DNA polymerase mutations as drivers of genome instability and cancer

Mareike Herzog
Wellcome Trust Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Christ’s College 2016
“Think, think, think.”
- Winnie-the-Pooh
Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This dissertation contains less than 60,000 words excluding appendices, bibliography, footnotes, tables and equations and has less than 150 figures.

Mareike Herzog
2016
Acknowledgements

First I would like to thank my supervisors Dave Adams, Steve Jackson and Thomas Keane for the opportunity to be a member of their teams, learn from them and earn my PhD with them. I want to thank them for the warm welcome I received, their patience, their kindness and their support. I am grateful to Dave for his many thoughts and advice on my work, for all his encouragement in stressful times and his trust in me, to Steve for never being too busy to talk to me and for the independence in exploring my projects and to Thomas for putting himself through the gruelling process of introducing me to bioinformatics.

I would like to thank all members of the Experimental Cancer Genetics Team at the Sanger Institute, for their friendship and their enthusiastic willingness to help me with anything I need: Daniela, James, Manu, Sofia, Mamun, Vivek, Martín, Clara, Stefan, Marco, Nicky, Chi, Louise, Rebecca, Marcela, Richard and Alistair. I want to especially thank James for supporting my with everything I could imagine: from getting me a desk, supplying me with reagents to getting me started in the world of tissue culture. Clara Alsinet Armengol and Manu Supper have my gratitude for teaching me how to take care of embryonic stem cells and for being great "babysitters" when I had to be away from the institute. My mouse work was greatly aided by Louise van der Weyden, who is the goddess of mouse work and a super friendly one at that. I am indebted to Mamun Rashid for his help with the EMu signature extraction and to Vivek Iyer for helping me manage import and storage of my mouse sequencing data. Martin Del Castillo Velasco Herrera and Stefan Dentro were providers of great advice for all things statistics and bioinformatics. And I cannot thank Daniela Robels Espinoza enough. Not only was she the patient and kind with all my questions, taught me endless programming tricks and helped me fix some tricky bugs, she also put up with me as a roommate. She has been an endless source of motivation, reassurance and encouragement for me.

I would also like to thank all of the members of the Jackson Group at the Gurdon Institute: Fabio, Israel, Josep, Yaron, Gabi, Rimma, Serena, Delphine, Paco, Paul, Will, Matt, Donna, Natasha, Pallavi, Chrstine, Andy, Sati, David, Carlos, Ryotaro, Jessica, Jon, Abdul, Linda, Nicola, Julia, Helen, Kate, Gopal, Matylda, Siyue, Muku and Ana. To highlight but a few, I will always remember Rimma Belotserkovskaya, because she was the first person from the lab
I met when she lectured me in my undergraduate days and got me interested in DNA repair, Julia Coates, because she shared a "bench" with me for years and we suffered the institute’s temperamental air conditioning together and because she helped me move house no questions asked, Serena Bologna, for involving me in her project, being a great friend and speaking Italian to me, Linda Baskcomb, for being such a competent lab manager, and Nicola Geisler for supervising me during my rotation project, which convinced me to do my PhD in this group. Special thanks goes to Josep Forment, who invited me to collaborate on a rewarding project with him. He is a joy to work with and I will miss him greatly. I am also deeply grateful to Israel Salguero Corbacho, who helped me a lot with designing my plasmids and my cloning for the yeast strain construction. He is kind and funny and always happy to help me with anything. My deepest gratitude goes to Fabio Puddu, who, more than anyone, has been my daily mentor on this journey. From teaching me lab techniques (pulse-field gels are fiddly) to discussing and planning experiments (why would we run this analysis?), he has tried his best to make me a better scientist and I can only hope that somewhere along the way I was able to teach him something in return. I am grateful for all the criticism, encouragement, laughter, patience, praise and support.

I would like to acknowledge the Wellcome Trust for funding my PhD. Without their generous support this work would not have been possible and their funding has allowed me the privilege of not having to worry about feeding myself, paying the bills and having a roof over my head.

My friends have filled my days with joy and were always there with advice when needed and ultimately contributed much, that is intangible, to this project. I am very happy to count Chrissey, Böcki, Stephie, Monica and Julia as some of my best friends. Finally, I want to thank my family. My parents, Martina and Stefan, and my "baby" brother Gunnar, for their never ending support, motivation and their infinite love. And to Löffelchen I would just like to say with all my love "I couldn’t have done it without you. Ich habe dich über alles lieb."
Genomic stability is essential to preserve the genetic information encoded in DNA, and many biochemical pathways are devoted to repair DNA damaged by external factors, or during the course of essential cellular processes such as transcription and DNA replication. Malfunctioning of these processes may alter the DNA, leading to abnormal cellular behaviour or cell death, which in multicellular organisms may be associated with disease. For this reason, the machineries that safeguard the integrity of eukaryotic genomes are of prime interest to research in the areas of ageing, rare disease and cancer. Every time a cell divides, duplication of the genome is principally carried out by two DNA polymerases — Pol δ and Pol ε — which are highly processive and accurate. Together with polymerase gamma, which is active in mitochondria, these are the only human polymerases known to possess "proofreading" activity, making them extremely accurate. In parallel, cells have also evolved a repair system for base mismatches, to identify and correct mispaired bases occasionally produced by DNA polymerases. While it has been known that defects in mismatch repair promote carcinogenesis, mutations in replicative DNA polymerases driving tumorigenesis in mismatch repair proficient cells have only been recently identified. Here, I report the interrogation of twelve such DNA polymerase mutations for their potential to alter genetic information and contribute to genomic instability using the budding yeast *Saccharomyces cerevisiae* as model system. Of all the polymerase mutations tested, a subset caused significant increases in mutation accrual, and a shift in the observed mutation patterns/signatures. Most intriguingly, I observed that these increases are more severe than those caused by mutations disrupting the proofreading activity of the corresponding DNA polymerase, with my results further indicating that in some cases the high mutagenic potential depends on the proofreading activity. These strong increases in mutation rates do not likely result from inhibition of mismatch repair, as combination of these mutations with loss of mismatch repair factors results in synthetic sickness or lethality. My results point to these DNA polymerase mutations as driving extensive alterations of the genetic information, and are consistent with them being drivers of colorectal and endometrial cancer. Future work will be required to determine the exact mechanisms by which these mutations impair the fidelity of DNA replication.
Contents

1 Genomic integrity and instability
1.1 Genome stability and maintenance ................................................................. 1
1.1.1 Genome replication ....................................................................................... 2
1.1.1.1 Structure of DNA, semiconservative replication and prokaryotic replication .............. 2
1.1.1.2 Replication initiation and prevention of re-replication in eukaryotes ......................... 5
1.1.1.3 DNA replication in eukaryotes .................................................................. 14
1.1.1.4 DNA polymerases ............................................................................... 21
1.1.2 DNA repair and Translesion Synthesis ......................................................... 31
1.1.2.1 Direct Damage Reversal ....................................................................... 31
1.1.2.2 Damage to one strand of the DNA ....................................................... 32
1.1.2.3 Double stranded breaks (DSBs) in the DNA ......................................... 37
1.1.2.4 Translesion synthesis (TLS) ................................................................ 39
1.1.2.5 Pausing the cell cycle: checkpoints ..................................................... 40
1.1.3 Dividing up the genome: chromosome segregation ...................................... 41
1.2 Genome variation ............................................................................................ 42
1.2.1 Large-scale genomic variation ..................................................................... 42
1.2.1.1 Whole-genome, segmental and gene duplications .................................... 44
1.2.1.2 Aneuploidy ......................................................................................... 46
1.2.1.3 Chromosomal translocation and chromoanagenesis ............................... 50
1.2.1.4 Mobile elements ........................................... 54
1.2.1.5 Exon/domain shuffling .................................. 54
1.2.1.6 Acquisition of foreign DNA ............................ 58
1.2.2 Small-scale mutations ........................................ 58
1.2.2.1 Point mutation instability (PIN) ....................... 58
1.2.2.2 Small insertions/deletions (INDELs) ................... 60
1.3 Causes of mutations ........................................... 61
1.3.1 Endogenous causes of mutation ............................ 61
1.3.2 Exogenous causes of mutations ............................ 65
1.4 Mutational processes and human disease ...................... 68
1.4.1 DNA repair deficiencies .................................... 68
1.4.2 Cancer ....................................................... 69
1.4.3 Mutational signatures ....................................... 70
1.4.4 DNA polymerase defects in cancer ......................... 72
1.5 DNA Sequencing ................................................ 72
1.6 The budding yeast *Saccharomyces cerevisiae* as a model organism .............................. 80

2 Analysis of cancer-associated polymerase mutations  85
2.1 Introduction .................................................... 87
2.2 Identification of polymerase mutations .......................... 89
2.2.1 Literature search for DNA polymerase mutations in cancer ........................................ 89
2.2.2 Query of COSMIC database, discarding single nucleotide polymorphisms and unconserved residues ....................... 89
2.3 Generation and propagation of polymerase mutants in *S. cerevisiae* .................. 96
2.3.1 Constructing single mutant polymerase strains .................. 96
2.3.2 Mutation accumulation experiment: Propagation of single mutant polymerase strains ........................................ 97
2.3.2.1 Single-colony bottleneck propagation of mutant polymerase strains .................. 100
2.3.2.2 Population bottleneck propagation of mutant polymerase strains .......................... 100
2.4 Establishing sequence analysis practices .......................... 102
2.4.1 Automating genomic DNA extraction and whole-genome sequencing of *Saccharomyces cerevisiae* strains ........................................ 102
2.4.2 Establishing sequencing analysis protocols for the identification of SNVs and INDELs ........................................ 103
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4.3</td>
<td>Testing analysis protocol on <em>Saccharomyces cerevisiae</em> genetic screens</td>
<td>108</td>
</tr>
<tr>
<td>2.4.4</td>
<td>Applying analysis protocols to mouse genetic screens</td>
<td>113</td>
</tr>
<tr>
<td>2.4.5</td>
<td>Establishing a sequencing analysis protocols for large genomic changes</td>
<td>120</td>
</tr>
<tr>
<td>2.4.6</td>
<td>Analysing repetitive DNA regions in the yeast genome</td>
<td>122</td>
</tr>
<tr>
<td>2.5</td>
<td>Summary</td>
<td>131</td>
</tr>
<tr>
<td>3</td>
<td>Analysis of populations of <em>S. cerevisiae</em> strains carrying simple polymerase mutations</td>
<td>135</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>135</td>
</tr>
<tr>
<td>3.2</td>
<td>Increased mutation rates for strains heterozygous diploid: <em>pol2-P301R, pol2-S312F, pol2-L439V, pol2-M459K</em> and <em>pol3-S483N</em></td>
<td>136</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Increased number of single-nucleotide variants for a subset of polymerase variants</td>
<td>136</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Single-nucleotide variants in haploid polymerase mutant strains</td>
<td>138</td>
</tr>
<tr>
<td>3.2.3</td>
<td><em>pol2</em> mutants grow at a similar rate to wild type strains</td>
<td>141</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Correlation of mutation rate estimates with mutations accrual</td>
<td>141</td>
</tr>
<tr>
<td>3.3</td>
<td>Patterns of single-nucleotide variants</td>
<td>144</td>
</tr>
<tr>
<td>3.4</td>
<td>Geographical mutation patterns</td>
<td>151</td>
</tr>
<tr>
<td>3.5</td>
<td>Large-scale variation: aneuploidy, CNVs and rDNA copy number</td>
<td>157</td>
</tr>
<tr>
<td>3.6</td>
<td>No increase in INDELs (compared to MMR mutants)</td>
<td>158</td>
</tr>
<tr>
<td>3.7</td>
<td>Summary</td>
<td>163</td>
</tr>
<tr>
<td>4</td>
<td>Polymerase mutations in mammalian systems and in combination with other mutations</td>
<td>167</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>168</td>
</tr>
<tr>
<td>4.2</td>
<td>Synthetic lethality with mismatch repair deficiency</td>
<td>168</td>
</tr>
<tr>
<td>4.3</td>
<td>Epistatic relationship of mutations with exonuclease deficiency</td>
<td>170</td>
</tr>
<tr>
<td>4.4</td>
<td>Observed mutagenesis in <em>pol2-P301R</em> strains is not due to increased participation of Pol<em>ζ</em> in DNA replication</td>
<td>171</td>
</tr>
<tr>
<td>4.5</td>
<td>Examining polymerase mutations in other organisms</td>
<td>176</td>
</tr>
<tr>
<td>4.5.1</td>
<td>The Pole and Pold1 mutations in mouse models</td>
<td>176</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Human POLE P286R mutant cell lines</td>
<td>178</td>
</tr>
<tr>
<td>4.6</td>
<td>Summary</td>
<td>178</td>
</tr>
<tr>
<td>5</td>
<td>Discussion and future directions</td>
<td>181</td>
</tr>
<tr>
<td>5.1</td>
<td>Whole-genome sequencing as a flexible tool to address problems in cell biology</td>
<td>181</td>
</tr>
</tbody>
</table>
5.2 Polymerase mutations as drivers of mutagenesis ........................................... 182
5.3 Future directions .................................................................................................. 187

6 Materials and Methods ......................................................................................... 189

6.1 Growth Medium .................................................................................................. 189
   6.1.1 *Escherichia coli* Growth Media ................................................................. 189
   6.1.2 *Saccharomyces cerevisiae* Growth Media .................................................. 190
6.2 Other solutions .................................................................................................... 193
6.3 Microbial Strains .................................................................................................. 195
   6.3.1 *Escherichia coli* strains ................................................................................. 195
   6.3.2 *Saccharomyces cerevisiae* strains ................................................................. 196
6.4 Oligonucleotides .................................................................................................. 198
6.5 Solutions .............................................................................................................. 201
6.6 Protocols .............................................................................................................. 202
6.7 Automated serial propagation platform .............................................................. 207
6.8 Illumina sequencing .............................................................................................. 207
6.9 Sequencing analysis ............................................................................................. 207
   6.9.1 Quality control of DNA sequencing ............................................................... 207
   6.9.2 Alignment of sequencing reads to the reference genome ......................... 208
   6.9.3 Variant Calling of SNPs and INDELs, Annotation and Filtering .................. 208
   6.9.4 Extracting mutational signatures ................................................................. 208
   6.9.5 Scripts written for this work ......................................................................... 208
   6.9.6 Step-by-step workflow of variant analysis ................................................... 211

References ................................................................................................................. 215

A List of Abbreviations .............................................................................................. 309

B Supplementary Tables, Electronic Files and Articles Published .......................... 313
   B.1 Supplementary figures, tables and notes ........................................................... 313
      B.1.1 Software tools and parameters used ............................................................. 313
         B.1.1.1 Software tools and parameters used for simulated genomes and capillary sequencing analysis .......................................................... 313
         B.1.1.2 Software tools and parameters used for sequencing analysis of *S. cerevisiae* .......................................................... 314
      B.1.2 Strains used in mutation accumulation (MA) experiments ........................ 315
B.1.2.1 Manual propagation of strains heterozygous diploid for candidate polymerase mutations ........................................ 315
B.1.2.2 Automated propagation of strains haploid and heterozygous diploid for candidate polymerase mutations .................. 315
B.1.3 6-Thioguanine suppressor screen of haploid mouse cells ................................................. 316
B.1.4 Custom filters for DNA sequencing Filters .................................................................. 316
B.2 Electronic files of supplementary information ................................................................. 317
B.2.1 Supplementary files for the mouse synthetic lethality screens ................................. 317
   B.2.1.1 6TG_mouse_Sup1.xlsx ................................................. 317
   B.2.1.2 6TG_mouse_Sup2.xlsx ................................................. 317
   B.2.1.3 6TG_mouse_Sup3.xlsx ................................................. 317
   B.2.1.4 6TG_mouse_Sup4.xlsx ................................................. 318
B.2.2 Supplementary files for the mouse synthetic lethality screens ................................. 318
   B.2.2.1 MA_SampleNames.pdf .................................................. 318
   B.2.2.2 S1-3.experiment_merge.vcf ............................................. 318
   B.2.2.3 S4.experiment_merge.vcf ................................................. 318
   B.2.2.4 S5.experiment_merge.vcf ................................................. 318
B.3 Articles published during my PhD ................................................................................. 318
# List of Figures

1.1 Structure of DNA ................................ 3
1.2 Replication initiation in *E. coli* ........................................ 4
1.3 Lagging strand DNA synthesis in *E. coli* .................................. 5
1.4 Overlapping replication cycles in *E. coli* .................................... 6
1.5 The eukaryotic cell cycle .......................................................... 7
1.6 Licensing of eukaryotic origins of replication ............................... 10
1.7 Degradation of Cdt1 during the cell cycle and in response to DNA damage 13
1.8 Regulation of Cdt1 by association with Geminin ............................ 14
1.9 Structure of DNA polymerase δ and DNA polymerase ε .................. 18
1.10 Structure and representation of replicative DNA polymerases .......... 22
1.11 Comparison of primer-template DNA bound to four DNA polymerases. . 23
1.12 Mechanism of DNA polymerization .......................................... 25
1.13 Replication fidelity ................................................................. 28
1.14 Fidelity of different DNA polymerases ....................................... 29
1.15 Base excision repair (BER) of oxidized DNA base lesions ............ 33
1.16 Nucleotide excision repair (NER) .............................................. 35
1.17 A general outline of the DNA damage signal transduction pathway .... 41
1.18 The mitotic spindle ................................................................. 43
1.19 Gene duplications: the duplication-degeneration (DDC) model .......... 46
1.20 Uniparental Disomy - A special case of aneuploidy ...................... 49
1.21 Consequences of chromosomal translocations .............................. 51
1.22 Chromothripsis ...................................................................... 53
1.23 Chromoplexy and Chromothripsis ............................................. 54
1.24 Classes of DNA transposons ..................................................... 55
1.25 Blood clotting cascade ............................................................ 57
1.26 Transitions and Transversions .................................................. 60
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.27</td>
<td>Codon table</td>
<td>61</td>
</tr>
<tr>
<td>1.28</td>
<td>Replication slippage</td>
<td>62</td>
</tr>
<tr>
<td>1.29</td>
<td>Unequal crossovers result in chromosome rearrangements</td>
<td>64</td>
</tr>
<tr>
<td>1.30</td>
<td>Mutational signatures leave their marks on the genome</td>
<td>71</td>
</tr>
<tr>
<td>1.31</td>
<td>Summary of known mutational signatures</td>
<td>73</td>
</tr>
<tr>
<td>1.32</td>
<td>Early sequencing techniques: Gilbert and Sanger</td>
<td>74</td>
</tr>
<tr>
<td>1.33</td>
<td>Sequencing by synthesis</td>
<td>76</td>
</tr>
<tr>
<td>1.34</td>
<td>Solid-phase bridge amplification and sequencing by synthesis (Illumina)</td>
<td>77</td>
</tr>
<tr>
<td>1.35</td>
<td>Third-generation Sequencing Techniques</td>
<td>79</td>
</tr>
<tr>
<td>1.36</td>
<td>Life cycle of the budding yeast <em>Saccharomyces cerevisiae</em></td>
<td>81</td>
</tr>
<tr>
<td>1.37</td>
<td>The budding yeast mating type locus</td>
<td>83</td>
</tr>
<tr>
<td>2.1</td>
<td>Methodology of the work carried out during my PhD</td>
<td>86</td>
</tr>
<tr>
<td>2.2</td>
<td>Locations of DNA polymerase mutations within the proteins</td>
<td>91</td>
</tr>
<tr>
<td>2.3</td>
<td>Prevalence of polymerase mutations of interest in COSMIC</td>
<td>92</td>
</tr>
<tr>
<td>2.4</td>
<td>Alignment of polymerase residues of interest to the yeast proteins</td>
<td>95</td>
</tr>
<tr>
<td>2.5</td>
<td>Rationale for plasmid construction</td>
<td>98</td>
</tr>
<tr>
<td>2.6</td>
<td>Exonuclease domains conserved in B family polymerases</td>
<td>99</td>
</tr>
<tr>
<td>2.7</td>
<td>Mutation accumulation experiment: manual propagation of mutated <em>S. cerevisiae</em> strains</td>
<td>101</td>
</tr>
<tr>
<td>2.8</td>
<td>DNA extracted using a high-throughput protocol produces high quality sequencing data</td>
<td>104</td>
</tr>
<tr>
<td>2.9</td>
<td>The number of variants in W303 strains compared to the S288c reference genome</td>
<td>105</td>
</tr>
<tr>
<td>2.10</td>
<td>Experimental strategy to identify acquired mutations</td>
<td>107</td>
</tr>
<tr>
<td>2.11</td>
<td>Sequencing analysis identifies mutations capable of suppressing <em>sae2Δ</em> DNA damage hypersensitivity</td>
<td>109</td>
</tr>
<tr>
<td>2.12</td>
<td>Mutations in <em>SIR3</em> and <em>SIR4</em> identified as the cause for the hypersensitivity of <em>tof1Δ</em> cells to camptothecin</td>
<td>111</td>
</tr>
<tr>
<td>2.13</td>
<td>Generation of mutagenized libraries</td>
<td>112</td>
</tr>
<tr>
<td>2.14</td>
<td>Identification of suppressor mutations</td>
<td>114</td>
</tr>
<tr>
<td>2.15</td>
<td>Using multiple controls and multiple variant callers to enrich for high confidence variants</td>
<td>116</td>
</tr>
<tr>
<td>2.16</td>
<td>Clinically-relevant and newly-identified suppressor mutations</td>
<td>117</td>
</tr>
<tr>
<td>2.17</td>
<td>EMS mutagenic action</td>
<td>119</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>2.18</td>
<td>Relationship between read pairs and structural variants</td>
<td>121</td>
</tr>
<tr>
<td>2.19</td>
<td>An overview of the SVMerge pipeline</td>
<td>123</td>
</tr>
<tr>
<td>2.20</td>
<td>Visualising aneuploidy in budding yeast</td>
<td>124</td>
</tr>
<tr>
<td>2.21</td>
<td>Ambiguities in read mapping</td>
<td>125</td>
</tr>
<tr>
<td>2.22</td>
<td>Next-generation sequencing data can be used to estimate rDNA copy number</td>
<td>128</td>
</tr>
<tr>
<td>2.23</td>
<td>Next-generation sequencing data could also be used to assess Ty element copy number</td>
<td>130</td>
</tr>
<tr>
<td>3.1</td>
<td>Number of single-nucleotide variants per sample in pol2 mutant strains</td>
<td>137</td>
</tr>
<tr>
<td>3.2</td>
<td>Number of single-nucleotide variants per sample in pol3 mutant strains</td>
<td>139</td>
</tr>
<tr>
<td>3.3</td>
<td>Number of single-nucleotide variants per line per haploid genome for selected haploid and heterozygous diploid pol2 mutant strains</td>
<td>140</td>
</tr>
<tr>
<td>3.4</td>
<td>Growth of <em>S. cerevisiae</em> mutant strains in rich medium</td>
<td>142</td>
</tr>
<tr>
<td>3.5</td>
<td>Correlation of mutation rate estimates and mutation accrual for pol2 mutant strains</td>
<td>143</td>
</tr>
<tr>
<td>3.6</td>
<td>Single-nucleotide variant patterns</td>
<td>145</td>
</tr>
<tr>
<td>3.7</td>
<td>Single-nucleotide variant patterns adjusted to frequencies of trinucleotides in <em>S. cerevisiae</em></td>
<td>146</td>
</tr>
<tr>
<td>3.8</td>
<td>SomaticSignatures: Determining the numbers of signatures</td>
<td>147</td>
</tr>
<tr>
<td>3.9</td>
<td>2-3 signatures are determined using Non-negative matrix factorization</td>
<td>149</td>
</tr>
<tr>
<td>3.10</td>
<td>Contribution of the signatures to the variant pattern</td>
<td>150</td>
</tr>
<tr>
<td>3.11</td>
<td>EMu: Validating Signature Analysis</td>
<td>152</td>
</tr>
<tr>
<td>3.12</td>
<td>No observed clustering of mutations acquired by pol2-P301R strains</td>
<td>153</td>
</tr>
<tr>
<td>3.13</td>
<td>Mutations falling inside and outside of genes in heterozygous diploid polymerase mutant strains</td>
<td>154</td>
</tr>
<tr>
<td>3.14</td>
<td>Percentage of mutations within genes in haploid strains</td>
<td>156</td>
</tr>
<tr>
<td>3.15</td>
<td>Number of total aneuploidy and segmental insertions/duplications identified</td>
<td>157</td>
</tr>
<tr>
<td>3.16</td>
<td>Example of aneuploidy in <em>S. cerevisiae</em></td>
<td>159</td>
</tr>
<tr>
<td>3.17</td>
<td>Example of segmental deletions and amplifications in <em>S. cerevisiae</em></td>
<td>160</td>
</tr>
<tr>
<td>3.18</td>
<td>rDNA copy number changes in polymerase mutants</td>
<td>161</td>
</tr>
<tr>
<td>3.19</td>
<td>No increase in the number of INDELs detected per sample across strains</td>
<td>162</td>
</tr>
<tr>
<td>3.20</td>
<td>Mutation accrual in strains with mismatch repair deficiencies</td>
<td>163</td>
</tr>
<tr>
<td>4.1</td>
<td>Tetrad dissection to generate double mutants and detect synthetic lethality</td>
<td>169</td>
</tr>
</tbody>
</table>
4.2 The mutagenesis observed in strong mutator strains is partially rescued by mutating critical residues in the exonuclease domain active site .......................... 172
4.3 Synergistic effects on mutation number between rev3Δ and pol2-P301R ........ 174
4.4 Mutational patterns observed in pol2-P301R cells and pol2-P301R rev3Δ cells are highly similar ........................................... 175
4.5 Constructs used for conditional knock-in mutations in mice ........................ 177
List of Tables

1.1 Eukaryotic replicative DNA polymerases .............................................. 15
1.2 Families of DNA polymerases ............................................................... 22
1.3 Error rates of DNA polymerases from different families .................... 26
1.4 Incidence of aneuploidy during development ...................................... 47
1.5 The origin of human trisomy ................................................................. 50
1.6 Standard Nomenclature for *S. cerevisiae* genetics using *POL2* as an example. 83
1.7 A selection of Nobel Prizes awarded for work using *S. cerevisiae* as a model organism. ................................................................. 84

2.1 Polymerase exonuclease domain mutations in *S. cerevisiae* .................. 88
2.2 Genomic locations of mutations in DNA polymerases in different human genome assemblies ................................................................. 90
2.3 Checking DNA polymerase mutations for common variants ................ 93
2.4 Polymerase mutations identified from the literature with predicted consequences 94
2.5 Budding yeast equivalents of human DNA polymerase mutations of interest . 96
2.6 Haploid copy number of rDNA repeats across Eukaryotic species ......... 126

3.1 Mutation number fold change of *pol2* haploid and heterozygous diploid mutant strains when compared to the *POL2* strain ........................ 141
3.2 Estimates of mutation rate increases using resistance to Thialysine .......... 142

4.1 Synthetic lethality of polymerase mutants and mismatch repair deficiency ... 171
Chapter 1

Genomic integrity and instability

1.1 Genome stability and maintenance

In 2013, the 60th anniversary of the proposed molecular structure of DNA[1–3] was celebrated all over the world by museums, television, and radio programs. While the discovery of the DNA double helix has probably been one of the most important milestones in the history of biology, the studies leading up to our understanding of DNA and genetic inheritance started a century earlier with the work of Charles Darwin, Gregor Mendel and Friedrich Miescher. In 1859, Darwin suggested that living organisms were the result of evolution via the combined effects of variation, heritability of traits and natural selection [4, 5]. Seven years later, Mendel defined the laws of genetic inheritance by studying how some visible characteristics of pea plants were passed on to the following generations[6]. Although his findings went largely unnoticed at the time, Mendel understood that heritable traits were transferred from parents to offspring in what he termed "bildungsfähige[] Elemente" (loosely translated to "elements capable of formation"). He made no concrete statements about the physical nature of these elements, but suggested that they were likely contained within cells. Almost at the same time, Friedrich Miescher first purified DNA from leucocytes naming the substance nuclein[7]. It would however take decades until these discoveries were united in the study of genetics. Indeed, it was only in 1944 that Avery, MacLeod, and McCarty demonstrated that DNA is the carrier of genetic information[8–10].

Every time a cell divides, it needs to duplicate its genome so that both resulting cells have a full complement of genetic information. This duplication needs to be highly accurate to ensure that no crucial information is lost or altered in a detrimental way. However, some degree of inaccuracy is tolerated and essential to generate the variation that evolutionary selection acts on. Correct cell division requires the duplication of the genome and the correct segregation
of the two copies between the two daughter cells without any loss or alteration of the genetic information. This is no simple feat. Even though DNA is a structure with a radius that only measures 10 Å — a millionth of a millimeter — its length can reach several centimeters[1]. It is estimated that the ~5 million base pairs of DNA from a bacterium residing in the human gut flora would be 1.6mm long when stretched out and that the entire human genome of a male, diploid cell laid end-to-end would be approximately 2m long[11]. The size of the human genome was estimated to be around 6000 megabases by physical and genetic measurements, which was confirmed and further refined by the Human Genome Project[12–15]. Maintaining and copying roughly 6 billion bases of DNA sequence represents a tremendous molecular challenge and the human genome is not the largest known by far. Paris japonica is a perennial plant from Japan with a haploid genome fifty times larger than human[16], but the current record-holder is a freshwater amoeboid called Polychaos dubius, whose genome size was estimated (albeit not reconfirmed with the most current methods available) to measure 670 000 megabases[17](reviewed in [18]).

Before considering the alterations genomes can experience, the consequences of such alterations and the processes that give rise to them, our current understanding of the mechanisms that maintain genome integrity will be summarized. This involves mechanisms that ensure the faithful duplication and segregation of the genome during mitosis and meiosis, as well as mechanisms that repair the ubiquitous damage to DNA.

1.1.1 Genome replication

1.1.1.1 Structure of DNA, semiconservative replication and prokaryotic replication

Though the structure of the DNA macromolecule was unknown, early experiments demonstrated that the occurrence of the four DNA nucleobases cytosine, guanine, adenine and thymine was not even and that the ratio of purines (adenine and guanine) to pyrimidines (cytosine and thymine) is very close to 1[19]. It was evidence like this, and extensive X-ray measurements by Franklin and Wilkins that led to the proposal of the double helical structure (Fig. 1.1)[1–3]. DNA strands are formed by two backbones of deoxypentose rings linked by phosphate residues that wind around each other and have an intrinsic directionality with strands running anti-parallel to each other. From these backbones, nucleobases project towards one another, perpendicular to the axis of the double helix and form pairs: adenine is paired with thymine via hydrogen bonds and cytosine is similarly interacting with guanine accounting for the ratio of purines to pyrimidines. Immediately, Watson and Crick provided a largely accurate hypothesis of how DNA could be replicated based on their structure[20]:
1.1 Genome stability and maintenance

Figure 1.1: Structure of DNA
A - Structure of the DNA double helix. Reproduced from [23] in accordance with the publisher’s terms of use. B - Pairing of the 4 DNA bases: adenine forms hydrogen bonds with thymine, guanine bonds with cytosine. From [24].

they suggested "that these two chains separate and that a new chain is formed complementary to each of them, the result will be two pairs of strands, each pair identical to the original parent duplex and identical to each other"[21]. This hypothesis was further strengthened by the elegant Meselson-Stahl experiment which showed that DNA replication proceeds in a semi-conservative manner meaning that after replication the two products are each formed of one of the template strands and one of the newly synthesized strands[22].

Many aspects of DNA replication were first studied in prokaryotes such as *Escherichia coli*[25–29] and other non-eukaryotic systems, and are best described in these systems. Replication in archaebacterial, bacteriophage, and viral systems has been studied, but will not be included here[28, 30–32]. Replication of the circular *E. coli* genome starts at a short, specific sequence known as the origin or replication (oriC in *E. coli*) and proceeds in both directions from there[11, 33]. This sequence is recognized by the initiator protein DnaA which starts unwinding the DNA to start a replication fork[34–36]. The replicative helicase (DnaB) keeps unwinding the parental DNA strands at the rate of synthesis[25, 37]. (Fig. 1.2-A) The separated DNA strands are then copied by proteins known as DNA polymerases, which work by pairing the appropriate incoming deoxynucleoside 5’-triphosphate (dNTP) to the template base and then catalyzing its addition to the 3’hydroxyl group (3’OH) of the nascent strand[21, 38, 39] As a consequence, DNA polymerases cannot start from single-stranded
Genomic integrity and instability

Figure 1.2: Replication initiation in *E. coli*
A - Unwinding of the DNA origin by oriC; B - Loading of the DNA helicases and DNA primases; C - Assembly of the replisome — Reproduced from [37] with permission from the publisher.

DNA but require a short RNA or DNA primer to extend. For that purpose, a specialized RNA polymerase called primase (DnaG in *E. coli*) synthesizes short RNA primers for the DNA polymerases to extend[25, 37, 40–42]. (Fig. 1.2-B)

After primer synthesis, the DNA polymerase III holoenzyme is assembled (Fig. 1.2-C). This protein complex is made up of three components: an enzymatic subunit synthesizing DNA, a sliding clamp (β2 clamp) and a clamp loader. The ring-shaped β2 clamp encircles DNA and slides along it, increasing the speed (∼750 nucleotides/s) and processivity (>50 kb) of the tethered DNA polymerase[25, 43]. By opening the β2 clamp, the clamp loader allows the passage of one DNA strand into the ring for the purpose of loading (or unloading) the holoenzyme on the DNA molecule to be replicated[44]. The need to replicate an entire genome with only a pair of DNA polymerases, might be the main reason why *E. coli* cells have evolved a sliding clamp[25]. In fact, without the β2 clamp the Pol III enzymatic subunit is slow (∼20nts/sec) and not nearly as processive[45]. (Fig. 1.2-C)

Because of the 5’-to-3’ directionality of DNA synthesis and the antiparallel nature of the two DNA strands, only one strand (the leading strand) is synthesized in a continuous fashion[25]. The other strand (the lagging strand) is synthesized discontinuously in the direction opposite to the movement of the DNA helicase[25, 46, 47]. The lagging strand will thus be generated as a series of short stretches, called Okazaki fragments, with the polymerase cores rapidly dissociating at the end of a stretch[25]. Since the primase needs to be associated with the helicase to function and the polymerases are also complexed with the helicase, coordinating the leading and lagging strand likely involves DNA looping[46, 47]. Experiments using the bacteriophages T7 and T4 have shown that leading and lagging strand synthesis can
1.1 Genome stability and maintenance

Figure 1.3: Lagging strand DNA synthesis in *E. coli*
Model of lagging strand DNA synthesis coordination: lagging strand polymerase action is thought to generate a “Trombone” loop to accommodate both polymerases and their opposing movement at the replication fork. Reproduced from [37] with permission from the publisher.

occur simultaneously if the lagging strand loops around[48–52] (Fig. 1.3).

To ensure that the genome is replicated only once every cell division, *E. coli* regulates the initiation of DNA replication by a process called origin sequestration and by regulating the activity of the initiator protein DnaA[34]. Origin sequestration takes advantage of the fact that the various GATC methylation sites in the oriC sequence will be hemimethylated in the time immediately after replication, which provides multiple high-affinity binding sites for the protein SeqA[54–56]. While the binding of SeqA to the origin sequesters it and causes it to remain inactive for about a third of the cell cycle, several mechanisms work to lower the activity of DnaA[57]. However, this is not a stable state and eventually oriC will be fully methylated by the Dam methyltransferase[54] making it available for the next round of DNA replication. Interestingly, when growth conditions are optimal *E. coli* can grow with overlapping replication cycles allowing for a population doubling time shorter than the time required to replicate the entire chromosome (Fig. 1.4).

1.1.1.2 Replication initiation and prevention of re-replication in eukaryotes

The cell cycle In contrast to prokaryotes, eukaryotes strictly separate the timing of replication (S-phase) and cell division/mitosis (M-phase) with two gap phases (G1- and G2-phase) and it is critical that the stages of the cell cycle occur in the right order and that one phase is completed before the next begins (Fig. 1.5). Building on other work, Nurse, Hartwell and Hunt found that the progression of the cell cycle is orchestrated by cyclin-dependent kinases (CDKs), protein kinases which activate critical processes by phosphorylating a variety of key proteins[58]. G1-CDKs phosphorylate targets to promote S-phase entry, S-CDKs are involved
in initiation of replication and M-CDKs regulate mitosis to ensure that accurate segregation of chromosomes can occur[59]. Though their levels are constant throughout the cell cycle, CDK activity is tightly controlled through post-translational modifications and by its association with proteins called cyclins whose levels oscillate through the cell cycle: they accumulate gradually and are degraded at key stages of the cell cycle, which vastly decreases the CDK activity they are associated with. For instance, mitotic cyclins critical for the onset of cell division are degraded at the end of mitosis due to activity of the E3-ubiquitin ligase Anaphase-Promoting Complex/Cyclosome (APC/C), a multi-subunit complex that polyubiquitylates different proteins marking them for degradation by the 26S proteasome[60]. Simply put, it is the alternating waves of CDK and APC/C activity that ensure that in eukaryotes each chromosome is normally replicated once and only once. Even though several CDKs and cyclins as well as other E3-ubiquitin ligases are known to participate in the cell cycle, in fission yeast a single cyclin–CDK pair has been shown to be able to drive a near-normal cell cycle[REF]. Even mouse embryos missing a subset or combination of cyclins or CDKs are surprisingly healthy (from only minor defects in cyclin D1-deficient mice to lethality in mid-gestation in mice lacking all D-cyclins), demonstrating the robustness of the cell cycle[61–73](reviewed in [74]).
1.1 Genome stability and maintenance

Figure 1.5: The eukaryotic cell cycle
A schematic of the eukaryotic cell cycle. S-phase (replication) and M-phase (mitosis followed by cytokinesis) alternate and are separated by two growth phases (G1 and G2).

