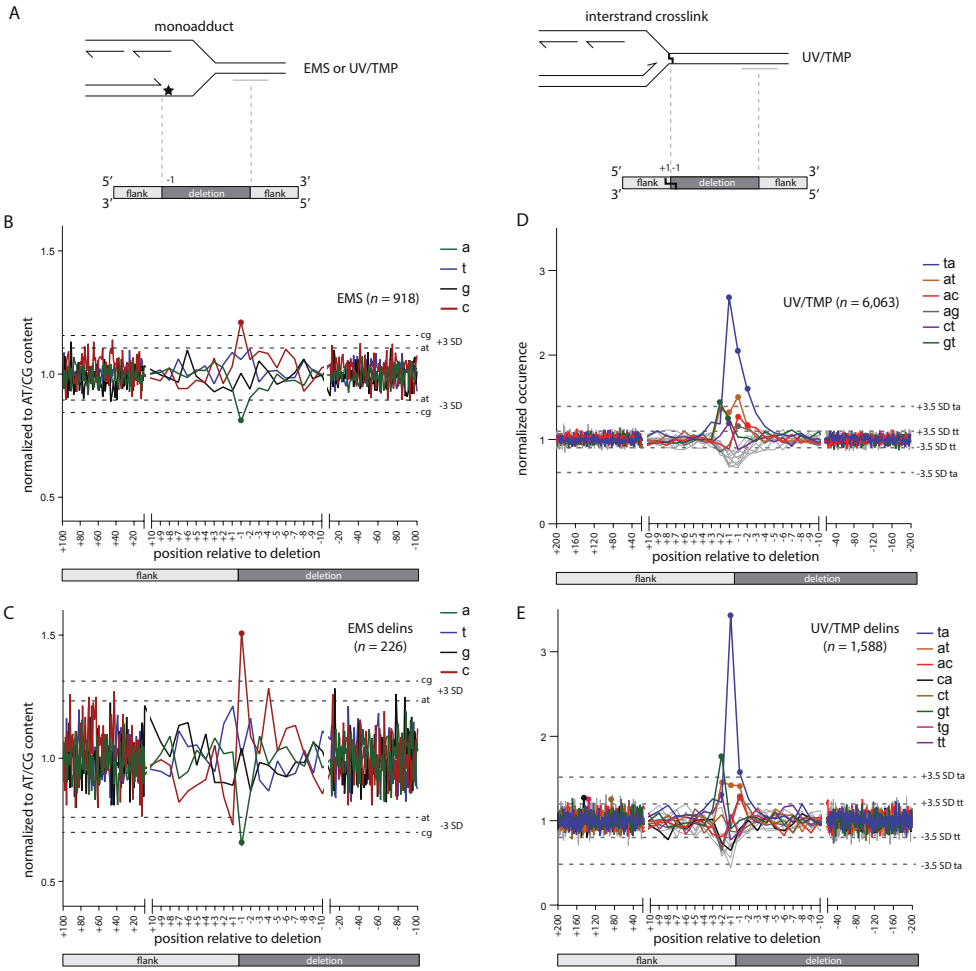
We thank the *C. elegans* Knockout Consortium, Shohei Mitani and the *C. elegans* Genetics Center for providing strains and sequence information of all deletion alleles.



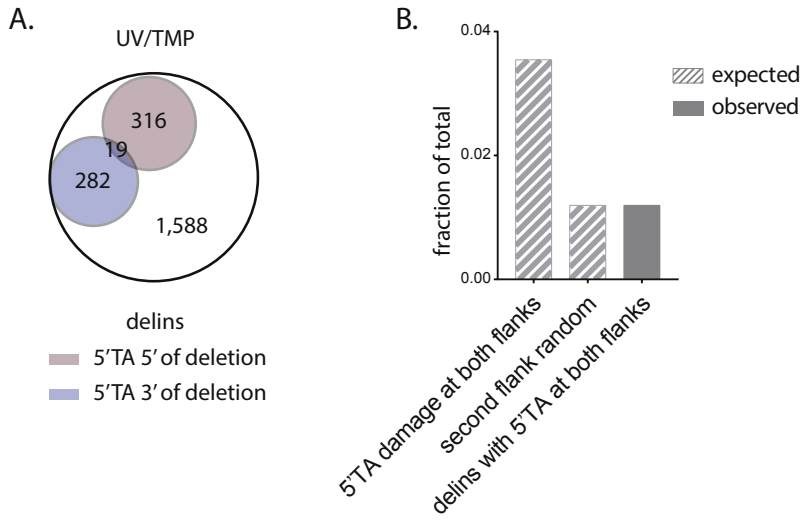
S1 Fig. Increased inheritable genetic defects in POLQ-deficient animals exposed to EMS and UV/TMP. A-B. The surviving fraction for the broods of P0 animals that were treated with either EMS (A) or UV/TMP (B) was determined. Lines represent the median for each dataset. C-D. The surviving fraction was determined for the broods of F1 animals that were born out of P0 animals treated with either EMS (C) or UV/TMP (D). Ten animals were analysed for untreated animals, while 50 treated animals were analysed for each genotype. Lines represent the median for each dataset.



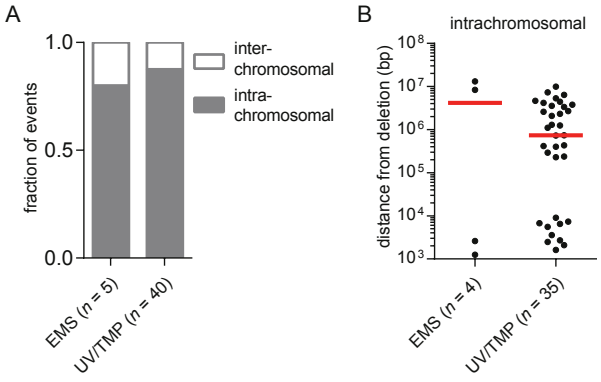
S2 Fig. Mutagen-induced deletions require POLQ. A. Schematic illustration of how mutagenesis libraries are constructed and screened for deletions at specific loci. For each genotype ~80,000 synchronized P0 L4 animals were mutagenized using EMS or UV/TMP. One day after exposure P0s were removed by hypochlorite treatment and eggs were allowed to hatch o/n in M9. In total ~100,000 F1 animals were used to generate a library in which each well of a 96-well plate contained ~100 animals. DNA of 10 plates was first pooled into one plate and subsequently all columns of this plate were pooled together to create 12 screen samples. We used a strategy based on the use of thermostable restriction enzymes to find deletion alleles [43]. This strategies employs the fact that wild-type template is digested prior to and during PCR amplification, while deletion alleles that lost the recognition site of the restriction enzyme are resistant to this digestion, leading to their preferred amplification. All initial hits were deconvoluted first by PCR of the pooled and then by PCR of the non-pooled samples. Hits that were confirmed (in duplicate) in the non-pooled samples were sequenced. **B-C**. Quantification of the screens for the indicated genotypes and conditions. Fisher's exact test was used to determine statistical significance: * represents $p < 0.05$. **D**. The size distribution of deletions that were isolated using the *unc-93* reversion assay; smaller than wild-type bands were sequenced by Sanger sequencing. Each triangle represents a deletion either in *unc-93*, *sup-9* or *sup-10*.



