THE UNIVERSITY
of EDINBURGH

Detection and mutational consequences of embedded ribonucleotides

Thomas Williams

PhD by Research

University of Edinburgh

2020

Declaration

This thesis has been composed entirely by me. Apart from where stated, the experiments were carried out solely by me. The work in this thesis has not been submitted for any other degree or professional qualification.

Thomas Williams

June 2020

Abstract

Mutation is a fundamental driver of evolution. It occurs non-randomly throughout the genome, influenced by factors such as chromatin architecture, DNA replication, transcription and repair. In human populations an important, but poorly understood, subset of mutations are short insertions and deletions (indels), the formation of which has traditionally been ascribed to deficiencies in mismatch repair to correct polymerase slippage events. However in *S. cerevisiae*, the formation of short 2 to 5 base pair deletions has been shown to arise as a consequence of the activity of Topoisomerase 1 (Top1) on DNA-embedded ribonucleotides. In normal cells, such ribonucleotides are the most common aberrant non-canonical nucleotides. They occur stochastically throughout the genome, mainly as a result of polymerase misincorporation. In this thesis I detail the development of a novel, nanopore based methodology to detect ribonucleotides embedded in DNA at single nucleotide resolution. I also describe the design and implementation of a highly sensitive reporter construct in *S.cerevisiae* to detect the Top1 dependent deletion signature. Transferring this reporter to HeLa cells, I show that the same deletion signature is also present in human cells, and confirm this through the orthogonal analysis of a whole genome sequencing mutation accumulation experiment in RPE1 cells. I identify the same mutation signature in *de novo* mutations from human populations, showing similarities between this mutational process and a newly described COSMIC (Catalogue of Somatic Mutations in Cancer) indel signature. This work demonstrates the presence of a novel mutational process in human genomes, distinct to that caused by polymerase slippage and deficient mismatch repair activity. This process may be an important cause of short deletions in human cells, and has implications for our understanding of the generation of genetic diversity, the formation of *de novo* mutations, and the development of cancer.

Lay Summary

Every time a cell divides, the DNA sequence that provides the instructions to the cell must be copied. This process is highly accurate, but not perfect, and thus new changes can arise. These new changes, which are passed on to subsequent generations of cells, are called mutations. Some of these mutations are responsible for inherited diseases, and others can lead to the development of cancer. A common error during the copying of DNA is the incorrect incorporation of ribonucleotide molecules. Ribonucleotides are the building blocks of RNA, and deoxyribonucleotides are the building blocks of DNA. RNA is similar to DNA, but has slightly different properties. Incorporated ribonucleotides makes the surrounding DNA more fragile, and in yeast can result in mutations: permanent changes to a DNA sequence. In this work I detail the development of a technique with which to identify the position of ribonucleotides incorporated into DNA. I also describe the development of a system that can efficiently detect ribonucleotide-induced mutations in yeast and mammalian cells. This allowed me to show for the first time that this mutation type also occurs in human cells. Understanding the process underlying the formation of such mutations will improve our ability to understand the origins of DNA changes in developmental disorders, cancer, and in human populations.

Acknowledgements

First of all, thank you to Martin (Reijns) and Martin (Taylor) whose curiosity and desire to tackle difficult scientific questions is infectious, and inspiring. From Andrew Jackson I have learned the benefits of drilling down to the heart of a question. Andy Bretherick helped me navigate many tricky computational challenges. Greg Kudla always left his office door open, and invariably offered excellent advice. The Taylor Lab (Lana Talmane, Craig Anderson, Kathryn Jackson-Jones, Rob Young, Juliet Luft and Harriet Kemp) were without exception enthusiastic and supportive and made me look forward to coming in to work. The Jackson Lab (Grainne Neary, Margaret Harley, Carol-Anne Martin, Olga Murina, Žygimantė Tarnauskaitė, David Parry, Patricia Heyn and Adeline Fluteau) were patient teachers, and their frequent acts of kindness helped me to juggle laboratory and family life.

Colleagues at the IGMM were always generous with their time and skilled advice: I'd like to thank Shelagh Boyle, Kiko Sanchez-Luque, MJ Kempen, Oscar Bedoya Reina, Neil Clark, Ilia Flyamer, Craig Nicol, Connor Warnock, Laura Murphy, Dickie Wood, Elisabeth Freyer, Laura Lettice, Alison Meynert, Tracy Ballinger, Graeme Grimes, Wendy Bickmore and any others that I have missed. The Technical Team at the IGMM provided professional service, always with a smile: thank you to Stephen Brown, Jeff Joseph, Stewart McKay, Sean O'Neil, Joan Flannigan, Pamela Stewart, Julie Morrison, Margaret McGurk, Stacey Thomson, Keith Rooney, Robbie Pineda and Gerald Dickson. Jamie Campbell and Fraser Millar listened to my progress over lunch; Nikki Hall provided in-house photography services; Paula Carroll was always on hand to eat a cinnamon bun.

Outwith the IGMM but still in Edinburgh, thank you to Alison Pidoux, Sander Granneman, Aine O'Toole and Amanda Warr for useful discussions. And outside Edinburgh thank you to Jared Simpson, Olaf Nielsen, Alison Gammie, Jessica Williams, Hannah Klein, Ruben van Boxtel and Matt Loose for answering emails and sharing materials, data and expertise.

Thank you to the Wellcome Trust for the funding to undertake this project. Thank you to my thesis committee: Chris Ponting and Sarah Walmsley, and to the ECAT directors who have helped create a great network of research focused clinicians in Edinburgh: Moira Whyte, Brian Walker, John Iredale and Neil Henderson. Without the ECAT administrator Jo Ness I wouldn't have been able to get anything done. And finally thank you to Lindsay, Fraser and Blaise for tolerating early starts and late finishes, and asking questions just as hard as the most demanding reviewer.

Table of Contents

List of figures

List of tables

Chapter 1 Introduction

1.1 DNA replication and ribonucleotides

1.1.1 Structure of DNA and RNA

DNA (deoxyribonucleic acid) is the building block of life; it provides the instructions for cells on how to grow and divide, and these instructions are passed from ancestral cell to descendent cell with incredibly high fidelity. DNA is composed of a nitrogenous base (one of adenine, cytosine, guanine and thymine), a deoxyribose sugar, and a phosphate backbone. The deoxyribose sugar is bound covalently to the base at its 1' carbon, and then to the upstream and downstream phosphates of adjacent nucleotides at its 5' and 3' carbons, respectively. The nitrogenous bases are classified as either purines (adenine, guanine) or pyrimidines (cytosine, thymine), and the purine-pyrimidine pairs adenine:thymine and guanine:cytosine connect the two strands of the DNA double helix through hydrogen bonds (Watson and Crick, 1953). Adenine:thymine pairs are held together by a double hydrogen bond, whilst guanine:cytosine are held together by 3 bonds, creating a stronger interaction (Figure 1).

In the central synthesis or dogma of cell biology, the main paradigm in the field since its formulation 50 years ago (Crick, 1970), information within cells is transmitted uni-directionally from DNA through the process of transcription to closely related messenger RNA (mRNA). This mRNA in turn through codons (a semi-redundant code of trinucleotides) is translated into a sequence of amino acids, the constituents of proteins. The complete collection of DNA present in a cell constitutes the genome, which in eukaryotes can vary widely in size. Eukaryotic genome can range in sizefrom the 12 million base pairs of the baker's yeast genome (*Saccharomyces cerevisiae)* to larger than the 6.4 billion base pairs of the diploid human genome (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001). Examples of very largegenomes include that of some salamander (an estimated 120 billion base pairs) (Sun *et al.*, 2012; Sclavi and Herrick, 2019) or plants such as *Paris japonica* (an estimated 150 billion base pairs)(Pellicer, Fay and Leitch, 2010).

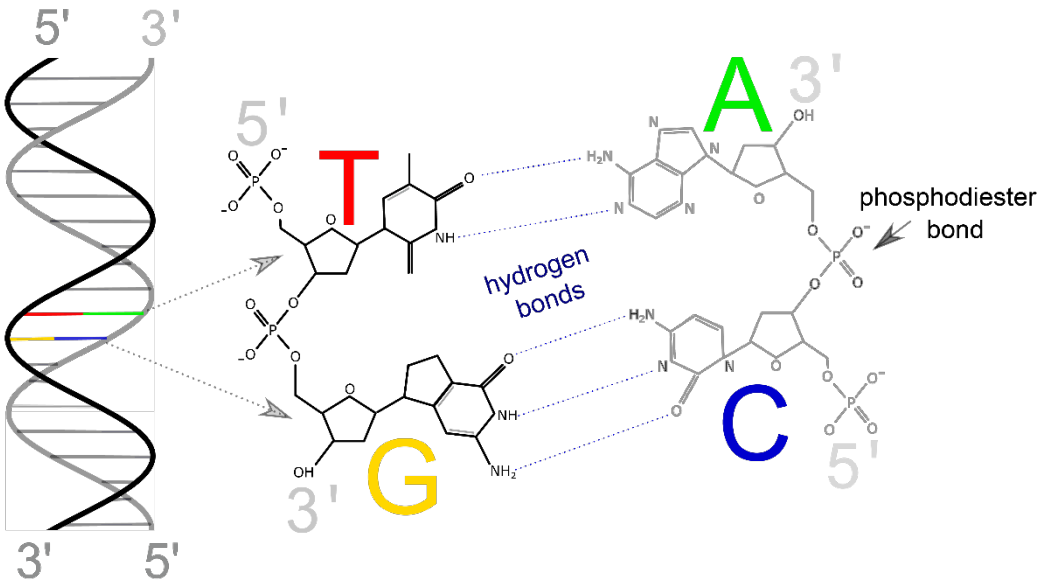


Figure 1.1. Structure of DNA. *In its simplest form deoxyribonucleic acid is composed of 2 nucleotide strands bound in an anti-parallel fashion into a double helix. Each nucleotide is composed of a pentose carbon sugar, a nitrogenous base (adenine, cytosine, guanine or thymine); and a phosphate group; the last of these bind together through covalent phosphodiester bonds to form a phosphate backbone. Nucleotides are bound to one another across the double helix with hydrogen bonds. Adapted with permission from* (Talmane, 2018).

There remains controversy about how DNA arose, and whether RNA (ribonucleic acid) preceded DNA, or both arose at the same time (Cech, 2012; Xu *et al.*, 2020). However, the structure of RNA and DNA is similar: both consist of a nitrogenous base, bound to a sugar and a phosphate group. RNA differs from DNA in two ways. Firstly, in the sugar in the sugar-phosphate backbone; in RNA this is ribose, rather than deoxyribose. Ribose differs from deoxyribose in having a hydroxyl group at the 2' C in the 5-carbon ring, rather than a hydrogen (Figure 1.2). This renders the ribose, and consequently any strand of RNA, up to 100,000 times more highly susceptible to hydrolysis than the equivalent DNA in representative physiological conditions (Li and Breaker, 1999). Secondly, they differ in the identity of one base, uracil, which takes the place of thymine in base pairing for RNA compared to DNA.

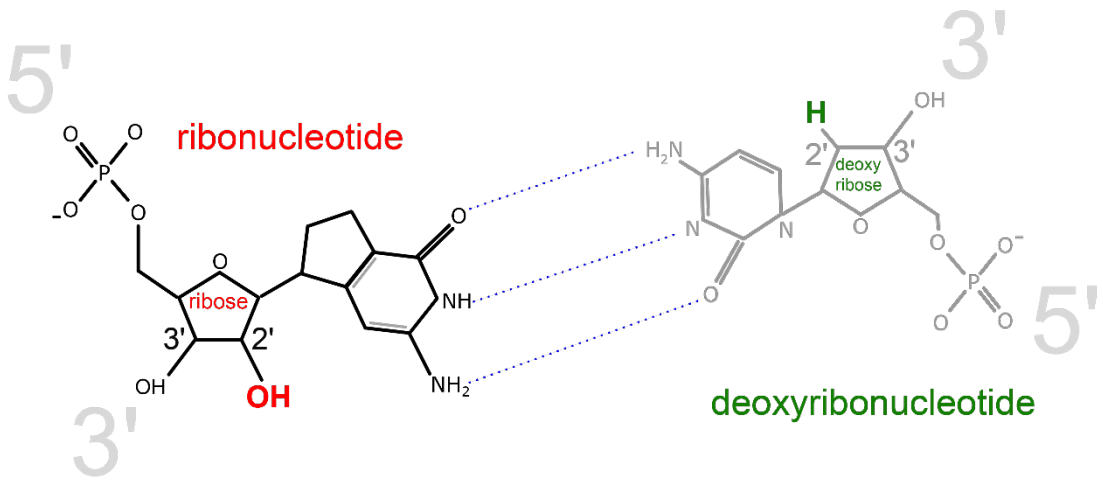


***Figure 1.2. Structure of RNA compared to DNA**. Ribonucleic acid (RNA) differs from de-oxyribonucleic acid (DNA) in the pentose sugar that forms part of the nucleotide. In RNA this is a ribose, with a reactive hydroxyl group (in red) at the 2' carbon of the pentose ring. In DNA there is a single hydrogen (in green) at the 2' carbon. RNA and DNA are able to bind to one another covalently through the sugar phosphate backbone (see Figure 1.1) or through the hydrogen bonds that connect the nitrogenous bases.*

1.1.2 Organisation of DNA and DNA replication

The nuclear genomes of eukaryotic organisms have further levels of DNA organisation beyond the secondary structure of the helix itself. Strands of DNA are wound around nucleosomes (Kornberg and Thomas, 1974), which are octameric complexes of proteins called histones (Luger *et al.*, 1997; Richmond and Davey, 2003). These nucleosomal structures are further organised into higher order chromatin structures (Figure 1.3) (Rao *et al.*, 2014). In order for the cell activities such as transcription to take place, DNA must be accessible as a single strand. This entails resolution of the structures of the chromosome and nucleosome, and breaking of the hydrogen bonds that bind the complementary purines and pyrimidines of the double helix.

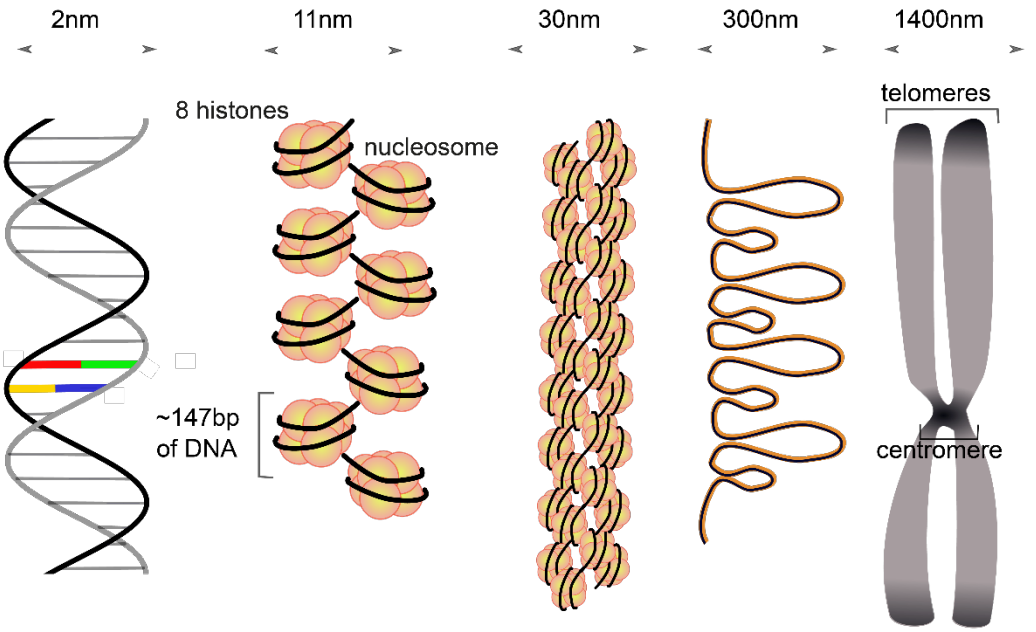


Figure 1.3. Structure of eukaryotic chromosomal DNA. *Approximately 147 base pairs of DNA wrap around 8 histone proteins to form a nucleosome. Nucleosomes are compacted to form 30 nm chromatin fibres, forming loops that are further compressed and coiled to form chromosomes at mitosis. These have a central centromere and distal telomeres. Adapted with permission from* (Talmane, 2018).

DNA is replicated by specialised DNA polymerases, the first of which was described in *E.coli* in 1958 by Lehman and colleagues (LEHMAN *et al.*, 1958), with eukaryotic equivalents subsequently described in *S. cerevisiae* (Chang, 1977). DNA polymerases synthesise a complementary strand of DNA from the template strand in order to allow DNA replication and subsequent cell division (Pursell *et al.*, 2007). This synthesis requires the unwinding of the DNA double helix and dissolution of the hydrogen bonds that hold the opposing strands of DNA together, so that polymerases can act on a single template strand of DNA. In the nuclear genomes of eukaryotic organisms, due to larger size of chromosomes and their generally linear nature, a whole chromosome cannot be unwound at once. Therefore DNA replication commences at multiple origins of replication along a chromosome (Brewer and Fangman, 1991; McGuffee, Smith and Whitehouse, 2013; Clausen *et al.*, 2015; Reijns *et al.*, 2015), where the double stranded DNA helix is unwound and synthesis of complements to the Watson (forward) and Crick (reverse) strand takes place bidirectionally (Figure 1.4). Commencing at origins of replication, the DNA double helix is unwound by a DNA helicase, in eukaryotes composed of a complex of 6

proteins called the MCM complex (Tye, 1999; Yeeles *et al.*, 2017; Meagher, Epling and Enemark, 2019). DNA replication is initiated by the polymerase α-primase complex (Perera *et al.*, 2013). The primase first synthesises a short 5-10 nucleotide RNA primer; polymerase α then takes over and synthesises a short tract of DNA (Hu, Wang and Korn, 1984).

DNA synthesis proceeds in an obligate 5' to 3' direction due to the additive nature of the action of the DNA polymerases: new nucleotides are added to the free 3'OH of the strand being synthesised. This means that on the leading strand replication can proceed uninterrupted until encountering activity from the adjacent origin of replication. However, for the opposing strand, replication must occur in a discontinuous processing to form a continuous DNA strand, whereby the gaps between fragments (Okazaki junctions) are ligated together (Sugimoto, Okazaki and Okazaki, 1968). The current understanding of this process in eukaryotes is that in the leading strand polymerase (Pol) ε continues the activity of polymerase α, and that polymerase δ has this role in the lagging strand (Pursell *et al.*, 2007; Nick McElhinny *et al.*, 2008; Larrea *et al.*, 2010; Lujan *et al.*, 2012; Reijns *et al.*, 2015) (Figure 1.4). It is important to note however that there is also dissenting work suggesting that polymerase δ follows on from polymerase α in both strands, either just at the origins of replication (Zhou *et al.*, 2019) or throughout replication (Johnson *et al.*, 2015).

In lagging strand synthesis, when the previously synthesised Okazaki fragment is encountered, its 5' end is displaced by the processive Pol δ which eventually dissociates. The displaced 5' sections incorporating the Pol α replicated DNA (10-30 nucleotides long) (Liu *et al.*, 2017) were thought to be processed and fully removed by the endonucleases FEN1 and DNA2 (Maga *et al.*, 2001; Perera *et al.*, 2013). As Pol α lacks an intrinsic proof reading capacity and is more error prone that Pol δ (Kunkel *et al.*, 1989) , the removal of Pol α synthesised DNA could be desirable in removing potential mutations. Most of the Pol α synthesised DNA is removed during this process, but work in mutagenesis assays showed that some of this was likely to be retained (McElhinny, Kissling and Kunkel, 2010; Lujan *et al.*, 2014). This was confirmed

in later work (Reijns *et al.*, 2015), which showed that DNA replicated by Pol α makes a small but significant contribution to the genome (an estimated 1.5%).

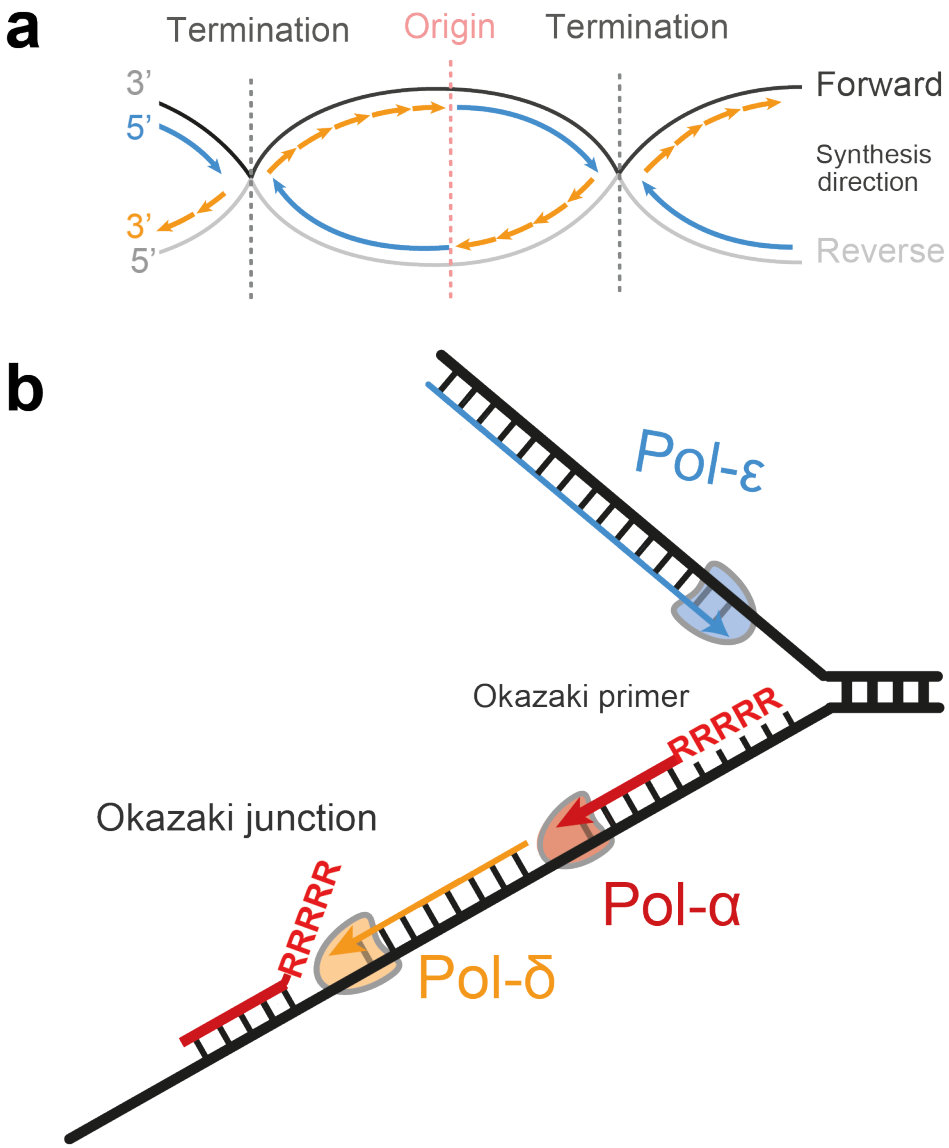


Figure 1.4. DNA replication. *a) Replication initiates bi-directionally at a number of origins across the genome. Nucleotides can only be added in a 5' to 3' manner by the main replicative polymerases α, δ and ε. This means that the leading strand can be synthesised continuously (blue arrows) until it encounters activity from the adjacent origin of replication, whereas the lagging strand is synthesised in a discontinuous fashion (yellow arrows). b) Activity of the main replicative polymerases at the replication fork. Pol ε synthesis the leading strand in a continuous manner. On the lagging strand, DNA synthesis is initiated by primase-polymerase α, which creates a short segment of RNA (the Okazaki fragment) as a primer for DNA synthesis. Pol α hands over to pol δ, which then displaces the downstream pol α fragment, including the ribonucleotide Okazaki fragment and the Okazaki junction. Figure adapted with permission from* (Reijns *et al.*, 2015).

1.1.3 Causes and frequency of incorporation of ribonucleotides into S. cerevisiae genomes

Whilst Watson and Crick speculated in their original paper on the double helix nature of DNA (Watson and Crick, 1953) that “it is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact”, it has become clear in recent years that RNA is frequently misincorporated into genomic DNA in eukaryotic genomes.

In addition to the Okazaki junction primer tracts synthesised by Pol α described above, work in RNase H2 knockout strains of *S. cerevisiae* has shown that single ribonucleotides are frequently incorporated into eukaryotic DNA by Pol α, δ and ε during replication (Nick McElhinny, Watts, *et al.*, 2010). This is likely driven by the much higher concentrations of ribonucleotide triphosphates (rNTPs, RNA precursors) than nucleotide triphosphates (dNTPs,DNA precursors) in the cell nucleus. It is estimated that in *S. cerevisiae* there is a 36- 190 fold molar excess of the four rNTPs compared to their corresponding dNTPs (Nick McElhinny, Watts, *et al.*, 2010). This excess of rNTPs relative to dNTPs means that despite enzyme selectivity for dNTPs they are occasionally misincoporated by the replicative polymerases. These replicative polymerases are able to distinguish dNTPs from rNTPs, due to the structure of a “steric gate” that selects nucleotides for incorporation into DNA (Brown and Suo, 2011). Here, an active side residue, usually one with a bulky side chain, collides with the 2’OH group on the ribose of an incoming rNTP. This is a highly efficient process, but polymerases appear to stochastically misincorporate rNTPs. The rate at which this occurs at least partly correlates with the concentration of rNTPs relative to dNTPs. Such embedded ribonucleotides are subsequently removed by ribonucleotide excision repair (RER), a process initiated by cleavage by an enzyme called RNase H2 (Lazzaro *et al.*, 2012; Sparks *et al.*, 2012; Williams, Lujan and Kunkel, 2016). In eukaryotes, RNase H2 is a complex containing 3 subunits (Rnh201, Rnh202, Rnh203 in yeast (Jeong *et al.*, 2004); RNASEH2A, RNASEH2B, RNASEH2C in higher eukaryotes) (Crow *et al.*, 2006) that functions to cleave single and multiple rNMPs (ribonucleoside

monophosphates) embedded within DNA. Loss of any of these subunits results in both *S. cerevisiae* and high eukaryotes results in loss of function of the entire complex. Subsequent to cleavage with RNase H2, a nick translation reaction occurs, allowing strand displacement synthesis carried out by a DNA polymerase (likely polymerase δ). The resulting ribonucleotide containing flap is then excised by the endonucleases Fen1 or Exo1, similar to the processing of Okazaki fragments, and the gap re-ligated by DNA ligase I (Figure 1.5) (Sparks *et al.*, 2012; Williams, Lujan and Kunkel, 2016).

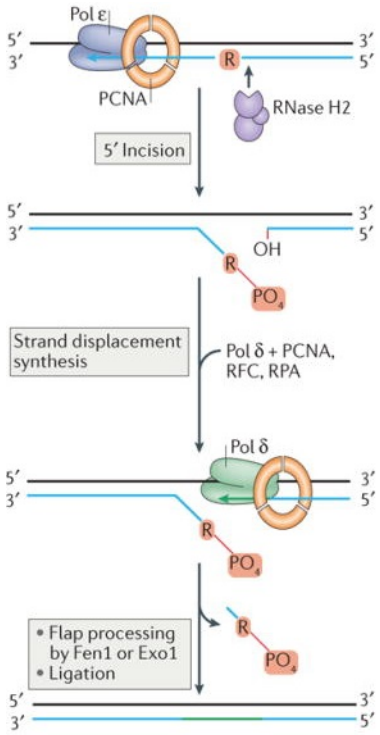


Figure 1.5. Ribonucleotide excision repair. *Repair is initial when RNase H2 incises at the 5' end of an embedded ribonucleotide. This is followed by a nick translation reaction, where the nicked strand is displaced, allowing Pol δ to bind and fill the gap. The flap generated by strand displacement is then processed by flap endonuclease 1 (Fen1) or exonuclease 1 (Exo1), followed by ligation. Figure reproduced with permission from* (Williams, Lujan and Kunkel, 2016)

In vitro work using purified yeast DNA polymerase complexes and physiological concentrations of dNTPs and rNTPs has demonstrated that the rates of ribonucleotide incorporation are dependent on the polymerase, with polymerase α being most likely to incorporate ribonucleotides (1 per 625

nucleotides), ε (1 per 1,250) and δ (1 per 5000) (Nick McElhinny, Watts, *et al.*, 2010) (Figure 1.6).

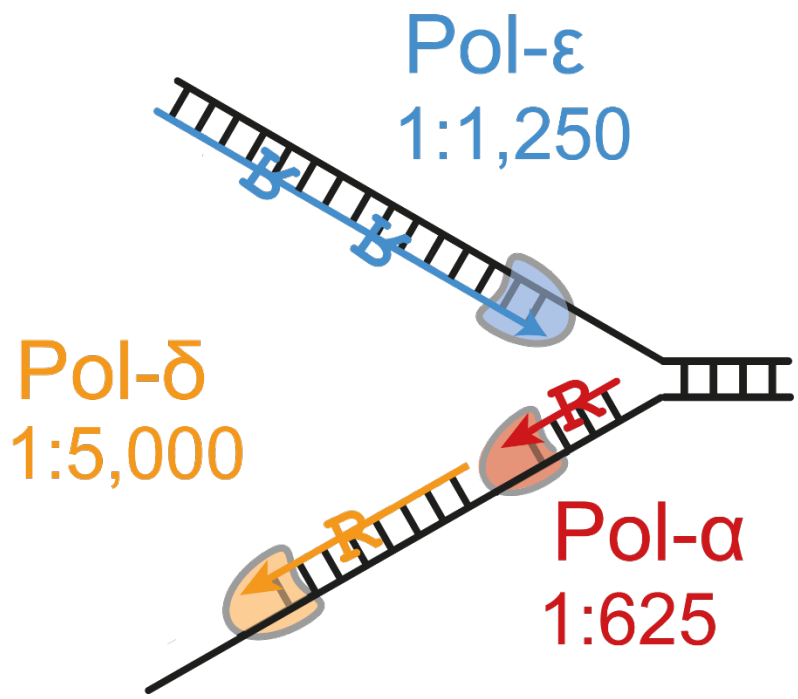


Figure 1.6. Rates of misincorporation of ribonucleotides by replicative polymerases. *Polymerase α, which lacks intrinsic 3' to 5' proofreading exonuclease activity, is the most likely to misincorporate ribonucleotides, followed by polymerase ε and polymerase δ. Data from S. cerevisiae, presented in (Nick McElhinny, Watts,* et al.*, 2010), figure adapted with permission from* (Reijns *et al.*, 2015).

Overall, an estimated 13,000 rNTPs are incorporated into the *S. cerevisiae* genome each round of cell division (Williams *et al.*, 2013). The vast majority of ribonucleotides misincorporated into *S. cerevisiae* genomic DNA are subsequently removed by RER. Current approaches to detect genome embedded ribonucleotides are based on alkali gels (Nick McElhinny, Kumar, *et al.*, 2010; Lujan *et al.*, 2013), nick translation assays (Rumbaugh *et al.*, 1997), or next generation sequencing (Clausen *et al.*, 2015; Daigaku *et al.*, 2015; Ding *et al.*, 2015; Koh *et al.*, 2015). The first two methods allows an estimation of the frequency of ribonucleotide incorporation at a genome wide level, but does not provide the resolution required to identify the location of those ribonucleotides. Next generation sequencing approaches have the capacity to identify genome embedded ribonucleotides as single base pair

resolution, but cannot directly quantify the frequency of incorporation, discriminate single ribonucleotides from tracts of ribonucleotides (such as Okazaki primers), or identify phasing of ribonucleotides along a single strand of DNA. There is therefore considerable interest in developing techniques that might be able to circumvent these limitations, and be applied to detecting non-canonical nucleotides more generally. Such a technique would permit insights into the potential physiological function of genome embedded ribonucleotides, and allow further investigation of fundamental cellular processes such as DNA replication.

1.1.4 Causes and frequency of incorporation of ribonucleotides into mammalian genomes

In addition to work in *S. cerevisiae*, it is clear that ribonucleotides are also incorporated into the genomes of mammals. Work by the Jackson lab (Reijns *et al.*, 2012) in mouse embryonic fibroblasts has shown that an estimated 1 million ribonucleotides are incorporated into the genome of each replicating mammalian cell, making them the most common aberrant nucleotide in nuclear genomic DNA. Again, here the predominant cause of ribonucleotide misincorporation is inferred, from the direct evidence in *S. cerevisiae* (Nick McElhinny, Watts, *et al.*, 2010), and the use of hydroxyurea in mammalian cells to modify dNTP:rNTP ratios (Reijns *et al.*, 2012) to be misincorporation by replicative polymerases. In addition, there are other polymerases implicated in the non-homologous end-joining repair pathway of double stranded breaks that incorporate high frequencies of ribonucleotides (Pryor *et al.*, 2018), including polymerase μ (Nick McElhinny and Ramsden, 2003) and terminal deoxynucleotidyl transferase (TdT) (Martin *et al.*, 2013).

Two recent studies, building on earlier work (Yang *et al.*, 2002; Reijns *et al.*, 2012) have examined the incorporation of ribonucleotides into mammalian *mitochondrial* genomes (Berglund *et al.*, 2017; Moss *et al.*, 2017). One of these demonstrated the incorporation of ribonucleotides into mitochondrial DNA in mouse tissues (Moss *et al.*, 2017), and the second in cultured human

cells (Berglund *et al.*, 2017). The second study estimated that approximately 36 and 54 ribonucleotides were incorporated into HeLa cells and fibroblasts respectively during each round of replication in the 16,569 bp human mitochondrial genome. In both cases, the cause of ribonucleotide incorporation is suggested to be *misincorporation* by the main polymerase in mitochondrial DNA replication, Polymerase γ. In humans, mutations in the *MPV17* gene cause hepato-cerebral disease, characterised by early progressive liver failure and neurological abnormalities, hypoglycemia and increased lactate in body fluids, a disease process driven by mitochondrial dysfunction (Mandel *et al.*, 2001; Spinazzola *et al.*, 2006). The pathology underlying this disorder is associated with aberrant and increased ribonucleotide misincorporation into the mitochondrial genome (Moss *et al.*, 2017). Therefore a technique to quantitatively identify sites of embedded mitochondrial ribonucleotides, would allow a more general investigation of the mechanisms underlying the consequences of ribonucleotides.

1.1.5 Potential physiological roles of ribonucleotide incorporation into genomes

The realisation that the steric gate within a polymerase was a potential determinant of the frequency of ribonucleotide misincorporation led researchers to introduce point mutations immediately adjacent to this steric gate (a tyrosine at position 645) to examine the effects of this on ribonucleotide incorporation. In Pol ε, they either introduced a point mutation at position 644, substituting the methionine (M) at this position in wild type for a leucine (M644L) or a glycine (M644G) (Nick McElhinny, Kumar, *et al.*, 2010). They found that a glycine (which is less bulky than the methionine) at this position led to an 11-fold *increase* in the rate of misincorporation of rNTPs compared to wildtype. They also found that replacement with a leucine led to a 3-fold *reduction* in the rate of rNTP misincorporation. The methionine at position 644 is highly conserved in eukaryotes (Pursell and Kunkel, 2008). Given the known deleterious consequences of ribonucleotide misincorporation (detailed below), this raises the question of why no organism has a Pol ε that incorporates fewer rNTPs, and the possibility that incorporated ribonucleotides may serve a

physiological function, either in relation to Pol ε activity, or separate to this. Three possible examples of the physiological roles of genome-embedded ribonucleotides are outlined below.

1.1.5.1 Mismatch repair

Two studies in 2013 found that incorporated ribonucleotides may serve a function in facilitating mismatch repair (Ghodgaonkar *et al.*, 2013; Lujan *et al.*, 2013). Mismatch repair (MMR) is an evolutionarily conserved DNA repair pathway that corrects DNA replication errors: base substitution mismatches, and insertion-deletion mismatches, on the newly synthesised strand (Kunkel and Erie, 2005). Defects in MMR in humans have been extensively studied and are associated with an increased risk of cancer due to a mutator phenotype (Boland and Lynch, 2013). On the lagging strand, the frequent 5' DNA ends between Okazaki fragments serve as a marker and entry point for the MMR machinery (McElhinny, Kissling and Kunkel, 2010).

The studies that investigated a potential physiological role of ribonucleotides in mismatch repair found that cleavage of ribonucleotides by the enzyme RNase H2 appeared to serve a similar function on the leading strand. They found that in the absence of RNase H2, the rate of mismatch repair specific mutations (single base transitions and indels) increased, in a manner specific to the leading strand (Lujan *et al.*, 2013). The authors hypothesised that RNase H2 may nick the DNA backbone at the positions of newly incorporated ribonucleotides, allowing these nicks to mark the newly synthesised strand, and as entry points for mismatch repair machinery. This hypothesis was supported by in vitro work with mammalian cell extracts (Ghodgaonkar *et al.*, 2013), where deficiency of RNase H2 in the cells lines used resulted in a reduced efficiency of the MMR machinery, which was rescued by the addition of recombinant human RNase H2.

1.1.5.2 Non homologous end joining

Another physiological role for genome embedded ribonucleotides is in non-homologous end joining (NEHJ), a mechanism that preserves genome stability by ligating the ends of broken chromosomes together. This process involves

the use of Polymerase μ or TdT (referred to above) to process the ends of the broken chromosomes for ligation. Experiments using DNA cut with CRISPR-Cas9 in mouse embryonic fibroblasts showed that NHEJ was promoted by the addition of excess levels of riboguanine triphosphate (rGTP), whereas it was inhibited by the lack of a ribonucleotide substrate (Pryor *et al.*, 2018). The authors of this paper showed that the repair of broken strands consisted of an initial first step of repair of the first strand with ribonucleotides, repair of the second strand with deoxyribonucleotides, and a subsequent RNAse H2 dependent excision of the ribonucleotides embedded when the first strand was repaired.

1.1.5.3 Mating type switch in Schizosaccharomyces pombe

In addition to the role of genome embedded ribonucleotides in mismatch repair and NHEJ, in the model organism *Schizosaccharomyces pombe* (*S. pombe*, fission yeast), ribonucleotides have been shown to have a key physiological role (Sayrac *et al.*, 2011). In *S. pombe*, cells alternate between a P and M mating type (Miyata and Miyata, 1981). In the haploid state, cells can exist as P and M types and clonally grow through mitotic division. However, when the organism encounters nutritional stress, cells of opposite mating type form a diploid zygote by fusing together. This diploid zygote then enters meiosis to produce 4 haploid spores, which can germinate when nutritional conditions improve to produce dividing haploid cells (Beach and Klar, 1984). The change in mating type occurs due to a replication-coupled recombination event that transfers genetic information from one of two silent donor loci (*mat2P* or *mat3M*) into the mating-type determining *mat1* locus. It is believed that this mating type switch (which occurs in 1 in 4 cells) is mediated by two consecutive ribonucleotides, possibly part of a retained Okazaki fragment, present in 25% of cells at a well-defined location on chromosome II (Vengrova and Dalgaard, 2006). These two ribonucleotides are believed to lead to stalling of the replication fork, which leads to a recombination event mediated by homology arms upstream and downstream of the mating type cassettes (Vengrova and Dalgaard, 2006; Yamada-Inagawa, Klar and Dalgaard, 2007; Sayrac *et al.*, 2011). This role of a retained ribonucleotide as a switch to control cellular

differentiation raises the possibility that similar, yet to be described ribonucleotide marks may also have a functional role in cellular development in higher eukaryotes (Dalgaard, 2012). However, in addition to the potential physiological roles of genome-embedded ribonucleotides, they can also have deleterious consequences, both for genome stability (demonstrated in both *S. cerevisiae* and mammalian genomes)(Conover *et al.*, 2015; Zimmermann *et al.*, 2018) and generating mutations (demonstrated in *S. cerevisiae*) (Williams *et al.*, 2015, 2019).

1.2 Mutation as a driver of evolution, inherited disease and cancer

The word mutation derives from the Latin *mutare*, to change. It has been used in the English language since the 17th century as a word to describe change, in a variety of contexts including natural or divine law, music and politics (Condit *et al.*, 2002). Darwin used it in On the Origin of Species (Darwin, 1859) to refer to changes in species, but it appears to have first been used in the currently understood genetic sense in 1894 (Condit *et al.*, 2002), and formalised as a biological term by De Vries at the turn of the 20th century (de Vries, 1909). Today it is defined by the Oxford English Dictionary as “a process in which the genetic material of a person, a plant or an animal changes in structure when it is passed on to children”, a definition in keeping with its use in biology (with the important caveat that the concept of “children” also apply to a descendent cell, to allow the concept to apply to unicellular organisms or cancer cells)

Mutations are a fundamental driver of evolution: they are how genetic variation is generated, to be subsequently distributed through populations by migration (Ramachandran *et al.*, 2005; Auton *et al.*, 2015), selection (Pollard *et al.*, 2006; Lindblad-Toh *et al.*, 2011) or drift (Pedersen *et al.*, 2017; Rivas *et al.*, 2018). But in addition to its role in evolution, mutagenesis also plays a fundamental role in causing human disease (Prohaska *et al.*, 2019): germline inherited

genetic variants can affect normal cell function, either unique to an individual or transmitted through families (Martin *et al.*, 2018), or acquired somatic mutations can lead to the development of cancer (Hanahan and Weinberg, 2000, 2011). There has therefore been considerable interest in understanding the molecular mechanisms that underlie mutagenesis. Although often the focus has been on evolution or disease in humans; many of these studies, due to constraints in studying humans or human cells, taken place in less complex model organisms such as *E.coli*, *S. cerevisiae* or *C. elegans*. Below I outline the main types of mutations seen in human cells and populations, their importance in evolution, genetic disorders and cancer, and postulated mechanisms.

1.2.1 Types of mutations and significance of short deletions

Mutations take place at a variety of scales, from single base pair changes to losses of entire chromosomes. The most common mutations seen in humans (compared to one another, and to closely related species) are single nucleotide polymorphisms or SNPs (Auton *et al.*, 2015; Eichler, 2019). Here, the DNA sequence is changed at a single position, so that (for example), a C in the ancestral cell is converted to a T in the descendent cell. In humans, an estimated 90% of mutations in the germline take the form of SNPs, 9% are small insertions or deletions, and 1% large structural changes (Montgomery *et al.*, 2013). Allowing for the fact that a number of mutations go undetected due to repetitive sequences, the average newborn child is expect to have ~100 mutations compared to their parents (Lynch, 2016) .

As outlined above, insertions anddeletions (indels) are estimated to make up around 10% of inherited mutations (Lynch, 2016). The biggest group within these are short indels (generally classified as < 50 base pairs) (Eichler, 2019). Structural variants (mutations >50 bp in length) are less common. Losses and gains of entire chromosomes can occur (for example, in Turner's syndrome or Down syndrome), but unless part of a balanced translocations (where there is

no overall gain or loss of chromosomal material) are rarely transmitted to future generations. Most new or *de novo* mutations do not have significant functional consequences, as they are most commonly located in non-coding regions of the human genome, or represents synonymous changes that do not affect amino acid sequence. Even if a mutation occurs in a coding sequence, it may not impact on the function of a protein, and therefore will not lead to a change in phenotype. However, a small percentage of mutations lead to phenotypic changes, and in this way can drive evolution, through a variety of mechanisms.

Indels, and short indels in particular, are important in the genome evolution of all organisms, not just humans. They can lead to changes in proteins by altering gene length by disrupting reading frames with the introduction of new stop codons, or by changing the amino acid sequence of the protein as a whole (Vakhrusheva *et al.*, 2011). Indels can also indirectly lead to mutations, with associations between indels and single base transitions/transversions in regions flanking them (Tian *et al.*, 2008). Finally, they may lead to knock on evolutionary adaptions in other parts of proteins affected by an initial indels (Leushkin, Bazykin and Kondrashov, 2012). Mutations in previously noncoding DNA may also lead to the formation of new protein coding genes (Zhang *et al.*, 2019), and indels in regulatory sites such as promoters and enhancers can also lead to phenotypic change (Young *et al.*, 2015).

Although insertions, deletions and structural changes are much less common than SNPs in humans, they are much more likely to result in disease. A recent analysis of the Human Gene Mutation Database (Eichler, 2019) (www.hgmd.cf.ac.uk/) suggested that 34% of all disease causing variation is made up of variants that are larger than a single base pair substitution; whilst short indels are implicated in one quarter of all Mendelian disorders. Short indels within open reading frames can cause disease by disrupting protein coding genes, by leading either to missense mutations, with a change of the sequence of amino acids in the protein, to a nonsense mutation with premature truncation of the translated protein, or to loss of protein expression due to nonsense-mediated decay (Pellicer, Fay and Leitch, 2010). Outside open

reading frames, indels may disrupt noncoding regulatory sequences such as promoters or enhancers, leading to disease (Spielmann and Mundlos, 2016).

As well as influencing evolution and causing inherited human disease, indels can contribute towards the development of cancer. Cancer is rarely triggered by a single event; it is more frequently is due to a build-up of a sequence of mutations in key genes that regulate the cell cycle, which once disrupted can lead to clonal expansion and with further dis-regulation the invasive features of cancerous cells (Hanahan and Weinberg, 2000, 2011). Compared to SNPs, indels are over-represented in mutations that form part of the sequential mutational steps that leads to the eventual development of malignancy (Yang *et al.*, 2010). Short indels may lead to the activation of oncogenes such as EGFR (Shigematsu and Gazdar, 2006), where short deletions lead to the activation of the receptor and increased tyrosine kinase activity (Kumar *et al.*, 2008; Gazdar, 2009). Conversely, frameshift or truncating mutations may lead to the inactivation of key tumour suppressors genes such as BRCA1/2 (Rebbeck *et al.*, 2015), and to a higher risk of malignancy.

1.2.1.1 Established causes of short deletions in human genomes

Traditionally, establishing the causes of different mutation types has been based on inference from experiments in model organisms (Shcherbakova and Kunkel, 1999), matching known exposures to tumour types (Jamal-Hanjani *et al.*, 2017), and to enrichment of certain mutation types amongst carriers of particular genetic variants (Boland and Lynch, 2013). In recent years however this field has been transformed by the generation of large quantities of whole exome and genome sequences from cancer samples. Here, sequencing of tumour samples and paired normal cells from the same patients has allowed calling of new, cancer specific somatic mutations and in depth investigation of indels as well as SNPs, and the development of the concept of cancer mutational signatures (Alexandrov and Stratton, 2014). This is the idea that the underlying mutagenic process that has led to malignancy has a characteristic pattern of mutations. This mutational signature may be due to

extrinsic mutation-inducing agents (mutagens) such as UV light or tobacco smoke, or cell intrinsic processes , for example as a result of defective DNA repair. Over the past decade exome and whole genome sequences from many thousands of cancer samples have been analysed to delineate a large number of such mutational signatures (Alexandrov *et al.*, 2020). These mutational signatures have captured distinctive patterns of indels for which a cause has been identified, and also highlighted others for which the aetiology remains unknown (see Chapter 5). Some signatures delineated encompass what was already known, including the distinctive 1 bp indels in mismatch repair deficient tumours (e.g. the colonic or endometrial cancers that occur as a result of Lynch syndrome (Boland and Lynch, 2013), and longer deletions in areas of microhomology seen with the loss/inactivation of the BRCA1/BRCA2 genes (Davies *et al.*, 2017). In other instances observations in cancer samples have been supported by evidence from mutation accumulation studies in human cell lines. A study published in 2017 took a colonic cell line where the researchers had inactivated MLH1, a key gene in the mismatch repair pathway. They found an increase in deletions in 1 bp deletions in long runs of As or Ts, matching what had been observed in cancer studies (Drost *et al.*, 2017).

However, up until the advent of large scale whole genome sequencing of cancer samples, the understanding of causes of short deletions in human genomes has relied largely on inference from model organisms. Most studies to date have postulated that the main cause of inherited short indels is slippage as the main replicative polymerases copy DNA, followed by a failure of mismatch repair machinery to detect and correct these errors (Taylor, Ponting and Copley, 2004; Montgomery *et al.*, 2013; Morganella *et al.*, 2016).

A mechanism for the formation of short insertions or deletions in repetitive sequences was first proposed in 1966, from studies on the virus phage T4 in *E.coli* (Streisinger *et al.*, 1966). The authors investigated frameshift mutations in the T4 lysozyme gene. T4 lysozyme is a monomeric protein of 164 amino acids, which liberates virions from their bacterial host by hydrolysing peptidoglycans in the bacterial cell wall (Poteete and Hardy, 1994). Mutations

in the lysozyme gene can be induced by the addition of proflavine, and identified by rescuing mutants by adding an extrinsic source of lysozyme (in this case, egg white). As genetic sequencing was not available at this point, they inferred the amino acid sequence from trypsinisation and chromatography of the lysozyme protein, and from this the underlying genetic sequence. They found that the most common mutations induced by proflavine in the T4 lysozyme coding sequence were frameshifts, most commonly indels of of 1 or 2 base pairs. Mutations were most commonly found in repetitive sequences of a single base or base doublet. Streisinger and colleagues proposed these being due to a slippage event leading to these short indels (Figure 1.7), where the mispairing of bases at a repeating sequence, and consequent extension of the mispaired segment, led to the addition or deletion of bases upstream of the mispaired sequence.

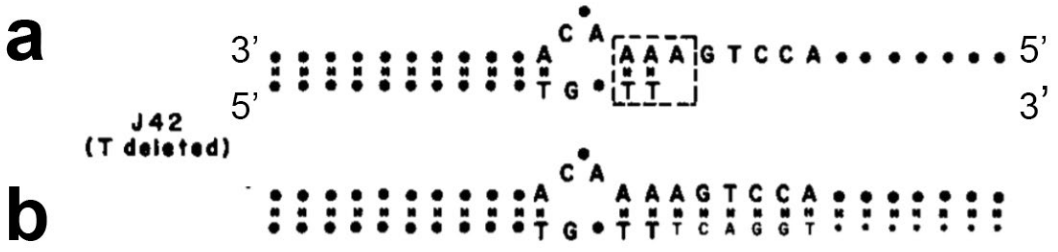


Figure 1.7. A general mechanism for frameshift mutations.*As DNA is synthesised in a 5' to 3' manner from the template strand, the replicative polymerase (area of activity outlined in dashed box) may slip as it proceeds across a repetitive sequence: in this example an AAAA tract. If the polymerase slips so that the template strand bulges out opposite the newly synthesised complementary strand, this will result in a deletion. b) The newly synthesised complementary strand has a T deletion relatively to the template. Figure adapted with permission from* (Streisinger *et al.*, 1966)*.*

This mechanism for the generation of short indels has subsequently been supported by biochemical evidence showing that misaligned DNA bases form extrahelical loops and intrahelical bulges with extra bases stacked within the helix (Joshua-Tor *et al.*, 1992), but that do not disrupt the overall structure of the double helix and are likely to allow ongoing synthesis of the complementary strand. This inference is supported by a study which was able to co-crystallise and visualise using crystallography a DNA polymerase in complex with a misaligned substrate (Garcia-Diaz *et al.*, 2006). Finally, there is observational

evidence showing increased mutation rates in repeat sequences in yeast and humans (Taylor, Ponting and Copley, 2004; Baptiste, Jacob and Eckert, 2015) which supports the findings originally seen in T4 phage.

Subsequently, two variations to this model have been proposed to explain short indels within repetitive sequences: the dNTP-stabilised and the misincorporation misalignment model. In the dNTP-stabilised scenario, an incoming dNTP is incorrectly paired to a nucleotide upstream of the correct position on the template strand; ongoing synthesis of the complementary strand downstream of this incorrectly paired dNTP results in a deletion in the newly synthesised strand (Figure 1.8, central panel). In the misincorporation-misalignment model, an incorrect base is inserted by the polymerase; this is proposed to slow down the catalytic rate of the polymerase and allow realignment with the correct neighbour (if a subsequent base happens to be the appropriate base pair (Figure 1.8, right panel). In support of this latter model, an in vitro study showed that template nucleotides were preferentially lost when they had as a 5' neighbour a nucleotide complementary to the dNTP provided in excess (Bebenek and Kunkel, 1990).

Another mechanism put forward for the formation of short indels at repetitive sequences is the absence of homologous recombination mediated repair of DNA double stranded breaks. In the absence of homologous recombination, double stranded breaks can be repaired by non-homologous end joining (NHEJ) or microhomology mediated end joining (MMEJ) (Ottaviani, LeCain and Sheer, 2014a). MMEJ relies on the presence of micro-homologies in the strands of DNA upstream and downstream of the break, and leads to deletions of greater than 3 base pairs (Chu, 1997; Davies *et al.*, 2017).

Establishing the likely mechanisms for the formation of short indels in a human populations or in cancer genomes is challenging. Similarities between processes demonstrated in vitro or in simple model systems must be matched to complex biological reality, and the underlying mechanisms inferred from these similarities. However, accepting these limitations, it has been proposed that ~75% of short indels in human populations are caused by strand slippage,

and the majority of the remainder by imperfect repair after double strand breaks (Montgomery *et al.*, 2013).

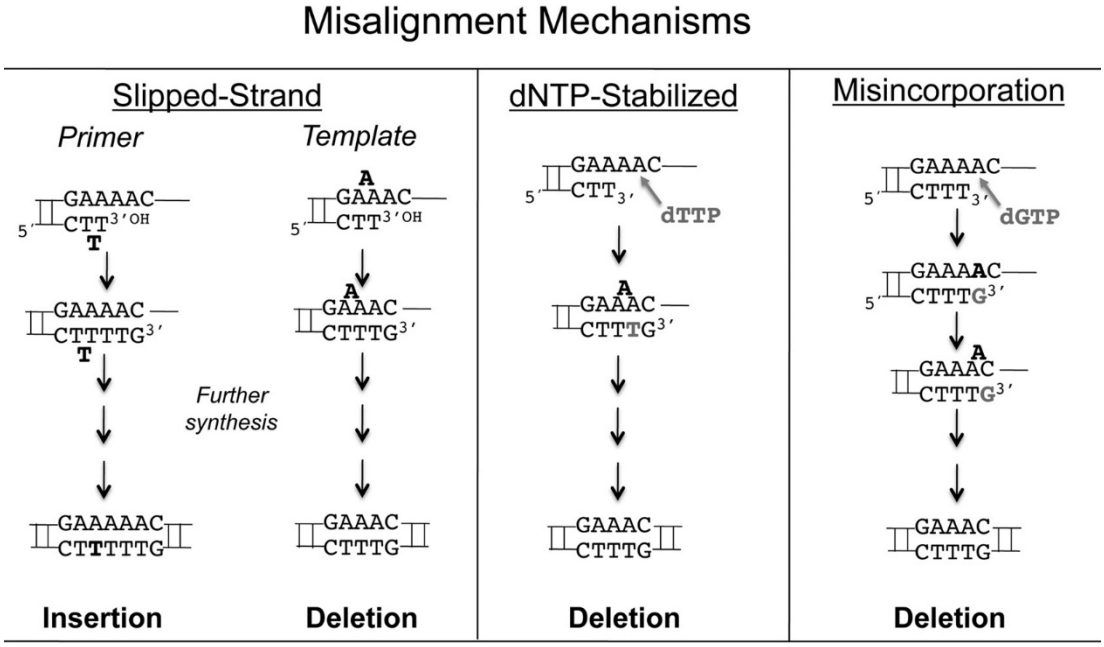


Figure 1.8. dNTP-stabilised and misincorporation-misalignment models. *Left hand panel: polymerase slippage hypothesis as proposed by Streisenger et al., showing how slippage and bulging out of the complementary and template strand can result in an insertion or deletion, respectively. Middle panel: In this scenario, a template base is skipped because the incoming dNTP correctly binds to the next 5' base, creating a bulged template and, after continued synthesis, a deletion mutation. Right panel: The mutation begins as a base misincorporation event; the newly added mismatched nucleotide can relocate to the next template position, if this is complementary to it. The resulting complementary strand has a deletion. Figure adapted with permission from (Baptiste, Jacob and Eckert, 2015).*

The mechanisms outlined above, apart from the misincorporation/ misalignment model, are most persuasive when applied to indels observed in long, repetitive sequences, rather than indels in short repeats, imperfect repeats, or not in repeats at all. In the last decade, findings from *S. cerevisiae* have highlighted another potential mechanism for the generation of short indels: the cleavage of embedded ribonucleotides by Topoisomerase 1 (Top1) (N. Kim *et al.*, 2011; Williams *et al.*, 2015, 2019). These results raise the question of whether this process is taking place in human cells. Below, I outline some of the known pathophysiological consequences of ribonucleotides in *S.*

cerevisiae and mammalian cells, followed by the mutational consequences of these ribonucleotides.

1.2.2 Pathophysiological consequences of ribonucleotide incorporation in S. cerevisiae and mammalian genomes

RNase H2 is not essential in *S.cerevisiae,* but yeast strains in which one of the subunits of the enzyme (generally RNH201, equivalent to human RNASEH2A) has been knocked out, when combined with a polymerase ε with mutations adjacent to the steric gate leading to higher rates of ribonucleotide incorporation, show slower growth than equivalent wild type strains (Nick McElhinny, Kumar, *et al.*, 2010). This is accompanied by activation of the genome integrity checkpoint and altered progression through the cell cycle (Nick McElhinny, Kumar, *et al.*, 2010; Williams *et al.*, 2013), and increased sensitivity to low levels of the replication stress-inducing agents hydroxyurea or methyl methanesulfonate (Lazzaro *et al.*, 2012). *S. cerevisiae* RNase H2 null strains also show chromosomal instability, with increased loss of heterozygosity (LOH) and elevated rates of non-allelic homologous recombination (Conover *et al.*, 2015; O'Connell, Jinks-Robertson and Petes, 2015).

In mammals, RNase H2 is essential; mice which are null for the enzyme exhibit embryonic lethality. This is due to elevated DNA damage and reduced cellular proliferation during gastrulation (Hiller *et al.*, 2012; Reijns *et al.*, 2012; Uehara *et al.*, 2018). Growth arrest was demonstrated to be a consequence of the p53 damage response and associated with substantial levels of genome embedded ribonucleotides. Mammalian cells have subsequently been shown to tolerate RNase H2 deletion when p53 is inactivated, including HeLa cells, where the activity of p53 is suppressed by HPV viral protein E6 (Schwarz *et al.*, 1985; Werness, Levine and Howley, 1990). In a recent study, HeLa cells where RNase H2 was knocked out by CRISPR-Cas9, were treated with EdU (5-ethynyl-2'-deoxyuridine) in order to identify dividing cells. Those that were EdU positive (i.e. actively replicating) showed evidence of increased replication dependent DNA damage, as shown by increased γ-H2AX foci

(Zimmermann *et al.*, 2018). This led in turn to elevated levels of sister chromatid exchange and a hyper-recombinant phenotype, reminiscent of what has been observed in yeast (Conover *et al.*, 2015). This effect was mediated at least in part by the activity of Top1.

Topoisomerases are enzymes that resolve torsional stress created by replication and transcription, where unwinding of the DNA double helix leads to increased supercoiling. There are two types, Type I, which create a single nick in DNA to provide a passage for one of the strands to unwind, and Type II, which cut the DNA on both strands to allow passage of duplex DNA (Cho and Jinks-Robertson, 2018). Top1, a type IB enzyme, had previously been implicated as enzyme that could mediate an alternative ribonucleotide excision repair pathway, due to ribonucleotide specific endonuclease activity (Sekiguchi and Shuman, 1997)

When Top1 acts on DNA, cleavage occurs as a transesterification reaction in which the phosphodiester bond between two nucleotides is attacked by the tyrosine of the enzyme, resulting in the formation of a DNA-(3'phosphotyrosyl)-enzyme intermediate. After unwinding of the strand, the topoisomerase then religates the bond that was originally cleaved (Shuman, 1992; Sekiguchi and Shuman, 1997). However, the equivalent Top1 cleavage reaction for RNA leads to the formation of a RNA-3'-phosphoryl-enzyme intermediate. This can then be attacked by the adjacent 2'OH of the ribose sugar to yield a free 2'3' cyclic phosphate product, releasing the topoisomerase. If the reaction forming this 2'3' cyclic phosphate is not reversed, which appears to be the case in some contexts (Sparks and Burgers, 2015), Top1 can cleave upstream of the cyclic phosphate, creating a short gap and a single stranded break.

It is these 2'3' cyclic phosphate intermediates and/or single stranded breaks that are then proposed to result in lesions that activate the DNA damage response in mammalian cells (Zimmermann *et al.*, 2018). In HeLa cells deficient in RNase H2 (and therefore with higher levels of genome embedded ribonucleotides), the causal role of Top1 in generating this DNA damage was confirmed by depleting TOP1 levels using short interfering RNAs (siRNAs).

This siRNA depletion led to a reduction in the number of γ-H2AX foci, showing that it was Top1 activity on genome embedded ribonucleotides leading to activation of the DNA damage response. Depletion of Top1 using siRNA also reduced the level of apoptosis when cells were further treated with a PARP inhibitor. PARP inhibitors act by interfering with the repair of breaks in DNA (Durkacz *et al.*, 1980; Rouleau *et al.*, 2010) normally promoted by PARP, which in turn leads to replication fork stalling, double stranded break formation and cell death. The synergistic activity of Top1 activity and PARP inhibition suggests that Top1 activity was leading to DNA lesions that engage PARP1 binding, creating a vulnerability to PARP trapping. Whether the exact stimulus for PARP1 binding is the 2'3' cyclic phosphates, covalent TOP1-DNA adducts in conjunction with single stranded gaps, or double stranded breaks that arise when these undergo replication, remains to be determined.

In addition to the effects documented in RNase H2 deficient cell lines and cancer, mutations in RNase H2 can lead to the autosomal recessive disease Aicardi-Goutières syndrome (AGS). AGS is an inflammatory encephalopathy that mimics congenital viral infections (Aicardi and Goutières, 1984; Rice *et al.*, 2007), and is associated with high levels of type I interferon production (Rice *et al.*, 2013). AGS is associated with partial loss of function mutations in all of the subunits of RNase H2 (Rice *et al.*, 2007; Reijns *et al.*, 2011). Inflammation in AGS caused by RNase H2 deficiency is due to increased levels of genome embedded ribonucleotides. Lymphoblastoid cells from patients with AGS accumulate genomic rNMPs when compared to WT controls (Pizzi *et al.*, 2015), as do patient fibroblasts (Günther *et al.*, 2015). Furthermore, reduced RER activity correlates with endogenous replication stress and cytokine production (Mackenzie *et al.*, 2016). This in turn may contribute to the rise in interferon levels that is a factor in disease progression (Crow, Shetty and Livingston, 2020).

1.2.3 Mutational consequences of ribonucleotide incorporation into the S. cerevisiae genome

Given the highly disruptive effects on DNA integrity of supra-physiological levels of genome embedded ribonucleotides in *S. cerevisiae* and mammalian cells, it is perhaps not surprising that these have also been associated in yeast with a distinctive mutational signature. *S. cerevisiae* lacking RNase H2 generate a characteristic short (2-5 bp) deletion signature, as assayed by mutation rates in reporter genes (Chen *et al.*, 2000; Clark *et al.*, 2011; N. Kim *et al.*, 2011). They are also prone to double stranded breaks, the repair of which can cause point mutations, insertion/deletions (indels) and large-scale genomic rearrangements (Conover *et al.*, 2015). The magnitude of these changes correlates with the degree of ribonucleotide incorporation, so that RNase H2 null strains with polymerase mutants with mutations adjacent to the steric gate that increase ribonucleotide incorporation (such as pol2-M644G) have a higher mutation rate than equivalent strains with a wild type polymerase (Nick McElhinny, Kumar, *et al.*, 2010).

1.2.4 Key role of Top1 activity in generating short deletions

A key realisation in this field came in 2011, when a study (N. Kim *et al.*, 2011) found that when Top1 is knocked out in addition to RNase H2, the rate of 2 bp deletions is reduced to baseline levels. The results from this study have subsequently been supported from further work in reporter constructs (Williams *et al.*, 2015) and whole genome mutation accumulation experiments (Williams *et al.*, 2019). Insight into the mechanisms for this pattern of mutagenesis was provided by the results from an elegant biochemical study in 2015 (Sparks and Burgers, 2015) (Figure 1.9). When Top1 cleaves at a ribonucleotide, binding to the 3’ phosphate, as described previously (Section 1.2.2), this can lead to the formation of a 2’3’ cyclic phosphate. This 2’3’ cyclic phosphate is difficult to resolve, but Top1 can cleave again, most commonly 2 base pairs upstream of the original site, releasing the 2’3’ cyclic phosphate and

associated 2 bases, but leaving a 2 nucleotide gap.

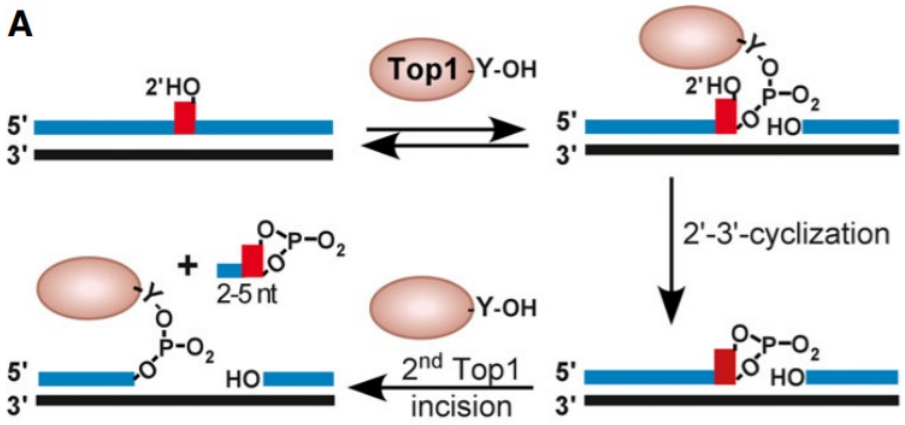


Figure 1.9. Formation of 2',3' cyclic phosphate after Topoisomerase 1 incision. *Topoisomerase 1 (Top1) cleaves a ribonucleotide (marked in red). During the course of Top1 action, the transient covalent RNA-3'-phosphoryl-Top1 intermediate is attacked by the adjacent 2'-hydroxyl, releasing Top1 and producing a 2'3'-cyclic phosphate intermediate. Top1 can cleave again, most commonly 2 base pairs upstream of the original cleavage site, releasing the 2',3'-cyclic phosphate and bound nucleotides, but leaving a gap. Figure reproduced with permission from* (Sparks and Burgers, 2015).

In the context of a repeat sequence, a slippage on the opposing strand can lead to realignment of the Top1 bound 3' end of the cleaved strand so that it is adjacent to the hydroxyl group downstream (Figure 1.10). These 2 ends can then be covalently linked through ligation through the action of DNA ligase. This avoids the creation of a permanent double stranded break, but results in a 2 base pair deletion.

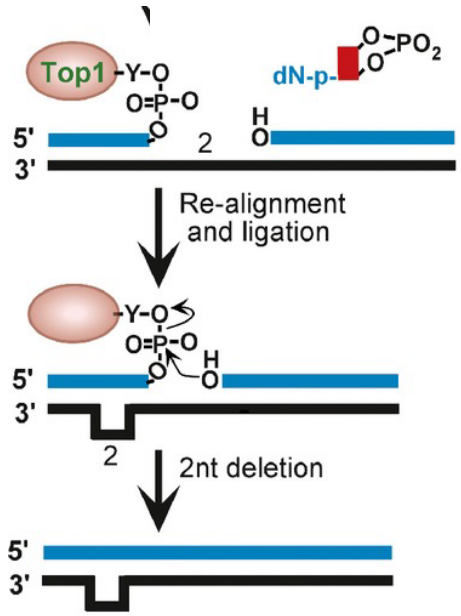


Figure 1.10. Model for generation of Topoisomerase 1 and ribonucleotide dependent short deletions. *Resolution of the 2',3'-cyclic phosphate "dirty end" leads to the formation of a gap, most commonly 2 nt in length. If the sequence upstream of the gap is repetitive, slippage along this can lead to re-alignment of the 2 opposing ends of the gap. These free ends can then re-ligate, forming a covalent bond and releasing topoisomerase 1. Figure reproduced with permission from* (Sparks and Burgers, 2015).

1.3 Aims and research outline

1.3.1 Specific motivations and importance

As outlined above, ribonucleotide misincorporation and removal represent fundamental biological processes that are conserved across all eukaryotes. As the example of the mating type switch in *S. pombe* suggests, it is possible that there are other biological functions of ribonucleotide in physiological settings that remain to be elucidated. Therefore techniques to better understand this would be of interest to the cell biology community

The presence of a ribonucleotide associated, Top1 mediated mutational signature in *S. cerevisiae*, in combination with evidence of conservation of 1 dependent DNA damage in mammalian cells, lead us to hypothesise that ribonucleotide incorporation might be associated with a distinctive indel signature in human cells. Investigating such a hypothesis was important as short deletions are key contributors to evolution, human inherited diseases, and cancer. A better understanding of the mechanisms that cause short

deletions, and an elucidation of the role of polymerase slippage and failure of mismatch repair as a cause of short deletions in human genomes, would be of interest to evolutionary biologists, human geneticists and cancer biologists.

1.3.2 Main research questions

The researched outlined in this thesis aims to provide answers to the following questions:

1. Can genome embedded ribonucleotides be detected quantitatively at single nucleotide resolution? Is the identification of *tracts* of ribonucleotides, and *phasing* of ribonucleotides along a single strand of DNA, possible?
2. Is a ribonucleotide-dependent indel mutational signature present mammalian cells?

1.3.3 Thesis structure

Subsequent to the Materials and Methods (Chapter 2), Chapters 3-5 contain an introduction, a results section and a discussion. In Chapter 3 (Quantitative detection of DNA-embedded ribonucleotides at single nucleotide resolution) I detail my approach to exploit nanopore sequencing to detect genome embedded ribonucleotides. I lay out the thinking behind my initial proof of concept experiment, the results from this, and a larger scale analysis which demonstrates the presence of both an amplitude and kinetic signature for a single embedded ribonucleotide in a strand of DNA. In Chapter 4 (Mutational consequences of ribonucleotide incorporation in *S*. cerevisiae) I outline the development of a highly sensitive reporter construct to identify 2 base pair equivalent deletions in *S. cerevisiae*, a construct that can be directly transferred to mammalian cells. In Chapter 5 I present results from the use of the reporter construct in HeLa cells in which RNase H2 has been knocked out. I also present my analysis of whole genome sequencing (WGS) from a mutation accumulation experiment in RNase H2 null retinal pigmented epithelium (RPE1) cell lines, conducted by Martin Reijns, a senior research

fellow in the Jackson Lab. I also present results from my analyses of other research groups' mutation accumulation experiments in mismatch repair deficient *S.cerevisiae* and human cells, RNase H2 null *S.cerevisiae*, and WGS of chronic lymphocytic leukaemia (CLL) patient samples. I demonstrate similarities between the as yet unexplained indel signature 4 in an extensive analysis of WGS in cancer (Alexandrov *et al.*, 2020) and the mutational signature seen in RNase H2 null *S.cerevisiae* and human cell lines.

Chapter 2 Materials and Methods

2.1 Note on data availability

Relevant reference sequences and the scripts used to interpret raw data have been made available in the following GitLab directory: https://git.ecdf.ed.ac.uk/twillia2/genome_embedded_ribonucleotides

This master directory contains a number of sub-directories which in turn contain the relevant files referred to in this chapter. They are referred to throughout in the following format: "/sub-directory/file", if navigating from the address above.

2.2 Laboratory Methods

2.2.1 General reagents

2.2.1.1 Sources of reagents

Chemicals were purchased from Sigma Aldrich, Fisher Chemicals and Amersham Biosciences (GE Healthcare). Enzymes were purchased from New England Biolabs and Takara Bio. Cell culture materials were purchased from Gibco (Invitrogen).

2.2.1.2 Common buffer solutions

All commonly used buffers (Table 2.1) were made using distilled water (dH_2O). Solutions were sterilised by autoclaving at 121°C for 15 min. Solutions that could not be autoclaved were passed through a 0.22 µm filter (Millipore).

Table 2.1. Commonly used buffers.

Buffer	Chemicals used for preparation
1x TE	1 Litre: 10 mM Tris-HCl (pH 8), 1 mM EDTA (pH8.0) (Fisher D/0700/53) 160 ml
20x TBE	1 Litre: Boric Acid (Fisher Scientific B/3800/53) 110 g, Tris Base (Fisher T3710/60) 216 g, 0.5M EDTA (pH8.0)(Fisher D/0700/53)80 ml
Lysis buffer	2% Triton X-100, 1% SDS, 0.5 M NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA

2.2.1.1 Drug stock solutions

Drugs stock solutions used are shown in Table 2.2 . They were added to tissue culture medium or yeast growth medium immediately prior to use in the working concentrations shown in Table 2.8 and Table 2.19. If not purchased as stock solution, they were prepared from powder in tissue culture hoods, and after dissolving in dH_2O passed through a 0.22 µm filter (Millipore) to sterilise.

Table 2.2. Drug stock solutions

Drug	Stock concentration	Manufacturer	Catalog number	Storage temperature
Puromycin	50 mg/ml	Sigma-Aldrich	BP2956-100	–20˚C
G418 (Neomycin)	250 mg/ml	Sigma-Aldrich	N6386	4˚C
Hygromycin	50 mg/ml	EMD Millipore	400052	4˚C
Nourseothricin (ClonNAT)	100 mg/ml	Werner BioAgents	5.001.000	-20 ˚C

2.2.1.2 Plasmids

Plasmids used in this study are detailed in Table 2.3. Those that were created throughout the course of this thesis (Table 2.4) were constructed using standard cloning techniques (Section 2.2.4.3) using the oligonucleotides detailed in Table 2.5.

Table 2.3. Summary of plasmids used in this thesis purchased commercially or previously generated by colleagues at the IGMM.

Plasmid	Description	Source (reference)
pSpCas9(BB)-2A-GFP (pX458)	Cas9 from *S. pyogenes* with 2A-EGFP, and BbsI cloning site for gRNA, used to introduce mutations into reporter construct	Addgene (plasmid number: 48138, gift from Principal Investigator Feng Zhang) (Ran *et al.*, 2013)
pBRNeo/pUC18 (pK18)	Kanamycin resistance gene and pUC18 multiple cloning site, used for creation of FISH probes	Courtesy of Laura Lettice, IGMM (Pridmore, 1987)
pXAT2	AAVS1 sgRNA expression vector, used to introduce reporter construct into Hela cells.	Addgene (plasmid number 80494, gift from Principal Investigator Knut Woltjen) (Oceguera-Yanez *et al.*, 2016)
pAAVS-Nst-CAG-Dest	Empty Backbone. Gateway Destination donor vector for AAVS1 targeting	Addgene (plasmid number 80489, gift from Principal Investigator Knut Woltjen) (Oceguera-Yanez *et al.*, 2016).
pFA6a-natMX6	Plasmid for yeast gene deletion using natMX6 selectable marker conferring nourseothricin resistance	Courtesy of Martin Reijns A.Jackson Laboratory; originally gift from Jean Beggs (Hentges *et al.*, 2005)

pFA6a-hisMX6	Plasmid for yeast gene deletion using his3MX6 selectable marker conferring growth in histidine deficient media	Courtesy of Martin Reijns A.Jackson Laboratory; originally gift from Jean Beggs (Longtine *et al.*, 1998)
pRS426	Plasmid for yeast gene deletion using URA3MX6 selectable marker conferring growth in deficient media	Courtesy of Martin Reijns, A.Jackson Laboratory; originally gift from Jean Beggs (Sikorski and Hieter, 1989)
pX461-RNASEH2A-gRNA1 (pMAR526)	Human RNASEH2A guide RNA 1 (G1F and R) in exon 1 cloned into BbsI of pX461	Created by Martin Reijns, A Jackson Laboratory (Benitez-Guijarro *et al.*, 2018; Zimmermann *et al.*, 2018). pX461 from Addgene (plasmid number 48140, gift from Principal Investigator Feng Zhang) (Ran *et al.*, 2013)
pX462-RNASEH2A-gRNA2 (pMAR527)	Human RNASEH2A guide RNA 2 (G2F and R) in exon 1 cloned into BbsI of pX462	Created by Martin Reijns, A Jackson Laboratory (Benitez-Guijarro *et al.*, 2018; Zimmermann *et al.*, 2018). pX462 from Addgene, plasmid number 48141, gift from Principal Investigator Feng Zhang) (Ran *et al.*, 2013)
pML104	Expresses Cas9, guide RNA expression cassette with BclI-SwaI cloning sites. Contains URA3 marker.	Courtesy of Martin Reijns. Addgene (plasmid number 67638, gift from Principal Investigator John Wyrick) (Laughery *et al.*, 2015)

Table 2.4. Plasmids generated for this thesis.