Consequences of eukaryotic genome structure  In contrast to *E. coli*, eukaryotes typically have several linear chromosomes of much larger size, necessitating more than one origin per chromosome. For example, while S-phase tends to take about 8 hours in human cells, if the largest human chromosome (Chromosome 1, 250 Mb) was replicated from only one origin, this would take more than 50 days[11]. Having more than one site to initiate DNA replication requires that initiation is simultaneously triggered at multiple origins at the right point of the cell cycle in the parental chromosomes, and that origins on the newly synthesised daughter are blocked from initiation until the next cell cycle. As a result, initiation of replication has to be coordinated with the rest of the cell cycle to ensure that replication and division alternate appropriately so that each entire chromosome is only replicated once per division. Additionally, cells contain checkpoints, which can interfere with the normal progression of the cell cycle in response to potentially devastating events such as DNA damage. As well as activating checkpoints, failure to regulate replication initiation can lead to re-replication which in turn causes gene amplification, polyploidy and other kinds of genome instability (see Chapter 1.2.1)[75–77]. Genetic and biochemical studies in model systems such as the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, egg extracts from the frog *Xenopus laevis*, the fly *Drosophila melanogaster* as well as mammalian cell lines have shown that initiation is regulated by a common set of conserved proteins and that are cell cycle regulated[78, 79].
Wrapping up the eukaryotic genome  The dimensions of DNA (up to 2m in length in a human cell) when compared to the dimensions of the nucleus where it is contained (approximately 6μm in diameter), poses a storage problem that the cell solves by the association of DNA with positively-charged histone proteins, to form an ordered structure called a nucleosome. Nucleosomes then associate with one another to form dynamic higher order structures that can change compaction from a relatively open structure in interphase to the highly compact metaphase chromosomes. Nucleosomes consist of 146bp of DNA wrapped around an octameric histone core, made up of two copies of each histone H2A, H2B, H3 and H4, and together with other proteins, such as histone H1 variants, allow assembly of nucleosomes into more complex structures[80]. The tails of the histone proteins extend out from the nucleosome and are subject to many post-translational modifications - methylation, acetylation, phosphorylation, ubiquitylation. These modifications allow, for instance, epigenetic control of gene expression regulating the compaction and subsequent availability of DNA to, for example, the transcription machinery[81]. Chromatin poses further challenges for the replication machinery as with each cycle of replication, chromatin and its associated epigenetic marks need to be replicated(see [82–84] for further reading).

Eukaryotic replication origins  Eukaryotic replication initiation is well studied in the budding yeast *S. cerevisiae*, whose origins - originally called autonomously replicating sequences (ARS) - are the best-characterised chromosomal origins[85]. The modular origins consist of an A element, the most important sequence block within the 100-200bp, and a variable number of B elements. The A block contains the short AT-rich ARS consensus sequence (ACS) which is found in all budding yeast origins[86, 87] and crucial for origin function as it is the most important binding site for the origin recognition complex (ORC; discussed in more detail below), a component of the initiation machinery[88]. The less conserved B elements, not easily identified by sequence conservation[89–91], show some functional conservation across many origins: they also contribute to ORC binding and provide a binding site for Abf1, which is known to stimulate initiation[89, 92–94]. While *E. coli* and budding yeast origins are short, well defined sequences, other eukaryotic origins are typically not specified by their sequence. Many are rich in adenine and thymine, presumably because opening up AT-rich sequences requires less energy to break hydrogen bonds. Additionally, the chromatin organization of the DNA is thought to specify origins in eukaryotes [11, 95–97]. For example, even though fission yeast (*S. pombe*) origins tend to have one or more functionally important segment of about 20-50bp, apart from their being AT rich no consensus sequence has been identified and they are not interchangeable with the smaller budding yeast origins[98–100]. This seems to hold
true for metazoan origins\cite{100, 101}: while chromosomal locations of replication origins have been pinpointed, essential sequences have not been identified. In fact, in many cases, the exact point of replication onset can occur within so-called initiation zones\cite{101–104} and *Xenopus* DNA seems to have little or no sequence requirement for replication initiation *in vitro*\cite{105}. Considering the disparity between origins across species, it is intriguing that the ORC and other proteins involved in initiation such as cell division cycle 6 (Cdc6) and Cdc10-dependent transcript factor 1 (Cdt1) are conserved among eukaryotes\cite{88, 106–112}.

**Replication Licensing** Preventing re-replication is a two step process in eukaryotes\cite{113–115}: Step 1 is the formation of so-called pre-replicative complexes (pre-RC) on DNA origins at the end of mitosis and the beginning of the next cell cycle in early G1 phase ("origin licensing"); Step 2 is the initiation of DNA replication during S-phase during which the pre-RC is disassembled. Step 2 cannot proceed without Step 1 having occurred. The two steps are temporally isolated from each other in different stages of the cell cycle, with the result that inactive pre-RCs are assembled during periods of low CDK and high APC/C activity, but are only functional and able to commence DNA replication when those activities are reversed, thus preventing re-usage of an origin in the same cell cycle\cite{59}.

The pre-RC complex is formed by sequential recruitment of the licensing proteins Cdt1 and Cdc6 onto origin-bound ORC followed by the assembly of the MCM2-7 complex on DNA, which is the replicative helicase in eukaryotes(Fig. 1.6). ORC was first identified as a protein binding to ARS in yeast \cite{116}. In metazoans, ORC preferentially binds to AT-rich sequences. In *S. pombe* the ORC4 subunit contains nine AT-hook motifs absent in the human protein\cite{117}. In vertebrates, ORC is potentially targeted to origins by HMGA1a which contains the AT-hook motif\cite{84}. In humans, CDC6 is recruited to ORC by MCM8 which interacts with ORC2 and CDC6\cite{118}. Another MCM family member, MCM9, binds Cdt1 directly and promotes recruitment of the MCM2-7 replicative helicase complex\cite{119–130}. Loading of the MCM2-7 is stimulated by ORC and CDC6 ATPases activities\cite{131, 132}. The MCM complex is loaded directly onto DNA and forms a double hexamer\cite{133–137}. Until recently, MCM2-7 helicase activity had not been detected *in vivo*, but immuno-depletion of the putative helicase from *Xenopus* egg extracts inhibited DNA unwinding, further suggesting its involvement in separating the DNA strands during replication\cite{138}.

Thus the end result of pre-RC formation is the loading of inactive helicase into replication origin DNA sequences at the beginning of the cell cycle, which once they are activated - not before the onset of S-phase - allows unwinding of the DNA and access of the polymerases to DNA for the start of replication \cite{114, 139, 140}. Activation of the helicase requires the
Figure 1.6: Licensing of eukaryotic origins of replication
A schematic detailing the ordered assembly of licensing factors on eukaryotic origins starting with the Origin Recognition Complex (ORC). Until the end of mitosis licensing is prevented by actions of cyclin-CDKs. After anaphase the licensing factors such as Cdc6 and Cdt1 assemble on DNA and the MCM2-7 complex is loaded. After activation, licensing factors dissociate, the MCM2-7 begins unwinding the DNA and replication commences.
assembly of the Cdc45–GINS–MCM2–7 (CMG) complex[114, 141, 142]. Cdc45 and GINS are loaded onto the MCM2-7 complex due to the activity of two protein kinases, CDK and DDK (DBF4 and DRF1-dependent kinase), in concert with other factors such as MCM10 and Dpb11[114, 143–155]. Important replication factors like RPA, DNA polymerase α, RFC, PCNA, and DNA polymerase δ are recruited subsequently to start DNA replication[156]. It is known that MCM2, MCM4, MCM6 are essential targets of DDK in vivo, but the exact mechanism of how DDK promotes origin activation remains unclear [157–159]. It could involve the recruitment of factors such as Sld3, Sld7 and Cdc45 possibly via the DDK-dependent phosphorylation on the MCM2-7 complex [160]. Vertebrate homologs of Dpb11 (TopBP1), Sld2 (RecQL4) and Sld3 (Treslin/Ticrr) have been identified and are involved in initiation [161–165]. Human CDT1 was also shown to be involved in activation of the MCM complex. It associates with the kinase CDC7 and recruits CDC45 [166, 167]. As MCM2-7 starts unwinding DNA, Ctd1 and Cdc6 are released from the origin [114, 168, 169]. It has been observed that many more origins are assembled than actually used during replication, with inactive origins possibly functioning as back-ups which can be fired later during S-phase to ensure completion of replication. Pre-RCs that are not activated are usually displaced by a moving replication fork, ensuring that replicated DNA is not licensed for replication [59].

Regulation of licensing in S. cerevisiae In yeast, the main inhibition of licensing in S phase is due to high CDK activity which negatively regulates licensing factors [169]. In S. cerevisiae, one example is Cdc6. Following its CDK-mediated phosphorylation, Cdc6 is marked for degradation by the E3 ligase SCF$^{Cdc4}$ [170–173]. Additionally, Cdc6 expression is blocked due to CDK regulation of the transcription factor Swi5 and CDK phosphorylation and subsequent binding of Cdc6 blocks its licensing activities [114, 174]. Upon completion of mitosis, CDK activity is lowered in two key ways: the mitotic cyclin Clb2 is degraded by the 26S proteasome, and the CDK inhibitor Sic1 inhibits G1-CDK activity [175–178]. Degradation of Clb2 releases Cdc6 to participate in licensing again [174]. The phosphatase Cdc14 is also involved in promoting preRC assembly. Among other things it removes CDK-dependent phosphorylation from Swi5, which can then activate expression of Cdc6 and Sic1 [179, 180]. Cdc14 dephosphorylates Sic1, which protects it from SCF$^{Cdc4}$-mediated degradation [179]. The CDK inhibitor Sic1 is also a key barrier to firing origins too early [178, 181, 182]. As the cell cycle progresses, phosphorylation of Sic1 by the CDK complexes Cln-Cdc28 and Clb-Cdc28 targets it for ubiquitin-mediated degradation [183–185].
**Negative regulation of licensing factors**  To prevent re-licensing of origins after initiation of replication, many of the licensing factors described above are subject to negative regulation often in multiple ways. This is likely to prevent reassembly of a preRC complex, subsequent reloading of the helicase and thus re-replication. Interfering with this negative regulation has been shown to be able to cause re-replication in many cases [114, 169, 186].

Apart from Cdc6 described above many other licensing factors are negatively regulated after initiation by CDK activity in *S. cerevisiae*. The CDK Clb-Cdc28 also phosphorylates Orc2 and Orc6, components of the ORC [187–189]. This inhibits interaction between Cdt1 and the complex, thus hampering recruitment and loading of the MCM complex [190, 191]. CDKs also promote the nuclear export of Cdt1 and MCM2-7 during S phase, G2 and early mitosis in budding yeast [114, 169, 192–195]. This prevents access of these factors to origin DNA. Finally, the activity of DDK is restricted to S phase due to the Dbf4 subunit being targeted for degradation by APC/C [196–198]. Similar mechanisms are known to regulate licensing factors in *S. pombe* [114, 169]. Proteolytic degradation regulates both Cdc6 and Cdt1 in S and G2 phase [199–203].

Unlike in yeast, it is not clear whether CDK regulation of licensing factors directly inhibits re-replication in metazoans. While mammalian CDT1, CDC6 and ORC have all been shown to be targets of CDK activity *in vitro*, it is not clear that these modifications prevent re-replication [204–208]. Furthermore, while in human and *Xenopus*, Cdc6 is CDK phosphorylated and the ectopically expressed protein is transported from the nucleus after phosphorylation, a significant portion of Cdc6 is bound to chromatin [205, 209–211]. Additionally, Cdc6 is also exported from the nucleus in a Cul4-mediated manner and subject to caspase-3-mediated cleavage [212, 213].

Cdt1 overexpression causes re-replication making Cdt1 regulation a key part of licensing regulation [214]. Targeting Cdt1 for degradation is conserved in higher eukaryotes including *Caenorhabditis elegans, Drosophila, Xenopus*, and mammals [114, 215]. There are multiple mechanisms to degrade Cdt1, highlighting the importance of this process [75, 114, 216] (Fig. 1.7). One CDK-dependent mechanism involves the SCF–Skp2 E3 ubiquitin ligase complex to mark Cdt1 for degradation [206, 217–221]. In human cells, Cdk2 and Cdk4 bind Cdt1 and phosphorylate it, thereby recruiting the E3 ligase to mark Cdt1 for degradation during S and G2 phase. While this pathway has been observed in human cells, it is not conserved in other metazoans suggesting it could be an evolutionarily recent addition [221]. Because impairment of this pathway still leads to Cdt1 degradation, another mechanism for Cdt1 degradation was identified involving the Cul4–Ddb1–Cdt2 complex and it has been demonstrated to be essential for Cdt1 degradation from *S. pombe* to metazoans [114, 203, 219, 221–228]. Degradation
Figure 1.7: Degradation of Cdt1 during the cell cycle and in response to DNA damage during S phase, Cdt1 is targeted for degradation by the SCF–Skp2 and by the PCNA–Cul4–Ddb1–Cdt2 pathways. Cyclin-CDK activity results in the phosphorylation of Cdt1, which in turn allows recruitment of SCF–Skp2, an E3 ubiquitin ligase complex. PCNA binds Cdt1 directly and recruits the Cul4–Ddb1–Cdt2 E3 ubiquitin ligase complex. DNA damage results in PCNA-mediated degradation of Cdt1 akin to its degradation in S-phase. Figure reproduced from [230] with permission from the publisher.

of Cdt1 is mediated by the sliding clamp proliferating cell nuclear antigen (PCNA), but only when it is bound to DNA. This couples Cdt1 degradation to active replication. When the PIP-motif (PCNA-interacting Protein motif) of Cdt1 that binds PCNA is mutated, Cdt1 levels are stabilised and re-replication occurs[203]. These pathways thus provide slightly distinct functions: SCF–Skp2 acts in both S and G2 phase, whereas Cul4–Ddb1–Cdt2 promotes Cdt1 degradation only in S phase[219]. APC/CCdh1 has also been demonstrated to promote proteolysis of Cdt1 in human cells[229].

Metazoans have also evolved CDK-independent pathways to prevent re-replication and they mostly involve Cdt1 regulation. The most striking of these involves the protein Geminin(Fig. 1.8). Discovered as an inhibitor of DNA replication in Xenopus, Geminin was identified as an inhibitor of Cdt1[231]. It binds to and thus sequesters Cdt1 on chromatin during S and G2 phase which prevents it from binding MCM2-7[232–236]. Loss of Geminin alone can be sufficient to induce re-replication[228, 231, 237–242]. Geminin is targeted for degradation by the APC/C<sup>Cab1</sup>, meaning that it is absent from cells from late mitosis until the end of G1 phase allowing licensing to occur in this time window[243]. Several studies also suggest that Geminin has a role promoting licensing and that the key factor is the stoi-
Geminin regulates Cdt1 by binding and sequestering it during S and G2-phase. Its degradation after mitosis allows Cdt1 participation in origin licensing in early G1 phase.

In summary, CDK-dependent mechanisms are mainly responsible for maintaining proper control of DNA replication initiation and CDK-independent pathways are important for preventing re-replication in metazoans. These mechanisms are critical for the maintenance of genome stability.

1.1.1.3 DNA replication in eukaryotes

While the replication machinery is fairly well defined in *E. coli*, the eukaryotic replisome requires more proteins to function and is presently less well understood. Much of our knowledge of the elongation phase of DNA replication comes from biochemical and structural analysis of replication factors, *in vitro* studies using the SV40 viral DNA system and genetics using the yeasts *S. cerevisiae* and *S. pombe*. While the core components of the replication machinery of the *E. coli* system have eukaryotic counterparts and are overall more similar than different, this core machinery is intertwined with a plethora of other factors that regulate replication and coordinate it with other cellular processes. (Table 1.1) (reviewed in [25, 78, 156, 247, 248]). For instance, as already discussed, the eukaryotic helicase differs from its prokaryotic counterpart in that it is extensively regulated. It is a target of phosphorylation, ubiquitylation and requires activation after assembly on DNA. Furthermore, its polarity is opposite to that of DnaB meaning it encircles the leading strand[25, 249]. Several of the single subunit proteins in prokaryotic replisomes, have multisubunit equivalents like RPA, the single-strand binding protein
1.1 Genome stability and maintenance

Table 1.1: Eukaryotic replicative DNA polymerases

The nomenclature for the cartoon depictions is for *S. cerevisiae* genes. For Pol d, a fourth subunit (p12) is shown, which is found in humans but not in *S. cerevisiae*. Specific subunit interactions are as shown. The largest subunit of each complex contains the polymerase activity and, for Pol d and Pol e, the 30'-exonuclease activity. The Pri1 subunit of Pol a is the catalytic primase subunit. Proposed replication functions and additional functions are as indicated. Reproduced from [255] with permission from the publisher.

(SSB) equivalent, and the Polα/primase complex[250, 251]. Similarly, many replisome components have functions beyond DNA replication. A prime example of this is the eukaryotic sliding clamp PCNA, which is involved in several cellular pathways such as DNA repair and translesion synthesis, DNA methylation, cell cycle regulation and chromatin dynamics[252–254]. Other components of the eukaryotic replication machinery have no known prokaryotic counterpart such as Cdc45, Dpb11 and the GINS complex[25].

**Eukaryotic replicative polymerases** In eukaryotes, the replication fork is propagated by three DNA polymerases: Polymerase α/primase (Pol α), DNA Polymerase δ (Pol δ), and DNA Polymerase ε (Pol ε), the latter of which are the only nuclear polymerases in eukaryotes that possess intrinsic proofreading (3’ exonucleyotic activity) ability (see 1.1)[256]. Until the discovery of Polδ, Polα was thought to be the main replicative polymerase in eukaryotes[25]. This polymerase has the unique capability to also initiate DNA replication in eukaryotic cells, because the primase and DNA polymerization abilities are both found in its four subunit complex[39, 78, 257, 258]. Its subunit structure is conserved among eukaryotes[25, 78]. The largest subunit Pol1 has polymerase ability[25]. The Pri1 subunit (p48) contains the primase activity and catalyzes the formation of the short RNA primers utilized for limited elongation by the polymerization function of Pol α[25, 78]. The other two subunits play roles in stabilising and regulating the catalytic subunits[78]. Pol α/primase is the only protein complex known to
Genomic integrity and instability

prime DNA replication in eukaryotes. Primase binds the single-stranded DNA template and starts RNA primer assembly[78]. The final size of the primer varies in eukaryotes between 8 and 12 nucleotides depending on the structure of the primase[78, 258, 259]. This primer is then extended by the Pol1 subunit by about 20 nucleotides[25, 256, 259, 260]. The polymerase is then switched in a process termed "polymerase switching" which is known to be mediated by RFC[260–264]. Due to its unique ability to initiate DNA synthesis, Pol α is tightly regulated via post-translational modifications, such as phosphorylation by CDKs, Cdk2/cyclin A (Cdc28/Clb in S. cerevisiae) during S and G2, and interactions with other proteins especially those involved in initiation, such as Cdc45[78, 113, 265]. Additionally, it cannot initiate on single stranded DNA coated in RPA on its own accord[266, 267]. DNA polymerase δ is the lagging strand polymerase and thus responsible for generating Okazaki fragments[78]. In budding yeast, Pol δ has three subunits: Pol3, Pol31/Hys2, and Pol32[268, 269]. Fission yeast and humans have an additional small fourth subunit, which likely stabilizes the complex[270, 271]. In all three organisms the subunits are assembled in a similar fashion: the catalytic and second largest subunit form a complex and the third subunit binds to the second[78]. Pol δ interacts with PCNA via at least two of its subunits[25]. The homotrimeric PCNA is located "behind" the polymerase on the DNA strand[272, 273]. As in E. coli this likely acts as a tether for the polymerase, decreasing its dissociation from DNA, thereby increasing the processivity of the polymerase[274]. The third, Pol ε, was first identified in yeast and most insights have been gained in this system[275]. In S. cerevisiae, Pol ε is a heterotrimer of the Pol2, Dpb2, Dpb3, and Dpb4 subunits[276]. In humans the catalytic subunit is called p261 and is encoded by the POLE gene. The small subunits have also been identified in other organisms[277]. While Dpb2 is essential in both budding and fission yeast, the other two subunits are non-essential (except Dpb3 in S. pombe)[78]. However, the phenotypes of deletions in S. cerevisiae suggest they provide stabilising functions to Pol ε and work in S. pombe suggests roles during initiation, elongation and cell separation[78]. POL2 itself is an essential gene and mutations in the catalytic site are lethal[278–280]. However, perhaps surprisingly, almost the entire catalytic domain is non-essential in S. cerevisiae and S. pombe[278, 279, 281]. These mutant strains show defects including a defect in elongation step of chromosomal DNA replication[278, 279, 281, 282]. The C-terminal region shows poor overall sequence identity between yeasts and human, but is contains two conserved cysteine-rich motifs that coordinate zinc fingers that interact with the other subunits[283–285]. This region is both essential for growth and required for the S-phase checkpoint in S. cerevisiae[279, 286]. One of the motifs contains a metallocenter that has been shown to be critical for subunit interaction[287]. Pol ε subunit interaction seems important for genome stability. Mutations in the yeast Dpb2 sub-
unit, that stabilise its interactions with other subunits, cause an increased mutation rate[259]. In fact, evidence suggests, that the presence of mutated Dpb2 protein in the cell does not only affect the intrinsic fidelity of Pol ε, but also promotes the increased participation of DNA polymerase zeta (Pol ζ; the catalytic subunit encoded by REV3 in S. cerevisiae), an error-prone polymerase, in DNA replication[288]. The inter-origin distance can be long in eukaryotes: in budding and fission yeast it is on average 38kb[289] and can be much longer in higher eukaryotes[259]. The eukaryotic replicative polymerases Pol δ and Pol ε have been found to be comparable in their high processivity in the presence of the sliding clamp PCNA[290, 291], reviewed in [259]. However, while Pol δ processivity requires its interaction with PCNA, Pol ε seems to be highly processive even without PCNA[292]. Pol ε shows high affinity for DNA, but a low affinity for PCNA; in contrast, Pol δ shows the opposite affinities for those two binding partners[291]. Recently, a structure for POLE has shed some light on this phenomenon: Pol ε has an extra domain (P domain) close to the DNA, allowing it to encircle the nascent double-stranded DNA, likely decreasing it "falling off" the DNA(Fig. 1.9). Lagging strand replication, like in E. coli, requires more steps to be initiated. Replication has to be initiated several times by primase and the primer elongated by Pol α[78]. Pol α is then switched to Pol δ which elongates the growing DNA strand until it encounters the previously synthesized Okazaki fragment[78]. Subsequently, the discontinuously synthesized fragments of DNA are joined up in a process called "Okazaki fragment maturation"[78]. This process needs to be highly accurate to avoid insertions and deletions (INDELs) and efficient so that none of the nearly 100,000 nicks in the DNA, generated during one budding yeast S phase, remain[293]. The former would alter the genetic information and the latter, if unrepaired and then replicated, would result in a double-strand break (DSB) in the DNA. And while DSBs can be repaired, a small number of lesions can overwhelm the repair system and cause cell death[294]. The nicks between fragments are processed by Pol δ and Rad27(FEN1) and ligated by DNA ligase I[293]. Pol δ displaces 2-3 nucleotides of any RNA or DNA of the next fragment that it meets[293]. Rad27, a 5’ flap endonuclease, efficiently processes the nick. If it is absent or not functioning at an optimal level, Pol δ idles (it backs up using its exonuclease activity)[293]. This process is thought to keep the length of displaced downstream nucleotides to a minimum[293]. This behaviour was not observed for Pol ε consistent with Pol δ as the lagging strand polymerase. At some point Pol δ will switch from idling to strand displacement[293]. In these cases, displaced DNA will be single-stranded and coated by RPA, which makes it an inefficient target for FEN1, especially if the DNA forms secondary structures[295]. As demonstrated in yeast, in these cases the essential Dna2 nuclease/helicase will cleave these flaps[295]. While it is thought to be the less common path to process Okazaki
Genomic integrity and instability

Replicating the ends of DNA: Telomeres and telomerase  As a consequence of the linear nature of eukaryotic chromosomes, telomeres are long stretches of repetitive sequences at the ends of each chromosome there to protect them from degradation and subsequent loss of genetic information. During replication, telomeres provide a conundrum to the replisome, termed the "end replication problem": because of the way lagging strand synthesis is achieved the 3’ end cannot be replicated in its entirety and some of the telomere sequence at the very end will be lost with each round of replication[297]. This successive shortening of chromosomes is buffered by a specialized reverse transcriptase (a molecule that synthesizes DNA from an RNA template), called telomerase, that takes advantage of its own RNA template to add short GT-rich repeats to extend telomere repeats[297]. While constitutively active in many organisms such as budding yeast, its activity in multicellular organisms is usually restricted to a few subsets of cells such as germ cells and stem cells, and progressive telomere shortening in somatic cells has been linked to senescence and aging[298] while telomerase reactivation is
considered a hallmark of cancer[299].

**Which polymerase replicates which strand?** In *E. coli*, both leading and lagging strand are essentially simultaneously replicated by two DNA polymerases of the same kind, namely PolIII. In eukaryotes, the answer is a lot less clear-cut, but extensive work especially in budding yeast in the past two decades has helped shed some light on the contributions of Pol δ and Pol ε on the replication of leading and lagging strand. The current and most widely espoused model of the replication fork names Pol ε the leading strand and Pol δ the lagging strand polymerase[78, 255, 300–305]. Substantial evidence, suggests Pol δ as the lagging strand polymerase. For instance, in *S. cerevisiae*, telomere addition is dependent on Pol α and Pol δ[306]. Additional to the ability of Pol δ to idle at Okazaki fragments, studies of *pol3 rad27* double mutants further suggest that Pol δ is involved in Okazaki fragment maturation and thus likely also in elongation[293, 307, 308]. Most *pol3 rad27* double mutants are lethal, and those that are viable accumulate small duplications, a common defect in Okazaki fragment maturation. Additionally, Pol δ directly interacts with Pol α via the Pol δ Pol32 subunit[309, 310]. The pol1-L868M allele reduces Pol α fidelity, but not its activity[311]. This mutator phenotype is exacerbated by inactivation of Pol δ proofreading, but not affected by loss of Pol ε proofreading[311]. This could mean that Pol δ could correct errors made by Pol α[311]. The dispensable nature of the *POL2* N-terminal polymerase domain calls the extent of Pol ε contribution to replication into question[279]. But the lethality of missense mutations of active site residues in Pol ε points to the significance of its polymerase activity[280]. Studies with mutated forms of Pol δ and Pol ε suggest that they proofread errors on opposite strands during chromosomal replication[272, 302].

Considering the evidence for Pol δ as the lagging strand polymerase, this would place Pol ε on the leading strand. However, this does not elucidate how much Pol δ contributes to leading strand replication. In fact, Pol δ could well replicate the vast majority of leading strand, which is also supported by the fact that in budding yeast the inactivation of Pol δ proofreading has a bigger effect than inactivation of Pol ε proofreading when measured in mutation rate reporter assays[311–313]. Work using a yeast genetic system tried to address the contribution of Pol δ. A reporter gene is inserted asymmetrically between two chromosomal origins of replication ARS306 and ARS307[300]. This experimental set-up allowed assignment of which strand would be leading and which lagging during replication and, by flipping the reporter, these assignments could be reversed. Using the *pol3-L612M* strain, which has wild-type activity, but an increased mutation rate, allowed the determination of which strand was copied by the faulty polymerase[300]. Critically, out of the 12 possible base substitution
errors, six are found at an increased frequency and the six base substitution error rates that increase and the six that do not can be paired as “reciprocal” mispairs[314]. For example, a T-A to C-G base substitution can occur either by mispairing of the T to a dGMP or the A to a dCMP, which occur. The pol3-L612M strain generates template T-dGMP mispairs at a much higher frequency than the other[314]. This allows determination of which strand was mutated. Regardless of the orientation of the reporter, mutations in the reporter gene accumulated almost exclusively (>90%) on the assigned lagging strand, suggesting that L612M Pol δ has at most a limited role in leading strand replication[300]. This is further corroborated by work where a Pol ε mutant was created that retains its replication ability but not fidelity. The authors analysed mutation patterns and frequencies in a mutational reporter gene and found that they depend on the orientation of the reporter and its location relative to origins of replication.

Taken together, under normal conditions, Pol δ is the lagging strand polymerase and Pol ε is the leading strand polymerase, though its absence can be compensated for by the replisome[305]. This seems to be conserved in S. pombe[315]. The division of labour between the two polymerases has not yet been clearly resolved in higher eukaryotes. Experiments with nuclear extracts of Xenopus leavis eggs, which are robust systems for biochemical analysis, showed that depletion of Pol δ or Pol ε resulted in a considerable decrease in DNA synthesis[316, 317]. Immunodepletion of Pol δ resulted in a significantly more severe defect in DNA synthesis than that of Pol ε and was associated with a defect in lagging strand synthesis, namely an accumulation of short nascent strands and gapped DNA[316]. In human cells, Pol ε foci co-localise with sites of active DNA synthesis, but not always with Pol δ[318, 319]. Pol ε is also not always present in replication forks containing PCNA though that could be due to its high processivity without PCNA[78, 318]. In vitro and in vivo replication of the SV40 virus genome can occur entirely with Pol α and Pol δ[156, 263, 320]. Pol ε is not detected on viral DNA - the other two polymerases are - but is present on chromosomal DNA[320]. However, SV40 is a virus that replicates quickly and independently of the cell cycle - due to it encoding its own initiation machinery[321]. Pol ε is known to also have roles in replication initiation and cell cycle checkpoints, which makes it likely that Pol ε is not required for replication of the SV40 virus DNA, but indispensable for chromosomal replication[259, 322]. While the current model of Pol δ as lagging strand and Pol ε as leading strand polymerase is likely broadly applicable, many questions remain about replication of certain parts of the genome, especially those that are difficult to replicate such as fragile sites and repetitive sequences[255]. The replication fork has been shown to be quite plastic and it is entirely possible that contributions of the different polymerases vary significantly depending on context.
1.1 Genome stability and maintenance

1.1.1.4 DNA polymerases

Structure of DNA polymerases The eukaryotic replicative polymerases are all members of the B-family of polymerases, while the prokaryotic replicative polymerase belongs to the C-family[256]. While similar in many ways, polymerases show marked differences within and between species. All polymerases must be able to move along the template as synthesis proceeds[323]. Additionally, all have some measure of and a mechanism for fidelity ensuring that the copied information is reasonably preserved[323]. Crystal structures obtained to date also show that all polymerases use the same two-metal-ion mechanism to catalyse the polymerization reaction[323]. DNA polymerases have been divided into 7 different families based primarily on the structure of the catalytic subunit and amino acid sequence[323, 324](1.2). The known DNA polymerases have conserved structures, especially in the catalytic subunits. However, catalytic subunits can range in size by about one order of magnitude (39-kDa human Pol β compared to 353-kDa human Pol ζ)[324]. The overall structure of a DNA polymerase has often been likened to a right hand with different protein domains designated "palm", "fingers" and "thumb"(Fig. 1.10). The subunits form a cleft with the palm domain at the bottom. This domain contains three catalytic amino acid residues which coordinate two divalent metal ions essential for catalysis[324]. Generally, the palm seems to be the location for catalysis of the polymerization reaction, the fingers play an important role in interactions with the template base and the incoming nucleoside triphosphate which will be added to the DNA chain, and the thumb is thought to be involved in positioning of the double stranded DNA and processivity, as well as the movement of the polymerase along the DNA[324]. While the palm domain appears relatively conserved across families, the structures that have been obtained so far show great variation in the finger domains between families(Fig. 1.11)[323]. Figure 1.11 shows that although thumb and finger structures are not homologous, they show at least minor similarities: in this example thumb domains are mostly made up of antiparallel α-helices of which at least one seems to interact with the minor groove of the primer-template product, and out of the finger domains three out of four an α-helix provides interaction with the incoming dNTP. In the fourth case this seems to be accomplished by a similarly positioned β-ribbon[323].

Mechanism of DNA polymerization It is believed that all DNA polymerases use a two-metal ion mechanism to catalyse the polymerization reaction[323, 324]. The reaction can only ever occur on the 3’ end of the new strand giving polymerases a 5’ to 3’ directionality[326]. DNA polymerases are incapable of assembling nucleotides de novo. They all require a primer of either DNA or RNA. For the polymerization reaction, the 3’-OH of a primer strand and the α-phosphate of a dNTP are adjacent to each other and oriented optimally for the reaction[326].
**Table 1.2:** Families of DNA polymerases
Reproduced from [324] with permission from the publisher.

<table>
<thead>
<tr>
<th>Name</th>
<th>Family</th>
<th>Bacterial gene</th>
<th>Human gene</th>
<th>Yeast gene</th>
<th>Mol. Wt. (kDa)$^a$</th>
<th>3’ Exo</th>
<th>Other activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC Pol I</td>
<td>A</td>
<td>pol A</td>
<td></td>
<td></td>
<td>103</td>
<td>+</td>
<td>5’ Exonuclease</td>
</tr>
<tr>
<td>γ (gamma)</td>
<td></td>
<td></td>
<td>POLG</td>
<td>MIP1</td>
<td>140</td>
<td>+</td>
<td>dRPlyase</td>
</tr>
<tr>
<td>δ (theta)</td>
<td></td>
<td></td>
<td>POLQ</td>
<td>—</td>
<td>250</td>
<td>—</td>
<td>ATPase, helicase</td>
</tr>
<tr>
<td>λ (mu)</td>
<td></td>
<td></td>
<td>POLN</td>
<td>—</td>
<td>100</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Ec Pol II</td>
<td>B</td>
<td>polB</td>
<td></td>
<td></td>
<td>89</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>α (alpha)</td>
<td></td>
<td></td>
<td>POLA</td>
<td>POL1 (CDC17)</td>
<td>165</td>
<td>—</td>
<td>Primase</td>
</tr>
<tr>
<td>δ (delta)</td>
<td></td>
<td></td>
<td>POLD1</td>
<td>POL3 (CDC3)</td>
<td>125</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>ε (epsilon)</td>
<td></td>
<td></td>
<td>POLF</td>
<td>POL2</td>
<td>225</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>ζ (zeta)</td>
<td></td>
<td></td>
<td>POLZ (REV3)</td>
<td>REV3</td>
<td>353</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Ec Pol III</td>
<td>C</td>
<td>dnaE</td>
<td></td>
<td></td>
<td>130</td>
<td></td>
<td>(separate subunit)</td>
</tr>
<tr>
<td>β (beta)</td>
<td>X</td>
<td></td>
<td>POLB</td>
<td>—</td>
<td>39</td>
<td>—</td>
<td>dRP lyase, AP lyase</td>
</tr>
<tr>
<td>λ (lambda)</td>
<td></td>
<td></td>
<td>POLL</td>
<td>POL4 (POLX)</td>
<td>66</td>
<td>—</td>
<td>dRP lyase, TdT</td>
</tr>
<tr>
<td>μ (mu)</td>
<td></td>
<td></td>
<td>POLM</td>
<td>—</td>
<td>55</td>
<td>—</td>
<td>TdT</td>
</tr>
<tr>
<td>TdT</td>
<td></td>
<td></td>
<td>TdT</td>
<td>TdT</td>
<td>56</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>σ (sigma)</td>
<td></td>
<td></td>
<td>POLS (TRF4-1)</td>
<td>TRF4</td>
<td>60</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Ec Pol IV</td>
<td>Y</td>
<td></td>
<td>dinB</td>
<td>umuC</td>
<td>40</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Ec PolV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>46</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>η (eta)</td>
<td></td>
<td></td>
<td>POLH</td>
<td>RAD30</td>
<td>78</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>ι (iota)</td>
<td></td>
<td></td>
<td>POLI (RAD30A, XPF)</td>
<td>—</td>
<td>80</td>
<td>—</td>
<td>dRP lyase</td>
</tr>
<tr>
<td>κ (kappa)</td>
<td></td>
<td></td>
<td>POLK (DNB)</td>
<td>—</td>
<td>76</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>RevI</td>
<td></td>
<td></td>
<td>REV1</td>
<td>REV1</td>
<td>158</td>
<td>—</td>
<td></td>
</tr>
</tbody>
</table>

$^a$Deduced from protein primary structure.

**Figure 1.10:** Structure and representation of replicative DNA polymerases
A - Surface crystal structure of a DNA polymerase complexed with DNA. B & C - Cartoon representation of the DNA polymerase structure in polymerization mode (B) and proofreading mode (C). Figure reproduced from [325]. Used by permission of the publisher.
Figure 1.11: Comparison of primer-template DNA bound to four DNA polymerases.
A - Taq DNA polymerase bound to DNA (co-crystal structure); B - the binary complex of HIV-1 RT and DNA (co-crystal structure); C - the model of DNA bound to RB69 gp43 (homology model); D - the ternary complex of rat pol β with DNA and dideoxy-NTP (co-crystal structure).
Figure reproduced from [323] in accordance with the publisher’s copyright permission policy.
Even though they are structurally different, the finger domains of the pol β, RT, pol I, and pol α DNA polymerases all use similar residues to stabilize the incoming dNTP[323]. In the presence of a correct template–primer duplex the finger domain undergoes a rotation and this conformational change "closes" the active site[323, 326]. 3’-OH and α-phosphate of dNTP are then properly aligned for the reaction using the two metal ions. In Fig. 1.12 Metal ion A affects the 3’OH of the primer, which is thought to lower the pKA of the OH enabling its attack on the incoming dNTP[323]. Both metal ions are also likely to stabilize the structure and charge of the reaction transition state[323]. Metal ion B interacts with the β- and γ-phosphates and is thought to facilitate their leaving[323]. This reaction only occurs efficiently if the two reaction partners are oriented correctly within the active site. Thus the intrinsic fidelity of the polymerase active site is achieved by two things: the induced fit conformational change of the finger domain, which detects the presence of a correct base pair, and the fact that an incorrect nucleotide will not hydrogen bond with the template easily and thus the optimal arrangement of substrates for the enzymatic reaction will not be achieved[323, 326–329]. While the basic mechanism of polymerization is conserved, polymerases vary considerably with regards to efficiency, fidelity and substrate preference. The efficiency of different DNA polymerases at inserting correct nucleotides varies over an astonishing 107-fold range, while their fidelity varies as much as 100,000-fold (Fig. 1.3)[324, 330]. Examples for variation in substrate preference include Pol β, which preferentially uses single-nucleotide gaps, and Pol η, which tends to replicate damaged DNA[324]. Some polymerases have additional enzymatic activities including, but not limited to, 5’-to-3’ exonuclease activity in Pol δ, Pol ε and Pol γ, ATPase capability in Pol θ and primase activity in Pol α[324]. These can be found in a different domain of the same polypeptide (e.g. Pol δ, Pol ε) or in separate, but tightly associated, subunits (e.g. Pol α)[324].