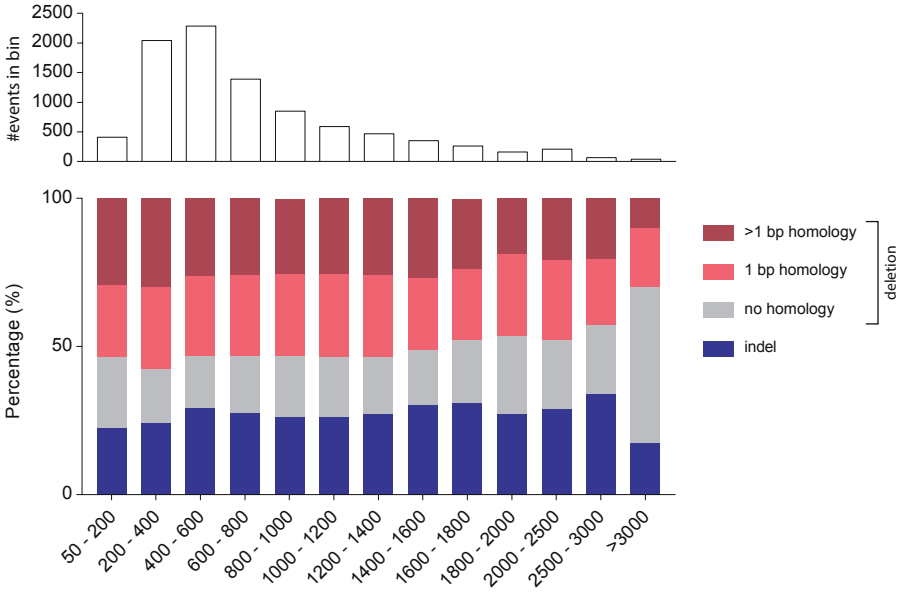
S3 Fig. Damage-specific signatures in deletion profiles. A. Schematic illustration of the concept that one junction of DNA-damage-induced deletions is defined by the nascent strand blocked at sites of DNA damage. In this hypothesis, the replication-blocking lesion may dictate position -1, being the outermost nucleotide of the lost sequence. **B**. The base composition of all breakpoints, normalized to the relative AT/CG content around the breakpoints (from +100 to -100) for EMS- induced deletion alleles. Position +100 to +1 reflects the sequence that is retained in the deletion alleles; position -1 to -100 reflects the sequence that is lost. Dashed lines represent three times the SD. Data points outside these boundaries are marked with a dot. **C**. As in B, but only for delins. **D**. The tandem base composition of all breakpoints, normalized to the relative di-nucleotide occurrence at position +200 to -200. For each indicated position (+ for retained sequence; - for lost sequence) the base composition is coupled to the composition of the immediate downstream base. Only dinucleotides that were found elevated (>3.5 times the SD) are depicted in the legend, with elevated data points marked with a dot. Only the largest and smallest variations for individual dinucleotides are shown. **E**. As in D, but only for UV/TMP-induced delins.



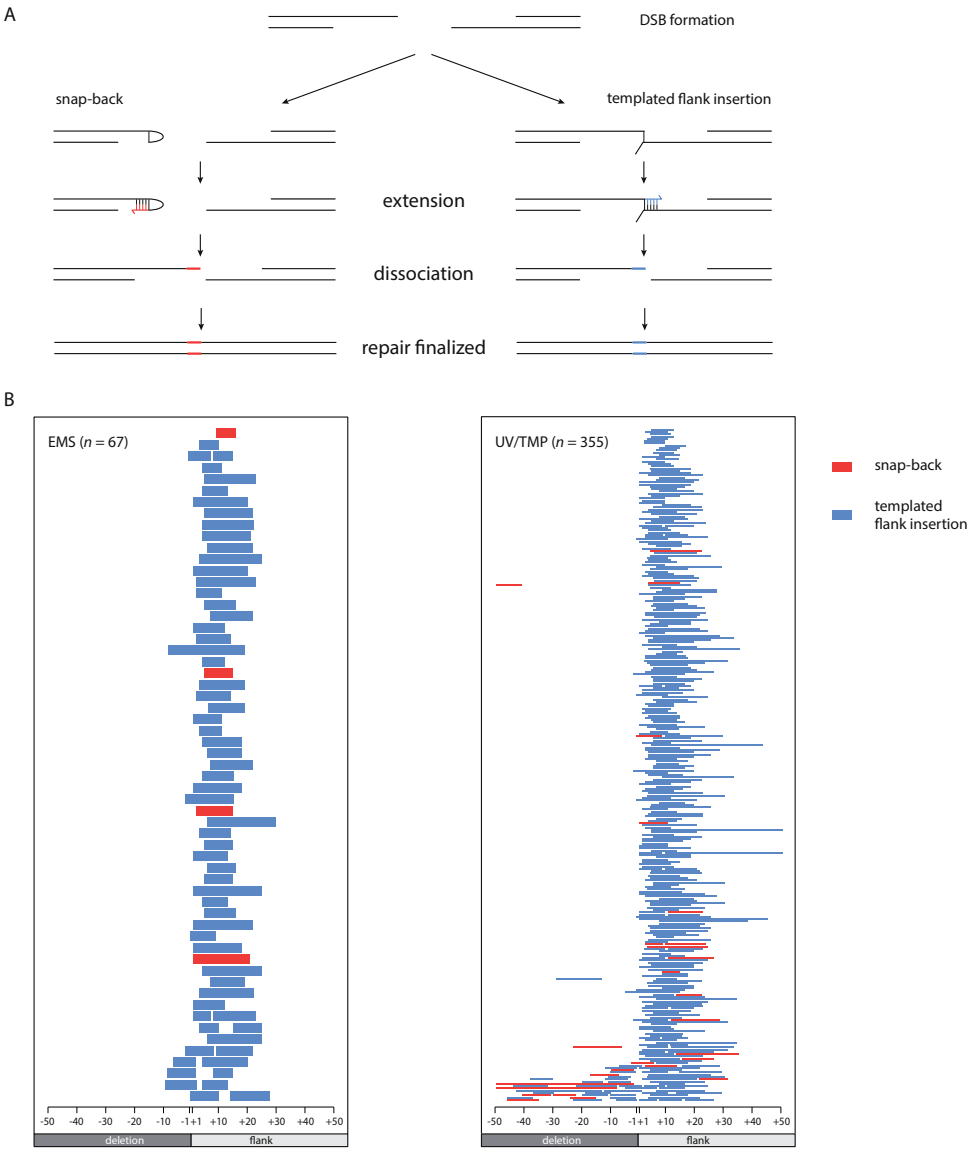
S4 Fig. Deletions result from a single replication-blocking lesion. A. A Venn-diagram representation of delins (*n*=1,588) in which the blue and red circles indicate the number of delins that have a 5'TA at position +1,-1 at the 5' or 3' side of the deletion (see Figures 3 and 5). The overlapping area represents the number of cases where a 5'TA was found at both sides of the deletion. UV/TMP-induced deletions are thus characterized by a single 5'TA at only one breakpoint. **B**. A histogram depicting the observed and the expected number of delins that are flanked by 5'TA. The expected number is calculated for two scenarios: i) the probability of finding a second 5'TA is equally overrepresented as finding a 5'TA at a given breakpoint (which would argue that a delins results from a crosslink at both breakpoints), or ii) the probability of finding a second 5'TA is equal to its probability for any given sequence (which would argue that only one crosslink underlies the genesis of a delins).



S5 Fig. The origin of insertions in EMS and UV/TMP-induced delins. A. Distribution of insertions that originate from inter- (>1,000 bp away from deletion) or intrachromosomal locations relative to the position of the delins. **B.** The distance between the positions of the delins and the location from where the templates originate is plotted (in bp) for delins with intrachromosomal inserts that do not map to the immediate vicinity.



S6 Fig. POLQ signature is diminished in large deletions. Distribution of all deletion alleles binned to size. For each bin the categories delins, no homology, 1 bp of homology and >1 bp homology are shown. The number of events in each bin is shown in the upper panel.



S7 Fig. Snap-back replication represents a minor part of delins. A) Schematic illustration of the concept that templated insertions can result from extending the 3' hydroxyl end of a DSB end using i) its flanking sequence in-cis through snapback interaction (left drawing), or ii) the other end of the DSB in trans (right drawing). Note that snap-back replication results in insertions that are of reverse-complement configuration with respect to the sequence in the flank that served as a template. **B.** Visual representation of the origins of flank insertions in both forward (blue) and reverse-complement (red) orientation. A single line represents one mapped flank insertion and is drawn relative to its cognate breakpoint with '-' for deleted and '+' for retained sequences. Only inserts where a significant part of the insert could be traced back (likelihood of finding a longest common substring of that particular size: $p < 0.05$) were represented.

