DNA precursor asymmetries, Mismatch Repair and their effect on mutation specificity

Robert J Buckland
DNA precursor asymmetries, Mismatch Repair and their effect on mutation specificity

Robert J Buckland

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie/medicine doktorsexamen framläggs till offentligt försvar i BIA201, Biologihuset, fredagen den 8 maj, kl. 09:00.
Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Dr. Juan Méndez,
Spanish National Cancer Research Centre, Madrid, Spain.
In order to build any structure, a good supply of materials, accurate workers and quality control are needed. This is even the case when constructing DNA, the so-called “Code of Life.” For a species to continue to exist, this DNA code must be copied with incredibly high accuracy when each and every cell replicates. In fact, just one mistake in the 12 million bases that comprise the genome of budding yeast, *Saccharomyces cerevisiae*, can be fatal. DNA is composed of a double strand helix made up of just four different bases repeated millions of times. The building blocks of DNA are the deoxyribonucleotides (dNTPs); dCTP, dTTP, dATP and dGTP. Their production and balance are carefully controlled within each cell, largely by the key enzyme Ribonucleotide Reductase (RNR). Here, we studied how the enzymes that copy DNA, the replicative polymerases α, δ and ε, cope with the effects of an altered dNTP pool balance. An introduced mutation in the allosteric specificity site of RNR in a strain of *S. cerevisiae, rnr1-Y285A*, leads to elevated dCTP and dTTP levels and has been shown to have a 14-fold increase in mutation rate compared to wild type. To ascertain the full effects of the dNTP pool imbalance upon the replicative polymerases, we disabled one of the major quality control systems in a cell that corrects replication errors, the post-replicative Mismatch Repair system. Using both the CAN1 reporter assay and whole genome sequencing, we found that, despite inherent differences between the polymerases, their replication fidelity was affected very similarly by this dNTP pool imbalance. Hence, the high dCTP and dTTP forced Pol ε and Pol α/δ to make the same mistakes. In addition, the mismatch repair machinery was found to correct replication errors driven by this dNTP pool imbalance with highly variable efficiencies. Another mechanism to protect cells from DNA damage during replication is a checkpoint that can be activated to delay the cell cycle and activate repair mechanisms. In yeast, Mec1 and Rad53 (human ATR and Chk1/Chk2) are two key S-phase checkpoint proteins. They are essential as they are also required for normal DNA replication and dNTP pool regulation. However the reason why they are essential is not well understood. We investigated this by mutating RAD53 and analyzing dNTP pools and gene interactions. We show that Rad53 is essential in S-phase due to its role in regulating basal dNTP levels by action in the Dun1 pathway that regulates RNR and Rad53’s compensatory kinase function if dNTP levels are perturbed.

In conclusion we present further evidence of the importance of dNTP pools in the maintenance of genome integrity and shed more light on the complex regulation of dNTP levels.

**Keywords**
DNA Replication Fidelity, Mutations, dNTP pools, Mismatch Repair, Checkpoint, Ribonucleotide Reductase, Msh2.
DNA precursor asymmetries, Mismatch Repair and their effect on mutation specificity

Robert J Buckland
"Everyday we should work out how to explain what we do and why we do it, as if we are talking to our grandmothers." – Jonathan Knowles

To my wonderful family.
# Table of Contents

Abstract iii  
Summary iv  
List of Abbreviations v  
List of Papers vi  
Introduction;  
  DNA Replication;  
    Cycles, brakes and forks 1  
    Polymerases 6  
    Maintaining Accuracy 7  
  DNA Errors and Repair  
    Maintenance Crews 11  
    Mutations and Damage 11  
    dNTP Pools 14  
Study Aims 15  
Tools  
  Yeast as a Model 16  
  Methods 16  
Results  
  Paper 1 19  
  Paper 2 21  
  Paper 3 24  
Discussion 26  
Future Directions 29  
Acknowledgements 30  
References 32  
Papers 37
Abstract

In order to build any structure, a good supply of materials, accurate workers and quality control are needed. This is even the case when constructing DNA, the so-called “Code of Life.” For a species to continue to exist, this DNA code must be copied with incredibly high accuracy when each and every cell replicates. In fact, just one mistake in the 12 million bases that comprise the genome of budding yeast, *Saccharomyces cerevisiae*, can be fatal. DNA is composed of a double strand helix made up of just four different bases repeated millions of times. The building blocks of DNA are the deoxyribonucleotides (dNTPs); dCTP, dTTP, dATP and dGTP. Their production and balance are carefully controlled within each cell, largely by the key enzyme Ribonucleotide Reductase (RNR). Here, we studied how the enzymes that copy DNA, the replicative polymerases α, δ and ε, cope with the effects of an altered dNTP pool balance. An introduced mutation in the allosteric specificity site of RNR in a strain of *S. cerevisiae*, *rnr1-Y285A*, leads to elevated dCTP and dTTP levels and has been shown to have a 14-fold increase in mutation rate compared to wild type. To ascertain the full effects of the dNTP pool imbalance upon the replicative polymerases, we disabled one of the major quality control systems in a cell that corrects replication errors, the post-replicative Mismatch Repair system. Using both the CAN1 reporter assay and whole genome sequencing, we found that, despite inherent differences between the polymerases, their replication fidelity was affected very similarly by this dNTP pool imbalance. Hence, the high dCTP and dTTP forced Pol ε and Pol α/δ to make the same mistakes. In addition, the mismatch repair machinery was found to correct replication errors driven by this dNTP pool imbalance with highly variable efficiencies.

Another mechanism to protect cells from DNA damage during replication is a checkpoint that can be activated to delay the cell cycle and activate repair mechanisms. In yeast, Mec1 and Rad53 (human ATR and Chk1/Chk2) are two key S-phase checkpoint proteins. They are essential as they are also required for normal DNA replication and dNTP pool regulation. However the reason why they are essential is not well understood. We investigated this by mutating *RAD53* and analyzing dNTP pools and gene interactions. We show that Rad53 is essential in S-phase due to its role in regulating basal dNTP levels by action in the Dun1 pathway that regulates RNR and Rad53’s compensatory kinase function if dNTP levels are perturbed.

In conclusion we present further evidence of the importance of dNTP pools in the maintenance of genome integrity and shed more light on the complex regulation of dNTP levels.
Summary; Layman’s Introduction.

Each individual cell of every single organism contains a strand of molecules with all the information required for life on it. This code controls what proteins are made in the cell and the cascade that follows accounts for a vast amount of what constitutes that organism, from its very species to how it reacts to its environment. In fact one could say the code is so important to each and every cell that it is written in its DNA.

Of course I am talking about DNA itself and more specifically dNTPs (deoxyribonucleoside triphosphates or deoxyribonucleotides), the individual letters that are used to build the code and form a library of information known as a genome. As a cell replicates, this vital information must also be reproduced and as the integrity of an organism’s genome is obviously vital to its continued survival, it must be copied very accurately. The DNA is also under pressure during normal ‘use’, external factors such as U.V. light or chemical exposure can damage it. A mistake or damaged code is known as a mutation, which can lead to a fault in the protein produced, which in turn can lead to a change in the cell or even whole organism.