**Fidelity of DNA polymerases** Replication is a very accurate process, especially in higher eukaryotes[331]. It is estimated that copying all 6000 Megabases in a human cell proceeds with about one error per cell division, resulting in an error rate between 1x10^{-9} and 1x10^{-10} errors per base pair in mammalian cells[259, 332]. This is mainly due to three things: intrinsic fidelity of the polymerase mechanism; 5’-to-3’ exonuclease activity of the replicative polymerases; and mismatch repair(Fig. 1.13)[259, 323, 333–335]. The prevailing model is that those three processes act in series[336–339]. The replicative polymerases misincorporate a nucleotide roughly every 10^4-10^5 nucleotides[334, 336]. Most of those errors are reversed by the exonuclease activity[340, 341]. The remaining mistakes are targeted by the mismatch repair system, accounting for the overall low error rate[341]. The 5’-to-3’ exonuclease activity
1.1 Genome stability and maintenance

Figure 1.12: Mechanism of DNA polymerization
Polymerisation uses a two-metal ion mechanism that is thought to stabilize the resulting penta-coordinated transition state. (The two essential aspartates are annotated with the *E. coli* DNA polymerase I numbers.) Figure reproduced from [323] in accordance with the publisher’s copyright permission policy.
Table 1.3: Error rates of DNA polymerases from different families
Reproduced from [324] with permission from the publisher.
allows DNA polymerases to excise wrongly incorporated nucleotides by the hydrolysis of the phosphodiester bond and subsequently reattempt incorporation of the correct nucleotide. The structure of the Klenow fragment, a large fragment of *E. coli* DNA pol I which is obtained after cleavage with subtilisin, showed that it is comprised of two separate domains[342]. One contains the active site for the polymerization and the other the active site for the exonuclease activity resulting in an approximately 30-40Å distance between the two active sites[259, 323]. When the structure is obtained with DNA that contains a 4 nucleotides long 3’ overhang the ssDNA is seen to bind the exonuclease site[343]. Extensive structural, biochemical and mutagenic studies of exonuclease-domain containing polymerases suggest that a 2-metal ion mechanism is utilised analogous to the polymerization mechanism[344–347]. The proposed mechanism is that both active sites compete for the 3’ end of the primer strand resulting in a rapid shuttling between them[345, 348, 349] (reviewed in [323]). The exonuclease site binds ssDNA while the polymerase active site preferentially associates with correctly Watson-Crick paired double-stranded DNA[323]. Mismatches in the dsDNA distort and destabilise it thus favouring the binding to the exonuclease site[323]. Furthermore, the polymerase is known to stall after a mismatch - likely due to the fact that the 3’ end that is to be added onto tends to be misoriented - which further increases the probability that the most recently formed phosphodiester bond will be hydrolised[323]. This means that the fidelity of a DNA polymerase is thus a combination of correct base-pairing in the polymerization active site and competition with the exonuclease active site which preferentially excises mismatched nucleotides and single stranded DNA.

There are a variety of assays that can and have been used to determine the intrinsic accuracy of a polymerase. Initially, assays used synthetic templates of only one or two bases and radioactive nucleotides[350]. More recently, the lacZ fidelity assay has been used effectively. It measures polymerase errors using a gapped DNA substrate *in vitro* that contains the wild-type lacZ-α complementation sequence[351]. This assay scores all 12 single base-base mismatches and different deletions in a variety of sequence contexts and has been used to measure the intrinsic fidelity of a range of polymerases[334, 352] This showed that the fidelity of polymerases can range in several orders of magnitude, but that, in general, replicative polymerases tend to be highly accurate(Fig. 1.14)[334].

In eukaryotes, only Pols ε, δ, and γ contain intrinsic exonuclease activity[259, 324]. This is beneficial because those polymerases replicate the majority of genomic DNA - Pol ε and Pol δ in the nucleus, and Pol γ in the mitochondria. POLE and POLD1, the catalytic subunits of Pol ε and Pol δ respectively, are known to contain three conserved motifs in their exonuclease domains called ExoI, ExoII and ExoIII[259]. The three motifs are regarded to
Nucleotide selectivity, exonuclease activity (proofreading) and mismatch repair all contribute to DNA replication fidelity in series to different degrees. The brackets indicate the magnitude of the contribution the different processes can make and examples of defects and conditions that result in reduced fidelity are shown on the right. Reproduced from [334] with permission from the publisher.

Contribute to exonuclease activity in different manners and quantities based on a collection of structural, genetic and biochemical work involving bacteriophage, prokaryotic and yeast polymerases[259]. The two divalent metal ions that are known to be critical for the hydrolysis reaction are coordinated by conserved acidic residues within these motifs[259]. ExoI contains a beta sheet with two absolutely conserved acidic residues (one glutamate and one aspartate in Pol ε and Pol δ) known to coordinate Metal A directly, while ExoII and ExoIII each contain a conserved aspartate that indirectly coordinates Metal B and Metal A respectively via water molecules[312, 346]. These residues are placed close to the terminal phosphate when complexed with DNA, allowing them to coordinate the two metal ions to efficiently catalyse hydrolysis of the 3-terminal phosphodiester bond[259]. The physical distance between the two active sites within the catalytic domain requires the mismatched primer to melt away from the other strand and switch active sites. Active site switching is promoted by what has been described as a hinge in the thumb domain and a β-hairpin[353–357]. Work in S. cerevisiae Pol δ has suggested that this β-hairpin eases strand separation and similar structures have been found in polymerases from T4 and RB69[358]. Recently, Hogg and co-workers showed that Pol ε is lacking this extended β-hairpin despite it being a high fidelity polymerase[296]. With the exception of DNA polymerase B1 from Sulfolobus sulfataricus, the extended β-hairpin is found in the exonuclease domains of all other B-type polymerases.
Figure 1.14: Fidelity of different DNA polymerases

“Eukaryotic DNA polymerase error rates for single base mutations. (A) Error rates for Homo sapiens Pol δ and S. cerevisiae polymerases α, δ and ε, for single base substitutions (BS; light grey bars) and one base deletions (dark grey). (B) Error rates for Homo sapiens polymerases η, κ, θ and ν and S. cerevisiae Pol ζ. Note the difference in scales between panels A and B. Error rates for each polymerase were obtained with the lacZ-α forward mutation assay. The assay measures error made when copying a template in a gapped DNA substrate in vitro that contains the wild-type lacZ-α complementation sequence. The assay [351] scores all 12 single base–base mismatches and different single-base deletion mismatches in numerous different sequence contexts.” Reproduced from [334] with permission from the publisher.
with available structures[296]. The authors speculate that Pol ε might be able to maintain high fidelity in the absence of the β-hairpin due to its P domain which increases polymerase association with DNA[296]. If the P domain were able to decrease dissociation during active site switching, this might account for the lack of mutator phenotype due to the lack of an extended β-hairpin[296]. In vitro and in vivo studies using a yeast mutant where both acidic residues in the ExoI motif were mutated show that exonuclease ability increases Pol ε fidelity by about an order of magnitude in mismatch repair proficient cells[359–361]. Together these two mutations abolish the exonuclease activity and depending on the reporter gene used tend to increase base-pair substitutions[259]. The equivalent mutation in Pol δ increases the mutation rate by about 100-fold(Fig. 1.14)[361]. Additionally, Pol δ has a lower fidelity for single- and multi-base deletions[362]. DNA polymerases can proofread their own mistakes (proofreading in cis) as well as sometimes correct errors made by other polymerases (proofreading in trans). For example, Pol δ is thought to be able to proofread for Pol α and Pol ε, whereas the reverse has not been demonstrated[259, 311].

**Other functions of DNA polymerases** Beyond DNA replication, DNA polymerases are involved in a variety of other cellular pathways. Many of them have specific roles in DNA repair pathways which are discussed in more detail below. Beyond that, some cell-cycle checkpoints depend on Pol ε[286, 363]. Similarly, primase and exonuclease deficient mutants of Pol δ show defects in DNA damage checkpoints[337, 364]. Additionally, when replication forks stall at DNA damage, they can be restarted by a "fork regression" process[365]. According to the model, in *E. coli*, the replication fork regresses providing an undamaged template strand for DNA Polymerase II. In eukaryotes this synthesis is probably performed by a major replicative polymerase which are also thought to conduct the DNA synthesis required during homologous recombination[324]. The synthesis activity of several DNA polymerases has been implicated in the development of the human immune system[324]. For instance, mammalian cells contain a template-independent polymerase called terminal deoxynucleotidyl transferase (TdT)[324]. TdT functions by inserting nucleotides at the junctions between the V, D and J elements in the recombination of immunoglobulin heavy-chain genes causing junctional diversity[366–368]. The somatic hypermutation (SHM) process that results in even more immunological diversity is likely initiated by activation-induced cytosine deaminase (AID), followed by replicative-type or repair-type DNA synthesis which may include members of family B, such as Pol ζ, Pol δ, and Pol ε, as well as members of family Y, such as Pol η or Pol ι[324].
1.1.2 DNA repair and Translesion Synthesis

DNA within a cell can experience different types of damage caused by a variety of endogenous and exogenous processes (see 1.3). DNA repair is a collective term to describe the plethora of mechanisms cells have evolved to identify and repair DNA damage. These processes are critical for genome maintenance: DNA damage can lead to mutations, fractured DNA and cell death. DNA damage comes in different types, varied severity and the repair pathway chosen depends heavily on the type of damage observed, as well as other factors such as cell cycle progression and transcription.

1.1.2.1 Direct Damage Reversal

If only a single base is damaged, direct damage reversal is one of the simplest and, in evolutionary terms, thought to be the oldest DNA repair mechanisms the cell can choose. These pathways rely on a single protein that can reverse DNA damage efficiently in a virtually error-free process with no need for a DNA template[369]. These proteins show high substrate specificity and act without the need for removal of the affected base or cutting of a DNA strand. Two well-studied types of DNA damage reversal are (i) the repair of pre-mutagenic pyrimidine dimers by photolyases and (ii) the reversal of alkylation damage by alkyltransferases. Pyrimidine dimers are molecular lesions where adjacent pyrimidine bases form covalent linkages that can distort the DNA helix[370, 371], and photoreactivation is the process by which pyrimidine dimers are returned to their original state[372]. The most common lesions - cyclobutane pyrimidine dimers (CPDs, including thymine dimers) and 6,4 photoproducts - are repaired by photolyases which, using 350-450 nm light as an energy source[371], inject one electron into the dimer, which undergoes spontaneous splitting into its monomers[370]. Because it requires light to function, direct damage reversal for example does not function in cells that are not reached by sunlight. In fact, in placental mammals, this type of damage is commonly repaired by nucleotide excision repair since photolyases are no longer functional in these organisms[372]. Alkylation damage is the addition of an alkyl group to DNA[373]. A well-known example of direct reversal of alkylation damage is the conversion of O6-methylguanine back to guanine by the O6-alkylguanine DNA alkyltransferases (also known as MGMT)[374]. O6-methylguanine is mutagenic to cells, because it base-pairs to thymine as well as cytidine, causing G:C to A:T transitions[375]. MGMT is not a true enzyme since it removes the methyl-group from the guanine in a stoichiometric manner using an SN2-type reaction[374]. Other examples of direct reversal of different types of alkylation damage are known such as the E. coli protein Ada which is an isozyme of MGMT[376] and can repair O4-methylthymine in
addition to O6-methylguanine, by the direct transfer of the methyl group from the affected base to a reactive cysteine residue[377, 378]. Direct damage reversal is thus a very efficient and useful process in cells, but the flip side of this specificity is the limited number of damage that can be repaired and that in some cases the repair proteins are used up in the process[369].

### 1.1.2.2 Damage to one strand of the DNA

If the DNA damage is confined to one strand only, then the other strand can be used as a template to repair the DNA correctly. There are a number of repair mechanisms that can excise a damaged nucleotide and direct its correct repair.

**Base excision repair (BER)**  Base excision repair is used to correct lesions that do not distort the structural integrity of the double helix (recently reviewed in [379, 380]). Commonly this damage involves oxidation, alkylation, deamination, depyrimidination or deprivation. To repair this damage BER relies on a variety of DNA N-glycosylases that recognize specific types of DNA damage and catalyse their removal by hydrolyzing the N-glycosidic bond anchoring the base to the phosphor-backbone[380]. This creates an basic or apurinic-apyrimidinic (AP) site which is recognized and cleaved by an AP endonuclease resulting in a single-strand break with a 5'-deoxyribose phosphate (5'-dRP) end that has to be removed[381]. In budding yeast, there is evidence that Rad27 removes the 5'-dRP[382], followed by resynthesis of the excised DNA by Pol2 (Pol ε) and ligation of the residual nick in the DNA strand by Cdc9[383]. In mammals, this break is repaired by one of two ways (Fig. 1.15). Most commonly, when only a single nucleotide needs to be repaired, a pathway called short-patch BER is utilised. In this case, DNA polymerase β causes the removal of the 5'-dRP and then re-synthesises the previously removed damaged nucleotide[380] and the residual nick in the DNA strand is sealed by XRCC1 in association with either DNA ligase I or DNA ligase III[380]. Alternatively, in about 10% of cases, the 5'-dRP is removed by the FEN1 endonuclease in a process called long patch base excision repair leading to replacement of between two and ten nucleotides[384]. In this case, to replenish the excised nucleotide track DNA polymerases β, δ, and ε are recruited and the process depends on both PCNA and FEN1[379, 380].

**Nucleotide excision repair (NER)**  Nucleotide excision repair is primarily utilised to address distortions in the DNA double helix caused by a variety of biochemical modifications[386]. Considering the wide variety of DNA damage recognised it is likely that this pathway does
1.1 Genome stability and maintenance

Figure 1.15: Base excision repair (BER) of oxidized DNA base lesions
BER demonstrated using 8-oxoguanine (8-oxoG) as an example. 8-oxoG is removed by the DNA glycosylase 8-oxoguanine DNA glycosylase (OGG1) leaving an AP site. AP endonuclease 1 (APE1) subsequently incises at the 5' side of the AP site sugar leaving either a native or oxidised sugar phosphate. The former can be repaired by single-nucleoside BER (1), whereas the latter can be repaired by long-patch BER (2a,2b).

(1) The polymerase (pol) β 5'-deoxyribose phosphate (dRP) lyase removes the native sugar phosphate leaving a single nucleotide gap which is filled by pol β and subsequently ligated.

(2a) An oxidised sugar phosphate cannot be removed by the lyase and is thus repaired by LP-BER, usually mediated by pol β gap-filling synthesis and flap endonuclease 1 (FEN1). This efficient pathway usually replaces only two nucleotides. (2b) Alternatively, repair can also occur by a LP-BER mechanism involving strand-replacement by pol β or pol δ/ε, followed by FEN1 cleavage, usually replacing three or more nucleotides.

Reproduced from [385] with permission from the publisher.
not leverage specific enzymes to recognize different DNA lesions, but rather detects distortions in the DNA double helix itself[387]. Once a DNA distortion is identified, a 25 to 30 base long stretch of DNA including the damage is excised and the gap filled by synthesis using the complementary strand as a template followed by ligation of remaining nicks in the DNA strand(Fig. 1.16)[388]. The versatility of NER allows it to act on a variety of DNA damage types including bulky adducts, photodimers, aromatic amine compounds and other lesions that distort the DNA double helix[386]. It is conserved from prokaryotes to eukaryotes, but in eukaryotes it is generally divided into two categories: transcription-coupled NER[386] and global genomic NER[389]. Global genome-wide NER (GG-NER) is thought to constantly scan the genome of eukaryotic cells for damage. In S. cerevisiae the Rad4-Rad23 protein complex (XPC-Rad23 in mammals) detects any structural changes in the DNA and binds such lesions[386]. Once bound, this complex recruits Rad3 (XPD) and Rad25 (XPB), two helicases with opposite polarity belonging to the general transcription factor TFIIH, which open a denaturation bubble around the damaged DNA[390]. The Rad1-Rad10 heterodimer (XPF-ERCC1) and Rad2 (XPG), structure specific endonucleases, subsequently excise the damaged DNA strand[390, 391]. DNA is synthesised by DNA polymerase δ or ε in co-operation with PCNA [392] after which the nicks are ligated by Cdc9 in yeast[393] and by XRCC1 with either DNA ligase I or DNA ligase III in humans[394]. Transcription-coupled NER (TC-NER) acts in a very similar manner, the main difference being that it acts more rapidly on lesions occurring on the transcribed strand of genes[395]. Unlike GG-NER, TC-NER does not require the Rad4-Rad23 (XPC-Rad23) complex to recognize a DNA lesion, but is initiated when the RNA polymerase II stalls after encountering a damaged DNA base while transcribing[396]. Once the polymerase recognises the damaged DNA, the process continues as for GG-NER[389]. TC-NER exclusively repairs damage occurring on the transcribed strand, meaning that damage is more efficiently repaired than on the untranscribed strand, in line with the observation of a mutational strand-bias present on a genome-wide scale in cancer cells which usually carry a high number of mutations[397, 398].

Mismatch repair (MMR) As previously mentioned, mismatch repair is the third process responsible for the high fidelity of DNA replication acting in conjunction with the intrinsic polymerase fidelity and proofreading[400–402]. There are two different kinds of mismatches: mispairings between two bases and IDLs (Insertion, Deletion, Loop), which, if left unaddressed, result in point mutations and insertions/deletions, respectively. MMR corrects DNA mismatches in two critical steps: (i) recognising a mismatch and (ii) directing the repair mechanisms towards the newly synthesized strand which carries the incorrectly inserted
1.1 Genome stability and maintenance

Figure 1.16: Nucleotide excision repair (NER)

A - Nucleotide excision repair (NER) repairs damaged DNA bases that distort the DNA helix structure (such as photoproducts resulting from UV exposure). B - The damage is recognized by XPC (bound to HHRAD23B). C - Binding of the XPC-HHRAD23B heterodimer is followed by binding of XPA, RPA, TFIIH and XPG, of which XPA and RPA are thought to allow specific recognition of the damage and TFIIH (a sub-complex of the RNA polymerase II transcription initiation complex) brings a helicase activity allowing unwinding of the duplex at the damaged site creating a bubble in the DNA. D - Subsequently, ERCC1–XPF binds. E - XPG is an endonuclease that cuts the damaged strand 3’ to the damage, while ERCC1–XPF cuts 5’ to the damage. F - Their combined action removes a 27-30nt fragment including the damaged bases and the gap is restored by repair synthesis followed by ligation. Reproduced from [399] with permission from the publisher.
In prokaryotes, this critical distinction between strands is made using methylation marks, whereas in eukaryotes the process remains unclear. Mismatch repair has been extensively studied in *E. coli* and it is known that in *E. coli* DNA is methylated at the N6 of adenine in short dGATC sequences. During and shortly after replication the nascent strand is transiently unmethylated (see 1.1.1.1) which MMR exploits to distinguish between the mother and daughter strand: the MutH type-II restriction endonuclease is able to recognize hemi-methylated DNA and specifically nicks the nascent strand to create an initiation site. The first step in MMR is the recognition and binding of this type of lesion by the MutS dimer, followed by the location of a hemi-methylated dGATC site and generation of a nick by the combined action of MutS, MutL, MutH and ATP. Several models have been proposed to explain how the binding of MutS leads to a nick. Subsequently, helicase II loads at the nick and unwinds the DNA towards the mismatch generating ss-DNA which is swiftly covered by SSB. Depending on the relative position of the mismatch to the nick different exonucleases excise the DNA. The resulting gap is repaired by the DNA polymerase holoenzyme and DNA ligase and, lastly, a deoxyadenosine methylase methylates the daughter strand.

Eukaryotic MMR is very similar, but not completely understood. Like prokaryotic MMR it shows substrate specificity, bidirectionality and dependence on a DNA nick and, while the hemi-methylated dGATC is not conserved, it is thought that eukaryotic MMR discriminates strands by a strand-specific nick. Many of the eukaryotic proteins involved in MMR have been identified by their homology to *E. coli* proteins. In *E. coli* MutS and MutL are heterodimers. The eukaryotic equivalents of MutS are formed by Msh2 and Msh3 (MutSα), which recognises base-base mismatches and 1-2 base indels, and by Msh2 and Msh6 (MutSβ), which recognises larger INDELs. Both are ATPases and involved in recognition of mismatches. Mlh1 heterodimerises to form MutL homologs: with Pms2 to form MutLα, Pms1 to form MutLβ and MLH3 to form MutLγ. Of these, MutLα has been shown to interact with both MutSα and MutSβ and is critical for eukaryotic MMR. Since PCNA has been shown to interact with Msh2 and Mlh1, as well as Msh6 and Msh3, it has been proposed that PCNA recruits MutSα and MutSβ to newly replicated DNA to monitor newly synthesised DNA for mismatches. Evidence from *S. cerevisiae* has identified Exo1 as the only exonuclease definitively involved. It can also bind Msh2 and Mlh1 and has been shown to catalyse 5’ directed mismatch excision in the presence of MutS and RPA. However, considering that exo1 null yeast and mice show only weak mutator phenotypes, there are likely other important exonucleases.
1.1.2.3 Double stranded breaks (DSBs) in the DNA

Double strand breaks are particularly hazardous to the integrity of the genome, because they can lead to genome fragmentation and rearrangements. While single-stranded breaks are likely much more widespread - estimates speak of thousands to tens of thousands of single-strand breaks occurring in every human cell every day - they are also almost all successfully repaired[440]. Current estimates suggest that ~1% of all single-strand lesions result in a DSB leading to approximately 50 DSBs per cell per cell cycle[441]. Considered the most toxic of all DNA lesions, there are three major pathways to repair DSBs[415]: (i) non-homologous end joining (NHEJ), which occurs throughout the cell cycle, (ii) microhomology-mediated end joining (MMEJ), which generally occurs during S phase and (iii) homologous recombination (HR), which competes with NHEJ in late S phase and the G2 phase of the cell cycle. Ideally, cells repair the break as soon as possible and preferentially by the more accurate HR, though NHEJ is considerably faster[415, 442].

Non-homologous end joining (NHEJ) Non-homologous end joining is the most straightforward way to repair a break in DNA: it pairs two broken ends of DNA and ligates them to restore the double helix. In budding yeast, repair of the DNA is guided by short (less than four bases) homologous sequences often located on single-stranded overhangs[415]. In the rare cases that those overhangs are matching perfectly, NHEJ is a non-mutagenic repair process; however, most likely NHEJ results in micro-insertions/micro-deletions or even translocations[415]. In budding yeast, the first step of NHEJ is binding of the broken DNA ends by the heterodimeric Ku70-Ku80 (KU) complex, which tethers the two DNA ends to one another[443] and helps to protect the integrity of the strands inhibiting repair by HR[415]. Subsequently, KU promotes the recruitment of other critical proteins such as Lif1 (XRCC4 in humans) and Dnl4 (DNA ligase IV), which facilitate the direct joining of the two broken ends[444]. Many DNA breaks cannot be mended this way, but require some processing of the broken ends. This processing is likely achieved by the Mre11-Rad50-Xrs2 (MRX) complex (Mre11-Rad50-Nbs1(MRN) in humans), the polymerase Pol4 and the flap endonuclease Rad27[445, 446]. In mammalian cells, Artemis is involved in processing. Like Ku, MRX is likely involved in bridging the broken ends [444], but additionally likely involved with cleaning up the DNA ends for ligation[447]. While its mutagenic potential may not be ideal, re-ligating DNA ends imperfectly is preferable to entering mitosis with fractured DNA which can lead to the loss of large segments of DNA and cell death.
Microhomology-mediated end joining (MMEJ) Another process for repairing DSBs is microhomology-mediated end joining. The exact mechanism behind MMEJ is currently under investigation, but it is known to repair DSBs by relying on small microhomologies of 5-20 nucleotides. Experimental evidence suggests the involvement of factors implicated in HR (MRX, Rad51, Rad52)[415].

Homologous recombination (HR) Homologous recombination is the repair pathway using identical or extremely similar sequence as a template. It is used for the majority of accurate repairs of DSBs and DNA inter strand crosslinks[448]. This template is usually the sister chromatid (after replication in S phase or in G2 phase) or less commonly the homologous chromosome. Different types of HR exist but the first steps are shared between all of them[448]: the MRX(budding)/MRN(human) complex binds the DNA on either side of the break to tether the ends of the break and induces checkpoint signaling (see 1.1.2.5). Binding of the MRX/MRN is followed by extensive resection of the 5’ end with involvement of proteins like Exo1/EXO1[449] and Sae2/CtIP[450], generating long 3’ single-stranded DNA ends which are recognized and coated with the Rad51/RAD51 recombinase. This makes a 3’ nucleoprotein filament which searches for a homologous DNA template and then invades the template duplex displacing one strand of the homologous duplex (displacement loop or D-loop) and pairing with the other resulting in a heteroduplex. A DNA polymerase then extends the end of the invading 3’ end resulting in a complex structure termed Holliday junction. Depending on the different pathways this structure is resolved in different ways (reviewed in [415, 451]). Briefly, classical double-strand break repair (DSBR) uses a two-end invasion, forming double Holliday junctions that can be resolved in a manner leading to a crossover or non-crossover product. In contrast, synthesis-dependent strand annealing (SDSA) also utilises two-end invasion, but produces only non-crossover products. Break-induced replication (BIR), which generally occurs at telomeres, the ends of chromosomes, or when a DSB is encountered by a polymerase, while highly inaccurate[452] does not require two-end invasion, but rather relies on unidirectional DNA synthesis from the location of strand invasion, which can lead to replicating a few hundred kilobases of DNA and is followed by cycles of separation, re-invasion and synthesis until the entire damage is repaired[452]. A slightly different mechanism, called single-strand annealing (SSA) is unique in that no invasion occurs and it is generally used to repair breaks between repeat sequences. During resection of the DNA ends, repeat sequences are recovered and the break is mended by annealing the two overhangs. This process can be highly mutagenic as any sequence that may have existed between the repeats used for annealing will be deleted[451].
1.1.2.4 Translesion synthesis (TLS)

Translesion synthesis is a DNA damage tolerance (DDT) mechanism that allows DNA replication to proceed past a DNA lesion such as a pyrimidine dimer[453] (reviewed in [454]). When one of the regular replicative polymerases encounters DNA damage it stalls[332], a state that cannot be remedied by excising the damage there at the fork as this would lead to DNA breaks. It is a far more sensible choice for the cell to replicate past the damage for the time being if possible and repair the DNA lesion later[332]. This can be achieved by TLS, which - even through it carries an increased risk for small-scale mutations - is preferable to possible large scale mutations[332]. While DNA damage tolerance pathways are not actually repairing DNA damage, they do provide a mechanism to cope with the DNA damage during replication, increasing genome stability and promoting cell survival[455]. Cells achieve DNA damage tolerance by employing specialized translesion polymerases[332], many of which belong to the Y-family of polymerases and whose often larger and more flexible active sites are major contributors to their ability to accommodate damaged nucleotides and incorporate bases opposite them[326]. Usage of these polymerases carries an increased risk of mutagenesis, not only because the damaged and distorted bases they deal with can lead them to mispair, but also because they are generally less reliable even when replicating undamaged DNA[332]. Their error rate during normal synthesis is 1-2 orders of magnitude higher than other polymerases from the A and B family even when one does not factor in any proofreading activity associated with exonuclease domains[324]. However, they can be ideal for a specific type of DNA lesion. For instance, while Pol ι induced mutations when replicating past pyrimidine dimers, Pol η accomplishes error-free bypass of such lesions[456] making it a buffer for NER allowing tolerance of dimers that were missed by the repair process [457, 458](reviewed in [459–461]). This divergence in the ability of the polymerases is due to the different active site geometries of the Y-family polymerases and the flexibility of some of their domains, giving them differences in the spectrum of DNA lesions they can process efficiently and the types of mutations they will induce inadvertently[326]. This is the main reason why activity of translesion polymerases is tightly limited to damaged DNA, with polymerases being switched in a highly deliberate manner with roles for proteins such as PCNA[326]. The first Y-family polymerase to be identified was REV1 which is unique in its ability to only incorporate dCMP[462]. Interestingly, when one compares its structure with Dpo4, a bacterial Y-family polymerase, Rev1 shows an N-terminal extension which forms a long helix, which will come from the minor groove side of the DNA, flip out the (damaged) template base and supply one of its own arginines as a faux-template to hydrogen bond with the dCTP[326]. Another example of a specialized translesion polymerase is the B-family polymerase Pol ζ (Rev3/Rev7 in S. cerevisiae) which
is unique in its ability to extend primers with a terminal mismatch\cite{463–466}. Recently, error-prone polymerases have also been implicated in the repair of DSBs: X-family polymerases have been shown to be involved in NHEJ\cite{467}, and Pol\eta contributes to DNA synthesis during HR\cite{468, 469}.

1.1.2.5 Pausing the cell cycle: checkpoints

In order for cell division to proceed properly and for pathological mistakes to be avoided, cells have developed the ability to interfere with the progression of the cell cycle. The term "checkpoint" was first used by Hartwell and Weinert, who identified them as control mechanisms enforcing dependency in the cell cycle in budding yeast (such as the dependency of mitosis on DNA replication)\cite{470}. They correctly stated, that elimination of checkpoint can result in cell death, improper distribution of chromosomes and other cellular structures such as organelles and increased sensitivity to environmental influences such as DNA damaging agents. A variety of checkpoints exist controlling that critical processes have been completed before cell cycle progression is allowed to proceed. Examples are the G2/M checkpoint, which ensures that M phase is only entered once replication has been completed\cite{471}, and the spindle assembly checkpoint, which does prevent mitosis until the mitotic spindle has been assembled and all chromosomes are properly attached\cite{472}(see 1.3.1). The DNA damage checkpoint is used as a surveillance system of the integrity of the genome. Activated upon detection of DNA damage, it coordinates a variety of cellular responses, most notably arrest of cell cycle progression. Depending on when activated, cell cycle progression is halted (G1, G2 and M phase) or slowed down (S phase) to give the cell time to repair the damage before attempting to continue with the cell cycle. DNA damage checkpoints occur at different cell cycle states: at the G1/S transition (G1/S checkpoint), which prevents the commencement of DNA replication when DNA has been damaged\cite{473, 474}, during S phase (intra-S checkpoint), which slows down S phase progression and promotes alternative replication mechanisms such as TLS\cite{475}, and at the metaphase/anaphase transition in M phase (G2/M checkpoint), which prevents division of damaged chromatids in the budding yeast \textit{S. cerevisiae}\cite{476}. DNA damage checkpoints have been highly conserved through eukaryotic evolution and much of the mechanism of action was identified in the budding and fission yeasts. DNA damage checkpoints work as signal transduction cascades with signal amplification along the cascade. At the beginning of the cascades, sensor proteins such as the apical kinases Mec1 (ATR in humans) and Tel1 (ATM) generate a signal to so-called adaptor proteins such as Rad9 by means of phosphorylation\cite{477}. These in turn propagate the signal to transducers, such as the checkpoint kinases Rad53 (Chk2) and Chk1 (Chk1), which further amplify the signal and activate effector proteins, most of which are
still unknown. These effector proteins are responsible for producing the variety of responses to checkpoint activation such as cell cycle arrest (Fig. 1.17). Checkpoints like these described budding yeast mechanisms also exist in mammalian cells though differences do exist (for a review see [478–480]). Key downstream targets of the checkpoint response in mammalian cells include p21 to inhibit CDKs to prevent cell cycle progression and p53 to induce apoptosis in cases when repair is unsuccessful[481]. This demonstrates the intricate interplay between DNA repair and the cell cycle: DNA repair processes can interfere with cell cycle progression, while in turn, the cell cycle may greatly influence the DNA repair pathway chosen to repair DNA damage.

### 1.1.3 Dividing up the genome: chromosome segregation

In M-phase of the cell cycle, chromosomes are distributed equally into two daughter cells: initially, the replicated chromosomes - each made up of two sister chromatids - condense and the so-called mitotic spindle begins to form(Fig. 1.18). This molecular machinery is
Genomic integrity and instability

based on a bipolar array of microtubules, a major component of the cell’s cytoskeleton, and microtubules projecting from the poles attach to the centromeres of the chromosomes (more specifically a complex protein structure that forms at the centrosome called the kinetochore), so that by the metaphase stage of M-phase each chromosome is attached to both poles with each pole contacting one of the two sister chromatids (bi-orientation)[483]. Microtubules emanating from the poles either attach to the cellular cortex (astral microtubules), a chromosome centromere (kinetochore microtubules) or to a microtubule of the opposite pole (interpolar microtubules) and together with microtubule-dependent motor proteins shape the spindle and govern the positioning of chromosomes. The molecular forces generated by microtubules and the motor proteins work in such a way that chromosomes are aligned at the equator of the spindle, midway between the two poles and tension builds with both poles pulling the still-attached chromatids towards them, while the poles push each other apart. Once this set-up has been satisfactorily achieved, the spindle assembly checkpoint (SAC) - which senses either unattached chromosomes, the tension at kinetochores when bi-orientation is achieved or both[472] - ceases its inhibition of the APC/C which in turn removes inhibition of the separase enzyme which severs the cohesin ties holding the sister chromatids together. The sudden loss of sister-chromatid cohesion leads to chromosome segregation where the chromatids rapidly move towards their respective poles and away from one another. This physical separation of chromosomal DNA into two virtually identical sets allows subsequent cytokinesis, the division of the cytoplasm.

1.2 Genome variation

As efficient DNA replication and repair are at keeping genomic information intact, genome variations do occur frequently within cells affecting the cell and potentially the whole organism. Reviewed here is a selection of the most common types of variations that can and do occur with examples of the consequences of such changes to a genome.

1.2.1 Large-scale genomic variation

Large scale genome variations are any that affect more than a few dozen basepairs on the DNA. They come in a variety of types and as a consequence, with a variety of effects on the cell and/or the organism.
1.2 Genome variation

Figure 1.18: The mitotic spindle

The mitotic microtubule (MT) spindle assembles to separate the chromosomes. Spindle architecture depends on molecular kinesin and dynein motors. A - minus-directed motors can slide MTs poleward and contribute to MT-clustering at spindle poles. B - Kinesin-5 motors can slide antiparallel overlapping MT and thereby push the poles apart. C - Kinesin-13 can depolymerise MTs and contribute to spindle length control. D/E/F - MT can be nucleated in three different ways: at the centrosomes, near chromatin or by branching off existing MTs. Reproduced from [484] with permission from the publisher.
Whole-genome, segmental and gene duplications

One of the most drastic changes in DNA content are whole-genome duplications, but segmental and gene duplications can also have significant effects on the cell or organism. While all three processes are duplication events, they differ in scale and the way they arise. However, all of them are considered important in the evolution of new genes: small or large-scale duplication events all result in pairs of similar or identical genes, which then can be lost, shuffled, rearranged and/or adapted to new functions[485]. Additionally, duplications - especially larger segmental duplications - can generate large regions of sequence homology which can further lead to chromosomal rearrangements like inversions and translocations between chromosomes[486].

Whole-genome duplication

Whole-genome duplications are usually the result of non-disjunction during meiosis - when chromosomes do not appropriately separate and a cell ends up with both copies of the genome after replication - or the skipping of a division. Whole-genome duplication has been common in plants, but it has also occurred in the evolution of animals[485]. However, only about 50 known vertebrate species are considered polyploid having retained most or all of their duplicated genome, such as salmonid fishes and certain frogs, most famously the African *Xenopus laevis*[487]. Pinpointing whole-genome duplication events in evolution is not trivial, but two rounds are assumed to have occurred in the vertebrate lineage to humans[488], while another is estimated to have occurred 110 million years ago in the branch that gave rise to all teleost fishes[485]. While receiving another complement of the genome might seem initially harmless, it has important consequences for evolution. On the one hand it can be an important factor in speciation - inbreeding between closely related organisms of different ploidy is not straightforward, for example diploid and tetraploid parents will produce triploid offspring, which poses problems during segregation in mitosis - and on the other hand the extra genetic material allows for drastic evolutionary changes. Much of the extra material may be lost due to fractionation, but retention of genes can allow adaptive innovation, such as the array of Hox genes critical for embryonic development. A famous, albeit extreme example of *de facto* whole-genome duplications common in insects are polytene chromosomes which are generated by many rounds of replication without subsequent division. This generates giant chromosomes whose many chromatids remain fused together, such as the silk glands of the commercial silkworm *Bombyx mori* whose silk-producing cells are effectively hecatommyria-ploid after roughly 17 or 18 whole-genome duplications[489], which is thought to allow the silkworm to produce $10^{15}$ molecules of silk fibroin in just 4 days[490].
Segmental duplication  Segmental duplications are large, nearly identical duplications of genomic DNA that can range in size from only 1kb to more than 200kb[486]. As opposed to whole-genome duplication, segmental duplications do not commonly arise from non-disjunction events but rather from duplicative transpositions of small portions of DNA (see [486] for a review of possible mechanisms). Evolutionary recent segmental duplications have been identified in humans, showing non-random distributions of such events, with many genes duplicated incompletely or in such a way that give rise to chimeric proteins[491], which has given rise to the suggestion that segmental duplications may play an important part in exon/domain shuffling, a process critical in generating the degree of protein diversity we can observe today (see 1.2.1.4).

Gene duplication  While DNA duplication was initially thought to be a rare event, since only about 1% of human genes have no similarity with the genes of other animals and only 0.4% of mouse genes have no human homolog, it has been proposed that in fact not many sequence changes are needed to evolve a new function[492], raising the estimates of how common these events are. Current estimates suggest that - by whichever mechanism - gene duplications arise at quite a high rate (approximately 0.01 events per gene per million years)[493]. Once a gene has been duplicated it has been thought that due to the functional redundancy one copy can evolve a new function free from selective pressure, while the second copy will retain the original function[485, 492]. The more likely outcome of a duplication is that one copy becomes inactive in a process known as non-functionalization[485] due to the accumulation of evolutionary neutral, loss-of-function mutations[494]. Even though it has been the subject of evolutionary models since 1970[495–497], classical rare neo-functionalisation co-occurring with common loss of non-functional copies, does not account for the large number of duplicated genes that seem to be retained in genomes[485]. The recent duplication–degeneration–complementation (DDC) model by Force and colleagues has suggested another fate for duplicated genes[498, 499](Fig: 1.19). They stipulate that rather than only one gene accumulating mutations, while the other is kept under selection, likely both genes will accumulate loss-of-function mutations in independent sub-functions causing the partition of the ancestral functions between them, rather than the evolution of an entirely new function[498]. This model predicts that duplicated genes lose their degree of pleiotropy by splitting functions between them, which changes the selection pressure on them and allows evolution of a more specialized gene function in a process termed sub-functionalisation (for more information the reader is referred to [485]). Prime candidates for DDC have been characterised in the plant Arabidopsis thaliana: the APETALA1 (AP1), CAULIFLOWER (CAL) and FRUITFULL (FUL)
The duplication-degeneration (DDC) model relies on complementary degenerative changes in two duplicated genes in a way that the two together retain the original function. The coloured boxes represent cis regulatory elements, but mutations in other functional elements such as a protein domain or splice site is possible. Reproduced from [485] with permission from the publisher.

The three genes are all transcriptional regulators with roles in flower meristem specification and their similar sequences and locations within regions of conserved synteny (in the case of AP1 and CAL) makes them good duplicated gene candidates. Their support for the DDC model comes from their mutant phenotypes: double mutants have a markedly synergistic phenotype that is not seen in single mutants and the triple mutant fails to generate any flowering organs at all, showing that these genes share a high level of partial functional redundancy which can explain why all three are still retained in the genome. In conclusion gene duplication is a key source of new gene evolution, but to what degree these are real new functions of just sub-functionalisation remains under investigation[492].