Plasmid	Description	Plasmid backbone	Details
pPTEF-HygroRv1-bait-P2A-neoR (pTCW3)	Plasmid containing version 1 of hygromycin resistance gene ($hygro^R$) for *S. cerevisiae*, with a "bait" sequence located before P2A	Ordered from GeneArt.	Figure 2.1; Section 2.2.2.3.5
pPTEF-HygroRv1-P2A-neoR (pTCW6)	Plasmid containing version 1 of $hygro^R$ with "bait" removed, but with additional *Kpn*I site.	pTCW3	Figure 2.2; Section 2.2.2.3.5
pPTEF-HygroRv1-P2A-neoR (pTCW9)	Plasmid with version 1 of *S. cerevisiae* reporter construct with additional *Kpn*I site removed.	pTCW6	Figure 2.2; Section 2.2.2.3.5
pPTEF-HygroRv2-P2A-NeoR_del (pTCW10)	Plasmid with version 2 of *S. cerevisiae* reporter construct, with C deletion at 3' terminus of $hygro^R$ version 2.	pTCW9	Figure 2.2; Section 2.2.2.3.5
pPTEF-HygroRv2-P2A-NeoR (pTCW12)*	Plasmid containing version 2 of $hygro^R$ and C deletion corrected.	pTCW10	Figure 2.2; Section 2.2.2.3.5

pHygroRv1-bait-P2A-PuroR (pTCW4)	Plasmid for insertion of construct into HeLa cells containing version 1 of hygro^R.	Ordered from GeneArt.	Figure 2.4; Section 2.2.3.4.1
pHygroRv2-P2A-PuroR (pTCW14)**	Backbone from pTCW4 with insert containing new Hygro^R gene (version 2) from pTCW12.4.	pTCW4	Figure 2.5; Section 2.2.3.4.1
pAAVS1-Nst-CAG-hygroRv2-P2A-PuroR (pTCW15)	Plasmid for introduction of human reporter construct into AAVS1 locus in HeLa cells. Product of LR reaction between pAAVS1-Nst-CAG-DEST and pTCW14.		Section 2.3.4.2
pCAG-HygroRv2-P2A-PuroR (pTCW16)	Plasmid with sequence matching human reporter construct, minus AAVS1 homology arms, for fluorescent in situ hybridisation.	pK18 (Table 2.3)	Section 2.2.4.4
pSpCas9(HygroRv2_sgRNA)-2A-GFP (pTCW17)	Plasmid to generate breaks in version 2 of the hygro^R coding sequence	pX458 (Table 2.3)	Section 2.2.3.4.100
pol2-M644_sgRNA-URA3 (pMAR765)	Plasmid to create point mutations in pol2-M644	pML104 (*Table 2.3*)	Section 2.2.2.3.10

** Reporter used for experiments in S. cerevisiae ** Reporter used for experiments in HeLa cells*

Table 2.5. Oligonucleotides used to generate the plasmids used in this study.

Primer	Orientation	Primer sequence (5' → 3')	Plasmid
SalI-P2A-F	Forward	AGAGGTCGACTGATTGGAAGCGGAGCTACTAACTTCAGCCTGCTGAAGCAGGCTGGAGAC	pTCW6
NdeI-P2A-R	Reverse	TACCCATATGAGGTCCAGGGTTCTCCTCCACGTCTCCAGCCTGCTTCAGCAGGCTGAAGT	pTCW6
KpnIdel-F	Forward	GAGGCCAGTTAATTAAGAGTACCTAGAATTCTCTAG	pTCW9
KpnIdel-R	Reverse	CTAGAGAATTCTAGGTACTCTTAATTAACTGGCCTC	pTCW9
pTCW10v2_2_C_ins_F	Forward	GCACACGTCCGCGCGCAAAAGAGG	pTCW10
pTCW10v2_2_C_ins_R	Reverse	CCTCTTTTGCGCGCGGACGTGTGC	pTCW10
HygroR-sgRNA-F	Forward	CACCGGAGAGGTTTGAGAGAGATCCG	pTCW17
HygroR-sgRNA-R	Reverse	AAACCGGATCTCTCTCAAACCTCTCC	pTCW17

2.2.2 Microbiological methods

2.2.2.1 Growth of bacteria

E.coli strains were grown in 37˚C in/on Luria-Bertani (LB) medium (10 g/l tryptone, 5 g/l yeast extract, 10 g/l NaCl, 1 g/l glucose). To maintain selection for plasmid DNA, antibiotics were added to the growth medium at the concentrations outlined in Table 2.6.

Table 2.6. Antibiotics used for bacterial selection in this study.

Antibiotic	Working concentration	Solvent	Storage
Ampicillin	50 µg/ml	dH_2O	-20 ˚C
Kanamycin	10 µg/ml	dH_2O	-20 ˚C
Chloramphenicol	34 µg/ml	EtOH	-20 ˚C

2.2.2.2 Transformation of E. coli

2.2.2.2.1 Preparation of chemically-competent bacterial cells for cloning

Preparation of chemically-competent *E. coli* was performed in-house by Martin Reijns and colleagues. *E. coli* DH5α cells were grown overnight on LB agar. The next day, a single colony was taken and 5 ml of LB medium with 20 mM $MgSO_4$ was inoculated and grown overnight to stationary phase. Following this, 250 ml of LB with 20 mM $MgSO_4$ was inoculated with 2 ml of the stationary phase culture and incubated at 23˚C in a shaking incubator at a minimum of 200 rpm until OD_{600} reached 0.4–0.6 (usually 8-10 h). The culture was then cooled on ice for about 15 min and cells were kept on ice for all subsequent steps. Cells were sedimented (10 min, 3,000 rpm, 4˚C) and gently resuspended in 80 ml of ice-cold sterilised TB buffer (10 mM PIPES-HCl pH 6.7, 15 mM $CaCl_2$, 0.25 M KCl, 55 mM $MnCl_2$). Cells were left on ice for 10 min, centrifuged for 10 min at 3,000 rpm at 4˚C, and gently resuspended in 20 ml of ice-cold TB buffer. After adding 1.5 ml of DMSO followed by a final 10 min incubation on ice, cells were dispensed into 200 µl aliquots in cold, sterile tubes and snap-frozen in liquid nitrogen. Aliquots were stored at –80˚C until required for transformation.

2.2.2.2.2 Transformation of chemically competent bacterial cells

For transformation of chemically-competent *E. coli*, 1 ng of plasmid DNA or 1 μl of a ligation reaction was added to 100 μl of competent DH5α cells. Cells and DNA were incubated on ice for 30 min before a 45 second heat-shock at 42˚C. Following 2 min recovery on ice, the cells were resuspended in 1 ml of LB medium. For selection in ampicillin, 100 μl of cells were plated immediately For selection in kanamycin, the resuspended cells were first incubated at 37˚C for 60 min with shaking at 250 rpm. 100 μl of cells were spread on LB-agar plates with the appropriate concentration of antibiotic (Table 2.6). The plates were incubated overnight at 37˚C to achieve discrete colonies.

2.2.2.3 S. cerevisiae methods

2.2.2.3.1 S. cerevisiae growth

Unless stated otherwise, *S. cerevisiae* strains were grown in a YPD (Yeast extract peptone dextrose) medium, described in Table 2.7. Unless stated otherwise, all plates used for *S. cerevisiae* growth were YPD, with the same composition as the equivalent broth solution except for the addition of agar. The ingredients were dissolved into 800 ml of H_2O, then made up to a final volume of 1 L. The solution was then autoclaved at 121 °C for 20 minutes at 15 psi. For strains containing the HIS3 gene, synthetic dropout media was produced by suspending Formedium (Ref No: DCS0071 Complete supplement Drop-out –HIS) in 1 litre of YPD (Table 2.7).

Table 2.7. Components of 1 L YP-glucose media. For broth, the agar was omitted, and for YP-galactose broth D-glucose substituted for Galactose (Sigma G0750)

Ingredient	Volume/mass
Peptone (Formedium PEP02)	20 g
Yeast Extract(Formedium YEA03)	10 g
Adenine hemisulfate salt,minimum 99% Sigma A9126-100g	0.1
D-Glucose (Fisher Scientific G/0500/53)	20 g
Agar (Formedium AGA03)	20 g
H_20	Made up to 1L

2.2.2.3.2 Antibiotics used for S. cerevisiae culture

The antibiotics detailed in Table 2.8 were used for selection of *S. cerevisiae* strains. Note that neomycin was used at different concentrations for different experiments as described in Chapter 4 (section 4.2.6.1).

Table 2.8. **Antibiotics used for *S. cerevisiae* selection in this study.**

Antibiotic	Working concentration	Solvent
Hygromycin	300 µg/ml	dH_2O
Neomycin (G418)	250-1000 µg/ml	dH_2O
Nourseothricin	100 µg/ml	dH_2O

2.2.2.3.3 S. cerevisiae transformations

The *S. cerevisiae* strain used for all experiments was the standard experimental strain BY4741 (Brachmann *et al.*, 1998), unless stated otherwise. For transformations, the protocol developed by Gietz and Woods was used (Gietz and Woods, 2006). The *S. cerevisiae* strain to be transformed was grown from a single colony in 5 ml of YPD at 30 ˚C overnight, shaken at 250 rpm. In the morning 1 ml of this culture was placed into 50 ml of YPD in a 250 ml flask, and incubated for a further 5 hours at 30 ˚C, shaking at 250 rpm. Once an optical density (OD595) of between 0.8 and 1.0 was reached, the culture was centrifuged at RT for 2 minutes at 3,500 g, and the cells washed twice with 20 ml of dH_2O, centrifuging at RT for 2 minutes at 3,500 g after each wash. The cells were reconstituted in 500 µl of 1x TE. 50 µl of these cells were added to an ice-fold mixture containing 240 µl PEG 50% 3350, 36 µl 1 M Lithium Acetate, 10 µl Salmon Sperm DNA, 14 µl dH_2O, and 10 µl of PCR product. The mixture was then incubated at 42 ˚C for 30 minutes. The cells were then centrifuged, and resuspended in 100 µl TE, and finally plated onto YPD plates using the Copacabana method (Worthington, Luo and Pelo, 2001). After overnight incubation, the plates were replica plated onto YPD plates containing antibiotic or nutrient selection medium.

2.2.2.3.4 Gene deletions

In general, genes were deleted in transformations as outlined in Section 2.2.2.3.3, replacing each gene with either an antibiotic resistance gene or a

nutritional selection marker. Oligonucleotide sequences are given in Table 2.5 and the template plasmids in Table 2.3

2.2.2.3.5 Principle of the reporter construct and subsequent modifications

The sequence for the *S. cerevisiae* experiments was designed as described in Chapter 4 (Section 4.2.4), and the sequence ordered from GeneArt as pTCW3 (GitLab: "reference_sequences/pTCW3_GeneArt_order.gb"; Figure 2.1). The plasmid then underwent a series of modifications.

The original plasmid (TCW3) contained a hygromycin resistance gene flanked by two restriction enzyme cut sites. At the end of the hygromycin gene was a "bait" with tandem repeats demonstrated to be hotspots previously for mutations (N. Kim *et al.*, 2011; Kim *et al.*, 2013). However, when we realised that this "bait" could accumulate mutations whilst the strain was growing in hygromycin with no phenotypic effects, it was removed. A restriction enzyme digest of pTCW3 with *Sal*I and *Nde*I created a backbone, which was gel purified (Section 2.2.4.3.4) to separate from the insert (153 bp long). Primers SalI-P2A-F and NdeI-P2A-R (Table 2.5) were annealed by heating to 95 °C for 2 minutes and allowing to cool to room temperature. They were subsequently extended using DNA Polymerase I (New England Biolabs) by incubating for 1 hour at 37°C. Subsequently the filled in duplex was ethanol precipitated (Section 2.2.4.3.1), restriction digested with *Sal*I and *Nde*I (Section 2.2.4.3.6), and ethanol precipitated again. The insert and backbone were ligated together (Section 2.2.4.3.10) to create pTCW6 (Figure 2.2). Subsequently, an unanticipated *Kpn*I cut site on the GeneArt backbone was eliminated by QuikChange mutagenesis (section 2.2.4.3.9) to yield plasmid pTCW9 (Figure 2.2) using primers KpnIdel-F and KpnIdel-R (Table 2.5).

Version 2 of the hygromycin resistance gene, which as described in Chapter 4 (Section 4.2.4.2) had a higher frequency of dinucleotide tandem repeats, and a lower frequency of mononucleotide tandem repeats, was ordered as a G-block (gBlocks® Gene Fragments from IDT, without phosphorylation, GitLab: "mutation_detection_construct/reference_sequences/ g_block_di_rep_only_20180204.fa"). This sequence was then inserted into

pTCW9 by subsequent restriction digestion and ligation. 2 µg of pTCW9 and 500 ng of the HygroR G-block respectively were digested with *Kpn*I-HF and *Sal*I-HF (Section 2.2.4.3.6). After PCR column purification (Section 2.2.4.3.4) of both reactions, the digested pTCW9 was eluted in 44 µl elution buffer, and treated with phosphatase (Section 2.2.4.3.12). The pTCW9 backbone was purified using gel electrophoresis and eluted in 50 uL of TE. 1 µl of the pTCW9 backbone was ligated with 5 µl of the G-block using a T4 DNA ligase reaction (Section 2.2.4.3.11) to create pTCW10, which was then transformed into *E. coli* (Section 2.2.2.2.2).

The creation of pTCW10 was complicated by difficulty in integrating the G-block in an error free fashion- a CG rich run at the 3'end of the Hygro^R led to insertions and deletions compared to the reference sequence of varying lengths. pTCW10 itself contained a C deletion at the 3' end, which was eliminated with a further round of QuickChange mutagenesis (Section 2.2.4.3.9) using primers pTCW10v2_2_C_ins_F and pTCW10v2_2_C_ins_R (Table 2.5) to create pTCW12 (Figure 2.2), the plasmid used for amplification of the reporter construct and insertion into *S. cerevisiae* for the fluctuation assays in Chapter 4. The reference sequence for pTCW12 (pPTEF-HygroRv2-P2A-NeoR.fa,Table 2.4) is available on GitLab:
("mutation_detection_construct/reference_sequences/pPTEF-HygroRv2-P2A-NeoR.fa").

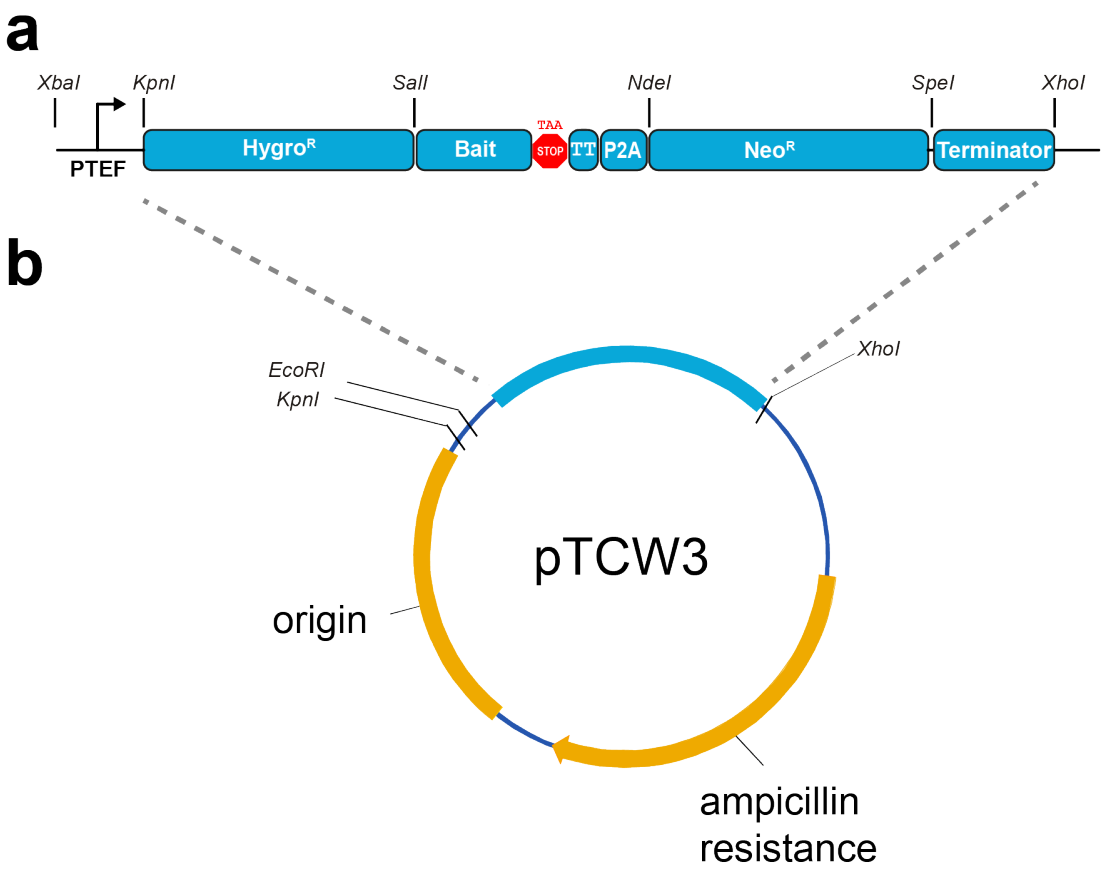


Figure 2.1 General schema for the reporter construct in S. cerevisiae.a) The reporter construct as originally designed. *This consists of a PTEF promoter, the hygromycin resistance gene (Hygro^R), followed in the original iteration by a "bait" sequence with tandem repeat sequences documented previously as being hotspots for short deletions* (N. Kim *et al.*, 2011; Kim *et al.*, 2013). *This bait sequence was followed by a stop codon, a two base pair frameshift, a self-cleaving peptide, a neomycin resistance gene (Neo^R) and a terminator sequence. Restriction enzymes are marked above the sequence.* ***b) GeneArt order.*** *The gene sequence described above was ordered from GeneArt, and was manufactured as the requested sequence flanked by XhoI and EcoRI cut sites, with an ampicillin resistance sequence to facilitate cloning. Due to the manufacturing process, there was an additional KpnI cut site upstream of the sequence.*

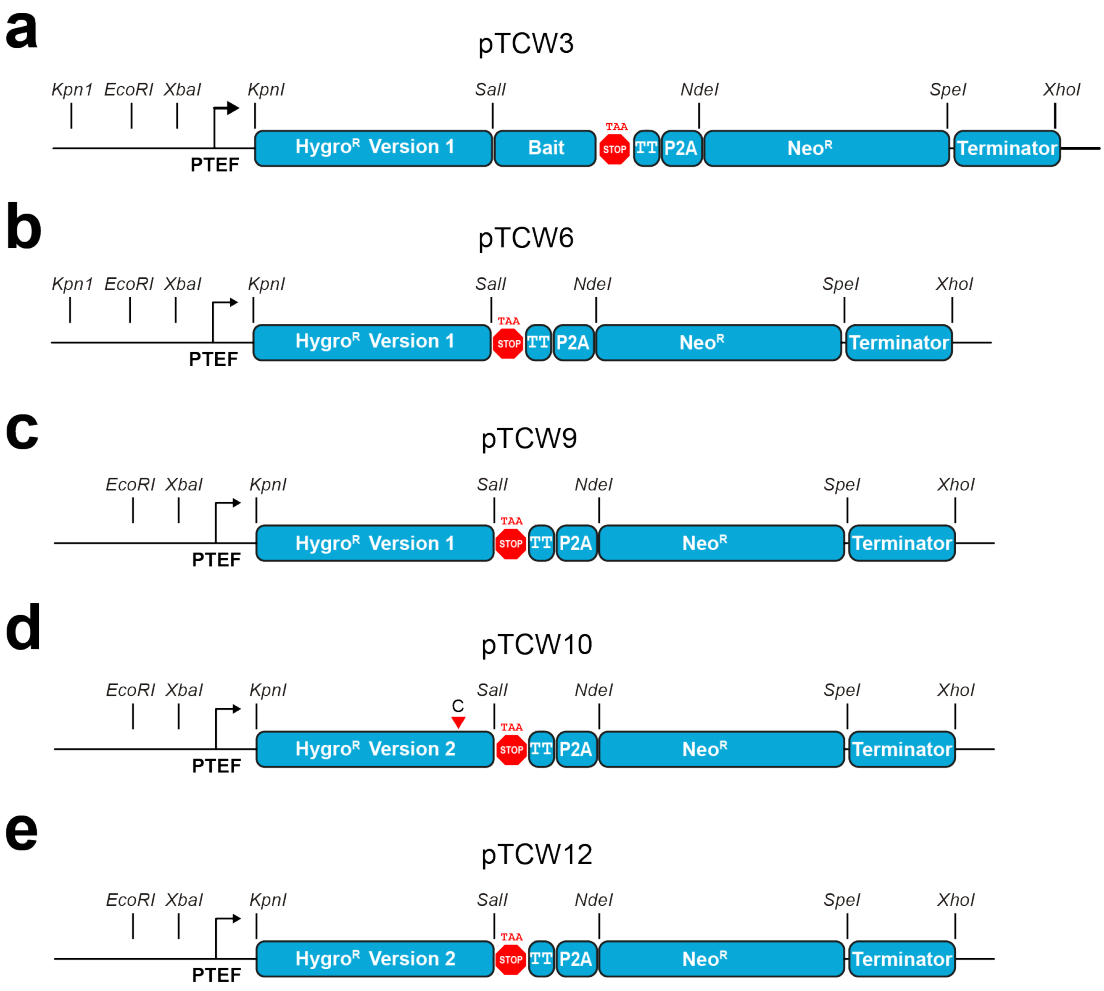


Figure 2.2. Modifications of plasmid to arrive at final construct. a) pTCW3. *Original plasmid ordered from GeneArt, described in detail in Figure 2.1.* ***b) pTCW6.*** *Plasmid after removal of "bait" sequence.* ***c) pTCW9.*** *Plasmid after removal of 5' KpnI cut site.* ***d) pTCW10.*** *Plasmid after insertion of version 2 of the hygromycin resistance gene, but with C deletion at 3' end of hygro^R.* ***e) pTCW12.*** *Final plasmid with C deletion corrected and version 2 of hygromycin resistance gene (for details on this refer to Figure 4.3, Chapter 4).*

2.2.2.3.6 Reporter construct insertion into S. cerevisiae

To integrate it into the BY4741 ancestral strain, primers agp1-MX6-F and -R (Table 2.9) were used, which contain sequences homologous to AGP1 (a non-essential gene) locus in BY4741, downstream of origin ARS306, using an approach outlined previously (Nick McElhinny *et al.*, 2008) (Chapter 4, Figure 4.3).

Table 2.9. Primers used to amplify reporter construct for integration into BY4741

Primer name	Primer sequence (5' → 3')
agp1-MX6-F	TGGGTTATTGGTCGGTAACGGTACCGCGTTGGTTCATGCGGGTCCAGCTGGACT ACTTATCGGATCCCCGGGTTAATTAAG
agp1-MX6-R	CTTCTTGCTTGATTAATTCTTCATCAAAGATTTGTCTATGAGAATCTAGGTCGATCT TGTGAATTCGAGCTCGTTTAAAC

The construct was PCR amplified using pTCW12 in a Phusion Flash reaction (Section 2.2.4.3.8.1) using the following settings:

Denaturation		98˚C	10 sec
34 x	Denaturation	98˚C	10 sec
	Annealing	60˚C	5 sec
	Extension	72˚C	90 sec
Extension		72˚C	90 sec

Subsequent to this the BY4741 strain was transformed as described in Section 2.2.2.3.3. After overnight incubation, the plates were replica plated onto YPD plates containing 300 µg/ml of hygromycin.

2.2.2.3.7 Colony PCR to confirm integration of reporter construct

Two FastStart colony PCR reactions (Section 2.2.4.3.8.1) were used to confirm successful insertion of the reporter construct at the AGP1 locus (Table 2.10) using the following settings.

Denaturation		95˚C	5 min
34 x	Denaturation	95˚C	30 sec
	Annealing	58˚C	30 sec
	Extension	72˚C	30 sec
Extension		72˚C	30 sec

Table 2.10. Primers used to confirm insertion of the construct into the AGP1 locus

Primer	Orientation	Location	Primer sequence (5' → 3')	Amplicon length
agp1-upF	Forward	Upstream	GTTCTGATGATTGCGTTGGG	262 bp
PTEF-R	Reverse	Upstream	GATGTATGGGCTAAATGTACG	
TTEF-F	Forward	Downstream	TCGCCTCGACATCATCTGC	224bp

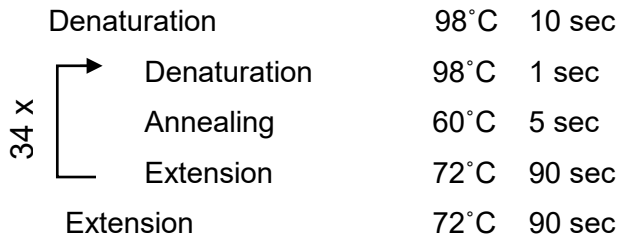
agp1-doR	Reverse	Downstream	CAAACGTTCCCTATATTCTTCG	

To confirm incorporation at the correct locus, and loss of the original AGP1 gene segment, extracted DNA (Section 2.2.4.3.1) from strains with a positive colony PCR underwent Phusion Flash PCR (Section 2.2.4.3.9.1) with primers agp1-upF and downstream reverse primer agp1-doR (Table 2.11), with a WT BY4741 control.

Table 2.11. Primers used to amplify entire reporter construct in S. cerevisiae

Primer	Orientation	Primer sequence (5' → 3')	Amplicon length
agp1-upF	Forward	GTTCTGATGATTGCGTTGGG	1.4 kb (WT)
agp1-doR	Reverse	CAAACGTTCCCTATATTCTTCG	2.7 kb (reporter)

PCR reactions were performed using the following settings:



PCR products were run on an agarose gel to confirm expected product size (Table 2.11). The initial strain with the construct was named TCWY1 (Figure 2.3).

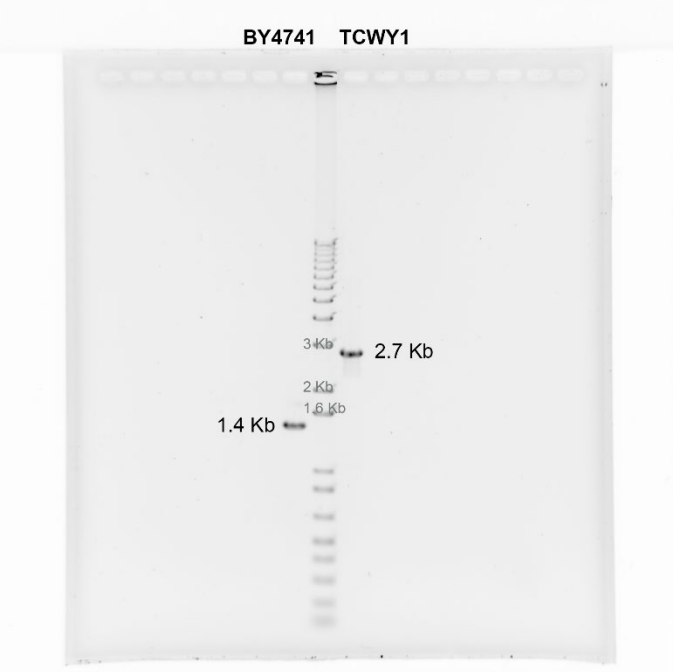


Figure 2.3. Confirmation of replacement of the AGP1 gene with reporter construct. *1% agarose gel, with 1Kb+ ladder. The amplification product for DNA extracted from the ancestral strain BY4741 is shown on the left, and for TCWY1 (AGP1 replaced by reporter construct) on the right.*

2.2.2.3.7.1 Sequencing of entire reporter construct in S. cerevisiae

For the initial construct (version1), which consisted on mononucleotide and dinucleotide repeats (for details on design see Chapter 4, Section 4.2.4), the primers numbered 1-12 in Table 2.12 were used. For the final construct (version 2, for details on design again see Chapter 4), primers 1-2,5-7, and 10-12 were used, but primers 3,4,8 were replaced with primers 13-16 (Table 2.12). The reference sequence is available on GitLab ("/mutation_detection_construct/reference_sequences/s.cerevisiae_reporter_construct.fa).

Table 2.12. Primers used for Sanger sequencing of initial (version 1) and final (version 2) mutation detection constructs.

Primer	Primer names	Primer sequence (5' → 3')	Construct
1	S24F	ACGGATCCCCGGGTTAAT	Both
2	S297F	CACAGACGCGTTGAATTGTC	Both
3	S776F	CTGTGCTGCAGCCTGTGG	Version 1 only
4	S1276F	AGAGTCTTGTGGACGGCAAT	Version 1 only
5	S1793F	AAAGGTAGCGTTGCCAATGA	Both
6	S2298F	ACCAGGATCTTGCCATCCTA	Both

7	S142R	GCACGTCAAGACTGTCAAGG	Both
8	S653R	GGCAAAATGCCGGTATACAT	Version 1 only
9	S1139R	CGCGACCTCGTATTGAGAAT	Version 1 only
10	S1658R	GCCTCGAAACGTGAGTCTTT	Both
11	S2136R	CCATTACGCTCGTCATCAAA	Both
12	S2652R	CGACAGCAGTATAGCGACCA	Both
13	S752F	CAAGATCTCCCAGAGACAGAGC	Version 2 only
14	S1258F	TGGTCTCGACCAACTATACCAG	Version 2 only
15	S588R	CATATCCTCTCCCTCCCACA	Version 2 only
16	S1113R	ACATCGCCTCTGACCAGTCT	Version 2 only

2.2.2.3.7.2 Deletion of RNH201 and topoisomerase 1 (TOP1)

Subunit A of the RNase H2 complex in *S. cerevisiae* (RNH201) was replaced with the antibiotic resistance gene for Nourseothricin (Nat) in a BY4741 strain containing the reporter construct. In a separate reaction on a BY4741 strain containing the reporter construct, the Topoisomerase 1 enzyme (Top1) was knocked out using the same resistance gene. Plasmid pFA6a-natMX6 (Table 2.3), which contains a PTEF promoter and a TTEF terminator region and Mx6 homology arms, was used as a template for the generation of a PCR amplicon for transformation. A Phusion Flash reaction (Section 2.2.4.3.8.1) was conducted, with the primers shown in Table 2.13, and transformation was conducted as detailed in Section 2.2.2.3.3. After transformation, colonies were grown overnight on YPD and then replica plated onto plates containing hygromycin and nourseothricin (to select for the $Hygro^R$ reporter and the deletion cassette).

Table 2.13. Primers used to generate amplicons to knock out RNH201, TOP1, MSH2

Primer	Gene to knock out	Orientation	Primer sequence (5' → 3')
rnh201-MX6-F	RNH201	Forward	CTAATGAGAGTGTCGAAAAACCTTGAAAACAACTACTGCACACCAAATTGATACGATTAACGGATCCCCGGGTTAATTAAG
rnh201-MX6-R	RNH201	Reverse	TGAAGTTATGACATATGTAGTATTACATGAAGATATATAGTATGTGCAAACTGGAGGTGAGAATTCGAGCTCGTTTAAAC
Top1-MX6-F	TOP1	Forward	AAACAGCAAATAAAAAAAATCTAAAGGGAGGGCAGAGCTCGAAACTTGAAACGCGTAAAACGGATCCCCGGGTTAATTAAG
Top1-MX6-R	TOP1	Reverse	CCTAATGCGAACTTGATGCGTGAATGTATTTGCTTCTCCCCTATGCTGCGTTTCTTTGCGGAATTCGAGCTCGTTTAAAC
msh2-MX6-F	MSH2	Forward	ATCAACTGTAAAAAATCTCTTTATCTGCTGACCTAACATCAAAATCCTCAGATTAAAAGTCGGATCCCCGGGTTAATTAAG
msh2-MX6-R	MSH2	Reverse	TAGAGCCTTATTGCCTTTAGGAAAATTGATCTATGACAGAGATTATTCTTTCGTGAACACGAATTCGAGCTCGTTTAAAC

In order to confirm insertion of the resistance gene at the correct locus, the primers from Table 2.14 were used, with the PTEF-R and TTEF-F primers shown in Table 2.10 as the upstream reverse and downstream forward primers for each colony PCR, which was carried out in a Colony PCR reaction as described in Section 2.2.4.3.8.1.

Table 2.14. Primers used to confirm knockout of S. cerevisiae genes in this study

Primer	Gene knocked out	Orientation	Location (primer pair)	Primer sequence (5' → 3')	Amplicon length
rnh201-upF	RNH201	Forward	Upstream (PTEF-R)	CACCAATAAACT ACGCTACGC	278bp
rnh201-doR	RNH201	Reverse	Downstream (TTEF-F)	CAAAGCATAGTA GCAGATGAC	235bp
Top1-upF2	TOP1	Forward	Upstream (PTEF-R)	TCTAATTACCTG AGTCCTATTC	318bp
Top1-doR2	TOP1	Reverse	Downstream (TTEF-F)	GTATAAGATATC TTCCTAGTAAC	224bp
msh2-upF	MSH2	Forward	Upstream (PTEF-R)	TTAAATGTTGAC ACTCTACTCC	252bp
msh2-doR	MSH2	Reverse	Downstream (TTEF-F)	GAATAAACTGTA CCTTGTCTAC	268bp

2.2.2.3.7.3 Double deletion of RNH201 and TOP1

For this double knockout, a BY4741 strain containing the reporter construct and an RNH201 knockout underwent a second transformation to replace the topoisomerase 1 gene with the HIS3 gene, which allows strains to grow in histidine deficient media. A PhusionFlash reaction (Section 2.2.4.3.8.1) was used, with the primers shown in Table 2.13 and plasmid pFa6a-hisMX6 (Table 2.3). After heat transformation, colonies were replica plated onto plates with hygromycin, noursethricin and histidine deficient growth media. For confirmation of the Top1 knockout, the primers for TOP1 in Table 2.14 were used, with PTEF-R and TTEF-F from Table 2.10.

2.2.2.3.8 Creation of topoisomerase 1 over-expression strain

This was performed by Martin Reijns. The Nat6MX6 cassette was amplified using the PTEF-F and TTEF-R primers (Table 2.10) and pFA6a-NatMX6 (Table 2.3) as a template. Yeast strain YAEH271 (El Hage *et al.*, 2010), a gift

from Aziz El Hage (Wellcome Centre for Cell Biology, Edinburgh) was transformed to replace KanMX6 with NatMX6, to make MRY141 (MATa his3Δ1 leu2Δ0 met15Δ ura3Δ0 PGAL1-3HA-TOP1 (NatMX6). Subsequently the cassette of version 2 of the reporter construct from PTEF to TTEF was amplified using agp1-MX6-F and agp1-MX6-R (Table 2.9) primers using pTCW12 (Table 2.4) as a template. MRY141 was then transformed to produce MRY146/TCWY21 and MRY147/TCWY22 (MATa his3Δ1 leu2Δ0 met15Δ ura3Δ0 PGAL1-3HA-TOP1 (NatMX6) agp1::version2 reporter construct (from pTCW12). Colony PCR was used to confirm reporter insertion at the AGP1 locus and sequenced to ensure no mutations were present.

2.2.2.3.9 Knock out of MSH2

This work was conducted by Martin Reijns. A Phusion Flash reaction (Section 2.2.4.3.8.1) was used, with the primers shown in Table 2-18, and plasmid pFA6a-natMX6 (Table 2.3). Colony PCR was conducted with the primers from Table 2.12, with PTEF-R and TTEF-F from Table 2.10.

2.2.2.3.10 Creation of pol2-M644G and pol2-M644L mutants

This work was conducted by Martin Reijns, using a previously published approach (Laughery *et al.*, 2015). Plasmid pML104 (Addgene: 67638, gift from John Wyrick) was used as a backbone. A gRNA sequence was cloned into pML104 (*SwaI/BclI*) by ligating with pol2-M644-gRNAF and pol2-M644-gRNAR duplex (Table 2.15) to create pMAR765 (Table 2.4). TCWY16 was then transformed with pMAR765 and a repair template duplex for M644G or M644L (Table 2.15) selecting single colonies on SD-URA. Mutations were confirmed by sequencing to make MRY152/153 (M644G) and MRY154/155 (M644L) (Table 2.16).

Table 2.15. Primers used by Martin Reijns to create S. cerevisiae strains for fluctuation assays.

Primer	Orientation	Primer sequence (5' → 3')
PTEF-F	Forward	GGATCCCCGGGTTAATTAAG
TTEF-R	Reverse	GCAGATGATGTCGAGGCGA
pol2-M644-gRNAF	Forward	GATCCATGATGTTTGGGTACATAGGTTTTAGAGCTAG
pol2-M644-gRNAR	Reverse	CTAGCTCTAAAACCTATGTACCCAAACATCATG

M644G_repair_F	Forward	TACTATCTGGTTGTAGTCTATTTGTAGTCATGATGTTTGGG TACCCAGATGCGACATCTACATGATAGATCAAAGGTAGTT CGTTTCTTA
M644G_repair_R	Reverse	TAAGAAACGAACTACCTTTGATCTATCATGTAGATGTCGCA TCTGGGTACCCAAACATCATGACTACAAATAGACTACAAC CAGATAGTA
M644L_repair_F	Forward	TACTATCTGGTTGTAGTCTATTTGTAGTCATGATGTTTGGG TACAGAGATGCGACATCTACATGATAGATCAAAGGTAGTT CGTTTCTTA
M644L_repair_R	Reverse	TAAGAAACGAACTACCTTTGATCTATCATGTAGATGTCGCA TCTCTGTACCCAAACATCATGACTACAAATAGACTACAACC AGATAGTA

2.2.2.3.11 Summary of strains used in fluctuation assays

The final strains used in the fluctuation assays presented in Chapter 4 are presented below (Table 2.16).

Table 2.16. S. cerevisiae strains used in fluctuation assays

Name	Genotype	Description
TCWY16	Wild type	BY4741 with integration of v2 reporter construct at the AGP1 locus. MATa his3Δ1 leu2Δ0 met15Δ ura3Δ0 agp1::v2-reporter-construct
TCWY20	*rnh201Δ*	BY4741 with v2 reporter construct and RNH201 (RNase H2 subunit A) knocked out and replace with Nat (nourseothricin resistance). MATa his3Δ1 leu2Δ0 met15Δ ura3Δ0 agp1::v2-reporter-construct rnh201::Nat
TCWY18	*Top1Δ*	BY4741 with v2 reporter construct and TOP1 (topoisomerase) knocked out and replace with Nat (nourseothricin resistance) MATa his3Δ1 leu2Δ0 met15Δ ura3Δ0 agp1::v2-reporter-construct Top1::Nat
TCWY19	*rnh201Δ Top1Δ*	BY4741 with v2 reporter construct and nrh201/Top1 double knockout. RNH201 replaced with Nat, TOP1 replaced with HIS3. MATa his3Δ1 leu2Δ0 met15Δ ura3Δ0 agp1::v2-reporter-construct rnh201::Nat Top1:his3
TCWY21	pGAL-Top1	MRY141 (Top1 expression when grown in galactose, no expression when grown in glucose) transformed with version 2 of reporter construct. MATa his3Δ1 leu2Δ0 met15Δ ura3Δ0 PGAL1-3HA-TOP1 (NatMX6) agp1::v2 reporter construct
TCWY22	pGAL-Top1	Independent clone, genetically identical to TCWY21
MRY144	*msh2Δ*	BY4741 with v2 reporter construct and MSH2 (a key mismatch repair pathway enzyme) knocked out and replaced with Nat (nourseothricin resistance). MATa his3Δ1 leu2Δ0 met15Δ ura3Δ0 agp1:: v2 reporter-construct msh2::NatMX6
MRY145	*msh2Δ*	Independent clone, genetically identical to MRY144
MRY152	pol2-M644G	BY4741 with v2 reporter construct and pol2-M644G amino acid substitution introduced using CRISPR-Cas9. MATa his3Δ1 leu2Δ0 met15Δ ura3Δ0 agp1:: v2-reporter-construct pol2-M644G (ATG>GGG; pol2-A642A, GCC>GCA)
MRY153	pol2-M644G	Independent clone, genetically identical to MRY152
MRY154	pol2-M644L	BY4741 with v2 reporter construct and pol2-M644L amino acid substitution introduced via CRISPR-Cas9. MATa his3Δ1 leu2Δ0 met15Δ ura3Δ0 agp1:: v2-reporter-construct pol2-M644L (ATG>CTG; pol2-A642A, GCC>GCA
MRY155	pol2-M644L	Independent clone, genetically identical to MRY154

2.2.2.4 S. cerevisiae fluctuation assay methods

All fluctuation assays were carried out as previously described (Spell and Jinks-Robertson, 2004). To generate individual colonies without mutations in the hygromycin resistance, before the start of the fluctuation assay, strains to be tested were grown overnight in 5 ml of YPD with hygromycin 300 g/ml at 30 °C shaking at 250 rpm. In the morning the culture spun down for 4 minutes at 3,500 g, YPD was removed, and cells reconstituted in 1 ml dH_2O. A 10^{-6} dilution was making serial dilutions in a 96 well plate (25 µl of cells added to 225 µl of dH_2O), and 100 µl of this dilution spread using the Copacabana method (Worthington, Luo and Pelo, 2001) on YPD plates containing hygromycin (Table 2.8). Plates were incubated at 30 °C for 72 hours.

Where possible (Top1 over-expression, *msh2Δ*, pol2 variants) two independent knock out strains were used for each fluctuation assay; for the remainder of strains (wildtype, *rnh201Δ, Top1Δ* and *rnh201Δ Top1Δ*) a single ancestral strain was used. 16 (or 2 x 8 if two strains were used) cultures were started using a single colony, added to 5 ml YPD, and grown for 3 days at 30 °C shaking at 250 rpm. Cells from each independent culture were pelleted by centrifugation at 3,500 g, and resuspended in 1 ml of dH_2O. For all strains except *rnh201Δ,* 100 µl of each suspension was plated on 2 YPD plates. Because of the large expected number of mutants, for *rnh201Δ* 100 µl of a 10^{-2} dilution of cells was plates onto 2 YPD plates containing G418. Initially, a concentration of 250 ng/uL of G418 was used; for the final experiments this was increased to 1000 ng/uL. All cultures were diluted 10^{-6} using serial 1:10 dilutions in a 96 well plate (25 uL cells to 225 uL dH_2O) and 100 µl of this dilution plated onto YPD plates to estimate the total number of cells in the initial culture for mutation rates calculation. Plates were incubated at 30 °C for 72 days, and photos taken as described in Section 2.3.3.1.1. To calculate mutation rates the Lea Coulson method of the median was used (Lea and Coulson, 1949), implemented in an Excel macro-enabled spreadsheet shared by Prof. Hannah Klein (New York University). With this macro mutation rates are calculated for each individual culture, and an overall rate for each strain

calculated using the Lea Coulson method, with 95% confidence intervals generated using tables published previously (Altman, 1990).

2.2.2.4.1 Independent cultures to generate large numbers of mutants for mutational spectra sequencing

In order to generate numbers of independent mutants for sequencing, in keeping with the 100-200 mutants per strain published in previous studies (results from these summarised in (Williams *et al.*, 2015), we grew independent cultures in 1 ml 96 well plates, and plated 25 µl of dilutions of these cultures in 10 x10 cm 5x5-well plates containing YPD agar with G418 (1000 ug/ml). This required exploratory studies to estimate the number of mutants per 1 ml culture per strain, as the relationship between the number of colonies expected in a 5 ml culture (the standard for the fluctuation assays) cannot linearly be related to what is expected in a 1 ml culture. In addition, the plates required a faster rotation speed than 5 ml cultures (320 rpm vs 250 rpm), due to a small volume within the plates to mix the culture media. In theory this change in growth conditions has the potential to affect metabolism and hence the mutation rate (see section on topoisomerase 1 up/down regulation, Chapter 4 , Section 4.2.8), but we assumed that this would have no effect on mutational spectra.

2.2.2.4.1.1 Amplification of Hygromycin Resistance gene to detect 2 bp equivalent mutations

The PCR mixture to screen for mutations in a *S. cerevisiae* fluctuation assay mutations was a 10 µl FastStart PCR Master Mix reaction as detailed in Section 2.2.4.3.7.1, using the primers detailed in Table 2.17 For each colony a 5’ and a 3’ amplicon were amplified, as the construct was too large to consistently amplify in a single colony PCR reaction. A colony PCR was necessary to avoid having to purify DNA from the large number of independent colonies required for mutational spectra analysis. PCR reactions were performed using the following settings:

	Denaturation	95˚C	5 min
34 x	Denaturation	95˚C	30 sec

Annealing	58˚C	30 sec
Extension	72˚C	45 sec
Extension	72˚C	45 sec

PCR products were separated on a 1.5% 0.5X TBE agarose gel to confirm correct amplification.

Table 2.17. Primers used to amplify 5' and 3' amplicons of hygromycin resistance gene within reporter construct in S. cerevisiae

Amplicon	Primer	Orientation	Primer sequence (5' → 3')	Amplicon length
Upstream	S297F	Forward	CACAGACGCGTTGAATTGTC	817bp
Upstream	S1113R	Reverse	ACATCGCCTCTGACCAGTCT	
Downstream	S752F	Forward	CAAGATCTCCCAGAGACAGAGC	833bp
Downstream	S1658R	Reverse	GCCTCGAAACGTGAGTCTTT	

2.2.2.4.1.2 Sequencing of 5' and 3' amplicons of the hygromycin resistance gene

For sequencing these amplicons to identify the mutations in neomycin (G418) resistant colonies, the 4 primers for PCR amplication were used (Table 2.17), and primers S588R and S1258F from Table 2.12. The reference sequence used is available on GitLab ("/mutation_detection_construct/reference_sequences/s.cerevisiae_reporter_construct.fa").

2.2.3 Cell culture methods

2.2.3.1 Preparation and growth of cell lines

Cell culture conditions used for the maintenance of the human cell lines in this thesis are summarised in Table 2.18. Cells were trypsinised in trypsin:versene (1:1, v/v) or in 1X TrypLE™ Express Enzyme (Thermo Fisher Scientific, catalog number 12605036) at 37°C for 5 min, and passaged at 1:6-1:12, as required.

Table 2.18. Cell culture conditions used for maintaining cell lines.

Cell line	Cell culture medium	Supplements of cell culture medium	Cell culture conditions	Source
HeLa	DMEM Medium (Thermo Fisher	10% fetal bovine serum, 100 U/ml penicillin, 100 µg/ml streptomycin	37°C, 5% CO_2, normoxic	Gift from G. Stewart (university of Birmingham); originally from ATCC

	Scientific, catalog number 41965039)			

2.2.3.1.1 Antibiotics used for HeLa cell culture

The antibiotics detailed in Table 2.19 were used for selection of HeLa strains.

Table 2.19. **Antibiotics used for HeLa cell selection in this study.**

Drug	Working concentration	Diluent	Storage temperature
Puromycin	0.5-20 µg/ml	dH_2O	–20˚C
G418 (Neomycin)	500 µg/ml	dH_2O	4˚C
Hygromycin	400 µg/ml	dH_2O	4˚C

2.2.3.2 Preservation of cell lines

For long term storage, 2–8 x 10^6 adherent cells were harvested and resuspended in 1 ml of FBS, containing 10% DMSO. The cells were stored in 2 ml cryostat tubes and frozen at –80˚C in Styrofoam containers, and after 24 hours transferred for long-term storage in liquid nitrogen.

2.2.3.3 Transfection of cultured mammalian cells

HeLa cells were transfected using Lipofectamine™ 2000 transfection reagent (Invitrogen) according to the manufacturer's instructions. One day before transfection, 5 x 10^5 cells were plated in 6-well plates in 2 ml of growth medium without antibiotics so that cells were 70-90% confluent at the time of transfection. The following day, for each reaction 250 ml OptiMEM (Gibco) was mixed with 15 uL lipofectamine 2000, and a further 250 ml of OptiMEM with 3 µgof the plasmid to be used for transfection. After leaving at RT for 5 minutes, these two solutions were mixed together, and 200 uL this added to a well in a 6 well plate of HeLa cells, washed with PBS. OptiMEM was added to a total volume of 1600 uL. After 4-6 hours the OptiMEM was replaced with 2 ml of DMEM growth medium. After 48 hours of growth, cells were single cell sorted using fluorescence-activated cell sorting (FACS) into a 96-well plate.

2.2.3.4 Generating human reporter cells to detect 2 bp deletions

2.2.3.4.1 Principles of the reporter construct

The sequence for the HeLa experiments was designed as described in Chapter 5 (Section 5.2.2) and the sequence ordered from GeneArt as pTCW4 (GitLab: “reference_sequences/pTCW4_GeneArt_order.gb”;Figure 2.4). The attL1 and attL2 homology arms upstream and downstream of the plasmid allowed integration of the a Gateway cloning compatible AAVS1 Destination vector (Chapter 5, Figure 5.3).

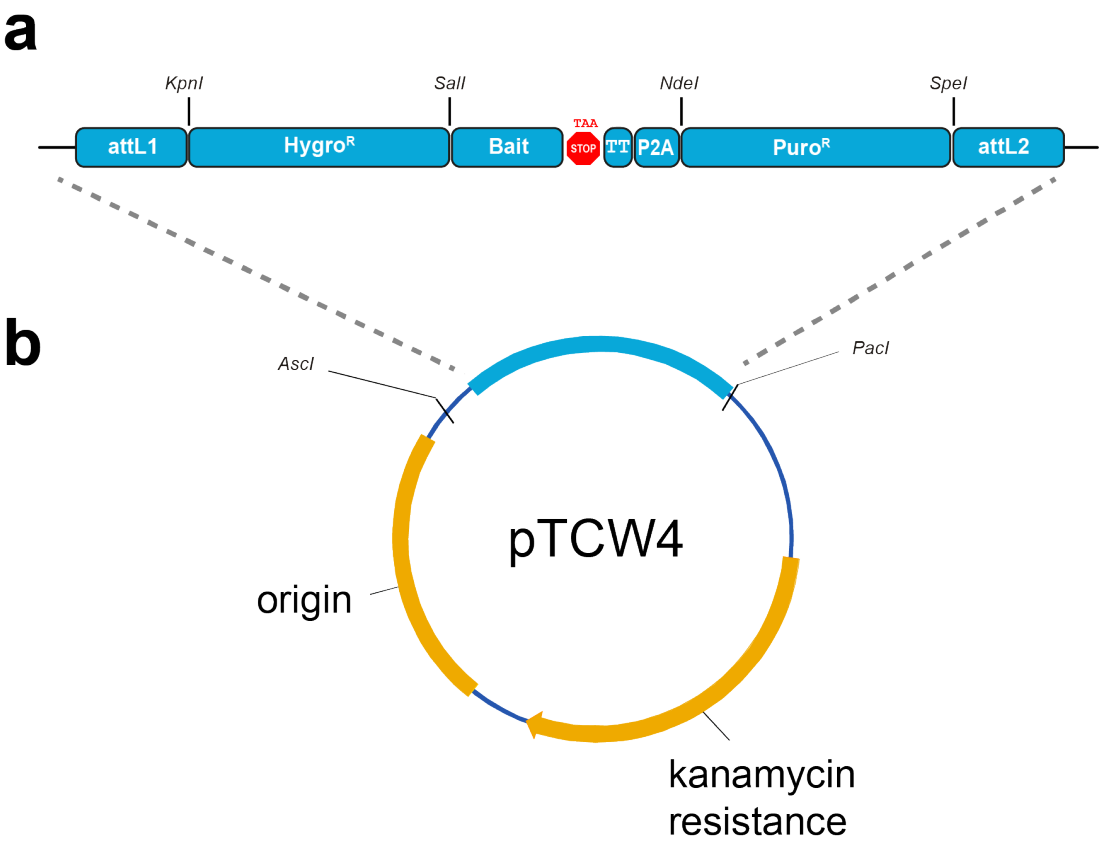


Figure 2.4 General schema for the reporter construct in HeLa cells. a) The reporter construct as originally designed. *This consists of an attL1 homology sequence, the hygromycin resistance gene (Hygro^R), followed in the original iteration by a “bait” sequence with tandem repeat sequences documented previously as being hotspots for short deletions. This bait sequence was followed by a stop codon, a two base pair frameshift, a self-cleaving peptide, a puromycin resistance gene (Puro^R) and an attL2 homology sequence. Restriction enzymes are marked above the sequence.* ***b) GeneArt order.*** *The gene sequence described above was ordered from GeneArt, and was manufactured as the requested sequence flanked by AscI and PacI cut sites, with a kanamycin resistance sequence to facilitate cloning.*

pTCW4 was then modified to replace the original hygromycin resistance gene with the version 2 used in the S. cerevisiae fluctuation assays (Figure 2.5). The backbone from pTCW4 ("/mutation_detection_construct/ reference_sequences/pTCW4_GeneArt_order.gb") was restriction digested using *Kpn*I-HF and *Nde*I, and the insert (length 1,106 bp) digested from pTCW12 (Section 2.2.4.3.8). The result of this reaction was pTCW14, which then underwent the LR reaction described below to create a vector to introduce the reporter construct into the HeLa genome (Chapter 5, Figure 5.3).

2.2.3.4.2 LR reaction to introduce human reporter construct into pAAVS1-Nst-CAG-DEST

An LR reaction to introduce the human reporter construct into the pAAVS1-Nst-CAG-DEST backbone was undertaken using a Gateway LR Clonase II reaction. The reaction consisted of 1 µl 100 ng/µl Entry clone (pTCW14), 1 µl 150 ng/µl pAAVS1-Nst-CAG-DEST, 6 µl TE buffer, pH8.0 and 2 µl of LR Clonase II. The reactions were incubated at 25 °C for 1 hour. Then 1 µl of Proteinase K solution was added to each sample to terminate the reaction, and incubated at 37 °C for 10 mins after briefly vortexing. Following this 1 µl of the reaction was transformed using chemically competent *E.coli* cells. The resulting plasmid (pTCW15, Figure 2.5c) was used to introduce the reporter construct into HeLa cells.

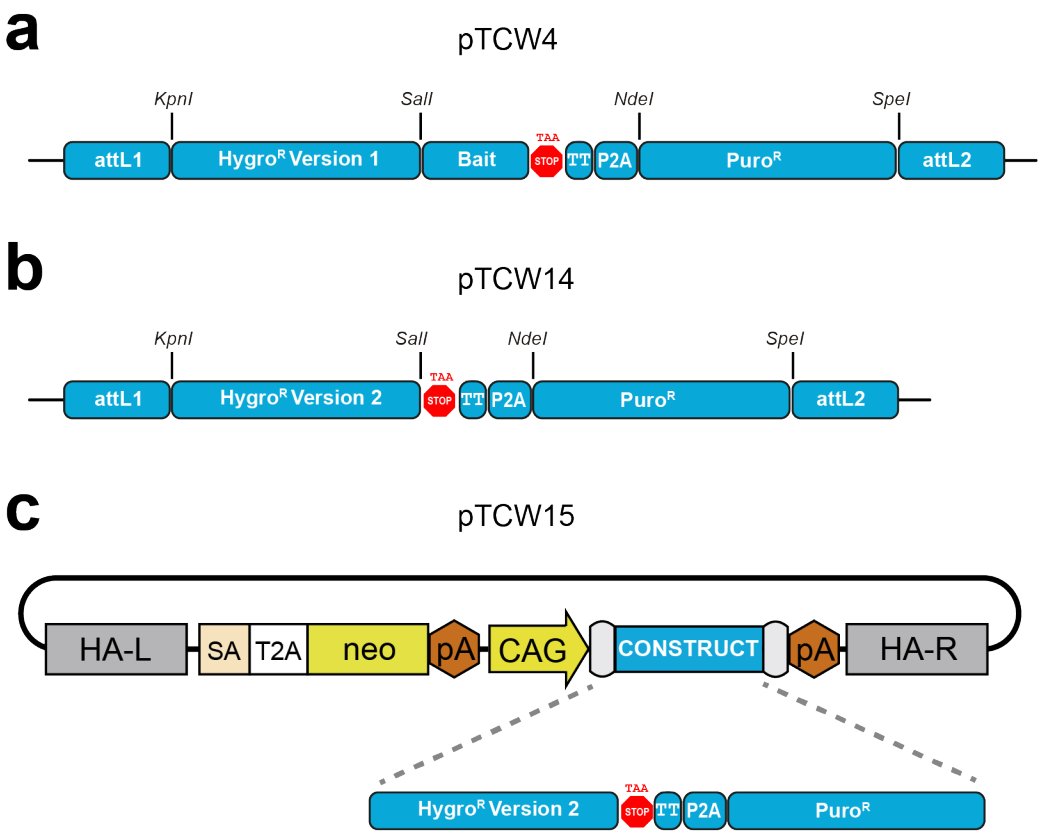


Figure 2.5. Modifications of plasmid to arrive at final construct for HeLa cells. a) pTCW4. *More details provided in Figure 2.4.* ***b) pTCW14.*** *Version 1 of the hygromycin resistance gene together with the "bait" sequence were replaced with Version 2 of the hygromycin resistance gene.* ***c) pTCW15.*** *Plasmid used to introduce reporter construct sequence into HeLa cells using homologous recombination. Further details in Chapter 5, Figure 5.3; adapted with permission from* (Oceguera-Yanez *et al.*, 2016).

2.2.3.4.2.1 Sequencing of pTCW15 (pAAVS1-Nst-CAG-hygroRv2-P2A-PuroR)

Prior to insertion by homologous recombination into the AAVS1 locus on chromosome 19 (Chapter 5, Figure 5.3) the entire construct (pTCW15) was sequenced using primer numbers 13-16 from Table 2.12 (S752F, S1258F, S588R, S1113R), primers in Table 2.24 (HygroR_up, PuroR_rev), and those in Table 2.20. Results were checked against the reference sequence (GitLab: "mutation_detection_construct/reference_sequences/pAAVS1-Nst-CAG-HumanReporterConstruct.fa").

Table 2.20. Primers used for Sanger sequencing of pAAVS1-Nst-CAG-hygroRv2-P2A-PuroR (pTCW15). Those starting with H are those designed for the purposes of this study; the remainder are from (Oceguera-Yanez *et al.*, 2016).

Primer number	Primer name	Primer sequence (5' → 3')

1	H1443F	GTCACCGAGCTGCAAGAACT
2	H1755F	GTCGAGGTGCCCGAAGGAC
3	H1327R	GTGGGCTTGTACTCGGTCAT
4	d218	CGAGTCAGTGAGCGAGGAAG
5	d216	TGTGCTGCAAGGCGATTAAG
6	d1131	TGAATGAACTGCAGGACGA
7	d378	ACAGATAAAAGTACCCAGAACCAG
8	pCAG_F	GCAACGTGCTGGTTATTGTG
9	Bglob_p_R	TTTTGGCAGAGGGAAAAAGA

2.2.3.4.2.2 Integration of human reporter construct into HeLa cells

Version 2 of the human reporter construct was integrated into the HeLa genome following a published protocol (Oceguera-Yanez *et al.*, 2016). The pXAT2 plasmid, expressing a guide RNA to enable a double stranded break at the AAVS1 locus (Chapter 5, Figure 5.3), was co-transfected with pTCW15 (Table 2.4, methods in Section 2.2.3.3). pTCW15 contained version 2 of the human reporter construct with homology arms for the AAVS1 locus. A double stranded break at the AAVS1 locus meant that the sequence between the AAVS1 homology arms in pTCW15 could be integrated into the HeLa genome by homologous recombination. After 48 hours of transfection cells were re-seeded into a 10-cm plate containing 500 µg/ml G418. After a further 48 hours of growth the cells were trypsinised and transferred to a T75 flask. They were then single cell sorted by FACS into 96-well plates. Once the colonies in the 96-well plate had reached confluency, the plate was trypsinised and split into two plates, with one plate frozen after packing with tissue paper in a Styrofoam container, and the second tested for integration of the construct as detailed below.

2.2.3.4.2.3 Screening for clones with correct integration of the reporter construct

A Prime Star PCR reaction (2.2.4.3.9.1) was conducted to confirm integration of the human reporter construct into the AAVS1 locus in the HeLa genome (Chapter 5, Figure 5.3) using lysates prepared as per Section 2.2.4.3.6. For detection of the construct, primers dna803 and dna804 were used; for detection of the wild type genotype, primers dna803 and dna183 were used (Table 2.21).

Table 2.21. Primers used to confirm integration of human reporter construct into HeLa cells, and confirmation of RNASEH2A knock out (Figure 5.4).

Primer	Orientation	Primer sequence (5' → 3')	Amplicon length (genotype)
dna803	Forward	TCGACTTCCCCTCTTCCGATG	dna803-dna804: 1.2 k (reporter) dna803-dna183: 1.4 kb (WT) HeLa_constr_R-dna183: 1.3 kb (reporter)
dna804	Reverse	GAGCCTAGGGCCGGGATTCTC	
dna183	Reverse	CTCAGGTTCTGGGAGAGGGTAG	
HeLa_constr_R	Forward	GTCACCGAGCTGCAAGAACTC	
RNASEH2A-ex1F	Forward	ACCCGCTCCTGCAGTATTAG	262 bp (WT)
RNASEH2A-ex1R	Reverse	TCCCTTGGTGCAGTGCAATC	

PCR reactions were performed using the following settings:

34 x	Denaturation	95˚C	10 sec
	Annealing	70˚C	15 sec
	Extension	72˚C	90 sec
Extension		72˚C	90 sec

Clones which contained both a wild type locus and copy of the reporter construct were then tested for genomic integration at the 3' end of the construct, using primers dna183 and HeLa_constr_R (Table 2.21), and the reaction detailed above.

2.2.3.4.3 Sequencing to confirm correct sequence of reporter construct in HeLa cells

2.2.3.4.3.1 Amplification of entire construct in HeLa cells

Once integration of the construct into the AAVS1 locus was confirmed, the sequence was checked. The construct was amplified using a Prime Star PCR reaction (Section 2.2.4.3.7.1) using forward primer HygroR_up and reverse primer PuroR_rev (Table 2.24), and sequenced using primer numbers 13-16 from Table 2.12 (S752F, S1258F, S588R, S1113R), the primers in Table 2.24 (HygroR_up, PuroR_rev), and in addition primer numbers 1-3 in Table 2.22 (H1443F, H1755F,H1327R). Results were checked against the reference sequence (GitLab: "/mutation_detection_construct/reference_sequences/human_reporter_construct.fa").

Table 2.22. Additional primers used for sequencing of reporter construct in HeLa cells

Primer name	Primer sequence (5' → 3')

H1443F	GTCACCGAGCTGCAAGAACT
H1755F	GTCGAGGTGCCCGAAGGAC
H1327R	GTGGGCTTGTACTCGGTCAT

2.2.3.4.4 Inactivation of RNASEH2A in HeLa cells containing reporter construct

This was conducted using Lipofectamine™ 2000 transfection (Invitrogen) in 6-well plates (Section 2.2.3.3). Plasmids pMAR527 and pMAR527 (Table 2.3) were used, which express the Cas9 enzyme and guide RNAs for exon1 of RNASEH2A respectively. In addition pMAR527 contains the coding sequence for GFP. After transfection and 48 hours of growth in DMEM cells expressing GFP were sorted by FACS into 96-well plates. Once confluency was reached, clones were replated and processed as described in Section 2.2.4.3.6. Lysates were used for PCR amplification of the genomic region targeted by guide RNAs. Primers RNASEH2A-ex1F and RNASEH2A-ex1R (Table 2.21) were used to amplify and then sequence exon 1 of RNASH2A, using a ReddyMix PCR reaction (Section 2.2.4.3.9.1) and the Touchdown protocol (Korbie and Mattick, 2008) outlined below:

	Denaturation	94˚C	5 min
2x	Denaturation	94˚C	30 sec
	Annealing	65˚C	30 sec
	Extension	72˚C	45 sec
6x	Denaturation	94˚C	30 sec
	Annealing	62˚C	30 sec
	Extension	72˚C	45 sec
10x	Denaturation	94˚C	30 sec
	Annealing	59˚C	30 sec
	Extension	72˚C	45 sec
34x	Denaturation	94˚C	30 sec
	Annealing	56˚C	30 sec

Extension	72˚C	45 sec
Extension	72˚C	5 min

PCR products were separated by agarose gel electrophoresis. Clones which showed evidence of possible frameshift mutations at the RNASEH2A locus were sequenced using Sanger sequencing (Section 2.2.4.3.7.1) and carried through to an RNase H2 assay (Section 2.2.5.2). Results from these are shown in Table 2.23.

Table 2.23. HeLa clones used in fluctuation assays

Name	RNase H2 status	Sequencing of RNase H2A exon 1	Description
A1	Wild type	No mutation	Ancestral HeLa clone with reporter construct incorporated
A1.P1.G1 (2)	Wild type	10 bp equivalent deletion	HeLa clone with reporter construct incorporated, underwent CRISPR-Cas9 gene editing to knock out RNASEH2A, but maintains RNase H2 activity (42%) on RNase H2 assay consistent with expression of wild type enzyme.
A1.P1.E2 (1)	RNase H2 null	Complex indel	HeLa clone with reporter construct incorporated, underwent CRISPR-Cas9 gene editing to knock out RNase H2, with low RNase H2 activity (9%) on RNase H2 assay consistent with complete loss of enzymatic activity.
A1.P2.H4 (5)	RNase H2 null	38 bp deletion	HeLa clone with reporter construct incorporated, underwent CRISPR-Cas9 gene editing to knock out RNase H2, with low RNase H2A activity (7%) on RNase H2 assay assay consistent with complete loss of enzymatic activity.

2.2.3.4.5 Human fluctuation assay

2.2.3.4.5.1 Growth of cell cultures

Cells were recovered from frozen stocks into T75 flasks. The following day the cells in each flask were trypsinised and cell numbers determined using a Moxi™ Z automated cell counter (Avantor). For each clone, 10 wells of a 96-well plate were each seeded with 2000 cells. These 96-well plates were then re-labelled to blind them (by Carol-Anne Martin, Jackson Lab), with clone numbers replaced with a letter. Cells were then moved from 96 to 24 then 6-well plates, and eventually to T75s. These cultures were then grown until confluency.

2.2.3.4.6 Plating

At this stage, cells in the T75s were trypsinised with 1.5 ml of TriplE. 18.5 ml of cell culture media was added, and cell suspensions (total volume of 20 ml)

moved to a Falcon 50 ml container. The cell concentration for each culture was measured using the Moxi automated cell counter. For each Falcon, a 1:50 dilution was made by adding 20 µl of cells to an Eppendorf with 980 µl of cell culture media. Based on the cell concentrations, the volume of this dilution required for 1000 cells was added to 30 ml of cell culture media, and 15 ml of this dilution plated into 2 x 10 cm plates. A total of 1000 cells was chosen, as plating efficiency was found to be between 25-50% in pilot experiments, leaving 250 to 500 cells to be counted in each plate.

The 20 ml of original cells were split between 2 selection plates, 10 ml in each 10-cm plate. After 4 hours, to allow the cells time to adhere to plate surface, 6 ml of cell culture media with 4 µl of 2 mg/ml puromycin was added to each selection place, for a final selection concentration of 0.5 µg/ml.

2.2.3.4.7 Growth of plating efficiency and selection plates

Plates used for calculation of the denominator of population size at the start of selection were grown without any change in media for 14 days. These were used to estimate the starting population of viable cells for each independent clone. For the selection plates, medium was changed every 2-3 days for 14 days, maintaining a puromycin concentration of 0.5 µg /ml.

2.2.3.4.8 Selection of colonies for sequencing, plate fixing and staining

From each selection plate, 2 colonies were taken for culture in separate wells of a 96-well plate (only 1 of the 4 would contribute towards mutation spectrum calculations), using a 20 µl pipette tip to scrape off each colony. After 5 days of growth DNA was extracted for cells in these plates as described in Section 2.2.4.3.6.

After removal of colonies for sequencing, both selection and non-selection plates were washed with 10 ml PBS. They were fixed by adding 5 ml of PBS with 2% formaldehyde for 10 minutes, and then rinsed with water. 5 ml of crystal violet 0.1% solution was added for 10 minutes, and then removed and the plate washed with water and left to dry. Plates were photographed as described in Section 2.3.3.1.2).

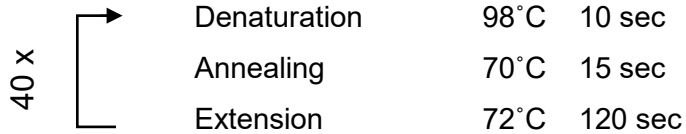
2.2.3.4.9 Sequencing of mutant colonies to determine underlying mutation

PCR amplification of DNA from mutant colonies was conducted as described in Section 2.2.4.3.7.1, using the primers in Table 2.24.

Table 2.24. Primers used to generate amplicon containing hygromycin resistance gene in HeLa cells.

Primer	Orientation	Primer sequence (5' → 3')	Amplicon length
HygroR_up	Forward	CAACGTGCTGGTTATTGTGC	1893bp
PuroR_rev	Reverse	CAGCCTGCACCTGAGGAGTG	

PCR reactions were performed using the following settings:



2.2.3.4.9.1 Sequencing of the hygromycin resistance gene in HeLa cells

The primers used to sequence the hygromycin resistance gene in mutant colonies were primer numbers 13-16 from Table 2.12 (S752F, S1258F, S588R,S1113R), HygroR_up from Table 2.24, and H1327R from Table 2.22. The reference sequence used is available on GitLab ("/mutation_detection_construct/reference_sequences/human_reporter_construct.fa").

2.2.3.4.10 CRISPR-Cas9 positive control experiment to create frameshift mutations in the human reporter construct

A positive control CRISPR-Cas9 experiment was designed to demonstrate proof of concept for the human reporter construct. Benchling (*CRISPR Guide RNA Design Software for Molecular Biology | Benchling*, no date) was used to select a gRNA target within the Hygro^R sequence. Primers HygroR-sgRNA-F and HygroR-sgRNA-R (Table 2.5) were annealed at concentration of 1 µM. Subsequently this insert was ligated (Section 2.2.4.3.11) into a backbone of pX458 (Table 2.3) digested with *Bbs*I (Section 2.2.4.3.8). The resulting plasmid (pTCW17, Table 2.4) was transfected using Lipofectamine™ 2000 (Invitrogen) in a 6-well plate (Section 2.2.3.3). After incubation for 48 hours, the wells were

split, half of the cells left in situ at dilutions of 1:2, 1:5 and 1:10 (Figure 5.4), and the other half diluted and reseeded into a 10-cm plate. Puromycin 2 µg/ml was added to the cells, and after 1 week single colonies trypsinised and placed using a 20 µl pipette tip into single 96 well plate wells. Once these had reached confluence, DNA was isolated (Section 2.2.4.3.6) and the region was amplified using a Prime Star Reaction (Section 2.2.4.3.7.1). Sequencing of a number of mutants (Section 2.2.4.3.7.1) confirmed 2 bp equivalent mutations: a 11 bp deletion, and two 1 bp insertions.

2.2.3.5 RPE1 mutation accumulation experiment

These experiments were conducted by Martin Reijns, using an approach outlined in Chapter 5, Figure 5.6. Descendent clones underwent Illumina Whole Genome Sequencing by Edinburgh Genomics, with the analysis described in Section 2.3.3.2.2.

2.2.4 Nucleic acid methods

2.2.4.1 Quantification

2.2.4.1.1 NanoDrop

The concentration of nucleic acids was determined by measuring the optical density at 260 nm using a NanoDrop 1000 UV-Vis Spectrophotometer (Thermo Fisher Scientific). 1.5 µl of each sample was used for each measurement. The purity of the nucleic acid sample was determined by measuring the absorbance at 230 nm, 260 nm and 280 nm. The 260/280 ratio of a sample free of protein contamination should be 1.8-2.2, and a 230/260 ratio ≥1.7 indicates a sample free of carbohydrates and lipids.

2.2.4.1.2 Invitrogen™ Qubit™ Flurometer

For nanopore sequencing, more accurate concentrations than those provided by the NanoDrop Spectrophotometer were required, in order to optimise sequencing conditions. Therefore nucleic acid concentrations were determined using a Qubit 1X dsDNA HS Assay Kit as per the manufacturers instructions.

2.2.4.2 Nucleic acid gel electrophoresis

2.2.4.2.1 Agarose gels

Nucleic acid samples were analysed by agarose electrophoresis on gels ranging from 0.7% to 1.5% agarose in 0.5 x TBE (w/v). Gels were prepared by dissolving agarose (Hi-Pure Low EEO agarose, Biogene) in 0.5X TBE buffer by boiling for 90 seconds in a microwave oven and adding either ethidium bromide (after cooling to ~50°C) to a final concentration of 0.5 µg/ml, or SYBR Safe (Invitrogen, Thermo Fisher Scientific, catalog number 1691992) in a 1:1000 concentration. DNA samples were mixed with 6X DNA loading buffer (30% (v/v) glycerol, 0.4% (w/v) Orange G), loaded on to the gel, and 5 volts/cm applied to resolve the nucleic acid fragments by size. The nucleic acids were visualised using a UV transilluminator (BioDoc-It System, UVP) or, for SYBR Safe stains, a blue light transilluminator. For reference, markers containing DNA fragments of known sizes were included (1kb DNA Ladder, Invitrogen or 100bp DNA Ladder, Promega).

2.2.4.2.2 Native PAGE

To make an 8% 50 ml gel, 10 ml of 40% (19:1) acrylamide: bis-acrylamide solution (Severn Biotech Ltd) was mixed with 40 µl TMED and 1.25 ml of 20x TBE buffer and made up to 50 ml with H_20. The solution was gently mixed and 400 µl of 10% APS added. Gels were run at 5W in a buffer solution of 800 ml 0.5 x TBE for period determined by length of DNA fragments to be visualised (generally 2-3 hours).

2.2.4.2.3 Denaturing PAGE

To prepare formamide loading dye (FLD) 960 µl formamide was mixed with 40 µl 0.5 M EDTA (pH 8) to give a final concentration of 20mM EDTA. To make a 10% gel, 21g of urea was mixed with 12.5 ml of 19 acrylamide:1 polyacrylamide, 2.5 ml of 20x TBE, 16.5 ml of H_2O and 40 µl of TEMED. After mixing, 400 µl 10% APS was added. Gels were run in a buffer solution of 800 ml 1x TBE, pre-heated by running at 25 W for 30 minutes prior to loading, and run at 25 W for a period determined by the length of the DNA fragments to be visualised (generally 1-2 hours).

2.2.4.3 DNA methods

2.2.4.3.1 Phenol/chloroform extraction

Genomic DNA was isolated from *S. cerevisiae* and RPE1 cells using phenol/chloroform extraction. For *S. cerevisiae*, genomic DNA was isolated as per the emRiboSeq protocol (Ding *et al.*, 2015). A pellet from a 5 ml yeast culture (5 $A600_{nm}$ units) was resuspended in 200 µl lysis buffer (Table 2.1). An equal volume of TE-equilibrated phenol and glass beads (0.40–0.60 mm diameter, Sartorius) were added, and cells lysed by vortexing for 2 min; 200 µl TE buffer was then added, followed by an additional 1 min of vortexing. Extraction of DNA from RPE1 cells was conducted by Martin Reijns as previously described (Benitez-Guijarro *et al.*, 2018). Total nucleic acids were isolated from cells by lysis in ice-cold buffer (20 mM Tris-HCl pH 7.5, 75 mM NaCl, 50 mM EDTA) and subsequent incubation with 200 µg/ml Proteinase K (Roche) for 10 min on ice followed by addition of sarkosyl (Sigma) to a final concentration of 1%. Following these steps, nucleic acids were sequentially extracted with TE-equilibrated phenol, phenol:chloroform:isoamyl alcohol (25:24:1), and chloroform, and then precipitated with isopropanol. At each stage 200-400 µl of each reagent was added, the mixture spun at 5,000g, and the aqueous phase taken into the next step of the extraction.