This study investigates what happens to the genome if the supply of dNTPs is skewed combined with loss of one of the main safety systems (Mismatch Repair) that is often mutated in cancers.
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDKs</td>
<td>Cyclin Dependent Kinases</td>
</tr>
<tr>
<td>DNA</td>
<td>DeoxyriboNucleic Acid</td>
</tr>
<tr>
<td>dNTPs</td>
<td>Deoxyribonucleotides (DeoxyriboNucleoside TriPhosphates)</td>
</tr>
<tr>
<td>dCTP</td>
<td>DeoxyCytidine TriPhosphate</td>
</tr>
<tr>
<td>dTTP</td>
<td>DeoxyThymidine TriPhosphate</td>
</tr>
<tr>
<td>dATP</td>
<td>DeoxyAdenosine TriPhosphate</td>
</tr>
<tr>
<td>dGTP</td>
<td>DeoxyGuanosine TriPhosphate</td>
</tr>
<tr>
<td>MMR</td>
<td>MisMatch Repair</td>
</tr>
<tr>
<td>Pol</td>
<td>Polymerase</td>
</tr>
<tr>
<td>PCNA</td>
<td>Proliferating Cell Nuclear Antigen</td>
</tr>
<tr>
<td>RNR</td>
<td>RiboNucleotide Reductase</td>
</tr>
<tr>
<td>RPA</td>
<td>Replication Protein A</td>
</tr>
</tbody>
</table>
Papers in this Thesis

I. Increased and Imbalanced dNTP Pools Symmetrically Promote Both Leading and Lagging Strand Replication Infidelity.
Robert J. Buckland, Danielle L. Watt, Balasubramanyam Chittoor, Anna Karin Nilsson, Thomas A. Kunkel, Andrei Chabes
PLOS Genetics 2014

II. Genome-wide analysis of the specificity and mechanisms of replication infidelity driven by a mutation in ribonucleotide reductase that imbalances dNTP pools.
Danielle L. Watt¹, Robert J. Buckland¹, Scott Lujan¹, Thomas A. Kunkel and Andrei Chabes
Manuscript
¹Equal contribution.

III. Molecular Basis of the Essential S Phase Function of the Rad53 Checkpoint Kinase.
Nicolas C. Hoch, Eric S.-W. Chen, Robert Buckland, Shun-Chung Wang, Alessandro Fazio, Andrew Hammet, Achille Pellicioli, Andrei Chabes, Ming-Daw Tsai, Jörg Heierhorst
Molecular and Cellular Biology 2013

Other Publication not included in thesis (Book Chapter)

Determination of Deoxyribonucleoside Triphosphate Concentrations in Yeast Cells by Strong Anion-Exchange High-Performance Liquid Chromatography Coupled with Ultraviolet Detection.
Shaodong Jia, Lisette Marjavaara, Robert Buckland, Sushma Sharma, and Andrei Chabes
Methods in Molecular Biology; DNA Replication 2015
Introduction

Whether you are a mouse or a moose, just four letters make up your genetic code: C, T, A and G (the bases of the dNTPs; dCTP, dTTP, dATP and dGTP). These exist in a complimentary double strand, two copies of everything - comparable to a mirrored server, with C always paired with G and T always paired with A. A genome is very long, for example the information needed for a simple cell of baker’s yeast (Saccharomyces cerevisiae) is 12 million base pairs, a human cell has 3 billion and the Norway Spruce 23 billion (Nystedt et al. 2013).

In eukaryotes the genome is usually not just one continuous chain, but is separated into multiple sections in a cell, known as chromosomes. However it is possible that just a single erroneous base can be fatal. Therefore there are highly conserved mechanisms involved in regulating and protecting genetic material both during replication and in situ where DNA is under considerable stress from many sources.

The paradox to the requirement for accurate replication is that an organism must adapt to its surroundings and genetic adaptation requires variation. This leads to selection and the famous survival of the fittest term coined from Charles Darwin’s work by Herbert Spencer in his Principles of Biology (1864). Following on from Darwin’s work, Hugo De Vries coined the term ‘mutation’ around 1900 as a driver of evolution, in a field hotly debated for years. All organisms have an inherent error rate within their DNA replication systems, which allows for variation. However, whether ‘allows’ here means gives permission or tolerates is open to debate, a balance is found between the costs of beneficial mutations, deleterious mutations and the cost of maintaining fidelity (reviewed in (Sniegowski et al. 2000)). Indeed the existence of anti-mutator phenotypes (for example mutant yeast strains that have a lower mutation rate than wild type (Herr et al. 2011)) would suggest that the ‘normal’ replication systems are not quite as accurate as they could be. So what are the systems in place that maintain this fine balance between accuracy and variation?

Replication;

Cycles, Brakes and Forks

Most cells divide and replicate (there are some notable exceptions, such as neurons and bone cells) as part of what is known as the cell cycle (Figure 1). Simply, there are four main phases; Gap/Growth 1 (or G1) where cells increase in size. Synthesis phase (S phase) is when the DNA of a cell is made
into two copies, also known as DNA replication. After this there is another phase of growth, G2, before the cell divides in Mitosis (M).

All this can happen in less than a day and billions of times in a person’s life (Tomasetti and Vogelstein 2015).

**Figure 1. The Cell Cycle.** Simplified general model of a cell cycle with major checkpoints shown as red dotted lines. For further description see text.

**Checkpoints**

Each of these cell cycle phases contains one or more checkpoints, cascades of signalling that can delay or stop cell development by putting the brakes on. It is a point of much debate as to how many actual individual checkpoints there are, and the field has come a long way since Hartwell and Weinert described it as "the arrest of a cell at a particular phase of the cycle due to a lack of appropriate signals for cell cycle progression" in 1989. The standard textbook suggests three checkpoints; G₁/S, G₂/M and Mitosis (or metaphase to anaphase), Figure 1. However, up to ten checkpoints have been described, although this is largely down to differences in definitions (Khodjakov and Rieder 2009). It is probably more accurate to say that the cell cycle is constantly monitored for problems. Some checkpoints are during
phases and some are a sort of pre-flight check before entering the next stage of the cell cycle. The G1/S checkpoint is often called the restriction point in animal cells, it checks that the cell has grown sufficiently and everything is ready for replication to start in S phase, so DNA damage can also activate it. Similarly, the G2/M checkpoint checks that replication has been completed (amongst other things) as it would be a disaster to start cell division with only a partial copy of the genome. Then the Mitosis checkpoint largely checks that mitotic spindle is correctly assembled to ensure the machinery is in place for division.

Checkpoints can be activated and controlled in several ways depending upon which phase they are monitoring. For example, the three major checkpoints mentioned above are all largely managed by cyclin-dependent kinases (CDKs) as part of the cell-cycle control system.

**Figure 2. Ribonucleotide Reductase (RNR) regulation** in *S. cerevisiae* (Adapted from (Tsaponina et al. 2011)).

For DNA replication, perhaps the most important checkpoint is another one, usually called the S-phase checkpoint. This checkpoint works through mechanisms that are heavily entwined with normal dNTP production (Figure 2); this is discussed further in Paper 3. In order to be replicated, the double strand of DNA must be untwined (see below) leaving unstable single stranded DNA which is then coated by Replication Protein A (RPA). The cell
can cope with repairing one or two small errors within the cell cycle, however a larger problem requires more time. These bigger faults lead to a large amount of single stranded DNA, so RPA is a sort of flag for damaged DNA. When a threshold accumulation of RPA is reached, the S-phase checkpoint is activated. The pathway (Figure 2) begins with RPA recruiting Mec1 and ultimately leads to increased dNTP pools, largely via a reduction in levels of inhibitors of RNR (Crt1, Sml1 and Dif1). Ribonucleotide reductase, RNR, is introduced below.

Whilst a delay is annoying, an aborted take-off is a fairly drastic piece of action, so there are mechanisms in place to ensure it is a rare event.

**The Replication fork**

The sugar-phosphate backbone of double stranded DNA makes it highly resistant to cleavage as there are no hydroxyl groups exposed. As a side note, this is why the incorporation of ribonucleotides into DNA makes the strand weaker as it introduces hydroxyl groups. The pairing and binding of the bases that allows formation of the famous double helix structure also gives the coil a great strength. However, this double strand presents a problem to replication. In order for the copying machinery to access the code it must first open up the strands; this is done by the formation of a replication fork (Figure 3).