1.2.1.2 Aneuploidy

Aneuploidy was first observed by Theodor Boveri, who also significantly speculated about the relationship between this type of genome aberration and malignancy[500–506]. As opposed
1.2 Genome variation

<table>
<thead>
<tr>
<th>Gestation (weeks)</th>
<th>Sperm</th>
<th>Oocytes</th>
<th>Pre-implantation embryos</th>
<th>Pre-clinical abortions</th>
<th>Spontaneous abortions</th>
<th>Stillbirths</th>
<th>Livebirths</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6-8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.4: Incidence of aneuploidy during development
Reproduced from [508] with permission from the publisher.

to polyploidy - an addition of a whole set of chromosomes (see 1.2.1.1) - aneuploidy involves an abnormal number of chromosomes in a cell[483], where the aneuploid set differs from the commonly observed wild-type set by only a few chromosomes[507]. Similar to whole-genome duplication, chromosomes can be lost or gained due to non-disjunction - the failure of chromosomes to separate correctly during cell division. Generally speaking, aneuploidy is much more detrimental than whole-genome duplications as the relative gene doses changes[507, 508] and aneuploidy is generally inviable. This type of genome aberration is relatively rare: in the yeast *S. cerevisiae* 99.25% of meiosis I and 96% of meiosis II occur without aneuploidy[509], in the fruit fly *Drosophila melanogaster* non-disjunction of chromosome X occurs in only ~0.02-0.06% of cases and in mice aneuploidy in fertilised eggs does not exceed 1-2%, meaning that non-disjunction can be as rare as 1 in 10,000 cases[508]. Intriguingly, in humans meiotic non-disjunction is more common, with an estimated 10-30% of fertilised human eggs being aneuploid[508], and the leading cause of pregnancy loss(Fig. 1.4). Additionally, chromosomal abnormalities occur in approximately 1 out of 160 live births in humans[510], making it also the leading cause of genetic disability and mental retardation[508].

**Nullisomy** The loss of the entire chromosome pair in a diploid (or all four in a tetraploid etc.) is known as nullisomy. In most species, any kind of nullisomy is lethal to the cell and/or organism, because a significant amount of genetic information is lost[507]. A few exceptions are known in plants, where *de facto* polyploids behave as diploids during mitosis. The bread wheat *Triticum aestivum* accounts for over 95% of wheat grown worldwide and is an allohexaploid species[511], which is a type of polyploidy where the chromosome sets derive from different species in this case likely due to multiple rounds of hybrid speciation[512]. *T. aestivum* contains three of the five known genomes in *Triticum* and contains three homeologous diploid sets of seven chromosomes[511]. Genetically, it behaves like a diploid[513], due to the Ph1 locus which reduces centromere associations between the different sets of chromosomes[514], meaning that during mitosis the two homologous chromosomes derived from the same genome pair up. *T. aestivum* can tolerate the loss of a pair of chromosomes from
one genome, since it contains two, not identical but homeologous, additional chromosome pairs which can compensate to allow survival. In fact, all possible bread wheat nullisomics have been generated[515–517] and while they show differences in growth and appearance, they are all viable and fertile[517].

**Monosomy**  Monosomy, carrying only one copy of a chromosome, is detrimental for two main reasons[507]: it results in differences in gene dosage, which perturb cellular functions and genes on the remaining chromosome are now hemizygous and normally recessive, deleterious mutations are now phenotypically visible. While all autosomal monosomics in humans are lethal, Turner’s syndrome - the loss of one X chromosome while retaining all 44 autosomes - is seen in 1 in 5000 female births[507]. The phenotype is relatively mild with sterility, short stature and a near normal intelligence (some specific cognitive shortcomings do occur) possibly due to the fact that in females who are diploid for the X chromosome, one of the two chromosomes is randomly inactivated in every cell.

**Disomy**  While disomy is the normal condition for diploid organisms, it is a type of aneuploidy for tetraploid organisms such as *Xenopus leavis*. A marginal case of disomy in humans is uniparental disomy (UPD), whereby offspring inherit both members of a chromosome pair from one parent and none from the other[518]. This can occur as either heterodisomy, where offspring receives both or parts of both homologs from the parent, or isodisomy, where only one or sequences of only one homolog are present(Fig. 1.20). Isodisomy is potentially harmful, because like monosomy, it allows mutations that a parent carries heterozygously to be expressed phenotypically (reminiscent of loss of heterozygosity in cancer)[518]. It has been demonstrated to be the cause for cases of cystic fibrosis[519, 520], Hemophilia A, Duchenne muscular dystrophy and Osteogenesis imperfecta[521]. In contrast, heterodisomy is not expected to be deleterious except in cases where genes concerned are subject to genomic imprinting[522], the epigenetic process in which genes are expressed depending on the parent who transmitted it. For instance, if a maternal copy of a gene is subject to imprinting, it will be silenced in the offspring and only the paternal copy will be expressed. If such a gene was affected by UPD the offspring would be phenotypically null for this gene despite carrying intact copies. Imprinted genes have been identified in plants, fungi and animals with roughly 150 known in mice and about half of that in humans[523]. The first demonstration of heterodisomy causing a defect was in a case of nondeletion Prader–Willi syndrome[522].
1.2 Genome variation

Figure 1.20: Uniparental Disomy - A special case of aneuploidy
An example of uniparental disomy: a non-disjunction event in maternal meiosis leads to the transmission of both copies of a particular chromosome pair to the gamete, which is then fertilised by a spermatocyte that is nullisomic for the pair in question. Due to meiotic recombination regions of heterodisomy and isodisomy are found across the chromosome. Homozygosity due to isodisomy is denoted by the asterisk and the solid bar represents an imprinted gene that - even though heterozygous and not detrimental in the mother - results in two inactive copies in the zygote. Reproduced from [518] with permission from the publisher.

Trisomy Trisomy is another condition of chromosomal imbalance which often causes abnormality and death in diploid organisms. While most human trisomies are fatal[508], extra copies of chromosomes 21 (1 in 800 births), 18 (1 in 6000 births) and 13 (1 in 10,000) account for the vast majority of viable autosomal trisomies, trisomies in sex chromosomes are also observed such as in Klinefelter syndrome (XYY, about 1 in 1000 male births)[508, 510](Fig. 1.5). Trisomy 21 (Down syndrome) is by far the most common viable human aneuploidy with affected individuals leading relatively long lives and its likelihood has been linked to maternal age[512]. Trisomy 13 (Patau syndrome) and trisomy 18 (Edwards syndrome), albeit viable, confer very low life expectancy (less than 10% of those affected reach 1 year of age)[510].

Somatic aneuploidy While the above are aneuploidies arising in meiosis and affect the entire organism, aneuploidy can also arise spontaneously in somatic cells giving rise to chromosomal mosaicism, the presence of two or more populations of cells with different genotypes. Mosaicism in humans exists in virtually every person as a consequence of the non-zero error rate of genome replication and repair. However, generally mosaicism refers to more substantive changes in the organism such as somatic aneuploidy. While general mosaicism is observed throughout an organism[524], confined mosaicism is only found in a certain area such as the brain[525]. Usually, the time in development when the mitotic event giving rise
Table 1.5: The origin of human trisomy
Reproduced from [508] with permission from the publisher.

To the mosaicism occurred determines whether the mosaicism is general or confined[526], with general mosaicism only occurring if the event occurred in the first few days of embryonic development[526]. At this stage, mosaicism can affect around 70% of all cells in the embryo[524]. However, euploid cells (those with a full complement of chromosomes) tend to divide more efficiently than aneuploid ones and thus their contribution to the organism can reduce over time, with initial general mosaicism becoming confined during development[526]. The best studied type of confined mosaicism is confined placental mosaicism which has been linked to many pregnancy complications such as intrauterine growth retardation, spontaneous abortion and stillbirth[526]. Additionally, aneuploidy has been found in nearly all major human tumor types[527], often reflecting the loss of a tumor suppressor gene or in other cases duplication of a gene that promotes tumor progression such as c-Met in renal carcinoma[528] (see 1.4.2). In general, clinical consequences of mosaicism can vary depending on which chromosomes are involved, the tissues affected and the extent of the mosaicism[526].

1.2.1.3 Chromosomal translocation and chromoanagenesis

Chromosomal translocation Chromosome translocations were first identified in cancers: Nowell and Hungerford in 1961 showed a "minute chromosome" that replaced one of the four smallest autosomes in chronic myeloid leukaemia (CML) cells[529], which in the early
Figure 1.21: Consequences of chromosomal translocations
Chromosomal translocations can result in the placement of genes near different regulatory elements (A) or in aberrant gene fusions (B). Reproduced from [534] with permission from the publisher.
Genomic integrity and instability

Chromoanagenesis  Next-generation sequencing has recently led to the identification of a phenomenon termed chromoanagenesis, where hundreds of genomic rearrangements occur in a limited genomic region[535]. Different types of these events have been identified among them chromothripsis, or chromosome shattering, and chromoplexy[536]. The mechanism by which such catastrophic events occur remains elusive, but several models exist including the micronuclei model, which stipulates that a mitotic chromosome segregation error can lead to the formation of a micronuclei containing whole or fragments of chromosomes explaining why chromothripsis is extensive in a confined region of the genome[535]. Aberrant replication, DNA repair and checkpoint activity in micronuclei are thought to lead to the shattering of the DNA. These fragments can then be re-ligated and re-incorporated into the cell’s nucleus(Fig. 1.22). Chromoplexy, a related but distinct process, in which DNA from one or more chromosomes becomes scrambled, differs from chromothripsis in the number of breakpoints (tens rather than hundreds) and their location (unclustered and located on multiple chromosomes rather than the confined locations in chromothripsis; Fig. 1.23)[536]. Additionally, chromothripsis is suspected to occur in one cataclysmic event, whereas chromoplexy can occur in sequential events as detected in heterogenous prostate cancer samples. While current data suggests chromothripsis to be relatively rare, chromoplexy has been identified in many prostate cancer samples[536].

Trinucleotide repeat expansion  A large fraction of a given genome, ~50% in case of humans, can be made up of repetitive sequences, the simplest of which are tandem microsatellite repeats of 1-6bp, which can be present with a few hundreds of copies to thousands. It has been known for roughly 25 years that expansion of these sequences can have severe consequences, though the mechanism of how these repeat expansions occur remains elusive. However, the propensity of these DNA stretches to form unusual secondary structures such as hairpins,triplexes, tetrplexes and slipped-strand structures has been linked to increased instability of these sequences and subsequent expansion during replication and repair[537]. To date more than 20 human syndromes, most notably Huntington’s disease, as well as many pathologies in animals and plants, are known to be attributable to repeat expansion[537]. The number of expanded repeats has been linked to the disease’s severity, onset and progression[538, 539].

Other large scale rearrangements  There are other kinds of large-scale genomic rearrangements that arise by similar mechanism to chromosomal translocations - initiated by double strand breaks followed by aberrant recombination - and can have similar consequences depending on the exact circumstances and the genomic regions involved. These include chro-
Figure 1.22: Chromothripsis
Chromothripsis is the shattering of one or more chromosomes, leading to the simultaneous generation of many double strand breaks, most of which are repaired by NHEJ in a manner leading to chromoanagenesis: the generation of a highly rearranged chromosome. Broken DNA fragments can also circularise to generate double minute chromosomes which are often amplified. Reproduced from [535] with permission from the publisher.
Genomic integrity and instability

Figure 1.23: Chromoplexy and Chromothripsis
Schematic representations of genomic rearrangements found in Chromoplexy (top) and Chromothripsis (bottom). Reproduced from [536] with the permission of the publisher.

mosomal inversions and interstitial insertions/deletions, the former of which can be generally harmless unless critical genomic regions such as genes or genes and their regulatory elements are interrupted and the latter of which can be deleterious depending on the DNA lost or gained and whether breakpoints generate aberrant products.

1.2.1.4 Mobile elements
In the genome, there are DNA sequences, termed mobile elements, that can move around, change their number or location and often affect the activity of close genes. A prominent type of mobile elements are transposons, which can change their position within the genome[540]. There are two distinct groups of transposons: retrotransposons (Class I) and DNA transposons (Class II). They differ in their mechanism of transposition, the former of which is often referred to as "copy and paste" and the latter as "cut and paste"[541](Fig. 1.24). While the vast majority of transposons appears to be epigenetically silenced to prevent their expansion[542], transposition of transposons can greatly affect the sequence they relocate to, depending mostly on where they insert: for example they can disrupt genes causing "knock-out mutations"[543, 544] or they can, if they do not excise perfectly, bring some genomic sequences with them greatly driving evolution in a process called exon shuffling[545].

1.2.1.5 Exon/domain shuffling
In 1978, Gilbert first speculated about the evolutionary utility of splicing: a single base change could change more than just one amino acid in a protein - it could could change splicing pat-
1.2 Genome variation

There are two types of transposable elements: retrotansposons (Class I) and DNA transposons (Class II). Class I move via an RNA intermediate (“copy and paste”), while the latter excise themselves from the DNA (“cut and paste”).

Reproduced from [548] with permission of the publisher.

terns and generate an entirely new protein[546]. Suggesting that splicing changes need not be 100% efficient, his hypothesis allowed for new gene functions without gene duplication and, going even further, suggested that, if exons correspond to protein functions, recombination in intron sequences could allow for independent rearrangement of these functions using repetitive intron sequences as recombination "hotspots". This mechanism is known as exon shuffling and can occur by two known mechanisms: illegitimate recombination, since recombination between non-homologous genes is more likely in intronic regions, repeats and transposon sequences[547](see 1.2.1.3), and retroposed exon insertion[492]. This mechanism was likely only significant after the evolution of spliceosomal introns (self-splicing introns are not as tolerant to recombination)[131, 132] and in the evolution of higher eukaryotes exon shuffling has been suggested as a common phenomenon[492]. Many proteins - especially those in metazoans - are modular in structure and particular domains contribute different aspects to the overall function of a protein. These are called mosaic proteins and many of the protein domains involved are mobile and found in many otherwise unrelated proteins suggesting they were subject to exon shuffling[116, 118]. While it has been observed in nematodes, hydrozoa and molluscs, it is especially common in metazoans and its increase likely coincided with the time of metazoan radiation[132]. Thus, intriguingly, it has been highly active at the time when many complex multicellular organisms evolved and, notably, most mosaic proteins, assumed to be the result of exon shuffling, are extracellular and involved in multicellularity[131, 132].
An analysis of mosaic proteins has revealed that there is a strong correlation between domain organization and intron-exon structure[549]. This gave rise to the “modularization hypothesis” which suggests that introns behave as "mobile genetic elements and transpose to other heterologous sites in the genome"[549–551]. This means that a protein domain can acquire mobility if introns of identical phase insert themselves on either side of the domain encoding sequence. Such a construct is called a "proto-module", which may then undergo tandem duplication and insert itself into other proteins to generate mosaic proteins[550]. Not every exon is an efficient contributor to exon shuffling due to splice-frame rules[552]. Exons will need to be in the same phase as its new neighbours to not cause a frameshift upon insertion and the flanking introns need to be of the same phase and many of the documented mosaic proteins are constructed from these so-called symmetrical exons[552]. There are four different types of introns: introns in UTRs, phase 0 introns, phase 1 introns and phase 2 introns[550, 552, 553]. Phase 0 introns lie between two codons, phase 1 introns lie between the first and second nucleotide of a codon and phase 2 introns lie between the second and third nucleotide of a codon[553]. Based on its flanking introns, exons can be classified into 9 classes: three symmetric exons (1-1, 2-2 and 0-0) and 6 asymmetric ones (0-1, 0-2, 1-0, 1-2, 2-0, and 2-1)[553]. Symmetric exons or a symmetric exon set (made by combining asymmetric exons in such a way that restores symmetry) are the only ones that can be inserted into an intron of the same phase without changing the reading frame[549]. That is why it is not surprising that most of the protein domains known to be mobile are encoded by symmetric exons or symmetric sets of exons and most modules are class 1-1, though why they are more common than modules of class 0-0 and class 2-2 is unclear[552].

A striking example of exon shuffling can be found in the group of hemostatic proteases that are involved in the blood clotting cascade. In this cascade inactive proteins are activated by proteolytic cleavage, which in turn allows the now activated protein to cleave another leading eventually to a stable fibrin clot(Fig. 1.25). All the hemostatic proteases involved have large extensions N-terminal to their serine protease domains, which include a number of discrete domains involved in functions such as substrate recognition[549]. These N-terminal domains include some that are also found in other, unrelated proteins as for example fibronectin. The strong correlation between exons and domains in these proteins combined with the fact that most exons are 1-1 symmetric exons, is highly suggestive of these proteins arising from exon shuffling. Recently, exon shuffling has been "re-created" in vitro making it interesting for pharmaceutical protein development[549].
1.2 Genome variation

Figure 1.25: Blood clotting cascade
Schematic representation of the blood-clotting cascade. Many of the involved factors are serine proteases which - by cleaving - activate another downstream serine protease. The amplification inherent in signal transduction cascades allows a small stimulus to generate a stable fibrin clot. It is thought that many of these proteases are the result of exon shuffling. XIII - Fibrin stabilising factor (transglutaminase), XII - Hageman factor (serine protease), XI - Plasma thromboplastin (serine protease), IX - Christmas factor (serine protease), VII - Stable factor (serine protease), PL - platelet membrane phospholipid, Calcium - Calcium ions, TF - Tissue factor.
1.2.1.6 Acquisition of foreign DNA

The acquisition of foreign DNA - or horizontal gene transfer - is a common process observed in prokaryotes and to a limited degree in plants. In bacteria we distinguish between transformation, transduction and conjugation. Transformation is the uptake of DNA directly from the environment and a natural process in some species of bacteria, but it can also be brought about by artificial means[554]. DNA from dead organisms is abundant in the environment and some species like Neisseria gonorrhoeae actively secrete DNA into the environment, where it can be taken up by other bacteria to spread useful genes[554]. More efficiently bacteria can share DNA directly in a process called conjugation, which involves cell to cell contact to share DNA, most commonly a plasmid or transposon[555]. Alternatively, bacteria can receive DNA from another bacteria via bacteriophage in a process known as transduction[554]. There have also been multiple examples of horizontal gene transfer in plants such as the transfer of chloroplast or mitochondrial DNA. However, evidence for gene transfer from bacteria to the nuclei of multi-cellular plants is rare[556]. It has, however, been described for Agrobacterium rhizogenes and the related bacterium A. tumefaciens, which can transfer DNA, called T-DNA, to the host genome that integrates into the genome via non-homologous recombination[556]. T-DNA sequences have been found in different plant species[556], including cultivated sweet potato plants[557]. Whether horizontal gene transfer in metazoans occurs is a matter debate - detection of Y chromosomes in human females is likely persistence of foreign cells rather than uptake of foreign DNA by the host[558, 559] -, recent genome sequence analysis studies provide some limited evidence that horizontal gene transfer from bacteria and viruses may have taken place in animals throughout evolution[560].

1.2.2 Small-scale mutations

While small types of variants are not visible using techniques such as fluorescence in situ hybridization (chromosome painting), they are no less significant and the effects they can have on an organism can be equally favourable or detrimental.

1.2.2.1 Point mutation instability (PIN)

Point mutations are single base substitutions and can be subdivided into transitions or transversions depending on the type of observed change[561, 562]. A transversion is a mutation changing a purine to a pyrimidine or vice versa, for instance a T to A or a T to G mutation, while in a transition a purine is replaced by another purine (for example a G to A mutation) or a pyrimidine is replaced with another pyrimidine (such as a C to T mutation) (Fig. 1.26).
Even though there are twice as many ways to achieve a transversion, transitions are much more common in most cases studied likely due to spontaneous, transient tautomeric shifts in DNA bases, which can result in altered bonding preferences. For instance, while the amino form of adenine pairs with thymine, the tautomeric imino form pairs with cytosine, which can cause a T to C transition. When a point mutation falls into coding regions of the genome it can also be classified by its functional consequence (note that mutations in regulatory sequences can also show effects, but their prediction and subsequent classification is more challenging).

Since genes code for proteins and proteins are chains of amino acid residues, the DNA sequence of the gene codes for the sequence of amino acids. Since four nucleotides cannot code for 21 amino acids, more than one DNA base at a time codes for an amino acid. In fact, triplets of DNA bases are used to signal the start, the end of a gene and the sequence of amino acids in between. A nonsense mutation is one that changes a triplet in such a way that it no longer codes for an amino acid, but signals the end of the protein and often causes a truncated protein or one that will be expressed at very low levels due to the action of the nonsense-mediated decay pathway, which is why nonsense mutations can be quite detrimental. Missense or non-synonymous mutations are those that change an amino acid in the resulting protein. The severity of such mutations is variable, dependent on how chemically similar the two amino acids are, and how critical the amino acid is for protein function. A single amino acid change could change the function, localisation, activity or stability of the protein. Lastly, synonymous or silent mutations are those that while changing a DNA triplet do not change the amino acid that will be inserted. This is due to redundancy within the triplet code: some amino acids are coded for by more than one triplet. The consequences of mutations in non-coding regions are less clear. While they are largely considered to be silent, they may affect regulatory regions for genes (such as promoters and enhancers), alter splicing patterns if they fall close to intron/exon boundaries or affect other genomic features such as miRNAs.

There are a myriad of examples of the effects of a single point mutation on cells or organisms. One example that shows the detrimental effects that point mutations can have in humans is heterozygous missense mutations in the \textit{FBN1} gene causing Marfan syndrome, an autosomal dominant disease affecting the connective tissue. Fibrillin-1, encoded by \textit{FBN1}, is an extracellular protein and a major component of 10-12 nm microfibrils of connective tissue, which have important structural properties as well as acting as a sequester for the growth hormone TGF\(\beta\). Point mutations in \textit{FBN1} are likely to cause a misshapen protein that is non-the-less incorporated into the connective tissue. Patients present with a variety of severe phenotypes: excessively tall stature, other skeletal abnormalities (such as arachnodactyly and
scoliosis), ectopia lentis and severe cardiovascular abnormalities (often mitral valve disease and progressive aortic root dilation leading to aortic dissection followed by aortic rupture with sudden death)[567].

An example of a point mutation that has benefited humans immensely comes from the world of plants: south-west Asia is generally considered the cradle of agriculture and many of the early cultivated plants such as barley were selected for their ability to flower in spring, when farmers could take advantage of abundant water from snowmelt, and be harvested in early summer, before drought would decimate the crop[568]. While perfect for the habitat, these plants were difficult to cultivate in higher latitudes where temperatures, day lengths and water availability was drastically different. A single point mutation in the gene Ppd-H1 causing a Gly-to-Trp change was shown to affect its flowering time, allowing the spread of this crop into Europe where it can be planted in the spring (to avoid injuries by frost) and be harvested in the autumn, taking advantage of the long moist summer[569]. Single point mutations such as this have likely had a significant impact on the spread and lifestyle of humans and their effect is not to be underestimated.

1.2.2.2 Small insertions/deletions (INDELs)

INDELs are a catch-all term for insertions and deletions[570] and they can vary in size from deletion or insertions of single nucleotides to many thousands. The boundary between a small INDEL and a large interstitial deletion is not very well defined. The consequences of small INDELs (less than 50bp) can be similar to those of point mutations. If located in non-coding regions they can disrupt essential features of the DNA sequence or have no discernible effect. Should they fall into coding regions they can affect the proteins to varying degrees. Inframe deletions or insertions (meaning the net nucleotide change is a multiple of three) can affect the protein function, but frameshift INDELs (those that cause a net nucleotide change that is not divisible by three) usually lead to a premature stop codon and the consequence is akin to that

Figure 1.26: Transitions and Transversions
1.3 Causes of mutations

Figure 1.27: Codon table
A table showing the relationship between an RNA triplet codon and the matched amino acid.

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>UUU = phe</td>
<td>UCU = ser</td>
<td>UAU = tyr</td>
<td>UGU = cys</td>
</tr>
<tr>
<td></td>
<td>UUC = phe</td>
<td>UCC = ser</td>
<td>UAC = tyr</td>
<td>UGC = cys</td>
</tr>
<tr>
<td></td>
<td>UUA = leu</td>
<td>UCA = ser</td>
<td>UAA = stop</td>
<td>UGA = stop</td>
</tr>
<tr>
<td></td>
<td>UUG = leu</td>
<td>UCG = ser</td>
<td>UAG = stop</td>
<td>UGG = trp</td>
</tr>
<tr>
<td>C</td>
<td>CUU = leu</td>
<td>CCU = pro</td>
<td>CAU = his</td>
<td>CGU = arg</td>
</tr>
<tr>
<td></td>
<td>CUC = leu</td>
<td>CCC = pro</td>
<td>CAC = his</td>
<td>CGC = arg</td>
</tr>
<tr>
<td></td>
<td>CUA = leu</td>
<td>CCA = pro</td>
<td>CAA = gln</td>
<td>CGA = arg</td>
</tr>
<tr>
<td></td>
<td>CUG = leu</td>
<td>CCG = pro</td>
<td>CAG = gln</td>
<td>CGG = arg</td>
</tr>
<tr>
<td>A</td>
<td>AUU = ile</td>
<td>ACU = thr</td>
<td>AAU = asn</td>
<td>AGU = ser</td>
</tr>
<tr>
<td></td>
<td>AUC = ile</td>
<td>ACC = thr</td>
<td>AAC = asn</td>
<td>AGC = ser</td>
</tr>
<tr>
<td></td>
<td>AUA = ile</td>
<td>ACA = thr</td>
<td>AAA = lys</td>
<td>AGA = arg</td>
</tr>
<tr>
<td></td>
<td>AUG = met</td>
<td>ACG = thr</td>
<td>AAG = lys</td>
<td>AGG = arg</td>
</tr>
<tr>
<td>G</td>
<td>GUU = val</td>
<td>GCU = ala</td>
<td>GAU = asp</td>
<td>GGU = gly</td>
</tr>
<tr>
<td></td>
<td>GUC = val</td>
<td>GCC = ala</td>
<td>GAC = asp</td>
<td>GGC = gly</td>
</tr>
<tr>
<td></td>
<td>GUA = val</td>
<td>GCA = ala</td>
<td>GAA = glu</td>
<td>GGA = gly</td>
</tr>
<tr>
<td></td>
<td>GUG = val</td>
<td>GCG = ala</td>
<td>GAG = glu</td>
<td>GGG = gly</td>
</tr>
</tbody>
</table>

of a nonsense mutation.

1.3 Causes of mutations

Variation arises continuously in biological systems, by sexual recombination, from one cell division to the next or in an instant. Mutation can be the consequence of internal processes of the cell or due to extrinsic influences. DNA damage repair plays a significant role in preventing and creating mutations and has been touched on before (see 1.1.2). Different examples of both types of causes will be mentioned here, though the list is by far not exhaustive and many aspects of mutagenesis from the identity of mutagens, to their mode of action and the extent of their effect are far from elucidated.

1.3.1 Endogenous causes of mutation

Mutations due to endogenous causes can arise in a multitude of ways: due to the intrinsic error-rate of DNA replication, errors in mitosis, failed or defective DNA repair, exposure to endogenous mutagens or enzymatic modification of DNA. While defects in DNA replication and/or repair would probably affect the integrity of genomic information (discussed later), the most common source of mutations in organisms with intact replication and repair machineries are assaults on DNA, which are mostly, but not always, repaired[571]. While some mutations
Figure 1.28: Replication slippage

Replication slippage involves the denaturation of the nascent strand from the template followed by missalignment during rehybridization. This leads to the new strand having a different length than the template and is especially common in repetitive regions of the genome. Reproduced from [576] with permission from the publisher.

are due to exogenous mutagens[572](see 1.3.2), a significant portion of DNA damage is due to mutagens which are generated by normal cellular processes[573, 574], and it is thought that it causes ~70,000 lesions and/or strand breaks per day per mammalian cell[575]. The best understood types of endogenous mutation causes include spontaneous reactions (mostly hydrolysis), chemicals generated during cellular metabolism (such as reactive oxygen species) and errors during cell division (including non-disjunction) and replication (due to polymerase infidelity).

**Replication and mitosis/meiosis error** A number of mutations arise during the cell cycle due to imperfections in the faithful replication and segregation of genomic material. Replication slippage is the best described mechanism for replication induced mutations: one DNA strand forms a little loop during replication which can result in the formation of small INDELs [577]. This is especially common in areas of repetitive sequences(Fig. 1.28). Other types of polymerase errors are discussed in 1.4. Errors in M-phase of the cell cycle can be often more severe, leading to gross chromosomal changes. During meiosis prophase I, many chromosomes recombine with their homologues forming crossovers which allow genetic exchanges between chromosomes during sexual reproduction and also acting as a critical tether of chromosomes during meiosis I, where many oocytes arrest for long periods (up to several decades in the case of humans). Crossovers were described by Thomas Hunt Morgan in his work on *Drosophila* genetics[578], and demonstrated by Harriet Creighton and Barbara McClintock in
1.3 Causes of mutations

Knowledge of their existence was extensively exploited to generate linkage maps to locate genes on chromosomes relative to one another. Crossovers usually exchange equal parts of the genome, however, sometimes homologous sequences are not paired precisely, especially when repetitive genomic regions such as transposons are involved due to their high similarity, which can result in unequal crossovers or chromosomal translocations (Fig. 1.29). Other gross abnormalities like aneuploidy and whole-genome duplication can be due to non-disjunction, the failure of homologous chromosomes or sister chromatids to separate properly in meiosis I, meiosis II or mitosis\[509\]. While the exact causes of non-disjunction are unclear, several mechanisms have been proposed and those that cause aneuploidy in female meiosis are of particular interest\[see 1.2.1.2\]. Of critical importance to all types of cell division is the spindle assembly checkpoint (SAC) which is critical to prevent cell division before all chromosomes are properly paired and attached to the spindle\[472\]. Only when this has happened will the SAC release its inhibition on the APC/C allowing cells to complete division, explaining how defects or errors in the proper function of the SAC can lead to aneuploidy. This, however, is not the only reason non-disjunction can occur and does not explain why it has been demonstrated to occur much more in female than in male meiosis and why fidelity of female meiosis seems to deteriorate with age (termed "Maternal Age Effect")\[580\]. Non-disjunction occurs more commonly in meiosis I than meiosis II and mitosis\[581\], due to the fact that here homologous chromosomes rather than sister chromatids are paired up and need to withstand the tensions of the spindle. The most favoured reason for the Maternal Age Effect is the prolonged arrest of oocytes in late stages of prophase I (in contrast to male gametes which proceed quickly through both meiosis I and II) which is thought to be vulnerable to deterioration of cohesion between the chromosomes and fluctuations in the activity of the SAC\[581–583\]. Cohesion along chromosome arms keeps paired homologs attached in meiosis I (and sister chromatid centromeres attached in meiosis II) and since experiments in mice have shown that cohesin is only deposited during S-phase before birth and cannot be replaced, cohesion proteins in humans have to endure some 40-50 years\[581\].

Smaller scale changes, mainly point mutations occur cell-cycle independently throughout a cell’s life and the affected DNA bases are often collateral damage of normal cellular metabolism. Other mutations can be attributed to the action of distinct DNA modifying enzymes.

**Depurination and depyrimidination** Hydrolysis is a common affliction of DNA and one of the most common types of hydrolysis is the cleavage of the N-glycosidic bond tethering the DNA base to the phosphor-backbone leading to an basic site. It is estimated that depriva-
Genomic integrity and instability

Figure 1.29: Unequal crossovers result in chromosome rearrangements
Crossovers sometimes occur between similar but not equivalent regions of the genome leading to an unequal exchange of DNA between chromosomes.

Crosstown occurs roughly 10,000 times per human cell per day[584], while depyrimidination is rarer with only 700 occurrences in a cell in the same timeframe[440]. Most of those are efficiently repaired by BER (see 1.1.2.2), but especially in S phase those lesions cause issues when replication forks are stalled due to the lacking genetic information[585]. In *S. cerevisiae* this type of lesion has been shown to be bypassed by Polδ and Polζ usually by inserting an adenine causing point mutations[586].

**Oxidative damage** Oxidative damage is a consequence of many metabolic processes of the cell, but can also be due to external mutagens such as air pollutants[587]. The majority of the estimated 12,000 lesions per cell per day in human cells[588] is due to reactive oxygen and nitrogen species, ROS and RNS, respectively[589]. RNS are oxides of nitrogen[590] and ROS include O₂-derived free radicals, compounds that easily convert to them or oxidizing agents[589] and they have been implicated in at least 25 distinct types of DNA lesions[591], including generation of abasic sites, DNA breaks and deamination[592, 593]. In spite of this plethora of damage, the mutagenic consequence of many ROS and RNS remains unclear, with only a few exact mechanisms having been elucidated such as the oxidation-damaged guanine variant 7,8-dihydro-8-oxoguanine (8-oxoG), which is more likely to pair to an adenine than to its usual partner cysteine[594].

**Deamination** Many DNA bases including cytosine, 5-methylcytosine, 5-hydroxymethylcytosine, guanine, and adenine can be spontaneously deaminated, with a variety of consequences for their hydrogen bonding preference and subsequent mutagenic potential. In human cells, around 500 times per day cytosine is deaminated and converted to uracil, which acts much
1.3 Causes of mutations

like a thymine (so much so that it is the base used in thymine’s place in RNA) due to its ability to hydrogen bond adenine[595]. Deamination of cytosine is catalysed by AID and the APOBEC family of enzymes, of which the former has a well-described role in the somatic hypermutation of immunoglobulins which greatly increases the variability of antibodies and the resilience of the immune system[596], while the latter are known to deaminate cytosine, but different members show different sequence context preferences and their role in cells is much less understood[597]. AID and APOBEC also deaminate 5-methylcytosine causing mutations[598]. DNA methylation is a widespread phenomenon and not in itself considered harmful. While N4- methylcytosine and N6-methyladenine are found almost exclusively in bacteria[599], 5-methylcytosine is the most common methylation observed in mammals[600] often followed by a guanine (CpG dinucleotide) except in embryonic stem cells where also non-CpG cysteines show a high degree of methylation[601]. This methylation is considered an epigentic mark which has been shown to be involved in many cellular functions such as regulation of gene expression, genetic imprinting and marking the template strand shortly after DNA replication (see 1.1.1.1)[523, 602]. However, 5-methylcytosine is very susceptible to deamination to a thymine which occurs ~1,500 times per human cell per day[603]. Other, less common forms of deamination that can occur are 5-hydroxymethylcytosine to 5-hydroxymethyluracil[604], adenine to hypoxanthine (which pairs preferentially with guanine) [584] and guanine to xanthine (which also pairs with cytosine and is thus not generally mutagenic, but rarely does pair with thymine)[605].

1.3.2 Exogenous causes of mutations

For most individuals endogenous mutagens are the main cause of mutations, however, significant contributions to mutation numbers can be made by exogenous mutagens if the individual is exposed to one. Most environmental mutagens have been identified due to their ability to cause cancer and more than 100 agents have been classified as "carcinogenic to humans" with an additional 300 and more with probable links to human cancer by the the International Agency for Research on Cancer (IARC), an arm of the World Health Organization[606]. These carcinogens can have genotoxic or non-genotoxic effects or both[607] and it is the estimated ~90% of mutagenic carcinogens we will consider here[608]. Included below is a selection of some of the most severe and well studied known genotoxic agents.

Tobacco Smoke, Coal and Soot  In 1930, it was first proposed that tobacco smoke could have a role in lung cancer, which was definitely confirmed in 1986[606] after decades of studies investigating lung cancer aetiology[609–611], including studies in which model organisms
who developed lung cancer after exposure to cigarette smoke[612]. While cigarette smoke contains more than sixty well-known carcinogens[613], it also contains benzo[a]pyrene, the first discovered chemical carcinogen[614]. Benzo[a]pyrene was first isolated by Alfred Winterstein in 1936 from coal tar[615] and when applied to mouse skin proved to be highly carcinogenic[616]. Coal tar and soot - the major exposures experienced by chimney sweeps - were the first occupational carcinogens identified[617, 618], which was confirmed when - after recommendation of daily baths - the incidence of scrotal cancer in this population was greatly reduced[619, 620]. After exposure, benzo[a]pyrene is quickly metabolised to the carcinogenic diol-epoxide 2[621], which is highly reactive and known to form bulky adducts on DNA with a high preference for guanines[621, 622].

Radiation: ionising radiation and ultraviolet-light  In physics, radiation means transmission of energy through time in space in the form of waves or particles and can include many types of radiation such as visible light, sound and radio waves. Often radiation is roughly separated into two categories: ionising (IR) and non-ionising (NIR), with the former having enough energy to displace an electron from an atom thus ionising it[623]. Both types of radiation can have genotoxic effects. Exposure to NIR can excite atoms - promoting an electron from ground state to a higher energy state - which among other things can lead to the generation of ROS[624]. IR is particularly damaging to cells because of its high energy and ability to ionise atoms[623], and includes α-particles, β-particles and γ-rays (as well as X-rays and the high energy end of UV light). All three types of IR have enough energy to break the DNA backbone, damage nucleotides or alter hydrogen bonds between bases[625]. Most importantly, IR generates double and many more single stranded breaks resulting in cell death if not repaired and often INDELs after successful repair[415]( see 1.1.2). Exposure to ionising radiation be it in the form of medical X-rays, exposure to radioactive material or cancer treatments can result in DNA damage and subsequent mutation[626]. UV light is positioned somewhere between the wavelength of IR and NIR and UV light can cause damage consistent with both types of radiation. Our sun emits UV-A, UV-B and UV-C light and of those all can reach the earth, though, all UV-C and most UV-B light is usually absorbed by the stratosphere and the ozone layer[627], meaning that ~95% of the UV light reaching the earth’s surface is UV-A and the rest UV-B light (with variation depending on the local depletion of the ozone layer). While UV-B can only penetrate the epidermis and reach the dermis layer of the skin[628], allowing it to cause skin reddening and sunburn, UV-A light can penetrate deeper into the skin reaching the subcutaneous layer and has been implicated in wrinkling and skin aging. Both types of UV light can be mutagenic and have been associated with cancer, but
the types of mutation resulting from UV exposure depend on the type of radiation[624, 629]. UV light can lead to the formation of pyrimidine dimers on the same strand such as cyclobutane pyrimidine dimers (CPDs) and (6,4)-photoproducts (6-4PPs)[627] with a preference for thymine-thymine dimers[630]. The cytosine bases of CPDs are unstable and often deaminate to generate uracil[631] or thymine if the cysteines were methylated[632]. It has been estimated that about 86% of all melanoma cases can be tracked back to exposure to UV light through the sun or devices such as tanning beds[633] with intermittent high exposure carrying a higher risk than chronic low exposure[634]. This has led to the classification of sunbed usage as a carcinogen and more severe regulations of its use in some countries[635–637].

