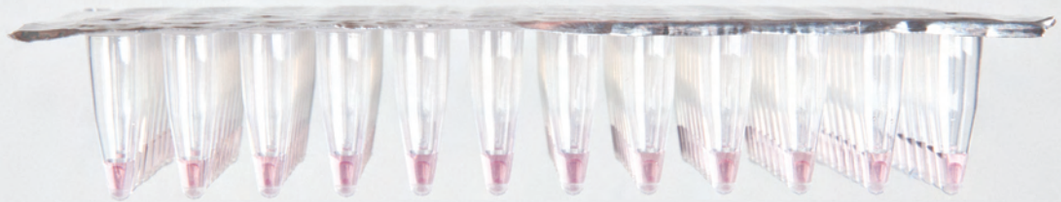
Microsatellite and G-quadruplex instability in worm, fish and man

Wouter Koole



Microsatellite and G-quadruplex instability in worm, fish and man

Wouter Koole

Microsatellite and G-quadruplex instability in worm, fish and man

Wouter Koole

Microsatellite and G-quadruplex instability in worm, fish and man

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 woensdag 13 mei 2015
klokke 11:15

door

Wouter Koole

geboren te Arcen en Velden

in 1982

Cover Design: Wouter Koole and Lotte Bronsgeest

Cover Photo: PCR PWR

Clarification: during my PhD I made use of a technique called Polymerase Chain Reaction (PCR) in which DNA molecules are amplified. Using this relatively simple but powerful technique I made a discovery that laid the foundation for chapter 4 and 5 of this thesis. The photo shows a plate in which 96 PCR-reactions were run.

Photography: Lotte Bronsgeest, www.lottebronsgeest.com

Layout & Printing: Off Page, Amsterdam, www.offpage.nl

ISBN: 978-94-6182-554-4

Promotiecommissie

Promotor:	Prof. dr. M. Tijsterman	
Overige leden:	Prof. dr. R. Kanaar	Erasmus Medical Center, Rotterdam
	Prof. dr. L.H.F. Mullenders	
	Dr. P. Knipscheer	Hubrecht Institute, Utrecht

TABLE OF CONTENTS

Chapter 1

General introduction

Cover Photo: Drop by Drop ▶

500,000,000,000,000,000,000,000 NUCLEOTIDES AT RISK

DNA encodes the genetic instructions for life. It is for every organism of vital importance to safeguard its genetic information and transmit this faithfully to its progeny. Unfaithful replication of the genome and erroneous repair of damaged DNA lead to mutations in this genetic information. These mutations can be seen as a double-edged sword: on the one hand mutations are the driving force behind evolution and result in genetic diversity, which is beneficial for maintenance of the species, but on the other hand mutations can lead to reduced fitness of an individual organism, for instance due to mutation-induced uncontrolled growth of cells (cancer).

Keeping in mind that a haploid human genome consists of approximately $3.3x10^9$ base pairs (and thus $1.3x10^{10}$ nucleotides for a diploid genome) and that a human body is estimated to consist of ~$3.7x10^{13}$ cells (Bianconi et al., 2013), this means that an average person contains at least ~$5x10^{23}$ nucleotides providing inheritable genetic information. It is overwhelming to realize that all these nucleotides are copied from a single original template (the genome of a fertilized oocyte) with an extremely low mutation-rate. A recent study, in which the sequenced genomes of 78 Icelandic parent-offspring trios was used, estimated that the *de novo* mutation rate is $1.2x10^{-8}$ per nucleotide per generation (Kong et al., 2012). This low mutation rate is even more mind-boggling when realizing that replicative polymerases encounter many obstacles during replication and nucleotides are under a constant attack of endogenous and exogenous DNA damaging sources. For instance, it has been anticipated that human cells may experience up to 10^5 spontaneous DNA lesions per cell per day (Hoeijmakers, 2009; Lindahl, 1993). How all these nucleotides are protected against mutations has been under investigation for decades. Particularly the fact that every cancer is the consequence of one or more mutations in a genome has led to extensive research in the fields of genome stability and cancer genomics. Nevertheless, despite an officially declared war on cancer (National Cancer Act 1971), and an army of scientists, many questions in these fields of research remain to be answered, since cancer is still responsible for one in eight deaths worldwide to date (ACS, 2013).

In the first part of this introduction I will focus on aspects related to genome stability; the sources of genome instability and the currently known pathways that protect the genome against instability. Special emphasis will be on DNA polymerases and helicases, which will be of relevance in particular for chapters 4 and 5 of this thesis. The second part of this introduction will provide an up-to-date overview about the current knowledge of the two main themes in this thesis: microsatellites (short tandem repeats with units of 1-8 base pairs long) and G-quadruplexes (stable secondary structures with guanines as main building blocks). Finally, since this thesis describes various model systems to study genome instability, also a short overview will be provided discussing the main advantages and limitations of each system.

ENDOGENOUS AND EXOGENOUS HAZARDS TO DNA

Genomic integrity is threatened by various types of DNA damage and processes that can lead to unwanted changes in the DNA and loss of genetic information. Below, an overview is provided of the most prominent threats to DNA. A distinction is made between exogenous and endogenous sources that can ultimately lead to genome instability. A schematic overview is provided in Figure 1.

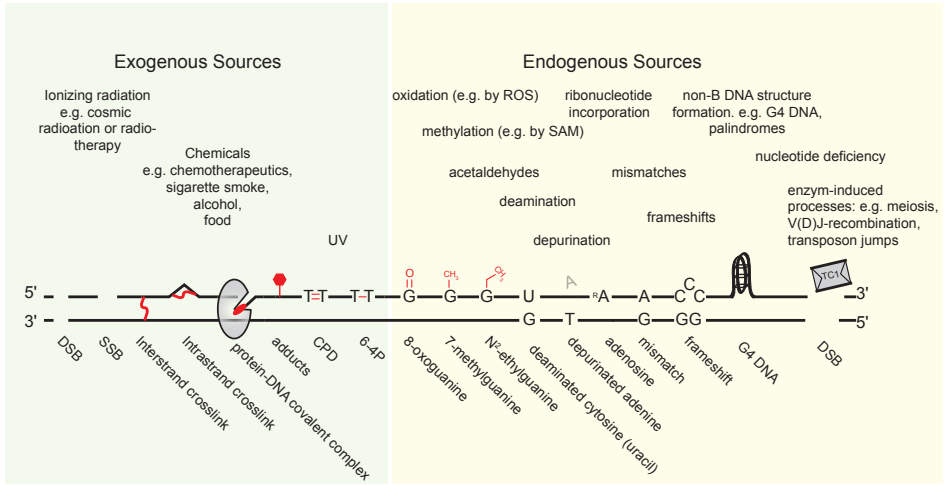


Figure 1 | Exogenous and endogenous sources that can lead to genomic instability. An overview of prominent exogenous and endogenous sources/processes that can lead to DNA modifications and ultimately to genomic instability. Examples of common types of DNA damages, replication errors and other replication blocking structures are shown that all need to be accurately resolved, repaired, or bypassed to prevent genomic instability. DSB, double-strand break; SSB, single-strand break; CPD, cyclobutane pyrimidine dimer; 6-4P, 6-4 photoproduct; ROS, reactive oxygen species; SAM, S-adenosylmethionine.

Exogenous Sources

Exogenous DNA damage is caused by physical and chemical sources from outside of the cell. Two well-known sources of physical DNA damage are ionizing radiation (IR) and ultraviolet (UV) light from sunlight. IR (for example from cosmic radiation or radiotherapy) can lead to oxidation of nucleotides and generate single-strand breaks (SSB) or, even more toxic, double-strand breaks (DSB) (Ciccia and Elledge, 2010). The most prominent lesions induced by UV light are pyrimidine dimers and 6-4 photoproducts. It has been estimated that cells exposed to the sun's UV light can suffer from up to 10^5 lesions per cell per day (Hoeijmakers, 2009).

An example of a chemical source that causes DNA damage is cigarette smoke. Estimates vary between 45 and 1,000 bulky aromatic DNA adducts per cell in tissues that are exposed to cigarette smoke (Ciccia and Elledge, 2010; Lindahl and Barnes, 2000). Also foods can contain DNA-damaging chemicals, such as aflatoxins in contaminated peanuts and heterocyclic amines in burnt meat (Wogan et al., 2004). Other chemical sources that can inflict severe DNA damage are chemical agents that are used in cancer chemotherapy. Commonly used therapies consist of toxic crosslinking agents such as mitomycin C (MMC) and cisplatin, which generate intrastrand and interstrand crosslinks (covalent links between nucleotides of the same or a different DNA strand, respectively). Other agents such as camptothecin (CPT) and etoposide trap topoisomerases in a covalent complex with DNA, which results in high numbers of SSBs and DSBs.

Endogenous sources

Besides threats from outside the cell, also endogenous reactive molecules and processes that take place during cellular and DNA metabolism endanger the genetic code. Since DNA is a chemically reactive molecule, it is exposed to processes like hydrolysis, oxidation and nonenzymatic methylation(Lindahl, 1993). A major threat of DNA hydrolysis is the generation of predominantly apurinic sites (~2,000 - 10,000 lesions/cell/day (Lindahl and Nyberg, 1972)) and deaminated cytosines (~100 - 500 lesions/cell/day (Lindahl and Barnes, 2000)). Oxidation contributes to the generation of oxidized base lesions, such as 8-oxoguanines. During normal cellular metabolism many reactive oxygen species (ROS) are generated, which, in turn, lead to oxidation of DNA bases. Non-enzymatic methylation of bases is mainly caused by the molecule *S*-adenosylmethionine (SAM). SAM is used as cofactor in many cellular transmethylation reactions and SAM-induced methylation of bases is estimated to occur at rates of ~600 and ~4000 lesions/cell/day (for 3-methyladenine and 7-methylguanine, respectively (Lindahl and Barnes, 2000)). Other reactive molecules are aldehydes, which are common byproducts formed during cellular metabolism (e.g. lipid peroxidation (O'Brien et al., 2005)) and histone demethylation (Rosado et al., 2011).

Also during DNA replication there are several processes that can lead to unwanted changes in the genetic code. Formation of secondary structures (e.g. palindromes and G-quadruplexes), mismatches and frameshifts are all associated with replication and require repair to prevent loss of genetic information. Until recently, an underrated class of DNA damage is the incorporation of ribonucleotides during DNA replication. A recent study in *Saccharomyces cerevisiae* describes that the leading-strand polymerase ε (pol-ε) incorporates one ribonucleotide monophosphate (rNMP) per 1,250 deoxyribonucleotide monophosphates (dNMP), and the lagging-strand polymerase δ (pol-δ) one rNMP per 5,000dNMPs (McElhinny et al., 2010; Nick McElhinny et al., 2010). Although it is thought that "mis"-incorporation of these rNMPs may have a biological function (as will be discussed later), these rNMPs need to be removed to prevent base-substitutions and fork collapse.

Finally, another prominent class of endogenous DNA damage is the formation of DSBs during active processes such as V(D)J-recombination, meiosis, transposon

jumping and uncoiling of DNA by topoisomerase II. Altogether, it has been estimated that a dividing human cell suffers from ~10 DSBs per day (Lieber, 2010).

THE DNA-DAMAGE RESPONSE

To counteract the threats posed by DNA damage, organisms have evolved an elaborate genome maintenance apparatus to sense DNA lesions, signal their presence and promote their repair, collectively often termed the DNA-Damage Response (DDR) (Jackson and Bartek, 2009). The DDR is a signal transduction pathway that consists of a well-orchestrated interplay between a plethora of enzymes which determine the cell's fate: survival, replicative senescence or death (Hoeijmakers, 2009). Below, a concise overview will be provided of the main signaling routes of the DDR and the specific repair-pathways that are recruited to resolve the damage.

DNA-damage signaling

Two key players in the DDR are the protein kinases ATM (ataxia-telangiectasia mutated) and ATR (ATM and Rad3-related) (Cimprich and Cortez, 2008; Shiloh, 2003). An important function of ATM together with its regulator the MRN-complex (Mre11, Rad50 and NBS1) is sensing the presence of DSBs, while ATR together with ATRIP (ATR Interacting Protein) senses replication protein A (RPA)-coated single-strand DNA (ssDNA) generated by resected DBSs or stalled replication forks (Matsuoka et al., 2007). Both kinases then phosphorylate proteins to set a signaling cascade in motion that includes the checkpoint kinases Chk1 and Chk2, which in turn activate a second wave of phosphorylation events. This whole cascade of signaling events is believed to be important for at least two particular reasons: first, it results in reduced cyclin-dependent kinases (CDK) activity which slows down or halts cell-cycle progression and allows more time for repair before going into the next phase of the cell-cycle. Second, ATM/ATR signaling enhances repair by recruiting repair proteins to the damage and activating DNA repair proteins by post-translational modifications (Jackson and Bartek, 2009). Besides ATM and ATR also PARP1 and PARP2 are important players in the DDR. Both poly(ADP-ribose) polymerases are believed to be one of the earliest responders in the DDR: within seconds they catalyze the addition of poly (ADP-ribose) chains on proteins at SSBs and DSBs and thereby recruit DDR factors to the chromatin at breaks (reviewed in (Pines et al., 2013)).

When the damage is repaired effectively, the DDR is inactivated and the cell can progress with its normal function. However, if the damage cannot be repaired, chronic DDR signaling can lead to genomic instability, cellular senescence or apoptosis. An important player in this process is the tumor suppressor TP53. A recent study in which 3,281 tumors across 12 tumor types were genome-wide sequenced, illustrated the importance of this transcription factor once again: in more than 40% of the tumors, a mutation was found in TP53 (Kandoth et al., 2013).

Base excision repair

Base excision repair (BER) is an important DNA repair pathway that is responsible for the removal of non-helix-distorting base lesions, such as oxidized, alkylated and deaminated bases. BER can be subdivided in two pathways: short-patch BER and long-patch BER. During short-patch BER only a single nucleotide is repaired, while during long-patch BER a repair tract of approximately two to eleven nucleotides is produced (Pascucci et al., 1999). The basis of BER can be summarized in four basic steps: first, the recognition and removal of the damage by a DNA glycosylase. Second, the cleavage of the DNA backbone by a DNA AP endonuclease or AP lyase resulting in a single nucleotide gap in the DNA. Next, this gap is filled by a DNA polymerase, and finally the gap is sealed by a DNA ligase (Robertson et al., 2009). This pathway was discovered in *Escherichia coli* nearly 40 years ago (Lindahl, 1974), but it became quickly apparent that this pathway was conserved among other species. In human BER, several glycosylases are involved such as 8-Oxoguanine DNA glycosylase (OGG1) and Uracil-DNA glycosylase (UNG). APEX1, APEX2 and Flap structure-specific endonuclease 1 (FEN1) function as endonucleases, and predominantly DNA polymerase beta (POLβ) and Ligase 3 (LIG3) act as required polymerase and ligase, respectively. In addition, also PARP1 and PARP2 are involved in BER and act as sensors and signal transducers for lesions. For an elaborate review about BER see reference (Robertson et al., 2009).

Nucleotide excision repair (NER)

Whereas BER is active at small base adducts, nucleotide excision repair (NER) targets the more bulky lesions that distort the structure of the DNA helix (Cleaver et al., 2009). Lesions that are repaired by NER are for instance pyrimidine dimers and 6-4 photoproducts induced by UV-light. NER is often subclassified into two branches: transcription-coupled NER (TC-NER) and global-genome NER (GG-NER). The main difference between these two classes is the way the lesion is detected, while subsequent repair is executed via a similar mechanism. In TC-NER, transcription-blocking lesions result in the stalling of RNA polymerase II (RNAPII) and the subsequent recruitment of Cockayne Syndrome protein A and B (CSA and CSB, respectively). In GG-NER, helix-distorting lesions are recognized by protein complexes that are encoded by the genes from Xeroderma Pigmentosum complementation group C and E (genes XPC and XPE, respectively). Upon lesion detection, the DNA is opened via the multifunctional protein complex TFIIH (transcription factor II human). Next, re-annealing of the DNA is prevented by RPA and XPA, followed by incision of the DNA on both sides of the lesion by endonucleases ERCC1-XPF and XPG. This results in the excision of the damage as part of a 22-30 base pair (bp) oligo. The resulting gap is filled predominantly by polymerases δ and ε, and ligation is performed by ligase I and III.

Mismatch repair (MMR)

The mismatch repair machinery deals primarily with misincorporated nucleotides and insertion and deletions loops (IDLs) which are formed during DNA replication.

MMR improves the fidelity of DNA replication several orders of magnitude and its importance is illustrated by patients that suffer from defective MMR (known as the Lynch syndrome or hereditary nonpolyposis colon cancer (HNPCC)): patients develop colon cancer at an early age (with an average onset of 45 years of age), but also other tissues are predisposed to tumor formation (e.g. 40-60% of the female carriers will develop endometrial cancer) (Lynch et al., 2009).

MMR's main task is to remove sections of nascent strands containing mispaired nucleotides or IDLs. Similar to other repair pathways MMR can be divided in distinct stages: recognition of the lesion, removal of the lesion and finally filling and ligation of the gap. Key players in MMR for detecting mismatches and IDLs are the MutSα and the Mutsβ complexes. The MutSα-complex is a heterodimeric complex consisting of MSH2 and MSH6 that recognizes base-base mismatches and IDLs of 1-3 nucleotides. The MutSβ-complex comprises MSH2 and MSH3 and is involved in recognizing IDLs of approximately 1-12 nucleotides (Peña-Diaz and Jiricny, 2012). After recognition of the lesion both MutS-complexes undergo an ATP-dependent conformational switch, which converts them into a sliding clamp on the DNA. Subsequently and in cooperation with PCNA and RFC1, the MutLα heterodimer (consisting of the nucleases MLH1 and PMS2) is recruited and activated. Due to the endonuclease activity of PMS2 nicks are introduced around the lesion, which serve as entry site for the 5′ to 3′ exonuclease EXO1. EXO1 activity results in excision of the mismatched nucleotide and a single-stranded gap of approximately 150 bps. This gap is filled by high-fidelity polymerases δ and ε. The sequential steps of canonical MMR are illustrated in Figure 2.

Despite the discovery of the first mismatch repair gene already more than 50 years ago (Siegel and Bryson, 1963), ample questions remain in the field of mismatch repair. For example, how does the MMR machinery know which base in a DNA mispair is the incorrect one, in other words, how is the nascent strand recognized? It is known that in prokaryotes like *Escherichia coli* (*E. coli*) the nascent strand is recognized by the presence of unmethylated adenines, however strand recognition by methylation is not used by eukaryotes. Until recently, many researchers favored the hypothesis that the MMR in eukaryotes was directed to the nascent strand by the presence of strand discontinuities such as gaps between Okazaki fragments. However, recent studies (Ghodgaonkar et al., 2013; Lujan et al., 2013) provide strong evidence that the incorporation of ribonucleotides into the nascent strand during DNA synthesis guides the MMR machinery towards the nascent strand: removal of ribonucleotides by RNAse H2-dependent ribonucleotide excision repair (RER) creates an initiation site for MMR, and in this way the MMR machinery is directed to the error-containing nascent strand.

Another long-standing puzzle in the field of MMR is the identification of several colorectal and other cancers that are characterized by high microsatellite instability (MSI), a hallmark for defective MMR, but for which no genetic or epigenetic defect is detected in the known players of MMR. This raises the question whether there

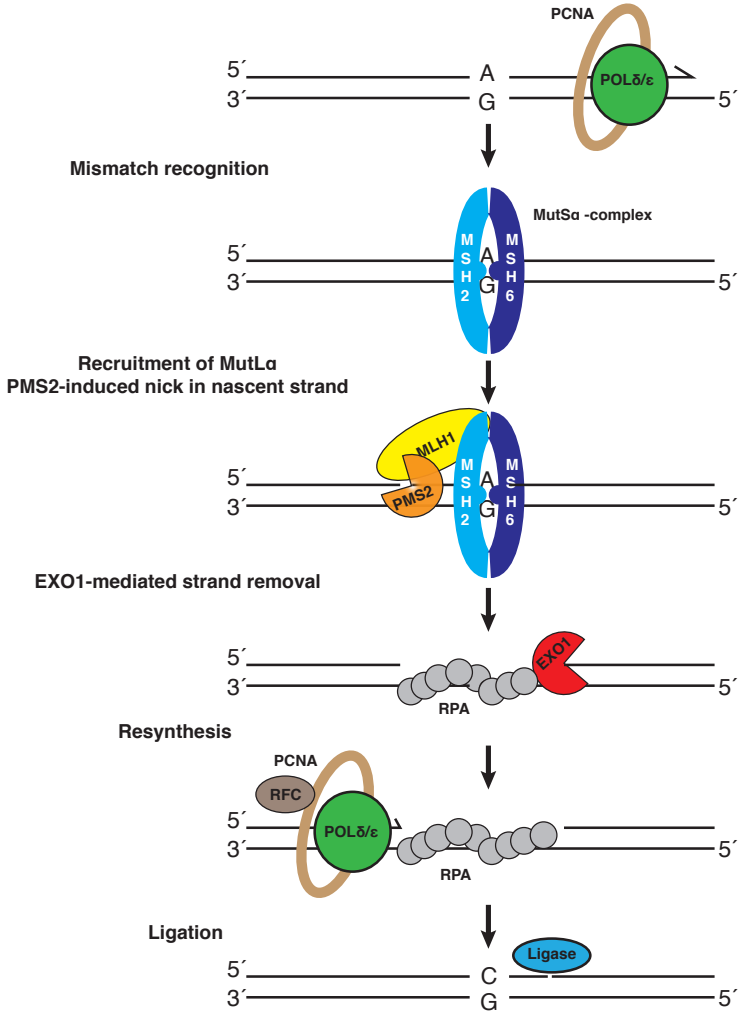


Figure 2 | Sequential steps of MMR in eukaryotes. See text for further details.

are more proteins involved in MMR than currently known. Indeed, a recent study by Li et al (Li et al., 2013) reports the involvement of such a novel player, namely the histone methyltransferase SETD2: cells lacking functional SETD2 display MSI and an elevated mutation frequency. The authors provide evidence that a SETD2-dependent epigenetic histone mark, H3K36me3, is required to recruit the Mutsα-complex to the chromatin.

These very recent discoveries showing important roles of chromatin organization and ribonucleotides incorporation/removal in MMR, underline that still a lot is incompletely understood in the field of MMR.

DNA damage tolerance pathway (DDT)

Despite the presence of numerous repair pathways, DNA lesions might escape repair. When the replication machinery encounters a blocking lesion, the replication fork is stalled which may ultimately result in replication fork collapse, genomic instability and toxicity. To avoid such a scenario, cells can trigger the DNA damage tolerance pathway leading to bypass of the lesion, thus tolerating the lesion to be unrepaired. In this way cells can proceed to complete DNA replication, which is of importance for cellular survival, while the lesion can be repaired at a later time point.

There are two major pathways implicated in DDT: translesion synthesis (TLS) and template switching (TS, also often termed damage avoidance (DA)). In TLS, specialized polymerases are recruited to the fork to replace the stalled replicative polymerase. In contrast to the normal replicative DNA polymerases δ and ε, TLS polymerases have a more open structure, and therefore they can directly bypass damaged bases or bulky adducts. This bypass, however, often comes with the cost of mutation induction, since TLS polymerases are notorious for operating in an error-prone fashion. Therefore TLS is considered an error-prone pathway. By contrast, template switching/damage avoidance is an error-free process: during template switching the undamaged sister chromatid is temporarily used.

The choice between using either TLS or template switching at a lesion is still not fully understood. It is clear though that the posttranslational modification of the homotrimer DNA sliding clamp PCNA plays an important role. It is generally believed that mono-ubiquitination of PCNA promotes TLS whereas polyubiquitination stimulates TS/DA. For more details and a comprehensive overview about DDT see (Ghosal and Chen, 2013).

Fanconi anemia (FA) pathway and the repair of DNA interstrand crosslinks

A form of damage that can never be bypassed is a DNA interstrand crosslink (ICL), a covalent chemical bond between two nucleotides of opposing DNA strands. ICL's are very toxic to cells since they prevent strand separation and therefore hamper essential biological processes such as DNA replication and transcription. The toxicity but also the therapeutically potential of ICL's became apparent when a ship, loaded with ± 60.000 tonnes of the ICL-agent nitrogen mustard, was bombed in the harbor of Bari during the Second World War (Deans and West, 2011). Many soldiers and civilians were exposed to the nitrogen mustard and autopsies on fatal casualties showed that the chemical specifically attacked the victims' white blood cells. This discovery led to the view that these kind of chemicals could be used as a potential treatment for patients that suffered from leukemia. More than 70 years later, ICL-agents such as cisplatin and mitomycin C (MMC) are still widely used in the clinic to treat leukemia and other types of cancer. Also in research laboratories these compounds are commonly used to investigate the molecular mechanisms that play a role in the repair of ICLs.

Many years of research has led to great insight into which genes and pathways are involved in ICL repair. Key in the dissection of the molecular mechanisms behind ICL repair was the identification of the underlying mutations in Fanconi anemia (FA) patients, as these patients showed severe sensitivity to ICL-agents. To date, 16 Fanconi anemia complementation groups (FANCA-FANCQ) have been connected to distinct genes and there are still patients in whom a mutation has yet to be identified (see for a complete list of identified genes and their functions references (Garaycoechea and Patel, 2014; Kottemann and Smogorzewska, 2013). Interestingly, several of these 16 genes were already known to play a role in other DNA repair pathways (e.g. FANCQ/XPF in NER, FANCO/RAD51C in homologous recombination). Studies in Xenopus extracts (Klein Douwel et al., 2014; Knipscheer et al., 2009) have lead to a model in which the interstrand crosslinks are removed by an elaborate interplay between FA proteins, excision repair, translesion synthesis and homologous recombination (HR) (see Figure 3) (Garaycoechea and Patel, 2014); i) Upon replication, two converging forks stall at the ICL. ii) Recruitment of the so-called "FA core complex" takes place and ensures iii) monoubiquitination of FANCD2 and FANCI. iv) Next, unhooking of the ICL is accomplished by the incision in one strand on both sides by the nucleases FANCQ/XPF-ERCC1 and FANCP/SLX4, which results in a broken chromatid (bottom strand in Figure 3c) and an intact chromatid containing the crosslink (top strand in Figure 3c). v) Translesion synthesis bypasses the lesion, and it is generally thought that NER removes the lesion in the top strand, and HR (involving FANCD1/BRCA2, FANCO/RAD51C, FANCN/PALB2 and FANCJ/BRIP1) is used to repair the broken chromatid, resulting in two complete repaired DNA duplexes.

Although it is nowadays clear why FA-defective patients are hypersensitive to exogenous administered ICL-agents, it is less well understood which endogenous source(s) are the cause of the clinical manifestations seen in FA-patients. FA patients are characterized by bone marrow failure, congenital abnormailities, infertility and a high risk to develop cancers. Remarkably, FA is phenotypically heterogeneous; some patients develop bone marrow failure at the onset of 3 years, whereas other patients with the exact same mutation may never suffer from bone marrow failure. Interestingly, recent studies by the group of Patel and Hira indicate that aldehydes may be an endogenous source that can result in genomic instability in the absence of a functional FA pathway ((Garaycoechea et al., 2012) and references therein). Aldehydes (like formaldehyde and acetaldehyde) can be formed during cellular metabolism (e.g. during DNA and histone methylation) and are able to form DNA adducts. Intriguingly, both mice and humans who are deficient for the FA-pathway and also in the breakdown of acetaldehyde due to a mutation in aldehyde dehydrogenase 2 (*Aldh2*), show dramatically increased manifestations of FA-related clinical features such as bone marrow failure (Hira et al., 2013; Langevin et al., 2011). This finding strongly suggest that adducts of aldehydes to the DNA are natural substrates for the FA-pathway. Whether adducts formed by these aldehydes are removed from the DNA in a similar way as ICLs is yet unknown. Furthermore, it will be interesting to find out

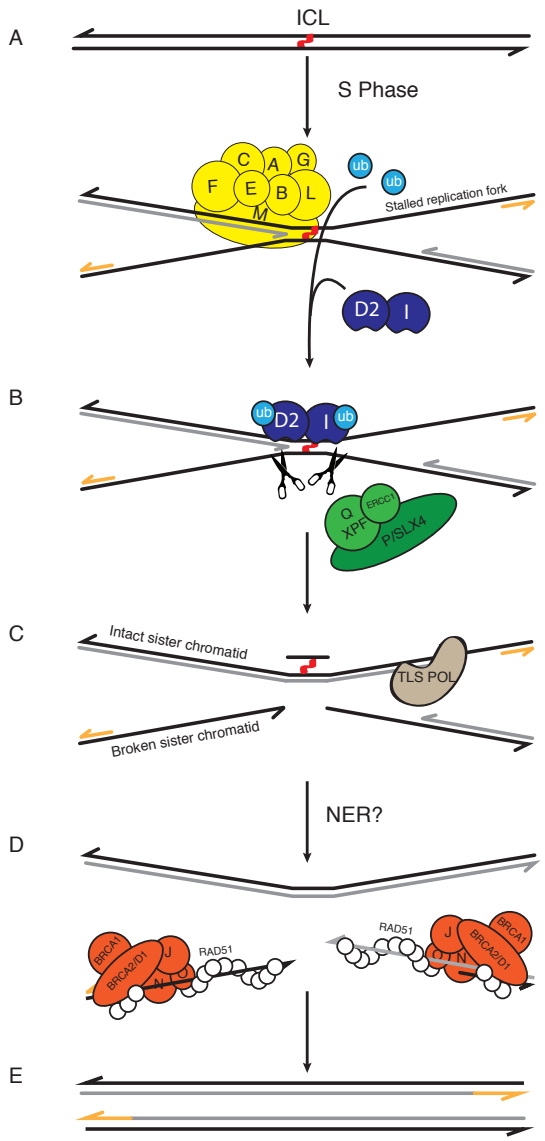


Figure 3 | Current view of Interstrand Crosslink Repair by the Fanconi Anemia Pathway. (A) Upon replication, the leading strands of two converging replication forks are blocked at an ICL (depicted in red). Recruitment of the FA core complex (indicated in yellow) takes place and ensures monoubiquitination of its substrates FANCD2 and FANCI. (B) Incision on both sides of the ICL is accomplished by the nucleases XPF (FANCQ), ERCC1 and SLX4 (FANCP), resulting in the uncoupling of the two sister chromatids. (C) Translesion synthesis ensures extension of the nascent strand beyond the ICL and it is thought that the crosslink is removed by NER. (D) HR is used to repair the broken chromatid. (E) Together, ICL-repair by the FA pathway results in two fully repaired DNA duplexes. Figure adapted from Garaycoechea & Patel, 2014.

whether exposure to different levels of these toxic metabolites could also explain the phenotypically heterogeneity found in FA-patients.

Double-strand break repair

DNA double-strand breaks (DSBs) are dangerous lesions for a cell. Inappropriate repair of DSBs can lead to loss of genomic information, inversions and, perhaps most hazardous of all, chromosome translocations, which can lead to the formation of oncogenic gene-fusions. Given that cells frequently endure DSBs and that DSBs are formed under various conditions (e.g. during replication, mitosis, meiosis), it is not surprisingly that cells have evolved several DSB repair pathways to preserve genomic integrity. Two prominent DSB repair pathways are homologous recombination (HR) and nonhomologous end-joining (NHEJ). HR is considered as an error-free process but requires the use of a homologous template. Therefore, HR can usually only take place for breaks that occur during or after DNA replication, when an identical sister chromatid is available as template (so during the S and G2 phase of the cell cycle). In contrast, the error-prone pathway NHEJ is able to ligate two broken ends of a chromosome without the need of a homologous template and can therefore also be active in cells in the G1-phase. A third pathway that is able to repair DSBs goes by a variety of names of which alternative NHEJ (alt-NHEJ) and microhomology-mediated end joining (MMEJ) are most commonly used. As the latter name implies, this pathway makes use of small pieces of homology to fuse two broken ends together. Below I will discuss these three pathways in more detail.

Homologous recombination (HR)

Crucial for homologous recombination is the generation of 3′ single-stranded DNA (ssDNA) tails that are required to find and invade a homologous template. Recent studies suggest a two-step model for the generation of these overhangs ((Symington, 2014) and references therein): first, the DSB is recognized, bound and processed by the MRE11-RAD50-NBS1 complex leaving a short 3′ overhang. In yeast, and likely in higher organisms, this 3′ short overhang is generated by endonuclease activity of Mre11 in cooperation with Sae2 (homolog of the human nuclease CtIP) 15-20nt downstream of the break, followed by 3′ to 5′ exonuclease activity of Mre11 towards the break (Cannavo and Cejka, 2014). The second step involves long-range resection by the nucleases EXO1 and DNA2 generating extensive tracts of ssDNA. To prevent the formation of secondary structures, the ssDNA becomes coated with the heterotrimeric complex RPA (RPA1, RPA2, RPA3).

After resection and coating of the ssDNA with RPA, loading of RAD51 takes place, aided by BRCA2, creating a nucleoprotein filament that is able to invade homologous duplex DNA (known as D-loop formation). After invasion, the strand can be extended by a DNA polymerase, dissociate (a process stimulated by helicase RTEL1) and re-anneal to the other end of the break (a process called synthesis dependent strand annealing, SDSA). Alternative to SDSA, double Holliday Junctions can be formed,

and endonucleases GEN1, MUS81/EME1, SLX1/SLX4 are required for resolving the intertwined DNA strands(Boulton et al., 2012; Ciccia and Elledge, 2010).

Unlike HR, which is error-free, an error-prone alternative when breaks are surrounded by repeat sequences is single strand annealing (SSA). Independent of RAD51, but catalyzed by RAD52, annealing of the resected strands can take place at the two repeats, followed by flap removal by XPF/ERCC1 and ligation (Ciccia and Elledge, 2010).

Classical Nonhomologous end-joining

In classical NHEJ (cNHEJ), the first step is the binding of the heterodimer Ku (Ku70 and Ku80). This step happens within seconds after DSB-formation, and binding of Ku prevents resection of the break ((Mahaney et al., 2009) and references therein). Next, Ku translocates inwards, allowing the recruitment of the protein kinase DNA-PKcs at the DNA termini, thereby assisting in tethering the broken ends together. To remove non-ligatable end groups or other lesions, processing may occur at the termini by among others the exonuclease ARTEMIS and polymerases λ and μ. Finally, the break is ligated by XRCC4 and Ligase IV (LIG4), with the help of XLF. Because frequently limited processing of the DNA ends takes places, NHEJ is characterized by small deletions or insertions and is therefore considered an error-prone repair pathway.

Alt-NHEJ/ MMEJ

Two main features characterize alt-NHEJ: first of all, it accomplishes the repair of a break without the requirement of the classical NHEJ factors such as Ku and Ligase IV. Second, the repair products are often characterized by excessive deletions, microhomology of 1-10 base pairs and templated insertions (Deriano and Roth, 2012). This pathway has very recently been subjected to increasing investigation and many questions remain to be addressed. For example, what determines whether a DSB is repaired via cNHEJ or Alt-NHEJ? One study suggests that perhaps PARP1 could play a role in this process by competing with Ku and thereby directing the repair of the break towards Alt-NHEJ (Wang et al., 2006). Furthermore it remains poorly understood which nucleases, polymerases and ligases are involved in processing and ligating the DNA termini.

Since many chromosomal translocations show characteristics of Alt-NHEJ (Decottignies, 2012), it is clear that better understanding of the process of Alt-NHEJ can be of great value.

DNA POLYMERASES AND HELICASES IN DNA REPAIR

DNA polymerases and helicases play an important role in DNA repair pathways. In chapter 4 and 5 of this thesis, I investigate the genetic consequences of G-quadruplex instability, and to which extent polymerases and helicases are involved in the prevention and the repair of G-quadruplex-induced DNA damage. In the following paragraphs, I provide a short introduction to the polymerases and helicases involved in DNA repair.

Polymerases in DNA repair

Both repair and bypass of damaged DNA often require DNA polymerase activity. The mammalian genome encodes at least 16 DNA polymerases, which can be subdivided in four main families (A, B, X and Y, see also table 1). Replication of undamaged DNA is performed by polymerases from the B-family, including Pol α, Pol δ and Pol ε. Initiation of DNA synthesis depends on the Pol α –primase complex. The primase synthesizes an oligo of 7-12 ribonucleotides, which is then elongated by pol α with ±20-30 deoxyribonucleotides (Muzi-Falconi et al., 2003). Next, the Pol α –primase complex is substituted by either Pol δ or Pol ε, which will in their turn elongate the synthesized RNA-DNA hybrid in the lagging or leading strand, respectively. For a long time it was thought that in human cells the Pol α –primase complex was the sole complex that could initiate *de novo* synthesis. However, a recent study discovered a second primase, named PRIMPOL, which is furthermore a TLS polymerase and is involved in the repriming of DNA synthesis at stalled replication forks (García-Gómez et al., 2013; Mourón et al., 2013).

Polymerases from the Y-family (Pol η, Pol κ, Pol ι, and REV1) are mainly involved in translesion synthesis. These Y-family polymerases have a more open catalytic active site (compared to Pol δ and Pol ε), which allows them to synthesize DNA past damaged nucleotides. A major function of Pol η is to bypass UV-induced CPDs and defective Pol η will lead to the cancer predisposition disease Xeroderma Pigmentosum. Pol κ and Pol ι are mainly involved in the bypass of N^2-dG adducted sites and dA templates, respectively (Sale et al., 2012). REV1 can only incorporate dC residues opposite abasic cites and dG. In addition, REV1 plays an important role in the bypass of G-quadruplex structures in chicken cells (Sarkies et al., 2010).

In higher eukaryotes, the X-family of polymerases consists of four members, namely Pol β, Pol λ, Pol μ and TdT. Remarkably, some organisms, such as *C. elegans* and *D. melanogaster*, appear to be devoid of any X-family polymerase (Uchiyama et al., 2009). Pol β plays an important role in BER, whereas the other three polymerase are implicated in NHEJ (Yamtich and Sweasy, 2010). Notably, Pol μ and TdT are mainly expressed in lymphoid tissues and are thought to play an important role in V(D)J-recombination.

Three polymerases belong to the group of A-family polymerases. One member, Pol γ, is dedicated for the replication of mitochondrial DNA. Pol ν is considered to be a proficient TLS polymerase for the accurate bypass of thymine glycols (Takata et al., 2006). The last member of the A-family polymerases is named Pol θ, also known as POLQ. Pol θ is a large 290kDA protein and is characterized by an N-terminal ATPase-helicase like domain and a C-terminal polymerase domain, flanking a large central

Table 1 | An overview of mammalian and *C. elegans* genes involved in DNA repair pathways. Also an overview is provided of all DNA polymerases and a selection of prominent helicases and their polarity.

Mammalian Homolog	(putative) *C. elegans* homolog		
HR		**ICL /Fanconi Anemia pathway**	
MRE11	*mre-11*	*FANCA*	-
RAD50	*rad-50*	*FANCB*	-
NBS1	-	*FANCC*	-
CtIP	*com-1*	*FANCD1/BRCA2*	*brc-2*
EXO1	*exo-1*	*FANCD2*	*fcd-2*
DNA2	*dna-2*	*FANCE*	-
RPA1-3	*rpa-1 - 3*	*FANCF*	-
RAD51	*rad-51*	*FANCG/XRCC9*	-
RAD52	-	*FANCI*	*fnci-1*
BRCA1	*brc-1*	*FANCJ/BRIP1/BACH1*	*dog-1*
BRCA2	*brc-2*	*FANCL/POG*	-
GEN1	*gen-1*	*FANCM*	*fncm-1*
SLX1	*slx-1*	*FANCN/PALB2*	-
SLX4/FANCP	*slx-4*	*FANCO/RAD51C*	-
MUS81	*mus-81*	*FANCP/SLX4*	*slx-4/him-18*
EME1	*eme-1*	*FANCQ/XPF*	*xpf-1*
		FAN1	*fan-1*
NHEJ			
KU70	*cku-70*	**NER**	
KU80	*cku-80*	*DDB1*	*ddb-1*
LIG4	*lig-4*	*DDB2*	-
DNAPK	-	*ERCC1*	*ercc-1*
XRCC4	-	*ERCC4/XPF/FANCQ*	*xpf-1*
Artemis	-	*ERCC5/XPG*	*xpg-1*
XLF	-	*CSA*	-
		CSB	*csb-1*
MMR		*LIG1*	*lig-1*
MSH2	*msh-2*	*LIG3*	K07C5.3
MSH3	-	*PCNA*	*pcn-1*
MSH6	*msh-6*	*RFC1-5*	*rfc-1 – 4*, F44B9.8
MLH1	*mlh-1*	*RPA1-3*	*rpa-1 -3*
PMS2	*pms-2*	*XPA*	*xpa-1*
EXO1	*exo-1*	*XPB/ERCC3*	Y66D12A.15
		XPC	*xpc-1*
		XPD/ERCC2	Y50D7A.2

Table 1 | An overview of mammalian and *C. elegans* genes involved in DNA repair pathways. Also an overview is provided of all DNA polymerases and a selection of prominent helicases and their polarity. (*Continued*)

Mammalian Homolog	(putative) *C. elegans* homolog
BER	
-	*apn-1*
APE1	*exo-3*
NTHL1	*nth-1*
UDG/UNG	*ung-1*
Polymerase β	-
LIG3	K07C5.3
FEN1	*crn-1*
PARP1	*pme-1*
PARP2	*pme-2*

Polymerases		**Helicases**	
Family A		RecQ-family	
Pol γ	*polg-1*	*RECQ1* (3′-5′)	K02F3.12
Pol ν	-	*BLM* (3′-5′)	*him-6*
Pol θ /POLQ	*polq-1*	*WRN* (3′-5′)	*wrn-1*
		RECQ4 (3′-5′)	-
Family B		*RECQ5* (3′-5′)	*rcq-5*
Pol α	*div-1*		
Pol δ	F10C2.4	Fe-S cluster	
Pol ε	F33H2.5	*FANCJ* (5′-3′)	*dog-1*
Pol ζ / REV3	Y37B11A.2	*XPD* (5′-3′)	Y50D7A.2
		RTEL1 (5′-3′)	*rtel-1*
Family X		*DDX11/CHL1* (5′-3′)	*chl-1*
Pol β	-		
Pol λ	-	others	
Pol μ	-	*PIF1* (5′-3′)	*pif-1*
TdT	-	*XPB* (3′-5′)	Y66D12A.15
		DNA2 (3′-5′)	*dna-2*
Family Y		*ATRX*	*xnp-1*
Pol η	*polh-1*	*HELQ* (3′-5′)	*helq-1*
Pol κ	*polk-1*		
Pol ι	-		
REV1	*rev-1*		

domain. Pol θ is highly expressed in the testis, placental tissue and hematopoietic cells (Seki et al., 2003; Shima et al., 2004) (Kawamura et al., 2004). Overexpression of Pol θ has been observed in several cancers and correlates with a lower patient survival rate(Higgins et al., 2010; Lemée et al., 2010). Pol θ is a low fidelity polymerase but has the capability to extend DNA from minimally paired primers(Arana et al., 2008; Seki et al., 2004; Yousefzadeh et al., 2014). As research progresses, more and more functions are ascribed to Pol θ. Pol θ is implicated in the bypass of abasic sites and thymic glycols (Seki et al., 2004; Yoon et al., 2014), functioning as a backup polymerase in BER (Asagoshi et al., 2012; Prasad et al., 2009; Yoshimura et al., 2006), linked to Alt-NHEJ and ICL repair in *D. Melanogaster* (Chan et al., 2010; Harris et al., 1996) and found to be involved in the timing of firing of origins of replication (Fernandez-Vidal et al., 2014).

Helicases in DNA repair

Helicases are ATP-dependent motor proteins that are able to unwind duplex nucleic acids. Various cancers and genetic disorders are linked to helicase defects, which illustrates their importance. Their prominence is furthermore marked by the great number of helicases found; a recent computational study reported 95 human genes encoding for helicases, of which 64 and 31 are thought to be RNA and DNA helicases, respectively (Umate et al., 2011). Based on motifs and consensus sequences, helicases have been classified in two larger superfamilies (SF1 and SF2) and four smaller superfamilies (SF3-6) (Singleton et al., 2007). Other classifications are based on whether the helicase acts on single or double-strand DNA (indicated by α and β, respectively) and by their polarity; type A helicases translocate in a 3′ to 5′ direction and type B helicases from 5′ to 3′. Most helicases discussed in this thesis belong to the SF2 family. Two prominent subclasses of the SF2 family are the RecQ family and Fe-S family. The RecQ family consists of five 3′ to 5′ helicases (RECQL1, BLM, WRN, RECQL4 and RECQL5). They are highly conserved and required for genome stability (Chu and Hickson, 2009). Defects in BLM, WRN and RECQL4 are linked to syndromes that predispose to cancer (Bloom syndrome, Werner syndrome and Rothmund-Thomson syndrome, respectively). RecQ helicases are primarily related to the repair of DSBs, fork regression and Holliday junction branch migration (Brosh, 2013). Furthermore there is biochemical and *in vivo* data that BLM and WRN are involved in the unwinding and bypass of G4 DNA (Fry and Loeb, 1999; Sarkies et al., 2012; Sun et al., 1998).

In contrast to the RecQ family, helicases belonging to the Fe-S family have a 5′ to 3′ polarity and are characterized by a conserved iron-sulfur (Fe-S) cluster. Although the exact role of the Fe-S cluster remains to be elucidated, it is thought that its redox properties are used to scan the genome for DNA damage (Wu and Brosh, 2012). Four DNA helicases belong to the Fe-S family (XPD, FANCJ/BRIP1/BACH1, RTEL1 and DDX11/CHL1/ChiR1) and are all implicated in autosomal recessive genetic disorders. XPD plays an important role in NER and is linked to Xeroderma Pigmentosum.

DDX11 is connected with the Warsaw Breakage Syndrome and is important for sister chromatid cohesion during DNA repair (van der Lelij et al., 2010). Regulator of telomere elongation helicase 1 (RTEL1) is indispensable for the maintenance of telomeres and for the dismantling of D-loop recombination intermediates (Barber et al., 2008; Vannier et al., 2013; 2012). Defective RTEL1 leads to dyskeratosis congenital (Ballew et al., 2013). Homozygous mutations in FANCJ lead to Fanconi Anemia (Levitus et al., 2005), whereas female carriers of monoallelic mutations in FANCJ have an elevated risk to develop breast cancer (Hiom, 2009). Several interacting proteins have been described for FANCJ of which the most prominent are BRCA1, MLH1, MRE11, RPA and FANCD2 (Cantor et al., 2001; Chen et al., 2014; Guillemette et al., 2014; Sommers et al., 2014; Suhasini et al., 2013). As research progresses, functions are described for FANCJ in ICL (Levitus et al., 2005), MMR, NER (Guillemette et al., 2014), displacement of DNA-protein blocks (Sommers et al., 2014) and unwinding of G-quadruplexes (Bosch et al., 2014; Cheung et al., 2002; Kruisselbrink et al., 2008; Sarkies et al., 2012; Schwab et al., 2013) and thereby maintaining genomic and epigenetic stability.

Other helicases that are considered as important genome caretakers are PIF1, ATRX, DNA2, and XPB. Although they function in different pathways such as Break-Induced Replication (PIF1) (Wilson et al., 2013) and NER (XPB), they share the property to unwind G4 DNA (Gray et al., 2014; Law et al., 2010; Lin et al., 2013; Ribeyre et al., 2009).

For a more detailed overview about DNA helicases and their role in DNA repair and cancer see reference (Brosh, 2013).

MICROSATELLITES AND G-QUADRUPLEX STRUCTURES

Apart from DNA damage, genomic integrity is endangered by DNA sequences that are difficult to replicate. In this thesis, I focus on two of such sequences: microsatellites and G-quadruplex sequences. In the next section, I will describe these mutagenic sequences in more detail.

Microsatellites

More than 40% of the human genome consists of repeats (Gemayel et al., 2010). Although historically these repeats were seen as “junk DNA” and therefore ignored, we now start to realize that their importance has been severely misjudged; changes in repeat-length can lead to phenotypic changes and many diseases, particularly neurodegenerative diseases.

Repeats can be categorized in two main groups: interspersed and tandem repeats (TR). Interspersed repeats are remnants of transposons and, as the name suggests, are interspersed throughout the genome. Tandem repeats consist of a short DNA sequence, named a “unit”, that is repeated several times right next to the other. Tandem repeats can

be subdivided in microsatellites, minisatelites and megasatellites. This classification is based on the length of the unit. Although definitions vary, microsatellites (also known as short tandem repeats) are repeats with units of 1-8 nucleotides, minisatellites are 9-135 nucleotides, megasatellites consist of untis greater than 135 nucleotides (Gemayel et al., 2010). With the completion of sequencing and assembly of the human genome, it appeared that approximately 17% of all genes contain a TR in their open reading frame (ORF). Of these TRs, microsatellites have been under great investigation, since they are thought to influence processes such as gene expression, chromatin organization and recombination at hotspots (Li et al., 2002).

Approximately 3% of the genome consists of microsatellites and the distribution of microsatellites appears to be nonrandom (Katti et al., 2001; Lander et al., 2001). The majority of microsatellites consist of mono-, di-, tri- and tetranucleotide repeats, of which the dinucleotide repeats are the most prominent. Microsatellites can be extremely unstable; mutation rates vary between 10^{-3} and 10^{-7} per cell division but rates above 10^{-2} have been described (Gemayel et al., 2010). Since their discovery in the early 1980s, it quickly became apparent that their instability and subsequent polymorphisms make microsatellites excellent markers for genome mapping and population genetics.

Two major models are proposed for microsatellite expansion and contraction: recombination and strand-slippage during replication (Gemayel et al., 2010). In the latter, the nascent strand denatures from the template strand during synthesis of the microsatellite and then pairs with another part of the repeat sequence. This can lead to a loop either in the nascent strand or in the template strand, resulting in expansion or contraction of the microsatellite, respectively. Notably, in most cases, these loops are recognized by the mismatch repair machinery, thereby preventing microsatellite instability. During a recombination event, unequal crossover or gene conversion can lead to expansion/contraction of a microsatellite.

Microsatellites are abundant in the genome, are very instable, linked to important biological functions and connected to disease. These aspects stress the need to elucidate the factors and mechanisms involved in microsatellite instability.

G-quadruplex structures

In 1910, the German chemist Ivar Bang made the observation that guanylic acid formed gels when kept at high concentrations (Bang, 1910). More than fifty years thereafter Gellert and colleagues (Gellert et al., 1962) found an explanation for this unusual physical property; using X-ray diffraction techniques they showed that guanylic acids can assemble into tetrameric structures, also named G-quartets (see Figure 4a+b). In this configuration each of the four guanine molecules is a donor and acceptor of two hydrogen bonds and in the central cavity a metal cation plays an important role in stabilizing the G-quartet. Stacking of multiple G-quartets forms a stable G4 structure (hereafter also named G-quadruplex and G4 DNA).

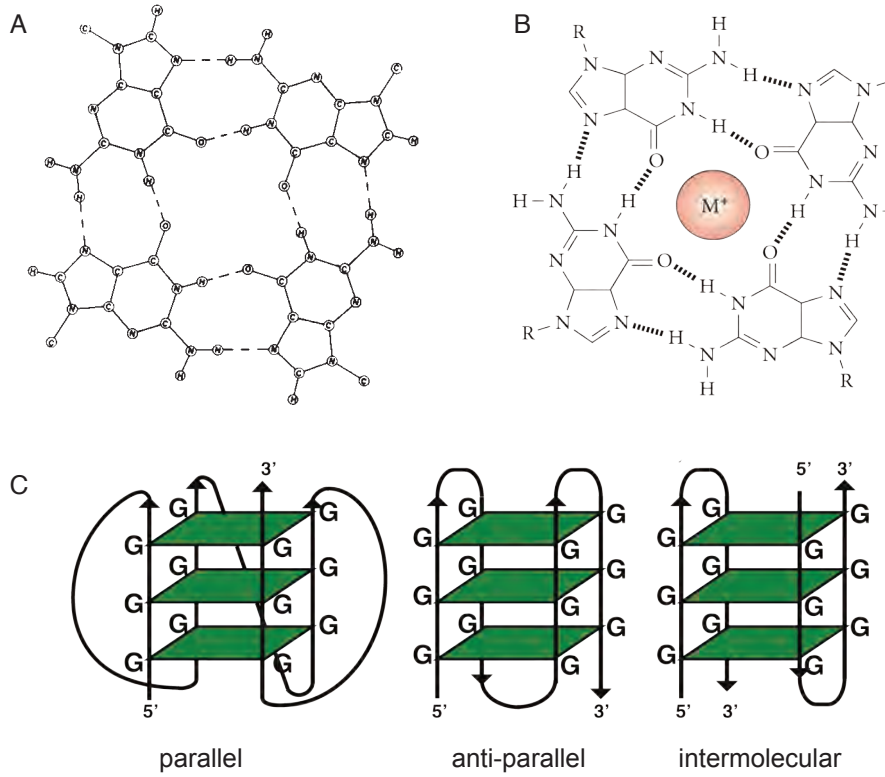


Figure 4 | G-quadruplex DNA. (A) Proposed arrangement of interacting guanines in a G-quartet by Gellert et al in 1962. (B) Current view of interactions in a G-quartet. M+ denotes a monovalent cation. Illustration adapted from Bochman et al, 2012. (C) Schematic representation of intramolecular G-quadruplex structures in a parallel and anti-parallel conformation (left and middle panels). The right panel illustrates the topology of an intermolecular G-qaudruplex structure formed by dimerization of two strands. Illustration adapted from Tarsounas & Tijsterman, 2013.

G-quadruplexes come in many flavours: they can form within one strand (intramolecular) or from two or more strands (intermolecular), strands can run in a parallel or antiparallel orientation (Figure 4c), and various loop structures can form by the nucleotide linkers between the stacks. Additionally, G-quadruplexes can consist of DNA or RNA molecules or a combination of both. The stability of a G-quadruplex depends on multiple factors: the number of G-quartets formed, the size of the loops and the nature of the stabilizing cation.

Although the formation of G-quadruplexes was shown *in vitro*, many researchers remained skeptic about their presence *in vivo* and about a potential biological function. However, this skepticism is likely greatly reduced with the publication of some seminal publications in the last couple of years. Below, I will briefly introduce some of these groundbreaking publications that demonstrate that G-quadruplexes are

present *in vivo*, cause genomic and epigenetic instability, have a biological function and are linked to several diseases.

G-quadruplexes and genomic instability

In 2002 the lab of Lansdorp identified a helicase-defective worm that triggered deletions upstream guanine-rich DNA (Cheung et al., 2002). They named the gene encoding the helicase *dog-1* (for deletion of guanine-rich DNA), which later turned out to be the homolog of human FANCJ (Youds et al., 2008). Six years later our lab demonstrated that solely guanine-rich sequences that match the G-quadruplex consensus motif ($G_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}$) lead to the induction of deletions (Kruisselbrink et al., 2008). Besides in worms, G-quadruplex sequences appeared to cause genomic instability in yeast (Piazza et al., 2012; Ribeyre et al., 2009) and in human cells (Rodriguez et al., 2012). In addition, it has been shown that G-quadruplex sequences are enriched in breakpoints in cancer genomes (De and Michor, 2011). Studies in chicken cells furthermore show that G-quadruplexes can lead to epigenetic instability (Sarkies et al., 2010; 2012; Schwab et al., 2013). Finally, a study in mice shows that G-quadruplex formation endangers telomere integrity (Vannier et al., 2012).

Evidence of G-quadruplex formation *in vivo*

Besides the previously described publications that (indirectly) imply that G-quadruplexes must form *in vivo*, several labs have attempted to visualize G-quadruplexes *in vivo* with help of antibodies. In 2001 a study provided evidence that telomeres of the single-celled eukaryote *Stylonychia lemnae* formed G-quadruplex structures *in vivo*. Recent studies with newly developed antibodies showed the presence of G-quadruplexes in mammalian cells *in vivo* (Biffi et al., 2013; Henderson et al., 2013). G-quadruplexes appear to be enriched in replicating cells and cancerous tissue (Biffi et al., 2013; 2014).

G-quadruplexes and their function

One of the first publications describing an *in vivo* function for G-quadruplexes was a study performed in the bacteria *Neisseria gonorrhoeae*. Here, Cahoon and Seifert showed that a G-quadruplex drives antigenic variation by serving as a recombination hotspot. Disruption of the G4 motif by changing only a single nucleotide blocked recombination and subsequent antigenic variation (Cahoon and Seifert, 2009).

More recently, G-quadruplexes have been found to be important in defining the origins of replication (Besnard et al., 2012; Hoshina et al., 2013; Valton et al., 2014). Furthermore G-quadruplexes have been implicated in transcription, RNA localization, translation, telomere protection and meiosis (see reference (Bochman et al., 2012) and references therein). Another argument that G-quadruplexes have a biological purpose is given by a genome-wide computational analysis, which shows that G4 motifs are evolutionary conserved in yeast (Capra et al., 2010).

G-quadruplexes and disease

The link between G-quadruplexes and disease is growing rapidly. One of the first diseases identified with a clear link to G-quadruplexes was the ATR-X syndrome, which is characterized by mental retardation and α-thalassaemia (Law et al., 2010). ATR-X was shown to bind G-quadruplexes and it was suggested that mutated ATR-X in combination with elevated G-quadruplex formation leads to epigenetic changes at the α-globin locus and subsequently anemia. A different recent study reported that G-quadruplex formation in DNA and RNA-molecules at the C9*orf*72 locus causes the neurodegenerative disease amyotrophic lateral sclerosis (ALS) (Haeusler et al., 2014). Another compelling study suggests that RNA-G-quadruplexes can control the translation of oncoproteins as MYC, NOTCH and BCL2 (Wolfe et al., 2014). Finally, since G-quadruplexes can drive genomic instability, G-quadruplexes have a clear link to cancer. Understanding how G-quadruplex structures can lead to genomic instability is therefore crucial.

MODEL ORGANISMS

Apart from human cell lines, I have used the nematode *Caenorhabditis elegans* and the zebrafish *Danio rerio* as model organisms for the work described in this thesis. Below, I will give a brief introduction about the strengths and limitations of *C. elegans* and *D. rerio*.

C. elegans

C. elegans is a multicellular organism of ± 1mm in size and consists of 959 cells. It has a short life cycle of approximately 3 days and a life span of three weeks. A fertilized egg develops into an adult worm via four larval stages named L1- L4. In the absence of food, larvae can switch to a stage called dauer, which allows the worm to survive up to several months. Worms can also be kept as frozen stock at -80 degrees for years, if not decades. In the laboratory the nematode is usually grown on agar plates or in liquid cultures and uses *E. coli* as a food source and is therefore relatively inexpensive to maintain. In a wildtype population two *C. elegans* sexes are found: the majority are self-fertilizing hermaphrodites and 0.1% are male.

C. elegans is highly appreciated for its genetics. The ability to self-fertilize enables the generation of genetically identical progeny, whereas the use of males allows the combination of mutations via crossings. Forward genetic screens can be easily performed via mutagenesis, whereas RNAi-based reverse genetic screens are applicable since the availability of genome-wide RNAi libraries (Kamath et al., 2003). Furthermore, many mutants are available and well documented at www.wormbase.org. The number of available mutant alleles was recently boosted by a so-called million mutation project in which more than 2000 mutagenized strains were sequenced, resulting in a library containing more than 800,000 unique single

nucleotide variants and 16,000 insertions/deletions (indels) (Thompson et al., 2013). Until recently, a limitation of *C. elegans* was the inability to perform targeted genome editing. However, this hurdle is now overcome by the availability of tools using zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs) and the clustered regulatory interspaced short palindromic repeat (CRISPR)/Cas9 system (Waaijers et al., 2013; Wood et al., 2011).

In 1998 the entire genome of *C. elegans* was sequenced and it quickly became apparent that many genes and pathways are strongly conserved. Also many genes involved in DNA repair are highly conserved and many human counterparts can be found back in the worm's genome as illustrated in table 1. Another feature that makes *C. elegans* an attractive model for the study of DNA repair is its germline, which has a spatio-temporal organization of mitotic and meiotic cells that can be easily monitored for the presence of damaged DNA (Lemmens and Tijsterman, 2011).

D. rerio

In recent years, the zebrafish has obtained a prominent role in biomedical research. Historically, the zebrafish has mostly been used to study developmental biology, however with evolving techniques the zebrafish has become also an excellent tool for studying disease mechanisms. Since the zebrafish is optically translucent for the first few weeks, the development from a single cell to a swimming fish can be monitored in great detail (see Figure 5a for the first 48 hours of embryonic development of a zebrafish). Recently, also transparent adult fish (Figure 5b) have been created, which makes it relatively easy to perform live *in vivo* imaging in full grown fish (White et al., 2008).

Within 3 months an embryo develops into a fertile adult, and its total life span usually varies between 2 to 3 years. On a weekly basis, females can spawn up to hundreds of eggs per clutch. These eggs can be easily collected and injected with transgenes or morpholinos (oligos that inhibit the expression of a protein). Until recent, targeted genome engineering was not possible and only forward genetic screens (mutagenesis or transposon-based) and target-selected mutagenesis (requiring extensive sequencing) (Wienholds et al., 2003) led to a relatively small library of mutants. However, in the last four years ZFN-, TALEN and CRISPR/Cas9-technology (Hwang et al., 2013) have proven to be exquisite tools for targeted genome editing in the zebrafish.

In the last decade, the zebrafish has been increasingly used as a model in cancer research (Amatruda and Patton, 2008). Many genetic cancer models and tools have been developed, but also xenotransplantations, in which human tumorigenic cells are injected, are becoming prevalent.

A drawback of working with zebrafish as a model organism is that it takes approximately five to six months before a homozygous animal is established, and that creating a homozygous mutant can sometimes be complicated because some segments of the zebrafish genome have been duplicated and thus for numerous genes there are duplicated copies.

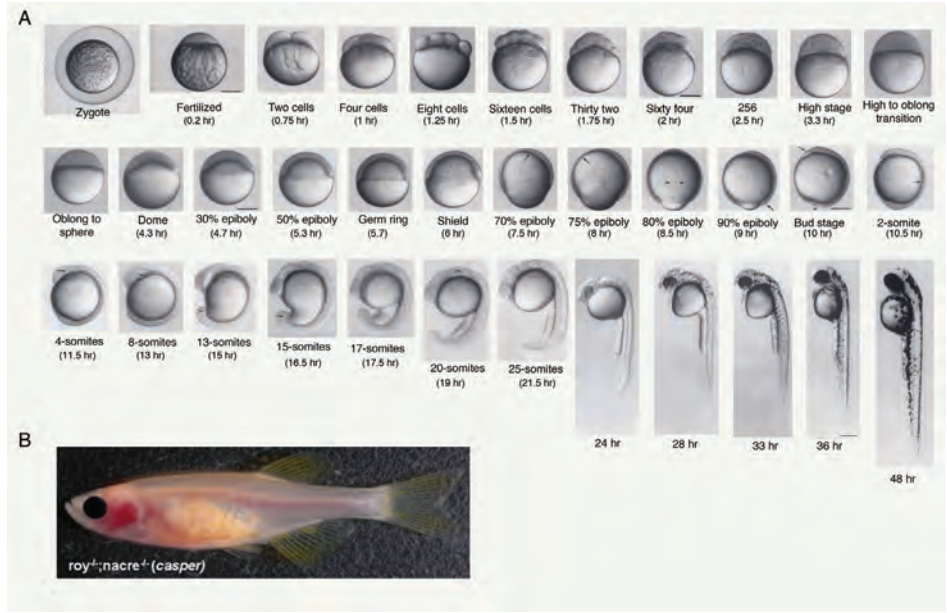


Figure 5 | The zebrafish *Danio rerio*. (A) Embryonic stages during zebrafish development. Figure adapted from Kimmel *et al*,1995. (B) Picture of an adult zebrafish of which the body is largely transparent due to the loss of melanocyte and iridophores. Figure adapted from White *et al*, 2008.

AIM OF THIS THESIS

Microsatellites and G-quadruplexes are sequences that are abundant in the genome and are linked to diseases such as cancer and neurodegenerative diseases. Unstable microsatellites and G-quadruplexes are thought to be an important underlying cause of these devastating diseases. However, many aspects about G4 DNA and microsatellite instability are incompletely understood. For example, what determines why some microsatellites and G-quadruplexes are more prone to induce mutations than others? Which genes and pathways prevent microsatellite and G4 DNA instability? What are the genetic consequences of microsatellite and G-quadruplex instability and which molecular mechanisms act to produce genomic changes at these sequences? The answers to these questions will be of great importance in the development of new and better treatments of microsatellite- and G4 DNA-related diseases. In this thesis I aim to provide new insights into the biology concerning microsatellite and G-quadruplex instability.