REFERENCES

1 R. van Schendel, et al. Polymerase Theta is a key driver of genome evolution and of CRISPR/Cas9-mediated mutagenesis Nat. Commun. 6, 7394 (2015).
2 S. H. Chan, A. M. Yu, and M. McVey Dual roles for DNA polymerase theta in alternative end-joining repair of double-strand breaks in Drosophila PLoS. Genet. 6(7), e1001005 (2010).
3 R. Ceccaldi, et al. Homologous-recombination-deficient tumours are dependent on Poltheta-mediated repair Nature 518(7538), 258 (2015).
4 P. A. Mateos-Gomez, et al. Mammalian polymerase theta promotes alternative NHEJ and suppresses recombination Nature 518(7538), 254 (2015).
5 S. F. Roerink, R. van Schendel, and M. Tijsterman Polymerase theta-mediated end joining of replication-associated DNA breaks in C. elegans Genome Res. 24(6), 954 (2014).
6 W. Koole, et al. A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites Nat. Commun. 5, 3216 (2014).
7 B. Lemmens, R. van Schendel, and M. Tijsterman Mutagenic consequences of a single G-quadruplex demonstrate mitotic inheritance of DNA replication fork barriers Nat. Commun. 6, 8909 (2015).
8 M. McVey and S. E. Lee MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings Trends Genet. 24(11), 529 (2008).
9 R. Ceccaldi, B. Rondinelli, and A. D. D'Andrea Repair Pathway Choices and Consequences at the Double-Strand Break Trends Cell Biol. 26(1), 52 (2016).
10 M. J. Yousefzadeh, et al. Mechanism of suppression of chromosomal instability by DNA polymerase POLQ PLoS. Genet. 10(10), e1004654 (2014).
11 T. Kent, et al. Mechanism of microhomology-mediated end-joining promoted by human DNA polymerase theta Nat. Struct. Mol. Biol. (2015).
12 S. F. Roerink, et al. A broad requirement for TLS polymerases eta and kappa, and interacting sumoylation and nuclear pore proteins, in lesion bypass during C. elegans embryogenesis PLoS. Genet. 8(6), e1002800 (2012).
13 D. M. Muzzini, et al. Caenorhabditis elegans POLQ-1 and HEL-308 function in two distinct DNA interstrand cross-link repair pathways DNA Repair (Amst) 7(6), 941 (2008).
14 M. Brauchle, K. Baumer, and P. Gonczy Differential activation of the DNA replication checkpoint contributes to asynchrony of cell division in C. elegans embryos Curr. Biol. 13(10), 819 (2003).
15 S. Brenner The genetics of Caenorhabditis elegans Genetics 77(1), 71 (1974).
16 C. Coulondre and J. H. Miller Genetic studies of the lac repressor. IV. Mutagenic specificity in the lacI gene of Escherichia coli J. Mol. Biol. 117(3), 577 (1977).
17 P. Anderson Mutagenesis Methods Cell Biol. 48, 31 (1995).
18 S. Flibotte, et al. Whole-genome profiling of mutagenesis in Caenorhabditis elegans Genetics 185(2), 431 (2010).
19 A. Wei, et al. Efficient isolation of targeted Caenorhabditis elegans deletion strains using highly thermostable restriction endonucleases and PCR Nucleic Acids Res. 30(20), e110 (2002).
20 G. Jansen, et al. Reverse genetics by chemical mutagenesis in Caenorhabditis elegans Nat. Genet. 17(1), 119 (1997).
21 M. Edgley, et al. Improved detection of small deletions in complex pools of DNA Nucleic Acids Res. 30(12), e52 (2002).
22 T. W. Harris, et al. WormBase: a comprehensive resource for nematode research Nucleic Acids Res. 38(Database issue), D463-D467 (2010).
23 P. A. Muniandy, et al. DNA interstrand crosslink repair in mammalian cells: step by step Crit Rev. Biochem. Mol. Biol. 45(1), 23 (2010).
24 F. Esposito, R. G. Brankamp, and R. R. Sinden DNA sequence specificity of 4,5',8-trimethylpsoralen cross-linking. Effect of neighboring bases on cross-linking the 5'-TA dinucleotide J. Biol. Chem. 263(23), 11466 (1988).
25 V. Boyer, E. Moustacchi, and E. Sage Sequence specificity in photoreaction of various psoralen derivatives with DNA: role in biological activity Biochemistry 27(8), 3011 (1988).
26 M. E. Arana, et al. Low-fidelity DNA synthesis by human DNA polymerase theta Nucleic Acids Res. 36(11), 3847 (2008).
27 M. Seki, et al. High-efficiency bypass of DNA damage by human DNA polymerase Q EMBO J. 23(22), 4484 (2004).
28 T. Kent, et al. Polymerase theta is a robust terminal transferase that oscillates between three different mechanisms during end-joining Elife. 5 (2016).
29 N. Kleinboelting, et al. The Structural Features of Thousands of T-DNA Insertion Sites Are Consistent with a Double-Strand Break Repair-Based Insertion Mechanism Mol. Plant 8(11), 1651 (2015).
30 G. A. Sega A review of the genetic effects of ethyl methanesulfonate Mutat. Res. 134(2-3), 113 (1984).

31 B. Sedgwick Repairing DNA-methylation damage Nat. Rev. Mol. Cell Biol. 5(2), 148 (2004).
32 M. Brauchle, K. Baumer, and P. Gonczy Differential activation of the DNA replication checkpoint contributes to asynchrony of cell division in C. elegans embryos Curr. Biol. 13(10), 819 (2003).
33 E. Wienholds, et al. Target-selected inactivation of the zebrafish rag1 gene Science 297(5578), 99 (2002).
34 E. Wienholds, et al. Efficient target-selected mutagenesis in zebrafish Genome Res. 13(12), 2700 (2003).
35 E. A. Greene, et al. Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in Arabidopsis Genetics 164(2), 731 (2003).
36 K. Nairz, et al. A reverse genetic screen in Drosophila using a deletion-inducing mutagen Genome Biol. 5(10), R83 (2004).
37 J. L. Cooper, et al. Retention of induced mutations in a Drosophila reverse-genetic resource Genetics 180(1), 661 (2008).
38 O. Thompson, et al. The million mutation project: a new approach to genetics in Caenorhabditis elegans Genome Res. 23(10), 1749 (2013).
39 K. Shirasawa, et al. Genome-wide survey of artificial mutations induced by ethyl methanesulfonate and gamma rays in tomato Plant Biotechnol. J. 14(1), 51 (2016).
40 T. Helleday, S. Eshtad, and S. Nik-Zainal Mechanisms underlying mutational signatures in human cancers Nat. Rev. Genet. 15(9), 585 (2014).
41 S. Nik-Zainal, et al. Mutational processes molding the genomes of 21 breast cancers Cell 149(5), 979 (2012).
42 S. Nik-Zainal, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences Nature (2016).
43 A. Wei, et al. Efficient isolation of targeted Caenorhabditis elegans deletion strains using highly thermostable restriction endonucleases and PCR Nucleic Acids Res. 30(20), e110 (2002).

6

GENERAL DISCUSSION AND FUTURE PERSPECTIVES

This dissertation discusses several aspects of double-strand break (DSB) repair in *C. elegans*. Faithfull repair of DSBs is crucial for cells to maintain genome stability and for that reason eukaryotic cells are equipped with a variety of DSB-repair mechanisms. Apart from homologous recombination (HR), which is considered to be an error-free pathway, most of the other DSB-repair pathways are intrinsically error-prone, frequently leading to small genetic changes, but occasionally leading to gross chromosomal aberrations. Cells that are compromised in their ability to repair DSBs are more likely to undergo malignant transformation. Although all cells within a single organism are generally equipped with the same mechanisms to repair DSBs, the contribution and availability of each repair mechanism depends on cell type (e.g. germ cells versus somatic cells) and cell stage. It is especially crucial for germ and stem cells to properly deal with genetic insults as these cells give rise to progenitors.

To investigate DSB-repair pathways I made use of whole-genome sequencing approaches, which enabled me to examine the entire genome of animals that were either wild type or carried a genetic defect in one or more DNA repair mechanisms. By probing the genomes of animals that accumulated mutations we identified specific signatures, one of which leading to the identification of a previously unknown error-prone DSB-repair pathway, which depends on the A-family polymerase Theta (POLQ). In essence, POLQ attempts to connect two DNA ends by using single base-pair of homology between the ends from which POLQ can extend. This frequently results in the repair of the break and the deletion of a small piece of genetic information. Occasionally, however, during extension the two DNA ends dissociate and the process of connecting and extension by POLQ is repeated until the break is repaired. The repeated action of POLQ leaves behind a smoking-gun for POLQ-mediated repair: a small piece of newly synthesized DNA, which is a carbon copy of part of the DSB flank is inserted between the two broken ends.

Although this thesis provides detailed mechanistic insight into how POLQ-mediated end-joining repairs a break *in vivo*, many questions still remain unanswered. Especially little is currently known about the spatial and temporal regulation of this pathway as well as the context in which this pathway operates. A selection of outstanding questions will be discussed in the following sections.

How is Polymerase Theta-mediated repair orchestrated?