Asbestos and other mineral fibers  Asbestos is a carcinogen implicated in the development of the majority of mesothelioma, a cancer in the outer lining of the lung[638]. The adverse health effects of asbestos exposure have been known since 1899, when Montague Murray diagnosed the first fatal case of asbestosis due to exposure at work[639]. Asbestos has been used extensively in the last century as a building material due to its desirable properties in construction ranging from sound proofing and inflammability to its inexpensiveness[640] meaning it can still be found in many buildings and exposure, especially considering its long latency, is still a major health challenge[638, 641] especially for construction workers and those processing materials[642–644]. Its directly genotoxic effects can range from DNA base oxidation and generation of double stranded breaks to deletions and aneuploidy[645] and non-genotoxic (or indirectly genotoxic) effects include the generation of ROS and RNS[646].

Chemotherapy  Many if not most classical chemotherapeutic agents - as well as radiation therapy - work by inducing DNA damage[626] and commonly used agents include alkylating agents and platinum base compounds. Alkylating agents work by adding an alkyl group to either the DNA base or backbone[647] either on one strand or in the case of bifunctional compounds in a manner creating inter-strand crosslinks[609]. While alkylating agents can arise from endogenous processes or be present in the environment - in tobacco smoke[648] and even in food (albeit at much lower concentrations)[649] - chemotherapy represents a deliberate use of these compounds. Most commonly bifunctional alkylating compounds are used that cause inter- or intra-strand DNA cross links that will lead among other things to DNA breaks and subsequent S-phase arrest followed by apoptosis[650]. Another major class of chemotherapeutics, platinum agents, work by forming adducts on DNA and also cause inter- and intra-strand crosslinks and are thus described as "alkylating-like"[651]. Other compounds commonly used in cancer therapy include agents like hydroxyurea, which deplete the dNTP
Genomic integrity and instability
pool required for replication\[652\], and intercalating agents which will insert themselves between two DNA strands thereby blocking replication\[653\].

1.4  Mutational processes and human disease

Genome integrity is fundamental to the health of an organism and failure to maintain the genome in an optimal balance results in a variety of diseases.

1.4.1 DNA repair deficiencies

Many key DNA repair proteins were actually identified due to diseases caused by mutations in them, a fact often reflected in their names. A variety of diseases exist, but a few will be introduced here. Common features of most DNA repair deficiencies are premature aging and a susceptibility to cancer. Defects in NER are responsible for several genetic human disorders and affected individuals have skin highly sensitive to sunlight due to NER’s involvement in repairing UV-induced pyrimidine dimers in humans\[654\]. The most prominent example of NER deficiency is Xeroderma pigmentosum(XP), an autosomal recessive disorder characterised by hypersensitivity to UV light, premature aging and cancer susceptibility. Many of the proteins involved in NER can be mutated causing XP, such as XPA, XPB and XPC\[415\]. Other genetic diseases with defects in NER are Cockayne syndrome, caused by mutations in ERCC8 and ERCC6 involved in TC-NER, and Trichothiodystrophy. Interestingly, a variant of XP is caused by mutations in POLH, the gene encoding Pol η, which can bypass photopyrimidine dimers during replication. Another rare, but severe DNA repair disorder is Ataxia telangiectasia (A-T), an autosomal recessive neurodegenerative disease affecting an estimated 1 in 300,000 to 1 in 90,000 people\[655\]. A-T is caused by mutations in the ATM gene, which stands for Ataxia telangiectasia mutated, and is involved in sensing DNA damage and coordinating the cellular response to such events, and affected individuals are afflicted by a variety of symptoms, from affected movement and coordination, a weakened immune system and a predisposition to cancer\[656\]. Werner’s syndrome, Bloom’s syndrome and Rothmund-Thomson syndrome are other DNA repair disorders caused by mutations in RecQ helicases: WRN, BLM and RTS/RECQ4, respectively\[657\]. These helicases are subject of active research but have been shown to be involved in critical steps of DNA damage repair such as DNA end resection, branch migration and the resolution of double Holliday junctions\[657\]. These diseases are characterised by premature aging and/or cancer predisposition.
1.4 Mutational processes and human disease

1.4.2 Cancer

Cancer is a disease of the genome\[503, 658, 659\] characterised by abnormal cellular growth and spread. This was suspected as early as 100 years ago, when David von Hansemann observed that "one can notice a certain 'disorder' in the karyokinetic processes of tumors"\[658\] and Boveri published his observations on sea urchins\[500–502\]. After the rediscovery of Mendel’s work, the latter together with Sutton\[660\] became one of the early proponents of chromosomes as carriers of genetic information. In 1914, he published highly controversial and speculative work proposing that cancer was due to abnormal genetic material\[503, 506\]. Since then much about cancer has been elucidated, which is beyond the scope of this chapter, but the reader is referred to \[299\], for further reading. In the simplest term, cancer is due to a disregulation of tissue growth pathways, many of which are key players in embryogenesis, and mutations in genes broadly classified into oncogenes and tumor suppressor genes allow these pathways to escape their tight regulation\[661\]. Most mutations that promote cancer are somatic, however, germline mutations can predispose an individual to cancer development. A single mutation is rarely enough to cause malignant tumors, and cancer was proposed to be a multi-step process early on\[366\]. Mathematical modeling in the 50s and 60s\[662, 663\] gave rise to the two-hit hypothesis\[664\]: cancer could be acquired by as little as two mutations of which one or both could be somatic and for many colorectal cancers a stereotypical progression of mutations could be identified\[665, 666\]. The first genes involved in cancer, were identified as those carried by viruses known to cause cancer\[667–669\] and homologues of those were later identified first in avian cells\[670\], then humans\[671\]. These viral genes were called oncogenes, genes who promote cancers, and in the late 70s and early 80s, oncogenes were identified to encode proteins that regulate cell growth\[672–680\]. At the same time, it became clear that mutations of the human homologues of viral oncogenes could transfer the same cancer-promoting properties and in 1982, Robert Weinberg, Michael Wigler and Mariano Barbacid cloned the first human oncogene \[365, 681–683\], which was later identified as ras \[684–686\]. It was found that a glycine to valine mutation in the 12th amino acid made the protein constitutively active\[687–689\]. The first suggestion that the dominantly acting oncogenes were not the whole story, were experiments by Harris and colleagues who observed that when a cancer and normal mouse cell were fused, the normal phenotype was dominant\[690\] leading to arguments that inherited tumors were the results of mutation in genes that suppressed tumor formation followed by somatic inactivation of the second allele\[691\]. This was confirmed in the 80s with the identification of Rb and TP53 as tumour suppressor genes\[692, 693\], and the observation that their inactivation would promote tumorigenesis\[694–698\]. The importance of such genes is further demonstrated by the HPV oncoproteins E6 and E7 which have been
found to bind and inactivate TP53 and pRB, respectively, to promote their own proliferation causing cancer in the process[699]. The fact that Rb and TP53 are involved in cell cycle progression and checkpoint control, respectively, demonstrated how critical proper regulation of these processes are to human health. Considering the fact that mutations in certain genes cause cancer and that people carrying a predisposing mutation have a much increased incidence of cancer, it is not surprising that just about anything that has been shown to cause mutations increases one’s risk for developing cancer: from radiation and tobacco smoke (see 1.3.2) to DNA repair deficiency (see 1.4.1). In fact, genetic instability is one of the hallmarks of cancer[299] and just about any type of DNA variation (see 1.2) can be involved in carcinogenesis from chromosomal translocation (for instance the Philadelphia chromosome) to a single point mutation (such as activation of ras). This is also exemplified by the discovery that hereditary non-polyposis colon cancer (HNPCC) is caused by predisposing germline mutations in genes involved in MMR[419].

1.4.3 Mutational signatures

It was known from in vitro studies that UV irradiation causes pyrimidine mutations[701, 702], but it was uncertain whether those types of mutations would also occur in cancers and contribute to carcinogenesis. Early studies sequencing exons of TP53 in cancers[703–705] provided evidence that UV and aflatoxin, a carcinogenic toxin on mold-affected crops such as peanuts, leave distinct mutation patterns on the genome. This was the first evidence that genotoxic carcinogens leave a more-or-less unique signature in the genomes of cells they affected[706–709], and the 90s saw a collection of studies sequencing more and more cancer samples sampling more and more genes[710–712]. The advent of next-generation sequencing and the subsequent drop in sequencing costs saw the advent of cancer exome and genome studies[713] and a multitude of cancers were sequenced and the profile of their mutations reported[714–760]. In the last years, work has focused on using computational methods to untangle these patterns into distinct "mutational signatures", each the remnant of a different process active at some point in the cancer’s past[761–764](Fig. 1.30). In the past years, dozens of signatures have been identified and attribution to endogenous and exogenous mutational processes is in progress(Fig. 1.4.3). For example, the mutational signature left by benzo[a]pyrene exposure is well described, as its tendency to form bulk adducts especially in guanines is well documented, and exposed cells show many C:G>T:A transversions with a transcriptional strand bias[398, 708]. Understanding how mutagens and mutagenic processes affect genomes and potentially identifying new critical carcinogens and genes involved in
Figure 1.30: Mutational signatures leave their marks on the genome
A schematic of how different mutational processes leave a characteristic imprint on a genome. The mutational patterns generated, length of exposure and intensity of the mutagenic process can vary highly which is reflected in the final mutational portrait. Reproduced from [700] with permission from the publisher.
Genomic integrity and instability are vital exercises, demonstrated by the fact that identification of potent carcinogens can be used in public health campaigns to drastically curb exposure to the substance and reduce cancer incidence[619], the ability of health care professionals to screen for predisposing mutations and thus identify high-risk individuals[765] and the identification of new drug targets as well as the advent of patient stratification and personalised medicine[766, 767].

1.4.4 DNA polymerase defects in cancer

Considering the importance of mutations in the development of cancer, it is not unreasonable to suspect that defects in DNA polymerases could give rise to cancer especially considering that absence of Pol η does predispose to cancer(see 1.4.1). Recent work has highlighted possible roles for non-null mutations in DNA polymerases δ and ε[768]. While replicative polymerases are still very accurate when proofreading is inactivated (error rates 1–5 × 10^{-5} depending on the mispairing measured)[334], mice engineered to have a homozygous proofreading deficiency (Exo−) in either Pol δ or Pol ε (equivalent of the budding yeast pol2-4 and pol3-01 strains (see Chapter 2))[769–771] develop tumors and show increased mortality while heterozygous mutants are indistinguishable from their wild-type parents. These mice show markedly different types of tumors with Pol εExo− mice developing mainly intestinal tumours with 50% survival of ~16 months and Pol δExo− mice exhibiting primarily thymic lymphomas with 50% survival of ~ 6 months. Considering that the error rates and specificities of Pol εExo− and Pol δExo− enzymes are reportedly very similar[334], the reason for the difference in tumor subtypes remains unclear. Sequencing of tumour genomes has now revealed a number of sporadic mutations in Pol ε, many of which are found in the proofreading exonuclease domain[259], while pedigree-sequencing identified two germline variants predisposing to CRC: PolE L424V and PolD1 S478N[768]. It remains unclear how these mutations affect exonuclease activity in these tumours, how they impact replication fidelity and how they mutagenise cells[259].

1.5 DNA Sequencing

DNA sequencing is a useful tool for biologists and health-care professionals and has a broad range of applications from cloning to evolutionary studies. In 1977, Walter Gilbert and Frederick Sanger developed methods to determine the sequence of a DNA molecule. The Gilbert-Maxam method was based on chemical cleavage at specific bases (Fig. 1.32-A)[772], while the Sanger method relied on dideoxy chain termination (Fig. 1.32-B)[773]. Due to its high
### 1.5 DNA Sequencing

#### Figure 1.31: Summary of known mutational signatures

Reproduced from [700] with permission from the publisher.

<table>
<thead>
<tr>
<th>DNA damage</th>
<th>DNA repair pathway</th>
<th>Replicational process</th>
<th>Mutational signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCG→TAA at methylated Cpgs correlates with age</td>
<td>Replicational polymerases</td>
<td>Deamination</td>
<td>Replication</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Signature 1A</td>
</tr>
<tr>
<td>APOBEC editing</td>
<td></td>
<td></td>
<td>Signature 2</td>
</tr>
<tr>
<td>Replication</td>
<td></td>
<td></td>
<td>Signature 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Signature 7</td>
</tr>
<tr>
<td>UV radiation on pyrimidines and pyrimidines</td>
<td>Transcription-coupled repair</td>
<td>Replication</td>
<td>Replication</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Signature 11</td>
</tr>
<tr>
<td>Topoisomerase-induced D- methylguanine lesions</td>
<td>Direct repair using methylguanine DNA methyltransferase</td>
<td>Replication</td>
<td>Replication</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Signature 4</td>
</tr>
<tr>
<td>Benzo(a)pyrene (BaP) adducts on guanine</td>
<td>Transcription-coupled repair</td>
<td>Replication</td>
<td>Replication</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Signature 22</td>
</tr>
<tr>
<td>Ethylated acid (EA) adducts on adenine</td>
<td>Transcription-coupled repair</td>
<td>Replication</td>
<td>Replication</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Signature 6</td>
</tr>
<tr>
<td>Natural errors</td>
<td>Mismatch repair enzyme</td>
<td>Replication</td>
<td>Replication</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Signature 10</td>
</tr>
<tr>
<td>Natural errors</td>
<td>Defective DNA polymerase ε</td>
<td>Replication</td>
<td>Replication</td>
</tr>
</tbody>
</table>

Nature Reviews | Genetics
Genomic integrity and instability

Figure 1.32: Early sequencing techniques: Gilbert and Sanger
Schematics to illustrate the principles of early DNA sequencing strategies. A| Dideoxy “Sanger” Sequencing: This type of sequencing required a DNA template, a primer, a DNA polymerase, all four standard dNTPs and one of the di-deoxy NTPs, which terminate DNA strand elongation. Ratios of dNTP to ddNTP were chosen to generate products of every length (dNTPs in approximately 100-fold excess). One reaction for each base is prepared with the appropriate ddNTP. Products are run on a denaturing polyacrylamide-urea gel with each reaction in a different lane to separate fragments by size. The DNA sequence can then be determined by reading the gel from the bottom up. B| Maxam–Gilbert sequencing: Chemical treatment of a radiolabelled fragment of DNA breaks it into fragments at specific bases. For instance, pyrimidines (C+T) are hydrolysed with hydrazine. In a separate reaction, the addition of salt inhibits hydrazine action on thymine (C only). Bases are separated on gels similar as well and the sequence can be inferred from the band pattern.
efficiency and relatively low use of radioactivity, Sanger sequencing was quickly adopted for routine sequencing. However, it was still a method that was laborious and did require radioactivity. In 1987, Applied Biosystems introduced the first automatic sequencing machine (namely AB370). Improvements such as capillary gels and fluorescent terminating nucleotides allowed this capillary sequencer to detect up to 500,000 bases a day and its read length could reach 600 bases. Its current model AB3730xl can generate an output of 2.88 million bases per day and since 1995 the read length can reach 900 bases. Automatic sequencing instruments and their software were the main tools used for the Human Genome completion in 2001[14]. This achievement stimulated the development of new sequencing instruments to increase the accuracy and power of sequencing, while simultaneously reducing the cost and labour involved. Next generation sequencing methods are characterised by massive parallel sequencing, high throughput and reduced costs[774]. The three most typical massively parallel sequencing systems were developed a decade ago: 454 was launched in 2005, Solexa the next year and SOLiD the year after[774]. As most of the sequencing data described in this work has been obtained with instruments from Illumina (who purchased Solexa), their next generation sequencing technique will be reviewed here.

Illumina sequencing relies on the "sequencing by synthesis" concept to produce short sequencing reads from tens of millions of surface-amplified DNA fragments simultaneously[777]. Sequencing by synthesis works by adding four differently fluorescently labelled, 3'-OH chemically inactivated nucleotides to a primed DNA strand. The chemical modification of the nucleotide prevents the addition of more than one nucleotide at a time (Fig. 1.33). Each base incorporation cycle is followed by washing off excess nucleotides and an imaging step to identify the base just incorporated. This is followed by a chemical step that reverses the chemical block and removes the fluorescent group, making the DNA fragment extendable again. Another base incorporation cycle follows. This process is carried out in a massively parallel fashion in an Illumina sequencer. A library is prepared from extracted DNA by shearing and size selection of DNA fragments, followed by ligation of specific adaptor sequences and indexes for sample identification to single-stranded DNA(in case of multiplexed libraries where up to 96 samples are mixed in one sequencing reaction). The library is then added to a lane of an eight-lane flow cell, whose surface is coated with oligonucleotides complementary to the adaptors that are ligated to the DNA fragments to be sequenced[777]. DNA fragments are thus hyrbidised to the surface of the flow cell and subsequently amplified in place by an isothermal polymerase resulting in discrete clusters of amplified DNA (Fig. 1.34-A). The flow cell is placed in the sequencer and sequencing by synthesis is carried out by flowing through reagents alternating with laser image acquisition (Fig. 1.34-B). Not the entire DNA fragment
Figure 1.33: Sequencing by synthesis

Schematic of Sequencing by Synthesis (SBS) | 1) Incorporation of a fluorescent dATP-PC-ROX, after washing and imaging 2) the terminator is photo-cleaved. 3) Next, dGTP-PC-Bodipy-FL-510 is incorporated, excess nucleotides washed off and the fluorophore imaged. 4) This is followed by another round of photocleavage. This proceeds to sequence the DNA molecule. Reproduced from [775] in accordance with the publisher’s policy. Copyright (2005) National Academy of Sciences.
1.5 DNA Sequencing

Figure 1.34: Solid-phase bridge amplification and sequencing by synthesis (Illumina)

A. In solid-phase bridge amplification, fragmented DNA is ligated to adapter sequences and bound to a primer immobilized on a solid support, such as a patterned flow cell. The free end can interact with other nearby primers, forming a bridge structure. PCR is used to create a second strand from the immobilized primers, and unbound DNA is removed.

B. After solid-phase template enrichment, a mixture of primers, DNA polymerase and modified nucleotides are added to the flow cell. Each nucleotide is blocked by a 3'-O-azidomethyl group and is labelled with a base-specific, cleavable fluorophore (F). During each cycle, fragments in each cluster will incorporate just one nucleotide as the blocked 3' group prevents additional incorporations. After base incorporation, unincorporated bases are washed away and the slide is imaged by total internal reflection fluorescence (TIRF) microscopy using either two or four laser channels; the colour (or the lack or mixing of colours in the two-channel system used by NextSeq) identifies which base was incorporated in each cluster. The dye is then cleaved and the 3'-OH is regenerated with the reducing agent tris(2-carboxyethyl)phosphine (TCEP). The cycle of nucleotide addition, elongation and cleavage can then begin again.

Figure and Figure Description reproduced from [776] with permission from the publisher.
is sequenced as base call quality drops off with each cycle limiting the read length. The reasons of this are numerous and while some can and have been addressed by improvements in fluorescent labels, optics and flowcell design, phasing is an intrinsic problem of sequencing clusters of DNA[778]. Phasing is the maintenance of synchronicity of synthesis is a given cluster. Each cluster is made up of millions of DNA strands, which are visualised as a single fluorescent dot. Identification of the added base depends on all DNA strands being extended in a synchronous manner as an "average" signal is detected[778]. Since the chemical steps involved in this process are not 100% efficient, synthesis on some templates lag behind that on others and quality typically drops after a number of cycles as the population loses synchrony. Initially, read length was limited to ~32-40bp (2007)[777], but read length capability has been rapidly improving and in this work the Illumina HiSeq 2500 was used to produce paired-end reads of 125bp each. Paired-end sequencing - sequencing the same DNA from both ends - allows to generate more high quality data than sequencing the same number of bases from a single end under the same conditions. Additionally, paired-end reads are useful for detection of large scale variation (see Chapter 2.4.5). Once sequencing the forward and the reverse strand of the DNA has been accomplished, the HiSeq machine itself will analyse the images and output base calls and quality scores for each cycle[774].

More recently, new, third-generation sequencing machines have been developed. They mainly differ from next-generation sequencing in that they do not need amplification of the template and that the signal is captured in real time[774]. Main advantages of these new sequencing techniques include shorter sample preparation times and significantly longer read lengths. The Pacific Biosciences sequencer works by visualising the fluorophores on labelled nucleotides as a polymerase replicates the DNA (Fig. 1.35-A), while the Oxford nanopore relies on characteristic disruptions of an electric current as a DNA molecule is threaded through a protein pore in a membrane (Fig. 1.35-B). Sequencing costs have been falling dramatically in the last decade with a human genome being sequenced for less than $5,000 in 2012 (as opposed to the more than $300 million the initial draft sequence cost[779]) and a budding yeast genome costing as little as £10 to sequence. If one accepts the premise that genetics is the pursuit to link genotype to phenotype then DNA sequencing will remain a cornerstone of genetics research.
1.5 DNA Sequencing

Figure 1.35: Third-generation Sequencing Techniques

**A** Pacific Biosciences (PacBio). Template fragments are processed and ligated to hairpin adapters at each end, resulting in a circular DNA molecule with constant single-stranded DNA (ssDNA) regions at each end with the double-stranded DNA (dsDNA) template in the middle. The resulting ‘SMRTbell’ template undergoes a size-selection protocol in which fragments that are too large or too small are removed to ensure efficient sequencing. Primers and an efficient ϕ29 DNA polymerase are attached to the ssDNA regions of the SMRTbell. The prepared library is then added to the zero-mode waveguide (ZMW) SMRT cell, where sequencing can take place. To visualize sequencing, a mixture of labelled nucleotides is added; as the polymerase-bound DNA library sits in one of the wells in the SMRT cell, the polymerase incorporates a fluorophore-labelled nucleotide into an elongating DNA strand. During incorporation, the nucleotide momentarily pauses through the activity of the polymerase at the bottom of the ZMW, which is being monitored by a camera. 

**B** Oxford Nanopore Technologies. DNA is initially fragmented to 8–10 kb. Two different adapters, a leader and a hairpin, are ligated to either end of the fragmented dsDNA. Currently, there is no method to direct the adapters to a particular end of the DNA molecule, so there are three possible library conformations: leader–leader, leader–hairpin and hairpin–hairpin. The leader adapter is a double-stranded adapter containing a sequence required to direct the DNA into the pore and a tether sequence to help direct the DNA to the membrane surface. Without this leader adapter, there is minimal interaction of the DNA with the pore, which prevents any hairpin–hairpin fragments from being sequenced. The ideal library conformation is the leader–hairpin. In this conformation the leader sequence directs the DNA fragment to the pore with current passing through. As the DNA translocates through the pore, a characteristic shift in voltage through the pore is observed. Various parameters, including the magnitude and duration of the shift, are recorded and can be interpreted as a particular k-mer sequence. As the next base passes into the pore, a new k-mer modulates the voltage and is identified. At the hairpin, the DNA continues to be translocated through the pore adapter and onto the complement strand. This allows the forward and reverse strands to be used to create a consensus sequence called a ‘2D’ read. 

Figure and Text reproduced from [776] with permission from the publisher.
1.6 The budding yeast *Saccharomyces cerevisiae* as a model organism

As should be clear from how much of the above presented knowledge was gained from experiments in budding yeast, *S. cerevisiae* is a valuable model organism to study DNA replication, repair, genome maintenance and other fundamental aspects of cell biology. Budding yeast is classified as a fungus or mold, and as a single-celled eukaryote contains membrane-bound organelles such as a nucleus and mitochondria. They get their common name of baker’s or brewer’s yeast from their many applications in generating just such foods, and their name budding yeast from the way they divide: a smaller daughter cells buds off its mother in a process that can be as fast as 90 minutes in optimal conditions[780]. *S. cerevisiae* shows a rudimentary sexual dimorphism with two different mating types in haploid cells called MATα and MATα. When in each other’s proximity cells of different mating types can mate and form a diploid cell. When nutrients are scarce, a diploid can then undergo meiosis and sporulation resulting in four haploid spores, two MATα and two MATα[780](Fig. 1.6). In nutrient-rich conditions, these then germinate back to haploid yeast and - if still present around their siblings of opposite mating type - re-mate to form a diploid. Used in laboratories since the 1930s[781], yeast cells are inexpensive and easy to culture and store: their cells are about ~5 μm in diameter (between bacteria and human cell sizes) and they can be easily grown on agar plates where they form colonies in 2-3 days at room temperature (more quickly in 30°C incubators). They can be stored short-term in fridges, long term at -80°C in glycerol or at room temperature when freeze-dried. One of the most commonly used experimental yeast strains, S288C, was constructed by Robert Mortimer[781] in the 50s primarily from EM93, which had been isolated from rotting figs in California and was suitable for genetic crosses, and S288C has been used as a parental strain for a plethora of mutants.

**Genetics and tools** Yeast genetics expanded exponentially after it was successfully transformed with a DNA plasmid that had been amplified in *E. coli*[782](reviewed in [781]). The key attractive feature of yeast cells for geneticists has been the pliable nature of its genome and the ever expanding array of tools, plasmids, selectable markers and DNA cassettes. Development of the polymerase chain reaction (PCR), a now standard laboratory tool to amplify DNA sequences, combined with the remarkable efficiency of homologous recombination in yeast[783, 784] - transformation of linearised DNA into yeast will cause its homology-directed insertion almost without fail - has led to the development of a myriad of custom-designed yeast strains and an extensive collection of deletion strains completed in
1.6 The budding yeast *Saccharomyces cerevisiae* as a model organism

Haploid and diploid yeast cells can reproduce by mitosis. A haploid Mat $a$ and a haploid Mat $\alpha$ can mate and form a diploid. A diploid cell can undergo meiosis and generate four haploid spores.

In 2002, the first and to-date only complete systematic deletion collection of any organism[785]. Yeast cells lend themselves to the isolation of mutants and suppressors (mutants that reverse a phenotype of another mutant)[786], genetic crosses, epistasis experiments, microscopy analysis[787], complementation cloning, efficient gene-replacement, Synthetic Gene Array (SGA) experiments[788], next-generation sequencing, chromatin immunoprecipitation (ChIP) experiments[789], tagging of proteins with fluorescent and other probes[787] and the determination of protein-protein interactions using the yeast two-hybrid system[790] to name but a few. Of great utility is also the fact, that after meiosis the four resulting spores stay attached to one another (called a tetrad) and using a dissection needle all four meiotic products can be recovered allowing genetic analysis of mutants and combinations of mutants[780].

The yeast genome In 1996, the *S. cerevisiae* S288C genome sequence was completed making it the third species to be sequenced and the first fully sequenced eukaryote[791]. This was not just a notable achievement in itself, but has provided the scientific community with a wealth of information. Combined with a detailed database of genes, their mutants and their phenotypes, the genome can be queried by anyone in the Saccharomyces Genome Database (SGD). Combined with Gos Micklem’s YeastMine tool to systematically search the database, the SGD website has been a helpful tool for detailed experimental planning. A haploid yeast genome is roughly 12 Megabases in size spread over 16 chromosomes - likely the result of a whole-genome duplication[792–794] - and contains 5820 verified genes/open reading
frames (ORFs) for which 4958 homologs can be identified in humans. Most budding yeast genes do not contain any intron (only ~4% do) partly explaining the high gene density in the genome. The non-protein coding genes in the genome include those that are transcribed to generate transfer RNA, (tRNAs, critical to decode the genomic triplets into amino acids) and ribosomal DNA (rDNA, a main component of ribosomes, the molecular complexes that assemble proteins), which can be found in 100-150 tandem repeats on chromosome XII. Other repetitive regions of the yeast genome are the so-called long terminal repeat (LTR) retrotransposons, or Ty elements, which are scattered across the entire genome. Chromeosome III is the chromosome in yeast cells that determines the cell’s mating type: it contains the MAT locus which can contain either the MATa or MATα allele. A diploid will usually contain one of each on its two homologous chromosome III. The two different alleles confer mating type behaviour in a slightly different, albeit quite complex, manner which is reviewed in. In contrast to other organisms, chromosome III is also carrying information for the other mating type: the HMRa locus contains a functional MATa allele and the HMLα contains a copy of the MATα allele. These loci (also known as silent mating-type cassettes) are silenced in heterochromatin, but act as "back-ups" that can actually allow haploid yeast cells to switch mating type by transferring the information of the cassette of the other mating type into the active MAT locus (Fig. 1.37). In populations in the wild this ability ensures that a single haploid cell can divide, progeny can switch mating type and a diploid population can form. This ability has been inactivated in most laboratory yeast strains to ensure that most strains are stable both in mating type and ploidy. With the advent of next generation sequencing techniques and the recent drop in sequencing costs, it is now possible to sequence the whole genome of a yeast for a few tens of Euro.

**Biological advances using S. cerevisiae as model organism** The past century has seen remarkable advances in our understanding of biology and many key insights have come from studying the budding yeast *S. cerevisiae*. Apart from advances such as insights into DNA replication, DNA repair and regulation of the cell cycle (see 1.1), yeast has been used to elucidate much about eukaryotic vesicle trafficking (Nobel Prize in 2013), initiation of transcription (Nobel Prize in 2006) and eukaryotic telomere structure (Nobel Prize in 2009), among many other landmark discoveries. *S. cerevisiae* continues to be a valuable and flexible organism in the study of cell biology and genetics and remains suitable to address questions about DNA replication and genome maintenance.
The mating type is determined by the genetic information contained within the mating type locus on Chromosome III. Yeast cells also contain inactive copies of the genetic information for both mating types in silent mating type cassettes at the ends of the chromosome. Budding yeast cells can use chromosomal recombination to replace the information in the mating type locus with that in one of the silent cassettes, though this ability is often inactivated in strains kept in the laboratory by mutations in the HO endonuclease, which makes the cut that initiates mating type switching.

Table 1.6: Standard Nomenclature for *S. cerevisiae* genetics using POL2 as an example.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Nomenclature</th>
</tr>
</thead>
<tbody>
<tr>
<td>YNL262W</td>
<td>Systematic Name for ORF</td>
<td>Each ORF has a systematic name. The convention is: Y (for yeast), N (chromosome number; A = chrl, B = chrl, ...), L (for left arm of the chromosome), 262 (ORFs are numberes starting from centromere), W (for the coding strand: W for Watson and C for Crick strand).</td>
</tr>
<tr>
<td><em>POL2</em></td>
<td>Wilde type (wt)</td>
<td>Italicised common gene name (three capital letters followed by numbers) with plus in superscript</td>
</tr>
<tr>
<td><em>POL2</em></td>
<td>Dominant allele (often used to mean wt)</td>
<td>Italicised common gene name (three capital letters followed by numbers)</td>
</tr>
<tr>
<td><em>POL2</em>-1, <em>POL2</em>-2, etc.</td>
<td>Specific dominant allele</td>
<td>Designation for dominant mutant allele followed by hyphen and number</td>
</tr>
<tr>
<td><em>pol2</em></td>
<td>Recessive allele</td>
<td>Three italicized lowercase letters and number</td>
</tr>
<tr>
<td><em>pol2</em>-1, <em>pol2</em>-2, <em>pol2</em>-3, <em>pol2</em>-4, etc.</td>
<td>Specific recessive allele</td>
<td>Designation for recessive mutant allele followed by hyphen and number</td>
</tr>
<tr>
<td><em>pol2Δ</em></td>
<td>Deletion of gene</td>
<td>Designation for recessive mutant allele followed by Δ</td>
</tr>
<tr>
<td><em>Pol2p</em></td>
<td>Protein Product of gene</td>
<td>Three letters, with the first being uppercase, followed by a number and optional lower case p; not italicized</td>
</tr>
<tr>
<td><em>Pol2</em></td>
<td>Protein Product of gene (Alternative)</td>
<td></td>
</tr>
</tbody>
</table>
Table 1.7: A selection of Nobel Prizes awarded for work using *S. cerevisiae* as a model organism.

<table>
<thead>
<tr>
<th>Area of science</th>
<th>Year</th>
<th>Nobel Prize</th>
<th>Principal Investigators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fermentation</td>
<td>1907</td>
<td>Chemistry</td>
<td>Eduard Buchner</td>
</tr>
<tr>
<td>Cell cycle regulation</td>
<td>2001</td>
<td>Medicine or Physiology</td>
<td>Leland H. Hartwell, Tim Hunt and Sir Paul M. Nurse</td>
</tr>
<tr>
<td>Transcription</td>
<td>2006</td>
<td>Chemistry</td>
<td>Roger D. Kornberg</td>
</tr>
<tr>
<td>Telomeres</td>
<td>2009</td>
<td>Medicine or Physiology</td>
<td>Elizabeth H. Blackburn, Carol W. Greider and Jack W. Szostak</td>
</tr>
<tr>
<td>Vesicle trafficking</td>
<td>2013</td>
<td>Medicine or Physiology</td>
<td>James E. Rothman, Randy W. Schekman and Thomas C. Südhof</td>
</tr>
</tbody>
</table>
Chapter 2

Analysis of cancer-associated polymerase mutations

Overarching hypothesis

DNA polymerase mutations identified in cancer samples can be constructed in the model organism *S. cerevisiae* to examine their relevance to tumour progression and whole-genome sequencing of budding yeast samples can yield relevant biological insights.

Aims:

- To compile a list of relevant mutations in DNA polymerases identified in cancer samples

- To prioritise mutations and determine their *S. cerevisiae* equivalents

- To conduct mutation accumulation experiments to identify the consequences of mutations in DNA polymerase on a genome-wide scale

- To establish sequence analysis protocols for budding yeast whole-genome sequencing data

- To show that these sequence analysis protocols are functional and can be applied beyond this project
Figure 2.1: Methodology of the work carried out during my PhD
The chapter in which each step is covered is indicated at the left. Projects loosely associated with the my principal DNA polymerase mutation project are highlighted in darker boxes.
2.1 Introduction

Methods and results detailed in sub-sections 2.4.3 and 2.4.4 have been published or are accepted for publication (see [801] and [802]). Figures and Figure legends have partially been reproduced from this work in accordance with the copyright provisions of the publisher.

Cancer is a disease of mutations and defects in the mechanisms that maintain replication fidelity are likely underlying mutations in genes involved in tumourigenesis. It has been described that germline mutations in the DNA mismatch repair (MMR) machinery predispose to hereditary colorectal cancer[803, 804], but the case has been less clear-cut for polymerases. Due to the relatively recent advent of tumour genome sequencing, we now have the tools to actually get information on which polymerase genes are commonly mutated, the frequency of such mutations, which tumour types are affected and the characteristics of such tumours. So far, sequencing of a number of cancers has revealed somatic mutations in \textit{POLE} coding for the catalytic subunit of \textit{Pole}[259]. At the same time, pedigree-sequencing of families with a history of colorectal cancer identified two predisposing germline variants \textit{POLE} L424V and \textit{POLD1} S478N[768]. The condition was termed polymerase proofreading-associated polyposis (PPAP)[768, 805] though there is currently no genome-wide association studies (GWAS) evidence for associated risk between polymerase SNPs and colorectal cancer[806].

It is known that mutations in the exonuclease domain (EDM) of \textit{Polε} and \textit{Polδ} in yeast cause a base substitution phenotype of varying severity. Mutations affecting the catalytic residues of the proofreading domain of \textit{POL3} (\textit{pol3-01}) cause a mutator phenotype with increased base substitution and frameshift mutations[338]. Similar mutations in \textit{Polε} (\textit{pol2-4}) reduce proofreading activity about 100-fold in vitro, while leaving polymerase activity at wild-type levels[312]. \textit{In vivo}, these mutations cause a mutator phenotype and using different reporter assays the increase in mutation rate was found to be between 5- and 43-fold(Table 2.1), highlighting the significance of proofreading for genome maintenance as well as the limitations of classical reporter assays to accurately describe mutator phenotypes.

Mice carrying mutations in the proofreading domain of polymerase ε (\textit{Polε}^{exo-/exo-}; the mouse equivalent of the yeast \textit{pol2-4} mutation) showed a predisposition to cancer, while \textit{Polε}^{exo-/exo} were virtually indistinguishable from wild-type in this respect[769]. Spontaneous mutations were more frequent in \textit{Polε}^{exo-} mice than in \textit{Polδ}^{exo-} mice, in contrast to the budding yeast, where the \textit{pol3-01} mutation causes a higher mutational frequency than \textit{pol2-4}. This either reflects a true discrepancy between yeast and mice or results from the fact that mutation frequency is estimated usually at single genetic loci (e.g. \textit{Atp1a1} and \textit{Hprt} in mice versus \textit{URA3}, \textit{CAN1} and \textit{SUP4-o} in yeast), further confirming the need for improved methods to assess mutation rate increases. Mice deficient for both \textit{Polε} and \textit{Polδ} proofreading activity
Analysis of cancer-associated polymerase mutations

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Assay gene</th>
<th>Fold change to wt</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>pol3-01</td>
<td>his7-2</td>
<td>240</td>
<td>[338]</td>
</tr>
<tr>
<td>pol3-01</td>
<td>URA3&lt;sup&gt;a&lt;/sup&gt;</td>
<td>130</td>
<td>[338]</td>
</tr>
<tr>
<td>pol3-01</td>
<td>URA3&lt;sup&gt;a&lt;/sup&gt;</td>
<td>52</td>
<td>[282]</td>
</tr>
<tr>
<td>pol3-01</td>
<td>lys2::InsLD</td>
<td>0.6</td>
<td>[308]</td>
</tr>
<tr>
<td>pol3-01</td>
<td>his7-2</td>
<td>74</td>
<td>[308]</td>
</tr>
<tr>
<td>pol3-01</td>
<td>CAN1</td>
<td>630</td>
<td>[282]</td>
</tr>
<tr>
<td>pol3-01</td>
<td>SUP4-o&lt;sup&gt;b&lt;/sup&gt;</td>
<td>32-106</td>
<td>[303]</td>
</tr>
<tr>
<td>pol3-01</td>
<td>trp1–289</td>
<td>100</td>
<td>[282]</td>
</tr>
<tr>
<td>pol3-01</td>
<td>lys2::InsE&lt;sup&gt;c&lt;/sup&gt;</td>
<td>26 - 188</td>
<td>[807]</td>
</tr>
<tr>
<td>pol2-4</td>
<td>CAN1</td>
<td>5</td>
<td>[312]</td>
</tr>
<tr>
<td>pol2-4</td>
<td>ade5-1</td>
<td>43</td>
<td>[312]</td>
</tr>
<tr>
<td>pol2-4</td>
<td>URA3&lt;sup&gt;a&lt;/sup&gt;</td>
<td>15</td>
<td>[282]</td>
</tr>
<tr>
<td>pol2-4</td>
<td>his7-2</td>
<td>24</td>
<td>[312]</td>
</tr>
<tr>
<td>pol2-4</td>
<td>his7-2</td>
<td>63</td>
<td>[282]</td>
</tr>
<tr>
<td>pol2-4</td>
<td>leu2-1</td>
<td>18</td>
<td>[312]</td>
</tr>
<tr>
<td>pol2-4</td>
<td>hom3-10</td>
<td>9</td>
<td>[312]</td>
</tr>
<tr>
<td>pol2-4</td>
<td>his1-7</td>
<td>31</td>
<td>[312]</td>
</tr>
<tr>
<td>pol2-4</td>
<td>SUP4-o&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.9</td>
<td>[303]</td>
</tr>
<tr>
<td>pol2-4</td>
<td>trp1–289</td>
<td>3.9</td>
<td>[282]</td>
</tr>
<tr>
<td>pol2-4</td>
<td>lys2::InsE&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1.2 - 6</td>
<td>[807]</td>
</tr>
<tr>
<td>pol2-16</td>
<td>URA3&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.6</td>
<td>[282]</td>
</tr>
<tr>
<td>pol2-16</td>
<td>his7–2</td>
<td>1.4</td>
<td>[282]</td>
</tr>
<tr>
<td>pol2-16</td>
<td>trp1–289</td>
<td>1.9</td>
<td>[282]</td>
</tr>
</tbody>
</table>

Table 2.1: Polymerase exonuclease domain mutations in <i>S. cerevisiae</i>

Figures were taken from publications as indicated. Fold change shows the ratio between mutant value and wild-type. All strain mutations are haploid unless otherwise indicated. As a comparison mutation rates for the strain pol2-16 are shown, in which all of POL2 except the non-catalytic C-terminus is deleted<sup>[285]</sup>. <sup>a</sup>Forward mutation of URA3. <sup>b</sup>SUP4-o orientation was altered to be both on leading and lagging strand, which gave vastly different mutation rates in the case of pol3-01. <sup>c</sup>lys2::InsE alleles contain various sizes of dA homonucleotide runs. For similar experiments, see <sup>[338, 339, 808]</sup>.
2.2 Identification of polymerase mutations

were viable, but died earlier of thymic lymphoma.