Along the genome there are points on each chromosome known as Origins, these are specific sequences that serve as start points of replication. In *S. cerevisiae* they are found roughly every 30kb and are known as Autonomously Replicating Sequences (ARS) which are about 100-200bp in length but all contain the same initial sequence (TTTTATGTTTA) or a slight variation on it (Nieduszynski et al. 2006). These ~400 origins fire at different times during S phase, depending upon DNA packing and other sequences after the initial conserved one. However in other eukaryotes it is thought that chromatin structure may be more important than sequence as conserved sequences have not been found. Things are different in Bacteria, where replication starts from one origin cluster on each chromosome (Morgan 2007) and Archaea are somewhere between the two, having several origins (Wu et al. 2014).
In yeast, origin firing actually gets ‘primed’ back in late M phase (Figure 1) when CDKs are inactive, this is when helicase (the enzyme that prises the two strands of DNA apart) begins to be loaded onto DNA (reviewed in (Tognetti et al. 2015)). However it is not fully assembled and activated until the G1-S transition. Origin firing begins when the aptly named Origin Recognition Complex (ORC) associates with Mini Chromosome Maintenance proteins 2-7 (MCM2-7), which form a ring-like core of the helicase. Then a pre-Replicative Complex (pre-RC) is formed with Cdc6 and Cdt1 (these are together known as the helicase loader) helping, so that two MCM2-7 hexamers open and then close around the twin strands of DNA. This stage is sometimes called ‘origin licensing.’

As it is vital that all sections of the genome are only copied once, helicase loading is inhibited in S-phase by destruction or displacement of the helicase loaders (Mendez and Stillman 2003). Whilst this prevents multiple copies of sections of the genome it is also crucial that all regions are replicated. Therefore this complex loading mechanism has to be highly efficient, although it is not quite a one shot chance; studies have shown that there is a
fail-safe and incorrectly loaded complexes are actively disassembled (Riera and Speck 2015).

Once the helicase has opened out the two strands of DNA (creating the ‘fork’), it is then followed by a primase (associated with Pol α, see below) which creates the initial stretch of product – a primer (actually RNA) for the main replicative construction machinery, the polymerases, to follow on from. There are also proteins required to hold the fork open, they help the polymerases attach and stay attached (the clamp loader and sliding clamp or PCNA) and to remove the RNA primer. This great collection of parts slides along the genome together, like a biological track-laying train.

**Polymerases**

The bulk of the work in the replication fork is carried out by a group of enzymes called DNA polymerases which make a copy of DNA by creating a nascent strand out of dNTPs, stringing them together as dNMPs. A lot of our understanding of these enzymes comes from the Nobel Prize winning work of Arthur Kornberg who was involved in the discovery and study of several examples (Kornberg 1989). To date 17 eukaryotic DNA polymerases have been identified (Bebenek and Kunkel 2004), (García-Gómez et al. 2013) with various functions, however our main interests here are the three major eukaryotic replicative polymerases:

**DNA Polymerase Alpha (α)** is a complex with both primase and polymerase components (Hu et al. 1984). As mentioned above, it initiates replication by making an RNA primer (and is therefore essential (Foiani et al. 1994)) and then extends it by adding around 20 deoxynucleotides. The majority of the genome is then copied by the highly efficient **Polymerases Epsilon (ε) and Delta (δ)**. Epsilon replicates the leading strand (see below) and Delta replicates the lagging (Pursell et al. 2007), (Miyabe et al. 2011); both also have roles in DNA repair. Although, it has recently been proposed that Pol δ may occasionally initiate leading strand synthesis (Daigaku et al. 2015).

These polymerases all have a very similar general structure, comprising of so-called palm, fingers and thumb domains. The DNA is gripped as if the polymerase makes a fist around it. An important difference between Pol α, Pol δ and Pol ε is that the latter two have a built in error-checking mechanism which is described below.

The nucleotides in a strand of DNA bind to each other such that all pentose sugar rings always sit the same way, for example with the $5^{\text{th}}$ carbon at the top and the $3^{\text{rd}}$ carbon at the bottom. This is called directionality and is
written as 5’ to 3’. As mentioned above, the two DNA strands are complimentary but this chemistry means that whilst one strand runs 5’ to 3’ the other must therefore run 3’ to 5’. Now DNA can only be replicated in a 5’ to 3’ direction as the polymerases add a nucleotide to the free hydroxyl group that is only found at the 3’ end. As two totally separate replication ‘machines’ running in opposite directions would make the produced strands end up unconnected/disassociated, another solution is required. What happens is that one strand is synthesised continuously (known as the leading strand), the other (the lagging strand) is made in short stretches (known as Okazaki fragments, Figure 3) which are then joined together by an extra enzyme, DNA ligase. Add in the fact that each stretch requires its own primer to be made, and then removed, and this makes lagging strand synthesis inherently more complicated.

Maintaining accuracy

![Diagram showing factors affecting genome integrity](image)

**Figure 4. Factors Affecting Genome Integrity.** The three major systems in place, each increase fidelity to such a degree that only about 1 in 10 billion bases is replicated incorrectly.

There are several processes that determine the fidelity of DNA replication (Figure 4), and all are highly conserved throughout almost all species.
**Skilled workers and a good supply chain**

The first factor is the DNA polymerase’s selectivity to insert the correct nucleotide when synthesizing DNA; that is the accuracy of the enzymes copying the genome. Although the major replicative polymerases, Pol δ and Pol ε, are high fidelity enzymes (see below), their accuracy is dependent upon the supply of dNTPs. Although the bricklayer is good, they cannot build a house without bricks! This is where the site manager comes in, to make sure there is an adequate supply of materials.

![Enzymes involved in dNTP synthesis](image)

**Figure 5. Enzymes involved in dNTP synthesis.** 3. dCMP kinase. 4. dCMP deaminase. 5. dUTPase. 6. dTMP synthase. 7. dTMP kinase. The chart represents the relative levels of each dNTP maintained in unsynchronised *S. cerevisiae* cells.

Of the seven main enzymes involved in dNTP synthesis (Figure 5), Ribonucleotide Reductase is key in controlling the levels of all four dNTPs in all cells. RNR catalyses the rate-limiting step in the production of all four dNTPs for the synthesis of nuclear and mitochondrial DNA (Reichard 1988),(Thelander 2007). In yeast, under normal conditions RNR consists of a complex of several subunits; a large subunit, which exists as a homodimer of Rnr1 proteins or a heterodimer of Rnr1/Rnr3 proteins (Rnr3 levels are...
hugely increased in response to DNA damage), and a small subunit comprised of Rnr2/Rnr4 proteins. However, different forms of RNR complex have been observed (Hofer et al. 2012). The large subunits contain allosteric specificity sites that modulate enzyme activity and control the balance of the four dNTPs by influencing the specific ribonucleoside diphosphate reduction reaction within the catalytic sites (Thelander and Reichard 1979). This means that there is a feedback mechanism whereby the enzyme’s own products control what it continues to make/process. A form of RNR exists in practically all organisms and in eukaryote RNR Class 1 a highly conserved loop of 13 amino acid residues (Loop 2) connects the allosteric specificity and catalytic sites and is crucial for the correct allosteric regulation of the enzyme (Hofer et al. 2012),(Xu et al. 2006). Interestingly, the four dNTPs are maintained in a non-equimolar balance (Figure 5) with levels of the highest (dTTP) being approximately five times higher than the lowest (dGTP).

Even if the supply of dNTPs is correct, it is important that the polymerases select the correct ones to match (compliment) the template DNA, and this is where the polymerase’s selectivity comes in. Their accuracy at this is approximately 99.99%, meaning they make only one error per 10,000 bases replicated (Figure 4). In case of an erroneous selection, the afore mentioned finger domain will not close correctly around the strand, so replication cannot usually proceed (Rais Ganai, Umeå University, PHD Thesis, 2015).

**The Site Supervisor?**

The second mechanism also involves the polymerases and is called proofreading, in which errors are removed from primer termini during replication by a 3’-5’ exonuclease activity. This is where the polymerase (both and δ and ε) itself recognises a mistake, pauses and goes back to remove the erroneous nucleotide before inserting the correct one. In order to do this the DNA must be shifted from the polymerase active site to the exonuclease site within the polymerase, this can happen whilst the polymerase is still attached to the DNA or with dissociation and re-attachment (Ganai et al. 2015). The mechanisms behind these actions are not fully understood but it is known that the thumb domain is involved as part of its role in stabilising DNA in the active site (Swan et al. 2009). However, this exonuclease ability increases the polymerases’ accuracy by over 10 fold for eliminating base substitutions and over 4 fold for Pol δ and over 100 fold for Pol ε when considering single base deletions (Pursell and Kunkel 2008). This equates to them only making about 1 mistake in every million bases replicated.
Pol α does not possess a proofreading site and is therefore much less accurate than fully functional δ and ε (at least 5 times higher rate for base substitions (Pursell and Kunkel 2008), however Pol δ has been reported to proofread and correct errors from α (Pavlov et al. 2006). A recent study however has estimated that 1.5% of the genome is replicated by Pol α (Reijns et al. 2015), a figure that is of potential significance given its higher error rate compared to the other two replicative polymerases.