Our laboratory has shown that POLQ plays an important role in maintaining genome stability, but it remains unknown how POLQ is recruited to sites of damage. The primary DNA-damage sensors ATM and ATR are conserved in *C. elegans* but it is currently an outstanding question whether the downstream targets of these signalling kinases are conserved as well. Although both ATM and ATR share many downstream targets, at least in higher eukaryotes, they respond to different types of damages. ATR primarily responds to stalled replication forks lesions, while ATM is activated by DSBs[1]. Mice with defects in both ATM and POLQ exhibit a more severe phenotype than either deficiency alone, suggesting that POLQ and ATM do not act in the same pathway[2]. Unfortunately, *C. elegans* ATR is an essential protein making it impossible to genetically address its involvement in POLQ-mediated repair.

Some of our data indicate that POLQ acts at replication-associated DSBs (Chapter 3 and [3]) as the absence of TLS polymerases pol eta and kappa as well as the helicase FANCJ result in a distinct class of deletions that for their formation depend on POLQ. One possibility would therefore be that POLQ is recruited to DSBs by factors involved in replication. However, both transposon and

CRISPR\Cas9-induced breaks, which are thought to form independent of replication also require POLQ activity for their repair (Chapter 4). This suggests that POLQ can be recruited to DSB outside the context of DNA replication. A candidate for this function is the ssDNA binding protein RPA that coats ssDNA of resected DSBs to protect it from degradation. Notably, RPA coats ssDNA in- and outside the context of replication, which fits with POLQ recruitment for replication-associated and replication-uncoupled breaks.

Some proteins have already been implicated in alternative end-joining (altEJ), an ill-defined category of DSB-repair pathways that includes POLQ-mediated repair. PARP1 is implicated in altEJ and is rapidly recruited to DSBs. In mammalian systems PARP1 was shown to act upstream of POLQ, though it is yet unclear whether PARP1 recruits POLQ directly or indirectly and in which context[4,5]. Surprisingly, our preliminary data in *C. elegans* suggest that animals deficient for *parp-1* do not show a DSB-repair defect and are still proficient in POLQ-mediated repair, arguing that in *C. elegans*, POLQ action does not depend on PARP.

A second question that is currently unanswered is which factor(s) are responsible for the finalization of repair in POLQ-mediated repair? The current model for POLQ-mediated repair requires a ligation step. A likely candidate is LIG3, also because this protein has previously been implicated in altEJ. In mice, LIG3$^{-/-}$ cells could be created but only when LIGI or LIG3 was targeted to the mitochondria. It was subsequently found that the frequency of altEJ-mediated DNA translocations was reduced in a nuclear LIG3-deficient mouse backgrounds when breaks were induced by a zinc-finger endonuclease, implicating LIG3 in altEJ[6]. No mutant allele of *C. elegans* LIG-3 (K07C5.3) is currently available, but I have recently used CRISPR\Cas9-induced mutagenesis to create one, which is currently being investigated for POLQ-mediated repair phenotypes.

Instead of using a candidate approach to identify factors that are involved in POLQ-mediated repair we can perform unbiased screens. The classical approach in *C. elegans* is to carry out a forward genetic screen combined with a phenotypical read out to identify mutants of interest. A pilot EMS screen was performed that identified two new alleles of POLQ but thus far no novel factors. Because this was a very small-scale being far from saturated I suggest increasing scale.

An alternative approach is to use a biochemical approach: immunoprecipitation (IP) of POLQ followed by mass spectrometry to identify proteins that co-precipitate, indicating a direct or indirect interaction with POLQ. For years it has been technically extremely challenging to endogenously tag proteins in *C. elegans*, but CRISPR\Cas9 technology made it feasible to tag proteins with for example GFP or FLAG, thus enabling us to IP POLQ. The latter approach would also allow the identification of essential genes that would be missed in forward genetics screens.

Which parameters determine the deletion size in Polymerase Theta-mediated repair?

One of the most enigmatic questions that thus far remains unanswered is what determines the deletion size in POLQ-mediated repair events? The heritable genomic changes seen after repair of transposition and CRISPR\Cas-9 breaks are typically <20 bp, while for replication-associated deletions they are 50 – 300 bp, sometimes larger, but almost never smaller. Can the difference between repair outcomes of direct breaks (e.g. via CRISPR\Cas-9 or transposition) and replication-associated breaks simply be explained by the context in which the break occurs? Moreover, we found subtle but clear differences between the deletion-size distribution of TLS-deficient and FANCJ-deficient animals: intriguingly, when we compare both distributions we find a median

deletion size of 110 and 138 bp respectively (n > 90 for both sets)[3,7]. This difference is most probably explained by the fact that a G-quadruplex motif, being 20-25 bases on average, is the replication-blocking obstacle in FANCJ-deficient animals, while a single damaged base blocks the fork in TLS-deficient animals. This notion argues that the context of the replication fork impediment is of direct influence to the resulting genomic change and is thus a factor of relevance in thinking about the mechanism.

At present, we do not know whether a G4-structure is more likely to occur in the leading or lagging strand. Our data demonstrates that replication can approach a replication block (e.g. a G4-structure or a psoralen cross-link) to within a few nucleotides ([3], Chapter 3 and 5), and as such determines one deletion breakpoint. But what determines the other breakpoint, and thus the size of the deletion? If the lesion is present in the lagging strand, the other breakpoint may be determined by the previous Okazaki fragment. Okazaki fragments are deposited at ~300 bp intervals[8] which would fit the ~50-300 bp deletion size distribution, about half an Okazaki fragment. On the other hand, if the replication block occurs on the leading strand we foresee two options that can lead to a deletion: re-priming of the leading strand behind the replication blocking lesion or the approach of a converging replication fork. PrimPol, a protein that contains both TLS and primase activity, has been shown to be able to bypass replication blocking lesions either by employing its TLS activity or by re-priming downstream of the blocking lesion[9]. *C. elegans* does not contain a homolog of human PrimPol which makes a jump-over model by re-priming the leading strand downstream of the replication block less likely. In a converging replication fork model a ssDNA gap results of a size that is dependent on how close a converging fork can approach an arrested fork. In the next cell cycle such a ssDNA gap can be converted into a DSB. We have recently provided strong experimental evidence for this scenario[10]. To demonstrate that Okazaki fragments are of relevance in deletion formation we need to perturb Okazaki fragment deposition. To address the question whether G4 structures are predominantly forming in leading or lagging strands, we require information on origins of replication.

Is the role of Polymerase Theta conserved in higher eukaryotes?

To understand DSB-repair in model organisms such as *C. elegans* is not our primary goal. POLQ is conserved in mouse and human, but only recently it became evident that the role of POLQ in DSB-repair is also functionally conserved[4,5,11-13]. It is thus of great interest to translate the findings observed in model systems to humans. Our laboratory found that most deletions that occur in *C. elegans* germ cells are brought about by the activity of POLQ. Sequencing of natural isolates of *C. elegans* have allowed us to examine genome diversification and to discover that genomic changes >1 bp are carrying the hallmarks of POLQ-mediated repair. This specific mutation profile was recapitulated in a small-scale evolution experiment where POLQ-deficient and proficient animals were grown in parallel for ~250 generations. The mutational spectrum observed in POLQ-proficient animals was nearly identical to the spectra observed in natural isolates, but was, however, completely altered in POLQ-deficient animals. From this we concluded that POLQ plays a major role in the genome diversification of *C. elegans*. It will now be of interest to address the contribution of POLQ-mediated repair in genome variations in mammals, either in germ cells leading to genetic variation or in somatic cells leading to cancer.

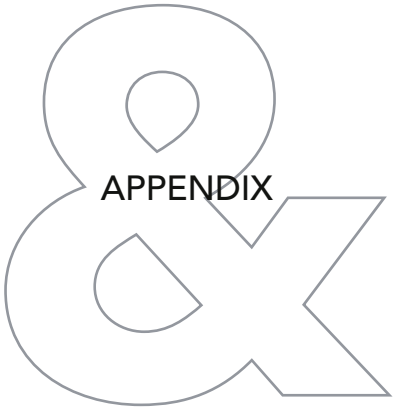
It is currently unknown why NHEJ does not act on breaks in *C. elegans* germ cells, while it appears to be functional in these cells[14,15]. Both studies show that NHEJ is actively suppressed to

prevent illegitimate repair between chromosomes during meiosis. Given that error-prone repair in germ cells of *C. elegans* almost exclusively rely on POLQ for the repair of DSBs it is of great interest to investigate whether germ cells of higher eukaryotes equally depend on POLQ. Several studies have already identified altered expression profiles for key DSB-repair proteins in germ cells of mice as well as germline mutations that hint towards repair activity by POLQ[16-18]. In the soma the situation is quite different as NHEJ is the dominant pathway to repair spontaneous DSBs that are replication-uncoupled, both in mammals and *C. elegans*[19,20]. It appears that in this context repair by POLQ is rather an alternative to NHEJ and HR as POLQ events can generally only be detected in the absence of one of these DSB-repair pathways. Interestingly, tumours that are HR-deficient rely on POLQ for their survival and knockdown of POLQ in HR-proficient cells upregulates HR activity indicating that they can act on similar substrates. POLQ is therefore considered to be an attractive novel druggable candidate target for cancer therapy[5].