OUTLINE OF THIS THESIS

Chapter 2 concerns microsatellite instability. We investigate whether factors such as tract length, orientation, nucleotide composition and transcription influence the stability of a microsatellite. To this end, we make use of newly-developed reporters that read out microsatellite instability in human cells. We furthermore test whether these reporter systems can be used to screen for microsatellite instability-inducing compounds as well as for genes that protect the genome against microsatellite instability.

Chapter 3 presents a new genetic tool that enables mosaic analysis in the zebrafish. We developed technology that employs MSI, which we show to be a powerful tool to trace single cells and also to study tumor induction in a living animal.

Chapter 4 focuses on G4 DNA instability in *C. elegans*. Previous studies have shown that G-quadruplexes can induce deletions in the genome typically 50-300bp in size. In this chapter, we reveal the molecular mechanism that explains the formation of these deletions.

In **Chapter 5** we examine G4 DNA instability in human cells. We address the question whether G-quadruplexes are fragile in human cells as well and whether they induce deletions through a similar mechanism as witnessed in *C. elegans*.

In **Chapter 6** I provide a summarizing discussion and include a number of future perspectives related to microsatellite and G-quadruplex instability and their link to disease.

REFERENCES

ACS, A.C.S. (2013). Cancer Facts & Figures 2013. 1–64.

Amatruda, J.F., and Patton, E.E. (2008). Genetic models of cancer in zebrafish. International Review of Cell and Molecular Biology *271*, 1–34.

Arana, M.E., Seki, M., Wood, R.D., Rogozin, I.B., and Kunkel, T.A. (2008). Low-fidelity DNA synthesis by human DNA polymerase theta. Nucleic Acids Res. *36*, 3847–3856.

Asagoshi, K., Lehmann, W., Braithwaite, E.K., Santana-Santos, L., Prasad, R., Freedman, J.H., Van Houten, B., and Wilson, S.H. (2012). Single-nucleotide base excision repair DNA polymerase activity in C. elegans in the absence of DNA polymerase β. Nucleic Acids Res. *40*, 670–681.

Ballew, B.J., Yeager, M., Jacobs, K., Giri, N., Boland, J., Burdett, L., Alter, B.P., and Savage, S.A. (2013). Germline mutations of regulator of telomere elongation helicase 1, RTEL1, in Dyskeratosis congenita. Hum. Genet. *132*, 473–480.

Bang, I. (1910). Untersuchungen über die Guanylsäure. Biochemische Zeitschrift.

Barber, L.J., Youds, J.L., Ward, J.D., McIlwraith, M.J., O'Neil, N.J., Petalcorin, M.I.R., Martin, J.S., Collis, S.J., Cantor, S.B., Auclair, M., et al. (2008). RTEL1 maintains genomic stability by suppressing homologous recombination. Cell *135*, 261–271.

Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J.-M., and Lemaitre, J.-M. (2012). Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nature Structural & Molecular Biology *19*, 837–844.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M.C., Tassani, S., Piva, F., et al. (2013). An estimation of the number of cells in the human body. Ann. Hum. Biol. *40*, 463–471.

Biffi, G., Tannahill, D., McCafferty, J., and Balasubramanian, S. (2013). Quantitative visualization of DNA G-quadruplex structures in human cells. Nature Chemistry *5*, 182–186.

Biffi, G., Tannahill, D., Miller, J., Howat, W.J., and Balasubramanian, S. (2014). Elevated Levels of G-Quadruplex Formation in Human Stomach and Liver Cancer Tissues. PLoS ONE *9*, e102711.

Bochman, M.L., Paeschke, K., and Zakian, V.A. (2012). DNA secondary structures: stability and function of G-quadruplex structures. Nat. Rev. Genet. 1–11.

Bosch, P.C., Segura-Bayona, S., Koole, W., van Heteren, J.T., Dewar, J.M., Tijsterman, M., and Knipscheer, P. (2014). FANCJ promotes DNA synthesis through G-quadruplex structures. Embo J *33*, 2521–2533.

Boulton, J.N.C.M.G.T.S., Taylor, M.R.G., and Boulton, S.J. (2012). Playing the End Game: DNA Double-Strand Break Repair Pathway Choice. Molecular Cell *47*, 497–510.

Brosh, R.M. (2013). DNA helicases involved in DNA repair and their roles in cancer. Nature Publishing Group *13*, 542–558.

Cahoon, L.A., and Seifert, H.S. (2009). An alternative DNA structure is necessary for pilin antigenic variation in Neisseria gonorrhoeae. Science *325*, 764–767.

Cannavo, E., and Cejka, P. (2014). Sae2 promotes dsDNA endonuclease activity within Mre11-Rad50-Xrs2 to resect DNA breaks. Nature *514*, 122–125.

Cantor, S.B., Bell, D.W., Ganesan, S., Kass, E.M., Drapkin, R., Grossman, S., Wahrer, D.C., Sgroi, D.C., Lane, W.S., Haber, D.A., et al. (2001). BACH1, a Novel Helicase-like Protein, Interacts Directly with BRCA1 and Contributes to Its DNA Repair Function. Cell *105*, 149–160.

Capra, J.A., Paeschke, K., Singh, M., and Zakian, V.A. (2010). G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae. PLoS Comput. Biol. *6*, e1000861.

Chan, S.H., Yu, A.M., and McVey, M. (2010). Dual roles for DNA polymerase theta in alternative end-joining repair of double-strand breaks in Drosophila. PLoS Genet. *6*, e1001005.

Chen, X., Wilson, J.B., McChesney, P., Williams, S.A., Kwon, Y., Longerich, S., Marriott, A.S., Sung, P., Jones, N.J., and Kupfer, G.M. (2014). The Fanconi anemia proteins FANCD2 and FANCJ interact and regulate each other's chromatin localization. J. Biol. Chem. *289*, 25774–25782.

Cheung, I., Schertzer, M., Rose, A., and Lansdorp, P.M. (2002). Disruption of dog-1 in Caenorhabditis elegans triggers deletions upstream of guanine-rich DNA. Nat. Genet. *31*, 405–409.

Chu, W.K., and Hickson, I.D. (2009). RecQ helicases: multifunctional genome caretakers. Nature Publishing Group *9*, 644–654.

Ciccia, A., and Elledge, S.J. (2010). The DNA Damage Response: Making It Safe to Play with Knives. Molecular Cell *40*, 179–204.

Cimprich, K.A., and Cortez, D. (2008). ATR: an essential regulator of genome integrity. Nat. Rev. Mol. Cell Biol. *9*, 616–627.

Cleaver, J.E., Lam, E.T., and Revet, I. (2009). Disorders of nucleotide excision repair: the genetic and molecular basis of heterogeneity. Nat. Rev. Genet. *10*, 756–768.

De, S., and Michor, F. (2011). DNA secondary structures and epigenetic determinants of cancer genome evolution. Nature Structural & Molecular Biology *18*, 950–955.

Deans, A.J., and West, S.C. (2011). DNA interstrand crosslink repair and cancer. Nature Publishing Group *11*, 467–480.

Decottignies, A. (2012). Alternative end-joining mechanisms: a historical perspective. Front Genet *4*, 48–48.

Deriano, L., and Roth, D.B. (2012). Modernizing the nonhomologous end-joining repertoire: alternative and classical NHEJ share the stage. Genetics *47*, 433–455.

Fernandez-Vidal, A., Guitton-Sert, L., Cadoret, J.-C., Drac, M., Schwob, E., Baldacci, G., Cazaux, C., and Hoffmann, J.-S. (2014). A role for DNA polymerase θ in the timing of DNA replication. Nature Communications *5*, 4285.

Fry, M., and Loeb, L.A. (1999). Human werner syndrome DNA helicase unwinds tetrahelical structures of the fragile X syndrome repeat sequence d(CGG)n. J. Biol. Chem. *274*, 12797–12802.

Garaycoechea, J.I., and Patel, K.J. (2014). Why does the bone marrow fail in Fanconi anemia? Blood *123*, 26–34.

Garaycoechea, J.I.J., Crossan, G.P.G., Langevin, F.F., Daly, M.M., Arends, M.J.M., and Patel, K.J.K. (2012). Genotoxic consequences of endogenous aldehydes on mouse haematopoietic stem cell function. Nature *489*, 571–575.

García-Gómez, S., Reyes, A., Martínez-Jiménez, M.I., Chocrón, E.S., Mourón, S., Terrados, G., Powell, C., Salido, E., Méndez, J., Holt, I.J., et al. (2013). PrimPol, an Archaic Primase/Polymerase Operating in Human Cells. Molecular Cell *52*, 541–553.

Gellert, M., Lipsett, M.N., and Davies, R.D. (1962). Helix formation by guanylic acid. Proc. Natl. Acad. Sci. U.S.a. *48*, 2013–2018.

Gemayel, R., Vinces, M.D., Legendre, M., and Verstrepen, K.J. (2010). Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. Genetics *44*, 445–477.

Ghodgaonkar, M.M., Lazzaro, F., Olivera-Pimentel, M., Artola-Borán, M., Cejka, P., Reijns, M.A., Jackson, A.P., Plevani, P., Muzi-Falconi, M., and Jiricny, J. (2013). Ribonucleotides misincorporated into DNA act as strand-discrimination signals in eukaryotic mismatch repair. Molecular Cell *50*, 323–332.

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

Gray, L.T., Vallur, A.C., Eddy, J., and Maizels, N. (2014). G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD. Nat. Chem. Biol. *10*, 313–318.

Guillemette, S., Branagan, A., Peng, M., Dhruva, A., Schärer, O.D., and Cantor, S.B. (2014). FANCJ localization by mismatch repair is vital to maintain genomic integrity after UV irradiation. Cancer Res. *74*, 932–944.

Haeusler, A.R., Donnelly, C.J., Periz, G., Simko, E.A.J., Shaw, P.G., Kim, M.-S., Maragakis, N.J., Troncoso, J.C., Pandey, A., Sattler, R., et al. (2014). C9orf72 nucleotide repeat structures initiate molecular cascades of disease. Nature *507*, 195–200.

Harris, P.V., Mazina, O.M., Leonhardt, E.A., Case, R.B., Boyd, J.B., and Burtis, K.C. (1996). Molecular cloning of Drosophila mus308, a gene involved in DNA cross-link repair with homology to prokaryotic DNA polymerase I genes. Mol. Cell. Biol. *16*, 5764–5771.

Henderson, A., Wu, Y., Huang, Y.C., Chavez, E.A., Platt, J., Johnson, F.B., Brosh, R.M., Sen, D., and Lansdorp, P.M. (2013). Detection of G-quadruplex DNA in mammalian cells. Nucleic Acids Res. *42*, 860–869.

Higgins, G.S., Harris, A.L., Prevo, R., Helleday, T., McKenna, W.G., and Buffa, F.M. (2010). Overexpression of POLQ confers a poor prognosis in early breast cancer patients. Oncotarget *1*, 175–184.

Hiom, K. (2009). FANCJ: Solving problems in DNA replication. DNA Repair *9*, 250–256.

Hira, A., Yabe, H., Yoshida, K., Okuno, Y., Shiraishi, Y., Chiba, K., Tanaka, H., Miyano, S., Nakamura, J., Kojima, S., et al. (2013). Variant ALDH2 is associated with accelerated progression of bone marrow failure in Japanese Fanconi anemia patients. Blood *122*, 3206–3209.

Hoeijmakers, J.H.J. (2009). DNA damage, aging, and cancer. N. Engl. J. Med. *361*, 1475–1485.

Hoshina, S., Yura, K., Teranishi, H., Kiyasu, N., Tominaga, A., Kadoma, H., Nakatsuka, A., Kunichika, T., Obuse, C., and Waga, S. (2013). Human origin recognition complex binds preferentially to G-quadruplex-preferable RNA and single-stranded DNA. J. Biol. Chem. *288*, 30161–30171.

Hwang, W.Y., Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Peterson, R.T., Yeh, J.-R.J., and Joung, J.K. (2013). Efficient genome editing in zebrafish using a CRISPR-Cas system. Nat. Biotechnol. *31*, 227–229.

Jackson, S.P., and Bartek, J. (2009). The DNA-damage response in human biology and disease. Nature *461*, 1071–1078.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature *421*, 231–237.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.

Katti, M.V., Ranjekar, P.K., and Gupta, V.S. (2001). Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evol. *18*, 1161–1167.

Kawamura, K., Bahar, R., Seimiya, M., Chiyo, M., Wada, A., Okada, S., Hatano, M., Tokuhisa, T., Kimura, H., Watanabe, S., et al. (2004). DNA polymerase theta is preferentially expressed in lymphoid tissues and upregulated in human cancers. Int. J. Cancer *109*, 9–16.

Klein Douwel, D., Boonen, R.A.C.M., Long, D.T., Szypowska, A.A., Räschle, M., Walter, J.C., and Knipscheer, P. (2014). XPF-ERCC1 acts in Unhooking DNA interstrand crosslinks in cooperation with FANCD2 and FANCP/SLX4. Molecular Cell *54*, 460–471.

Knipscheer, P., Räschle, M., Smogorzewska, A., Enoiu, M., Ho, T.V., Schärer, O.D., Elledge, S.J., and Walter, J.C. (2009). The Fanconi anemia pathway promotes replication-dependent DNA interstrand cross-link repair. Science *326*, 1698–1701.

Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. Nature *488*, 471–475.

Kottemann, M.C., and Smogorzewska, A. (2013). Fanconi anaemia and the repair of Watson and Crick DNA crosslinks. Nature *493*, 356–363.

Kruisselbrink, E., Guryev, V., Brouwer, K., Pontier, D.B., Cuppen, E., and Tijsterman, M. (2008). Mutagenic capacity of endogenous G4 DNA underlies genome instability in FANCJ-defective C. elegans. Curr. Biol. *18*, 900–905.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Langevin, F.F., Crossan, G.P.G., Rosado, I.V.I., Arends, M.J.M., and Patel, K.J.K. (2011). Fancd2 counteracts the toxic effects of naturally produced aldehydes in mice. Nature *475*, 53–58.

Law, M.J., Lower, K.M., Voon, H.P.J., Hughes, J.R., Garrick, D., Viprakasit, V., Mitson, M., De Gobbi, M., Marra, M., Morris, A., et al. (2010). ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner. Cell *143*, 367–378.

Lemée, F., Bergoglio, V., Fernandez-Vidal, A., Machado-Silva, A., Pillaire, M.-J., Bieth, A., Gentil, C., Baker, L., Martin, A.-L., Leduc, C., et al. (2010). DNA polymerase theta up-regulation is associated with poor survival in breast cancer, perturbs DNA replication, and promotes genetic instability. Proc. Natl. Acad. Sci. U.S.a. *107*, 13390–13395.

Lemmens, B.B.L.G., and Tijsterman, M. (2011). DNA double-strand break repair in Caenorhabditis elegans. Chromosoma *120*, 1–21.

Levitus, M., Waisfisz, Q., Godthelp, B.C., de Vries, Y., Hussain, S., Wiegant, W.W., Elghalbzouri-Maghrani, E., Steltenpool, J., Rooimans, M.A., Pals, G., et al. (2005). The DNA helicase BRIP1 is defective in Fanconi anemia complementation group J. Nat. Genet. *37*, 934–935.

Li, F., Mao, G., Tong, D., Huang, J., Gu, L., Yang, W., and Li, G.-M. (2013). The Histone Mark H3K36me3 Regulates Human DNA Mismatch Repair through Its Interaction with MutS&alpha. Cell *153*, 590–600.

Li, Y.-C., Korol, A.B., Fahima, T., Beiles, A., and Nevo, E. (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol. Ecol. *11*, 2453–2465.

Lieber, M.R. (2010). The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. Annu. Rev. Biochem. *79*, 181–211.

Lin, W., Sampathi, S., Dai, H., Liu, C., Zhou, M., Hu, J., Huang, Q., Campbell, J., Shin-ya, K., Zheng, L., et al. (2013). Mammalian DNA2 helicase/nuclease cleaves G-quadruplex DNA and is required for telomere integrity. Embo J *32*, 1425–1439.

Lindahl, T. (1974). An N-glycosidase from Escherichia coli that releases free uracil from DNA containing deaminated cytosine residues. Proc. Natl. Acad. Sci. U.S.a. *71*, 3649–3653.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. Nature *362*, 709–715.

Lindahl, T., and Barnes, D.E. (2000). Repair of endogenous DNA damage. Cold Spring Harb Symp Quant Biol *65*, 127–133.

Lindahl, T., and Nyberg, B. (1972). Rate of Depurination of Native Deoxyribonucleic Acid. Biochemistry *11*, 3610–&.

Lujan, S.A., Williams, J.S., Clausen, A.R., Clark, A.B., and Kunkel, T.A. (2013). Ribonucleotides Are Signals for Mismatch Repair of Leading-Strand Replication Errors. Molecular Cell 1–7.

Lynch, H.T., Lynch, P.M., Lanspa, S.J., Snyder, C.L., Lynch, J.F., and Boland, C.R. (2009). Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. Clin. Genet. *76*, 1–18.

Mahaney, B.L., Meek, K., and Lees-Miller, S.P. (2009). Repair of ionizing radiation-induced DNA double-strand breaks by non-homologous end-joining. Biochem J *417*, 639–650.

Matsuoka, S., Ballif, B.A., Smogorzewska, A., McDonald, E.R., Hurov, K.E., Luo, J., Bakalarski, C.E., Zhao, Z., Solimini, N., Lerenthal, Y., et al. (2007). ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. Science *316*, 1160–1166.

McElhinny, S.A.N., Kissling, G.E., and Kunkel, T.A. (2010). Differential correction of lagging-strand replication errors made by DNA polymerases {alpha} and {delta}. Proc. Natl. Acad. Sci. U.S.a. *107*, 21070–21075.

Mourón, S., Rodriguez-Acebes, S., Martínez-Jiménez, M.I., García-Gómez, S., Chocrón, S., Blanco, L., and Méndez, J. (2013). Repriming of DNA synthesis at stalled replication forks by human PrimPol. Nature Structural & Molecular Biology *20*, 1383–1389.

Muzi-Falconi, M., Giannattasio, M., Foiani, M., and Plevani, P. (2003). The DNA polymerase alpha-primase complex: multiple functions and interactions. ScientificWorldJournal *3*, 21–33.

Nick McElhinny, S.A., Kumar, D., Clark, A.B., Watt, D.L., Watts, B.E., Lundström, E.-B., Johansson, E., Chabes, A., and Kunkel, T.A. (2010). Genome instability due to ribonucleotide incorporation into DNA. Nat. Chem. Biol. *6*, 774–781.

O'Brien, P.J., Siraki, A.G., and Shangari, N. (2005). Aldehyde sources, metabolism, molecular toxicity mechanisms, and possible effects on human health. Crit Rev Toxicol *35*, 609–662.

Pascucci, B., Stucki, M., Jónsson, Z.O., Dogliotti, E., and Hübscher, U. (1999). Long patch base excision repair with purified human proteins. DNA ligase I as patch size mediator for DNA polymerases delta and epsilon. J. Biol. Chem. *274*, 33696–33702.

Peña-Diaz, J., and Jiricny, J. (2012). Mammalian mismatch repair: error-free or error-prone? Trends in Biochemical Sciences *37*, 206–214.

Piazza, A., Serero, A., Boulé, J.-B., Legoix-Né, P., Lopes, J., and Nicolas, A. (2012). Stimulation of Gross Chromosomal Rearrangements by the Human CEB1 and CEB25 Minisatellites in Saccharomyces cerevisiae Depends on G-Quadruplexes or Cdc13. PLoS Genet. *8*, e1003033.

Pines, A., Mullenders, L.H., van Attikum, H., and Luijsterburg, M.S. (2013). Touching base with PARPs: moonlighting in the repair of UV lesions and double-strand breaks. Trends in Biochemical Sciences *38*, 321–330.

Prasad, R., Longley, M.J., Sharief, F.S., Hou, E.W., Copeland, W.C., and Wilson, S.H. (2009). Human DNA polymerase theta possesses 5'-dRP lyase activity and functions in single-nucleotide base excision repair in vitro. Nucleic Acids Res. *37*, 1868–1877.

Ribeyre, C., Lopes, J., Boulé, J.-B., Piazza, A., Guédin, A., Zakian, V.A., Mergny, J.-L., and Nicolas, A. (2009). The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. PLoS Genet. *5*, e1000475–e1000475.

Robertson, A.B., Klungland, A., Rognes, T., and Leiros, I. (2009). DNA repair in mammalian cells: Base excision repair: the long and short of it. Cell Mol Life Sci *66*, 981–993.

Rodriguez, R., Miller, K.M., Forment, J.V., Bradshaw, C.R., Nikan, M., Britton, S., Oelschlaegel, T., Xhemalce, B., Balasubramanian, S., and Jackson, S.P. (2012). Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. Nat. Chem. Biol. *8*, 301–310.

Rosado, I.V., Langevin, F., Crossan, G.P., Takata, M., and Patel, K.J. (2011). Formaldehyde catabolism is essential in cells deficient for the Fanconi anemia DNA-repair pathway. Nature Structural & Molecular Biology *18*, 1432–1434.

Sale, J.E., Lehmann, A.R., and Woodgate, R. (2012). Y-family DNA polymerases and their role in tolerance of cellular DNA damage. Nature Publishing Group *13*, 141–152.

Sarkies, P., Murat, P., Phillips, L.G., Patel, K.J., Balasubramanian, S., and Sale, J.E. (2012). FANCJ coordinates two pathways that maintain epigenetic stability at G-quadruplex DNA. Nucleic Acids Res. *40*, 1485–1498.

Sarkies, P., Reams, C., Simpson, L.J., and Sale, J.E. (2010). Epigenetic instability due to defective replication of structured DNA. Molecular Cell *40*, 703–713.

Schwab, R.A., Nieminuszczy, J., Shin-ya, K., and Niedzwiedz, W. (2013). FANCJ couples replication past natural fork barriers with maintenance of chromatin structure. J. Cell Biol. *201*, 33–48.

Seki, M., Marini, F., and Wood, R.D. (2003). POLQ (Pol θ), a DNA polymerase and DNA-dependent ATPase in human cells. Nucleic Acids Res. *31*, 6117–6126.

Seki, M., Masutani, C., Yang, L.W., Schuffert, A., Iwai, S., Bahar, I., and Wood, R.D. (2004). High-efficiency bypass of DNA damage by human DNA polymerase Q. Embo J *23*, 4484–4494.

Shiloh, Y. (2003). ATM and related protein kinases: safeguarding genome integrity. Nat Rev Cancer *3*, 155–168.

Shima, N., Munroe, R.J., and Schimenti, J.C. (2004). The mouse genomic instability mutation chaos1 is an allele of Polq that exhibits genetic interaction with Atm. Mol. Cell. Biol. *24*, 10381–10389.

Siegel, E.C., and Bryson, V. (1963). Selection of resistant strains of Escherichia coli by antibiotics and antibacterial agents: role of normal and mutator strains. Antimicrob Agents Chemother (Bethesda) *161*, 629–634.

Singleton, M.R., Dillingham, M.S., and Wigley, D.B. (2007). Structure and mechanism of helicases and nucleic acid translocases. Annu. Rev. Biochem. *76*, 23–50.

Sommers, J.A., Banerjee, T., Hinds, T., Wan, B., Wold, M.S., Lei, M., and Brosh, R.M. (2014). Novel Function of the Fanconi Anemia Group J or RECQ1 Helicase to Disrupt Protein-DNA Complexes in a Replication Protein A-stimulated Manner. Journal of Biological Chemistry *289*, 19928–19941.

Suhasini, A.N., Sommers, J.A., Muniandy, P.A., Coulombe, Y., Cantor, S.B., Masson, J.-Y., Seidman, M.M., and Brosh, R.M. (2013). Fanconi Anemia Group J Helicase and MRE11 Nuclease Interact To Facilitate the DNA Damage Response. Mol. Cell. Biol. *33*, 2212–2227.

Sun, H.H., Karow, J.K.J., Hickson, I.D.I., and Maizels, N.N. (1998). The Bloom's syndrome helicase unwinds G4 DNA. J. Biol. Chem. *273*, 27587–27592.

Symington, L.S. (2014). DNA repair: Making the cut. Nature *514*, 39–40.

Takata, K.I., Shimizu, T., Iwai, S., and Wood, R.D. (2006). Human DNA Polymerase N (POLN) Is a Low Fidelity Enzyme Capable of Error-free Bypass of 5S-Thymine Glycol. Journal of Biological Chemistry *281*, 23445–23455.

Thompson, O., Edgley, M., Strasbourger, P., Flibotte, S., Ewing, B., Adair, R., Au, V., Chaudhry, I., Fernando, L., Hutter, H., et al. (2013). The million mutation project: a new approach to genetics in Caenorhabditis elegans. Genome Res. *23*, 1749–1762.

Uchiyama, Y., Takeuchi, R., Kodera, H., and Sakaguchi, K. (2009). Distribution and roles of X-family DNA polymerases in eukaryotes. Biochimie *91*, 165–170.

Umate, P., Tuteja, N., and Tuteja, R. (2011). Genome-wide comprehensive analysis of human helicases. Commun Integr Biol *4*, 118–137.

Valton, A.-L., Hassan-Zadeh, V., Lema, I., Boggetto, N., Alberti, P., Saintomé, C., Riou, J.-F., and Prioleau, M.-N. (2014). G4 motifs affect origin positioning and efficiency in two vertebrate replicators. Embo J *33*, 732–746.

van der Lelij, P., Chrzanowska, K.H., Godthelp, B.C., Rooimans, M.A., Oostra, A.B., Stumm, M., Zdzienicka, M.Z., Joenje, H., and de Winter, J.P. (2010). Warsaw Breakage Syndrome, a Cohesinopathy Associated with Mutations in the XPD Helicase Family Member DDX11/ChlR1. The American Journal of Human Genetics *86*, 262–266.

Vannier, J.B., Sandhu, S., Petalcorin, M.I., Wu, X., Nabi, Z., Ding, H., and Boulton, S.J. (2013). RTEL1 Is a Replisome-Associated Helicase That Promotes Telomere and Genome-Wide Replication. Science *342*, 239–242.

Vannier, J.-B., Pavicic-Kaltenbrunner, V., Petalcorin, M.I.R., Ding, H., and Boulton, S.J. (2012). RTEL1 dismantles T loops and counteracts telomeric G4-DNA to maintain telomere integrity. Cell *149*, 795–806.

Waaijers, S., Portegijs, V., Kerver, J., Lemmens, B.B.L.G., Tijsterman, M., van den Heuvel, S., and Boxem, M. (2013). CRISPR/Cas9-Targeted Mutagenesis in Caenorhabditis elegans. Genetics *195*, 1187–1191.

Wang, M., Wu, W., Wu, W., Rosidi, B., Zhang, L., Wang, H., and Iliakis, G. (2006). PARP-1 and Ku compete for repair of DNA double strand breaks by distinct NHEJ pathways. Nucleic Acids Res. *34*, 6170–6182.

White, R.M., Sessa, A., Burke, C., Bowman, T., LeBlanc, J., Ceol, C., Bourque, C., Dovey, M., Goessling, W., Burns, C.E., et al. (2008). Transparent adult zebrafish as a tool for in vivo transplantation analysis. Cell Stem Cell *2*, 183–189.

Wienholds, E., van Eeden, F., Kosters, M., Mudde, J., Plasterk, R.H.A., and Cuppen, E. (2003). Efficient target-selected mutagenesis in zebrafish. Genome Res. *13*, 2700–2707.

Wilson, M.A., Kwon, Y., Xu, Y., Chung, W.-H., Chi, P., Niu, H., Mayle, R., Chen, X., Malkova, A., Sung, P., et al. (2013). Pif1 helicase and Polδ promote recombination-coupled DNA synthesis via bubble migration. Nature *502*, 393–396.

Wogan, G.N., Hecht, S.S., Felton, J.S., Conney, A.H., and Loeb, L.A. (2004). Environmental and chemical carcinogenesis. Semin Cancer Biol *14*, 473–486.

Wolfe, A.L., Singh, K., Zhong, Y., Drewe, P., Rajasekhar, V.K., Sanghvi, V.R., Mavrakis, K.J., Jiang, M., Roderick, J.E., Van der Meulen, J., et al. (2014). RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. Nature *513*, 65–70.

Wood, A.J., Lo, T.-W., Zeitler, B., Pickle, C.S., Ralston, E.J., Lee, A.H., Amora, R., Miller, J.C., Leung, E., Meng, X., et al. (2011). Targeted genome editing across species using ZFNs and TALENs. Science *333*, 307.

Wu, Y.Y., and Brosh, R.M.R. (2012). DNA helicase and helicase-nuclease enzymes with a conserved iron-sulfur cluster. Nucleic Acids Res. *40*, 4247–4260.

Yamtich, J., and Sweasy, J.B. (2010). DNA polymerase Family X: Function, structure, and cellular roles. Biochimica Et Biophysica Acta (BBA) - Proteins and Proteomics *1804*, 1136–1150.

Yoon, J.-H., Roy Choudhury, J., Park, J., Prakash, S., and Prakash, L. (2014). A role for DNA polymerase θ in promoting replication through oxidative DNA lesion, thymine glycol, in human cells. Journal of Biological Chemistry *289*, 13177–13185.

Yoshimura, M., Kohzaki, M., Nakamura, J., Asagoshi, K., Sonoda, E., Hou, E., Prasad, R., Wilson, S.H., Tano, K., Yasui, A., et al. (2006). Vertebrate POLQ and POLbeta cooperate in base excision repair of oxidative DNA damage. Molecular Cell *24*, 115–125.

Youds, J.L., Barber, L.J., Ward, J.D., Collis, S.J., O'Neil, N.J., Boulton, S.J., and Rose, A.M. (2008). DOG-1 is the Caenorhabditis elegans BRIP1/FANCJ homologue and functions in interstrand cross-link repair. Mol. Cell. Biol. *28*, 1470–1479.

Yousefzadeh, M.J., Wyatt, D.W., Takata, K.-I., Mu, Y., Hensley, S.C., Tomida, J., Bylund, G.O., Doublié, S., Johansson, E., Ramsden, D.A., et al. (2014). Mechanism of suppression of chromosomal instability by DNA polymerase POLQ. PLoS Genet. *10*, e1004654–e1004654.

◀ This photo shows the head of a stained worm (*C. elegans*). The blue staining in a cell reports the presence of a mutation in which a G-quadruplex (an unusually folded piece of DNA) and surrounding DNA was deleted. In this photo the blue cells mark the pharynx (foregut) of the worm.

Cover Photo: Plasmid 91 ▶

Chapter 2

A versatile microsatellite instability reporter system in human cells

Wouter Koole[1*], Henning S. Schäfer[2*], Reuven Agami[2], Gijs van Haaften[2], Marcel Tijsterman[1]

*These authors contributed equally to this work
[1] Department of Toxicogenetics, Leiden University Medical Center, Leiden, The Netherlands
[2] Division of Gene Regulation, The Netherlands Cancer Institute, Amsterdam, The Netherlands

Adapted from Koole et al. Nucleic Acids Res. 41, e158 (2013)

ABSTRACT

Here, we report the investigation of microsatellite instability (MSI) in human cells with a newly developed reporter system based on fluorescence. We composed a vector into which microsatellites of different lengths and nucleotide composition can be introduced between a functional copy of the fluorescent protein mCherry and an out-of-frame copy of EGFP; *in vivo* frameshifting will lead to EGFP expression, which can be quantified by Fluorescence Activated Cell Sorting (FACS). Via targeted recombineering, single copy reporters were introduced in HEK293 and MCF-7 cells. We found predominantly -1 and +1 base pair frameshifts, the levels of which are kept in tune by mismatch repair. We show that tract length and composition greatly influences MSI. In contrast, a tracts' potential to form a G-quadruplex structure, its strand orientation or its transcriptional status is not affecting MSI. We further validated the functionality of the reporter system for screening microsatellite mutagenicity of compounds and for identifying modifiers of MSI: using a retroviral miRNA expression library, we identified miR-21, which targets MSH2, as a miRNA that induces MSI when overexpressed. Our data also provide proof of principle for the strategy of combining fluorescent reporters with next generation sequencing technology to identify genetic factors in specific pathways.

INTRODUCTION

The human genome is full of DNA repeats. One abundant class of repeats, making up for ~3% of the human genome (Lander et al., 2001), are microsatellites, which are often defined as repetitive runs of DNA sequences consisting of 1-8 base pairs (bp) long units (Richard et al., 2008). Soon after their discovery in the early 1980s it became apparent that these tandem repeats are highly polymorphic in length, and have mutation rates even up to 10^{-2} per locus per generation (Ellegren, 2004). It is their repetitive nature that makes microsatellites prone to mutagenesis; due to strand slippage during DNA replication or unequal recombination, microsatellites can expand or contract. Microsatellites can be found all over the genome, present even in protein-coding sequences (Woerner et al., 2009). In fact, 17% of human genes contain tandem repeats in their open reading frames (ORFs) (Gemayel et al., 2010), and microsatellites have been shown to affect biological processes such as chromatin organization, recombination, DNA replication, transcription and translation (reviewed in (Li et al., 2002)). It is therefore of no surprise that microsatellites are thought to play a significant role in evolution, and that many diseases, including several neurodegenerative diseases, and cancer are linked to variations in the length of genomic microsatellites.

The stability of microsatellites is influenced by several factors. An important factor is the status of Mismatch Repair (MMR). This pathway is well-conserved among species and consist of a delicate interplay of many proteins (for a review see for example (Peña-Diaz and Jiricny, 2012) and references therein). In brief, mis-incorporated nucleotides or small insertion-deletions loops are recognized by a heterodimeric protein complex consisting of MSH2 and MSH3 or MSH6. These mutS complexes interact with the mutL proteins MLH1 and PMS2, which are essential for incision and subsequent removal by EXO1 of the newly synthesized DNA. Numerous other proteins (e.g. PCNA, RFC, polymerase-δ, RPA and DNA ligase I) are required to complete the faithful repair of a mismatch or loop. Another important determinant that affects the stability of microsatellites is the length (the number of repeat-units) of the tract. While a correlation between the length of the microsatellite and the mutation rate has been seen in numerous organisms (Beck et al., 2003; Brinkmann et al., 1998; Brohede et al., 2002; Harr and Schlötterer, 2000; Primmer et al., 1998; Wierdl et al., 1997; Yamada et al., 2002), thus far there is no consensus whether this is a linear, quadratic or exponential relationship (Bhargava and Fuentes, 2009; Brinkmann et al., 1998; Lai and Sun, 2003). Also, the genomic environment of the microsatellite is an important determinant for MSI: ample evidence exists that the locus where the microsatellite is situated is greatly affecting its stability (Harfe and Jinks-Robertson, 2000; Hawk et al., 2005; Kondrashov and Rogozin, 2004; Ma et al., 2012). For example, a recent report showed that the presence of other repeats in close proximity of a microsatellite decreases its stability (Ma et al., 2012). Other factors like nucleotide composition, possible formation of secondary structures such as G-quadruplex structures and

levels of transcription of the locus have also been implicated in the stability of microsatellites (as reviewed in ref (Bhargava and Fuentes, 2009)).

Many aspects on microsatellite dynamics have been studied in a plethora of organisms. However, several aspects have not been addressed in human cells, despite the notion that microsatellite dynamics clearly vary between organisms (even between humans and chimpanzees) (Webster et al., 2002). To gain full insight into MSI in human cells, we developed an experimental setup that is able to quantify MSI in human cells. We monitor MSI using a modular fluorescent reporter system in combination with FACS. To exclude the influence of the genomic environment, we targeted different microsatellites to the same genomic locus. We addressed the influence of length, orientation, nucleotide composition, secondary structure, the transcriptional status of the locus as well as compound exposure. In addition, we show how this system aids to identify and characterize genetic regulators of MSI by assaying ~450 miRNAs. This methodology can be easily adapted to read out other genome instability phenotypes in mammalian cells to uncover novel regulators in a specific pathway.

MATERIAL AND METHODS

Plasmid construction and sequencing

Standard molecular cloning techniques were used to obtain the constructs described in this manuscript. Briefly, using polymerase chain reaction (PCR), we amplified 3 DNA fragments: mCherry (from plasmid pRSET-B mCherry) without termination codon, flanked by a NheI and a HindIII restriction-site, a coding stuffer fragment of 215 basepairs (bp) flanked by a BamHI and a EcoRI restriction site, EGFP without start-codon flanked by EcoRI and EcoRV restriction sites (from pEGFP-N2, Clontech). Pieces were sequentially cloned into plasmid pcDNA5/FRT/TO (Life Technologies). Microsatellites were subsequently placed into the HindIII and BamHI site using oligo-cloning. All plasmids sequences were checked by Sanger sequencing according to standard procedures, but with the following adjustments: we used a 1:3 mix of ABI Prism dGTP BigDye terminator v3.0 and BigDye terminator v3.0, 200ng plasmid per sequencing reaction, and no more than 25 cycles.

Cell culture

Flp-in T-Rex-293 cells (Life Technologies) were cultured on poly-l-lysine coated surface at 37°C and 5% $C0_2$ in DMEM (ref 41966, Gibco) supplemented with 10% fetal bovine serum (Bodinco BV) and 1% penicillin-streptomycin. Stable polyclonal Flp-in T-Rex 293 cell lines were generated via integration of the plasmids using Flp recombinase-mediated DNA recombination according to manufacturer's protocol. Cells were cultured in the presence of 100µg/ml hygromycin (ref 10843555001, Roche) and 15µg/ml blasticidin (ref A11139-03, Gibco) and 0,1µg/ml doxycycline hyclate (ref D9891, Sigma). MCF-7 cells with a single copy integrated pFRT/*lac*Zeo and the murine ecotrophic receptor were cultured under similar conditions, however selected with hygromycin and neomycin.

Transient knockdown of MSH2

150.000 Cells were seeded per well (12-well plate) per condition. Cells were transfected with a mix of 4ul DharmaFECT1, 5ul 20µM ONTARGETplus smartpool MSH2 siRNA, or non-targeting siGENOME Control pool (REF T-2001-03, L-003909-00, D-001206-13-05, Thermo Scientific Dharmacon) and 200ul optimem after 24 hours. 48 hours later, cells were sorted and 25.000 mCherry+EGFP- cells per well were seeded in a 96 well plate. A second round of knockdown was performed 24 hours after and cells were analysed 6 days later.

Fragment analysis

mCherry-EGFP- cells were isolated from HEK 293 cells carrying plm405 (G_{14}-repeat) by FACS. 48 Hours later flow cytometry was used to sort one cell per well in a 96-well plate. Cells were grown to large colonies and checked by eye for expression of mCherry and EGFP. Next, DNA was isolated by NaCl/EtOH precipitation from single cell colonies (mCherry+EGFP+, mCherry+EGFP-, mCherry-EGFP-). A standard PCR was performed using GoTaq DNA polymerase (REF M3175, Promega) and oligos CAGTCATAGCCGAATAGCCTCT and 6-FAM-labeled GACCACCTACAAGGCCA AGA. Samples were run on an ABI 3730 analyser (Applied Biosystems) and analysed with Peak Scanner software v1.0.

Tests for transcription, ICR191 and Phen-DC_6

Transcription

Cells were cultured in medium containing 0,1µg/ml doxycycline hyclate and charcoal stripped fetal bovine serum (ref A15-119, PAA laboratories). At day 1 mCherry+EGFP- cells with plm212 (C_{23}-repeat) were sorted and cultured in doxycycline hyclate free medium. At day 3 a similar sort was performed but now 4 cells per well were plated in a 96-well plate with or without doxycycline hyclate. At day 25, when wells were confluent, doxycycline hyclate was added to all plates. Cells were analysed for mCherry and EGFP expression by flow cytometry after 3 days.

Phen-DC_6

5000 mCherry+GFP- cells of a polyclonal cell line with plm273 ($G_{11}AG_{11}$) were plated per well (96 well), and incubated with 5µM Phen-DC_6 (kindly provided by M.P. Teulade-Fichou, C Guetta and A. Nicolas) or control (DMSO). Cells were quantified by flow cytometry after 6 days.

ICR191

5000 mCherry-GFP- cells (with plm315, G_{20}-repeat) were plated per well (96 well plate). 24 Hours later cells were incubated with various concentrations of ICR 191(ref I3636,

Sigma) dissolved in 0.01M HCL and optimem for two hours. Cells were analysed by flow cytometry after 9 days.

miRNA screen

An MCF-7 reporter cell line was created by single copy integration of plm191 (G_{23}-repeat). Next, a miRNA expression library (miR-Lib) (Voorhoeve et al., 2006) was used to transduce the reporter cell line with ±450 different miR-Vecs by retroviral transduction in a 96-well format in triplicate. Per experiment, cells were drug-selected for 5 days and all stable cell lines with integrated miR-Vecs were combined in one large pool. Next, mCherry+EGFP- cells were selected by flow cytometry and after 21 days mCherry+EGFP+ and mCherry+EGFP- populations were sorted per pool and genomic DNA was isolated (DNeasy kit, Qiagen). The miRNA coding regions from the retroviral (genomically integrated) miR-Vecs were amplified using IlluSeq_IndXX_Mirvec_f and P7_MirVec_r primers, followed by a second amplification step using P5_illuseq and P7_MirVec_r primers (sequences can be found in Supplementary Table S1). The resulting library was deep-sequenced on an Illumina Hiseq platform according to manufacturers protocol. The resulting reads were aligned to the miR-lib resulting in 2.5 million aligned reads divided over 6 barcodes (triplicate measurements of two cell populations).

Flow cytometry

Cells were sorted with flow cytometer BD FACSAria III and analysed with flow cytometers BD FacsDiva (BD biosciences) and guava easyCyte HT (Millipore) and their respective software. To set gates we used control polyclonal cell lines that consisted out of solely mCherry+EGFP- or mCherry+EGFP+ cells.

RESULTS AND DISCUSSION

Construction and validation of a fluorescent microsatellite reporter

To study which factors contribute to the stability of microsatellites in human cells, we wished to develop a fast and reliable system that measures MSI independent of its genomic context. Therefore, we decided to make use of FACS enabling measurements of 10^6 to 10^7 events per hour and site-specific recombination to target microsatellites at the same specific locus, which ensures the identical genomic environment for each fragile tract, and thus complements and further extends previously described MSI-reporters that require episomal replication or random integration (Cejka et al., 2003; Gasche et al., 2003; Hile et al., 2000).

Accordingly, we created a transgene reporting MSI by expression of the green fluorescent protein EGFP upon a frameshift; we placed the sequence of EGFP downstream of a microsatellite containing ORF such that a -1bp frameshift at the microsatellite leads to in frame EGFP. The upstream ORF encodes functional red fluorescent protein mCherry (illustrated in Figure 1A). mCherry expression visualizes

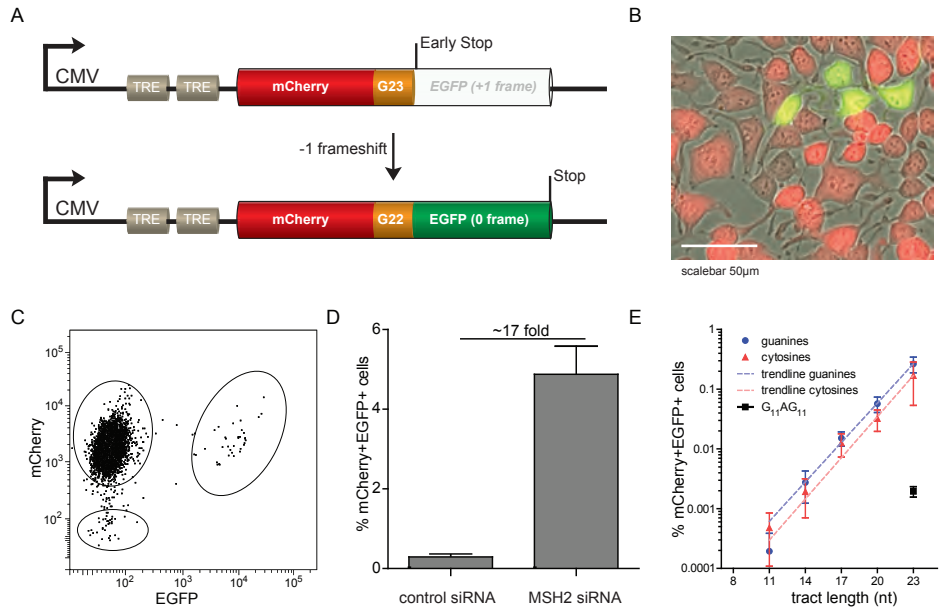


Figure 1 | A fluorescent-based MSI reporter system at a fixed position in the human genome. (A) Schematic representation of the reporter transgene: the coding sequence of EGFP is placed in the +1 frame downstream of 23 guanines (G23) and the coding sequence of mCherry. A -1 frameshift brings the EGFP ORF in frame with the upstream mCherry. As a result, cells will express both markers. TRE = Tetracycline responsive element. (B) A representative image of an MSI event. In cultures, MSI events will manifest as single cells or a small clonal group of cells. Notably, mCherry+EGFP+ cells showed reduced expression levels of mCherry compared to mCherry+EGFP- cells. This image is an overlay of pictures taken in bright-field, green and red channel. (C) Representative FACS-plot of a population Flp-In T-Rex-293 cells with an integrated copy of the reporter transgene. The vast majority are mCherry+EGFP-, but also mCherry+EGFP+ are seen. (D) Knockdown of MSH2 by siRNA results in increased numbers of mCherry+EGFP+ cells (7 days after pre-sorting mCherry+EGFP- cells). Error bars represent the standard error of the mean (s.e.m.) of three experiments. (E) Exponential relationship between the number of nucleotides in a microsatellite (X-axis) and the incidence of frameshifting, as measured by the percentage of mCherry+EGFP+ cells (Y-axis) 2 days after pre-sorting mCherry+EGFP- cells. Values represent the means and s.e.m. of three independent experiments.

the presence of the reporter as well as the transcriptional activity of the transgene. Unique restriction sites flanking the microsatellite allow for easy insertion of different fragile sites. The sequence of this dual-fluorescent fusion product was cloned into the vector pcDNA5/FRT/TO (LifeTechnologies) that allows tetracycline-inducible expression under the control of a cytomegalovirus promotor (Andersson et al., 1989; Boshart et al., 1985; Hillen and Berens, 1994; Hillen et al., 1983; Nelson et al., 1987). The vector also contains a single FLP Recombination Target (FRT) site that is required for Flp recombinase-mediated integration of the vector into a genome that contains a single copy integrated FRT site (Craig, 1988; Sauer, 1994). This system allowed us to efficiently create stable cell lines with a single copy integrated reporter containing a microsatellite.

Firstly, we tested a reporter containing a microsatellite of 23 guanines in the coding strand (plm191, Figure 1A) in Flp-In T-REX-293 cells that contain a single copy integrated Flp Recombination Target site (LifeTechnologies). As expected, the majority of cells expressed mCherry but not EGFP (mCherry+EGFP-) after integration of the reporter, however a small percentage (±0.2%) also expressed EGFP (mCherry+EGFP+), as visualized by microscopy and FACS (Figure 1B and 1C). As a control we integrated a similar reporter but without a microsatellite (plm184). This cell line showed only mCherry+GFP- cells (data not shown). Together, this indicated that EGFP expression was the consequence of the presence of the microsatellite. To further substantiate that the EGFP+ cells were the result of MSI, we performed a transient knockdown of the MMR-factor MSH2 by siRNA on pre-sorted mCherry+EGFP- cells. Because the MSH2 protein was reported to be stable (Diouf et al., 2011), two rounds of siRNA-treatment were performed, and FACS-analysis was performed after 7 days of culturing. This resulted in a ±17-fold increase in the number of mCherry+EGFP+ cells, indicative for MSI-dependent expression of EGFP (Figure 1D). Together, these data indicate that EGFP expression is a read-out for MSI. The advantage of the mCherry internal control lies in the fact that by pre-sorting mCherry+EGFP- cells we can i) discard cells that became EGFP positive before an experimental condition was applied, and ii) get rid of cells that lost the reporter or its expression due to silencing.

Exponential correlation between length of a monotract and MSI

To determine the type of correlation between the length of the microsatellite tract and its instability, we made transgenes carrying monotracts of lengths 8, 11, 14, 17, 20 and 23 nucleotides, of all four different nucleotides. On integration in 293-T-REX cells stable polyclonal lines were obtained. Unexpectedly, we found that adenine and thymine monotracts diminished EGFP expression levels (when put in frame as a control), making MSI analysis of these tracts impossible. Therefore, we pursued analysis with the lines containing monotracts of guanines and cytosines. We established populations of 100% mCherry+EGFP- cells (by FACS) and determined the percentages of mCherry+EGFP+ after 48 h of culturing. We started to detect MSI in cell lines containing monotracts of 11bp or longer; in cells with monotracts of 8 nucleotides (either guanine or cytosine) no mCherry+EGFP+ cells were detected, corresponding to a mutation frequency that is $< 10^{-5}$. This result may be explained by the inability of a small monotract to induce slippage events: several *in vitro* and computational studies have argued a threshold for slippage being ~7-9 basepairs (Kelkar et al., 2010; Lai and Sun, 2003; Paoloni-Giacobino et al., 2001; Rose and Falush, 1998; Shinde et al., 2003). Figure 1E indicates that the correlation of MSI and the length of a monotract is best fitted by an exponential trend line; for every extra guanine or cytosine in the monotract the number of mCherry+EGFP+ cells increased with 1.66- and 1.70-fold, respectively (R^2= 0.80 and 0.43). We found no significant difference in MSI between cell lines containing similar sized guanine or cytosine monotracts, suggesting no influence of strand orientation, similar to what was reported in yeast (Henderson and Petes, 1992).

Interruption of pure repeats

To explain the distribution and stability of microsatellites across genomes, several groups have postulated mutational models (reviewed in (Bhargava and Fuentes, 2009)). Experimental evidence, as described above, helps to improve these mutational models. Rates for expansions and contractions are obviously important parameters in these models, however another factor of great influence is the effect of point mutations that split up a monotract into two smaller ones. To measure the effect of such an event experimentally, we analysed a reporter (plm273) with an instability locus that is similar to a G23 monotract, apart from having an A instead of a G at position 12, hence comprising two monotracts of 11 Gs ($G_{11}AG_{11}$). We found that the MSI-rate of this fragile site was > 100-fold lower than that of a G23 monotract (Figure 1E). These results show that a single mutation can lead to stabilization of a repeat with more than two orders of magnitude. Similar observations were found in human mismatch repair-deficient cells (Boyer et al., 2008). Interestingly, the $G_{11}AG_{11}$ repeat was still 10-fold more unstable than a single G11 repeat, suggesting that two repeats in close proximity are more unstable than the sum of two single G11 repeats.

Contraction exceeds expansion of longer monotracts

During the course of our experiments we noticed the presence of mCherry-EGFP- cells by microscopy and FACS (Figure 1C, lower gate). These cells were resistant to the transgene encoded selection-marker which is located directly upstream of the fluorescent reporter, making silencing or loss of the locus unlikely. Interestingly, we noticed that the fraction of mCherry-EGFP- cells was similar to the fraction of mCherry+EGFP+ cells. We thus wondered whether mCherry-EGFP- cells are the result of +1 frameshift events: although the monotracts were cloned downstream of mCherry to not interfere with its expression, it may be that the -1 frame of GFP encodes an amino acid sequence that interferes with proper mCherry expression.

To test this explanation, we constructed transgenes with EGFP in the -1 reading frame. Indeed, cells with integration of such constructs resulted in three populations of cells: the vast majority of cells being mCherry-GFP-, and two smaller populations of mCherry+EGFP- and mCherry+EGFP+ cells (Figure 2A). To confirm that the latter two populations represented -1bp and +1bp frameshifts, respectively, we performed fragment analysis on these populations. A cell line containing a G14 monotract was used because analysis of long repeats is technically challenging. Indeed, fragment analysis on colonies derived from single cells confirmed that mCherry+EGFP- represented -1 frameshifts, whereas mCherry+EGFP+ represented +1 frameshifts (see Supplementary Figure S1). We thus serendipitously constructed a transgenic reporter that allows the quantification of both contractions and expansions of microsatellites.

Subsequently, we used this experimental setup to investigate whether -1 and +1 events are induced with equal frequencies. We isolated mCherry-EGFP- cells of transgenes containing a G_{23} monotract and analysed the populations 48 hours later. We found 12 times more mCherry+EGFP- cells than mCherry+EGFP+, indicating

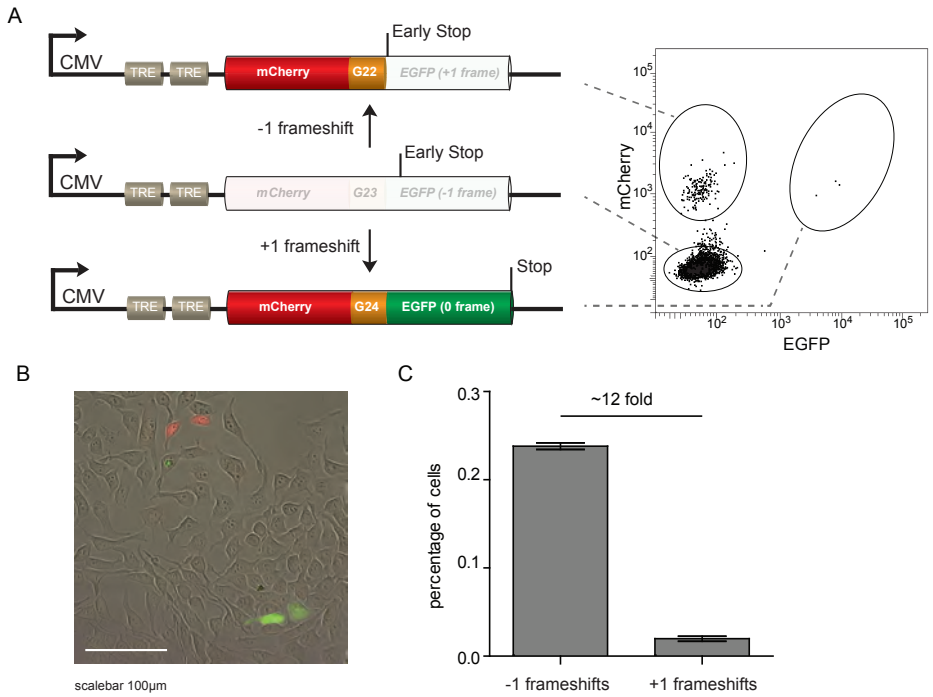


Figure 2 | A fluorescent-based reporter system that reads out both -1 and +1 frameshifts. (A) Schematic representation of the reporter construct. When the coding sequence of EGFP is placed in the -1 frame neither mCherry nor EGFP is expressed. A -1 frameshift results in mCherry+EGFP- cells whereas a +1 frameshift results in mCherry+EGFP+ cells. The right panel illustrates the three different populations in a representative FACS-plot. (B). Representative image of HEK293 cells carrying a single copy integrated reporter transgene. (C) Quantification of percentages mCherry+EGFP- and mCherry+EGFP+ cells representing -1 and +1 frameshifts, respectively, 2 days after pre-sorting mCherry-EGFP- cells. Error bars denote standard deviation (s.d.) of three independent experiments.

that 1 bp contractions are more prevalent than 1bp expansions. This result supports studies reporting that longer microsatellites are more prone to contracting (Harr and Schlötterer, 2000; Xu et al., 2000).

No role for on-going transcription on MSI levels

Studies in bacteria and yeast have shown that transcription can destabilize simple repetitive DNA sequences (Kiyama and Oishi, 1994; Wierdl et al., 1996). We questioned if and to which extent transcriptional activity affects MSI in human cells by monitoring frameshifting at monotracts in cells that were cultured in the presence or absence of doxycycline, a regulator of the Tet repressor that controls the levels of transcription over the transgene (Yao et al., 1998). To avoid transcriptional activity of the transgene at the time of pre-sorting, cells were grown in doxycycline-free medium for three days, resulting in low but still detectable levels of mCherry. Then, four single mCherry+EGFP- cells were seeded and the cultures were grown till confluency, while the transcriptional status of the transgene was visualized by mCherry expression (Figure 3A). To be able to use EGFP as a marker, doxycycline was also added to the non-transcribed conditions three days prior to reading out MSI. In contrast to studies in bacteria and yeast, MSI is not elevated by ongoing transcription in human cells (Figure 3B).

MSI-reporter applicable for testing compounds

One of the applications of the system we developed resides in testing compounds for possible mutagenic characteristics in mammalian cells. As proof of principle we used the compound 6-chloro-9-[3-2(2-chloro-ethylamino)propylamino]-2-methoxy-acridine (ICR191), an acridine derivate that is a known inducer of frameshift mutations (Ferguson and Denny, 1990). We exposed mCherry-EGFP- cells with a G_{20}-repeat to various concentrations of ICR191 and allowed them to grow for several days, before analysis by flow cytometry. Figure 3C shows a dose-dependent increase of both -1 and +1 frameshifts. In addition, we observed, in line with previous studies (Calos and Miller, 1981; Cariello et al., 1990; Chen et al., 2000; Ferguson and Denny, 1990; Taft et al., 1994), that +1 frameshifts were more frequently induced by ICR191 than -1 frameshifts (53- versus 12-fold, respectively).

Quadruplex formation and MSI

Non-B-DNA conformations of microsatellites are thought to facilitate slippage and thus influence the frameshift mutation rate (Bhargava and Fuentes, 2009). One of such non-B-DNA structures is a G-quadruplex structure; a secondary structure that consists of minimally 3 stacked planar arrays of 4 guanines held together by Hoogsteen hydrogen bonding (Bochman et al., 2012). In this study, we have used several microsatellites that have the ability to form a G-quadruplex structure. To address a possible influence of G-quadruplex formation on MSI, we assayed cells containing a $G_{11}AG_{11}$-repeat, a sequence that has the potential to form a G4-quadruplex structure *in vivo* (Koole et al., unpublished), in the presence and absence of Phen-DC_6 (De Cian et al., 2007) an effective stabilizer of G4-structures in yeast and human cells (Halder et al., 2011; 2012; Lopes et al., 2011; Piazza et al., 2010). We found, however, similar level of MSI in cells treated or mock-treated with to Phen-DC_6 (Figure 3C), suggesting that quadruplex-formation does not stimulate frameshift mutations in monotracts.

Role of miRNAs on MSI

Besides testing intrinsic factors, such as the sequence context of microsatellites, and exogenous influences, such as mutagenic compounds, the reporter system also allows systematic testing for genetic factors and their effect on MSI (as demonstrated by siRNA knockdown of MSH2). Moreover, also unbiased screens are possible to identify genetic regulators of MSI: as proof of principle we used an unbiased screening approach to identify miRNAs that influence MSI (Figure 4A). To this end, we integrated a single copy of a transgene containing a G_{23}-repeat (plm191) into a human breast carcinoma cell line MCF-7. The resulting cell line was transduced with

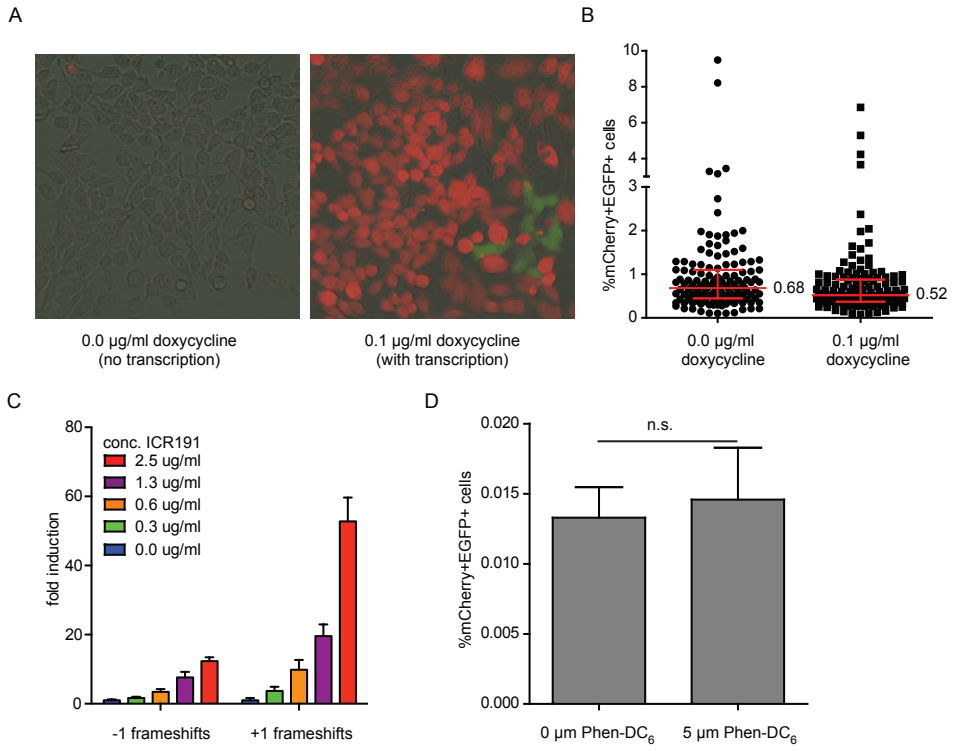


Figure 3 | Structural and chemical determinants of MSI. (A) Representative images of HEK293 cells carrying a single copy C23 reporter transgene in the absence or presence of doxycycline. Images are overlay of pictures taken in bright-field, green and red channel. (B) Percentage of mCherry+EGFP+ HEK293 cells carrying a single copy C23 reporter after 28 days of culturing. Three days prior to flow cytometry, doxycycline was added to all cells to read out reporter expression. The graph represent the data of ~140 individually-grown populations per condition derived from three independent experiments. Error bars denote the median with interquartile range. Percentage of median is shown. (C) mCherry-GFP- HEK 293 cells carrying a single copy G20 reporter, that reads out -1 and +1 events, were treated with the indicated concentrations of ICR191 for two hours and analysed by flow cytometry 9 days later. The data represent 8 replicates derived from two independent experiments. Error bars denote the s.e.m. (D) mCherry+EGFP- HEK 293 cells containing a single copy G11aG11 reporter were grown for 6 days in the presence or absence of the G4 stabilizer Phen-DC6 and analysed by flow cytometry. Data represent 32 replicates derived from 2 independent experiments. Error bars denote the s.e.m, n.s= not significant by unpaired two-tailed t-test.

a miRNA expression library consisting of expression constructs (miR-Vecs) for most human miRNAs (Voorhoeve et al., 2006). Stable cell lines with integrated miR-Vecs were pooled and mCherry+EGFP- cells were selected by flow cytometry to remove cells with mutation events that occurred during the transduction and selection process. After three weeks of culturing, mCherry+EGFP- and mCherry+EGFP+ cells were sorted and genomic DNA was extracted and subjected to next generation sequencing to determine the relative abundance of each miRNA insert per population. Out of a total of 301 datapoints for individual miR-Vecs (Figure 4B), we selected 17 miRNAs

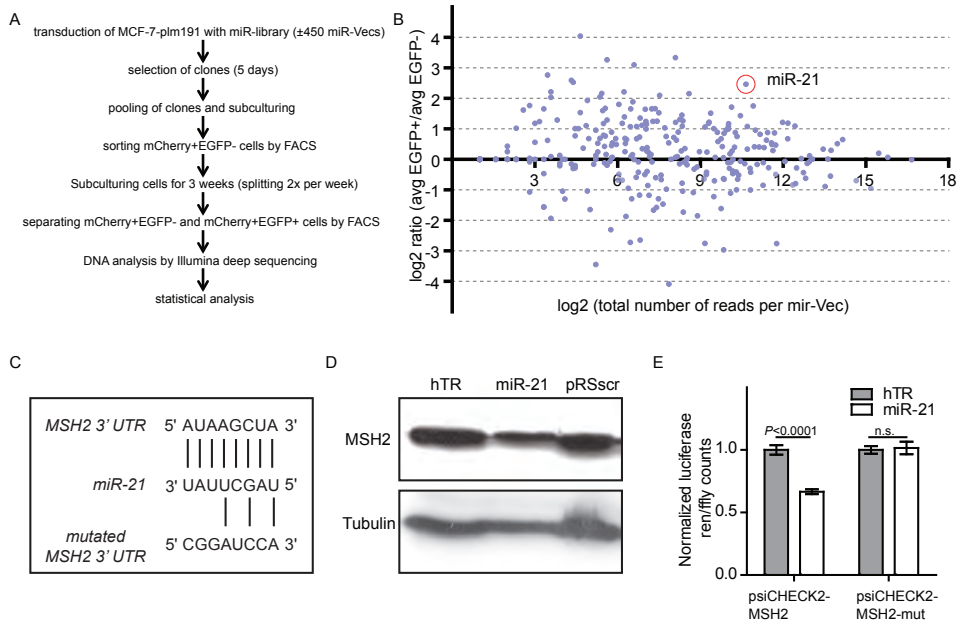


Figure 4 | An unbiased screen for miRNAs affecting MSI reveals miR-21 as negative regulator by affecting MSH2 levels. (A) Flow chart of the micro-RNA screen using MCF-7 cells that contain a single copy integrated MSI reporter. (B) A dot plot of the relative abundance of miR-Vec constructs in cells that were selected for an MSI event. The X-axis represents the Log2 count per miR-Vec, as quantified by deep sequencing. The Y-axis represents the Log2 ratio of the relative abundance of a miR-Vec in EGFP(+) over EGFP(-) cells. (C) Predicted conserved binding site in 3' UTR of *MSH2* for miR-21. The seed sequence of miR-21 is shown (1-8 bp). Lower sequence illustrates the mutated version of the 3' UTR of *MSH2* that was used for psiCHECK2-MSH2-mut. (D) Immuno-detection of MSH2 protein levels in MCF-7 cells with integrated miR-Vec miR-21 or control vectors hTR or pRSscr. Tubulin was stained to control for protein loading. (E) Dual-luciferase-assay (Promega) using vectors psiCHECK2-MSH2 and psiCHECK2-MSH2-mut in combination with co-transfection of miR-21 or control plasmid (hTR). Error bars denote s.d., statistical analysis: unpaired two-tailed t-test.

for independent validation experiments. Out of these 17 miRNAs, only miR-21 reproducibly induced MSI. Cell growth and plating efficiency was not affected by miR-21 overexpression (data not shown). Several target genes have been described for miR-21, including the mismatch repair gene MSH2 (Valeri et al., 2010). The microRNA binding prediction program targetscan predicts a binding site for miR-21 in the 3' UTR of human *MSH2* mRNA (Figure 4C). Using immunoblotting experiments, we indeed observed a visible reduction of MSH2 protein levels after miR-21 overexpression (Figure 4D) arguing that MSH2 is a *bona fide* target of miR-21. To further substantiate this interaction, we cloned the *MSH2* 3'UTR downstream of the Renilla luciferase gene. Co-transfection of miR-21 with this reporter construct (psiCHECK2-MSH2) in HEK 293 cells (which have low miR-21 levels (Zhu et al., 2008)) led to a ±40% reduction in luciferase counts, supporting the notion of negative regulation of MSH2 by miR-21 (Figure 4D). To demonstrate the specificity of miR-21 in targeting the

3'UTR of MSH2, we mutated its binding site by site-directed mutagenesis (Figure 4C), resulting in reporter construct psiCHECK2-MSH2-mut. These mutations killed the down-regulatory effect of miR-21 overexpression on luciferase expression (Figure 4E), thus providing proof for specificity in miR-21's genetic interaction with the evolutionary conserved binding site in the 3`UTR of MSH2. These results also provide proof of principle that our newly developed reporter system can be used to find (novel) *bona fide* genetic modifiers of MSI.