Not much is known about whether these mutations are passenger mutations or promote
tumour progression. Additionally, it is unclear whether these mutations affect polymerase
fidelity and to what degree. In my thesis, I will explore these questions, first, by assembling a
list of mutations in DNA polymerases, then, using the budding yeast *S. cerevisiae* to test the
effects of altered DNA polymerases on genomes, I will identify the most striking candidates
to explore further in yeast, mouse and human (Fig. 2.1).

### 2.2 Identification of polymerase mutations

#### 2.2.1 Literature search for DNA polymerase mutations in cancer

Whole-exome and whole-genome sequencing of cancer samples has identified mutations in
DNA polymerases and the list is growing with little follow-up work on the nature of these
variants. The Cancer Genome Atlas (TCGA), a project to catalogue genetic mutations re-
sponsible for cancer, has identified DNA polymerase mutations in 3% of colorectal cancers
(CRC)[718] and 7% of endometrial cancers they sequenced[809]. While recurrent mutations
in *POLE* could be identified, none were found for *POLD1*. A different CRC project identi-
fied another recurrent change p.Pro286Arg[751]. Only a minority of tumours show LOH or
inactivating mutations for *POLE* or *POLD1*[806].

For this project, the mutations described in the work from Palles and co-workers[768],
Church and co-workers[810] and the TGCA endometrial sequencing project[809] were assem-
bled into a list of mutations and, in order to properly locate these mutations in whole-genome
datasets, amino acid changes were converted to their genomic coordinates(Table 2.2). The
mutations are all found within the N-terminal exonuclease domains of the polymerases (Fig.
2.2), which may reflect a real increased prevalence of mutations in this part of the protein,
but is more likely due to the identification of several mutations by specifically sequencing the
exonuclease domain of Polε and Polδ[810].

#### 2.2.2 Query of COSMIC database, discarding single nucleotide polymorphisms and unconserved residues

The availability of vast amounts of cancer sequencing data allows the assessment of the recur-
rence of individual mutations as a base for further prioritisation as well as their distribution
among different types of cancer. The Catalogue of Somatic Mutations in Cancer (COSMIC)
Table 2.2: Genomic locations of mutations in DNA polymerases in different human genome assemblies

<table>
<thead>
<tr>
<th>Gene</th>
<th>AA change</th>
<th>Chr</th>
<th>Pos(37)</th>
<th>Pos(38)</th>
<th>REF</th>
<th>ALT</th>
</tr>
</thead>
<tbody>
<tr>
<td>POLD1</td>
<td>p.Arg311Cys</td>
<td>19</td>
<td>50905959</td>
<td>50402702</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>POLD1</td>
<td>p.Gly426Ser</td>
<td>19</td>
<td>50909472</td>
<td>50406215</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>POLD1</td>
<td>p.Pro327Leu</td>
<td>19</td>
<td>50906319</td>
<td>50403062</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>POLD1</td>
<td>p.Ser370Arg</td>
<td>19</td>
<td>50906449</td>
<td>50403192</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>POLD1</td>
<td>p.Ser478Asn</td>
<td>19</td>
<td>50907113</td>
<td>50406456</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>POLD1</td>
<td>p.Val392Met</td>
<td>19</td>
<td>50906786</td>
<td>50403529</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Met444Lys</td>
<td>12</td>
<td>133250189</td>
<td>132673603</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Ala456Pro</td>
<td>12</td>
<td>133249857</td>
<td>132673271</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Arg446Gln</td>
<td>12</td>
<td>133250183</td>
<td>132673597</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Asp275Val</td>
<td>12</td>
<td>133253217</td>
<td>132676631</td>
<td>T</td>
<td>A</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Gln453Arg</td>
<td>12</td>
<td>133250162</td>
<td>132673576</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Leu424Val</td>
<td>12</td>
<td>133250250</td>
<td>132673664</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Pro286Arg</td>
<td>12</td>
<td>133253184</td>
<td>132676598</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Pro436Arg</td>
<td>12</td>
<td>133250213</td>
<td>132673627</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Ser297Phe</td>
<td>12</td>
<td>133253151</td>
<td>132676565</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Val411Leu</td>
<td>12</td>
<td>133250289</td>
<td>132673703</td>
<td>C</td>
<td>A</td>
</tr>
</tbody>
</table>

Genomic locations and nucleotide changes for the DNA polymerase mutations were identified using the human reference genome assemblies GRCh37 and CRCh38. Re-mapping between assemblies was done using the NCBI Genome Remapping Service[811]. Locations and nucleotide changes were computed using the reference genomes, their annotations and the codon table (see Fig. 1.27). AA Change stands for amino acid change, Chr for chromosome, Pos(37) for the position along the chromosome in genome assembly GRCh37, Pos(38) reflects the position in assembly GRCh38, REF is the base found in the reference genome and ALT is the base identified in the cancer samples. The source for the mutation can be found in the last column.
2.2 Identification of polymerase mutations

Figure 2.2: Locations of DNA polymerase mutations within the proteins

The locations of the mutations within the protein with reference to the domain structure is given. Plot was generated by Dr. Carla Daniela Robles Espinoza using a custom written script.
Analysis of cancer-associated polymerase mutations

Figure 2.3: Prevalence of polymerase mutations of interest in COSMIC
The list of DNA polymerase mutations were cross-referenced with the COSMIC whole-genome data (v74) [812]. Recurrence of mutations in the whole dataset is displayed with information about the tissue of origin. For comparison, the composition of tumour origins across the whole database for the relevant tissue types is featured.

is a vast database of somatic changes observed in human cancer samples [812]. To assess the prevalence of these mutations, I accessed their curation of 22,690 whole cancer genomes and analysed mutation recurrence and tumour origin (Fig. 2.3). Recurrence indicates that DNA polymerase mutations to prioritise for testing include POLE S297F, POLE P286R, POLE V411L and POLE A456P. Indeed, DNA polymerase mutations are enriched in endometrial cancers and to a lesser extent colorectal cancers, which is not due to an overrepresentation of those cancer types in the dataset as a whole (endometrial cancers are 2.7% of all samples, colorectal cancers are 5.8%).

None of these variants were excluded from the list of candidates on the basis of occurrence in sequencing projects aiming to capture common variation in the human population (Table 2.3) considering the most common variant was found in 0.03% of the population. To get preliminary information on the severity of these mutations I ran bioinformatic predictions soft-
<table>
<thead>
<tr>
<th>Gene</th>
<th>AA change</th>
<th>dbSNP</th>
<th>1000Genomes</th>
<th>500Exomes &amp; CGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>POLD1</td>
<td>p.Arg311Cys</td>
<td>rs201010746 T=0.000001 (ExAC)</td>
<td>T=0.00021</td>
<td>rs201010746</td>
</tr>
<tr>
<td>POLD1</td>
<td>p.Gly426Ser</td>
<td>-</td>
<td>-</td>
<td>lowQual</td>
</tr>
<tr>
<td>POLD1</td>
<td>p.Pro327Leu</td>
<td>rs397514633 (OMIM)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>POLD1</td>
<td>p.Ser370Arg</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>POLD1</td>
<td>p.Ser478Asn</td>
<td>rs397514632 (OMIM)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>POLD1</td>
<td>p.VAL392Met</td>
<td>rs778843530 A=0.000008 (ExAC)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Met444Lys</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Ala456Pro</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Ala465Val</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Arg446Gln</td>
<td>rs151273553 T=0.0003 (ExAC)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Asp275Val</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Gln453Arg</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Leu424Val</td>
<td>rs483352909 A=0.000008 (ExAC)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Pro286Arg</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Pro436Arg</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Ser297Phe</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>POLE</td>
<td>p.Val411Leu</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2.3: Checking DNA polymerase mutations for common variants
DNA polymerase mutations were cross-referenced with dbSNP, build 139 [813], 1000 Genomes, release May 2013[814] and in-house common variation sequencing projects (500 Exome Project and 300 control exomes of the cancer genome project). The submitter to db-SNP is denoted in parentheses. The minor allele frequency (MAF) is denoted in the Table with the minor allele. MAF refers to the frequency of the least common allele in a given population.
Table 2.4: Polymerase mutations identified from the literature with predicted consequences
Polymerase mutations were identified from the literature [768, 809, 810] and their potential effects on protein structure and function was predicted using bioinformatic mutation prediction software. Scores are judged as follows: PROVEAN | If the score is <= -2.5 (predefined threshold), the protein variant is predicted "deleterious". SIFT | Score ranges from 0-1 and any score <0.05 is considered "deleterious". Poly-Phen2 | The score is the probability of the substitution being deleterious. PredictProtein(PPopen) | Scores range from -100 to 100 and score > 50 indicated a "strong signal for effect", a score between 50 and -50 indicates a "weak effect" and scores below -50 signify "no effect".

To overcome the limitations of single-gene reporter assays, a strategy employing mutation accumulation followed by whole-genome sequencing was developed. Rather than testing mutations in human cells, mutations were to be tested in budding yeast. The evolutionary conservation of Polε and Polδ makes this approach possible, as the routine methods for strain construction, short doubling time, expertly curated reference genome and low sequencing costs makes it advantageous. Alignment of human POLD1 and POLE with S. cerevisiae POL2 and POL3, respectively (Fig. 2.4), shows that most candidates can be constructed in yeast as the residues in question are conserved. Four variants from the list of DNA polymerase mutations
### 2.2 Identification of polymerase mutations

Figure 2.4: Alignment of polymerase residues of interest to the yeast proteins

Sequences were aligned using Clustal Omega version 1.2.1 [827–829]. Sequences used for alignment (uniprot ID in parenthesis): *Homo sapiens* POLE (Q07864), *Saccharomyces cerevisiae* POL2 (P21951), *Homo sapiens* POLD1 (P28340), *Saccharomyces cerevisiae* POL3 (P15436), *Schizosaccharomyces pombe* POL2 (P87154) and *Schizosaccharomyces pombe* POL3 (P30316). The residue identified as mutated in [768],[810] and [809] is encircled and unconserved residues are marked red. The amino acid change identified in the human samples is given at the top of each alignment.

<table>
<thead>
<tr>
<th><strong>POLD1</strong></th>
<th>Arg311Cys</th>
<th>Pro327Leu</th>
<th>Ser370Arg</th>
<th>Val392Met</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>P L R V L G I P E P V Q S Y E P D Y I T</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>P L R I M G V P E P I F S H A P D V I I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Schizosaccharomyces pombe</em></td>
<td>P L R I M G V P D P V Y E F Q P D V L I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>Gly426Ser</td>
<td>Ser478Asn</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>L K T V K A V S A H</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Schizosaccharomyces pombe</em></td>
<td>I H N F F A V C S Q</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>POLE</strong></th>
<th>Asp275Val</th>
<th>Pro286Arg</th>
<th>Ser297Phe</th>
<th>Val411Leu</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>A F D I E K F P D A M I S Y M R W V K R</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>A F D I E K F P D S M I S Y M R W V K R</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Schizosaccharomyces pombe</em></td>
<td>A F D I E K F P D S M I S Y M R W V K R</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Schizosaccharomyces pombe</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>POLE</strong></th>
<th>Leu424Val</th>
<th>Ala428Thr</th>
<th>Met444Lys</th>
<th>Arg446Gln</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>H N L K A A A A K A E D M C R M C R M A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>Q G L K A A V T Q S E L M T P M T P Y A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Schizosaccharomyces pombe</em></td>
<td>Q G L K A A V T Q S E L M T P M T P Y A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Schizosaccharomyces pombe</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>POLE</strong></th>
<th>Ala456Pro</th>
<th>Gln483Arg</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>T L A T Y Q P Q T K</td>
<td></td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>H L S E Y K P Q H L</td>
<td></td>
</tr>
<tr>
<td><em>Schizosaccharomyces pombe</em></td>
<td>V L A Q Y K P Q V L</td>
<td></td>
</tr>
</tbody>
</table>
### Table 2.5: Budding yeast equivalents of human DNA polymerase mutations of interest

Using protein alignments equivalents of human DNA polymerase mutations were determined when possible. In cases where the affected amino acid is not conserved, the amino acid found in the budding yeast protein at that position is given in brackets.

<table>
<thead>
<tr>
<th>Human variant</th>
<th>Conserved</th>
<th>S. cerevisiae variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>POLD1 p.Arg311Cys</td>
<td>Yes</td>
<td>pol3 p.Arg311Cys</td>
</tr>
<tr>
<td>POLD1 p.Gly426Ser</td>
<td>No, (T)</td>
<td>-</td>
</tr>
<tr>
<td>POLD1 p.Pro327Leu</td>
<td>Yes</td>
<td>pol3 p.Pro322Leu</td>
</tr>
<tr>
<td>POLD1 p.Ser370Arg</td>
<td>Yes</td>
<td>pol3 p.Ser375Arg</td>
</tr>
<tr>
<td>POLD1 p.Ser478Asn</td>
<td>Yes</td>
<td>pol3 p.Ser483Asn</td>
</tr>
<tr>
<td>POLE p.Met444Lys</td>
<td>Yes</td>
<td>pol2 p.Met459Lys</td>
</tr>
<tr>
<td>POLE p.Ala456Pro</td>
<td>No, (S)</td>
<td>-</td>
</tr>
<tr>
<td>POLE p.Ala428Thr</td>
<td>No, (T)</td>
<td>-</td>
</tr>
<tr>
<td>POLE p.Ala465Val</td>
<td>Yes</td>
<td>pol2 p.Ala480Val</td>
</tr>
<tr>
<td>POLE p.Arg446Gln</td>
<td>No, (P)</td>
<td>-</td>
</tr>
<tr>
<td>POLE p.Asp275Val</td>
<td>Yes</td>
<td>pol2 p.Asp290Val</td>
</tr>
<tr>
<td>POLE p.Leu424Val</td>
<td>Yes</td>
<td>pol2 p.Leu439Val</td>
</tr>
<tr>
<td>POLE p.Pro286Arg</td>
<td>Yes</td>
<td>pol2 p.Pro301Arg</td>
</tr>
<tr>
<td>POLE p.Ser297Phe</td>
<td>Yes</td>
<td>pol2 p.Ser312Phe</td>
</tr>
</tbody>
</table>

Table 2.5: Budding yeast equivalents of human DNA polymerase mutations of interest

To test, including the POLE A456P variant, were removed due to lack of conservation (Table 2.5).

### 2.3 Generation and propagation of polymerase mutants in *S. cerevisiae*

#### 2.3.1 Constructing single mutant polymerase strains

All polymerase mutations were introduced into a W303 MAT a haploid *S. cerevisiae* strain generating twelve single mutants and mating them to the isogenic Mat α strain generating heterozygous diploid strains. Point mutations were introduced by plasmid integration: two different plasmid constructs were made for POL2 and POL3(Fig. 2.5-A). Integration of each plasmid results in a functional copy of the gene carrying the mutation and a truncated, non-functional fragment(Fig. 2.5-B), C-terminal for POL3 and N-terminal for POL2.

To allow wild-type expression of the ensuing mutated POL2 gene, we also included 1kb of the upstream region containing the promoter. This does, however, lead to an N-terminal
truncation which is likely transcribed, but also targeted by nonsense-mediated decay (NMD). See YMH8-YMH41 in 6.3.2 for genotypes of all strains generated.

As reference, strains deficient for the proofreading activity of POL2 and POL3 were generated by introducing mutations in the exonuclease domain. As discussed earlier, the exonuclease domain is crucial for the preferential hydrolysis of non-complementary nucleotides at the 3’-terminus of a nascent DNA strand. Elimination of the exonuclease activity of yeast polδ or ε is known to result in a mutator phenotype and can thus act as a positive-control[339]. Three conserved amino acid motifs (called Exo I, II and III) in the N-terminal regions of the proteins form the active site of the exonuclease domain and are conserved in polymerases[313]. The alleles pol3-01 and pol2-4 (see Table 2.1) contain mutations of two acidic amino acids (one aspartic acid and one glutamic acid), thought to be involved in metal ion coordination, to alanines, which are known to affect proofreading, but not polymerase activity of these proteins (see red triangles in Fig. 2.6). I introduced these two point mutations using my plasmid constructs to generate haploid pol2-4 (YMH28) and pol3-01 (YMH32) equivalents.

### 2.3.2 Mutation accumulation experiment: Propagation of single mutant polymerase strains

There are several classical reporter gene assays to measure mutagenic activities in yeast. Assays measuring resistance to thialysine (Thia) or canavanine (Can) measure different types of mutation events inactivating the lysine permease (LYP1) or arginine permease (CAN1) genes, respectively[830, 831]. Beyond that other constructs have been used to study frame shifts (reversion of hom3–10 or lys2ΔBgl) [832]. Proxies for gross chromosomal rearrangements and aneuploidy events are also available[833, 834]. These assays have been instrumental in identifying mutator phenotypes (Table 2.1), but they do have considerable limitations. For instance, counting resistant colonies provides no measure of phenotypically silent, synonymous mutations. Furthermore, usually only a specific type of mutation in a single gene in a single locus of the genome is used as a proxy for the whole-genome, neglecting factors such as sequence composition and context, variable DNA damage and repair frequencies across the genome, as well as chromatin states and physical conformation of the DNA. Additionally, if one wanted to study the whole mutational spectrum, one would have to combine a vast array of assays to cover the entire catalogue of mutation types. Additionally, forward mutation assays do not allow the experimenter to distinguish between frameshifts and single base changes unless reporter genes are sequenced, which is labour intensive and relatively expensive. Recent work indicates that when compared to whole-genome sequencing measurements of particular muta-
Figure 2.5: Rationale for plasmid construction

A| Two different types of vectors were designed - one for POL2 and one for POL3 mutations - which contain a selectable marker and a fragment of the gene. The vector pRS306 was modified to generate appropriate integrating plasmids. This vector contains an ampicillin resistance for selection in *E. coli* and *URA3* for selection and counter-selection in *S. cerevisiae* and no centromere allowing integration after linearisation. The red arrows denote approximate sites for linearisation, the black vertical lines at either side of the vectors symbolise that they are circular.

B| Linearised vectors (here the N-terminal example is shown) will insert into the gene by HR creating a truncated gene as well as a functional gene fusion carrying the mutation introduced in the plasmid.
2.3 Generation and propagation of polymerase mutants in \textit{S. cerevisiae}

Figure 2.6: Exonuclease domains conserved in B family polymerases
Sequences were aligned using Clustal Omega version 1.2.1[827–829]. Sequences used for alignment (uniprot ID in parenthesis): \textit{Homo sapiens} POLE (Q07864), \textit{Saccharomyces cerevisiae} POL2 (P21951), \textit{Homo sapiens} POLD1 (P28340), \textit{Saccharomyces cerevisiae} POL3 (P15436), \textit{Enterobacteria phage T4} 43 (P04415), \textit{Enterobacteria phage RB69} 43 (Q38087), \textit{Bacillus phage phi29} 2 (P03680). The three exonuclease motifs (ExoI, ExoII and ExoIII) are underlined. The residues mutated to generate exonuclease deficient strains are highlighted by red triangles. The polymerase mutations in \textit{POLD1} are highlighted by black triangles, those in \textit{POLE} by blue triangles.
tions, some reporter assays provided reasonably accurate results, while others were not optimal proxies for the whole-genome[835]. With this in mind, I have decided to test the effects of the polymerase mutations by propagating the strains carrying mutated DNA polymerases and detecting mutations acquired during the process by whole-genome sequencing.

2.3.2.1 Single-colony bottleneck propagation of mutant polymerase strains

To obtain a significant number of mutations per strain, mutations were allowed to accumulate in parallel over 26 passages through single colony bottlenecks while cells were grown on non-selective rich medium for a total of three months. As illustrated in Fig. 2.7, in each case the starting strain was sequenced as well as each parallel line that was propagated. To determine the number of parallel lines needed to obtain sufficient mutations, I considered the fact that in a similar experiment, wild-type yeast cells accumulated on average 10.25 mutations after 100 passages[835]. Considering that for examination of mutational spectra, a significantly higher number of mutations is needed, the wild-type YMH9 strain was propagated in 72 parallel lines (projected to result in ~180 mutations in total), the YMH29 strain (carrying the pol2-4 variant) in 54 parallel lines and all others in 18 parallel lines (see Table B.1.2.1). The shorter time span (25 instead of 100 passages) is aimed to reduce any contributions from secondary arising mutations. However, even in the case of 100 passages (using the Canr assay) no change in mutation rate between starting and final strains was detected[835], suggesting that alterations in mutation frequencies are most likely due to the query mutation rather than secondary mutations.

2.3.2.2 Population bottleneck propagation of mutant polymerase strains

The main drawback of using single-colony bottlenecks, is that, if sequencing reveals an insufficient number of mutations, one cannot simply sequence more strains. Instead, the experiment would have to be repeated. As an alternative, the same strains as well as the haploid precursors (Table B.1.2.2) were propagated using population (10^4 cells) bottlenecks. This avoids the need for extensive parallel lines and more than one sample from the final population can be sequenced. Final populations can also be stored frozen and more colonies sequenced later. However, since these samples are not independent (as they are in the case of parallel lines with single colony bottlenecks) the actual number of independent mutations depends on the complexity of the final population.

In this experiment, the strains were propagated automatically by a serial-propagation platform in conjunction with our collaborators Ville Mustonen (WTSI) and Jonas Warringer (Uni-
2.3 Generation and propagation of polymerase mutants in *S. cerevisiae*

Figure 2.7: Mutation accumulation experiment: manual propagation of mutated *S. cerevisiae* strains

Experimental strategy: the heterozygous diploid polymerase mutant strains (all derived from the same wild-type W303 strain) were patched onto YPAD. From each patch 18 different parallel mutation accumulation lines were derived, by streaking small amounts of cells for single colonies on fresh YPAD plates. The remainder of the patch was frozen for later DNA extraction and serves as a starting points. The cells were grown to single colonies at 25°C (~20-25 generations) and cells were moved to a fresh plate using single-colony bottlenecks for 25 passages. Starting colonies and 2 colonies from each parallel line were whole-genome sequenced.
versity of Gothenburg, Sweden). This involved using a robot to transfer populations of cells onto new agar plates every two-three days for three months (see 6.7 and [836]). Twenty-eight colonies each for YMH8 and YMH9 (wild-type background) and YMH28 and YMH29 (pol2-4 mutation) and eighteen each for all other strains were sequenced.

2.4 Establishing sequence analysis practices

The majority of the work in this thesis uses the budding yeast *Saccharomyces cerevisiae* and next-generation sequencing. This chapter also describes the establishment of DNA sequencing analysis protocols in budding yeast and their application to other projects as a validation of the analysis strategy.

2.4.1 Automating genomic DNA extraction and whole-genome sequencing of *Saccharomyces cerevisiae* strains

Extracting high quality genomic DNA (gDNA) from yeast cultures by standard protocols is a low throughput method for extracting DNA for sequencing (see 6.6 for protocol). For the scale of this and other work a more high-throughput protocol for extracting gDNA was needed. Dr. Fabio Puddu, with the assistance of Nicola Geisler, developed a protocol to extract gDNA from 96 samples at a time using a robot, which I tested for sequencing by comparing the sequencing data I generated from samples extracted by phenol-chloroform extraction and those that were extracted using the robot (see 6.6 for protocols).

To assess whether the sequencing data obtained from DNA extracted using this high-throughput protocol was of similar high quality as the data acquired from DNA obtained by conventional phenol–chloroform extraction, samples subjected to either of these methods were compared for quality using key quality control measurements.

The Sequencing Facility at the Wellcome Trust Sanger Institute assesses all DNA for concentration, volume and total amount. From over 1000 samples prepared with the high throughput method 96% passed their quality control thresholds to proceed to library preparation and sequencing. For whole-genome deep sequencing, a mean genome-wide coverage of at least 30× is ideal and so far all samples that were sequenced after DNA extraction using this protocol have a coverage of at least that (Fig. 2.8-A). DNA sequencing of samples extracted using the high-throughput protocol is of comparable quality to sequencing of DNA extracted using phenol-chloroform in metrics regarding read alignment (Fig. 2.8-B), coverage of the entire genome (Fig. 2.8-C) and insert size distribution (Fig. 2.8-D) as well as GC content. Thus,
DNA extraction using this high-throughput extraction protocol allows us to obtain DNA of sufficient quantity and concentration for sequencing and the data obtained after sequencing compares favourably to previously sequenced samples in key quality measures. DNA extraction using this protocol was used for the remainder of this work.

2.4.2 Establishing sequencing analysis protocols for the identification of SNVs and INDELs

One of the main issues with identifying mutations from sequencing data is that one has to make decisions about which variants to retain as true variants and which to filter out as likely artifacts or errors, all the while usually not knowing what the true answer is. To tackle this problem, I developed variant calling and filtering strategies while continuously monitoring the approximate false negative and false positive rate under the supervision of Dr. Thomas.

Comparing to a capillary sequence reference The yeast reference genome was generated from a strain of the S288c background, whereas most of the strains featured in this work are of the W303 background, a strain generated in the 1970s. The genome of W303 is 85.4% identical to the S288c background and divergent sequences resemble those of Σ1278b. 799 proteins differ between the W303 and S288c strains, but most of the time only one or two residues differ[837]. Running variant calling and filtering on previously generated sequencing data from the Jackson lab of 22 strains from the W303 background, showed that, on average, MATa W303 lab strains carry 9,534 variants before filtering and 9192 after default filtering when compared to the S288c reference genome(Fig. 2.4.2). It also confirmed that the rad5-535 allele (a G535R missense mutation in RAD5 carried by the original W303 strain) has been corrected in our K699 and K700 strains.

The Saccharomyces Genome Resequecing Project completed ABI sequencing on a haploid W303 strain to a depth of between 1x and 3x which is freely available to download[838]. Compared to Illumina HiSeq data, ABI or capillary sequencing produces high quality long reads with a high degree of accuracy[774]. Comparing my W303 background data to the capillary sequencing data can provide some insight into the accuracy of my variant calling and filtering strategy. Due to the fact that they are not the exact same strain, discrepancies are expected, but I will be able to get an estimate for the false negative rate. False-positive rate estimates are much more problematic, due to the low coverage of the ABI sequencing data (2.3X), meaning that there will be regions of zero coverage, and won’t be calculated here.

Dr. Thomas Keane performed long-read alignment on the capillary sequencing data (see
Figure 2.8: DNA extracted using a high-throughput protocol produces high quality sequencing data

A | Mean genome wide coverage of 1577 samples sequenced after DNA was extracted using the high-throughput extraction protocol.  
B | Comparison of the percentage of reads that could be mapped to the reference genome and the percentage of reads that were paired between the 1577 samples whose DNA was extracted using the high-throughput extraction protocol and 168 samples extracted manually using phenol-chloroform.  
C | The same samples were compared for which fraction of the reference genome was sequenced to more than a depth of 5 and more than a depth of 10, respectively.  
D | Representative examples of insert size distributions for a high-throughput extraction (see 6.6) and a manual extraction (see 6.6) are shown.
Figure 2.9: The number of variants in W303 strains compared to the S288c reference genome. Aligned sequencing data from 22 *S. cerevisiae* strains of the W303 background strains was used to identify the number of background mutations to be expected when sequencing W303 *S. cerevisiae* strains. The samples are all control samples taken from other sequencing projects performed in the lab (see Table 6.3.2). Variant calling was carried out with samtools mpileup using parameters as specified in Table B.1.1.2 and filtering was done by vcf-annotate using its default filtering parameters. Total numbers of mutations per sample before and after filtering were counted and plotted.
Table B.1.1) and I performed variant calling as well as filtering on the ABI sequenced sample as well as 10 Illumina sequenced samples. Initially, when intersecting the variants called from the ABI sequencing with different samples of the Illumina sequenced set, we found 44.5%-50.8% of INDELs and 77.7%-78.5% of SNPs from the W303 Capillary data in the Illumina calls. Using the Integrative Genomics Viewer (IGV)[839] to look at the alignments in regions where the variant calling called a variant for the capillary sequencing data, but not the Illumina sequencing, suggested sensible ways to “tweak” the filtering step of the analysis. The alignments revealed that many of those variants were not captured due to mapping quality and depth filter thresholds (as well as many variants mapping to mitochondrial DNA) and adjustments of those reduced the approximate false negative rate to 2.3% meaning we can capture >97% of variants identified in capillary sequencing in the Illumina sequencing data. Running the GATK indel realignment tool to account for misalignment around an INDEL did not improve the calling sensitivity.

Simulated genome data Another, albeit imperfect, approach is to include simulated sample data in every analysis. Simulated data effectively avoids the issue of unknown results: the mutations in the samples are known and analysis should find them with minimal false negatives and false positive rates. The major shortcoming of the technique is, clearly, that it is simulated and can only approximate the realities of next-generation sequencing. Most of my project’s analysis will involve experimental samples and controls. Both sets of samples will have their variants called in relation to the reference genome and in order to identify the mutations experimental samples acquired during the experiment, mutations identified in control samples should be discarded from the experimental data (Fig. 2.10-A). This set-up is also reflected in the simulated data set we generated. Using pIRS (profile-based Illumina pair-end reads simulator)[840], several simulated samples were generated: control samples and experimental samples (containing all control sample mutations and additional ones). The control dataset had 8000 mutations inserted. This dataset was further mutated computationally to simulate experimental settings. The number of mutations to add was chosen considering the wild-type mutation rate (base-substitutional mutation rate: $0.33 \times 10^{-9}$ per site per cell division, [841]) and the suggested fold increase for a polymerase exonuclease deficient strain[768]. At the chosen parameter, around 200-300 SNPs were introduced. An INDEL dataset with 800 INDELs was also generated. After alignment and variant calling a false-negative and a false-positive frequency were determined. The false-negative frequency for SNV calls was 4-5.5% and 39.2% for INDELs. When the same adjustments for mapping quality, low depth and mitochondrial mutations as before were made, this number drops to less than 1% for SNVs
2.4 Establishing sequence analysis practices

Figure 2.10: Experimental strategy to identify acquired mutations

A | In most sequencing experiments performed in this work, single nucleotide variants and small INDELs have been called with respect to a reference genome for both a control (pretreatment) and an experimental (post-treatment) sample. The list of identified mutations will be intersected to identify mutations only present in the latter, giving us a list of acquired mutations.

B | In some cases a mutation may not be detected due to sequencing errors or filtering of low quality even though it is present in the DNA (see stricken out A mutation). This will lead to an apparent false positive in the list of acquired mutations (see bold A mutation).
and 12.6% for INDELs. The false-positive frequency for INDELs was found to be ~25%, whereas for SNVs the false-negative frequency was less than 1%. However, interestingly, not a single case of a true false positive was found (a variant call where no variant was present). Instead, variants that were mistakenly not called or filtered out from the control sample (false negative), could then not be removed from experimental samples creating effectively a false positive (Fig. 2.10-B). This highlights the case for more lax filtering to be applied to control samples and/or using more than one control sample to minimise the number of “false positives” generated this way. Sequencing was also simulated at different coverages (20X, 30X, 40X and 50X) and no difference in variant calling accuracy was found at these coverage levels.

### 2.4.3 Testing analysis protocol on *Saccharomyces cerevisiae* genetic screens

Screens in budding yeast have been used extensively and successfully to identify gene interactions. One example is synthetic lethality where two mutations result in lethality when co-occurring in one cell while cells carrying only one of the two are viable. Possibly more interesting are suppressor mutations (synthetic viability), where a mutation results in a phenotype which is reversed by a second mutation. While synthetic lethality can occur due to the inactivation of two parallel important pathways and not reflect true genetic interaction, suppressor mutations are often more informative about underlying molecular processes. Until recently, identifying a suppressor mutation involved laborious cloning of the suppressor loci. However, with the advances in sequencing technology and the associated reduction in costs, high-throughput synthetic viability genomic screening has become more and more feasible. To address a long-standing question in yeast DNA repair biology - the DNA damage sensitivities of sae2Δ cells - Dr. Tobias Oelschlägel performed a synthetic viability genomic screening identifying sae2Δ cells spontaneously resistant to camptothecin (CPT). 48 suppressor were sent for sequencing at the Wellcome Trust Sanger Institute as detailed in [801] and Chapter 6.8.

Since CPT is an inhibitor of DNA enzyme topoisomerase I (*TOP1*), stabilising the *TOP1*-DNA complex and resulting in replication-dependent DSBs, we expected that inactivating mutations of *TOP1* would likely be among the suppressor mutations. Such expectations, together with the fact that this project would likely involve confirming identified suppressor mutation with an orthogonal sequencing technology, this screen was ideal to test our analysis strategy. Together with Dr. Thomas Keane, I analysed the bwa-aligned bam files using the filtering strategy we developed in Chapter 2.4.2 (see Chapter 6.9 for more details). Similar to the set-up of my mutation accumulation experiments, this work involved sequencing a sensitive
2.4 Establishing sequence analysis practices

Figure 2.11: Sequencing analysis identifies mutations capable of suppressing sae2Δ DNA damage hypersensitivity

A | Outline of the screening approach that was used to identify suppressors of sae2Δ camptothecin (CPT) hypersensitivity. B | Validation of the suppression phenotypes; a subset (sup25–sup30) of the suppressors recovered from the screening is shown along with mutations identified in each clone. C | Summary of the results of the synthetic viability genomic screening (SVGS) for sae2Δ camptothecin (CPT) hypersensitivity. The ORF and the type of mutation are reported together with the number of times each ORF was found mutated and the number of clones in which each ORF was putatively driving the resistance. Figure and text reproduced from [801] in accordance with the terms of the Creative Commons Attribution License.
starting strain and multiple suppressors. Retaining only mutations found in the suppressors and not in the starting strain will ideally reveal the suppressor mutations. We found that 24 of the clones possessed *TOP1* mutations and, interestingly, 10 contained either *mre11-H37R* or *mre11-H37Y* mutations (Fig. 2.11). Further strengthening our hypothesis, that these were real suppressors, was the fact that *MRE11* and *TOP1* mutation never occurred in the same samples and, intriguingly, the 10 colonies with *MRE11* mutations were not just resistant to CPT, but also other DNA damaging agents: phleomycin, which generates DSBs, the replication inhibitor hydroxyurea (HU), DNA-alkylating compound methyl methanesulphonate (MMS) and ultraviolet light (UV). Follow-up work to characterize the *mre11-H37R* mutant and elucidate its role as a suppressor of *sae2*Δ-dependent CPT hypersensitivity was largely carried out by Dr. Fabio Puddu and the work has been published[801].

We have extended this method to other questions in yeast DNA replication biology. For instance, the absence of the Tof1/Csm3 complex causes hypersensitivity of cells to CPT. To identify mutations that can alleviate this hypersensitivity, Dr. Fabio Puddu carried out a suppressor screen as above for *sae2*Δ cells and sequenced 16 suppressors of *tof1*Δ cells’ hypersensitivity to CPT (Fig. 2.12). I performed the analysis as described above and in [801](see Chapter 6.9 for more details). Two of the strongest suppressors were found to have *TOP1* mutations. Two different inactivating nonsense mutations in the *SIR3* gene were found in three clones, while eight other suppressor clones carried a nonsense mutation in the *SIR4* gene. Further work by Dr. Puddu confirmed that inactivating members of the Sir complex mediated suppression of camptothecin hypersensitivity and this is likely due to disruption of sir-dependent heterochromatin. We suggest a model that Topoisomerase 1 inhibition in proximity of sir-dependent heterochromatin causes intense topological stress that leads to DNA hypercatenation, especially in the absence of the Tof1/Csm3 complex.

We have also applied this approach to phenotypes outside of replication in collaboration with other researchers and are pursuing the molecular mechanism behind the suppression of other replication stress associated phenotypes. This demonstrates that, not only does the bioinformatical analysis I carried out retrieve relevant mutations that we can confirm by other techniques in the lab, but, while designed for mutation accumulation experiments, it can also be applied to a wide variety of genetic experiments and will be used to generate biological insights beyond the realm of its initial conception.
Figure 2.12: Mutations in SIR3 and SIR4 identified as the cause for the hypersensitivity of tof1Δ cells to camptothecin.