2.2.4.3.2 Bead extraction

After phenol/chloroform extraction the pellets were dissolved in 50 µl H_2O (*S. cerevisiae*) and 0.5M NaCl (RPE1). 1 µl of DNase-free RNase (Roche, 0.5 µg/ul) was added and the solution incubated for 1 hour at RT (20–25 °C). 50 µl of Ampure XP beads (Beckman Coulter, cat. no. A63880) was added and incubated for >1 minute, the mixture was placed on a magnetic stand, and the supernatant discarded. Beads were washed twice with 500 µl of 75% EtOH, and beads left to air dry before re-suspension in 50 µl of TE (10 mM Tris-HCl pH8.0, 1 mM EDTA). DNA concentration and quality were determined using NanoDrop (section2.2.4.1.1), Qubit dsDNA BR Assay (Thermo Fisher Scientific, Section 2.2.4.1.2) and gel electrophoresis (Section 2.2.4.2.1).

2.2.4.3.3 Ethanol precipitation

For some nanopore experiments DNA was ethanol precipitated to allow buffer change. The DNA solution to be purified was mixed with 1 µl glycogen, 10 µl Sodium Acetate, and 250 µl 100% ethanol. This solution was stored for >1 hour at -20 °C, then spun at 13,300 g for 30 minutes at 4 °C. The ethanol was removed and the pellet was washed twice with 250 µl of 75% ethanol. After air drying, the pellet was dissolved in the appropriate volume of Elution Buffer (Nanopore Technologies).

2.2.4.3.4 Purification of DNA from nanopore sequencing experiments, PCR reactions and restriction digests

To purify the products of extension reactions for nanopore experiments, PCR products for sequencing, column or gel purification was conducted. For extraction of products of restriction digests, gel purification was conducted. For column purification, the QIAquick PCR Purification kit (Qiagen) was used according to the manufacturer's instructions. For gel purification, DNA fragments produced by restriction digestion were resolved by agarose gel electrophoresis using SYBR Safe (Life Technologies).The desired DNA fragment was excised from the gel using a scalpel and purified using the QIAquick Gel Extraction kit (Qiagen) according to the manufacturer's instructions. For both gel and column purification, DNA was eluted in 30 µl of Elution Buffer (Qiagen or Nanopore Technologies) unless stated otherwise and stored indefinitely at –20˚C.

2.2.4.3.5 Purification of plasmid DNA

Plasmid DNA was prepared using the QIAprep Spin Miniprep Kit (Qiagen) following the manufacturer's instructions. DNA was extracted from 5 ml of stationary phase *E. coli* culture and eluted in 50 µl of elution buffer (10mM Tris-HCl pH 8.5) or DNase/RNase-free distilled water (Gibco). For larger scale purifications, the ZymoPURE II Plasmid Maxiprep Kit (Zymo Research) was used. In this case, DNA was extracted from 100-150 ml of stationary phase *E. coli* following the manufacturer's instructions and eluted in 400 µl of the elution buffer provided.

2.2.4.3.6 Preparation of DNA from HeLa cells for amplification and Sanger sequencing

A confluent well in a 96-well plate was emptied and the cells lysed with 75 µl of DirectPCR (yolk sac) Lysis Buffer (Viagen Biotech) with Proteinase K (20 mg/mL, Roche) added (20 µl added to 1 ml of Lysis Buffer). The plate was incubated at 55 °C overnight, then heated to 85 °C for 45 minutes to inactivate the Proteinase K.

2.2.4.3.7 DNA sequencing

2.2.4.3.7.1 Sanger Sequencing

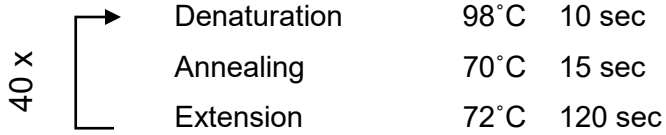
Dye terminator sequencing reactions (ABI) were performed and processed by the Institute of Genetics and Molecular Medicine (IGMM) sequencing service on a 3130/3730 genetic analyser (Applied Biosystems). DNA sequencing data was analysed using either Sequencher 5.4.6 (Gene Codes Corp.) or Mutation Surveyor® (SoftGenetics).

A typical PCR reaction to screen for mutations in a *S. cerevisiae* fluctuation assay contained template DNA (picked from a single yeast colony with a sterile pipette tip), 5 µl FastStart PCR Master Mix (Roche, catalog number 04710436001), 0.5-1 µM forward primer, 0.5-1 µM reverse primer and dH2O up to a total reaction volume of 10 µl. PCR reactions were performed on a DNA Engine Tetrad 2 thermal cycler (MJ Research) using the following settings:

	Denaturation	95˚C	5 min
34 x	Denaturation	95˚C	30 sec
	Annealing	58˚C	30 sec
	Extension	72˚C	45 sec
	Extension	72˚C	45 sec

A typical PCR reaction to amplify the hygromycin resistance gene for detection of mutations accumulating during the HeLa fluctuation assay contained: Prime Star Max PCR Master Mix (Takara Bio), 0.5-1 µM forward primer, 0.5-1 µM reverse primer, 1-100 ng of DNA and H_2O to a final volume of 10 µl.

A typical PCR reaction for a Prime Star PCR Master Mix reaction was performed on a DNA Engine Tetrad 2 thermal cycler (MJ Research) using the following settings:



PCR products were separated on a 1.5% 0.5X TBE agarose gel to confirm correct amplification before Sanger sequencing.

2.2.4.3.8 Restriction digests

Plasmid DNA was digested with restriction endonucleases in buffer supplied by the manufacturer (NEB). Digests were performed in 100 µl of buffer with 2–5 µg of DNA and 20 U of the appropriate enzyme(s) overnight,at the appropriate temperature. For double digests, optimal buffer conditions were selected for both enzymes using guidelines from the manufacturer. DNA fragments were purified as described in Section 2.2.4.3.4

2.2.4.3.9 Amplification of DNA by polymerase chain reaction (PCR)

Primers were designed using the Primer3Plus application (Untergasser *et al.*, 2007), specifying an optimal melting temperature (T_m) of 60˚C, which did not differ by more than 6°C between the forward and reverse primers in a pair. The oligo melting temperature formula used was that from (Breslauer *et al.*, 1986).

2.2.4.3.9.1 Polymerase chain reaction (PCR)

Specific regions of DNA were amplified using a polymerase chain reaction (PCR). Different DNA polymerase containing master mixes were used: ThermoPrime 2X ReddyMix PCR Master Mix with 1.5 mM MgCl2 (Thermo Fisher Scientific, catalog number AB0575DCLDB) for standard PCR, FastStart PCR Master Mix (Roche, catalog number 04710436001) for colony PCR, and Phusion Flash High-Fidelity PCR Master Mix (Thermo Fisher Scientific, catalog number F548S) for when a high level of polymerase accuracy was required.

A typical PCR mixture using Phusion Flash Master Mix contained 10 µl Phusion Flash PCR Master Mix (ThermoFisher, catalog number F548S), 0.5-0.1 µM forward primer, 0.5-1 µM reverse primer, 1-100 ng template DNA (plasmid or genomic) and dH_2O up to a total reaction volume of 20 µl.

A typical PCR reaction using Phusion Flash Master Mix was performed on a DNA Engine Tetrad 2 thermal cycler (MJ Research) using the following settings:

Denaturation		98˚C	10 sec
34 x	Denaturation	98˚C	1 sec
	Annealing	60˚C	5 sec
	Extension	72˚C	90 sec
Extension		72˚C	90 sec

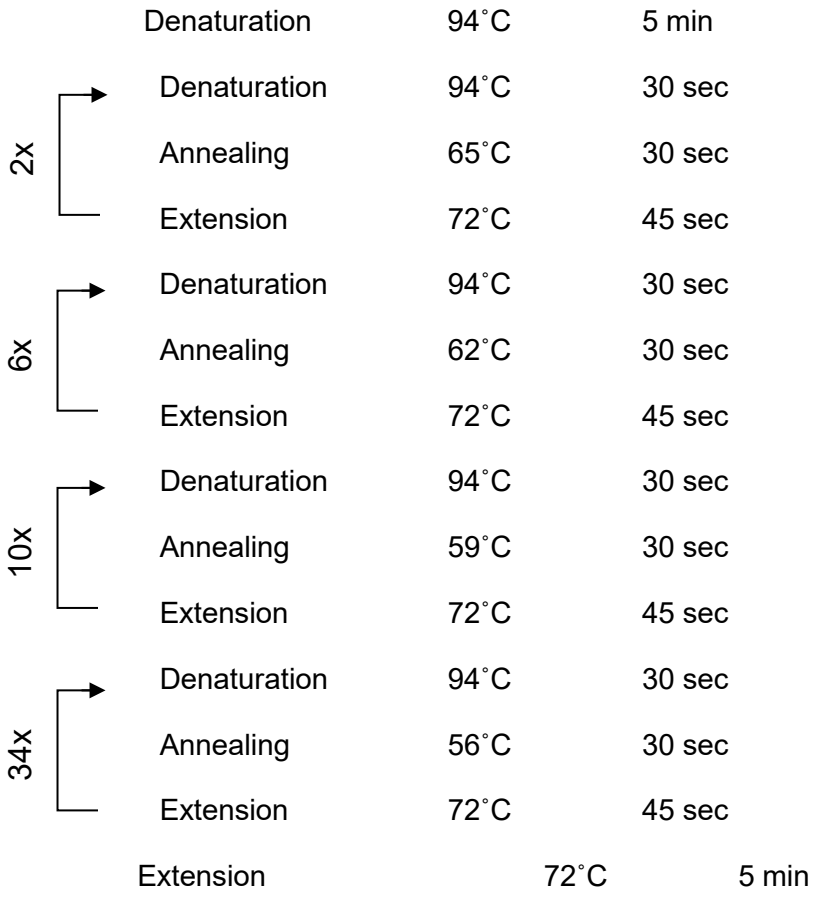
A typical colony PCR reaction was prepared using template DNA (from a an entire single bacterial colony, or a sample from a yeast colony, both picked with a sterile pipette tip), 5 µl FastStart PCR Master Mix (Roche, catalog number 04710436001), 0.5-1 µM upstream primer, 0.5-1 µM downstream primers, and 4 µl H_2O.

A typical colony PCR reactions was performed on a DNA Engine Tetrad 2 thermal cycler (MJ Research) using the following settings.

Denaturation		95˚C	5 min
34 x	Denaturation	95˚C	30 sec
	Annealing	58˚C	30 sec
	Extension	72˚C	30 sec
Extension		72˚C	30 sec

A typical PCR using ReddyMix consisted of 5 µl ThermoPrime 2X ReddyMix PCR master mix (Thermo Scientific) with 1.5 mM $MgCl_2$, 0.5-1 µM forward primer, 0.5-1 µM reverse primer, 1-100 ng template DNA (genomic) and dH_2O up to a total reaction volume of 10 µl.

A typical ReddyMix PCR reaction was performed using a Touchdown PCR program on a DNA Engine Tetrad 2 thermal cycler (MJ Research) using the following settings.



2.2.4.3.10 Site-directed mutagenesis

Primers for use in site-directed mutagenesis were designed with the following conditions: between 25 and 55 bases in length, with a melting temperature (T_m) of ≥78˚C (see below), and with desired mutation located near the middle of the primer with a minimum of ~10–15 bases of correct sequence on either side. The T_m of primers was calculated using the following formula:

$$T_m = 81.5 + 0.41(\%GC) – 675/N – \%mismatch$$

in which N = the primer length in bases,

percentage GC = (#G bases + #C bases)/oligonucleotide length, and

percentage mismatch = #mismatches/oligonucleotide length.

Point mutations were introduced into plasmid vectors using the PCR based QuikChange method (Stratagene). Primers containing the desired mutation were designed to anneal to the same sequence on opposite strands of the plasmid and were extended by PCR, generating a mutated plasmid. The PCR contained: 50 ng plasmid, 0.25 mM dNTPs, 0.2 µM mutagenic primers (forward and reverse), 1X DNA polymerase buffer with $MgCl_2$ (Stratagene) and 1.25 U of *Pfu*Turbo DNA polymerase (Stratagene). Cycling parameters for site-directed mutagenesis were as follows:

Denaturation		95˚C	1 min
18x	Denaturation	95˚C	30 sec
	Annealing	55˚C	1 min
	Extension	68˚C	1 min per kb
Extension		68˚C	10 min

The PCR product was treated with 10 U of *Dpn*I (NEB) at 37˚C overnight to digest the methylated parental vector DNA and thereby select for the mutation-containing synthesised DNA. Following incubation, 0.5–1 µl of the *Dpn*I treated DNA was used for transformation into DH5α *E. coli* (Section2.2.2.2.2). Resulting colonies were mini-prepped and screened by Sanger sequencing (see Section 2.2.4.3.7.1).

2.2.4.3.11 Ligation of DNA molecules

For the nanopore experiments, Blunt/TA Ligase Master Mix purchased from New England Biolabs, (Catalog number M0367S). Reactions consisted of 10-25 µl of solution containing DNA to be ligated and an equal volume of Blunt/TA Ligase Master Mix. The reaction was left at RT for 1 hour. For the creation of the plasmids used to transform yeast and HeLa cells. Ligation reactions contained: 100–200 ng of vector DNA, 2–3X this molar amount of insert DNA, 1 U T4 DNA Ligase (Roche) and 1X Ligation Buffer (Roche). Following incubation for 4–5 h at room temperature, 1 µl ligation mixture was used to transform *E. coli* (Section 2.2.2.2.2).

2.2.4.3.12 Phosphatase reactions

A typical reaction comprised 1 µg of plasmid DNA, 5 µl 10x Antartic phosphatase buffer, and 1 µl Antarctic phosphatase (New England Biolabs, catalogue number M0289S) to a total reaction size of 50 µl, with the reaction run for 2 hours at 37˚C.

2.2.4.4 Fluorescence in situ hybridisation (FISH)

FISH was used to make decisions on which cell lines to take forward into fluctuation assays: to determine the copy number for reporter construct insertion into HeLa cells. There was a high risk of duplication of the chromosome containing the reporter construct, given the high levels of genomic instability in HeLa cell lines.

To generate a plasmid for the probe, the coding sequence within the homology arms integrated into the HeLa genome was required. pTCW15 (Table 2.4) was digested with *Hind*III and *Sca*I, and pK18 (Table 2.3) was digested with *Hind*III. The heaviest bands from both reactions (5,279 and 2,661 base pairs, respectively) were gel purified and ligated, and heat transformed to create pTCW16. The sequence was confirmed using primers M13F and M13R. FISH was conducted by Shelagh Boyle, using methods described in (Boyle *et al.*, 2020). Any modifications to this protocol are described below.

2.2.4.4.1 Harvesting cells for metaphase spreads

Colcemid, a colchicine analogue which inactivates spindle fibre formation and limitis microbule formation, leading to the arrest of cells in metaphase, was used to prepare cells for metaphase spreads. HeLa cells were cultured overnight in order to yield an ~80% confluent T75 flask with a high proportion of replicating cells. Colcemid (KaryoMAX colcemid 15210-040 10µg/ml) 1:100 was then added to media for 45 mins prior to harvesting.

2.2.4.4.2 Nick translation

To identify the DNA sequence from the reporter construct, labelled dUTP was used. Nick translation was used to generate labelled dUTP containing probe DNA that would bind to the reporter construct. Plasmid pTCW16 (Table 2.4)

was used as the substrate for nick translation, with Green496-dUTP (ENZ-42831L, ENZO Lifesciences) as the labelled dUTP.

2.2.5 Protein methods

2.2.5.1 Protein preparation from HeLa cells

2.2.5.1.1 Whole cell extracts preparation

For whole cell extract preparation, cultured HeLa cells were harvested from a confluent well of 6-well plate by trypsinisation. 4.5 ml of DMEM were added and the cells were spun down at 1,200 rpm. The cells were washed with 1 ml of PBS, and spun down for 1 minute at 0 °C at 5,000 rpm. The supernatant was removed and the pellet stored at –80°C until required.

Whole cell extracts were prepared as described in (Benitez-Guijarro *et al.*, 2018). Cells were incubated in lysis buffer [50 mM Tris–HCl pH 8.0, 280 mM NaCl, 0.5% NP-40, 0.2 mM EDTA, 0.2 mM EGTA, 10% glycerol (vol/vol), 1 mM DTT and 1 mM phenylmethyl-sulfonyl fluoride (PMSF)] for 10 min on ice, followed by the addition of an equal volume of 20 mM HEPES pH 7.9, 10 mM KCl, 1 mM EDTA, 10% glycerol (vol/vol), 1 mM DTT and 1 mM PMSF for an additional 10 min. Whole-cell extracts were cleared by centrifugation (17,000 g for 10 min at 4°C), and protein concentration was determined as described below.

2.2.5.1.2 Protein quantification

Total protein concentrations in cell extracts or purified recombinant protein eluates were determined using Quick Start Bradford Protein Assay (Bio-Rad). Initially, a standard curve was drawn using a BSA concentration range of 0.2, 0.4, 0.6, 0.8 and 1.0 mg/ml, provided by manufacturer. 10 µl of each solution was mixed with 190 µl of 1X Bradford dye reagent and the absorbance at 595 nm (A_{595}) was measured after 5 min. Absorbance readings were plotted against protein concentration, and a line of best fit was used. To measure unknown protein concentrations, 1 µl of protein sample was mixed with 9 µl dH_2O and 190 µl of 1X Bradford dye reagent, and A_{595} was measured after 5 min. The absorbance reading was compared to the BSA standard curve to calculate protein concentration.

2.2.5.2 RNase H2 assay

Whole cell lysates were prepared as detailed in Section 2.2.5.1.1; the final protein concentration used per reaction was 100 ng per µl. The assay was performed as per (Reijns *et al.*, 2011). To measure enzyme activity, substrate was formed by annealing a 3′-fluorescein-labelled oligonucleotide (GATCTGAGCCTGGGaGCT, DRD–DNA; uppercase DNA, lowercase RNA) to a complementary 5′-dabcyl-labelled DNA oligonucleotide (Eurogentec). RNase H2 specific activity was determined against the DRD–DNA substrate. Activity against a double-stranded DNA substrate of the same sequence was measured and used to correct for non-RNase H2 activity against the DRD–DNA substrate. Reactions were performed in 100 µl of buffer (60 mM KCl, 50 mM Tris–HCl pH 8.0, 10 mM MgCl2, 0.01% BSA, 0.01% Triton X-100) with 250 nM substrate in 96-well flat-bottomed plates at 25 °C. Fluorescence was read for 100 ms using a VICTOR2 1420 multilabel counter (Perkin Elmer), with a 480-nm excitation filter and a 535-nm emission filter.

2.2.6 Nanopore laboratory methods

2.2.6.1 General note

After creation of the DNA substrate for sequencing, the Oxford Nanopore Technologies standard protocol for 1D^2 sequencing of genomic DNA with SQK-LSK308 (GitLab: “/nanopore/nanopore_protocols/1D2_protocol.pdf”) was followed unless stated otherwise. Of note, the 1D^2 sequencing kit is now longer available.

2.2.6.2 Nanopore specific reagents

Klenow (DNA Polymerase I, Large Fragment) was purchased from New England Biolabs (Catalog number M0210S). Klenow Fragment 3’-5’ exonuclease negative was purchased from New England Biolabs (Catalog number M0212S).

2.2.6.3 Fill in and sequencing of an oligonucleotide maximising nucleotide discrimination during sequencing

2.2.6.3.1 Oligonucleotide sequences and order details

The sequences for these oligonucleotides are shown in Table 2.25. The ribonucleotide is denoted with an r prior to the base, in this case rC. The DRD60 oligonucleotide has a single embedded ribonucleotide, and the D60C with an identical sequence serves as a DNA control. Barcode 1 (D120BC1) is the sequence annealed to the RNA containing oliogonucleotide, and barcode 2 (D120BC2) is the sequence annealed to the equivalent DNA control (Chapter 3, Figure 3.6). As we found in initial experiments that the quality of the oligonucleotides was low, the DRD60 oligonucleotide was ordered with RNase Free HPLC (high performance liquid chromatography) purification, and the DNA oligonucleotides with PAGE (poly-acrylamide gel electrophoresis) purification, both from Integrated DNA Technologies (IDT).

Table 2.25. Primers used for proof of concept experiment for detection of ribonucleotides.

Oligonucleotide	Oligonucleotide sequence (5' → 3')
DRD60	/5Phos/TATAGCGTCTACGAGTGAGTAGATCGTrCTGCTAGTGCGATACTATCGTCATCGATCTATG
D60C	/5Phos/TATAGCGTCTACGAGTGAGTAGATCGTCTGCTAGTGCGATACTATCGTCATCGATCTATG
D120BC1	/5Phos/ATGTCTGACTGTCAGTCTCAGATATCACTACATCTTAGTGCACATCACGCAGCATGTAGTCGTGATGCTCATCGTCACGCGCAGTGACAGCTACGCATAGATCGATGACGATAGTATCGC
D120BC2	/5Phos/ATGTCTGACTGTCAGTCTCAGATATCACTACATCTTAGTGCACATCACGCAGCACACGACTACTATGCTCATCGTCACGCGCAGTGACAGCTACGCATAGATCGATGACGATAGTATCGC

2.2.6.3.2 Synthesis of filled in duplexes for nanopore sequencing

The oligonucleotides listed in Table 2.25 were diluted using dH_2O to a concentration of 100 µM. Following this, 1 µl of DRD60 and 2 µl D120BC1, and 1µl D60C and 2µl DRD120BC2, respectively, were annealed in a reaction containing 5 µl of Klenow buffer without magnesium, with 42 dH_2O added to make a total of 50 µl. A 1 ml stock of this Klenow buffer was created by adding 10 µL of DTT (1M) to 100 uL 5 M NaCl, 100 µL 1 M Tris pH8 and 790 uL H2O. The DNA oligonucleotides were heated to 95 °C for 2 minutes, and then allowed to cool slowly to room temperature. Following this, 43 µL of each oligo

duplex reaction was mixed with 5 µL of 330 µM dNTPs, 1 µL of Klenow 3'-5' exo -, and 1 µL Mg^{2+} 500 mM. This reaction was incubated at 37 °C for 2 hours, purified using Ampure beads, washed twice with 200 µL ethanol, and eluted in 42 µL. Following this the duplexes were A-tailed in a 50 µl reaction with 42 µL of duplex, 3 µL Klenow 3'-5' exo -,and 5 µL of dA from the NEBNext® dA-Tailing Module. A further Ampure bead purification step was conducted with elution in 33 µl of H20. 30 uL of this product was taken forward into a Oxford Nanopore 1D genomic DNA by ligation (SQK-LSK108) sequencing reaction (GitLab: "nanopore/nanopore_protocols/nanopore_ONT_1D_experimental_protocol.pdf").

2.2.6.4 Ligation and sequencing of oligonucleotides to test 64 possible trinucleotide combinations

The sequences of the 64 different oligonucleotides containing a central ribonucleotide are available on GitLab ("/nanopore/oligo_design/64_trinucleotides_with_central_ribo.tsv").They were ordered from IDT as 100 nmol scale (standard desalted; ship dry; full yield). The sequences for the upstream and downstream duplexes used to bind to the ribonucleotide containing oligo are given in Table 2.26; these were ordered as 100 nmol scale, PAGE purified. The conceptual set up for ligation and sequencing is outlined in Chapter 3, Figure 3.10.

Table 2.26. Oligonucleotides used to form upstream and downstream duplexes to bind to the ribonucleotide containing oligonucleotide.

Oligonucleotide	Oligonucleotide sequence (5' → 3')
five_prime_for	/5Phos/ATAGCGTCTACGAGTGAGTAGATCGTCTGCTAGTGCGATACTATCGTCAT
five_prime_rev	AGATCGATGACGATAGTATCGCACTAGCAGACGATCTACTCACTCGTAGACGCTATA
three_prime_for	/5Phos/TCACGACTACATGCTGCGTGATGTGCACTAAGATGTAGTGATATCTGAGA/3SpC3/
three_prim_rev	TCTCAGATATCACTACATCTTAGTGCACATCACGCAGCATGTAGTCGTGATGTCAT/3SpC3/

All 64 ribonucleotide containing oligonucleotides were reconstituted at a concentration of 100 µM using TE, and 10 µl of each added to create a mastermix. The 5' (five_prime_for/five_prime_rev) and 3'

(three_prime_for/three_prime_rev) duplexes were annealed by heading 10 µl of each of the primer pairs to 95 °C and allowing them to cool slowly to room temperature. Subsequently 10 µl of each duplex was added to 5 µL of the oligonucleotide mastermix, together with 25 µl of Blunt/T4 Ligase Master Mix to create a 50 ul ligation reaction, which was incubated for 1 hour at room temperature. This reaction was column purified (Section 2.2.4.3.4) and eluted in 30 µl dH_2O. Following this the eluent was run on a 1.5% agarose gel with SYBR-Safe, gel purified (Section 2.2.4.3.4), and eluted into 25 µl of dH_2O. This was then taken forward into a 1D^2 sequencing reaction (Section 2.2.6.1).

2.2.6.5 Acylation of embedded ribonucleotides

The duplex used was a modification of C12 (Section 2.3.2.2.2) 110 nucleotides in length, with a ribonucleotide (rA) at position 83 (Table 2.27).

Table 2.27. Oligonucleotide used in 2'-hydroxyl acylation experiments

Oligonucleotide	Oligonucleotide sequence (5' → 3')
C12_duplex_forward	/5Phos/ATAGCGTCTACGAGTGAGTAGATCGTCTGCTAGTGCGATACTA TCGTCATCGATCTAGTCAGTCATGCGTAGCTGTCTGACGrATGATGAC TGCGCGTGACGATGATGACA

NAI was purchased from Sigma-Aldrich (product code 129887) and NMIA from Merck (product code 03-310). Sixty-nine mg of NMIA was diluted in 3 ml of DMSO to create a 130 mM solution; NAI arrived prepared at a concentration of 2M, and was diluted 2:1 in DMSO to prepare a stock concentration of 1M.

For acylation, 1 µl of a 10 µM solution of the ribonucleotide containing duplexes were added to a 9 µl reaction containing TE and NAI/NMIA to a final concentration of 100 mM and 130 mM respectively. Following incubation for variable time periods (15, 30 and 60 minutes), they were treated with NaOH at a final concentration of 1M at 95° C for 2 mins to perform alkali hydrolysis as per the original SHAPE protocol (Wilkinson, Merino and Weeks, 2006). Following this, control oligonucleotides containing a ribonucleotides, both untreated, and treated with NaOH or a NaCl control, were run with the adducted oligonucleotide using denaturing PAGE (Section 2.2.4.2.3).

2.2.6.6 Sequencing of oligonucleotides connected by hairpin loop

Initially, a long and short hairpin were designed, as there was uncertainty about whether the length of the hairpin would affect the accuracy of sequencing for the complementary strand. The design of the hairpins was based on sequences previously published for high accuracy duplex sequencing (long hairpin) (Taylor, Cinquin and Cinquin, 2016) and bisulfite sequencing (short hairpin) (Laird *et al.*, 2004). The long hairpin sequence was visualised in Unafold (Markham, Zuker and Keith, 2008) and GC bonds that interfered with hairpin formation modified to create a 12 nucleotide long complementary stem for the hairpin loop, with a 6 nucleotide overhang to bind to the ribonucleotide containing duplex. The sequences for both hairpins are given in Table 2.28.

For this series of experiments one of the 64 oligonucleotides with a single embedded ribonucleotide (B2 from Section 2.3.2.2.2) was modified. This oligonucleotide ("B2_ext_sh_C", Table 2.28) incorporated the upstream sequence from previous experiments to create a 110 nucleotide long template oligonucleotide, with the ribonucleotide now at position 32. A complementary sequence to this template (B2_ext_sh_C_rev) contained an adenine overhang at the 3' end of the complement (to allow binding to the ONT adapter) and a 6 nucleotide long overhang on the 3' end of the template to allow binding of the hairpin (Table 2.28).

To anneal the duplex to which the hairpin would bind, 10 µl of 100 µM B2_ext_sh_C and 10 µl of 100 µM B2_ext_sh_C_rev were mixed, heated to 95°C for 2 minutes, then allow to cool slowly. To form the hairpins 10 µl of 100 µM solutions of each were heated to 95°C on a hot plate, held at 95 °C for 5 minutes, then rapidly cooled to 4 °C by placing the samples on ice as described by Oxford Nanopore Technologies in the patent relating to their initial 2D technology (ONT, 2011).

Following annealing, 5 µl 100 µM B2_ext_sh_C/ B2_ext_sh_C_rev annealing product was added 2.5 µl 100 µM of annealed hairpin (short or long), 2.5 µl 1D^2 barcode adaptor, and 10 µl Blunt/TA ligase MM. Both reactions were incubated for 30 minutes at RT. The reactions were column purified (Section 2.2.4.3.4), eluted in 30 µl H20, and taken into a standard 1D^2 ONT sequencing reaction (Section 2.2.6.1).

Table 2.28. Oligonucleotides used for hairpin experiments.

Oligonucleotide	Oligonucleotide sequence (5' → 3')
B2_ext_sh_C	/5Phos/GATCTGTCAGTCGATGCGTAGCTGTCTGACArUCGATGACTGC GCGTGACATCGATGACATCACGACTACATGCTGCGTGATGTGCACTAA GATGTAGTGATATCTGAGC
B2_ext_sh_C_DNA	/5Phos/GATCTGTCAGTCGATGCGTAGCTGTCTGACATCGATGACTGCG CGTGACATCGATGACATCACGACTACATGCTGCGTGATGTGCACTAAG ATGTAGTGATATCTGAGC
B2_ext_sh_C_rev	/5Phos/ATATCACTACATCTTAGTGCACATCACGCAGCATGTAGTCGTG ATGTCATCGATGTCACGCGCAGTCATCGATGTCAGACAGCTACGCATC GACTGACAGATCA
3'_rev_49	/5Phos/ATATCACTACATCTTAGTGCACATCACGCAGCATGTAGTCGTG ATGTCA
long_hairpin	/5Phos/AGATCGGAAGAGCACAACTCTGAACTCCAGTCTACACTCTTTC CCTACACGACGCTCTTCCGATCTGCTCAG
short_hairpin	/5Phos/AGATCGGAAGAGCCACGACGCTCTTCCGATCTGCTCAG

2.2.6.7 Synthesising a sequence complementary to the ribonucleotide containing template with a ligated hairpin

Initially a long hairpin was ligated to the template oligonucleotide (B2_ext_sh_C, Table 2.28), and the 3' end of the long hairpin (composed from hairpin stem and the 6 nt overhang) functioned as a primer from which to synthesise the complementary strand. However, this was unsuccessful. Subsequently, a 49 nucleotide long complementary primer was used to create a double stranded oligonucleotide, which bound to the template oligonucleotide, and also to the 3' end of the hairpin once the 6 nucleotide long overhang had bound to the template sequence (Figure 3.18).

Initial trials with a New England Biolabs repair/extension mix (NEBNext® Ultra™ II End Repair/dA-Tailing Module, catalog E7546S) showed that it contained 5'->3' exonuclease activity that removed the overhang which bound the hairpin. Therefore all further synthesis reactions were conducted with Klenow 3' –> 5' exonuclease negative (Klenow exo-) enzyme.

A long hairpin loop was annealed by heating to 95 °C for 5 minutes, then placed on ice. A duplex for extension was created by mixing 5 µl of 100µM B2_ext_sh_C and 5 µl 100 µM 3'_rev_49 (Table 2.28), heating to 95 °C for 2 minutes, then allowing to cool slowly. In an extension step, 5 µl 100 uM B2_ext_sh_C/3'_rev_49 annealing product, 2 µl H20, 1 µl 10 mM dNTPs, 1 µl 10x NEB2 buffer, 1 µl Klenow 3'-5' exo- were added together, and heated at 37°C for 1 hour. The reaction was column purified (Section 2.2.4.3.4) and eluted in 30 µl H20. 24 µl of this reaction product was mixed with 1 µl long_hairpin 100 uM and 25 µl Blunt/TA ligase MM, and taken into a standard $1D^2$ ONT sequencing reaction (Section 2.2.6.1).

2.2.6.8 Testing efficiency of hairpin formation

In order to confirm hairpin self-complementarity, 4 further oligonucleotides were adapted from the ones used to investigate efficiency of synthesis across a ribonucleotide (Chapter 3, Section 3.2.3; Table 2.29).

Table 2.29. Sequences used in testing degree of hairpin self-complementarity.

Oligonucleotide	Oligonucleotide sequence (5' → 3')
D109a	/5Phos/GATCTGTCAGTCGATGCGTAGCTGTCTGACTGATATGACTGCG CGTGACTGACATGACATCACGACTACATGCTGCGTGATGTGCACTAAG ATGTAGTGATATCTGAGC
D109b	/5Phos/GATCTGTCAGTCGATGCGTAGCTGTCTGATCTCGATGACTGCG CGTGAGCGATATGACATCACGACTACATGCTGCGTGATGTGCACTAA GATGTAGTGATATCTGAGC
D109c	/5Phos/GATCTGTCAGTCGATGCGTAGCTGTCTGAGTCGTATGACTGCG CGTGAGTCAGATGACATCACGACTACATGCTGCGTGATGTGCACTAA GATGTAGTGATATCTGAGC
D109d	/5Phos/GATCTGTCAGTCGATGCGTAGCTGTCTGAGTATGATGACTGCG CGTGATCGATATGACATCACGACTACATGCTGCGTGATGTGCACTAAG ATGTAGTGATATCTGAGC

A complement for these was synthesised using 3'_49_rev as a primer (Table 2.28), hairpins were ligated, and the products sequenced using $1D^2$ ONT workflow after library prep carried out as outlined below. Initially 2.5 µl of 100 µM 109a, 109b, 109c, 109d were added to 10 µl of 100 µM 3'_rev_49. The solution was heated to 95 °C for 2 minutes, then allowed to cool slowly.To fill in the duplexes, 5 µl 109/3'49 annealing product was mixed with with 2 µl H_20, 1 µl 10 mM dNTPs,1 µl NEB2 buffer, and 1 µl Klenow 3'-5' exo-, and incubated

at 37 °C for 1 hour. The reaction was column purified (Section 2.2.4.3.4) and eluted in 30 µl H20. 24 µl of this reaction product was mixed with 1 µl long_hairpin 100 µM and 25 µl Blunt/TA ligase MM and taken into a standard $1D^2$ ONT sequencing reaction (Section 2.2.6.1).

2.3 Computational Methods

2.3.1 Reference genomes

2.3.1.1 S. cerevisiae data

For analyses of emRiboSeq data, analyses were performed using the sacCer3 (V64) S. cerevisiae reference genome, downloaded from the Saccharomyces Genome Database (www.yeastgenome.org). For analyses of HydEnSeq data and whole genome sequencing in yeast strains, the reference genome used by Lujan and colleagues, GSE56939_L03_ref_v2 (Lujan *et al.*, 2014), was downloaded from the NCBI Gene Expression Omnibus portal (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56939). For direct comparisons between the two, Liftover (v261) (Kuhn, Haussler and Kent, 2013) was used to create chain files to project between datasets mapped to each of the reference genomes.

2.3.1.2 Human data

Analyses were performed using either the hg19 (GRCh37) or GRCh38 human reference genome, downloaded from the UCSC Genome Browser (https://hgdownload.soe.ucsc.edu/downloads.html).

2.3.2 Computational analysis

Analyses related to *S. cerevisiae* emRiboSeq (Reijns *et al.*, 2015) and HydEnSeq (Clausen *et al.*, 2015) data were conducted in RStudio version 1.0.44 (RStudio, 2015). For analyses using nanopolish (see below), Python 2.7.10 was used, for all other analyses Python 3.7.4 was used.

2.3.2.1 Analysis of emRiboSeq and HydEnSeq data

For emRiboSeq data, reads were available as processed in (Reijns *et al.*, 2015). They were directly compared to HydEnseq data available from (Clausen *et al.*, 2015).

2.3.2.1.1 Origins of replication analysis

The background and development of the code for this analysis is laid out in Chapter 3, Section 3.2.4. The code for these analyses is available in the GitLab directory "/nanopore/origin_replication_analysis/"

A script named "/defining_midpoints_origins_replication.py" was used to define the origins of replication in the analysis. Separate scripts were written to perform trinucleotide normalisation outwith of the origins of replication for emRiboSeq ("EmRiboSeq_trinucleotide_context_whole_genome_excluding_oris.py") and "HydEnSeq_trinucleotide_context_whole_genome_excluding_oris.py").

The R code used to generate the plot for emRiboSeq at origins of replication is called "plotting_emRiboSeq_signal_across_origins_of_replication.R" with an equivalent for plotting HydEnSeq origins also available in the same directory ("plotting_HydEnSeq_signal_across_origins_of_replication.R").

Code for analysis looking at sum of signal at origins of replication is again available separately for emRiboSeq: ("plotting_total_emRiboSeq_signal_across_origins_of_replication.R")

and HydEnSeq data: (origin_replication_analysis/plotting_total_HydEnSeq_signal_across_origins_of_replication.R).

2.3.2.1.2 Okazaki junction analysis

The background and development of the code for this analysis is laid out in Chapter 3, Section 3.2.1.2. The code for this analysis (presented in Figure 3.4) is available at: "nanopore/okazaki_junction_analysis/identifying_okazaki_junctions_pol_alpha_emRiboSeq_HydEnSeq.R".

2.3.2.2 Design of oligonucleotides for Nanopore experiments

2.3.2.2.1 Design of an oligonucleotide maximising nucleotide discrimination during sequencing

Oxford Nanopore models:
(GitLab:"/nanopore/oligo_design/DNA_5mer_models.txt") were used to create a list of possible oligonucleotides as described in Chapter 3, Section 3.2.4:

(GitLab: "oligo_design/design_proof_concept_nanopore_part1.py"). This script generated 10 000 possible oligonucleotides of a length of 155 bases. Using the Mfold (Zuker, 2003) and Unafold packages (Markham, Zuker and Keith, 2008), a score was assigned for the chance of self-dimerisation and unwanted primer-primer dimers to all of these 10,000 options, these scores were added for each option, and ranked (GitLab: "/nanopore/oligo_design/design_proof_concept_nanopore_part2.py"). The 10 oligonucleotides with the lowest chance of dimerization were then manually inspected, including the barcode, and what was considered to be the pair of oligos with the lowest chance of dimerization selected.

2.3.2.2.2 Design of oligonucleotides to test 64 possible trinucleotide combinations

The script used to design barcodes for use on the 64 ribonucleotide containing oligonucleotides is available on GitLab:
("/nanopore/oligo_design/designing_barcodes.py"). The kmeans function in R was implemented to choose the 64 most discriminatory barcodes from the 515 options outputted by this script (GitLab: "/nanopore/oligo_design/kmeans_to_select_barcodes.R"); the concept underlying this approach is illustrated in Figure 3.11, Chapter 3. The final sequences used are available on GitLab:
("/nanopore/oligo_design/64_trinucleotides_with_central_ribo.tsv").

2.3.2.3 Nanopore computational analyses

2.3.2.3.1 General note on sequencing analysis

For all experiments, Albacore version 2.0.1 (Oxford Nanopore Technologies) was used to basecall reads. BWA-MEM (Li, 2013) was used to align the

sequence reads to a reference sequence, and to create a sorted bam file. Subsequent to this nanopolish v 0.7.1 (Loman, Quick and Simpson, 2015) was used to create eventalign files for the reads for the RNA containing sequences and the equivalent DNA controls. The information for amplitude measurements was compressed (using the mean of events) for each *k*-mer using Python, and signal compared in aggregate for the RNA containing sequence versus the equivalent DNA control, over the region containing the ribonucleotide. Following this, output files were imported into R for further analysis. Scripts are available in the "/nanopore/" directory on GitLab.

2.3.2.3.2 Analysis of output from proof of concept experiment

After base calling, reads were mapped to the two possible reference sequences, labelled as "rna.fasta" (DRD60/D120BC1) and as "dna.fasta" (D60C/D120BC2) (GitLab: "oligo_analysis/dna.fasta" and "oligo_analysis/rna.fasta"). Mean read amplitudes, and time taken to pass through the pore, were calculated for each *k*-mer in R (Figure 3.7). For *k*-mer 25 (with ribonucleotide in central position in the 5mer) the "density" function in R was used to visualise the distribution of all the amplitude readings for the RNA containing sequence compared to the DNA control (Figure 3.8), and a Kolmogorov-Smirnov test used to compare the two distributions (GitLab: "/oligo_analysis/analysis_initial_ribo_containing_oligo.R", Figure 3.9).

2.3.2.3.3 Analysis of sequencing of 64 possible trinucleotide combinations with a central ribonucleotide

Sequence outputs were mapped to the 64 possible different reference sequences (Gitlab: "/oligo_analysis/reference_fasta_test10.fasta"). After creation of eventalign files, the amplitude and kinetic signal for the *k*-mers for each of the 64 oligonucleotides were condensed as detailed in Section 2.3.2.2.2 (Figures 3.12, 3.13). A further analysis was conducted plotting the mean delta in amplitude versus kinetic signal for the 64 trinucleotide combinations (Figure 3.14, Gitlab: "oligo_analysis/ analysis_64_trinucleotide_combinations.R")

2.3.2.3.4 Analysis of experiment to test efficiency of hairpin formation

Sequence outputs were mapped to the reference fasta for the ribonucleotide containing oligo, the long and short hairpins, and the complement of the ribonucleotide containing oligo (GitLab: "/oligo_analysis/hairpin1_ref.fas"). After creation of the eventalign file, *k*-mers for each of the successful reads were condensed as detailed above, imported into R, analysed and plotted (GitLab: "/oligo_analysis/analysis_long_vs_short_hairpins.R").

2.3.2.3.5 Analysis of experiments to test synthesis over ribonucleotide after hairpin ligation

The analyses were the same as detailed in Section 2.3.2.2.4, except reads were mapped to a single reference fasta (GitLab: "oligo_analysis/extended_hairpin_ref.fas"). The mapped reads were imported into and plotted in R (GitLab: "oligo_analysis/ analysis_extension_over_ribo.R").

2.3.2.3.6 Analysis of testing self-complementarity of hairpins

As there were 16 (4x4) possible ligation products (4 references and 4 complements), a reference fasta with these 16 possibilities was used to map sequencing products. (GitLab: "109_16_references_20190623.fasta"). The proportion of total reads that mapped to each reference sequence were then compared, with the expectation that if there was 100% hairpin efficiency, 100% of the sequences would map to a sequence consisting of reference1 + long_hairpin + complement of reference1. After aggregation of data using python (GitLab: "querying_eventalign_109_oligos.py") summary data was uploaded to and visualised in R (GitLab: "oligo_analysis/analysing_109_oligos.R").

2.3.3 Next Generation Whole Genome Sequencing Data

2.3.3.1 Alignment of raw sequencing reads

Bwa mem (Li, 2013) was used in all analyses to align fastq reads to a reference genome and create bam files. For analysis of *S. cerevisiae* whole genome

sequencing data, Samblaster (Faust and Hall, 2014) was used to deduplicate bam files.

2.3.3.2 Calling variants

2.3.3.2.1 S. cerevisiae whole genome sequencing data

GATK (without Base Quality Score Recalibration- BQSR) (Van der Auwera *et al.*, 2013) was used to create vcf files from the bam files. The GATK Haplotype Caller was used to directly call variants for each strain, and then “Hard Filters” for SNPs (--filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0") and indels (--filterExpression "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0”) to classify variants as high quality or not.

To identify variants unique to each strain in the mutation accumulation experiments conducted by the Kunkel lab (Lujan *et al.*, 2014; Conover *et al.*, 2015), the filters used in their publications to call mutations were implemented, with some modifications, writing a Python script to do so (GitLab: “COSMIC/generic_template_for_processing_Lujan_data.py”). The filters used were:

1) Eliminating any mutations shared with either ancestral clonal isolate, with < 1% of reads supporting the variant allele in the ancestor
2) At least 20 reads supporting the variant allele
3) Excluding repetitive regions. These are a list of coordinates representing 4.7% of the yeast genome, which I converted into a bed file from the supplementary material in the Lujan paper.
4) Setting a ratio of between 0.4 and 0.6 between the reference and variant allele (to which I added a condition of < 0.4 if the variant allele had become homozygous).

2.3.3.2.2 RPE1 whole genome sequencing data

VCF files were provided by Edinburgh Genomics. Mutations were called by Dave Parry (Jackson Lab) in a customised Perl script (GitLab: “/COSMIC/vase_per_sample.pl”) filtering on parameters including read quality

and depth, variant allele frequency, and the absence of reads supporting the variant in the ancestor to identify high confidence unique mutations for each cell line. The filters applied were:

1) Eliminating any mutations shared with either ancestral clonal isolate, with < 5 % of reads supporting the variant allele in the ancestor
2) At least 20 reads supporting the variant allele
3) Excluding repetitive regions: this step was not used, as the VCFs were provided by Edinburgh Genomics
4) Setting a ratio of between 0.25 and 0.75 between the reference and variant allele

2.3.4 Image analysis of fluctuation assay outputs

2.3.4.1 Image acquisition

2.3.4.1.1 S. cerevisiae

Images of *S. cerevisiae* plates from the fluctuation assays were captured using a BioDoc-It System, UVP.

2.3.4.1.2 HeLa cells

Plates were photographed in batches of 8 plates with the help of Craig Nicol, IGMM Design and Publication Studio, on a D800 Nikon 36 megapixel SLR with AF-S Micro NIKKOR 60mm f/2.8 lens, using jpg fine format.

2.3.4.2 Image analysis

Previous studies have relied on manual counting of colonies to calculate *r* for each individual culture. However, given the large numbers of plates being generated for each fluctuation assay (64 individual cultures, 4 plates per culture, 256 plates with *S.cerevisiae* colonies) I, with help from Laura Murphy (IGMM Imaging Services), created an automated approach on detecting colonies using ImageJ Software (US National Institutes of Health) (Abràmoff, Magalhães and Ram, 2004; Schneider, Rasband and Eliceiri, 2012). These scripts automatically detected colonies above a predefined size and percentage circularity, and incorporated a manual checking step to ensure that all colonies had been captured. Representative scripts have been uploaded to

the GitLab (“imaging_analyses/s_cerevisiae_imaging_analysis.ijm” and “imaging_analyses/ HeLa_imaging_analysis.ijm”).

2.3.5 Analysis of Sanger Sequencing Data for Mutation Detection Construct

All sequencing traces were annotated using a standardised format (<strain number><yyyymmdd><number><amplicon><direction_primer>), so that the forward trace for the upstream amplicon of the first TCW16 strain sequenced on 1st Jan 2020 would be labelled “16_20200101_1_U_F”.

Using these standardised formats, they were uploaded to Mutation Surveyor® (SoftGenetics) and analysed again a reference genome (GitLab: “mutation_detection_construct/sequencing_analysis/saccharomyces_2bpdel_detection_construct_dinuc_rep_only_no_bait_analysis_version_20200109.seq”) starting 20 bp upstream from the start codon for the Hygromycin resistance gene. This reference was started upstream because point mutations upstream of the start codon ATG could lead to alternative start sites, and therefore a functional puromycin resistance gene downstream.

A python script then categorised mutations as 2 bp deletions in perfect, imperfect or not in tandem repeats, >2 bp deletions, 1 bp insertions, >1 bp insertions, or SNPs. The scripts for the *S. cerevisiae* and HeLa samples are available on GitLab (“mutation_detection_construct/sequencing_analysis/python_analysis_saccharomyces_sequencing.py” and “mutation_detection_construct/sequencing_analysis/python_analysis_HeLa_sequencing.py”). If two reads supported a mutation then this was accepted. However, if there were not two reads supporting a mutation, the reads were manually checked using Sequencher 5.4.6 (Gene Codes Corp).

2.3.5.1 Cosine similarities analysis

In order to compare mutational spectra in different strains, cosine similarities were used. Mutation rates for each strain were converted into a vector, with the ordered values in each vector representing the proportion of total mutations that were of each type (2 base pair deletions in perfect repeat, 2 base pair

deletions not in perfect repeat, and so forth). To compare 2 vectors the following formula was used:

```
v1<- vector1 #strain 1

v2<-vector2  #strain 2

cosSim<-sum(v1*v2)/sqrt(sum(v1**2)*sum(v2**2))
```

I then wrote a script in R to make this calculation for all strains compared in a pairwise fashion, and used the gplots function in R (Galili, 2020) to create a visual representation of these comparisons (GitLab: "/mutational_spectra_analyses/cosine_similarities.R").

Chapter 3 Quantitative detection of DNA-embedded ribonucleotides at single nucleotide resolution

3.1 Introduction

3.1.1 Why study RNA incorporated into DNA?

As described in the introduction, RNA is frequently misincorporated into DNA. An estimated 13,000 rNTPs are incorporated into genomic DNA in each round of cell division in *S. cerevisiae* (Williams *et al.*, 2013), and over one million in mammalian cells (Reijns *et al.*, 2012). In some contexts these ribonucleotides appear to serve a physiological function. In *S. cerevisiae* they lead to the creation of RNase H2 mediated nicks in the backbone of double stranded DNA to allow entry of mismatch repair machinery (Lujan *et al.*, 2013); in mammalian cells they promote non-homologous end joining (Pryor *et al.*, 2018); and in *Schizosaccharomyces pombe* a retained di-ribonucleotide leads to replication fork stalling and the mating type switch (Yamada-Inagawa, Klar and Dalgaard, 2007). Ribonucleotide incorporation can also be used to study fundamental cellular processes such as replication, for example by defining origins of replication (Clausen *et al.*, 2015; Koh *et al.*, 2015; Reijns *et al.*, 2015) and establishing the contributions of different polymerases to leading and lagging strand replication (Reijns *et al.*, 2015). An excess of incorporated ribonucleotides leads to genomic instability in *S. cerevisiae* (Conover *et al.*, 2015) and mammalian cells (Reijns *et al.*, 2012; Zimmermann *et al.*, 2018), and in *S. cerevisiae* to a characteristic mutational signature of short (2-5 bp) deletions, predominantly in repeat sequences (N. Kim *et al.*, 2011; Williams *et al.*, 2019).

These ribonucleotides have been studied using a variety of methods, including alkali gels and next generation sequencing technologies. However, further investigation in their physiological and pathophysiological roles would be facilitated by techniques that allow quantitative detection of ribonucleotides at single base resolution; that allow identification of a tract of RNA as well as

single misincorporated ribonucleotides; and that permit an understanding of phasing of ribonucleotides along a single strand of DNA.

Below I outline the principles of DNA sequencing as described by Frederick Sanger and colleagues, and how these methods, together with the use of alkali gels, have subsequently been applied to the detection of ribonucleotides and other non-canonical nucleotides. Following on from this I describe my analyses of existing emRiboSeq (Ding *et al.*, 2015; Reijns *et al.*, 2015) and HydEnSeq (Clausen *et al.*, 2015) data in *S.cerevisiae*, and discuss the limitations of these next generation sequencing based technologies in the investigation of RNA misincorporated into DNA. I outline my investigation of the use of nanopore sequencing to directly detect RNA incorporated into DNA, demonstrating that ribonucleotides are associated with a distinctive amplitude and kinetic signature. I conclude by outlining the steps that could be taken to use nanopore sequencing to detect embedded ribonucleotides in any sequence context, and datasets to which this sequencing could be applied.

3.1.2 Sanger sequencing, Next Generation Sequencing, and use of these technologies to detect non-canonical nucleotides

What is now referred to as Sanger sequencing was developed in the 1970s by Frederick Sanger, Steve Nicklen and Alan Coulson (Sanger and Coulson, 1975; Sanger, Nicklen and Coulson, 1977). The original approach used an *E.coli* DNA polymerase I to extend a primer along the DNA sequence to be determined, in 4 separate lanes, each containing a mixture of all four normal deoxynucleotidetriphosphates (dNTPs) and a single modified di-deoxynucleotidetriphosphates (ddNTPs), in a ratio of ~100:1. In the initial description of the method, lane 1 contained ddGTP, lane 2 ddATP, lane 3 ddTTP, and lane 4 ddCTP. The ddNTPs lack a 3' hydroxyl group, and thus lead to termination of the extension reaction when they are incorporated. As complementary DNA is synthesised from the template, this will stochastically terminate (based on the ratio of dNTPs to ddNTPs), leading to a series of DNA sequences that terminate at the respective G, A T and C. To infer the DNA sequence being examined, the results from each reaction is fractionated by

electrophoresis on a denaturing acrylamide gel, and the sequence can be read off from the pattern of bands in the four lanes.

Subsequent modifications of the technique used fluorophore labelling to determine sequence: so-called dye-terminator sequencing (Smith *et al.*, 1986). Here, each of the chain terminating ddNTPs is labelled with a fluorescent dye which emit light at different wavelengths. The DNA to be sequenced is incubated with a polymerase and a mixture of normal dNTPs and labelled ddNTPs. Following this the reaction mixture is separated using capillary electrophoresis, and the template sequence inferred from the chromatograms produced as the DNA fragments pass through a laser.

The main limitation of these technologies is the limited length of DNA whose sequence can be determined, and the necessity of having relatively large quantities of DNA, normally the product of a single PCR reaction, in order to generate sufficient fluorophore activity to be detected by the sequencer. With the advent of whole genome shotgun sequencing, developed in part in order to accelerate progress of sequencing the entire human genome (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001), the basic principle of Sanger sequencing was retained, but the throughput significantly increased.

In contemporary approaches to next generation sequencing, the sequencing by synthesis approach is used (Bentley *et al.*, 2008). Here, after DNA is isolated, adaptors are ligated to sheared DNA strands. These adaptors hybridise primers tethered to a solid substrate (the flow-cell) and through multiple PCR cycles are amplified through “bridge amplification” to generate clonal clusters of amplified DNA fragments. Following this, a primer binding to the adaptor sequence is added to the chip, and a polymerase creates a copy of the clonal strands. Reversible 3′-O-azidomethyl 2′-deoxynucleoside triphosphates, each labelled with a different fluorophore, are added for each cycle of incorporation. As with the ddNTPs used in Sanger sequencing, these cause termination of extension once incorporated by the polymerase. After each cycle of incorporation, the identity of each base is determined by laser

induced excitation of the fluorophores and imaging, again similar to the method used for dye-terminator sequencing. Following this step, a reagent (in the original method, tris(2-carboxyethyl)phosphine) is added, which removes the fluorescent dye and side arm from a linker attached to the base, and also generates a 3' hydroxyl group, which allows progression to the next cycle. The use of a large number of parallel reactions to create copies of individual strands of DNA means that whole genomes, rather than individual genes, can be sequenced at once. However, the direct use of Next Generation Sequencing is limited to the detection of the canonical nucleotides, or strictly speaking nucleotides which will pair to a canonical nucleotide: it cannot directly detect the nucleotide strand being sequenced. The same is true of Sanger sequencing.

Although Sanger sequencing and NGS (based on the same underlying principles) can only detect canonical nucleotides, after its introduction a large number of techniques have since capitalised on the potential for this technology to *indirectly* detect strands of non-canonical nucleotides such as RNA transcripts or methylated cytosines. RNA-Seq (Nagalakshmi *et al.*, 2008) and CAGE (cap analysis gene expression) (Nagalakshmi *et al.*, 2008) use reverse transcription reactions to create a DNA copy of the original RNA transcript which can then be sequenced, and the composition of the transcripts inferred. Bisulfite sequencing utilises the conversion of unmethylated cytosines to uracil to infer the presence of methylated cytosines (Frommer *et al.*, 1992). Here, bisulfite is added to DNA, and unmethylated cytosines are converted to uracil by deamination of the C6 position to create 5,6-dihydrouracil-6-sulfonate. A subsequent alkylation step eliminates the sulfonate group and regenerates the double bond, creating uracil. In subsequent amplification reactions the uracil binds to a complementary adenine, which in turn binds to thymine; hence the original cytosine is converted to a thymine. Methylated cytosines are protected from deamination and PCR amplified strands retain the original guanine pairing. The relevance of these two conceptual approaches: the complementary pairing of RNA to equivalent DNA bases, and

the different biochemical properties of non-canonical nucleotides (for example 5-methylcytosine), will become clear in the sections below.

3.1.3 Previous methods used to detect and quantify genome embedded ribonucleotides

3.1.3.1 Alkaline gels

Alkaline gels have been used to infer the position and frequency of ribonucleotides incorporated into DNA, for example in experiments looking at the roles of different polymerases in synthesising the lagging and leading strands (Lujan *et al.*, 2013). In alkaline gels, the susceptibility of the 2' hydroxyl group of the ribose to cleavage by an alkali is exploited. Here, when DNA which includes incorporated ribonucleotides is treated with an alkali solution, the 2'hydroxyl can bind to the adjacent phosphodiester bond, releasing one end with a 2'-3' cyclic phosphate, and the other with a 5' hydroxyl (Li and Breaker, 1999). This process is an acceleration of the primary pathway for the uncatalyzed degradation of RNA polymers. The alkali accelerates this because the free hydroxide ions in the solution deprotonate the 2'OH of the ribose, making the reaction base-catalysed, and increasing the likelihood of the reaction occurring.

This hydrolysis reaction at the position of the RNA leads to a nick in the DNA duplex. When the DNA is run in a denaturing gel, the two strands separate, and the strand which contained the RNA and has been cleaved runs as a shorter product than an equivalent composed purely of DNA (Nick McElhinny, Kumar, *et al.*, 2010). These products can then be analysed in two ways. Firstly, to infer the position of the embedded ribonucleotide, the DNA can be transferred to a membrane and probed by Southern blot analysis to determine the location of the alkali sensitive site. Secondly, the alkali treated DNA can be compared in aggregate to a non-alkali treated control, and the mean length and distribution of the alkali treated DNA compared to the control used to infer the frequency of incorporated ribonucleotides. These experiments have been used to estimate of frequency of ribonucleotide misincorporation in *S. cerevisiae*, and in mammalian cells.

In order to estimate the frequency of ribonucleotide misincorporation into mammalian cells, researchers (Reijns *et al.*, 2012) performed targeted homologous recombination to generate embryonic stem (ES) cells with a premature stop codon in Exon 7 of the RNaseh2b subunit. They then injected these ES into mouse host blastocysts to generate germline chimeras and subsequently mice heterozygous for the *RNaseh2b* premature stop codon. They found that mice homozygous for the premature stop codon showed embryonic lethality, and demonstrated that this was due to cell cycle arrest secondary to p53 activity. Subsequent generation of *Rnaseh2b−/−;p53−/−* mouse embryonic fibroblasts (MEFs) allowed the isolation of total nucleic acids from these cells and controls, and alkali gel electrophoresis and densitometry of the traces suggested that the average size of RNase H2 null DNA fragments was between 3.7 and 11 kb, leading to the estimate of ~ 1 million ribonucleotides per mouse genome.

Although alkaline gel hydrolysis has proven useful in identifying hotspots of RNA misincorporation into *S. cerevisiae* genomes (Nick McElhinny, Kumar, *et al.*, 2010) and for quantifying the overall misincorporation of RNA into yeast and mammalian genomes (Reijns *et al.*, 2012), it has proved too cumbersome and imprecise a technique to detect ribonucleotide incorporation at a whole genome level. It is not a high throughput technique; cannot distinguish single from multiple ribonucleotides, and does not provide single base resolution or the identity of a ribonucleotide base. Finally, because alkali is used for hydrolysis, it does not distinguish between embedded ribonucleotides and abasic sites, which will also be cleaved.

3.1.3.2 Next Generation Sequencing based approaches to detect genome embedded ribonucleotides

As with other NGS technologies that leverage the principles of Sanger sequencing to infer the presence of non-canonical nucleotides, with time there arose the realisation that this could be used to identify single nucleotide resolution the presence of ribonucleotides embedded in DNA. In 2015 papers describing 4 techniques which leveraged NGS to identify ribonucleotides were published: emRiboSeq (Ding *et al.*, 2015), HydEnSeq (Clausen *et al.*, 2015),

Pu-Seq (Daigaku *et al.*, 2015) and Ribose-Seq (Koh *et al.*, 2015). These technologies make use of either enzymatic (RNase H2, emRiboSeq) or biochemical (an alkali such as sodium hydroxide, for the remaining 3 techniques) cleavage of either the 5' or 3' end of an embedded ribonucleotide, with subsequent ligation of adapters, degradation of remaining DNA, and high-throughput sequencing of adaptor ligated sequences. These sequences can then be aligned to the reference genomes for the DNA that has been sequenced (*S.cerevisiae* for emRiboSeq, HydenSeq, Ribose-Seq; *Schizosacharomyces pombe* for Pu-Seq) The start, or end, position of the sequence, depending on whether it was the 5' or 3' end that was cleaved, is then used to infer the position of the original ribonucleotide.

These technologies have been used for a variety of purposes- to confirm the contribution of different DNA polymerases (α, δ and ε) to the lagging and leading strands (Reijns *et al.*, 2015) , and to confirm the location of origins of replication (Clausen *et al.*, 2015). It has been used on mammalian DNA to examine mitochondria for evidence of ribonucleotide incorporation(Moss *et al.*, 2017), with studies in human DNA showing the presence of increased ribonucleotide incorporation at origin-L, where a RNA primer is thought to be used to commence the replication of mitochondrial DNA, but unusually retained after replication has been completed in order to serve as a primer for further rounds of mitochondrial division (Berglund *et al.*, 2017). However despite the insights that these new techniques have yielded into our understanding of replication biology (both nuclear and mitochondrial), there are a number of limitations to these NGS based approaches to the detection of genome embedded ribonucleotides, which are discussed in further detail below.

3.1.3.3 Limitations of NGS based approaches to detecting ribonucleotides

3.1.3.3.1 Susceptible to yet to be defined biases, not directly comparable

The first limitation is multi-step protocols are used to generate the free end of DNA used for sequencing. This means that although at a high level, for example for determining laggings vs leading strand preferences genome-wide

results have been concordant (Clausen *et al.*, 2015; Koh *et al.*, 2015; Reijns *et al.*, 2015), at single nucleotide resolution it remains unclear how comparable results are.

3.1.3.3.2 Not quantitative

Another limitation of the NGS approaches to ribonucleotide detection is that sequencing is only performed on the strands of DNA that have been ligated ,and thus that at some point have had a ribonucleotide adjacent to the start/end of the read. Therefore the methods do not give a quantitative read out of ribonucleotide incorporation. Instead, the reads can only be compared relative to one another: the number of reads mapping to a ribonucleotide at a particular location only give a relative increase proportional to the entirety of all the reads for the sequencing run.

3.1.3.3.3 Do not allow detection of runs of embedded ribonucleotides

It is known from previous work that runs of ribonucleotides are transiently incorporated into DNA, for example the Okazaki primers that are used by Pol α to start replication of the discontinuous DNA fragments that replicate the lagging strand (Okazaki *et al.*, 1968; Zheng and Shen, 2011; Balakrishnan and Bambara, 2013). Similarly, in *Schizosaccharomyces pombe*, a presumed di-ribonucleotide repeat at the mat1 locus on chromosome II (Vengrova and Dalgaard, 2006; Yamada-Inagawa, Klar and Dalgaard, 2007) has been identified as leading to stalling in Pol ε, a double stranded DNA break, and a subsequent recombination event between two cassettes with homologous up- and down- stream sequences, that moves information from the transcriptionally silenced donor locus (mat2P or mat3M) into the transcriptionally expressed mat1 locus. This in turn leads to a change in the mating type. NGS approaches would not be able to detect such a run of embedded ribonucleotides, as the cleavage of the ribonucleotide would mark either the start (for emRiboSeq) or end (for alkali cleavage based approaches) of a run of ribonucleotides.

3.1.3.3.4 Do not allow examination of phasing along a single strand

Another feature which would be of interest in the study of genome embedded ribonucleotides would be to determine whether there is phasing of ribonucleotides along a single strand of DNA; that is, a single strand of DNA which contains not one but several embedded ribonucleotides, or even several tracts of ribonucleotides, along its length. Because each read detects the presence of a single embedded ribonucleotide, even if there were a number of embedded ribonucleotides in proximity to each other, only one of these would be detected. Given the limitations to NGS based approaches to the detection of genome embedded ribonucleotides, the possibility of using nanopore sequencing to identify embedded ribonucleotides was explored.

3.1.4 Nanopore sequencing

3.1.4.1 Principles of nanopore sequencing

Nanopore sequencing is different to Sanger sequencing in that it uses an artificial membrane embedded with pores that allow the passage of a single strand of DNA. Strands are bound to a motor protein that feeds the strand through the pore and slows its passage so that accurate measurements can be made as it passes through (Loman and Watson, 2015; Jain *et al.*, 2016) (Figure 3.1). On either side of the artificial membrane are sensitive pH meters able to detect picoAmp (pA) level changes in current. As the DNA strand passes through this pore, this affects the flow of current through the pore, and this is recorded by the sequencer. As more than one nucleotide is passing through the pore at any one time, nucleotide calling is based on "models" (amplitude changes for a particular *k*-mer, or string of nucleotides of length *k*), conventionally for *k*-mers of length 5 or 6. The sequencer samples the current thousands of times per second for each pore (Simpson, 2015); these measurements are then processed by the event detection software, which attempts to detect points where the current level changes; this change in current indicates the presence of a new *k*-mer in the pore as the DNA strand passes through.

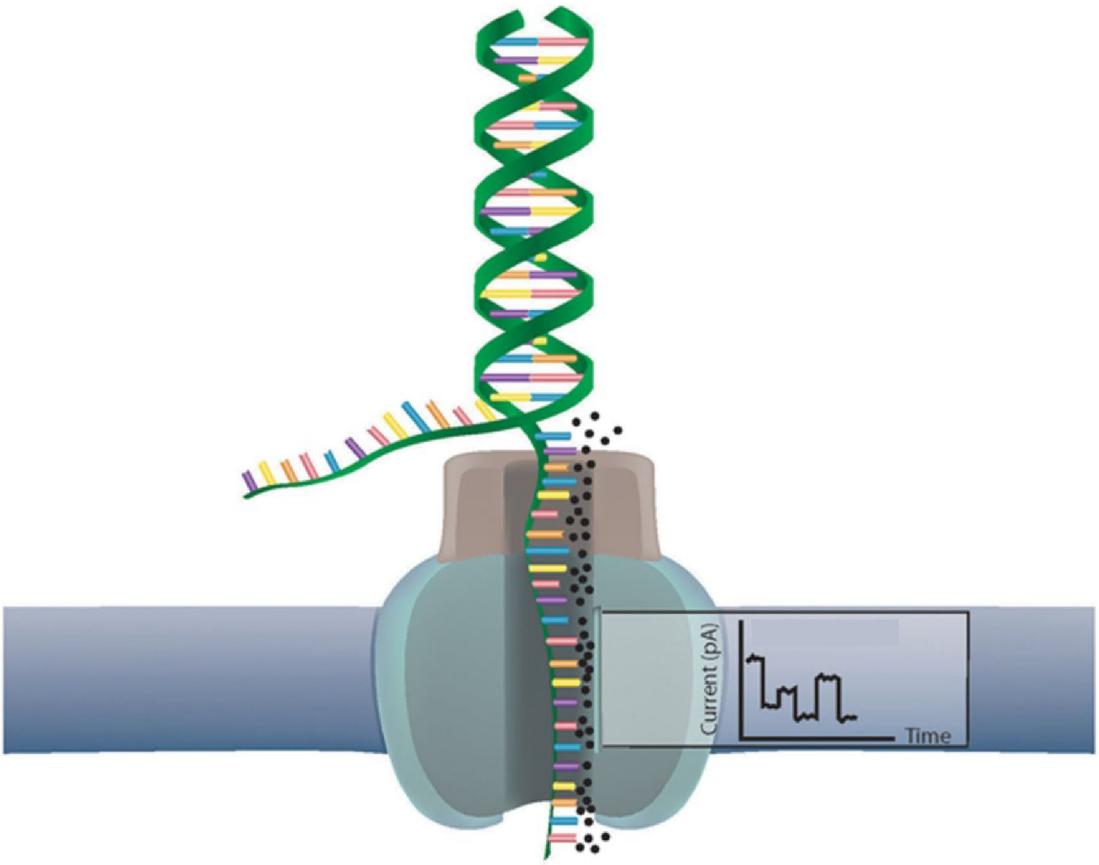


Figure 3.1. Principles of nanopore sequencing. *A single strand of DNA is unwound and passed through a pore which crosses a synthetic membrane. As the DNA strand passes through it affects the current passing through the pore and leads to an amplitude change detected by a sensitive pH meter. These amplitude changes ("events") are then interpreted by a base caller which assigns the events to a number of possible k-mers. Figure adapted with permission from* (Churko *et al.*, 2013).

One of the main limitations of nanopore sequencing is the inherent inaccuracy of the system- nucleotides are called on the basis of a large number of *k*-mer (1024 using a 5-mers) which occur in a relatively narrow amplitude window. These limitations have been addressed in two ways. Firstly, by adapting the software that is used to call the sequence that is being tested, which has been through a series of iterations in the years since the technology was introduced. Secondly, Oxford Nanopore Technologies (ONT) have introduced a system called $1D^2$, which matches the forward and reverse strands of a DNA sequence to construct a consensus sequence which is more accurate than sequencing either individually. The $1D^2$ system relies on the fact that as the end of the first strand of a DNA double strand passes through the pore, the strand that is most likely to pass through next is the complementary strand: strands passing

through a single pore can then be matched to one another, and if found to be complementary, a consensus sequence calculated.

3.1.4.2 Use of nanopore sequencing to detect non-canonical bases

3.1.4.2.1 Methylated cytosines

Nanopore sequencing has recently been used to not only sequence DNA, but to detect the presence of non-canonical nucleotides in a strand of DNA. The principle here is that a non-canonical nucleotide in a given *k*-mer will create a different amplitude change across the pore compared to the *k*-mer composed only of DNA nucleotides. The first application of this principle was to the detection of methylated cytosines (Rand *et al.*, 2017; Simpson *et al.*, 2017), where *E.coli* genomes with and without methylated cytosines were sequenced and the *k*-mer models with and without a methylated cytosine compared to identify amplitude signals associated with methylation. This study found that a 5-methylcytosine (5-mC) led to a shift in the electrical signal for many *k*-mers, with the shift in signal depending on the position of 5-mC within the *k*-mer.

3.1.4.2.2 Detection of 5-bromodeoxyuridine using nanopore sequencing

Subsequent to this, nanopore sequencing has been developed to be able to detect 5-bromodeoxyuridine (BrdU) embedded within DNA (Hennion *et al.*, 2018; Müller *et al.*, 2019). BrdU is a thymidine analogue, which, because of its limited cytotoxicity, is commonly used to study the cell cycle, as it substitutes for thymidine during replication and can be detected using an antibody. As with 5-mC, the authors of these two studies found that BrdU generated *k*-mers whose distribution of signal events was distinct to that of the DNA control.

3.1.4.3 Using nanopore sequencing to quantitatively detect ribonucleotides at single nucleotide resolution

The success of nanopore sequencing in detecting methylated cytosines (and our knowledge that attempts were being made to detect BrdU, now published) led to us to hypothesise that ribonucleotides embedded in DNA might behave differently to the equivalent deoxyribonucleotides (we knew that native RNA itself could also be directly sequenced) (Garalde *et al.*, 2018; Workman *et al.*, 2019). If ribonucleotides were to behave differently to the equivalent DNA,

nanopore sequencing would offer the potential to *quantitatively* detect single or runs of multiple ribonucleotides along a strand of DNA, to detect phasing of ribonucleotides along this strand, and allow their identification at single nucleotide resolution.

Below, I outline the analyses of existing data that prompted me to investigate nanopore technology as an option to provide further insight into genome embedded ribonucleotides. Accurate, quantitative determination of these genome embedded ribonucleotides would provide direct information on the location and possible function of these genome embedded ribonucleotides, and also help to inform our understanding of fundamental cellular processes such as location and timing of DNA replication, and Okazaki fragment incorporation and removal during this process.

3.2 Results

3.2.1.1 Comparability of emRiboSeq and HydEndSeq at S. cerevisiae origins of replication

As described earlier, the comparability of the different NGS approaches for detecting genome embedded ribonucleotides remains incompletely understood. As at a high level they are concordant in detecting origins of replication in *S. cerevisiae,* I decided to compare raw sequencing data over the same locations for 2 of these approaches (emRiboSeq and HydEnSeq) to identify whether there was a concentration or depletion of genome embedded ribonucleotides at these origins, and whether there was evidence for Pol δ activity on the leading (as opposed to lagging) strand at these origins, as subsequently proposed using HydEnseq data (Zhou *et al.*, 2019). In order to perform this analysis I took raw sequencing data from emRiboSeq runs published in 2015 (Reijns *et al.*, 2015) archived on the Taylor Lab data repository and also available from NCBI Gene Expression Omnibus (GEO) (GSE64521, 2015), and compared it to raw sequencing data for HydEnSeq, also available from NCBI GEO (GSE62181, 2015). I examined sequencing for a number of *S. cerevisiae* strains with different polymerase mutations: in pol α, Pol δ and Pol ε.

I normalised read counts for reads ending at each position in the *S. cerevisiae* genome to the total number of reads in a sequencing run. I then took the read coordinates and used Liftover (Kuhn, Haussler and Kent, 2013) to map both sets of reads to the SacCer3 reference genome. To identify the origins of replication in *S. cerevisiae*, I downloaded data from OriDB (Nieduszynski *et al.*, 2007), an online database of origins of replication in *S. cerevisiae*. These origins of replication were defined using a combination of techniques including chromatin-immunoprecipitation, microarrays, 2D gel electrophoresis, sites of BrdU incorporation, and accumulation of single stranded DNA when cells challenged with the antimetabolite hydroxyurea (Siow *et al.*, 2012). I used the subset of these annotated as “Confirmed”, and selected those that were > 10,000 nucleotides from each other. I mapped read counts for emRiboSeq sequencing runs to the mid-point of each origin of replication, and superimposed the read data for both sequencing methods, centred over the midpoint. In order to visualise changes in ribonucleotide incorporation over the origins of replication I plotted the ratio of signal for the forward over the forward + reverse strand. I also carried out a further normalisation step by dividing each position against mean signal for each of the 96 ribonucleotide trinucleotides across the genome (excluding the origins). This was to ensure that signal across the origins of replication was not being determined by trinucleotide context of the origins themselves affecting ribonucleotide incorporation, rather than misincorporation rates related to differential polymerase activity at the origins. Links to code for these calculations available in Materials and Methods (Section 2.3.2.1.1).

As the start of each emRiboSeq read in principle indicates RNase H2 cleavage of an embedded ribonucleotide, and sequencing from this point, a value of > 0.5 indicates a preponderance of ribonucleotides on the forward strand, and a value of < 0.5 a preponderance of ribonucleotides on the reverse strand. These initial plots confirmed published observations that ribonucleotide misincorporation by mutant Pol α, δ and ε serves as a marker of origins of replication. Pol ε mutator strains demonstrate higher ribonucleotide misincorporation signal on the forward strand downstream to origins of

replication (in the visualisation in Figure 3.2a), to the right of the origin), and α/δ mutator strands showing increased signal upstream of replication origins on the forward strand (see Figure 1.4). I also generated equivalent plots for the HyEnSeq data, which supports the observations seen in the emRiboSeq data (Figure 3.2b), although the patterns are less clear.

I was interested by the increase in pol α and δ signal seen just upstream to the origin of replication on the forward strand in the emRiboSeq data, and wondered whether this might be related to increased Pol α and δ activity around the origins of replication. There remains uncertainty as to whether at the origins of replication synthesis of the leading strand is initiated by Pol ε. It is possible instead that this initiation is carried out by Pol α and δ, with subsequent synthesis of the leading strand undertaken by Pol ε. A paper published since I conducted this analysis, using HydEnSeq data (with the modification of using *E. coli* Rnase HII rather than alkali hydrolysis), concluded in favour of this second option (Zhou *et al.*, 2019).