CONCLUSION

In this study we describe a functional fluorescence-based reporter that reads out MSI *in vivo* at one specific locus and can be measured by flow cytometry allowing for high-throughput analysis. To prove the functionality of this reporter and meanwhile obtaining biological informative data we tested several repeat-intrinsic and environmental factors that can influence MSI in human cells. In addition, candidate genes can be screened for factors that are potentially implicated in MSI. We combined retroviral transduction of several hundred miRNA-overexpressing constructs with next generation sequencing to identify a regulator of MSI when overexpressed. Although we made use of a miRNA-library and an MSI-specific reporter, we would like to point out that this approach is applicable to any other type of library (e.g. compound or knockdown libraries) and is easily adaptable to readout other potential fragile sequences or other specific pathways when using adjusted fluorescent reporters.

ACKNOWLEDGEMENTS

The authors thank M.P. Teulade-Fichou, C. Guetta and A. Nicolas for providing the compound Phen-DC_6.

FUNDING

European Research Council [203379, DSBrepair, to M.T]; European Commission (DDR response); Zorg Onderzoek Nederland/Medische Wetenschappen/Netherlands Genomics Initiative-Horizon to M.T.; Koningin Wilhelmina Fonds [Hubr2008-4107 to M.T.]; VENI-NWO (Nederlandse Organisatie voor Wetenschappelijk Onderzoek) to G.v.H.; Forschungsstipendium der Deutsche Forschungsgemeinschaft to (H.S.S.). Funding for open access charge: European Research Council

Conflict of interest statement. None declared.

REFERENCES

Andersson, S., Davis, D.L., Dahlbäck, H., Jörnvall, H., and Russell, D.W. (1989). Cloning, structure, and expression of the mitochondrial cytochrome P-450 sterol 26-hydroxylase, a bile acid biosynthetic enzyme. J. Biol. Chem. *264*, 8222–8229.

Beck, N.R., Double, M.C., and Cockburn, A. (2003). Microsatellite evolution at two hypervariable loci revealed by extensive avian pedigrees. Mol. Biol. Evol. *20*, 54–61.

Bhargava, A., and Fuentes, F.F. (2009). Mutational Dynamics of Microsatellites. Mol Biotechnol *44*, 250–266.

Bochman, M.L., Paeschke, K., and Zakian, V.A. (2012). DNA secondary structures: stability and function of G-quadruplex structures. Nat. Rev. Genet. 1–11.

Boshart, M., Weber, F., Jahn, G., Dorsch-Häsler, K., Fleckenstein, B., and Schaffner, W. (1985). A very strong enhancer is located upstream of an immediate early gene of human cytomegalovirus. Cell *41*, 521–530.

Boyer, J.C., Hawk, J.D., Stefanovic, L., and Farber, R.A. (2008). Sequence-dependent effect of interruptions on microsatellite mutation rate in mismatch repair-deficient human cells. Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis *640*, 89–96.

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., and Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. The American Journal of Human Genetics *62*, 1408–1415.

Brohede, J., Primmer, C.R., Møller, A., and Ellegren, H. (2002). Heterogeneity in the rate and pattern of germline mutation at individual microsatellite loci. Nucleic Acids Res. *30*, 1997–2003.

Calos, M.P., and Miller, J.H. (1981). Genetic and sequence analysis of frameshift mutations induced by ICR-191. J. Mol. Biol. *153*, 39–64.

Cariello, N.F., Keohavong, P., Kat, A.G., and Thilly, W.G. (1990). Molecular analysis of complex human cell populations: mutational spectra of MNNG and ICR-191. Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis *231*, 165–176.

Cejka, P., Marra, G., Hemmerle, C., Cannavo, E., Storchova, Z., and Jiricny, J. (2003). Differential killing of mismatch repair-deficient and -proficient cells: towards the therapy of tumors with microsatellite instability. Cancer Res. *63*, 8113–8117.

Chen, W.D.W., Eshleman, J.R.J., Aminoshariae, M.R.M., Ma, A.H.A., Veloso, N.N., Markowitz, S.D.S., Sedwick, W.D.W., and Veigl, M.L.M. (2000). Cytotoxicity and mutagenicity of frameshift-inducing agent ICR191 in mismatch repair-deficient colon cancer cells. J Natl Cancer Inst *92*, 480–485.

Craig, N.L. (1988). The mechanism of conservative site-specific recombination. Annu. Rev. Genet. *22*, 77–105.

De Cian, A., Delemos, E., Mergny, J.-L., Teulade-Fichou, M.-P., and Monchaud, D. (2007). Highly efficient G-quadruplex recognition by bisquinolinium compounds. J. Am. Chem. Soc. *129*, 1856–1857.

Diouf, B., Cheng, Q., Krynetskaia, N.F., Yang, W., Cheok, M., Pei, D., Fan, Y., Cheng, C., Krynetskiy, E.Y., Geng, H., et al. (2011). Somatic deletions of genes regulating MSH2 protein stability cause DNA mismatch repair deficiency and drug resistance in human leukemia cells. Nature Medicine *17*, 1298–1303.

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. *5*, 435–445.

Ferguson, L.R., and Denny, W.A. (1990). Frameshift mutagenesis by acridines and other reversibly-binding DNA ligands. Mutagenesis *5*, 529–540.

Gasche, C., Chang, C.L., Natarajan, L., Goel, A., Rhees, J., Young, D.J., Arnold, C.N., and Boland, C.R. (2003). Identification of frame-shift intermediate mutant cells. Proc. Natl. Acad. Sci. U.S.a. *100*, 1914–1919.

Gemayel, R., Vinces, M.D., Legendre, M., and Verstrepen, K.J. (2010). Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. Genetics *44*, 445–477.

Halder, K., Largy, E., Benzler, M., Teulade-Fichou, M.-P., and Hartig, J.S. (2011). Efficient suppression of gene expression by targeting 5'-UTR-based RNA quadruplexes with bisquinolinium compounds. Chembiochem *12*, 1663–1668.

Halder, R., Riou, J.-F., Teulade-Fichou, M.-P., Frickey, T., and Hartig, J.S. (2012). Bisquinolinium compounds induce quadruplex-specific transcriptome changes in HeLa S3 cell lines. BMC Res Notes *5*, 138.

Harfe, B.D., and Jinks-Robertson, S. (2000). Sequence composition and context effects on the generation and repair of frameshift intermediates in mononucleotide runs in Saccharomyces cerevisiae. Genetics *156*, 571–578.

Harr, B.B., and Schlötterer, C.C. (2000). Long microsatellite alleles in Drosophila melanogaster have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. Genetics *155*, 1213–1220.

Hawk, J.D., Stefanovic, L., Boyer, J.C., Petes, T.D., and Farber, R.A. (2005). Variation in efficiency of DNA mismatch repair at different sites in the yeast genome. Proc. Natl. Acad. Sci. U.S.a. *102*, 8639–8643.

Henderson, S.T., and Petes, T.D. (1992). Instability of simple sequence DNA in Saccharomyces cerevisiae. Mol. Cell. Biol. *12*, 2749–2757.

Hile, S.E., Yan, G., and Eckert, K.A. (2000). Somatic mutation rates and specificities at TC/AG and GT/CA microsatellite sequences in nontumorigenic human lymphoblastoid cells. Cancer Res. *60*, 1698–1703.

Hillen, W., and Berens, C. (1994). Mechanisms underlying expression of Tn10 encoded tetracycline resistance. Annu. Rev. Microbiol. *48*, 345–369.

Hillen, W., Gatz, C., Altschmied, L., Schollmeier, K., and Meier, I. (1983). Control of expression of the Tn10-encoded tetracycline resistance genes. Equilibrium and kinetic investigation of the regulatory reactions. J. Mol. Biol. *169*, 707–721.

Kelkar, Y.D., Strubczewski, N., Hile, S.E., Chiaromonte, F., Eckert, K.A., and Makova, K.D. (2010). What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. Genome Biol Evol *2*, 620–635.

Kiyama, R., and Oishi, M. (1994). Instability of plasmid DNA maintenance caused by transcription of poly(dT)-containing sequences in Escherichia coli. Gene *150*, 57–61.

Kondrashov, A.S., and Rogozin, I.B. (2004). Context of deletions and insertions in human coding sequences. Hum. Mutat. *23*, 177–185.

Lai, Y., and Sun, F. (2003). The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol. Biol. Evol. *20*, 2123–2131.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Li, Y.-C., Korol, A.B., Fahima, T., Beiles, A., and Nevo, E. (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol. Ecol. *11*, 2453–2465.

Lopes, J., Piazza, A., Bermejo, R., Kriegsman, B., Colosio, A., Teulade-Fichou, M.-P., Foiani, M., and Nicolas, A. (2011). G-quadruplex-induced instability during leading-strand replication. Embo J *30*, 4033–4046.

Ma, X., Rogacheva, M.V., Nishant, K.T., Zanders, S., Bustamante, C.D., and Alani, E. (2012). Mutation Hot Spots in Yeast Caused by Long-Range Clustering of Homopolymeric Sequences. CellReports *1*, 36–42.

Nelson, J.A., Reynolds-Kohler, C., and Smith, B.A. (1987). Negative and positive regulation by a short segment in the 5'-flanking region of the human cytomegalovirus major immediate-early gene. Mol. Cell. Biol. *7*, 4125–4129.

Paoloni-Giacobino, A., Rossier, C., Papasavvas, M.P., and Antonarakis, S.E. (2001). Frequency of replication/transcription errors in (A)/(T) runs of human genes. Hum. Genet. *109*, 40–47.

Peña-Diaz, J., and Jiricny, J. (2012). Mammalian mismatch repair: error-free or error-prone? Trends in Biochemical Sciences *37*, 206–214.

Piazza, A., Boulé, J.-B., Lopes, J., Mingo, K., Largy, E., Teulade-Fichou, M.-P., and Nicolas, A. (2010). Genetic instability triggered by G-quadruplex interacting Phen-DC compounds in Saccharomyces cerevisiae. Nucleic Acids Res. *38*, 4337–4348.

Primmer, C.R., Saino, N., Møller, A.P., and Ellegren, H. (1998). Unraveling the processes of microsatellite evolution through analysis of germ line mutations in barn swallows Hirundo rustica. Mol. Biol. Evol. *15*, 1047–1054.

Richard, G.F., Kerrest, A., and Dujon, B. (2008). Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. Microbiology and Molecular Biology Reviews *72*, 686–727.

Rose, O., and Falush, D. (1998). A threshold size for microsatellite expansion. Mol. Biol. Evol. *15*, 613–615.

Sauer, B. (1994). Site-specific recombination: developments and applications. Curr. Opin. Biotechnol. *5*, 521–527.

Shinde, D., Lai, Y., Sun, F., and Arnheim, N. (2003). Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. Nucleic Acids Res. *31*, 974–980.

Taft, S.A., Liber, H.L., and Skopek, T.R. (1994). Mutational spectrum of ICR-191 at the hprt locus in human lymphoblastoid cells. Environ. Mol. Mutagen. *23*, 96–100.

Valeri, N.N., Gasparini, P.P., Braconi, C.C., Paone, A.A., Lovat, F.F., Fabbri, M.M., Sumani, K.M.K., Alder, H.H., Amadori, D.D., Patel, T.T., et al. (2010). MicroRNA-21 induces resistance to 5-fluorouracil by down-regulating human DNA MutS homolog 2 (hMSH2). Proc. Natl. Acad. Sci. U.S.a. *107*, 21098–21103.

Voorhoeve, P.M.P., le Sage, C.C., Schrier, M.M., Gillis, A.J.M.A., Stoop, H.H., Nagel, R.R., Liu, Y.-P.Y., van Duijse, J.J., Drost, J.J., Griekspoor, A.A., et al. (2006). A Genetic Screen Implicates miRNA-372 and miRNA-373 As Oncogenes in Testicular Germ Cell Tumors. Cell *124*, 13–13.

Webster, M.T., Smith, N.G.C., and Ellegren, H. (2002). Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. Proc. Natl. Acad. Sci. U.S.a. *99*, 8748–8753.

Wierdl, M., Dominska, M., and Petes, T.D. (1997). Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics *146*, 769–779.

Wierdl, M., Greene, C.N., Datta, A., Jinks-Robertson, S., and Petes, T.D. (1996). Destabilization of simple repetitive DNA sequences by transcription in yeast. Genetics *143*, 713–721.

Woerner, S.M., Yuan, Y.P., Benner, A., Korff, S., Knebel Doeberitz, von, M., and Bork, P. (2009). SelTarbase, a database of human mononucleotide-microsatellite mutations and their potential impact to tumorigenesis and immunology. Nucleic Acids Res. *38*, D682–D689.

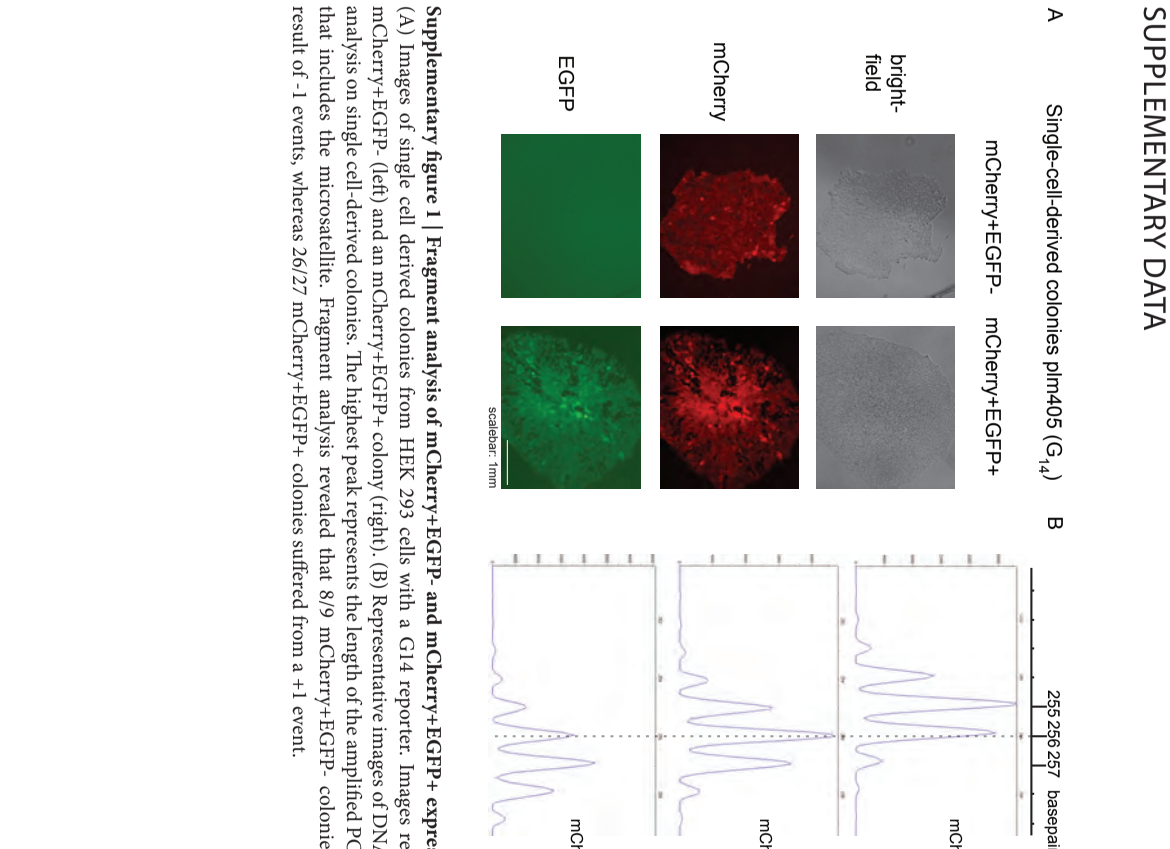
Xu, X., Peng, M., and Fang, Z. (2000). The direction of microsatellite mutations is dependent upon allele length. Nat. Genet. *24*, 396–399.

Yamada, N.A., Smith, G.A., Castro, A., Roques, C.N., Boyer, J.C., and Farber, R.A. (2002). Relative rates of insertion and deletion mutations in dinucleotide repeats of various lengths in mismatch repair proficient mouse and mismatch repair deficient human cells. Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis *499*, 213–225.

Yao, F., Svensjö, T., Winkler, T., Lu, M., Eriksson, C., and Eriksson, E. (1998). Tetracycline repressor, tetR, rather than the tetR-mammalian cell transcription factor fusion derivatives, regulates inducible gene expression in mammalian cells. Hum. Gene Ther. *9*, 1939–1950.

Zhu, S., Wu, H., Wu, F., Nie, D., Sheng, S., and Mo, Y.-Y. (2008). MicroRNA-21 targets tumor suppressor genes in invasion and metastasis. Cell Res. *18*, 350–359.

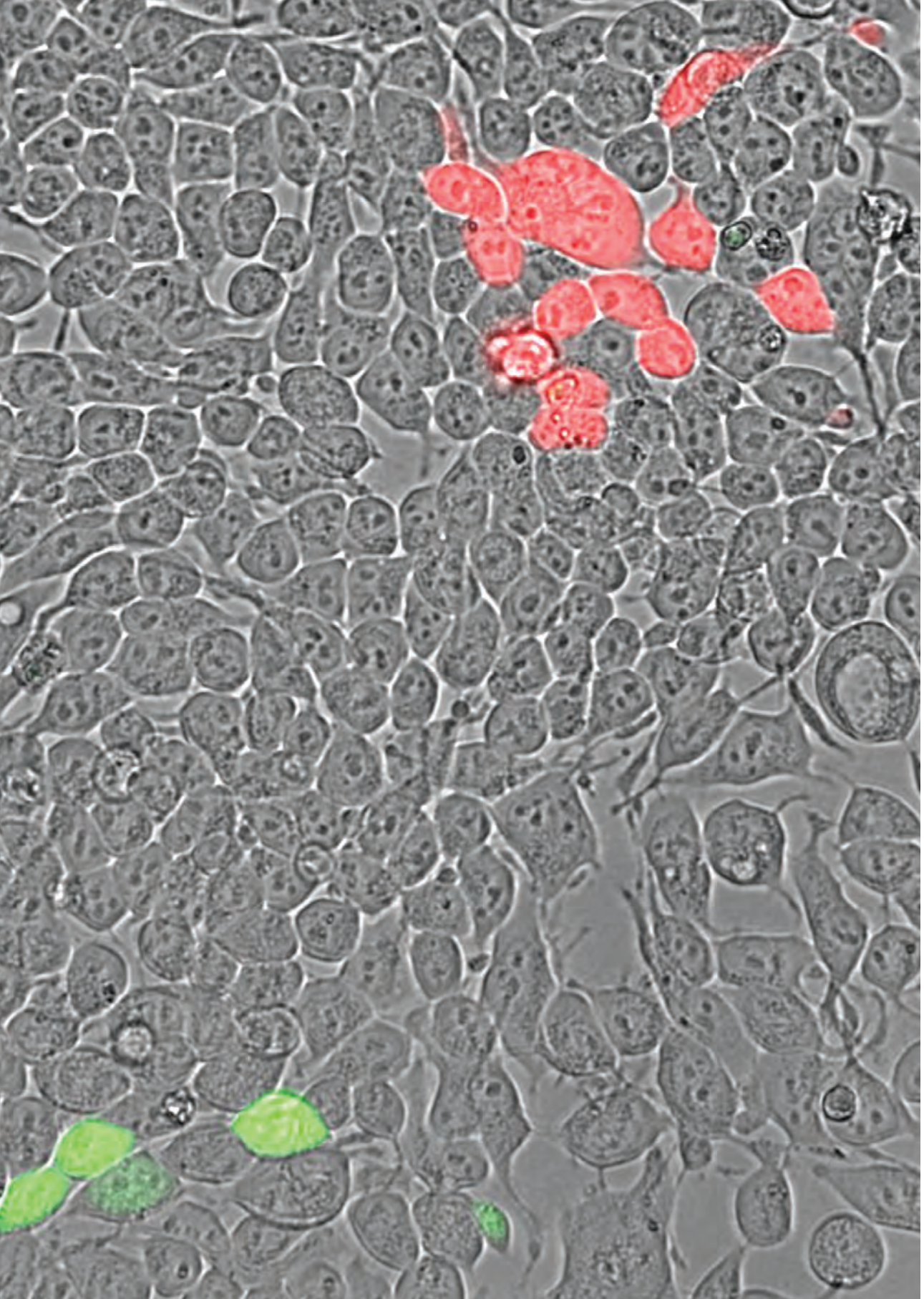
SUPPLEMENTARY DATA



Supplementary figure 1 | Fragment analysis of mCherry+EGFP- and mCherry+EGFP+ expressing cells. (A) Images of single cell derived colonies from HEK 293 cells with a G14 reporter. Images represent an mCherry+EGFP- (left) and an mCherry+EGFP+ colony (right). (B) Representative images of DNA fragment analysis on single cell-derived colonies. The highest peak represents the length of the amplified PCR-product that includes the microsatellite. Fragment analysis revealed that 8/9 mCherry+EGFP- colonies were the result of -1 events, whereas 26/27 mCherry+EGFP+ colonies suffered from a +1 event.

Supplementary Table 1 | Primers used for amplification of miR-Vecs

Primer	Sequence
IlluSeq_Ind01_MirVec_f	ACACTCTTTCCCTACACGACGCTCTTCCGATCTaaGCTTGGTACCGAGCTCGGATC
IlluSeq_Ind02_MirVec_f	ACACTCTTTCCCTACACGACGCTCTTCCGATCTatGCTTGGTACCGAGCTCGGATC
IlluSeq_Ind03_MirVec_f	ACACTCTTTCCCTACACGACGCTCTTCCGATCTagGCTTGGTACCGAGCTCGGATC
IlluSeq_Ind04_MirVec_f	ACACTCTTTCCCTACACGACGCTCTTCCGATCTacGCTTGGTACCGAGCTCGGATC
IlluSeq_Ind05_MirVec_f	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtaGCTTGGTACCGAGCTCGGATC
IlluSeq_Ind06_MirVec_f	ACACTCTTTCCCTACACGACGCTCTTCCGATCTttGCTTGGTACCGAGCTCGGATC
IlluSeq_Ind07_MirVec_f	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtgGCTTGGTACCGAGCTCGGATC
IlluSeq_Ind08_MirVec_f	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtcGCTTGGTACCGAGCTCGGATC
IlluSeq_Ind09_MirVec_f	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgaGCTTGGTACCGAGCTCGGATC
IlluSeq_Ind10_MirVec_f	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgtGCTTGGTACCGAGCTCGGATC
IlluSeq_Ind11_MirVec_f	ACACTCTTTCCCTACACGACGCTCTTCCGATCTggGCTTGGTACCGAGCTCGGATC
IlluSeq_Ind12_MirVec_f	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgcGCTTGGTACCGAGCTCGGATC
P5_IlluSeq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
P7_MirVec_r	CAAGCAGAAGACGGCATACGAGATAACCTACAGGTGGGGTCTTTC
P7	CAAGCAGAAGACGGCATACGAGAT
P5	AATGATACGGCGACCACCGAGATCT



◀ This image shows kidney cells, that contain a reporter construct, resulting in expression of green fluorescent protein upon lengthening of a microsatellite (a repetitive DNA sequence) or expression of red fluorescent protein upon shortening of the microsatellite.

Cover Photo: Rocket Science ▶

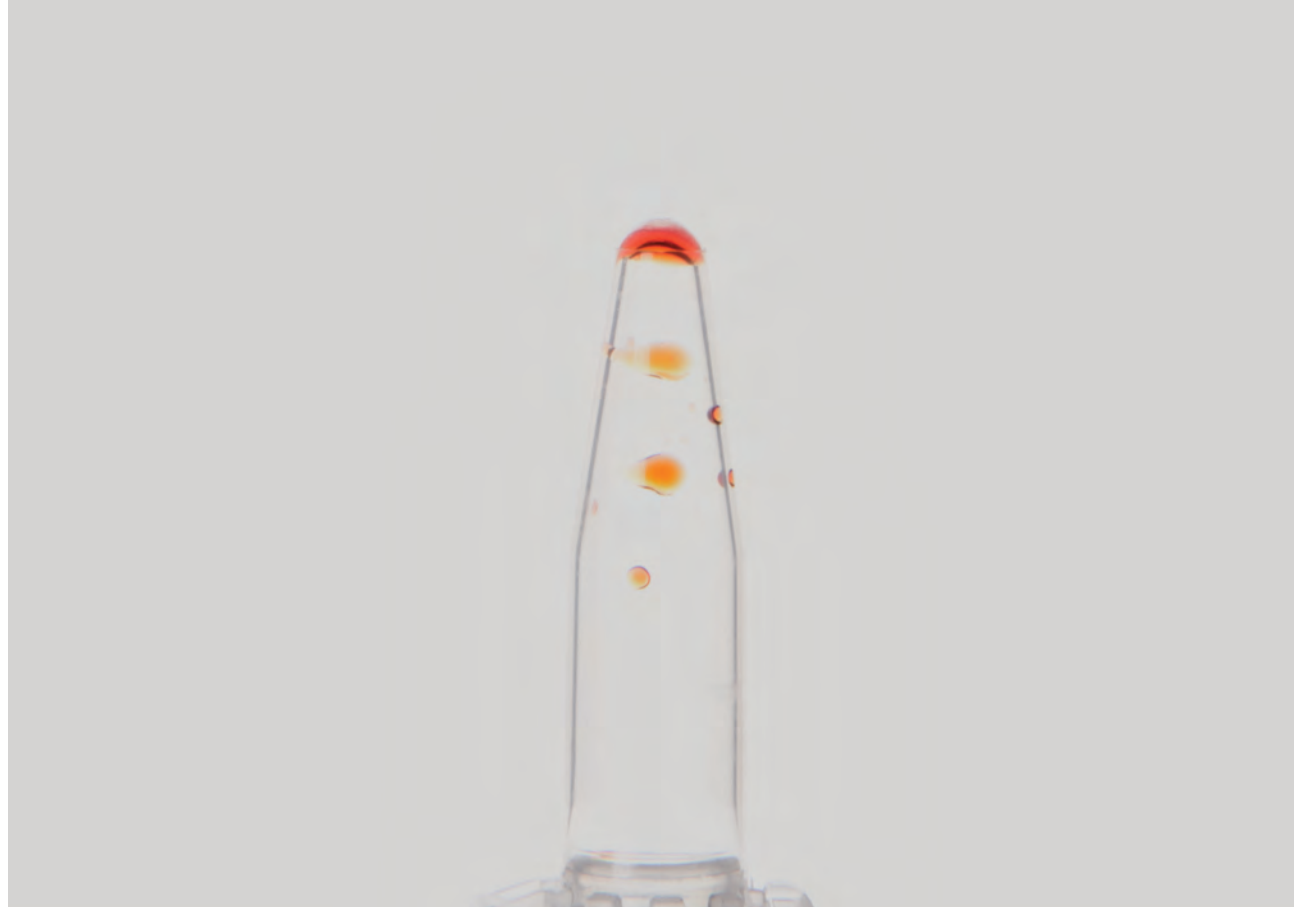
Chapter 3

Mosaic analysis and tumor induction in zebrafish by microsatellite instability-mediated stochastic gene expression

Wouter Koole and Marcel Tijsterman

Department of Toxicogenetics, Leiden University Medical Center, Leiden, The Netherlands

Adapted from Koole and Tijsterman Dis. Model. Mech. 7, (7) pp. 929-936 (2014)



ABSTRACT

Mosaic analysis, in which two or more populations of cells with differing genotypes are studied in a single animal, is a powerful approach to study developmental mechanisms and gene function *in vivo*. Over recent years, several genetic methods have been developed to achieve mosaicism in zebrafish, but despite their advances, limitations remain and different approaches and further refinements are warranted. Here, we describe an alternative approach for creating somatic mosaicism in zebrafish that relies on the instability of microsatellite sequences during replication. We placed the coding sequences of various marker proteins downstream of a microsatellite and out-of-frame; *in vivo* frameshifting into the proper reading frame results in expression of the protein in random individual cells that are surrounded by wild-type cells. We optimized this approach for the binary Gal4-UAS expression system by generating a driver line and effector lines that stochastically express Gal4-VP16 or UAS:H2A-EGFP and self-maintaining UAS:H2A-EGFP-Kaloop, respectively. To demonstrate the utility of this system, we stochastically expressed a constitutively active form of the human oncogene H-RAS and show the occurrence of hyperpigmentation and sporadic tumors within 5 days. Our data demonstrate that inducing somatic mosaicism through microsatellite instability can be a valuable approach for mosaic analysis and tumor induction in *Danio rerio*.

INTRODUCTION

Somatic mosaicism is a widely used term to describe the presence of two genetically different cell populations in a single individual. Mosaic animals arise from genetic alterations or epigenetic changes (e.g. X-chromosome inactivation) in a subset of cells during development. Mosaicism can also be obtained when cells are transplanted from one animal to another, although technically this is termed chimerism. The importance of mosaic analysis was evident from the moment the first naturally occurring mosaic animals were discovered, nearly 100 years ago (examples are given in (Xu and Rubin, 2012)). Since then, investigators have developed various techniques to stimulate somatic mosaicism, enabling experiments to trace cell lineage and the study of developmental processes and gene function in whole animals (for an overview of established techniques, see (Buckingham and Meilhac, 2011)).

An advantage of mosaic analysis in the zebrafish over other model organisms is that the development of the embryos occurs *ex utero* and the embryos are transparent. Therefore, individual cells can be studied during embryonic development and imaged *in vivo* with relative ease. Additionally, the identification of 'casper' fish (White et al., 2008), which are almost entirely transparent because of the lack of two pigment cell types – melanocytes and iridophores, allows live imaging in adult animals.

Several methods exist in zebrafish to create mosaic animals (for reviews, see (Blackburn and Langenau, 2010; Carmany-Rampey and Moens, 2006; Weber and Köster, 2013)). For example, transplantation assays and DNA and/or mRNA injection at the one-cell stage can be used, but these are invasive, time-consuming and often technically challenging; therefore, non-invasive genetic approaches are preferred. In the last few years, several such approaches have been developed (Boniface et al., 2009; Collins et al., 2010; Emelyanov and Parinov, 2008; Esengil et al., 2007; Gerety et al., 2013; Hans et al., 2011; 2009; Knopf, 2010; Pan et al., 2011), most of which rely on Cre recombinase-controlled lox site recombination (Cre-lox system), and are controlled either through heat shock or administration of the ligand 4-hydroxytamoxifen (4-OHT). Despite their promise and advances, these techniques also have limitations and drawbacks, such as the leakiness of the estrogen receptor variant (ER^{T2}) that is used to modulate Cre (Boniface et al., 2009; Gerety et al., 2013; Mosimann and Zon, 2011) and the known toxicity and side-effects of Cre recombinase and the drug 4-OHT (Anastassiadis et al., 2010; Schmidt-Supprian and Rajewsky, 2007). Furthermore, there are reservations as to whether these drugs can penetrate all tissues, especially in adult fish. Moreover, the available number of Cre-lox lines in zebrafish is currently limited and restricts the application of these systems.

The binary Gal4-UAS expression system is a powerful, and commonly used, transgenic tool in the zebrafish. Since the introduction of the Gal4-UAS system in zebrafish more than a decade ago (Scheer and Campos-Ortega, 1999), hundreds of so called 'driver lines' have become available that express the transcriptional activator Gal4

under the control of a specific enhancer or promoter. Furthermore, the repertoire of 'effector-lines', which express Upstream Activated Sequence (UAS)-linked transgenes in specific tissues when activated by Gal4 binding to the UAS, is large and rapidly expanding. Animals that make use of this system express a specific gene in all cells of a certain type of tissue (depending on the Gal4-driver and UAS-effector line), and the surrounding tissues remain wild type. However, the ability to trace a single (often mutant) cell within a wild-type tissue is preferred for cell lineage tracing, gene function experiments and cancer modeling studies. To achieve this goal, we developed a system in which single cells express a gene – e.g. *Gal4* or oncogenic *H-RAS* – only when the ORF is placed in-frame after an *in vivo* frameshift mutation. Here, we show that the random activation of genes through microsatellite instability can be a valuable tool for mosaic analysis in zebrafish.

RESULTS

Microsatellite-dependent mosaicism

To investigate whether we could activate genes stochastically *in vivo* through microsatellite instability, we designed various reporter constructs in which we placed the coding sequence of LacZ downstream of a microsatellite-containing ORF. The constructs were designed in such a way that a frameshift-mutation within the microsatellite could bring the coding sequence of LacZ in-frame with the upstream ORF, for which we used the coding sequence of the enhanced green fluorescent protein (EGFP).

As a control, we placed the coding sequence of LacZ out-of-frame downstream of a random sequence (5′-GATTCTGCCAAGT-3′) that was not prone to frameshift mutations (Fig. 1A, upper reporter). Using Tol2-transposase-mediated transgenesis, which causes transient expression (Balciunas et al., 2006; Kawakami et al., 2004), we injected this reporter into one-cell stage embryos and analyzed the embryos for staining of LacZ two days later. We did not find any LacZ staining in any of the embryos that had been injected (n=100) (Fig. 1B, upper image). Although the upstream coding sequence of EGFP was in-frame, no EGFP expression was detected either, probably because the out-of-frame sequence of LacZ did not serve as a proper 3′UTR (comparable observations have been made in human cells; (Koole et al., 2013)). Next, we injected similar reporters in which we had replaced the random sequence with microsatellites that comprised 22, 23 or 24 guanines (G_{22}, G_{23} and G_{24}, respectively) resulting in reporters that had the ORF of LacZ in all three reading frames (Fig.1). The length of these microsatellites were chosen based on experiments in human cells, which have shown that microsatellites with a similar tract-length are highly prone to frameshift mutations in wild-type cells (Koole et al., 2013). Embryos that had been injected with an in-frame LacZ coding sequence displayed LacZ-staining in the majority of cells (Fig. 1, microsatellite G23). Importantly, in contrast with the reporter

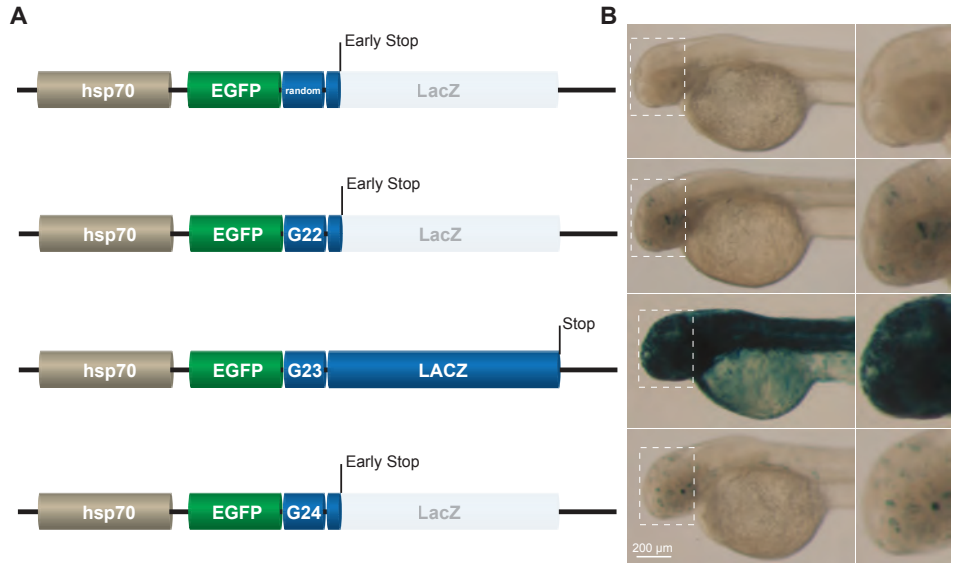


Figure 1 | Microsatellite-dependent stochastic gene activation in Danio rerio. (A) Schematic representation of reporters in which the coding sequence of LacZ was placed downstream of an EGFP ORF; a random sequence or a frameshift-prone G_{22}, G_{23} or G_{24} microsatellite tract was placed in between these two elements. The random sequence, G_{22} and G_{24} put the LacZ ORF out-of-frame. The reporter was under the control of the *hsp70* promoter. (B) Images of LacZ-stained embryos (2 dpf) that had been injected with the corresponding reporters shown in A. The embryos show microsatellite-dependent stochastic LacZ expression (blue) *in vivo*. One-cell stage embryos were rendered transgenic by using Tol2-transposase-mediated transgenesis. Inset images on the right correspond to the dashed box areas shown in the left-hand images.

that contained a random sequence, we observed stochastic expression of LacZ when its ORF was placed out-of-frame downstream of a microsatellite that comprised 22 or 24 guanines (Fig. 1). Different animals had different LacZ spots, both in terms of location and in the number of cells per spot, which reflects the stochastic nature of microsatellite-dependent gene activation. To further substantiate that reporter activation is the result of microsatellite instability, we assayed LacZ restoration in the reporters when they were injected into mismatch-repair-deficient animals (Feitsma et al., 2007). Previously, we have shown that the rate of frameshifts at microsatellites increases profoundly in mismatch-repair-defective *Caenorhabditis elegans* and human cells, using a similar type of microsatellite instability reporters (Koole et al., 2013; Pothof et al., 2003). Indeed, we also found increased rates of reporter activation in mismatch-repair-compromised zebrafish embryos (supplementary material Fig. S1), which provides strong support for causative frameshift events. Taken together, these data indicate that placing the coding sequence of a gene out-of-frame downstream of a microsatellite results in its stochastic expression *in vivo* due to microsatellite instability.

UAS-effector lines that stochastically express genes

To broaden the application of stochastic gene activation in the study of zebrafish biology, we adapted it for combined use with the binary Gal4-UAS system. The available number of Gal4-driver lines for the zebrafish community is large and growing, mainly because of enhancer trap screens (Asakawa et al., 2008; Davison et al., 2007; Distel et al., 2009; Scott et al., 2007). In order to be able to exploit this collection of driver lines and to perform lineage tracing in living animals, we established the UAS-effector line *Tg(UAS:H2A-G$_{23}$-EGFP)hu6243*, which stochastically marks individual cells with nuclear EGFP in tissues where Gal4 is expressed. This line carries the construct plm74, in which the coding sequences for histone 2A (H2A) and EGFP were placed downstream of a UAS-cassette. However, a G23 microsatellite was introduced in between H2A and EGFP, such that EGFP is out-of-frame (a schematic overview of the construct is given in Fig. 2A). We crossed this effector line *Tg(UAS:H2A-G$_{23}$-EGFP)hu6243* with a driver line *Et(E1b:Gal4-VP16)s1101t* (Scott et al., 2007), which widely expresses Gal4-VP16 in many tissues, with central nervous system neurons exhibiting the strongest expression (Schoonheim et al., 2010). In the progeny of that cross, we found stochastic expression of nuclear EGFP, which started at the same time (around the 10-somite stage) and location as Gal4-VP16 would be expressed (Fig. 2B; supplementary material Movie 1). Importantly, to show that this nuclear EGFP can be used as a stable marker to trace cells and their descendants in a living animal, we performed time-lapse imaging (Fig. 2B and supplementary material Movie 1), which demonstrated that nuclear EGFP was stably inherited upon cell division and that individual cells, and their descendants, could be monitored over time.

In order to further facilitate cell fate mapping experiments, we also generated fish containing a construct (plm76) that marked all of the tissue in which Gal4 was expressed. This construct also stochastically marked individual cells with nuclear EGFP. We modified a previously described bicistronic reporter construct (Trichas, 2007) that encodes a membrane-bound red fluorescent protein (myr-TdTomato), a viral 2A peptide that allows bicistronic expression (Szymczak et al., 2004) and a nuclear green fluorescent protein (H2A-EGFP). Additionally, here, we placed the reporter downstream of a UAS-cassette and replaced H2A-EGFP with H2A-G$_{23}$-EGFP so that the coding sequence of EGFP was placed out-of-frame (see Fig. 2C for a schematic overview of the construct). The transgenic founder fish *Tg(UAS:Myr-TdTomato-H2A-G$_{23}$-EGFP)hu6242* were crossed to *Et(E1b:Gal4-VP16)s1101t* fish.

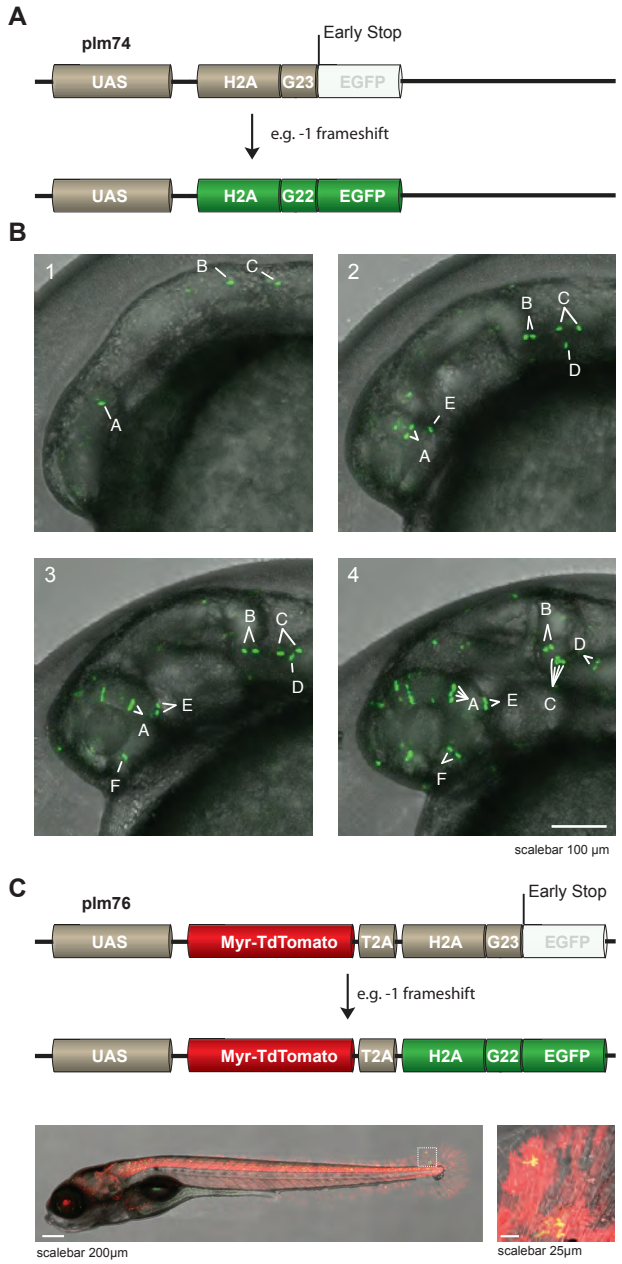


Figure 2 | UAS-reporters stochastically express H2A-EGFP. (A) (Upper panel) Schematic representation of reporter plm74 in which frameshifts at a G_{23} microsatellite can result in EGFP expression. Note that, for illustration purposes, we used the example of a −1 frameshift, however other, bigger, frameshifts can also lead to in-frame EGFP. UAS, upstream activator sequences. (B) Time-lapse images of the head region of an embryo derived from crossing *Et(E1b:Gal4-VP16)s1101t*fish to *Tg(UAS:H2A-G$_{23}$-EGFP)hu6243* fish. Stochastic expression of nuclear EGFP (green) in cells was observed, and individual cells and their descendants can be traced over time (denoted by the letters A-F in the images). Images 1–4 represent ▶ pictures taken around 16, 21, 25 and 31 hpf and correspond to supplementary material Movie 1. (C) (Upper panel) Schematic representation of reporter plm76 in which Myr-TdTomato is placed downstream of a UAS-cassette, followed by a T2A sequence and H2A-G$_{23}$-EGFP, meaning that EGFP is out-of-frame. Transgenic F1 animals [*Et(E1b:Gal4-VP16)s1101t* fish crossed to *Tg(UAS:Myr-TdTomato-H2A-G$_{23}$-EGFP)hu6242*] exhibit cells that express fluorescent membrane-labeled TdTomato and, in a mosaic pattern caused by *in vivo* stochastic frameshifting, EGFP-fluorescent nuclei (lower panels, an enlarged image of the boxed area is shown in the right-hand image).

We found that, in the Gal4-expressing tissues of progeny fish, the majority of cells expressed membrane localized TdTomato, some cells were also found to express nuclear H2A-EGFP (Fig. 2C). Individual cells could be distinguished easily by their red fluorescent membranes, and lineage tracing was facilitated by the random cells that were marked with green nuclear H2A-EGFP. We thus describe two UAS-effector lines – *Tg(UAS:H2A-G$_{23}$-EGFP)hu6243* and *Tg(UAS:Myr-TdTomato-H2A-G$_{23}$-EGFP)hu6242* – that can be used to perform cell fate mapping experiments in Gal4-expressing tissues in living animals.

Mosaic labeling with self-maintaining Kaloop

In order to follow the fate of all descendants of a single cell, permanent cell-labeling is required to avoid loss of signal when, for example, a promoter is only temporarily activated during embryogenesis. Indeed, a recognized drawback of the binary Gal4-UAS system is that cells and their descendants lose their label once the Gal4-expression is diminished (which depends on the tissue-specific promoter that drives Gal4 expression). This problem has been addressed in a recent report (Distel et al., 2009) through the establishment of a sophisticated self-maintaining system termed 'Kaloop'. In that study, a bicistronic reporter system was used that included KalTA4 (an optimized version of the Gal4-activator) downstream of a 4×UAS-cassette with EGFP and a T2A sequence, allowing expression of two proteins from the same single transcript (Szymczak et al., 2004). Once activated, KalTA4 maintains its own expression, and that of EGFP, through a positive-feedback loop, because KalTA4 binds in *cis* to its own UAS-promoter and leads to labeling of the complete cell lineage (Distel et al., 2009). We reasoned that lineage tracing by microsatellite instability using Gal4-UAS genetics can be further complemented when combined with the Kaloop system, so that a single cell and its descendants can be followed, even when driver-dependent Gal4 expression is lost. To test this, we adapted plm74 (UAS:H2A-G$_{23}$-EGFP) by placing the coding sequence of KalTA4 and a viral 2A sequence in the same reading frame as that of EGFP (see Fig. 3B for a schematic representation of this reporter; herein termed plm78). After a frameshift mutation, KalTA4 should maintain its own expression, as well as that of H2A-EGFP. Embryos that had been injected with *gal4* mRNA with plm74 or plm78 displayed stochastic labeling with EGFP of individual nuclei at 1 day post-fertilization (dpf) (Fig. 3A,B). Strikingly, embryos that had been injected with plm74 and *gal4* mRNA progressively lost EGFP expression, and at 5 dpf, nuclei that had been labeled previously were undetectable. By contrast, embryos that had been injected with plm78 and mRNA *gal4* maintained

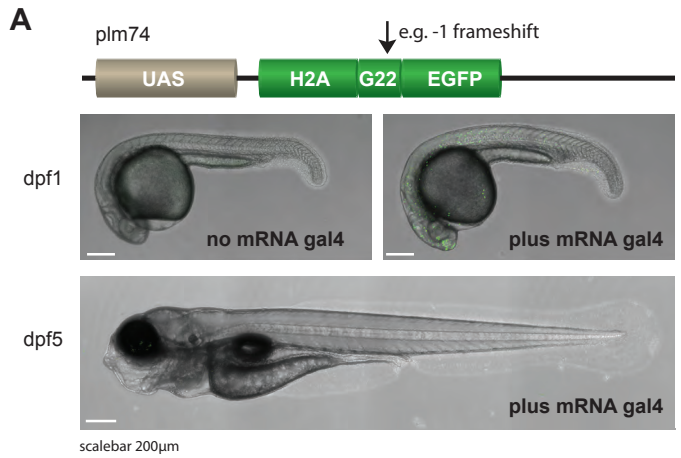


Figure 3 | Mosaic labeling with self-maintaining Kaloop. (A) Schematic representation of plm74 (upper panel) in which a –1 frameshift at a G$_{23}$microsatellite results in an ORF of H2A-G22-EGFP. The reporter was injected into one-cell stage embryos with or without *gal4* mRNA to temporarily activate the reporter. Stochastic expression of nuclear EGFP was observed 1 day after injection (right image) when *gal4* mRNA was co-injected, but not in the absence of *gal4* mRNA (left image). EGFP expression diminished within 5 days (lower image; quantified in C). (B) Schematic representation of plm78 in which, owing to a –1 frameshift, the ▶ coding sequence of EGFP, together with the linker peptide T2A and transcription activator KalTA4, becomes in-frame with the ORF of H2A-G22. Co-injection of *gal4* mRNA resulted in stochastic activation of nuclear EGFP in embryos at 1 dpf, and expression was still observed in larvae at 5 dpf. (C) Quantification of injected embryos (n=100 per condition) that express nuclear EGFP.

EGFP labeling of individual nuclei until, at least, 5 dpf (when animals were killed), indicating that self-maintained labeling of the cells had been established by the Kaloop system. Embryos that had been injected with plm74 without *gal4* mRNA did not show any expression of H2A-EGFP (Fig. 3A, left image). In six out of 100 embryos that had been injected with only plm78 (without co-injection of *gal4* mRNA), we observed a few cells that expressed H2A-EGFP, suggesting self-activation of the Kaloop construct. These results suggest that we have established a UAS-effector line that mosaically labels individual cells by harnessing microsatellite instability. Use of the Kaloop system maintains cell labeling, even when initial driver-dependent Gal4 expression is diminished, and optimizes cell fate mapping for Gal4-driver lines.

Stochastic activation of Gal4-VP16

In addition to Gal4 driver-lines, many UAS-effector zebrafish lines have become available. To establish a line that can stochastically activate these effector lines through microsatellite instability, we used a similar approach to that described above – the coding sequence for Gal4-VP16 was placed out-of-frame behind mCherry and a microsatellite of 23 guanines (Fig. 4A). The construct was placed downstream of a heat-shock-inducible promoter (*hsp70*) and used to establish the stable transgenic line *Tg(hsp70:mCherry-G_{23}-Gal4-VP16)hu7161* that was crossed with a UAS-Kaede line. Without heat shock, mCherry expression was observed in the lens (supplementary material Fig. S2, second left panel); therefore, transgenic animals could easily be recognized and selected. A similar observation has been made previously when using the same promoter (Blechinger et al., 2002; Boniface et al., 2009). After heat shock, we observed strong mCherry expression in the lens and weak, but detectable, expression

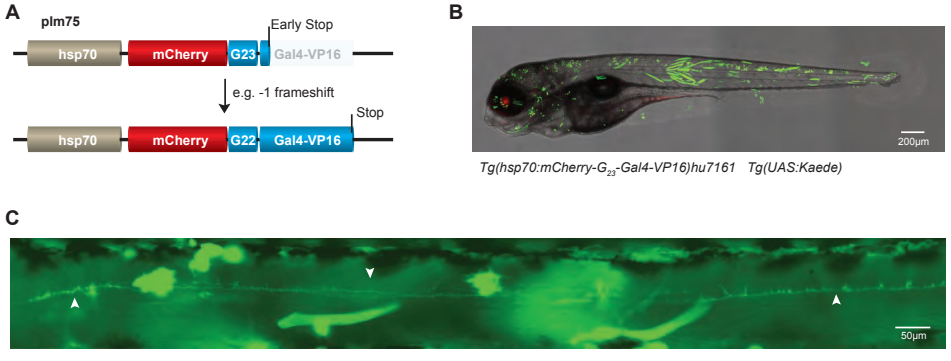


Figure 4 | Stochastic activation of Gal4-VP16. (A) (Upper panel) Schematic representation of plm75 in which the coding sequence of Gal4-VP16 is placed downstream of that of mCherry, but out-of-frame because of a G_{23} intervening sequence, all under the control of the heat shock *hsp70*promoter. (B) A representative (merged) image shows stochastic expression of Kaede in an embryo at 5 dpf from a cross of the transgenic line *Tg(hsp70:mCherry-G_{23}-Gal4-VP16)hu7161* with *Tg(UAS:Kaede)*. (C) Image of a single neuron that was activated stochastically (indicated with white arrows).

of mCherry in other embryonic tissues (supplementary material Fig. S2, second right panel). Importantly, in the same embryos, we identified mosaic expression of Kaede (Fig. 4B), often providing clear single cell resolution (Fig. 4C). These data show that the *Tg(hsp70:mCherry-G_{23}-Gal4-VP16)hu7161* line can be used to stochastically activate UAS-effector lines through microsatellite instability.

Sporadic tumor induction in zebrafish

Modeling human cancers in animals has yielded important findings, and sophisticated murine models have been established to study tumor development. The zebrafish is an emerging model organism in which to study tumorigenesis, mainly because of its ease of *in vivo* imaging and drug screening possibilities. Importantly, pathways that are involved in tumorigenesis – i.e. cell proliferation, angiogenesis, apoptosis – are highly conserved between human and zebrafish, and a wide range of cancers that resemble human malignancies have been identified in zebrafish (Amatruda and Patton, 2008). In many studies, tumor-induction is established through overexpression of an oncogene under the control of a tissue-specific promoter, resulting in tissue-wide overexpression of the oncogene. The current dogma for carcinogenesis is that tumor development starts with a genetic event (often caused by a replication defect) in a single cell that is surrounded with normal cells within that tissue. We reasoned that activation of an oncogene through microsatellite instability *in vivo* should better mimic sporadic carcinogenesis because it is stochastic, cell-division-dependent and restricted to individual cells. To demonstrate the importance of stochastic activation, we compared embryos that overexpressed a constitutively active form of the human oncogene H-RAS in complete tissues or stochastically. For these experiments, we made use of the transgenic line *Tg(UAS:EGFP-HRAS-G12V)io6* (Santoriello et al., 2009), which carries a glycine to valine mutation (G12V) in H-RAS. This mutation locks the protein in an active state and is a common mutation found in patients with the rare genetic disease Costello syndrome (Costello, 1977). H-RAS was tagged with EGFP at the N-terminus in order to visualize cells that express oncogenic H-RAS (Santoriello et al., 2009). First, we used the driver line *Tg(hsp70:Gal4)* (Scheer and Campos-Ortega, 1999) to overexpress EGFP-HRAS-G12V in complete tissues after heat shock. F1 embryos showed, predominantly, strong expression of EGFP-HRAS-G12V in the lens and trunk muscle cells, and most embryos died within 72 hours (Fig. 5A, left panel), which constrained further analysis. To exclude that the heat shock caused the lethality, we also subjected embryos from the same clutch that were not double transgenic to heat shock; all embryos appeared normal, indicating that the lethality that was observed in double transgenic animals was, indeed, owing to the overexpression of EGFP-HRAS-G12V. Next, we crossed the driver line *Tg(hsp70:mCherry-G_{23}-Gal4VP16)hu7161* with *Tg(UAS:EGFP-HRAS-G12V)io6* and subjected the F1 embryos to heat shock. Stochastic activation of EGFP-HRAS-G12V was observed in animals that had been heat shocked and, in contrast with larvae from a cross with hsp70:Gal4, the

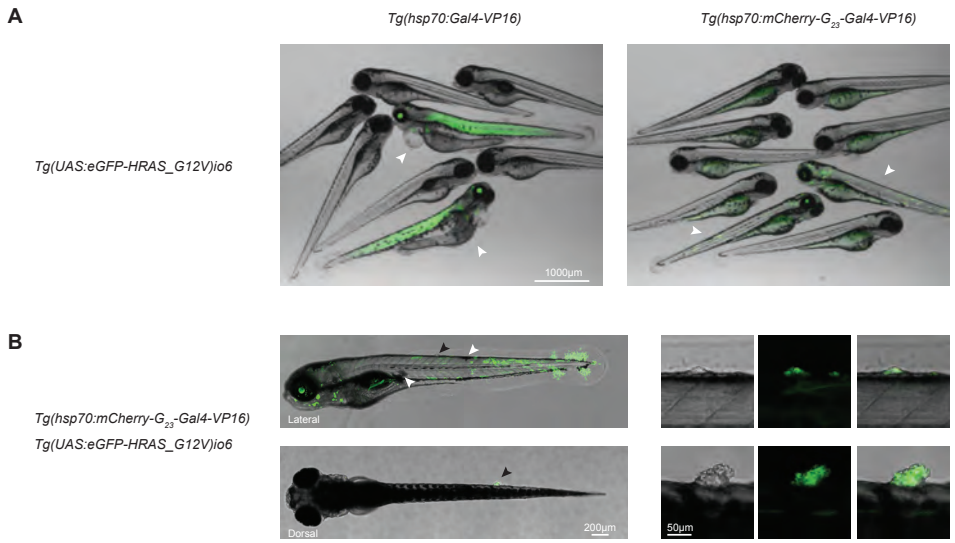


Figure 5 | Sporadic tumor induction. (A) Merged images show the difference between the expression of UAS:EGFP-HRAS-G12V in a tissue-wide (left) and stochastic (right) manner in embryos; embryos die at 3 dpf upon tissue-wide overexpression of EGFP-HRAS-G12V using *hsp70:Gal4* as an activator, whereas embryos with stochastic activation of EGFP-HRAS-G12V are healthy. White arrowheads indicate embryos that are transgenic for both alleles (the driver and effector alleles), other embryos carried neither of the alleles, or only one of them. (B) Lateral (top images) and dorsal (bottom images) view of a larvae (5 dpf) with stochastic activation of EGFP-HRAS-G12V. Hyperpigmentation (white arrowheads) and tumor formation (black arrowheads) are indicated. Right panels are enlarged images of the areas indicated by the black arrowheads. Brightfield, EGFP and merged channels are displayed.

larvae appeared healthy. Interestingly, within 5 days, we observed hyperpigmentation (Fig. 5B, white arrowheads; Table 1; supplementary material Fig. S2) – an early sign of melanoma development (Santoriello et al., 2010). Furthermore, we observed the abnormal growth of cells, which indicated the early onset of tumor formation in 47 out of the 130 embryos that we analyzed (Fig. 5B, black arrowheads and insets; Table 1). All of the early tumors that were observed were EGFP-positive, strongly suggesting that expression of EGFP-HRAS-G12V was driving the abnormal growth of these cells. Additionally, non-heat-shocked animals that were transgenic for EGFP-HRAS-G12V and hsp70:mCherry-G_{23}-Gal4-VP16 did not show any tumor formation or hyperpigmentation, again indicating that the overexpression of the oncogene was driving these processes (Table 1). Taken together, these data indicate that, by using microsatellite instability, we activated oncogenic H-RAS-G12V in individual cells, resulting in hyperpigmentation and the onset of sporadic tumor formation. These results highlight the benefit of the stochastic activation of oncogenes in tumor models because the competition between mutant cells and their surrounding wild-type cells can be easily monitored in the same tissue of a single organism, which is impossible when tissue-wide activation of oncogenes is employed.

Table 1 | Tumor induction by stochastic activation of oncogenic H-RAS through heat shock

Genotype	larvae analysed	Heat shock	larvae with tumors	larvae with hyper-pigmentation
Tg(hsp70:mCherry-G23-Gal4-VP16), Tg(UAS:EGFP-H-RAS_G12V)	130	30, 54, 78, 102 hpf	47	63
Tg(hsp70:mCherry-G23-Gal4-VP16), Tg(UAS:EGFP-H-RAS_G12V)	51	30 hpf	1	8
Tg(hsp70:mCherry-G23-Gal4-VP16), Tg(UAS:EGFP-H-RAS_G12V)	82	-	0	0
Tg(hsp70:mCherry-G23-Gal4-VP16) / *Tg(UAS:EGFP-H-RAS_G12V)* / wild type (same clutch) *	198	30, 54, 78, 102 hpf	0	0

* The three control zebrafish lines were analyzed cumulatively.

DISCUSSION

Here, we describe an alternative approach to create mosaicism in zebrafish, which depends on microsatellite instability and avoids labor intensive invasive techniques and limitations that are associated with techniques that require the administration of drugs (Emelyanov and Parinov, 2008; Esengil et al., 2007; Gerety et al., 2013; Hans et al., 2009; 2011; Knopf, 2010). Furthermore, we designed our system in a way that it is fully compatible with the large available collection of Gal4-driver and UAS-effector lines.

Although we optimized our approach for the binary Gal4-UAS system, stochastic expression of any gene is possible when the coding sequence is placed behind a microsatellite. Our constructs are designed such that the promoters and coding sequence can be exchanged with relative ease. Additionally, the length of the microsatellite can be altered. Because the frequency of frameshifts depends on the length of the microsatellite (for an example, see (Koole et al., 2013)), it is possible to increase or decrease the level of stochastic events when varying the tract-length of the microsatellite.

Microsatellite stability is greatly affected by the status of the mismatch repair (MMR) pathway (for a recent review about MMR, see (Jiricny, 2013)) , and, as illustrated in supplementary material Fig. S1, the established lines that contain a microsatellite-reporter offer the potential to study MMR *in vivo*. A study in *C. elegans* using a similar type of reporter has been valuable in the identification of novel genes that are involved in microsatellite instability, allowing the screening of animals that showed enhanced activation of a microsatellite instability reporter (Pothof et al., 2003). Our established transgenic lines, which use microsatellite instability as a read out, provide the potential to find possible new alleles that are involved in MMR when

using forward genetic screens, to test candidate genes by reverse genetics approaches – for example, using morpholinos or CRISPR technology (Hwang et al., 2013) – to test MMR-related compounds or to investigate in whole animals those tissues or cells that are more prone to microsatellite instability. These opportunities to study MMR and microsatellite instability *in vivo* in zebrafish is an attractive approach to gain new insights into the MMR-associated Lynch syndrome.