Thus the exonuclease proofreading ability of Pol ε and δ brings the fidelity of the replicative polymerases up to around just one mistake per million bases copied.

The Building Inspector

Errors that escape proofreading can still be repaired post-replication, through the mismatch repair system (MMR) which corrects mismatched bases incorporated during replication (reviewed in (Kunkel and Erie 2005). It is not yet fully understood exactly when and how MMR recognises the new DNA strand in eukaryotes. In Bacteria the old strand is methylated whereas the newly synthesized one is not, however this is not the case in eukaryotes. One theory is that nicks in the DNA strand due to ribonucleotide incorporation help strand identification (Lujan et al. 2013) and another is that MMR is connected to the replication machinery (for example (Hsieh 2012)).

The major components of MMR are the bacterial homolog MutS proteins, a heterodimer of either Msh2-Msh6 (MutSα) or Msh2-Msh3 (MutSβ) that recognises and binds to the mismatch. MutSα is mainly responsible for repairing single bases-base mismatches, short insertions and deletions (indels) and small loops whereas MutSβ is involved in larger loop repair. Therefore, Msh2 is essential for MMR (Harfe and Jinks-Robertson 2000) and loss of this activity elevates mutation rates (Johnson et al. 1996).

A fully functioning Mismatch Repair system brings the fidelity of DNA replication up to near 100%, incredibly making just one mistake per 10 billion bases replicated.
DNA Errors and Repair

The Maintenance Crews

These exceptionally rare mistakes are still not necessarily there to stay. There are also several other systems that can find and repair errors even after DNA has been replicated (perhaps including errors from replication but also errors that appear for other reasons, see below). In fact “several” is a bit of an understatement. Even by the turn of the century 130 genes had been identified that had roles in human DNA damage repair (Wood et al. 2001), and perhaps now people have stopped estimating the total number. These repair systems include Excision Repair (which detects distortions in the helix due to an error on one strand, MMR can be included here with Base and Nucleotide Excision Repair (BER and NER), Lesion Repair (a chemical repair of faulty (not incorrect) bases), Post Replication Repair (which includes the BRCA proteins) and Double Strand Break (DSB) repair (which involves proteins including ATM). There are also translesion polymerases that allow DNA replication to bypass faulty bases if it has not been detected and repaired.

Mutations and DNA Damage

So what happens when these DNA integrity systems fail? This could potentially be as seemingly minor as a single base substitution or as major as loss of part of, or even a whole chromosome.

Common Types of Error;

Mutations from Replication

The most common types of mutation arising from mistakes of the replicative machinery are shown in Table 1. This could be just the 1 in 10 billion long shot or due to disease or ‘external’ pressure (see below) on the replisome. All can have varying effects upon the organism depending upon many factors. The consequences are largely dependent upon the position of the mutation; at its simplest this could be whether the mutation is in a coding or non-coding section of the genome. Then there are other factors, for example if the organism is haploid it only has one copy of each gene per cell, so a mutation is more likely to have an effect (for example on gene expression) than in diploids that have two copies. A similar situation can arise in diploid cells due to haploinsufficiency or in the human sex chromosomes.
<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Location</th>
<th>Cause</th>
<th>Spotted by;</th>
<th>Disease examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indels</td>
<td>Insertion or Deletion of a base</td>
<td>Most common at repeat sequences</td>
<td>Primer/template slippage</td>
<td>MMR</td>
<td>Cancer (MMR often mutated in cancer)</td>
</tr>
<tr>
<td>Base Substitution</td>
<td>Changing of one base to another Transition or transversion</td>
<td>Most likely sequence context is important</td>
<td>Polymerase error, e.g. enhanced due to imbalanced dNTP pools</td>
<td>Proof-reading?</td>
<td>Polymerase mutations</td>
</tr>
<tr>
<td>Complex</td>
<td>Insertions or deletions of more than one base</td>
<td>Microsatellite repeats (multibase repeated units)</td>
<td>Mis-alignment of template</td>
<td>MMR</td>
<td>Cancers; Microsatellite instability</td>
</tr>
</tbody>
</table>

Table 1. Mutations from DNA replication. Showing most common location in the genome, what is likely to have caused the error, what mechanism is usually responsible for correcting it, and diseases that could increase it.

The above could be described as macro-positional effects, however there are also more localised or micro-positional factors. Nucleotides in a sequence code for amino acids, with a specific batch of three nucleotides coding for a specific amino acid. Amino acids are the building blocks of proteins and if one of the nucleotides changes (a substitution) then the amino acid and therefore protein produced may also change. There is some redundancy in the system meaning that there are several different combinations for some amino acids. For example if the triplet of nucleotides CGC were to mutate to CGA or CGG, then the amino acid produced would always be Arginine. If there is a change in sequence but not amino acid then this is known as a silent mutation. However an insertion or deletion will shift the amino acid code along, a so-called frameshift. Think of it like a queue; if someone in front of you gets a friend to take their place then you are still in the same position in the queue, whereas, if the person in front of you leaves or someone else pushes in, then your position in the queue changes. As the amino acids are coded in triplets from the start of a gene, a frameshift therefore almost always alters a gene. This is probably why the major repair system, MMR has evolved as a repeat sequence specialist, in order to protect against the highest risk mutation class.
These mutations are largely formed by misinsertion or misalignment of the DNA and replication machinery. These mechanisms are discussed more in Papers 1 and 2 and later in this thesis.

**Post Replicative Causes of mutations and damage**

DNA is under constant pressure from many external sources. One only needs to pick up a newspaper to see that a new foodstuff or household chemical has been discovered to be mutagenic. As the name suggests, mutagens cause mutations and an accumulation of these generally leads to cancer, so these chemicals are, more often than not, referred to as carcinogenic in the press. However, these substances are often not particularly dangerous in normal everyday low doses, what is usually important is the exposure (amount and time).

**Radiation:** It was proven around a hundred years ago that radiation from many common sources can damage DNA, examples include X-rays, radiology treatments or radon gas in houses in certain areas. However, we are all exposed to a daily radiation source; DNA is sensitive to Ultra Violet radiation as it can cause two neighbouring pyrimidines to bind together (for example, the two thymine groups in dTTPs binding together). Thus over-exposure to sunlight can lead to skin cancer, one of the most common cancers of the Western world.

**Chemicals:** There are many toxins that are proven or suspected mutagens found in everyday life. As long ago as 1761, John Hill suggested that use of snuff led to cancer but it was not really until the 1940s that a chemical was proven to be mutagenic. However, things progressed with the advent of the Ames test in the 1970s (Maron and Ames 1983), the first reproducible (cross-species) test of mutagenicity on cells. Previously, results had shown variability between species, for example, chemicals that were strongly suspected to be mutagens in animals had no effect upon cells. Ames realised that most chemicals are metabolised in the liver and so preactivated chemicals by mixing them with a liver extract before adding them to bacteria. This gave much more comparable results. Known carcinogens or mutagens are far too numerous to list and range from medicines and foods (e.g. alkylating agents in smoked meats) to industrial chemicals, cigarettes, and far too many things used in molecular biology (such as sodium azide or ethidium bromide). Many comprehensive lists of carcinogens and potential carcinogens are available, for example at www.cancer.org.
Infection; Several infections are known to increase incidence of cancer, the notable ones being the bacteria *Helicobacter pylori* which causes gastric cancer in some infected individuals and Human Papillomavirus (HPV) which is responsible for most cases of cervical cancer.