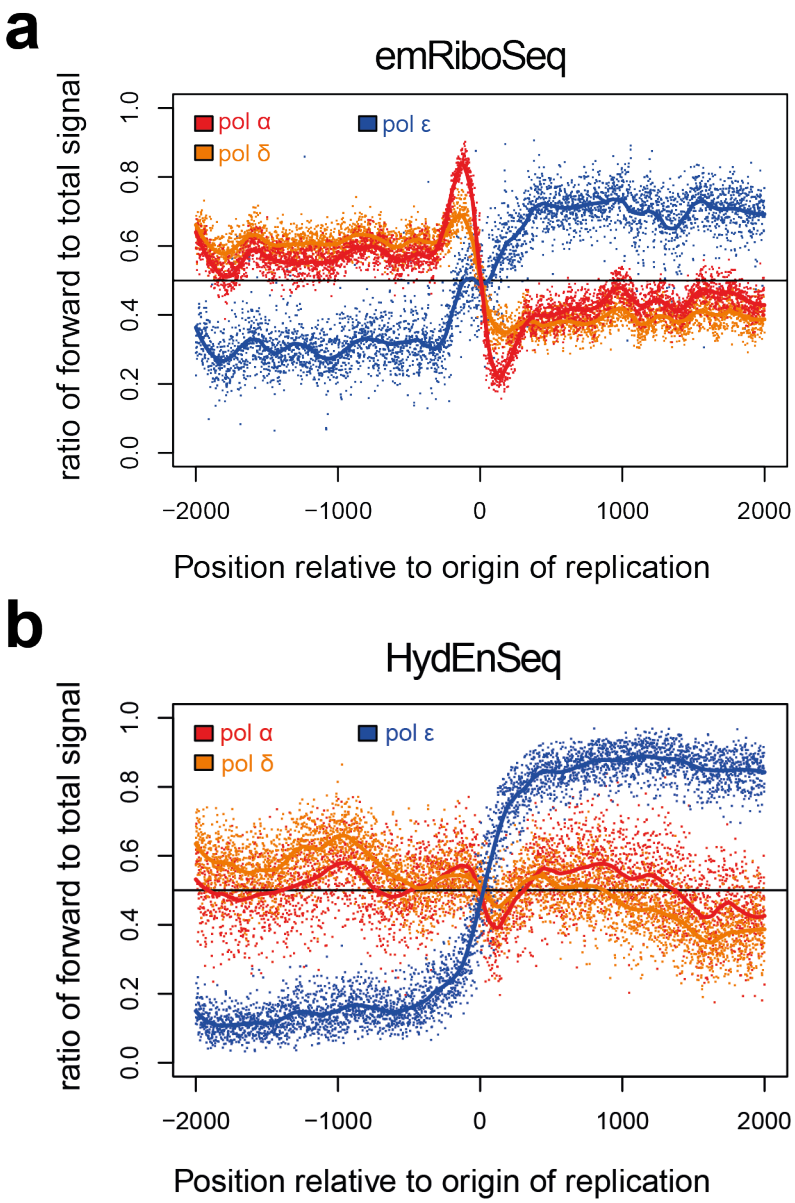


Figure 3.2. Tracking polymerase activity over origins of replication in S. cerevisiae. a) emRiboSeq signal over S. cerevisiae origins of replication shows increase in Pol ε signal downstream of the origin, and reciprocal changes for pol α and δ. *Data is shown for Pol α, δ and ε mutator strains, all in an RNase H2 null background. The y axis shows the ratio of forward strand to total signal for each strain, superimposed over 164 confirmed origins of replication, normalised for genome wide trinucleotide signal. The x-axis shows position relative to the mid-point of the origin of replication. The line crossing the y-axis at 0.5 indicates where signal would be if equal on the forward and reverse strand. Raw data for each position relative to the origins is shown, and a smooth curve has been fitted using Loess (local polynomial) regression.* ***b) A similar pattern is seen using HydEnSeq data.*** *Data is shown as for a); it can be seen that whilst for Pol **ε** the pattern across the origin of replication is similar to that for emRiboSeq, for Pol α and δ the change across the origin of replication is less clearly delineated.*

In order to gain deeper insight into the activity of the polymerases over the origins of replication, I used the same data to look for *overall* activity of the polymerases over the origins of replication, to determine whether in fact there appeared to be increased Pol α and δ activity relative to Pol ε over the origins. For both approaches (emRiboSeq and HydEnSeq) I normalised for overall ribonucleotide incorporation for each technique across the *S. cerevisiae* genome, so that signal over the origins of replication was not a reflection of sequence context only (both emRiboSeq and HydEnSeq have shown that riboC and riboG are preferentially incorporated at a genome wide level).

It can be seen that whilst the frequency of ribonucleotide incorporation appears to *decrease* at the origins of replication for emRiboSeq, with HydEnSeq there appears to be an *increase* in overall ribonucleotide incorporation. I wanted to rule out an effect of position in cell cycle on results: the majority of HydEnSeq data is from cells grown in log phase growth, whereas the majority of that for emRiboSeq is from cells in stationary phase. I therefore analysed the data available for emRiboSeq for cells in log phase growth, but again found a *decrease* in overall ribonucleotide deposition around the origins of replication.

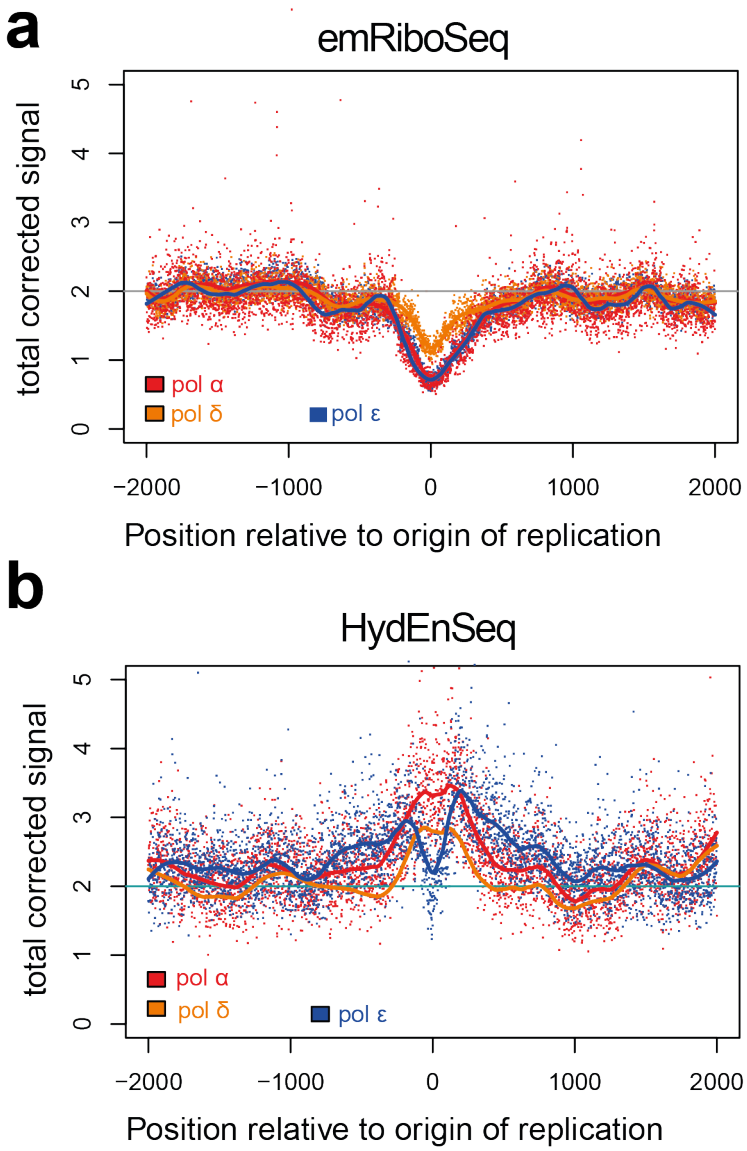


Figure 3.3. Comparison of likelihood of total emRiboSeq and HydEnSeq signal over origins of replication in SacCer3 genome. a) Overall emRiboSeq signal decreases over origins of replication. *Sum of signal for emRiboSeq (in reads per million) over origins of replication for mutant polymerases. The horizontal line intersecting the y axis at 2 represents mean ribonucleotide incorporation across the entire genome (value of 1 for the forward and reverse strand). Each data point is corrected for mean ribonucleotide incorporation at that trinucleotide at a whole genome level. Data for 164 origins of replication. b)* ***) Conversely, overall HydEnSeq signal increases over origins of replication*** *Sum of signal on forward and reverse strand for HydEnSeq for mutant polymerases.*

Overall, these analyses suggested to me that there may be differences in the results given by each technique related to the methodology (RNase H2 vs alkali cleavage, post-cleavage processing and sequencing steps) rather than the underlying biology. Of note, the recently published paper on the relative role of Pol δ vs epsilon at replication origins (Zhou *et al.*, 2019) in essence uses the peak in Pol δ activity at the origins of replication (Figure 3.3) to argue for Pol α-to-δ-to-ε leading strand initiation. This conclusion is not supported by the emRiboSeq data. These results emphasise the importance of an orthogonal technique to study ribonucleotide incorporation at a genome wide level.

3.2.1.2 Location of genome embedded ribonucleotides relative to Okazaki junction sites

Another outstanding question in DNA replication field is the degree to which Okazaki fragments are completely removed during replication. An analysis using alkali gel hydrolysis (Reijns *et al.*, 2015) suggested that 1.5% of the *S. cerevisiae* genome is synthesised by primase-Pol α, which raises the possibility that the RNA primers synthesised by this enzyme complex might also be retained. An approach to this problem could be to compare results for emRiboSeq to HydEnSeq at prespecified genomic locations, and whether there are two distinct peaks at, for example, the presumed location of Okazaki fragments, to infer whether there was at any point a run of retained genome embedded ribonucleotides at these locations. This is because emRiboSeq would cleave at the 5’ end of any RNA run, whereas emRiboSeq would cleave at the 3’ end; a gap between the 2 peaks at the same location would suggest the presence of an intervening RNA segment.

To identify the locations of Okazaki fragments, data was downloaded from a paper (Smith and Whitehouse, 2012) which had used a *S. cerevisiae* strain defective in DNA ligase I. This enzyme normally ligates together the Okazaki fragments which make up the lagging strand during replication. Based on a methodology devised by Harriet Kemp and Martin Taylor (Reijns *et al.*, 2015), I calculated the probability of any position in the *S. cerevisiae* genome constituting the start of an Okazaki fragment. I then selected the 10,000

highest values peaks for emRiboSeq data for a Pol α mutant, and then superimposed data for the HydEnSeq signal for the same strain, normalised as per previously as reads per million. I also superimposed the Okazaki Junction Rate as calculated for each of the peak locations. Link to code available in Materials and Methods (Section 2.3.2.1.2).

Both techniques show an increased likelihood at the same locations for the retention of ribonucleotides in a Pol α mutant. They also show an increased probability of the formation of an Okazaki junction (Figure 3.4a). This suggests that ribonucleotides may be consistently retained at the location of Okazaki fragments, which, as described by Smith and Whitehouse, co-locate with the presence of nucleosomes. The mechanism for this apparent ribonucleotide retention remains unclear: they may be ribonucleotides which are stochastically misincorporated by polymerases and subsequent protected from excision by the RNase H2 machinery by nucleosome binding, or they may be retained elements of the Okazaki RNA primer that initiates synthesis of the lagging strand.

Another result from this analysis is that it provides some support for the pol-alpha retention hypothesis put forwards previously (Reijns *et al.*, 2015) (Figure 3.4b). Here, an increased mutation rate seen in human populations downstream of Okazaki junctions is posited to be due to the retention of DNA synthesised by the more error prone primase-Pol α. Because Okazaki junctions coincide with the location of nucleosomes and other DNA binding proteins, it is possible that Pol δ activity is impeded by these proteins, leading to retention of Pol α synthesised tracts downstream of the proteins. If this model were correct one would expect a peak of Pol α* synthesised DNA downstream of the location of Okazaki junctions, which is seen in this analysis; however one might expect a longer tail downstream of the location of the junctions.

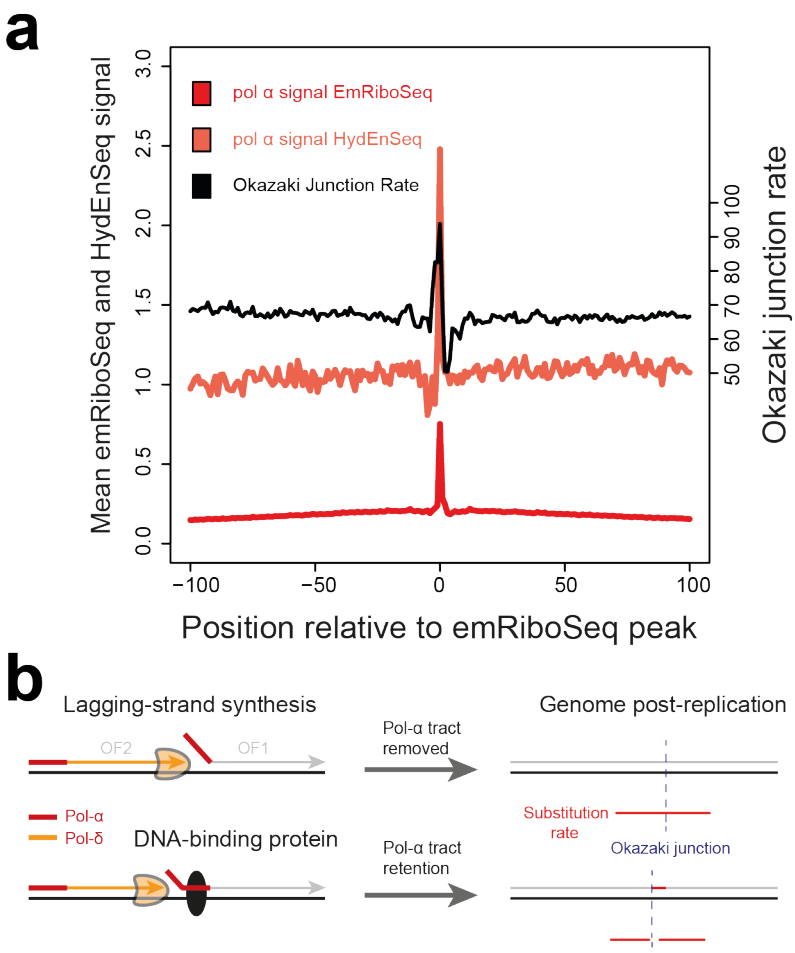


Figure 3.4 Ribonucleotide co-location with Okazaki junctions. a) Comparing peak signal from emRiboSeq to HydEnSeq signal and Okazaki Junction Rates at the same locations. *The locations of the 10,000 peaks with highest signal in the emRiboSeq Pol* ***α*** *library were compared to signal from the same genomic locations in the Pol α HydEnSeq library, and to the Okazaki Junction rate as computed from data in a previous study in S. cerevisiae* (Smith and Whitehouse, 2012). ***b) Model for Pol α retention hypothesis*** (Reijns *et al.*, 2015). *Here DNA binding proteins, which co-locate with the location of Okazaki junctions, are hypothesised to affect removal of Pol α synthesised DNA, leaving this more-error prone DNA embedded in the replicated genome. Reproduced with permission from* (Reijns *et al.*, 2015).

The colocation of the main peak for both emRiboSeq and and HydEnSeq data suggests that in most cases the retained ribonucleotide at these locations represent a single nucleotide, rather than a tract such as that created by a retained Okazaki primer. An orthogonal, high throughput technique would allow for further investigation of this process. Accurate determination of whether tracts of RNA are incorporated into DNA at any stage of cell division would allow mechanistic insights into the frequency of their incorporation, and mechanisms involved in their removal, for example the removal of Okazaki fragments during lagging strand replication.

As described above, if successful, nanopore sequencing would *quantitatively* detect single or runs of multiple ribonucleotides along a strand of DNA, detect phasing of ribonucleotides along this strand, and allow their identification at single nucleotide resolution. A first necessary step was to demonstrate proof of concept for the technique, and this in turn entailed consideration of how to build a model library for the large number of possible *k*-mers with an embedded ribonucleotide.

3.2.2 Key problem of calculating values for ribonucleotide in deoxyribonucleotide context

In order to be able to detect a single ribonucleotide in DNA context, assuming *k*-mers of a length of 5, models (the amplitude change assigned to each *k*-mer) would need to be calculated for a large number of *k*-mer possibilities. For the instance below, where N signifies a DNA base, and **R** an RNA base:

NNNN**R**NNNN

The decoding of the nanopore single works from what are called “squiggle profiles” which intrinsically encode information on flanking bases (the 4 Ns up and downstream of the ribonucleotide). A training set to establish signal for a central ribonucleotide would require information on what both the DNA and RNA signal profiles look like; this would entail a full combination of 4 upstream and 4 downstream nucleotides. With a single ribonucleotide such as rC this would represent 4^8 permutations of the flanking -4 and +4 nucleotides =

65,536 different model templates, multiplied by 4 for any central ribonucleotide, and by 2 for a DNA version to act as a control. This would consist of a total of 524,288 permutations.

A simpler starting dataset to build a library of 5mers would consist of 5 x 1024 (4^5) *k*-mers with a central ribonucleotide (**R**NNNN, **N**RNNN and so forth) plus a final 1024 DNA controls, for a total of 6144 *k*-mers. This would provide a full training set to obtain pore measurements for a 5-mer library, but would entail the loss of wider sequence context variability information.

We reasoned that if one were able to replicate across from a region with a ribonucleotide in a determined position, with stretches of DNA inserted stochastically in equimolar concentrations both before and after the ribonucleotide, this would give a read out (strictly speaking, the complement) of the true sequence, which, if sufficient oligonucleotides were generated and then sequenced, would allow for the determination of the full number of models needed to detect the presence of a single embedded ribonucleotide. For this method to be successful, it would require both the forward strand (containing the embedded ribonucleotide) and the reverse strand (the complement of the forward strand) to be sequenced and matched to one another. Oxford Nanopore Technologies had recently introduced the $1D^2$ system, which meant that this should be possible in principle.

3.2.3 Design of a ribonucleotide containing oligonucleotide to identify differences in k-mer amplitude signals

Initially, I set out to determine whether it was possible to sequence with the necessary accuracy a short, synthesised oligonucleotide with a single ribonucleotide, and compare the amplitude signal for this oligonucleotide to an equivalent DNA control. To develop this system, it would prove necessary to synthesise the complement of the original oligonucleotide. I therefore created an overlapping pair of oligonucleotides that could be filled in in both directions to create a double stranded sequence of sufficient length for nanopore sequencing, but composed of two segments that could be easily synthesised

by commercial manufacturers using current technology, with lengths of 60 nucleotides (for oligonucleotide with embedded ribonucleotide, the longest available using commercial suppliers) and 120 nucleotides (for oligonucleotide composed entirely of DNA, again the length limit for high quality DNA oligonucleotides from commercial suppliers). To distinguish between the RNA containing oligonucleotide sequences and a DNA control, a barcode was introduced downstream of the ribonucleotide containing sequence, and the complement of the barcode for the RNA-containing sequence used as the barcode for the DNA control (Figure 3.5).

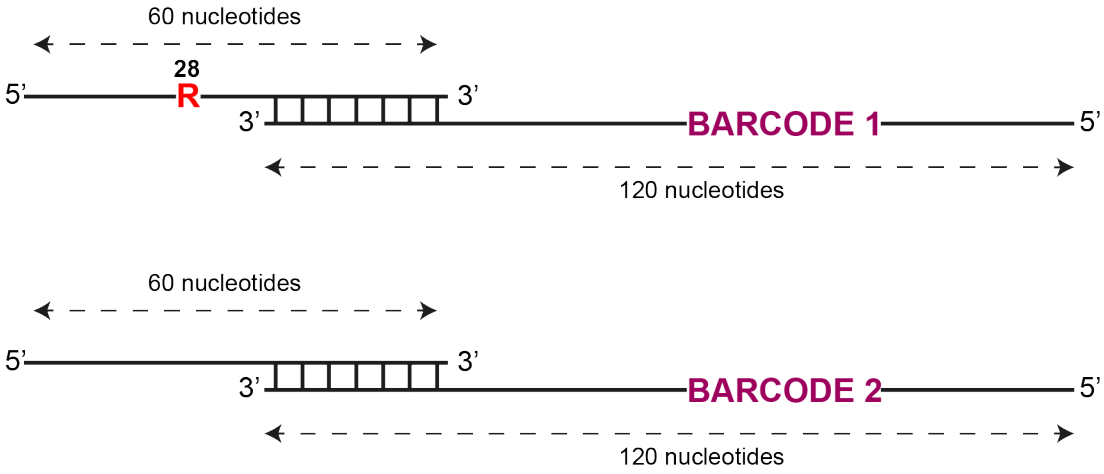


Figure 3.5. Schema for overlapping duplex to examine difference in nanopore signal for single ribonucleotide vs DNA control. a) Ribonucleotide containing duplex. *The forward strand is 60 nucleotides long, and contains a riboC (depicted with an R) at position 28. The reverse strand is 120 nucleotides long, contains a 20 nt long barcode (barcode 1), and has a 20 nucleotide complementary sequence at the 3' end.* ***b) Equivalent DNA control.*** *Without a ribonucleotide, and a different barcode to allow discrimination from the ribonucleotide containing equivalent.*

3.2.4 Design of an oligonucleotide maximising nucleotide discrimination during sequencing

One of the main shortcomings of nanopore technology is its low per-nucleotide accuracy when compared to Next Generation Sequencing. Due to the manner in which *k*-mers are detected, nanopore sequencing is prone to skips or stays, particularly when there are repetitive homopolymer runs (for example AAAAA). The signal as this homopolymer run passes through the pore is very similar for

a number of consecutive *k*-mers, transitions between nucleotides are not readily detected, therefore one of the nucleotides in the run may be missed (a skip), or more nucleotides than are actually present inferred from the signal (a stay). In order to circumvent this, I used the models for each *k*-mer provided by Oxford Nanopore Technologies (details in Materials and Methods, Section 2.3.2.2.1) I designed an oligonucleotide sequence with a high probability of being accurately sequenced, and thus the location of a single embedded ribonucleotide being accurately compared between oligonucleotides. Using the models from ONT, I wrote a Python script with the following conditions:

- No repeat sequences of any nucleotides (eg AA, TT, or longer homo-polymer run)
- No dinucleotide repeats (eg TATA), as these lead to increased chance of primer self-dimerisation
- No trinucleotide repeats, for the same reason
- Greatest possible delta in amplitude between each subsequent *k*-mer
- No repeat of previous *k*-mer (unless this condition is implemented, the script creates an oligonucleotide of repeat sequences which meet the first 2 conditions)
- In many situations, there was only a single *k*-mer that met the above conditions, but if there was more than one option which met all these conditions, the script weighted the options by the distance in amplitude from the preceding *k*-mer, and selected a *k*-mer from the list of options.

This script (Materials and Methods, Section 2.3.2.2.1) generated 10 000 possible oligonucleotides of a length of 155 bases. Using the Mfold (Zuker, 2003) and Unafold packages (Markham, Zuker and Keith, 2008) a score was assigned for the formation of dimers. The 10 oligos with the lowest chance of dimerization were manually inspected, and the pair of oligos with the lowest chance of dimerization selected.

These oligonucleotides were then annealed, filled in using Klenow, A-tailed, and sequenced as per the standard ONT workflow (Figure 3.6; Materials and Methods, Section).

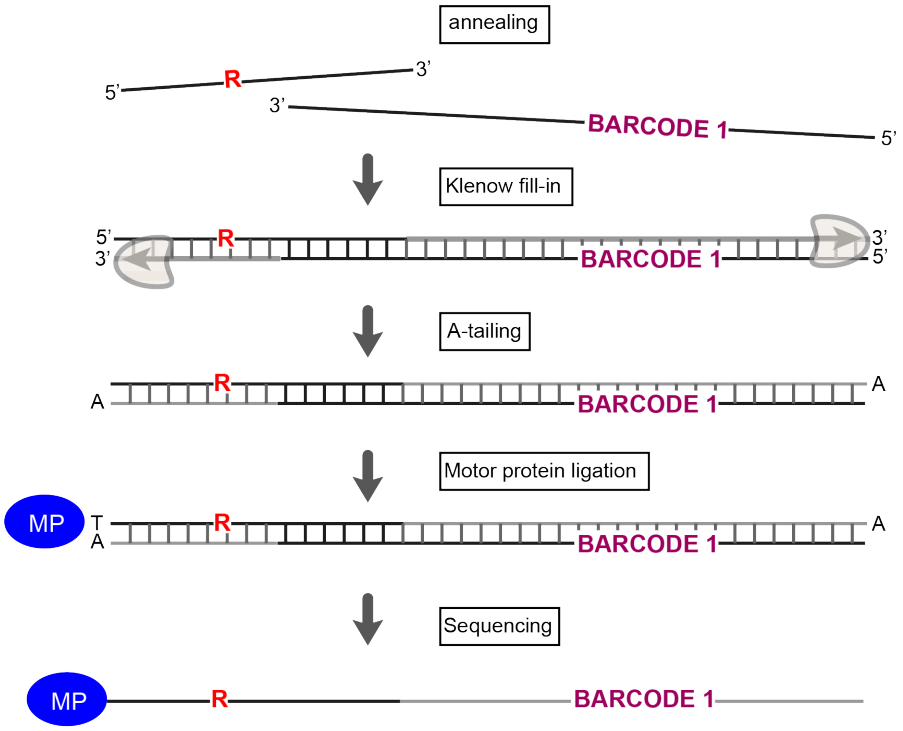


Figure 3.6. Schema for synthesis of duplexes for sequencing to detect single embedded ribonucleotide. *The example shown is for a ribonucleotide containing oligonucleotide, but the principle is the same for the DNA control. The oligonucleotide containing the ribonucleotide (DRD60) is annealed to a oligonucleotide with an overlapping 20 nucleotide segment and a barcode. Klenow is used to fill in the 3' ends, and after this both end are A-tailed. Motor protein (MP) is ligated and then the forward strand containing the oligonucleotide is sequenced.*

3.2.5 Demonstration of proof of concept for nanopore sequencing using oligonucleotides with single embedded ribonucleotide

In my proof of concept experiment, the presence of the riboC as opposed to a canonical cytosine led to a difference in the mean amplitude for a number of the *k*-mers containing the riboC (Figure 3.7). The mean amplitude signal for *k*-mers from the control DNA sequence is shown in blue, compared to the mean amplitude signal for equivalent *k*-mers with a riboC, which is shown in red.

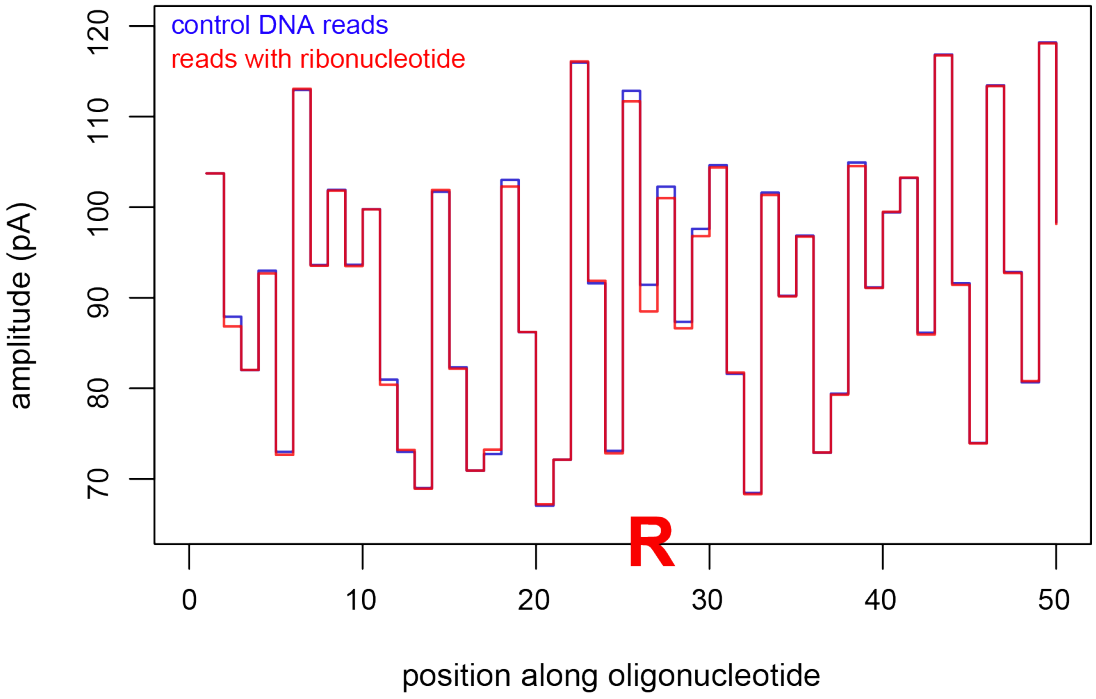


Figure 3.7. Comparison of mean amplitude signals over the relevant k-mers for riboC base compared to DNA control. *The location of the ribonucleotide is indicated with a red R on the x-axis. It can be seen that there is a shift in the mean amplitude for the k-mers surrounding the C.*

I next examined the distributions for the *k*-mers surrounding the ribonucleotide (Figure 3.8). The rC was incorporated at position 28 in the oligonucleotide, so that the 6-mers which include this ribonucleotide are 23-28. None of the distributions are wholly separate from one another, and they vary in their degree of differentiation depending on the *k*-mer. The patterns of these results: overlapping distributions, and amplitude differences varying on the *k* -mer, are similar to those seen in experiments looking at 5-mC and BrdU (Rand *et al.*, 2017; Simpson *et al.*, 2017; Hennion *et al.*, 2018).

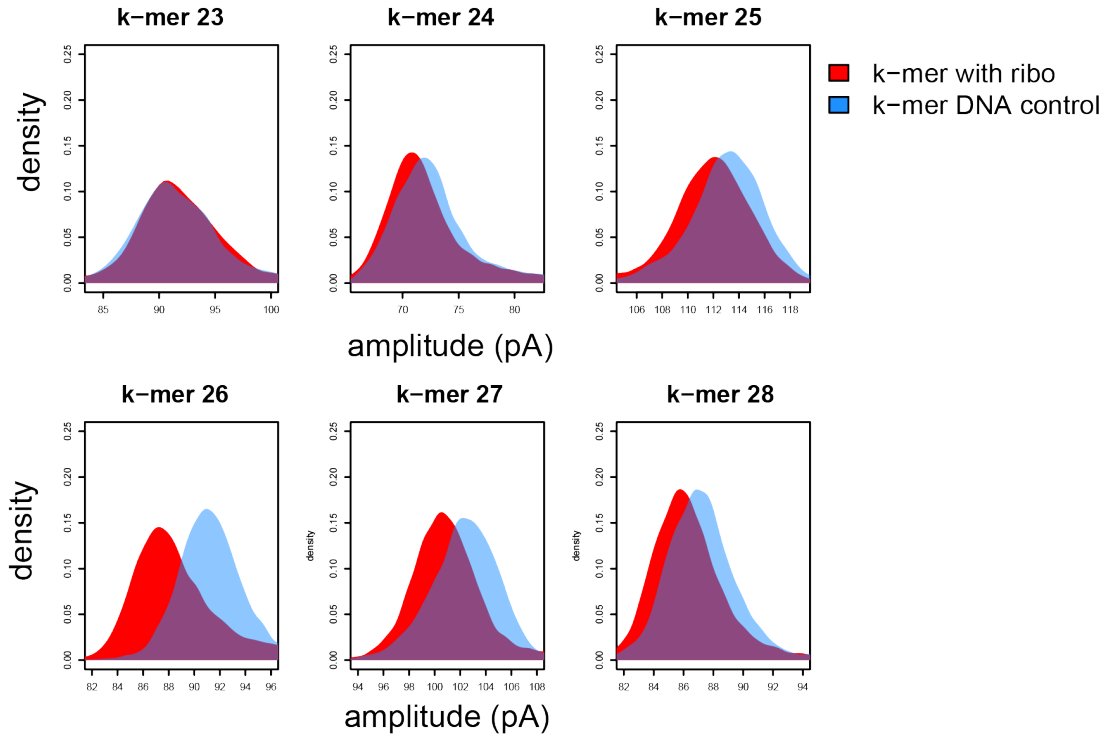


Figure 3.8. Distribution of amplitude signal at 6 k-mers surrounding position 28 for ribo C compared to DNA control. *The amplitude distribution for the k-mer containing a ribonucleotide are shown in red, and for the DNA control in blue. Distributions calculated using the kernel density function in R.*

3.2.6 Identification of a distinctive kinetic signature associated with DNA embedded ribonucleotides

As part of the analysis we also examined the time that each *k*-mer took to pass through the pore. Our prediction was that there might be an impact on the time that the ribonucleotide containing *k*-mer took to pass through the pore, with the ribonucleotide containing complex passing through either faster or slower than the DNA equivalent. However instead we found that in fact the *k*-mers affected were those *upstream* of where the ribonucleotide was located. With the ribonucleotide at position 28, the *k*-mer which showed the greatest increase in the time taken to pass through the pore was 14, leaving a distance of 14 nucleotides between the ribonucleotide and the increase in transit time (Figure 3.9a and Figure 3.9b). The median value for a fully DNA control compared to a central rC shifted from 0.004 seconds to 0.01 seconds (Figure 3.9c, two sided Kolmogorov-Smirnov test for independence p-value < 2.2e-16).

Examining the *E. coli* outer membrane lipoprotein CsgG (Goyal *et al.*, 2014), which is the template for the R9 nanopore used in our experiments, suggested the reason for this shift in the time taken for the *k*-mer to pass through the pore (Figure 3.9d).

Forty ångström is equivalent to 4 nm, and the distance between nucleotide bases in B-form DNA is 0.33 nm. Four divided by 0.33 is 12.1, which is in rough agreement with the likely distance between the ribonucleotide and the *k*-mer in the channel at the point at which the ribonucleotide engages with the motor protein to start entering the pore. Direct RNA nanopore sequencing uses a different motor protein than DNA sequencing (Garalde *et al.*, 2018), as the protein used for DNA sequencing lacks efficient binding activity or processivity on RNA. We therefore reasoned that it was some feature of the interaction between the ribonucleotide and the motor protein that lead to this slowing down of the oligonucleotide as it was being sequenced.

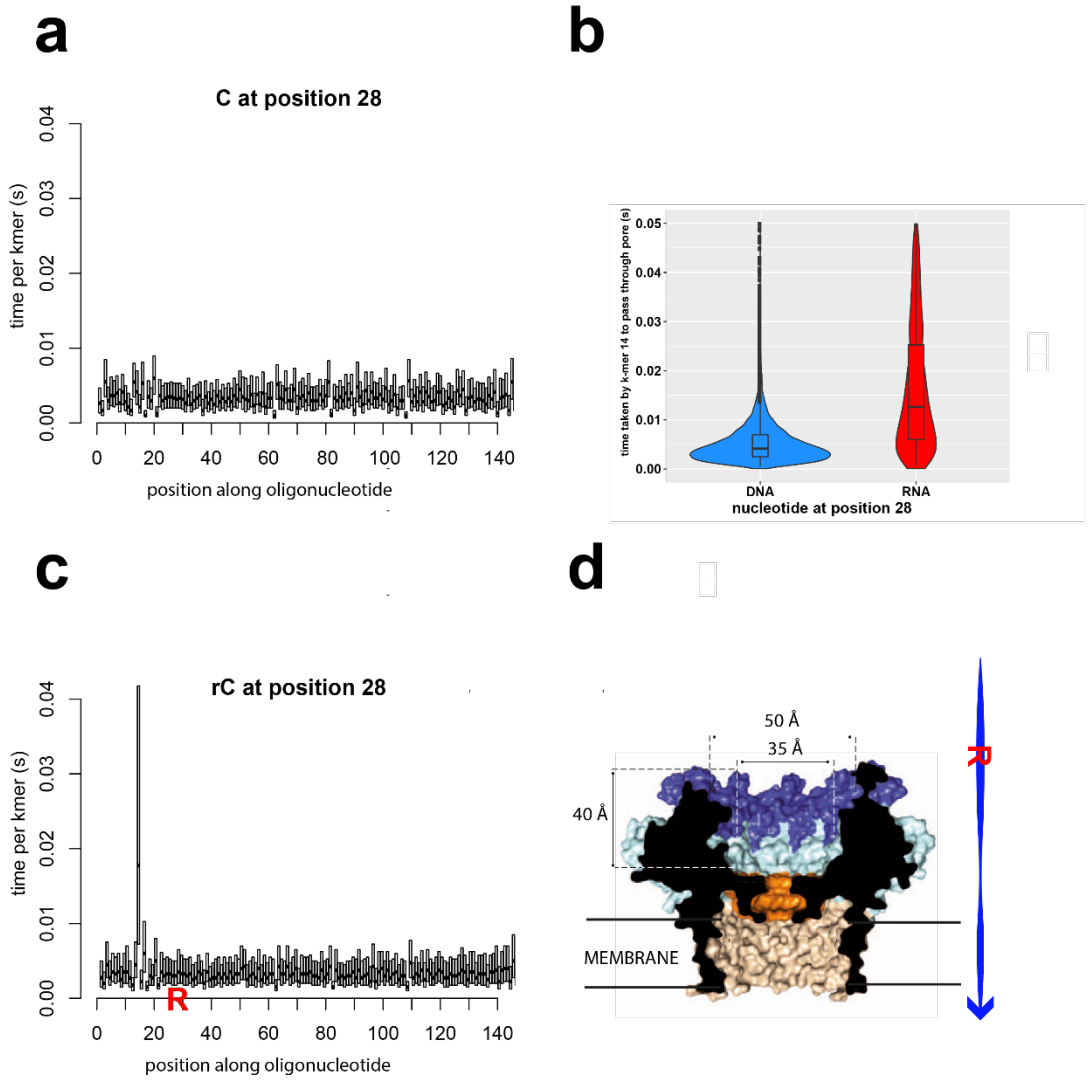


Figure 3.9. Demonstration of a kinetic signature associated with an embedded ribonucleotide. a) Kinetics for DNA oligonucleotide. *Median and interquartile range for time taken for each k-mer to pass through the sequencing pore for an entirely DNA oligonucleotide.* ***b) Kinetics for equivalent RNA containing oligonucleotide.*** *Kinetic signature for an identical oligonucleotide, except for replacement of cytosine at position 28 with rC (marked with red R on x axis).* ***c) Kinetic signal for k-mer 14.*** *Violin plot with median and interquartile range for k-mer 14, 14 nucleotides upstream from the position of the riboC.* ***d) Model of E.coli outer membrane lipoprotein CsgG****. Adapted from* (Goyal *et al.*, 2014)*, showing the 40 ångström distance between the entry to the pore and the central channel, roughly equivalent to 14 nt in length; the blue arrow represents the oligonucleotide passing through the channel with an embedded ribonucleotide at the point of entry into the pore.*

3.2.7 Designing oligonucleotides to test 64 possible trinucleotide combinations

Following demonstration of proof of principle for the detection of a ribonucleotide embedded in DNA, before moving on to the design of a system to detect a ribonucleotide in any context, we decided to test for the detectability of a discriminatory signal for a central ribonucleotide in any of the possible 64 trinucleotide contexts (NRN). Using the same sequence designed previously to maximise *k*-mer identification, I included the DNA control for each ribonucleotide on the same oligonucleotide, preceded by a barcode sequence which would allow matching of all oligonucleotides sharing the same RNA containing trinucleotide sequences. The original DRD60 sequence was used as a template for the design of these oligonucleotides (Figure 3.10).

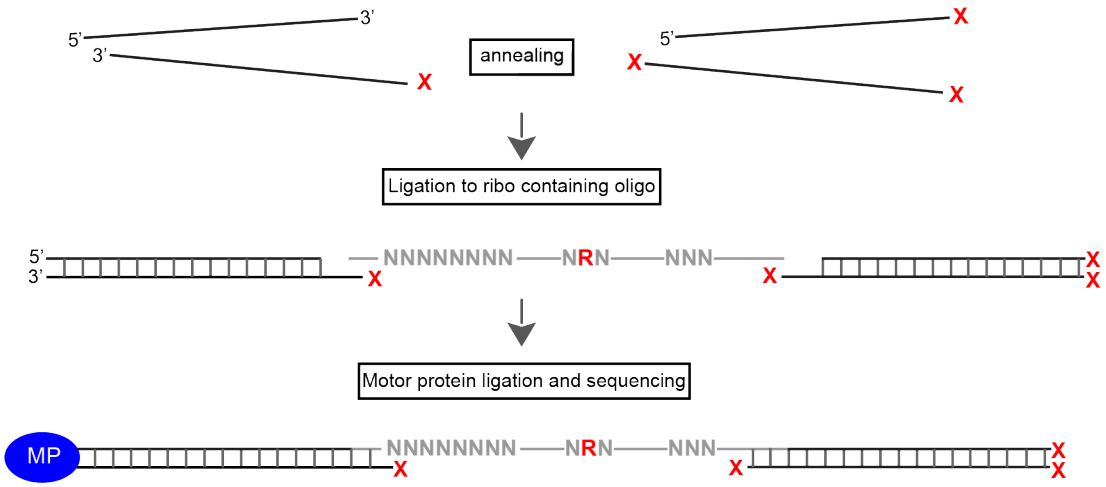


Figure 3.10. Format of ribonucleotide containing oligonucleotide to allow testing of 64 different ribonucleotide containing trinucleotide sequences. *Workflow for assembly of composite sequence containing ribonucleotide. Two flaking segments up and downstream of the ribo containing segment were annealed and then ligated to one of 64 different oligonucleotides with a unique central ribonucleotide (N**R**N). Upstream of this trinucleotide combination was a barcode to map the ribo to the reference, and a downstream DNA control for the trinucleotide combination being tested. **X** refers either to a 5' end functionally blocked by lack of 5'phosphate group, or 3' ends blocked with a C3 Spacer phosphoramidite.*

In order to create the 8 base long barcodes, ONT 6-mer models were used and the same principles as outlined above on page 110 applied. The script

(Chapter 2, Section 2.3.2.2.2) outputted 515 possible barcodes. In order to select the 64 *most* discriminatory barcodes, after discussion with Neil Clark (Chris Ponting Lab) I used *k*-means in order to cluster the possible barcodes into the 64 groups most different to one another. The principle underlying *k*-means is to be able to partition a number of *n* observations into *k* sets, so that each set *k* is composed of the *n* observations that are most similar to one another (Macqueen, 1967). For each barcode I created a vector of the 13 *k* -mers that made up the barcode (this vector is the *n* described above). Following this I used the *k*-means function in R (Hartigan and Wong, 1979) to select the 64 groups (*k*) of *n* most distinct from one another. The principle of clustering vectors in multi-dimensional space is illustrated in Figure 3.11, choosing 3 discriminatory *k*–mers for illustration of the vectors in 3-dimensional (as opposed to the final 13-dimensional) space.

For the group of vectors within each set, I then selected the barcode with the maximum sum of deltas between each of the *k*-mers that made up the barcode, reasoning that this was the most idiosyncratic barcode within that set. Preceding and following the 60 nucleotide RNA containing oligonucleotide were 2 duplexes with a 6 nt long overhang to bind to the oligonucleotide (Figure 3.10). These served 2 purposes: firstly to lengthen the sequence of DNA that passed through each pore, and increase the likelihood of successful sequence, and secondly to stabilise the oligonucleotide as it passed through the pore. The ends of the primer sequences were blocked using a SpC3 molecule to avoid concatemerisation of the duplexes.

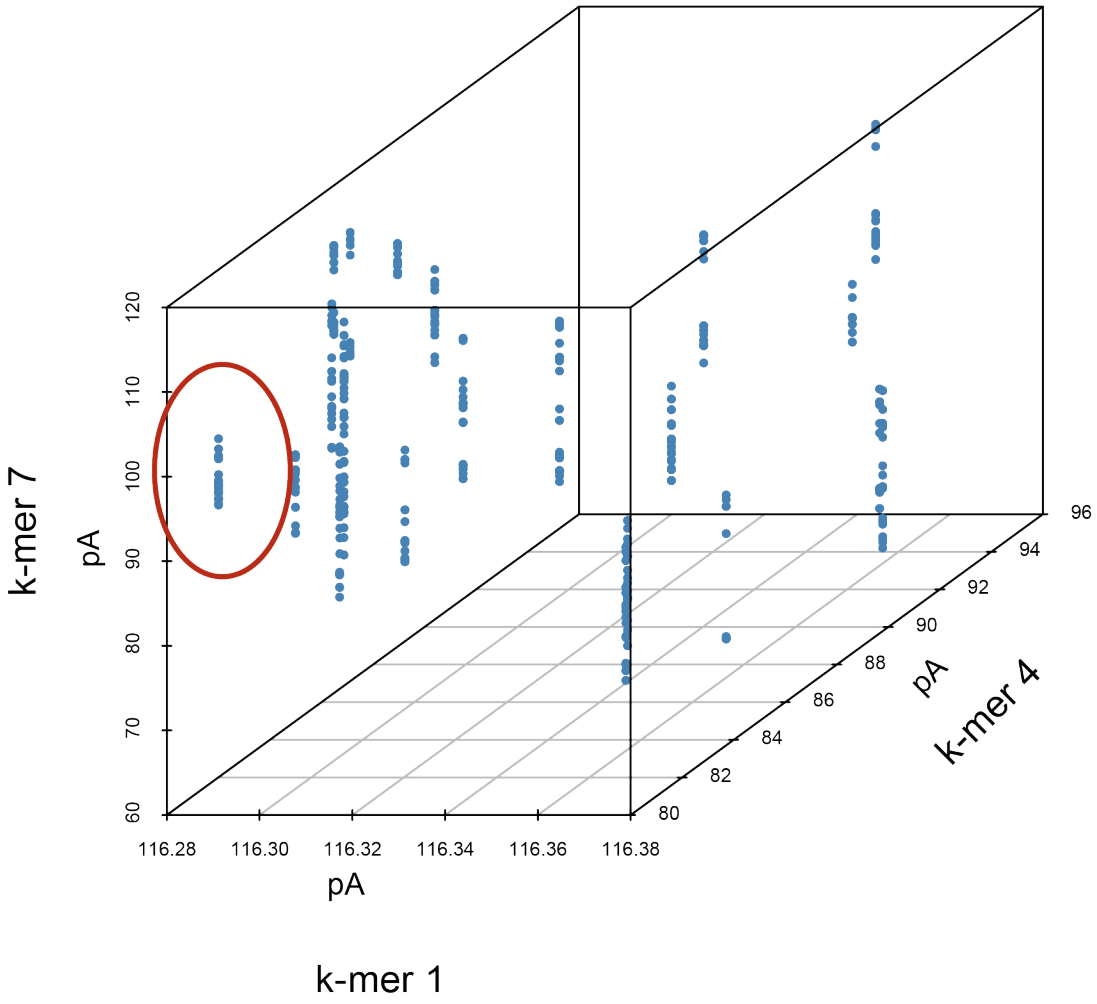


Figure 3.11. Graphical representation of 515 barcode vectors in 3-dimensional space. *Each point is one of the 515 barcodes, which contains a total of 13 k-mers. The kmeans function divides these into a predefined group of clusters (in this case 64) based on similarity to one another. For the purposes of illustration, 3 of the k-mers for all the barcodes are clustered in 3 (rather than 13) dimensional space: k-mer 1 on the x axis, k-mer 7 on the y axis, and k-mer 4 on the z axis. A possible cluster is circled in red.*

3.2.8 Confirmation of a distinctive, but variable, amplitude signature for 64 trinucleotide combinations with central ribonucleotide

Analysis for mean amplitude signal for the 64 different trinucleotide combinations with a central ribonucleotide showed variable degrees of difference with respect to the equivalent DNA control trinucleotide (Figure 3.12).

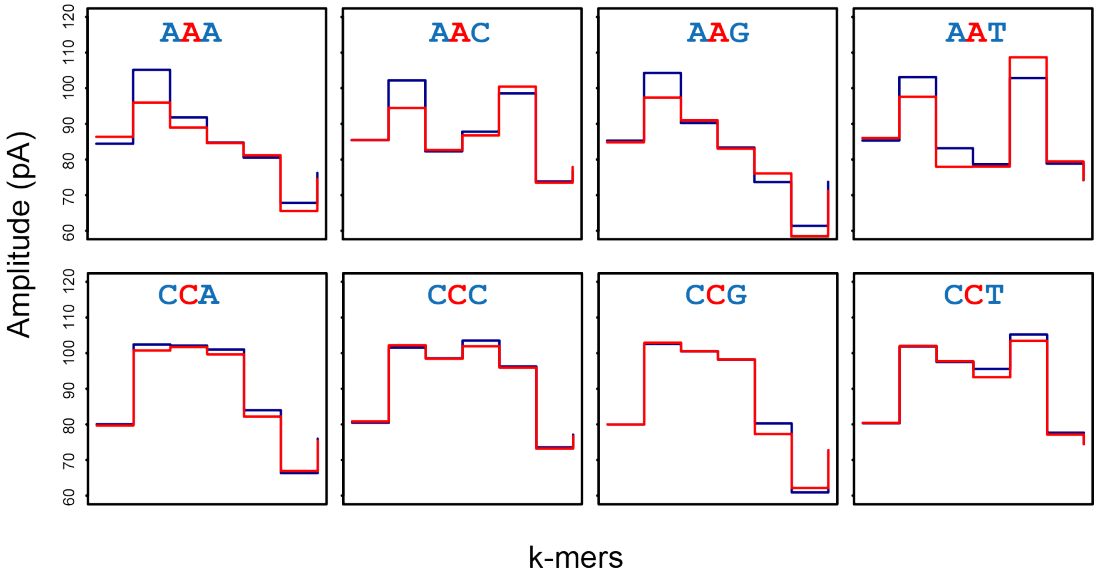


Figure 3.12. Representative trinucleotide amplitudes. *Representative examples of amplitude signal for trinucleotides with a central A and central C. The y-axis represents amplitude, and the step changes along the x-axis k-mers for the same ribonucleotide, as per Figure 3.7. Full amplitude data for all 64 trinucleotide combinations available on GitLab ("/nanopore/oligo_analysis/64_trinucleotides_amplitudes.pdf").*

3.2.9 Confirmation of a distinctive, but variable, kinetic signature for 64 trinucleotide combinations with a central ribonucleotide

Supporting results in the initial proof of concept experiments, some ribonucleotides appeared more likely than others to lead to an increase in the time taken to pass through for a *k*-mer, with the effect being most marked for a central riboC, and least pronounced for a central riboA (Figure 3.13).

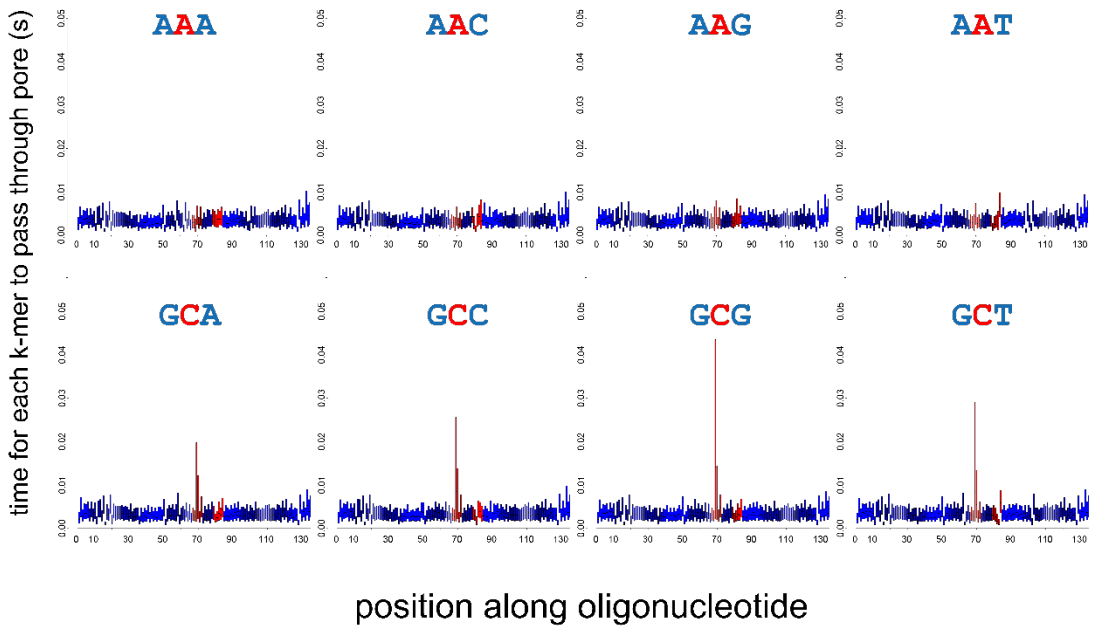


Figure 3.13. Representative plots of increase in transit time upstream of single embedded ribonucleotide. *The x axis shows the k-mers along the oligonucleotide, with the sections highlighted in red representing the location of the ribo and the k-mers 12 nucleotides upstream of the ribo. Median time for each k-mer to pass through the pore on the y axis, with interquartile range. Full kinetic data for all 64 trinucleotide combinations available on GitLab ("/nanopore/oligo_analysis/64_trinucleotides_kinetics.pdf").*

3.2.10 Discriminatory power of amplitude and kinetic signature for each trinucleotide combination

As each ribonucleotide appeared to have different amplitude and kinetic signatures, I calculated the maximum discriminatory *k*-mer for the amplitude (6 *k*-mers incorporating the ribonucleotide) and kinetic (for the *k*-mers between 8-18 nucleotides upstream of the ribo). For the kinetic signature, I calculated the percentage increase in time taken to pass through the pore compared to the mean time for the oligo as a whole; for the amplitude signature the pA difference for the most discriminatory *k*-mer when compared to the DNA control. I found that the trinucleotides with a central ribo C showed high discriminatory power in terms of their kinetic signature, with a limited impact on the amplitude signature. Conversely, those with a central ribo-A showed a large increase in the amplitude signature but a less marked kinetic signature (Figure 3.14). A central ribo G and U appeared to have an intermediate impact in trinucleotide discrimination.

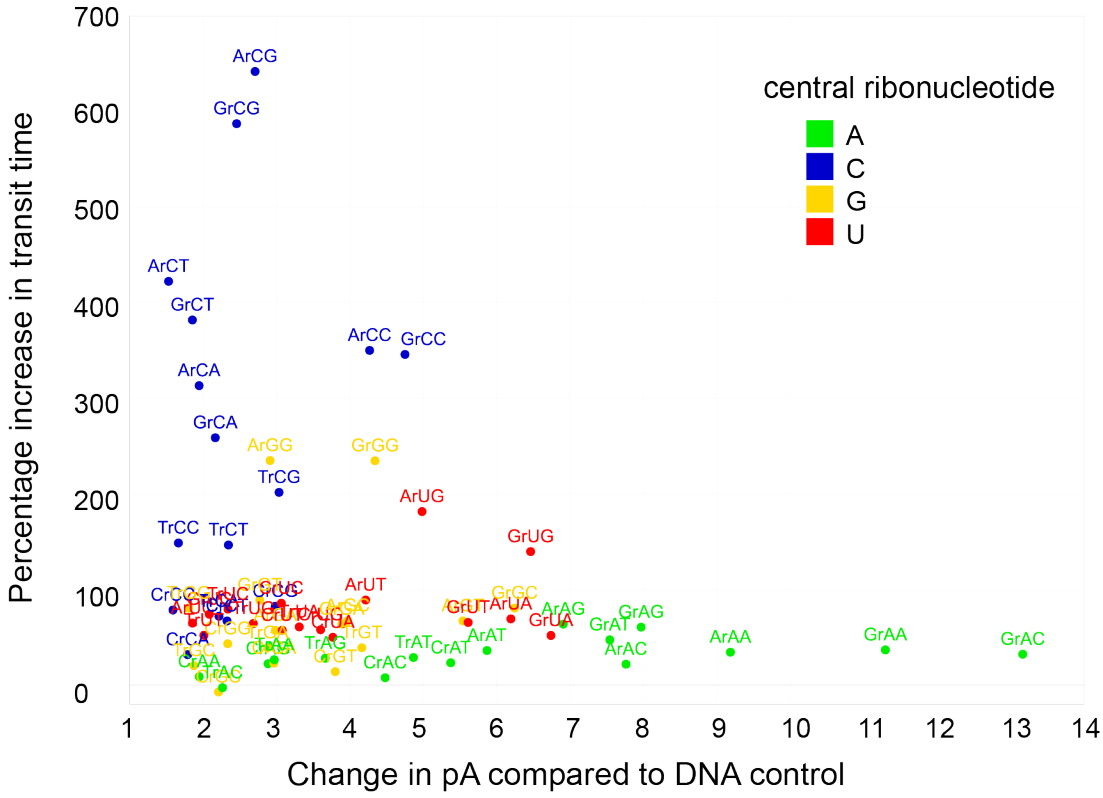


Figure 3.14. Comparison of amplitude to kinetic signatures for 64 trinucleotide combinations. *The y axis shows the percentage change in time taken for the most discriminatory upstream k-mer from each group of the same trinucleotide to pass through the pore, compared to the mean for that class of oligonucleotides. The x axis shows the difference between the most discriminatory RNA containing k-mer for each trinucleotide compared to the DNA control.*

3.2.11 Borrowing from SHAPE technology to enhance the signal from embedded ribonucleotides

Given the possibility that the signal from an embedded ribonucleotide, compared to a DNA control, might be too small to be consistently discriminatory, we considered whether it might be possible to modify the ribonucleotide in some way to enhance the difference in signal when compared to a deoxyribonucleotide. The same principle has been used in RNA biology to map the secondary structure of RNA, in a technique called SHAPE (selective 2′-hydroxyl acylation and primer extension) (Merino *et al.*, 2005). Here, an electrophilic adduct such as NAI (2-methylnicotinic acid imidazolide) or NMIA (N-methylisatoic anhydride) (Lee *et al.*, 2017) is used to conjugate to the accessible 2'OH group of a ribonucleotide. The bulky 2'-O-adduct can then be detected as an impediment to cDNA synthesis by a reverse transcriptase

enzyme, which is unable to synthesize across the adduct. The newly synthesised fragments are then sequenced using NGS, and the ends of the traces using to infer the location of accessible RNA, and hence its secondary structure.

To investigate whether it was possible to conjugate the adduct to a ribonucleotide embedded in a DNA strand, I followed the protocols published by Wilkinson et al and Spitale et al (Wilkinson, Merino and Weeks, 2006; Spitale *et al.*, 2013) to conjugate NMIA and NAI to a single embedded ribonucleotide (Chapter 2, Section). We reasoned that an RNA molecule whose 2' hydroxyl was now bonded with an adduct would be resistant to alkali hydrolysis, and tested this hypothesis using denaturing PAGE (poly-acrylamide gel electrophoresis).

3.2.12 Unsuccessful attempt to adduct acylating agents to a DNA embedded ribonucleotide

Treatment of DNA oligonucleotides containing a single embedded ribonucleotide with the acylating agents NAI and NMIA did not lead to any prevention of alkali hydrolysis, suggesting an absence of adduct formation to the 2' hydroxyl group within the rA used in the experiment (Figure 3.15) Even in this in vitro context there was no evidence of adduct binding, perhaps due to a lack of DNA flexibility in the upstream and downstream arms surrounding the ribonucleotide (McGinnis *et al.*, 2012). We therefore opted to continue to explore the possibility of ribonucleotide base calling based on a combination of amplitude changes and kinetics. However, given the variability of these depending on nucleotide context, we decided it was necessary to investigate a greater proportion of the *k*-mer signal patterns with an incorporated ribonucleotide.

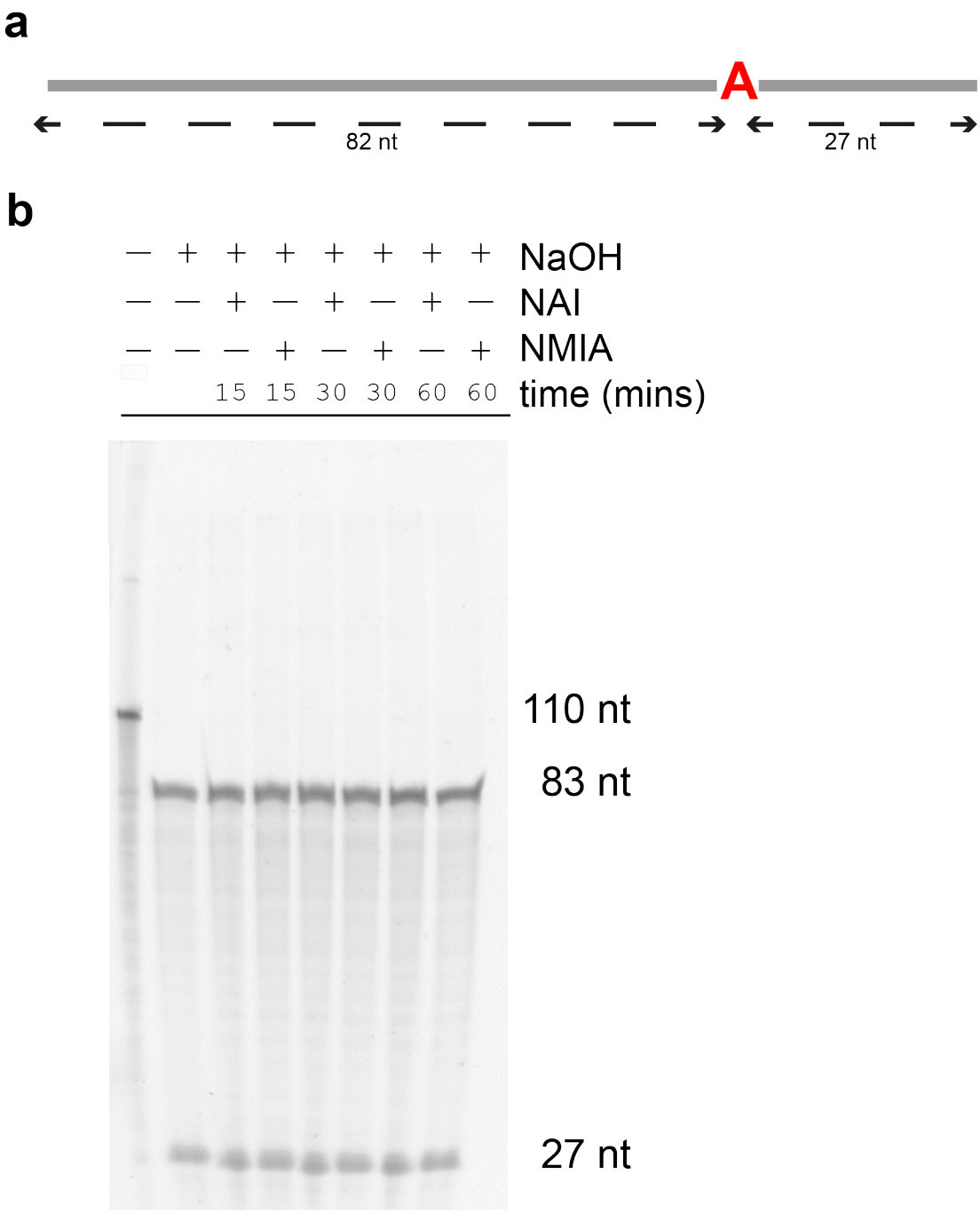


Figure 3.15. Gel showing alkali hydrolysis of DNA oligonucleotide with single ribonucleotide.a)Starting oligonucleotide*.Depiction of a 110 nucleotide long oligonucleotide containing an embedded rA at position 83, with 82 nucleotides upstream and 27 nucleotides downstream****. b) Denaturing PAGE.*** *Gel showing treatment of oligonucleotide in a) with acylating agents NAI and NMIA, followed by treatment with NaOH at 1M concentration. There is no evidence of acylation of the 2'OH in the ribonucleotide and subsequent protection against alkali hydrolysis.*

3.2.13 Design of a system to synthesise across a single ribonucleotide with a hairpin

Discussions with Oxford Nanopore Technology lead to the realisation that their $1D^2$ system, which uses matching of the forward to the reverse strand to increase accuracy of base-calling, relied on length of read to pair the forward strand to its reverse. When analysing native genomic DNA, shearing of fragments in preparation for library sequencing meant that the combination of length discrimination and detection of two strands passing through the same pore in quick succession meant that this strategy was highly likely to result in the selection of the forward and reverse strand of the same original duplex. However, given that any system we designed would consist of short oligonucleotides of the same length, very similar in overall sequence context, I realised that this approach was not going to be viable. I therefore set out to design a system with a hairpin that would connect the ribonucleotide containing sequence to its complement, which could then be sequenced as a single entity, and the RNA containing section mapped to its complement.

I initially designed a long and short hairpin, as there was uncertainty about whether the length of the hairpin would affect the accuracy of sequencing for the complementary strand. I based the design of the hairpins on sequences previously published for high accuracy duplex sequencing (long hairpin) (Taylor, Cinquin and Cinquin, 2016) and bisulfite sequencing (short hairpin) (Laird *et al.*, 2004) (Chapter 2, Section 2.3.2.2). One of the 64 oligonucleotides with a single embedded ribonucleotide was modified to create a duplex to which the hairpin could anneal. A trial of sequencing using both the long and short hairpins was conducted, using the duplex described above, with a ligation step to ligate the hairpin and the ONT $1D^2$ adapters (Figure 3.16).

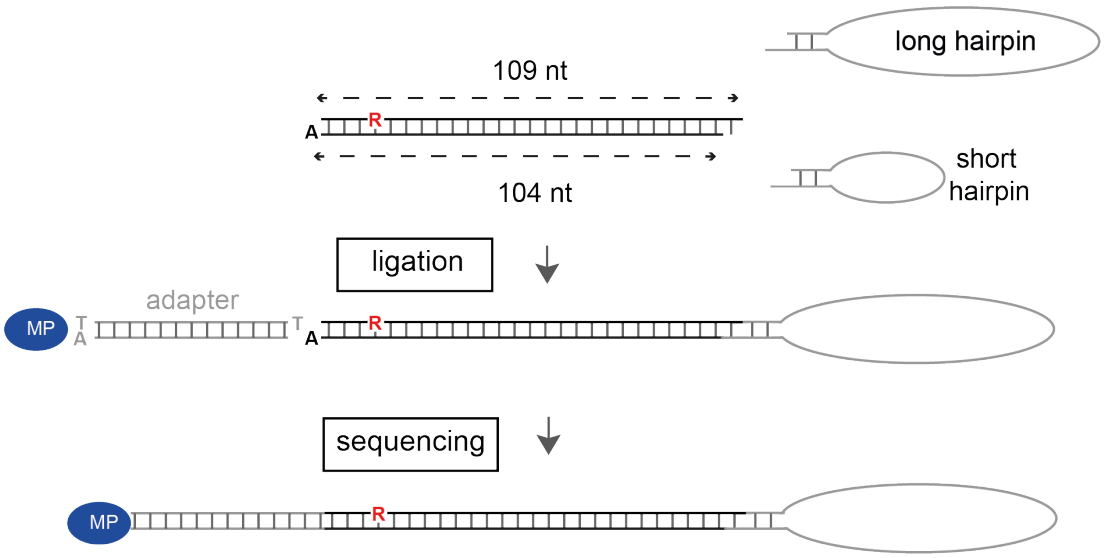


Figure 3.16. Schematic of approach to ligate hairpin and sequence ribonucleotide containing DNA template and complement. *Initially 2 complementary oligonucleotides, 109 and 104 nucleotides long, are annealed together, creating an A overhang on the 3' strand of the complement, and a 6 nt long overhand at the 3' end of the template. After this either a long or short hairpin is ligated. Subsequent to this 1D² sequencing adapters are ligated to the A-tail, and the duplex with ligated hairpin taken forwards into a sequencing reaction. MP: motor protein.*

3.2.14 Investigating the feasibility of using a hairpin to link template and complement

I was able to successfully sequence duplexes ligated to both the long and short hairpins (Figure 3.17). Results from this sequencing showed that the long hairpin appeared to be more likely to result in a full length read, and therefore in subsequent experiments we used this.

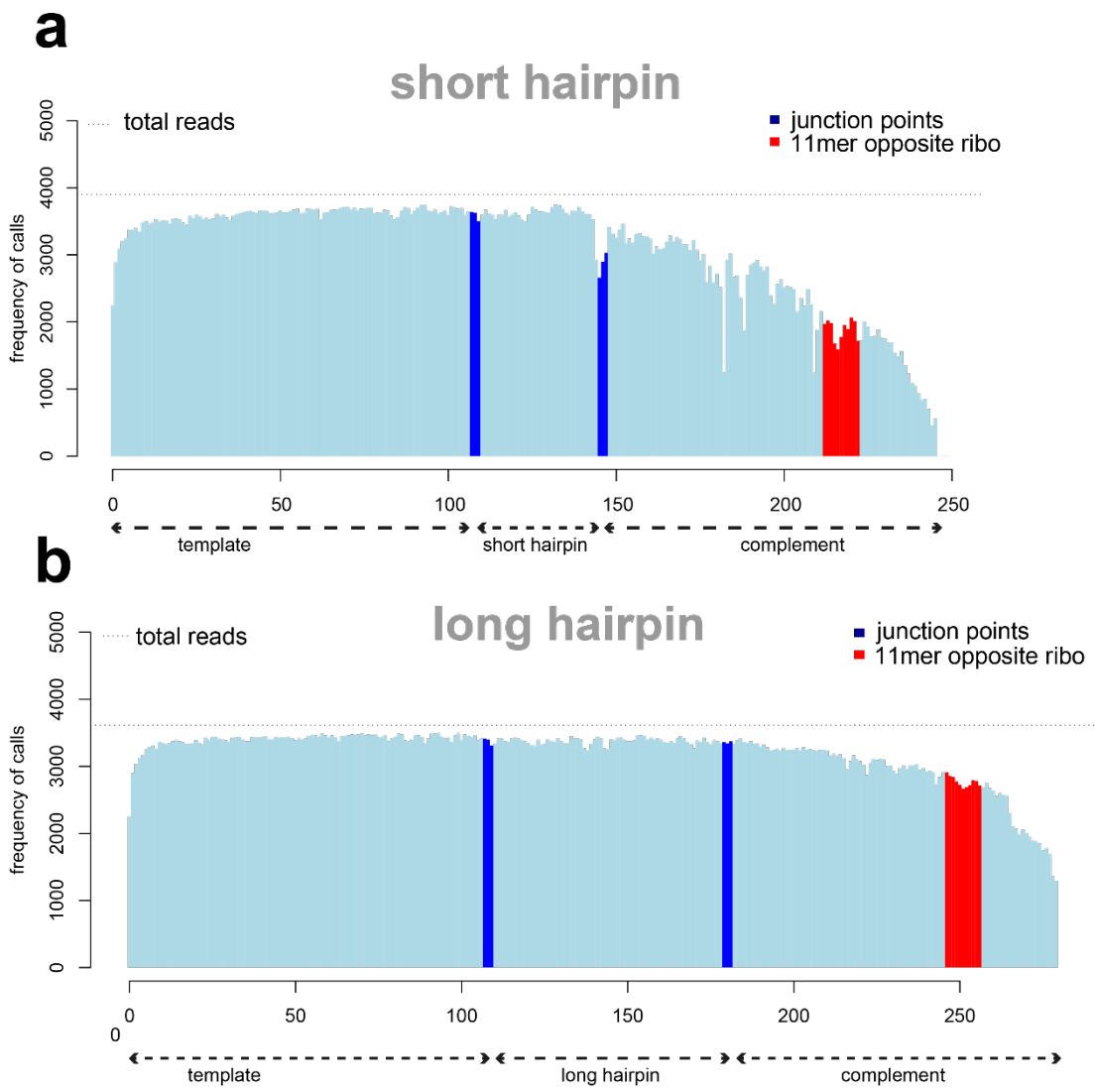


Figure 3.17. Figure showing read counts successfully mapped to short and long hairpin ligated reads. a) Read counts mapping to each position along a duplex ligated to a hairpin. *The grey dotted line over the histogram of read counts represent the total read count for the run, and each light blue bar in the histogram the number of sequences that have a read matching to each k-mer. Below the x-axis are delineated the reference points for the ligated sequence, starting with B2_ext_sh_C, a ribonucleotide containing sequence, followed by the short short hairpin and complementary sequence (B2_ext_sh_C_rev). Junction points between the different components are marked in dark blue. Following the end of the short hairpin sequence there is a drop-off in the number of reads matching the reference sequencing* ***b) Results for sequencing of a duplex ligated to a longer hairpin****. Whilst there is still a drop-off towards the 3' end of the sequence, this is less marked than for the short hairpin*

The next step was to synthesise a new sequence complementary to the template with an embedded ribonucleotide. I initially ligated the hairpin to the template oligonucleotide, and attempted to use the 3' end of the hairpin (composed from hairpin stem and the 6 nt overhang) as a primer from which to synthesise the complementary strand, but found that this was unsuccessful.

Subsequent to this I used a 49 nucleotide long complementary primer to create a double stranded oligonucleotide, which bound to the template oligonucleotide, and also to the 3' end of the hairpin once the 6 nucleotide long overhang had bound to the template sequence (Figure 3.18).

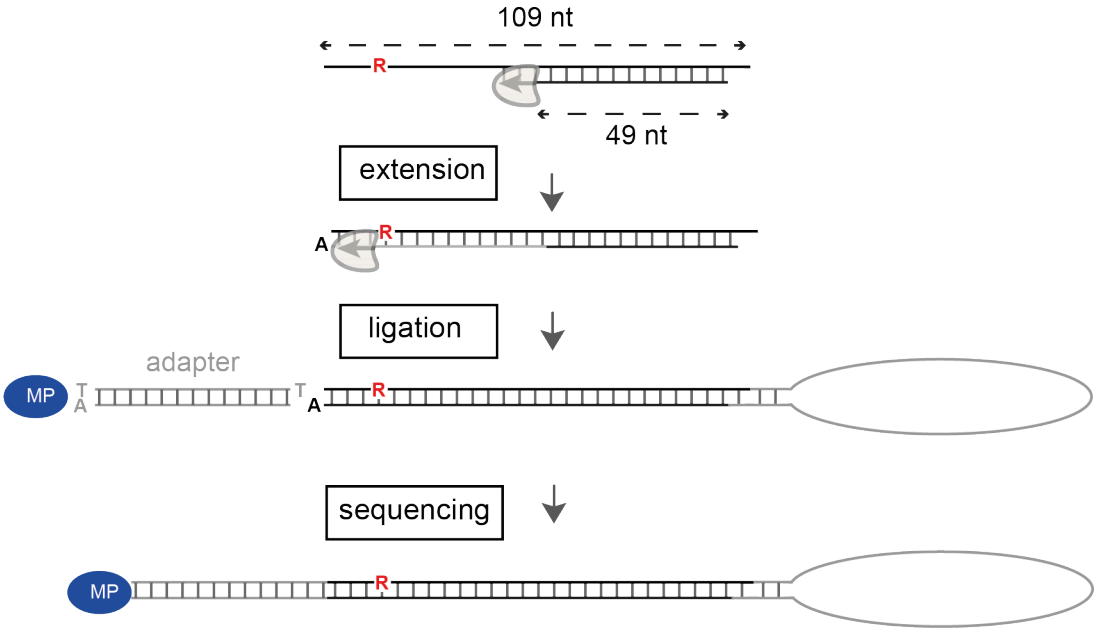


Figure 3.18. Extension across a ribonucleotide containing oligo. *Annealing of 109 nt and 49 nt long oligos is followed by extension across a single ribonucleotide. After this a hairpin and adapter are ligated, and the whole complex sequenced. Ribonucleotide shown as red R.*

3.2.15 Synthesising a sequence complementary to the ribonucleotide containing template

Products of an extension reaction over a single ribonucleotide were successfully sequenced (Figure 3.19). These results showed a similar pattern to previously. However, one aspect of these results was puzzling, and also seen in previous sequencing reactions. This is the drop off in read frequencies progressing downstream. Due to the sequencing set-up, a duplex should only be sequenced if an A-tail was present at the end of the newly synthesised strand, and therefore one would expect a fairly even distribution of reads across the sequence, rather than a decline towards the downstream end going into the adapter sequence. This raised the concern that there might be hairpin

oligonucleotide dimerization rather than self-annealing. This unexpected finding led to the design of a further experiment to examine this hypothesis.