Microsatellite instability is dependent on replication, and thus frameshift mutations can also occur in replicating germ cells. One possible concern is the inheritance of a germline frameshift event that results in progeny having full expression of the transgene in all cells instead of mosaic expression. During the course of our experiments and maintenance of our transgenic lines we found one fish (out of ~250 analyzed) that inherited a germline event and gave rise to embryos in which all cells were labeled. We reason that the frequency of such germline events is low and, therefore, should not provide complications for the maintenance of stable lines over many generations.

In this study, we used microsatellite-dependent activation of oncogenic EGFP-HRAS-G12V to mimic the initial steps of tumorigenesis. Mosaic analysis through microsatellite-dependent oncogene activation has several advantages over other available strategies. First, microsatellite instability is dependent on cell division, and therefore oncogene activation only occurs in proliferative cells. Because 'driver mutations' are required for tumorigenesis (Vogelstein et al., 2013) and DNA is most vulnerable to mutations during replication (Aguilera and García-Muse, 2013), it is generally believed that most tumors arise owing to a mutational event in proliferative cells. Other techniques that stochastically activate genes also activate non-dividing cells, which might interfere with accurate reproduction of the early steps of tumorigenesis. Second, microsatellite-dependent activation is restricted to single cells; the chance that a neighboring cell obtains a frameshift mutation at the same time is negligible. This ensures that groups of activated cells are clonally derived. When using other techniques, for example treatment with 4-OHT, there is a chance that neighboring cells are also affected at the same time, which can complicate clonal analysis. Finally, combining this microsatellite instability technique with other techniques allows for the design of new experimental setups. For example, most tumors contain at least two 'driver' mutations (Vogelstein et al., 2013), and in order to be able to model two consecutive stochastic events, it is desirable that both events can be induced by two separate techniques that do not interfere with each other.

Modeling sporadic cancers with the help of microsatellite-dependent gene activation has also proven valuable in mouse models (Akyol et al., 2008; Miller et al., 2008). In those studies, the coding sequence of Cre was placed downstream of a microsatellite, enabling stochastic bi-allelic inactivation of floxed tumor suppressor genes or the activation of oncogenes. Although inducible inactivation of tumor suppressor genes is still restricted in zebrafish, we show that stochastic activation of oncogenes is possible in zebrafish. Interestingly, we also show that tumors can be observed easily within days, whereas the timeframe in which tumor development is studied in mice is usually weeks up to months. This temporal advantage, together with the relatively low cost of fish maintenance, the development of high-throughput screening and imaging techniques (Pardo-Martin et al., 2010) to test small molecule libraries (Stern et al., 2005) and the ease of imaging tumor development, makes zebrafish an attractive model in which to study cancer development and perform related drug discovery *in vivo*.

The first mosaic experiments that were employed in zebrafish used invasive techniques; however, in recent years, several genetic non-invasive techniques have been developed to improve the mosaic labeling of cells in zebrafish. Because all techniques have advantages and limitations, it is of great importance that scientists have a variety of tools from which they can choose (and then combine) to best suit their experiments. The use of the microsatellite-dependent activation of transgenes will expand the 'toolbox' for mosaic analysis in zebrafish and provide new opportunities to perform cell lineage tracing experiments, gene function studies and research into tumor biology.

MATERIAL AND METHODS

Plasmid construction. All plasmids comprised six elements (with the exception of plm78 and plm76) – (1) the pminiTol2 plasmid (pDB739) as backbone (Balciunas et al., 2006); (2) a *hsp70*promoter (Halloran et al., 2000) or 14×UAS sequences, both flanked with a *Swa*I and a *Kpn*I restriction site; (3) a coding sequence (e.g. mCherry, EGFP or H2A) starting with a Kozak sequence but lacking a stop codon and flanked with a *Kpn*I and an *Nhe*I restriction site; (4) a microsatellite or random sequence (5′-GATTCTGCCAAGT-3′) with flanking *Nhe*I and *Bam*HI restriction sites; (5) a coding sequence [EGFP or Gal4-VP16 (Köster and Fraser, 2001)] without a start codon but including a stop codon that was flanked with *Bam*HI and *Xba*I restriction sites; (6) a SV40 3′UTR flanked with *Xba*I and *Hin*dIII restriction sites. Elements two, three and five were obtained through PCR-amplification, element four was obtained through the cloning of DNA oligos. All elements allow easy substitution because most of the flanking restriction sites are unique. Plm78 was created by swapping KGFP from the plasmid 4×Kaloop (pTolmini-4xUASKGFP-T2A-KalTA4GI) (Distel et al., 2009) with H2A-G_{23}-EGFP using the restriction sites *Brg*I and *Eco*RI. To create plm76, we used the construct Tom-2A-GFP (Trichas, 2007) in which we replaced the promoter with a UAS cassette, and H2B-GFP with H2A-G_{23}-EGFP.

Fish maintenance, transgenesis and transgene-induction. Wild-type and transgenic embryos were obtained by natural spawning of adult fish that were maintained at 28.5°C. For transgenesis, one-cell stage embryos (F0) were co-injected with 5–20 pg of plasmid DNA and 20 pg of transposase mRNA, as described previously (Balciunas et al., 2006; Kawakami et al., 2004; Kawakami, 2005). Injected embryos were grown until adulthood, crossed with Gal4- or UAS-lines, and the F1 progeny were examined by fluorescent microscopy. Positive embryos were selected and raised. For heat-shock treatment, embryos were transferred to a 50 ml tube and heat shocked for 30 minutes at 37°C in a water bath.

Microscopy. For microscopy, embryos were anesthetized with Tricaine (Sigma) and fixed using 0.5% low-melting-point agarose in glass-bottomed dishes (MatTek). Images were taken by using a Leica TCS SP5-STED microscope. Image stacks were taken with a 10× objective and 1.3 digital zoom. For time-lapse imaging, embryos (kept in the chorion) were fixed in 0.5% low-melting-point agarose, and a heated-stage chamber was used to keep embryos at 28.5°C. Image stacks were taken every 25 minutes. Images were processed using Leica software for automatic stitching and ImageJ to create overlays and maximum projections.

X-gal staining. Embryos were heat shocked for 30 minutes at 37°C, left to recover for 6 hours, fixed on ice for 30 minutes in freshly prepared fixing solution (PBS, 1% formaldehyde, 0.2% gluteraldehyde, 0.02% IGEPAL) and then washed three times for 20 minutes in PBS at room temperature. Next, the embryos were stained for β-galactosidase expression overnight in X-gal solution (0.16 M Na_2HPO_4, 0.03 M NaH_2PO_4, 0.2 mM $MgCl_2$, 0.1 mM SDS, 5 mM $[Fe(CN)_6]^{3-}$, 5 mM $[Fe(CN)_6]^{4-}$, 1 mM X-gal). All fixation, washing and staining steps were performed on a shaking platform.

Tumor induction. To induce expression of EGFP-HRAS-G12V, we heat shocked embryos (which had been transferred to a 50 ml tube) for 30 minutes at 37°C at 30 hpf, 54 hpf, 78 hpf and 102 hpf, unless stated otherwise.

ACKNOWLEDGEMENTS

We thank Marina Mione, Jeroen Bakkers, Karli Reiding, Peter Schoonheim, Reinhard Köster, Bas Ponsioen, Shankar Srinivas, Federico Tessadori, Pim Toonen and Edwin Cuppen for sharing ideas, reagents and transgenic lines; Joop Wiegant and Annelies van der Laan for support with imaging; and Jeroen Bussmann for comments on this manuscript.

FUNDING

European Research Council [203379, DSBrepair]; European Commission [DDR Response]; Zorg Onderzoek Nederland, Medische Wetenschappen, Netherlands Genomics Initiative – Horizon.

AUTHOR CONTRIBUTIONS

W.K. and M.T. conceived and designed the experiments, W.K. performed the experiments. W.K. and M.T. analyzed the data. W.K. and M.T. wrote the paper.

REFERENCES

Aguilera, A., and García-Muse, T. (2013). Causes of genome instability. Annu. Rev. Genet. *47*, 1–32.

Akyol, A., Hinoi, T., Feng, Y., Bommer, G.T., Glaser, T.M., and Fearon, E.R. (2008). Generating somatic mosaicism with a Cre recombinase-microsatellite sequence transgene. Nature Methods *5*, 231–233.

Amatruda, J.F., and Patton, E.E. (2008). Genetic models of cancer in zebrafish. International Review of Cell and Molecular Biology *271*, 1–34.

Anastassiadis, K., Glaser, S., Kranz, A., Bernhardt, K., and Stewart, A.F. (2010). Chapter Seven-A Practical Summary of Site-Specific Recombination, Conditional Mutagenesis, and Tamoxifen Induction of CreERT2. Meth. Enzymol. *477*, 109–123.

Asakawa, K., Suster, M.L., Mizusawa, K., Nagayoshi, S., Kotani, T., Urasaki, A., Kishimoto, Y., Hibi, M., and Kawakami, K. (2008). Genetic dissection of neural circuits by Tol2 transposon-mediated Gal4 gene and enhancer trapping in zebrafish. Proc. Natl. Acad. Sci. U.S.a. *105*, 1255–1260.

Balciunas, D., Wangensteen, K.J., Wilber, A., Bell, J., Geurts, A., Sivasubbu, S., Wang, X., Hackett, P.B., Largaespada, D.A., McIvor, R.S., et al. (2006). Harnessing a high cargo-capacity transposon for genetic applications in vertebrates. PLoS Genet. *2*, e169–e169.

Blackburn, J.S., and Langenau, D.M. (2010). aMAZe-ing tools for mosaic analysis in zebrafish. Nature Methods *7*, 188–190.

Blechinger, S.R., Evans, T.G., Tang, P.T., Kuwada, J.Y., Warren, J.T., and Krone, P.H. (2002). The heat-inducible zebrafish hsp70 gene is expressed during normal lens development under non-stress conditions. Mechanisms of Development *112*, 213–215.

Boniface, E.J., Lu, J., Victoroff, T., Zhu, M., and Chen, W. (2009). FlEx-based transgenic reporter lines for visualization of Cre and Flp activity in live zebrafish. Genesis *47*, 484–491.

Buckingham, M.E., and Meilhac, S.M. (2011). Tracing cells for tracking cell lineage and clonal behavior. Developmental Cell *21*, 394–409.

Carmany-Rampey, A., and Moens, C.B. (2006). Modern mosaic analysis in the zebrafish. Methods *39*, 228–238.

Collins, R.T., Linker, C., and Lewis, J. (2010). MAZe: a tool for mosaic analysis of gene function in zebrafish. Nature Methods *7*, 219–223.

Costello, J.M. (1977). A new syndrome: mental subnormality and nasal papillomata. Aust Paediatr J *13*, 114–118.

Davison, J.M.J., Akitake, C.M.C., Goll, M.G.M., Rhee, J.M.J., Gosse, N.N., Baier, H.H., Halpern, M.E.M., Leach, S.D.S., and Parsons, M.J.M. (2007). Transactivation from Gal4-VP16 transgenic insertions for tissue-specific cell labeling and ablation in zebrafish. Dev. Biol. *304*, 811–824.

Distel, M., Wullimann, M.F., and Köster, R.W. (2009). Optimized Gal4 genetics for permanent gene expression mapping in zebrafish. Proc. Natl. Acad. Sci. U.S.a. *106*, 13365–13370.

Emelyanov, A., and Parinov, S. (2008). Mifepristone-inducible LexPR system to drive and control gene expression in transgenic zebrafish. Dev. Biol. *320*, 113–121.

Esengil, H., Chang, V., Mich, J.K., and Chen, J.K. (2007). Small-molecule regulation of zebrafish gene expression. Nat. Chem. Biol. *3*, 154–155.

Feitsma, H., Leal, M.C., Moens, P.B., Cuppen, E., and Schulz, R.W. (2007). Mlh1 deficiency in zebrafish results in male sterility and aneuploid as well as triploid progeny in females. Genetics *175*, 1561–1569.

Gerety, S.S., Breau, M.A., Sasai, N., Xu, Q., Briscoe, J., and Wilkinson, D.G. (2013). An inducible transgene expression system for zebrafish and chick. Development *140*, 2235–2243.

Halloran, M.C.M., Sato-Maeda, M.M., Warren, J.T.J., Su, F.F., Lele, Z.Z., Krone, P.H.P., Kuwada, J.Y.J., and Shoji, W.W. (2000). Laser-induced gene expression in specific cells of transgenic zebrafish. Development *127*, 1953–1960.

Hans, S., Freudenreich, D., Geffarth, M., Kaslin, J., Machate, A., and Brand, M. (2011). Generation of a non-leaky heat shock-inducible Cre line for conditional Cre/lox strategies in zebrafish. Dev Dyn *240*, 108–115.

Hans, S., Kaslin, J., Freudenreich, D., and Brand, M. (2009). Temporally-controlled site-specific recombination in zebrafish. PLoS ONE *4*, e4640.

Hwang, W.Y., Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Peterson, R.T., Yeh, J.-R.J., and Joung, J.K. (2013). Efficient genome editing in zebrafish using a CRISPR-Cas system. Nat. Biotechnol. *31*, 227–229.

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

Kawakami, K.K. (2005). Transposon tools and methods in zebrafish. Dev Dyn *234*, 244–254.

Kawakami, K., Takeda, H., Kawakami, N., Kobayashi, M., Matsuda, N., and Mishina, M. (2004). A transposon-mediated gene trap approach

identifies developmentally regulated genes in zebrafish. Developmental Cell *7*, 133–144.

Knopf (2010). Dually inducible TetON systems for tissue-specific conditional gene expression in zebrafish. Proc. Natl. Acad. Sci. U.S.a. *107*, 19933–19938.

Koole, W., Schäfer, H.S., Agami, R., van Haaften, G., and Tijsterman, M. (2013). A versatile microsatellite instability reporter system in human cells. Nucleic Acids Res. *41*, e158.

Köster, R.W., and Fraser, S.E. (2001). Tracing transgene expression in living zebrafish embryos. Dev. Biol. *233*, 329–346.

Miller, A.J., Dudley, S.D., Tsao, J.-L., Shibata, D., and Liskay, R.M. (2008). Tractable Cre-lox system for stochastic alteration of genes in mice. Nature Methods *5*, 227–229.

Mosimann, C., and Zon, L.I. (2011). Advanced zebrafish transgenesis with Tol2 and application for Cre/lox recombination experiments. Methods Cell Biol. *104*, 173–194.

Pan, Y.A., Livet, J., Sanes, J.R., Lichtman, J.W., and Schier, A.F. (2011). Multicolor Brainbow imaging in zebrafish. Cold Spring Harb Protoc *2011*, pdb. prot5546.

Pardo-Martin, C., Chang, T.-Y., Koo, B.K., Gilleland, C.L., Wasserman, S.C., and Yanik, M.F. (2010). High-throughput in vivo vertebrate screening. Nature Methods *7*, 634–636.

Pothof, J., van Haaften, G., Thijssen, K., Kamath, R.S., Fraser, A.G., Ahringer, J., Plasterk, R.H.A., and Tijsterman, M. (2003). Identification of genes that protect the C. elegans genome against mutations by genome-wide RNAi. Genes Dev. *17*, 443–448.

Santoriello, C.C., Gennaro, E.E., Anelli, V.V., Distel, M.M., Kelly, A.A., Köster, R.W.R., Hurlstone, A.A., and Mione, M.M. (2010). Kita driven expression of oncogenic HRAS leads to early onset and highly penetrant melanoma in zebrafish. PLoS ONE *5*, e15170–e15170.

Santoriello, C., Deflorian, G., Pezzimenti, F., Kawakami, K., Lanfrancone, L., d'Adda di Fagagna, F., and Mione, M. (2009). Expression of H-RASV12 in a zebrafish model of Costello syndrome causes cellular senescence in adult proliferating cells. *2*, 56–67.

Scheer, N., and Campos-Ortega, J.A. (1999). Use of the Gal4-UAS technique for targeted gene expression in the zebrafish. Mechanisms of Development *80*, 153–158.

Schmidt-Supprian, M., and Rajewsky, K. (2007). Vagaries of conditional gene targeting. Nat. Immunol. *8*, 665–668.

Schoonheim, P.J.P., Arrenberg, A.B.A., Del Bene, F.F., and Baier, H.H. (2010). Optogenetic localization and genetic perturbation of saccade-generating neurons in zebrafish. J Neurosci *30*, 7111–7120.

Scott, E.K., Mason, L., Arrenberg, A.B., Ziv, L., Gosse, N.J., Xiao, T., Chi, N.C., Asakawa, K., Kawakami, K., and Baier, H. (2007). Targeting neural circuitry in zebrafish using GAL4 enhancer trapping. Nature Methods *4*, 323–326.

Stern, H.M., Murphey, R.D., Shepard, J.L., Amatruda, J.F., Straub, C.T., Pfaff, K.L., Weber, G., Tallarico, J.A., King, R.W., and Zon, L.I. (2005). Small molecules that delay S phase suppress a zebrafish bmyb mutant. Nat. Chem. Biol. *1*, 366–370.

Szymczak, A.L., Workman, C.J., Wang, Y., Vignali, K.M., Dilioglou, S., Vanin, E.F., and Vignali, D.A.A. (2004). Correction of multi-gene deficiency in vivo using a single "self-cleaving" 2A peptide-based retroviral vector. Nat. Biotechnol. *22*, 589–594.

Trichas (2007). Use of the viral 2A peptide for bicistronic expression in transgenic mice. BMC Biol *6*, 40–40.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558.

Weber, T., and Köster, R. (2013). Genetic tools for multicolor imaging in zebrafish larvae. Methods *62*, 279–291.

White, R.M., Sessa, A., Burke, C., Bowman, T., LeBlanc, J., Ceol, C., Bourque, C., Dovey, M., Goessling, W., Burns, C.E., et al. (2008). Transparent adult zebrafish as a tool for in vivo transplantation analysis. Cell Stem Cell *2*, 183–189.

Xu, T., and Rubin, G.M. (2012). The effort to make mosaic analysis a household tool. Development *139*, 4501–4503.

RESOURCE IMPACT

Background

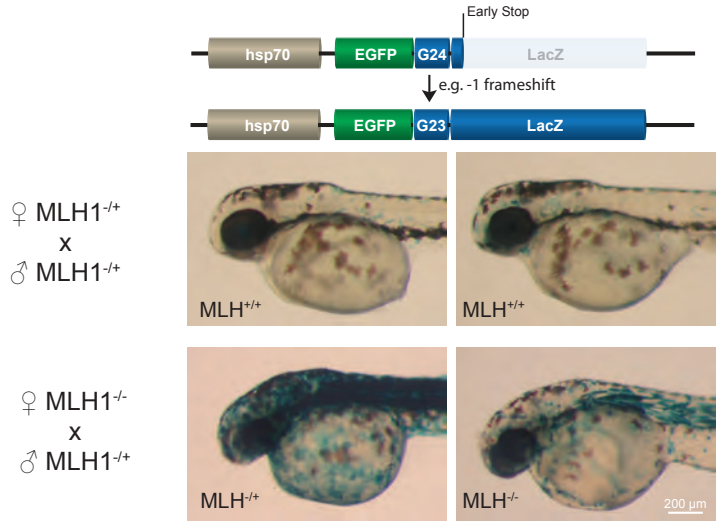
The zebrafish is an elegant and powerful vertebrate model system that is increasingly being used to study diseases and their underlying molecular mechanisms. Its small size, its fast rate of reproduction, the ease and relatively low costs of culturing, its striking anatomical and physiological similarities to mammals, and its transparency make the zebrafish a valuable model in which to study human diseases and to test drugs. However, mostly because of a lack of appropriate reagents and technology, the use of zebrafish as a model system in which to mark individual cells that are genetically different from surrounding cells and then follow their fate has been under-addressed. Such an experimental model would be particularly relevant to the study of cancer, which is a process in which single cells grow out to become malignant tumors through a process of stochastic mutations followed by selection for increased growth within an organism that is itself not genetically compromised or challenged.

Results

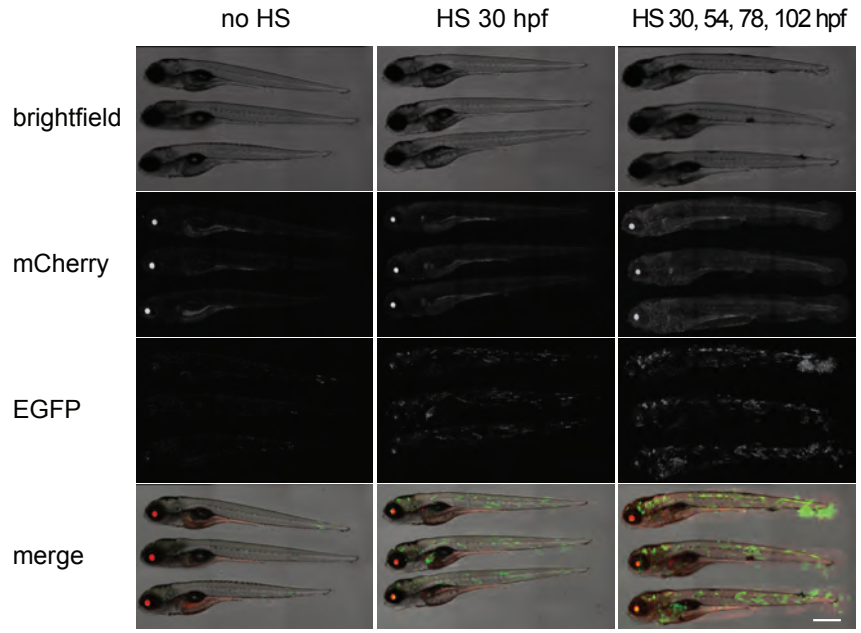
In this article, the authors describe a technology to genetically alter and then trace individual transformed cells in living zebrafish. They show that genes can be stochastically activated *in vivo* by placing their coding sequences out-of-frame downstream of microsatellite sequences, which are prone to frameshifts during DNA replication. Because the gene of interest is initially out-of-frame, low frequency *in vivo* stochastic frameshifting activates the gene of interest in occasional cells, which can be traced if the gene of interest is tagged with a fluorescent marker protein. Thus, using this approach, the fate of single altered cells that are surrounded by wild-type cells can be determined in living animals. The authors demonstrate that this method can also be used to mimic tumor development. Specifically, they show that microsatellite-dependent stochastic activation of oncogenic H-RAS results in the formation of tumors within 5 days.

Implications and future directions

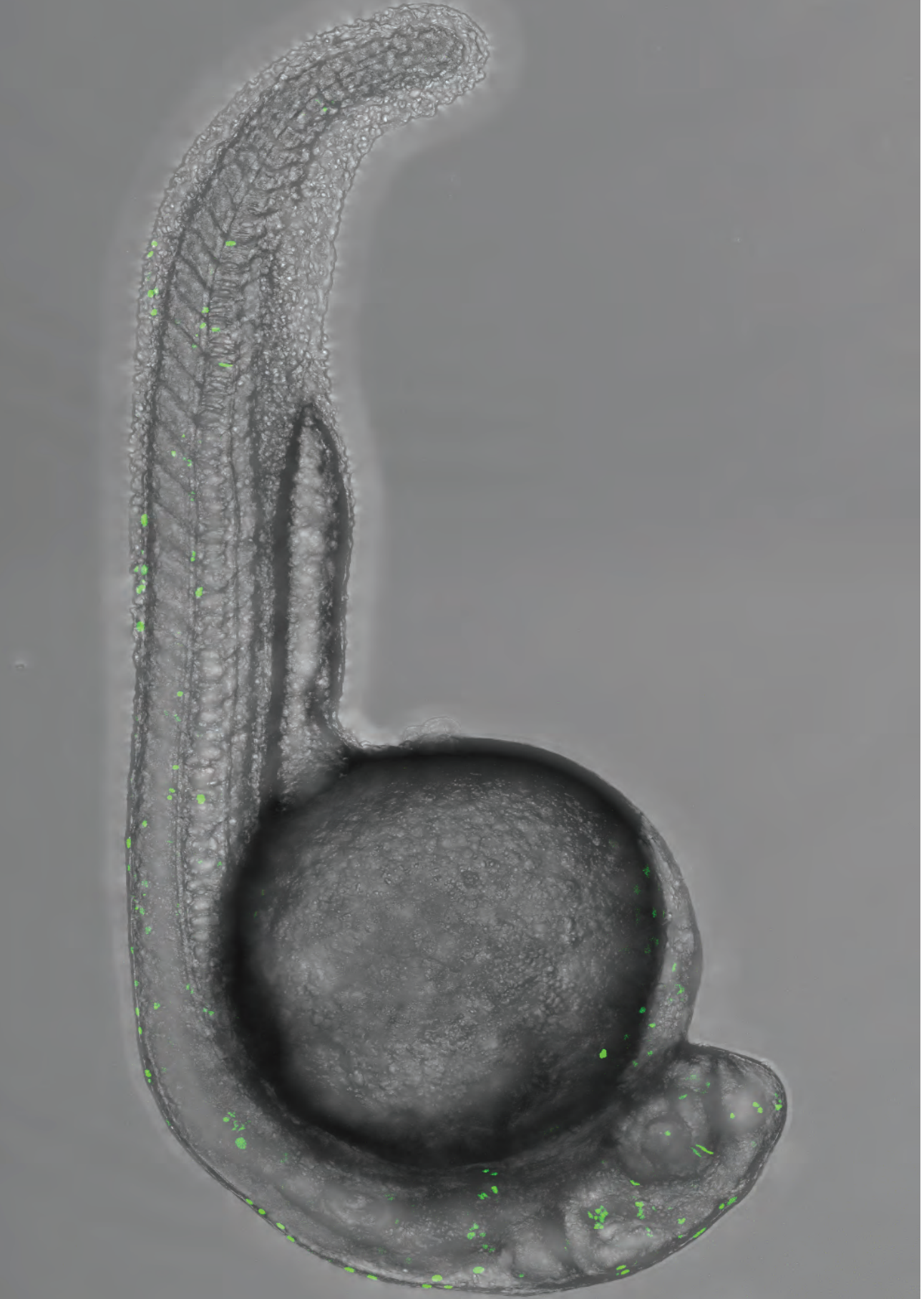
This study describes a new model system in which to trace single cells in living animals and to induce and monitor tumorigenesis. Because of its modular nature, this system can be easily adapted to study any (disease-related) protein of interest. Thus, the technology can be used to study and monitor the oncogenic effect of any (onco)gene of choice. Moreover, the fish system will also facilitate the search for cognate drugs, thus ultimately leading to a better understanding of the pathology of cancer and other diseases, and to the development of new therapeutics.



Supplementary figure 1 | Increased microsatellite instability reporter activation in mismatch repair deficient embryos. Images of embryos (2dpf) stained for LacZ expression. One-cell stage embryos were made transgenic via Tol2-transposase mediated transgenesis with the indicated microsatellite instability construct that reports LacZ expression when the LacZ-coding sequence becomes in-frame after a frameshift. Strong LacZ-staining (examples shown in bottom panels) was seen in 20 out of 22 injected viable embryos that were obtained from a cross between an MLH1$^{-/-}$ mother and an MLH1$^{-/+}$ father. Genotyping of animals showed that also heterozygous embryos from this cross had strong LacZ-staining, suggesting that the compromised mismatch repair pathway (due to MLH1-deficiency in the yolk) is not rescued by the paternal MLH1 wildtype allele during early development. Injected embryos obtained from MLH1$^{-/+}$ parents, showed LacZ-staining similar to wild type levels (top panels and Fig. 1).



Supplementary figure 2 | Random activation of oncogenic H-RAS by stochastic Gal4 expression. Larvae were heat shocked (HS) at various time points, as indicated. Larvae were derived from a cross *Tg(hsp70:mCherry-G$_{23}$-Gal4VP16)hu7161* with *Tg(UAS:EGFP-H-RAS_G12V)io6*. Larvae subjected to heat shock treatment showed stochastic expression of EGFP-H-RAS(G12V), hyper-pigmentation and tumor formation. The level of these phenotypes was depended on the number of heat shock treatments (see Table 1 for quantification). Images were taken at 126 hpf. Scalebar indicates 0.5 mm.



◀ This image shows a one-day-old zebrafish embryo. Individual cells that produce a green fluorescent protein in their nucleus can be distinguished and tracked over time during embryo development.

Cover Photo: 384 ▶

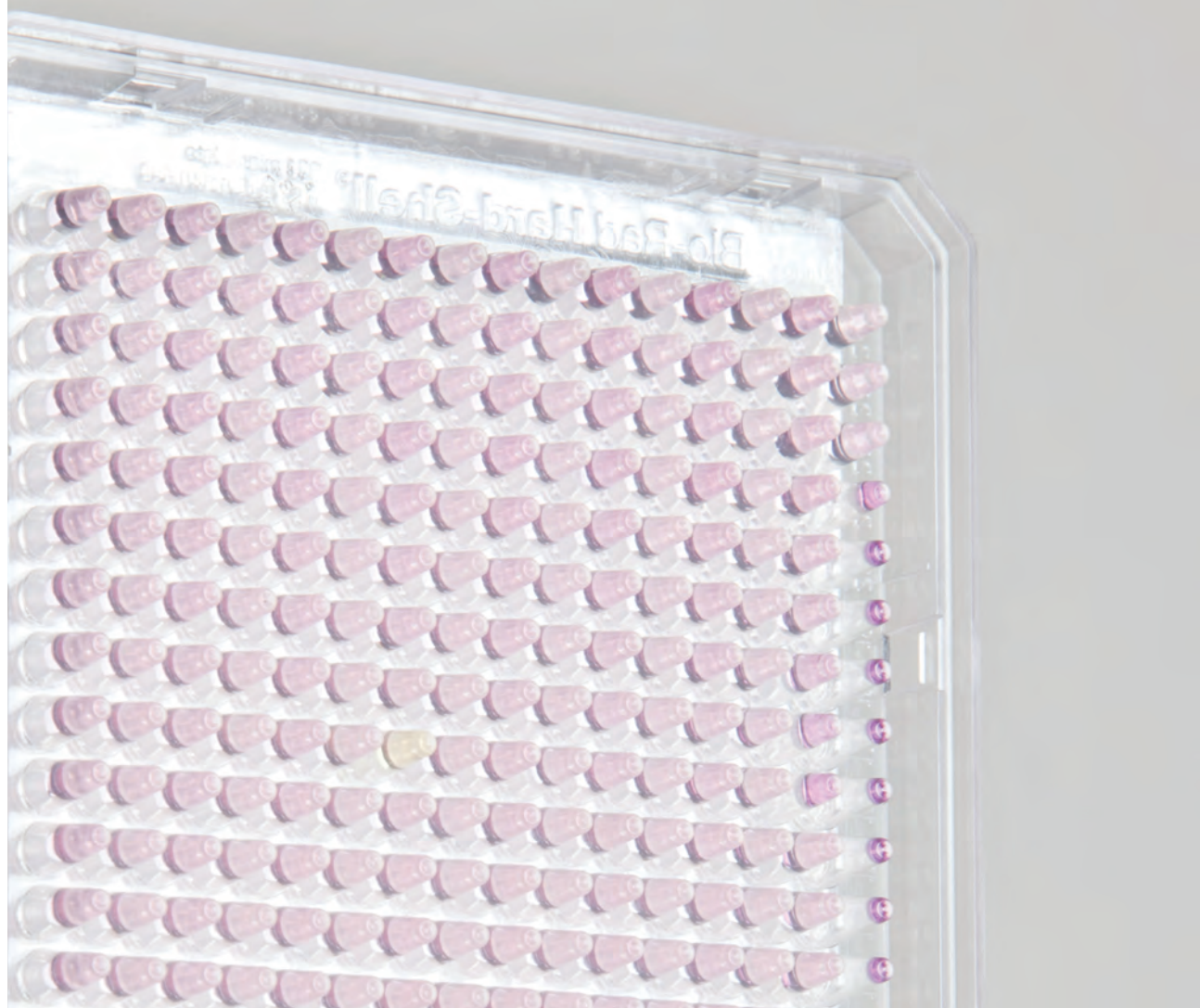
Chapter 4

A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites

Wouter Koole[1], Robin van Schendel[1], Andrea E. Karambelas[2], Jane T. van Heteren[1], Kristy L. Okihara[2] and Marcel Tijsterman[1]

[1]Department of Toxicogenetics, Leiden University Medical Center, Leiden, The Netherlands.
[2]Hubrecht Institute - KNAW - Utrecht University Medical Center, Utrecht, The Netherlands

Adapted from Koole et al. Nature Commun. 5:3216 (2014)



ABSTRACT

Genomes contain many sequences that are intrinsically difficult to replicate. Tracts of tandem guanines, for instance, have the potential to adopt stable G-quadruplex structures, which are prone to cause genome alterations. Here, we describe G4 DNA-induced mutagenesis in *C. elegans* and identify a non-canonical DNA break repair mechanism that generates deletions characterized by an extremely narrow size distribution, minimal homology of exactly one nucleotide at the junctions, and by the occasional presence of templated insertions. This typical mutation profile is fully dependent on the A-family polymerase Theta, the absence of which leads to profound loss of sequences surrounding G4 motifs. Theta-mediated end joining prevails over non-homologous end joining and homologous recombination and prevents genomic havoc at replication fork barriers at the expense of small deletions. G4 DNA-induced deletions also manifest in the genomes of wild isolates of *C. elegans*, indicating a protective role for this pathway during evolution.

INTRODUCTION

To preserve the integrity of genetic information, DNA replication needs to be accurate but also very efficient, as the entire cellular DNA needs to be duplicated before chromosomes can reliably be segregated to the two daughter cells during mitosis. DNA damage, as well as alternative DNA structures that obstruct replication fork progression, thus represent threats to chromosome integrity and to the faithful inheritance of genomes. One non-Watson-Crick DNA structure that is thermodynamically very stable under physiological conditions is a so-called G-quadruplex (Fig. 1a). While guanine-rich single-stranded DNA can readily adopt such a structure *in vitro*, its *in vivo* formation has remained a subject of debate ever since the structure was first described (Gellert et al., 1962). Recent years have, however, seen a greatly increased interest in G-quadruplexes due to new insights into their possible roles in various biological processes such as gene expression, epigenetic regulation, telomere maintenance and DNA replication initiation (Besnard et al., 2012; Law et al., 2010; Sarkies et al., 2012; 2010; Schwab et al., 2013; Smith et al., 2011; Vannier et al., 2012). While genome-wide *in silico* analyses have identified more than 300,000 G4 motifs (sites with G-quadruplex-forming potential) in the human genome (Huppert and Balasubramanian, 2005), perhaps the best *in vivo* evidence thus far for a functional role of G4 DNA comes from the bacterium *Neisseria gonorrhoeae*. This human pathogen evades its host's immune system by changing the identity of its surface antigen via a G4 DNA-induced recombination event (Cahoon and Seifert, 2009). Importantly, however, this example also illustrates the recombinogenic and thus potentially pathological nature of this highly thermal stable non-B-DNA structure – a concern further fuelled by the notion of hampered DNA replication near G4 sequences in yeast (Lopes et al., 2011; Paeschke et al., 2011; Ribeyre et al., 2009) and the observed association of G4 motifs with structural genomic variations in human cancers (De and Michor, 2011).

G4 DNA-induced genomic alterations were first identified in *C. elegans*, where the DOG-1/FANCJ helicase (Youds et al., 2008) was shown to suppress deletions initiated at G-rich DNA sequences (Cheung et al., 2002). Later work demonstrated that only G-rich DNA sequences that match the signature G4 motif G3-5N1-3G3-5N1-3G3-5N1-3G3-5 define sites of genomic instability (Kruisselbrink et al., 2008); G-rich DNA unable to adopt a G-quadruplex structure remained perfectly stable in DOG-1-deficient worms. An evolutionary conserved role for DOG-1/FANCJ in preventing G4 DNA-induced genome alteration was suggested by the mapping of large genomic deletions accumulating in a FANCJ-deficient human patient cell line to G4 DNA-containing regions (London et al., 2008). In line with a proposed role for this helicase in resolving G-quadruplexes *in vivo* to help the fork replicate the affected strand (Cheung et al., 2002), purified human FANCJ protein was shown to behave as a structure-specific DNA helicase that can unwind G4 DNA with 5′ to 3′ polarity (Wu et al., 2008). While the molecular nature of genomic alterations in

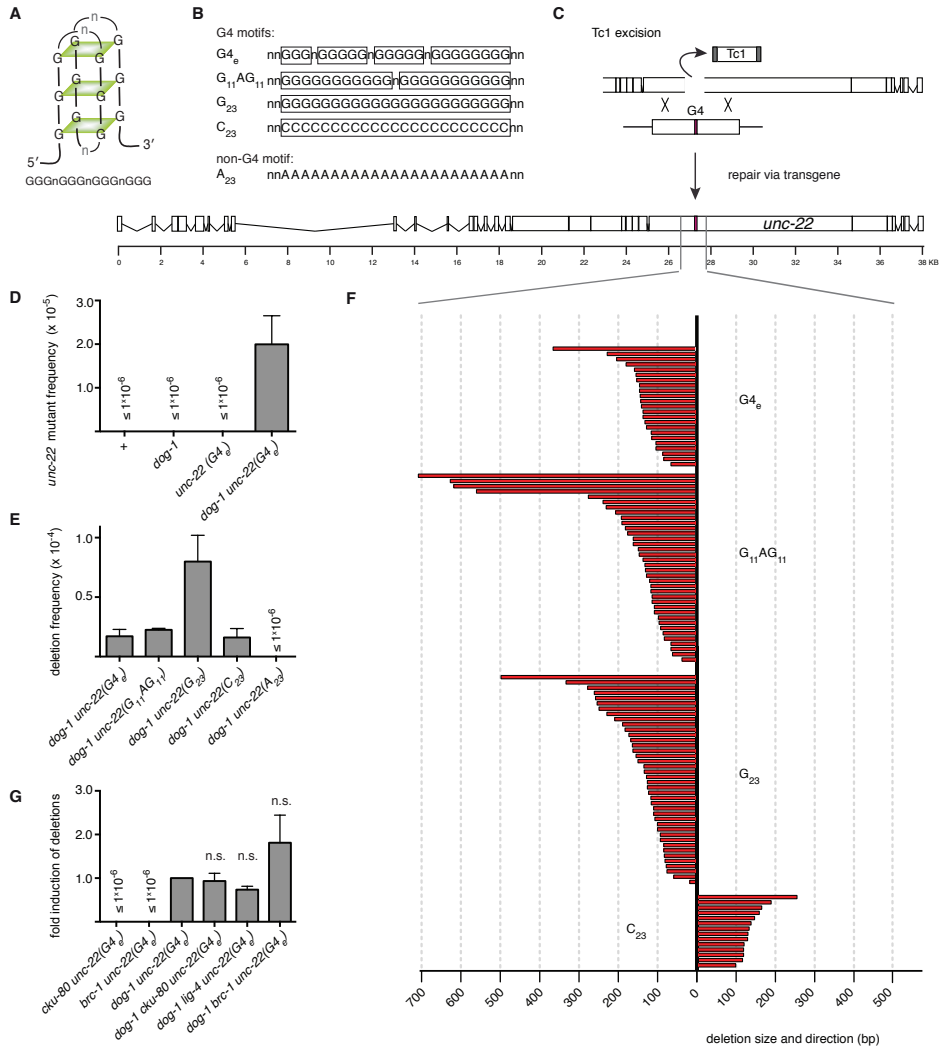


Figure 1 | A novel genomic selection-based assay to measure G4 DNA instability. (A) Schematic representation of a G-quadruplex structure. (B) DNA sequences that were targeted to the endogenous *unc-22* locus. The guanines that can participate in the formation of a G-quadruplex structure are boxed. (C) Schematic representation of the targeting strategy in which a Tc1 transposon-induced double-strand break in the *unc-22* locus is allowed to repair via an extra-chromosomal array carrying ~ 3kb *unc-22* sequences interspersed with a G4 motif. The intervening sequence was designed to result in a functional *unc-22* ORF upon integration. (D) *dog-1*- and G4 DNA-dependent mutation induction for the endogenous *unc-22*. The frequency is based on ±40 independent populations per genotype; the mean of at least two experiments is shown. Error bars indicate s.e.m.. (E) Comparison of G4 DNA-induced deletion frequencies at the *unc-22* locus for different G4 sequence motifs. The frequency is based on ±40 independent populations per genotype; the mean of at least two experiments is shown. Error bars indicate s.e.m.. (F) Graphic illustration of the G4 deletion profiles of different G4 motifs; each bar represents one mutant. (G) G4 DNA-induced deletion frequencies for the indicated genetic backgrounds. The fold induction with respect to *dog-1 unc-22(G4 $_e$)* is represented, and is based on at least two independent experiments. Error bars indicate s.e.m.. No significant (n.s.) difference was found between *dog-1* single and double mutants (paired two tailed t-test).

C. elegans demonstrated a strikingly atypical mutation profile (Cheung et al., 2002; Kruisselbrink et al., 2008), little is known about the molecular mechanisms that underlie G4 DNA-induced genome rearrangements. Here, we identify an alternative DNA double-strand break repair pathway, which requires Polymerase Theta (θ), and which is the pathway of choice to cope with these structural replication fork barriers.

RESULTS

An endogenous target to monitor G4 DNA instability

To investigate the mutagenic mechanism responsible for the generation of deletions at G4 sequences, we first engineered the *C. elegans* genome to create a selectable system for G4 DNA-induced mutations. Via transposon-mediated insertional mutagenesis (Plasterk and Groenen, 1992), we targeted several distinct G4 motifs to the coding region of the endogenous *unc-22* locus, which itself is devoid of G4 motifs. The inserted G4 DNA sequences did not perturb its host reading frame or function (Fig. 1b,c). With its 22 kb ORF and a clearly recognizable mutant phenotype, the *unc-22* locus provides an endogenous mutational sink that is largely non-discriminatory to size: all alterations that change the reading frame (whether one bp or tens of kb) will render worms insensitive to the muscle hyper-contracting effect of the cholinergic agonist levamisole (Brenner, 1974). Earlier work revealed that sequences that match the G4 DNA consensus $G_{3-5}N_{1-3}G_{3-5}N_{1-3}G_{3-5}N_{1-3}G_{3-5}$ induce deletions in *dog-1*-deficient animals; equienergetic G-rich DNA sequences that do not match this motif (*e.g.* G3C DNA or CG repeats) are perfectly stable (Kruisselbrink et al., 2008). However, because of the relatively high, size-related, background level of spontaneous mutagenesis in *unc-22* (~10^{-6}), we chose to target only those G4 sequences that we previously found to have high mutagenic potential (Fig. 1b).

Figure 1d demonstrates that a single G4 motif does not profoundly elevate the spontaneous mutation rate in *unc-22* in wild type animals. However, the mutation rate increases ~15-65 fold when introduced into a *dog-1*-deficient background, with the fold increase depending on the nature of the G-tract (Fig. 1d,e). We found the mutagenicity of the G4 motifs to be dependent on both the G4 sequence and the genomic context: a G_{23} monotract was a more potent inducer of deletions than non-monoG G4 sequences, whereas simply changing the polarity of the G_{23} tract with respect to the direction of the *unc-22* locus dampened its mutagenic potential, indicating that the mutagenic potential is not solely dictated by the nucleotide composition of the G4 motif.

Molecular characteristics of G4 DNA-induced genome alterations

To determine the features of G4 DNA-induced mutagenesis, we systematically characterized large numbers of mutants (Fig.1,2, Supplementary Table 1). These profiles augment and further refine previously determined deletion characteristics

but importantly also lead to novel insight hinting towards an error-prone repair mechanism that generates these deletions.

Firstly, we establish that the previously observed atypical and strikingly narrow deletion size distribution is intrinsic to G4 DNA instability: while the G4 motifs are located in a largely size non-discriminatory locus, 92% of deletions (104 out of 113) are between 50 and 300 bp, with only 9 cases being shorter or longer - the shortest being 14 bp and the longest being 704 bp (Fig. 1f). The nucleotide composition of the G4 motif does not significantly affect deletion size: median sizes being 137, 128, 123 and 128 bp for $G4_e$, $G_{11}AG_{11}$, G_{23} and C_{23}, respectively.

Secondly, we strengthen the suggestion that the premutagenic lesion is in fact a quadruplex fold. While all deletions have their 3′ junction close to the start of the G4 motif, the exact position is greatly influenced by the exact sequence of the motif: whereas the 3′ junctions of deletions induced by the sequence G4e centres around the outermost G of the G4 motif, the 3′ junctions at G23 and G11AG11 are located more internally (Fig. 2a). This increased spread is in line with the notion that the latter motifs are able to adopt many different quadruplex conformations that satisfy the G-quadruplex consensus G3-5N1-3G3-5N1-3G3-5N1-3G3-5. A quadruplex fold comprising only the fifteen 5′ guanines of a G23 monotract would allow replication to progress through eight 3′ guanines before being blocked, ultimately resulting in a junction positioned within the tract. We further tested whether the first blocking guanine of a quadruplex fold determines the position of the 3′ deletion junction by establishing via PCR a 3′ junction profile for a minimal G4 tract, qua739:GGGtGGGaGGGtGGG, which can adopt only one possible three-stacked quadruplex configuration (see Fig. 1a). Because of the low mutagenic capacity of minimal G4 motifs (Kruisselbrink et al., 2008) we obtained qua739 deletions at its original genomic location, by a PCR-based approach (see methods section). Indeed, deletions triggered by this motif have a very sharply positioned 3′ junction, with none mapping within the motif (Fig. 2a).

We then unexpectedly found that also the position of the 5′ junction is not random, and in fact is linked to the nucleotide composition of the 3′ junction. At first glance, the 5′ junctions appeared evenly distributed over a 50-300 bp region upstream of the G4 motif, without any preferential site or sequence (Fig. 1f). However, upon close examination, we observed what could be termed as single-nucleotide homology: ~60-80% of the 81 simple deletions (without inserts) had at least one nucleotide that could be mapped to either junction (Fig. 2b,c), which is profoundly more than the 47% chance if the deletions would be randomly distributed or when compared to a randomly generated set of ~18,000 deletions (P<0.0001, chi-square test; see Supplementary Fig. 1 for details). This overrepresentation is not restricted to deletions induced by G4 motifs at the *unc-22* locus: by whole genome sequencing of three *dog-1* strains that have been clonally grown for 50 generations, we found 59 unique simple deletions mapping to different genomic G4 motifs, 73% of which have homology at their most terminal nucleotide (P<0.0001, chi-square

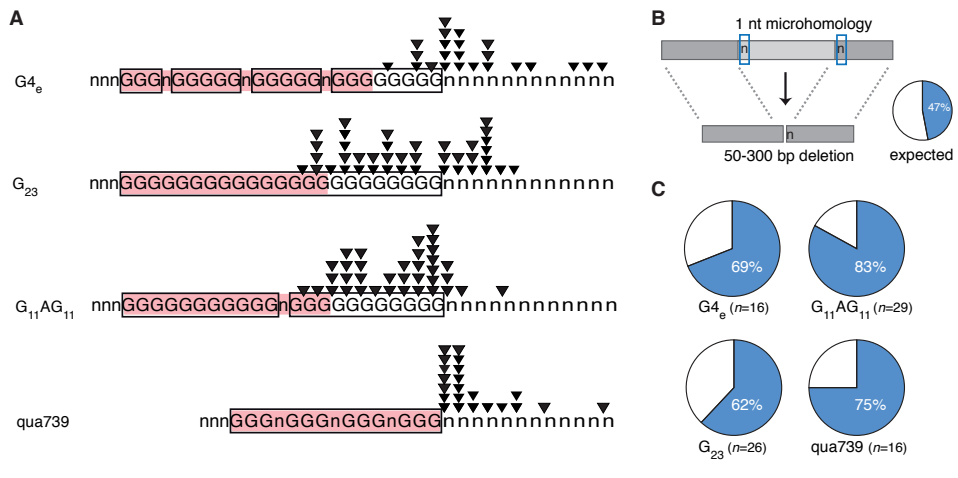


Figure 2 | Molecular characteristics of G4 DNA-induced deletions. (A) Position of the 3′ junctions with respect to the cognate G4 motif (boxed). Triangles indicate junctions of separately isolated deletion alleles. For each motif, the minimal 5′ sequence that complies to the G4 motif consensus is shown in pink, to visually emphasize the notion that almost all junctions (94%) map 3′ of a possible G4 structure. (B) Illustration defining single-nucleotide homology. The chance that the outermost nucleotide on one deletion junction is identical to the first deleted base at the other junctions was calculated as well as empirically determined to be 47% (Supplementary Fig. 1). (C) Pie charts displaying the overrepresentation of single-nucleotide homology for G4 DNA-induced simple deletions (in blue). (D) Templated insertions coinciding with G4 DNA-induced deletion formation. On the left this phenomenon is graphically represented. The right panel displays seven examples. Matching sequences are underlined. Lines 1-4 are simple inserts templated on the 5′ flank. Lines 5 and 6 represent cases where the same flank is used twice as a template. Line 7 represent a case where both the 5′ and 3′ flanks contribute to the insert, but also illustrates that the flank 3′ to the G-tract can serve as a template (see Supplementary Figure 2,4 for more cases).

test, Supplementary Fig. 2 Supplementary Table 2). This phenomenon which may suggest the use of microhomology can, however, be completely attributed to the first nucleotide: while the greatly increased level of homology at the terminal position is highly significant, this is not the case for the neighbouring bases. In fact, in more than 40% of the simple deletions, the homology is restricted to a single nucleotide; the number of observed cases having two bases of microhomology is statistically not different from a random distribution (P= 0.3, chi-square test), taken into account the increased incidence of one nucleotide homology.

This fact, together with the notion that more prominent sequence homology can frequently be found in the immediate vicinity of the 5′ junctions, argues that homology itself is not a driving force in the molecular events leading to deletion formation. Instead, we envisage this single-nucleotide homology to reflect the action of a DNA polymerase acting on and extending a one-base-pair intermediate in an alternative end joining repair mechanism. The observation that genetically inactivating NHEJ or HR affected neither G4 DNA-induced mutation rate nor spectra (Fig. 1g, Supplementary Fig. 3 and reference (Kruisselbrink et al., 2008)) supports this hypothesis.

Another novel characteristic that points towards the involvement of a DNA polymerase is the presence of insertions accompanying 28% of all *unc-22* deletions (32/113). For cases where insertions are larger than four nucleotides, their origin can be traced back to the sequence immediately flanking the junction (Fig. 2d, Supplementary Fig. 2,4, Supplementary Table 1-2). While many inserts map 5′ to the G-tract, we also found cases where newly inserted DNA matched the 3′ flank (Fig. 2d, line 7; Supplementary Fig. 2,4, Supplementary Table 1-2), suggesting that a free 3′ extendable end is available at either side of the G-tract. The latter observation is important because it argues for the existence of a DNA double-strand break (DSB) as an intermediate in the generation of deletions at G4 DNA sites, a conclusion that is strengthened by an increased number of foci of the DSB marker RAD-51 observed in both germ and somatic cells in *dog-1* animals, as compared to wild type animals (Supplementary Fig. 5). Two observations suggest that the 3′ ends of these breaks are fairly stable and refractory to trimming: firstly, templated inserts were typically found very close to their cognate template, suggesting that newly made DNA was joined to the DNA that served as its template, and secondly, we found cases in which the same flanking sequence was used as a template twice consecutively (Fig. 2d, lines 5-7; Supplementary Fig. 2), indicative of iterative cycles of extension and re-priming.

The typical G4 DNA-induced deletions require polymerase θ

After having established the signature of polymerase activity in G4 DNA-induced deletions, we set out to identify the responsible DNA polymerase via a candidate approach. We assayed mutants of the translesion synthesis polymerases η and κ, as these Y-family polymerases have the ability to replicate through damaged DNA and have previously been suggested to play a role in G4 DNA-mediated deletion induction (Youds et al., 2006). We found, however, no evidence for their involvement using multiple approaches and combinations of alleles (Table 1, Supplementary Fig. 6a,b).

Another candidate is polymerase θ, an A-family DNA polymerase implicated in the repair of interstrand crosslinks (Harris et al., 1996; Muzzini et al., 2008; Shima et al., 2004). Mammalian Pol θ has the ability to bypass DNA lesions and to extend matched as well as mismatched primer termini (Hogg et al., 2012; Seki et al., 2004). In addition, Drosophila Pol θ was recently implicated in the repair of DNA double-strand breaks induced by DNA transposition (Chan et al., 2010). To test a possible involvement of the *C. elegans* homolog of Pol θ, POLQ-1, we first used a PCR-based assay that detects

Table 1 | Quantification G4 DNA-induced deletions using PCR-assays.

Genotype	No. of animals assayed	No. of deletions	Percentage	Significance* (P<0.01)
Qua830 (GGGGGGGGGGGGGGGGaGGtGGG)				
polh-1(lf31)	95	0	0	-
polk-1(lf29)	96	0	0	-
polq-1(tm2572)	94	0	0	-
polq-1(tm2606)	91	0	0	-
dog-1(pk2247)	94	13	13.8	-
dog-1(gk10)	94	13	13.8	-
dog-1(pk2247) polh-1(lf31)	95	16	16.8	no
dog-1(gk10) polh-1(ok3317)	96	13	13.5	no
dog-1(gk10) polk-1(lf29)	90	15	16.7	no
dog-1(pk2247) polq-1(tm2572)	94	0	0	yes
dog-1(gk10) polq-1(tm2026)	93	0	0	yes
Qua1894 (GGGGGGGGGGGGGGGGGGGGGG)				
polh-1(lf31)	80	0	0	-
polk-1(lf29)	96	0	0	-
polq-1(tm2572)	96	0	0	-
polq-1(tm2606)	96	1	1.0	-
dog-1(pk2247)	96	13	13.5	-
dog-1(gk10)	96	8	8.3	-
dog-1(pk2247) polh-1(lf31)	96	19	19.8	no
dog-1(gk10) polk-1(lf29)	96	5	5.2	no
dog-1(pk2247) polq-1(tm2572)	96	1	1.0	yes
dog-1(gk10) polq-1(tm2026)	96	1	1.0	yes

PCR, polymerase chain reaction
*Significance was calculated by two-sided Z-proportion test. Double mutants were compared to the single *dog-1* mutant of the corresponding allele.

G4 DNA-induced deletions in isolated DNA by preferential amplification of smaller than wild type bands of sequences containing G4 motifs (Kruisselbrink et al., 2008). We found that while 14% of *dog-1* animals sustained at least one small-sized deletion at the G4 motif qua830, none were detected in animals that in addition were also deficient for *polq-1* (Fig. 3a). The same outcome was found for different G4 DNA-containing loci, and when different alleles of *dog-1* and *polq-1* were tested (Table 1, Supplementary Fig. 6c), together indicating that deletion induction at endogenous G4 motifs is completely dependent on functional POL θ.

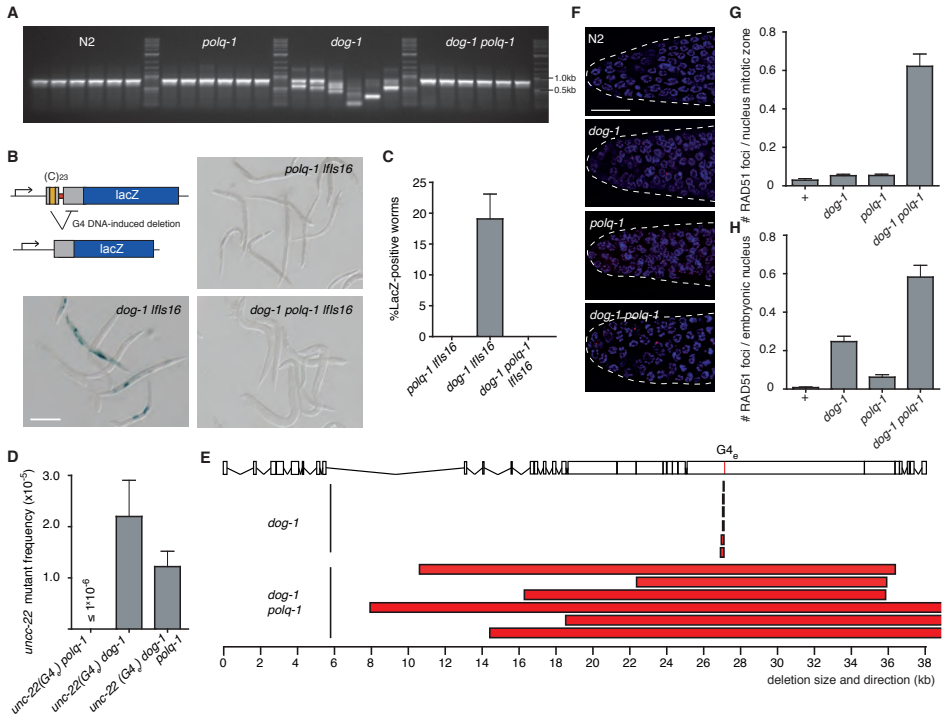


Figure 3 | Polymerase Theta-mediated end joining G4 DNA-induced DNA breaks. (A) PCR-based assay to measure G4 DNA-induced deletion formation. Animals of the indicated genotype (five per lane) are lysed and PCR amplified with primers flanking the endogenous G4 motif, qua830. Somatic deletions will manifest as shorter than wild type product (which is present in great excess). (B) Reporter-based assay to measure G4 DNA-induced deletion formation using LacZ expression as read out. The panels display 5-10 animals stained for LacZ and their indicated genotypes. See Methods section for detailed description of reporter and assay. Scale bar indicates 0.25mm. (C) Quantification of reporter LacZ expression by scoring animals (n>200 per experiment) of the indicated genotype for the presence of ≥1 blue cell. The average percentage of at least 4 independent experiments is shown. Error bars indicate s.e.m.. (D) Mutation frequency at the genomic *unc-22* (G4e) allele for the indicated genotypes. The frequency is based on ±50 independent populations per genotype; the mean of at least two experiments is shown. Error bars indicate s.e.m.. (E) Graphical representation of the *unc-22* mutations isolated from the indicated genotype. Bars represent the size and location of independently derived *unc-22* deletions. (F) Immunohistochemical analysis of proliferative germ cells in the mitotic zone of the *C. elegans* gonad. RAD-51 foci in red; DAPI in blue. Scale bar indicates 15µm. (G-H), Quantification of RAD-51 foci in the proliferative pre-meiotic germline, n≥14 germlines per genotype (G) and in developing embryos, n≥18 embryos per genotype (H). (A-H), N2, *dog-1(gk10)* and *polq-1(tm2026)* alleles were used; error bars indicate s.e.m..

We next visualized G4 DNA instability directly in animals using transgenes that express LacZ when a G4 DNA-induced deletion brings the reporter ORF in frame with the upstream ATG start codon (Fig. 3b). Twenty percent of *dog-1*-deficient animals stochastically express LacZ in various cell types with patterns indicative of G4 DNA-induced deletion events occurring at different stages of embryonic development (Fig. 3b). In contrast, no LacZ-expressing cells were observed in *dog-1 polq-1*-mutant animals, again indicating that the generation of the characteristic asymmetrical 50-300 bp deletions at G4 motifs is completely dependent on POL θ functionality (Fig. 3b,c). Apart from null alleles of *polq-1*, we also tested an allele (generated via random mutagenesis in the *C. elegans* million mutation project (Thompson et al., 2013)) that has a mutation in its polymerase domain. A change of an evolutionarily highly conserved proline residue at position 1417 into a serine (P1417S) led to a 50% reduction in the number of deletions at endogenous and transgenic G4 sites, which supports a direct role for the polymerase function of POLQ-1 in the generation of G4 DNA-induced deletions (Supplementary Fig. 7).

To address the fate of G4 DNA-induced breaks in the absence of functional POL θ, we assayed mutation induction at G4e within the *unc-22* locus. Remarkably, the G4e-related mutation frequency was only slightly reduced in *dog-1 polq-1* animals as compared to *dog-1* animals (Fig. 3d) indicating that POLQ-1 loss does not affect the mutagenicity of G4 sequences *per se*. The *unc-22* mutants isolated in this genetic background are, however, of a completely different nature: deletions were still observed but were all >10 kb and bidirectional with respect to the G4 motif (Fig. 3e). This outcome likely goes together with substantial DNA end resection, as we also observed a profound increase of RAD-51 foci in *dog-1 polq-1* double mutant animals as compared to either single mutant (Fig. 3f-h, Supplementary Fig. 8).

G4 DNA instability in natural isolates

Thus far, G-quadruplex-induced DNA rearrangements have been observed only in *dog-1*-deficient backgrounds; we have not identified events in wild type animals using five different assays (*i.e.* aCGH, whole genome sequencing, PCR on endogenous loci, *unc-22::G4* mutation induction and G4 DNA-reporter transgenes). Given the threshold levels of detection for these assays, we estimated that G4 sequences are at least 1000-fold more stable in wild type than in *dog-1*-deficient animals.

We hypothesized that G4 DNA instability may be apparent when analysed on an evolutionary scale and indeed found, by comparative genome analysis of natural isolates of *C. elegans* (Fig. 4a), that G4 DNA-induced deletions occur also during normal growth in genetically non-compromised animals. The Hawaiian strain CB4856 suffered fourteen deletions that contained one of the ~1700 G4 motifs present in the genome of the Bristol N2 strain, from which it is estimated to be ~600,000 generations separated (Supplementary Table 3 and Methods). In the majority of these cases, the G4 motif is located within few bases of the deletion's 3′ junction (Supplementary Table 3), a distribution that is highly non-random (Fig. 4b). We found another twelve such cases in the genomes of three other natural isolates (CB4857, RC301 and AB2) that are less diverged from Bristol N2 and that live in similar habitats (Supplementary Table 3). Also here, the non-symmetrical deletion fingerprint previously established for *dog-1*-deficient animals is apparent (Fig. 4c), arguing that G4 DNA-induced deletions in *dog-1* and wild type animals are generated via the same error-prone mechanism.

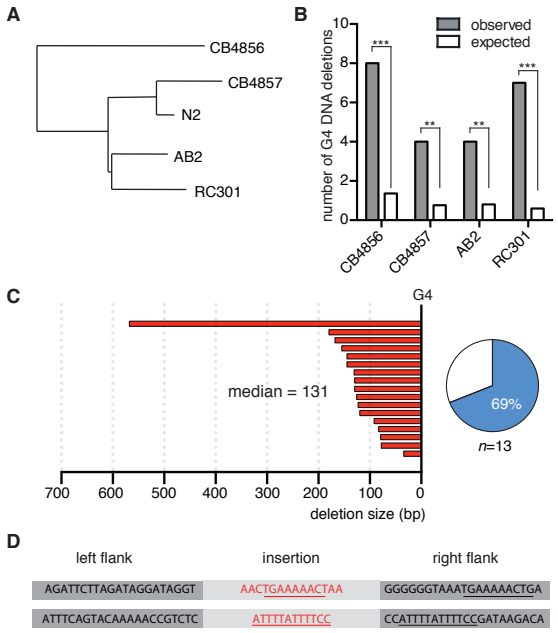


Figure 4 | G4 DNA instability during *C. elegans* evolution. (A) A phylogenetic tree of the *C. elegans* natural isolates used in this study, with (B) the observed versus expected number of G4 DNA-induced deletions in their genomes. Bristol N2 was used as a reference genome (see Methods for further details) *** $P<1x10^{-5}$, ** $P<1x10^{-2}$, binomial test. (C) Graphic representation (analogous to Fig. 1f) of the sizes of the G4 deletions found in *C. elegans* natural isolates, the G4 motif is set at 0 (bp). The pie chart displays the overrepresentation of single-nucleotide homology for G4 DNA-induced simple deletions (in blue). (D) Templated insertions coinciding with G4 DNA-induced deletion formation found in *C. elegans* natural isolates. Matching sequences are underlined. In both cases the G4 motif was located near the right junction, further strengthening the notion that also the flank 3′ to the G-tract can serve as a template.

Moreover, these deletions are characterized by single nucleotide homology (69%) and templated insertions (2 out of 17, Fig. 4d), strongly suggesting that they have been generated via polymerase θ-mediated end joining.