Cancer; Cancer is a disease of replication, consisting of uncontrolled dividing cells. Therefore it is unsurprising that many of the proteins and their associated genes mentioned in this thesis are linked to cancers. Cancer and pre-cancer can be a snowball effect of mutations leading to more mutations with cause and effect becoming hard to separate. Genes in the mismatch repair system are commonly mutated in many cancers, including HNPCC (Hereditary Non-Polyposis Colorectal Cancer) (Kunkel and Erie 2005) and gall bladder cancer (Srivastava et al. 2010). ATM mutations lead to the disease Ataxia telangiectasia which often leads to cancers and of course the BRCA genes are perhaps the most famous due to their well-established link to breast cancer. *In vivo* studies have shown that both Pol ε and δ exonuclease deficient mice develop cancer (Albertson et al. 2009), interestingly each deficiency led to different types of tumours.

dNTP pool imbalances lead to mutations

Previous work using the CAN1 reporter locus has shown that both raised (Chabes et al. 2003) and imbalanced dNTP pools (Kumar et al. 2010) lead to increased mutation rates in *S. cerevisiae*. The latter study introduced single amino acid substitutions into the specificity site of RNR1. These small changes led to the creation of several mutant strains with various dNTP pool imbalances.

A functional mutation within the CAN1 gene inactivates it, enabling cells to grow on media containing the toxin canavanine (see methods section), thus survival in the presence of canavanine can be used to quantify mutation rates. All of these dNTP imbalanced strains had increased mutation rates (between 4 and 14 times higher than wild-type). When the mutated can1 gene was sequenced in many independently formed colonies growing on canavanine, it was also found that two of the strains (with extreme dNTP imbalances) had unique mutational hotspots (Kumar et al. 2011). This meant that the dNTP imbalances were not uniformly increasing the mutation rate along the DNA, they were causing mutations at the same points in the genomic sequence again and again in many independent individuals.
One particularly interesting strain (called \textit{rnr1-Y285A}) had dCTP and dTTP levels around 15 times higher than normal which led to a ~14 fold increase in mutation rates and yet had no pronounced S-phase checkpoint activation.

When the DNA sequence surrounding the mutation was examined in these mutational hotspots it was seen that a dNTP imbalance leads to misinsertion or frameshift followed by impaired proofreading (discussed further in Paper 1). This supported the model proposed from our collaborator's findings from \textit{in vitro} work (Bebenek et al. 1992).

RNR activity increases in cancer cells (reviewed in (Aye et al. 2014)), leading to increased dNTP pools. However, as these cells are rapidly dividing this is to be expected. Again it is hard to separate cause from effect here and more research is required as there is a lot of speculation in the field. Conversely, a shortage of dNTPs in early cancer has been shown to lead to genome instability (Bester et al. 2011). Therefore RNR and its associates have long been targets for cancer therapy.

However, with such complex relationships (Figure 5), it is vital that we gain more information as targeting these pathways has the potential to cure or cause cancer. For example, there are cases of cancer caused by Hydroxyurea, an RNR inhibitor used to treat cancers (Aumont et al. 2015).

\textbf{Study Aims}

Many studies look at the accumulated mutations, the so-called “mutational signature” that forms in cancer patients (reviewed in (Hellday et al. 2014)). What we are interested in here is turning the timeline around. We want to understand the basic pathways involved in a cell’s reaction to disturbances in the DNA replication integrity systems, to try to understand cancer development from the very beginning. For example, what happens when you introduce dNTP pool imbalances or shut down MMR? We can do this by monitoring and analysing many variables and processes, like levels of checkpoint proteins (Figure 2 and Paper 3) and cell cycle progression (all papers).

We can also analyse the final output, studying the DNA sequences that are subsequent from the disturbances we introduced (Papers 1 and 2). Amongst other things, this can tell us whether certain sequences are particularly at risk of mutation under these conditions and also to infer what enzyme is making the mistake. Perhaps we can even match our induced mutational signatures to those known for cancers and help connect the disease start to the end.
The Tools of the Trade;

Yeast as a model

In order to understand a complex system one must break it down into simpler parts. The human body is a highly complicated mix of tissue and cell types with countless interactions. As empirical science demands as few variables as possible, this makes the human being a very challenging subject. Therefore most basic research takes place on the level of cell cultures. However, mammalian cells are difficult to genetically manipulate (although the recent development of tools such as CRISPR, reviewed in (Doudna and Charpentier 2014) is rapidly changing this), so an alternative has been required.

Researchers turned to a simple, every day, single celled organism that was easy and safe to culture, namely Bakers’ Yeast (\textit{Saccharomyces cerevisiae}). The first eukaryotic organism to have its genome sequenced in 1996, it was initially estimated that over 30% of its functional genes had human equivalents or homologs (Botstein et al. 1997). Indeed, you may have noticed during reading this thesis that most of the genes mentioned have their human homologs mentioned in brackets after them. Yeast has an added advantage of existing in stable states as both haploids and diploids (one or two copies of the DNA) and highly efficient homologous recombination, which help make it easier to genetically modify. When working on dNTPs another benefit of yeast is that, unlike mammalian cells, it has no salvage pathway for nucleotides, and this effectively makes each cell a closed system. Yeast’s importance in cancer research was recognised in 2001’s Nobel Prize for Medicine shared by Hartwell, Nurse and Hunt, awarded for work on the cell-cycle largely carried out in yeast.

Methods

Yeast; As mentioned above, it is very easy to get a mutation into a gene in yeast, and widely accepted protocols exist (for example (Longtine et al. 1998)) using PCR and cassettes. When a mutation has been introduced, it can be transferred between strains by simply crossing them by mixing two strains (of different mating type) on media. The mutation can then be tracked by various markers introduced with it, such as auxotrophic markers.

HPLC; High Performance Liquid Chromatography is a well-established technique for the separation, identification and quantification of molecules/compounds. We extract NTPs and dNTPs from cells (see Papers for more information) and then analyse them on a SAX HPLC column. This
separates the different (d)NTPs by their charge and they are then quantified by a U.V. detector.

**CAN1:** The Canavanine Forward Mutation Assay is a widely accepted method for mutation analysis (Hoffmann 1985) and references within). Briefly, canavanine is a toxic analogue (from a plant) of the amino acid arginine. Therefore, when present, it is taken into the cell as if it was arginine, leading to cell death. However a mutation in the gene (CAN1) that codes for the protein responsible can deactivate the protein as it is synthesised wrong or not at all. Thereby meaning the toxin will not be taken into the cell and the cell will survive and grow into a colony. If cells are spread out on a canavanine plate then each individual cell that has a functional mutation in CAN1 will grow into an independent colony. The mutation rate can then be calculated by comparing the number of colonies with the number of cells spread (using survival on normal media as a control).

The 1772 base pairs (bp) of the mutant can1 gene in resistant cells can then be sequenced. It is then simply a case of looking through the sequence and identifying a change (compared to the wild type sequence) which is the mutation responsible for resistance. It is like finding a needle in a small haystack. We sequence a large number of independent colonies (around 150-200) in order to build up a mutation spectrum of the strain and analyse whether mutations keep forming in the same positions in different individuals, which we call hotspots. Here we use a ‘point mutation’ rate 10 fold higher than that in the wild type strain as the definition of a hotspot. We then analyse the DNA sequence on either side of the mutation (10bp) in order to further understand the underlying mechanism of mutation formation.

**Whole Genome Sequencing:** Even with a high mutation rate one does not get many mutations forming in a single or few rounds of replication. Therefore to study spontaneous mutations formed without selection pressure one must grow cells over many generations. To do this we grew yeast strains for a few days (streaked on standard media to give individual colonies) and then took a single colony and re-streaked it. This was then repeated every two or three days for just over two months. After this period, we had cells that have accumulated mutations through approximately 800 generations and passed through 30 bottle-necks.

We then took samples from generations 0 and 800 and compared the changes in their genomes. This was done by extracting DNA and sequencing by Illumina sequencing. This method works by first fragmenting the whole genome into 200-800bp sections and adding paired end tags which help
alignment within the genome. The fragments are then sequenced in small over-lapping segments which can then be pieced back together using software.