In Chapter 5 we described the *in vivo* mechanism and identified several hallmarks of POLQ-mediated DSB repair. Especially templated flank insertions, where a small piece of DNA identical to nearby sequences is found inserted into a candidate DSB site is a smoking-gun for POLQ-mediated repair. It will therefore be of great interest to query datasets (*e.g.* human tumour datasets and/or *de novo* mutations) for POLQ signatures. A number of reports already anecdotally describe the presence of small insertions that resemble the immediate flank[21-23]. Human dataset generally consists of a mixture of mutational signatures generated by several repair pathways[24]. Dissecting the contribution of each mutational process, including POLQ-mediated repair, will be an interesting challenge for the years to come.

REFERENCES

1 A. M. Weber and A. J. Ryan ATM and ATR as therapeutic targets in cancer *Pharmacol. Ther.* **149**, 124 (2015).
2 N. Shima, R. J. Munroe, and J. C. Schimenti The mouse genomic instability mutation chaos1 is an allele of Polq that exhibits genetic interaction with Atm *Mol. Cell Biol.* **24**(23), 10381 (2004).
3 W. Koole, *et al.* A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites *Nat. Commun.* **5**, 3216 (2014).
4 P. A. Mateos-Gomez, *et al.* Mammalian polymerase theta promotes alternative NHEJ and suppresses recombination *Nature* **518**(7538), 254 (2015).
5 R. Ceccaldi, *et al.* Homologous-recombination-deficient tumours are dependent on Poltheta-mediated repair *Nature* **518**(7538), 258 (2015).
6 D. Simsek, *et al.* DNA ligase III promotes alternative nonhomologous end-joining during chromosomal translocation formation *PLoS. Genet.* **7**(6), e1002080 (2011).
7 S. F. Roerink, Schendel R. van, and M. Tijsterman Polymerase theta-mediated end joining of replication-associated DNA breaks in C. elegans *Genome Res.* **24**(6), 954 (2014).
8 G. Abdurashidova, *et al.* Start sites of bidirectional DNA synthesis at the human lamin B2 origin *Science* **287**(5460), 2023 (2000).
9 S. Garcia-Gomez, *et al.* PrimPol, an archaic primase/polymerase operating in human cells *Mol. Cell* **52**(4), 541 (2013).
10 B. Lemmens, Schendel R. van, and M. Tijsterman Mutagenic consequences of a single G-quadruplex demonstrate mitotic inheritance of DNA replication fork barriers *Nat. Commun.* **6**, 8909 (2015).
11 T. Kent, *et al.* Mechanism of microhomology-mediated end-joining promoted by human DNA polymerase theta *Nat. Struct. Mol. Biol.* **22**(3), 230 (2015).
12 K. E. Zahn, *et al.* Human DNA polymerase theta grasps the primer terminus to mediate DNA repair *Nat. Struct. Mol. Biol.* **22**(4), 304 (2015).
13 M. J. Yousefzadeh, *et al.* Mechanism of suppression of chromosomal instability by DNA polymerase POLQ *PLoS. Genet.* **10**(10), e1004654 (2014).
14 A. Adamo, *et al.* Preventing nonhomologous end joining suppresses DNA repair defects of Fanconi anemia *Mol. Cell* **39**(1), 25 (2010).
15 B. B. Lemmens, N. M. Johnson, and M. Tijsterman COM-1 promotes homologous recombination during Caenorhabditis elegans meiosis by antagonizing Ku-mediated non-homologous end joining *PLoS. Genet.* **9**(2), e1003276 (2013).
16 C. M. Carvalho, *et al.* Replicative mechanisms for CNV formation are error prone *Nat. Genet.* **45**(11), 1319 (2013).
17 G. Hamer, *et al.* Function of DNA-protein kinase catalytic subunit during the early meiotic prophase without Ku70 and Ku86 *Biol. Reprod.* **68**(3), 717 (2003).
18 M. J. Ashwood-Smith and R. G. Edwards DNA repair by oocytes *Mol. Hum. Reprod.* **2**(1), 46 (1996).
19 A. J. Hartlerode and R. Scully Mechanisms of double-strand break repair in somatic mammalian cells *Biochem. J.* **423**(2), 157 (2009).
20 D. B. Pontier and M. Tijsterman A robust network of double-strand break repair pathways governs genome integrity during C. elegans development *Curr. Biol.* **19**(16), 1384 (2009).
21 J. M. Kidd, *et al.* A human genome structural variation sequencing resource reveals insights into mutational mechanisms *Cell* **143**(5), 837 (2010).
22 D. F. Conrad, *et al.* Mutation spectrum revealed by breakpoint sequencing of human germline CNVs *Nat. Genet.* **42**(5), 385 (2010).
23 C. M. Carvalho, *et al.* Replicative mechanisms for CNV formation are error prone *Nat. Genet.* **45**(11), 1319 (2013).
24 T. Helleday, S. Eshtad, and S. Nik-Zainal Mechanisms underlying mutational signatures in human cancers *Nat. Rev. Genet.* **15**(9), 585 (2014).


APPENDIX

Summary

DNA is arguably the most important molecule found in any organism, as it contains all information to perform cellular functions and enables continuity of species. It is continuously exposed to DNA-damaging agents both from endogenous and exogenous sources. To protect DNA against these sources of DNA damage various DNA repair mechanisms have evolved. If not properly repaired, DNA damage can lead to mutations that may eventually lead to cell-death or tumorigenesis. One of the most dangerous types of DNA damage is a DNA double-stranded break (DSB), in which a DNA molecule is broken into two pieces. Cells are equipped with several DSB-repair mechanisms to deal with this type of damage. Some of these mechanisms repair DSBs in an error-free fashion, while others are inherently error-prone and can lead to the accumulation of mutations. Although accumulating many mutations in cells can lead to severely reduced cellular fitness, perfect DNA repair is less desirable in the long term as mutations allow for speciation and evolution to take place.

The key question addressed in my thesis is which DSB repair mechanisms organisms use to protect their genome against DSBs. We tried to answer this complex question by whole-genome approaches as this allows us to examine the entire genome in an unbiased way. Most of the work I present here has been performed in a small nematode species: *C. elegans*, which is a 1mm long worm that lives in soil. Many of the DNA repair mechanisms found in vertebrates are also found in this small animal, which makes it an excellent model organism to study and to genetically dissect the contribution of various DNA repair mechanisms to genome stability.

In **Chapter 1** I introduce DNA repair mechanisms and next-generation sequencing approaches that I have used during my research. Then I will introduce the model organism *C. elegans* and finally the central research question of this thesis will be introduced.

In **Chapter 2** I investigate genomes of related nematode or fly species for genomic alterations that occurred during evolution. More specifically, I investigate the genomic loss of introns: non-coding DNA sequences that in eukaryotes interrupt protein-coding exons and are removed prior to translation. Intron loss was found to be highly correlated with sequence homology at the borders, suggesting the involvement of a DSB repair mechanism that uses microhomology to repair spontaneous DNA breaks within an intron.

In **Chapter 3** I present our findings on the contribution of translesion synthesis (TLS) to genome stability. TLS is a damage avoidance mechanism that allows replication to continue past damaged nucleotides, often by incorporating of a wrong nucleotide opposite the damage. We noticed that during culturing of animals that were defective for the TLS polymerase pol eta and pol kappa, animals with apparent phenotypes arose in these populations while they were absent in populations lacking either single TLS polymerase. By next-generation sequencing of propagated populations of these double mutants we uncovered a very narrow mutational spectrum of deletions that were between 50 and 300 base pairs (bp) in size. By thorough analysis of a subset of deletions that were accompanied by insertions that originated from the flanks we inferred the involvement of another DNA polymerase. Genetic dissection of this DSB-repair pathway led to the involvement of Polymerase Theta (POLQ), an A-family polymerase which is required to create these 50-300 bp deletions. In POLQ-deficient animals these small deletions are completely absent and, instead, only large deletions spanning thousands of nucleotides are detected.