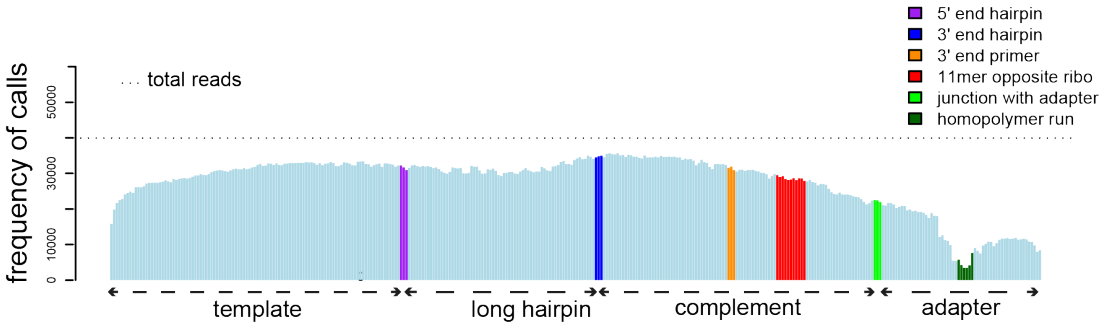


Figure 3.19. Results of sequencing extension over ribonucleotide product. *The sequenced product consists of a template oligonucleotide with a ribonucleotide (B2_ext_sh_C,), followed by a long hairpin, followed by a 49 nt reverse primer (3'rev49), and then the complement of the template oligonucleotide. The green band indicates the intersection of the filled in duplex with the 1D² adapter. The dip in read counts following the intersection with the adapter is likely to represent read inaccuracy over a homopolymer run in the adapter (dark green, C**AAAAA**G).*

3.2.16 Design of an experiment to test efficiency of hairpin formation

In order to confirm hairpin self-complementarity, I adapted 4 further oligonucleotides from the ones used to investigate efficiency of synthesis across a ribonucleotide, keeping the main sequence, but each with a section corresponding to 4 of the 64 barcodes designed and selected previously. The 8 nucleotide long barcode were split into two 4 nt long segments (ID1 and ID2, Figure 3.20a), extended to 5 nt, and then inserted into the location of previous ribonucleotide-containing segment and DNA control. After annealing of a 49 nucleotide long complement as per previously, the remainder of the overhang was synthesised, ligated to a hairpin, and then sequenced (Figure 3.20b).

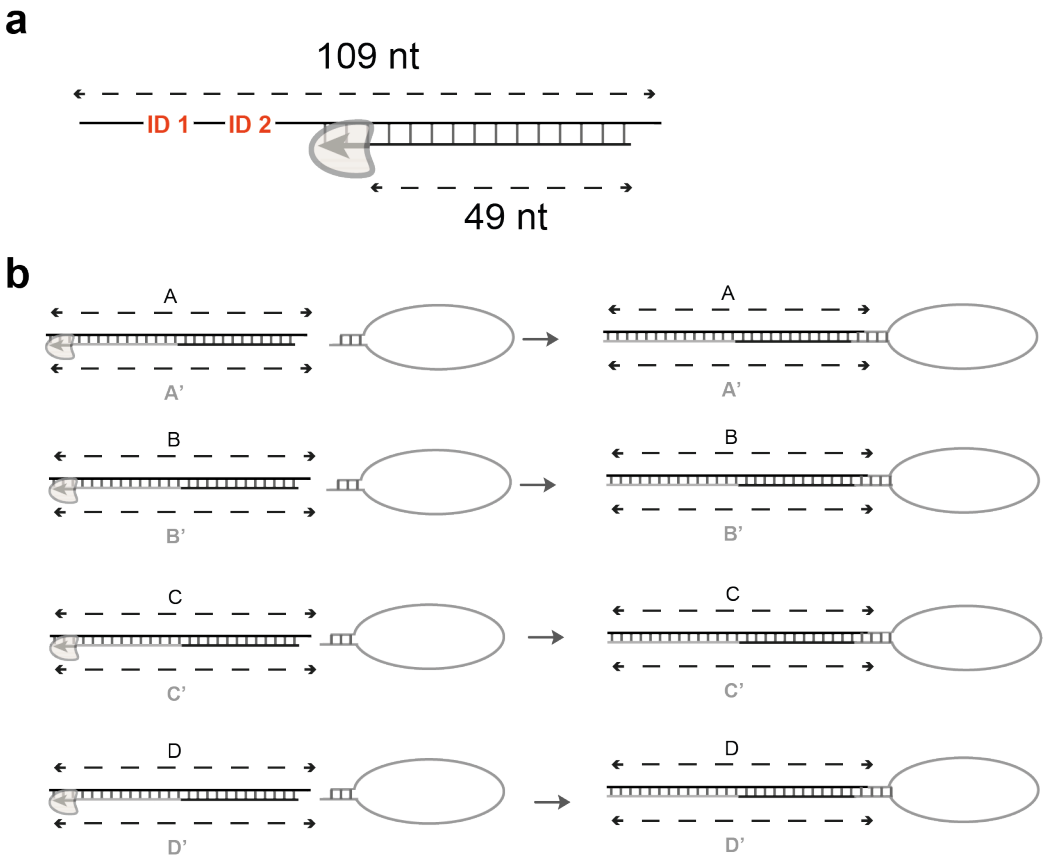


Figure 3.20. Experiment to test efficiency of hairpin formation. Template sequence. *The 109 nucleotide template contains 2 ID sequences, which distinguish oligonucleotides A-D. A 49 nucleotide complement is annealed and then extended using Klenow.* ***b) Extension****. Extension with Klenow over the unique IDs creates 4 unique duplexes: A-A', B-B',C-C',D-D', which are then ligated to a hairpin and sequenced.*

3.2.17 Testing the degree of hairpin self-complementarity

In a situation in which there was perfect hairpin self-complementarity, one would expect that the template strand sequence would always be followed by its complement. In this case, there are 4 possible template strand sequences (A-D). Therefore if this situation were correct one would expect close to 100% of sequences to match the reference A-hairpin-complement of A, or A-A' for short (Figure 3.21a). However, if there was a possibility of hairpins annealing to another hairpin, one would expect the sequencing of combinations of templates and complements of other templates, for example A-B', A-C' or A-D' (Figure 3.21a). Looking at A as a template, whilst products A-A' were the most common product sequenced after ligation to the hairpins, products A-B',

A-C’ and A-D’ were also generated at rates of between 14.65 and 16.38 % (Figure 3.21c ; Table 3.1). This suggests that around 1/3 of the time the hairpin bound to itself, whereas for the remaining 2/3 (66%) of annealing events the hairpin bound to an adjacent hairpin with a 1 in 4 chance (16.7%) of binding A, B, C or D (Figure 3.21b).

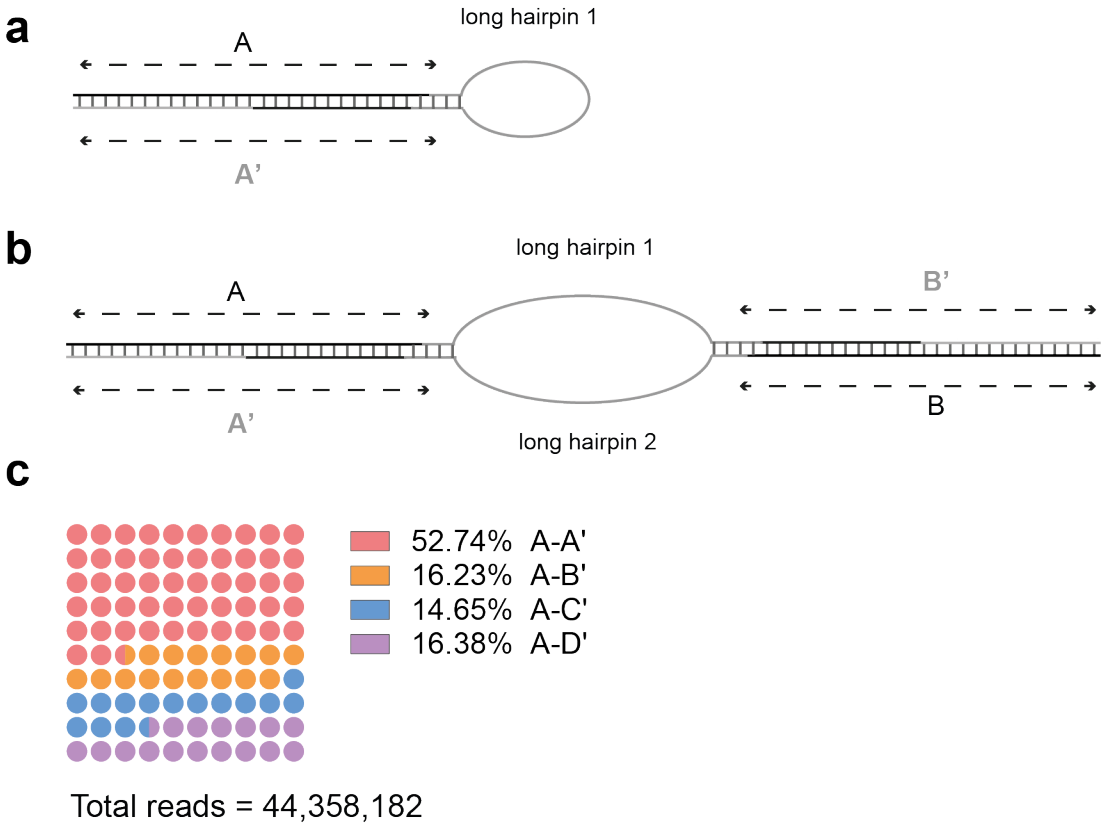


Figure 3.21. Testing of degree of hairpin self-complementarity. a) Schematic of experimental workflow. *A hairpin (labelled as hairpin 1) self anneals to itself, creating stem and overhang that binds template duplex.* ***b) Outcome of hairpin dimerization.*** *Two hairpins (1 and 2) dimerise, forming a duplex with central bubble that has overhangs that can bind to two template duplexes.* ***c) Results from sequencing of filled in duplexes and ligated hairpins.*** *Just over 50% of the reads are correct combination of template sequence and A and expected complement A’. However, the other possibilities A-B’, A-C’ and A-D’ are also present in equivalent numbers, suggesting that hairpins are dimerising and leading to the situation outlined in b).*

Table 3.1. Sequence outputs matching A reference template

Sequence	Number of reads	Percentage
A-A'	23,393,148	52.74%
A-B'	7,199,963	16.23%
A-C'	6,500,251	14.65%
A-D'	7,264,820	16.38%
Total	44,358,182	100.00 %

3.3 Discussion

3.3.1 Implications of the results for future work

The work presented in this chapter demonstrates proof of concept for the use of nanopore sequencing to detect genome embedded ribonucleotides. I demonstrate that, as predicted from work looking at methylated cytosines and BrdU, a single ribonucleotide embedded in a sequence of DNA has a distinctive amplitude signature, most marked for riboA (Figure 3.14). I also find that there is a distinctive kinetic signature for some ribonucleotides embedded in DNA, which I postulate to be due to an interaction between the motor protein used in nanopore sequencing and the ribonucleotide. The kinetic signature identified is most marked for a riboC (Figure 3.14). This opens up the possibility of using nanopore sequencing to identify single ribonucleotides or tracts of ribonucleotides in genomic DNA, and use of this approach to identify any non-canonical base embedded in DNA.

3.3.2 Creating model libraries for any non-canonical base

After the successful design of a system where hairpins reliably anneal to each other, the next step would be to create a library of oligonucleotides to allow the examination of the amplitude and kinetic signature for a much larger number of *k*-mers with an incorporated ribonucleotide. Here, one could design a system where the forward and reverse strand of the ribonucleotide containing duplex are linked using a unique barcode created by a stretch of “doped” DNA (Figure 3.22). If sufficient oligonucleotides were sequenced, it would be possible to start to build models not just for the amplitude signal for each ribonucleotide in a *k*-mer context, but also for the kinetic signature associated

with the nucleotides upstream of the ribonucleotide. In addition to creating model libraries for a single embedded ribonucleotides, it would also be possible to synthesise oligos with 2 or more adjacent ribonucleotides, to investigate whether tracts of ribonucleotides are present in genomic DNA. Finally, this system could be adapted to build model libraries for any non-canonical nucleotides that is covalently incorporated into genomic DNA.

Of note, since I designed and performed these experiments, the approach used for training base-calling by Oxford Nanopore Technologies has moved from a *k-mer* based system to one based on trained neural networks(Aria and Yeeles, 2019) . This means that the models I used to design the experiments and interpret the results are unlikely to be available anymore. However, the same principle of generating a wide repertoire of sequences incorporating an embedded ribonucleotide to train a bespoke model (e.g. using Bonito https://github.com/nanoporetech/bonito) is likely to hold true.

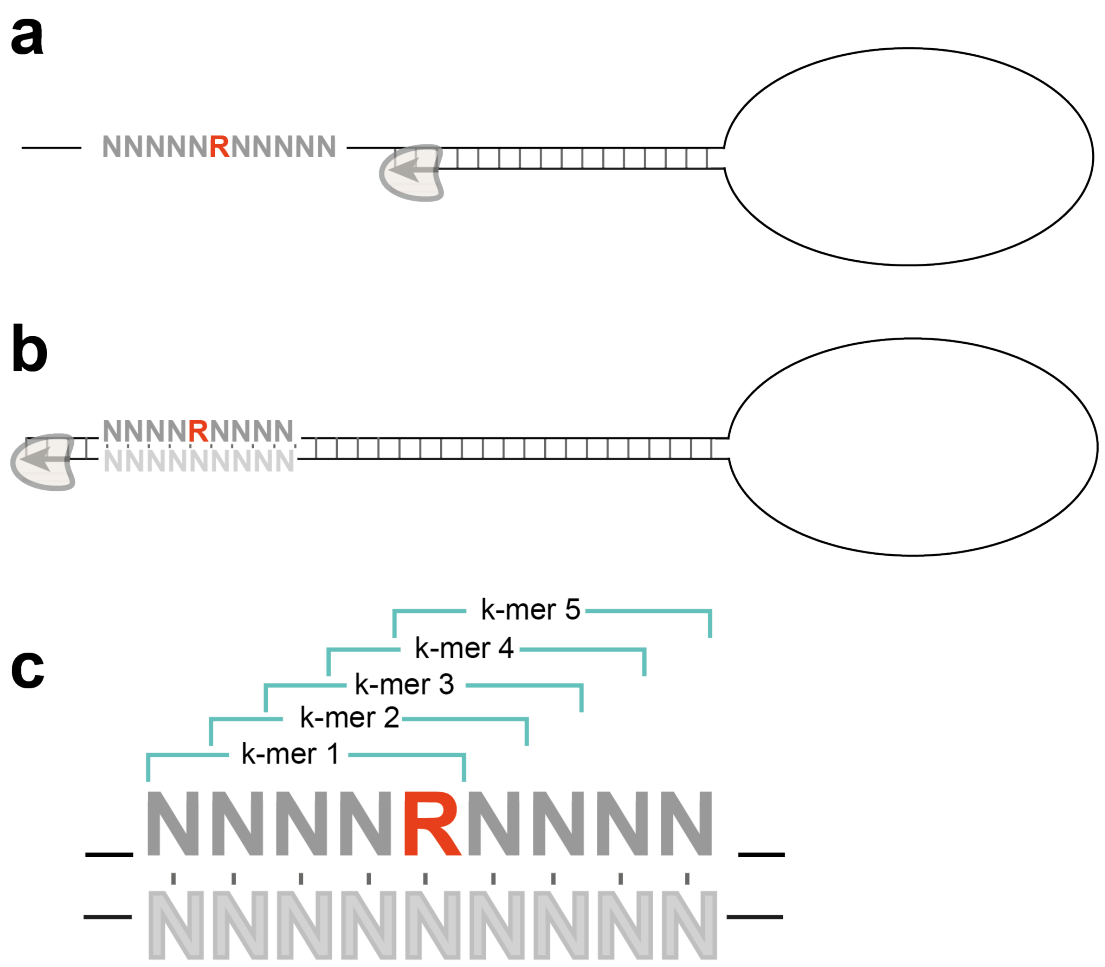


Figure 3.22. Schema for creating model libraries from hairpin and newly synthethised duplexes. a) Creating oligonucleotides for sequencing*. An oligonucleotide with a central ribonucleotide (or any other non-canonical nucleotide that can be synthethised across) is surrounded upstream by "doped" DNA nucleotides; in this case 4 up and downstream to constitute a 5-mer.* ***b) Final DNA duplex for sequencing.*** *The tract with the ribonucleotide is synthethised across, and the template, hair and complement sequenced.* ***c) Constructing model libraries.*** *A model library is constructed by matching the DNA k-mers for the complementary strand to the signal data from the template strand.*

3.3.3 Datasets that could be used to search for evidence of embedded ribonucleotides

The preliminary data outlined above suggests that there is a distinctive amplitude and kinetic signature that could be combined to identify the presence of ribonucleotides in genome DNA. I envisage four datasets that could be interrogated to look for biological confirmation of this signal. Even if a complete model library cannot be built to look for every possible combination of *k*-mer, it might be possible to use just a subset of *k*-mers with a known distinctive kinetic or amplitude signature. This distinctive signature (for

example, the marked kinetic signature for ArCG (Figure 3.14), could be searched for in genomic context to see if there is evidence of embedded ribonucleotides at this genomic location.

3.3.3.1 *S. cerevisiae* Pol ε mutant

As described in Chapter 1, *S. cerevisiae* strains deficient for RNase H2 and expressing a Pol ε mutation (POL2-M644G) with increased incorporation of ribonucleotides (Nick McElhinny, Kumar, *et al.*, 2010) have very high levels of RNA on the leading strand. Preparation of genomic DNA would be performed using the initial steps of the emRiboSeq protocol (Ding *et al.*, 2015), designed to prevent hydrolysis of embedded ribonucleotides by performing RNAse degradation of nucleic acids in a high NaCl concentration. Previous data suggests that in this strain one would expect an incorporation of a ribonucleotide roughly every 92 bases (Nick McElhinny, Kumar, *et al.*, 2010) with a preference for incorporation of Gs and Cs .

3.3.3.2 *Schizosaccharomyces pombe* mutant

The second dataset would be WGS of a *Schizosaccharomyces pombe* strain (courtesy of Olaf Nielsen, Copenhagen University), deleted for the mat 2 and mat 3 loci. This strain (EG530) (Nielsen and Egel, 1989) is expected to have a retained diribonucleotide on chromosome II, position 2133728 in every fourth cell (Vengrova and Dalgaard, 2006). Again, one could perform the start of the emRiboSeq protocol (Ding *et al.*, 2015) to preserve genome embedded ribonucleotides. The di-ribonucleotide in this situation however is a TT, and from our kinetic experiments C and G appear to most consistently have a kinetic signature. However it would be possible to identify the kinetic and amplitude signature for a double TT in the appropriate genomic context using the approach outlined above, and then search for this in whole genome sequencing data.

3.3.3.3 Human RPE1 RNase H2 mutant cell line

The third dataset is DNA isolated from an RPE1 cell line (Figure 5.6). Martin Reijns has extracted DNA, again using the first steps of the emRiboSeq protocol to preserve genome embedded ribonucleotides, and libraries from

p53 null (control) and p53 null/RNase H2 null have been sequenced on the PromethION nanopore platform. Although the frequency of genome embedded ribonucleotides in genomic DNA is likely to be low, from previous studies (Berglund *et al.*, 2017; Moss *et al.*, 2017) one would predict higher levels of RNA incorporation in mitochondrial DNA, which could be detectable on sequencing.

3.3.4 Linking the mutational signature associated with ribonucleotide misincorporation directly to rates of ribonucleotide incorporation

As outlined in Chapter 1, there is distinct mutational signature in *S. cerevisiae* which consists of an increase in short deletions associated with retained ribonucleotides. In Chapter 4 I will describe a sensitive and specific construct designed to detect 2 base pair equivalent mutations in *S. cerevisiae*; in Chapter 5 this reporter construct is carried into human cells. Nanopore sequencing, using an approach to target a specific sequence of DNA, such as bait-capture (Eckert *et al.*, 2016), would allow quantitative assessment of RNA misincorporation across the mutated sequence, and direct correlation of this misincorporation with mutation sequence context and rates.

Chapter 4 Mutational consequences of ribonucleotide incorporation in S. cerevisiae

4.1 Introduction

As discussed in Chapter 1, the presence of supra-physiological levels of ribonucleotides in the model organism *S. cerevisiae* has been shown to result in a distinctive mutational signature associated with an increased rate of short (2-5 base pair) deletions in directly repeated sequences. Mechanistic work conducted previously (N. Kim *et al.*, 2011) has shown that this is mediated by the action of the enzyme Top1. The aim of the work presented in this chapter was to recapitulate these findings in *S. cerevisiae* using an approach that could then be directly transferred to human cells. This would allow me to investigate whether a ribonucleotide associated, topoisomerase mediated mutagenesis is

also present in human genomes. In this introduction I will outline some of the model organisms used to study the process of mutagenesis, the theory underlying the fluctuation assay used to quantify mutation rates, and previous experiments conducted in *S. cerevisiae* to look at the mechanisms underlying short deletions in eukaryotic genomes.

4.1.1 Use of model organisms to study evolutionary processes

For a variety of reasons, including cost, speed, and the potential for genetic modification, a variety of non-human, so-called “model organisms” have been used to study the mechanisms of how cells function, and evolution more generally. These models have ranged from viral phages with faster replicating and smaller genomes, through to prokaryotic bacteria such as *E.coli*, the eukaryotic yeasts *S. cerevisiae* and *Schizosaccharomyces pombe* (S. pombe), the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, and small mammals such as mice (*Mus musculus*) and rats (*Rattus norvegicus domestica*). With the use of each organism there is an intrinsic trade-off between ease of use and relevance for human physiology and disease: in general the simpler the model, the further removed in evolutionary time from human biology. However, these models have proven powerful tools in understanding a variety of fundamental processes within eukaryotic cells, and, for the purpose of this thesis, the fundamental molecular process that shapes evolution: mutagenesis.

4.1.2 Principles of the fluctuation assay: Luria and Delbrück

As we currently understand it, evolution is shaped by the interplay of mutagenesis and selection. New mutation are produced and their frequency and ultimate survival in a population determined by stochastic sampling and selection. A question that was formulated in the 19th century is whether these mutations occur adaptively (in response to the conditions of a particular individual, increasing adaptation to a particular environment), or spontaneously, to be acted on by selection in subsequent generations.

Darwin's formulation (Darwin, 1859) was that natural selection acted on traits that already existed within a population; Lamarck posited that traits could develop and change within an organism's lifetime, and crucially be passed on to their offspring(Lamarck, 1809). Demonstrating the power of a simple model organisms to answer fundamental biological questions, one of the simplest available was used in the 1940s by Max Delbrück and Salvador Luria to answer this question.

Delbrück and Luria used a model system consisting of *E.coli* with lambda phage α. They noted that when *E.coli* was grown in the presence of the lambda phage, the cell culture solution would initially clear, then become turbid again as bacterial variants resistant to phage infection and lysis grew. This is because when *E.coli* without any resistance to the phage becomes infected this leads to bacterial cell lysis and clearing of the cell culture solution. However, with time mutant *E.coli* cells with resistance to the phage would arise, allowing subsequent descendants of these mutant *E.coli* to grow, and leading eventually to a turbid cell culture solution. The question was therefore whether the viruses by "direct action induced the mutants" (the mutation hypothesis), or whether bacteria which already had the mutation inducing resistance were selected for (hypothesis of acquired hereditary immunity) (Luria and Delbrück, 1943). Luria and Delbrück reasoned that if the hypothesis of acquired hereditary immunity were correct there would be a fixed small chance for each bacterium to survive an attack by the virus, and they would expect a binomial distribution in the number of resistant bacteria.

In conducting their experiment, they found that when they subsampled a culture to look for resistant clones, they got very variable, non-reproducible results, which they realised was due to the underlying mutational process rather than experimental variability. They state "large fluctuations are a necessary consequence of the mutation hypothesis and … the quantitative study of the fluctuations may serve to test the hypothesis."

They developed a mathematical formulation (now called the P_0 method) to extrapolate the mutation rate from the number of mutants found in a culture,

and the final number of cells in the culture. In their experiments they found highly variable numbers of mutants in each culture. There were many more cultures with no mutants than expected under the hypothesis of acquired hereditary immunity. With this hypothesis, if selective pressure (in the form of the phage) lead to immunity in a certain proportion of all cells, one would expect roughly similar numbers of mutants in all cultures of the same size. They also found higher numbers of cultures than expected with a very large number of mutants, in keeping with events early on the culture that led to a large number of descendants. They concluded that this lack of a binomial distribution was proof that resistance to the virus was due a heritable change which occurred independently of the action of the virus. In short, their findings support a Darwinian rather than Lamarckian view of evolution.

In a fluctuation assay, one aims to establish the parameter μ, which is the number of mutations per cell per generation (Luria and Delbrück, 1943). As outlined above, several cultures are grown in parallel, all inoculated with a small number of cells (N_0). At the end of a period of growth (with the easiest method growing the cells until saturation), all cells from the culture are plated onto a selective medium. One can then determine the number of mutant cells by counting the number of colonies which survive on selective medium: this is the variable *r*. The total number of viable cells (N_t) is calculated by diluting the final culture and growing this on a non-selective plate, counting the number of colonies that grow, and multiplying by the dilution factor to determine the total number in the final culture. To calculate a mutation rate, *r* (the number of mutants counted) must be converted to *m* (the total number of mutations that have occurred in the culture). These are not the same value, because if a mutation happened early on in the history of a culture then there will be many more clones of that mutant present than if the mutation had happened late on in the history of the culture.

A variety of statistical methods have been developed to calculate *m* from *r*, and thus μ. These include the original P_0 method described by Luria and Delbrück (Luria and Delbrück, 1943), the Lea Coulson method of the median (Lea and

Coulson, 1949), the Drake formula (Drake, 1991) and the Ma Sandri Sarkar (MSS) Maximal Likelihood Method (Ma, Sandri and Sarkar, 1992). All these methods, by generating a value of mutation rate per cell division, allow comparison of rates between different experimental systems, and between different mutational processes. Fluctuation assays have been used to investigate mutation rates for the normally rare mutational events that become more common when key DNA repair pathways in the cell such as mismatch repair or, as discussed next, ribonucleotide excision repair, are lost. One of the key model systems in this field has been *S. cerevisiae* (although μ can also be applied to the study of evolution in human tissues) (Werner *et al.*, 2020).

4.1.3 *S. cerevisiae* as a model system for DNA replication and mutagenesis

Although humans and the budding yeast *S. cerevisiae* are estimated to have diverged over 1 billion years ago (Douzery *et al.*, 2004), they have retained many similarities. One third of yeast genes are estimated to have human orthologs (O'Brien, Remm and Sonnhammer, 2005), and a significant proportion of these remain so well conserved that the human gene can be experimentally introduced into the *S. cerevisiae* genome and replace the function of the yeast ortholog (Kachroo *et al.*, 2015). This degree of conservation, in association with its well annotated genome and short life cycle, has meant that it has been a widely used model organism to study cell biology mechanisms, human diseases (Bassett, Boguski and Hieter, 1996) and the process of mutagenesis.

4.1.4 Mutation reporter constructs used previously in *S. cerevisiae*

Previous assays to detect mutations in a *S. cerevisiae* model system have involved the use of a variety of reporter genes including URA3, CAN1 and LYS2. In general, these consist of genes that are toxic to cells if grown in a particular medium. For the URA3 assay, for example, if a yeast strain is grown in a solution containing 5-FOA (5-Fluoroorotic Acid), the product of the URA3 gene Orotidine 5'-phosphate decarboxylase (ODCase) will produce the toxic

intermediate 5- fluorouracil which leads to cell death. However, an inactivating mutation in the URA3 gene (such as a non-synonymous SNP or frameshift) means that a clone that contains that mutation will not produce 5-fluorouracil and will not be sensitive to 5-FOA, and therefore these mutants can be selected for by growth on 5-FOA containing growth medium.

As outlined in Chapter 1, these assays have been used to effectively elucidate the effects of RNase H2 and Top1 deficiency in the generation of short indels in *S. cerevisiae*. However, for a number of reasons they were not suitable for transfer to a human experimental system. The genes used in these assays are not relevant for human cells: there is no exact equivalent of a URA3 or a CAN1 gene in mammals, although the HPRT gene has been used in a similar way to yeast for positive and negative selection in mammalian cells (Glaab and Tindall, 1997) . Often the reporter sequence that the assay relies on is very short, for example just 150 bp in the Lys frameshift assay (Lehner and Jinks-Robertson, 2009), which means that a large number of cell divisions are needed to capture these often very rare events. Additionally, many of the assays are based on negative selection using growth media, where neither the genes nor the media are directly comparable/transferrable between *S. cerevisiae* and mammalian cells.

I also knew that to be able to demonstrate a ribonucleotide dependent signature in human cells I would need a highly sensitive assay, as the number of human cells that can be grown in culture over several weeks is far less than can be achieved in yeast cell culture overnight. I therefore set out to establish previous estimates of absolute mutation rate (mutation rate per base pair per cell division) for ribonucleotide/Top1 dependent mutagenesis in *S. cerevisiae*. Once I had estimates of these, this would permit a judgement of whether it might be feasible to establish a reporter construct that could be validated in yeast, and then moved into a human system.

4.2 Results

4.2.1 Mutational signature associated with RNase H2 deletion in *S. cerevisiae*

To determine an estimate of the absolute mutation rate of ribonucleotide associated mutagenesis in *S. cerevisiae*, I systematically searched the published literature using a standardised approach. The details for the search strategy are outlined below.

4.2.1.1 Inclusion criteria and search strategy

I searched for studies in *S. cerevisiae* that had introduced a mutation into RNase H2 or replicative polymerase that results in supra-physiological incorporation of ribonucleotides in DNA, and used a wild type S. cerevisiae or equivalent strain without that mutation. I included studies that quantitatively summarised the mutation patterns seen in the mutant strain compared to the control. I searched for studies in PubMed and Embase, using two sets of search terms: "*S. cerevisiae* AND Mutation* OR Mutagenesis AND RNase*", and "*S. cerevisiae* AND polymerase* AND mutation* OR mutagenesis AND ribonucleotide*"

4.2.1.2 Results from search

286 unique studies were retrieved using the search terms in November 2016, 36 papers were screened on basis of title and abstract, and 18 met the inclusion criteria. A further 2 papers were identified from the references of included studies, giving a total of 20 studies which met the inclusion criteria.

All the studies had knocked out subunits of RNase H2. All except two papers used gene reporter constructs: CAN1, HOM3, URA3, TRP5, and variations on a LYS reporter construct. All papers used fluctuation assays to calculate a mutation rate. Additionally, one paper used whole genome sequencing (on a pol2-M644G strain), and the other employed a Yeast Artificial Chromosome (YAC) to measure loss of heterozogosity.

4.2.1.3 Increased mutation rates in RNase H2 deficient *S. cerevisiae* strains

11 studies examined the effects of knocking out RNase H2 in isolation (Table 4 .1). All of these used fluctuation assays to calculate mutation rate. As described in Chapter 1, RNase H2 in yeast has 3 subunits (rnh201, rnh202 and rnh203); the loss of any of these subunits leads to loss of RNase H2 activity. In these 11 studies, in general knocking out one of the subunits of RNase H2 led to a moderate increase in overall mutation rate, with a median ratio between the mutation rate in the knockout and control of 1.6 (IQR 1.44-2.45). Seven papers examined the proportion of mutations of a particular type. To do so they sequenced the reporter gene for a number of mutant colonies. All that performed sequencing reported an increased proportion of 2-5 bp deletions, although this effect was more marked in some studies than others: in one with the smallest increase the proportion of mutations that were short deletions increase from 0 to 0.05 (Allen-Soltero *et al.*, 2014), whilst in the one with the largest increase this rose from 0 to 0.83 (Chen *et al.*, 2000).

Table 4.1. Mutation rates in RNase H2 knockouts with WT polymerases

Reference	RNase H2 subunit deleted	Relative increase in mutation rate in RNase H2 KO[1]	SNPs control [2]	SNPs RNase H2 KO[3]	2-5 bp deletions control [4]	2-5 bp deletions RNase H2 KO [5]
Qiu et al., 1999	rnh201	0.9-6.1	n.d.[6]	n.d.	n.d.	n.d.
Chen et al., 2000	rnh201	5.07	n.d.	n.d.	0	0.83
Nick McElhinny et al., 2010	rnh201	1.1 -2.1	n.d.	n.d.	n.d.	n.d.
Clark et al., 2011	rnh201	1.1-1.6	0.57	0.32	0.01	0.35
N. Kim et al., 2011	rnh201	2.6	0.83	0.40	0.02	0.40
Shen et al., 2012	rnh201	2.5	n.d.	n.d.	n.d.	n.d.
Kim et al., 2013 (low transcription of reporter construct)	rnh201	1.6- 2.2	n.d.	n.d.	0 - 0.06	0.28 – 0.81
Kim et al., 2013 (high transcription of reporter construct)	rnh201	3.9 -9.1	n.d.	n.d.	0.01 – 0.14	0.35- 0.54
Ghodgaonkar et al., 2013	rnh201	1.48-1.62	n.d.	n.d.	n.d.	n.d.
Allen-Soltero et al., 2014	rnh201, rnh203	0.6 -2.3	n.d.	n.d.	0 (2 bp deletions)	0.05 (2 bp deletions)

[1] Ratio of mutation rate in RNase H2 knockout strain compared to control. If a range is presented, this represents the reporter construct in different contexts, often its position relative to an origin of replication
[2] Proportion of sequenced mutants in the control strain that contained a single nucleotide variant compared to the reference sequence
[3] Proportion of sequenced mutants in the RNase H2 knockout strain that contained a single nucleotide variant compared to the reference sequence
[4] Proportion of sequenced mutants in the control strain that contained a 2-5 base pair deletion compared to the reference sequence
[5] Proportion of sequenced mutants in the RNase H2 knockout strain that contained a 2-5 base pair deletion compared to the reference sequence
[6] not done (n.d)

Potenski et al., 2014	rnh202	1.5	1.0	0.39	0	0.31 1 bp indels, 0.30 ≥ 2 bp indels
Williams et al., 2015	rnh201	1.75	0.57	0.33	0.01	0.35

Table 4.2. Mutation rates in pol2-M644G mutants[7]

Reference	**RNase H2 subunit deleted**	**Relative increase in mutation rate in pol2 -M644G with RNase H2 KO** [8]	**SNPs control**	**SNPs RNase H2 KO**	**2-5 bp deletions control**	**2-5 bp deletions RNase H2 KO**
Nick McElhinny et al., 2010	rnh201	17-31	n.d.	n.d.	n.d.	n.d.
Lujan et al., 2011	rnh201	n.d	0.35-0.40	n.d.	0.02- 0.05	n.d.
Chon et al., 2013	rnh201	3.9	0.71	0.30	0.05	0.55
Lujan et al., 2013	rnh201	4.5-16	n.d.	n.d.	n.d.	n.d.
Conover et al., 2015	rnh201	2.1	n.d.	n.d.	n.d.	n.d.
Williams et al., 2015	rnh201	12	0.40	0.06	0.03	0.89

[7] See Table 4 .1 for explanation of column headers

[8] Compared to control strains with pol2-M644G mutation but wild type RNase H2

4.1.1.1 Further increases in mutation rates when RNase H2 knocked out in combination with mutant polymerases

Six studies report knocking out RNase H2 in combination with the pol2 – M644G mutant. Pol2-M644G is a Pol ε strain with a steric gate mutation that leads to a dramatic increase in the misincorporation of ribonucleotide (up to 1 in 92 bases, as outlined in Chapters 1 and 3). This increase in misincorporation is because the active site is less able to discriminate deoxyribose NTP from the corresponding ribose NTP, which are generally far more abundant in cells (Nick McElhinny, Kumar, *et al.*, 2010). The studies identified in my review found a marked increase in mutation rates compared to control (i.e. a pol2-M644G strain with normal RNase H2 activity). The median relative increase was 17-fold (IQR 8.3-24,Table 4.2). All of these apart from the study by Conover *et al* (Conover *et al.*, 2015) used fluctuation assays to calculate mutation rates; the Conover study used whole genome sequencing. All the studies that reported a proportion of 2-5 bp deletions found this to be increased in the RNase H2 strains relative to those which were RNase H2 proficient. Of note, the presence of the M644G mutant in itself did not seem to be sufficient to lead to a change in the mutation profile in the presence of physiological levels of RNase H2. One study (Williams *et al.*, 2015) examined *S. cerevisiae* strains with pol1 (α) and pol3 (δ) mutants that like the pol2-M644G mutant incorporate ribonucleotides at greater frequency. Pol1 (α) as discussed in Chapter 1 synthesises Okazaki primers and the start of the lagging strand, activity that is continued by pol3 (δ). The mutants used in this experiment incorporate high rates of ribonucleotides, but not as high as the 1 in 92 incorporated by the pol2-M644G mutant strain (Nick McElhinny, Kumar, *et al.*, 2010). The experiments found an increase in the overall mutation rate that was smaller than that seen for the Pol ε mutants (a 1.1 and 1.2 fold increase for the α and δ mutants, respectively). This could be because these two polymerase mutants have lower relative rates of ribonucleotide incorporation, or because removal of ribonucleotides by Top1 is leading, rather than lagging strand

specific (Williams *et al.*, 2015), in the absence of RNase H2. Therefore, Pol ε, which is the predominant leading strand polymerase, would if mutated lead to greater incorporation of ribonucleotides on this strand, and therefore a higher mutation rate due to greater Top1 activity on the leading strand. Conversely, ribonucleotide incorporation by Pol α and δ mutants would be predicted to lead to less Top1 activity due to fewer incorporated ribonucleotides and therefore a lower mutation rate.

In summary, loss of RNase H2 by itself was associated with a moderate increase in the overall mutation rate in this dataset, which consisted mainly of studies in reporter constructs. In addition to a moderate increase in the overall mutation rate (a median of 1.6-fold), supra-physiological with levels of ribonucleotides were associated with a change in the mutational spectra from SNPs being the dominant mutation type towards an increased prominent of 2-5 bp mutations. A Pol ε mutant strain (pol2-M644G) that increased the levels of misincorporated ribonucleotides in an RNase H2 null background led to a more dramatic rise in the relative mutation rate compared to strains with wild type polymerases and mutated polymerases with intact RNase H2.

4.2.1.4 Sequence context of short deletions in RNase H2 null *S. cerevisiae* reporter constructs

As discussed in Chapter 1, a factor that appeared to influence the likelihood of short deletions was the sequence context, which I examined further. From the twenty studies identified in the systematic review, four documented the nucleotide context and content of the short indels found in RNase H2 null *S. cerevisiae* strains with a wild type polymerase. These used three genes to detect mutations: URA3 (Clark *et al.*, 2011), CAN1 (N. Kim *et al.*, 2011; Potenski *et al.*, 2014) and LYS2 (Kim *et al.*, 2013). A total of 299 2-5 bp deletions were documented by the studies. Of these 247 (86.0%) occurred in perfect tandem repeats, for example an AG deletion in an AGAG repeat. Two base pair deletions constituted 78.9% (236) of the total, and of these 200

(66.9% of the total) occurred in perfect tandem repeats. Tandem repeats here were defined as a repetitive sequence with at least 2 units, each unit being 2 bp in length. Two base pair deletions could be further classified based on the nucleotide context of the deletion (Table 4 .3). The most common context for these 2 bp deletions was a CT-CA deletion (82.6% of all deletions, Figure 4.1).

Table 4.3. Classification of sequence context of 2 bp deletions. *Repeat context is shown in bold, and deleted segment in red, underlined.*

Category	Deletions subsumed	Example (deletion type)
AA-AT	AA,TT,AT,TA	**AT**AT**CC** (perfect repeat)
CT-CA	CT,TC,AC,CA,GA,AG,GT,TG	**AG**AG**TT** (perfect repeat)
CC-CG	CC,GG,CG,GC	**TT**CG**TT** (not in repeat)

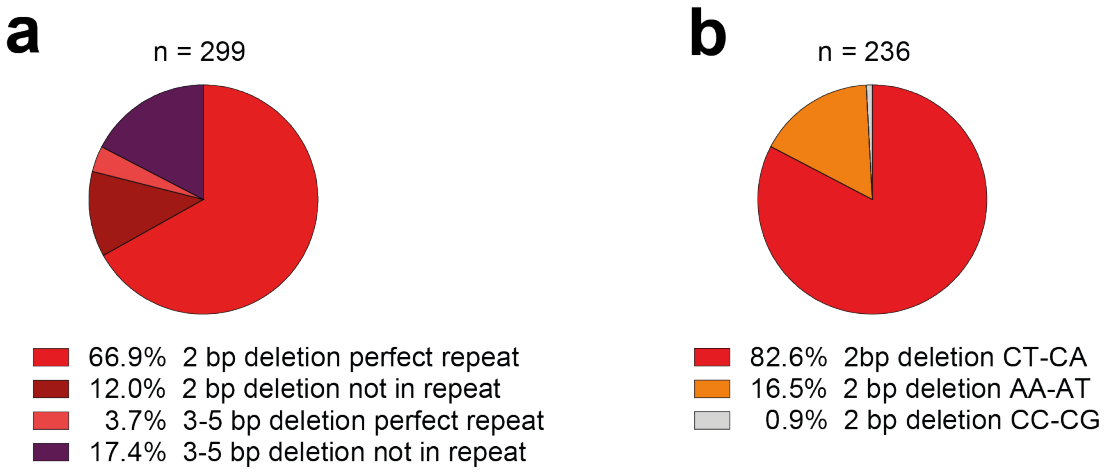


Figure 4.1. Repeat and sequence context of 2-5 base pair deletions in previous reporter constructs. a) The most common type of 2-5 bp deletion is a 2 base pair deletion in a perfect repeat context. b) The most common nucleotide context for a 2 bp deletion is CT-CA (see Table 4.3 for definitions of each mutation type).

Given that 84.7% of the 2 bp deletions occurred in perfect tandem repeats, I concluded that the context in which short deletions were most likely to happen were a CT-TG. Going forward, I refer to these as dinucleotide couplet repeats, as opposed to mononucleotide repeats, which were defined as consisting of runs of a single nucleotide longer than ≥4 nucleotides in length.

4.2.2 Sequence context of short deletions in RNase H2 null *S. cerevisiae* whole genome sequencing

One of the studies identified in the systematic review provided data on whole genome sequencing from a mutation accumulation experiment in a *S. cerevisiae* pol2-M644G RNase H2 null cell line (pol2-M644G *rnh201*∆) (Conover *et al.*, 2015). Here, samples are taken from a starting population of cells grown from a single progenitor, at least one sample is subject to whole genome sequencing (WGS) and at least one other sample is serially passaged so that it grows through a defined number of cellular generations, after which it is subject to WGS. Comparison of the WGS data then reveals the new mutations accumulated through subsequent cell generations as well as the mutation rate. I examined data from this experiment to gain further information on sequence context for ribonucleotide dependant mutations, and to make a decision on whether it would be worth conducting my own mutational accumulation experiment in an RNase H2 null *S. cerevisiae* strain with WT polymerases, the situation most comparable to that for mammalian cells.

A control for this strain (pol2-M644G with WT RNase H2, pol2-M644G RNH201) had been published in a previous paper (Lujan *et al.*, 2014) , and I compared the mutational spectra for both mutation accumulation experiments, alongside those from an entirely wild type control (see Methods for details on analysis of raw read data). The data from these experiments should be directly comparable as they were conducted using the same strains, varying only in the presence or absence of RNase H2, and grown under standardised conditions for a defined number of generations. Despite this, there remains the possibility of errors from batch effects, related to media or changes in other experimental set-up such as incubators.

Experiments for all 3 genotypes (WT, pol2-M644G RNH201 and pol2-M64G *rnh201*∆) consisted of two ancestral populations and a number of descendant

clones. For the purposes of analysis, new mutations in the descendent clones were aggregated for each genotype. The entirely wild type control had very low mutational rates, with 30 mutations, all SNPs, across a cumulative total of 7200 generations. The pol2-M644G RNH201 had higher mutation rate, with 414 mutations (again across 7200 generations). This is to be expected, as the mutation that widens the steric gate of the polymerase, allowing ribonucleotides to be misincorporated, also results in an increased frequency of mispaired deoxyribonucleotides (Nick McElhinny, Kumar, *et al.*, 2010). Almost all of these mutations were SNPs, with only 12 indels, and no 2-5 bp deletions (Table 4.4).My analysis of the pol2-M644G *rnh201*Δ mutation accumulation experiments (6,300 generations), now published by the Kunkel Lab (Williams *et al.*, 2019) uncovered an issue with a large number of shared terminal event mutations- that is, mutations called in cell lines that were meant to have been grown independently from one another. I found that over half (439 out of 768, 57.2%) of mutations were shared by two or more presumed to be independent clones. There is a very low likelihood of an identical mutation having occurred at the same location in a 12 million base pair genome, not just once but 439 times. Therefore I included these mutations in the analysis but treated them as a unique event, reducing the total number of mutations from 768 to 538. This total of 538 compares to a figure of 593 in the published paper (Williams *et al.*, 2019), which is likely due to my more conservative filters during mutation calling (see Chapter 2, Section). There was a striking increase in the rates of 2-5 bp deletions (125 mutations, 23.2% of the total accumulated mutations) compared to none in the pol2-M644G RNH201 control (Table 4.4, Figure 4.2a/b).

Table 4.4. Mutations in a pol2-M644G rnh201Δ mutation accumulation experiment from the Kunkel lab.

Type of mutation	pol2-M644G RNH201		pol2-M644G *rnh201Δ*	
	Number of events	**Percentage of total**	**Number of events**	**Percentage of total**
SNP	402	97.1	394	73.2
1 bp insertion	4	1.0	5	0.9
1 bp deletion	4	1.0	11	2.0
2 5 bp insertion	1	0.2	1	0.9
2-5 bp deletion	0	0	125	23.2
> 5 bp insertion	2	0.5	0	0
> 5 bp deletion	1	0.2	2	0.9
Total	414	100	538	100.0

Similar to the pattern seen in the reporter constructs, the most common mutation type was a 2 base pair deletion in a perfect repeat context (Figure 4.2c), followed by a 3-5 bp deletion in a perfect repeat context. Examining the sequence context of the deletions, of the 125 2 bp deletions (Figure 4.2d), 92 (73.6%) of the deletions were 2 bp in length, and the 2 bp deletions were again predominantly in a CT-CA context (95.7% of all 2 bp deletions).

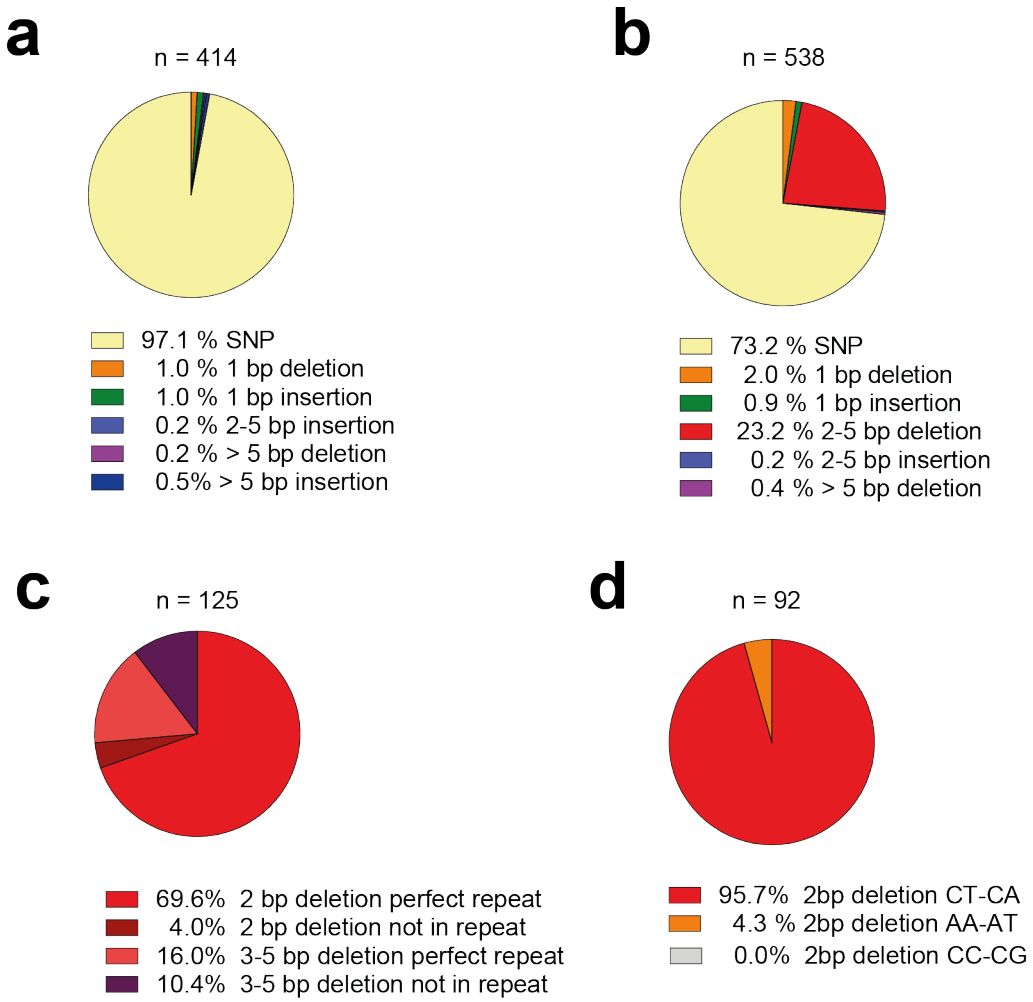


Figure 4.2. Deletion of RNase H2 in a pol2-M644G S. cerevisiae strain results in an increase in 2-5 base pair deletions enriched in a CA-CT context. a) Whole genome sequencing (WGS) of a mutation accumulation experiment in a pol2-M644G RNH201 strain shows predominant mutations are SNPs. b)An equivalent mutational accumulation in a pol2-M644G rnh201Δ (loss of RNase H2) shows an increase in 2-5 base pair deletions (23.2%). c) The most common category of 2-5 base pair deletion in the pol2-M644G rnh201Δ strain is a 2 base pair deletion in a perfect repeat (69.6%). d)The majority of 2 base pair deletions are equivalent to loss of CA-CT (see Table 4.3 for explanation of classification). *My analysis of data published in (Lujan et al., 2014; Conover et al., 2015); see Materials and Methods for details.*

4.2.2.1 Absolute mutation rates

In addition to establishing the relative increase in mutation rate, and the sequence context for short deletions, in RNase H2 null *S. cerevisiae*, a subset of the studies permitted calculation of the absolute mutation rate. These studies were those that provided the total number of mutations and details about the sequence length of the reporter construct used. The figure for the absolute mutation rate (μ_{BP}) is calculated using the following formula:

$$\mu BP = \frac{\mu 1}{lr}$$

where $\mu 1$ represents the mutation rate for the construct, and l*r* the length of the reporter construct, returns a rate of expected number of mutations per base pair per generation. Of course, the rate calculated using this method can only be regarded as a minimum rate, as there are likely to be mutations (synonymous substitutions, or indels that do not result in a frameshift) that do not affect protein function and therefore result in detectable mutants. The whole genome sequencing mutation accumulation experiments also permit the calculation of an equivalent rate:

$$\mu BP = \frac{total\ mutations}{total\ generations\ x\ length\ genome}$$

Here the total number of mutations in the experiment is divided by the number of generations of cell growth and the length of the mappable genome. At the time of performing these calculation, an absolute mutation rate for a mutation accumulation experiment in an RNase H2 null cell line was not available, so I

inferred a mutation rate by multiplying the median increase in mutation rate seen in my systematic review by the mutation rate[9].

Table 4.5. Absolute mutation rates in RNase H2 null* S. cerevisiae *experiments.

Study (reporter)	μ_{BP} POL RNH201	μ_{BP} POL *rnh201*Δ
Allen-Soltero et al., 2014 (CAN1)	3.4×10^{-10}	2.0×10^{-10}
N. Kim et al., 2011(CAN1)	5.6×10^{-11}	1.5×10^{-10}
Potenski et al., 2014 (CAN1)	5.6×10^{-11}	8.5×10^{-11}
Clark et al., 2011(URA3)	1.1×10^{-11}	2.0×10^{-11}
Lujan et al., 2014 (WGS)	1.7×10^{-10}	2.7×10^{-10} (inferred)

Making the assumption that the ribonucleotide associated mutation rate in *S. cerevisiae* was equivalent per base pair to that in human genomes, I calculated how many cell divisions would be needed to generate a statistically significant difference between a wild type and RNase H2 cell lines. I performed a power calculation based on a 1000 base pair long reporter construct, using the absolute rates from the WGS in *S. cerevisiae*. I estimated that during each cell division in a WT human cell there would be a 1.7^{-7} ($1.7x10^{-10}$ *1000) chance of a mutation occurring in the hypothetical construct, and a 2.7^{-7} ($2.7x10^{-10}$ *1000) chance of a mutation occurring in the hypothetical construct in an RNase H2 null background (Table 4.5). Using an α value of 0.05, and a prespecified power of 0.80, a power calculation in R (power.prep.test) showed that 345,350,631 cell divisions, or 41 flasks of T75 cells, would be needed in each group to demonstrate a significant difference in the mutation rates. This demonstrated the importance of designing a highly sensitive assay if there was any chance of a reporter construct detecting a ribonucleotide associated mutational signature in human cells. However, if one focused on 2-5 bp deletions only, assumed a proportion of 0.05 in wild type cells, and a ten-fold

[9] In the intervening period this experiment has been performed and published by the Kunkel Lab (Williams et al., 2019). Here, the estimated overall mutation rate was 3.1 times that of wild-type, for an absolute μ_{BP} of 6.3 x 10-10 (as compared to my prediction of 2.7x 10-10).

increase in this in RNase H2 null cells (the median of the proportion seen in the reporter constructs of 2-5 base pair deletion was 0.54), a lower number of cells (70,384,736; ~ 8 T75 flasks) would be needed.

4.2.3 Implications of results of literature review for experimental planning

These preliminary analyses established four key points for further experimental work. Firstly, they confirmed that a mutational signature of short, 2 to 5 base pair deletions was a distinctive feature of *S. cerevisiae* strains that fail to remove genome embedded ribonucletides.

Second, in absolute terms the overall rate for this mutational signature was low. It seemed probable that undertaking a whole genome sequencing mutation accumulation experiment in RNase H2 null *S. cerevisiae* wild type polymerases would yield few mutations and be unlikely to inform further work in human cells. This prediction was confirmed in a recently published paper (Williams *et al.*, 2019), which showed 81 mutations over 6300 generations in an RNase H2 null *S. cerevisiae* strain, of which only 12 were deletions > 1 bp in length).

Thirdly, due to this low absolute mutation rate, in order to demonstrate an increase in mutation rate in human cells, a very sensitive reporter would be needed. This would ideally be one specific for short deletions, which would increase the power of detecting a difference between a control and an experimental strain with increased levels of genome-embedded ribonucleotides. Finally, in keeping with the model put forward by Kim et al (N. Kim *et al.*, 2011), the short deletions seen in this mutational signature are enriched in repeat sequences. Therefore an experimental system would be more likely to detect this mutational signature if it was enriched in some way for repetitive sequences.

We therefore decided to design a reporter construct, specific for short deletions, and also highly sensitive, that would allow us to recapitulate previous findings in yeast. We designed a system that would be directly transferrable from a yeast to a human experimental model, as the overall goal of this research was to identify a mutational signature in human cells. Results in yeast would allow an estimation of the likely chances of detecting rare events for similar experiments in human cells.

4.1.2 Design of a reporter construct designed specifically to detect 2 bp equivalent deletions

4.2.3.1 Principles of the reporter construct

As described the majority of previous reporter constructs (e.g.URA3, CAN1) rely on inactivating mutations that affect gene function and can be used to select mutants. However, these systems are not specific for the mutational signature that was of most interest to us: short (2 to 5 bp) deletions. They are not specific because any number of mutations, including point substitutions, could inactivate the gene, and even those reporter constructs with the highest rates of 2-5 bp deletions (Table 4.1,Table 4.2) also have a high proportion of SNPs. They are not as sensitive as they could be (in the context of identifying 2-5 bp deletions) as short deletions in the 3' of the coding sequence may be tolerated with retention of a normal phenotype, and because of the relatively low frequency of tandem repeats in their coding sequence. We therefore set out to design a system that was both sensitive and specific for the detection of short deletions associated with the failure to remove genome embedded ribonucleotides. Based on the data analysis above, we focused on 2 bp deletions, the most common mutation type identified in both reporter constructs and the available whole genome sequencing data. As we wanted the system to be transferable between *S. cerevisiae* and human cells, we chose hygromycin B as the antibiotic with which to introduce our construct into cells. Hygromycin B is an aminoglycoside antibiotic that kills prokaryotic and

eukaryotic cells by inhibiting protein synthesis (Rao *et al.*, 1983). A gene conferring resistance to Hygromycin (Hygromycin-B 4-O-kinase) was initially found in *E.coli*¸ and the essentially identical gene hph is now commonly used in cloning studies to confer resistance to the antibiotic. Given that we knew deletion frequencies were increased in repeat sequences, we leveraged redundancy in the amino acid code to increase the proportion of repeats in the gene (see next section).

We designed the reporter construct (Figure 4.3a) so that the hygromycin resistance gene ($Hygro^R$) served both to insert the construct into *S. cerevisiae*/human cells, and as a sequence to detect mutations once integrated into the host genome. At the start of a mutation accumulation experiment it is important to know that there are no frameshift mutations in the $Hygro^R$ gene, and therefore strains used in all experiments were grown in Hygromycin B up until the experimental start point to ensure this was the case. We chose to introduce the construct at the AGP1 locus on Chromosome III (Figure 4.3b), a location used previously by the Kunkel Lab (Nick McElhinny *et al.*, 2008) for the introduction of the URA3 reporter construct. The AGP1 locus is just downstream of the high frequency replication origin ARS306, and codes for a non-essential gene that can be removed without impacting on cell growth, probably due to redundancy with the very similar gene GNP1 (Schreve, Sin and Garrett, 1998). The construct was introduced at this locus by homologous integration using the lithium acetate method (Gietz *et al.*, 1995).

When designing the construct, we introduced a +2 base pair frameshift immediately downstream of the stop codon at the end of the $Hygro^R$ gene sequence, followed by a sequence for porcine teschovirus-1 2A peptide (P2A) (J. H. Kim *et al.*, 2011). P2A is a short peptide sequence which when translated in frame leads to the creation of two separate proteins from the same transcript. P2As work by leading to the skipping of ribosomal synthesis of a

glycyl-prolyl peptide bond at the downstream C-terminus of the peptide, leading to an upstream peptide sequence terminating with the majority of the amino acids from the P2A, and a downstream peptide starting with the terminal proline from the P2A. Downstream of the P2A sequence we inserted the coding sequence for the Neomycin/G418 resistance gene (Wang, Wang and Silva, 1996), which could be used to select for mutants with an upstream frame-shifting mutation in the $Hygro^R$ once integrated into the host genome. The bacterial neomycin resistance gene (Neo^R) confers resistance to the aminoglycoside antibiotic neomycin and also the closely related G418 (also known as geneticin). These aminoglycoside antibiotics act by blocking peptide synthesis; the Neo^R gene inhibits them by phosphorylation of the antibiotic (Thompson *et al.*, 2002).

Once the construct is inserted into host cells, without any mutations those cells will still be susceptible to Neomycin/G418 as Neo^R is in the wrong reading frame for translation. However, mutations in the $Hygro^R$ gene sequence which lead to a 2 base pair deletion, or an equivalent frameshift, will eliminate the stop codon and bring the P2A into frame. This will result in translation of the P2A and Neo^R, with subsequent cleavage of the NeoR from the HygroR peptide sequence, and a functional NeoR gene product. Mutant cells would now be resistant to Neomycin/G418, and thus could be selected for by growth in one of these antibiotics. The mutation that lead to a gain in resistance could be a 2 bp deletion, or another insertion or deletion that results in a -2 shift of the reading frame before the P2A sequence, such as a 5bp deletion or 1 bp insertion.

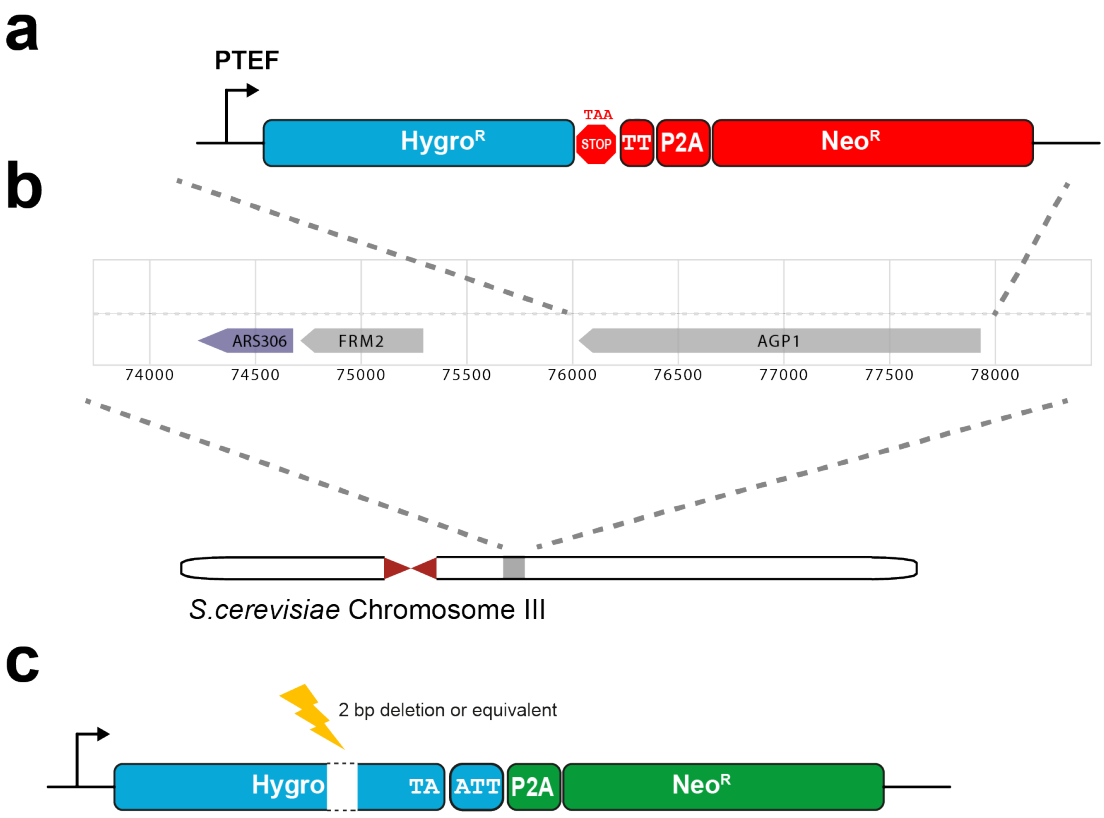


Figure 4.3. Design of a detection reporter construct for the detection of 2 base pair deletions. a) The construct consists of 2 antibiotic resistance genes separated by a frameshift and a self cleaving peptide. *The hygromycin resistance gene (HygroR) is used to select for introduction of the construct into the host genome. Upstream of HygroR is the constitutively activate promoter PTEF. Immediately downstream are an in-frame stop codon (TAA) and 2 base pair frameshift (TT). These ensure that a self-cleaving peptide (P2A) and the Neomycin/G418 resistance gene (NeoR) are not in frame and thus not translated. Therefore prior to any mutational event the host cells are susceptible to Neomycin/G418.* ***b).The construct is inserted into the AGP1 locus.*** *The whole construct was PCR amplified, and the lithium acetate method used to insert the construct using homologous integration by replacement of the AGP1 gene, located just downstream of the ARS306 origin of replication.* ***c) Detection of 2 base pair equivalent mutations****. At the start of an experiment, cells which incorporate the reporter construct without any mutations are not susceptible to Neomycin/G418. However, a 2 bp deletion or equivalent (5 bp deletion, 1 bp insertion, and others) will lead to a frameshift, resulting in loss of the stop codon and and translation of a functional P2A and NeoR protein. Selection using Neomycin/G418 will identify the descendants of any cells which have undergone a 2 bp deletion or equivalent mutation.*

4.2.3.2 Re-engineering the nucleotide sequence of the HygroR gene to increase tandem repeats and eliminate stop codons

To increase the proportion of repeats, I then re-engineered the nucleotide sequences of the Hygro^R gene (Figure 4.4). I categorised repeats as being repetitive sequences of at least 2 bp units (e.g. AAAA or ATAT). I wrote an algorithm to progress through the gene sequence for Hygro^R. Using longer sequences as subunits to progress through the gene became computationally very intensive. Examining the unmodified sequences of previous genes (Neo^R, Hygro^R) used for antibiotic resistance, the maximum length of repeat present was 10 bases (4 codons). I therefore chose a 5 codon unit to be sure that these repeats would be captured in any re-engineered sequence. As the script progressed through the Hygro^R gene, it calculated all possible permutations of codon sequences that did not alter the encoded amino acid, and then ranked these in terms of length of tandem repeat sequence. Once all the possible 5-codon tandem repeat sequences had been calculated, with start and finish site delineated, these were ranked and re-inserted into the Hygro^R gene, one at a time, replacing whole codons only, and censoring those codons that had been edited from subsequent changes (GitLab: “designing_construct/creating_reporter_construct_v1.py”).

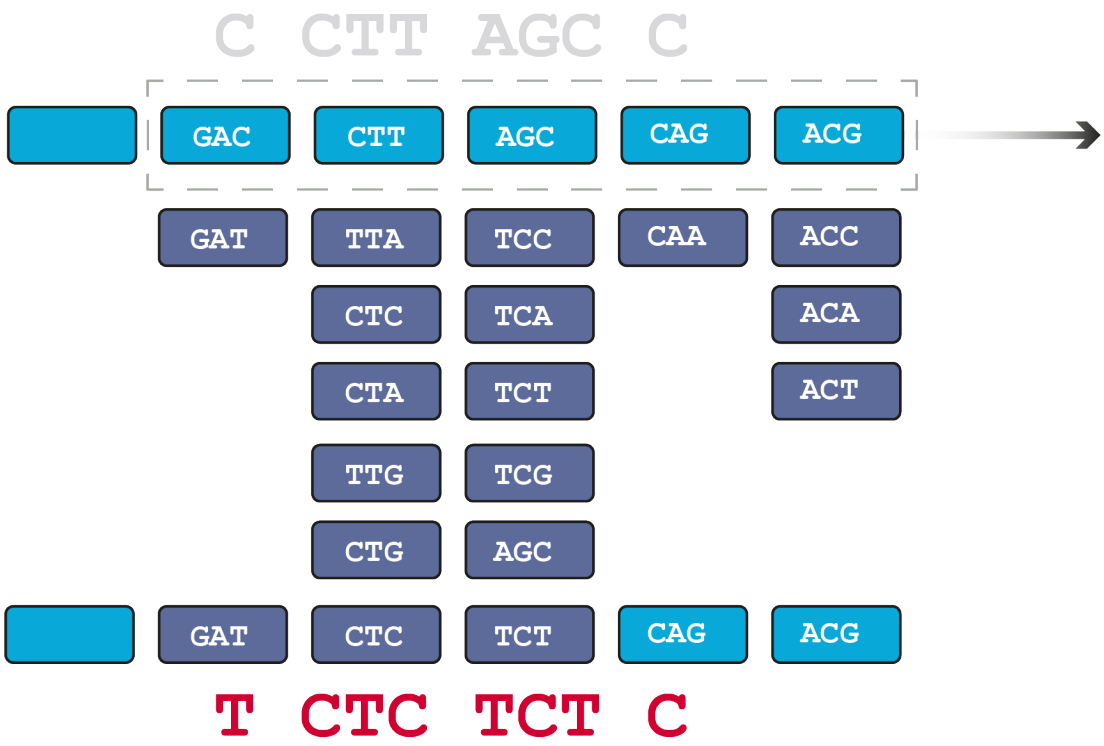


Figure 4.4. Increasing the frequency of tandem repeat sequences in the HygroR gene. *The frequency of tandem repeats in the HygroR gene was increased by making use of redundancy in amino acid codon trinucleotides. A Python script processed through the DNA sequence for the HygroR gene in a stepwise fashion, 5 codons at a time. At each codon position the script permuted all the possibilities for that 5 codon long sequence, based on the synonymous codon sequences at each position. The permuted combinations which contained any 2 bp tandem repeats (couplets) were stored for each codon position. After computing the possible couplets within any 5 codon sequence for the entire HygroR gene, these were ranked in size order, and re-inserted back into the original sequence. In this example, the first four codons, which did not contain a dinucleotide repeat sequence, before re-engineering now consist of an 8 nucleotide long TC repeat.*

A subsequent essential step was to also re-engineer out any stop codons which might be created by a 2 bp or equivalent mutation upstream of the sequence for P2A and Neo^R. Introducing a 2bp deletion at the start of the Hygro^R sequence led to 13 stop codons prior to the start of P2A, so I developed a second algorithm to take the output from the tandem repeat optimisation and perform a second round of stop codon elimination. A 2 bp deletion was introduced at the start of the sequence, and all stop codons identified. The newly formed stop codons were then repermuted to prevent their formation,

preserving a tandem repeat sequence if present, or removing it if it was necessary to prevent the formation of a stop codon (GitLab: “designing_construct/eliminating_stop_codons_v1.py”).).

4.2.4 A reporter construct with a re-engineered hygromycin resistance sequence mainly detects 1 bp insertions, rather than 2 bp deletions

The initial reporter construct (Version 1) had a higher percentage of mononucleotide runs than the original hygromycin sequence (9.1% vs 2.7%). When trialled in WT *S. cerevisiae* we found that it was more likely to detect 1 base pair insertions than 2 bp deletions. In our proof of concept initial assay in WT *S. cerevisiae* with the reporter construct, we detected 5 mutations. Of these 4 were single base pair insertions in mononucleotide repeats, and the final one a 179 base pair deletion. This suggested to us that the construct was more likely to detect polymerase slippage events (Baptiste, Jacob and Eckert, 2015) than 2 base pair deletions caused by Top1 activity. I therefore re-engineered the hygromycin gene using a modification to the algorithm described above (GitLab: “designing_construct/creating_reporter_construct_v2.py”). In this approach dinucleotide couplet repeats were prioritised over mononucleotide repeats, yielding a final product that consisted of 50.9 % dinucleotide repeats, and 0.4% mononucleotide repeats (Figure 4.5).

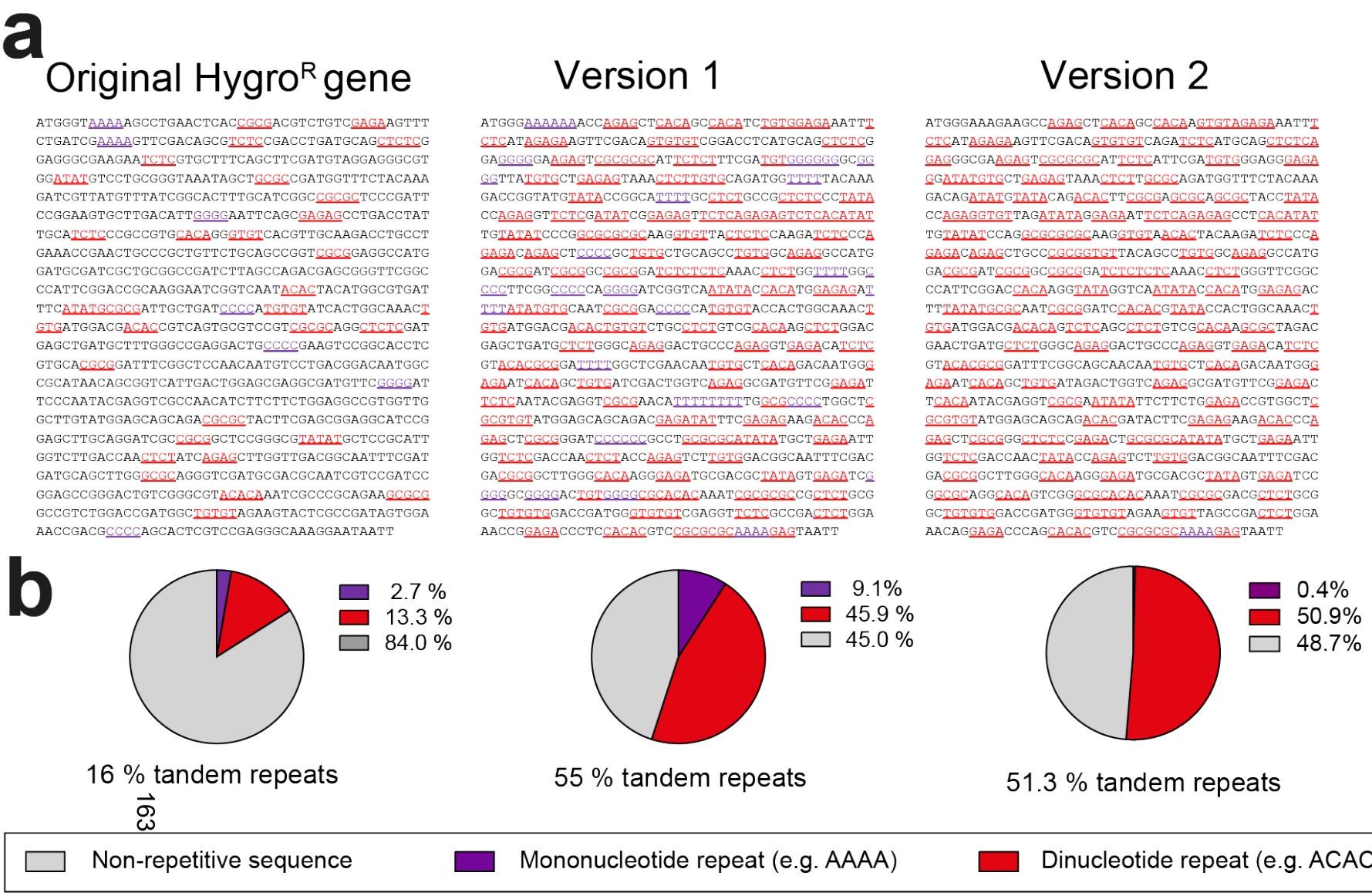

a
Original Hygro^R gene
Version 1
Version 2
b
2.7 %
13.3 %
84.0 %
16 % tandem repeats
9.1%
45.9 %
45.0 %
55 % tandem repeats
0.4%
50.9%
48.7%
51.3 % tandem repeats
163
Non-repetitive sequence
Mononucleotide repeat (e.g. AAAA)
Dinucleotide repeat (e.g. ACAC)

Figure 4.5. Results of HygroR gene re-design. Schematic showing coding sequence of original HygroR gene, and two re-designs (Version 1 and Version 2). *Mononucleotide couplet repeat sequences such AAAA are shown in purple, and dinucleotide couplet repeat sequences such as ACAC in red. Version 1 is the output of a script prioritising total number of couplet repeat sequences, and Version 2 the output when prioritising dinucleotide couple repeats over mononucleotide repeat sequences.* ***b) Pie charts showing proportion of repeat sequences in each version of the HygroR gene****. The original sequence consists of 16% tandem repeats, with a combination of mono- and di-nucleotide couplet repeats. In Version 1, the proportion of both mono- and di-nucleotide couplet repeats is increased, with 55 % of the gene consisting of couplet repeats. In Version 2, the proportion of dinucleotide couplet repeats is increased, with a reduction in the total proportion of repeats, but also a reduction in the proportion of mononucleotide couplet repeats, with only a single AAAA stretch (0.4% of total sequence) remaining.*

4.2.5 A reporter construct designed to detect 2 base pair equivalent deletions in *S. cerevisiae* successfully recapitulates published findings, with higher sensitivity

4.2.5.1 Introductory note

In analysing the outputs from the *S. cerevisiae* fluctuation assays, after several rounds of experiments outlined below, we detected a change in the number of final cells growing on the plates used as the denominator for the mutation calculation. This rendered inter-experimental comparison problematic, particularly when establishing a baseline wild-type rate. However, because within each experiment the final denominator was similar both within and between strains, the analyses below are presented as a series of results, each of which uses a WT strain for each set of experiments as a control. However, when interpreting the results, it should be noted that the mutation rate calculated for the WT strain in different experiments may not be directly comparable. Reasons for this inter-experimental variability in growth of denominator plates could include a change in media (produced by the IGGM kitchens), in conditions in the incubators, or in the centrifuges used to spin down the cells prior to plating.