Analysis – finding the needles in a really big haystack! Whereas the CAN1 assay required the analysis of 1772 bp of sequence for each sample, whole genome analysis requires the checking of 12 million bases. Therefore a little help is needed from computers and algorithms. The output is a little different too: in order to produce a map of hotspots throughout the entire genome we would need to sequence a huge amount of individual samples. This is because the chance of the same mutation forming at the same point in two individuals with no selection pressure is incredibly low. So what is done instead is to use algorithms to generate a logo of the most common sequence around each type of mutation.
Results and discussion;

Paper 1; Increased and Imbalanced dNTP Pools Symmetrically Promote Both Leading and Lagging Strand Replication Infidelity.

The bulk of DNA replication is carried out by three polymerases, all working in a 5’ to 3’ direction. Polymerase ε replicates the leading strand in a near continuous (for the sake of argument) fashion. Polymerase α and δ are responsible for the lagging strand (although α is required to prime leading strand replication), which is inherently more complex as it has to be synthesised in small pieces (Okazaki fragments) which then need to be joined. The differences mean that each polymerase leads to a recognisable error profile meaning certain mutations display a strand bias. Of the very few mistakes they make, Pol ε mainly makes mainly base substitution errors, whereas with Pol δ it is mainly indels (Korona et al. 2011; Lujan et al. 2012).

Previous work in the group showed that introducing an imbalance in the *S. cerevisiae* dNTP pool by mutating the *RNR1* specificity loop (Kumar et al. 2010) led to increased *CAN1* mutation rates and specific mutation spectra (Kumar et al. 2011). However, these experiments could not tell us the whole story as the strains used had functioning MMR repairing some mistakes. We also wondered if the polymerases responsible for replicating each strand were affected in the same way by dNTP pool imbalances.

In order to investigate the latter, we created a strain with the *CAN1* reporter gene and promoter in reversed orientation; thereby switching the leading and lagging strand synthesis. We used the interesting *rnr1-Y285A* strain as its elevated dCTP and dTTP (Paper figure 1) led to a relatively high mutation rate (14 times wildtype) but with no detectable checkpoint activation. Then we also disabled MMR by deleting its essential *MSH2* gene in order to see the full effect and spectra of mutations made by the polymerases and to determine the repair system’s efficiency.

When the orientation of the *CAN1* gene was reversed we found no alteration in the mutation rate (OR1 = 57x10⁻⁷, OR2 = 60x10⁻⁷) in *rnr1-Y285A*. Then, when we sequenced the mutant *can1* gene in 170 independent colonies from each strain, we also found no difference in mutational spectrum between either orientations. Although this might seem at first like a non-result, it is in fact very interesting, as it meant that the pool imbalance was driving almost identical mistakes on the leading and lagging strands. This means that the dNTP imbalance overrides inherent differences (and mutation profiles) between the polymerases and becomes the mutational
force majeure (Figure 6). However, there were a couple of exceptions where hotspots were unique to either Pol α/δ or Pol ε, implying that there may still be some difference between the polymerases at certain sequences contexts. When these sections of sequence (10bp either side of the mutation) were analysed, they were found to be have very similar nucleotide sequences to ones which did not show this strand bias. As this did not give enough information to explain these site differences, this could be a limitation of the CAN1 assay and we hypothesized that whole genome sequencing might shed further light on these exceptions.

![Figure 6](image)

**Figure 6. The strand symmetry of mutational hotspots from a yeast strain with elevated dCTP and dTTP.** The horizontal bar represents the length of the CAN1 reporter gene and the vertical bars represent the frequency of point mutations in approximately 170 individuals with the gene in natural (OR1) and reversed (OR2) orientation.

We also found that, whilst the elevated dCTP and dTTP or the loss of MMR each individually raised the CAN1 mutation rate by about 14-fold, if they were combined then the rate increased over 500-fold. When we analysed the mutational hotspots it was evident that, whilst they appeared to be a combination of dNTP pool and MMR driven sites, each could be explained by the sequence context and pool imbalance. Briefly, a mismatch is formed when the required dNTP is in relative shortage, the polymerase
struggling to find the correct one and instead inserting a different one. Normally the polymerase (ε or δ) might recognise its own mistake and shift the template to its exonuclease site to remove the error. However, when the following nucleotides required in the sequence to be synthesised are dNTPs that are in hugely elevated concentrations, the polymerase is effectively pushed forward and rapidly extends past the mismatch.

MMR is known to be a specialist at repairing mistakes in repeats, thus in its absence, if the sequence context features a repeat next to a sequence that is sensitive to this dNTP imbalance then the mutation rate increases vastly due to synergy between the two genotypes.

The efficiency with which MMR functioned also varied from repairing 799 out of 800 mistakes at some sites to just 3 of 4 at others. A pattern was evident in the hotspots in that G-C base pairings in mononucleotide repeats were repaired mainly with correction factors of less than 10 whereas for A:T base pairs it was over 200. The reason for this disparity could be that MMR itself requires a ‘normally’ balanced dNTP pool. When MMR recognises and removes a mistake, it may then recruit a polymerase that makes the same mistake again due to it having to use the same perturbed dNTP pool. This is supported by the fact that, at least at some hotspots, MMR was more efficient in the presence of wildtype pools.

**Paper 2:** Genome-wide analysis of the specificity and mechanisms of replication infidelity driven by a mutation in ribnucleotide reductase that imbalances dNTP pools. (Manuscript)

Whilst the previous paper gave valuable insights into the effects of the high dCTP and dTTP coupled with loss of MMR in the *S. cerevisiae rnr1-Y285A* strain, there are certain limitations to the CAN1 assay. First of all, it is only a tiny fraction of even the yeast genome; this means that it cannot contain all possible sequence contexts. For example it does not contain many heteronucleotide microsatellites (multiple repeats of small mixed nucleotide motifs). Another flaw is the assay itself. The fact that a mutation is required for resistance to a toxin means that the results could be potentially selective. As mentioned previously indels are much more likely to be functional mutations whereas substitutions have a higher chance of being silent. The natural progression from Paper 1 was therefore the first whole genome sequencing study with imbalanced dNTP pools.

Here, we created a diploid version of the *rnr1-Y285A msh2Δ* yeast strain and grew it on normal media for 800 generations. We made the strain
diploid in order to reduce the loss of individuals with mutations in essential genes and inactivation of MMR also meant removal of any mutational bias it creates. The lack of selective media meant that the mutational stress to the cells came primarily from the imbalanced dNTP pool and the loss of MMR. As the growing phase of the experiment proceeded, the \textit{rnr1-Y285A msh2}\textDelta{} cells became less and less healthy, with slower growth, disturbed cell cycle (Figure 1B in the manuscript) and also showed morphological changes. Some appeared spikulated and others developed large vacuole-like structures. However, dNTP pool imbalances remained approximately the same in cells between the start and end of the experiment.

The genome wide mutation rates from sequencing showed the huge importance of MMR on a genome scale; whereas a mutation would appear once in every 250 replicated genomes in wild type yeast this increased 100-fold to 1 error in every 2.6 genomes when MMR was inactivated (Lujan et al. 2014). With the \textit{RNR1} mutation that imbalanced the dNTP pool (20x dCTP and 16x dTTP compared to wild type) on top of the loss of MMR there were 3.6 errors per genome! Thus the mutation rate of \textit{rnr1-Y285A msh2}\textDelta{} was 900-fold higher than that of wild type. This figure is quite comparable to the >500-fold increase the haploid strain gave using the \textit{CAN1} assay (Paper 1).

On a genome-wide scale, the mutations were dispersed throughout all chromosomes (Figure 2A in paper). However 84% of the mutations in the whole genome were base substitutions compared to just 50% in the \textit{CAN1} assay supporting our theory that the latter is missing silent mutations.