(A) Loss of Tof1 and Csm3 but not Mrc1 causes hypersensitivity to camptothecin in a Top1-dependent manner. (B) Loss of pausing at the replication fork barrier on rDNA does not cause camptothecin hypersensitivity. (C) Outline of the procedure for a synthetic viability screen. (D) Synthetic viability screening identifies sir3 and sir4 alleles as suppressors of the camptothecin hypersensitivity of tof1Δ strains. (E) sir3 and sir4 deletions suppress the hypersensitivity of tof1Δ cells. (F) Deletion of SIR2 (encoding the third member of complexes containing Sir3p and Sir4p) also suppresses the hypersensitivity of tof1Δ cells and reduces the sensitivity of a wild-type strain. Drop tests were performed by Dr. Fabio Puddu, the assignments of mutations depicted in D were added by me.
Figure 2.13: Generation of mutagenized libraries
(a) Experimental workflow. (b) Schematic of 6-TG metabolism and genotoxicity. Inactivating mutations in the genes highlighted in red have been shown to confer resistance to 6-TG. (c) Number of suppressors recovered at increasing concentrations of ethyl methanesulfonate (EMS) treatment. (d) Mutation consequences identified by whole-exome sequencing of 7 suppressor clones.
2.4 Establishing sequence analysis practices

2.4.4 Applying analysis protocols to mouse genetic screens

The main advantage haploid yeast cells have for suppressor screens is that their haploid genome makes phenotypes, that would be recessive in a diploid, visible and selectable. Carrying out suppressor screens in diploid cells requires the appearance of dominant mutations, or mutations in both alleles of the same gene for a phenotype to be visible, making identification of suppressors more difficult. The success with next-generation sequencing of suppressors in haploid yeasts induced us to explore options in mammalian systems. Forward genetic screening in human cell lines has been feasible with the discovery of RNA interference (RNAi)[842], and more recently with insertional mutagenesis[843] and CRISPR/Cas9 libraries in near-haploid human cell lines[844–846]. And while loss-of-function (LOF) approaches like these are powerful, they have their limitations. Suppressor phenotypes caused by separation-of-function, gain-of-function or by mutations in essential genes[801, 847] are unlikely identifiable in these types of screens. The development of H129-3 haploid mouse embryonic stem cells (mESCs)[848] allowed us to circumvent the problems posed by diploid genomes. In collaboration with Dr. Josep Forment, haploid cells were treated with varying doses of the DNA-alkylating agent ethylmethanesulfonate (EMS) and 196 suppressors to the toxic nucleotide precursor 6-TG were isolated (Fig. 2.13-a). HPRT is known to initiate the cytotoxic mechanism of 6-thioguanine (6-TG) conversion to 2’deoxy-6-thioguanosine triphosphate (a cytotoxic nucleotide) in cells(Fig. 2.13-b)[849]. HPRT is thus a prime candidate for suppressor mutations since the loss of HPRT abolishes the cytotoxic effects of 6-TG. To test whether we could identify suppressors in the mouse genome, which is much larger than that of the budding yeast (2,716Mbp as opposed to 12Mbp in the reference genome), DNA from seven of these resistant clones and from a control mESC sample not treated with EMS was subjected to whole-exome sequencing.

Similar to the suppressor screen analysis detailed in Chapter 2.4.3, I performed variant calling (see Chapter 6.9 for details and Table B.1.1.2 for all parameters) on sequencing data aligned to the GRCm38 mouse reference genome by the Sanger Institute. I used my own scripts to remove any variants detected outside the bait regions and heterozygous variants where appropriate (see Chapter 6.9.5 for a list, description and location of Scripts). Low quality variants were filtered using standard and custom filters, variants present in the control sample were discarded and the remaining variants annotated for their functional consequences. Due to the much larger size of the genome, this alone proved not enough to remove all background mutations from the samples. This is likely due to the phenomenon described in Chapter 2.4.2, where false negatives in the control sample lead to an accumulation of apparent false positives in the data. To further filter the variants, Dr. Thomas Keane from the Vertebrate
Figure 2.14: Identification of suppressor mutations
(a) Genes harboring independent mutations in different clones. Mutations were assigned as deleterious or neutral according to PROVEAN and SIFT software. (b) Distribution of homozygous mutations identified in suppressor gene candidates; numbers of independent clones are in brackets and types of Hprt mutations are shown in detail. (c) Examples of sequencing reads obtained for heterozygous mutations affecting the Dnmt1 gene. SNVs causing missense mutations G1157E or G1157R (top panel) and G1477R or affecting the splicing donor sequence on intron 36 (bottom panel; see also Supp. Fig. 2), were never detected in the same sequencing read, indicating that they locate to different alleles. (d) Distribution of suppressor gene mutations identified, including heterozygous deleterious mutations.
Resequencing Team at the Sanger provided data from sequencing of a strain from the 129S5 background[850]. While this helped to dramatically reduce the number of likely incorrect single nucleotide variants (SNVs), the number of small INDELs remained unreasonably high, especially since EMS is a DNA-alkylating agent mainly producing SNVs. While SNV detection can generally be very reliable, INDEL detection has been less accurate[851, 852]. In order to retain only high-confidence INDEL variants I supplemented the alignment-based variant calling, with Scalpel, an INDEL caller that uses micro-assembly to identify INDELs and supports "somatic" mutation detection, whereby the algorithm will only report variants found in the sample, but not the control[853]. INDELs that were not identified by both callers were discarded from the dataset. This allowed a drastic reduction of the number of likely incorrect variants in our dataset (Fig. 2.15, for a more detailed description of the workflow see Chapter 6.9.6). Analysis of the 7 suppressors identified 189 different mutations that were either missense mutations, nonsense mutations, frameshift variants, inframe insertions or mutations affecting splice sites (Fig. 2.13-d).

To evaluate candidates for suppressor mutations, genes that were mutated in more than one sample, ideally carrying different mutations, were identified. To further aid in the determination of causative suppressor mutations, PROVEAN and SIFT[815–819] mutation prediction tools were used to evaluate mutations. Taking all these methods into account, the most striking candidate for a suppressor gene was, interestingly, Hprt (Fig. 2.14-a). In four of the samples three different missense mutations and one nonsense mutation were identified, and a fifth sample (D3) carried a mutation affecting a splice donor site, which can also have severe consequences at the protein level. While Hprt is a known suppressor gene, this clearly shows that even without prior knowledge of the 6-TG mechanism of action we would have identified Hprt as a candidate gene for suppression and we would have been able to assign causative mutations in 5 out of seven cases. In addition to Hprt, inactivating mutations of genes encoding for mismatch repair (MMR) proteins Msh2, Msh6, Mlh1 and Pms2 are also known to confer resistance to 6-TG[854], as well as mutations in DNA methyltransferase Dnmt1[855], and in fact the two remaining clones from our initial analysis of 7 carried nonsense mutations in Msh6 and Pms2.

To analyze the frequency of these mutations in suppressors, the remaining 189 suppressor clones were subjected to targeted sequencing of known suppressor mutations (see Table B.1.3). Deleterious mutations in most of these genes were identified (Fig. 2.14-b), confirming that if we had carried out whole-exome sequencing, as for the first 7 clones, we would have identified Hprt, Msh2, Msh6, Mlh1 and Pms2 as strong suppressor candidates, confirming that this approach is feasible for other screens with little or no prior knowledge of suppressor genes.
Effects of using more than a sequenced control sample to clear samples of background mutations and using more than one INDEL caller to enrich for high confidence INDEL calls using 74 WES mouse samples. SNVs are labelled green, INDELs are labelled blue and median values are represented as horizontal lines. From left to right the data shows successive intersection steps. "No intersection": Number of variants after variant calling and filtering to remove low quality variants are shown. These variants are mostly differences between the 129S5 and the reference background. ".- Ctrl": All variants also identified in an untreated mESC sample were removed from the samples. ".- Ctrl - 129S5": Additionally, any variants identified in a 129S5 background strain sequenced at the Sanger Institute were removed [850] ".- Ctrl - 129S5 + Scalpel": Since INDEL calling tends to be more error prone than SNV calling, we only included variants that were called by a second variant caller "Scalpel"[853].
2.4 Establishing sequence analysis practices

Figure 2.16: Clinically-relevant and newly-identified suppressor mutations
(a) Distribution of point mutations on Dnmt1, Hprt and MMR proteins; each square represents an independent clone. Asterisks (*) denote STOP-codon gains. (b) Predicted consequences of potential new suppressor mutations. Consequences were predicted as in Fig. 1e. (c) De novo introduction of new mutations Dnmt1 G1157E and Mlh1 A612T confers cellular resistance to 6-TG. (d) Hprt, Mlh1 and Msh6 mRNA expression levels (fragments per kilobase per million reads). Black dots indicate wild-type (WT) samples, red dots represent clones with already identified mutations (controls), and white dots represent samples for which no causative mutations were identified. Error bars represent uncertainties on expression estimates. (e) Reduced Hprt mRNA levels correspond to reduced protein production as detected by western blot.
sors. Intriguingly, a subset of clones presented heterozygous deleterious mutations in known suppressor genes. While these cells are sorted for haploid clones on a regular basis, diploid cells do remain and these particular cases could have arisen in the small diploid population or spontaneously after EMS treatment in a diploidized cell. Regardless, in order to be true suppressors these clones would each have to carry heterozygous mutations affecting both alleles of the gene, resulting in homozygous loss of the protein function. While our sequencing data is not phased, we have identified examples, where mutations occurred in such a way that they could be covered by a read (they are less than 150bp apart) or by the different members of a pair. Examples are shown in Fig. 2.14-c which demonstrate that these heterozygous mutations do not co-occur in the same reads indicating that, indeed, these cases are compound heterozygotes. Their scores in PROVEAN and SIFT predictions indicated that they are likely causing the 6-TG sensitivity suppression. When the clones carrying heterozygous mutations were also taken into account, we could also include Dnmt1 in the list of identified suppressor genes (Fig. 2.14-d).

When searching the literature, Dr. Josep Forment was able to assign many of the missense and nonsense variants to clinically-relevant mutations in Hprt (causing Lesch-Nyhan syndrome and its variants[856]) and DNA MMR (linked to Lynch Syndrome[803, 804])(Fig. 2.16-a), as well as previously not identified variants that are predicted deleterious(Fig. 2.16-b), highlighting the ability of this method to identify critical regions of a protein. Mutations affecting splicing donor and acceptor residues were also identified and confirmed by Dr. Josep Forment to reduce total protein level. To test whether some of the newly identified mutations are as deleterious as predicted, Dr. Josep Forment introduced the A612T and G1157E mutations in Mlh1 and Dnmt1 (which I identified as heterozygous mutations), respectively, into wild-type mESCs as homozygous mutations by CRISPR/Cas9 gene editing and showed that cells carrying these mutations were resistant to 6-TG treatment (Fig. 2.16-c).

For a small group of clones, no mutation in the targeted genes could be identified (Fig. 2.14-a,c) and we subjected the clones to whole-exome DNA sequencing and RNA sequencing (and included some in which we were able to identify potential causative mutations as controls). This allowed the production of an unprecedented description of EMS mutagenic preferences on the whole exome level, confirming its preference for producing SNVs, especially C:G>T:A transitions (Fig. S2.17-a,b,c), which could explain the high number of mutants affecting splice sites we recovered. While I was able to successfully retrieve previously identified mutations in the control samples, the DNA sequencing data identified no other obvious gene candidate. However, the RNA sequencing analysis carried out by Dr. Tomasz Konopka revealed significant reductions in expression levels of Hprt, Msh6 or Mlh1 in several clones
Figure 2.17: EMS mutagenic action
(a) Distribution of mutation types identified by whole-exome sequencing of 66 suppressor clones. SNV, single-nucleotide variant. INDEL, insertion or deletion. Only homozygous mutations were considered. (b) Distribution of identified SNVs. (c) EMS mutational pattern. (d) Number of mutations per chromosome in sequenced clones. Mutation numbers (both homozygous and heterozygous) were normalized to exon bait coverage.
Analysis of cancer-associated polymerase mutations

(Fig. 2.16-d), which could explain the 6-TG resistance of these samples. Further work may help to elucidate whether in such clones epigenetic alterations or mutations in regulatory regions not covered by exome-sequencing could explain the suppression mechanism in these clones.

Taken together, my work with Dr. Fabio Puddu and Dr. Josep Forment has shown, not only that we can exploit next-generation sequencing to unravel complex genetic interactions in haploid *S. cerevisiae* and mouse cells, with the potential to extend to human cells and essential gene biology, but also that our bioinformatical analysis is robust and recovers SNVs with high fidelity. Moreover, by using more than one variant caller strategy we can efficiently reduce INDEL false positive levels.

2.4.5 Establishing a sequencing analysis protocols for large genomic changes

We have established that this analysis can identify SNVs and small INDELs with a satisfactory sensitivity and accuracy. While polymerases with a low fidelity are not known for causing large-scale genomic rearrangements, a comprehensive genome analysis will address such changes. A structural variant (SV) is any form of rearrangement in chromosome structure and includes any or a combination of translocations, inversions, copy number variation (CNVs) as well as large insertions and deletions. These changes are critical as contributors to genetic diversity and evolution, but are also frequently involved in disease (see 1.2.1). Several methods exist to detect SVs such as microscopy-based chromosome banding and fluorescence in situ hybridisation (FISH), pulse-field gel electrophoresis, microarrays and sequencing-based mate-pair sequencing (sequencing the ends of large, kilobase-long DNA fragments) and whole-genome sequencing (WGS). Next-generation sequencing can detect many SVs by analysis of the mate pairs: for instance, in the event of a translocations the two mates of a pair (which by definition originated from the same DNA fragment) will align to different chromosomes of the reference genome and in the case of insertions or deletions the mate pairs will be much closer or further apart, respectively, than the average insert size dictates (Fig. 2.18). Since read pairs are a key source of evidence for SV detection, the quality of the underlying sequencing library is key and routine quality control (QC) of measures such as insert size is required. A second line of evidence for SVs can be split reads, reads that span a breakpoint and thus only align to parts of the reference in a continuous manner, and this depends highly on the alignment program and its ability to process split reads. A third source of evidence for SVs, especially CNVs is the read depth, following the assumption that an increase in copy number will be accompanied by a roughly proportional increase in coverage. There is a plethora of available
2.4 Establishing sequence analysis practices

Figure 2.18: Relationship between read pairs and structural variants

A | Schematic of Illumina paired-end sequencing: a fragment of DNA is sequenced from both sides inwards for 150bp (may vary depending on sequencing machine) leaving a fragment in the middle unsequenced. Its size depends on the library prep, but should be similar for all DNA fragments in the library. B | Large insertions and deletions: large insertions and deletions will be visible in the sequencing by alterations in the distance between the paired reads. C | An inversion most striking effect on a pair of reads is that they will now both be aligning to the forward or the reverse strand. D | In a translocation event members of a read pair may now align to different chromosomes.
SV callers that use one of those evidence sources (e.g. BreakDancer[857] uses read-pair information) or a combination (e.g. Lumpy[858] uses read pair, depth and split read information). Since SV callers can usually not detect the full spectrum of SVs and each one has advantages and limitations, Dr. Kim Wong in David Adams’ lab developed SVMerge a meta SV calling pipeline[859], which uses a variety of callers to make SV predictions (Fig. 2.19).

To complement the use of a program like this, we wanted to be able to visualise aneuploidy and large copy number changes in budding yeast WGS data. To this end, Dr. Puddu and I wrote compact scripts, that extract positional genome coverage data from bam files. The coverage values are normalised to the whole-genome median and ploidy information given by user input. As a control, this tool was used to visualise aneuploidy in a haploid strain that is diploid for Chromosome IX (Fig. 2.20).

### 2.4.6 Analysing repetitive DNA regions in the yeast genome

One of the biggest technical challenges facing NGS analysis are repetitive DNA sequences, sequences that are similar or often identical to other regions of the genome. That is especially problematic, because most genomes are abundant in repetitive sequences: about half of the human genome and >80% of the maize genome are covered by repeats[861]. From a computational point of view, repeats create uncertainty in alignments (as well as de novo assembly, which will not be further discussed), which can lead to errors when analysing sequences for genome variation. The main computational challenges are due to repeats that are >97% identical across more than one copy and that are longer than typical NGS read length (typically longer than 100-200bp).

After alignment of deep sequencing data, one major challenge remains: how to deal with reads that align to more than one location (multi-reads). In the human genome, the number of short reads (25bp or longer) that can be uniquely mapped tends to be around 70-80% even though the repeat content of the human genome is about 50 %[860]. This level of accuracy can be achieved due to the fact that repeats are often non-identical and many reads will have a unique "best match" (Fig. 2.21). "Best match" alignments are a simple way to resolve a significant portion of reads but this is not always correct[860]. Structural and copy number variant detection in unique regions has become relatively reliable, but the short read length of NGS sequencing data prevents similarly accurate detection in repetitive regions[860]. The most reliable sources for SV, coverage and read-pairs, pose more of a challenge in repetitive regions. Suppose an example of two transposable elements (TE), one on chromosome II another on chromosome V. Reads from either will relative equally be distributed between both
2.4 Establishing sequence analysis practices

Figure 2.19: An overview of the SVMerge pipeline

"SVMerge uses a suite of software tools to detect structural variants (SVs) from mapped reads. The calls are filtered, merged and then validated computationally by local de novo assembly. The output is in BED format, allowing for easy downstream analysis or viewing in a genome browser. The SVMerge pipeline is extendable so that calls made by other software can be included in the downstream analysis. BAM, Binary Alignment/Map format." Figure and text reproduced from [859] in accordance with the publisher’s terms of use.
Figure 2.20: Visualising aneuploidy in budding yeast
Output of script to visualise aneuploidy: normalised coverage plotted by position for each of *S. cerevisiae*'s 16 chromosomes. Here the DNA content of a haploid strain diploid for Chromosome IX is shown.
2.4 Establishing sequence analysis practices

Figure 2.21: Ambiguities in read mapping

A Read-mapping confidence versus repeat-copy similarity. As the similarity between two copies of a repeat increases, the confidence in any read placement within the repeat decreases. At the top of the figure, we show three different tandem repeats with two copies each. Directly beneath these tandem repeats are reads that are sequenced from these regions. For each tandem repeat, we have highlighted and zoomed in on a single read. Starting with the leftmost read (red) from tandem repeat X, we have low confidence when mapping this read within the tandem repeat, because it aligns equally well to both X1 and X2. In the middle example (tandem repeat Y, green), we have a higher confidence in the mapping owing to a single nucleotide difference, making the alignment to Y1 slightly better than Y2. In the rightmost example, the blue read that is sequenced from tandem repeat Z aligns perfectly to Z1, whereas its alignment to Z2 contains three mismatches, giving us a high confidence when mapping the read to Z1. B Ambiguity in read mapping. The 13 bp read shown along the bottom maps to two locations, a and b, where there is a mismatch at location a and a deletion at b. If mismatches are considered to be less costly, then the alignment program will put the read in location a. However, the source DNA might have a true deletion in location b, meaning that the true position of the read is b. Figure and Text reproduced from [860] with permission from the publisher.
when aligned. There are cases when the aligner will distribute members of the same pair on different chromosomes when the mapping quality is 0, suggesting a translocation where there is none. Also, suppose another example of these two TEs: the genome was sequenced to a mean depth of 30x and the two TEs show a coverage of about 60x. One may suppose that this means instead of two, this sample contains four copies of the TE. However, the coverage varies considerably across the genome, making the distinction between N and N+1 a low confidence proposition[860]. To cope with multi-reads (those with a reported mapping quality of 0) some prefer to discard them with unmapped read pairs and many SV detection programs ignore them in their analysis (though some allow the manual setting of mapping quality thresholds).

The budding yeast *S. cerevisiae* contains three major repetitive regions: ribosomal DNA (rDNA), Ty retrotransposons and telomeres. The rDNA genes encode ribosomal RNAs, major components of ribosomes, and rRNA makes up about 80% of RNA in budding yeast cells[862]. To cope with the high biosynthetic demand, eukaryotic cells tend to have hundreds of rDNA copies organised into clusters. In budding yeast, they exist in a single cluster located on chromosome XII (accounting for almost 2/3 of the chromosome’s length and 10% of the entire genome)[863]. Their highly repetitive nature makes the rDNA locus a highly fragile region of the genome and copies are continuously lost for example due to recombination events[862]. However, under normal conditions, cells can maintain a characteristic number of repeats and counteract loss by gene amplification (Table 2.6; see [864] for a review).

The maintenance of rDNA clusters involves many factors that are required generally for genome maintenance (such as replication, DNA repair and chromatin dynamics) and the de-

<table>
<thead>
<tr>
<th>Species</th>
<th>Copy number</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>150</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>55</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>−240</td>
</tr>
<tr>
<td><em>Xenopus laevis</em></td>
<td>−600</td>
</tr>
<tr>
<td><em>Gallus domesticus</em> (chicken)</td>
<td>200</td>
</tr>
<tr>
<td><em>Mus musculus</em> (mouse)</td>
<td>100</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>350</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>570</td>
</tr>
<tr>
<td><em>Pisum sativum</em> (pea)</td>
<td>3,900</td>
</tr>
<tr>
<td><em>Triticum aestivum</em> (wheat)</td>
<td>6,350</td>
</tr>
</tbody>
</table>

Table 2.6: Haploid copy number of rDNA repeats across Eukaryotic species
Selection of rDNA repeats observed in different eukaryotic species [865].
mand the rDNA cluster places on these factors means that perturbations in rDNA stability and copy number affect the availability of these factors in other regions of the genome[862]. Additionally, rDNA instability has been linked to aging in budding yeast[862] and a reduced copy number of rDNA repeats was shown to increase sensitivity to DNA damage[863]. Apparently, cells require a copy number of rDNA genes in excess of transcriptional demand to allow for DNA repair to proceed effectively[866]. In low-rDNA-copy-number cells the locus reportedly shows more genetic instability and this instability extends to other parts of the genome[862]. While the exact contributions and mechanism of the rDNA locus and its effects on genome instability and aging are still under active investigation, it is clear that reductions in rDNA copy number are detrimental to genomic stability and the cell as a whole and should be assessed when assessing effects of polymerase mutations on genome stability.

In *Saccharomyces cerevisiae*, the rDNA locus on Chromosome XII consists of approximately 150 repeats of a 9.1kbp unit (Fig. 2.22-A)[860] and when one plots the Illumina sequencing coverage along the chromosome the locus can be clearly identified as a sharp peak(Fig. 2.22-B). The rDNA unit contains genes for the 5S rRNA and the 35S rRNA, which are separated by two intergenic spacers (IGS1, 2). IGS2 contains the rARS, an origin of replication and IGS1 contains EXP, an expansion sequence made up of the replication fork barrier (RFB) and E-pro, a bi-directional promoter for non-coding RNAs that functions in regulating the rDNA repeat number[860]. The RFB ensures the unidirectionality of replication forks by the association with the protein Fob1[867], preventing head-on collisions between the replication and the transcription machinery in this highly transcribed region[867–869]. In the *S. cerevisiae* S288c reference genome assembly (R64-1-1/EF4) contains two copies of the 9.1kb rDNA repeat unit separated by the IGS1 to indicate the repetitive nature of the rDNA locus(Fig. 2.22-C). To estimate the amount of rDNA repeats present Dr. Fabio Puddu and I wrote a script, that measures the sequencing coverage in the first of the two rDNA unit copies and compares it to coverage upstream of the locus (Fig. 2.22-C). The upstream region was chosen because four copies of the 5Svariant, four copies of the ASP3, and a transposon are located downstream of the rDNA locus. In biology, the efficacy of any measurement method depends on two things: (1) when measuring the same sample more than once, does it give the same answer (technical reproducibility) and (2) how accurately does it measure the thing it purports to measure (how does it compare to other widely used measurement methods)? To answer the first question, Dr. Fabio Puddu and I sequenced 116 yeast strains twice starting from the same genomic DNA. The results show a strong correlation between the two independent measures (Fig. 2.22-D), with the divergence between the two measures increasing with the size of the rDNA locus but remaining almost always contained within +/-5% of the
Figure 2.22: Next-generation sequencing data can be used to estimate rDNA copy number reliably.

A | rDNA exist in ~150 repeats of a 9.1 Kbp unit on Chromosome XII. B | The coverage of Illumina sequencing reads across Chromosome XII. C | The *S. cerevisiae* S288c assembly (R64-1-1/EF4) contains two copies of the rDNA unit with an origin of replication each (red circle) separated by one copy of the replication fork barrier (RFB, red circle with white bar). Measurements of rDNA enrichment are derived from coverage over rDNA repeats (blue box) over the average genomic coverage. D | Reproducibility of the measurement using the same sample. E | Accuracy of the measurement: strains stable for their rDNA copy number with estimates of rDNA copy number by pulse-field gel electrophoresis (20, 40, 60, 110, and 150 repeats) [866] had their rDNA copy number estimated using NGS coverage. F | rDNA copy number measured in wild-type W303 and BY4743 laboratory strains that are diploid, MATa and MATa. The W303 diploid strain was sporulated, four tetrads (biological replicates) were dissected and the resulting haploid cells sequenced and assessed for rDNA copy number (fall four spores of the same tetrad are labelled in the same colour).
2.4 Establishing sequence analysis practices

average of the two measures. To answer the second question and assess the precision of this method four colonies each derived from 5 yeast strains carrying stable rDNA loci of known length were sequenced[866]. The rDNA copy number was estimated in these strains by Ide et al. using pulsed-field gel electrophoresis (PFGE) to be 20, 40, 60, 110 and 150 repeats, respectively. Fig. 2.22-E shows that estimating rDNA loci size using whole-genome sequencing produces results in agreement with PFGE and considering that PFGE also produces estimates at best, WGS estimates of rDNA copy number perform just as accurately. By employing this method, I found that wild-type laboratory strains in the W303 background have consistently bigger rDNA loci (~180 units) compared to wild-type strains in the BY4743 background (~120 units) (Fig. 2.22-G, left). We also observed that in both backgrounds, haploid strains of the mating type a, seemed to have slightly bigger loci than the corresponding Matα strains (Fig. 2.22-G, left). To determine if this is generally true, we sporulated a wild-type W303 strain and analysed the rDNA length in the progeny. In these conditions, Mat a and Matα strains did not show any significant difference between each other, but they showed a greater variability in rDNA length. When the four spores coming from a single meiotic event (marked with the same color in Fig. 2.22-G, right) show rDNA loci of different size, the data observed are compatible with Mendelian inheritance of this trait, in the presence or in the absence of unequal sister chromatid exchange. In sum, this tool can be used as a read-out in a screen identifying genes that regulate and maintain rDNA copy number.

Beyond the rDNA locus we have investigated the other two repetitive DNA regions in the yeast genome: Ty elements and telomeres. Ty elements are retrotransposons pervasive in the yeast genome (3.1% of the genome) characterised by their flanking long terminal repeats (LTRs). There are five distinct retrotransposon families (Ty1–Ty5)[870]. Their success at colonising the yeast genome varies greatly and while the numbers observed can vary greatly, the consensus is that Ty1 elements occur most and Ty5 elements least often[870–872]. Considering these fluctuations and that, in principle, Ty elements are very similar to the rDNA locus in that their length greatly exceeds read length, we extended our approach to measuring Ty element copy number. Because, unlike rDNA, Ty elements are spread throughout the genome, a custom "Ty reference genome", a fasta file principally made up of the Ty element sequences and surrounding control sequences, was constructed and NGS reads aligned to it. This redistributes Ty element reads back onto a single locus allowing estimates of Ty element number within the genome (Fig. 2.23). In principle this approach works, but as of yet we have not completed sequencing of strains to show that our approach accurately determines Ty element number (akin to Fig. 2.22-E), but the general trend of Ty elements reported in the literature is reflected in our measurements. Telomeres provide a greater challenge to copy
Figure 2.23: Next-generation sequencing data could also be used to assess Ty element copy number. To estimate copy number of Ty elements a custom “reference genome” was built mainly consisting of a single copy of each Ty element and control surrounding sequences. NGS data was aligned to this reference and coverage plots normalised to the genomic median are shown.
number estimation. Their repeat size is smaller than a single read and, thus, requires another approach for telomere length estimates. While a program exists for the estimation of human telomere length from NGS data\cite{873}, it relies entirely on the fact that the human repeat is invariable (a TTAGGG tandem repeat). In contrast, the \textit{S. cerevisiae} telomere repeat is degenerate with the consensus sequence G_{2-3}(TG)_{1-6}\cite{874}. This poses a great challenge to budding yeast telomere length measurements using next-generation sequences. Together with Zhihao Ding, I have been trying to adjust his program to measure \textit{S. cerevisiae} telomere repeat number, but so far we are still underestimating yeast telomere length, likely because we don’t capture the degenerate nature adequately yet.

In summary, together with Dr. Puddu, I developed a simple program to measure rDNA repeat number in the budding yeast \textit{S. cerevisiae} and we can show that our method is a suitable alternative to classic laboratory approaches. We are now employing this method as a tool in our array of methods to document genomic changes, but are also using it to identify factors involved in rDNA copy number maintenance (see Chapter 4). Work to accurately estimate Ty element number and telomere length in budding yeast is ongoing.

\section*{2.5 Summary}

During this phase of my work, I compiled a list of mutations in DNA polymerases delta and epsilon identified in sequencing of human cancer samples. After assessment of the occurrence of these in the wider population, the corresponding residues in the budding yeast \textit{S. cerevisiae}’s replicative polymerases were identified and those that affect residues that are evolutionarily conserved were retained. These mutations were then introduced into yeast cells and mutation accumulation experiments performed where cells were propagated to let any effects of polymerase mutations manifest in the genome. DNA was extracted at the beginning and end of these experiments and sent for whole-genome sequencing. In the meantime, I developed and tested a sequencing analysis strategy using existing datasets that had the advantages of positive controls and follow-up validation. This allowed the development of accurate protocols and tools for identifying SNVs, small INDELs, changes in rDNA repeat number and to a lesser extent structural variants. In the process, I contributed to projects unraveling complex genetic interactions in budding yeast and the proof-of-concept application of our yeast synthetic viability screens to mouse genetics.
Evaluation of hypotheses

Aims:

• To compile a list of relevant mutations in DNA polymerases identified in cancer samples

  A list of DNA polymerase mutations found in colorectal and endometrial cancers was assembled from the literature.

• To prioritise mutations in DNA polymerases and determine their *Saccharomyces cerevisiae* equivalents

  Recurrence in cancer sample, bioinformatic predictions and the alignment of the human and yeast protein sequences identified a list of mutations with priority for the variants POLE S297F, POLE P286R and POLE V411L, which were tested for effects such as mutation rate increases first.

• To conduct mutation accumulation experiments to identify the consequences of DNA polymerase mutations on a genome wide scale

  After construction of all remaining mutations in budding yeast, they were subjected to mutation accumulation experiments for three months in several parallel lines. Starting and final yeast colonies were sent for whole-genome sequencing to identify acquired mutations and characterize any changes in numbers, locations and patterns compared to wild-type.

• To establish sequence analysis protocols for budding yeast whole-genome sequencing data

  Whole-genome sequencing data analysis in budding yeast was developed for single nucleotide variants, insertions/deletions, aneuploidy and copy number changes in repetitive regions. Determinations of false negative and false positive rates were estimates and measuring rDNA repeat copy number was validated with published southern blot data.

• To show that these sequence analysis protocols are functional and can be applied beyond this project

  My whole-genome sequencing analysis protocols were applied to suppressor screens in budding yeast and identified both expected mutations (for instance mutations in TOP1 as suppressors for camptothecin sensitivity), which act as a positive control, and previously unknown mutations, which were shown to be biologically relevant by further
experiments. Taking this work successfully into suppressor screens with haploid mouse embryonic stem cells shows that overall this analysis protocol is robust, produces validated results and can be used for applications beyond its initial conception.
Chapter 3

Analysis of populations of *S. cerevisiae* strains carrying simple polymerase mutations

Overarching hypothesis

DNA polymerase mutations can contribute to tumour progression likely by elevating the mutation rate.

**Aims:**

- To assess all candidate DNA polymerase mutations for the number of mutations acquired in the same amount of time
- To identify the type of mutations caused by mutated DNA polymerases
- To determine whether candidate mutations leave a distinct mutation pattern on the genome
- To compare mutations acquired in mutated polymerase strains to those resulting from mismatch repair deficiency

3.1 Introduction

In the previous chapter, a list of mutations in DNA polymerases identified in colorectal and endometrial cancer was collated and after alignment to the yeast proteins, where possible, those
mutations were introduced into diploid *Saccharomyces cerevisiae* as heterozygous mutations. Determining the effect of these mutations on a genome - for instance to identify mutations in DNA polymerases that raise the mutation rate - will assist in differentiating between passenger mutations and those that promote tumourigenesis. To this end, strains carrying these mutations were subjected to mutation accumulation experiments and whole-genome sequenced before and after the experiment to collect information about acquired mutations in this timeframe. Analysis protocols for budding yeast whole-genome sequencing data were developed and tested and then used to analyse mutation accumulation experiment data.

3.2 Increased mutation rates for strains heterozygous diploid: *pol2-P301R, pol2-S312F, pol2-L439V, pol2-M459K* and *pol3-S483N*

3.2.1 Increased number of single-nucleotide variants for a subset of polymerase variants

After propagation, the polymerase mutant strains were sequenced at the Wellcome Trust Sanger Institute (see Chapter 6.8) and the sequencing data aligned to the yeast reference genome. Variant calling for single-nucleotide variants and small insertions and deletions was performed to detect any changes in mutation accrual. First, samples were checked for their polymerase genotype: any sample not identified to carry the expected polymerase mutation was discarded from the dataset. While it is possible, that a missing polymerase mutation is a case of a false negative, these samples were discarded in case of contamination or early reversion of the mutation. The remaining samples were analysed as described in Chapter 1.4.3: filtered samples were intersected with the initial starting strains to remove any background mutations and only retain those mutations acquired during the experiment (for a detailed description of the analysis workflow, the software, scripts, and commands used see Chapter 6.9.5 and 6.9.6).

Strains not carrying a mutation in the DNA polymerases acquired a mean of 6.8 single-nucleotide variants (SNVs) during the course of the experiment translating to roughly $5.5 \times 10^{-10}$ SNV mutations per generation per base (assuming 500 generations), which is consistent with recently published mutation rates in vegetative diploid *S. cerevisiae*[875]. The exonuclease deficient *pol2-4* mutants acquired on average 11.8 SNVs, meaning 1.7× the number of mutations as the wild-type(Fig. 3.1). Of the *pol2* candidate mutations four showed significant
3.2 Increased mutation rates for strains heterozygous diploid: *pol2-P301R*, *pol2-S312F*, *pol2-L439V*, *pol2-M459K* and *pol3-S483N*

![Figure 3.1: Number of single-nucleotide variants per sample in *pol2* mutant strains](image)

*Pol2* mutant strains were whole-genome sequenced before and after a three-month propagation on rich medium. The number of acquired single-nucleotide variants was determined for each parallel line that was propagated. The number of samples for each strain are as follows: n=65 (*POL2*), n=49 (*pol2-4*), n=18 (*pol2-A480V*), n=17 (*pol2-S312F, pol2-V426L, pol2-L439V*), n=16 (*pol2-M459K*), n=15 (*pol2-P301R, pol2-Q468R*), n=14 (*pol2-D290V*). All strains are heterozygous diploid for the mutation in question. The median is denoted by a black line. Student’s T-test was used to determine which samples are significantly different from wild-type at *** p < 0.001.
increases in mutation accrual: pol2-P301R (27.9×), pol2-S312F (4.3×), pol2-L439V (3.4×) and pol2-M459K (10.3×). Strikingly, their increase in mutation number exceeds that observed for the exonuclease deficient strain. Also, interestingly, while POLE p.Val411Leu is one of the most frequently observed mutations in POLE in sequenced cancer samples (Fig. 2.3), the equivalent budding yeast variant, pol2-V426L, does not lead to an increase in SNV accumulation. Whether this holds true for the human mutation remains unclear.

In the case of pol3 mutants, the exonuclease deficient strain also shows increased mutation accumulation when compared to wild-type with an average of 22.3 SNVs per strain (3.3×, Fig. 3.2). Of the pol3 candidate mutations tested, one, the pol3-S483N strain, accumulated a mean of 230 SNVs per strain, meaning it accumulated 33.3× the number of SNVs as the wild-type strains. Again, this is a striking increase compared to the mutational increase observed in exonuclease deficient cells. Why these mutations in the exonuclease domain produce an effect stronger than mutating the catalytic residues of this domain is currently unclear.

### 3.2.2 Single-nucleotide variants in haploid polymerase mutant strains

In the heterozygous diploid strains, the effects of the polymerase mutations are likely mitigated by the presence of a wild-type copy of the polymerase on the other chromosome. In a haploid setting, only the mutated polymerase would be present and the genome would be half the size. Theoretically, if the polymerases - the mutated and wild-type one - are available in cells at similar levels and share the burden of copying the genome equitably in the heterozygous strains, then in the haploid strain the mutant polymerase would copy roughly the same amount of DNA each division. To examine mutation numbers in haploid strains, four of the pol2 mutant strains - pol2-P301R, pol2-S312F, pol2-A480V and pol2-M459K - were propagated as haploids alongside the wild-type and exonuclease deficient strain for 13 passages using single colony bottlenecks as described in Chapter 2.3.2.1.

Fig. 3.3 depicts the SNVs per haploid genome accumulated in each line after the propagation for the haploid and heterozygous diploid strains. While the number of SNVs accumulated in wild type strains is fairly similar between haploid and heterozygous diploid strains, a difference can be seen for all pol2 mutant strains (see Fig. 3.3 and Table 3.1). For some, such as pol2-S312F, the fold change in mutation numbers compared to wild type is 14× bigger in the haploid than in the heterozygous diploid, suggesting the increase in mutation accrual is not likely simply due to a wild type polymerase replicating half the genome with high fidelity. Similarly, while pol2-4 and pol2-A480V show similar mutation numbers in a heterozygous diploid setting, the absence of a wild type polymerase and half the genome have a markedly
3.2 Increased mutation rates for strains heterozygous diploid: pol2-P301R, pol2-S312F, pol2-L439V, pol2-M459K and pol3-S483N

Figure 3.2: Number of single-nucleotide variants per sample in pol3 mutant strains
Pol3 mutant strains were whole-genome sequenced before and after a three-month propagation on rich medium. The number of acquired single-nucleotide variants was determined for each parallel line that was propagated. The number of samples for each strain are as follows: n=65 (POL3), n=18 (pol3-S483N, pol3-S375R), n=17 (POL3, pol3-01, pol3-R316C), n=16 (pol3-P322L). All strains are heterozygous diploid for the mutation in question. The median is denoted by a black line. Student’s T-test was used to determine which samples are significantly different from wild-type at *** p < 0.001.
Analysis of populations of *S. cerevisiae* strains carrying simple polymerase mutations

Figure 3.3: Number of single-nucleotide variants per line per haploid genome for selected haploid and heterozygous diploid *pol2* mutant strains

The number of detected single-nucleotide variants (SNVs) is plotted for heterozygous diploid strains similar to in Fig. 3.1: each dot is the measurement of an independent line. To account for differences in genome size, measurements were normalised to a haploid genome. For haploid strains the number of detected SNVs after 13 passages is depicted in the same manner. The black bars indicate mean and standard deviation.
3.2 Increased mutation rates for strains heterozygous diploid: *pol2-P301R, pol2-S312F, pol2-L439V, pol2-M459K* and *pol3-S483N*

<table>
<thead>
<tr>
<th></th>
<th>HAPLOID</th>
<th>DIPLOID</th>
</tr>
</thead>
<tbody>
<tr>
<td>pol2-4</td>
<td>6.1x</td>
<td>1.7x</td>
</tr>
<tr>
<td>pol2-A480V</td>
<td>17.1x</td>
<td>1.7x</td>
</tr>
<tr>
<td>pol2-M459K</td>
<td>54x</td>
<td>10.3x</td>
</tr>
<tr>
<td>pol2-P301R</td>
<td>89.3x</td>
<td>27.9x</td>
</tr>
<tr>
<td>pol2-S312F</td>
<td>60.6x</td>
<td>4.3x</td>
</tr>
</tbody>
</table>

Table 3.1: Mutation number fold change of *pol2* haploid and heterozygous diploid mutant strains when compared to the *POL2* strain. After propagation, the number of detected single-nucleotide variants in the *pol2* mutant strains is normalised to the number detected in the wild-type *POL2* strain of the same ploidy.