4.2.5.2 Colonies identified in initial assays are heterogenous, with a combination of large and small colonies

In our initial fluctuation assay, in addition to finding that the majority of mutations were 1 bp insertions, we also found that the colonies that grew on the G418 resistance media plates were of two types: large and small. Sequencing the smaller colonies showed no mutations within the $Hygro^R$ gene. A review of the literature led the identification of a paper (Chenevert *et al.*, 1984) that had identified a relatively high spontaneous mutation rate in *S. cerevisiae* leading to resistance to neomycin, at concentrations of 250 µg per ml. We therefore increased the concentration of G418 in these experiments to 1000 µg per ml, which led to disappearance of the small colonies.

4.2.5.3 Fluctuation assays confirm a Top1 dependent increase in mutation rates after deletion of RNase H2

An initial fluctuation assay compared mutation rates in WT *S. cerevisiae* , RNase H2 null *S. cerevisiae* (*rnh201*∆), and a double knockout strain with RNase H2 and Top1 deletions *(rnh201*∆ *top1* ∆*)* (Figure 4 .6). This assay confirms findings from previous reporter constructs, with an increase in the mutation rate relative to wild type in the RNase H2 null strain. There was a 37 fold increase in the overall mutation rate, with a rise from a rate of 1.65×10^{-10} mutations per base per generation (95% confidence interval $1.51\text{-}1.76 \times 10^{-10}$) to a rate of 6.15×10^{-9} ($5.4\text{-}6.9 \times 10^{-9}$) (Figure 4 .6a). Top1 activity is the cause of this increased rate of deletions, as when Top1 is knocked out the mutation rate falls to below baseline (from 1.65×10^{-10} to 3.22×10^{-11}). Representative results from the fluctuation assay demonstrate a high numbers of mutant colonies in the RNase H2 null strain, but no different in the plating efficiency denominator plates (Figure 4 .6b). For the actual assay, the number of cells plated for this strain was diluted 100-fold to allow accurate counting of colonies.

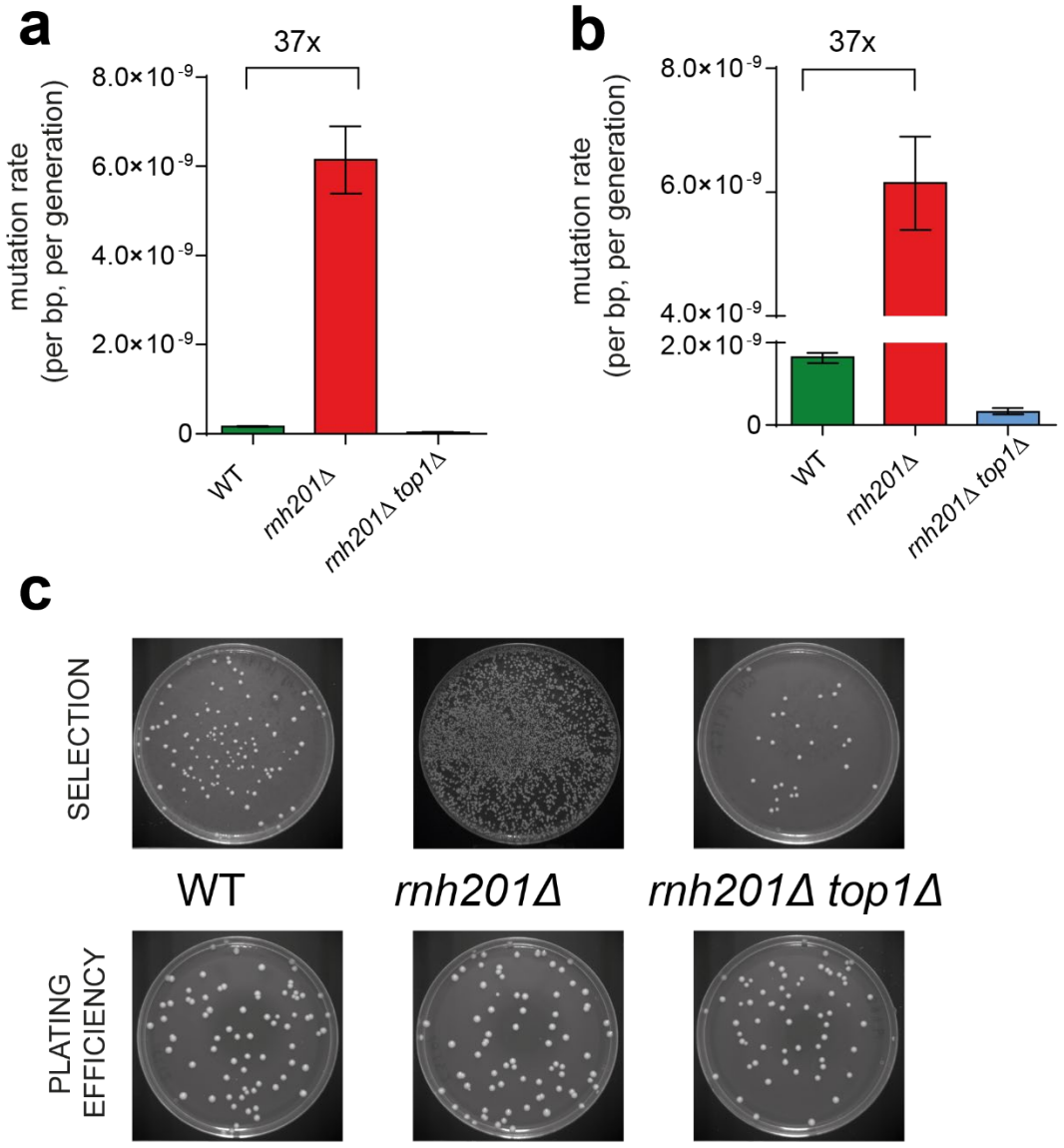


Figure 4.6. A S. cerevisiae reporter construct successfully recapitulates previous findings of an elevated rate of short deletions in an RNase H2 null strain. a) The mutation rate in an RNase H2 null strain is 37 times higher than an equivalent wildtype strain. *Mutation rates in wild type (WT), RNase H2 null (rnh201Δ) and RNase H2 null and Top1 null (rnh201Δ top1Δ) strains. The mutation rate rises 37-fold with knockout of the RNase H2 gene, and falls to below baseline with additional knock out of Top1, indicating the causal role of this enzyme in creating 2 bp equivalent mutations. Mutation rates for each strain are calculated from 16 independent cultures using the Lea-Coulson method of the median, with 95% confidence intervals.* ***b) Same figure as in a), but with interrupted y-axis. c) Representative images for each strain****. On the top row (selection plates) each colony growing on G418 medium represents the descendant of a cell where a 2 bp equivalent mutation has occurred. The bottom row shows a representative plating efficiency plate, used to calculate the denominator of total number of cells in the final culture.*

4.2.5.4 The reporter construct is much more sensitive in detecting 2 bp equivalent deletions than previous assays

The increased mutation rate in the RNase H2 null strain when compared to WT, recapitulates findings in previous work, although with much higher absolute mutation rates in the mutant than previously documented. I also found higher mutation rates in the WT strain than most previous experiments; figures for these are compared in Table 4.6.

Table 4.6. Comparison of mutation rates with previous assays.

Study	μ_{BP} POL RNH201	μ_{BP} POL *rnh201*Δ	Fold increase RNH201	Fold increase *rnh201*Δ
This thesis	1.65×10^{-10}	6.15×10^{-9}	NA	NA
CAN1 (Allen-Soltero *et al.*, 2014)	3.4×10^{-10}	2.0×10^{-10}	0.5	30.8
CAN1 (N. Kim *et al.*, 2011)	5.6×10^{-11}	1.5×10^{-10}	2.9	41.0
CAN1 (Potenski *et al.*, 2014)	5.6×10^{-11}	8.5×10^{-11}	2.9	72.4
URA3 (Clark *et al.*, 2011)	1.1×10^{-11}	2.0×10^{-11}	15.0	307.5

I interpret this increase in mutation rate relative to previous findings as being driven by three factors. The first is the increased proportion of sequence in the construct composed of tandem repeats. The construct consisted of 51.3% repeat sequences repeats, compared to 21.0% in the URA3 gene and 21.5 percent in the CAN1 gene. In addition, almost all of the repeats in the reporter construct are dinucleotide repeats, rather than mononucleotide runs. Dinucleotide repeats are known to be more mutagenic in this context. The second is the relationship between transcription and 2-5 bp deletion frequency, known to be topoisomerase1 mediated. As the construct we designed is constitutively activated with a PTEF promoter, the high transcription rate, in conjunction with the increased proportion of mutagenic tandem repeats, could explain the high rate compared to previous studies. The studies using URA3 for example utilised the endogenous promoter for the gene (Pavlov, Newlon and Kunkel, 2002), rather than the constitutively activated PTEF promoter.

Finally, our reporter construct is very sensitive, so that *any* 2 bp deletion will be detected, regardless of its position in the Hygro^R gene, whereas deletions close to the 3'UTR in the URA3 or CAN^R genes might not be detected because such mutations may not allow growth on selective medium.

4.2.6 Top1 activity is mutagenic even in RNase H2 proficient *S. cerevisiae*

4.2.6.1 Knocking out Top1 leads to a reduction in the mutation rate

As outlined above, the reporter construct we created sensitively detected 2 bp deletions. We also noted that Top1 activity itself appears to be mutagenetic, so that when Top1 is knocked out in RNase N2 proficient yeast, there is a reduction in the mutation rate (Figure 4.7a) compared to wildtype. Our initial fluctuation assay had shown that whilst the construct was designed to detect 2 bp deletions, other mutations causing equivalent frameshifts, such as 1 bp insertions, also resulted in G418 resistant colonies. Sequencing throughout a series of fluctuation assays revealed that other mutation types could also lead to the translation of a functional neomycin resistance gene product, including longer insertions and deletions, and point substitutions upstream of the original start codon that lead to the creation of an alternative reading frame. In order to determine the 2 bp deletion rate, rather than an overall mutation rate, we conducted sequencing of a larger number mutants from our fluctuation assays. As the colonies on each selection plate are potentially related to one another and not independent, we only sequenced 1 mutant per independent culture, which necessitated the development of a method to grow a large number of *S. cerevisiae* in parallel (see Materials and Methods, Section 2.2.2.4.1).

I found that using our reporter construct, in WT *S. cerevisiae* the predominant mutation type are 2 bp deletions, and within this group 2 bp deletions in perfect repeat sequences are the most common (Figure 4.7). In the RNase H2 null strain, almost all the mutations detected are now 2 bp deletions, again

predominantly in perfect repeat sequences. This is in keeping with results from my systematic review (Table 4.1), where there is a shift towards a greater proportion of short deletions in the RNase H2 null strains. As only 2 bp deletions and equivalent mutations are detected by the construct, it is not surprising that the total proportion of mutations that are 2 bp deletions in both WT and RNase H2 null strains is higher than previous studies using different reporters; the findings are also in keeping with results from WGS of the pol2-M644G *rnh201*∆ strain (Table 4 .4). I conducted a cosine similarities analysis for the mutations for all 4 strains to identify which of the other strains the RNase H2 null mutation profile most closely resembled. In a cosine similarities analysis, the mutation profile for each strain is converted into a vector, which is then compared pairwise in *n*-dimensional space, where *n* represents the number of mutation categories. A result of 1 shows that the vectors are perfectly aligned, whilst a results of -1 shows that they are opposite to one another (for details see Materials and Methods section 2.3.5.1).

For the RNase H2 and Top1 double knockout strain (*rnh201*∆ *top1*∆) and the Top1 knockout (*top1*∆) strain the mutational spectra is different. In both these strains there is a shift in the mutational spectra, so that 1 bp and longer insertions are the most common type of mutation. When rates of 2 bp deletions are plotted in absolute terms (Figure 4.7d), these are very rare in the absence of Top1 activity. The absolute rate of 2 bp deletions for the double knockout compared to wild type is 5%, and that for the topoisomerase knockout 4%. From these findings two inferences can be made. Firstly, that Top1 is the main cause of 2 bp deletions in this experimental system. Secondly, that in the absence of Top1, embedded ribonucleotides don't cause mutations that can be detected by this reporter assay.

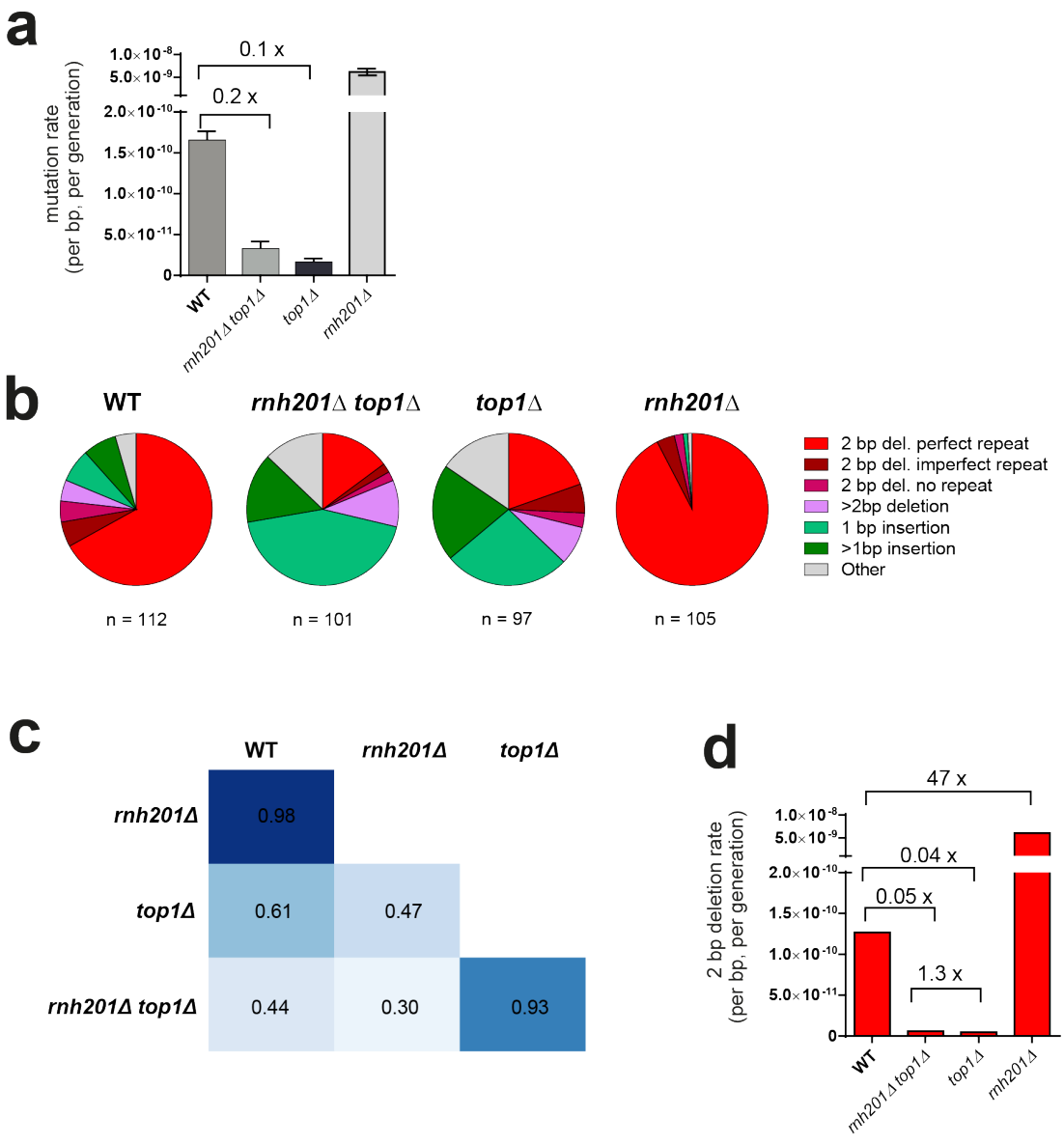


Figure 4.7. Top1 is the main cause of 2 bp deletions in this experimental system. a) The mutation rate for an RNase H2 null strain is much higher than that of WT, and falls to below the WT rate with additional loss of Top1. *Mutation rates for wild type, RNase H2 null and Top1 null (rnh201Δ top1Δ), Top1 null (top1Δ) and Rnase H2 null (rnh201Δ) strains. The mutation rate when Top1 is knocked out falls below that for wild type. Mutation rates for each strain are calculated from 16 independent cultures using the Lea-Coulson method of the median, with 95% confidence intervals. b)* ***Mutational spectra for mutant colonies in all 4 strains****. The mutations seen in the wild type strain are predominantly 2 bp deletions in perfect repeats, with a number of other types of mutations. For strains without Top1, mutations are predominantly 1 bp or longer insertions. In the rnh201Δ strain, the majority of all mutations are 2 bp deletions in perfect dinucleotide couplet repeats. total numbers of sequenced colonies below each pie-chart. All colonies from independent cultures* ***c) A cosine similarities***

analysis shows that the WT and RNase H2 null mutational spectra are most closely related. *Cosine similarities were calculated by converting mutation categories into a vector for all strains, and a pairwise comparison was conducted for all the strains. A cosine similarity of 1 indicates that the vectors are perfectly aligned, and a value of -1 that they are opposite to one another (for details of calculations see Material and Methods Section 2.3.5.1).* ***d) The rate of 2 bp deletions in the two strains without Top1 activity is very low****. Rates of 2 bp deletions deletions only in all strains, calculated by multiplying proportion of 2 bp deletions from sequencing results shown in b) by overall mutation rate in a). The low rates in the top1Δ and rnh201Δ top1Δ supporting a causal role for Top1 in creating such short deletions. The similar rates for both (1.3x higher in the double knockout strain) suggests that in the absence of Top1 the presence of genome embedded ribonucleotides does not substantially increase the chance of a 2 bp deletion occurring.*

4.2.7 Increased Top1 expression leads to an increase in short deletions

To further investigate the role of Top1 in generating short deletions in the context of physiological levels of genome embedded ribonucleotides, I performed further experiments. Given the causal role of Top1 in generating short deletions, we predicted that *increased* levels of Top1 would lead to an increased rate of 2 base pair deletions in our reporter construct. I conducted a fluctuation assay in a *S. cerevisiae* strain with a glucose/galactose regulatable Top1 expression system modified from that developed for a previous study (El Hage *et al.*, 2010). In this system, galactose leads to an upregulation of Top1 expression, and glucose to downregulation. In order to use an appropriate control, we grew the wildtype control for the up/down regulation experiments in the same growth media, so that rates for a wild type strain grown in galactose were compared to the Top1 strain grown in galactose, and likewise for the strain grown in glucose.

When the strain with an inducible promoter for Top1 is grown in the presence of galactose, which increases transcription of topoisomerase, there is a 2.5-fold increase in the overall mutation rate (Figure 4 .8a). Conversely, when this strain is grown in a glucose containing medium, the overall mutation rate falls to 0.4-fold (Figure 4 .8b). Comparing the mutational spectra from the fluctuation assays (Figure 4.8c), when Top1 activity is low (top1 o/e, YP-glucose) the proportions of 2 bp deletions is also low, whereas when Top1 activity is high (top1 o/e, YP-galactose) almost all mutations detected are 2 bp deletions. Examining specifically the absolute mutation rates for 2 bp deletions, (Figure 4.8d) these show a 2.8-fold increase with Top1 activation, and a 10-fold decrease when Top1 transcription is suppressed, consistent with findings from previous studies (Sloan, 2016; Sloan *et al.*, 2017).

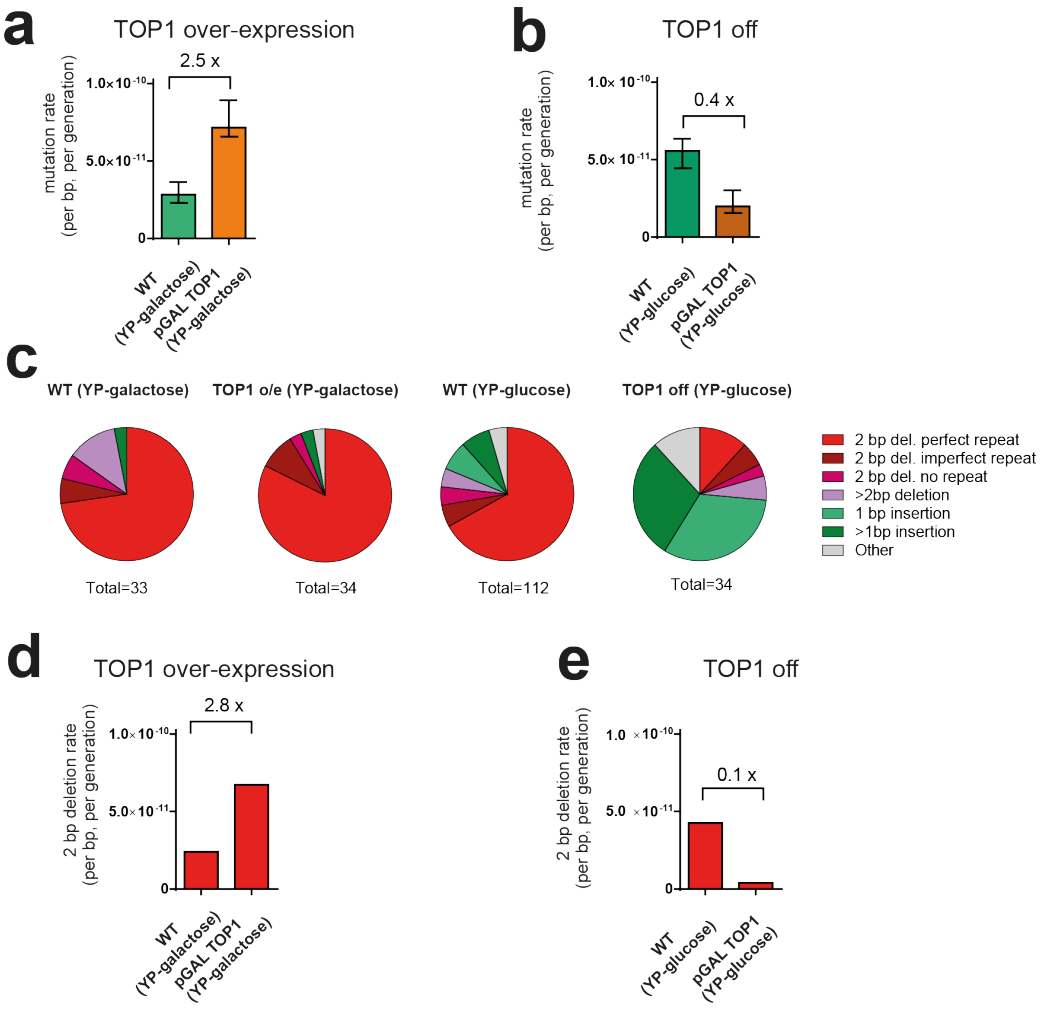


Figure 4.8 Topoisomerase over-expression in RNase H2 proficient S. cerevisiae increases the rate of 2 bp deletions.Top1 overexpression*. The endogenous promoter for Top1 has been replaced by a promoter that is upregulated in galactose containing medium and downregulated by glucose containing medium. When pGAL TOP1 is grown in galactose, there is an increase in the rate of 2 bp equivalent mutations when compared to a WT strain grown in galactose.* ***b) Mutation rates for the same strains grown in glucose (TOP1 off).*** *Mutation rate for the WT strain is higher than for equivalent grown in galactose, suggesting metabolic changes related to the carbon source may indirectly affect mutation rate. The pGALTOP1 strain shows a decreased mutation rate compared to control, consistent with causal role for Top1 in generation of short deletions.* ***c) Mutational spectra****. The predominant mutation in all strains (except for pGAL TOP1 off) are 2 bp del in perfect repeats. In pGAL TOP1 strain grown in glucose, 1 bp and greater insertions are most common, similar to the top1Δ strain (Figure 4.7) confirming the importance of Top1 in generating short deletions.* ***d) and e) Absolute rates of 2 bp deletions in all strains.*** *Calculated by multiplying proportion of 2 bp deletions from sequencing results shown in a) and b) by overall mutation rates shown in c). An increase in Top1 activity leads to an increase in the rate of 2 bp deletions, and suppression of Top1 activity to a 10 fold reduction in the rate of 2 bp deletions.*

4.2.8 Mechanisms of Top1 mutagenesis in the presence of RNase H2: is it leading strand specific?

The experiments above showed that there is Top1 mediated short deletion mutagenesis even in the presence of physiological levels of ribonucleotides. The mechanisms for this remain poorly understood (Cho and Jinks-Robertson, 2018): it is not known whether this process is mediated by ribonucleotides during the period that they are transiently misincorporated, or by the presence of another base lesion that interacts with Top1. Additionally, previous work has shown that Top1 mutagenesis is strand specific, predominantly affecting the leading strand, which is synthesised by Pol ε (Williams *et al.*, 2015). Importantly, these experiments only examined this phenomenon in the *absence* of RNase H2. I sought to investigate this process further by using strains of the reporter construct with a Pol ε (pol2) with steric gate mutations which incorporate more (pol2-M644G) and fewer (pol2-M644L) (Nick McElhinny, Kumar, *et al.*, 2010) ribonucleotides into the leading strand during DNA replication. If embedded ribonucleotides were the main driver of Top1 mutagenesis in the presence of RNase H2, one would expect an increase in mutation rate with the pol2-M644G mutant, and a decrease in mutation rate with a pol2-M644L mutant.

However, results from the experimental data do not support this hypothesis: mutation rates are equivalent in both Pol ε mutant strains and the wild type control (Figure 4.9a). Contrary to expectation, an increase of ribonucleotide incorporation into the leading strand did not result in more Top1 mediated mutagenesis. In addition, the mutational spectra for all three strains appear to be very similar, with no shift in the presence of less (pol2-M644L) or more frequent (pol2-M644G) ribonucleotide incorporation (Figure 4.9b), as might be expected if leading strand incorporation and Top1 mediated mutagenesis were more common for the pol2-M644G strain.

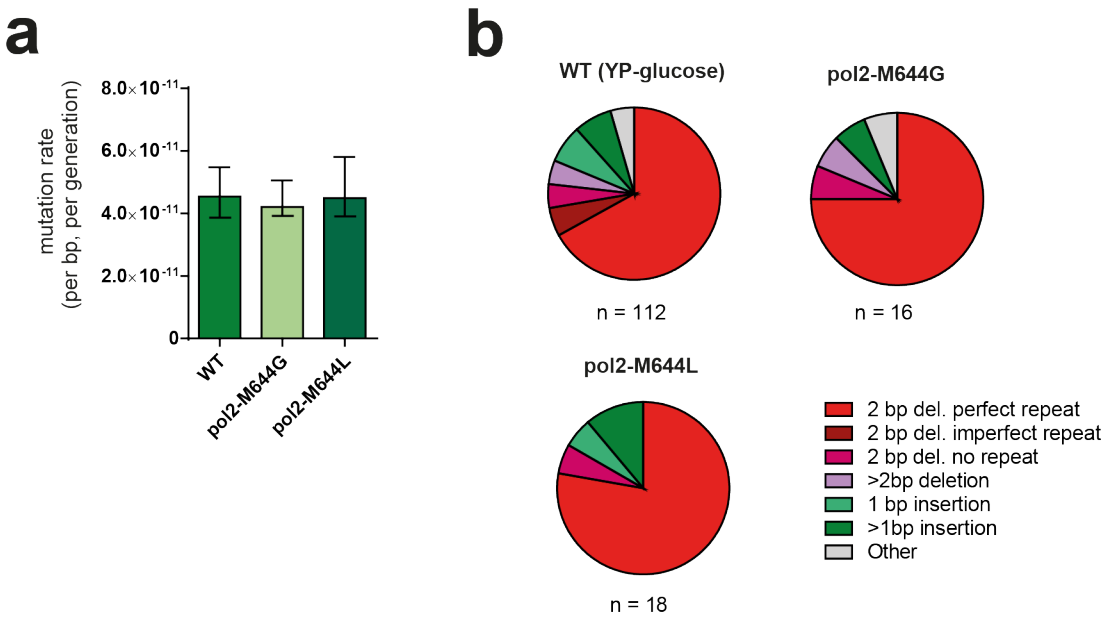


Figure 4.9. Mutation rates and spectra for Pol ε mutants. a) Mutation rates*. Results from a fluctuation assay comparing mutation rates in S. cerevisiae strains containing a wild type Pol ε (WT), a mutant which incorporates fewer ribonucleotides (pol2-M644L) and a mutant which incorporates more ribonucleotides (pol2-M644G). Results for all 3 strains are similar, with overlapping 95% confidence intervals.* ***b) Mutational spectra for the strains in (a).*** *In all groups the most common mutation type is 2 bp deletions in a perfect repeat context. Mutation rates for each strain are calculated from 16 independent cultures using the Lea-Coulson method of the median, with 95% confidence intervals.*

4.2.9 Confirming specificity of the construct for Top1 mediated mutagenesis

Another well described cause of short deletions in repetitive sequences in *S. cerevisiae* is mismatch repair deficiency (Lujan *et al.*, 2014). We wanted to check the specificity of the reporter construct for ribonucleotide associated short deletion mutagenesis, so I performed a fluctuation assay in an MSH2 knockout strain created by Martin Reijns. MSH2 (MutS homolog 2) is an enzyme in the mismatch repair pathway, shared by both *S. cerevisiae* and humans. MSH2 interacts with other components of the pathway MSH3 and MSH6 to identify and repair mismatched bases (Cerretelli *et al.*, 2020). In combination with MSH6 and MSH2 forms the MutSα complex, which scans DNA for mispaired bases, and mispaired indels of 1-10 bases (Lyer *et al.*,

2006). MSH2 also binds with MSH3 to form the MutSβ complex, which primarily supports repair of indels 1-10 bases in length.

This fluctuation assay showed a 3.4-fold increase in the overall mutation rate, compared to wildtype (Figure 4.10b). This is in keeping with an expectation of an increase in mutation rate, as the mismatch repair machinery is known to identify and repair short insertions and deletions (Lang, Parsons and Gammie, 2013; Serero *et al.*, 2014). However, the magnitude of the increase is much less than that seen with the RNase H2 null strain, which was perhaps surprising given the documented propensity for mismatch repair strains to generate 1 bp insertions (Lujan *et al.*, 2014). However, the relatively low numbers of 1 bp insertions (which would be detected by the reporter construct) is likely due to the infrequent occurrence of mononucleotides runs ≥4 base pairs in length in version 2 of the Hygro^{R} sequence (0.4% of total, Figure 4 .4), as the frequency of 1 bp insertions is known to be related to the length of these runs (Baptiste, Jacob and Eckert, 2015)

In fact, the mutational spectrum and proportion of mutations which are 2 base pair deletions for the *msh2Δ* strain is very similar to that seen in the wild type (Figure 4 .10b and c). That the reporter construct detects an increase in 2 bp deletions does suggest that mismatch repair does has a role in preventing 2 bp deletions. However, the magnitude of this difference (3.1-fold) is much less marked than that seen in the RNase H2 null strain (47-fold), indicating that the contribution of ribonucleotides in this context is more significant.

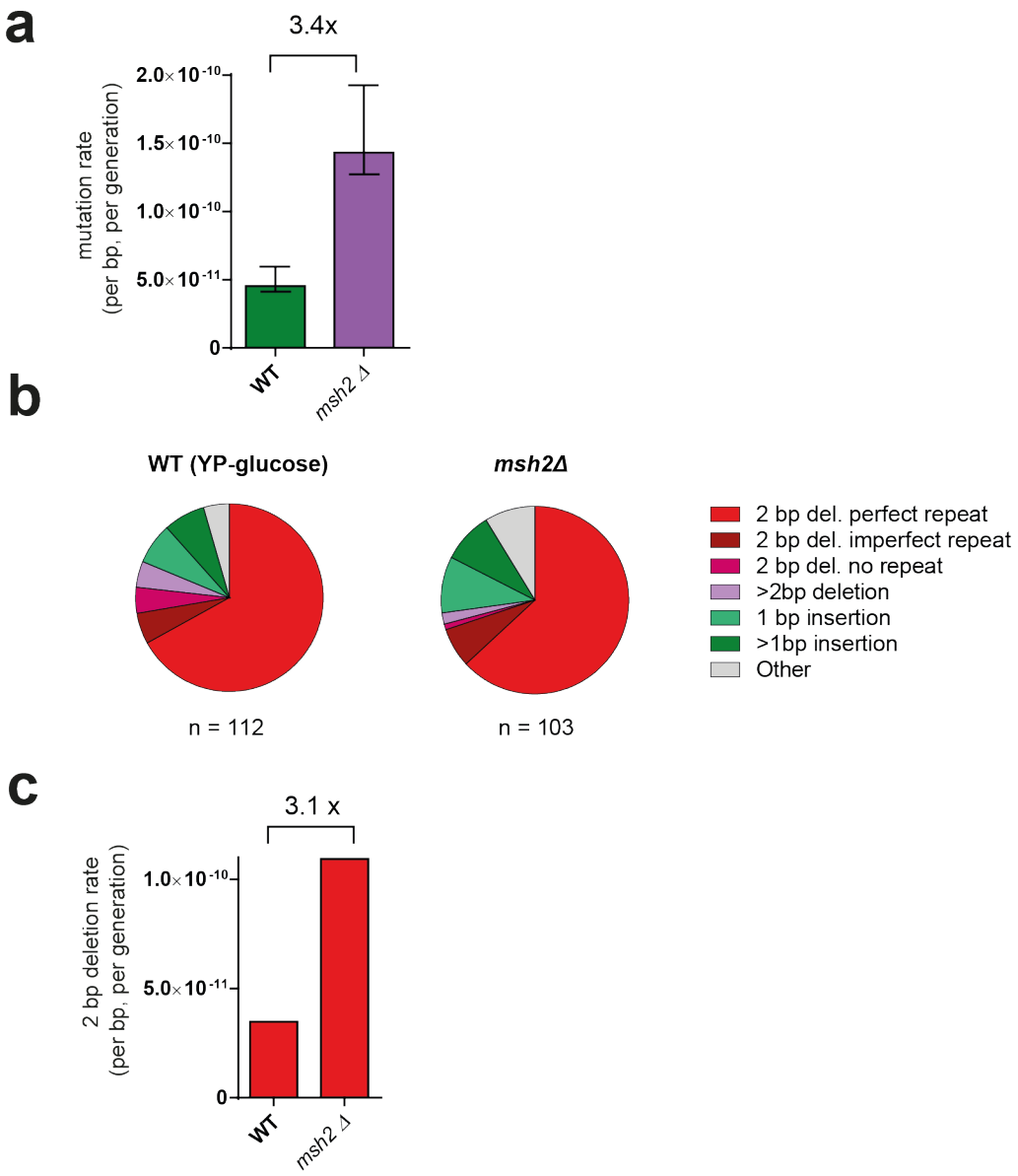


Figure 4.10. Mutation rates in a mismatch repair knockout. a) Mutations rates for WT compared to a mismatch repair deficient strain (msh2Δ). *The mutation rate increases 3.4-fold compared to WT.* ***b) Mutational spectra for the strains in (a).*** *For both strains the most common mutation type is 2 bp deletions in a perfect repeat context. There is no marked increase in 1 bp insertions in the msh2Δ strain. c)* ***Absolute rates of all 2 bp deletions in WT compared to an msh2Δ strain****. The increase of 3.1-fold is in keeping with the overall increase in mutation rate. Mutation rates for each strain are calculated from 16 independent cultures using the Lea-Coulson method of the median, with 95% confidence intervals.*

4.3 Discussion

4.3.1 Top1 is an important cause of 2 base pair deletions in this highly sensitive and specific system

My experiments confirm that, in this model system, an important cause of 2 base pair deletions is the action of Top1. In the presence of supra-physiological levels of ribonucleotides, Top1 acts on genome embedded ribonucleotides to cause short deletions in repetitive sequences (Figure 4.11).

The model system I have developed is highly sensitive for the detection of 2 base deletions, with a higher 2 base pair deletion rate detected than seen in previous reporter constructs (Table 4 .6). It is also specific for this mutation type: in wild type *S. cerevisiae* over three quarters (76.8%) of all mutations detected are 2 bp deletions, rising to 98.1% of all deletions in the absence of RNase H2 is knocked out (Figure 4.7b).

The model system is also sensitive enough to detect a *decrease* in 2 bp deletions in the absence of Top1, even when RNase H2 is present (Figure 4.7c). The rate of 2 bp mutations is directly related to Top1 levels in the presence of physiological levels of genome-embedded ribonucleotides. If Top1 levels are increased with an inducible galactose promoter, there is a consequent increase in the rate of 2 bp deletions; conversely if Top1 levels are lowered, there is a decrease in the rate of these deletions (Figure 4.8).

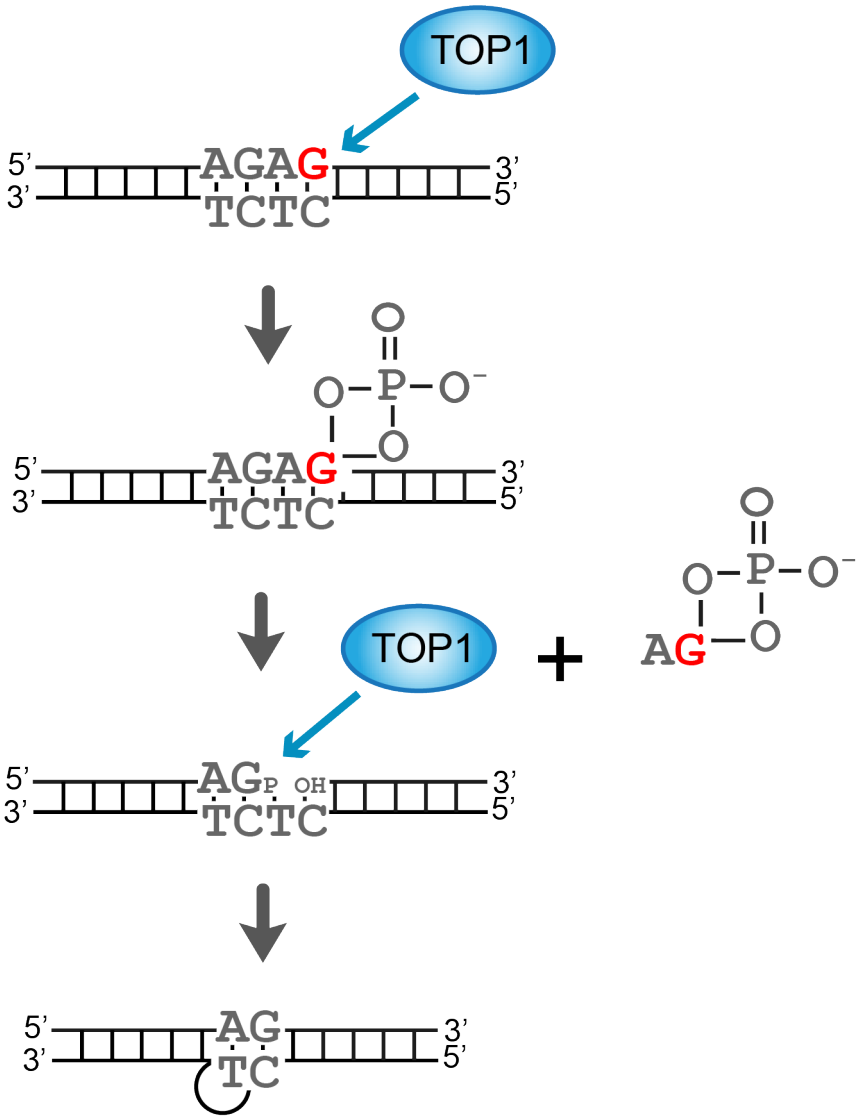


Figure 4.11. A model for Top1 mediated mutagenesis in perfect tandem repeats. *Based on* (Sparks and Burgers, 2015). *An embedded ribonucleotide (red G) is cleaved by Top1 (TOP1); however the presence of a 2'-OH group leads to the formation of a 2',3'-cyclic phosphate and dissociation of TOP1. TOP1 cleaves 2 nt upstream of the original incision, releasing the 2',3'-cyclic phosphate and creating a 2 nt gap. If the ribonucleotide is located in a perfect tandem repeat there is slippage at the upstream TC on the complementary strand and re-ligation, leading to resolution of the gap but a 2 bp deletion in the DNA of the cells descending from those on the template strand.*

The system does not only detect Top1 mediated mutagenesis. The deletion on MSH2, part of the mismatch repair machinery, also leads to an increase (3.1-fold) in the rates of 2 bp deletions compared to a wild type control (Figure 4.10). However, this increase in the rate of 2 bp deletions is much less marked than that seen in the absence of RNase H2 (47-fold). Two questions arise from this comparison. Firstly, what is the relative contribution of Top1 activity as compared to deficient mismatch repair in the generation of short deletions? Secondly, are the deletions seen in mismatch repair deficiency due to the inefficient repair of lesions due to polymerase slippage (Figure 1.7), or are they instead due to the inability of the mismatch repair machinery to correct lesions created by aberrant Top1 activity? To examine this, one could conduct fluctuation assays in an *msh2Δ rnh201Δ* strain. If 2 bp deletions in a mismatch repair deficient background are also driven by Top1 activity, one would expect a substantial rise in 2 bp deletions above the level seen in the *rnh201Δ* strain, due to the unavailability of mismatch repair to repair these. Conversely, if the 2 bp deletions seen in the *msh2Δ* strain are due to polymerase slippage, then one would expect a small but not dramatic rise in the rate of overall deletions. The total number of deletions seen in the *rnh201Δ* strain would be added to those seen in the *msh2Δ* background, but one would not expect a synergystic interaction between the two processes and a substantial rise in 2 bp deletions above what is seen in the *rnh201Δ* strain. Top1 could be knocked out in both strains to confirm the causal role of the enzyme in generating these deletions.

4.3.2 Does Top1 act on ribonucleotides to cause mutations in the presence of RNase H2?

Two papers, published in 2011 (Lippert *et al.*, 2011; Takahashi *et al.*, 2011) at the same time as the studies showing the mutational consequences of Top1 activity acting on ribonucleotides, showed that Top1 activity could lead to 2-5 bp deletions even in the presence of physiological levels of ribonucleotides. These 2 studies both used reporter constructs : CAN1 in both, and an

additional *lys2ΔA746NR* frameshift assay in the first (Lippert *et al.*, 2011).They found that whilst mutations did occur in low transcription conditions, the rate increased with upregulating transcription of the reporter, supporting a role for Top1 activity (which is active in highly transcribed regions). The paper from the same year (N. Kim *et al.*, 2011) that looked at *ribonucleotide* associated mutations also noted that there appeared to be a sequence bias in the hotspots seen when RNase H2 was knocked out. Here, there was an increase in deletions in $(AG)_4$ and $(TC)_3$ repeats in a *lys2ΔA746NR* frameshift assay when RNase H2 was knocked out. However, in the same system there was no increase in mutations when RNase H2 was knocked out at an $(AT)_2$ hotspot, as compared to those for those seen with no RNase H2. Thus, there appeared to potentially be two Top1 mediated mutational processes, one linked to Top1 acting on ribonucleotides, and another acting via a ribonucleotide independent process, with rates associated with transcriptional activity.

However, the substrate for Top1 binding and end processing in ribonucleotide independent short deletion mutagenesis has remained elusive. A study published in 2017 (Sloan *et al.*, 2017) showed that knocking out Top1 led to a reduction in 2 bp deletions in hotspots in a CAN1 reporter construct, and that overexpression lead to an increase in the rate of 2 bp deletions. However, the mechanism for this process remains to be delineated. Another study found that a mutant of Top1 with reduced re-ligation activity showed an increased mutation rate in dinucleotide repeats (Kim *et al.*, 2013), suggesting that the excision of trapped Top1 cleavage complexes (cc) leads to short deletions, but no cause has been found for the trapping of the Top1 cc that initiates this process (Cho and Jinks-Robertson, 2018).

Our experiment looking at mutation rates in pol2-M644G was designed to provide insight into this question. If embedded ribonucleotides are the main driver of Top1 mutagenesis in the presence of RNase H2, one would expect

an increase in mutation rate with the pol2-M644G mutant, which incorporates more ribonucleotides, and a decrease in mutation rate with a pol2-M644L mutant, which incorporates fewer ribonucleotides. However, this is not what we see in this experiment. Mutation rates for WT and the 2 polymerase mutants are equivalent (Figure 4.9a), with no change in the mutational spectra. This raises 2 possibilities. The first is that ribonucleotides do not play a role in Top1 mediated short deletion mutagenesis. The second is that this mutagenesis is strand specific. If the process were lagging strand specific, then one expect that mutations in Pol α and δ that led to increased ribonucleotide incorporation on the lagging strand might lead to an increase in mutation rate. Preliminary experiments conducted by Martin Reijns suggest that this may in fact be the case, as there is a rise in the mutation rate with Pol α and δ mutator strains. This perhaps supports a model where RNase H2 travels with the leading strand (Lujan *et al.*, 2013), and Top1 travels with the lagging strand. In this model the presence of RNase H2, ribonucleotides are efficiently removed from the leading strand, and Top1 is more likely to act on ribonucleotides incorporated into the lagging strand, as they are slower to be removed by RNase H2. However, arguing against this, a previous study found that Top1 activity is leading strand specific, with no change in the mutation rate in Pol α and δ mutator strains in the presence of RNase H2 (Williams *et al.*, 2015). Clearly more work is needed to fully delineate the mechanism of Top1 mediated mutagenesis in the *presence* of RNase H2.

Elucidating the relationship between genome-embedded ribonucleotides in the presence of RNase H2 is of broader interest beyond the mutagenesis field. Recently published work has shown that, in the absence of RNase H2, topoisomerase cleavage of embedded ribonucleotides leads to PARP trapping lesions that impede DNA replication, endanger genome stability, and render these cells more vulnerable to treatment with PARP-inhibitors (Zimmermann *et al.*, 2018). If topoisomerase does act on genome embedded ribonucleotides

even in the presence of RNase H2, that is in a physiological context, this opens up the possibility of a mechanism of action of these agents in a context outwith that of deficient ribonucleotide excision repair.

4.3.3 The construct can be used to search for the presence of a Top1 dependent mutational signature in mammalian cells

The development of a highly sensitive and specific reporter construct achieved a number of goals. It recapitulated previous findings that there is a distinctive mutational process associated with genome-embedded ribonucleotides and Top1 activity. It raised interesting questions about the role of Top1 activity in generating short deletions in the setting of physiological levels of genome-embedded ribonucleotides. And for the purposes of this research question underlying this thesis, it suggested that it might be possible to directly ask the question of whether Top1 mediated mutagenesis takes place in human cells. The construct was designed to be directly transferrable into human cells, unlike previous constructs which due to being based on nutritional markers are limited to use in yeast. The dramatic (47-fold) rise in 2 bp mutation rate seen in the *rnh201Δ* strain suggested that if there was also a Top1 mediated mutational signature in mammalian cells, the reporter would be able to detect it. The higher absolute mutation rates seen in our construct compared to previous studies also meant that fewer human cells would have to be grown to generate sufficient cell divisions to yield what were likely to be very rare mutational events, with a rate of 1.65×10^{-10} mutations per base pair per generation.

Chapter 5 Mutational consequences of ribonucleotide incorporation in human genomes

5.1 Introduction

As outlined in Chapter 1, short indels (<50 base pairs in length) are important in shaping evolution, causing inherited disease, and in cancer. In this chapter, building on the work presented in Chapter 4, I will examine the hypothesis that Top1 mediated short deletion mutagenesis may play a role in the formation of *de novo* mutations in human populations and in cancer.

5.1.1 Short indels in human genomes

De novo ("starting from the beginning") mutations are those that are unique to an individual, acquired during the formation of gametes or post-zygotically (Acuna-Hidalgo, Veltman and Hoischen, 2016). These are how all the mutations that define us as a species or as individuals are acquired; insight of what drives them is also fundamental to our understanding of human inherited disease and cancer.

Whilst population level datasets such as the 1000 Genomes project (Altshuler *et al.*, 2012) provide an overview of mutational processes, they document shared mutations that are likely to have been shaped by selection in different populations. Therefore our understanding of mutations in humans is driven by studies of *de novo* mutations, cancer genome sequencing, and from experimental datasets.

In humans, an estimated ~100 *de novo* mutations occur in every newborn child (Lynch, 2016). Parental age is correlated with mutational load in the child, and studies of trios show that this mutational load is mainly concentrated in the male germline (Jónsson *et al.*, 2017), which is presumed to be due to the higher number of cell divisions that spermatogonial cells undergo compared to

oocytes. By the time of fertilisation, it is estimated that a primary oocyte has undergone 31 cell divisions (Drost and Lee, 1995). Unlike the female germline, the male germline undergoes continuous replication cycles. In the male germline at the time puberty is reached and spermatogenesis commences 34 cell divisions have already occurred; after this sperm are continuously produced through asymmetrical division of self-renewing spermatogonial stem cells every 16 days. Therefore, by the age of 30 an estimated 429 cell divisions will have taken place, and by the age of 40 an estimated 659 (Ségurel, Wyman and Przeworski, 2014). There is debate about the relative contribution of different mutagenic mechanisms to the generation of mutations in the male germline (Gao *et al.*, 2019). However, mutations that occur as a consequence of faulty DNA replication are widely accepted to constitute a significant proportion of all germline mutations (Crow, 2000; Kong *et al.*, 2012; Acuna-Hidalgo, Veltman and Hoischen, 2016).

5.1.2 The COSMIC classification of mutational signatures

Two factors have led to the widespread adoption of sequencing technologies in cancer. The first is the availability, and falling costs, of the technology available (Schwarze *et al.*, 2018). The second is the realisation that understanding the evolutionary history of tumours could offer insights into how cancers developed, and guiding therapy (Aparicio and Caldas, 2013). The Cancer Genome Atlas (TCGA) was set up in 2006, with initial funding from the National Institutes of Health in the United States. Initially, the project used PCR-based gene panels to look at genes frequently mutated in cancer; with time this progressed to hybrid capture methods of large gene panels (Hodges *et al.*, 2007) and by 2009 to whole exome and whole genome sequencing technologies (*The Cancer Genome Atlas - Timeline and Milestones - National Cancer Institute*, 2020). In 2010 the data was made publicly accessible via the International Cancer Genome Consortium (ICGC) Data Portal. By this stage there were ~5000 cancer genome sequences (a combination of gene panels,

hybrid capture exonic sequences, and whole exome sequences) available. Recently, ICGC has expanded to include WGS from centres globally, and is called PCAWG: PanCancer Analysis of Whole Genomes (*PCAWG | ICGC Data Portal*, 2020). At the time of my analyses, 1950 curated whole genome sequences had summary data on indels available from the PCAWG data portal. Importantly, these cancer whole genome sequences are generated using tumour:normal pairs. The mutations unique to each individual (when compared to the reference genome) are subtracted from those called in the tumour to give a dataset of mutations likely to be unique to the each tumour.

Faced with these large datasets, it was recognised that in order to understand these cancers in a systematic way it might be helpful to classify these cancers based on so-called “mutational signatures” (Alexandrov *et al.*, 2013; Alexandrov and Stratton, 2014). The concept of the mutational signature is to define the recurring mutational patterns present in a set of tumours. These mutational signatures can then be mapped to known mutational processes, to identify the contribution of each type of mutational process to a cancer or cancer type, estimate the timing of a mutational process in the development of a cancer (Nik-Zainal *et al.*, 2012), or guide therapy (Davies *et al.*, 2017).

The COSMIC (Catalogue of Somatic Mutations in Cancer) consortium has developed a number of mutational signatures in order to analyse and better understand the large volume of cancer genomes described above. In its first iteration, the consortium identified 30 single base substitution (SBS) mutational signatures (Alexandrov *et al.*, 2013). An example of this, SBS1 (Figure 5 .1a) is proposed to be caused by the deamination of 5-methylcytosine to thymine, which generates G:T mismatches in double stranded DNA (Coulondre *et al.*, 1978); failure to detect and remove these prior to DNA replication results in a permanent change from the original C to a T (Pfeifer, 2006; Alexandrov *et al.*, 2013).

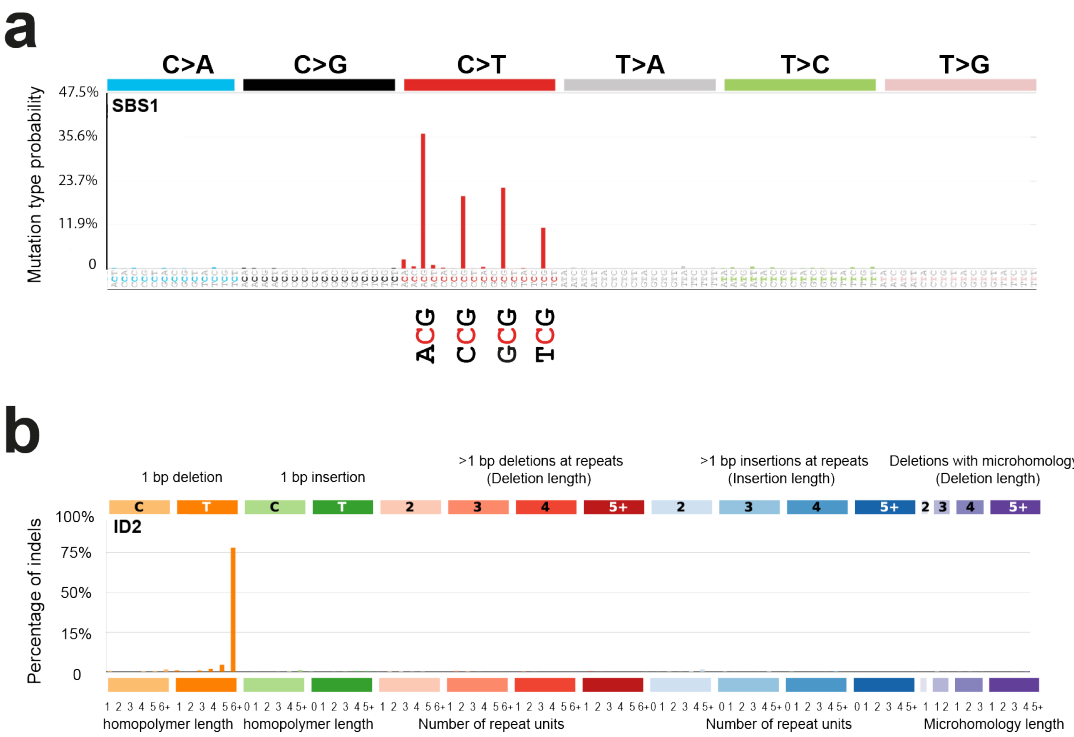


Figure 5.1. Examples of COSMIC mutational signatures. a) A single base substitution signature: C to T transitions (SBS signature 1) caused by methyl-cytosine deamination. *Cytosines are predominantly methylated in a CG context; this methylation can lead to methyl-cytosine deamination and transition from a C to a T. Therefore one would expect mutations to be concentrated in trinucleotides in a CG context (NCG) which is indeed the case. Mutational profile using the conventional 96 mutation type classification. This classification is based on the six substitution subtypes: C>A, C>G, C>T, T>A, T>C, and T>G (all substitutions are referred to by the pyrimidine of the mutated Watson—Crick base pair). Further, each of the substitutions is examined by incorporating information on the bases immediately 5' and 3' to each mutated base generating 96 possible mutation types (6 types of substitution x 4 types of 5' base x 4 types of 3' base). The y axis shows the percentage of all mutations within a particular signature that are of each trinucleotide type; there is no correction for underlying sequence composition.* ***b) An indel mutational signature: increase in 1 bp deletions in T (equivalent to A) repeats (ID signature 2) due to defects in mismatch repair.*** *This is seen as the orange bar in the T section of 1 bp deletions. As with the SBS signatures, 1 bp indels are referred to by the pyrimidine of the mutated Watson—Crick base pair. Insertions at non-repeat sequences are indicated as Number of repeat units=0. Deletions at non-repeat sequences are indicated as Number of repeat units=1. Numbering of repeat units follows on from the initial 0 for insertions, and 1 for deletions. Hence, the peak for ID2 for T deletions at homopolymer length 6+ means that these are 1 bp deletions occurring in repeats of 6 or more Ts (or deletion of an A in tract of 6 or more As). For deletions >1 bp in length only, deletions that do not occur within perfect repeats are assessed for regions of microhomology in the sequence adjacent to the deletion, and if this exists they allocated to the "Deletions with microhomology" category, depending on the size of the deletion and the microhomology. Again, there is no correction for underlying sequence composition. Figures adapted from (Alexandrov et al., 2020) under a Creative Common Licence (CC BY 4.0, 2020).*

Subsequent to the identification of mutational signatures for single base substitutions, signatures have also been proposed for short insertions and deletions (Alexandrov *et al.*, 2020). An example of one of these, COSMIC indel signature 2 (ID2, Figure 5 .1b), which would correspond to the loss of mismatch repair machinery: deletions of Ts in homopolymer repeat tracts, consistent with the slippage model outlined in the Introduction.

5.1.3 Difficulties in inferring causation from observational studies

Inferring causation from the studies outlined above is problematic, due to their observational nature. Therefore the study of mutagenesis has drawn on a second major source of information on human mutational processes: inference from experimental systems. These can take two forms:

1. Biochemical experiments, or those from non-human model systems (phages, prokaryotes, *S. cerevisiae*, *Drosophila* and others) (Streisinger *et al.*, 1966; Kunkel, 1985; Bebenek and Kunkel, 1990; Sia *et al.*, 1997)
2. Mutation accumulation experiments in human cell lines with knockout of genes postulated to be associated with a mutational signature(Drost *et al.*, 2017; Kucab *et al.*, 2019).

Point 1) is addressed in the Introduction and in the results from Chapter 4, which confirms previous findings of a distinctive Top1 mediated short deletion signature in *S. cerevisiae*, and demonstrates that this correlates with the degree of Top1 activity and frequency of ribonucleotide incorporation. A number of studies have examined Point 2), and again data from some of these experiments will be presented later on. However, as will become clear, there is little experimental effort which has been directed specifically towards elucidating the causes of short deletions in human genomes, and the original work presented here is a contribution towards that corpus of knowledge.

5.1.4 Approach

In this chapter I examine data from *de novo* mutations (DNMs) to estimate the relative frequency of 2-5 bp deletion in the context of all new short indels in human populations. *De novo* mutations are those which have arisen for the first time in an individual, rather than being inherited from their mother or father. I transfer my reporter construct from *S. cerevisiae* into HeLa cells and use this to assess indel mutation frequencies in mammalian cells. I present my analysis of a mutation accumulation experiment conducted by Martin Reijns in retinal pigmented epithelial (RPE1) cells. Finally, I present analyses of previous mutation accumulation experiments, in both *S. cerevisiae* and human cell lines, and in cancer samples, to establish whether the mutational signature associated with Top1 activity may be present in human DNMs and in cancer.

5.2 Results

5.2.1 De novo mutations in human population studies

5.2.1.1 Gene4Denovo dataset

A number of studies (Kong *et al.*, 2012; Besenbacher *et al.*, 2015; Jónsson *et al.*, 2017) have generated large scale whole genome and whole exome sequencing of trios (mother, father and child) to identify *de novo* mutations in the child. While it is hard to directly access this data, the Gene4Denovo database (Zhao *et al.*, 2020) brings together published DNMs from 59 studies with standardised annotation. I downloaded all short indels (<50 bp) annotated by Gene4Denovo, and identified 2,316 indel mutations. To further examine the nature of *de novo* indels in this dataset I modified the Python script used by the COSMIC consortium, called SigProfilerMatrixGenerator (Bergstrom *et al.*, 2019) (script available at GitLab, "/COSMIC/ generating_COSMIC_plot_DNMs.py"). In order to assess the frequency of 2-5 base pair deletions, I modified the classification used by COSMIC to identify such 5 bp deletions as a separate entity. In order to maintain consistency, the

COSMIC approach to plotting indels is preserved in subsequent figures, but additional elements such as pie charts incorporate this extra analysis step.

5.2.1.2 2-5 bp deletions are the largest single category of human exonic short indels

I first assessed the overall frequency of different indel types. I found that 2 to 5 bp deletions constitute the largest single category of short indels in human exonic *de novo* mutations: 33.9% of the total (Figure 5 .2a). To gain insight into the sequence context of the indels, I plotted these using the COSMIC classification (Figure 5 .2b). Within the COSMIC classification, a 2-5+ base pair deletion can be situated in no repeat context, an imperfect repeat, a short perfect repeat (<5 repeat units) or a long perfect repeat (≥5 repeat units) (Figure 5 .2c and d). For exonic DNMs, my analysis established that the largest group of 2-5 bp deletions are those in imperfect repeats (Figure 5 .2d); the second largest group after this are deletions in short (<5 repeat unit) perfect repeats. Therefore a mutational process that explains the creation of short deletions in the human germline would be an important contributor to human mutagenesis. Given that a 2-5 base pair deletion signature is associated with Top1 mutagenesis in yeast, it is important to examine whether the same process also contributes to mammalian mutagenesis.

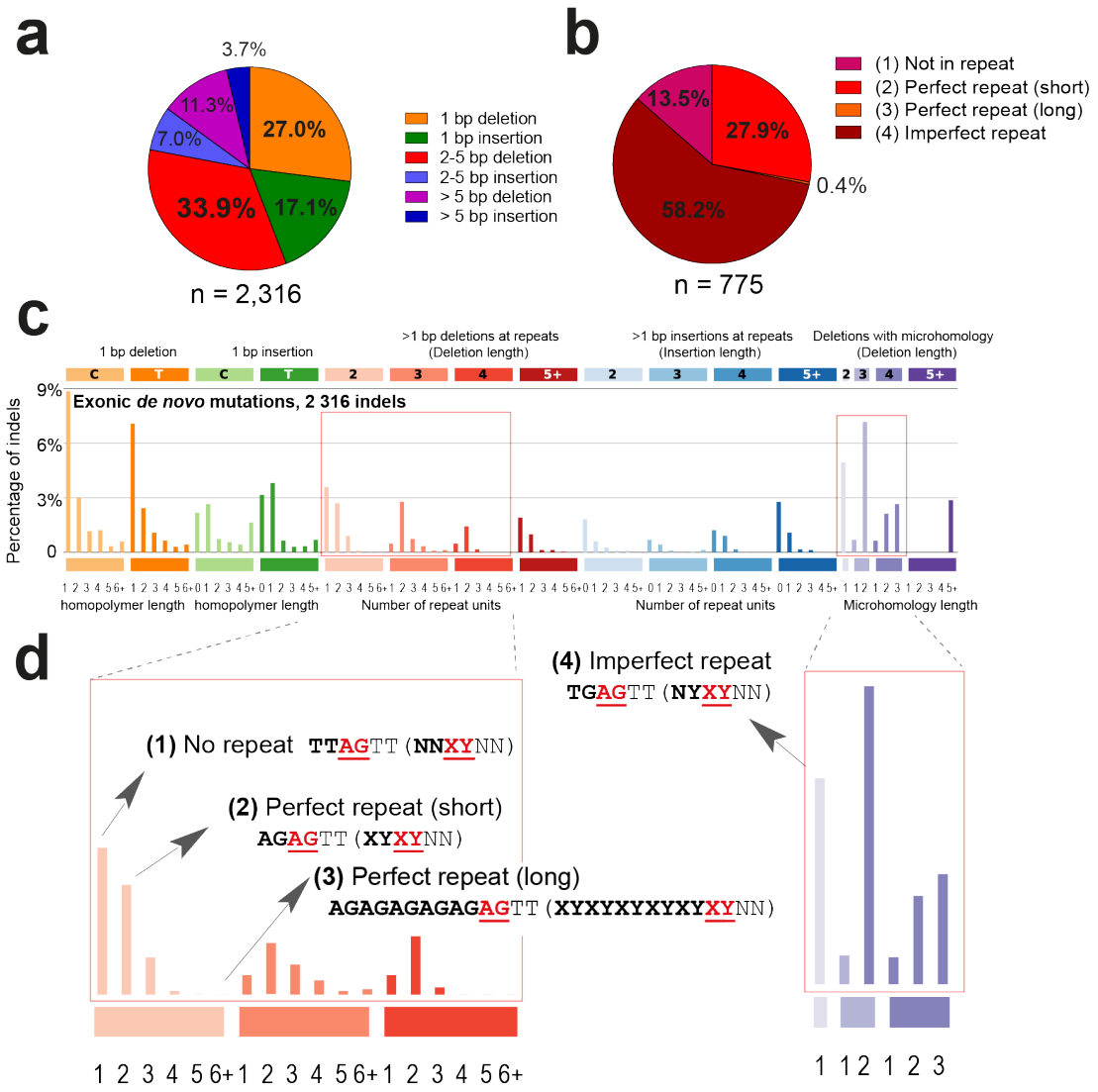


Figure 5.2. The most common category of short indels (<50 bp) in human de novo mutations (DNMs) are 2-5 bp deletions at short perfect and imperfect repeats. a) 2-5 bp deletions constitute 33.9% of exonic DNMs. *My analysis of exonic de novo mutations (DNMs) from the Gene4Denovo dataset* (Zhao *et al.*, 2020). *2,316 short indels were available from the database.* ***b) Visualisation of exonic DNMs using the COSMIC classification shows 2-4 bp deletions are located in non-repeat, imperfect repeat and perfect repeat contexts.*** *Plot of the same analysis as a) using the COSMIC classification; the logic underlying these plots is explained in detail in Figure 5 .1. Panel created using the SigProfilerPlotting function* (Bergstrom *et al.*, 2019). *2-4 base pair deletions are shown in pink squares; due to the COSMIC classification subsuming 5 bp deletions within the "5+" category it is not possible to show 5 bp deletions separately.* ***c) Explaining the COSMIC classification for >1 bp deletions in more detail.*** *Explanation with examples of 1> bp deletions within repeats. Deletions are shown in red, underlined, and the repeat context in bold; a generic example follows a specific one. 1) AG deletion not in a repeat context ("not in repeat", 1 repeat unit).*

2) AG deletion in an AGAG dinucleotide repeat couplet ("perfect repeat, short", 2 repeat units). 3) AG deletion in an $[AG]_6$ repeat ("perfect repeat, long", 6 repeat units. 4) AG deletion with a microhomology of 1: the downstream G matches a G immediately upstream of the deletion. This could also be described as an imperfect repeat. ***d) Representation of the 2-5 bp deletions in the Gene4Denovo dataset using the classification shown in c).*** *The largest group of 2-5 bp deletions are those in imperfect repeats (58.2%); the COSMIC plot in (b) shows that the largest groups within this category are 2 bp and 3 bp deletions with microhomology. The second largest group are those in short perfect repeats (27.9%).*

5.2.2 A HeLa fluctuation assay to detect an RNase H2 specific mutational signature

The high sensitivity of the reporter construct designed and tested in *S. cerevisiae* (Chapter 4) suggested it could have utility in demonstrating similar effects in human cells. HeLa cells were chosen as RNase H2 knockouts are viable (Zimmermann *et al.*, 2018), due to the suppression of p53 activity by the HPV protein E6 (Schwarz *et al.*, 1985; Werness, Levine and Howley, 1990). This means a reduction in the number of experimental steps to create a viable RNase H2 knockout cell line. Retinal pigmented epithelial (RPE1) cells, a more karyotypically normal (and stable) cell line were also considered, but I found them to be constitutively resistant to both hygromycin and puromycin as a result of immortalisation with hTERT (human telomerase reverse transcriptase) (Bodnar *et al.*, 1998; ATCC, 2020a) using a plasmid containing both these resistance cassettes (ATCC, 2020b).This precluded their use, in order to allow comparability with the *S. cerevisiae* experiments it was necessary to retain the modified hygromycin resistance gene; and in order to select cells quickly enough to make the selection experiments feasible we selected the rapidly acting antibiotic puromycin.

The generation of the HeLa reporter cell line (Figure 5.3) was based on a previously published method (Oceguera-Yanez *et al.*, 2016), utilising CRISPR-Cas9 technology to knock in the construct into the AAVS1 locus on chromosome 19 using homologous recombination.

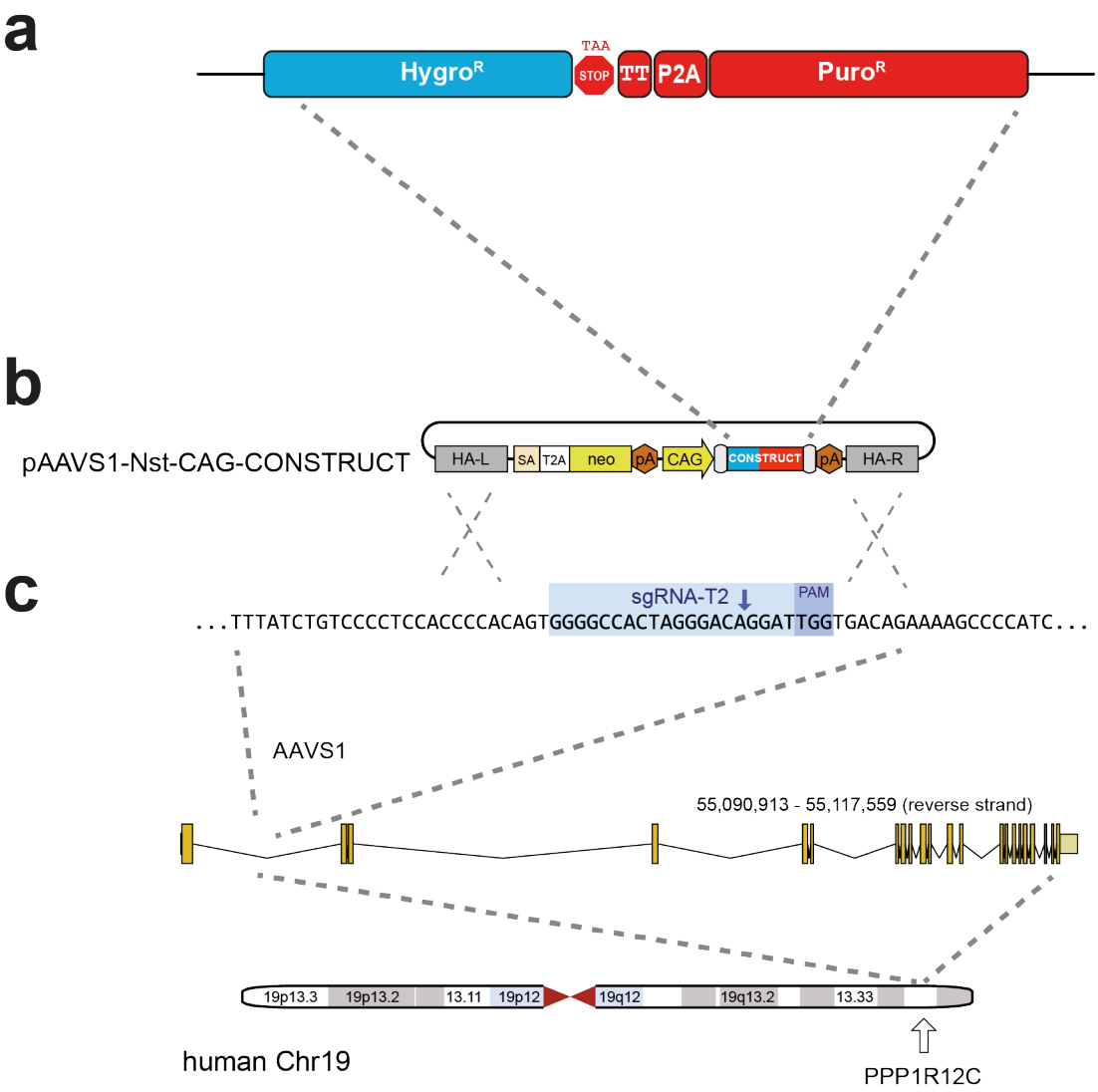


Figure 5.3. **Modification of the reporter construct used in S. cerevisiae and insertion at AAVS1 safe harbour locus using CRISPR-Cas9 into HeLa cells**. The approach used to introduce the construct into the human genome is based on that described by *(Oceguera-Yanez* et al.*, 2016)*. **a) Modification of the reporter construct for use in human cells.** The same reporter construct as used in S. cerevisiae, with 2 modifications: replacement of the hygromycin gene with the puromycin resistance gene (Puro^R), and removal of the PTEF promoter. **b) The construct is cloned into the pAAVS1-Nst-CAG vector.** This vector includes homology arms for the AAVS1 locus, a neomycin resistance gene followed by a poly-Adenine (pA) tail, and a constitutively active CAG promoter, which leads to transcription of the Hygro^R and Puro^R resistance genes. **c) CRISPR-Cas9 mediated insertion of the reporter in the AAVS1 locus by homology directed repair.** Cells were co-transfected with the pAAVS1-Nst-CAG-CONSTRUCT vector and a plasmid containing CRISPR-Cas9 with single guide RNA (sgRNA-T2). The sgRNA-T2 creates a double stranded break at the AAVS-1 locus, the location of a major hotspot for adeno-associated virus (AAV) integration. The sequence flanked by the left and right homology arms was inserted into AAVS1 through homologous

directed repair at the site of the double stranded break; the length of each homology arm is ~800 bp The AAVS1 hotspot is located in intron 1 of the protein phosphatase 1, regulatory subunit 12C (PPP1R12C) gene on human chromosome 19. Panels (b) and (c) adapted from *(Oceguera-Yanez* et al*., 2016)* under a Creative Commons licence *(CC BY 4.0, 2020)*.

This "safe harbour" locus was initially identified as a site where the adenovirus associate virus (AAV) (Kotin, Linden and Berns, 1992) integrated into the human genome. It has subsequently been widely utilised in cloning experiments as coding sequences integrated here are maintained in an open chromatin conformation (Ogata, Kozuka and Kanda, 2003) that allows persistent expression of these transgenes (Smith *et al.*, 2008).

For the reporter construct version 2 of the Hygro^R gene was used (Figure 4.5), as was the P2A self-cleaving peptide. The Neo^R gene was replaced by a puromycin resistance gene. The AAVS1 vector already contained a CAG promoter, which the Hygro^R-P2A-Puro^R was cloned downstream of (Figure 5.3).

5.2.2.1 Confirmation of insertion of reporter construct into HeLa cell lines

The construct was co-transfected into Hela cells (gift of Grant Stewart, University of Birmingham) using a CRISPR-Cas9 expressing plasmid (Figure 5 .4). Insertion at the AAVS1 locus was confirmed by PCR (Figure 5 .4a); clone A1 was selected for further experimental work. Fluorescent in situ hybridisation (FISH), conducted by Shelagh Boyle (MRC Human Genetics Unit, Edinburgh) was used to additionally confirm that the construct was incorporated at a single genomic location on one copy of Chromosome 19 (Figure 5 .4b) in clone A1. A positive control experiment using CRISPR-Cas9 targeting of the Hygro^R gene in the construct for clone A1 showed that these results in frameshift mutations and growth in puromycin (Figure 5.4c). Clone A1 then underwent CRISPR-Cas9 targeting of RNASEH2A, and clones for the subsequent fluctuation assay were selected using an RNase H2 assay. In

order to control for off-target effects of the CRISPR-Cas9, Clone A1.P1.G1 was selected as a wildtype control (RNASEH2A +), and clones A1.P1.E2 and A1.P2.H2 as representative RNase H2 null (RNASEH2A-KO) clones, as their RNase H2 levels approximated those seen in a cell line (#74, courtesy of Olga Murina, Jackson Lab) used previously (Zimmermann *et al.*, 2018) (Figure 5.4e). These clones were used for experiments to compare 2 bp deletion rates between WT and RNase H2 null mammalian cells.