The mutation rates for specific errors increased in a manner that was consistent with the dNTP imbalance, i.e. a base in excess being mispaired in place of the correct base that was in relative short supply. For example, the rate increased 86-fold for C-G to A-T mutations via a C-dTTP mispair, 27-fold for T-A to G-C mutations via a T-dCTP mispair, and 46-fold for deletion of a C-G base pair via an unpaired template C. The mutational pressure of the \textit{rnr1-Y285A} dNTP pool towards mispairs was evident in that total base substitutions increased 26-fold in rate when compared to the MMR deficient single mutant whereas the rate of indels only rose 2-fold.

Algorithmic binning of mutations and their surrounding sequences revealed that errors were predominantly induced at C-G base pairing, which were located in a sequence motif of six nucleotides that facilitated increased misinsertion and also increased mismatch extension at the expense of proofreading (Paper Figure 3). The motif for the most common mutation, a substitution, was 5'-AAAGCT-3' where C was the location of the mutation, note that the nucleotides required after the site of the mismatch (CTTT) are the ones in huge excess in this strain. This is very similar to the motif for the
The most common deletion of a lone G-C pair (i.e. excluding mononucleotide runs where deletions are most common); 5'-GAAACT-3'. It could be that the position of the G in the sequence helps determine if there is a mismatch or a deletion.

The mutation rates were then analysed relative to replication origins and this showed that the error rates were again very similar between leading and lagging strand replication. This confirmed our findings in Paper 1 that this dNTP pool imbalance over-rides the replicative polymerases’ intrinsic error profiles (see Paper 1 section, above). The analysis also enabled us to show that the rates with which errors were made also became greater as S phase progressed, suggesting that later replicated genes could be more at risk of mutation in the presence of a dNTP pool imbalance.

When the wide-scale positions of errors were analysed it was found that both single base substitution and deletion rates increased more than 3-fold in coding sequences compared to intergenic regions of the genome. This is partly explained by the fact that deletions were predominantly in short homonucleotide runs which are enriched in coding sequences.

Another interesting observation was that the dNTP pool imbalance increased single base deletions at G-C base pairings much more than it did at A-T pairings, despite both dCTP and dTTP being raised. This could be due to dGTP being the most limiting of the nucleotides in this strain.

A final interesting result with single base deletions was that when the length of mononucleotide repeat was analysed, the dNTP pool imbalance increased both types of deletions most at 4bp long repeats, the most common motif being 5’-GACCCCA-3’. This could be related to shape of the polymerase gripping the strand and exonuclease activity has previously been shown to be less efficient in runs of four C’s compared to 4 A’s (Greene and Jinks-Robertson 2001).

The pool imbalance increases all mutations but hugely increases base substitutions, swinging the latter from 38% of all mutations in msh2Δ to 84% in rnr1-Y285A msh2Δ. Our data suggest that this particular dNTP pool imbalance places coding sequences and late replicating genes more at risk of genome instability. This could be highly significant as it means that parts of the genome that are normally evolutionarily protected are under threat of mutation in the presence of dNTP disturbances such as those that could be seen in cancer or infection.
Paper 3; Molecular Basis of the Essential S Phase Function of the Rad53 Checkpoint Kinase.

A large amount of replication fork stalling leads to checkpoint activation via (in yeast) the essential kinases Mec1 and Rad53. This prevents additional origins from firing when DNA is damaged or DNA replication is blocked. Like traffic control, it slows or stops the motorway to prevent a multiple pile up and it also shuts the sliproads in order to stop more traffic joining the motorway.

Whilst it is known why Mec1 and Rad53 are vital for S-phase checkpoint activation following DNA damage, it is not known why they are required for normal DNA replication.

Mec1 phosphorylates Rad53 in one of two ways, either via Mrcl (the replication checkpoint pathway) or via Rad9 (the DNA damage checkpoint pathway). Rad53 then phosphorylates Dun1, a downstream checkpoint kinase, which in turn leads to activation of ribonucleotide reductase and elevation of dNTP pools.

To investigate the role of Rad53 in normal replication, our collaborators created a RAD53 mutant that lacked the N-terminal phosphorylation site cluster; rad53-4AQ. This mutation was found to be synthetic lethal with the deletion of RAD9 but not with mrc1, suggesting that the Rad9 pathway is the more important of the two in activation of Rad53 in rad53-4AQ. Known as a DNA damage checkpoint adapter, Rad9 allows Mec1 to phosphorylate and activate Rad53. We showed that this lethality was because rad53-4AQ cannot activate the downstream kinase Dun1.

To create a docking site to bind Dun1, two Mec1 motifs in its SQ/TQ cluster domain (SCD1) must be phosphorylated on Rad53 and the rad53-4AQ mutation prevented this. This inability to activate Dun1 led to decreased basal dNTP pools, replication fork stalling and subsequent activation of the S-phase checkpoint. The latter then led to a compensatory increase in dNTP levels and survival, meaning the strain became dependent upon the checkpoint even for undamaged DNA replication.

Therefore, we found that the primary function of Rad53 in normal S phase is its Dun1 mediated regulation of basal dNTP pools. This requires a very low kinase activity, in order to slightly decrease Sml1 levels in contrast to full-blown checkpoint activation that requires almost total Sml1 degradation. Rad53 kinase activity appears to become essential in cells where dNTP levels dip below about 60% of those in wildtype cells. This leads to Mec1 activation which then leads to Rad53 SCD1 phosphorylation which
in turn promotes Dun-1 dependent Sml1 degradation and subsequent rise in pools.

**Figure 4. Simple model of DNA replication in normal S-phase; living on the edge of checkpoint activation.**

These results give a picture of a normal cell hovering around sufficient dNTP levels in S phase. Indeed the concentration of dNTPs in a cell has been estimated to be around 4% of that required to construct a genome (Kumar et al. 2010). The reason for this is obvious: as mentioned before, too high a concentration of dNTPs leads to increased mutation rates. So the cell is operating on a “supply as demanded” production line. It stands to reason that a slight hindrance in this supply (or maybe even an increase in demand) is a distinct possibility during normal DNA replication. Perhaps a cell encounters favourable growth conditions and the replisome gets ahead of its suppliers (thereby depleting pools) or perhaps an external substance mildly inhibits RNR. Therefore the cell has evolved a safety net; if the dNTPs become limiting, the S-phase checkpoint kicks in, like an emergency generator but just running on tick-over. This gives a carefully controlled boost to return the dNTP pools back to normal without full checkpoint activation.
Discussion

A cell is always living in a delicate equilibrium between accuracy of replication and genetic variation. The DNA integrity system is a highly complex collection of interactions, for example yeast Msh2 (which is essential for MMR) has around 150 genetic and physical interactions identified to date (yeastgenome.org). Central to reliable DNA replication is the accurate supply of dNTPs and the whole replicative system is finely balanced with slight dNTP pool imbalances leading to increased mutation rates. However the paradox to this is that dNTP pools are not in a steady state during the cell cycle. It has been reported that dNTPs rise in S phase (relative to G1 levels) with some estimating a 10 fold rise in this period (Pontarin et al. 2012), although they are maintained in approximately the same non-equimolar balance relative to each other (Chabes et al. 2003). This suggests that dNTPs are raised when needed as part of the natural cell cycle, without causing a problem. However, if they are raised at other times does this cause problems, or is it the alteration of the balance that is important?

Our work here supports the latter theory (but does not discount the first), showing that the balance between the dNTPs is very important to genome integrity. Imagine a boat with four engines (the dNTPs); if you open the throttle and increase the fuel going into the two engines on the right the boat will not run straight, this is like the effect of a dNTP imbalance on polymerase selectivity. A similar analogy can be used for the effect of a dNTP pool imbalance upon proofreading; if polymerases have too much fuel (high dNTPs are the equivalent to pushing the accelerator) – the steering does not have time to work, i.e. there is less chance of moving DNA to the exonuclease site.

Thus we can form a model of dNTP pool imbalances driving polymerases to make mistakes (Figure 7). What is important is a nucleotide being required that is in relative shortage of supply having a following sequence that requires nucleotides that are in excess.