DISCUSSION

Due to their ability to obstruct replication fork movement, G-quadruplex structures have recently emerged as DNA sequences at risk for replication fork arrest (Lopes et al., 2011; Paeschke et al., 2011) and spontaneous chromosomal rearrangements (De and Michor, 2011). Whether these structures pose a block to lagging strand synthesis (Cheung et al., 2002) or leading strand synthesis, as was suggested by data derived in yeast (Lopes et al., 2011), is unknown, and there are no indications in our experiments hinting towards one or the other. We can nevertheless rule out a requirement for ongoing transcription across a G4 locus for it to become mutagenic,

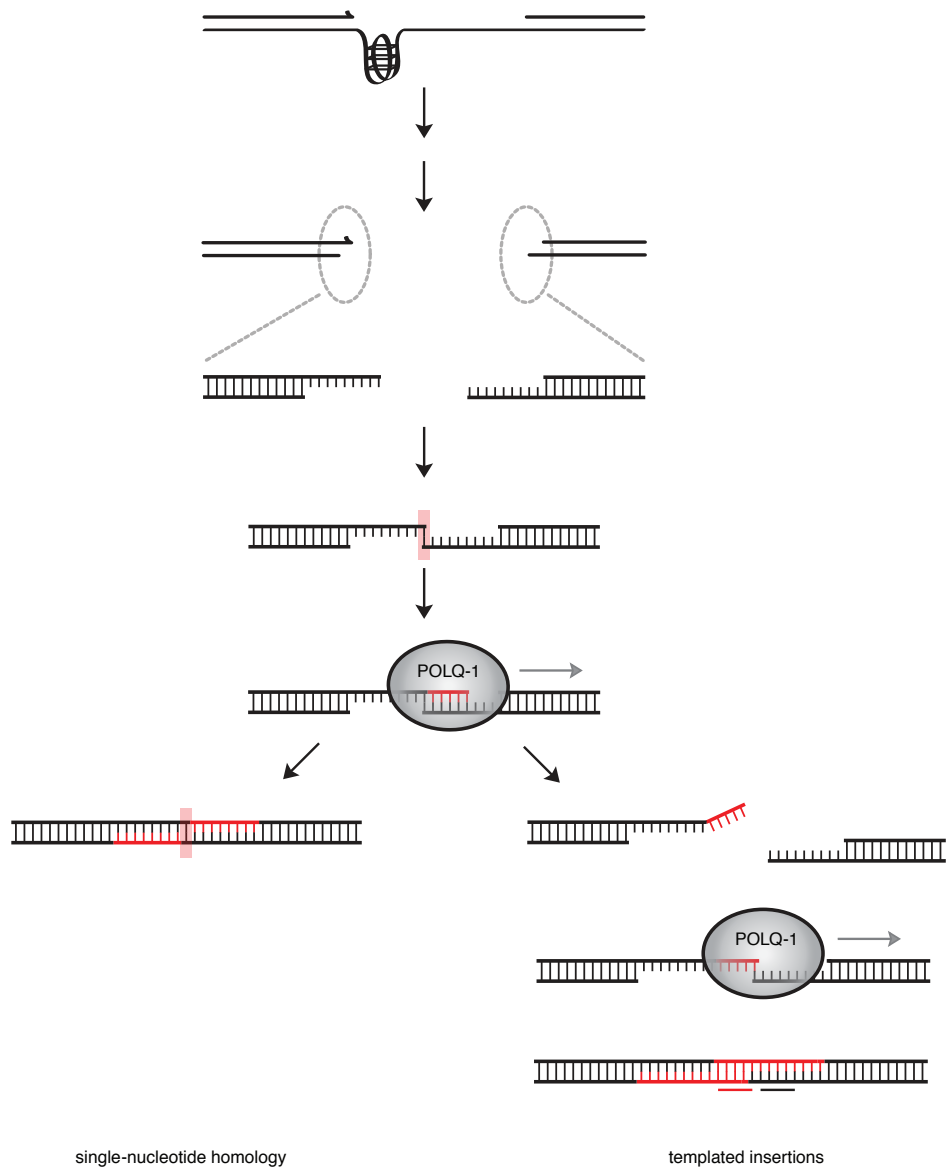


Figure 5 | Tentative model for polymerase Theta-mediated end joining. A replication fork block at a G4 structure results in a DSB (see also Supplementary Fig. 9). The broken ends are joined by TMEJ resulting in two types of deletions: firstly, simple ones characterized by single-nucleotide homology, and secondly, deletions with associated templated insertions. The generation of the latter class is simply an iteration of the steps leading to the simple deletions, with one exception, being the dissociation of both ends after initial Pol Theta-mediated templated DNA synthesis.

as was recently shown to be the case for G4-induced antigenic variation in *N. gonorrhoeae (Cahoon and Seifert, 2012)*: some of the G4-containing loci we studied, including the one in *unc-22,* are not transcribed in germ cells (Wang et al., 2009); still they give rise to inheritable genome alterations. Of interest, the inclusion of highly mutagenic G4 sequences, in either orientation, into the coding strand of *unc-22* did not affect its expression, which argues that single G4 motifs are not strong modifiers of transcription or translation.

Here, we show that in *C. elegans*, a POL θ-dependent pathway acts to prevent extensive loss of sequences near G4 DNA sites but at the expense of generating small deletions. These deletions are typified by a limited size distribution, single-nucleotide homology and the occasional inclusion of templated insertions, and can be recognized in genetic backgrounds that are compromised for their ability to resolve G-quadruplexes through dedicated helicase actions (*i.e. dog-1*), as well as in wild type strains isolated from geographically different regions of the globe. The deletion spectra provide hints as to how POL θ acts to generate deletions of 50-300 bp. The notion of single-nucleotide homology may reflect an extension reaction primed on a one-base-pair intermediate. To us, this interpretation is more plausible than to assume that annealing of a single base provides sufficient stability to seal the breaks by other means. Instead, we propose that POL θ extends this single base pair intermediate, and as such is the creator of extended homology, that is subsequently sufficient to guide break repair. The occasional presence of templated insertions associated with deletion formation may testify to this scenario by reflecting repair false starts, in which an initial primed extension reaction (leading to the incorporation of a number of templated nucleotides) is abrogated, and subsequently restarted by re-annealing and extension, but now at a new position that is defined by the 3′ end of the templated insert (Fig. 5).

From the nature of the deletion junctions, as well as from the analysis of RAD-51 foci, we infer that POL θ acts to connect DNA ends of DSBs that arise at replication fork barriers. Direct demonstration of these DSBs via molecular means has thus far been unsuccessful, probably because of the very low rates of deletion formation (10^{-5} per animal generation). The notion that templated inserts are derived from sequences located both upstream as well as downstream of the G4 argues that extendable 3′ hydroxyl ends are available on either side of the G4, hence a DSB. This notion also disfavours a model in which POL θ would act as a G4 translesion synthesis polymerase by facilitating the nascent G4-blocked strand to jump ahead 50-300 nucleotides, as this would predict that only the sequence ahead of the G4 can act as a template.

While current and previous work showed that G4 DNA-induced deletion formation in *C. elegans* is independent of canonical NHEJ (Fig. 1g, Supplementary Fig. 3 and references (Kruisselbrink et al., 2008; Youds et al., 2006)), the outcome is similar: it safeguards genetic integrity by minimizing the loss of genetic information at break sites, via joining the ends. We thus propose to term this alternative end joining pathway, TMEJ, for polymerase Theta-Mediated End Joining.

Why TMEJ generates deletions specifically in the range 50-300 bp is not known, but may reflect the asymmetric generation of stable free 3′ hydroxyl ends around G-quadruplexes that subsequently serve as substrates for POL θ activity. The nascent strand blocked at the G4 structure provides one obvious point of entry, but the origin of the more distal end is less clear. Notably, the size distribution of G4 DNA-induced deletions correlates to predicted distances between replication blocks and upstream Okazaki fragments (Smith and Whitehouse, 2012), thus to ssDNA gaps containing a replication fork barrier. Next-round replication of such gaps in rapidly dividing tissues would predict the formation of a DSB with 3′ hydoxyl ends 50-300 bp apart, across from a sister chromatid that still contains the obstructing lesion (Supplementary Fig. 9). Evoking such a scenario for DSB formation may also explain why we thus far failed to find any (structure specific) nuclease to be required for G4 DNA-induced deletion formation (references (Kruisselbrink et al., 2008; Youds et al., 2006); data not shown). The unavailability of the sister chromatid as the preferred repair donor could explain why homologous recombination cannot operate to repair these replication-associated DNA breaks, by that very fact providing the biological raison d'être of TMEJ.

Whether TMEJ acts to prevent genomic catastrophe at replication-blocking structures (or lesions) in mammalian systems is an outstanding question that future work needs to address. The outcome of the dynamic interplay between available repair systems can be context and species specific, and the notion that expression of mammalian POLQ expression appears to be highest in testis and human placental tissue (Seki et al., 2003) may suggests cell lineage specific use of a POLQ pathway, perhaps favouring error-prone repair over cell death or arrest in situations where rapid cycles of proliferation is critical, like in the *C. elegans* embryo (Holway et al., 2006).

We anticipate that ectopic activation of TMEJ may be of high clinical significance, also considering the recent finding that Polymerase Theta up-regulation is associated with poor survival in cancer (Higgins et al., 2010a; Lemée et al., 2010): if not properly controlled, POL θ's ability to tie DNA ends together can have very undesirable effects, as it provides cells with the means to proliferate in the presence of increased replication stress (Higgins et al., 2010b). Blocking this activity may thus constitute a potent strategy towards preventing cancerous growth.

METHODS

Strains and culturing. Nematodes were cultured according to standard protocols (Brenner, 1974). The following alleles were used in this study: *unc-22(lf39)* [G23], *unc-22(lf72)* [G4e], *unc-22(lf73)* [G11AG11], *unc-22(lf95)* [A23], *unc-22(lf96)* [C23], *dog-1(pk2247)*, *dog-1(gk10)*, *rde-2(pk1657)*, *polk-1(lf29)*, *polh-1(lf31)*, *polh-1(ok3317)*, *cku-80(ok861)*, *lig-4(ok716)*, *brc-1(tm1145)*, *polq-1(tm2026)*, *polq-1(tm2572)*, *polq-1(gk765752)*, *lfIs16[hsp::ATG-C_{23}-stops-GFP-LacZ(prp3019);rol-6^D(su1006)]*, *lfIs055 [myo-2::C_{23}-stops-GFP-LacZ(pLM20); rol-6^D(su1006)]*, *lfIs177[myo-2::ATG-C23-stops-NLS-GFP-LacZ(pLM88);*

pGH8;pCFJ104;rol-6^{D}(su1006)]. Alleles were generated in our laboratory or kindly provided by the *C. elegans* Genetics Center and the laboratory of Dr. Shohei Mitani.

Transposon-mediated insertional mutagenesis. Cloning details and plasmid sequences are available upon request. In brief, targeting vectors contained G4 sequences flanked by ~ 1-2 kb of sequence identical to the *unc-22* genomic sequence flanking *unc-22(st136::Tc1)*. Plasmids were co-injected with marker plasmid pRF4 into N2 according to standard protocols. Extrachromosomal arrays were crossed into a mutator background (*rde-2*) also carrying *unc-22(st0136::Tc1)*. Populations were screened for *unc-22* revertants that were subsequently molecularly analysed. Animals that had reverted because of successful targeting of the G4 sequence to the endogenous *unc-22* gene were out-crossed to N2.

***Unc-22* forward mutation frequency assay.** The *unc-22* forward mutation frequency was determined as described previously (Mori et al., 1988). In brief, for each data point, single L4-stage worms were transferred to 20-100 9 cm plates and populations were grown until the F2-F3 generation. A sub-fraction of each population was inspected for *unc-22* mutants that are resistant to the paralysing effect of (2mM) levamisole. Independently derived mutants were molecularly analysed by PCR and sequencing.

Deletions at endogenous G4 DNA loci. We assayed endogenous G4 DNA loci using a PCR-based approach that was described previously (Gengyo-Ando and Mitani, 2000; Pontier et al., 2009). In brief: genomic DNA was isolated either from single worms or pools of worms and subjected to nested rounds of PCRs with primers that flank a G4 motif; amplicons are typically 1 kb in size. Smaller than wildtype bands (deletions) are preferentially amplified a) because they have the intrinsic bias to amplify small over large DNA segments and b) because the G4 motif in non-deletion carrying fragments also hamper DNA replication in vitro (Pontier et al., 2009). To determine the frequency of deletions, L4 stage worms were used (1 animal in a 15ul lysis reaction of which 1ul was used in a PCR). The following primers were used: qua830 5′-ctagttcagggtatctggac-3′; 5′-gattgcgggcactttacctcg-3′; 5′-ccttctctcgaagcgcgacc-3′; 5′-gattttattgactctccgtccg-3′. qua1894 5′-attgtgggaaaaatccgacg-3′; 5′-tttgccatcaaggttccagac-3′; 5′-gtataagagttcctggtcggc-3′; 5′-ggatttcacagcgtcaagag-3′; qua739 5′-aacggacaattatgagctacgc-3′; 5′-gataagagaaacgcaaattacgg-3′; 5′-ccttggcttggatttcttcg-3’; 5′- aaggcgcacagattttaagc-3′.

LacZ-based transgenic reporter assay. Transgenic animals were generated that carry a multicopy array of the reporter construct pRP3019 (*lfIs16*). pRP3019 was constructed using the backbone of pRP1821 (Tijsterman et al., 2002). As illustrated in Fig. 3b a G4 motif (yellow) was placed, in reversed orientation, downstream of the start codon of a heatshock-driven LacZ reporter. Immediately downstream of the tract, stop codons (red) were introduced followed by a non-selective ORF (grey) that is in frame with the LacZ ORF (blue) and functions as a deletion buffer. This reporter will only express LacZ when the stop codons are deleted *in vivo*, *e.g.* via a typical G4 DNA-induced deletion, and the downstream ORF is brought into frame with the upstream ATG. To read-out G4 DNA instability, *lfIs16* animals were synchronized by bleaching and overnight hatching and then transferred to new plates and heat-shocked (34°C for 2 times 2 hours with 30 minutes recovering time in between) when animals reached the L3-L4 stage. LacZ expression was visualized with X-gal staining protocol (Tijsterman et al., 2002). Experiments were performed at least in triplo for each genotype; each experiment consisting of ≥ 4 independent populations having more than 100 worms. G4 DNA reporters pLM20 and pLM88, used to generate alleles *lfIs055* and *lfIs177*, respectively, deviate slightly from pRP3019 but have the same principle and outcomes. Maps for these reporters are available upon request. Injection marker plasmids pGH8 and pCFJ104 (used for *lfIs177*) are described in reference (Frøkjær-Jensen et al., 2008).

Immunostainings and RAD-51 foci quantification. Germlines were dissected from young adults (1 day post-L4 stage) and processed for immunostaining (Roerink et al., 2012). Samples were incubated overnight at room temperature with primary anti-RAD-51 antibodies (Novus Biologicals #29480002) diluted 1:200 in PBSTB (PBS with 0.1% Tween 20 and 1% BSA), followed by antibody AlexaFluor 488 (Invitrogen #A11008, diluted 1:1000 in PBSTB); DNA was stained with 0.5ug/ml DAPI. Samples were mounted with Vectashield. Microscopy was performed with a Leica DM6000 microscope.

Whole genome sequencing of *dog-1* mutation accumulation lines. *dog-1(pk2247)* animals were substantially out-crossed to wildtype N2 (Bristol) and mutation accumulation (MA) lines were generated by cloning out F1 animals from one hermaphrodite. Each generation, three worms were transferred to new plates and MA lines were maintained for 50 generations. A single animal was then cloned out and propagated to obtain a full plate for DNA isolation. Worms were rinsed off with M9 and incubated for one hour at room temperature while shaking. 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. The genomes of 3 independently grown *dog-1(pk2247)* MA-strains were sequenced.

Bioinformatic analysis was performed as follows: paired-end whole genome sequence data of *dog-1(pk2247)* MA-lines were mapped to the reference genome (Wormbase release 225) using bwa with normal settings. Sorted BAM files were created by Samtools and subsequently analyzed using Pindel (Ye et al., 2009). A deletion was considered only if it was uniquely seen in one of the three sequenced strains and covered at least five times, thereby excluding events that were already present in the starting strain. Raw sequences have been made publicly available at NCBI SRA (Accession code SRP032440).

G4 DNA deletions in natural isolates. Paired-end whole genome sequence data were downloaded from the NCBI Short Read Archive (SRP011413), and sequence reads were mapped to the *C. elegans* reference genome (Wormbase release 225). The average base coverage was 176x, 164x, 166x and 75x for AB2, CB4857, RC301 and CB4856, respectively. Pindel (Ye et al., 2009) was used to detect structural variations (SVs) in the natural isolates as compared to the N2 reference genome. We included only SVs that had at least 10 reads supporting the SV and no reads supporting the reference genome. We used the samtools mpileup command to include only those events that showed a coverage drop for the sequence within the structural variation (average deletion coverage < 5x and surrounding flanks (100bp) coverage > 10x). As a third criterion, we collected only those SVs that were N2-like in one of the other 3 natural isolates. We found 1626, 913, 962 and 714 deletions of at least 10bp for CB4856, CB4857, AB2 and RC301, respectively. In this collection of SVs, we searched for deletions that contained a G4 motif. We tested 7 cases with Sanger sequencing; all of these confirmed the Pindel junction prediction.

Statistical Analysis. We determined the statistical significance of elevated G4 DNA deletion induction in the natural isolates as follows: we first determined the probability of a deletion junction to be within 50 nucleotides of a G4 motif, as 100% of G4 DNA-induced deletions in *dog-1* comply to this definition. This probability is $8.4x10^{-4}$: 50*1680 (the number of G4 motifs in the genome)/$1x10^{8}$ (the size of the genome). The expected number of G4 DNA-induced deletions, as displayed in Figure 4, is then set to n*$8.4x10^{-4}$, where n is the total number of deletions found in a strain. We then used binomial distributions to calculate the p-values for the observed number of G4 DNA-induced deletions, given the probability of $8.4x10^{-4}$ and the total number of deletions per strain: 1626, 913, 962 and 714 deletions in CB4856, CB4857, AB2 and RC301, respectively. For other experiments, statistical significance was determined with a two-tailed unpaired Student's *t*-test, unless otherwise stated.

REFERENCES

Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J.-M., and Lemaitre, J.-M. (2012). Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nature Structural & Molecular Biology *19*, 837–844.

Brenner, S. (1974). The genetics of Caenorhabditis elegans. Genetics *77*, 71–94.

Cahoon, L.A., and Seifert, H.S. (2009). An alternative DNA structure is necessary for pilin antigenic variation in Neisseria gonorrhoeae. Science *325*, 764–767.

Cahoon, L.A.L., and Seifert, H.S.H. (2012). Transcription of a cis-acting, noncoding, small RNA is required for pilin antigenic variation in Neisseria gonorrhoeae. PLoS Pathog *9*, e1003074–e1003074.

Chan, S.H., Yu, A.M., and McVey, M. (2010). Dual roles for DNA polymerase theta in alternative end-joining repair of double-strand breaks in Drosophila. PLoS Genet. *6*, e1001005.

Cheung, I., Schertzer, M., Rose, A., and Lansdorp, P.M. (2002). Disruption of dog-1 in Caenorhabditis elegans triggers deletions upstream of guanine-rich DNA. Nat. Genet. *31*, 405–409.

De, S., and Michor, F. (2011). DNA secondary structures and epigenetic determinants of cancer genome evolution. Nature Structural & Molecular Biology *18*, 950–955.

Frøkjær-Jensen, C., Davis, M.W., Hopkins, C.E., Newman, B.J., Thummel, J.M., Olesen, S.-P., Grunnet, M., and Jorgensen, E.M. (2008). Single-copy insertion of transgenes in Caenorhabditis elegans. Nat. Genet. *40*, 1375–1383.

Gellert, M., Lipsett, M.N., and Davies, R.D. (1962). Helix formation by guanylic acid. Proc. Natl. Acad. Sci. U.S.a. *48*, 2013–2018.

Gengyo-Ando, K., and Mitani, S. (2000). Characterization of mutations induced by ethyl methanesulfonate, UV, and trimethylpsoralen in the nematode Caenorhabditis elegans. Biochem. Biophys. Res. Commun. *269*, 64–69.

Harris, P.V., Mazina, O.M., Leonhardt, E.A., Case, R.B., Boyd, J.B., and Burtis, K.C. (1996). Molecular cloning of Drosophila mus308, a gene involved in DNA cross-link repair with homology to prokaryotic DNA polymerase I genes. Mol. Cell. Biol. *16*, 5764–5771.

Higgins, G.S., Harris, A.L., Prevo, R., Helleday, T., McKenna, W.G., and Buffa, F.M. (2010a). Overexpression of POLQ confers a poor prognosis in early breast cancer patients. Oncotarget *1*, 175–184.

Higgins, G.S., Prevo, R., Lee, Y.-F., Helleday, T., Muschel, R.J., Taylor, S., Yoshimura, M., Hickson, I.D., Bernhard, E.J., and McKenna, W.G. (2010b). A small interfering RNA screen of genes involved in DNA repair identifies tumor-specific radiosensitization by POLQ knockdown. Cancer Res. *70*, 2984–2993.

Hogg, M., Sauer-Eriksson, A.E., and Johansson, E. (2012). Promiscuous DNA synthesis by human DNA polymerase θ. Nucleic Acids Res. *40*, 2611–2622.

Holway, A.H., Kim, S.-H., La Volpe, A., and Michael, W.M. (2006). Checkpoint silencing during the DNA damage response in Caenorhabditis elegans embryos. J. Cell Biol. *172*, 999–1008.

Huppert, J.L., and Balasubramanian, S. (2005). Prevalence of quadruplexes in the human genome. Nucleic Acids Res. *33*, 2908–2916.

Kruisselbrink, E., Guryev, V., Brouwer, K., Pontier, D.B., Cuppen, E., and Tijsterman, M. (2008). Mutagenic capacity of endogenous G4 DNA underlies genome instability in FANCJ-defective C. elegans. Curr. Biol. *18*, 900–905.

Law, M.J., Lower, K.M., Voon, H.P.J., Hughes, J.R., Garrick, D., Viprakasit, V., Mitson, M., De Gobbi, M., Marra, M., Morris, A., et al. (2010). ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner. Cell *143*, 367–378.

Lemée, F., Bergoglio, V., Fernandez-Vidal, A., Machado-Silva, A., Pillaire, M.-J., Bieth, A., Gentil, C., Baker, L., Martin, A.-L., Leduc, C., et al. (2010). DNA polymerase theta up-regulation is associated with poor survival in breast cancer, perturbs DNA replication, and promotes genetic instability. Proc. Natl. Acad. Sci. U.S.a. *107*, 13390–13395.

London, T.B.C., Barber, L.J., Mosedale, G., Kelly, G.P., Balasubramanian, S., Hickson, I.D., Boulton, S.J., and Hiom, K. (2008). FANCJ is a structure-specific DNA helicase associated with the maintenance of genomic G/C tracts. J. Biol. Chem. *283*, 36132–36139.

Lopes, J., Piazza, A., Bermejo, R., Kriegsman, B., Colosio, A., Teulade-Fichou, M.-P., Foiani, M., and Nicolas, A. (2011). G-quadruplex-induced instability during leading-strand replication. Embo J *30*, 4033–4046.

Mori, I., Moerman, D.G., and Waterston, R.H. (1988). Analysis of a mutator activity necessary for germline transposition and excision of Tc1 transposable elements in Caenorhabditis elegans. Genetics *120*, 397–407.

Muzzini, D.M., Plevani, P., Boulton, S.J., Cassata, G., and 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.

Paeschke, K., Capra, J.A., and Zakian, V.A. (2011). DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. Cell *145*, 678–691.

Plasterk, R.H., and Groenen, J.T. (1992). Targeted alterations of the Caenorhabditis elegans genome by transgene instructed DNA double strand break repair following Tc1 excision. The EMBO Journal *11*, 287–290.

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

Ribeyre, C., Lopes, J., Boulé, J.-B., Piazza, A., Guédin, A., Zakian, V.A., Mergny, J.-L., and Nicolas, A. (2009). The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. PLoS Genet. *5*, e1000475–e1000475.

Roerink, S.F., Koole, W., Stapel, L.C., Romeijn, R.J., and 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.

Sarkies, P., Murat, P., Phillips, L.G., Patel, K.J., Balasubramanian, S., and Sale, J.E. (2012). FANCJ coordinates two pathways that maintain epigenetic stability at G-quadruplex DNA. Nucleic Acids Res. *40*, 1485–1498.

Sarkies, P., Reams, C., Simpson, L.J., and Sale, J.E. (2010). Epigenetic instability due to defective replication of structured DNA. Molecular Cell *40*, 703–713.

Schwab, R.A., Nieminuszczy, J., Shin-ya, K., and Niedzwiedz, W. (2013). FANCJ couples replication past natural fork barriers with maintenance of chromatin structure. J. Cell Biol. *201*, 33–48.

Seki, M., Marini, F., and Wood, R.D. (2003). POLQ (Pol θ), a DNA polymerase and DNA-dependent ATPase in human cells. Nucleic Acids Res. *31*, 6117–6126.

Seki, M., Masutani, C., Yang, L.W., Schuffert, A., Iwai, S., Bahar, I., and Wood, R.D. (2004). High-efficiency bypass of DNA damage by human DNA polymerase Q. Embo J *23*, 4484–4494.

Shima, N., Munroe, R.J., and Schimenti, J.C. (2004). The mouse genomic instability mutation chaos1 is an allele of Polq that exhibits genetic interaction with Atm. Mol. Cell. Biol. *24*, 10381–10389.

Smith, D.J., and Whitehouse, I. (2012). Intrinsic coupling of lagging-strand synthesis to chromatin assembly. Nature *483*, 434–438.

Smith, J.S., Chen, Q., Yatsunyk, L.A., Nicoludis, J.M., Garcia, M.S., Kranaster, R., Balasubramanian, S., Monchaud, D., Teulade-Fichou, M.-P., Abramowitz, L., et al. (2011). Rudimentary G-quadruplex-based telomere capping in Saccharomyces cerevisiae. Nature Structural & Molecular Biology *18*, 478–485.

Thompson, O., Edgley, M., Strasbourger, P., Flibotte, S., Ewing, B., Adair, R., Au, V., Chaudhry, I., Fernando, L., Hutter, H., et al. (2013). The million mutation project: a new approach to genetics in Caenorhabditis elegans. Genome Res. *23*, 1749–1762.

Tijsterman, M., Pothof, J., and Plasterk, R.H.A. (2002). Frequent germline mutations and somatic repeat instability in DNA mismatch-repair-deficient Caenorhabditis elegans. Genetics *161*, 651–660.

Vannier, J.-B., Pavicic-Kaltenbrunner, V., Petalcorin, M.I.R., Ding, H., and Boulton, S.J. (2012). RTEL1 dismantles T loops and counteracts telomeric G4-DNA to maintain telomere integrity. Cell *149*, 795–806.

Wang, X., Zhao, Y., Wong, K., Ehlers, P., Kohara, Y., Jones, S.J., Mara, M.A., Holt, R.A., Moerman, D.G., and Hansen, D. (2009). Identification of genes expressed in the hermaphrodite germ line of C. elegans using SAGE. BMC Genomics *10*, 213.

Wu, Y.Y., Shin-ya, K.K., and Brosh, R.M.R. (2008). FANCJ helicase defective in Fanconia anemia and breast cancer unwinds G-quadruplex DNA to defend genomic stability. Mol. Cell. Biol. *28*, 4116–4128.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and 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.

Youds, J.L., Barber, L.J., Ward, J.D., Collis, S.J., O'Neil, N.J., Boulton, S.J., and Rose, A.M. (2008). DOG-1 is the Caenorhabditis elegans BRIP1/FANCJ homologue and functions in interstrand cross-link repair. Mol. Cell. Biol. *28*, 1470–1479.

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

ACKNOWLEDGEMENTS

We thank the *C. elegans* Genetics Center (CGC, St. Paul, MN, USA), the *C. elegans* Gene Knockout Consortium and Shohei Mitani (National Bioresource Project for the Nematode, Tokyo, Japan) for providing strains. The CGC is supported by NIH Office of Research Infrastructure Programs (P40 OD010440). We thank B. Lemmens and H. Vrieling for suggestions and comments on the manuscript. MT is supported by grants from the European Research Council (203379, DSBrepair), the European Commission (DDResponse), ZonMW/NGI-Horizon and KWF (Hubr2008-4107).

Author contributions

W.K., R.v.S. and M.T. conceived and designed the study. W.K., R.v.S, K.O., A.K, and J.v.H. performed the experiments. R.v.S. performed the bioinformatic analysis. All authors interpreted the experimental data. W.K. and M.T. wrote the manuscript.

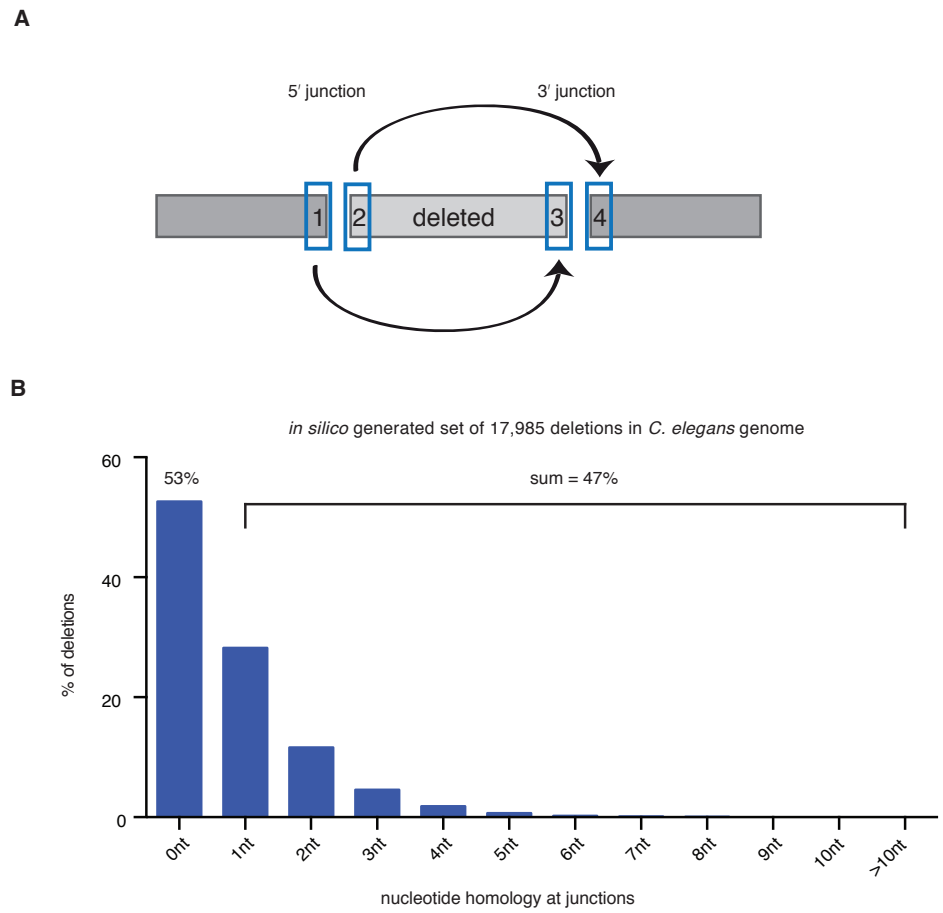
Competing financial interest

The authors declare no competing financial interest.

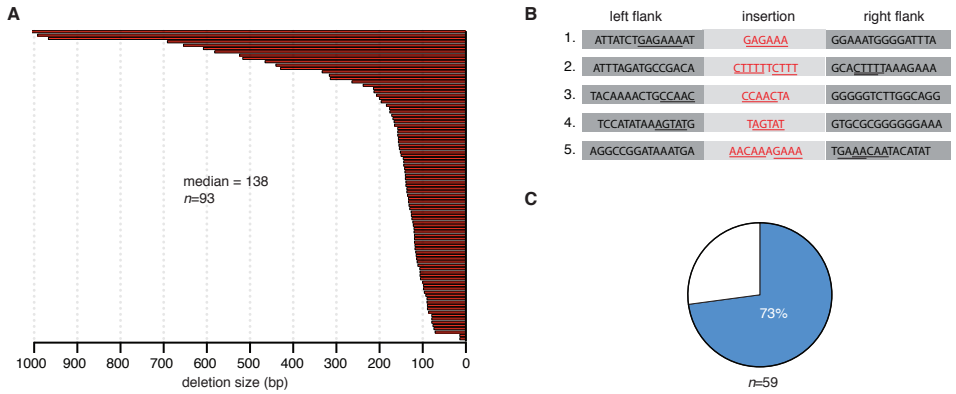
Additional information

Accession codes: Raw sequences have been made publicly available at NCBI SRA (Accession code SRP032440).

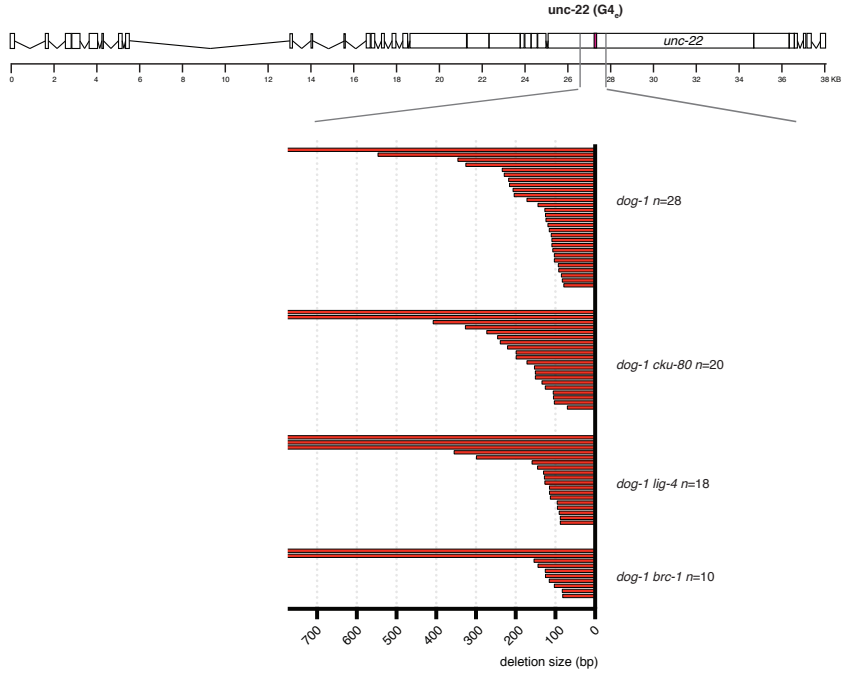
SUPPLEMENTARY MATERIALS



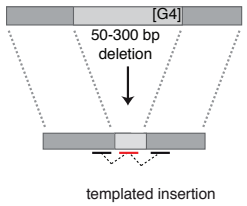
Supplementary Figure 1 | Expected chance on single nucleotide homology at junction of deletion. We determined the probability of finding one nucleotide homology at simple deletions in two different ways: (A) By calculation and according to the following rational: any deletion footprint can be reconstructed into a 5′ and 3′ junction, together giving rise to 4 fixed nucleotide positions (named 1-4 in figure); one at each junction flanking the deletion (position 1 and 4), and one at each junction adjacent to this base but within the deletion (position 2 and 3). From the 256 possible nucleotide combinations, 112, thus 44%, have the same base composition at the flank position of one junction and the deleted position of the other junction (thus position 1 with 3, and 2 with 4). This 44% likelihood of having one nucleotide homology at the junctions increases to 47% when we correct for the fact that the *C. elegans* genome is 64.6% A/T-rich. (B) By generating a set of 17,985 random deletions *in silico* using the *C. elegans* reference genome. Also here we found that 47% of all breakpoints have homology at the terminal position.



Supplementary Figure 2 | G4 DNA-induced deletions in *dog-1* mutation accumulation lines. (A) Graphical representation of the size of G4 DNA-induced deletions in basepairs (bp) found in *dog-1* mutation accumulation lines. (B) Examples of templated insertions originating from the 5′ and 3′ flanks of G4 motif. (C) Pie chart displaying the overrepresentation of single-nucleotide homology for G4 DNA-induced simple deletions (in blue).

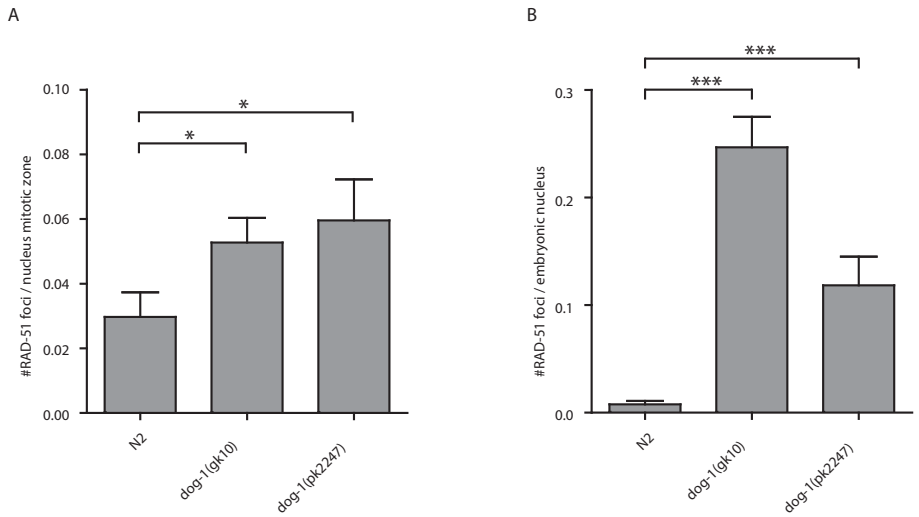


Supplementary Figure 3 | G4 DNA-induced deletion formation is not affected in animals with mutated NHEJ (*cku-80* and *lig-4*) or HR (*brc-1*). Size of G4 DNA-induced deletions at *unc-22(G4$_e$)* for the indicated genotypes.

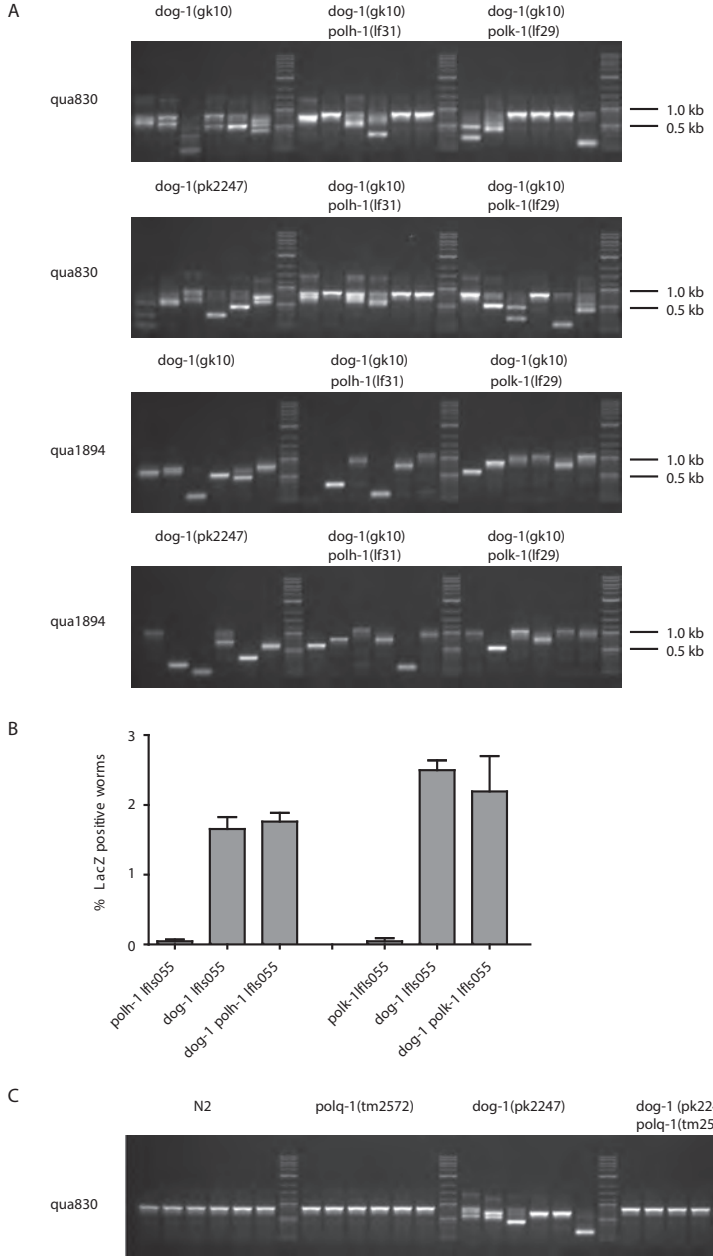


	left flank	insertion	right flank
1.	ATTTTCAAATTTTTCTGTTAAAAA	ATTTTTCTGTTAAAA	AGGGGATTTCAAAG
2.	TTATTGCATAAAAGCTCTATGAAA	ATATTGCATAAAAGCTCTATAAAAAT	GGGGGGGGAGTGGT
3.	GGATCTGTAAAAACAAATATATGA	AAAATATT	GAGTACTGGAGAAA
4.	AAGGCGCCAAGATCAATTTCAGT	AATTTCA	GGGGGTAAGTTTAT
5.	TTCTCCTAGCTGCCCCTTCTTCAC	TTCTTC	GCGCGACGTGAAAA
6.	AGTTTTATAAAATAAGAATTAATCA	CACAAAAATAA	CACAAAAACCCGGC
7.	CTCTTGGGAACATTTAGATGCCGA	CACTTTTCTT	GCACTTTTAAAGAA
8.	TTGGCTTTACAGTAGTCAATAAAA	AATGTTTTTT	GGAATGTTTTGAAT

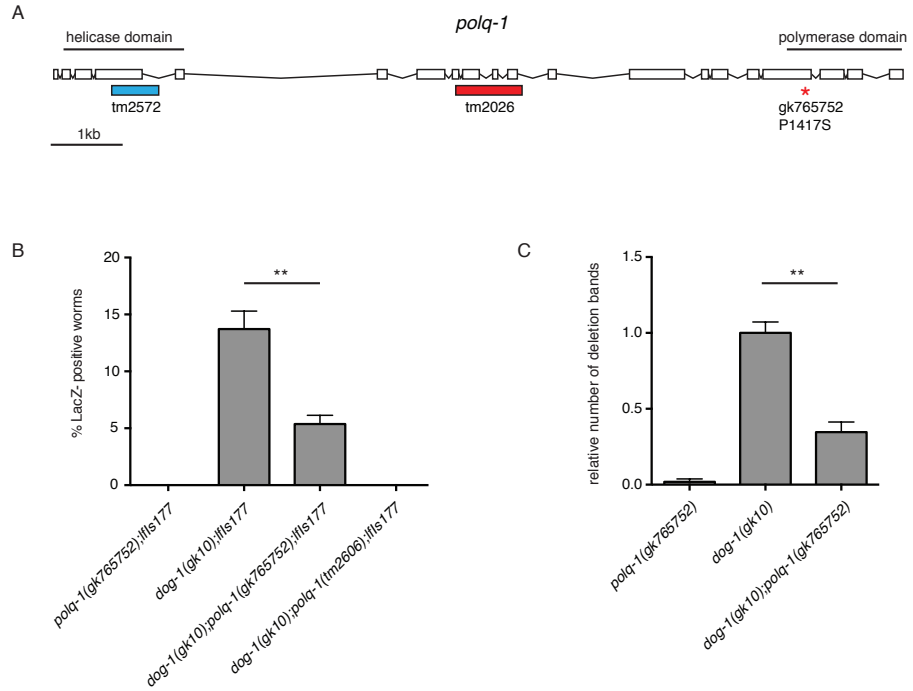
Supplementary Figure 4 | Templated insertions are derived from sequences positioned both 5′ and 3′ to a G4 motif. Sequences adapted from reference [17].



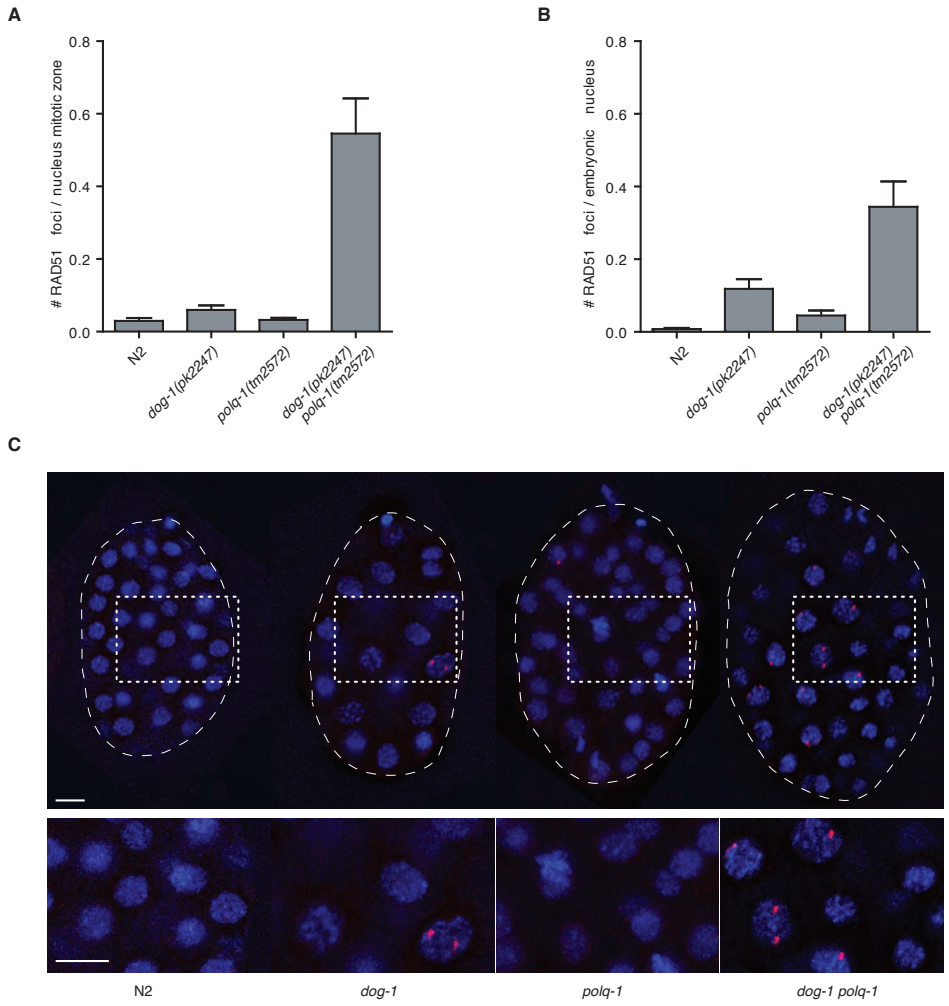
Supplementary Figure 5 | Increase of RAD-51 foci in DOG-1 deficient worms. (A) Average number of RAD-51 foci per nucleus in the mitotic zone of the germline ($n \geq 11$ germlines per genotype). (B) Average number of RAD-51 foci per nucleus in developing embryos ($n \geq 8$ embryos per genotype). Error bars indicate s.e.m., *= $P < 0.05$, ***= $P < 0.0001$ (unpaired two-tailed t-test).



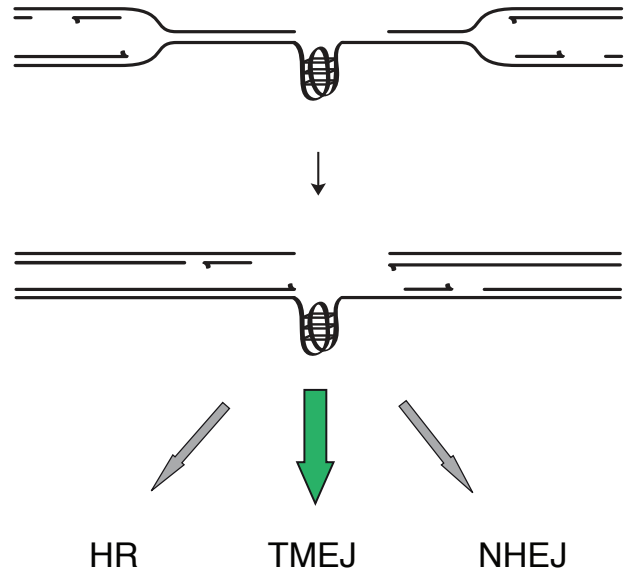
Supplementary Figure 6 | Involvement of *polq-1*, but not *polh-1* and *polk-1*, in G4 DNA-induced deletion formation. (A) Examples of deletion PCRs on qua830 and qua1894 for the indicated genotypes (5 worms per lane). Quantification can be found in Table 1. (B) No increase in deletion formation is observed in *dog-1(gk10) polh-1(ok3317)* and *dog-1(gk10) polk-1(lf29)* double muntants compared to single *dog-1(gk10)* mutants when using a LacZ-based reporter assay that reads out G4 DNA-induced deletions. Error bars, s.e.m. (*n*= 10-15 independent populations). (C) Suppresion of typical G4 DNA-induced deletion formation in *dog-1(pk2247) polq-1(tm2572)* double mutants.



Supplementary Figure 7 | Reduced G4 DNA-induced deletion formation in *polq-1* polymerase point mutant. (A) Schematic representation of the *polq-1* locus including the localisation of the helicase and polymerase protein domains. The position and size of the *polq-1* mutations and their allele names are indicated underneath. *tm2572* and *tm2606* are deletions that result in premature stops. Allele *gk765752* results in a Proline to Serine amino acid change at position 1417 of the encoded protein. (B) Quantification of reporter LacZ expression by scoring animals (*n*>150 per experiment) of the indicated genotype for the presence of ≥1 blue cell. The average percentage of at least 3 independent experiments is shown. Error bars indicate s.e.m. ** $P < 0.001$, by unpaired two tailed t-test. (C) G4 DNA-induced deletion formation at qua1894 as determined by PCR analysis (normalised to *dog-1* single mutants). Analyses were carried out in quadruplicate on 80 single L4-staged animals per genotype. Error bars indicate s.e.m. ** $P < 0.001$, by unpaired two tailed t-test.



Supplementary Figure 8 | Increased RAD-51 foci formation in *dog-1 polq-1* animals in germline and developing embryos. (A) Average number of RAD-51 foci per nucleus in the mitotic zone of the germline. Error bars indicate s.e.m. ($n \geq 10$ germlines per genotype). (B) Average number of RAD-51 foci per nucleus in developing embryos. Error bars indicate s.e.m. ($n \geq 16$ embryos per genotype). (C) Representative images of embryos of the indicated genotype stained for RAD-51 (in red) and DAPI (in blue). Blow-ups of insets are shown underneath. Scale bars indicate 5μm.



Supplementary Figure 9 | Tentative model to explain the formation of DSBs at G4 motifs. Second round replication of an unresolved ssDNA gap that still holds a replication obstructing G4 structure can result in a DSB. In this scenario, the DSB cannot be repaired by HR using the sister chromatid as a non-damaged template and thus totally relies on TMEJ for repair. Note also that this model explains why deletions are 50-300 nucleotides in size, as they reflect the size of the ssDNA gap generated in the initial replication stalling event.

Supplementary Table 1
G4 DNA-induced deletions at *unc-22* locus and qua739

G4 motif	Size deletion	Left flank	Deleted (left)	Deleted (right)	Right flank	Insertion
$G4_e$	155	CATTGTTGAAAATGA	GCATGGAAAAGATACTGCTC	– AGGGGGGGGTGAAATCAGGC	CTCCCGGGTCGACGC	–
$G4_e$	62	TTGACGAGATCCATA	AGGAAGGATGTAGGGATCAT	– GGGGAGGGGGGGGTGAAATC	AGGCCTCCCGGGTCG	–
$G4_e$	80	CAGAAGGACCACTCA	AGATTGACGAGATCCATAAG	– GGGGAGGGGGGGGTGAAATC	AGGCCTCCCGGGTCG	–
$G4_e$	133	ATGGAAAAGATACTG	CTCAGTGCAATGTTACTGTT	– GGGGGAGGGGGGGGTGAAAT	CAGGCCTCCCGGGTC	–
$G4_e$	100	TGTTCTTGATGTACC	AGGAACTCCAGAAGGACCAC	– GTGGGGGAGGGGGGGGTGAA	ATCAGGCCTCCCGGG	–
$G4_e$	124	AAGATACTGCTCAGT	GCAATGTTACTGTTCTTGAT	– GGTGGGGGAGGGGGGGGTGA	AATCAGGCCTCCCGG	ACAGTC
$G4_e$	149	ATCATTGTTGAAAAT	GAGCATGGAAAAGATACTGC	– GGTGGGGGAGGGGGGGGTGA	AATCAGGCCTCCCGG	CATTG
$G4_e$	83	CAGGAACTCCAGAAG	GACCACTCAAGATTGACGAG	– GGGGTGGGGGAGGGGGGGGT	GAAATCAGGCCTCCC	–
$G4_e$	112	TCAGTGCAATGTTAC	TGTTCTTGATGTACCAGGAA	– GGGGTGGGGGAGGGGGGGGT	GAAATCAGGCCTCCC	–
$G4_e$	137	AAAATGAGCATGGAA	AAGATACTGCTCAGTGCAAT	– GGGGTGGGGGAGGGGGGGGT	GAAATCAGGCCTCCC	–
$G4_e$	362	AACTTAACCTCGCCG	GACTTTTGGATCTCCGTATC	– GGGGTGGGGGAGGGGGGGGT	GAAATCAGGCCTCCC	–
$G4_e$	100	TTACTGTTCTTGATG	TACCAGGAACTCCAGAAGGA	– GGGGGTGGGGGAGGGGGGGG	TGAAATCAGGCCTCC	–
$G4_e$	111	TCAGTGCAATGTTAC	TGTTCTTGATGTACCAGGAA	– GGGGGTGGGGGAGGGGGGGG	TGAAATCAGGCCTCC	ACAAATGTTAT
$G4_e$	142	TTGTTGAAAATGAGC	ATGGAAAAGATACTGCTCAG	– GGGGGTGGGGGAGGGGGGGG	TGAAATCAGGCCTCC	GT
$G4_e$	176	GCTGTTCGTGGAGAT	ACTGGAGTTTACAAAATCAT	– GGGGGTGGGGGAGGGGGGGG	TGAAATCAGGCCTCC	–
$G4_e$	133	AATGAGCATGGAAAA	GATACTGCTCAGTGCAATGT	– AGGGGGTGGGGGAGGGGGGG	GTGAAATCAGGCCTC	–
$G4_e$	139	GTTGAAAATGAGCAT	GGAAAAGATACTGCTCAGTG	– AGGGGGTGGGGGAGGGGGGG	GTGAAATCAGGCCTC	–
$G4_e$	151	TACAAAATCATTGTT	GAAAATGAGCATGGAAAAGA	– AGGGGGTGGGGGAGGGGGGG	GTGAAATCAGGCCTC	–
$G4_e$	200	ATCCTCAGCAATTCA	CATCTTCTCTGCTGTTCGTG	– AGGGGGTGGGGGAGGGGGGG	GTGAAATCAGGCCTC	T
$G4_e$	128	AGCATGGAAAAGATA	CTGCTCAGTGCAATGTTACT	– GAGGGGGTGGGGGAGGGGGG	GGTGAAATCAGGCCT	AAGATGT
$G4_e$	142	CATTGTTGAAAATGA	GCATGGAAAAGATACTGCTC	– GAGGGGGTGGGGGAGGGGGG	GGTGAAATCAGGCCT	AAATGA
$G4_e$	224	GAGCTGATGTTACGA	ACACACCAACATCCTCAGCA	– GAGGGGGTGGGGGAGGGGGG	GGTGAAATCAGGCCT	–
$G4_e$	140	ATCATTGTTGAAAAT	GAGCATGGAAAAGATACTGC	– GAGGGAGGGGGTGGGGGAGG	GGGGGGTGAAATCAG	–
$G_{11}AG_{11}$	104	ATGTTACTGTTCTTG	ATGTACCAGGAACTCCAGAA	– GGGGGGGAGGGGGGGGGGGA	TCCATCAGGCCTCCC	TCACTGTTCTTGTGTTCT
$G_{11}AG_{11}$	110	AGTGCAATGTTACTG	TTCTTGATGTACCAGGAACT	– GGGGGGGAGGGGGGGGGGGA	TCCATCAGGCCTCCC	–
$G_{11}AG_{11}$	116	CTGCTCAGTGCAATG	TTACTGTTCTTGATGTACCA	– GGGGGGGAGGGGGGGGGGGA	TCCATCAGGCCTCCC	–
$G_{11}AG_{11}$	187	CATCTTCTCTGCTGT	TCGTGGAGATACTGGAGTTT	– GGGGGGGAGGGGGGGGGGGA	TCCATCAGGCCTCCC	–
$G_{11}AG_{11}$	272	TTGCTAAATGGAAGG	CCAACGATGCAACAATCGAT	– GGGGGGGAGGGGGGGGGGGA	TCCATCAGGCCTCCC	–
$G_{11}AG_{11}$	113	GCTCAGTGCAATGTT	ACTGTTCTTGATGTACCAGG	– GGGGGGGGAGGGGGGGGGGG	ATCCATCAGGCCTCC	–
$G_{11}AG_{11}$	113	GCTCAGTGCAATGTT	ACTGTTCTTGATGTACCAGG	– GGGGGGGGAGGGGGGGGGGG	ATCCATCAGGCCTCC	–
$G_{11}AG_{11}$	124	GAAAAGATACTGCTC	AGTGCAATGTTACTGTTCTT	– GGGGGGGGAGGGGGGGGGGG	ATCCATCAGGCCTCC	–
$G_{11}AG_{11}$	128	CATGGAAAAGATACT	GCTCAGTGCAATGTTACTGT	– GGGGGGGGAGGGGGGGGGGG	ATCCATCAGGCCTCC	T
$G_{11}AG_{11}$	133	ATGAGCATGGAAAAG	ATACTGCTCAGTGCAATGTT	– GGGGGGGGAGGGGGGGGGGG	ATCCATCAGGCCTCC	CATGGAAA
$G_{11}AG_{11}$	158	GGAGTTTACAAAATC	ATTGTTGAAAATGAGCATGG	– GGGGGGGGAGGGGGGGGGGG	ATCCATCAGGCCTCC	–
$G_{11}AG_{11}$	556	TTGAGAAGAAGGACA	AGTTTGGAGACTGGGTTGAG	– GGGGGGGGAGGGGGGGGGGG	ATCCATCAGGCCTCC	–
$G_{11}AG_{11}$	62	ACTCAAGATTGACGA	GATCCATAAGGAAGGATGTA	– GGGGGGGGGAGGGGGGGGGG	GATCCATCAGGCCTC	–
$G_{11}AG_{11}$	82	CCAGGAACTCCAGAA	GGACCACTCAAGATTGACGA	– GGGGGGGGGAGGGGGGGGGG	GATCCATCAGGCCTC	–
$G_{11}AG_{11}$	110	TCAGTGCAATGTTAC	TGTTCTTGATGTACCAGGAA	– GGGGGGGGGAGGGGGGGGGG	GATCCATCAGGCCTC	C

$G_{11}AG_{11}$	145	ATCATTGTTGAAAAT	GAGCATGGAAAAGATACTGC	– GGGGGGGGGAGGGGGGGGGG	GATCCATCAGGCCTC	–
$G_{11}AG_{11}$	178	TCTGCTGTTCGTGGA	GATACTGGAGTTTACAAAAT	– GGGGGGGGGAGGGGGGGGGG	GATCCATCAGGCCTC	–
$G_{11}AG_{11}$	227	AAGAGCTGATGTTAC	GAACACACCAACATCCTCAG	– GGGGGGGGGAGGGGGGGGGG	GATCCATCAGGCCTC	–
$G_{11}AG_{11}$	614	TGATCTTGAATGGAA	GCCACCAGCAAACGACGGTG	– GGGGGGGGGAGGGGGGGGGG	GATCCATCAGGCCTC	TTCTTGAACGATCTTGAAC
$G_{11}AG_{11}$	62	GGACCACTCAAGATT	GACGAGATCCATAAGGAAGG	– CTAGGGGGGGGGGGAGGGGG	GGGATCCATCAGGCC	CGAA
$G_{11}AG_{11}$	92	GTTCTTGATGTACCA	GGAACTCCAGAAGGACCACT	– GGGGGGGGGGGAGGGGGGGG	GGGATCCATCAGGCC	–
$G_{11}AG_{11}$	143	ATCATTGTTGAAAAT	GAGCATGGAAAAGATACTGC	– GGGGGGGGGGAGGGGGGGGG	GGGATCCATCAGGCC	–
$G_{11}AG_{11}$	188	ATTCACATCTTCTCT	GCTGTTCGTGGAGATACTGG	– GGGGGGGGGGAGGGGGGGG	GGGATCCATCAGGCC	–
$G_{11}AG_{11}$	88	CTTGATGTACCAGGA	ACTCCAGAAGGACCACTCAA	– AGGGGGGGGGGGAGGGGGG	GGGGATCCATCAGGC	–
$G_{11}AG_{11}$	94	GTTACTGTTCTTGAT	TACCAGGAACTCCAGAAGGA	– AGGGGGGGGGGGAGGGGGG	GGGGATCCATCAGGC	–
$G_{11}AG_{11}$	104	GTGCAATGTTACTGT	TCTTGATGTACCAGGAACTC	– AGGGGGGGGGGGAGGGGGG	GGGGATCCATCAGGC	–
$G_{11}AG_{11}$	58	ACTCAAGATTGACGA	GATCCATAAGGAAGGATGTA	– TAGGGGGGGGGGAGGGGGG	GGGGGCAGAAGGACC	–
$G_{11}AG_{11}$	127	CATGGAAAAGATACT	GCTCAGTGCAATGTTACTGT	– CTAGGGGGGGGGGGAGGGGG	GGGGGATCCATCAG	–
$G_{11}AG_{11}$	203	ACACCAACATCCTCA	GCAATTCACATCTTCTCTGC	– CTAGGGGGGGGGGGAGGGGG	GGGGGGATCCATCAG	–
$G_{11}AG_{11}$	235	TCGATACAGGAGCAA	GAGCTGATGTTACGAACACA	– CTAGGGGGGGGGGGAGGGGG	GGGGGGATCCATCAG	–
$G_{11}AG_{11}$	79	GTACCAGGAACTCCA	GAAGGACCACTCAAGATTGA	– CCTAGGGGGGGGGGGAGGGG	GGGGGGGATCCATCA	–
$G_{11}AG_{11}$	158	AGATACTGGAGTTTA	CAAAATCATTGTTGAAAATG	– CCTAGGGGGGGGGGGAGGGG	GGGGGGGATCCATCA	–
$G_{11}AG_{11}$	172	TCTGCTGTTCGTGGA	GATACTGGAGTTTACAAAAT	– CCTAGGGGGGGGGGGAGGGG	GGGGGGGATCCATCA	–
$G_{11}AG_{11}$	704	TTGACTTCTGATCAT	GGCAATTGTTGCTAAGAATC	– CCTAGGGGGGGGGGGAGGGG	GGGGGGGATCCATCA	–
$G_{11}AG_{11}$	623	GATAAGGATCATGTT	GATCTTGAATGGAAGCCACC	– GCCTAGGGGGGGGGGGAGGG	GGGGGGGATCCATC	TTTTTTTT
$G_{11}AG_{11}$	34	ATAAGGAAGGATGTA	GGGATCATATGAATTGCCTA	– TGCCTAGGGGGGGGGGGAGG	GGGGGGGGATCCAT	–
G_{23}	250	AACAATCGATACAGG	AGCAAGAGCTGATGTTACGA	– GGGGGGGGGGGGGGGGATGC	ATCAGGCCTCCCGGG	–
G_{23}	89	CCAGGAACTCCAGAA	GAAGGACCACTCAAGATTGA	– GGGGGGGGGGGGGGGGATG	CATCAGGCCTCCCGG	–
G_{23}	164	ACTGGAGTTTACAAA	ATCATTGTTGAAAATGAGCA	– GGGGGGGGGGGGGGGGATG	CATCAGGCCTCCCGG	T
G_{23}	494	GCAACTGCTGATAAT	TTGACTCCAGGAGAGACTTA	– GGGGGGGGGGGGGGGGATG	CATCAGGCCTCCCGG	CAT
G_{23}	80	AACTCCAGAAGGACC	ACTCAAGATTGACGAGATCC	– GGGGGGGGGGGGGGGGGAT	GCATCAGGCCTCCCG	–
G_{23}	130	GCATGGAAAAGATAC	TGCTCAGTGCAATGTTACTG	– GGGGGGGGGGGGGGGGGGA	TGCATCAGGCCTCCC	–
G_{23}	106	CAATGTTACTGTTCT	TGATGTACCAGGAACTCCAG	– GGGGGGGGGGGGGGGGGGA	TGCATCAGGCCTCCC	–
G_{23}	112	TCAGTGCAATGTTAC	TGTTCTTGATGTACCAGGAA	– GGGGGGGGGGGGGGGGGGA	TGCATCAGGCCTCCC	–
G_{23}	77	CTCCAGAAGGACCAC	TCAAGATTGACGAGATCCAT	– GGGGGGGGGGGGGGGGGGA	TGCATCAGGCCTCCC	–
G_{23}	169	TGGAGATACTGGAGT	TTACAAAATCATTGTTGAAA	– GGGGGGGGGGGGGGGGGGA	TGCATCAGGCCTCCC	AT
G_{23}	257	GCCAACGATGCAACA	ATCGATACAGGAGCAAGAGC	– GGGGGGGGGGGGGGGGGGG	ATGCATCAGGCCTCC	–
G_{23}	113	GCTCAGTGCAATGTT	ACTGTTCTTGATGTACCAGG	– GGGGGGGGGGGGGGGGGGG	ATGCATCAGGCCTCC	–
G_{23}	103	ATGTTACTGTTCTTG	ATGTACCAGGAACTCCAGAA	– GGGGGGGGGGGGGGGGGGG	ATGCATCAGGCCTCC	–
G_{23}	159	TGGAGTTTACAAAAT	CATTGTTGAAAATGAGCATG	– GGGGGGGGGGGGGGGGGGG	ATGCATCAGGCCTCC	A
G_{23}	160	CTGGAGTTTACAAAA	TCATTGTTGAAAATGAGCAT	– GGGGGGGGGGGGGGGGGGG	ATGCATCAGGCCTCC	AAAAAA
G_{23}	245	ACAATCGATACAGGA	GCAAGAGCTGATGTTACGAA	– GGGGGGGGGGGGGGGGGGG	ATGCATCAGGCCTCC	–
G_{23}	72	CCAGAAGGACCACTC	AAGATTGACGAGATCCATAA	– GGGGGGGGGGGGGGGGGGG	GGATGCATCAGGCCT	–
G_{23}	225	AGAGCTGATGTTACG	AACACACCAACATCCTCAGC	– GGGGGGGGGGGGGGGGGGG	GGATGCATCAGGCCT	GT
G_{23}	55	AGATTGACGAGATCC	ATAAGGAAGGATGTAGGGAT	– GGGGGGGGGGGGGGGGGGG	GGGATGCATCAGGCC	CTCCACATAACTGTGATCTTGATGTTTGCTTCAGCTTCTCC