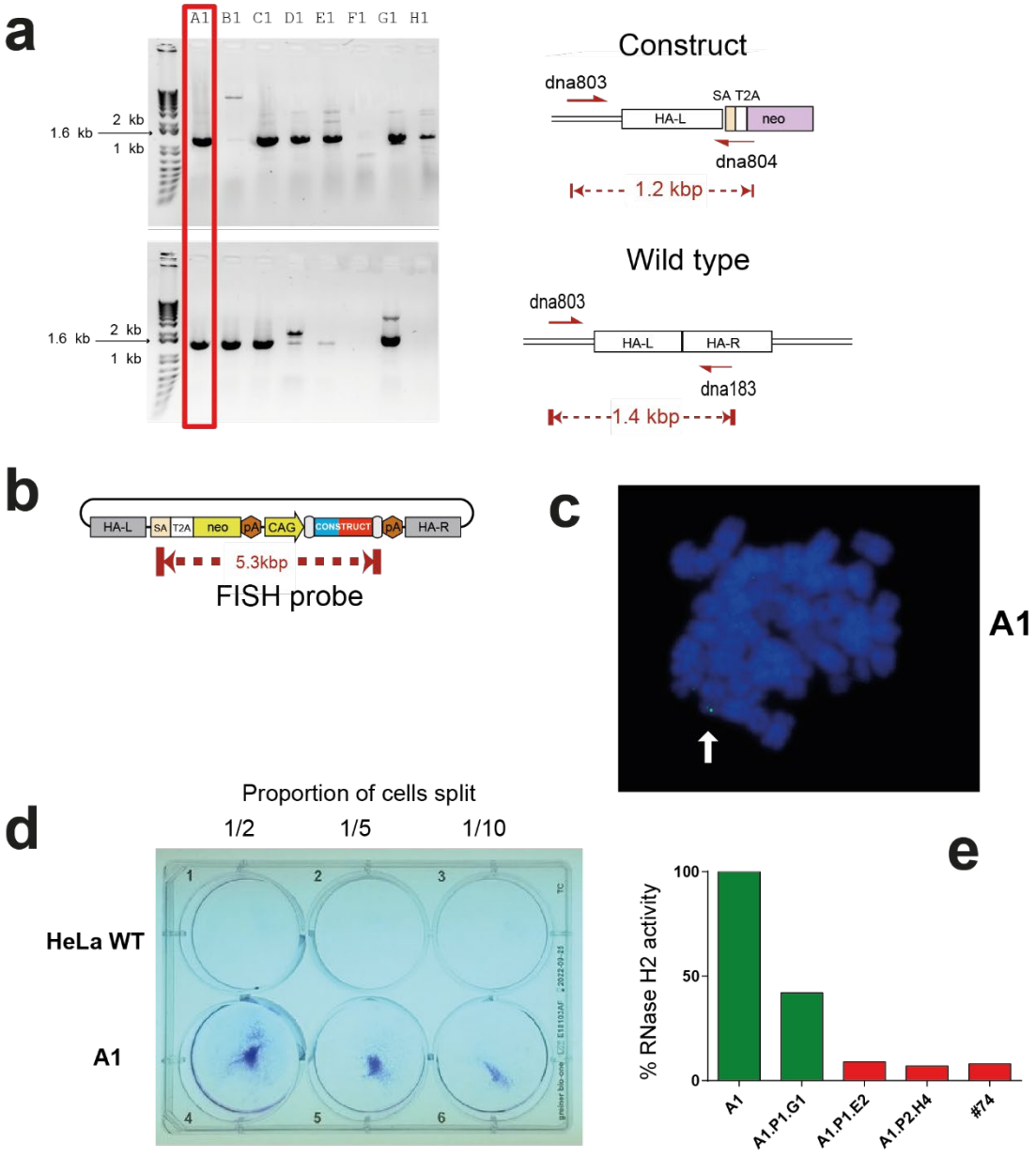


Figure 5.4.Validation of reporter construct in HeLa cells.a) Confirmation of insertion of the reporter construct using PCR. *PCR primers in flanking genomic DNA and within the neomycin gene (Figure 5 .3b) amplified a 1.2 kb fragment consistent with integration at the AAVS1 locus. Amplification of the WT locus was also possible for Clone A, which was taken forward into subsequent experiments. The schematic to the right shows the coordinates and expected size of the PCR product using primers dna803 and dna804 for detection of the construct, and dna803 and dna183 for confirmation of a wildtype AAVS1 locus* ***b) Schematic showing design of FISH probe.*** *The probe was designed so that only regions of the reporter construct without the homology arms were detected.* ***c)Confirmation of the reporter construct using FISH.*** *White arrow shows a single green probe for the reporter construct, indicating a single locus identified by FISH (DAPI, blue).* ***d) Positive control experiment for***

reporter construct using CRISPR-Cas9 targeting. *The plates on the top row were seeded with ½, 1/5 and 1/10 of a confluent culture from a single well of HeLa WT cells transfected with a CRISPR-Cas9 sgRNA targeting the Hygromycin resistance gene. The plates on the bottom row were seeded using an equivalent number of cells from a culture of HeLa cells with the reporter construct integrated. The cells were then treated with 14 days of puromycin, and fixed and stained with crystal violet. There are a number of surviving cells from the culture that contained the construct, in proportion to the initial number seeded, but none in the wildtype control.* ***e) RNase H2 assay to select clones for fluctuation assay.*** *The ancestral clone A1 was used a positive control for RNase H2 activity, with other levels represented as a percentage of this. All clones underwent CRISPR-Cas9 editing of the RNASEH2A subunit. The WT clone used for the fluctuation assays showed 42% of activity compared to the ancestral clone A1; those that had levels approximating a clone used previously as an RNase H2 knockout control (courtesy of Olga Murina) showed levels of 9 and 7% respectively. Panels (a) and (b) adapted from* (Oceguera-Yanez *et al.*, 2016) *under a Creative Commons licence* (CC BY 4.0, 2020).

5.2.3 A HeLa fluctuation assay shows an increase in 2 bp deletions in RNASEH2A-KO cell lines

High puromycin concentrations in culture media led to the eventual death of all RNASEH2A-KO cells, even those with 2 base pair equivalent frameshift mutations in the Hygro^{R} gene. Seeding the RNASEH2A-KO cells at higher plating densities ameliorated this effect, but led to resistance to puromycin activity in the RNase H2 proficient (RNASEH2A +) cells due to the formation of multiple layers of cells in the culture plates. Therefore a number of pilot experiments were required to determined optimal puromycin selection conditions and plating densities. The experiments presented in this section were conducted using puromycin for selection at a concentration of 0.5 µg/ml, and seeding 1,000 cells per selection plate. This combination of conditions led to survival of the RNASEH2A-KO clones, without development of resistance in the plates containing RNASEH2A + cells due to overgrowth and formation of multiple layers of cells.

Results from the optimised fluctuation assay showed an increased mutation rate in the RNASEH2A-KO clones, giving an estimate of a mutation rate 3.1x higher than in the RNASEH2A + control (Figure 5.5a and b), supporting the results seen in *S. cerevisiae*. When analysing mutant colonies (Chapter 2,

Section 2.3.5) sequencing traces indicated that the clones used in the experiments appeared to have developed two copies of the reporter construct, likely as a result of the high levels of genome instability in HeLa cells. To ensure that all experimental strains contained the same number of copies of the construct, qPCR was conducted by Martin Reijns, which showed equivalent levels for the single RNASEH2A + and 2 RNASEH2A-KO clones.

Examining the mutational spectra for all cell types (Figure 5.5c), 2 bp deletions are also seen in the RNASEH2A + clone, but alongside a wider variety of other mutation types. When rates of only 2 bp deletions are compared, there is a 4.4 fold increase in the RNASEH2A-KO clones, compared to WT (Figure 5.5d). The estimate of the absolute 2 bp deletion rate is similar for the 2 RNASEH2A-KO clones (A1.P1.E2 and A1.P2.H4, Figure 5.5e).

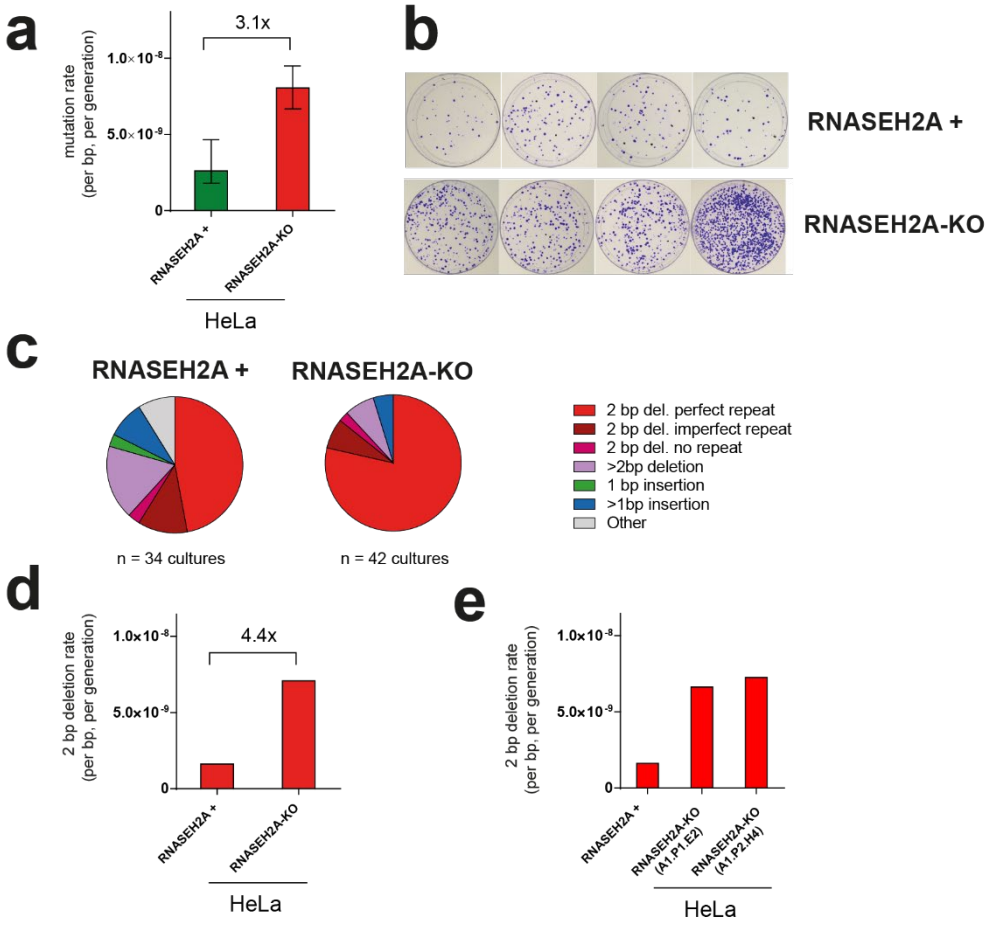


Figure 5.5. Two bp deletions are increased in HeLa cells in the absence of RNase H2. a) Frameshift mutations are significantly increased in RNASEH2A-KO strains. *RNASEH2A + represents HeLa cells with the reporter construct, and RNASEH2A-KO HeLa cells with RNASEH2A additionally inactivated by CRISPR-Cas9 genome editing. Data shown from 9 independent cultures from a single RNASEH2A + clone, and 16 independent cultures from 2 RNASEH2A-KO clones (10 from clone A1.P1.E2, and 6 from clone A1.P2.H4). The RNASEH2A + HeLa clone had also undergone CRISPR-Cas9 gene editing for RNASEH2A, but retained RNase H2 activity. Mutation rate with 95% confidence intervals calculated using the Lea-Coulson method of the median.* ***b) Representative final culture plates with colonies surviving puromycin selection fixed and stained using crystal violet.*** *The top row shows HeLa RNASEH2A + cells, and the lower row RNASEH2A-KO cells (n= 4 independent cultures from a single clone). All plates are from independent cultures.* ***c) The predominant mutation type in RNASEH2A-KO cells with the reporter construct are 2 bp deletions.*** *Plots showing mutational spectra of deletions detected by the mutation construct (single colonies from 34 independent cultures sequenced for RNASEH2A + clones, 42 single colonies from independent cultures for RNASEH2A-KO clones. Categories as defined in Figure 5.2; the "other" category encompasses single base substitutions which lead to a new start codon and a frameshift, complex mutations with single base substitutions and frameshift indels, and mutants where no mutations were detected in the Hygro^R gene. The mutations come from the experiments presented in a), and in addition cultures used in control experiments (using the clones from a) and others) and those used to in precursor assays to*

determine optimal antibiotic concentrations and cell plating densities. ***d) Overall two base pair deletion rate.*** *Calculated by multiplying proportion of mutations that were 2 bp deletion in (c) by overall mutation rates (a). There was a 4.4-fold increase in the 2 base pair mutation rate in RNASEH2A-KO cells compared to RNASEH2A + cells.* ***e) Two base pair deletion rate for each RNASEH2A-KO cell line****. Calculated as in d) for rates estimated from each of the 2 lines (n= 10 independent cultures for A1.P1.E2, n=6 for A1.P2.H4).*

5.2.4 An RPE1 human cell line mutation accumulation assay

To confirm that the initial findings in the reporter construct were applicable genome-wide, Martin Reijns designed and conducted a whole genome sequencing mutation accumulation experiment to identify whether the RNase $H2^{-/-}$ mutational signature could also be identified in RPE1 cells. Cas9-expressing hTERT-RPE1 cells were made $p53^{-/-}$ (necessary in order to be able to tolerate RNase H2 knockout) by the Durocher Lab (Lunenfeld-Tanenbaum Research Institute,Toronto)(Zimmermann *et al.*, 2018). Subsequently this cell line underwent CRISPR-Cas9 editing to create $p53^{-/-}$ RNASE $H2A^{-/-}$ and $p53^{-/-}$ $RNASEH2B^{-/-}$ cells. Starting cell lines for the experiment were selected by Martin Reijns by single cell cloning of these lines (2 x $p53^{-/-}$, referred to as WT; 1 x $p53^{-/-}$ RNASE $H2A^{-/-}$; 1 x $p53^{-/-}$ $RNASEH2B^{-/-}$) (Figure 5.6).

Cell lines were then grown in non-selective conditions for ~ 96 generations, and sequenced to identify mutations that had accumulated during this period; both the starting population (at t=0) and final population for each cell line was sequenced to identify mutations that had been generated during the experiment. David Parry (Jackson Lab) applied an algorithm used by the Jackson Laboratory (Materials and Methods, Section 2.3.3.2.2) to validate de novo mutations in patients with genetic disorders. Variants were filtered on parameters including read quality and depth, variant allele frequency, and the absence of reads supporting the variant in the ancestor to identify high confidence unique mutations for each cell line.

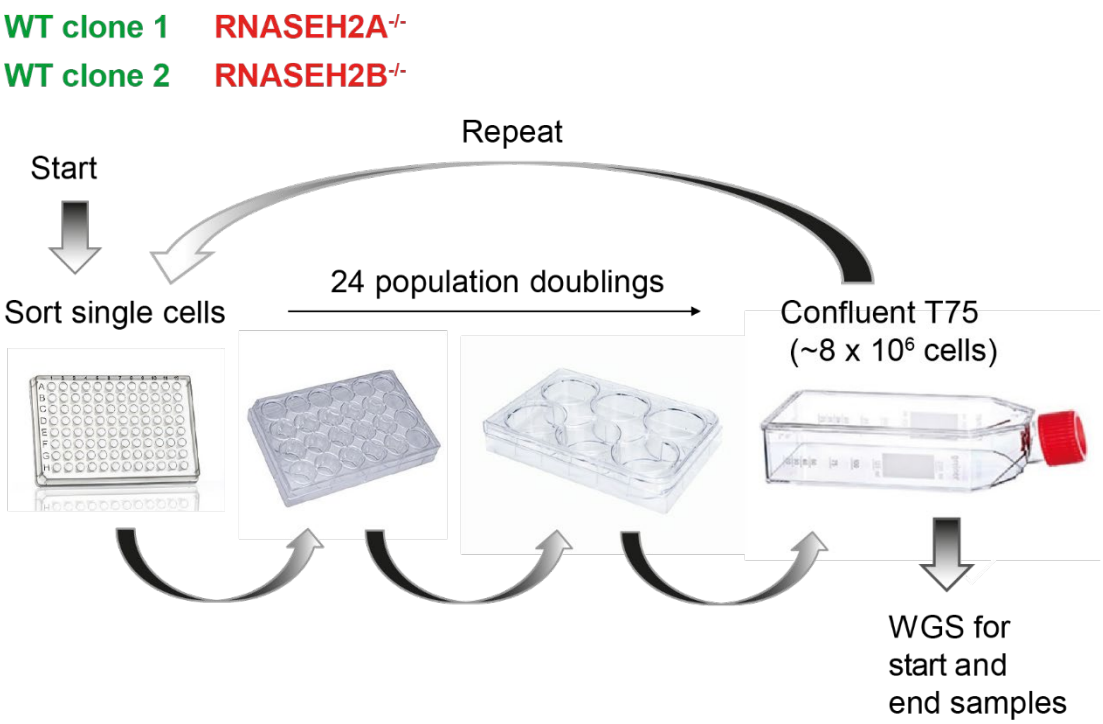


Figure 5.6. Design of a mutation accumulation experiment in RPE1 cells. *The experiment consisted of the growth of 4 clones over approximately 96 generations. 2 cell lines with p53 inactivated using CRISPR-Cas9 served as "wild type" controls for 2 RNase H2 null strains, one with the RNASEH2A gene inactivated using CRISPR-Cas9 (RNASEH2A$^{-/-}$), and the second with the RNASEH2B gene inactivated (RNASEH2B$^{-/-}$). Growth of each cell line was started by single cell sorting. Cells were transferred into progressively larger growth volumes (96 well plate -> 24 well plate -> 6 well plate -> T75 flask) and then single cell sorted using FACs once the flask had reached confluence to create a bottleneck for mutations that occurred over these ~24 population doublings. After 4 rounds, whole genome sequencing (WGS) of samples from the ancestral and descendent clones was performed on an Illumina short read platform. Experimental design and cell culture all the work of Martin Reijns.*

5.2.5 RPE1 mutation accumulation experiment confirms an increased rate of 2-5 base pair deletions in the absence of RNase H2.

Once unique descendent mutations had been identified, I assessed whether there was a difference in mutation patterns between the WT and RNase H2 null cell lines. There were a limited number of 2-5 base pair deletions in the WT lines (Figure 5.7c). Therefore to increase statistical power mutations for the 2 WT cells lines were pooled, and compared the totals for the WT and RNase H2 null cell lines. Results supported those seen in HeLa cells. In the absence of RNase H2, the number of 2-5 base pair deletions, but not those of other mutation types, was significantly increased when using a Fisher's exact test for enrichment (Figure 5 .7a and b). This increase in numbers of 2-5 bp deletions was seen in both the RNASEH2A$^{-/-}$ and RNASEH2B$^{-/-}$ cell lines (Figure 5.7c), in keeping with knockout of both subunits leading to overall enzyme loss and accumulation of genomic ribonucleotides.

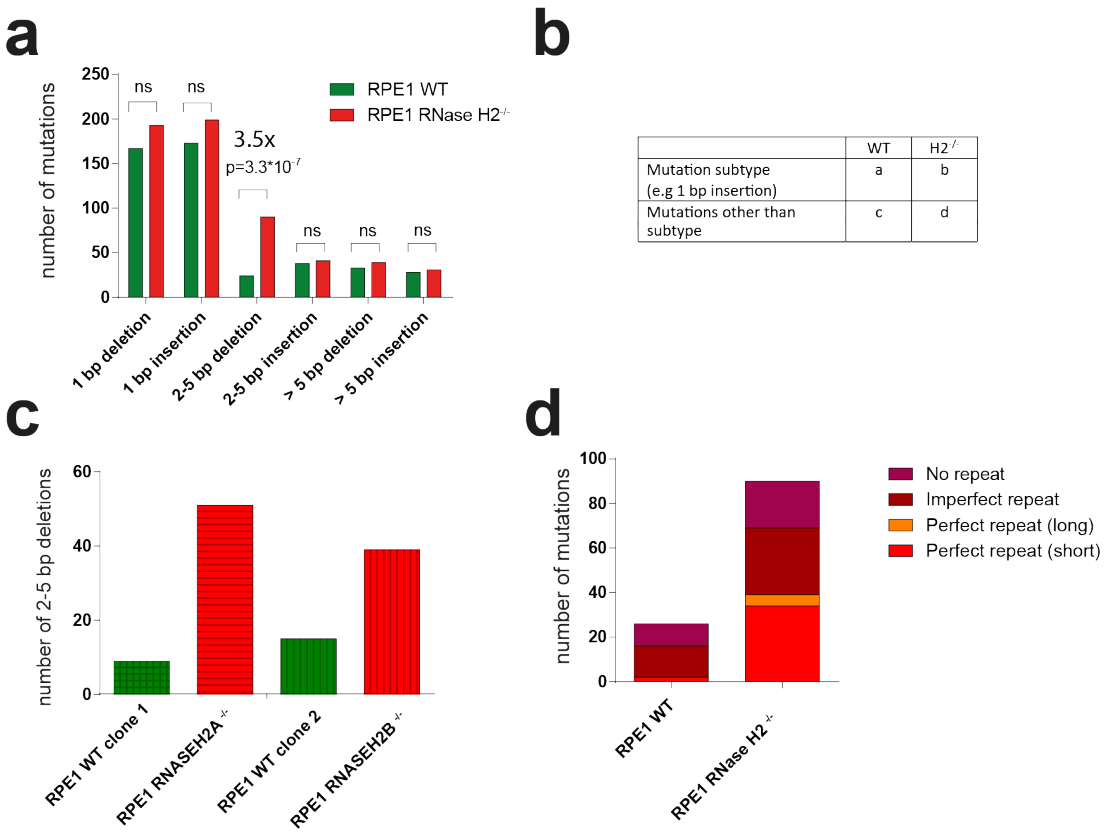


Figure 5.7. There is an increase in 2-5 bp deletions in RPE1 cells in the absence of RNase H2. a) Results from whole genome sequencing of RPE cells deficient in p53 only (RPE1 WT) compared to those with an additional loss of RNase H2 (RPE1 RNase H2$^{-/-}$). *Comparison of total number of mutations from 2 RPE1 WT clones over ~96 generations , compared to total of number of mutations in the 2 RPE1 RNase H2$^{-/-}$ cell lines also grown over ~96 generations. All cell lines grown in absence of selective pressure. Ancestral population and descendent population sequenced for each clone, and only mutations present in descendent population counted. The number of 2-5 bp deletions is 3.5-fold higher in the RNase H2$^{-/-}$ cell lines. Proportions of each type of mutation for each group (RPE1 WT and RPE1 RNase H2$^{-/-}$) compared using a Fisher's exact test. P-value robust to Bonferroni correction for 6 statistical tests (threshold for significance p = 0.008)* ***b) Set up of Fisher's exact tests.*** *For each mutation subtype, the proportion of each subtype within either the WT (a/c) or RNase H2$^{-/-}$ group (b/d) was compared, and a p-value calculated for the likelihood of the odds ratio not being equal to 1.* ***c) Comparison of the number of 2-5 bp deletions in the RNase H2$^{-/-}$ mutation accumulation clone compared to the WT controls. d) Types of 2-5 bp deletion for RPE1 WT and RPE1 RNase H2$^{-/-}$ cell lines.*** *For examples of deletions not in repeats, in imperfect repeats and in perfect repeats refer to Figure 5-2. The main type of 2-5 bp deletion in the H2 null cell lines are deletions in short perfect repeats, although all types of 2-5 bp deletions are more common in the RNase H2$^{-/-}$ than WT.*

In the RNase $H2^{-/-}$ cell lines, the most common 2-5 bp deletion type are deletions in short perfect repeats, but all 2-5 bp mutation types are more common than in WT(Figure 5.7d). This suggests that the absence of RNase H2 is associated with an increase in all 2-5 bp deletion mutation types. As from my work and others' in *S. cerevisiae*, deletions in an RNase H2 null context are Top1 dependent, suggesting that Top1 mutagenesis could explain many short deletions seen in humans in perfect *or* imperfect repeats. To establish this likely contribution further, I set out to compare the signature I had identified to that for the postulated most common cause of short deletions: deficiencies in mismatch repair.

5.2.6 Identifying COSMIC signatures in MMR deficient S. cerevisiae and human cell line data

COSMIC had already reported several mutational signatures as being associated with deficiencies in mismatch repair in cancer genomes. These include ID2 (Figure 5.1b), that is distinct to that which I had observed above. However, for more direct comparability I wanted to examine the signatures seen in cell line experiments equivalent to the RPE1 mutation accumulation one described above, and compare them to what was seen with RNase H2 deficiency, both in human cell lines and *S. cerevisiae*.

5.2.6.1 Description of datasets

5.2.6.1.1 S. cerevisiae mutation accumulation experiments

As described in Chapter 4, in the last few years the Kunkel Lab have published the results of a number of mutation accumulation experiments in *S. cerevisiae* looking at the effects of loss of mismatch repair machinery (Lujan *et al.*, 2014) or ribonucleotide misincorporation (Conover *et al.*, 2015) on mutational spectra. I chose to re-analyse their data for a pol2-M644G Pol ε mutator strain in which the mismatch repair enzymes MSH2, or MSH3 and MSH6, had been

knocked out (*GEO Accession Viewer GSE56939*, 2014), and called mutations as described in Material and Methods (Section 2.3.3.2.1). This was to be able to compare the results to those for the same mutator strain in which RNase H2 had been knocked out (Conover *et al.*, 2015). As the COSMIC mutational signature software cannot analyse bespoke reference genomes (in this case the assembly used by the Kunkel Lab to analyse their mutational accumulation experiments) I modified the script I had used to analyse the human DNMs dataset to create the COSMIC indel plot (scripts available at GitLab, "/COSMIC/generating_COSMIC_plot_RNaseH2_null_S.cerevisiae.py" and "/COSMIC/generating_COSMIC_plot_MMR_def_S.cerevisiae.py").

5.2.6.1.2 MLH1 null colonic epithelial cell line dataset

As a comparison for the RPE1 mutation accumulation experiment, I identified an experiment in which the MLH1 gene, a component of the mismatch repair pathway, had been inactivated in a colonic epithelial cell line. Subsequently, control organoids with wildtype MLH1 and mismatch deficient ($MLH1^{-/-}$) organoids were grown and underwent whole genome sequencing (Drost *et al.*, 2017). VCF files from this whole genome sequencing were kindly shared with me by one of the authors (Professor Ruben van Boxtel, UMC Utrecht). The WT and $MLH1^{-/-}$ organoids in the experiment had been grown for the same time period, so I merged the vcf files for 3 WT organoid WGS, and those for 3 $MLH1^{-/-}$ organoids, to compare them to one another (scripts available at GitLab, "/COSMIC/generating_COSMIC_plot_MLH1_null_organoids.py" and "/COSMIC/generating_COSMIC_plot_WT_organoids.py").

5.2.7 The majority of indels resulting from MMR deficiency are 1 bp deletions in T/A homopolymer tracts

For both MMR deficient *S. cerevisiae* and human colonic epithelial cells, the predominant mutation type when visualised using the COSMIC schema are 1 bp deletions, and particularly 1 bp deletions in T/A tracts of repeat length 5 or more (Figure 5.8a and b). This pattern corresponds to COSMIC signature ID2 (Figure 5.2b). The second largest group of mutations are 1 bp insertions in T/A tracts of length 5 or more (*ID1 - COSMIC Mutational Signatures*, 2020; Alexandrov *et al.*, 2020).

When each mismatch repair deficient cell line is compared to its control, it can be seen that the mutational spectrum shifts from a variety of short deletions to a predominance of 1 bp deletions. This is in keeping with the Streisinger polymerase slippage model (Streisinger *et al.*, 1966) (Figure 1.7). Here, slippage and bulging out on the template strand on a run of As or Ts (where the bonds between paired nucleotides are more likely to be broken, due to the 2 hydrogen bonds connecting them, rather than the 3 connecting C:G (Figure 1.1). This slippage and bulging out leads to a 1 bp deletion on the complementary strand.

When comparing the mutational signature for each mismatch repair deficient cell line to its equivalent wildtype control, it can be seen that the mutational spectra for both the pol2-M644G *mmr-* and $MLH1^{KO}$ represents a change from a combination of short indel types in the wildtype control, to a predominance of 1 bp deletions (Figure 5.8c). I conclude that the MMR signature is distinctive and comparable in both *S. cerevisiae* and human cells, but does not explain the 2-5 bp deletions seen in human populations.

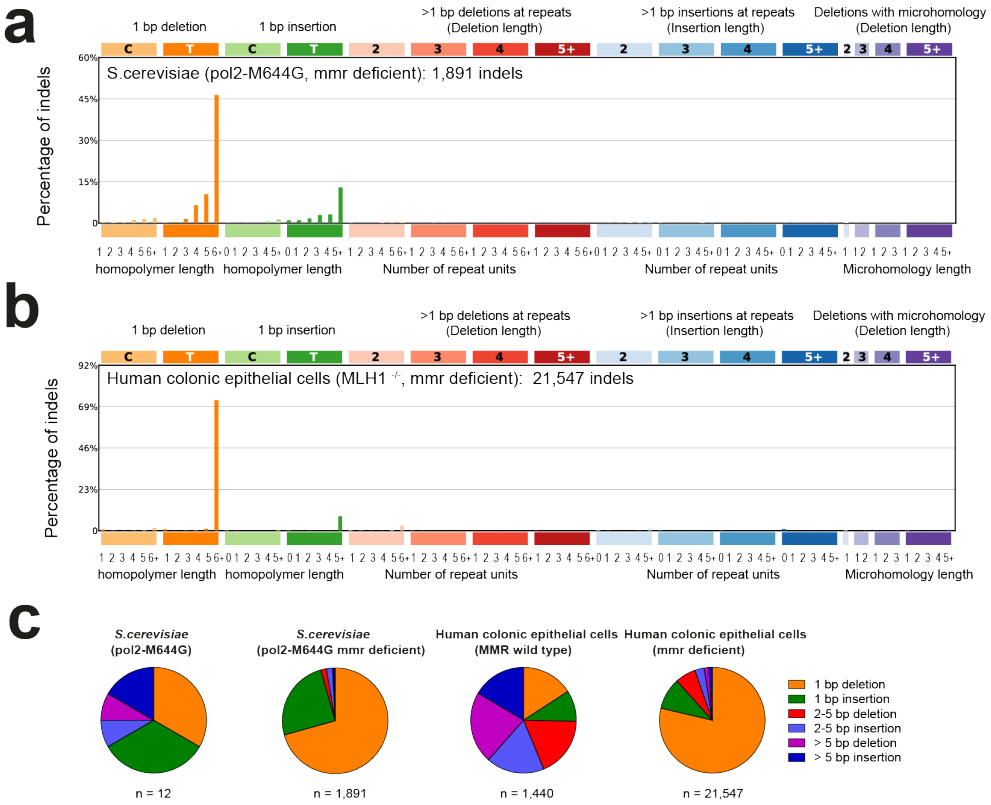


Figure 5.8. The predominant mutational signature in mismatch repair deficiency in S. cerevisiae and human cells are 1 base pair deletions. a) A 1 bp deletion signature in S. cerevisiae with deficient mismatch repair. *A COSMIC plot of short indels (<50bp) from my re-analysis of a whole genome sequencing (WGS) mutation accumulation experiment* (Lujan *et al.*, 2014)***,*** *showing that the main indel mutation type in mismatch repair machinery deficient cells is a deletion of a T (equivalent to an A) in a homopolymeric run of >5 As/Ts. WGS is shown for a pol2-M644G mutator strain to allow comparison to a pol2-M644G strain with knockout of RNase H2 (Figure 5 .9).* ***b) A 1 bp deletion signature in human cells with deficient mismatch repair****. COSMIC plot for WGS from a mutation accumulation experiment in human intestinal colonic epithelial cells. Again, the predominant signature is 1 bp deletions in T/A homopolymer runs. My re-analysis of an experiment conducted by* (Drost *et al.*, 2017)*; unpublished vcf files kindly shared by Professor Ruben van Boxtel, UMC Utrecht.For a) and b) approach to classifying indels based on that of the COSMIC consortium* (Alexandrov *et al.*, 2020)*; plots created using the SigProfilerPlotting function* (Bergstrom *et al.*, 2019)*. Fastq reads downloaded from GEO and processed as described in Methods.* ***c) When compared to the mutational spectra in controls, the loss of mismatch repair results in a shift in mutational spectra towards 1 bp deletions.*** *Comparison of proportions of types of short indel MMR wild type vs MMR deficient strains in S. cerevisiae and human cells. In both S. cerevisiae and colonic epithelial cells with intact mismatch repair machinery, there is a variety of different types of mutations, including insertions and deletions. When mismatch repair machinery is lost longer indels, including 2-5 base pair deletions, become proportionally much less common.*

5.2.8 The COSMIC profile for RNase H2 deficiency

Given the COSMIC profile for MMR deficiency does not match the patterns for 2-5 bp deletions seen in human DNMs, I next examined whether Top1 dependent mutagenesis could account for these mutations. The mutational profile when RNase H2 is knocked out in *S. cerevisiae* is clearly different to that seen with the loss of components of the mismatch repair pathway (Figure 5.9a). Instead of an increase in 1 bp deletions, the main mutation group are short deletions, most of which are in a repeat context of 2 or more repeat units. These are mutations that correspond for instance to an AG deletion from $[AG]_2$ repeat (Figure 5.2c).

When the short indels seen in the RPE1 WT ($p53^{-/-}$) mutation accumulation experiment are subtracted from those identified in the RPE1 RNase $H2^{-/-}$ cell lines, to remove RNase $H2^{-/-}$ independent events, a similar profile is seen (Figure 5.9b), with the most common mutations being 2 bp deletions in a repeat context of 2 units. Distinct from yeast, what are also common are what COSMIC classifies as 2bp deletions with microhomology length = 1.The COSMIC classification for indels is a new one; as the authors acknowledge "we recognise that different classifications of IDs may be preferred by others"(Alexandrov *et al.*, 2020). The most challenging category is this one of deletions with microhomology, as this presupposes an aetiological mechanism: a failure of homologous recombination, and repair by NHEJ or MMEJ. However, the minimum length of microhomology needed for either of these 2 mechanisms remain unclear (Ottaviani, LeCain and Sheer, 2014b). The situation in which these mechanisms are most likely to come into play are with failure of homologous recombination, as is seen in BRCA1/2 deficient tumours. The indel mutation type seen most commonly in this context are long (5+ base pairs) in regions with microhomology (COSMIC ID6) (Davies *et al.*, 2017; Alexandrov *et al.*, 2020). The COSMIC consortium have also identified another signature (ID8) leading to non-homologous DNA end-joining activity,

but again this signature is characterised by long (5+ base pair) deletions in regions with and without microhomology (Alexandrov *et al.*, 2020).

However, these mutations with short microhomologies that do not fit with either the pattern for MMR deficiency or failure of homologous recombination can also be conceptualised as imperfect repeat (Figure 5.2c). The 2-5 bp deletion profile in RNase $H2^{-/-}$ cells is similar to that of the COSMIC ID4 signature (Figure 5.9c), for which no cause has yet been attributed.

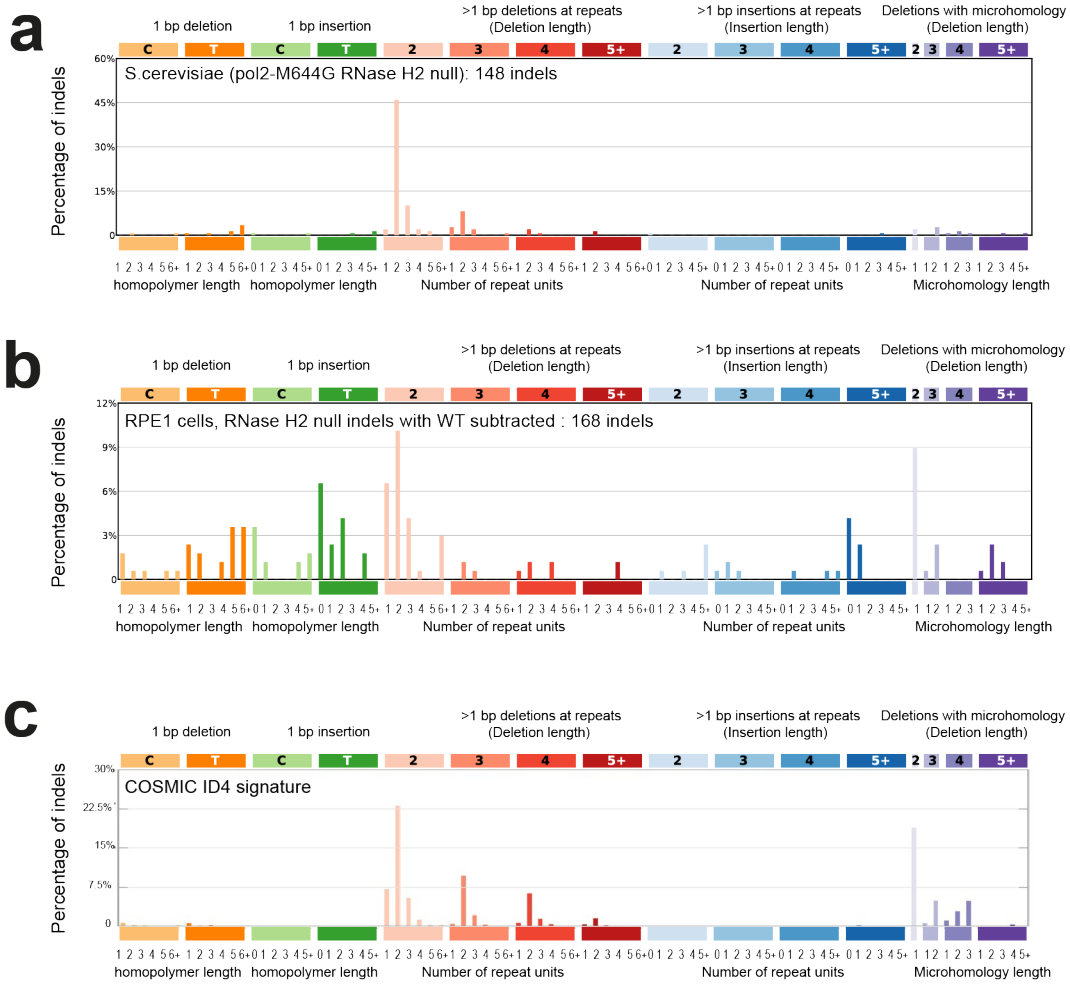


Figure 5.9. COSMIC plots for RNase H2 deficient S. cerevisiae and RPE1 cell lines show similarities with the COSMIC ID4 signature. a) COSMIC plot for S. cerevisiae pol2-M644G mutation strain null for RNase H2 (rnh201Δ). *The predominant mutation type is short (2-5bp) deletions in a repeat context. My re-analysis of data from* (Conover *et al.*, 2015). *Unlike for b), the number of generations for the control and RNase H2 knockout are not equivalent and control mutations cannot be subtracted to produce a more RNase H2 null specific signature; however the number of these is small (12, Figure 5 .8b) with no 2-5 bp deletions. b)* ***The mutational signature for RNase H2 deficiency in an RPE1 human cell line****. Here, the two main mutation groups are 2 bp deletions in short repeats, and 2 bp deletions with a micro-homology of 1 nt. For each mutation category, the total number of mutations in the p53$^{-/-}$ WT control cell line was subtracted from the total in the RNase H2 null cell line; negative values are not plotted. Cell lines were grown for an equal number of generations (see Figure 5 .6). Plots for a) and b) created using SigProfilerPlotting function* (Bergstrom *et al.*, 2019). *c)* ***Mutational signature ID4 from the COSMIC analysis of 4,645 whole-genome and 19,184 exome sequences.*** *Figure adapted from* (Alexandrov *et al.*, 2020) *under a Creative Common Licence* (CC BY 4.0, 2020). *Again, the predominant mutations are short deletions in a repeat context, and 2 bp deletions with a microhomology of 1 nt, equivalent to an imperfect repeat.*

5.2.9 Proposing a model for short indels in human genomes

The observations of increased rates of short deletions in perfect repeats (in *S. cerevisiae*, HeLa cells, and RPE1 cells) and imperfect repeats (*S. cerevisiae* WGS, RPE1 cells) leads me to propose a modification for the existing model for topoisomerase1 (Top1) mediated generation of short indels in mammalian genomes (Figure 5 .10). As described in Chapter 1 (Section 1.2.4), an elegant biochemical study (Sparks and Burgers, 2015) has shown that cleavage of a single ribonucleotide by Top1 can lead to the formation of a difficult to resolve 2'3' cyclic phosphate intermediate. Further cleavage by Top1 2 base pairs upstream of the initial nick leads to the release of the 2'3' cyclic phosphate. Sparks and Burgers in their *in vitro* study found that Top1 created a nick 2 bp upstream of the initial ribonucleotide in 86.7% of cases (Sparks and Burgers, 2015), with second cleavages at 3 (5.0%), 4 (7.4%) , and 5 (1.0%) nt upstream being less frequent. No second cleavage 1 nt upstream of the ribonucleotide was observed. In addition, they found that a G-C base pair (with 3 rather than 2 hydrogen bonds) was required to generate a slippage re-alignment product with enough stability for re-ligation. Given that 2-5 bp deletions in *imperfect* repeats are the ones most commonly seen in the RPE1 mutation accumulation experiment, I propose that these may be sufficient to create a template for stabilisation and repair, and result in a short deletion, most commonly 2 bp long, due to the tendency of Top1 to cleave most frequently at this position.

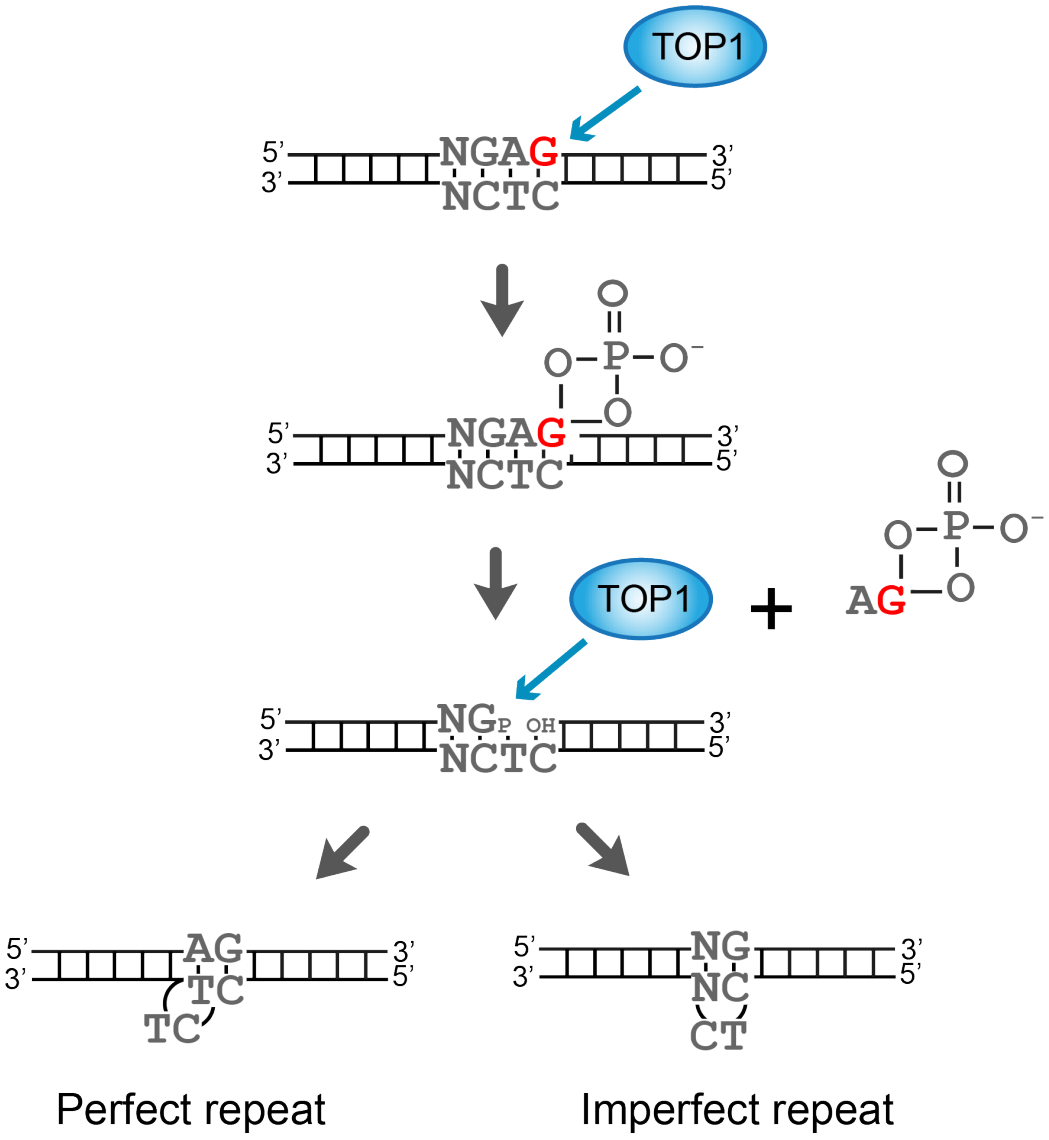


Figure 5.10. Proposed model for Top1 mediated mutagenesis at perfect and imperfect repeats. *Based on* (Sparks and Burgers, 2015). *An embedded ribonucleotide (red G) is cleaved by Top1 (TOP1); however the presence of a 2'-OH group leads to the formation of a 2',3'-cyclic phosphate and dissociation of TOP1. TOP1 cleaves 2 nt upstream of the original incision, releasing the 2',3'-cyclic phosphate and creating a 2 nt gap. If the ribonucleotide is located in a perfect tandem repeat (N on the template strand = A) there is slippage at the upstream TC on the complementary strand and re-ligation, leading to resolution of the gap but a 2 bp deletion in the cells descending from the template strand. If the nucleotide is located in an imperfect tandem repeat (N on the template strand ≠A), the central 2 nt on the complementary strand may bulge out, leading to binding of the most downstream C in the dinucleotide couplet repeat to the upstream G, and re-ligation, again resulting in a 2 nt deletion on the template strand.*

This model makes several predictions. Firstly, that short deletions will be predominantly 2 bp in length. Secondly, that although they will be concentrated in imperfect or perfect repeat sequences, that the frequency of deletions will not be related to the length of a repeat. Thirdly, that the deletions will be concentrated in imperfect or perfect repeat sequences which contain a G-C bond.

5.2.10 Do the 2-5 bp deletions seen in RNase H2 deficiency fulfil the predictions from the model?

The short deletions seen in human *de novo* mutations are most commonly proposed to be due to a failure of mismatch repair. I therefore set out to investigate the predictions of the model in relation to datasets for MMR and RNase H2 deficiency that I assembled for *S. cerevisiae* and human cells. For these datasets, I examined the length of the deletion, the length of repeat in which the deletion had occurred, and the G-C content of the deletion and its repeat context.

5.2.10.1 Deletion length

The length of deletions for MMR and RNase H2 deficient S. cerevisiae and human cells can be seen from previous analyses (Figure 5 .8 and Figure 5 .9). For MMR deficiency, the most common short indel type are 1 bp deletions (Figure 5 .8a and b) , whereas for RNase H2 deficiency the most common type are 2 bp deletions (Figure 5 .9a and b).

5.2.10.2 Length of repeat

The number of 2-5 base pair deletions for which to examine the repeat context for MMR deficient *S. cerevisiae* strains was small. I therefore curated a dataset of short 2-5 bp deletions from published (Lujan *et al.*, 2014; Serero *et al.*, 2014) and unpublished (Lang, Parsons and Gammie, 2013) (mutations shared by Professor Alison Gammie, NIH) datasets. I examined the types of mutations seen in these, classifying them as being not in a repeat, in an imperfect repeat,

or in a short (<5 units) or long (≥5 units) perfect repeat (Figure 5 .2c). I found that the majority of 2-5 deletions in this dataset were located in long perfect repeats (Figure 5 .11a). In comparison, data from the mutation accumulation experiment in RNase H2 deficient *S. cerevisiae* (Conover *et al.*, 2015) showed that the majority of 2-5 base pair deletions were located in short perfect repeats (Figure 5 .11a). For the same comparison in human cells, I took the mutations shared for MLH1 $^{-/-}$ intestinal epithelial organoids (Drost *et al.*, 2017) and performed a similar analysis. In order to arrive at an approximation of an MMR deficiency specific signature, I subtracted the mutations seen in WT cells from the MLH1$^{-/-}$ cell line, as they were grown for an equivalent number of generations. Whilst the results seen are not as dramatic as in *S. cerevisiae*, the majority of 2-5 bp deletions are those in long repeats (Figure 5.11b). For RNase H2, on the other hand, the most common mutation type seen in the RNase H2 null specific signature (again, mutations in the WT control were subtracted) are deletions in short perfect repeats (Figure 5.11b).

5.2.10.3 Sequence context

In order to examine the sequence context of mutations, I examined the subset of 2 bp mutations, which can be subsumed into categories for comparison between strains/cell lines to look at the sequence context of mutations (Table 5 -1). In MMR deficient *S. cerevisiae* cell lines, the predominant category of 2 bp deletions is loss of an AA-AT (Table 5.1) in a long repeat (Figure 5.11c), again consistent with Streisinger's polymerase slippage model (Streisinger *et al.*, 1966). In the *S. cerevisiae* RNase H2 null mutation accumulation experiment however the most common deletion is a CT-CA equivalent deletion in a short repeat, supporting the model proposed previously (N. Kim *et al.*, 2011; Sparks and Burgers, 2015) (Figure 5.11c).

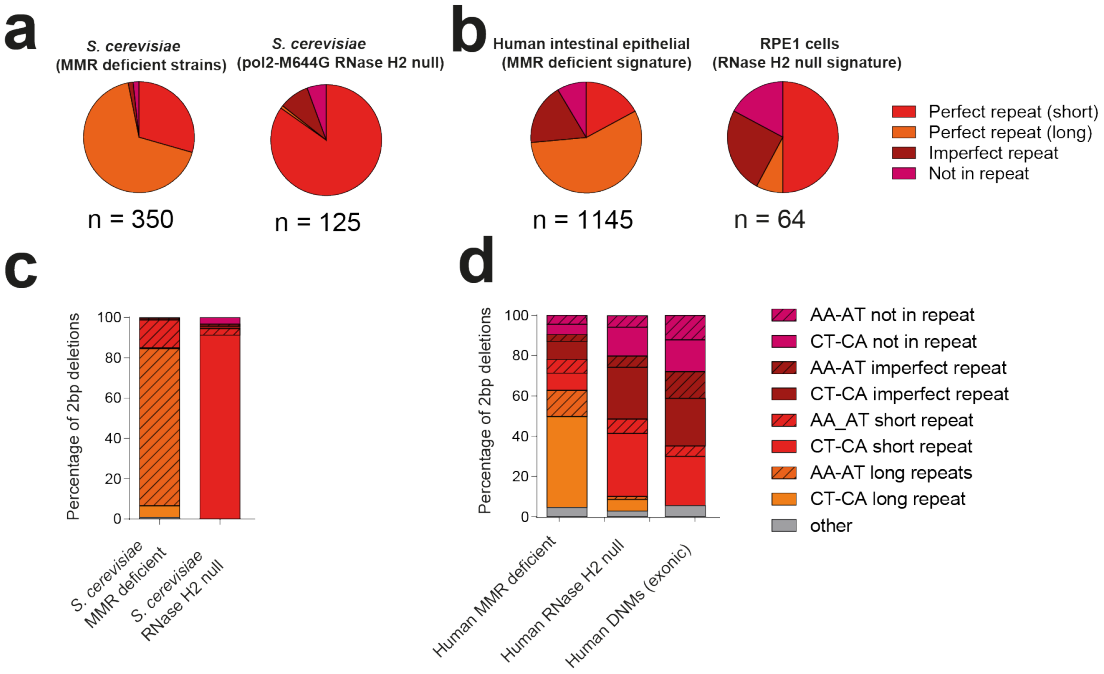


Figure 5.11. The pattern of 2-5 bp deletions seen in human de novo mutations more closely approximates an RNase H2 mutational signature than that of MMR deficiency. *a) MMR deficiency in S. cerevisiae leads to 2-5 bp deletions in long repeats, whilst RNase H2 deficiency leads to deletions in short repeats. Proportion of 2-5 bp deletions subtypes 1-4 from (a) for MMR and RNase H2 deficient S. cerevisiae mutation accumulation experiments. For both Saccharomyces MMR deficient lines, the predominant category of 2-5 bp deletions is a deletion in a long (≥5) perfect repeat. For RNase H2 deficient Saccharomyces and RPE1 cells, the most frequent type of mutation is a deletion in a short (≤5) perfect repeat sequence. For Saccharomyces 2-5 bp deletions in MMR deficient cells lines, mutations pooled from 3 studies (Lang, Parsons and Gammie, 2013; Lujan et al., 2014; Serero et al., 2014). For mutations in RNase H2 deficient Saccharomyces (pol2-M644G mutant line) my analysis of data from (Conover et al., 2015) is shown. b) In human cell lines, the predominant 2-5 bp deletions in an MMR deficient cell lines are those in long repeats, whilst in an RNase H2 deficient cell line they are those in short repeats. For mutations in MMR deficient human cell line, my analysis of unpublished data from (Drost et al., 2017), courtesy of Prof. Ruben van Boxtel, UMC; mutations in the WT control have been subtracted from those deficient in MMR (cells grown for equivalent number of generations). For RPE1 mutations, mutations within each mutational*

class for the WT control subtracted from the total number in the RNase H2 null cell lines (cells grown for equivalent number of generations).Experimental work for RPE1 cells conducted by Martin Reijns. c) The main class of 2 bp deletions in MMR deficient S. cerevisiae are AA-AT equivalent deletions in long repeats, whilst with RNase H2 deficiency they are CT-CA equivalent deletions in short repeats. For explanation of deletion types see Table 5 -1; mutations are subset of datasets in (c). d) The 2 bp deletions in a human DNM dataset more closely approximate an RNase H2 deficiency cell line than an MMR deficient one. The most common 2 bp deletion type in MMR deficient human cells are CT-CA and AA-AT equivalent deletions in long repeats, wherease in RNase H2 null RPE1 cell lines predominant type of mutations are CT-CA deletions in short perfect, imperfect and not in a repeat context. In exonic human de novo mutations , deletions within long perfect repeats are uncommon, and the most common mutation types are again CT-CA deletions in short perfect, imperfect and not in a repeat context. e) A cosine similarities analysis shows that exonic human DNMs most closely resemble those seen in human RNASEH2A-KO deficiency. DNMs from Gene4denovo database (Zhao et al., 2020); remainder of mutations subset of datasets in (c).

In the human experiments in the MMR deficient cell lines the predominant 2 bp deletion subtype is again deletions within long repeats (Figure 5 .11d); however these are now within CA-CT equivalent repeats rather than AA-TA equivalent repeats. For the RNase H2 null RPE1 cell line, the predominant 2 bp deletions are CA-CT equivalent mutations within short repeats.

***Table 5.1.* Categories of 2 bp deletions including repeat context**. Repeat context is shown in bold, and the deleted sequence in red, underlined.

Category	Deletions subsumed	Example
AA-AT not in repeat	AA,TT,AT,TA	**CC**AA**CC**
AA-AT imperfect repeat		**AT**AA**CC**
AA-AT short repeat		**AA**AA**CC**
AA-AT long repeat		**AAAAAAAA**AA**CC**
CT-CA not in repeat	CT,TC,AC,CA,GA,AG,GT,TG	**TT**AG**TT**
CT-CA imperfect repeat		**TG**AG**TT**
CT-CA short repeat		**AG**AG**TT**
CT-CA long repeat		**AGAGAGAG**AG**CC**
Other	CC,GG,CG,GC	**TT**CG**TT**

Returning to the human DNMs, I conducted a cosine similarities analysis comparing the mutational spectra for all 5 datasets (see Chapter 4, Section 2.3.5.1 for more detailed explanation of the methodology). This showed that the human *de novo* 2-5 bp deletions were most closely related to the human RNase $H2^{-/-}$ mutations (cosine similarity 0.92), and after this to the S.

cerevisiae RNase H2 null mutations (cosine similarity 0.72). Determining causality for this category of 2-5 bp deletions in the human population is clearly challenging, but based on the data presented so far, Top1 mediated short deletion mutagenesis (as evidenced from the mechanistic studies in *S. cerevisiae* presented in Chapter 4) is a more likely explanation for these, rather than the failure of mismatch repair to detect and repair polymerase slippage events.

5.2.11 Examining CLL genomes for evidence of an RNase H2 associated mutational signature

Whilst the raw sequencing data for the PanCancer Analysis of Whole Genomes (*PCAWG | ICGC Data Portal*, 2020) is hosted by a variety of servers globally, the summary data for each tumour sample in the ID83 format is available from the central ICGC data portal. In addition, for the 1950 whole genome sequenced tumour samples ICGC also makes available an analysis of copy number variants for all coding genes (*Copy Number Variants | ICGC Data Portal*, 2020). Interrogating this dataset, I identified 7 tumour samples with loss of one of the 3 RNase H2 subunits (A, B or C; Figure 5.12a). The most common tumour type with loss of RNase H2 was CLL, with 3 samples losing the RNASEH2B subunit (Figure 5.12a). This small number of RNase H2 null tumours is likely an under-estimate, given the difficulty of identifying biallelic deletions in whole genome sequencing, due to the often heterogenous cellularity of tissue samples, e.g. (Li *et al.*, 2017).

Analysing the dataset of ID83 summary counts for the CLL tumours, the mean number of 2-4 bp deletions (as described previously, the ID83 system does not allow extraction of 2-5 bp deletions) was higher in the RNASEH2B homozygous null tumours than those annotated as having a copy number of 2 or more (Figure 5.12c).

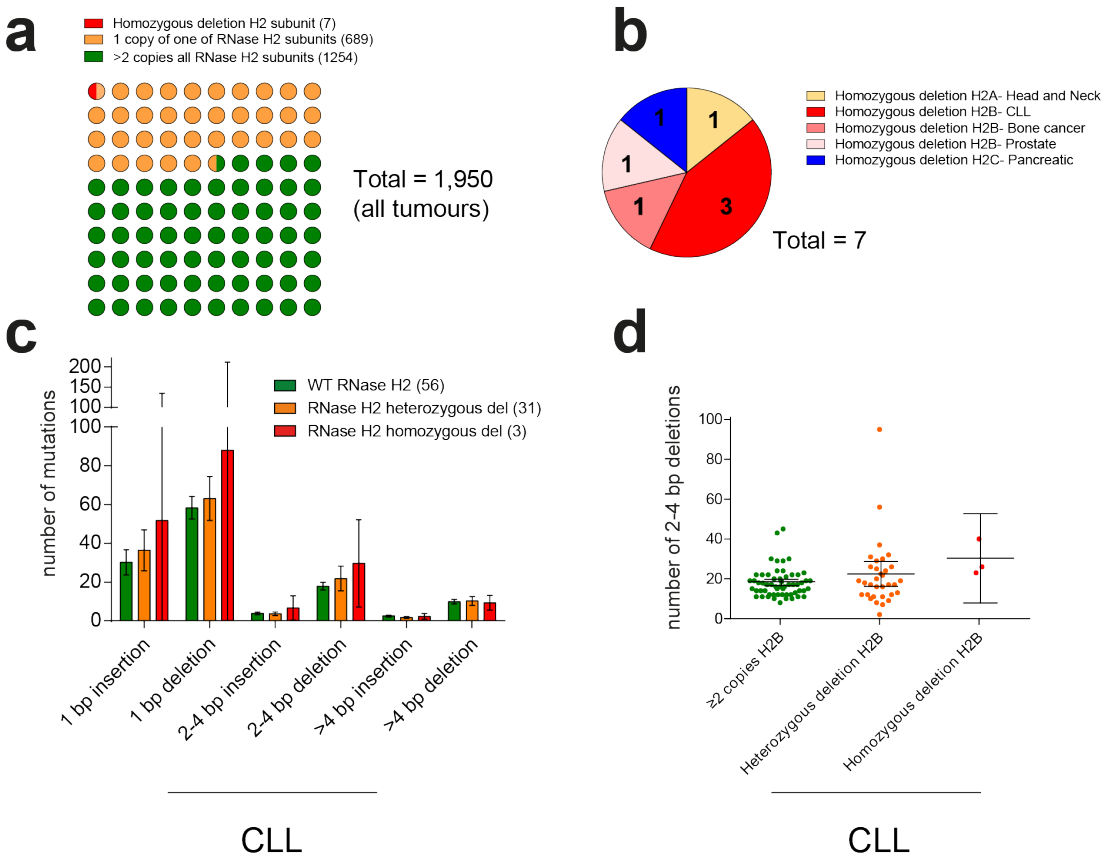


Figure 5.12. Examining the PCAWG dataset for evidence of the mutational signature associated with RNase H2 deficiency. *Copy number variant (CNV) from the PCAWG dataset (PCAWG | ICGC Data Portal, 2020) was examined for evidence of homozygous deletions of any of the 3 RNase H2 subunits: A, B or C.* ***a)Out of the 1,950 whole genome sequences available with CNV data, 7 were annotated as having a complete homozygous deletion of one of the subunits****. **b) The most common cancer type with homozygous deletion of an RNase H2 subunit was chronic lymphocytic leukaemia (CLL). c) Analysis of PCAWG mutational signature data for CLL dataset shows no significant difference for number of 2-5 bp deletions in homozygous RNase H2 samples compared to heterozygous or wildtype RNase H2 tumours****. 3 CLL tumours annotated as having a homozygous knockout of RNASEH2B, 31 tumours annotated as having a heterozygous deletion (either through complete loss of the gene, or a structural variant affecting the coding sequence), and 56 tumours annotated as having 2 or more intact copies of the subunit. The mean rate of 2-4 bp (using the data available it was not possible to categorise mutations as 2-5 bp) deletions is higher in the RNASEH2B null tumours, but with wide and overlapping confidence intervals.* ***d) Individual data points (number of 2-4 bp deletions) for each tumour.*** *Means with 95% confidence intervals calculated using the sample mean and lower and upper Gaussian confidence limits based on the t-distribution.*

However, the confidence intervals are wide and overlapping. A detailed plot of 2-4 bp deletions shows some marked outliers in the distributions for both the wild type and heterozygous tumours (Figure 5 .12d). In summary, this dataset is underpowered to be able to examine the hypothesis that RNase H2 deficiency is associated with an increased number of short deletions in CLL tumours. Larger numbers of RNASEH2B null tumours are needed to show whether or not an RNase H2 deficiency/Top1 mutational signature is present in CLL.

5.3 Discussion

5.3.1 Can one infer that Top1 activity drives short deletions in human genomes?

In the work described in this Chapter, I explored whether the genome embedded ribonucleotide/Top1 mediated mutational signature described in *S. cerevisiae* could also be identified in human genomes. Using a reporter construct designed for use in both *S. cerevisiae* and human cells, and validated in yeast, I show that supra-physiological levels of ribonucleotides in human cells are also associated with an increase in 2 base pair equivalent deletions (Figure 5 .5). These findings are supported by my analysis of results from a mutation accumulation experiment in RPE1 cells, which also show an increase in 2-5 bp mutations when RNASEH2A or B is knocked out (Figure 5.7). From these results, two interlinked questions arise: is this mutational pattern also driven by Top1 activity in human cells, and why is the increase in mutations seen less marked than what is observed in *S. cerevisiae*?

A key difference between mammalian and yeast cells is that in the former Top1 is essential (Morham *et al.*, 1996), whereas in yeast it can be knocked out without discernible effects on the viability of cells(Yanagida and Sternglanz, 1990). Therefore whilst the experiments outlined in this chapter show that is

possible to detect the presence of a ribonucleotide related signature (as it is possible to knock out RNase H2, as long as p53 is also inactivated), demonstrating a Top1 related signature in mammalian cells will be more challenging. Demonstrating the causal role of Top1 might require transient knockdown of Top1 activity using short interfering RNAs (Zimmermann *et al.*, 2018), or the use of protein silencing using a degron system (Röth, Fulcher and Sapkota, 2019). These would probably have to be introduced in the final round of cell division of a fluctuation assay, during which half of the total number of mutations in the experiment would be expected to occur (as outlined in Chapter 4). It is not clear however what the other unintended consequences of Top1 inhibition might be for cell replication and survival, and how these might affect the results of the fluctuation assay.

In addition, the fluctuation assay results seen in human cells were much less marked than those seen in *S. cerevisiae*, with an increase of 4.4, rather than 47-fold, in the rate of 2 base pair deletions with the loss of RNase H2. Demonstrating a causal role for topoisomerase would involve showing a reduction in the 2 base pair mutation rate when Top1 is possibly knocked down for possibly just a single round of cell division, limiting the power of the analysis for an already very laborious and time-consuming experiment. Another option would be to demonstrate an *increase* in the 2 base pair (or overall short) deletion rate by activating the Top1 promoter in a manner analogous to that presented in Chapter 4 (Figure 4.8). However, even in *S. cerevisiae* the increase in mutation rate seen with constitutive activation of the promoter was only 2.5-fold, suggesting that even if the effect were of the same magnitude in human cells, demonstrating a significant increase may be challenging. These considerations regarding the difficulties in demonstrating the causal role of Top1 in short-deletion mutagenesis raise the important question of why the effect seen in human cells is less marked than that in *S. cerevisiae*.

5.3.2 Why are there differences between the results in *S. cerevisiae* and human cells?

The difference in the magnitude of the effect in 2 base pair deletion rates seen when RNase H2 is lost in *S. cerevisiae* compared to human cells could be related to a variety of factors. With specific reference to the fluctuation assays, scoping experiments showed that the RNase H2 cells were much less robust than then WT controls, and much more sensitive to puromycin dosage than the WT controls, even when a mutation was present in the construct conferring puromycin resistance. Visualisation of colonies using direct microscopy through the 2 week time course of selection showed that a number of small colonies present in plates from the RNASEH2A-KO clones did not survive through to the staining steps. Another possible factor is differences in the activity of the CAG promoter used in the HeLa cells lines compared to the PTEF promoter used in *S. cerevisiae*; the literature review presented in Chapter 4 shows that short mutation rates are related to transcriptional activity. Finally, it may be that in human cells there are other mechanisms for the removal of single embedded ribonucleotides, or resolution of the 2'3' cyclic phosphate intermediates created by Top1.

5.3.3 Further work to identify a ribonucleotide/TOP1 mediated mutational signature in a CLL dataset

As highlighted in the Introduction to this chapter, an understanding of mutational signatures in cancer genomes can serve a variety of roles. It can help to understand the drivers of particular cancer types (for example deficiencies in mismatch repair), and stratify types into those which are likely to benefit from particular therapeutic interventions, for example the use of PARP-I in those with deficient homologous recombination (Davies *et al.*, 2017). Cancer genomes can also be used to better understand the roles of particular enzymes, and the consequences of deficiency, or absence of those enzymes, as demonstrated in Figure 5 .1. However, I found that only a small

proportion (7/1950, 0.4%) of the PCAWG WGS were annotated as having lost both copies of one of the RNase H2 subunits. Analysing these, the small number of RNase H2 null genomes meant that it was difficult to draw strong conclusions about the role of RNase H2 loss in the generation of 2-5 bp deletions. A larger dataset of CLL tumours exists as part of the NIHR/NHS England 100,000 genomes project (*The 100,000 Genomes Project | Genomics England*, 2020). Work is ongoing in collaboration with the Palles Lab (University of Birmingham) to combine the data from PCAWG with those from the 100,000 genomes project to see if an association between RNase H2 loss and an increased frequency of 2-5 bp deletions can be demonstrated. The Jackson Laboratory is also conducting whole genome sequencing of RNase H2 null intestinal epithelial cells (Aden *et al.*, 2019), to see if the COSMIC ID4 mutational profile is also seen in this context.

Chapter 6 Conclusions and General Discussion

6.1 Key findings and implications

6.1.1 Nanopore sequencing to quantitatively detect phasing of ribonucleotides at single nucleotide resolution

Next Generation Sequencing based approaches already exist to detect ribonucleotides at single base pair resolution, and have yielded insights into the division of labour for polymerases at the replication fork (Clausen *et al.*, 2015; Reijns *et al.*, 2015; Zhou *et al.*, 2019). In my analysis comparing emRiboSeq (Ding *et al.*, 2015) and HydEnSeq (Clausen, Williams and Kunkel, 2015) over Okazaki junctions (Smith and Whitehouse, 2012), I add to these findings. I show that there is evidence for retention of single ribonucleotides at these locations, with little support for the retention of ribonucleotide tracts (Chapter 3, Figure 3.4). My analysis also provides some support for the Pol α tract retention hypothesis (Reijns *et al.*, 2015). The expectation from this hypothesis is that protein binding at Okazaki junctions will lead to an increased likelihood of the retention of more error prone pol α synthesised DNA. The retention of lower fidelity Pol α synthesised DNA is posited to lead to the increased likelihood of mutations at these locations. Whilst a peak of retained ribonucleotides is seen at Okazaki junctions at the same location for emRiboSeq and HyEnSeq, what is not seen is the shoulder that one might expect from retained Pol α synthesised DNA if this model is correct.

The example above shows concordance between these two NGS based methodologies. Another outstanding question regarding DNA polymerase activity is the role of different polymerases at origins of replication: what are the relative roles of Pol α, δ and ε at these locations in initiating and continuing synthesis of the leading and lagging strands? Use of NGS based technologies

has already been used to confirm Pol ε as the main leading strand polymerase, and Pol α and δ as the main lagging strand polymerases (Clausen *et al.*, 2015; Reijns *et al.*, 2015). I sought to use data from these experiments to investigate at higher resolution whether Pol δ might also have a role in the initiation of leading strand synthesis, as suggested by biochemical (Aria and Yeeles, 2019) fluctuation assay (Johnson *et al.*, 2015) and sequencing (Garbacz *et al.*, 2018; Zhou *et al.*, 2019) studies. However, comparing results for both methods over *S. cerevisiae* origins of replication, I find that these are inconsistent. Whilst the RNase H2 cleavage based emRiboSeq shows a decrease in ribonucleotide incorporation at the origins of replication, the alkali hydrolysis based HydEnSeq shows an *increase* at these locations (Figure 3.3). These discrepant results suggested to me that differences in methodology for each technique (RNase H2 vs alkali cleavage, post-cleavage processing and sequencing steps), rather than the underlying biology, might explain the different results. These discrepancies, and the other limitations of NGS based approaches described in Chapter 3 (Section 3.1.3.3), prompted me to investigate nanopore technology as a means to quantitatively detect phasing of single, or potentially tracts of, ribonucleotides in DNA.

In developing this technology I demonstrated that ribonucleotides not only have a distinct amplitude signature when compared to the equivalent DNA kmer, but also a kinetic signature that could be used to call a ribonucleotide in genomic context. The location of this kinetic signature, upstream of the embedded ribonucleotide, is in keeping with what is understood about the structure of the pore. In addition, I developed a method that would permit the construction of a k-*mer* model library for the large number of combinations required to be able to detect ribonucleotides in the full combination of possible 5-*mers* used to deconvolute the squiggle trace. However, full prototyping of this approach was limited by inconsistent hairpin formation. Experiments showed that in a approximately two thirds of cases, rather than forming the

intended hairpin structure, the hairpin oligonucleotides instead dimerised with a second hairpin oligonucleotide. This meant that what was sequenced was two unrelated DNA strands, rather than a template and its complement. If this technical difficulty could be resolved, the approach outlined could be applied to the detection of any non-canonical nucleotide that it is possible to replicate across, such as 8-Oxo-2'-deoxyguanosine (Gajewski *et al.*, 1990).

6.1.2 New questions about the relationship between RNase H2 and Top1 activity

To develop a high sensitivity reporter construct that could be prototyped in yeast and directly transferred to a human model system, I leveraged the redundancy of the genetic code, making use of the observation that ribonucleotide/Top1 mediated short deletion mutagenesis was enriched in short couplet repeat sequences. I found that this reporter construct re-capitulated previously published findings showing that loss of RNase H2 was associated with an increase in short (2-5 bp) deletions, mediated by the activity of Top1. I additionally found that deletion of Top1 led to a *decrease* in the rate of 2 bp equivalent deletions, even in RNase H2 proficient cells, supporting the observation that Top1 has mutational consequences in the presence of physiological levels of ribonucleotides (Lippert *et al.*, 2011). Upregulating levels of Top1 using a galactose inducible promoter led to an increase in the mutation rate, and downregulation (by growth of the same construct in glucose containing media) to a reduction in the mutation rate, supporting previous findings (Sloan *et al.*, 2017). My results expand on these findings by showing a distinctive Top1 dependent mutational signature in a wider sequence context, and confirming the specificity of these mutations for a dinucleotide repeat context. In both my experiments and those by Sloan et al, the mutational signature is transcription dependent: future experiments could examine this dependency with a regulatable promoter for the reporter construct itself.

Whether this Top1 mediated mutational signature in the presence of physiological levels of ribonucleotides is due to the formation of 2'3' cyclic phosphate intermediates after ribonucleotide cleavage, or to the formation of Top1 cleavage complexes, not related to embedded ribonucleotides, which are resolved by a second upstream cleavage event, remains unclear (Cho and Jinks-Robertson, 2018). In order to investigate the role of embedded ribonucleotides, I performed fluctuation assays with a Pol ε mutant (M644G) that incorporates ribonucleotides at an estimated rate of 1 in 92 bases (Nick McElhinny, Kumar, *et al.*, 2010), alongside a Pol ε mutant (M644L) that incorporates fewer ribonucleotides than the wild type. This showed no difference in the mutation rate, suggesting that ribonucleotides on the leading strand are not the cause of Top1 mediated mutagenesis in the presence of RNase H2. Previous experiments in *S. cerevisiae* have shown that in the absence of RNase H2 Top1 activity appears to be *leading* strand specific (Conover *et al.*, 2015; Williams *et al.*, 2015), contrary to the findings in the fluctuation assays presented in Chapter 5. However these experiments have to our knowledge not previously conducted in a wild type RNase H2 context. These experiments confirm that Top1 mutagenesis may occur due to the action of the enzyme on non-ribonucleotide substrates in the presence of RNase H2 (Cho and Jinks-Robertson, 2018).

However, subsequent work by Martin Reijns has been conducted using the same reporter construct with Pol α and pol δ mutator strains, in a wild type RNase H2 background. These show that there is an increase in 2 base pair deletion rates with these strains, that incorporate more ribonucleotides than the wild type equivalent. This is in contrast to my findings with the Pol ε mutator strain. These support a model where in physiological conditions (RNase H2 present) the lagging strand is where mutagenesis occurs. This could be due to either increased Top1 activity on the lagging strand, or increased RNase H2

activity on the leading strand, leading to more efficient removal of ribonucleotides misincorporated by Pol ε.