The S-phase checkpoint is a regulatory feedback system that can increase the dNTP levels up to 8 fold (in yeast (Chabes et al. 2003)) if DNA damage is encountered. This raising of the dNTP pools in response to stress could be seen as switching on an “evolutionary fire”, allowing infidelity to enable survival through times of trouble. However it is a bit like a factory production line where the quality drops in times of high demand. The rnr1-Y285A strain is particularly interesting because there appears to be no activation of the S-phase checkpoint (Kumar et al. 2010) and data not shown here, despite large rises in dCTP, dTTP and mutation rates. However, when we ran Western blots of 800 generation old diploid rnr1-Y285A msh2Δ cells,
we found some interesting variability. The protein expression profile (for Sml1 and the Rnr proteins) did not always fit that expected for yeast cells with either activated or unactivated S-phase checkpoint. One isolate in particular had a large increase in Rnr3 levels (Rnr3 is induced after DNA damage or replication stress) and yet normal Sml1 and Rnr1 levels. A large increase in Rnr3 would suggest the S-phase checkpoint has been activated (see Figure 2) and Dun1 is inhibiting Crt1. However, if Dun1 is activated, then Sml1 should be degraded. It could be that the genes for some of these proteins have themselves been mutated and are not functioning properly. Alternatively, it could be that the cells are trying to adjust their dNTP pools without success. The observation of large vacuoles in the cells could be due to sequestering of excess Rnr in vacuoles where it will then not be available to the cell as has been previously reported (Ma et al. 2011). It is quite interesting that even after all those generations the cells had not managed to normalise their dNTP pools. It appears that, despite the many enzymes and feedbacks involved in dNTP synthesis RNR is vital to their balance.

The final consideration is the mismatch repair system. It is the icing on the cake of the incredible system of DNA replication that only makes 1 mistake in every 10 billion bases. Deficiency in MMR is now synonymous with cancer which is not surprising since its loss increases mutations to 1 every 2.6 genomes (at least in yeast). We show that there is a synergy between the loss of MMR and imbalanced dNTP pools. Recent studies have shown that infection can affect dNTP pools, with HIV increasing them (Amie et al. 2013) and HPV decreasing them (Bester et al. 2011). Also, a recent study demonstrated that a mutation in polymerase δ (Mertz et al. 2015) found in cancer patients led to raised dNTP pools when introduced into yeast. There are undoubtedly mutations in many other parts of the replicative machinery with similar potential effects. If such a mutation was combined with MMR deficiency then the results of the above mentioned synergy would be drastic. The *rnr1-Y285A msh2Δ* yeast strain makes around 4 mistakes every time it copies its genome, by simple unscientific scaling up this would equate to about 1000 mistakes per cell division in humans.

Preliminary clinical data seems to suggest this situation is indeed found in humans, as many cancer patients with replicative polymerase mutations also have MMR mutations.
Figure 7. Model for dNTP pools driving mutations. The roller-coaster cars symbolize a replicative polymerase and the track the DNA template. The relative excess leads to a misinsertion opposite the C, the following bases required are in excess and push the polymerase into rapid extension before proofreading has a chance to act (the passenger in the back of each car?).

We have created a library of yeast strains with different dNTP pool imbalances, these are a good tool to analyse replication pathways and mutation formation in cells in the laboratory. However, it is important to note some important differences between the mechanisms of the model we study and humans. One is that the allosteric activity site in yeast RNR is relaxed, whereas in mammals it strictly sets a limit for dNTP pools (Chabes et al. 2003). This means that S-phase checkpoint activation in yeast can lead to a much greater rise in dNTP pools than in mammalian cells. The reason behind this is that, as a unicellular organism, yeast has to try everything to survive. If a yeast cell’s DNA is damaged, raised dNTP pools can increase tolerance to this damage but at the cost of higher rates of mutation (Sabouri et al. 2008). However a multicellular organism does not have the same pressure to tolerate mutations in each cell after DNA damage, and therefore it can afford to sacrifice a faulty or damaged cell for the greater good.
Potential future directions

As mentioned before, yeast is a nice closed system. However cells in a human body are anything but. Mammalian cells have a nucleotide salvage system and exist in a cellular matrix within a circulatory system with conditions often in high degrees of flux. Even in a yeast cell there are many enzymes other than RNR involved in dNTP synthesis (Figure 5), what happens if we mutate or over-express some of these? It could be that they produce very different dNTP imbalances or alter ribonucleotide levels and the ensuing effects on the cell and mutations could be very different. There are also new techniques of analysing genome-wide ribonucleotide incorporation (Clausen et al. 2015), perhaps imbalanced dNTP pools affect this incorporation due to their effect on polymerase selectivity.

Another potential area of research would be to move into a more complex model such as a mouse RNR1 mutant, this would allow the study of the effects of dNTP imbalances upon multicellular organisms, such as tumour formation. At the moment we have looked at the abacus, next we have to study the super-computer!
Acknowledgements

There are so many people to thank, I apologise if I miss anyone. Anyway, this may be the only opportunity I get to write an Oscar speech, so here goes...

First of all, thanks to my amazing family. Thanks for giving me something to come home to on the good and the bad days! Also Phil and Tina, for everything – especially the fika and for providing me with my Hulk Tank! ;-) Of course also to my family who couldn’t escape the UK, thanks for all your support. I love you all.

Then, I would like to add a second dedication to all the technicians I have worked with over many years in academia (21 years, bloody hell!). Let us not forget that most of the real science takes place on the technician’s bench and not in professors’ heads. However, almost all of them feel unappreciated. So thank you all, you have taught and helped me so much! I hope that didn’t sound patronising... ;-)

Thank you Andrei, especially for giving me the opportunity to work with some wonderful people, in such a beautiful place. To all my group members, past and present – it has been a pleasure to work with you. Especially Anna Karin and Lisette, you two are just amazing! Danielle (my genius co-author), Sushma and Andrea – you are all so helpful, knowledgeable and nice! Jörgen – you left me the msh2Δ strain but on the other hand you introduced me to innebandy. I also appreciate your attempts to teach me what a ‘proper’ Swedish man should be like! Phong, there’s never a dull moment with you around. ;-) Also, thanks to Vladimir, Olga, Dinesh, Anna Lena, Paulina, Daniil, Yulia and finally to my fallen comrade, Balu – I hope you found somewhere to be happy.

Our collaborators over the pond; Tom and Scott. Once Danielle translated what you meant, it was great! ;-) 

Erik and his group. Especially Ulf, Rais and Peter (my Brain Trust); I often go to their office to quiz them about polymerases but they just want to talk about films, cricket and whatever (respectively) which was much more fun! ;-) You guys are all as awesome as Rais!

Clas, you are so nice and helpful and a good friend – our corridor is so hollow without you! Ingrid and Anna, you are also great.
Björn, when you, me and Ulf get together and talk films we are a danger to ourselves. Although, people should listen to us more... for the greater good.

Yue Chen, I miss you buddy – when are you coming back to the promised land?

Teaching Labs, special thanks to all the guys I have taught labs with (and of course the students!) – it was so much fun! Just a shame these damn PHDs get in the way of it! ;-) Special mention to my smiling ‘Incan’ friend: Melissa, one day we will perfect the good cop, bad cop teaching method!

Although he only just arrived, Sjoerd – it’s great to finally have someone to talk football with in this land of figure-skating with sticks! ;-) The guys in Endocrinology and Oncology at the hospital, thanks for keeping me running at a normal rate, although your radioactive cocktails suck!

Tor; you have much wisdom and a great sense of humour. Thanks for all the advice, amongst other things you told me to keep my acknowledgements short. How’s this? ;-)

Finally, to everyone else at MedChem, past and present; Thank you, I will miss you all.

“There’s a time when a man needs to fight, and a time when he needs to accept that his destiny is lost, that the ship has sailed, and that only a fool will continue. The truth is, I’ve always been a fool”

Edward Bloom, Big Fish.


Tomasetti, C. and B. Vogelstein (2015). "Variation in cancer risk among tissues can be explained by the number of stem cell divisions." Science 347(6217): 78-81.