The hallmark of POLQ-mediated repair of DSBs is the creation of templated insertions. Surprisingly, when we investigated whole-genome data from a few natural isolates of *C. elegans*,

we noticed that this hallmark was also frequently present in small (<50 bp) events, which were completely absent in the mutational spectrum of TLS mutants. To investigate this further we made use of assays that directly induce DSBs in the germ-line of *C. elegans*. In **Chapter 4** I investigate the contribution of POLQ in repairing direct DSBs, either via the recently developed CRISPR/Cas9 system or via transposition, in which a mobile DNA element releases itself from the DNA leaving behind a DSB that is repaired by the cell's machinery. In both assays we observed that DSB repair depends on POLQ and in its absence a completely different mutational spectrum was observed at the DSB sites. Additionally, we show in a small-scale evolution experiment that propagated worm populations that are either proficient (i.e. wild-type) or deficient for POLQ have a completely different deletion spectrum genome-wide. Whereas wild-type animals exclusively show small deletions (median of ~7 bp), POLQ-deficient animals only show very large deletions (median of ~15,000 bp), arguing that POLQ is necessary for the protection of genome integrity at the expense of small mutations instead of catastrophic extensive deletions. Additionally, analysis of tens of sequenced genomes of natural isolates of *C. elegans* predominantly showed POLQ-like footprints, suggesting that POLQ is a key driver of nematode evolution.

Finally, in **Chapter 5** I investigate the molecular mechanism of POLQ-mediated repair of DSBs *in vivo*. I show that POLQ-deficient animals are hypersensitive to the commonly used mutagens ethyl methanesulfonate (EMS) and photoactivatable trimethylpsoralen (UV/TMP). Furthermore, I show that mutagen-induced deletions in wild-type worms are the result of POLQ-mediated repair. Protocols used for inducing and detecting genomic deletions has been used in *C. elegans* research for over four decades and has resulted in ~10,000 deletion alleles that I show to be induced by POLQ-mediated repair. By in-depth bioinformatic analysis of this public available dataset I dissect the mechanism by which the POLQ polymerase repairs mutagen-induced DNA breaks. The data indicates that single nucleotide homology between two break ends is sufficient for POLQ to initiate repair. The extension process is occasionally interrupted and dissociation of the break ends occurs, triggering additional rounds of priming and extension until the break is sealed. In addition to a deletion, this results in an insertion (delins) that is copied from the immediate deletion flank.

Nederlandse Samenvatting

Dit proefschrift bevat een aantal studies die we hebben uitgevoerd om inzicht te krijgen in hoe organismen hun erfelijk materiaal, het DNA, beschermen tegen invloeden van buiten en binnen de cel. Ik heb met name onderzoek gedaan naar DNA herstelmechanismen die in werking treden zodra er een DNA dubbelstrengsbreuk optreedt. Om dit te kunnen onderzoeken heb ik gebruik gemaakt van een model organisme genaamd *Caenorhabditis elegans*. Omdat velen die dit proefschrift lezen wellicht niet bekend zijn met deze materie, zowel met de rondworm *C. elegans* als met DNA dubbelstrengsbreuken, volgt nu een korte introductie.

DNA in het kort

Ieder mens is ooit begonnen als een enkele bevruchte eicel. Een opeenstapeling van celdelingen zorgt er uiteindelijk voor dat elk mens uit ongeveer 37.000 miljard cellen bestaat. Dat zijn enorm veel cellen en om een beeld te vormen van de hoeveelheid kunnen we deze cellen achter elkaar leggen en dan vormt er een rij cellen die naar de maan en bijna terug reikt. Deze 37.000 miljard cellen zijn allemaal kopieën van één enkele bevruchte eicel. Zo worden tijdens de celcyclus alle onderdelen (organellen) van een cel verdubbeld en bij de uiteindelijk splitsing van de cel verdeeld over beide dochtercellen. De celkern (nucleus) is een organel dat binnen een cel weer een afgezonderde ruimte vormt waar het erfelijk materiaal, het DNA, zich bevindt. DNA bestaat uit twee lange strengen van nucleotiden die samen de bekende DNA dubbele helix vormen. Nucleotiden zijn de bouwstenen van DNA en deze kunnen vier verschillende soorten basen bevatten: A(denine), T(hymine), G(uanine) en C(ytosine). In het DNA paart A altijd met T en C met G. Een heel lang DNA molecuul dat verpakt is tot een compacte structuur heet een chromosoom. Van elk chromosoom heb je twee kopieën: een van je moeder en een van je vader. In totaal hebben mensen 23 chromosoomparen, maar dit varieert tussen verschillende soorten. Op elk chromosoom liggen verschillende genen, coderende gebieden DNA waarin staat hoe een eiwit gemaakt moet worden. Een menselijk genoom bevat ongeveer 20.000 eiwit-coderende genen en van elk gen heb je dus twee kopieën (een van je vader en een van je moeder). Tijdens een celcyclus verdubbelt de cel het complete DNA in een proces dat replicatie heet. Tijdens de replicatie wordt de dubbele helix van het DNA als het ware opengeritst en bouwen speciale eiwitten genaamd polymerases nieuwe nucleotiden in tegenover de bestaande. Het polymerase bouwt een A tegenover een T in, een G tegenover een C, enz. Doordat dit voor beide strengen gebeurd bestaat elk DNA molecuul aan het einde van de replicatie voor de helft uit bestaand DNA en voor de helft uit nieuw gerepliceerd DNA. Het is dus belangrijk dat polymerases uitermate nauwkeurig zijn in het inbouwen van nucleotide (een 0,001% foutmarge resulteert bijvoorbeeld al in 12.000 foute nucleotiden). Een interessant gegeven is overigens dat om van 1 naar 37.000 miljard cellen te gaan er maar ongeveer 45 replicatieronden nodig zijn. De kracht van verdubbelen is dat het steeds sneller gaat, denk maar aan de reeks: 1 – 2 – 4 – 8 – 16 – 32 – 64 – 128 – 256, etc. (2^{45} = ~ 35.000 miljard).

Mutaties in het DNA

Door allerlei invloeden van buiten (bijv. zonlicht, kosmische straling, tabaksrook) en binnen de cel (oxidatie van basen, metabole processen) beschadigt het DNA continu en zijn er herstelmechanismen nodig die de integriteit van het DNA kunnen waarborgen door het te repareren. Een misvatting is dat DNA schade altijd leidt tot mutaties. In veruit de meeste gevallen is een cel prima in staat de schade weg te halen, zonder verlies van informatie. Als een T bijvoorbeeld beschadigd is kan deze

uit de ruggengraat van het DNA worden gesneden en omdat er een onbeschadigde A tegenover staat kan de cel dit 'lezen' en vervolgens een onbeschadigde T inbouwen. In sommige gevallen lukt het echter niet om de beschadiging te repareren en treedt er een mutatie op. Mutaties kunnen veroorzaakt worden door beschadigt DNA, maar kunnen bijvoorbeeld ook optreden tijdens DNA replicatie (bijv. door het inbouwen van een foutief nucleotide). Over het algemeen zijn mutaties geen direct probleem voor de cel, omdat het merendeel van het DNA bestaat uit niet-coderend DNA. Mutaties leiden dus zelden tot foutieve eiwitten. Daarnaast worden in cellen die bijvoorbeeld afwijkend gedrag vertonen als gevolg van een of meerdere mutaties een zelfvernietigingsprogamma geactiveerd. Echter kunnen er in sommige gevallen mutaties optreden die leiden tot de ontwikkeling van kanker. In het geval van kanker heeft een cel een ongelukkige combinatie van mutaties opgelopen die haar in staat stelt zich ongeremd te delen, hetgeen leidt tot een wildgroei van weefsel: een tumor. Tegenwoordig is het mogelijk om alle mutaties in tumoren uit te lezen en dit heeft ons veel inzicht verschaft over de DNA herstelmechanismen in menselijke cellen en de oorzaken van kanker.