Another question relates to the relationship between deficient mismatch repair mediated short deletions, and those caused by the action of Top1. In the *S. cerevisiae* reporter constructs, an increase in 2 bp deletions is seen with both the loss of MMR and RNase H2, although the increase is more dramatic with the loss of RNase H2. In the *S.cerevisiae* literature, these two processes are assumed to be distinct, with MMR deficiency leading to 1 bp indels associated with polymerase slippage, and RNase H2 deficiency leading to an increase in 2-5 bp deletions (Nick McElhinny, Kumar, *et al.*, 2010; Lujan *et al.*, 2013). However, in the context of this reporter construct, and its applicability to mutational processes in mammalian genomes, it would interesting to perform a fluctuation assay with *S. cerevisiae* strain deficient for both MMR and RNase H2. As detailed in Chapter 4, if the mutations identified and repaired by the mismatch repair machinery are caused mainly by polymerase slippage, one would expect an additive increase in the mutation rate compared to WT . The addition of the RNase H2 specific (47-fold) to the MMR specific (3.1-fold) would lead to an expected ~50-fold increase in the overall mutation rate. However, if a proportion of the Top1 dependent 2-5 base pair mutations being detected and repaired by mismatch repair machinery are not polymerase slippage events, but ribonucleotide/Top1 mediated mutations, then one would expect a multiplicative, rather than additive, increase in the mutation rate. These findings would have implications for the interpretation of 2-5 base pair deletions seen in mammalian cells deficient for mismatch repair (Figure 5.8): can these all be attributable to failure of polymerase slippage, or is Top1 mutagenesis one of a number of additional underlying mutational processes that are normally identified and correct by the mismatch repair machinery?

6.1.3 A ribonucleotide dependent mutational signature can also be detected in human cells

Transfer of the reporter construct into a HeLa cell fluctuation assay confirmed the findings seen in *S. cerevisiae*, but with a less marked relative increase (3.4 vs 47-fold) in the RNase H2 null strains. Possible reasons for this discrepancy between the results in the two model systems include the increased relative fragility of human RNase H2 null cells (meaning reduced survival of clones with mutations), differences in the activity of the promoters used in each model system (if higher relative activity in *S. cerevisiae*), or the presence of additional ribonucleotide removal pathways in mammalian cells (Malfatti *et al.*, 2017). The findings in the HeLa fluctuation assay were supported by my analysis of an orthogonal mutation accumulation experiment conducted by Martin Reijns, which showed a 3.5-fold increase in the number of 2-5 base pair deletions in the RNase H2 null strain when compared to a WT equivalent grown over an equivalent number of generations.

Comparison was made of the mutational signature found in the RPE1 mutation accumulation experiment with results from *S. cerevisiae* and human cell lines. Plots based on the COSMIC classification showed similarities between the mutation profiles of RNase H2 null *S. cerevisiae* and RPE1 cells, and COSMIC indel signature 4. Cosine similarity testing showed that the short deletion mutation profile seen in the RNase H2 null RPE1 cells was most similar to that of the RNase H2 null *S. cerevisiae* cell lines, and the *de novo* mutations seen in human populations. This bore out the similarities seen between the RNase H2 null RPE1 cells and the human *de novo* mutations in terms of deletion length, deletion composition, and the length of the repeat within which deletions took place. Investigation of the mutational signature seen in CLL tumours deficient for RNase H2 lacked power to either confirm or refute the presence of this mutational signature in cancer samples. It is possible that combining these results with those from the 100,000 Genomes Project (*The*

100,000 Genomes Project | Genomics England, 2020) (work ongoing with the Palles Lab, University of Birmingham) may be able to address this question more comprehensively. This work with cancer genomes would support the hypothesis that ribonucleotide/Top1 dependent mutagenesis occurs in human cells outwith laboratory conditions.

6.1.4 Top1 mediated mutagenesis in human populations

As discussed in Chapter 5 (section), whilst the experiments above demonstrate the plausibility of ribonucleotide associated mutagenesis in human genomes, they do not address whether Top1 plays the same causal role as seen in *S.cerevisiae*, and whether this mutational process has relevance in the presence of physiological levels of ribonucleotides. Studies in human cells demonstrating Top1 activity on genome embedded ribonucleotides raise the expectation that this is a biochemically conserved process (Zimmermann *et al.*, 2018). Demonstrating a causal role for Top1 would likely involve a transient knockdown of Top1 using short interfering RNAs or a degron system (5.3.1).

At a genome wide level, the fact that Top1 cleavage has been linked to transcriptional activity (Baranello *et al.*, 2016; Ahmed *et al.*, 2017), suggests that the rate of short indels could also be associated with transcriptional activity, as shown in previous (Lippert *et al.*, 2011) reporter constructs. This is another avenue of evidence that could be explored to investigate whether Top1 activity is associated with short deletion mutagenesis in human cancers or at a population level. One could take 2-5 bp deletions from the PCAWG or Gene4Denovo datasets, and correlate the location of these in relation to transcriptional activity from a dataset such as ENCODE (Abascal *et al.*, 2020). Finding a correlation between increased transcriptional activity and increased short deletion rates would add support to the hypothesis that Top1 activity is associated with short deletion mutagenesis in human cells.

As well as contributing towards a deeper understanding of the processes driving the development and progression of cancer, mutational signatures have potential roles in diagnostics (identifying the causes of subtypes of cancer) and therapeutics. A tool used to detect evidence of deficient homologous recombination (HR) called HRDetect (Davies *et al.*, 2017) established that a larger proportion than expected of breast and ovarian tumours showed evidence of deficient HR than would be predicted based on patients who had a germline mutation in their BRCA1/2 genes only. The researchers went on to identify patients with previously undiagnosed BRCA1/2 pathogenic germline variants, and tumours with a functional deficiency of BRCA1/2 due to methylation of the BRCA1 promoter. As HR deficient tumours are known to be sensitive to poly(ADP-ribose) polymerase (PARP) inhibitors (Bryant *et al.*, 2005; Farmer *et al.*, 2005), this opens up the possibility of personalised cancer therapies on the basis of mutational signatures. Of particular interest for in the context of a possible Top1 mediated mutational signature are the recent demonstration that RNase H2 deficiency *also* appears to render cancers more vulnerable to treatment with PARP-inhibitors (Zimmermann *et al.*, 2018) .

6.2 Final remarks: the limits of observational inference

The work outlined in this thesis has highlighted several lines of evidence that support the possibility that a ribonucleotide/Top1 mediated mutational process may lead to short deletions in human genomes. In a series of experiments with reporter constructs I show that ribonucleotides are mutagenic not only in *S.cerevisiae* but in human cells. I demonstrate the same increase in short deletions associated with retained ribonucleotides in an orthogonal analysis of a whole genome sequencing mutation accumulation experiment in RPE1 cells. These experiments further support the importance of ribonucleotide excision repair (RER) as a key factor in maintaining mammalian genome integrity

(Reijns *et al.*, 2012), alongside other pathways such as mismatch repair (Kunkel and Erie, 2015), nucleotide excision repair (Spivak, 2015) and base excision repair (Krokan and Bjørås, 2013). From inference from *S. cerevisiae* (Cho and Jinks-Robertson, 2018) and related experiments in human cells (Zimmermann *et al.*, 2018), the likely mechanism of mutations in the absence of effective RER is Top1 activity on retained genome embedded ribonucleotides. The distribution of these ribonucleotides across the S. cerevisiae and mammalian genome remains poorly understood. In this thesis I have additionally shown that nanopore sequencing is a methodology that could be applied to search for phased single, and potentially tracts, of ribonucleotides in genomic DNA. In the future these new methodologies could be brought together to gain *in vivo* insight into the mechanisms of Top1 mediated mutagenesis in mammalian cells.

However, a fundamental issue lies at the heart of determining the *significance* of this process in human populations, or in cancer cells. This is the difficulty of demonstrating causality from observational data. This problem underlies any interpretation of the COSMIC mutational signatures, which have undeniably yielded insights into the biology that underlies cancer, and have been used to infer the aetiology of different cancer subtypes. However they remain observational in nature, and thus demonstration of causality is challenging (Hume, 1748). Similarly, the 2-5 base pair deletions in human *de novo* mutations share features with that seen in the RPE1 RNase H2 null cell line dataset, including mutation length, repeat context, and nucleotide context, that group them closer to the RNase H2 null mutational signature than the mismatch repair deficiency signature. However, this does not *prove* that ribonucleotide/Top1 mediated mutagenesis is a cause of mutations in human populations. Instead, I demonstrate in this thesis that it appears to be a better explanation (Popper, 1959) for many of the mutations observed than

polymerase slippage and mismatch repair deficiency (Taylor, Ponting and Copley, 2004; Montgomery *et al.*, 2013; Baptiste, Jacob and Eckert, 2015).

References

Abascal, F. *et al.* (2020) ‘Expanded encyclopaedias of DNA elements in the human and mouse genomes’, *Nature*. Nature Research, 583(7818), pp. 699–710. doi: 10.1038/s41586-020-2493-4.

Abràmoff, M. D., Magalhães, P. J. and Ram, S. J. (2004) ‘Image processing with imageJ’, *Biophotonics International*, 11(7), pp. 36–41. doi: 10.1201/9781420005615.

Acuna-Hidalgo, R., Veltman, J. A. and Hoischen, A. (2016) ‘New insights into the generation and role of de novo mutations in health and disease’, *Genome Biology*. BioMed Central Ltd., 17(1). doi: 10.1186/s13059-016-1110-1.

Aden, K. *et al.* (2019) ‘Epithelial RNase H2 Maintains Genome Integrity and Prevents Intestinal Tumorigenesis in Mice’, *Gastroenterology*. W.B. Saunders, 156(1), pp. 145-159.e19. doi: 10.1053/j.gastro.2018.09.047.

Ahmed, W. *et al.* (2017) ‘Transcription facilitated genome-wide recruitment of topoisomerase I and DNA gyrase’, *PLoS Genetics*. Public Library of Science, 13(5). doi: 10.1371/journal.pgen.1006754.

Aicardi, J. and Goutières, F. (1984) ‘A Progressive familial encephalopathy in infancy with calcifications of the basal ganglia and chronic cerebrospinal fluid lymphocytosis’, *Annals of Neurology*. Ann Neurol, 15(1), pp. 49–54. doi: 10.1002/ana.410150109.

Alexandrov, L. B. *et al.* (2013) ‘Signatures of mutational processes in human cancer’, *Nature*. Nature Publishing Group, 500(7463), pp. 415–421. doi: 10.1038/nature12477.

Alexandrov, L. B. *et al.* (2020) ‘The repertoire of mutational signatures in human cancer’, *Nature*. Nature Research, 578(7793), pp. 94–101. doi:

10.1038/s41586-020-1943-3.

Alexandrov, L. B. and Stratton, M. R. (2014) ‘Mutational signatures: The patterns of somatic mutations hidden in cancer genomes’, *Current Opinion in Genetics and Development*. Curr Opin Genet Dev, pp. 52–60. doi: 10.1016/j.gde.2013.11.014.

Allen-Soltero, S. *et al.* (2014) ‘A saccharomyces cerevisiae RNase H2 interaction network functions to suppress genome instability.’, *Molecular and cellular biology*. United States, 34(8), pp. 1521–1534. doi: 10.1128/MCB.00960-13.

Altman, D. (1990) *Practical statistics for medical research*. London: Chapman and Hall.

Altshuler, D. M. *et al.* (2012) ‘An integrated map of genetic variation from 1,092 human genomes’, *Nature*. Nature Publishing Group, 491(7422), pp. 56–65. doi: 10.1038/nature11632.

Aparicio, S. and Caldas, C. (2013) ‘The Implications of Clonal Genome Evolution for Cancer Medicine’, *New England Journal of Medicine*. Massachussetts Medical Society, 368(9), pp. 842–851. doi: 10.1056/NEJMra1204892.

Aria, V. and Yeeles, J. T. P. (2019) ‘Mechanism of Bidirectional Leading-Strand Synthesis Establishment at Eukaryotic DNA Replication Origins’, *Molecular Cell*. Cell Press, 73(2), pp. 199-211.e10. doi: 10.1016/j.molcel.2018.10.019.

ATCC (2020a) *hTERT RPE-1 ATCC ® CRL-4000™*. Available at: https://www.lgcstandards-atcc.org/Products/All/CRL-4000.aspx?geo_country=gb#characteristics (Accessed: 10 September 2020).

ATCC (2020b) *pGRN145 plasmid in E. coli GC10 ATCC ® MBA-141™*. Available at: https://www.lgcstandards-atcc.org/products/all/MBA-141.aspx (Accessed: 10 September 2020).

Auton, A. *et al.* (2015) 'A global reference for human genetic variation', *Nature*. Nature Publishing Group, 526(7571), pp. 68–74. doi: 10.1038/nature15393.

Van der Auwera, G. A. *et al.* (2013) 'From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline', *Current Protocols in Bioinformatics*. John Wiley and Sons Inc., 43(SUPL.43). doi: 10.1002/0471250953.bi1110s43.

Balakrishnan, L. and Bambara, R. A. (2013) 'Okazaki fragment metabolism.', *Cold Spring Harbor perspectives in biology*, 5(2). doi: 10.1101/cshperspect.a010173.

Baptiste, B. A., Jacob, K. D. and Eckert, K. A. (2015) 'Genetic evidence that both dNTP-stabilized and strand slippage mechanisms may dictate DNA polymerase errors within mononucleotide microsatellites', *DNA Repair*. Elsevier, 29, pp. 91–100. doi: 10.1016/j.dnarep.2015.02.016.

Baranello, L. *et al.* (2016) 'RNA Polymerase II Regulates Topoisomerase 1 Activity to Favor Efficient Transcription', *Cell*. Cell Press, 165(2), pp. 357–371. doi: 10.1016/j.cell.2016.02.036.

Bassett, D. E., Boguski, M. S. and Hieter, P. (1996) 'Yeast genes and human disease', *Nature*. Nature, pp. 589–590. doi: 10.1038/379589a0.

Beach, D. H. and Klar, A. J. (1984) 'Rearrangements of the transposable mating-type cassettes of fission yeast.', *The EMBO Journal*. John Wiley & Sons, Ltd, 3(3), pp. 603–610. doi: 10.1002/j.1460-2075.1984.tb01855.x.

Bebenek, K. and Kunkel, T. A. (1990) ‘Frameshift errors initiated by nucleotide misincorporation’, *Proceedings of the National Academy of Sciences*, 87(13).

Benitez-Guijarro, M. *et al.* (2018) ‘RNase H2, mutated in Aicardi-Goutières syndrome, promotes LINE-1 retrotransposition.’, *The EMBO journal*, 37(15). doi: 10.15252/embj.201798506.

Bentley, D. R. *et al.* (2008) ‘Accurate whole human genome sequencing using reversible terminator chemistry’, *Nature*. Nature Publishing Group, 456(7218), pp. 53–59. doi: 10.1038/nature07517.

Berglund, A.-K. *et al.* (2017) ‘Nucleotide pools dictate the identity and frequency of ribonucleotide incorporation in mitochondrial DNA.’, *PLoS genetics*. Public Library of Science, 13(2), p. e1006628. doi: 10.1371/journal.pgen.1006628.

Bergstrom, E. N. *et al.* (2019) ‘SigProfilerMatrixGenerator: A tool for visualizing and exploring patterns of small mutational events’, *BMC Genomics*. BioMed Central Ltd., 20(1), p. 685. doi: 10.1186/s12864-019-6041-2.

Besenbacher, S. *et al.* (2015) ‘Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios’, *Nature Communications*, 6, p. 5969. doi: 10.1038/ncomms6969.

Bodnar, A. G. *et al.* (1998) ‘Extension of life-span by introduction of telomerase into normal human cells’, *Science*. Science, 279(5349), pp. 349–352. doi: 10.1126/science.279.5349.349.

Boland, C. R. and Lynch, H. T. (2013) ‘The history of Lynch syndrome.’, *Familial cancer*. NIH Public Access, 12(2), pp. 145–157. doi:

10.1007/s10689-013-9637-8.

Boyle, S. *et al.* (2020) ‘A central role for canonical PRC1 in shaping the 3D nuclear landscape’, *Genes & Development*. Cold Spring Harbor Laboratory, 34(13–14), pp. 931–949. doi: 10.1101/gad.336487.120.

Brachmann, C. B. *et al.* (1998) ‘Designer deletion strains derived from Saccharomyces cerevisiae S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications’, *Yeast*. Yeast, 14(2), pp. 115–132. doi: 10.1002/(SICI)1097-0061(19980130)14:2<115::AID-YEA204>3.0.CO;2-2.

Breslauer, K. J. *et al.* (1986) ‘Predicting DNA duplex stability from the base sequence.’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 83(11), pp. 3746–50. doi: 10.1073/pnas.83.11.3746.

Brewer, B. J. and Fangman, W. L. (1991) ‘Mapping replication origins in yeast chromosomes’, *BioEssays*. John Wiley & Sons, Ltd, 13(7), pp. 317–322. doi: 10.1002/bies.950130702.

Brown, J. A. and Suo, Z. (2011) ‘Unlocking the sugar “steric gate” of DNA polymerases’, *Biochemistry*. NIH Public Access, 50(7), pp. 1135–1142. doi: 10.1021/bi101915z.

Bryant, H. E. *et al.* (2005) ‘Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase’, *Nature*. Nature, 434(7035), pp. 913–917. doi: 10.1038/nature03443.

CC BY 4.0 (2020) *Creative Commons Public Licenses*. Available at: https://creativecommons.org/licenses/by/4.0/legalcode (Accessed: 22 July 2020).

Cech, T. R. (2012) ‘The RNA worlds in context’, *Cold Spring Harbor Perspectives in Biology*. Cold Spring Harbor Laboratory Press, 4(7), pp. 1–5. doi: 10.1101/cshperspect.a006742.

Cerretelli, G. *et al.* (2020) ‘Molecular pathology of Lynch syndrome’, *The Journal of Pathology*. NLM (Medline), 250(5), pp. 518–531. doi: 10.1002/path.5422.

Chang, L. M. S. (1977) ‘DNA polymerases from Bakers’ yeast’, *Journal of Biological Chemistry*. Elsevier, 252(6), pp. 1873–1880. doi: 10.1016/s0021-9258(18)71839-5.

Chen, J. Z. *et al.* (2000) ‘Mutational spectrum analysis of RNase H(35) deficient Saccharomyces cerevisiae using fluorescence-based directed termination PCR.’, *Nucleic acids research*. Oxford University Press, 28(18), pp. 3649–56. doi: 10.1093/NAR/28.18.3649.

Chenevert, J. M. *et al.* (1984) ‘Use of plasmid-mediated resistance to the antibiotic G418 for the rapid screening of a yeast mutant library.’, *Journal of experimental pathology*, 1(4), pp. 307–13.

Cho, J. E. and Jinks-Robertson, S. (2018) ‘Topoisomerase I and genome stability: The good and the bad’, in *Methods in Molecular Biology*. Humana Press Inc., pp. 21–45. doi: 10.1007/978-1-4939-7459-7_2.

Chu, G. (1997) ‘Double strand break repair’, *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology, pp. 24097–24100. doi: 10.1074/jbc.272.39.24097.

Churko, J. M. *et al.* (2013) ‘Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases’, *Circulation Research*. Lippincott Williams & WilkinsHagerstown, MD, pp.

1613–1623. doi: 10.1161/CIRCRESAHA.113.300939.

Clark, A. B. *et al.* (2011) ‘Mismatch repair-independent tandem repeat sequence instability resulting from ribonucleotide incorporation by DNA polymerase ε.’, *DNA repair*. Netherlands, 10(5), pp. 476–82. doi: 10.1016/j.dnarep.2011.02.001.

Clausen, A. R. *et al.* (2015) ‘Tracking replication enzymology in vivo by genome-wide mapping of ribonucleotide incorporation’, *Nature Structural & Molecular Biology*. Nature Publishing Group, 22(3), pp. 185–191. doi: 10.1038/nsmb.2957.

Clausen, A. R., Williams, J. S. and Kunkel, T. A. (2015) ‘Measuring ribonucleotide incorporation into DNA in vitro and in vivo.’, *Methods in molecular biology (Clifton, N.J.)*, 1300, pp. 123–39. doi: 10.1007/978-1-4939-2596-4_9.

Condit, C. M. *et al.* (2002) ‘The changing meanings of ?mutation:? A contextualized study of public discourse’, *Human Mutation*. John Wiley & Sons, Ltd, 19(1), pp. 69–75. doi: 10.1002/humu.10023.

Conover, H. N. *et al.* (2015) ‘Stimulation of Chromosomal Rearrangements by Ribonucleotides’, *Genetics*, 201(3), pp. 951–961. doi: 10.1534/genetics.115.181149.

Copy Number Variants | *ICGC Data Portal* (2020). Available at: https://dcc.icgc.org/releases/PCAWG/consensus_cnv/ (Accessed: 15 July 2020).

Coulondre, C. *et al.* (1978) ‘Molecular basis of base substitution hotspots in Escherichia coli’, *Nature*. Nature Publishing Group, 274(5673), pp. 775–780. doi: 10.1038/274775a0.

Crick, F. (1970) ‘Central dogma of molecular biology’, *Nature*. Nature, 227(5258), pp. 561–563. doi: 10.1038/227561a0.

CRISPR Guide RNA Design Software for Molecular Biology | Benchling (no date). Available at: https://www.benchling.com/crispr/ (Accessed: 2 July 2020).

Crow, J. F. (2000) ‘The origins, patterns and implications of human spontaneous mutation’, *Nature Reviews Genetics*. European Association for Cardio-Thoracic Surgery, pp. 40–47. doi: 10.1038/35049558.

Crow, Y. J. *et al.* (2006) ‘Mutations in genes encoding ribonuclease H2 subunits cause Aicardi-Goutières syndrome and mimic congenital viral brain infection’, *Nature Genetics*. Nat Genet, 38(8), pp. 910–916. doi: 10.1038/ng1842.

Crow, Y. J., Shetty, J. and Livingston, J. H. (2020) ‘Treatments in Aicardi–Goutières syndrome’, *Developmental Medicine & Child Neurology*. Blackwell Publishing Ltd, 62(1), pp. 42–47. doi: 10.1111/dmcn.14268.

Daigaku, Y. *et al.* (2015) ‘A global profile of replicative polymerase usage.’, *Nature structural & molecular biology*, 22(3), pp. 192–8. doi: 10.1038/nsmb.2962.

Dalgaard, J. Z. (2012) ‘Causes and consequences of ribonucleotide incorporation into nuclear DNA.’, *Trends in genetics : TIG*, 28(12), pp. 592–7. doi: 10.1016/j.tig.2012.07.008.

Darwin, C. (1859) *The Origin of Species by Means of Natural Selection: or, the Preservation of Favoured Races in the Struggle for Life*. First. London: John Murray.

Davies, H. *et al.* (2017) ‘HRDetect is a predictor of BRCA1 and BRCA2

deficiency based on mutational signatures', *Nature Medicine*. Nature Publishing Group, 23(4), pp. 517–525. doi: 10.1038/nm.4292.

Ding, J. *et al.* (2015) 'Genome-wide mapping of embedded ribonucleotides and other noncanonical nucleotides using emRiboSeq and EndoSeq.', *Nature protocols*, 10(9), pp. 1433–44. doi: 10.1038/nprot.2015.099.

Douzery, E. J. P. *et al.* (2004) 'The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils?', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 101(43), pp. 15386–15391. doi: 10.1073/pnas.0403984101.

Drake, J. W. (1991) 'A constant rate of spontaneous mutation in DNA-based microbes.', *Proceedings of the National Academy of Sciences of the United States of America*, 88(16), pp. 7160–4.

Drost, J. *et al.* (2017) 'Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer', *Science*. American Association for the Advancement of Science, 358(6360), pp. 234–238. doi: 10.1126/science.aao3130.

Drost, J. B. and Lee, W. R. (1995) 'Biological basis of germline mutation: Comparisons of spontaneous germline mutation rates among drosophila, mouse, and human', *Environmental and Molecular Mutagenesis*. Environ Mol Mutagen, 25(2 S), pp. 48–64. doi: 10.1002/em.2850250609.

Durkacz, B. W. *et al.* (1980) '(ADP-Ribose)n participates in DNA excision repair ', *Nature*. Nature, 283(5747), pp. 593–596. doi: 10.1038/283593a0.

Eckert, S. E. *et al.* (2016) 'Enrichment by hybridisation of long DNA fragments for Nanopore sequencing', *Microbial genomics*. Microb Genom,

2(9), p. e000087. doi: 10.1099/mgen.0.000087.

Eichler, E. E. (2019) ‘Genetic Variation, Comparative Genomics, and the Diagnosis of Disease’, *New England Journal of Medicine*. Massachussetts Medical Society, 381(1), pp. 64–74. doi: 10.1056/NEJMra1809315.

Farmer, H. *et al.* (2005) ‘Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy’, *Nature*. Nature, 434(7035), pp. 917–921. doi: 10.1038/nature03445.

Faust, G. G. and Hall, I. M. (2014) ‘SAMBLASTER: fast duplicate marking and structural variant read extraction’, *Bioinformatics*, 30(17), pp. 2503–2505. doi: 10.1093/bioinformatics/btu314.

Frommer, M. *et al.* (1992) ‘A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 89(5), pp. 1827–31. doi: 10.1073/pnas.89.5.1827.

Gajewski, E. *et al.* (1990) ‘Modification of DNA Bases in Mammalian Chromatin by Radiation-Generated Free Radicals’, *Biochemistry*. American Chemical Society, 29(34), pp. 7876–7882. doi: 10.1021/bi00486a014.

Galili, T. (2020) *GitHub - talgalili/gplots*. Available at: https://github.com/talgalili/gplots (Accessed: 27 November 2020).

Gao, Z. *et al.* (2019) ‘Overlooked roles of DNA damage and maternal age in generating human germline mutations’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 116(19), pp. 9491–9500. doi: 10.1073/pnas.1901259116.

Garalde, D. R. *et al.* (2018) ‘Highly parallel direct RNA sequencing on an

array of nanopores', *Nature Methods*. Nature Publishing Group, 15(3), pp. 201–206. doi: 10.1038/nmeth.4577.

Garbacz, M. A. *et al.* (2018) 'Evidence that DNA polymerase δ contributes to initiating leading strand DNA replication in Saccharomyces cerevisiae', *Nature Communications*. Nature Publishing Group, 9(1), pp. 1–11. doi: 10.1038/s41467-018-03270-4.

Garcia-Diaz, M. *et al.* (2006) 'Structural analysis of strand misalignment during DNA synthesis by a human DNA polymerase', *Cell*. Cell Press, 124(2), pp. 331–342. doi: 10.1016/j.cell.2005.10.039.

Gazdar, A. F. (2009) 'Activating and resistance mutations of EGFR in non-small-cell lung cancer: Role in clinical response to EGFR tyrosine kinase inhibitors', *Oncogene*. NIH Public Access, pp. S24–S31. doi: 10.1038/onc.2009.198.

GEO Accession Viewer GSE56939 (2014). Available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56939 (Accessed: 14 July 2020).

Ghodgaonkar, M. M. *et al.* (2013) 'Ribonucleotides misincorporated into DNA act as strand-discrimination signals in eukaryotic mismatch repair.', *Molecular cell*, 50(3), pp. 323–32. doi: 10.1016/j.molcel.2013.03.019.

Gietz, R. D. *et al.* (1995) 'Studies on the transformation of intact yeast cells by the LiAc/SS-DNA/PEG procedure', *Yeast*. Yeast, 11(4), pp. 355–360. doi: 10.1002/yea.320110408.

Gietz, R. D. and Woods, R. A. (2006) 'Yeast transformation by the LiAc/SS Carrier DNA/PEG method.', *Methods in molecular biology (Clifton, N.J.)*. Methods Mol Biol, 313, pp. 107–120. doi: 10.1385/1-59259-958-3:107.

Glaab, W. E. and Tindall, K. R. (1997) ‘Mutation rate at the hprt locus in human cancer cell lines with specific mismatch repair-gene defects.’, *Carcinogenesis*, 18(1), pp. 1–8.

Goyal, P. *et al.* (2014) ‘Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG.’, *Nature*, 516(7530), pp. 250–3. doi: 10.1038/nature13768.

GSE62181 (2015) *GSE62181*. Available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62181.

GSE64521 (2015) *GSE64521*. Available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64521 (Accessed: 2 June 2020).

Günther, C. *et al.* (2015) ‘Defective removal of ribonucleotides from DNA promotes systemic autoimmunity’, *Journal of Clinical Investigation*. American Society for Clinical Investigation, 125(1), pp. 413–424. doi: 10.1172/JCI78001.

El Hage, A. *et al.* (2010) ‘Loss of Topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis’, *Genes and Development*. Cold Spring Harbor Laboratory Press, 24(14), pp. 1546–1558. doi: 10.1101/gad.573310.

Hanahan, D. and Weinberg, R. A. (2000) ‘The hallmarks of cancer’, *Cell*. Cell, pp. 57–70. doi: 10.1016/S0092-8674(00)81683-9.

Hanahan, D. and Weinberg, R. A. (2011) ‘Hallmarks of cancer: The next generation’, *Cell*. Elsevier, pp. 646–674. doi: 10.1016/j.cell.2011.02.013.

Hartigan, J. A. and Wong, M. A. (1979) ‘Algorithm AS 136: A K-Means Clustering Algorithm’, *Applied Statistics*, 28(1), p. 100. doi: 10.2307/2346830.

Hennion, M. *et al.* (2018) ‘Mapping DNA replication with nanopore sequencing’, *bioRxiv*. Cold Spring Harbor Laboratory, p. 426858. doi: 10.1101/426858.

Hentges, P. *et al.* (2005) ‘Three novel antibiotic marker cassettes for gene disruption and marker switching in Schizosaccharomyces pombe’, *Yeast*. Yeast, 22(13), pp. 1013–1019. doi: 10.1002/yea.1291.

Hiller, B. *et al.* (2012) ‘Mammalian RNase H2 removes ribonucleotides from DNA to maintain genome integrity’, *The Journal of Experimental Medicine*, 209(8), pp. 1419–1426. doi: 10.1084/jem.20120876.

Hodges, E. *et al.* (2007) ‘Genome-wide in situ exon capture for selective resequencing’, *Nature Genetics*. Nat Genet, 39(12), pp. 1522–1527. doi: 10.1038/ng.2007.42.

Hu, S.-Z., Wang, T. S. and Korn, D. (1984) ‘DNA primase from KB cells. Evidence for a novel model of primase catalysis by a highly purified primase/polymerase-alpha complex.’, *Journal of Biological Chemistry*. ASBMB, 259(4), pp. 2602–2609.

Hume, D. (1748) *An enquiry concerning human understanding*. Oxford: Oxford University Press.

ID1 - COSMIC Mutational Signatures (2020). Available at: https://cancer.sanger.ac.uk/cosmic/signatures/ID/ID1.tt (Accessed: 30 November 2020).

International Human Genome Sequencing Consortium (2001) ‘Initial sequencing and analysis of the human genome’, *Nature*. Nature Publishing Group, 409(6822), pp. 860–921. doi: 10.1038/35057062.

Jain, M. *et al.* (2016) ‘The Oxford Nanopore MinION: delivery of nanopore

sequencing to the genomics community.', *Genome biology*. Genome Biol, 17(1), p. 239. doi: 10.1186/s13059-016-1103-0.

Jamal-Hanjani, M. *et al.* (2017) 'Tracking the Evolution of Non–Small-Cell Lung Cancer', *New England Journal of Medicine*. New England Journal of Medicine (NEJM/MMS), 376(22), pp. 2109–2121. doi: 10.1056/nejmoa1616288.

Jeong, H.-S. *et al.* (2004) 'RNase H2 of Saccharomyces cerevisiae is a complex of three proteins.', *Nucleic acids research*. England ([Erratum appears in Nucleic Acids Res. 2004 Feb 24;32(4):1616]), 32(2), pp. 407–414.

Johnson, R. E. *et al.* (2015) 'A Major Role of DNA Polymerase δ in Replication of Both the Leading and Lagging DNA Strands', *Molecular Cell*. Cell Press, 59(2), pp. 163–175. doi: 10.1016/j.molcel.2015.05.038.

Jónsson, H. *et al.* (2017) 'Parental influence on human germline de novo mutations in 1,548 trios from Iceland', *Nature*. Nature Research, 549(7673), pp. 519–522. doi: 10.1038/nature24018.

Joshua-Tor, L. *et al.* (1992) 'Three-dimensional structures of bulge-containing DNA fragments', *Journal of Molecular Biology*. Academic Press, 225(2), pp. 397–431. doi: 10.1016/0022-2836(92)90929-E.

Kachroo, A. H. *et al.* (2015) 'Systematic humanization of yeast genes reveals conserved functions and genetic modularity', *Science*. American Association for the Advancement of Science, 348(6237), pp. 921–925. doi: 10.1126/science.aaa0769.

Kim, J. H. *et al.* (2011) 'High Cleavage Efficiency of a 2A Peptide Derived from Porcine Teschovirus-1 in Human Cell Lines, Zebrafish and Mice', *PLoS ONE*. Edited by V. Thiel. Public Library of Science, 6(4), p. e18556. doi:

10.1371/journal.pone.0018556.

Kim, N. *et al.* (2011) ‘Mutagenic processing of ribonucleotides in DNA by yeast topoisomerase I.’, *Science (New York, N.Y.)*. NIH Public Access, 332(6037), pp. 1561–4. doi: 10.1126/science.1205016.

Kim, N. *et al.* (2013) ‘RNA:DNA hybrids initiate quasi-palindrome-associated mutations in highly transcribed yeast DNA.’, *PLoS genetics*, 9(11), p. e1003924. doi: 10.1371/journal.pgen.1003924.

Koh, K. D. *et al.* (2015) ‘Ribose-seq: global mapping of ribonucleotides embedded in genomic DNA.’, *Nature methods*, 12(3), pp. 251–7, 3 p following 257. doi: 10.1038/nmeth.3259.

Kong, A. *et al.* (2012) ‘Rate of de novo mutations and the importance of father’s age to disease risk.’, *Nature*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 488(7412), pp. 471–5. doi: 10.1038/nature11396.

Korbie, D. J. and Mattick, J. S. (2008) ‘Touchdown PCR for increased specificity and sensitivity in PCR amplification’, *Nature Protocols*. Nat Protoc, 3(9), pp. 1452–1456. doi: 10.1038/nprot.2008.133.

Kornberg, R. D. and Thomas, J. O. (1974) ‘Chromatin structure: Oligomers of the histones’, *Science*. Science, 184(4139), pp. 865–868. doi: 10.1126/science.184.4139.865.

Kotin, R. M., Linden, R. M. and Berns, K. I. (1992) ‘Characterization of a preferred site on human chromosome 19q for integration of adeno-associated virus DNA by non-homologous recombination’, *EMBO Journal*. Genetic Therapy Inc, 11(13), pp. 5071–5078. doi: 10.1002/j.1460-2075.1992.tb05614.x.

Krokan, H. E. and Bjørås, M. (2013) ‘Base excision repair’, *Cold Spring Harbor Perspectives in Biology*. Cold Spring Harb Perspect Biol, 5(4), pp. 1–22. doi: 10.1101/cshperspect.a012583.

Kucab, J. E. *et al.* (2019) ‘A Compendium of Mutational Signatures of Environmental Agents’, *Cell*. Cell Press, 177(4), pp. 821-836.e16. doi: 10.1016/j.cell.2019.03.001.

Kuhn, R. M., Haussler, D. and Kent, W. J. (2013) ‘The UCSC genome browser and associated tools’, *Briefings in Bioinformatics*, 14(2), pp. 144–161. doi: 10.1093/bib/bbs038.

Kumar, A. *et al.* (2008) ‘Structure and clinical relevance of the epidermal growth factor receptor in human cancer’, *Journal of Clinical Oncology*. NIH Public Access, pp. 1742–1751. doi: 10.1200/JCO.2007.12.1178.

Kunkel, T. A. (1985) ‘The mutational specificity of DNA polymerase-beta during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations.’, *Journal of Biological Chemistry*, 260, pp. 5787–5796.

Kunkel, T. A. *et al.* (1989) ‘Fidelity of DNA polymerase I and the DNA polymerase I-DNA primase complex from Saccharomyces cerevisiae.’, *Molecular and Cellular Biology*. American Society for Microbiology, 9(10), pp. 4447–4458. doi: 10.1128/mcb.9.10.4447.

Kunkel, T. A. and Erie, D. A. (2005) ‘DNA mismatch repair’, *Annual Review of Biochemistry*. Annual Reviews, 74(1), pp. 681–710. doi: 10.1146/annurev.biochem.74.082803.133243.

Kunkel, T. A. and Erie, D. A. (2015) ‘Eukaryotic Mismatch Repair in Relation to DNA Replication’, *Annual Review of Genetics*. Annual Reviews Inc., 49, pp. 291–313. doi: 10.1146/annurev-genet-112414-054722.

Laird, C. D. *et al.* (2004) ‘Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules.’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 101(1), pp. 204–9. doi: 10.1073/pnas.2536758100.

Lamarck, J. B. de (1809) *Philosophie zoologique*. Paris: Musée d’Histoire Naturelle (Jardin des Plantes).

Lang, G. I., Parsons, L. and Gammie, A. E. (2013) ‘Mutation rates, spectra, and genome-wide distribution of spontaneous mutations in mismatch repair deficient yeast.’, *G3 (Bethesda, Md.)*, 3(9), pp. 1453–65. doi: 10.1534/g3.113.006429.

Larrea, A. A. *et al.* (2010) ‘Genome-wide model for the normal eukaryotic DNA replication fork’, *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 107(41), pp. 17674–17679. doi: 10.1073/pnas.1010178107.

Laughery, M. F. *et al.* (2015) ‘New vectors for simple and streamlined CRISPR-Cas9 genome editing in *Saccharomyces cerevisiae*’, *Yeast*. John Wiley and Sons Ltd, 32(12), pp. 711–720. doi: 10.1002/yea.3098.

Lazzaro, F. *et al.* (2012) ‘RNase H and postreplication repair protect cells from ribonucleotides incorporated in DNA.’, *Molecular cell*, 45(1), pp. 99–110. doi: 10.1016/j.molcel.2011.12.019.

Lea, D. E. and Coulson, C. A. (1949) ‘The distribution of the numbers of mutants in bacterial populations’, *Journal of Genetics*. J Genet, 49(3), pp. 264–285. doi: 10.1007/BF02986080.

Lee, B. *et al.* (2017) ‘Comparison of SHAPE reagents for mapping RNA

structures inside living cells.', *RNA*. Cold Spring Harbor Laboratory Press, 23(2), pp. 169–174. doi: 10.1261/rna.058784.116.

LEHMAN, I. R. *et al.* (1958) 'Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from Escherichia coli.', *The Journal of biological chemistry*. Elsevier, 233(1), pp. 163–170. doi: 10.1016/S0021-9258(19)68048-8.

Lehner, K. and Jinks-Robertson, S. (2009) 'The mismatch repair system promotes DNA polymerase-dependent translesion synthesis in yeast', *Proceedings of the National Academy of Sciences*, 106(14), pp. 5749 –5754.

Leushkin, E. V., Bazykin, G. A. and Kondrashov, A. S. (2012) 'Insertions and deletions trigger adaptive walks in Drosophila proteins', *Proceedings of the Royal Society B: Biological Sciences*. Royal Society, 279(1740), pp. 3075–3082. doi: 10.1098/rspb.2011.2571.

Li, H. (2013) 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM'.

Li, W. *et al.* (2017) 'Major challenges related to tumor biological characteristics in accurate mutation detection of colorectal cancer by next-generation sequencing', *Cancer Letters*. Elsevier Ireland Ltd, 410, pp. 92–99. doi: 10.1016/j.canlet.2017.09.014.

Li, Y. and Breaker, R. (1999) 'Kinetics of RNA Degradation by Specific Base Catalysis of Transesterification Involving the 2'-Hydroxyl Group', *J. Am. Chem. Soc*. American Chemical Society, 121(23), pp. 5364–5372. doi: 10.1021/JA990592P.

Lindblad-Toh, K. *et al.* (2011) 'A high-resolution map of human evolutionary constraint using 29 mammals', *Nature*. Nature Publishing Group, 478(7370),

pp. 476–482. doi: 10.1038/nature10530.

Lippert, M. J. *et al.* (2011) ‘Role for topoisomerase 1 in transcription-associated mutagenesis in yeast’, *Proceedings of the National Academy of Sciences*, 108(2), pp. 698–703. doi: 10.1073/pnas.1012363108.

Liu, B. *et al.* (2017) ‘Direct visualization of RNA-DNA primer removal from okazaki fragments provides support for flap cleavage and exonucleolytic pathways in eukaryotic cells’, *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology Inc., 292(12), pp. 4777–4788. doi: 10.1074/jbc.M116.758599.

Loman, N. J., Quick, J. and Simpson, J. T. (2015) ‘A complete bacterial genome assembled de novo using only nanopore sequencing data’, *Nature Methods*. Nature Publishing Group, 12(8), pp. 733–735. doi: 10.1038/nmeth.3444.

Loman, N. J. and Watson, M. (2015) ‘Successful test launch for nanopore sequencing’, *Nature Methods*. Nature Publishing Group, 12(4), pp. 303–304. doi: 10.1038/nmeth.3327.

Longtine, M. *et al.* (1998) ‘Additional Modules for Versatile and Economical PCR-based Gene Deletion and Modification in Saccharomyces Cerevisiae’, *Yeast (Chichester, England)*. Yeast, 14(10). doi: 10.1002/(SICI)1097-0061(199807)14:10<953::AID-YEA293>3.0.CO;2-U.

Luger, K. *et al.* (1997) ‘Crystal structure of the nucleosome core particle at 2.8 Å resolution’, *Nature*. Nature, 389(6648), pp. 251–260. doi: 10.1038/38444.

Lujan, S. A. *et al.* (2012) ‘Mismatch repair balances leading and lagging strand DNA replication fidelity.’, *PLoS genetics*. Edited by C. E. Pearson,

8(10), p. e1003016. doi: 10.1371/journal.pgen.1003016.

Lujan, S. A. *et al.* (2013) ‘Ribonucleotides are signals for mismatch repair of leading-strand replication errors.’, *Molecular cell*, 50(3), pp. 437–43. doi: 10.1016/j.molcel.2013.03.017.

Lujan, S. A. *et al.* (2014) ‘Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition.’, *Genome research*. Cold Spring Harbor Laboratory Press, 24(11), pp. 1751–64. doi: 10.1101/gr.178335.114.

Luria, S. E. and Delbrück, M. (1943) ‘Mutations of Bacteria from Virus Sensitivity to Virus Resistance.’, *Genetics*, 28(6), pp. 491–511.

Lyer, R. R. *et al.* (2006) ‘DNA mismatch repair: Functions and mechanisms’, *Chemical Reviews*. Chem Rev, 106(2), pp. 302–323. doi: 10.1021/cr0404794.

Lynch, M. (2016) ‘Mutation and Human Exceptionalism: Our Future Genetic Load.’, *Genetics*. Genetics Society of America, 202(3), pp. 869–75. doi: 10.1534/genetics.115.180471.

Ma, W. T., Sandri, G. vH. and Sarkar, S. (1992) ‘Analysis of the Luria–Delbrück distribution using discrete convolution powers’, *Journal of Applied Probability*. Cambridge University Press (CUP), 29(2), pp. 255–267. doi: 10.2307/3214564.

Mackenzie, K. J. *et al.* (2016) ‘Ribonuclease H2 mutations induce a cGAS/STING-dependent innate immune response’, *The EMBO Journal*. EMBO, 35(8), pp. 831–844. doi: 10.15252/embj.201593339.

Macqueen, J. (1967) ‘Some methods for classification and analysis of multivariate observations’, in *In 5th Berkeley Symposium on Mathematical*

Statistics and Probability, pp. 281–297.

Maga, G. *et al.* (2001) ‘Okazaki fragment processing: Modulation of the strand displacement activity of DNA polymerase δ by the concerted action of replication protein A, proliferating cell nuclear antigen, and flap endonuclease-1’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 98(25), pp. 14298–14303. doi: 10.1073/pnas.251193198.

Malfatti, M. C. *et al.* (2017) ‘Abasic and oxidized ribonucleotides embedded in DNA are processed by human APE1 and not by RNase H2’, *Nucleic Acids Research*. Oxford University Press, 45(19), pp. 11193–11212. doi: 10.1093/nar/gkx723.

Mandel, H. *et al.* (2001) ‘The deoxyguanosine kinase gene is mutated in individuals with depleted hepatocerebral mitochondrial DNA’, *Nature Genetics*. Nat Genet, 29(3), pp. 337–341. doi: 10.1038/ng746.

Markham, N. R., Zuker, M. and Keith, J. M. (2008) ‘UNAFold: software for nucleic acid folding and hybridization., pp. 3--31’, *Bioinformatics*. Humana Press Totowa, NJ.

Martin, H. C. *et al.* (2018) ‘Quantifying the contribution of recessive coding variation to developmental disorders’, *Science*. American Association for the Advancement of Science, 362(6419), pp. 1161–1164. doi: 10.1126/science.aar6731.

Martin, M. J. *et al.* (2013) ‘Ribonucleotides and manganese ions improve non-homologous end joining by human Polm’, *Nucleic Acids Research*. Oxford University Press, 41(4), pp. 2428–2436. doi: 10.1093/nar/gks1444.

McElhinny, S. A. N., Kissling, G. E. and Kunkel, T. A. (2010) ‘Differential

correction of lagging-strand replication errors made by DNA polymerases α and δ', *Proceedings of the National Academy of Sciences of the United States of America*. PNAS, 107(49), pp. 21070–21075. doi: 10.1073/pnas.1013048107.

McGinnis, J. L. *et al.* (2012) 'The mechanisms of RNA SHAPE chemistry', *Journal of the American Chemical Society*. American Chemical Society, 134(15), pp. 6617–6624. doi: 10.1021/ja2104075.

McGuffee, S. R., Smith, D. J. and Whitehouse, I. (2013) 'Quantitative, Genome-Wide Analysis of Eukaryotic Replication Initiation and Termination', *Molecular Cell*, 50(1), pp. 123–135. doi: 10.1016/j.molcel.2013.03.004.

Meagher, M., Epling, L. B. and Enemark, E. J. (2019) 'DNA translocation mechanism of the MCM complex and implications for replication initiation', *Nature Communications*. Nature Publishing Group, 10(1), pp. 1–13. doi: 10.1038/s41467-019-11074-3.

Merino, E. *et al.* (2005) 'RNA Structure Analysis at Single Nucleotide Resolution by Selective 2'-Hydroxyl Acylation and Primer Extension (SHAPE)', *Journal of the American Chemical Society*. American Chemical Society, 127, pp. 4223–4231. doi: 10.1021/JA043822V.

Miyata, H. and Miyata, M. (1981) 'Mode of conjugation in homothallic cells of Schizosaccharomyces pombe', *The Journal of General and Applied Microbiology*. Applied Microbiology, Molecular and Cellular Biosciences Research Foundation, 27(5), pp. 365–371.

Montgomery, S. B. *et al.* (2013) 'The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes', *Genome Research*. Genome Res, 23(5), pp. 749–761. doi: 10.1101/gr.148718.112.

Morganella, S. *et al.* (2016) ‘The topography of mutational processes in breast cancer genomes’, *Nature Communications*, 7, p. 11383. doi: 10.1038/ncomms11383.

Morham, S. G. *et al.* (1996) ‘Targeted disruption of the mouse topoisomerase I gene by camptothecin selection.’, *Molecular and Cellular Biology*. American Society for Microbiology, 16(12), pp. 6804–6809. doi: 10.1128/mcb.16.12.6804.

Moss, C. F. *et al.* (2017) ‘Aberrant ribonucleotide incorporation and multiple deletions in mitochondrial DNA of the murine MPV17 disease model’, *Nucleic Acids Research*. Oxford University Press, 45(22), p. 12808. doi: 10.1093/NAR/GKX1009.

Müller, C. A. *et al.* (2019) ‘Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads’, *Nature Methods*. Nature Publishing Group, 16(5), pp. 429–436. doi: 10.1038/s41592-019-0394-y.

Nagalakshmi, U. *et al.* (2008) ‘The transcriptional landscape of the yeast genome defined by RNA sequencing.’, *Science*. American Association for the Advancement of Science, 320(5881), pp. 1344–9. doi: 10.1126/science.1158441.

Nick McElhinny, S. A. *et al.* (2008) ‘Division of Labor at the Eukaryotic Replication Fork’, *Molecular Cell*. NIH Public Access, 30(2), pp. 137–144. doi: 10.1016/j.molcel.2008.02.022.

Nick McElhinny, S. A., Watts, B. E., *et al.* (2010) ‘Abundant ribonucleotide incorporation into DNA by yeast replicative polymerases.’, *Proceedings of the National Academy of Sciences of the United States of America*, 107(11), pp. 4949–54. doi: 10.1073/pnas.0914857107.

Nick McElhinny, S. A., Kumar, D., *et al.* (2010) ‘Genome instability due to ribonucleotide incorporation into DNA.’, *Nature chemical biology*, 6(10), pp. 774–81. doi: 10.1038/nchembio.424.

Nick McElhinny, S. A. and Ramsden, D. A. (2003) ‘Polymerase Mu Is a DNA-Directed DNA/RNA Polymerase’, *Molecular and Cellular Biology*. American Society for Microbiology, 23(7), pp. 2309–2315. doi: 10.1128/mcb.23.7.2309-2315.2003.

Nieduszynski, C. A. *et al.* (2007) ‘OriDB: a DNA replication origin database’, *Nucleic Acids Research*, 35(suppl_1), pp. D40–D46. doi: 10.1093/nar/gkl758.

Nielsen, O. and Egel, R. (1989) ‘Mapping the double-strand breaks at the mating-type locus in fission yeast by genomic sequencing.’, *The EMBO journal*. European Molecular Biology Organization, 8(1), pp. 269–76.

Nik-Zainal, S. *et al.* (2012) ‘The life history of 21 breast cancers’, *Cell*. Cell Press, 149(5), pp. 994–1007. doi: 10.1016/j.cell.2012.04.023.

O’Brien, K., Remm, M. and Sonnhammer, E. (2005) ‘Inparanoid: a comprehensive database of eukaryotic orthologs’, *Nucleic Acids Research*, 33(1), pp. 476–480. doi: https://doi.org/10.1093/nar/gki107.

O’Connell, K., Jinks-Robertson, S. and Petes, T. D. (2015) ‘Elevated Genome-Wide Instability in Yeast Mutants Lacking RNase H Activity.’, *Genetics*, 201(3), pp. 963–75. doi: 10.1534/genetics.115.182725.

Oceguera-Yanez, F. *et al.* (2016) ‘Engineering the AAVS1 locus for consistent and scalable transgene expression in human iPSCs and their differentiated derivatives’, *Methods*. Academic Press Inc., 101, pp. 43–55. doi: 10.1016/j.ymeth.2015.12.012.

Ogata, T., Kozuka, T. and Kanda, T. (2003) ‘Identification of an Insulator in

AAVS1, a Preferred Region for Integration of Adeno-Associated Virus DNA', *Journal of Virology*. American Society for Microbiology, 77(16), pp. 9000–9007. doi: 10.1128/jvi.77.16.9000-9007.2003.

Okazaki, R. *et al.* (1968) 'Mechanism of DNA chain growth. I. Possible discontinuity and unusual secondary structure of newly synthesized chains.', *Proceedings of the National Academy of Sciences of the United States of America*, 59(2), pp. 598–605.

ONT (2011) *WO2013014451A1 - Hairpin loop method for double strand polynucleotide sequencing using transmembrane pores - Google Patents*. Available at: https://patents.google.com/patent/WO2013014451A1 (Accessed: 5 June 2020).

Ottaviani, D., LeCain, M. and Sheer, D. (2014a) 'The role of microhomology in genomic structural variation', *Trends in Genetics*. Elsevier Current Trends, pp. 85–94. doi: 10.1016/j.tig.2014.01.001.

Ottaviani, D., LeCain, M. and Sheer, D. (2014b) 'The role of microhomology in genomic structural variation', *Trends in Genetics*. Trends Genet, 30(3), pp. 85–94. doi: 10.1016/j.tig.2014.01.001.

Pavlov, Y. I., Newlon, C. S. and Kunkel, T. A. (2002) 'Yeast origins establish a strand bias for replicational mutagenesis', *Molecular Cell*. Cell Press, 10(1), pp. 207–213. doi: 10.1016/S1097-2765(02)00567-1.

PCAWG | ICGC Data Portal (2020). Available at: https://dcc.icgc.org/pcawg (Accessed: 9 July 2020).

Pedersen, C. E. T. *et al.* (2017) 'The effect of an extreme and prolonged population bottleneck on patterns of deleterious variation: Insights from the Greenlandic Inuit', *Genetics*. Genetics Society of America, 205(2), pp. 787–

801. doi: 10.1534/genetics.116.193821.

Pellicer, J., Fay, M. F. and Leitch, I. J. (2010) ‘The largest eukaryotic genome of them all?’, *Botanical Journal of the Linnean Society*. Oxford Academic, 164(1), pp. 10–15. doi: 10.1111/j.1095-8339.2010.01072.x.

Perera, R. L. *et al.* (2013) ‘Mechanism for priming DNA synthesis by yeast DNA Polymerase α’, *eLife*, 2013(2). doi: 10.7554/eLife.00482.

Pfeifer, G. P. (2006) ‘Mutagenesis at methylated CpG sequences’, in *DNA methylation: basic mechanisms*. Springer, pp. 259–281.

Pizzi, S. *et al.* (2015) ‘Reduction of hRNase H2 activity in Aicardi-Goutières syndrome cells leads to replication stress and genome instability.’, *Human molecular genetics*, 24(3), pp. 649–58. doi: 10.1093/hmg/ddu485.

Pollard, K. S. *et al.* (2006) ‘Forces Shaping the Fastest Evolving Regions in the Human Genome’, *PLoS Genetics*. Edited by M. Przeworski. Public Library of Science, 2(10), p. e168. doi: 10.1371/journal.pgen.0020168.

Popper, K. (1959) *The Logic of Scientific Discovery*. London: Hutchinson.

Poteete, A. R. and Hardy, L. W. (1994) ‘Genetic Analysis of Bacteriophage T4 Lysozyme Structure and Function’, *Journal of Bacteriology*, pp. 6783–6788.

Potenski, C. J. *et al.* (2014) ‘Avoidance of ribonucleotide-induced mutations by RNase H2 and Srs2-Exo1 mechanisms.’, *Nature*. England, 511(7508), pp. 251–254.

Pridmore, R. D. (1987) ‘New and versatile cloning vectors with kanamycin-resistance marker’, *Gene*. Gene, 56(2–3), pp. 309–312. doi: 10.1016/0378-1119(87)90149-1.

Prohaska, A. *et al.* (2019) ‘Human Disease Variation in the Light of Population Genomics’, *Cell*. Cell Press, 177(1), pp. 115–131. doi: 10.1016/j.cell.2019.01.052.

Pryor, J. M. *et al.* (2018) ‘Ribonucleotide incorporation enables repair of chromosome breaks by nonhomologous end joining’, *Science*. American Association for the Advancement of Science, 361(6407), pp. 1126–1129. doi: 10.1126/science.aat2477.

Pursell, Z. F. *et al.* (2007) ‘Yeast DNA polymerase epsilon participates in leading-strand DNA replication.’, *Science*. NIH Public Access, 317(5834), pp. 127–30. doi: 10.1126/science.1144067.

Pursell, Z. F. and Kunkel, T. A. (2008) ‘Chapter 4 DNA Polymerase ε. A Polymerase of Unusual Size (and Complexity)’, *Progress in Nucleic Acid Research and Molecular Biology*. NIH Public Access, pp. 101–145. doi: 10.1016/S0079-6603(08)00004-4.

Ramachandran, S. *et al.* (2005) ‘Support from the relationship of genetic and geographic in human populations for a serial founder effect originating in Africa’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 102(44), pp. 15942–15947. doi: 10.1073/pnas.0507611102.

Ran, F. A. *et al.* (2013) ‘Genome engineering using the CRISPR-Cas9 system’, *Nature Protocols*, 8(11), pp. 2281–2308. doi: 10.1038/nprot.2013.143.

Rand, A. *et al.* (2017) ‘Mapping DNA Methylation With High-Throughput Nanopore Sequencing’, *Nature methods*. Nat Methods, 14(4). doi: 10.1038/NMETH.4189.

Rao, R. N. *et al.* (1983) ‘Genetic and enzymatic basis of hygromycin B resistance in Escherichia coli’, *Antimicrobial Agents and Chemotherapy*. American Society for Microbiology Journals, 24(5), pp. 689–695. doi: 10.1128/AAC.24.5.689.

Rao, S. S. P. *et al.* (2014) ‘A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping’, *Cell*. Cell Press, 159(7), pp. 1665–1680. doi: 10.1016/j.cell.2014.11.021.

Rebbeck, T. R. *et al.* (2015) ‘Association of type and location of BRCA1 and BRCA2 mutations with risk of breast and ovarian cancer’, *JAMA - Journal of the American Medical Association*. American Medical Association, 313(13), pp. 1347–1361. doi: 10.1001/jama.2014.5985.

Reijns, M. A. M. *et al.* (2011) ‘The Structure of the Human RNase H2 Complex Defines Key Interaction Interfaces Relevant to Enzyme Function and Human Disease’, *Journal of Biological Chemistry*, 286(12), pp. 10530–10539. doi: 10.1074/jbc.M110.177394.

Reijns, M. A. M. *et al.* (2012) ‘Enzymatic removal of ribonucleotides from DNA is essential for mammalian genome integrity and development.’, *Cell*, 149(5), pp. 1008–22. doi: 10.1016/j.cell.2012.04.011.

Reijns, M. A. M. *et al.* (2015) ‘Lagging-strand replication shapes the mutational landscape of the genome.’, *Nature*, 518(7540), pp. 502–6. doi: 10.1038/nature14183.

Rice, G. *et al.* (2007) ‘Clinical and molecular phenotype of Aicardi-Goutières syndrome’, *American Journal of Human Genetics*. Cell Press, 81(4), pp. 713–725. doi: 10.1086/521373.

Rice, G. I. *et al.* (2013) ‘Assessment of interferon-related biomarkers in

Aicardi-Goutières syndrome associated with mutations in TREX1, RNASEH2A, RNASEH2B, RNASEH2C, SAMHD1, and ADAR: A case-control study', *The Lancet Neurology*. Lancet Neurol, 12(12), pp. 1159–1169. doi: 10.1016/S1474-4422(13)70258-8.

Richmond, T. J. and Davey, C. A. (2003) 'The structure of DNA in the nucleosome core', *Nature*. Nature, 423(6936), pp. 145–150. doi: 10.1038/nature01595.

Rivas, M. A. *et al.* (2018) 'Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population', *PLoS Genetics*. Public Library of Science, 14(5), p. e1007329. doi: 10.1371/journal.pgen.1007329.

Röth, S., Fulcher, L. J. and Sapkota, G. P. (2019) 'Advances in targeted degradation of endogenous proteins', *Cellular and Molecular Life Sciences*. Birkhauser Verlag AG, 76(14), pp. 2761–2777. doi: 10.1007/s00018-019-03112-6.

Rouleau, M. *et al.* (2010) 'PARP inhibition: PARP1 and beyond', *Nature Reviews Cancer*. Nat Rev Cancer, 10(4), pp. 293–301. doi: 10.1038/nrc2812.

RStudio (2015) 'RStudio: integrated development for R', *RStudio, Inc., Boston, MA URL http://www. rstudio. com*, 42, p. 14.

Rumbaugh, J. A. *et al.* (1997) 'Creation and removal of embedded ribonucleotides in chromosomal DNA during mammalian okazaki fragment processing', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology, 272(36), pp. 22591–22599. doi: 10.1074/jbc.272.36.22591.

Sanger, F. and Coulson, A. R. (1975) 'A rapid method for determining

sequences in DNA by primed synthesis with DNA polymerase', *Journal of Molecular Biology*. Academic Press, 94(3), pp. 441–448. doi: 10.1016/0022-2836(75)90213-2.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 74(12), pp. 5463–7. doi: 10.1073/pnas.74.12.5463.

Sayrac, S. *et al.* (2011) 'Identification of a Novel Type of Spacer Element Required for Imprinting in Fission Yeast', *PLoS Genetics*. Edited by J. E. Haber. Public Library of Science, 7(3), p. e1001328. doi: 10.1371/journal.pgen.1001328.

Schneider, C. A., Rasband, W. S. and Eliceiri, K. W. (2012) 'NIH Image to ImageJ: 25 years of image analysis', *Nature Methods*, 9(7), pp. 671–675. doi: 10.1038/nmeth.2089.

Schreve, J. L., Sin, J. K. and Garrett, J. M. (1998) 'The Saccharomyces cerevisiae YCC5 (YCL025c) gene encodes an amino acid permease, Agp1, which transports asparagine and glutamine', *Journal of Bacteriology*. American Society for Microbiology (ASM), 180(9), pp. 2556–2559. doi: 10.1128/jb.180.9.2556-2559.1998.

Schwarz, E. *et al.* (1985) 'Structure and transcription of human papillomavirus sequences in cervical carcinoma cells', *Nature*. Nature, 314(6006), pp. 111–114. doi: 10.1038/314111a0.

Schwarze, K. *et al.* (2018) 'Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature', *Genetics in Medicine*. Nature Publishing Group, 20(10), pp. 1122–1130. doi: 10.1038/gim.2017.247.

Sclavi, B. and Herrick, J. (2019) ‘Genome size variation and species diversity in salamanders’, *Journal of Evolutionary Biology*. Blackwell Publishing Ltd, 32(3), pp. 278–286. doi: 10.1111/jeb.13412.

Ségurel, L., Wyman, M. J. and Przeworski, M. (2014) ‘Determinants of Mutation Rate Variation in the Human Germline’, *Annual Review of Genomics and Human Genetics*. Annual Reviews, 15(1), pp. 47–70. doi: 10.1146/annurev-genom-031714-125740.

Sekiguchi, J. A. and Shuman, S. (1997) ‘Site-specific ribonuclease activity of eukaryotic DNA topoisomerase I’, *Molecular Cell*. Cell Press, 1(1), pp. 89–97. doi: 10.1016/S1097-2765(00)80010-6.

Serero, A. *et al.* (2014) ‘Mutational landscape of yeast mutator strains’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 111(5), p. 1897. doi: 10.1073/pnas.1314423111.

Shcherbakova, P. V and Kunkel, T. A. (1999) ‘Mutator phenotypes conferred by MLH1 overexpression and by heterozygosity for mlh1 mutations.’, *Molecular and cellular biology*, 19(4), pp. 3177–83.

Shigematsu, H. and Gazdar, A. F. (2006) ‘Somatic mutations of epidermal growth factor receptor signaling pathway in lung cancers’, *International Journal of Cancer*. John Wiley & Sons, Ltd, pp. 257–262. doi: 10.1002/ijc.21496.

Shuman, S. (1992) ‘DNA strand transfer reactions catalyzed by vaccinia topoisomerase I’, *Journal of Biological Chemistry*, 267(12), pp. 8620–8627.

Sia, E. A. *et al.* (1997) ‘Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes.’, *Molecular and Cellular*

Biology. American Society for Microbiology, 17(5), pp. 2851–2858. doi: 10.1128/mcb.17.5.2851.

Sikorski, R. S. and Hieter, P. (1989) 'A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae.', *Genetics*, 122(1).

Simpson, J. (2015) *Aligning Nanopore Events to a Reference*. Available at: http://simpsonlab.github.io/2015/04/08/eventalign/ (Accessed: 3 June 2020).

Simpson, J. T. *et al.* (2017) 'Detecting DNA cytosine methylation using nanopore sequencing', *Nature Methods*. Nature Publishing Group, 14(4), pp. 407–410. doi: 10.1038/nmeth.4184.

Siow, C. C. *et al.* (2012) 'OriDB, the DNA replication origin database updated and extended', *Nucleic Acids Research*. Oxford Academic, 40(D1), pp. D682–D686. doi: 10.1093/nar/gkr1091.

Sloan, R. *et al.* (2017) 'Effects of camptothecin or TOP1 overexpression on genetic stability in Saccharomyces cerevisiae.', *DNA repair*. DNA Repair (Amst), 59, pp. 69–75. doi: 10.1016/j.dnarep.2017.09.004.

Sloan, R. S. (2016) *Topoisomerase 1 (Top1)-associated genome instability in yeast: Effects of persistentcleavage complexes or increased Top1 levels*. Duke University.

Smith, D. J. and Whitehouse, I. (2012) 'Intrinsic coupling of lagging-strand synthesis to chromatin assembly.', *Nature*, 483(7390), pp. 434–8. doi: 10.1038/nature10895.

Smith, J. R. *et al.* (2008) 'Robust, Persistent Transgene Expression in Human Embryonic Stem Cells Is Achieved with AAVS1-Targeted Integration', *Stem Cells*. Wiley, 26(2), pp. 496–504. doi: 10.1634/stemcells.2007-0039.

Smith, L. M. *et al.* (1986) ‘Fluorescence detection in automated DNA sequence analysis’, *Nature*. Nature Publishing Group, 321(6071), pp. 674–679. doi: 10.1038/321674a0.

Sparks, J. L. *et al.* (2012) ‘RNase H2-Initiated Ribonucleotide Excision Repair’, *Molecular Cell*. Cell Press, 47(6), pp. 980–986. doi: 10.1016/j.molcel.2012.06.035.

Sparks, J. L. and Burgers, P. M. (2015) ‘Error-free and mutagenic processing of topoisomerase 1-provoked damage at genomic ribonucleotides.’, *The EMBO journal*. England, 34(9), pp. 1259–1269.

Spell, R. M. and Jinks-Robertson, S. (2004) ‘Determination of mitotic recombination rates by fluctuation analysis in Saccharomyces cerevisiae’, in *Genetic Recombination*. Springer, pp. 3–12.

Spielmann, M. and Mundlos, S. (2016) ‘Looking beyond the genes: the role of non-coding variants in human disease’, *Human molecular genetics*. Oxford University Press, 25(R2), pp. R157--R165.

Spinazzola, A. *et al.* (2006) ‘MPV17 encodes an inner mitochondrial membrane protein and is mutated in infantile hepatic mitochondrial DNA depletion’, *Nature Genetics*. Nat Genet, 38(5), pp. 570–576. doi: 10.1038/ng1765.

Spitale, R. C. *et al.* (2013) ‘RNA SHAPE analysis in living cells’, *Nature Chemical Biology*. NIH Public Access, 9(1), pp. 18–20. doi: 10.1038/nchembio.1131.

Spivak, G. (2015) ‘Nucleotide excision repair in humans’, *DNA Repair*. Elsevier B.V., 36, pp. 13–18. doi: 10.1016/j.dnarep.2015.09.003.

Streisinger, G. *et al.* (1966) ‘Frameshift mutations and the genetic code’, *Cold*

Spring Harbor symposia on quantitative biology. Cold Spring Harb Symp Quant Biol, 31, pp. 77–84. doi: 10.1101/SQB.1966.031.01.014.

Sugimoto, K., Okazaki, T. and Okazaki, R. (1968) 'Mechanism of DNA chain growth, II. Accumulation of newly synthesized short chains in E. coli infected with ligase-defective T4 phages.', *Proceedings of the National Academy of Sciences of the United States of America*, 60(4), pp. 1356–62.

Sun, C. *et al.* (2012) 'LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders', *Genome Biology and Evolution*. Oxford University Press, 4(2), pp. 168–183. doi: 10.1093/gbe/evr139.

Takahashi, T. *et al.* (2011) 'Topoisomerase 1 provokes the formation of short deletions in repeated sequences upon high transcription in Saccharomyces cerevisiae', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 108(2), pp. 692–697. doi: 10.1073/pnas.1012582108.

Talmane, L. (2018) *Protein Binding as a Selective Filter for New Mutations at Regulatory Sites in the Germline and in Cancers*. University of Edinburgh.

Taylor, M. S., Ponting, C. P. and Copley, R. R. (2004) 'Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes', *Genome Research*. Cold Spring Harbor Laboratory Press, 14(4), pp. 555–566. doi: 10.1101/gr.1977804.

Taylor, P. H., Cinquin, A. and Cinquin, O. (2016) 'Quantification of in vivo progenitor mutation accrual with ultra-low error rate and minimal input DNA using SIP-HAVA-seq.', *Genome research*. Genome Res, 26(11), pp. 1600–1611. doi: 10.1101/gr.200501.115.

The 100,000 Genomes Project | Genomics England (2020). Available at:

https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/ (Accessed: 5 October 2020).

The Cancer Genome Atlas - Timeline and Milestones - National Cancer Institute (2020). Available at: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history/timeline (Accessed: 9 July 2020).

Thompson, P. R. *et al.* (2002) 'Mechanism of aminoglycoside antibiotic kinase APH(3′)-IIIa: Role of the nucleotide positioning loop', *Biochemistry*. Biochemistry, 41(22), pp. 7001–7007. doi: 10.1021/bi0256680.

Tian, D. *et al.* (2008) 'Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes', *Nature*. Nature, 455(7209), pp. 105–108. doi: 10.1038/nature07175.

Tye, B. K. (1999) 'MCM proteins in DNA replication', *Annual Review of Biochemistry*. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA , 68, pp. 649–686. doi: 10.1146/annurev.biochem.68.1.649.

Uehara, R. *et al.* (2018) 'Two RNase H2 Mutants with Differential rNMP Processing Activity Reveal a Threshold of Ribonucleotide Tolerance for Embryonic Development', *Cell Reports*. Elsevier B.V., 25(5), pp. 1135-1145.e5. doi: 10.1016/j.celrep.2018.10.019.

Untergasser, A. *et al.* (2007) 'Primer3Plus, an enhanced web interface to Primer3.', *Nucleic acids research*. Oxford University Press, 35(Web Server issue), pp. W71-4. doi: 10.1093/nar/gkm306.

Vakhrusheva, A. A. *et al.* (2011) 'Evolution of prokaryotic genes by shift of stop codons', *Journal of Molecular Evolution*. J Mol Evol, 72(2), pp. 138–146.

doi: 10.1007/s00239-010-9408-1.

Vengrova, S. and Dalgaard, J. Z. (2006) ‘The wild-type Schizosaccharomyces pombe mat1 imprint consists of two ribonucleotides’, *EMBO reports*, 7(1), pp. 59–65. doi: 10.1038/sj.embor.7400576.

Venter, J. C. *et al.* (2001) ‘The sequence of the human genome.’, *Science*. American Association for the Advancement of Science, 291(5507), pp. 1304–51. doi: 10.1126/science.1058040.

de Vries, H. (1909) *The Mutation Theory*. Open Court Publishing Company.

Wang, X., Wang, Z. and Silva, N. A. Da (1996) ‘G418 Selection and stability of cloned genes integrated at chromosomal δ sequences of Saccharomyces cerevisiae’, *Biotechnology and Bioengineering*. John Wiley & Sons, Ltd, 49(1), pp. 45–51. doi: 10.1002/(SICI)1097-0290(19960105)49:1<45::AID-BIT6>3.0.CO;2-T.

Watson, J. D. and Crick, F. H. C. (1953) ‘Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid’, *Nature*. Nature, 171(4356), pp. 737–738. doi: 10.1038/171737a0.

Werner, B. *et al.* (2020) ‘Measuring single cell divisions in human tissues from multi-region sequencing data’, *Nature Communications*. Nature Research, 11(1). doi: 10.1038/s41467-020-14844-6.

Werness, B. A., Levine, A. J. and Howley, P. M. (1990) ‘Association of human papillomavirus types 16 and 18 E6 proteins with p53’, *Science*. Science, 248(4951), pp. 76–79. doi: 10.1126/science.2157286.

Wilkinson, K. A., Merino, E. J. and Weeks, K. M. (2006) ‘Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution’, *Nature Protocols*. Nature Publishing

Group, 1(3), pp. 1610–1616. doi: 10.1038/nprot.2006.249.

Williams, J. S. *et al.* (2013) ‘Topoisomerase 1-mediated removal of ribonucleotides from nascent leading-strand DNA.’, *Molecular cell*, 49(5), pp. 1010–5. doi: 10.1016/j.molcel.2012.12.021.

Williams, J. S. *et al.* (2015) ‘Evidence that processing of ribonucleotides in DNA by topoisomerase 1 is leading-strand specific.’, *Nature structural & molecular biology*. NIH Public Access, 22(4), pp. 291–7. doi: 10.1038/nsmb.2989.

Williams, J. S. *et al.* (2019) ‘Genome-wide mutagenesis resulting from topoisomerase 1-processing of unrepaired ribonucleotides in DNA’, *DNA Repair*. Elsevier B.V., 84, p. 102641. doi: 10.1016/j.dnarep.2019.102641.

Williams, J. S., Lujan, S. A. and Kunkel, T. A. (2016) ‘Processing ribonucleotides incorporated during eukaryotic DNA replication’, *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 17(6), pp. 350–363. doi: 10.1038/nrm.2016.37.

Workman, R. E. *et al.* (2019) ‘Nanopore native RNA sequencing of a human poly(A) transcriptome’, *Nature Methods*, 16(12), pp. 1297–1305. doi: 10.1038/s41592-019-0617-2.

Worthington, M. T., Luo, R. Q. and Pelo, J. (2001) ‘Copacabana method for spreading E. coli and Yeast colonies’, *BioTechniques*. Eaton Publishing Company, 30(4), pp. 738–742. doi: 10.2144/01304bm05.

Xu, J. *et al.* (2020) ‘Selective prebiotic formation of RNA pyrimidine and DNA purine nucleosides’, *Nature*. Nature Research, 582(7810), pp. 60–66. doi: 10.1038/s41586-020-2330-9.

Yamada-Inagawa, T., Klar, A. J. S. and Dalgaard, J. Z. (2007)

‘Schizosaccharomyces pombe switches mating type by the synthesis-dependent strand-annealing mechanism.’, *Genetics*, 177(1), pp. 255–65. doi: 10.1534/genetics.107.076315.

Yanagida, M. and Sternglanz, R. (1990) ‘DNA topology and its biological implications.’, in Cozzarelli, N. R. and Wang, J. C. (eds). New York: Cold Spring Harbor Laboratory Press, pp. 299–320.

Yang, H. *et al.* (2010) ‘Important role of indels in somatic mutations of human cancer genes’, *BMC Medical Genetics*. BioMed Central, 11(1), p. 128. doi: 10.1186/1471-2350-11-128.

Yang, M. Y. *et al.* (2002) ‘Biased incorporation of ribonucleotides on the mitochondrial L-strand accounts for apparent strand-asymmetric DNA replication’, *Cell*. Cell Press, 111(4), pp. 495–505. doi: 10.1016/S0092-8674(02)01075-9.

Yeeles, J. T. P. *et al.* (2017) ‘How the Eukaryotic Replisome Achieves Rapid and Efficient DNA Replication’, *Molecular Cell*. Cell Press, 65(1), pp. 105–116. doi: 10.1016/j.molcel.2016.11.017.

Young, R. S. *et al.* (2015) ‘The frequent evolutionary birth and death of functional promoters in mouse and human’, *Genome Research*. Cold Spring Harbor Laboratory Press, 25(10), pp. 1546–1557. doi: 10.1101/gr.190546.115.

Zhang, L. *et al.* (2019) ‘Rapid evolution of protein diversity by de novo origination in Oryza’, *Nature Ecology and Evolution*. Nature Publishing Group, 3(4), pp. 679–690. doi: 10.1038/s41559-019-0822-5.

Zhao, G. *et al.* (2020) ‘Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans’, *Nucleic acids research*. Oxford

University Press, 48(D1), pp. D913--D926.

Zheng, L. and Shen, B. (2011) ‘Okazaki fragment maturation: nucleases take centre stage.’, *Journal of molecular cell biology*, 3(1), pp. 23–30. doi: 10.1093/jmcb/mjq048.

Zhou, Z. X. *et al.* (2019) ‘Roles for DNA polymerase δ in initiating and terminating leading strand DNA replication’, *Nature Communications*. Nature Publishing Group, 10(1), pp. 1–10. doi: 10.1038/s41467-019-11995-z.

Zimmermann, M. *et al.* (2018) ‘CRISPR screens identify genomic ribonucleotides as a source of PARP-trapping lesions’, *Nature*. Nature Publishing Group, 559(7713), pp. 285–289. doi: 10.1038/s41586-018-0291-z.

Zuker, M. (2003) ‘Mfold web server for nucleic acid folding and hybridization prediction’, *Nucleic Acids Research*, 31(13), pp. 3406–3415. doi: 10.1093/nar/gkg595.