Supplementary Table 2
G4 DNA-induced deletions found in *dog-1* mutation accumulation lines using whole genome sequencing

Chromosome	Start	End	Size	G4 motif	LeftFlank	Deleted (left)		Deleted (right)	RightFlank	Insertion
CHROMOSOME_I	11438825	11438838	13	GGGGGGGGGGGGGGGG	CGACGAGGATGGGGG	GGGGGG	–	GGGGGGG	TAGATTCTTTTCAGA	T
CHROMOSOME_X	17231098	17231111	13	GGGGGGGGGGGGGGGGGGG	TTGAAAAACATTTTT	GGGGGG	–	GGGGGGG	GGGGGGGGAGAAAT	TAT
CHROMOSOME_X	6756169	6756239	70	GGGGGGGGGGGGGGGGGGGTCTTAGGG	AATATCATAACAAAA	CATTCCAAAATGTTTG	–	GGGGGGGGGGGGGGG	GGGGTCTTAGGGAAC	ATGC
CHROMOSOME_I	11271598	11271671	73	GGGGGAGGTAGGGGGGGGGGGGGG	TGATGAGAAAATGGG	AATGATTGGATTTATT	–	GGTAGGGGGGGGGGG	GGGGTTTCACGTTAT	–
CHROMOSOME_III	8292413	8292488	75	GGGGGGGGGGGGGGGGGG	TTTTGGGACGTTATT	CTCTCTTTGTATGTGT	–	GGGGGGGGGGGGGGG	GGGTCCAAAGAAACG	T
CHROMOSOME_III	3967711	3967789	78	GGGGGGGGGGGGGGGGTGGG	CTCCTACTCGTTTTT	CCTTCTTGCTTCTACT	–	GGGGGGGGGGGGTGG	GTCTCTTCAGAGAGC	–
CHROMOSOME_V	19074326	19074404	78	GGGGGGGGGGGGGG	AAAATTTCATATTTT	TCCCGTGTTTTTTCTA	–	GGGGGGGGGGGCACG	TGAAGTCATTTTTCG	–
CHROMOSOME_V	19666731	19666809	78	GGGGGGGGGGGGGGGGGGG	TATGAAAGCATTAAA	AAACCTAACAAACGCA	–	GGGGGGGGGGGGGGG	AGTGGTTAGTATGCT	–
CHROMOSOME_III	10291918	10292003	85	GGGGGGGGGGGGGG	TTCAAACCAAAAAAA	AACTCACAGTACAAAG	–	TGGGGGGGGGGGGGG	GAAAGAATTGGGGAT	–
CHROMOSOME_IV	4243102	4243190	88	GGGGGGGGGGGGGGGG	AAAAAAAAGACGCGA	TTGATGTTGAAGTGTA	–	GGGGGGGGGGGGGGG	GTAAGAAAAATTGCC	–
CHROMOSOME_IV	2657144	2657232	88	GGGGGGGGGGGGGGGGG	GGAACGAAATTGTTC	CCCACACCGATCCATC	–	GGGGGGGGGGGGGGC	AGAAAAACATTAAAT	–
CHROMOSOME_I	11657955	11658044	89	GGGGGGGGGGGGGG	GACGCAGACAGTTGA	AGGGGGAGGAGAACTG	–	GGGGGGGGGGGGGGA	CAATTGGAAATTGAG	–
CHROMOSOME_I	4719617	4719706	89	GGGCCGAGGGGGGGGGGGGGGGG	TATGTATTATTTTCT	TGTTTTATTTTTTGAA	–	GGGGGGGGGGGGGGG	TGACCTTTGAATACT	–
CHROMOSOME_X	648570	648662	92	GGGGGGAGGGGGGGGGGG	AAACTGATGATATTA	AACTGCAAACAACTTT	–	GGAGGGGGGGGGGGT	AAATCTGGGACAAAA	–
CHROMOSOME_I	14325553	14325648	95	GGGGGGGGGGGGGGGGGTGGG	GTATAGAGAATTGAT	GGAGAACGAGATAAAC	–	GGGGGGGGGGGTGGG	TCCCCGATGCACCAT	ATT
CHROMOSOME_IV	3628992	3629089	97	GGGGAGTGGGGGGGGGGGGGGGGGGGGGGGGGG	CCAACGGAGAGCTGA	GGAGCCACCCCTGTAA	–	GGGGGGGGGGGGGGG	GGATTTGCTACCGGT	–
CHROMOSOME_IV	12048766	12048863	97	GGGGTGGGGGGGGGGGGGCGGGGGG	CAGGGAAATCATGAT	TGTGATGGTTCCTACA	–	GGGGCGGGGGGAAGA	ATCTTTCTTTATTAT	GGG
CHROMOSOME_I	11247148	11247247	99	GGGGGGGGGGGGG	TATCTACAACATGAC	ACTTGAGTACTCATAC	–	TGGGGGGGGGGGGGG	CAGATAGCTCAGTCG	TCTCTGTTATCTACAACATGA
CHROMOSOME_I	3623822	3623926	104	GGGAGAGGGGGGGGGGGGGGG	TTCTTTCAAAACTAC	GTAGCCATTTCACACG	–	AGGGGGGGGGGGGGG	GGTATTACATATTAG	–
CHROMOSOME_X	5140352	5140457	105	GGGGGGGGGGGGGGGGTGGGGGTGTTGGG	TCTCTTTAGAAGCGG	AAAAAGGACAACGCTT	–	GGGGGGGGGTGGGGG	TGTTGGGCTCTAGTA	TA
CHROMOSOME_I	13859686	13859791	105	GGGGGGGGGGGGGGGGGGGGGGGGAAATGGGG	ATTATCTGAGAAAAT	TCGAGAGAAAAAACAA	–	GGGGGGGGGGGGGGG	GGAAATGGGGATTTA	GAGAAA
CHROMOSOME_V	16693069	16693175	106	GGGGGGGGAGGGGGGGGGGGGGGGGG	ACACAAGGTCTCGCA	GGGTCTTGCGTGATAG	–	GAGGGGGGGGGGGGG	GGGGCAGTGGGTCTC	–
CHROMOSOME_V	980795	980905	110	GGGGGGGGGGGGGGGGGGGGG	CACCACGTGACAACA	CGATTTTCGCGTAAAA	–	GGGGGGGGGGGGGGG	GAAATCAAGAAATCG	AAA
CHROMOSOME_V	4498844	4498956	112	GGGGGGGGGGGGGGGGGGGGG	ATTCTTGGCAGCCAA	ATTCCATGTCGCCTAT	–	GGGGGGGGGGGGGGA	GCGATGATGAGACGA	–
CHROMOSOME_V	3185937	3186050	113	GGGTCCTGGGGGGGGAGGGGGGGGGGGGG	CAATTTCAAACGTTT	GATAACTCACAGTGGT	–	GGGGGGGGGGGGGGT	GGTGTCTAGACAGTC	–
CHROMOSOME_I	1779244	1779358	114	GGGGGGGGGGGGG	GGTGCAATAACAAGT	GTCCCATTTTGGTTTG	–	TGGGGGGGGGGGGGG	AAATTCCCGCATTTT	TTTTTTTTTTCCCGCATTTTTATTTTT
CHROMOSOME_X	2301994	2302109	115	GGGGGGGGGGGGGGGGGGGGG	GGATTTTCTTTCTCT	TTTACACATTCTTATT	–	GGGGGGGGGGGGGTT	GAGAGAGAAACACTG	–
CHROMOSOME_I	33456	33573	117	GGGGGGGGGGGGAGGGGGGGGGGG	AAACCCGCTATCACA	GACTCAATGCGCACCG	–	AGGGGGGGGGGGGGT	GAAAAAAGAGAAATA	–
CHROMOSOME_I	10370627	10370744	117	GGGGGGGGGGGGGTGGGG	ATTTAGATGCCGACA	GATATCAATAGCCGTA	–	GGGGGGGGGGGTGGG	GCACTTTTAAAGAAA	CTTTTTCTTT
CHROMOSOME_X	16966755	16966873	118	GGGGGGGGGGGGG	CACTGTATAGGTTTT	GGCTCAAAGAACTGCT	–	TGGGGGGGGGGGGGG	TAAATGGTAGTTTGA	TTTTT
CHROMOSOME_X	15511810	15511928	118	GGGGTGGGGGGGGGGGGGGGGGGGGGGGGG	CTTCACCAACACCAC	CACACTATACCATTTG	–	GGGGGGGGGGGGGGG	CAGTCAAATGGTGTG	–
CHROMOSOME_II	12951642	12951760	118	GGGTGGGGGGGGGGGGG	TACAAAACTGCCAAC	TAAAACTTCAAATCAT	–	ATGGGTGGGGGGGGG	GGGGGTCTTGGCAGG	CCAACTA
CHROMOSOME_II	12826964	12827083	119	GGGGGGTGGGGGGGGGGGG	TCGGATAATTGTAGT	TTCGATAATTGTTATA	–	TGTGAGGGGGGTGGGG	GGGGGGGGGTGATAA	C
CHROMOSOME_X	3094437	3094556	119	GGGGGGGGGGGGGGGGGG	TCCTCTATATCTCCC	GTATTGTTACGCGCGC	–	TAGGGGGGGGGGGGGG	GGGGCCCAAAAGGCA	–
CHROMOSOME_I	14230955	14231076	121	GGGGGGGGGGGGGGGGGGGGGGG	GTTTTTTAAATTAAA	GTTCACGGTTTAGAAA	–	GGGGGGGGGGGGGGG	GTGTCTTGGAAGAGC	TTAAC
CHROMOSOME_V	8996750	8996872	122	GGGGGGGGGGGGGGG	ATTTCCGGGGCGCC	AGAATAGGCTTATTGG	–	TGGGGGGGGGGGGGG	GGAGCAGAAATCGGA	T
CHROMOSOME_V	14952691	14952815	124	GGGGGGGGGGGGGGGGGGG	CAACATGTTTTTTTG	CTGTTTTTTTTGTTTT	–	GGGGGGGGGGGTGGTC	CTGGGGATTTAGAAC	–
CHROMOSOME_V	20756348	20756473	125	GGGGGGGGGGGGGGGGGGGGGGGGG	TTTCAACTGACTCAT	GGCTGAAAATAGTATA	–	GGGGGGGGGGGGGGT	TCCTCGTCGTCGTCG	–
CHROMOSOME_I	2393304	2393430	126	GGGGGGGGGGGGGGGTGTCTGGG	ATGACCTAGAGCACC	GGGACCGTGGTGACTC	–	AAACGGGGGGGGGGG	GGGGGTGTCTGGGCG	–
CHROMOSOME_III	7564930	7565059	129	GGGGGGGGGGGGGGGGGGG	ATAAGTGTCTAAAGT	GGTTCGCAACTTTTGA	–	GGGGGGGGGGGGGGG	GTTCCTACAATTTTT	CTT
CHROMOSOME_IV	3628955	3629086	131	GGGGAGTGGGGGGGGGGGGGGGGGGGGG	ATAATTGGTTAGAAA	GGGGGATTTCTAGAGA	–	GGGGGGGGGGGGGGG	GGGGGATTTGCTACC	–
CHROMOSOME_IV	17154640	17154772	132	GGGGGGGGGGGGGGGGGG	TGAGCGAACATATGT	GAAATCCCTAGTTATT	–	GGGGGAAAGATATATA	GGCATAAAATGTTAC	–
CHROMOSOME_I	817788	817920	132	GGGGGGGGGGGGGGGGG	CACTACCTCCTAGTA	TTAAGGCTCCAGCTAA	–	GGGGGGGGGGGGGCAT	TCGTACAGTGTCAAC	–
CHROMOSOME_X	13775569	13775703	134	GGGGGGGGGGGGGGGGGGGGGGGTGGGG	AAGAAAAGAAGGTCG	CTTCAGTTTTTTGTGT	–	GGGGGGGGGGTGGGG	TTTGCGTCGAATCGT	–
CHROMOSOME_III	11474785	11474921	136	GGGCGGGGGGGGGGGGGGGGG	TCCGAATTATTTTCT	CGAATATCTAAAAAAC	–	GGGGGGGGGGGCTTAG	TAGGACCCCCCTTCG	TTTT
CHROMOSOME_IV	16646624	16646760	136	GGGGGGGGGGGGG	AAAACACAAAAAAAA	ACAAATATTTAGGGAA	–	GTCCAGGGGGGGGGG	GGGAATAAAAAGAAA	–

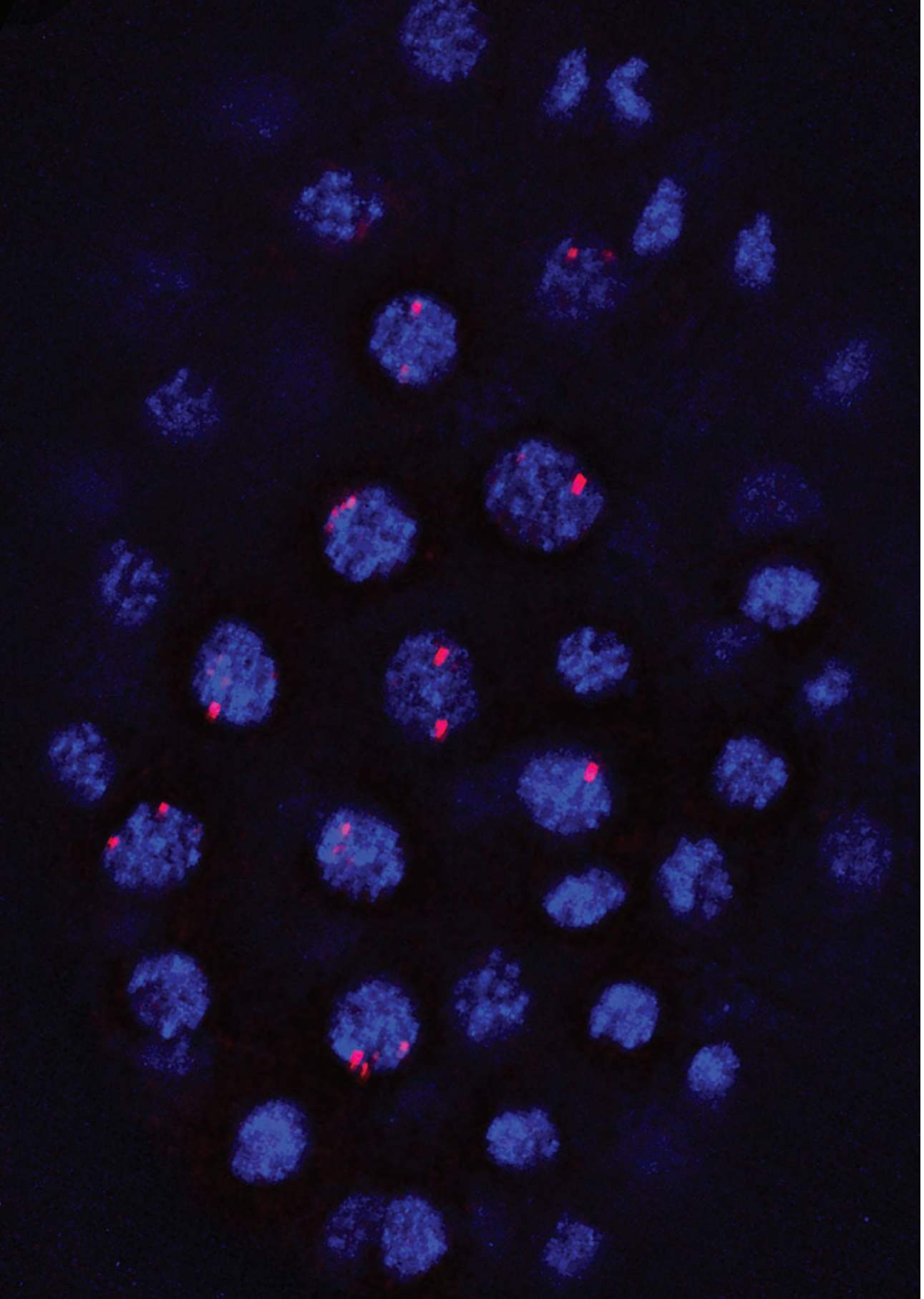
Supplementary Table 2 (*Continued*)
G4 DNA-induced deletions found in *dog-1* mutation accumulation lines using whole genome sequencing

CHROMOSOME_II	1325641	1325779	138	GGGGGGGGGGGGGGGGTGGG	AGGGTAGCGTCTTCT	TCACACTAATCTCCCG	–	GGGGGGGGTGGGGAGG	TGCAGGTTTTTTTCG	A
CHROMOSOME_II	8534695	8534833	138	GGGGGGGGGGGGG	AATAAACAACAACAA	TAAACAACAACAATAA	–	GGGGGGGGGGGGGCA	GAAATCAGAAAAGTG	–
CHROMOSOME_X	3903025	3903164	139	GGGGGGGGGGGGGGGGG	GATAGTAAGCACAGA	GACCGTATAAACCATG	–	GGGGGGGGGGGTAGAA	CTAAAACTGGAAGGG	–
CHROMOSOME_III	13534793	13534932	139	GGGGGGGGGGGGGGGGGGG	AAAAAAGGAAAAAAA	GTACGGTTTTATAGAT	–	GGGGGGGGGGGGGGT	GAATTTGTCAATTTT	A
CHROMOSOME_X	14025342	14025482	140	GGGGGGGGGGGGG	CATTAATGAAACGTG	AACTTTTTGTAATTAC	–	TTGGGGGGGGGGGGG	GCAGAAATCGGAAAA	–
CHROMOSOME_IV	16759342	16759483	141	GGGGTGGGTGGGCGGG	ATCCTGCCTGTCTAC	TCTGGTGGAATTGCGG	–	GGGTGGGTGGGCGGGA	ACTCACACGTAGAAT	CGGAGGAATCCTGCCTGTCT
CHROMOSOME_II	3244696	3244839	143	GGGGGGGGGGGGGG	AAAATTAATTTCTGA	ATTTTTCAACAAATTT	–	AGGGGGGGGGGGGGG	GTCAAACTTCTAGTA	–
CHROMOSOME_I	11387516	11387659	143	GGGCGGGGCCGGTGGGCGGGG	ATGTCATATTCCAGT	CGGTTTCCTCGACTCT	–	GGTGGGCGGGGCCAGT	TTTATGGGCTCCGCG	–
CHROMOSOME_IV	7304350	7304494	144	GGGGGGGGGGGGGGGGGGG	AAGTCAATGTAAAAG	AAAACAAGAAGCGTAT	–	GGGGGGGGGGGGGTTG	TGGTGACCTCCAAAA	–
CHROMOSOME_I	1342669	1342818	149	GGGGGAGGGGGGGGGGGGGGGGGGGGG	GCCGGTGGCGGCGAT	GCAGGTGTAGGCGAGA	–	GGGGGGGGGGGGGGG	GGGGGTCTCGCCACG	–
CHROMOSOME_I	969738	969889	151	GGGGGGGGGGGGGAGGG	ATCAGATTCAACGAC	GTCTCGCGATCCCGAC	–	GGGGGGGGGGGGGAG	GTCTTACGTTGTGAT	–
CHROMOSOME_V	16693028	16693181	153	GGGGGGGGAGGGGGGGGGGGGGGGGG	GTAAAGCTGAAATAT	GGAATTTGGTTTCCCA	–	GGGGGGGGGGGGGGA	GTGGGTCTCGAAGCA	–
CHROMOSOME_X	17425380	17425534	154	GGGGGGGGGGGGGGGGTGGGG	ATGCCGCAAAAACTT	GACACCTCCAGTTTCC	–	GGGGGGGGGGTGGGG	GAGCGGAGAGAGAAA	–
CHROMOSOME_X	3102846	3103001	155	GGGGGGGGGGGGGGGGGG	GGCCACCGCACACAA	TTGGTTTGACATGAAC	–	GGGGGGGGGGGGGAG	AGACACTCATACGTT	–
CHROMOSOME_III	3877334	3877489	155	GGGGGGGGGGGGGGGGGGGGGGGGGGGGG	TCCATATAAAGTATG	CATTCAAAACGTGCAA	–	GGGGGGGGGGGGGAG	GTGCCGCGGGGGAAA	TAGTAT
CHROMOSOME_II	9578699	9578856	157	GGGGGAGTAGGGCCGGCGGGG	TATTCAAAAATGTTT	AGCACAAAAAGTGAAT	–	AGTAGGGCGGGCGGGG	CAGCAGCAATCACAC	–
CHROMOSOME_X	9647081	9647238	157	GGGGGGGGGGGGGG	GCTGGTTAATCCGAA	GGTCACGGGTTTGTTC	–	GACGGGGGGGGGGGG	GGCTCACTTTTCCAC	–
CHROMOSOME_I	12222738	12222895	157	GGGGGGGGGGGGGGGGGGGGGTGGG	ATTATATCAAAAAAT	GTTCAACTAGAGAATT	–	GGGTGGGCTGAGAGAC	GCGCAGACACTCTCG	AAAAAAAAAAGA
CHROMOSOME_X	13775568	13775732	164	GGGGGGGGGGGGGGGGGGGGGGGGTGGGG	TCTTAAATTCTCACC	GGAAGTTAAGTAAAAA	–	GGGGGGGGGTGGGGT	TTGCGTCGAATCGTC	–
CHROMOSOME_V	3817937	3818102	165	GGGGGGGGGGGGGGGGGGGG	TTTGTAAAACGGAGG	TAAATTACCGGATGTG	–	GGGGGGGGGGGGGGT	TTGCTAATTTTTCCT	–
CHROMOSOME_I	10395371	10395539	168	GGGCGGGAGGGGGGGGGGGGGGGGGG	GCTCCCTACCTTCAA	TAAATTTAGAAGTACT	–	GGGGGGGGGGGGGGG	GGGTGCTTCTTGTTT	–
CHROMOSOME_V	5261808	5261979	171	GGGGGGGGGGGGGGGGGGGG	TGTAACGACGATTAC	ATCTTTCAGTGGAAGA	–	GGGGGGGGGGGGGCGT	AAAATGCCGACCAGG	GTGAAAAA
CHROMOSOME_V	4809207	4809382	175	GGGAGGGGGGGGGGGG	ATTCCGGCAATTTGC	CGATTTGCCGGAAAAA	–	GAGGGGGGGGGGGGT	CAAAAGTGAATGTAC	–
CHROMOSOME_X	8250930	8251106	176	GGGAGGGGGTGGGGGAGGGGGGGG	AGGTCAAGAATGGAA	AATTTCTCCAAGTGTC	–	GGGGGAGGGGGGGGG	TAATTTGAGAGGTTTT	–
CHROMOSOME_X	17231087	17231270	183	GGGGGGGGGGGGGGGGGGGG	TTTGGAATTGTCTTA	GGTTCTGAACGTCAAC	–	GGGGGGGGGGGGGGA	AAATCGGAAAAGTAC	A
CHROMOSOME_X	3825212	3825407	195	GGGGGGGGGGGGGGG	AGTTTAACTATGAAA	CCAAAAAAAAAACAAT	–	AGTGGGGGGGGGGGG	GGGCAGAGAACAGAG	AT
CHROMOSOME_X	2333414	2333613	199	GGGGGGGGGGGGGGGGGGGG	AGGCCGGATAAATGA	CCCAGTTTCACTGGGC	–	GGGGGGGGGGGGGGT	TGAAACAATACATAT	AACAAAGAAA
CHROMOSOME_V	12242066	12242272	206	GGGAGGGGGGGGGGGGG	AAAAAAGAGGATCTA	GAGTATTTTTGGAGAC	–	GGGAGGGGGGGGGGG	GGTGGCAGTAAAATT	–
CHROMOSOME_X	12140024	12140236	212	GGGGGGGGGGGGGGG	ATAGGTCCAGGGACC	AATATCCCTAATTCCA	–	TGGGGGGGGGGGGGG	GCACCATGTACCTGG	T
CHROMOSOME_III	11207014	11207227	213	GGGGGGGGGGGGGG	ATTTGCACAAAATCG	AATATTTTTCAACGGA	–	GGGGGGGGGGGGGCG	CCCCCGCCAAACGAC	–
CHROMOSOME_I	11782132	11782369	237	GGGGGGGGGGGGGGGGG	TCCCCCAAAAAAAAG	TGGAAAGGTGGCTGGC	–	GGGGGGGGGGGGGGG	GACCCAACACAAGCA	C
CHROMOSOME_III	13183780	13184043	263	GGGGGAGGGGGGGGGGG	GATGAATGGAAATTA	TTAACAAAAAATTATC	–	GGAGGGGGGGGGGGA	GCTCAGTGCAGACAG	–
CHROMOSOME_I	11797981	11798294	313	GGGGGGGGGGGGGGGGGGGG	AGAATACATAGGTGA	GATTAGTAGTCGGGGA	–	GGGGGGGGGGGGGGG	GGAATTGAAATTTTT	–
CHROMOSOME_II	4983411	4983726	315	GGGGGGGGGGGGGGGGGGG	AACCCCCAACCACAC	AATAAAATCTATAAAA	–	GGGGGGGGGGGGGGG	TGAATTGAGATTCTT	–
CHROMOSOME_I	33234	33566	332	GGGGGGGGGGGAGGGGGGGGGGGG	GATTTCTCGCAAGTA	GAAGGAGAAAAAATGT	–	GGGGGGGAGGGGGGG	GGGGGGTGAAAAAAG	–
CHROMOSOME_V	18246628	18247056	428	GGGTTGAAGGGGGGGGGGGGGGG	ATGCACAGAGCTGAG	AAATTGTCTGAAGTGC	–	GTTGAAGGGGGGGGG	GGGGGGCTTGATTTG	–
CHROMOSOME_X	16980971	16981410	439	GGGGGGGGGGGGGGG	GACGGGGCAACCGAA	TGAGAAGATATTCCGT	–	GGGGGGGGGGGTAAA	TGAAAAACTGAATGT	C
CHROMOSOME_V	20122343	20122807	464	GGGGGGGGGGGGGGAGGGG	ATTCGAACCCGCGAC	CTTTTGCTCCAATGTC	–	GGGGGGGAGGGGAGGA	GATGGAGAAGAAAAG	–
CHROMOSOME_V	5845189	5845705	516	GGGGGGGGGGGGGGGGGGGG	TATTTGTATATTTGC	AGGCTCAGCGCAGTCT	–	GTGGGGGGGGGGGGG	GGGGGGAGTTACTG	T
CHROMOSOME_IV	3291812	3292335	523	GGGATGGGGGGGGGGGGGGGGGGGGGGGGG	CCATTACGGTTCGAT	CTACAAAAAATGCGGG	–	GGGGGGGGGGGGGGG	GGAACCGATTCGTAT	–
CHROMOSOME_I	14087948	14088529	581	GGGGGGGGGGGGGGGGGGG	AGTAATTATAAATAT	TTTGTTAGTTGTAGAT	–	TTCGGGGGGGGGGGG	GGGGGGGTTGGATGC	–
CHROMOSOME_V	2975323	2975930	607	GGGGGGGGGGGGGGGGGGGG	TGCAAAGTGATTTTG	CCGTCGATTTTTGAAT	–	GGGGGGGGGGGGGGG	CAGTTTTTTGTGATC	AATTGCAAAGTGATTTTTTT
CHROMOSOME_X	10008372	10009025	653	GGGAGGGGGGGGGGGGGGGG	GTCTTAAGATCAAAT	GGTTAGTTGGCACTAT	–	CGGGAGGGGGGGGGG	GGGGGGGTCCGAGAA	–
CHROMOSOME_I	533079	533769	690	GGGGGGGGGGGG	TCGCCTGCCGGACCA	CGCGTGAGAAAGTCGT	–	AGTGGGGGGGGGGGG	GTGTTACGTGGGGGG	T
CHROMOSOME_IV	4473957	4474923	966	GGGCTGGGGGGGGGGGGGGG	GACACAATTTTCCGT	CCCCATTGGCCCAAAA	–	3GGCTGGGGGGGGGGG	GGGGGGAGTTGTGGG	–
CHROMOSOME_X	15025385	15026376	991	GGGAGGGGGGGGGGGGGGGGGGG	TGACACCACGTGAGT	TTGAAACTAGAACCTT	–	3GGGGGGGGGGGGGGG	GTGAAAAAGGAGTTG	–
CHROMOSOME_V	6022847	6025316	2469	GGGGGAGCGGGGGGGGGGG	TGGTTCTGAGCACAT	CGATGTGGCCATGATA	–	3GAGCGGGGGGGGGG	GGTCAAAACTTAAAG	–

Supplementary Table 3
G4 DNA-induced deletions in wild isolates of *C. elegans*

deletion found in: (N2 as reference genome, version WB225)

CB4856	CB4857	RC301	AB2	Chr	Start	End	Size	G4 motif	Left Flank	Deleted (left)	Deleted (right)	Right Flank	Insertion	G4 motif at 3' junction
✔				CHROMOSOME_III	4125895	4125923	28	GGGAAGGGATGGGGGGGGGGGGGGGG	TGGAAAAAAGGAATG	TCGTGGTGGGAAGG	GATGGGGGGGGGGG	GGGGGACAATAGGGA	–	Yes
✔				CHROMOSOME_I	12769955	12769990	35	GGGGGGGGGGGGGGGGGGGGG	CGGAAGAACATGTCT	AAGGAGAGAGGGGGGGG	GGGGGGGGGGGGGAGT	GGCGGTTCTTTAGTT	–	Yes
	✔	✔	✔	CHROMOSOME_II	12770069	12770143	74	GGGGGGGGGGGGGGGGGG	AAACGAAAAAATAGG	GCCCACCAATTAGCGTTTTGCT	GGGGGGGGGGGGGGGGGGCGG	AGTCGAAGACGCTGA	–	Yes
✔				CHROMOSOME_X	6756164	6756244	80	GGGGGGGGGGGGGGGGGGGGTCTTAGGG	GACAGAATATCATAA	CAAAACATTCCAAAATGTTTGA	TGGGGGGGGGGGGGGGGGGT	CTTAGGGAACTTTGG	–	Yes
✔				CHROMOSOME_II	4399129	4399222	93	GGGGGGGGGGGGGGGGGGG	CCCCAAGAACTATGC	CAGCCCGTGAGCTGTCTGCTGC	TTAGGGGGGGGGGGGGGGGGG	GCCTGAAACTGGATA	–	Yes
		✔		CHROMOSOME_X	16981316	16981410	94	GGGGGGGGGGGGGGGG	TTAGATAGGATAGGT	TTTATTTTGAAGATTACTGTAG	ATGATCTTGATGGGGGGGGGG	GGGGGGTAAATGAAA	AACTGAAAAACTAA	Yes
✔				CHROMOSOME_II	11660304	11660409	105	GGGGGGGGGGGGGGGGG	GAAAAAAATCAGAAT	GGCGTGAAACGCGAAAAAATAG	AAAAAATGTGCAAAACCACAAT	TTGCCAAAAGTTTTT	–	-
✔				CHROMOSOME_III	1866421	1866528	107	GGGGGGGGGGGGGGGG	GTTCCATCTCGAAAT	TTGAATTTCCCGGGGGGGGGGG	TTGCTTAATTCTCGTGGTTTAG	CCCAAAAAATGGGTT	–	-
		✔		CHROMOSOME_III	5940142	5940263	121	GGGGGGGGGGGGGG	AACGACGCACGTATT	ATTGAAGGCCACAAAGTAAAAA	GGGGGGGGGGGGGGTCGCTGACT	CAAACGGCACTCTTT	–	Yes
✔				CHROMOSOME_II	12770076	12770200	124	GGGGGGGGGGGGGGGGGGG	ATCTGCAAAAAAAAA	TATTTAAAAATTGGTTGAAATT	GTTGAGTGGGGGGGGGGGGGG	GGGGCGGAGTCGAAG	AA	Yes
✔				CHROMOSOME_X	12176926	12177053	127	GGGGGGTCGCGGGGGGGGGGGGGGG	CGAAAAAAAAACGTT	GGCAAGGTCATGGACATTGTTC	AAGGGGGGTCGCGGGGGGGGG	GGGGGCAACGTTCT	–	Yes
	✔			CHROMOSOME_X	15126882	15127013	131	GGGGGGGGGGAGGAGGGGGGGGG	TACAAAAACCGTCTC	TGCTGGTCTAATTCCAAAACGT	GAGGAGGGGGGGGGGGACCCGG	CCATTTTATTTTCCG	ATTTTATTTTCC	Yes
	✔	✔	✔	CHROMOSOME_I	55057	55188	131	GGGGGGGGGGGGGGG	ATGTGCGGGTGGATA	CCGTGGATATTCCACACAAAAA	GCCCCTTTTCGGGGGGGGGGG	GGGGCGTATTACGGG	–	Yes
✔				CHROMOSOME_III	6923340	6923485	145	GGGGGGGGGGGGGGGGGGG	TTGGTTAAGAGTTAG	TGGGGGGGGGGGGGGGGGGGG	TTAGAAGTTTTTCATTATTTTT	GCTTCCGGAATATTC	GTTAAGAGAATAGGCATTTGGTTAAGA	-
✔				CHROMOSOME_X	2187897	2188063	146	GGGGGGGGGGGGGG	AAAAAATTTTATAAC	CCACGTTCTTGAAATGTTATTT	TTTTTTGGGGGGGGGGGGGGG	TCAGGCGCCAATTTT	–	Yes
	✔	✔	✔	CHROMOSOME_II	13712305	13712451	146	GGGGGGGGGGGGGGAGGGG	CATCTGAAAAAATGA	GAGGGTTTTAGTGAGGGGTACG	AGGGGGGGGGGGGGGAGGGGT	ACTGTAGGGACACTG	–	Yes
			✔	CHROMOSOME_II	10784296	10784452	156	GGGGGGGGGGGGGGCGGG	CTTCATTTTTGTTGA	GACCCCACGCCACCTTGATGAC	ATGGGGGGGGGGGGGGGCGGGT	CACAATCGTGTGATG	–	Yes
		✔		CHROMOSOME_III	6138272	6138441	169	GGGGGGGGGGGGGG	CTCTGATACATGACG	GAATCTATGTAATAGGTTTTAC	GAAGGGGGGGGGGGGGGGTAAA	GTTCGAGTTGCCAAG	–	Yes
		✔		CHROMOSOME_III	1991720	1991901	181	GGGGGGGGGGGGGG	AATCCGCGCATCCCA	TGCGCCTTTAAACAATGTATCT	GAACAGGGGGGGGGGGGGGCT	TTCTAGTCTTTCTAG	–	Yes
	✔	✔	✔	CHROMOSOME_I	269409	269613	204	GGGGGGGGGGGGGGG	GTAGGAGTACTGTAG	AGGTAGTGTAGGAGCACTGTAG	GAGTACTGTAAGAACACTGTAG	GAATATTTTAAGAGT	–	-
✔				CHROMOSOME_V	19803064	19803423	359	GGGCTTCGGGCTTCGGGCTTCGGG	TTGAAAGCACTTTCC	CACCGAATTGTCTGCAATTCGA	GGTTCCCATATTCATTACACAG	ATATTATGCTACATA	AT	-
✔				CHROMOSOME_X	12062251	12062820	569	GGGGGGGGGGGGGGGGGG	CATATGTATCATTCT	AGAGATAGAGTAAAGAGAAATG	GGGGGGGGGGGGGGGGGGGAGT	GAATGAATGTGTCGG	–	Yes
		✔		CHROMOSOME_II	2753522	2754699	1177	GGGGGGGGGGGGGGG	AAATCGCAAAGGTCC	CCCCTGAGTCTGGCCACCTGAC	CCGAGCACTTTTCGATGCATTC	TTTGAAATTTTTGAC	–	-
✔				CHROMOSOME_X	3128769	3130544	1775	GGGGGGGGGGGGGGGG	AAACATTATAGAAAA	TACAGTGTGGGAAAGTTCTATA	CCTATAGAACTTTCCCACACTG	TATGTTGAGTTGCCC	–	-
✔				CHROMOSOME_V	15990570	15994311	3741	GGGGGGGGGGGGGGGGGG	TTTCGAAGAATCTTG	ACTGATCGCAAGATTGAAGCAA	TTCCATTTTGCGAAGAATCTTG	GCTGATCTTGCGGTT	–	-
			✔	CHROMOSOME_III	698698	705434	6736	GGGTTCGGGTCCGCGGGGTGGTGGG	TTTCAGTGTTCAAAC	ACAGTTAAATCGCGGGGTATAC	TCCGCGATTTAACTGTGTTAAC	TGCCAACCGTCTGCG	–	-



◀ An image is shown of a stained *C. elegans* embryo of approximately twenty minutes after fertilization. The blue staining marks the nucleus (containing DNA) of each cell of the embryo. The red staining indicates damaged DNA. Extensive DNA damage is observed in embryos that lack the G-quadruplex unwinding enzyme DOG-1 and the DNA repair enzyme polymerase Theta.

Cover Photo: Bluetube ▶

Chapter 5

G4 DNA instability in human cells

Wouter Koole[1], Jane T. van Heteren[1], Victor Guryev[2] and Marcel Tijsterman[1]

[1] Department of Toxicogenetics, Leiden University Medical Center, Leiden, The Netherlands
[2] Laboratory of Genome Structure and Ageing; European Research Institute for the Biology of Ageing; RuG and UMC Groningen, Groningen, The Netherlands

ABSTRACT

G4 DNA sequences, defined by tandem tracts of guanines, have the potential to adopt stable secondary structures that can perturb DNA replication. In *C. elegans*, G4 DNA can induce deletions, typically of 50-300 base pairs (bp), when animals are defective for *dog-1*/FancJ. Here, we investigate the mutagenicity of G4 DNA sequences in human cells. We assayed for G4 DNA-induced deletions at endogenous G4 motifs in DNA obtained from 25 FANCJ-proficient and -deficient cell populations clonally grown for 40 doublings, but no size alterations were observed. To investigate G4 DNA-instability with more sensitivity we developed various G4 DNA specific reporters and integrated these in HEK 293 cells, in a single-copy manner. Using fluorescent-based reporters that read out homology-directed repair between direct repeats, we found that G4 DNA increases the frequency of recombination, which can even be further increased by exposing cells to G4-stabilizing ligands. We also found that transcription across the G4 motif triggers recombination. Finally, using fluorescent-based reporters specific for reading out small deletion events, we found evidence for deletion induction similar to that found in *C. elegans*. In summary, we describe novel tools for investigating G4 DNA instability in human cells and provide the first examples of G4 DNA-induced genomic instability at the sequence level.

INTRODUCTION

During normal DNA replication, the replication machinery encounters many polymerase-blocking obstacles, such as base damages and single-strand breaks. Also intra- and interstrand crosslinks and double-strand breaks can block fork progression but are less common. Besides these DNA damages, also intrinsic features of DNA itself, such as the ability to adopt secondary structures, can hamper ongoing replication. If not dealt with in an accurate manner, such structures can cause unwanted mutagenic events. One example is a G-quadruplex structure (also known as G4 DNA), which is a stable secondary structure formed from single-stranded DNA by stacked planar arrays of four guanines, held together by Hoogsteen base pairing. These G-quadruplex structures, predicted to fold in ssDNA sequences complying to the consensus motif $G_{3-5}N_{1-7}G_{3-5}N_{1-7}G_{3-5}N_{1-7}G_{3-5}$, where G_{3-5} defines the number of guanines and N_{1-7} reflects the loop consisting of 1-7 basepairs that separates the runs of guanines (Burge et al., 2006; Huppert, 2010). The human genome contains ~400.000 of these motifs that have the potential to cause replication problems (Alan K Todd, 2005; Huppert and Balasubramanian, 2005). The high prevalence of G4 DNA sequences in the genome, while being a potential threat for normal replication progression, has led to many speculations about a possible biological function of G-quadruplexes. Recent studies have provided both supporting evidence for the existence of G-quadruplexes *in vivo* (Biffi et al., 2013; Henderson et al., 2013), and have as well provided strong arguments that G-quadruplexes are involved in a variety of biological processes. In *Neisseria gonorrhoeae*, for example, G4 DNA formation is required to induce DNA recombination of the pilin locus to promote antigenic variation: single base pair mutations in guanines that disrupt G4 DNA formation destroys pilin antigenic variation (Cahoon and Seifert, 2009). Other studies have implicated G4 structures in the regulation of transcription (Eddy and Maizels, 2006; 2008), in defining origins of replication (Besnard et al., 2012), and in the protection and elongation of telomeres (Oganesian et al., 2006; Smith et al., 2011; Zhang et al., 2010).

To safeguard these biological functions and maintain DNA integrity, it is vital that G4 motifs are correctly replicated. Despite the existence of several helicases that have been implicated in protecting genomes from loss of G4-motif-containing sequences, genomic instability has been linked to G4 DNA. A recent study shows that genome alterations in cancer cells are frequently found near G4 motifs, suggesting that this motif may underlie these genomic changes (De and Michor, 2011). Another study showed increased levels of γH2AX foci (an indicator of DNA damage) in cells exposed to the G4 DNA stabilizer pyridostatin. Specificity of this drug was demonstrated by showing an enrichment of G4 motifs in chromatin immunoprecipitation experiments (Rodriguez et al., 2012).

In this study we aim to investigate G4 DNA instability in human cells. Previous studies in *C. elegans* revealed that the existence of G4 sequences is warranted by the action of the helicase *dog-1* (the homologue of human FANCJ) (Youds et al.,

2008). In *dog-1* mutants G4 motifs trigger the formation of small deletions which are typically 50-300 base pairs in size, including the G4 motif and flanking 5′ upstream DNA (Cheung et al., 2002; Kruisselbrink et al., 2008). Besides their distinct size and unidirectionality, these G4 DNA-induced deletions are further characterised by single-nucleotide homology at the junctions and the occasional presence of templated insertions. We recently found that G4-induced deletions require the activity of the A-family polymerase Theta (θ), in a process we named Theta-Mediated End Joining (TMEJ) (Koole et al., 2014). It is however currently unknown how human cells deal with the ~ 400.000 G4 DNA motifs and which repair pathways operate to protect the genome against G4 DNA induced mutagenesis.

RESULTS

No evidence for G4 DNA instability in clonally propagated FANCJ-defective cells using PCR analysis of endogenous G4 loci

Studies in *C. elegans* have previously identified deletions, typically 50-300 bp long, occuring at endogenous G4 motifs in animals deficient in *dog-1*/FANCJ (Kruisselbrink et al., 2008; Pontier et al., 2009). In one of the assays PCR-technology is employed to amplify a locus that contains a G4 motif, and the resulting fragments (usually designed to be 1000-1500 bp long) are resolved on agarose gels. Molecules that, by deletion, have lost the G4 motif (and flanking DNA) are preferentially amplified and can be visualized as a smaller-than-wild type fragment. We used this assay to test whether deletions will be induced at endogenous G4 motifs in FANCJ-deficient human cells. To this end, we obtained two cell lines: immortalized fibroblasts (EUFA030/hTERT) that were derived from a FANCJ-defective patient and a subclone from this line that was complemented with cDNA of FANCJ (Levitus et al., 2005). We cultured 25 clonally-derived subpopulations of both cell lines for 40 doublings. Next, we generated cultures from a single cell and grew these for another 15 cell doublings before isolating DNA (schematically illustrated in Fig. 1a). For molecular analysis of these DNA samples, we designed 96 primer sets to amplify an equal number of endogenous loci, all containing a monotract of guanines (which are strong inducers of G4 deletions in *C. elegans*). Amplicons of 1000-1500 bp including the G4 motif were chosen. We subsequently analysed the DNA from the propagated subclones as well as the DNA from two control samples that were retrieved from anonymous healthy persons. Using this assay, we found no evidence for deletion induction: no smaller-than-wild type products were found in the propagated FANCJ-deficient cells. For one locus (documented as ggg137), we found a smaller amplicon in both FANCJ-deficient and FANCJ-corrected cells, as compared to control DNA (Fig. 1b). Sanger sequencing revealed that 47 bp were deleted in the cells that originally were derived from a FANCJ patient. This deletion included the G4 motif (Fig. 1c). We next tested DNA obtained from the patient's skin fibroblasts and lymphoblasts, as well as DNA from lymphoblasts of the patient's parents and

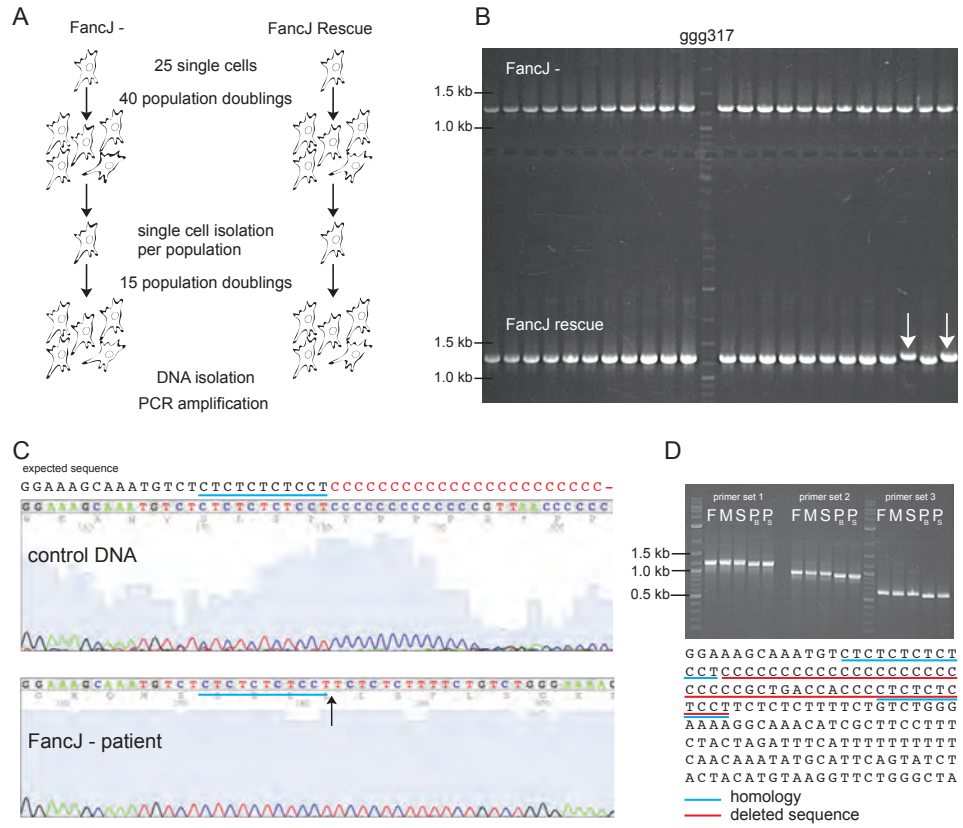


Figure 1 | PCR-based approach does not show *de novo* G4 DNA-induced deletions in FANCJ-deficient cells. (A) Flowchart of culturing and of DNA isolation of single cell-derived cultures of FANCJ-deficient and complemented cells. (B) Image of PCR-products resolved on a 2% agarose gel. Each lane shows a PCR-product amplified at locus ggg137 from DNA of an independently grown culture. Arrows indicate control DNA samples from healthy anonymous persons. (C) Sanger sequence results of PCR-products of control DNA and FANCJ-deficient cells at locus ggg137. Expected sequence (according reference genome, version hg17) is shown on top. The arrow indicates the position of the start of the deletion. The blue line underlines nucleotides that share homology with nucleotides 47 base pairs downstream (as shown in Fig. 1d) (D) Top panel shows image of PCR-products amplified at locus ggg137 using three different sets of primers and DNA samples obtained from the FANCJ-patient's blood (P_b) and skin (P_s), his father (F), mother (M), and sibling (S). Bottom panel shows sequence of locus ggg137. The red line denotes the nucleotides that were deleted in the FANCJ-patient. The blue lines mark the homology found at the junction of the deletion.

sibling to determine whether this deletion happened during somatic cell growth within the patient. While we initially validated the presence of the deletion in DNA samples of the patient and not in DNA samples of his family (supporting a *de novo* G4 DNA-induced deletion at that locus within the patient; Fig. 1d), we were not able to confirm this result unambiguously.

G4 sequences are highly polymorphic

To verify whether all selected amplicons that were tested above were similar to the reference genome, and thus contained G4 motifs we Sanger sequenced the PCR products obtained from the FANCJ patient cells and from Human Embryonic Kidney 293 cells (HEK 293). Although monotracts of guanines are difficult to replicate and sequence *in vitro*, we found G4 motifs at all tested loci. Interestingly, as exemplified in Fig. 2a and 2b, in many cases where we were able to reliably read through the complete G4-motif, we noted the presence of several sequence alterations as compared to the human reference genome (version hg17, release 2004). Moreover, we sometimes found two different sequences in one DNA sample, which together, and given the small sample set, argues that G4 sequences are highly polymorphic in the human population. We systematically tested this assumption further for non-monotract G4 motifs (matching the G4 consensus $G_{3-5}N_{1-7}G_{3-5}N_{1-7}G_{3-5}N_{1-7}G_{3-5}$), also to avoid the intrinsic difficulties of sequencing guanine monotracts, which could corrupt molecular analysis. We obtained reliable sequence-reads for 27 different loci and found that only 8 of these had the identical G4 motif compared to the reference genome (an example is shown in Fig. 2c). For the other 19 loci, we found various polymorphisms varying from single-nucleotide to multi-nucleotide polymorphisms (examples are shown in Fig. 2 d-f). These data provide evidence that G4 motifs are highly polymorphic.

Development of fluorescent-based reporters that read out G4 DNA instability

The PCR-based approach has only very limited sensitivity for identifying *de novo* G4 DNA-induced deletions. We thus aimed to develop other methods that are able to measure G4 DNA instability via different means. Previously, we have assayed frameshift mutations at microsatellites in human cells using reporters based on fluoresence (Koole et al., 2013). Fluorescent Activated Cell Sorting (FACS) of genetically engineered cell lines enabled us to reliably measure frameshift mutations in 10^6 to 10^7 cells, making it a sensitive and robust method to measure infrequent mutagenic events. In addition, we have previously shown the use of G4 DNA-specific reporters to read out G4 DNA instability in worms (Kruisselbrink et al., 2008): a G4 motif placed between two repeated sequences proved to be a substrate for homology-driven DNA repair pathways (*e.g.* single strand annealing (SSA) or homologous recombination (HR)). These reporters were robust indicators of G4 DNA instability in worms. We combined these previous strategies to develop recombination reporters aimed to read out G4 DNA instability in human cells by fluorescence. Reporters were cloned such that functional fluorescent mCherry could be expressed only after a repair-event that used the repeated sequences as donors (schematically illustrated in Fig. 3a.). In between the repeated sequences we placed a G4 motif (G26- or C26-repeat), or for control purposes, a non-G4 motif, as illustrated in Fig. 3a. Immediately downstream of the G4 motif we also cloned a recognition site for the rare-cutting endonuclease I-SceI, to be able to validate our reporter system: cleavage of this site by I-SceI leads to a double-strand break and provides a substrate for recombination

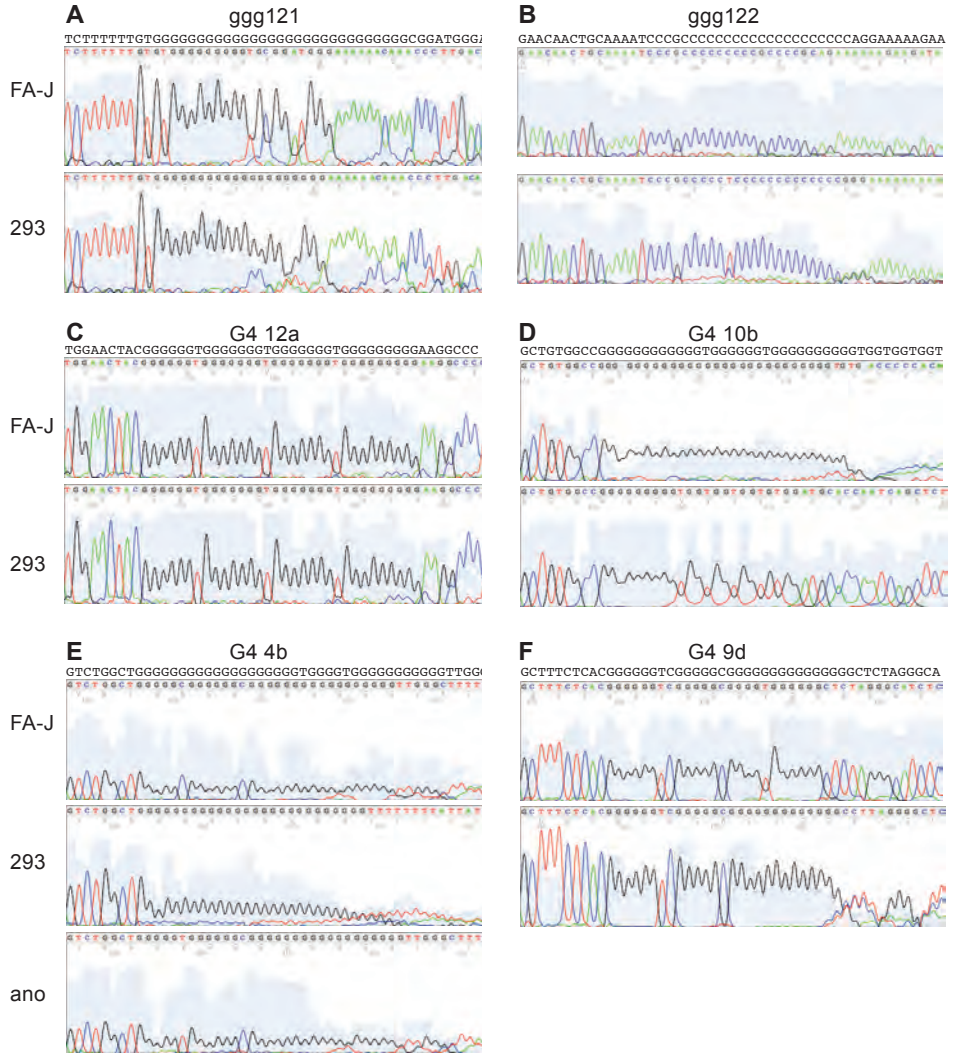


Figure 2 | G4 motifs are polymorphic sequences. Images of Sanger sequencing results are shown of various G4 motifs. DNA was used from a FANCJ-patient (FA-J), HEK 293 cells (293) and a healthy anonymous person (ano). Expected sequence (according to reference genome hg17) is shown on top per locus. Loci containing a guanine-monotract are indicated with "ggg", while non-monotract G4 sequences are indicated with "G4". See methods section for more details (*i.e.* genomic location and primers) about the specific loci.

pathways like SSA or HR to act upon. With these reporters, we generated stable cell lines of HEK 293 cells. Importantly, to ensure the same genomic environment and to avoid copy-number variation, we targeted the reporters in a single-copy manner to the same genomic location via Flp-FRT-mediated recombineering, as previously described (Koole et al., 2013).

To test possible functionality of the experimental system we first confirmed that the presence of a double-strand break in between the two repeated sequences led to detectable recombination events: expression of I-SceI resulted into a profound increase of mCherry-expressing cells within 48 hours after induction (Supplementary Fig. S1a) and recombination between the repeated sequences was confirmed by PCR and Sanger sequencing. Second, we showed that the expression level of mCherry was sufficiently high to reliably distinguish and quantify mCherry-positive and -negative cells by flow cytometry (Fig. 3b).

Next, we tested whether the presence of a G4 motif in between the two repeated sequences influenced the number of homology-directed repair events. We seeded

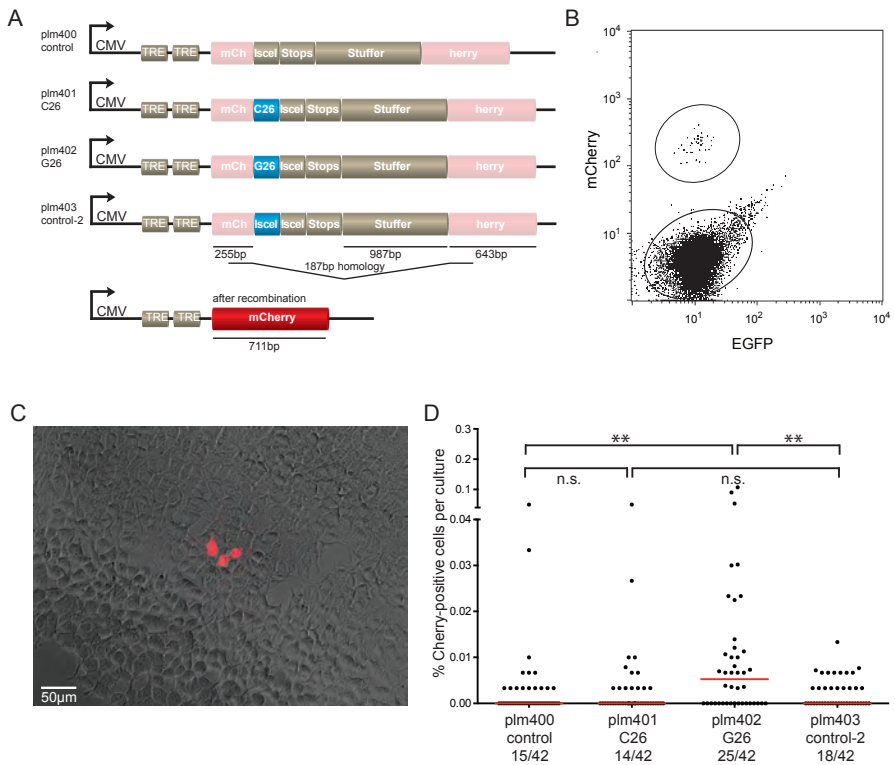


Figure 3 | G4 DNA-specific fluorescent reporters show increased recombination events at G4 sequences. (A) Schematic representation of G4 DNA specific (plm401 and plm402) and control reporters (plm400 and pl403). Only after recombination between the two homologous sequences is expression of functional mCherry possible. TRE, Tet Responsive Element. Stops, termination codons in all three reading frames. (B) Representative FACS-plot of a plm402(G_{26})-cell culture. mCherry-positive cells can be clearly distinguished and quantified. (C) Representative image of a recombination event in plm402(G_{26})-cells. This image represents an overlay of pictures taken in the red- and bright field channel. (D) Quantification of mCherry-positive cells in independent cell cultures of respective cell lines. The fraction of wells in which mCherry-positive cells were found is indicated. Red bars indicate the median. n.s., not significant. ** P<0.01 (Mann-Whitney test).

mCherry-negative cells in wells and measured for each cell line 42 independent cultures for the presence of mCherry-positive cells by flow cytometry after 7-10 days of culture (see Methods for further experimental detail). In all four different cell lines, we found that a fraction of the cultures contained clonal regions of mCherry-positive cells (an example is shown in Fig. 3c), indicating that recombination between the repeated sequences also occurred without the presence of a G4 motif or an I-SceI-induced double-strand break. However, we found that cells containing plm402 (G_{26}-repeat, hereafter named plm402(G_{26})-cells), showed significantly (P<0.01, Mann- Whitney test) more independent recombination-events compared to control cell lines containing reporters lacking a G4 motif (plm400 and plm403, hereafter named plm400($Ctrl_1$)-cells and plm403($Ctrl_2$)-cells, respectively) (Fig. 3d). This relatively small increase in mutation-frequency (1.7-fold, see Supplementary Fig. S1b for its calculation) proved consistent in other experiments (Fig. 4a and b, and data not shown). Remarkably, cell lines containing reporter plm401 (C_{26}-repeat, hereafter named plm401(C_{26})-cells), did not show elevated levels of recombination-events compared to the control cell lines, suggesting that the orientation of the G4 motif can be of influence for the stability of G4 sequences in the genome. Taken together, using fluorescent-based recombination reporters, we show that G4 sequences can lead to increased genomic instability, and their stability can be influenced by the orientation of the G4 sequence.

G4 DNA-stabilizing ligands cause increased G4 DNA instability

In recent years, various G-quadruplex-stabilizing ligands have been developed. Although it has been shown that most ligands stabilize G4 DNA effectively *in vitro*, minimal experimental data is available about their stabilizing effect in human cells. We questioned whether we could test the effect of these ligands on G4 DNA stability in human cells using our reporter-system. To this end, we cultured mCherry-negative cells in the presence and absence of the G4-stabilizing compound Phen-DC_6 (De Cian et al., 2007). Interestingly, we found an increased number of recombination events in plm401(C_{26})- and plm402(G_{26})-cells when cultured in the presence of Phen-DC_6 (Fig. 4a,b). Importantly, control cells (plm400($Ctrl_1$)- and plm403($Ctrl_2$)-cells) did not show significant elevated levels of recombination events in the presence of Phen-DC_6 (Fig. 4a,b), suggesting specificity of the ligand for G4 sequences *in vivo*. As shown in Fig. 4b, similar data were obtained when cells were grown in the presence of Phen-DC_3, a different G4 DNA-stabilizing compound but chemically closely related to Phen-DC_6 (De Cian et al., 2007). Together, these data show that G4 DNA-stabilizing compounds can cause increased G4 DNA instability in human cells.