&

Breuken in het DNA

Een van de ernstigste DNA beschadigingen is een dubbelstrengsbreuk. De naam zegt het al: beide strengen in de DNA helix zijn gebroken waardoor het DNA nu uit twee stukken bestaat. Dit is een ernstig probleem en leidt tot de activatie van allerlei signalering in de cel die vervolgens weer leidt tot de activatie van DNA dubbelstrengsbreuk herstelmechanismen. Er bestaan een aantal van deze mechanismen: Homologous recombination (HR), non-homologous end-joining (NHEJ) en alternative end-joining. HR is een proces dat gebruikt maakt van de identieke DNA kopie die tijdens DNA replicatie gemaakt is. Een van de breukeinden kopieert een deel van de kopie en komt weer los. Dan worden de breukeinden aan elkaar vast gemaakt en de breuk hersteld. Dit wordt ook wel foutloos herstel genoemd. NHEJ zet beide breukeinden aan elkaar vast zonder oog voor mogelijk verlies van DNA. Dit resulteert meestal in een DNA mutatie doordat er een paar nucleotiden verloren gaan (deletie) of soms extra nucleotiden worden ingebouwd (insertie). Alternatieve end-joining is ooit gevonden in de afwezigheid van NHEJ. Daardoor draag het de naam 'alternatief' alhoewel uit dit proefschrift blijkt dat het helemaal geen alternatief hoeft te zijn, maar misschien wel de belangrijkste route waarlangs DNA breuken hersteld worden. Anders dan HR en NHEJ bestaat alternative end-joining niet uit een enkel proces, maar is het een verzameling van een aantal reparatieprocessen. Alternatieve end-joining processen laten zich veelal kenmerken door het gebruik van homologe sequenties tijdens de reparatie en is, net als NHEJ, een mutageen proces. HR is namelijk niet altijd beschikbaar (er moet een DNA kopie aanwezig zijn) en dan zijn NHEJ en alternatieve end-joining de enige optie om de schade te herstellen. De beschikbaarheid van verschillende DSB reparatieroutes is afhankelijk van het cel stadium waarin de cel zich bevindt en bijvoorbeeld ook van het celtype. Zo weten we dat de capaciteit om breuken te repareren in somatische cellen (cellen die niet bijdragen aan het doorgeven van erfelijk materiaal aan de volgende generatie) anders is dan in kiemcellen.

C. elegans als modelorganisme

Tot zover is alles vanuit het menselijk perspectief geschreven, dus waarom wordt de rondworm *C. elegans* gebruikt in dit proefschrift? Het antwoord hierop is dat mensen en wormen niet zoveel van elkaar verschillen. Dat wil zeggen met betrekking tot DNA reparatie mechanismen. Het blijkt dat in vele soorten die qua uiterlijk in niets op elkaar lijken dat fundamentele mechanismen zoals

DNA herstelmechanismen wel geconserveerd is tijdens de evolutie. Van plant tot gist tot mens tot worm. Allemaal bevatten ze vrijwel hetzelfde arsenaal aan mechanismen om het erfelijk materiaal te beschermen. Dat doet vermoeden dat deze mechanismen al bestonden in een oercel waaruit later allerlei multicellulaire organismen, zoals de mens en *C. elegans* zijn geëvolueerd. *C. elegans* is een hermafrodiet, een dier dat zowel mannelijk als vrouwelijk is en zichzelf kan voortplanten. Binnen vijf dagen heeft een enkele worm 300 "kinderen" gekregen, die genetisch gezien vrijwel hetzelfde zijn. Soms worden er wel mannetjes geboren, die anders dan bij mensen, geen XY chromosomen bevatten, maar een enkel X chromosoom (X0) in plaats van twee (XX). Deze mannetjes kunnen wel paren met een hermafrodiet en kan op deze manier zijn erfelijk materiaal doorgeven. Wij onderzoekers maken hier gebruik van om vervolgens mutanten te combineren. Ik kan bijvoorbeeld een mannetje met mutatie A met een hermafrodiet met mutatie B kruisen om nageslacht met zowel mutatie A als B te maken). Daarnaast kunnen we het genoom van *C. elegans* wijzigen door de worm te injecteren met DNA of door de worm in mutagene stoffen te laten groeien. Ook kunnen we ontwikkeling van een eicel naar een larve goed observeren en eventuele afwijkingen vinden in het voorplantingsorgaan van deze worm. Ook is *C. elegans* het eerste multicellulaire diertje waarvan het genoom volledig uitgelezen is. Dit alles maakt de worm tot een zeer nuttig en relevant modelorganisme.

Dit proefschrift

In **hoofdstuk 2** analyseer ik de genomen van verschillende rondwormen en fruitvliegen. Specifiek kijken we naar genen en in het bijzonder naar intronen. Een gen bestaat uit eiwit coderende sequenties (exonen) en niet-coderende sequenties (intronen). Tijdens de evolutie van soorten gaan er soms intronen volledig verloren. Alhoewel dit een grote verandering in het DNA is, leidt intronverlies niet tot problemen, omdat het eiwit waarvoor dit gen codeert nog steeds gemaakt kan worden. Onze analyses van intronen die verloren zijn gegaan duiden sterk op een proces waarbij er een deletie optreedt van veelal korte intronen die homologie bevatten tussen de exon-intron juncties. Het gebruik van homologie is een karakteristieke eigenschap van microhomologie gemedieerde end-joining (MMEJ) wat onder alternatieve end-joining valt. Wij vermoeden dat een spontane breuk in een intron door MMEJ hersteld wordt en dat dit soms leidt tot verlies van het volledige intron.

In **hoofdstuk 3** kijken we naar wormen die een defect hebben in translesie synthese (TLS). TLS is een proces waarbij een replicatief polymerase (delta en epsilon) niet voorbij een beschadigde DNA base kan. TLS is het proces waarbij in plaats van een replicatief polymerase er een TLS polymerase (die veel toleranter is voor DNA schades) wordt gebruikt om over de schade heen te gaan. Hierdoor blijft de schade wel bestaan, maar kan DNA replicatie wel door gaan. In veel gevallen is het belangrijker dat DNA replicatie kan doorgaan dan dat de schade direct wordt gerepareerd. In *C. elegans* is de timing van celdeling en dus ook DNA replicatie tijdens de ontwikkeling heel belangrijk en enige verstoring hierheen kan leiden tot aberrante ontwikkeling en de dood van het wormembryo. Door het gebruik van whole-genome sequencing (WGS) van TLS-mutanten die ongeveer een jaar in ons lab gekweekt waren (om spontane mutaties te accumuleren) zagen we dat deze wormen een heel nauw-gedefinieerd deletie spectrum van 50 tot 300 basenparen bevatten. Verder onderzoek wees uit dat deze deleties werden veroorzaakt doordat een eiwit genaamd POLQ-1 (Polymerase theta in mensen).

Zoals dat wel vaker gaat in de wetenschap leidde data uit hoofdstuk 3 tot **hoofdstuk 4** waarin we een redelijk spectaculaire ontdekking deden: vrijwel alle DNA dubbelstrengsbreuken die tot

mutaties leiden in kiemcellen (cellen die leiden tot de volgende generatie wormen) zijn het gevolg van POLQ-gemedieerd herstel. En dat geldt niet alleen voor wormen die gekweekt zijn in het laboratorium, maar ook voor wormen die in de natuur voorkomen en onafhankelijk van elkaar zijn geëvolueerd. Dat betekent dat POLQ-gemedieerd herstel een grote rol speelt tijdens de evolutie doordat foutief herstel van DNA dubbelstrengsbreuken significant bijdraagt aan de genetische variatie tussen individuen: de basis van nieuwe soortvorming.

In **hoofdstuk 5** komen we tot de ontdekking dat wormen die gemuteerd zijn door twee veelgebruikte mutagenen (EMS en UV/TMP) die deleties veroorzaken ook het gevolg zijn van POLQ-gemedieerd herstel. EMS en UV/TMP leiden tot problemen tijdens DNA replicatie en in sommige gevallen tot deleties. *C. elegans* consortia gebruiken al 40 jaar EMS en UV/TMP om systematisch mutanten te maken voor allerlei genen. Onderzoekers kunnen vervolgens deze wormen opvragen en analyseren voor hun onderzoeken. Al deze deletie-informatie wordt minutieus bijgehouden in databases. Nadat wij vastgesteld hadden dat dit soort deleties in POLQ mutanten niet meer voorkwamen, en dus hadden laten zien dat EMS en UV/TMP geïnduceerde deleties afhankelijk zijn van POLQ, konden we alle deleties in de databases analyseren om zo te ontrafelen hoe POLQ breuken hersteld. Zo konden we zien dat in het merendeel van de gevallen er een stuk sequentie weg was (deletie) en dat de breuk was gerepareerd met een enkele base microhomologie. Daarnaast zagen we vele voorbeelden van situaties waarin een deel van de flank van de deletie op de plek van de deletie was ingekopieerd.