Role of transcription on G4 DNA instability

G4 motifs have found to be enriched at transcription start sites, 5′ UTRs and 5′ ends of the first intron of genes (Du et al., 2009; Eddy and Maizels, 2008; 2009; Eddy et al., 2011; Huppert and Balasubramanian, 2007; Maizels and Gray, 2013). Yet, transcription across these sites may pose a threat for the stability of the host gene;

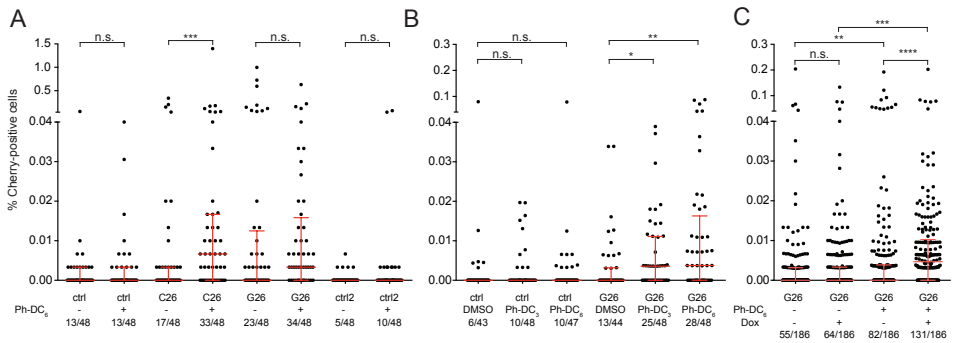


Figure 4 | G4 DNA-stabilizing compounds and transcription increase G4 DNA instability. (A,B) Quantification of the percentage mCherry-positive cells in the presence or absence of the G4-stabilizing compounds Phen-DC_6 and Phen-DC_3. The fraction of wells in which mCherry-positive cells were found, is indicated. (C) Quantification of the percentage mCherry-positive cells in the presence or absence of transcription (+ or - doxycycline (dox), respectively). (A-C) n.s., not significant. *P<0.05, ** P<0.01, *** P<0.001, **** P<0.001 (Mann-Whitney test). Red bars indicate the median and interquartile range.

during transcription a small segment of DNA transiently becomes single-stranded which can provide the opportunity for a motif to adopt a G4 quadruplex structure, which in turn can form a blocking lesion for both RNA and DNA polymerases. Furthermore, transcription over G-rich sequences can potentially lead to the formation of DNA:RNA hybrids, named R-loops (reviewed by (Aguilera and García-Muse, 2012) (Kim and Jinks-Robertson, 2012)) and to the formation of recently described DNA:RNA hybrid G-quadruplexes (called HQ) (Zheng et al., 2013). We investigated to which extent ongoing transcription influenced G4 DNA instability. Our reporters contained two Tet-Responsive Elements (TRE) at their CMV-promoter (Fig. 3a), which allowed us to control transcription over the transgene by culturing the cells in the absence or presence of doxycycline (Koole et al., 2013), a regulator of the Tet repressor (Yao et al., 1998). We cultured mCherry-negative plm402(G_{26})-cells in the presence or absence of doxycycline (with or without transcription, respectively) for 8 days. To make use of mCherry as a read out for FACS analysis, doxycycline was added to the non-transcribed conditions two days before quantification (see methods for more details). We found no significant difference in recombination frequencies for cells cultured in the presence or absence of doxycycline (Fig. 4c). However, when cells were cultured in the presence of the G4 ligand Phen-DC_6, transcription elevated the recombination frequency (P<0.001, Mann-Whitney test, Fig. 4c), an effect that may point towards transient formation of mutagenic G4 structures in transcribed DNA that can be stabilized by Phen-$D6_6$.

G4 DNA-induced deletion reporter

In *C. elegans* unresolved G4 structures can trigger the formation of deletions, typically 50-300 bp in size. We have recently shown that these deletions are formed

by Theta-mediated end joining (TMEJ) in *C. elegans* (Koole et al., 2014), however the question remains whether TMEJ also acts downstream of replication-blocking G4 structures in higher eukaryotes. Fuelled by the notion that I) we established a method and conditions to detect G4 DNA instability in HEK 293 cells using recombination reporters and II) we detected the presence of relatively high mRNA levels of polymerase Theta in HEK 293 cells (Supplementary Fig. S2), we questioned whether we could find TMEJ-like deletions in HEK 293 cells using differently designed fluorescent reporters. Therefore, we designed reporters such that the fluorescent marker EGFP would only be expressed after small deletions of ~5-700 basepairs. We cloned termination codons in all three reading frames downstream of the coding sequence of mCherry, followed by a G_{23}-repeat, a non-selective Open Reading Frame (ORF) of 216 basepairs and the EGFP ORF (illustrated as reporter plm299 in Fig. 5a). Both the sequences of mCherry and the non-selective ORF served as a buffer for possible deletions. We reasoned that only in the event of an in-frame deletion that takes away the termination-codon(s), EGFP is expressed, which results in cells that express mCherry and EGFP (mCherry+EGFP+ cells), or, when also a part of the coding region of mCherry is deleted, in mCherry-EGFP+ cells (illustrated in Fig. 5a). A similar reporter (plm310) was made in which the G_{23}-repeat was replaced by a C_{26}-repeat. We obtained stable cell lines containing reporters plm299 or plm310 (hereafter named plm299(G_{23})-cells and plm310(C_{26})-cells, respectively). We cultured cells in the presence of Phen-DC_6 and inspected cells for EGFP-expression by eye and by automated microscopy. Interestingly, we found in each cell line one clonal group of cells that expressed EGFP (an image for plm310 is shown in Fig. 5b). We isolated the cells by flow cytometry and found that the EGFP+ cells containing plm310(C_{26}) did not express mCherry (Fig. 5b, right panels). PCR-analysis and Sanger sequencing confirmed that EGFP-expression in these cells was the result of a 202 basepair in-frame deletion that included the G4 motif, the termination codons and part of the coding sequence of mCherry (Fig. 5c,d), explaining the loss of mCherry expression. In contrast, EGFP+ cells, containing reporter plm299(G_{23}), still expressed mCherry (data not shown). A smaller deletion (46 bp, including the G4 motif and termination-codon) was found in these mCherry+EGFP+ plm299(G_{23})-cells (Fig. 5c,e). Interestingly, similar to TMEJ-events described in *C. elegans*, both identified deletions showed single-nucleotide homology at the junction, suggesting that these events may have been the result of polymerase Theta activity.

To further investigate whether the identified deletions were G4 DNA-dependent, we designed various reporters with other G4 motifs and with corrupted motifs containing the same G-C-ratio but unable to form a secondary structure (sequences are shown in Supplementary Fig. 3). We obtained for each construct a polyclonal cell line, cultured these cells in 96 independent wells in the presence or absence of Phen-DC_6 and inspected wells for EGFP+ cells by eye and by automated microscopy. We found the majority of EGFP+ cells in cell lines carrying reporters with an intact G4 motif, although we also found EGFP+ cells using reporters with a corrupted G4 motif (Supplementary Fig. 3). Further experiments are required to determine whether

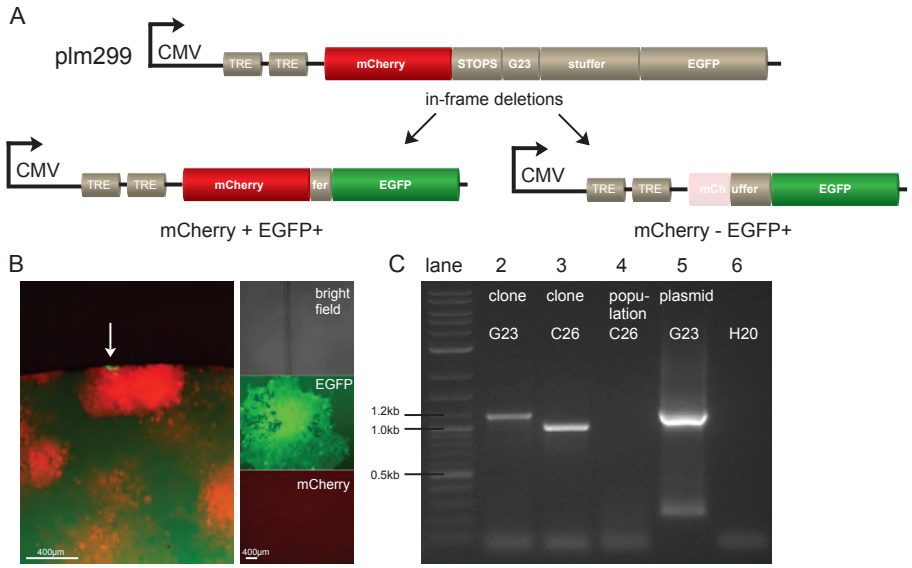


Figure 5 | Fluorescent reporter detects typical G4 DNA-induced deletion events. (A) Schematic representation of fluorescent reporter plm299, which can express EGFP only after a small deletion (<720 bp) removes the termination codon(s) (stops) are deleted and restores the reading frame of EGFP. The mCherry sequence and stuffer region serve as buffer for deletions. In-frame deletions can lead to mCherry+EGFP+ cells (bottom left panel), or, in cases where the deletion has taken away coding nucleotides of mCherry, mCherry-EGFP+ cells (bottom right panel). Plm310 is similar to plm299 but contains a C_{26}-repeat instead of a G_{23}-repeat. (B) Left panel, image of EGFP-expressing plm299(C_{26})-cells (indicated with a white arrow) surrounded by mCherry+EGFP- cells. EGFP+ cells were isolated by FACS and single-cell-derived colonies were grown. Right panels are pictures of such a mCherry-EGFP+ plm299(C_{26}) single-cell-derived colony, taken in bright field, green and red channel (top to bottom, respectively). (C) Agarose gel showing PCR products obtained by amplification of DNA samples using primers that flank the G4 motif. A clear lower PCR product was obtained from DNA samples of mCherry+EGFP- plm310(G_{23})-cells (lane 2) and mCherry-EGFP+ plm299(C_{26})-cells (lane 3), compared to control plasmid DNA (lane 5). No band was observed when using DNA from a population mCherry+EGFP- plm299(C_{26})-cells (lane 4), indicating the difficulty of replicating long G/C monotracts. (D,E) Representation of deletions found in mCherry+EGFP- plm310(G_{23})-cells and mCherry-EGFP+ plm299(C_{26})-cells. Sequences including G4 motif and flanking DNA are shown of plm299 (D) and plm310 (E). Deleted nucleotides are depicted in blue. Coding sequences are underlined with various colours, as described in the figure.

these EGFP+ cells are the result of a G4 DNA-induced deletion and polymerase Theta activity. Taken together, these data show that the reporters we designed successfully identify *de novo* deletions in human cells and thus pave the way to investigate the molecular mechanisms and genetic requirements of G4- induced mutagenesis in human cells.

Discussion

Here, we show that specific G4 DNA-instability reporters enabled the detection of *de novo* deletion events in FANCJ-proficient HEK 293 cells at G4 motifs, whereas a PCR-based approach in the FANCJ-deficient cell line UFA030/hTERT did not. We consider several explanations for not observing *de novo* deletion events using a PCR-based approach. First, the PCR-based approach may not have been sufficiently sensitive; we analysed ±100 G4 motifs in 2x25 independent DNA samples obtained from cells that were cultured for at least 40 cell doublings. We thus investigated ~ $2x10^5$ *in vivo* replicated G4 motifs. Using the fluorescent reporter system we can easily analyse up to ±$1x10^7$ replicated genomes.

Second, the cell line used might not be completely deficient in FancJ: western blot analysis on lysates of UFA030/hTERT cells, revealed a product around the expected height of FANCJ (Supplementary Fig. 4b) when using an antibody raised against the C-terminus of FANCJ. Interestingly, we did not see a product when using a different FANCJ-specific antibody that was raised against the N-terminus of FANCJ (Supplementary Fig. 4c). In addition, we noted the disappearance of this product when EUFA030/hTERT cells were transduced with a lentivirus expressing a FANCJ-specific shRNA (Supplementary Fig. 4d). These data suggest that the EUFA030/hTERT cell line may contain low levels of a splice-variant of FANCJ that lacks the N-terminus, but may be sufficient to unwind unresolved G-quadruplexes. A third possible explanation is that the cells we used do not express functional polymerase Theta (under the assumption that G4 DNA-induced deletions in human cells are polymerase Theta-dependent). Although we did observe similar mRNA levels of polymerase Theta as found in polymerase Theta-proficient HeLa cells (Supplementary Fig. 2), we were unable to confirm the presence of polymerase Theta on the protein level, due to the absence of a functional antibody.

So far, we have been able to obtain the footprints of only two deletion events, using the fluorescent reporters plm299 and plm310. Although these data show that our approach is successful to find and investigate deletion events, many more events must be analysed to reach any firm conclusions about the molecular mechanisms behind the formation of these deletion events in human cells. Preliminary data point towards increased mutagenesis when reporters contain G4 motifs (*e.g.* mono G tracts) that are highly unstable in *C. elegans* (Supplementary Figure 3), but future research is essential to investigate the extent and nature of G4 DNA-induced genome alteration in human cells.

Another question that remains unanswered is whether deletion-formation is increased in human cells upon complete depletion of FANCJ. We have attempted to knockdown FANCJ by transfection of siRNAs and lentiviral transduction of shRNAs, however, both approaches did not give sufficient knockdown of FANCJ at the protein level (Supplementary Fig. 4b,c and data not shown). Transfection of our fluorescent reporters into EUFA030/hTERT cells was unsuccessful because of the cells' low transfection efficiency, random integration (making it difficult to compare control reporters with reporters containing G4 motifs) and background EGFP expression in EUFA030 cells (these cells were immortalized with an hTERT-EGFP fusion construct), which interfered with EGFP as read-out for deletion events. However, the recently developed CRISPR/Cas9-targeted mutagenesis technology in mammalian cells (Cho et al., 2013; Cong et al., 2013; Mali et al., 2013) opens up new ways to investigate the role of FANCJ in resolving G-quadruplexes in human cells. This technique also provides an excellent opportunity to address the question whether the type of deletions that were obtained depend on polymerase Theta activity.

To further increase the number of small deletion events and reduce the labour and time required to retrieve these specific events, slight adaptations of the current constructs can be made. The constructs described here depend on expression of EGFP, which requires labour intensive inspection by eye, by automated microscopy, and/or by flow cytometry to isolate single EGFP+ cells, and time (±2 weeks) to grow up the isolated single EGFP+ cells. This process can be accelerated when, for example, downstream of EGFP a viral 2A peptide is cloned (to allow bicistronic expression) (Szymczak et al., 2004), followed by the coding sequence of the puromycin N-acetyl-transferase gene. Using these constructs, cells will become EGFP-positive upon an in-frame deletion event, and importantly, also resistant to the antibiotic puromycin. This way inspection for EGFP+ cells followed by single cell isolation becomes unnecessary, which saves labour and time and is therefor favourable when high numbers of deletion events are desired.

ACKNOWLEDGEMENTS

We would like to thank Johan de Winter and Jurgen Steltenpool for providing the FANCJ-deficient cell lines and additional DNA samples. Maartje van Kregten for comments on this manuscript. Furthermore we would like to thank A. Nicolas, M.P. Teulade-Fichou and C. Guetta for providing the compounds Phen-DC_6 and Phen-DC_3.

MATERIAL AND METHODS

Cell culture. FANCJ-deficient (EUFA30/hTert) fibroblasts (Levitus et al., 2005) were cultured in DMEM supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin at 37°C and 5% $C0_2$. FANC-J complemented cells (with cDNA of FANCJ using pIRESneo-vector) were cultured under similar conditions but in the presence with 5µg/ml G418. Flp-in T-Rex-293 cells (Life Technologies) were cultured and genetically modified as described earlier (Koole et al., 2013). To quantify recombination events in cells containing plm400-403, mCherry-negative cells were selected by flow cytometry and 1500 cells were seeded per well of a 96-well plate. When wells were confluent, 20,000 cells per well were analysed for expression of mCherry by flow cytometry. To test the effect of G4 ligands, cells were cultured in the presence of 2.5µM Phen-DC_6/Phen-DC_3 (kindly provided by M.P. Teulade-Fichou, C. Guetta and A. Nicolas) or control (DMSO). Fresh medium (including G4 ligands) was added after 4 days. To test the effect of transcription on G4-induced recombination events, cells were transferred to doxycycline-free medium containing charcoal stripped fetal bovine serum (ref A15-119, PAA laboratories) and grown for 3 days. Next, mCherry-negative cells were selected by flow cytometry, grown as one population for 3 days and split (1500 cells/well) in 96-well plates and cultured with or without the presence of 0.1µg/ml doxycycline hyclate and 2.5µM Phen-DC_6. When wells were nearly confluent (±6 days after last split), doxycycline was added to all plates and wells were analysed by flow cytometry after two days.

Cells containing plm299, plm310 and plm410-plm420 were grown in the presence or absence of 2.5µM Phen-DC_6 and inspected for EGFP-positive cells by eye and using a BD Pathway imager (BD Biosciences). EGFP-positive cells were selected manually and DNA was isolated using a DNA-isolation kit (Qiagen).

FANCJ knockdown and Western blot analysis. FANCJ-protein knockdown was performed via lentiviral transduction of the following shRNAs according to manufacturer's protocol (Sigma): sh1-3 are respectively TRCN0000049915 - TRCN0000049915 from the Mission Library shRNA library (Sigma), shC = vector SHC002. Transient transfection of siRNAs was performed using DharmaFECT1. The following siRNAs were used: ON-TARGET*plus* Non-targeting pool (D-001810-10-20), siGENOME BRIP1 (M-010587-00), ON-TARGET*plus* SMARTpool human BRIP1 (L-010587-00) (Thermo Scientific Dharmacon). Plm117 was made by cloning FANCJ-specific oligos gatcccGTACAGTACCCCACCTTATttcaagagaATAAGGTGGGGTACTGTACtttttggaaa and agcttttccaaaaaGTACAGTACCCCACCTTATtctcttgaaATAAGGTGGGGTACTGTACgg in backbone pTER+ (kindly provided by Marc van de Wetering). The following antibodies raised against human FANCJ were used for Western blot analysis: polyclonal anti-BACH1 produced in rabbit (Sigma, B1310, lot# 014K4843, dilution: 1:20,000) and monoclonal anti-hBRIP1/FANCJ (Roche, MAB6496, Clone 652747, lot CEFX0110091, dilution: 1:5,000)

PCR analysis on endogenous G4 DNA sites. First, DNA was obtained from clonally-derived FANC-J-deficient (and complemented) cells; per cell line, 5 9 cm dishes were seeded with ±100 single cells each and cells were cultured until single cell-derived colonies were observed. Second, 40 independent colonies per cell line were picked manually and transferred to a 96-well plate. Cells were cultured and passaged for in total 40 population doublings. Next, in a similar manner, a single-cell-derived colony was obtained from each independent population and grown for another total 15 population doublings, followed by DNA isolation using NaCl/EtOH precipitation. At the end of the experiment, we were able to retrieve DNA from 25 independent colonies per cell line.

A set of primers was designed covering 100 endogenous genomic loci containing a monotract of guanines, and generating amplicons of 1000-1500 bp (list of primer sequences is available upon request). Using these primers, we PCR-amplified, using homemade Taq polymerase and standard PCR conditions, 50ng of DNA per reaction for all DNA samples and two control DNA samples (of anonymous persons). PCR products were run on a 2% agarose gel. A set of nested primers was designed covering 100 loci containing a G4 DNA sequence according to the G4 consensus $G_{3-5}N_{1-7}G_{3-5}N_{1-7}G_{3-5}N_{1-7}G_{3-5}$ (list of primer sequences is available upon request). This set of primers was used to perform PCR-reactions on DNA obtained from the following cell lines: EUFA30/

hTert, EUFA30/hTert complemented with FANCJ cDNA, 293 HEK, and HELA. Also DNA from an anonymous person was used. PCR-reactions were performed using Amplitaq Gold Master Mix (LifeTechnologies) according to the manufacturer's protocol. PCR-products were resolved on 1% agarose gels and sequenced by Sanger sequencing. Below, a list with primers is provided for the G4 DNA-containing loci that were described in this manuscript in more detail:

Locus	Chr.	position (hg17)	Forward Primer	Reverse primer
ggg137	x	147636436	GAATTTGGAAATATGCCGC	GGGGCTCCCTTCAAAGAGAGC
			GCAGGAATAGATCCTTCAGC	TGTTTGCATCAAGAAGCACTG
			CTGGAGTCAGTCCTGTTCTGC	GGGGCTCCCTTCAAAGAGAGC
ggg121	7	105335835	GCCTGAGTATCCTCCCAAC	AATTACCACTGCTTGACACAG
ggg122	8	103868873	AGCACAAGCCACATTCTAAG	GAGTTATGCAGCCTCAAGAC
G4 12a	18	46819848	ACCAGGACTCTTGACTTTGC	TTGAGGTTCTTCAAACTGAGG
G4 10d	14	59355886	CACATCCTGCTGATTGGTC	GAGGTGTGGAGGGAGAGG
G4 4b	2	153904330	GCTTTGAGACACCAGAAACC	TCCTGGTGGAATCTGAGG
G4 9d	12	63563933	CCAGTGACTCAAACCTCTCC	AACAAGTATGCTCCCAGCAG

Plasmid construction and sequencing. Standard molecular cloning techniques were used to obtain the constructs described in this manuscript. Briefly, for plasmids plm400-plm404, we PCR-amplified 3 DNA fragments: a fragment containing the first 255 bp mCherry (from plasmid pRSET-B mCherry) flanked by NheI and HindIII-KpnI restriction-sites, a stuffer fragment of 987 bp (retrieved from the *C. elegans unc-22* gene) flanked by a *Kpn*I and a *Bam*HI restriction site, a fragment containing the last 643 bp of mCherry flanked by *Bam*HI and *Eco*RI restriction sites. Pieces were sequentially cloned into plasmid pcDNA5/FRT/TO (Life Technologies). I-SceI, G4 DNA sequences and termination-codons were subsequently placed into the HindIII and KpnI restriction site via cloning of DNA oligos. For plasmids plm299, plm310 and plm410-420, we modified plm188 (described in (Koole et al., 2013)) via cloning of DNA oligos. All plasmids were checked by Sanger sequencing according to standard procedures, but with the following adjustments: we used a 1:3 mix of ABI Prism dGTP BigDye terminator v3.0 and BigDye terminator v3.0, 200ng plasmid per sequencing reaction, and no more than 25 cycles.

Flow cytometry. Cells were sorted with the BD FACSAria III flow cytometer and analysed with the Guava easyCyte HT flow cytometer (Millipore), and using their respective software.

REFERENCES

Aguilera, A., and García-Muse, T. (2012). R loops: from transcription byproducts to threats to genome stability. Molecular Cell *46*, 115–124.

Alan K Todd, M.J.S.N. (2005). Highly prevalent putative quadruplex sequence motifs in human DNA. Nucleic Acids Res. *33*, 2901.

Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J.-M., and Lemaitre, J.-M. (2012). Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nature Structural & Molecular Biology *19*, 837–844.

Biffi, G., Tannahill, D., McCafferty, J., and Balasubramanian, S. (2013). Quantitative visualization of DNA G-quadruplex structures in human cells. Nature Chemistry *5*, 182–186.

Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K., and Neidle, S. (2006). Quadruplex DNA: sequence, topology and structure. Nucleic Acids Res. *34*, 5402–5415.

Cahoon, L.A., and Seifert, H.S. (2009). An alternative DNA structure is necessary for pilin antigenic variation in Neisseria gonorrhoeae. Science *325*, 764–767.

Cheung, I., Schertzer, M., Rose, A., and Lansdorp, P.M. (2002). Disruption of dog-1 in Caenorhabditis elegans triggers deletions upstream of guanine-rich DNA. Nat. Genet. *31*, 405–409.

Cho, S.W., Kim, S., Kim, J.M., and Kim, J.-S. (2013). Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat. Biotechnol. *31*, 230–232.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. Science *339*, 819–823.

De Cian, A., Delemos, E., Mergny, J.-L., Teulade-Fichou, M.-P., and Monchaud, D. (2007). Highly efficient G-quadruplex recognition by bisquinolinium compounds. J. Am. Chem. Soc. *129*, 1856–1857.

De, S., and Michor, F. (2011). DNA secondary structures and epigenetic determinants of cancer genome evolution. Nature Structural & Molecular Biology *18*, 950–955.

Du, Z., Zhao, Y., and Li, N. (2009). Genome-wide colonization of gene regulatory elements by G4 DNA motifs. Nucleic Acids Res. *37*, 6784–6798.

Eddy, J., and Maizels, N. (2006). Gene function correlates with potential for G4 DNA formation in the human genome. Nucleic Acids Res. *34*, 3887–3896.

Eddy, J., and Maizels, N. (2008). Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. Nucleic Acids Res. *36*, 1321–1333.

Eddy, J., and Maizels, N. (2009). Selection for the G4 DNA motif at the 5' end of human genes. Mol. Carcinog. *48*, 319–325.

Eddy, J., Vallur, A.C., Varma, S., Liu, H., Reinhold, W.C., Pommier, Y., and Maizels, N. (2011). G4 motifs correlate with promoter-proximal transcriptional pausing in human genes. Nucleic Acids Res. *39*, 4975–4983.

Henderson, A., Wu, Y., Huang, Y.C., Chavez, E.A., Platt, J., Johnson, F.B., Brosh, R.M., Sen, D., and Lansdorp, P.M. (2013). Detection of G-quadruplex DNA in mammalian cells. Nucleic Acids Res. *42*, 860–869.

Huppert, J.L. (2010). Structure, location and interactions of G-quadruplexes. Febs J. *277*, 3452–3458.

Huppert, J.L., and Balasubramanian, S. (2005). Prevalence of quadruplexes in the human genome. Nucleic Acids Res. *33*, 2908–2916.

Huppert, J.L., and Balasubramanian, S. (2007). G-quadruplexes in promoters throughout the human genome. Nucleic Acids Res. *35*, 406–413.

Kim, N., and Jinks-Robertson, S. (2012). Transcription as a source of genome instability. Nat. Rev. Genet. *13*, 204–214.

Koole, W., Schäfer, H.S., Agami, R., van Haaften, G., and Tijsterman, M. (2013). A versatile microsatellite instability reporter system in human cells. Nucleic Acids Res. *41*, e158.

Koole, W., van Schendel, R., Karambelas, A.E., van Heteren, J.T., Okihara, K.L., and Tijsterman, M. (2014). A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. Nature Communications *5*, 3216.

Kruisselbrink, E., Guryev, V., Brouwer, K., Pontier, D.B., Cuppen, E., and Tijsterman, M. (2008). Mutagenic capacity of endogenous G4 DNA underlies genome instability in FANCJ-defective C. elegans. Curr. Biol. *18*, 900–905.

Levitus, M., Waisfisz, Q., Godthelp, B.C., de Vries, Y., Hussain, S., Wiegant, W.W., Elghalbzouri-Maghrani, E., Steltenpool, J., Rooimans, M.A., Pals, G., et al. (2005). The DNA helicase BRIP1 is defective in Fanconi anemia complementation group J. Nat. Genet. *37*, 934–935.

Maizels, N., and Gray, L.T. (2013). The G4 Genome. PLoS Genet. *9*, e1003468.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-guided human genome engineering via Cas9. Science *339*, 823–826.

Oganesian, L., Moon, I.K., Bryan, T.M., and Jarstfer, M.B. (2006). Extension of G-quadruplex DNA by ciliate telomerase. Embo J *25*, 1148–1159.

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

Rodriguez, R., Miller, K.M., Forment, J.V., Bradshaw, C.R., Nikan, M., Britton, S., Oelschlaegel, T., Xhemalce, B., Balasubramanian, S., and Jackson, S.P. (2012). Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. Nat. Chem. Biol. *8*, 301–310.

Smith, J.S., Chen, Q., Yatsunyk, L.A., Nicoludis, J.M., Garcia, M.S., Kranaster, R., Balasubramanian, S., Monchaud, D., Teulade-Fichou, M.-P., Abramowitz, L., et al. (2011). Rudimentary G-quadruplex-based telomere capping in Saccharomyces cerevisiae. Nature Structural & Molecular Biology *18*, 478–485.

Szymczak, A.L., Workman, C.J., Wang, Y., Vignali, K.M., Dilioglou, S., Vanin, E.F., and Vignali, D.A.A. (2004). Correction of multi-gene deficiency in vivo using a single "self-cleaving" 2A peptide-based retroviral vector. Nat. Biotechnol. *22*, 589–594.

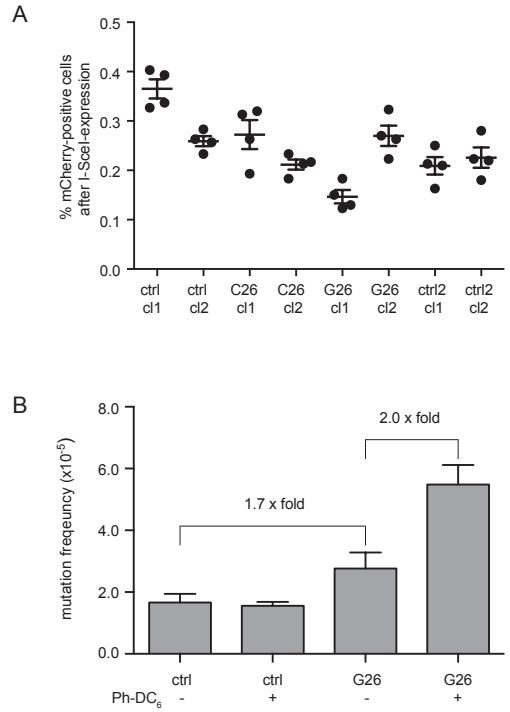
Yao, F., Svensjö, T., Winkler, T., Lu, M., Eriksson, C., and Eriksson, E. (1998). Tetracycline repressor, tetR, rather than the tetR-mammalian cell transcription factor fusion derivatives, regulates inducible gene expression in mammalian cells. Hum. Gene Ther. *9*, 1939–1950.

Youds, J.L., Barber, L.J., Ward, J.D., Collis, S.J., O'Neil, N.J., Boulton, S.J., and Rose, A.M. (2008). DOG-1 is the Caenorhabditis elegans BRIP1/FANCJ homologue and functions in interstrand cross-link repair. Mol. Cell. Biol. *28*, 1470–1479.

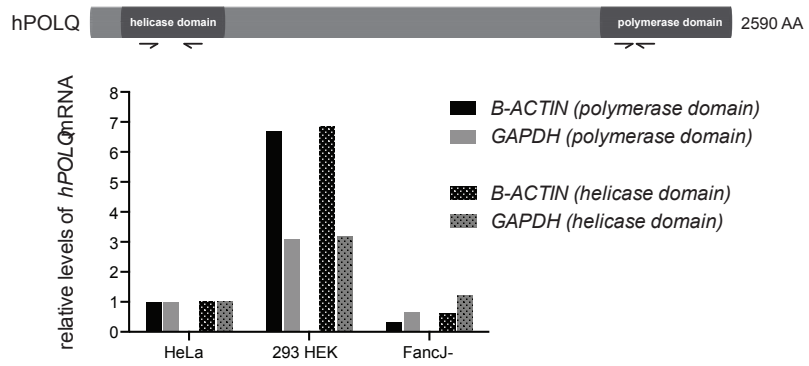
Zhang, M.-L., Tong, X.-J., Fu, X.-H., Zhou, B.O., Wang, J., Liao, X.-H., Li, Q.-J., Shen, N., Ding, J., and Zhou, J.-Q. (2010). Yeast telomerase subunit Est1p has guanine quadruplex-promoting activity that is required for telomere elongation. Nature Structural & Molecular Biology *17*, 202–209.

Zheng, K.-W., Xiao, S., Liu, J.-Q., Zhang, J.-Y., Hao, Y.-H., and Tan, Z. (2013). Co-transcriptional formation of DNA:RNA hybrid G-quadruplex and potential function as constitutional cis element for transcription control. Nucleic Acids Res. *41*, 5533–5541.

SUPPLEMENTARY FIGURES



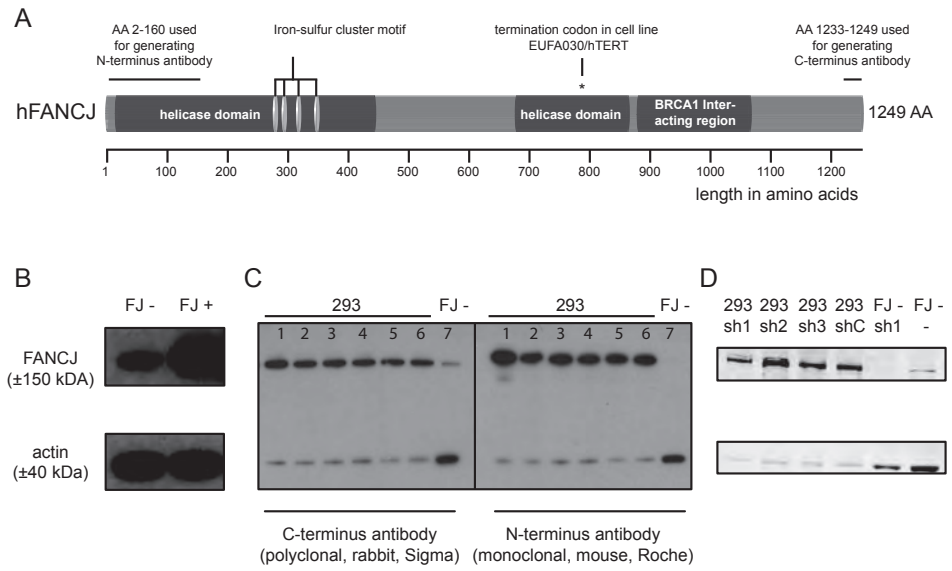
Supplementary Figure S1 | Frequency determination of recombination events using G4 DNA instability reporters. (A) Percentage of mCherry-positive cells measured 48 hours after IsceI-expression. For each construct (plm400-403), two independent clones were picked and 4 independent populations assayed. (B) Determined mutation frequency in plm400($Ctrl_1$)-cells and plm400(G_{26})-cells in absence or presence of Phen-DC_6. The frequency was calculated as follows: Assuming the Poisson distribution of recombination events, the fraction (P_0) of the wells analysed in which none mCherry-positive cell were found, was used to estimate the average number (m) of mutation events ($m = -\ln(P_0)$). Then, the estimated mutation frequency was obtained as m/n, where n = the number of cells analysed per well (±20.000 cells). Error bars denote s.e.m. $n \geq 4$.



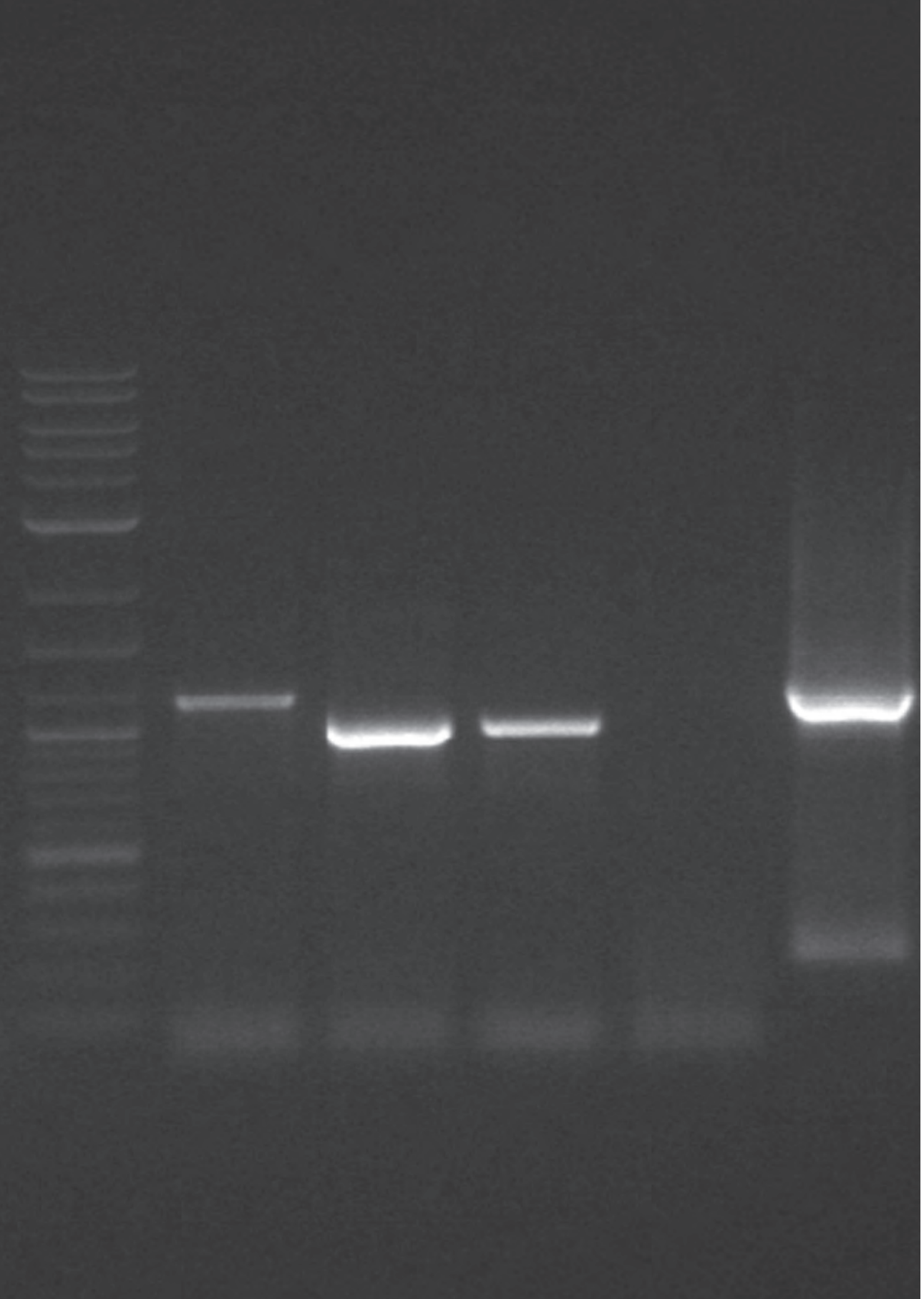
Supplementary Figure S2 | hPOLQ mRNA expression in HeLa, 293 HEK and FA-J cell lines. *hPOLQ* mRNA levels were measured using quantitative RT-PCR in indicated cell lines. Two sets of primers were used for *hPOLQ*, located at the helicase domain and the polymerase domain. B-actin and GAPDH were taken along as reference genes. Data was normalized to the ratio *hPOLQ/B-ACTIN* or *hPOLQ/GAPDH* mRNA levels observed in HeLa cells, since it has been shown that HeLa cells contain functional hPOLQ protein (shown by Higgins et al 2012, and data not shown).

	DMSO	Phen-DC_6
plm410 stops GGGGGAGGGGGAGGGGGAGGGGG	1	0
plm411 CCCCCTCCCCCTCCCCCTCCCCC stops	0	0
plm412 CCCCCTGGGGGTCCCCCTCCCCC stops	0	1
plm414 CCCTCCCACCCTCCC stops	3	0
plm415 CGCTCGCACGCTCGC stops	1	0
plm418 stops GGGGGGGGGGGGGGGGGGGGGGG	1	0
plm419 stops CCCCCCCCCCCCCCCCCCCCCCC	0	1
plm420 CCCCCCCCCCCCCCCCCCCCCCC stops	5	7

Supplementary Figure S3 | G4 DNA instability reporters specific for reading out G4 DNA-induced deletion events. Additional G4 DNA instability reporters that have been constructed to identify G4 DNA-induced deletion events. All reporters have a similar build up as schematized in the above two constructs, but vary in their G4 motif and the position of the stop-codons (stops) as indicated. plm412 and plm415 comprise of non-G4 motifs but contain the same G/C-content as plm411 and plm414, respectively. Constructs were integrated (single-copy) in HEK 293 cells and resulting stable cell lines were each grown in 96 wells with or without the presence of 2.5 μm Phen-DC_6. Cells were inspected for EGFP-expression and the number of wells in which ≥1 EGFP-positive cells is found, is indicated.



Supplementary Figure S4 | FANCJ-protein expression in EUFA030/hTERT cell lines and HEK 293 cells. (A) Schematic representation of the human FANCJ protein. (B) Western blot of lysates from EUFA030/hTERT (FJ-) and EUFA030/hTERT FANCJ complemented cells. The presence of a band running at the predicted height of FANCJ, suggests low levels of expression of FANCJ in FJ- cells. For this blot an antibody was used that was raised against the C-terminus of hFANCJ. (C) Western blot of HEK 293 (293) and FJ- lysates probed with an N-terminus and a C-terminus specific antibody. The C-terminus specific antibody detects the presence of a product in FJ- cells, which is not detected by a N-terminus specific antibody, suggesting expression of an alternatively spliced FANCJ-product which lacks the N-terminus. For both blots the exact same lysates were used and gels were run at the same time under identical conditions. HEK 293 lysates were obtained from cells that were transfected with various siRNA (Dharmacon) or shRNAs and loaded as follows: Lane 1= control siRNA, Lane 2= FANCJ siRNA , Lane 3+4= FANCJ siRNA OT+, Lane 5+6= FANCJ shRNA (plm117). (D) Western blot of lysates from HEK 293 and FJ- cells that stably express FANCJ-specific shRNAs. No profound reduction of FANCJ-protein levels was seen in HEK 293 cells. However, in EUFA030/hTERT (FJ-) cells a reduced band was seen of the presumably alternatively spliced FANCJ-product. For this blot the FANCJ C-terminus specific antibody was used. shRNAs obtained from the Mission Library (Sigma) were introduced via lentiviral transduction. shC= scrambled shRNA, sh1-3 are FANCJ-specific shRNAs. See material and methods for further details.



◀ This photo depicts an agarose gel in which in each lane amplified DNA molecules were run to detect the size of the molecules. A fluorescent stain was used to visualize the DNA molecules. Using this method, I was able to visualize that in human cells sometimes a small part of the DNA is lost due to the presence of a G-quadruplex sequence.

Chapter 6

Summarizing discussion & future perspectives

Cover Photo: Red, Green and Blue ▶

SUMMARIZING DISCUSSION & FUTURE PERSPECTIVES

In this thesis I investigate the stability of microsatellites and G-quadruplexes. Using various model organisms and newly developed tools, I characterize to which extent microsatellites and G-quadruplexes are maintained in the genome, and which genes, repair pathways and factors (such as tract length and composition) are of importance for their stability.

Using a microsatellite reporter system based on fluorescence and optimized for mammalian cells, I demonstrate in **Chapter 2** that microsatellite instability is greatly influenced by the tract length and the nucleotide composition of the microsatellite. Other factors, such as its strand orientation or its transcriptional status, appeared not to affect the stability of a microsatellite. Furthermore, I show that the MSI-reporter system can be used as a functional tool to screen for other modifiers of MSI and as proof of principle, I make use of compounds and a miRNA expression library. Next, I demonstrate in **Chapter 3** that inducing somatic mosaicism through MSI can be a powerful approach for mosaic analysis and tumor induction *in vivo*.

In **Chapter 4** I focused on G-quadruplex-induced mutagenesis in *C. elegans* and demonstrate that the mutagenicity of G-quadruplexes depends on the nucleotide composition, strand orientation and the presence of the helicase DOG-1/FANCJ. Importantly, I show that polymerase Theta plays a crucial role in the repair of G-quadruplex-induced DNA damage. Polymerase Theta-Mediated End-Joining (TMEJ) of G-quadruplex-induced DNA damage results in deletions characterized by an extremely narrow size distribution between 50-300bp, one nucleotide homology at the junctions, and occasionally insertions of up to 20 base pairs templated from the flanking sequences. In the absence of polymerase Theta, G-quadruplexes can lead to large deletions spanning several kilobases. In **Chapter 5** I present data that G-quadruplexes can also lead to genomic instability in human cells. I provide preliminary evidence that G-quadruplex sequences can be highly polymorphic between humans and G-quadruplex-dependent genomic instability is increased by transcription and stabilizing ligands. Finally, I provide preliminary data that G-quadruplexes can also lead to small genomic deletions in human cells.

Microsatellite and G4 DNA instability reporters

In this thesis a substantial repertoire of reporters is described that read out microsatellite and G4 DNA instability. The majority of the reporters are designed such that genomic events (such as frame-shifts, deletions or recombinations between direct repeats) result in a functional gene encoding a fluorescent protein. This allows for the easy detection of genomic alterations by eye (using a fluorescence microscope) and by FACS. When using FACS, the number of genomic events can be determined in a quick and reliable unbiased manner, and approximately 10^6-10^7 events (fluorescent cells) per hour can be measured. However, in the course of our experiments we noted that FACS is less preferred to detect events when the mutation frequency is lower than $\pm 10^{-6}$ per

cell division (which was the case for the detection of G4 DNA-induced deletions in mammalian cells) mainly for practical reasons. For example, inspection by eye and automated microscopy allowed us to analyze ± 16 plates (containing 96 wells each) in a day while FACS only allowed ± 6 plates. When using FACS, rate-limiting factors are the capacity of the FACS-machine, together with the required time for single-cell preparation of the cells. Furthermore, in our experience, isolating a single positive cell had a much lower (estimated 25% versus 90%) success rate in obtaining growing clones for DNA analysis than manually picking positive cells by eye, and, obviously, this low success rate is not favored when mutation-events are rare. As discussed in Chapter 5, an attractive alternative approach for detecting and isolating cells after a rare genomic event would be the generation of a reporter that allows selection on both fluorescence and a fast-selecting antibiotic such as puromycin.

In Chapter 2 we show that our newly developed microsatellite instability reporter can be a valuable and functional tool to identify novel modifiers of MSI. As proof of principle, we made use of a retroviral miRNA overexpression library and identified miR-21 as a modifier of MSI. Recently, various genome-wide libraries have become available that allow for systematically knocking out genes in human cells by making use of CRISPR-Cas9 technology (http://www.addgene.org/CRISPR/libraries/). It would be of great interest to test these libraries, or at least sub-libraries, on our ready-to-go cell lines containing microsatellite and G4 DNA instability reporters. Additionally, the CRISPR-Cas9 technology now facilitates a way to test for redundancy and it would be worthwhile trying to knock out two genes at the same time in our established cell lines. For example, would the combined knockout of FANCJ and PIF1 or any other helicase lead to increased G4 DNA instability?

In Chapter 3 we show that MSI can be used to stochastically express practically any kind of protein *in vivo*. Using various reporters we demonstrate MSI-induced stochastic expression of proteins such as EGFP, GAL4, LacZ and oncogenic HRAS. An attractive other candidate for MSI-induced stochastic expression *in vivo* would be the nuclease Cas9. Creating a zebrafish line that stochastically expresses Cas9 will likely generate knockout cells in a mosaic fashion when combined with the presence of a functional guide-RNA (the guide-RNA can be provided, for instance, upon injection at the one-cell stage). This can be a powerful approach to create mosaic knockout zebrafish.

Theta-Mediated End-Joining

An important finding described in this thesis is the critical role of polymerase Theta in the repair of G4 DNA-induced damage. Theta-mediated end-joining of G-quadruplex-induced breaks results in deletions characterized by several features: i) a narrow size distribution; the vast majority of deletions are between 50 and 300 bp in size ii) templated insertions from flanks iii) a predominance of minimal homology of exactly one nucleotide at the junction iv) the deletion starts in close proximity to the 3′ site of the G4 motif.

Currently, it remains unknown what defines the narrow size distribution of G4 DNA-induced deletions. Although various models can be envisioned, I will describe two models that are, to my opinion, the most plausible. The first model advocates that the distinct size is determined by the formation of Okazaki fragments (Fig. 1a, left and middle panel). Data from a recent study in yeast indicate that the distance between a replication block and a 5′ upstream Okazaki-fragment will leave a ssDNA gap of approximately 100 - 500 bp, resembling the size of G4 DNA-induced deletions (Smith and Whitehouse, 2012). A next round of replication of a ssDNA gap leads to a DSB-intermediate, and upon polymerase Theta-dependent repair, a deletion is formed having the size of the initial ssDNA gap (see bottom panel in figure 1). Since Okazaki-fragments are formed during lagging-strand synthesis, it is tempting to speculate that the G4 DNA-induced deletions are caused during lagging- strand synthesis. However, recent reports indicate that G4 sequences snap into a stable secondary structure during leading strand synthesis (Lopes et al., 2011; Schiavone et al., 2014). Nonetheless, this observation does not exclude that Okazaki fragments can still determine the size of deletions; when during leading strand synthesis the replisome is stalled or collapsed at a G4 structure, a convergent or activated dormant replication fork can encounter from the other direction (see middle panel of Fig. 1). In this way the replication of the initial leading strand will be finalized by lagging strand synthesis from a converging fork, leaving a ssDNA gap 5′ upstream of the G4 structure. The second model proposes a repriming mechanism; when leading-strand synthesis is stalled by a G-quadruplex, repriming and synthesis may take place on the template strand, leaving a ssDNA gap 5′ upstream of the G-quadruplex (Fig. 1, right panel). Analogous to the first model, upon a next round of replication the ssDNA gap will lead to a DSB-intermediate, which is then repaired by polymerase Theta. Evidence for a repriming mechanism is supported by studies using bacteria and yeast (Heller and Marians, 2006; Lopes et al., 2006; Yeeles and Marians, 2011). Moreover, the recent discovery of a second primase, named PrimPol, indicates that a repriming mechanism is also present in higher eukaryotes (García-Gómez et al., 2013; Mourón et al., 2013). Noteworthy, Lopes *et al* visualized by electron microscopy the formation of ssDNA gaps in the leading strand after exposing yeast to UV-irradiation (Lopes et al., 2006). Intriguingly, the majority of the formed ssDNA gaps were less than 400bp in size, reminiscent of the size of G4 DNA-induced deletions. Future studies are required to show whether one or both of these models can explain the size of typical G4 DNA-induced deletions. One of such studies would be the assessment of the size of G4 DNA-induced deletions in the absence of a functional CAF-1 complex (Chromatin Assembly Factor 1), a multi-subunit histone chaperone complex that is involved in the assembly of histones on nascent DNA: a study by Smith and Whitehouse showed that the size of Okazaki fragments is coupled to chromatin assembly (Smith and Whitehouse, 2012). They found an average increase in the size of Okazaki-fragments upon deletion of subunits of the CAF-1 complex. An increase in the size of G4 DNA-induced deletions in CAF-1 mutant worms would argue that their size is determined by Okazaki-fragment formation.

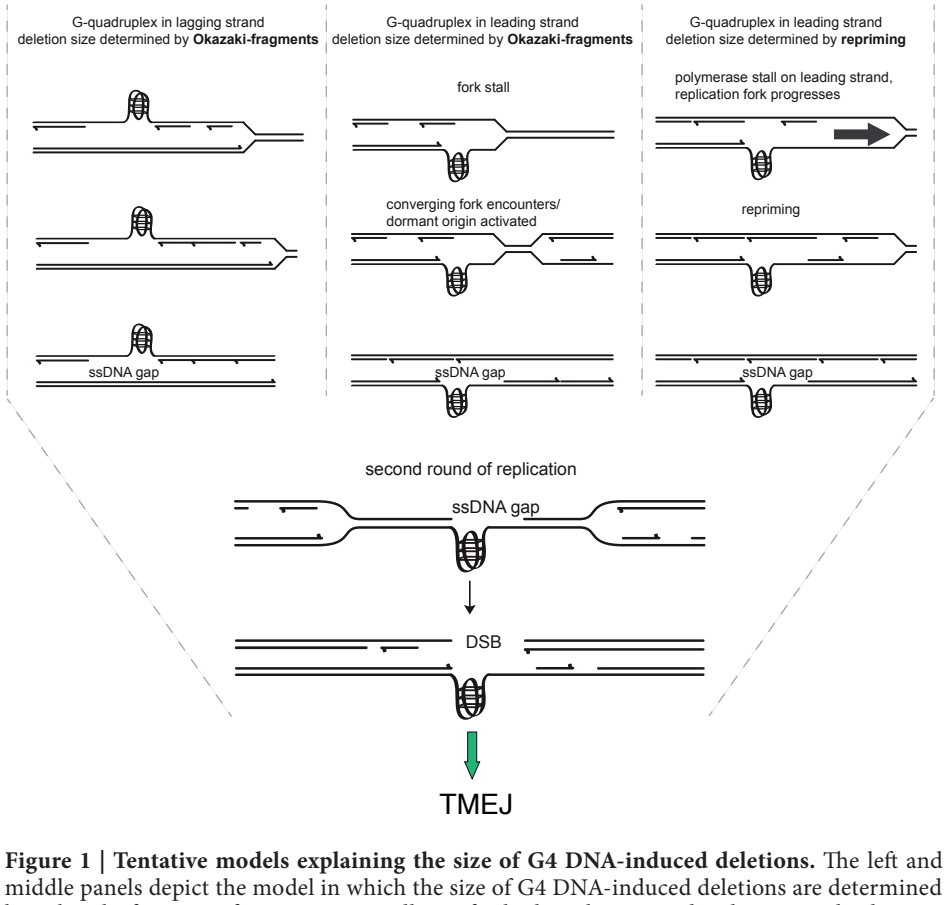


Figure 1 | Tentative models explaining the size of G4 DNA-induced deletions. The left and middle panels depict the model in which the size of G4 DNA-induced deletions are determined by Okazaki-fragment formation, regardless of whether the G-quadruplex is in the lagging or leading strand. In the right panel a model is shown in which a ssDNA gap is formed via a repriming mechanism. The bottom panel illustrates how a second round of replication of an unresolved ssDNA gap can result in a DSB-intermediate, which serves as a substrate for TMEJ. Note that the size of the ssDNA gap generated in the initial replication stalling event dictates the size of the G4 DNA-induced deletion.

In a fraction of the G4 DNA-induced deletions, we found the presence of templated insertions. The origin of the insertions can be traced back to the sequence immediately flanking the junction of the deletion. Remarkably, we found some cases where the same flank was used twice as a template, or even cases where both flanks served as a template (see for example Chapter 4, Fig. 2d). These observations suggest that initiated DNA synthesis by polymerase Theta was suddenly aborted, followed by re-annealing and a second (or sometimes third and fourth) attempt for extension of the 3´ hydroxyl end. Why synthesis by polymerase Theta is occasionally stopped so abruptly remains puzzling. Perhaps polymerase Theta encounters in these cases a protein or protein-complex that inhibits further synthesis. The Ku70/80 complex

could potentially form such a block, however, analysis of G4 DNA-induced deletions in *dog-1 cku-80*-deficient worms, did not diminish templated insertions (data not shown). Another interesting candidate would be the heterotrimeric complex RPA. Strikingly, a recent study in yeast indicates that the RPA-complex plays a crucial role in the suppression of MMEJ (Deng et al., 2014). Besides a protein block, also the collision with a RNA primer (of an upstream Okazaki-fragment), dsDNA at the transition-point from ssDNA into dsDNA, a hairpin or any other secondary structure in the template strand could stop the synthesis of polymerase Theta and cause dissociation of the newly synthesized strand followed by iterative rounds of repriming, extension and dissociation.

In this thesis I show that TMEJ is involved in the repair of G4 DNA-induced damage. An unresolved question is whether TMEJ is also involved in the repair of DNA damage induced by other type of lesions. A recent study in our lab shows that this is indeed the case; TMEJ-dependent deletions were found in pol eta- and pol kappa-deficient worms, arguing that damaged bases (such as 8-oxo-guanines) that are normally substrates for these TLS-polymerases can cause TMEJ-dependent deletions (Roerink et al., 2014). A different study in our lab (Roerink et al, unpublished) shows that TMEJ also plays a critical role in the repair of transposon-induced breaks, and similar results were observed in *Drosophila melanogaster* (Chan et al., 2010). Future studies will elucidate whether TMEJ may be involved in the repair of other types of damage such as UV-, IR- or CRISPR-Cas9-induced DNA damage.

TMEJ: creation, expansion and deletion of microsatellites

In chapter 2 of this thesis I have described several factors that contribute to the length variation of a microsatellite. However, little is known about the emergence and disappearance of microsatellites (Kelkar et al., 2011). Although there is ample evidence that base substitutions can cause the “birth” and “death” of microsatellites (Kelkar et al., 2011; Messier et al., 1996), and although it is well accepted that the presence of a fraction of the microsatellites can be explained by random chance, other mechanisms must be present to explain the high abundance of microsatellites found in organism’s genomes. Here, I propose the model that TMEJ can account for the creation, expansion and deletion of microsatellites. In support of this proposal, we present in Chapter 4 many examples where microsatellites (e.g. a G_{23}-repeat) are deleted due to TMEJ. Note that, similar to these G-quadruplex-forming microsatellites, many other microsatellites are prone to form a secondary structure, which may result in their disappearance via TMEJ. Strikingly, we also found various examples where a new microsatellite was born via a templated insertion (see examples 1-3 in Figure. 2). Remarkably, we found several events in which a monotract of adenosines/thymines was formed, of which in one case a monotract was born of even 18 nucleotides (example 3 in Figure 2). In all of these cases the flanks surrounding the deletion contained a small tract of 3-4 adenosine/thymines suggesting that

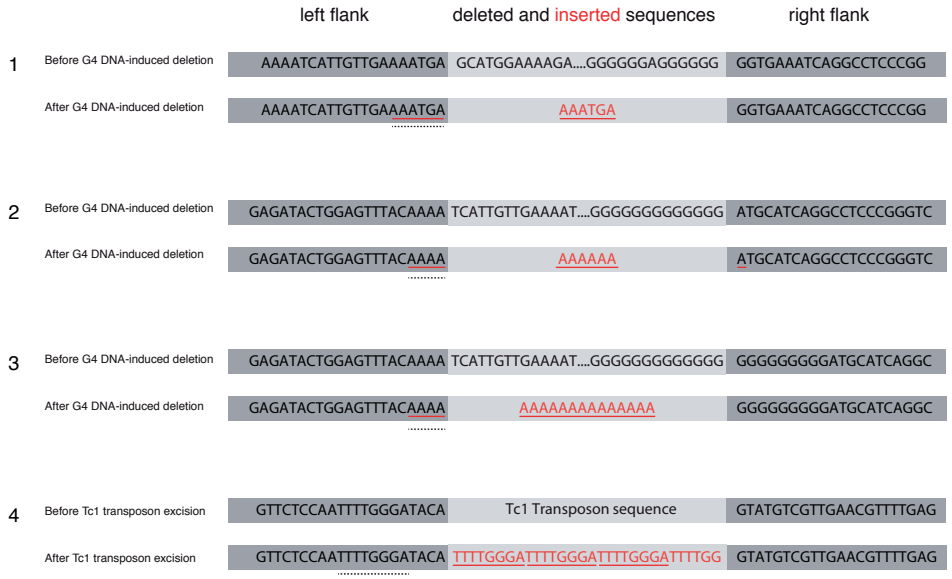


Figure 2 | Examples of microsatellites that are born via TMEJ. The first three examples present the repair-products of G-quadruplex-induced deletions (data from Chapter 4). Example 4 presents the repair-product of a transposon-induced DSB (Roerink *et al*, unpublished). Inserted nucleotides are depicted in red. The born microsatellites are underlined with a red line, and the proposed template used for the insertion is underlined with a dashed black line. All footprints represent germline mutations found in *C. elegans*.

the formation of the monotract could have been the result of templated synthesis. However, a study by Hogg *et al* shows that polymerase Theta is also able to extend 3′ termini in a template-independent manner(Hogg et al., 2012). Therefore, we cannot rule out that the formation of these A/T monotracts can be ascribed to template-independent synthesis in which the incorporation of nucleotides is restricted to adenosines/thymines. Yet, such preference/restriction for incorporation of solely adenosines/thymines by polymerase Theta during template independent synthesis has not been seen *in vitro* (Hogg et al., 2011).

Although we here show examples in which the birth of a microsatellites is accompanied with the loss of a G4 sequence and flanking DNA, one can reason that TMEJ-dependent repair of other DSB-intermediates (e.g. induced by IR) may lead to solely the birth of a microsatellite without the loss of (flanking) DNA. The creation of TMEJ-dependent microsatellites is further substantiated by a study from our lab in which we show the birth of microsatellites after TMEJ-dependent repair of transposon-induced DSBs (an example is given in Figure 1, Roerink et al, unpublished). This example nicely illustrates how a microsatellite of 3 units (each unit consisting of 8 nucleotides) is born after only a single DSB.

Besides the involvement in the birth and deletion of microsatellites, it can be hypothesized that TMEJ contributes in a similar manner to the expansion of a microsatellite when the microsatellite is in the direct vicinity of a DSB and is used as a template. Could this for example be the molecular mechanism behind the yet unexplained expansions of GGGGCC repeats seen at the C9*orf*72 locus of ALS-patients? A tentative model of how TMEJ can lead to this expansion of GGGGCC repeats at the C9*orf*72 locus is presented in Figure 3.

TMEJ and disease

An important question that is raised by the research described in thesis is to what extent TMEJ can be related to disease and moreover, whether the inhibition or absence of polymerase Theta can be used as a therapeutic approach.

To this point, there is compelling evidence that TMEJ can be linked to cancer. First of all, overexpression of polymerase Theta has been found in breast and colon cancers, and importantly, overexpression is correlated with a poor clinical outcome (Higgins et al., 2010a; Kawamura et al., 2004; Lemée et al., 2010; Pillaire et al., 2010). It will be worthwhile to test whether the inhibition of polymerase Theta in addition to current therapies, such as irradiation, leads to synergistic eradication of polymerase Theta-overexpressing tumors in patients. In concert with this suggestion, a recent study showed tumor-specific radiosensitization by knockdown of polymerase Theta when using a panel of various tumor cell lines (Higgins et al., 2010b).

Second, oncogenic chromosomal translocations often show footprints that are reminiscent of polymerase Theta-dependent repair products (Kidd et al., 2010; Murga Penas et al., 2010; Welzel et al., 2001) arguing that TMEJ can lead to chromosomal instability. Against this view, a recent study suggests that TMEJ in fact prevents the formation of chromosomal translocations (Yousefzadeh et al., 2014). Third, a large cohort study in which 3,281 tumours across 12 tumour types were sequenced, revealed that in ±10% of multiple cancer types polymerase Theta was found mutated (Kandoth et al., 2013). Rationally, screens for synthetic lethal compounds in polymerase Theta-deficient cancerous cells will offer potential interesting drug candidates. The work described in this thesis suggests that G-quadruplex stabilizing compounds might be an interesting class of compounds to test for eliminating polymerase Theta-deficient tumors.

Besides cancer, also other diseases might be the result of polymerase Theta-dependent repair. For example, footprints of 42 microdeletions of the *FOXL2* locus show the presence of insertions and usage of minimal homology at the junctions, resembling polymerase Theta-dependent footprints (Verdin et al., 2013). These microdeletions at the *FOXL2* locus lead to a disease named blepharophimosis-ptosis-epicanthus inversus syndrome (BPES), which is characterized by malformed eyelids and dysfunctional ovaries. Finally, as described in Figure 3, it will be interesting to investigate to which extent TMEJ contributes to the birth and expansion of microsatellites and related diseases.

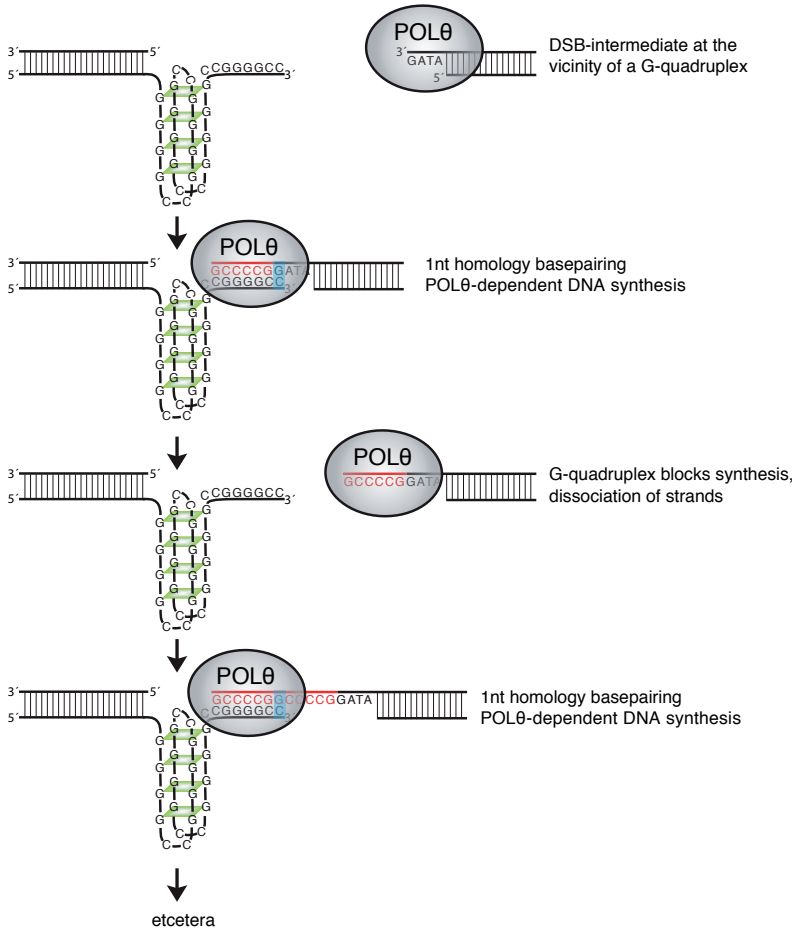


Figure 3 | Proposed model for repeat expansion at the C9ORF72 locus. The expansion of the G-quadruplex-forming repeat GGGGCC at the *C9ORF72* locus is associated with the neurodegenerative disease ALS. Normal human alleles have 2 to 25 GGGGCC repeats, whereas up to thousands of repeats have been found in patients (Haeusler et al, 2014). Here, a model is proposed of how repeat expansion can take place when a DSB at the *C9ORF72* contains a G-quadruplex structure in its immediate flank. As described in this thesis, G-quadruplex sequences can lead to DSB-intermediates, and therefore DSB-formation at a locus with several G-quadruplex-forming sequences in a row may lead to a DSB that is flanked with one (or more) G-quadruplex at its junction (illustrated in the top panel). *De novo* DNA synthesis by polymerase Theta is aborted by the G-quadruplex structure (second panel). Dissociation of strands takes place and a next attempt for repair by polymerase Theta will be made (panel 3 and 4). In this way, endless iterative cycles can take place causing enormous expansion of this hexanucleotide repeat. *De novo* synthesized DNA is presented in red. 1nt homology basepairing is colored in blue.

G-quadruplexes; friend or foe?

In recent years it has become convincingly evident that G-quadruplexes do exist *in vivo* and influence biological processes such as replication, transcription, translation and RNA localization. Strikingly, the majority of studies involving G-quadruplexes (including the work presented in this thesis) focuses on the downside of having G-quadruplexes in our genome: G-quadruplexes are linked to genomic instability, oncogene translation, epigenetic instability, RNA mislocalization and associated diseases such as cancer, anemia and neurodegenerative diseases such as ALS. However, as often said, your worst enemy can be your best friend: G-quadruplex are thought to be excellent therapeutic targets for small molecules (Ohnmacht and Neidle, 2014). Below I will set out several strategies on how G-quadruplexes may be used as a therapeutic intervention. First of all, it is thought that inducing G-quadruplex formation at single-stranded telomeric DNA via small molecules results in the inhibition of telomerase activity, DNA damage responses and cellular apoptosis and thereby prevents further expansion of cancerous cells that have too high levels of telomerase activity (Ohnmacht et al., 2013). A second approach is the genome-wide induction of G-quadruplex stabilization by small molecules, which will result in replication-associated stress and subsequently cellular death. Fast replicating cancerous cells will likely be more sensitive than non-dividing or slow dividing wild type cells. In favor of this concept, a recent study shows that the G4-stabilizing agent pyridostatin inhibits the growth of BRCA2-deficient cancerous cells (McLuckie et al., 2013). In addition, G-quadruplex stabilizing agents show promising synthetic lethal effects when combined with other DNA damaging compounds (see for examples reference (Ali and Bhattacharya, 2014) and references therein). A third mode of action of G4-stabilizing molecules could be at the transcriptional level: numerous oncogenes and cancer-related genes are candidates for small molecule-dependent inhibition of transcription by stabilizing the G4 sequence in their promotors. Promising targets are for example c-*KIT*, c-*MYC*, k-*RAS*, B-RAF, BCL-2, VEGF growth factor and hTERT. Recently it has been described that G-quadruplexes can influence the translation of several oncogenes, which opens up possibilities to interfere at the translational level by means of small molecules (Wolfe et al., 2014). In addition, it has been reported recently that DNA and RNA G-quadruplexes cause toxic aborted transcripts with ultimately the neurodegenerative disease ALS as outcome (Haeusler et al., 2014), pointing to G-quadruplex-binding small molecules as a potential cure for this deadly disease. Conceivably, for this particular and likely other diseases it would be of great interest to see whether molecules can be designed that rather suppress than stimulate the stabilization of G-quadruplex structures. Finally, and perhaps needless to say, all proteins (DNA and RNA helicases in particular) involved in binding and unwinding of G-quadruplexes can be considered as promising therapeutic targets as well. In fact, screens for G4-specific helicase-targeting molecules are underway (Hale et al., 2014).

Currently, various G-quadruplex stabilizing agents are available and many more are yet to come. Although, their stabilizing effects and other characteristics are often tested

in vitro, assays to test their stabilizing capacity *in vivo* and *ex vivo* are underdeveloped. Systems as for example described in Chapter 5 that allow for functional testing of G4 DNA-stabilizing compounds and helicase-inhibitors *ex vivo*, will proof valuable tools for the development of effective therapeutic small molecules in the near future.

The research and developed tools described in this thesis provide new insights into the genomic stability of microsatellites and G-quadruplexes and will contribute to better understanding and treatment of diseases such as cancer and other microsatellite- and G-quadruplex-related diseases.

REFERENCES

Ali, A., and Bhattacharya, S. (2014). DNA binders in clinical trials and chemotherapy. Bioorganic & Medicinal Chemistry *22*, 4506–4521.

Chan, S.H., Yu, A.M., and McVey, M. (2010). Dual roles for DNA polymerase theta in alternative end-joining repair of double-strand breaks in Drosophila. PLoS Genet. *6*, e1001005.

Deng, S.K., Gibb, B., de Almeida, M.J., Greene, E.C., and Symington, L.S. (2014). RPA antagonizes microhomology-mediated repair of DNA double-strand breaks. Nature Structural & Molecular Biology *21*, 405–412.

García-Gómez, S., Reyes, A., Martínez-Jiménez, M.I., Chocrón, E.S., Mourón, S., Terrados, G., Powell, C., Salido, E., Méndez, J., Holt, I.J., et al. (2013). PrimPol, an Archaic Primase/Polymerase Operating in Human Cells. Molecular Cell *52*, 541–553.

Haeusler, A.R., Donnelly, C.J., Periz, G., Simko, E.A.J., Shaw, P.G., Kim, M.-S., Maragakis, N.J., Troncoso, J.C., Pandey, A., Sattler, R., et al. (2014). C9orf72 nucleotide repeat structures initiate molecular cascades of disease. Nature *507*, 195–200.

Hale, T.K., Norris, G.E., Jameson, G.B., and Filichev, V.V. (2014). Helicases, G4-DNAs, and drug design. ChemMedChem *9*, 2031–2034.

Heller, R.C., and Marians, K.J. (2006). Replication fork reactivation downstream of a blocked nascent leading strand. Nature *439*, 557–562.

Higgins, G.S., Harris, A.L., Prevo, R., Helleday, T., McKenna, W.G., and Buffa, F.M. (2010a). Overexpression of POLQ confers a poor prognosis in early breast cancer patients. Oncotarget *1*, 175–184.

Higgins, G.S., Prevo, R., Lee, Y.-F., Helleday, T., Muschel, R.J., Taylor, S., Yoshimura, M., Hickson, I.D., Bernhard, E.J., and McKenna, W.G. (2010b). A small interfering RNA screen of genes involved in DNA repair identifies tumor-specific radiosensitization by POLQ knockdown. Cancer Res. *70*, 2984–2993.

Hogg, M., Sauer-Eriksson, A.E., and Johansson, E. (2012). Promiscuous DNA synthesis by human DNA polymerase θ. Nucleic Acids Res. *40*, 2611–2622.

Hogg, M., Seki, M., Wood, R.D., Doublié, S., and Wallace, S.S. (2011). Lesion bypass activity of DNA polymerase θ (POLQ) is an intrinsic property of the pol domain and depends on unique sequence inserts. J. Mol. Biol. *405*, 642–652.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.

Kawamura, K., Bahar, R., Seimiya, M., Chiyo, M., Wada, A., Okada, S., Hatano, M., Tokuhisa, T., Kimura, H., Watanabe, S., et al. (2004). DNA polymerase theta is preferentially expressed in lymphoid tissues and upregulated in human cancers. Int. J. Cancer *109*, 9–16.

Kelkar, Y.D., Eckert, K.A., Chiaromonte, F., and Makova, K.D. (2011). A matter of life or death: how microsatellites emerge in and vanish from the human genome. Genome Res. *21*, 2038–2048.

Kidd, J.M., Graves, T., Newman, T.L., Fulton, R., Hayden, H.S., Malig, M., Kallicki, J., Kaul, R., Wilson, R.K., and Eichler, E.E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell *143*, 837–847.

Lemée, F., Bergoglio, V., Fernandez-Vidal, A., Machado-Silva, A., Pillaire, M.-J., Bieth, A., Gentil, C., Baker, L., Martin, A.-L., Leduc, C., et al. (2010). DNA polymerase theta up-regulation is associated with poor survival in breast cancer, perturbs DNA replication, and promotes genetic instability. Proc. Natl. Acad. Sci. U.S.a. *107*, 13390–13395.

Lopes, J., Piazza, A., Bermejo, R., Kriegsman, B., Colosio, A., Teulade-Fichou, M.-P., Foiani, M., and Nicolas, A. (2011). G-quadruplex-induced instability during leading-strand replication. Embo J *30*, 4033–4046.

Lopes, M., Foiani, M., and Sogo, J.M. (2006). Multiple mechanisms control chromosome integrity after replication fork uncoupling and restart at irreparable UV lesions. Molecular Cell *21*, 15–27.

McLuckie, K.I.E., Di Antonio, M., Zecchini, H., Xian, J., Caldas, C., Krippendorff, B.-F., Tannahill, D., Lowe, C., and Balasubramanian, S. (2013). G-quadruplex DNA as a molecular target for induced synthetic lethality in cancer cells. J. Am. Chem. Soc. *135*, 9640–9643.

Messier, W.W., Li, S.H.S., and Stewart, C.B.C. (1996). The birth of microsatellites. Nature *381*, 483–483.

Mourón, S., Rodriguez-Acebes, S., Martínez-Jiménez, M.I., García-Gómez, S., Chocrón, S., Blanco, L., and Méndez, J. (2013). Repriming of DNA synthesis at stalled replication forks by human PrimPol. Nature Structural & Molecular Biology *20*, 1383–1389.

Murga Penas, E.M., Callet-Bauchu, E., Ye, H., Gazzo, S., Berger, F., Schilling, G., Albert-Konetzny, N., Vettorazzi, E., Salles, G., Wlodarska, I., et al. (2010). The t(14;18)(q32;q21)/IGH-MALT1 translocation in MALT lymphomas contains templated nucleotide insertions and a major breakpoint region similar to follicular and mantle cell lymphoma. Blood *115*, 2214–2219.

Ohnmacht, S.A., and Neidle, S. (2014). Small-molecule quadruplex-targeted drug discovery. Bioorg Med Chem Lett *24*, 2602–2612.

Ohnmacht, S.A., Varavipour, E., Nanjunda, R., Pazitna, I., Di Vita, G., Gunaratnam, M., Kumar, A., Ismail, M.A., Boykin, D.W., Wilson, W.D., et al. (2013). Discovery of new G-quadruplex binding chemotypes. Chem. Commun. *50*, 960.

Pillaire, M.-J., Selves, J., Gordien, K., Gourraud, P.-A., Gouraud, P.-A., Gentil, C., Danjoux, M., Do, C., Negre, V., Bieth, A., et al. (2010). A "DNA replication" signature of progression and negative outcome in colorectal cancer. Oncogene *29*, 876–887.

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

Schiavone, D., Guilbaud, G., Murat, P., Papadopoulou, C., Sarkies, P., Prioleau, M.-N., Balasubramanian, S., and Sale, J.E. (2014). Determinants of G quadruplex-induced epigenetic instability in REV1-deficient cells. Embo J *33*, 2507–2520.

Smith, D.J., and Whitehouse, I. (2012). Intrinsic coupling of lagging-strand synthesis to chromatin assembly. Nature *483*, 434–438.

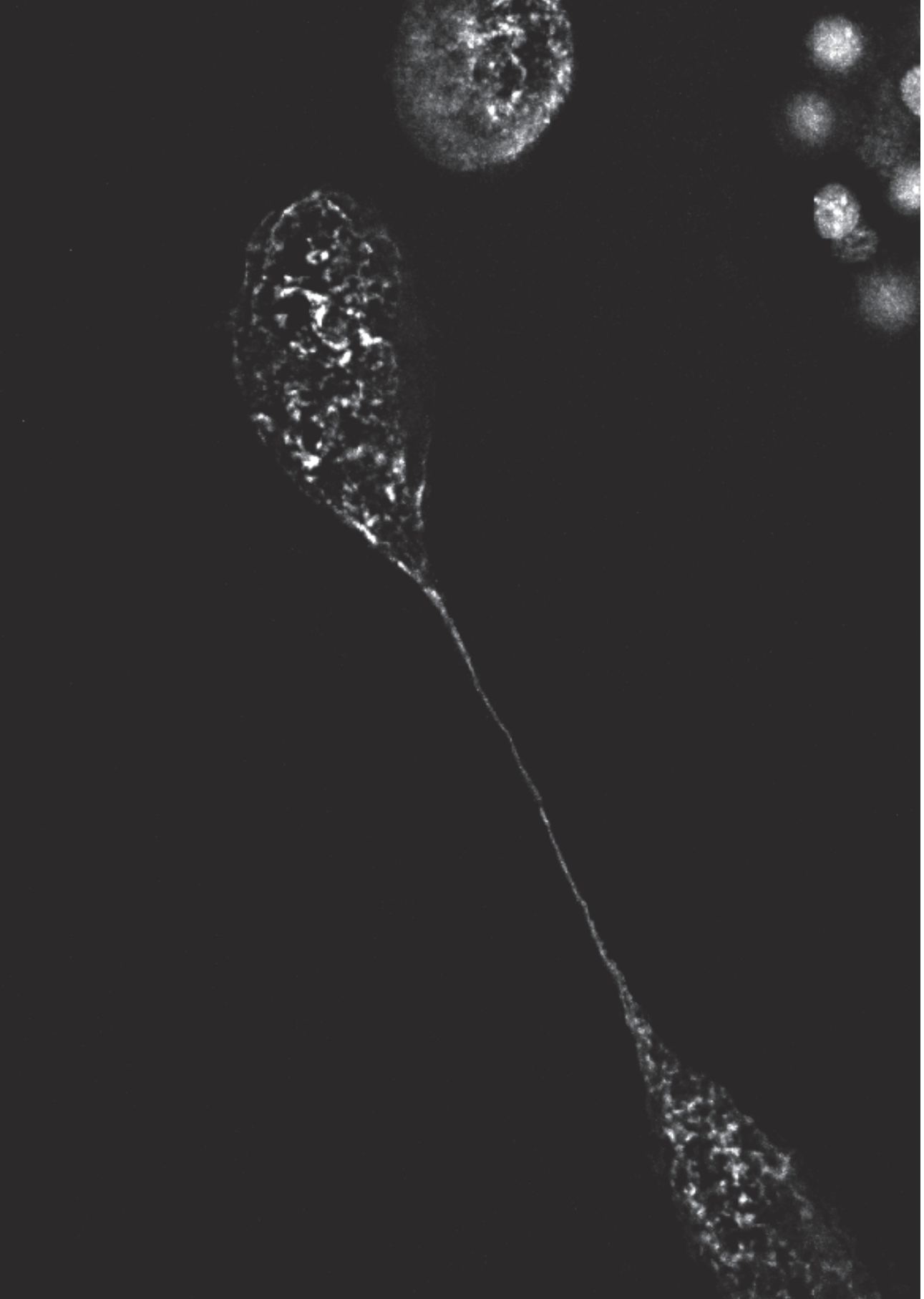
Verdin, H., D'haene, B., Beysen, D., Novikova, Y., Menten, B., Sante, T., Lapunzina, P., Nevado, J., Carvalho, C.M., and Lupski, J.R. (2013). Microhomology-mediated mechanisms underlie non-recurrent disease-causing microdeletions of the FOXL2 gene or its regulatory domain. PLoS Genet. *9*, e1003358.

Welzel, N., Le, T., Marculescu, R., Mitterbauer, G., Chott, A., Pott, C., Kneba, M., Du, M.Q., Kusec, R., Drach, J., et al. (2001). Templated nucleotide addition and immunoglobulin JH-gene utilization in t(11;14) junctions: implications for the mechanism of translocation and the origin of mantle cell lymphoma. Cancer Res. *61*, 1629–1636.

Wolfe, A.L., Singh, K., Zhong, Y., Drewe, P., Rajasekhar, V.K., Sanghvi, V.R., Mavrakis, K.J., Jiang, M., Roderick, J.E., Van der Meulen, J., et al. (2014). RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. Nature *513*, 65–70.

Yeeles, J.T.P., and Marians, K.J. (2011). The Escherichia coli replisome is inherently DNA damage tolerant. Science *334*, 235–238.

Yousefzadeh, M.J., Wyatt, D.W., Takata, K.-I., Mu, Y., Hensley, S.C., Tomida, J., Bylund, G.O., Doublié, S., Johansson, E., Ramsden, D.A., et al. (2014). Mechanism of suppression of chromosomal instability by DNA polymerase POLQ. PLoS Genet. *10*, e1004654–e1004654.



◀ This picture shows two nuclei in the gut of a worm that are connected with a thin thread of DNA (a so-called chromatin bridge). Such an event is very unusual in normal worms, while it is frequently observed in worms that lack the G-quadruplex unwinding enzyme DOG-1.

Addendum

Cover Photo: Iron G4 ▶

THESIS SUMMARY

Microsatellites and G-quadruplex motifs are DNA sequences that are prevalent throughout the genome and are linked to diseases such as cancer and neurodegenerative disorders. Microsatellites can be defined as short tandem repeats with units of 1-8 base pairs (bp) long. G-quadruplex motifs (also known as G4 DNA) can be defined as sequences that contain four tracts of three or more guanines interspaced by at least one random nucleotide. Although DNA is organized as a double helical structure, G-quadruplex motifs have the unique capacity to form a four-stranded DNA structure (named a G-quadruplex structure) in which the guanines interact with each other through Hoogsteen base pairing. Microsatellites and G-quadruplexes motifs are sequences that are intrinsically difficult to replicate because of their repetitive nature and capacity to form alternative DNA structures. As a result, they can trigger mutations upon replication. Microsatellites and G-quadruplexes play a significant role in the initiation of the aforementioned devastating diseases. However, many aspects about G-quadruplex and microsatellite instability (MSI) are incompletely understood. For example, what determines the degree of their instability, since some microsatellites and G-quadruplexes are more mutagenic than others? Which genes and pathways prevent microsatellite and G-quadruplex instability? What are the direct genetic consequences, and which molecular mechanisms act to produce genomic changes at these sequences? Answers to these questions will be of great importance in the development of new and better treatments of microsatellite- and G-quadruplex-related diseases. In this thesis I provide new insights into the biology behind microsatellite and G-quadruplex instability, by making use of various model organisms and newly developed molecular tools.

Using an MSI-reporter system based on fluorescence and optimized for mammalian cells, I demonstrate in **Chapter 2** that MSI is greatly influenced by the tract length and the nucleotide composition of the microsatellite. Other factors, such as its strand orientation or its transcriptional status, appeared not to affect the stability of a microsatellite. Furthermore, I show that the MSI-reporter system can be a useful tool to screen for MSI-inducing compounds, as well as to screen for genes that protect the genome against MSI. By testing a library of ±450 different miRNAs (small RNA-molecules that influence gene expression), I show that overexpression of one of these miRNAs, named miR-21, results in increased MSI. By additional experiments I show that miR-21 targets and thereby reduces the expression of the Lynch syndrome- and MSI-associated gene MSH2, explaining the increased levels of MSI observed in miR-21 overexpressing cells.

In **Chapter 3** I present a new genetic tool that enables us to trace single cells and also to study tumor development in living zebrafish (). I show that genes can be stochastically activated by placing their coding sequence downstream of a microsatellite. The gene of interest is placed such that only after a mutation in the microsatellite, a so-called frameshift, the gene becomes expressed. Low frequency frameshifting stochastically,

but occasionally, activates the gene of interest in cells. I show that when the activated gene of interest in these cells is tagged with a fluorescent protein, these cells and their progeny can be followed over time in a living animal. Using the same principle I also describe that microsatellite-dependent stochastic activation of an oncogene, in this case oncogenic H-RAS, results in the formation of tumors within 5 days. Since zebrafish are transparent during embryonic development, this technique provides the opportunity to study the early stages of tumor development in a living animal.

In **Chapter 4**, I focus on G-quadruplex instability in the nematode. Previous studies have shown that G-quadruplexes can induce 50-300 bp deletions in the genome of mutant worms that lack the helicase DOG-1 (the worm counterpart of the human Fanconi anemia-associated gene FANCJ), however, the underlying mechanism of this process was unknown. In this study, I reveal the molecular mechanism that explains the formation of these G-quadruplex-induced deletions. I show that a polymerase, named polymerase theta, plays an essential role in the formation of these G-quadruplex-induced deletions. Polymerase Theta-Mediated End-Joining (TMEJ) of G-quadruplex-induced DNA damage results in deletions that are characterized by a) an extremely narrow size distribution between 50-300 bp, b) one nucleotide homology at the junctions and c) occasional insertions of up to 20 base pairs templated from the flanking sequences. In the absence of polymerase theta, G-quadruplexes can lead to large deletions spanning several kilobases. By comparing the genomes of worms that lived geographically separated for millions of years, I provide evidence that G-quadruplex-induced genomic deletions occurred also during the normal evolution of wild type worms.

In **Chapter 5**, I investigate the stability of G-quadruplexes in human cells. I provide preliminary data that G-quadruplex motifs can be highly polymorphic between humans. Using fluorescence-based G-quadruplex-instability reporters, I further show that G-quadruplex-dependent genomic instability is increased upon transcription and upon treatment with G-quadruplex-binding ligands. Finally, I provide data that argues that G-quadruplexes can lead to small genomic deletions in human cells. Although it is unclear at the moment whether this genomic instability phenotype is FANCJ and polymerase theta dependent, the newly developed G-quadruplex-instability reporters presented in this chapter will, in combination with functional genetic screens, facilitate an answer to this question in the future.

My thesis ends with a summarizing discussion and future perspectives on microsatellite and G-quadruplex instability and their link to disease. For example, I speculate how polymerase theta-activity can lead to the expansion of a specific G-quadruplex motif that is associated with Amyotrophic Lateral Sclerosis (ALS) and how small molecules that bind G-quadruplexes or inhibit polymerase theta activity can aid in the treatment of cancer and other microsatellite- and G-quadruplex-related diseases.

NEDERLANDSE SAMENVATTING (VOOR NIET-INGEWIJDEN)

Bij het lezen van de titel “Microsatellite and G-quadruplex instability in the worm, fish and man” zullen velen zich even achter de oren moeten krabben. De eerste drie woorden klinken waarschijnlijk als abracadabra en hoewel de meeste lezers kunnen herleiden uit de titel dat er onderzoek is gedaan in wormen, vissen en mensen, zullen sommigen zich afvragen wat het nut is om onderzoek te doen in van die vieze regenwormen, en lieve visjes.

Hieronder zal ik uiteenzetten wat “microsatellites” en “G-quadruplexes” zijn, waarom we niet alleen onderzoek doen in menselijke cellen maar ook in model-organismes zoals de rondworm (geen regenworm, maar een wormpje van één millimeter groot) en de zebravis (een visje van ongeveer drie centimeter lang) en het belangrijkste: waarom bestuderen we dit en wat zijn de bevindingen beschreven in dit proefschrift?

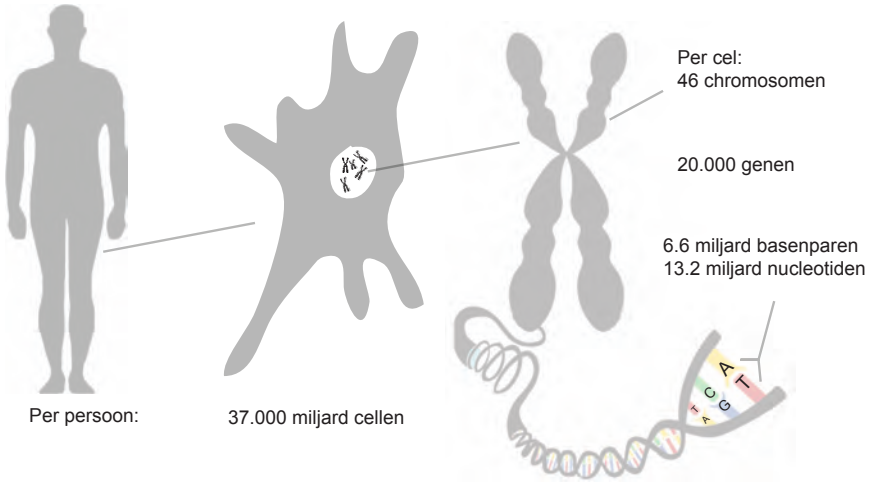
Even terug naar de schoolbanken…

Iedereen die een beetje heeft opgelet tijdens de lessen biologie weet dat de mens is opgebouwd uit een heleboel cellen. Om precies te zijn, gemiddeld genomen bestaat een mens uit ongeveer 37.000 miljard cellen! En dan te bedenken dat al deze cellen ontstaan uit één bevruchte eicel. Hoe kan het dat één zo’n minuscuul celletje zich kan vermenigvuldigen tot zoveel cellen en een volwaardig en gezond persoon?

Uiteraard zijn daarvoor voedingsstoffen zoals suikers, eiwitten, vetten, water en zuurstof voor nodig, maar daar alleen red je het niet mee. Op een of andere manier moeten cellen instructies krijgen die bepalen dat de ene cel bijvoorbeeld een zenuwcel wordt, een andere cel insuline aanmaakt en de ander een stamcel in de darm wordt. De blauwdruk voor al deze ingewikkelde processen ligt vastgelegd in het DNA van een cel.

Het DNA bevindt zich in de vorm van chromosomen in een cel. Een gezonde menselijke cel bevat 46 **chromosomen** en elk chromosoom is opgebouwd uit 2 strengen die met elkaar verbonden zijn (zie figuur 1). Elke streng is op zijn beurt weer opgebouwd uit **4 bouwstenen** (ook wel **nucleotiden** ofwel basen genoemd) die afgekort worden met de letters **A, T, C en de G**. De A kan in principe alleen een interactie aangaan met de T en dit wordt een basepaar genoemd. De C kan alleen een basepaar vormen met de G. Door deze baseparen worden de twee strengen bij elkaar gehouden. In totaal heeft een menselijke cel ongeveer 6,6 miljard baseparen (en dus 13.2 miljard nucleotiden) verdeeld over de 46 chromosomen. Als je je dan bedenkt dat een menselijk lichaam 37.000 miljard cellen heeft, kom je uit op ongeveer 500.000.000.000.000.000.000.000 nucleotiden/bouwstenen die belangrijke genetische informatie bevatten. Dit is natuurlijk een duizelingwekkend getal, helemaal als je nagaat dat met een beetje pech een verandering van één enkele nucleotide in een cel ineens kan leiden tot ongecontroleerde celdelingen met **kanker** als gevolg.

Kortom, het is voor elke cel en organisme dus van groot belang dat er niet zomaar foutjes, ook wel **mutaties** genoemd, in het DNA sluipen!



Figuur 1 |Een mens is opgebouwd uit ongeveer 37.000 miljard cellen. Elke gezonde lichaamscel heeft 46 chromosomen. Een chromosoom bestaat op zijn beurt uit twee DNA strengen welke opgebouwd zijn uit vier bouwstenen/nucleotiden: A, T, C, G. De A kan een basepaar vormen met de T en de C vormt een basepaar met de G. De mens heeft verspreid over de chromosomen ± 20.000 genen liggen. Elk gen, bestaande uit gemiddeld 10.000-15.000 baseparen, bevat de genetische code voor een eiwit. Elk eiwit heeft zijn eigen functie in de cel en eiwitten kunnen gezien worden als de werkpaarden van de cel.

Het ontstaan en voorkomen van mutaties in DNA

Er zijn verschillende manieren waarop er mutaties in DNA kunnen ontstaan. Allereerst doordat DNA beschadigd raakt door factoren van buitenaf. Bijvoorbeeld UV-straling afkomstig van de zon, radioactieve straling (denk aan de kernrampen in Tjernobyl en Fukushima), en sigarettenrook zijn voorbeelden waardoor DNA beschadigd kan worden. Echter, DNA kan ook beschadigd worden door processen binnen in de cel zoals het vrijkomen van zuurstofradicalen tijdens de stofwisseling. Wanneer DNA beschadigd raakt hoeft het niet per definitie te betekenen dat een blijvende mutatie optreedt. Cellen zijn namelijk zo geavanceerd dat ze **herstelmechanismen** hebben ontwikkeld om schade aan DNA te kunnen repareren. Om een indruk te krijgen hoe geavanceerd dit is, zie je in tabel 1 van hoofdstuk 1 een overzicht van verschillende herstelmechanismen en de belangrijkste eiwitten die daarbij betrokken zijn. Hoewel deze herstelmechanismen erg accuraat zijn, zijn ze niet geheel feilloos en wil er tijdens het herstel van DNA-schade wel eens een foutje optreden met als gevolg een mutatie in het DNA. Niet alleen door beschadigd DNA kunnen er mutaties ontstaan, ook tijdens een **celdeling** is er een verhoogd risico op het ontstaan van mutaties in DNA. Aangezien elke cel zijn genetische informatie nodig heeft zal voordat een cel deelt eerst het DNA gekopieerd moeten worden zodat beide cellen voorzien zijn van DNA. Dit kopiëren, ook wel repliceren genoemd, gaat met een enorme snelheid. Om een voorbeeld te geven: tijdens de vroege ontwikkeling van een zebravis-embryo kunnen cellen hun DNA (± 1,5 miljard basenparen) binnen 15 minuten kopiëren. Dit kopiëren gaat wel eens gepaard met foutjes. Ook voor dit soort foutjes heeft de cel herstelmechanismes paraat. Een herstelmechanisme dat voornamelijk betrokken is bij het herstellen van replicatie-foutjes heet het **mismatch-herstel (MMH)** mechanisme. Wanneer een cel dit MMH-mechanisme mist neemt het aantal mutaties in de cel met een duizendvoud toe!

Hoewel een enkele mutatie in DNA vaak niet tot gevaarlijke situaties leidt, wordt het wel gevaarlijk wanneer je duizend keer zoveel mutaties in je DNA krijgt. Zo wordt de kans veel groter dat een cel door mutaties verkeerde instructies krijgt en ongecontroleerd gaat delen. Dit zien we dan ook gebeuren in mensen die een defect MMH-mechanisme hebben (ook wel **Lynch syndroom** genoemd). Deze patiënten hebben een sterk verhoogde kans op kanker, waarbij de tumoren vaak voorkomen in weefsels waar cellen veel delen. De darm is zo'n weefsel waar cellen veel delen en Lynch sydroom patiënten krijgen vrijwel altijd darmkanker. Wat verder opvalt bij deze patiënten is dat de mutaties in hun DNA vaak optreden in **microsatellieten** (microsatellites) en daarom worden deze microsatellieten vaak bekeken voor het diagnosticeren van het Lynch syndroom. Maar wat zijn microsatellieten en waarom bestuderen we microsatellieten?

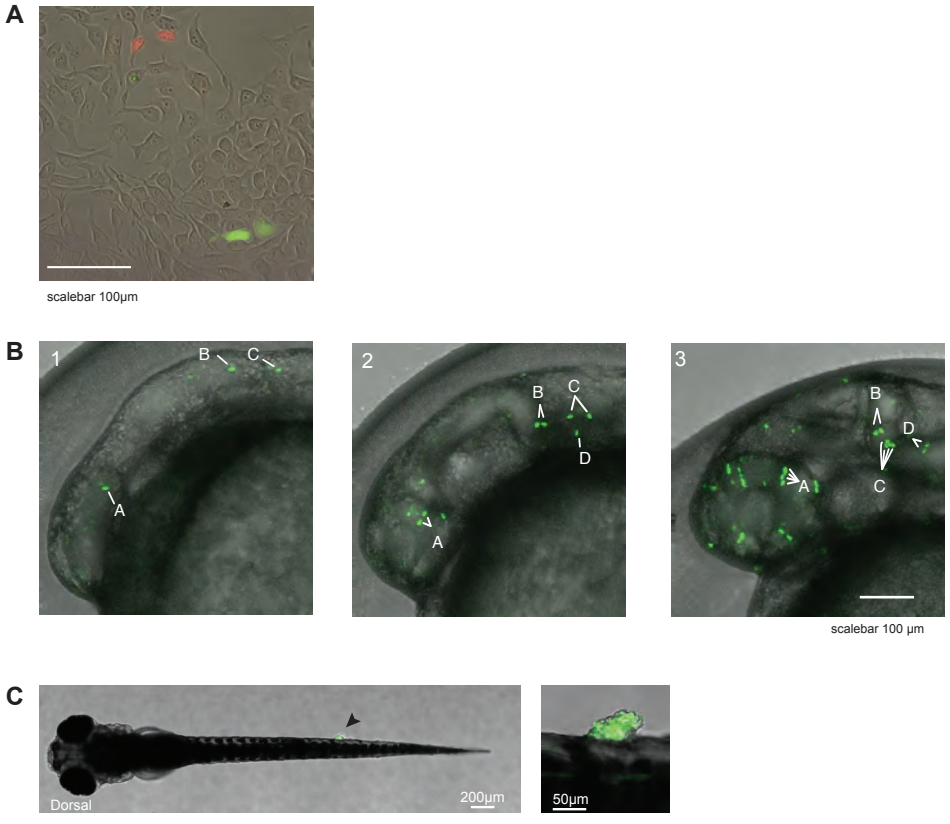
Microsatellieten

Microsatellieten bestaan uit repeterend DNA waarbij eenheden van 1 tot 8 baseparen herhaaldelijk achter elkaar voorkomen. Bijvoorbeeld de DNA volgordes AAAAAAAAAAA (waarbij de A de repeterende eenheid is), CAGCAGCAGCAG CAGCAG (waarbij CAG de repeterende eenheid is), GGGGCCGGGGCCGGGG CCGGGGCCGGGGCC (waarbij GGGGCC de repeterende eenheid is) enzovoorts, kunnen allemaal als microsatellieten bestempeld worden. Naar nu blijkt bestaat het menselijk genoom uit ongeveer 3% van deze microsatellieten en opvallend is dat deze microsatellieten erg variëren in lengte per persoon. Deze microsatellieten variëren dusdanig in lengte dat vrijwel ieder persoon een uniek patroon van microsatellieten heeft en daarom worden ze ook gebruikt in forensisch onderzoek.

Microsatellieten kunnen dus handig zijn voor onder andere forensisch onderzoek en het diagnosticeren van Lynch syndroom-patiënten. Echter, veranderingen in de lengtes van microsatellieten worden ook in verband gebracht met ziektes, zoals bepaalde spierziektes en neuronale ziektes. Kortom, genoeg redenen om alles te weten te komen over microsatellieten! Hoe meer we te weten komen over microsatellieten des te beter we kunnen voorspellen of een bepaalde persoon verhoogd risico heeft op een desbetreffende ziekte, maar nog belangrijker: des te beter we uiteindelijk medicijnen kunnen ontwikkelen voor deze microsatelliet-gerelateerde ziektes.

In **hoofdstuk 2** van dit proefschrift heb ik onderzoek verricht naar welke factoren een rol spelen bij de stabiliteit van deze microsatellieten. Om dit uit te zoeken heb ik een methode ontwikkeld waarbij we in menselijke cellen, gekweekt in een schaaltje,

kunnen uitlezen wanneer de lengte van een microsatelliet is veranderd. Deze menselijke cellen heb ik zodanig genetisch gemodificeerd dat ze bij het korter worden van een specifieke microsatelliet een fluorescerend eiwit (genaamd “mCherry”) aanmaken waardoor de cel rood kleurt. Bij het langer worden van de microsatelliet maakt de cel een groen fluorescerend eiwit aan: **G**reen **F**luorescent **P**rotein (GFP). In figuur 2A zie je een foto, waarbij je de groene en rode cellen kunt zien.



Figuur 2 | (A) Een foto van menselijke cellen is te zien waarbij de rode cellen aangeven dat een microsatelliet korter is geworden. In de groene cellen is een microsatelliet juist langer geworden. Met behulp van deze cellijnen kunnen we analyseren hoe vaak bepaalde microsatellieten instabiel zijn. (B) Deze foto’s geven het hoofd van een zebravis-embryo weer van respectievelijk 16, 21 en 31 uur oud. Doordat in deze vissen de code van het fluorescerende eiwit GFP achter een microsatelliet is gezet, zal alleen bij het korter/langer worden van de microsatelliet een cel groen worden. Dit gebeurt willekeurig en niet vaak. Maar als het gebeurt kan je heel duidelijk de cel volgen in de loop der tijd en kan je ook na een celdeling de dochtercel blijven volgen in de tijd. De letters A-D geven een individuele cel en dochtercellen weer. (C) Door de code van een kankergen te plaatsen achter een microsatelliet kunnen we ook random gezonde cellen veranderen in kankercellen. Doordat we het kankergen ook nog eens gelabeld hebben met GFP kunnen we zien in welke cellen het kankergen actief is en hoe een tumor groeit. In het linker plaatje zie je een 5 dagen oude zebravis waar aan de zijkant een kleine tumor begint te groeien. Het rechter plaatje zoomt in op de tumor.

Door gebruik te maken van een geavanceerde machine die één voor één de cellen analyseert op kleur kunnen we miljoenen cellen tellen in minder dan een uur en precies uitrekenen hoe stabiel een microsatelliet is. Hierdoor heb ik kunnen bepalen dat de stabiliteit van een microsatelliet erg afhangt van de lengte. Bijvoorbeeld: een microsatelliet van 23 C’s is bijna 100x meer instabiel dan een microsatelliet van 14 C’s.

Maar met deze methode kunnen we meer! Zo hebben we ook uitgezocht of we nieuwe genetische factoren kunnen vinden die betrokken zijn bij het instabiel worden van microsatellieten. Een van die mogelijke factoren blijken micro-RNA's te zijn. Dit zijn kleine RNA-moleculen die net als DNA opgebouwd zijn uit 4 bouwstenen, maar dan A, C, G, en U. De functie van miRNA’s is de aanmaak van bepaalde eiwitten te reguleren. Met geavanceerde technieken hebben we zodoende ±450 micro-RNA’s getest en ontdekt dat wanneer micro-RNA 21 (miR-21) aanwezig is in de cel, microsatellieten meer instabiel zijn. Vervolgens hebben we aangetoond dat deze miR-21 het mismatch-herstel mechanisme aantast en deze micro-RNA dus goed in de gaten gehouden moet worden, want teveel van miR-21 kan uiteindelijk leiden tot kanker.

De methode beschreven in dit hoofdstuk kan in de toekomst verder bijdragen aan het vinden van nieuwe genetische factoren die betrokken zijn bij het instabiel worden van microsatellieten en geeft inzicht in gerelateerde ziektes zoals het Lynch syndroom.

In **hoofdstuk 3** beschrijf ik een nieuwe methode waarbij we heel nauwkeurig specifieke cellen in de **zebravis** kunnen volgen over de tijd. Bij deze methode maak ik gebruik van hetzelfde principe als beschreven in hoofdstuk 2: door de DNA code van het fluorescerend eiwit GFP achter een microsatelliet te zetten zal bij het korter of langer worden van de microsatelliet de cel groen worden. Het korter/langer worden van de microsatelliet gebeurt random en vindt alleen plaats in delende cellen. Aangezien de mutatie die leidt tot het korter/langer worden van de microsatelliet blijvend is, zullen ook bij een volgende deling de cellen (wat we dochtercellen noemen) groen blijven. Op deze manier kunnen we heel precies een groepje cellen blijven volgen (lineage tracing). In figuur 2B zie je een voorbeeld waar we in de loop der tijd een cel en bijbehorende dochtercellen volgen. Daarnaast is hiervan een mooi filmpje te zien op https://www.youtube.com/watch?v=4D7dLzA8qrk.

In dit hoofdstuk laat ik niet alleen zien dat we met de ontwikkelde techniek cellen kunnen aankleuren en volgen, maar ook cellen kunnen veranderen in een kankercel. De aanpak beschreven in dit hoofdstuk heeft het voordeel dat telkens maar één cel en vervolgens de dochtercellen ongecontroleerd beginnen te delen, terwijl de omliggende cellen gezond zijn. Dit bootst het ontstaan van een tumor goed na, omdat een tumor vaak begint met maar één op hol geslagen cel. In figuur 2C zie je een voorbeeld hoe een groepje kankercellen in de zebravis is ontstaan uit één cel. Met dit model kunnen we in de zebravis binnen vijf dagen tumoren laten ontwikkelen. Een voordeel van de zebravis is dat ze transparant zijn tijdens de vroege embryonale ontwikkeling en zodoende kunnen we heel goed tumor-ontwikkeling bestuderen. Dit

biedt veel voordelen, waaronder het kunnen testen van nieuwe medicijnen die tumor-ontwikkeling tegen gaan.

G-quadruplexes

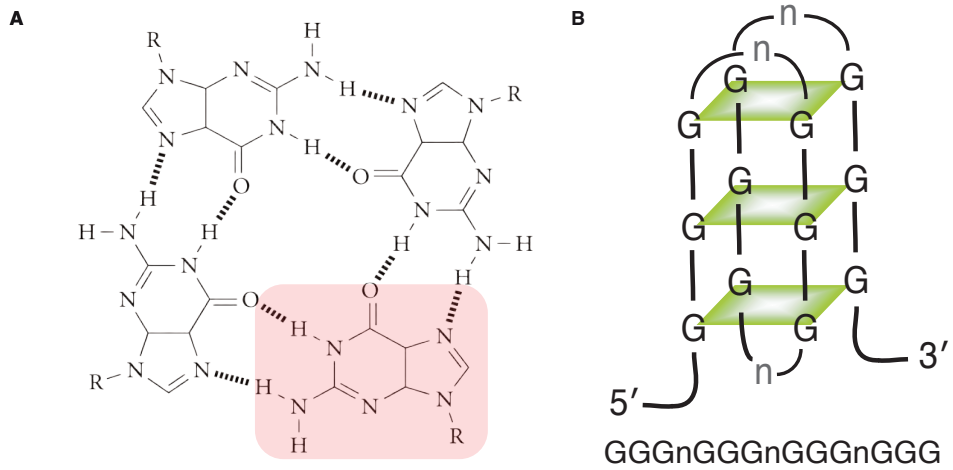
Naast microsatellieten zijn er ook andere stukken DNA die tot problemen kunnen leiden in een cel. Een G-quadruplex is zo'n stuk DNA.

Zoals eerder beschreven, is DNA opgebouwd uit 4 bouwstenen/nucleotiden (A, C, T, G) waarbij de A een interactie aan kan gaan met de T en waarbij de C interacteert met de G. Echter, meer dan 50 jaar geleden werd ontdekt dat de G als bouwsteen ook een interactie aan kan gaan met drie andere G's, waardoor een soort vierkant vlak ontstaat (ook wel een G-quartet genoemd, zie figuur 3A). Wanneer je drie vlakken op elkaar stapelt, spreken we van een **G-quadruplex**. (zie figuur 3B). Nu is gebleken dat een DNA-volgorde waarbij je minimaal vier rijtjes van drie G's hebt een G-quadruplex kan vormen (zie figuur 3B). Deze secundaire structuur is energetisch zeer stabiel en naar nu blijkt (onder andere uit de studies beschreven in hoofdstuk 4 en 5 in dit proefschrift), kunnen G-quadruplexes voor grote problemen zorgen.

Onlangs zijn G-quadruplexes bijvoorbeeld in verband gebracht met de ziekte Amyotrofische Laterale Sclerose (**ALS**). Dit is een spierziekte welke in 2014 extra aandacht kreeg door de "Ice-bucket-challenge". Daarnaast is er steeds meer bewijs dat de vorming van G-quadruplexes in DNA tot genomische instabiliteit kan leiden met kanker als gevolg. Daarom willen we alles weten over hoe en wanneer G-quadruplexes vormen, hoe ze tot genomische instabiliteit kunnen leiden en welke processen betrokken zijn bij het voorkomen of repareren van G-quadruplex-geïnduceerde DNA schade. In hoofdstuk 4 en 5 van dit proefschrift wordt het onderzoek over de instabiliteit van G-quadruplexes en bijbehorende processen en gevolgen beschreven.

Om de instabiliteit van G-quadruplexes te bestuderen hebben we gebruik gemaakt van **de rondworm *C. elegans***. Deze worm heeft een aantal voordelen, en één daarvan is dat de DNA herstelmechanismen voor een groot deel gelijk zijn als bij de mens. In tabel 1 van hoofdstuk 1 kun je zien welke herstelmechanismes en betrokken eiwitten zowel in de mens als in de worm voorkomen. Bijvoorbeeld, de 'borstkankergenen' BRCA1 en BRCA2 blijken ook aanwezig te zijn in de worm, net als de 'darmkankergenen' die betrokken zijn bij het eerder besproken Lynch syndroom. Een andere belangrijke reden waarom we voor deze studie *C. elegans* hebben gebruikt is dat we verschillende methoden hebben ontwikkeld, waarbij we de instabiliteit van G-quadruplexes kunnen waarnemen.

Eerdere studies toonden aan dat een G-quadruplex met ±150 naastgelegen baseparen soms zomaar uit het DNA van de worm verdwijnt. Dit wordt een deletie genoemd. Zo'n deletie vaagt dan in één klap behoorlijk wat genetische informatie weg. Dit is natuurlijk niet wenselijk voor een organisme. Tevens was bekend dat wanneer een worm het enzym genaamd *dog-1* miste (wat staat voor *d*eletion *o*f *g*uanine-rich DNA), ongeveer duizend keer vaker een G-quadruplex-geïnduceerde deletie voorkwam. Dit



Figuur 3 | (A) Waar normaal de nucleotide G een interactie aangaat met de nucleotide C, is gebleken dat 4 G's ook een interactie kunnen aangaan met elkaar. Dit wordt een G-quartet genoemd. Het rode vlak geeft de structuur van 1 G-nucleotide weer. (B) Wanneer je 3 G-quartets op elkaar stapelt spreken we van een G-quadruplex structuur. Een illustratie van een G-quadruplex is hier afgebeeld. Om een G-quadruplex te vormen heb je dus minimaal 4 rijtjes van 3 G's nodig en tussen elk rijtje kunnen één of meerdere nucleotides zitten (in deze figuur weergegeven met de letter "n"). Het blijkt dat er in het menselijk genooom maar liefst 300.000 plekken zijn waar zo'n G-quadruplex kan vormen!

enzym *dog-1* behoort tot een groep enzymen die vouwen in DNA kunnen ontwinden. Je kunt het vergelijken met een **strijkijzer** (*dog-1*) die kreukels (G-quadruplexes) in een hemd (DNA) kan glad strijken. Nota bene, net als bij strijken is ook bij het ontwinden van G-quadruplexes energie nodig om de kreukels eruit te halen. Hoewel een paar kreukels in een hemd niet veel kwaad kan, kan een kreukel in je DNA wel tot gevaarlijke situaties leiden. Vandaar ook dat iedere cel goed is uitgerust met een hoop verschillende strijkijzers (helicases) om deleties te voorkomen. Echter, voor een lange tijd bleef het een groot mysterie hoe en waarom die specifieke deleties van ±150 baseparen vormden. Welk mechanisme lag hier aan ten grondslag?

Na een lange zoektocht hebben we dit mechanisme ontdekt hetgeen beschreven staat in **hoofdstuk 4**. Cruciaal in de vorming van deze G-quadruplex-geïnduceerde deleties is het eiwit **polymerase Theta**. We laten zien dat wanneer een G-quadruplex niet goed plat gestreken wordt er uiteindelijk een breuk ontstaat in het DNA, ook wel een dubbel-strengs breuk genoemd (**DSB**). Uit ons onderzoek blijkt dat polymerase Theta in staat is deze DSB te repareren door de twee losse DNA-einden weer aan elkaar te plakken. Dit proces hebben we dan ook **polymerase Theta-mediated end-joining (TMEJ)** genoemd. Bijzonder aan TMEJ is dat door toedoen van slechts één eiwit klaarblijkelijk twee DNA uiteindes aan elkaar geplakt kunnen worden. Er zijn ook twee andere processen bekend die DSB-en kunnen repareren, maar bij deze complexe processen zijn veel meer eiwitten betrokken. TMEJ is in vergelijking met de andere

twee processen dus erg efficiënt. Een grote vraag in het veld is wanneer welk proces nu in actie komt? Wanneer wordt er gebruik gemaakt van TMEJ en wanneer komt een van de andere twee DSB-reparatie mechanisme aan bod? Vervolg onderzoek zal dit moeten gaan uitwijzen.

In hoofdstuk 4 laten we ook zien dat wanneer een mutant-worm Polymerase Theta mist, er veel meer DNA verloren gaat (±20.000 basenparen in plaats van 150 basenparen). We denken dat het evolutionair gezien voor een organisme voordeliger is om dan maar 150 basenparen te verliezen in plaats van 20.000 basenparen. Deze gedachte is mede gebaseerd op de ontdekking die we hebben gedaan dat over miljoenen jaren heen regelmatig hetzelfde soort deleties zijn ontstaan in het DNA van de worm. Dit zijn we te weten gekomen door bijvoorbeeld het genoom van een worm uit Engeland met het genoom van een worm uit Hawaii te vergelijken. Deze wormen hebben ooit dezelfde voorouder gehad, maar hebben een hele lange tijd afzonderlijk van elkaar geleefd en hebben onafhankelijk mutaties gekregen in hun DNA. Na vergelijking van het DNA van deze wormen zagen we dat op sommige plekken in de ene worm een G-quadruplex aanwezig was terwijl bij de andere de G-quadruplex plus ±150 naastgelegen basenparen verdwenen bleek te zijn. Het feit dat TMEJ dus al miljoenen jaren actief is in de worm, is een goede indicatie dat dit een belangrijk proces is.

Is TMEJ dan ook belangrijk in de mens en leiden G-quadruplexes ook tot deleties in het DNA bij de mens? Deze vragen heb ik geprobeerd te beantwoorden in **hoofdstuk 5**. Met verschillende methoden heb ik geprobeerd aan te tonen dat G-quadruplexes ook instabiel zijn in de mens. Net als in hoofdstuk 2 heb ik cellijnen gemaakt die rood dan wel groen kleuren na het verdwijnen van een G-quadruplex. Met deze cellijnen heb ik aangetoond dat ook in menselijke cellen G-quadruplexes instabiel kunnen zijn. Door het toevoegen van een G-quadruplex-bindend molecuul toon ik aan dat we de instabiliteit van G-quadruplexes kunnen verhogen. Daarnaast heb ik, hoewel preliminair, laten zien dat net als in de worm kleine deleties kunnen ontstaan in de buurt van G-quadruplexes. Of er een rol is weggelegd voor de menselijke variant van "strijkijzer" *dog-1* (in de mens heet dit eiwit FANCJ) bij het plat strijken van G-quadruplexes in de mens is helaas onduidelijk gebleven. Ook de vraag of polymerase Theta in de mens betrokken is bij het repareren van G-quadruplex-geïnduceerde schade, blijft vooralsnog onbeantwoord. Maar met behulp van de reeds ontwikkelde cellijnen kunnen deze vragen hopelijk snel worden beantwoord.

Hoe klinisch relevant is de data die is vergaard in hoofdstuk 4 en 5? Mensen die FANCJ missen lijden aan Fanconi Anemia. Dit is een zeldzame ziekte, waarbij de patiënten lijden aan onder andere bloedarmoede en kanker. Hoe deze ziektebeelden precies ontstaan is nog niet geheel duidelijk. Daarnaast is er voor deze patiënten nauwelijks een therapie voorhanden. Het onderzoek in hoofdstuk 4 laat zien dat wanneer *dog-1* in de worm afwezig is, een heleboel mis gaat op het gebied van G-quadruplexes en dat polymerase Theta een belangrijke rol hierbij speelt. Nu we dit beter begrijpen kunnen we kijken of FANCJ patiënten ook problemen hebben met G-quadruplexes en of we bijvoorbeeld drugs/medicijnen kunnen ontwikkelen die aangrijpen op G-quadruplexes of polymerase Theta. De wormen, vissen en cellijnen en bijbehorende methoden die beschreven staan in dit proefschrift bieden in ieder geval een uitstekende mogelijkheid om het effect van potentiële medicijnen veilig en snel te kunnen evalueren in het laboratorium, voordat er testen in patiënten gedaan worden.

Recentelijk zijn er duidelijke links gevonden tussen polymerase Theta en kanker. Bij bijvoorbeeld borstkanker- en eierstokkanker-patiënten is gebleken dat wanneer er teveel polymerase Theta aanwezig is, deze patiënten een kleinere kans hebben op het aanslaan van een chemokuur. Mede door het onderzoek dat is beschreven in hoofdstuk 4 hebben we nu een idee waarom dit zo is. Polymerase Theta is niet alleen goed in het aan elkaar plakken van DSB-en (de eerder beschreven breuken in het DNA) veroorzaakt door G-quadruplexes, maar ook van DSB-en ontstaan door andere processen. Zo zorgt een chemokuur ervoor dat er een heleboel toxische DSB-en ontstaan in snel delende cellen. Een chemokuur werkt dus voornamelijk op kankercellen aangezien deze continu delen. Echter, als al die breuken snel gerepareerd worden door polymerase Theta is de chemokuur niet meer effectief en worden de kankercellen niet uitgeroeid. We denken dus dat het belangrijk is om in dit geval in deze kankercellen polymerase Theta uit te schakelen, zodat de DSB-en niet meer hersteld kunnen worden en de cellen beter uitgeroeid worden tijdens een chemokuur. Kortom, mede door het onderzoek beschreven in hoofdstuk 4 zijn we erachter gekomen dat polymerase Theta een interessant drug-target is om effectiever kankercellen te doden. Zo zie je maar hoe onderzoek in een klein beestje als de rondworm toch klinisch erg relevant kan zijn!

DANKWOORD

Eindelijk is het zover: mijn proefschrift is af! Ik ben enorm trots op het eindresultaat en kijk met veel plezier terug op deze bijzondere tijd. Ik besef me terdege wat een voorrecht het is om wetenschap te mogen en kunnen bedrijven en om met zoveel interessante en intelligente mensen samen te werken. Nu is het tijd om de mensen te bedanken die hebben bijgedragen aan het tot stand komen van dit proefschrift.

Marcel, door jouw kalmte, inzicht, humor, positivisme en winnaars-mentaliteit creëer je een stimulerende omgeving voor een OIO om zich te kunnen ontwikkelen tot een zelfstandig wetenschapper. Ik ben je erg dankbaar dat je mij alle vrijheid en kansen hebt gegeven om mijn eigen onderzoek uit te voeren en me nooit beperkt hebt in het doen en laten van welke proef dan ook.

Beste Tijsterman-groep, mede door jullie was elke dag gezellig op het lab en geen proef te saai, zelfs niet wanneer er dagenlang door een microscoop getuurd moest worden op zoek naar bibberende wormen. Karin, Evelina, Jennemiek, en Evelien, het was erg fijn om jullie als collega te hebben. Kristy, bedankt voor alle hulp met de unc-22-assays en je ontvangst in Singapore. Marijn, Nick en Daphne, dankzij jullie kon er een fiets-trein gevormd worden die altijd de voorkeur genoot boven de NS-trein. Nick, thanks for the interesting conversations we had in the train or on the bike. Robin, hoofdstuk 4 zou significant anders zijn geweest zonder jouw expertise in bioinformatica en statistiek. Bennie, jouw passie voor de wetenschap is aanstekelijk en dank voor al je input. Ivo, mede door jouw hulp heb ik mooie plaatjes weten te maken van RAD-51 stainings. Jordi, thanks for helping me with the BD-pathway. Ron, leuk om na een tussenpose van een aantal jaren weer van je enthousiasme, grappen en werklust te hebben mogen genieten. Maartje, hopelijk laat pol θ je niet in de steek in de plant. Jane, bedankt voor al het werk dat je hebt verzet. Mede door jouw hulp zijn er twee mooie publicaties tot stand gekomen. Ook erg bedankt voor het nakijken van mijn hoofdstukken. Verder wil ik ook mijn studenten Andrea, Joya en Karli bedanken voor de bijdrage die ze hebben geleverd aan mijn onderzoek. Sophie, dank dat je al die jaren mijn klankbord wilde zijn voor alle perikelen die bij een promotie-traject naar boven komen. We zaten in vrijwel hetzelfde schuitje en dat heeft een mooie band geschept, ik ben dan ook heel blij dat je me tot op het laatste moment van dit traject wilt bijstaan.

Vervolgens wil ik al mijn andere collega's van de (voormalige) afdeling Toxicogenetica bedanken. Allereerst mijn kantoorgenoten Jaap, Anastasia, Poppie, Fabienne en Sophie voor de goede en gezellige atmosfeer in ons kantoor. Speciale dank ook naar mijn mede-AIO's en post-docs voor o.a. de leuke borrels in Lemmy's en AIO-retraite in Keulen. De Feco 2011 voor de memorabele dag in Den Haag en andere geslaagde borrels die we georganiseerd hebben. Wouter W, Antoine, Godelieve, Suming, Thomas en Bennie voor het bijstaan van de LUMC avond-maaltijden. Martijn voor de altijd bruikbare tips and tricks. Matthieu voor het immer paraat staan voor allerlei klusjes en bestellingen. Frans voor de talloze agar-platen. Ingrid voor de hulp met alle formulieren.

Verder wil ik binnen het LUMC Joop en Annelies bedanken voor de hulp met de confocal en Nico voor het sorteren van de talloze cellen. Sabine voor het uit de wind houden richting Leiden/Utrecht. Dirk, voor de nodige breaks overdag. Daarnaast wil ik Davy en Annemerie bedanken voor de goede verzorging van de zebravissen.

Ook ben ik een groot aantal mensen dank verschuldigd met wie ik heb samengewerkt tijdens mijn Hubrecht-periode. Pim, Ewart, Edwin, Rene, Jeroen Ba, Federico, Bas, Bert en Jeroen Bu, mede door jullie hulp is hoofdstuk 3 een mooi afgerond en gepubliceerd stuk geworden. Pau, Sandra and Puck, thank you for the pleasant and fruitful collaboration. The same holds true for Gijs and Henning from the NKI. Dank aan alle vrijdagmiddag-borrelaars, de Movember-clan en mijn mede-organisatoren van de AIO-retraite. Deze borrels, feesten en retraite staan in mijn geheugen gegrift! Also many thanks to all my former and new colleagues of Erasmus MC.

Uiteraard wil ik ook de mensen bedanken die niet direct met mijn werk te maken hebben gehad, maar wel voor de nodige ontspanning en momenten van relativering hebben gezorgd, altijd interesse in mijn werk hebben getoond en ervoor hebben gezorgd dat ik niet 24/7 in het lab zat. Allereerst mijn studievrienden. Ik waardeer het enorm hoe we al sinds 2000 al onze frustraties en successen binnen de wetenschap delen. Frank, het feit dat je vanwege mijn promotie komt overvliegen uit San Francisco zegt genoeg. Fantastisch dat je mijn paranimf wilt zijn.

Vervolgens wil ik de mannen van Bolwerk United bedanken, zaterdag blijft de mooiste dag! Mijn oud-huisgenoten Lotte en Marjolijn, voor hun belangstelling in de blauwe en groene wormen en vissen. Lotte, enorm bedankt voor de foto's die dit proefschrift hebben verrijkt! Mijn jaarclub en vrienden uit Venlo voor de vriendschap en de broodnodige afleiding.

Tot slot, lieve Lotte, wat ben ik blij dat ik jou heb ontmoet en wat is het fijn om met jou samen te zijn. Erg bedankt voor je geduld, zorgzaamheid, steun, interesse, vrolijkheid en liefde! Pa, ma, Stijn en Annebeth, Rolf en Chantal, Daan, Ties en de jongste telg, Stef, wat ben ik gelukkig met jullie. Een weekendje weg in Venlo of in Zeeland bij de familie Koole doet meer dan goed!

CURRICULUM VITAE

Wouter Koole was born in Arcen en Velden, the Netherlands on 16 May 1982. He attended the Valuas College in Venlo where he completed the Gymnasium programme and received his VWO diploma in June 2000. In September 2000, he started studying Biomedical Sciences at Utrecht University and acquired his Bachelor of Science degree in 2003. Later that year he enrolled in the Cancer Genomics and Developmental Biology Master's programme at the Utrecht University and graduated with merits in October 2006. As part of the Master's programme he completed a one-year rotation project in the lab of Prof. Dr. Ronald Plasterk at the Hubrecht Institute, the Netherlands. During this time he performed a genome-wide RNAi screen in *C. elegans* to identify novel genes that protect against alkylating damage. In November 2005 he started a second internship at Stanford University, Palo Alto, USA, in the lab of Prof. Dr. Roel Nusse, where he studied the role of Wnt-signaling in mouse and human embryonic stem cells. After his graduation in 2006, he worked in the Nusse-lab for another year to complete two of his projects, the results from which were subsequently published.

In September 2007, he started his Phd in the lab of Prof. Dr. Marcel Tijsterman at the Hubrecht Institute. In June 2009 his research continued in the department of Toxicogenetics, which later joined with the department of Human Genetics at the Leiden University Medical Center in Leiden, the Netherlands. During his PhD Wouter studied microsatellite and G-quadruplex instability by making use of various model organisms such as the nematode *C. elegans* and the zebrafish *Danio rerio*. The key results that were obtained during his PhD are described in this thesis.

In January 2014, Wouter was appointed a postdoctoral fellow in the lab of Dr. Derk ten Berge at the Erasmus Medical Center in Rotterdam, the Netherlands, where he investigates the use of adult liver stem cells for gene therapy.

LIST OF PUBLICATIONS

FANCJ promotes DNA synthesis through G-quadruplex structures.
Bosch PC, Segura-Bayona S, Koole W, van Heteren JT, Dewar JM, Tijsterman M and Knipscheer P.
(2014) *Embo J*, **33**, 2521-2533

A Polymerase Theta-dependent repair pathways suppresses extensive genomic instability at endogenous G4 DNA sites.
Koole W, van Schendel R, Karambelas AE, van Heteren JT, Okihara KL and Tijsterman M.
(2014) *Nature Commun.*, **5**, 3216

Mosaic analysis and tumr induciton in zebrafish by microsatellite instabiliy-mediated stochastic gene expression.
Koole W and Tijsterman M.
(2014) *Disease models & Mechanisms*, **7**, 929-936

A versatile microsatellite instability reporter in human cells.
Koole W*, Schäfer H*, Agami R, van Haaften G. and Tijsterman M.
(2013) *Nucleic Acid Research*, **41**, e158
* Shared First Authorship

A Broad Requirement for TLS Polymerases η and κ, and Interacting Sumoylation and Nuclear Pore Proteins, in Lesion Bypass during *C. elegans* Embryogenesis.
Roerink SF, Koole W, Stapel LC, Romeijn RJ and Tijsterman M.
(2012) *PLoS Genet.*, **8**, e1002800.

Embryonic stem cells require Wnt proteins to prevent differentiation to epiblast stem cells.
Ten Berge D, Kurek D, Blauwkamp T, Koole W, Maas A, Eroglu E, Siu RK and Nusse R.
(2011) *Nat. Cell Biol.*, **13**, 1070–1075.

Wnt signaling mediates self-organization and axis formation in embryoid bodies.
Ten Berge D*, Koole W*, Fuerer C, Fish M, Eroglu E and Nusse R.
(2008) *Cell Stem Cell,* **3**, 508-518
* Shared First Authorship

Identification of conserved pathways of DNA-damage response and radiation protection by genome-wide RNAi.
van Haaften G, Romeijn R, Pothof J, Koole W, Mullenders LHF, Pastink A, Plasterk RHA and Tijsterman M.
(2006) *Curr. Biol.*, **16**, 1344–1350.