Recentelijk zijn er een aantal artikelen over POLQ gepubliceerd door andere onderzoekers. Wat blijkt? POLQ speelt niet alleen een cruciale rol in *C. elegans*, maar ook in mensen. Zo vonden onderzoekers dat in verschillende soorten kanker POLQ verhoogd tot expressie kwam en dat dit gepaard ging met een lagere overlevingskans voor patiënten. Daarnaast zijn er tumoren gevonden waarin homologie recombinatie (HR) niet meer functioneert die volledig afhankelijk zijn geworden van POLQ. Als de productie van POLQ vervolgens werd verstoord overleefde deze tumoren niet. POLQ blijkt dus een potentieel anti-kanker target te zijn en onderzoekers zijn druk bezig om chemische stoffen te testen die POLQ kunnen uitschakelen zonder van invloed te zijn op andere cellulaire processen. Zo zien we dat fundamenteel onderzoek in model organismen kan bijdragen aan een levensbedreigende ziekte zoals kanker.

&

Curriculum Vitae

Robin van Schendel was born on November 21, 1983 in Rijswijk, the Netherlands. He successfully completed pre-university education (VWO) at Interconfessioneel Makeblijde College (IMC) in Rijswijk in 2001. In September 2001, he started the study of Computer Science at the Delft University of Technology (TUDelft) and obtained his Bachelor degree in 2004. His first internship was performed at TNO Defence, Safety & Security which involved designing and developing software that would aid in protecting sea mammals against damage caused by high-power sonar devices. In 2005 he enrolled for the Master Software Engineering, which he completed in 2007. During his internship he focused on detecting software vulnerabilities in commonly used software packages by investigating input paths. After obtaining his master degree he enrolled for the higher laboratory school (HLO) in Leiden with the specialization Molecular Biology. During his internship at the department of Toxicogenetics in the Leiden University Medical Center (LUMC) under the supervision of Dr. Albert Pastink he focused on setting up a quantitative assay for the detection of double-strand break repair by the appearance of fluorescent cells. He obtained his Bachelor degree with honors (cum laude) in June 2009. In February 2010 he started working as a PhD student at the department of Human Genetics in the LUMC, under supervision of Prof. dr. Marcel Tijsterman.

Publications

Microhomology-mediated intron loss during metazoan evolution
van Schendel R, Tijsterman M
Genome Biology & Evolution, 2013

A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites
Koole W, van Schendel R, Karambelas AE, van Heteren JT, Okihara KL, Tijsterman M
Nature Communications, 2014

Polymerase theta-mediated end joining of replication-associated DNA breaks in *C. elegans*
Roerink SF*, van Schendel R*, Tijsterman M
Genome Research, 2014

Polymerase Theta is a key driver of genome evolution and of CRISPR/Cas9-mediated mutagenesis
van Schendel R, Roerink SF, Portegijs V, van den Heuvel S, Tijsterman M
Nature Communications, 2015

Mutagenic consequences of a single G-quadruplex demonstrate mitotic inheritance of DNA replication fork barriers
Lemmens B, van Schendel R, Tijsterman M
Nature Communications, 2015

High density of REC8 constrains sister chromatid axes and prevents illegitimate synaptonemal complex formation
Agostinho A, Manneberg O, van Schendel R, Hernandez-Hernandez A, Kouznetsova A, Blom H, Brismar H, Hoog C
EMBO reports, 2016

Repression by H3K9me is dispensable for *C. elegans* development, but suppresses RNA:DNA hybrid-associated repeat instability
Zeller P*, Padeken J*, van Schendel R, Kalck V, Tijsterman M, Gasser S
Nature Genetics, 2016

T-DNA integration in plants results from Polymerase Theta-mediated DNA repair
van Kregten M, de Pater S, Romeijn R, van Schendel R, Hooykaas P, Tijsterman M
Nature Plants, 2016

Genomic scars generated by polymerase Theta reveal the versatile mechanism of alternative end-joining
van Schendel R, van Heteren J, Welten R, Tijsterman M
PLOS Genetics, 2016

*: Co-first authors

Acknowledgements

Finished! After a scientific expedition that started almost seven years ago I have reached the finish line. I am very proud of this achievement and feel privileged that I was given the opportunity to work in science. The negative results, the disappointments, the struggles to understand biology were all overcome by occasional moments of clarity which combined together, lead to this thesis. However, no man is an island and I have many, many people to thank.

First, and foremost, Marcel. Your passion for science, your never-ending enthusiasm and desire to understand biology is a driving force for all of us. Combined with your competitive nature and your creative ideas make you an ideal mentor.

I owe many thanks to all current and former members of the Tijsterman group: Daphne, Evelien, Sophie, Wouter, Jennemiek, Marijn, Nick, Jordi, Evelina, Jane, Juliëtte, Ivo, Maartje, Joost S, Hanneke and Ron. Daphne and Jennemiek, you were there to guide me through the first weeks of my PhD, when I could not tell a hermaphrodite from a male, let alone transfer him to another plate. A special thanks also to Jennemiek for teaching me how to microinject *C. elegans* worms. Nick, thank you for scientific inspiration in the early years of my Phd and for Paul Kelly. Wouter and Sophie, we published three fantastic papers together, which I am very proud to be a part of. Jane, nothing can stop you and without your efforts this thesis and my PhD would surely have been much more boring. Jordi, my dear Catalan, one day we will finish the screen as well as the cava. Ron, thanks to you the whole lab and my experiments keep running. I also wish to thank my students Erika, Eline, Tessa, Richard for helping me during my PhD.

I share my office with some bright young people that make my daily life at the lab pleasant and stimulating as well as providing me with many distractions. Pierre, Jenny, Bharath, Suming, Leonie and Juliëtte it is great to have you around.

Seven years is a long time and I have met many people during this period that contributed in one way or another to this time: Aude, Godelieve, Angela, Alex P, Wouter W, Thomas, Mark, Joost M, Mirna, Albert, Kees V. and Eleni.

The whole Human Genetics department would collapse if it was not for the secretaries and logistic staff that help us in our moments of need and I especially want to thank Matthieu and Ingrid.

I feel privileged to have worked beside my two paranymphs who together comprise an entire think tank and are as close to me as brothers: Bennie and Dimitris, thank you for being a source of inspiration and thank you for standing by my side on this special occasion. Behind every man there is a great woman and you two are no exception; Ana and Ileana thank you for being great friends.

Behalve mijn collega's en vrienden op het lab zijn er nog vele anderen die een bijdrage hebben geleverd, soms zonder het te weten, en voor de nodige afleiding hebben gezorgd. Ivo, sporten met jou zorgt altijd voor de broodnodige afleiding, ontspanning en uitputting. Erik and Agata, thanks again for dragging me to Poland for your wedding. Alexander en Sophie, de avonden squashen en etentjes bij jullie deden meer dan goed. Kees en Joost, we gaan al een tijd terug en hebben vele avonturen beleefd en daar ben ik jullie meer dan dankbaar voor. Dan rest mij nog mijn familie te bedanken voor hun steun en liefde. Moja kochana rodzino: Ulu, Pawle, Alicjo, Aniu, Magdaleno, François oraz dzieci. Czas spędzony z Wami jest dla mnie zawsze wyjątkowym wydarzeniem, prawdopodobnie większym niż Wam się wydaje. Papa en Mama, jullie staan altijd en onvoorwaardelijk voor mij klaar, zelfs in tijden waarin jullie het zelf moeilijk hebben en daar kan ik jullie niet genoeg voor bedanken. Irene, Koen, Dennis, Angela en alle neefjes en nichtjes, onze uitstapjes samen

vormden een welkome en heerlijke afleiding voor mij.

Mijn leven zou er heel anders uitzien zonder de steun van mijn allerliefste Milena. Jij bent mijn inspiratie en steun. Jij zorgt ervoor dat ik me elke dag weer realiseer hoe mooi de wereld is. Daarvan is onze Julia nog het mooiste voorbeeld, die zonder het zich te realiseren mij inspireert en opnieuw het leven leert.