3.2.3 *pol2* mutants grow at a similar rate to wild type strains

To correlate acquired mutation numbers over time with a mutation rate, the cells would need to grow at a similar rate and undergo a similar number of divisions in a given amount of time. In order to test whether the heterozygous diploid polymerase mutant strains grow at similar rates to the wild type, I monitored cell growth rates by measuring the absorbance at 595nm wavelength culturing cells in rich medium from stationary phase for 450 minutes.

All *pol2* heterozygous diploid polymerase mutant strains grow similar to the *POL2* wild-type strain suggesting that the mutation numbers obtained at the end of the mutation accumulation experiments can be compared and that any increases in mutation rate cannot be explained by differences in proliferation speed (Fig. 3.4).

3.2.4 Correlation of mutation rate estimates with mutations accrual

Additional to the propagation experiments, mutation rate estimates were also obtained using the resistance to thialysine (Thia\(^r\)) (Table 3.2.4). While Thia\(^r\) measurements are much more variable than mutation accumulation experiments (see Table 3.2.4 and Fig. 3.1 & 3.2), there is a positive linear relationship (R\(^2\) = 0.912) between mutation rate estimates using Thia\(^r\) and the number of SNVs detected by mutation accumulation experiments followed by NGS (Fig. 3.5).
Analysis of populations of *S. cerevisiae* strains carrying simple polymerase mutations

Figure 3.4: Growth of *S. cerevisiae* mutant strains in rich medium

Cell growth of heterozygous diploid mutant strains in rich medium was monitored by measuring absorbance at 595nm in a spectrophotometer. Cells were grown to saturation overnight at 30°C and released them into fresh rich medium at a dilution of 1:200. Measurements were taken every 30 minutes for 450min. Data from one experiment shown.

Table 3.2: Estimates of mutation rate increases using resistance to Thialysine

Numbers of resistant colonies were obtained from seven independent cultures for each tested strain. Fluctuation analysis was used to determine median mutation rate estimates as well as a 95% confidence interval (in brackets). Fold change with respect to the wild type is given.
3.2 Increased mutation rates for strains heterozygous diploid: *pol2-P301R, pol2-S312F, pol2-L439V, pol2-M459K* and *pol3-S483N*

Figure 3.5: Correlation of mutation rate estimates and mutation accrual for *pol2* mutant strains
For all *pol2* mutant strains and a wild type control, mutation rate estimates for haploid mutants obtained using thialysine resistance were plotted against the number of mutations detected after propagating heterozygous diploids for 25 passages. A regression line was drawn and the regression coefficient $R^2$ is given.
3.3 Patterns of single-nucleotide variants

The recent advances in identifying mutational processes and the patterns that they leave in cancer cells - mutational signatures - have focussed on base substitutions considering the affected base as well as the ones immediately upstream and downstream. There are six different kinds of single-nucleotide mutations (Fig. 1.26), which leads to 96 different triplet changes.

The mutations identified in the polymerase mutants can be visualised in the same format to show whether any preferences for certain mutations exist. The mutation pattern for the wild-type and pol2-P301R and pol3-S483N is shown, because of the high mutation accumulation in those samples (Fig. 3.6). These patterns are further normalised to the occurrence of triplets in the genome to show mutational preference independent of abundance of each triplet (Fig. 3.7). The pol2-P301R strain shows a stark preference for CTC>CAC mutations as well as for ACA>AAA and TCT>TAT. Adjustment to triplet occurrence in the genome makes the latter two less prominent, but highlights the enrichment for CTC>CAC mutations considering the low abundance of CTC triplets in the *S. cerevisiae* genome (Fig. 3.7-A). While the pol3-S483N pattern is a lot more similar to the one observed in the wild-type, adjustment to the genome-wide triplet distribution highlights a preference for T>C mutations, especially ATC>ACC and CTC>CCC changes.

Cancer cell mutational profiles are usually much more complex than this in that they are composites of often many mutational processes acting at different times, for varying lengths with diverse intensities. To deconvolute this multidimensional dataset and identify common underlying patterns several mathematical approaches have been employed, among them principal component analysis (PCA) and non-negative matrix factorization (NMF). NMF is a method from linear algebra allowing the deconstruction of a matrix into two smaller matrices, whose product approximates the original matrix, with the property that all values be non-negative[876]. One of its most well-known uses is to use NMF to decompose an object into its parts: NMF can be successfully used to represent faces as a composite of eyes, mouths, noses and so on or can be used to find semantic features in an encyclopedia and recombine them to reconstruct encyclopedic features[876]. In many ways, identifying underlying patterns of mutations in sequenced cancer samples is analogous to these examples and NMF has been used successfully to extract mutational signatures from cancer data and can also estimate the relative contribution of each mutational signature to a particular cancer[700]. NMF can be computationally expensive when the dataset is very large in which case PCA is more suitable. Considering the size of my dataset, NMF is a suitable method for this analysis.

To extract the mutational signatures from the sequencing data, the SomaticSignatures[763]
3.3 Patterns of single-nucleotide variants

The mutational landscape of single-nucleotide variants found after propagation in wild-type strains (A), *pol2-P301R* strains (B) and *pol3-S483N* (C) strains was visualised by considering the base change itself (all changes were transformed to have a pyrimidine base) as well as the immediately flanking residues. For all variants the sequence context was extracted based on the genomic location within the reference sequence. The trinucleotide changes that diverge most from the wild-type as tested by $\chi^2$ are marked by a “red dot”.

![Figure 3.6: Single-nucleotide variant patterns](image)
Figure 3.7: Single-nucleotide variant patterns adjusted to frequencies of trinucleotides in *S. cerevisiae*

A| The abundance of each triplet in the W303 genome was determined. Triplets with a central cytosine are shown in blue, those with a central thymidine are shown in green. B-D| The mutational landscape of single-nucleotide variants found after propagation adjusted for the genome wide occurrence of triplets in budding yeast. * The scale ends at 0.05, but the CTC>CAC mutation was measured to contribute at an adjusted value of 0.085.
3.3 Patterns of single-nucleotide variants

Figure 3.8: SomaticSignatures: Determining the numbers of signatures
The residuals sum of squares (RSS) and the explained variance between the observed matrix and fitted mutational spectrum for 2 to 8 signatures. The number of signatures can be chosen so that the addition of one more signature does not yield a sufficiently better approximation to the data. The first inflexion point has been suggested as an appropriate measure for the number of signatures [877].
package was used to apply NMF to the data and plot the results (see 6.9.4). The algorithm
determined the residuals sum of squares (RSS) and the explained variance for 2 to 8 signatures
(Fig. 3.8). The RSS is a measure of the discrepancy between the data and the model and a
small RSS and a large value for explained variance indicate a tight fit of the model to the data.
The likely number of signatures can be chosen by looking for the number where little improve-
ments in RSS and explained variance are made by adding another signature. The first inflexion
point has also been proposed as a measure to determine the number of signatures[877]. The
values for RSS and explained variance obtained for the mutator strains displayed in Fig. 3.8 in-
dicates that two signatures explain nearly 94% of the variance. Adding a third signature would
explain roughly another 3.5% of the variance, while adding a fourth signature only improves
the explained variance from 97.5% to 99%. The algorithm thus extracted two-three signatures
from the aggregated data of all sequenced pol2 and pol3 heterozygous diploid mutant strains
and the wild-type control. In the case of two signatures, Signature 1 shows a striking peak in
the C>A mutations (the same TCT>T A T also preferred in pol2-P301R samples), while Signa-
ture 2 is very similar to the pattern observed for the wild-type and the pol3-S483N strain(Fig.
3.9-A). Contributions of each of the two signatures to the mutation patterns observed in each
strain are also estimated by SomaticSignatures(Fig. 3.10-A). In accordance with the SNV
patterns displayed in Fig. 3.6 and the mutation accrual displayed in Fig. 3.1 and Fig. 3.2,
Signature 1 is estimated to mainly contribute to mutations in pol2 mutant strains. For the four
strains with significantly increased mutation numbers, pol2-P301R, pol2-S312F , pol2-L439V
and pol2-M459K, Signature 1 has an estimated contribution of almost 100%. For pol3-S483N,
a strain with a significant increase in mutation numbers, the Signature 1 contribution is approx-
imately 10%, in line with the observed mutation pattern in these strains. When determining
three signatures, Signature 1 remains unchanged, while the previous Signature 2 is split into
two distinct signatures (Fig. 3.9-B). In this case Signature 2 is the main contributor to mu-
tations in the pol3-S483N strain, while Signature 3 is the main contributor to the wild-type
strain, which is consistent with it contributing to half the mutations acquired by the pol2-4
strain(Fig. 3.10-B).

To obtain further evidence for the number of likely signatures, another signature extraction
algorithm EMu was applied to the collection of acquired mutations identified in Section 3.2.
EMu identifies the number of mutational signatures using expectation-maximization (EM) and
model selection criteria, such as the Bayesian information criterion (BIC)[764]. EMu identi-
fied three Signatures in the data. When comparing the output of EMu and SomaticSignatures,
Signature 1 and Signature B are similar and show similar contributions to sample mutations,
while Signature 2 and Signature A as well as Signature 3 and Signature C show similarities
3.3 Patterns of single-nucleotide variants

Figure 3.9: 2-3 signatures are determined using Non-negative matrix factorization. Two and three signatures were extracted from mutation data of all mutator strains and the wild type combined using SomaticSignatures with NMF. The signatures extracted are displayed in the 96 trinucleotide-change channel format indicating each mutation’s contribution to the overall pattern. **A** Two signatures extracted from mutator strains. **B** Three signatures extracted from mutator strains.
Two and three signatures were extracted from mutation data of all mutator strains and the wild type combined using SomaticSignatures with NMF. The contribution of each of the two and three signatures to the mutational landscape of each strain is displayed. A | Contributions of two signatures to mutations in both samples. B | Contributions of three signatures to mutations in both samples.
(Fig. 3.3).

As part of the Catalogue of somatic mutations in cancer (COSMIC), human exome and whole-genome sequencing data from cancers was subjected to mutational signature extraction [700, 738, 761, 762, 878]. Currently, the collection holds 30 Signatures assembled from “an analysis of 10,952 exomes and 1,048 whole-genomes across 40 distinct types of human cancer”[879]. To understand how these signatures from human cancers compare to those extracted from the yeast samples in this work, the similarity between Signature 1, Signature 2 and Signature 3 with all 30 COSMIC human signatures was assessed. Signature 1 (the most common contributor to pol2 mutated samples) was found to be most similar to COSMIC Signature 10 (cosine similarity = 0.63) and COSMIC Signature 8 (cosine similarity = 0.62). COSMIC Signature 10 was found most commonly in colorectal and uterine cancer and is statistically associated with the presence of POLE mutations, notably Pro286Arg and Val411Leu. Both signatures feature C>A mutations at TpCpT, but discrepancies in C>T and T>A mutations. COSMIC Signature 8 shows similarities to the yeast Signature 1 for C>A mutations at TpCpT and T>A mutations, however many peaks seen in COSMIC Signature 8 are absent in the yeast signature. Signature 3 (the signature observed in the wild-type yeast strains) is most similar to COSMIC Signature 5 (cosine similarity = 0.82), which is of unknown aetiology. And Signature 2 (the signature observed in the pol3-S483N strain) is most similar to COSMIC Signature 12 (cosine similarity = 0.78), Signature 5 (cosine similarity = 0.77) and Signature 16 (cosine similarity = 0.75). This signature and Signature 12 both show increased amounts of T>C mutations compared to all other mutation types.

3.4 Geographical mutation patterns

Apart from mutation numbers or mutational signatures, where in the genome mutations are located can provide more information on the mutagenic process at work and its possible effects. Do mutations cluster? Are there regions of the genome particularly prone to mutation? How do mutations occur with respect to features of the genome such as genes or origins of replication? In this next section, I have attempted to identify striking differences between wild type strains and polymerase mutants when it comes to the locations of the mutations within the genome.

**Kataegis** Kataegis describes localised hypermutation, sometimes observed in cancer samples [738]. While kataegis is linked to the APOBEC deaminases[880], we can nonetheless
Figure 3.11: EMu: Validating Signature Analysis

Signature analysis was also performed using EMu. Using model selection criteria, such as the Bayesian information criterion (BIC), EMu determined that a model with three signatures has the strongest statistical support. 

A | Signatures displayed in their trinucleotide mutation pattern.

B | Contribution of signatures displayed in A to the total mutations observed in mutator strains.
Figure 3.12: No observed clustering of mutations acquired by *pol2-P301R* strains
Mutations acquired across 15 parallel lines of propagated *pol2-P301R* strains were pooled, sorted and inter-mutation distances were determined and plotted.
look for regions of the genome where mutations are clustered or common. To that end, mutations across all parallel lines for each strain were pooled and sorted. The distance between consecutive mutations was calculated and plotted. This results in a “rainfall plot”, where most mutations will appear as a “cloud” at the value of the mean inter-mutation distance. Where mutations cluster the inter-mutation distance will be significantly smaller and data points will appear as “raindrops” making them easy to spot. Across all heterozygous diploid polymerase mutant strains no striking examples of mutation clustering was observed (see Fig. 3.12 for an example, Appendix B for all plots).
Genic versus intergenic locations of mutations  If mutations were acquired randomly across the genome, regardless of the effect on genomic information, we would expect the fraction of mutations within gene sequences to be concordant with the fraction of the genome covered by genes. The *S. cerevisiae* reference genome contains 12071326 nucleotides. The 5133 verified open reading frames (ORFs) listed in the Saccharomyces Genome Database cover 7905244 nucleotides or 65.49% of the genome. Each of the mutations acquired in the propagation experiment of heterozygous polymerase mutant strains was checked against all verified ORFs to determine whether it falls within a gene or outside them. If mutations were acquired in a truly random fashion, then one would expect roughly 65% of acquired mutations to fall into a gene. Interestingly, for wild-type propagated strains this percentage is quite low with 53.7%, while for strains with a significantly raised mutation number the percentage approaches or surpasses 65.5% (Fig. 3.13). The percentage of mutations that fall within genes is expected to be lower in haploid samples, due to the absence of a second copy for genes. Indeed, the fraction of mutations observed within genes in wild-type samples does go down to approximately 44.8% (Fig. 3.14). However, for strains that show an increased mutation accrual, the difference between haploid and heterozygous diploid strains in this respect is almost negligible.

Mutations around origins of replication  Because leading and lagging strand switch at origins of replication and because polymerases ε and δ are thought to replicate each, respectively, the mutational patterns in polymerase mutant strains could show interesting behaviours around origins of replication (ARS elements in *S. cerevisiae*).

ARS sequences and their locations were obtained from the Saccharomyces Genome Database. To include only high confidence DNA replication origin sites this list was compared to the OriDB database and only confirmed ARS elements were retained. The location data between the two databases varies as origin location was determined differently. Here, the SGD origin locations were used. For each origin, the center was determined as well as the coordinates 500bp upstream and downstream. For each mutation acquired by polymerase mutants, that falls within that window, the nucleotide change and distance from origin center was determined. Using the *pol2-P301R* samples, 2782 different SNVs were tested for their proximity to 352 different ARS elements. Within 500bp of the center of the origin only 4 mutations were identified, the same number for the 4015 mutations acquired by the *pol3-S483N* strains is 3 mutations, only. If the window is extended to 2500bp either side, 8 mutations are detected in the case of both strains.

Thus, while potentially, an interesting feature of the acquired mutations, currently, the number of mutations is not sufficient to determine patterns of mutations around origins of
Figure 3.14: Percentage of mutations within genes in haploid strains
The percentage of mutations that fall within genes is shown for both haploid and heterozygous diploid strains for a selection of *pol2* mutants.
Figure 3.15: Number of total aneuploidy and segmental insertions/duplications identified. The number of events of aneuploidy and large deletions/amplification was determined from the coverage data of all strains and counted. The number of samples available for each strain is listed below the figure legend.

Aneuploidy and large insertions and deletions

All samples pre- and post-propagation were checked for variations in chromosome number and large segmental deletions as well as amplifications. Aneuploidies and segmental deletions are relatively rare. In 65 wild-type samples, one variation in chromosome number (1.5%) and 5 instances of segmental deletions/amplifications (7.7%) were detected. Small increases in number of aneuploidies can be seen for the strains pol2-04, pol2-P301R and pol2-V426L (Fig. 3.15). However, with
mutations that occur this rarely, the sample size would need to be bigger to make conclusive statements about significant increases. Examples of aneuploidy and segmental deletions/amplification are shown in Fig. 3.16 and Fig. 3.17, respectively, and the full set of figures can be found in Appendix B.

**rDNA copy number** To screen for changes in rDNA copy number, rDNA copy number estimate analysis was performed for all pol2 and pol3 heterozygous mutants after their three month propagation. All strains were derived from the same starting cells and any increase or decrease in rDNA copy number could likely be due to the polymerase mutation, though that would have to be reconfirmed by an independent introduction of the mutation into a new wild-type strain. Because data from the start of the experiment are not available due to data corruption, these numbers are only indications, but increases in mean rDNA copy number for strains like pol3-S483N will be followed up with later experiments.

### 3.6 No increase in INDELs (compared to MMR mutants)

While there are clear differences in mutation accrual with respect to SNVs, no striking increases in INDELs were detected (Fig. 3.19). After three months most strains (including the wild-type) will have accumulated a mean $1.18 \pm 0.4$ INDELs. The only increase can be seen for pol3-S483N with a mean of 4.45 INDELs for each strain. However, compared to the average of 200 SNVs this strain accumulates the increase is minor.

This is in stark contrast to other mutations that affect DNA replication fidelity found in cancer. As discussed in Chapter 1.4.2, loss-of-function mutations in mismatch repair proteins are known to predispose to colorectal cancer[803, 804] and as a pilot experiment using the automated robot propagation set-up (see 6.7), 150 strains carrying mutations in known DNA repair genes were propagated for 3 months by our group and the Warringer group in Sweden as described. These included strains deleted for mismatch repair genes: MSH2, MSH6, MLH1 and PMS1. Ten colonies each had their DNA extracted, were sequenced and the data was aligned as described in Chapter 2 and 6. One Mlh1 sample did not pass sequence quality control.

Figure 3.20 shows that in the case of most mismatch repair mutants, the mutation increase is fairly evenly split between SNVs and INDELs with the exception of of Msh6 (part of MutSα), whose absence, as expected, results mainly in single-nucleotide mismatches. Loss of Msh2 (part of both MutSαand MutSβ), Mlh1 (MutL homolog) and Pms1 (which forms a heterodimer with Mlh1) leads to high numbers of SNVs and small INDELs.
3.6 No increase in INDELs (compared to MMR mutants)

Figure 3.16: Example of aneuploidy in *S. cerevisiae*

This example shows the coverage profile of a *pol2-04* propagated heterozygous diploid strain with an aneuploidy for chromosome XI (3n). At repetitive regions the coverage drops to 0 (for instance see chromosome XII) due to mapping quality thresholds in the program.
Figure 3.17: Example of segmental deletions and amplifications in *S. cerevisiae*

This example shows the coverage profile of a *pol3-01* propagated heterozygous diploid strain with a segmental deletion on chromosome V and an amplification of a region of chromosome III. At repetitive regions the coverage drops to 0 (for instance see chromosome XII) due to mapping quality thresholds in the program.


Figure 3.18: rDNA copy number changes in polymerase mutants

rDNA copy number estimates of all post-propagation samples is shown. The median is denoted by a black line. Each dot represents a post-propagation sample. A) rDNA repeat number for POL2 wild type samples and all heterozygous diploid pol2 mutant strains B) rDNA repeat number for POL3 wild type samples and all heterozygous diploid pol3 mutant strains.
Figure 3.19: No increase in the number of INDELs detected per sample across strains
The number of INDELs per sample was determined from the heterozygous diploid polymerase mutant strains that were propagated for three months using single colony bottlenecks. The back bars indicate mean and standard deviation.
3.7 Summary

In this part of this work, I have assessed the mutagenic potential of all candidate polymerase mutations in heterozygosity \textit{in vivo} using the budding yeast system. Strains carrying the mutations \textit{pol2-P301R}, \textit{pol2-S312F}, \textit{pol2-L439V}, \textit{pol2-M459K} or \textit{pol3-S483N} show significant

Figure 3.20: Mutation accrual in strains with mismatch repair deficiencies

Strains with deletions in mismatch repair genes were propagated for three months using population bottlenecks. Ten colonies each were sequenced and single-nucleotide variants as well as INDELs were identified. Results for strains carrying deletions in \textit{MLH1}, \textit{MSH2}, \textit{MSH6} and \textit{PMS1} are shown. Unique mutations across all colonies were counted and divided by number of sequenced colonies to obtain mean mutation numbers per sample.

These differences in mutational patterns between polymerase mutants and MMR deficient cells, is also expected to be reflected in the respective cancer genomes. Indeed, MMR deficient tumours (most commonly deficient for Mlh1 due to hypermethylation of the MLH1 promoter) commonly show high frequency of mutations, either mismatches in single bases or in regions of short tandem DNA repeats (microsatellites), the former leading to SNVs and the latter to INDELs[881]. Microsatellite instable (MSI) tumours show mutation loads ranging from 10 to 100 mutations per Mb. Polymerase epsilon mutated tumors, on the other hand, often show a mutation incidence exceeding 100 mutations/Mb and are mostly microsatellite stable (MSS)[882], characterised by mostly point mutations. The data we have collected in \textit{S. cerevisiae}, has thus been confirmed by the data collected in human tumours.

3.7 Summary

In this part of this work, I have assessed the mutagenic potential of all candidate polymerase mutations in heterozygosity \textit{in vivo} using the budding yeast system. Strains carrying the mutations \textit{pol2-P301R}, \textit{pol2-S312F}, \textit{pol2-LA39V}, \textit{pol2-M459K} or \textit{pol3-S483N} show significant
increases in single-nucleotide variants and, intriguingly, these increases are more pronounced than those resulting from mutating catalytic residues of the polymerase exonuclease domains. We further show that the pattern of SNVs is distinct from the wild-type and differs between pol2 and pol3 mutant strains. Striking geographical patterns or increases in large-scale mutations such as aneuploidy were not detected. Furthermore, in contrast to most mismatch-repair deficient cells, polymerase mutated strains show no increase in INDEL incidence. This disparity is also reflected in comparisons between MMR deficient tumours and those carrying polymerase epsilon mutations.

Evaluating hypotheses

Aims:

• To assess all candidate DNA polymerase mutations for the number of mutations acquired in the same amount of time

*Candidate DNA polymerase mutations lead to varying degree of mutation rate increases in budding yeast. Significant increases in mutation rates were identified for strains carrying: pol2-P301R, pol2-S312F, pol2-L439V, pol2-M459K and pol3-S483N. The increase in mutation numbers in these cases exceeds that observed in exonuclease deficient control strains.*

• To identify the type of mutations caused by mutated DNA polymerases

*Across all mutation types examined, striking increases in the number of mutations were shown for single-nucleotide variants. Comparatively, other types of mutations were acquired rarely by polymerase mutant strains tested here.*

• To determine whether candidate mutations leave a distinct mutation pattern on the genome

*Where the numbers of mutations allow, trinucleotide mutation patterns were plotted, adjusted to genome wide trinucleotide frequencies and mutation signature extraction was attempted. While the pattern of mutations in pol3-S483N strains looks similar to that observed in wild-type samples, it is subtly distinct with an increase in T>C mutations. Additionally, the pol2-P301R mutation results in a distinctive mutation pattern with key peaks in C>A and T>A mutations.*

• To compare mutations acquired in mutated polymerase strains to those resulting from mismatch repair deficiency
It is well-documented that mismatch repair deficiency predisposes to colorectal cancer and polymerase mutations are implicated in predisposition to colorectal cancer. Here, mutations in yeast accumulated to both are explored. While a near-complete mismatch repair deficiency results in roughly equal amounts of single-nucleotide variants and insertions/deletions, insertions/deletions make no or negligible contributions to any increases in mutation accrual due to a mutated DNA polymerase.
Chapter 4

Polymerase mutations in mammalian systems and in combination with other mutations

Overarching hypothesis

DNA polymerase mutations exert their mutagenic function in a manner separate from exonuclease deficiency and mutation accumulation increases identified in *Saccharomyces cerevisiae* can be confirmed in mammalian systems.

Aims:

- To determine whether polymerase mutations and mismatch repair deficiency act synergistically when present in the same cell.

- To examine whether the mutagenesis observed in mutant strains is independent of Polɛ exonuclease activity.

- To determine the mutagenesis observed in mutant strains may be due to increased involvement of a translesion polymerase in DNA replication.

- To investigate candidate polymerase mutations identified as mutators in *Saccharomyces cerevisiae* in mammalian systems.
4.1 Introduction

In the previous chapter, candidate polymerase mutations were assessed for their mutagenic potential. The exonuclease domain mutations \( \text{pol2-P301R}, \text{pol2-S312F}, \text{pol2-L439V}, \text{pol2-M459K} \) or \( \text{pol3-S483N} \) were all shown to significantly increase the number of acquired mutations when in heterozygosis. Strikingly, the \( \text{pol2-4} \) exonuclease deficient allele does not confer a comparably strong increase in mutation numbers. To get a better understanding of how these candidate polymerase mutations lead to more mutagenesis, in this chapter a series of double mutant experiments is described.

Additionally, while the budding yeast \( \text{Saccharomyces cerevisiae} \) was chosen to investigate the genome-wide effects of these mutations in a cost and time effective manner in an organism that has been shown to yield insights into DNA replication and repair that shows remarkable conservation, it remains to validate these in mammalian systems. Thus, this chapter also describes the introduction of the \( \text{pol2-L439V} \) and \( \text{pol3-S483N} \) mutations into a mouse model. These two mutations were identified in the germline of families with a predisposition to colorectal cancer and the whole organism aspect of the mouse model may give insights into the tumorigenesis, which a single-celled organism cannot provide. The \( \text{pol2-P301R} \) mutation, resulting in similarly striking mutation increases as \( \text{pol3-S483N} \) and one of the most common DNA polymerase alterations identified in human cancers, is introduced into a human cell line to show that the insights gained in yeast can be translated into a human system, while acquired at a fraction of the cost.

4.2 Synthetic lethality with mismatch repair deficiency

Given the differences between the mutational patterns observed in polymerase mutant cells and mismatch repair deficient cells and the fact that both are promoting tumor progression and predispose to colorectal cancer, the question is how the genome is affected when both fidelity systems are impaired simultaneously.

To obtain double mutants, a \( \text{MAT} \alpha \) strain deleted for \( \text{MSH2 (msh2Δ)} \) was created (see Chapter 6.6) and mated to haploid \( \text{MATa cells: pol2-P301R, pol2-S312F, pol2-A480V and pol2-M459K} \) (see Chapter 6.6). The strains now heterozygous diploid for \( \text{msh2Δ} \) and the polymerase mutation were kept on sporulation medium to undergo meiosis. Tetrads were isolated and dissected. After two days of growth tetrads were replicated onto selection plates selecting for the polymerase mutation (ura- plates), the \( \text{msh2Δ} \) mutation (G418 plates) and plates selecting for either mating type. This allowed confirmation that tetrads were haploid.
Figure 4.1: Tetrad dissection to generate double mutants and detect synthetic lethality

A) For tetrad dissection, two haploid single mutant strains of opposite mating types are mated and the resulting diploid forced to undergo meiosis resulting in a tetrad. Alternatively, an already existent diploid can be used. Replication followed by segregation results in each allele (here depicted as a “blue” wild-type and a “red” mutant allele) being present in two out of the four tetrads. This is mutually exclusive (for instance a tetrad spore will under regular circumstances contain either the “red” or the “blue” allele, not both or neither as that would result in aneuploidy). Using a micromanipulator tetrads are dissected on rich medium plates (YPAD) and allowed to germinate and expand to colony size. Testing growth in selective conditions, can reveal mating type or, in cases where a mutant allele is marked by for instance an antibiotic resistance, which spores contain the mutant allele. Double mutants can be generate analogously, by mating two single mutant haploid strains. If both mutant alleles are unlinked (on different chromosomes), they should be randomly assorted during meiosis, resulting in wild-type, single mutant and double mutant cells, which can be identified by tetrad analysis. B) An example, of a tetrad dissection of a mating of msh2Δ and pol2-P301R. When not all four spores result in a colony, microscopy can be used to confirm the spores germinated, but ceased to divide after a few division cycles. Should all missing colonies have been double mutants, synthetic lethality is a likely conclusion. C) If not all spores reached full colony size (denoted by “?”), then their genotype might be inferred by using selection to determine the genotype of the remaining colonies. Since every allele (wild-type and mutant) should occur twice, the genotype of the failed colonies can be determined. In the case of synthetic lethality, double mutants will not be among the full-sized colonies and most if not all failed colonies have an inferred double mutant genotype.
(either MATa or MATα) and whether they carried one or both mutations. Among all strains I sporulated, I observed spores that germinated, but didn’t grow to form a colony. Across a tetrad, if the two mutations are not linked, each mutation should be observed twice, randomly assorted (Fig. 4.1). As Fig. 4.1 illustrates that means if three meiotic products grew to a colony and one germinated and underwent extinction, the genotype of the extinct cells can be inferred. Should these cells be double mutants and should no living double mutants be recovered, a case for synthetic lethality can be made. Thus, the number of double mutants we should recover was determined from genotype information and the number of double mutants actually recovered was obtained (Table 4.1). As can be seen, pol2-P301R, pol2-S312F and pol2-M459K are likely synthetic lethal with msh2Δ in a haploid background. pol2-A480V msh2Δ double mutants can be viable, but grow visibly slower on rich medium plates.

Since of all four polymerase mutants, the one causing the lowest increase in mutation number (Fig. 3.1) is the only one to allow for a double mutant with msh2Δ in a haploid context, the data is in line with the notion of a threshold for lethal mutagenesis. This postulates that a certain elevation of mutation rate will overwhelm the population and cause extinction. If this is the explanation, then predictions about the viability of other polymerase mutants in combination with mismatch repair deficiency can be made. For instance pol2-Q468R msh2Δ double mutant should be viable, a pol3-S483N msh2Δ double mutant should not be and the viability of pol2-04 msh2Δ and pol2-L439V msh2Δ double mutants is uncertain. This will be the next experiments carried out to further underpin the relationship between polymerase mutants and mismatch repair deficiency.

4.3 Epistatic relationship of mutations with exonuclease deficiency

To gain further understanding of why some candidate pol2 and pol3 mutant strains acquire more mutations than the pol2-4 and pol3-01 exonuclease deficient strains, the relationship between pol2 candidate mutations and the exonuclease function was further explored. Pol2 candidate mutations (pol2-P301R, pol2-S312F, pol2-A480V, pol2-M459K) were combined with the mutations in the strain (pol2-D290A,E292A) in the same gene, to construct MATa haploid yeasts trains each carrying three point mutations in the POL2 gene. Wild-type, single mutant and combined strains were propagated using single colony bottlenecks for 13 passages (1.5 months) in 18 parallel lines each. Strains were sequenced and analysed for acquired SNVs and INDELs as in Chapter 2.4.2 and 3.2. While there is no apparent difference be-
### 4.4 Observed mutagenesis in *pol2-P301R* strains is not due to increased participation of Polζ in DNA replication

<table>
<thead>
<tr>
<th>polymerase mutant</th>
<th>tetrads dissected</th>
<th>double mutants expected</th>
<th>double mutants recovered</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>pol2-A480V</em></td>
<td>4</td>
<td>4</td>
<td>3*</td>
</tr>
<tr>
<td><em>pol2-M459K</em></td>
<td>8</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td><em>pol2-P301R</em></td>
<td>10</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td><em>pol2-S312F</em></td>
<td>15</td>
<td>14</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.1: Synthetic lethality of polymerase mutants and mismatch repair deficiency

Diploid strains heterozygous for both *msh2Δ* and a polymerase mutant (see first column) were sporulated and dissected after meiosis allowing the recovery of all four meiotic products. The number of tetrads dissected is the number of tetrads where all four meiotic products germinated. Replica-plating on selective medium plates allowed the identification of double mutants. The number of “double mutants expected” is the number of double mutants we expect to obtain from all tetrads considering the genotypes in the tetrad, the number of “double mutants recovered” is the number of double mutants that actually made it from germination to colony. A left-tailed 2X2 Fisher’s Exact Test was performed to determine the significance of the negative association between the polymerase mutations and mismatch repair deficiency.

* Three double mutants were recovered. However, they were smaller than all single mutant and wild-type strains.

between the number of SNVs acquired by *pol2-A480V* samples and the number acquired by *pol2-D290A,E292A,A480V* samples, a significant difference between the number of SNVs acquired by *pol2-P301R, pol2-S312F, pol2-M459K* samples and their *pol2-D290A,E292A* combined counterparts is observed (Fig. 4.2). In each of those cases the number of acquired mutations is much less when the *pol2* candidate point mutation is combined with the other two point mutations, however, the numbers still exceed those acquired by a *pol2-4* strain, suggesting that the observed mutagenesis is at least in part dependent on the exonuclease function of the protein.

### 4.4 Observed mutagenesis in *pol2-P301R* strains is not due to increased participation of Polζ in DNA replication

To investigate the high number of mutations in polymerase mutant strains (*pol2-P301R, pol2-S312F, pol2-L439V, pol2-M459K* and *pol3-S483N*) compared to the *pol2-4* and *pol3-01* exonuclease deficient mutant strains, the possibility of the involvement of another polymerase was explored. Recent reports indicate that in cases of replisome instability polymerase ζ, unique in its ability to extend primers with a terminal mismatch, can take over more of the replication burden and account for some of the mutagenesis[883, 884]. It was described that cells carrying the *dpb2-100* mutant allele have decreased interaction with the CMG (Cdc45-
Combinations of candidate pol2 mutations and the pol2-ΔD290A,E292A mutations in the same gene were achieved by site directed mutagenesis of the plasmids used to construct the strains. When haploid strains expressing Polε with all three missense mutations were obtained, they were propagated alongside haploid single mutants for 13 passages using single-colony bottlenecks. Samples were sequenced before and after propagation and acquired single nucleotide variants (SNVs) and INDELs were determined as detailed in Chapter 2. The number of parallel lines are as follows: n=18 (POL2, pol2-S312F, pol2-P301R, pol2-4 A480V, pol2-4 M459K, pol2-4 P301R), n=17 (pol2-4 A480V, pol2-4) and n=16 (pol2-4 S312F). Significance of the difference in number of SNVs was determined by two-tailed, unpaired t-Tests. n.s. P > 0.05; *** P ≤ 0.001
4.4 Observed mutagenesis in \textit{pol2-P301R} strains is not due to increased participation of Pol\(\zeta\) in DNA replication

MCM-GINS complex, which is thought to bring the polymerase \(\varepsilon\) into the replisome. Polymerase \(\zeta\), recently found to also interact with polymerase \(\delta\) subunits Pol31 and Pol32\cite{885, 886}, has been proposed to be the fourth polymerase in the replisome and is expected to take over replication when the main replicase has a problem with primer extension. Deletion of the catalytic subunit of Pol\(\zeta\), \textit{REV3}, substantially decreases the mutator phenotype found in \textit{dpb2-100} cells, but has little effect on the mutator phenotype observed in \textit{pol2-4} cells\cite{883}.

To investigate whether this could partly explain the apparent divergence between \textit{pol2-4} mutant strains and the other polymerase mutations, whose presence results in higher mutation numbers than exonuclease deficiency, a \textit{rev3}Δ strain was constructed and double mutants of polymerase mutations with \textit{rev3}Δ were generated by mating and tetrade dissection (see Chapter 4.2). The resulting haploid double mutant strains, the single mutant precursor strains and a wild-type control were propagated in 18 parallel lines using single-colony bottlenecks for 13 passages (1.5 months) on non-selective rich medium. Starting and final colonies were sent for whole-genome sequencing and analysed for single nucleotide variants (SNVs) and small insertions/deletions (INDELs) as before (see Chapter 3.2).

As expected, the mutation numbers observed in the \textit{rev3}Δ strains is lower than those in wild-type strains with a mean number of mutations of 1.89 compared to 3.55 in the wild-type (Fig. 4.3). However, unlike the work with \textit{dpb2-100} cells\cite{883}, where a double mutant with \textit{rev3}Δ decreased mutation numbers, leading them to conclude that Pol\(\zeta\) is responsible for a substantial part of the mutagenesis in \textit{dpb2-100} cells, here no such drop in mutation numbers can be observed. While there is only small increases in mutation number when one compares \textit{pol2-A480V} strains with the corresponding \textit{pol2-A480V rev3}Δ double mutant (a mean of 38.6 versus a mean of 49.1 mutations) and the \textit{pol2-S312F} with the corresponding \textit{pol2-S312F rev3}Δ cells (a mean of 134.6 versus a mean of 161.6 mutations), the mutation increase observed in \textit{pol2-P301R} strains upon combination with a deletion of \textit{REV3} is striking: \textit{pol2-P301R} cells acquire a mean of 116.8 mutations, while \textit{pol2-P301R rev3}Δ cells acquire a mean of 347.1 mutations in the same time frame.

To get an understanding of the process at work here, the mutation patterns of the SNVs acquired in \textit{pol2-P301R} cells and \textit{pol2-P301R rev3}Δ cells were compared (Fig. 4.4). Even though the number of mutations acquired by the \textit{pol2-P301R rev3}Δ cells is more than double that of the \textit{pol2-P301R} cells, there are no striking differences between the two patterns.
Polymerase mutations in mammalian systems and in combination with other mutations

Figure 4.3: Synergistic effects on mutation number between \textit{rev3}\textDelta and \textit{pol2-P301R}

Polymerase mutation and \textit{rev3}\textDelta haploid double mutants were obtained by crossing and propagated alongside haploid single mutants for 13 passages using single colony bottlenecks. Samples were sequenced before and after propagation and acquired single nucleotide variants (SNVs) and INDELs were determined as detailed in Chapter 2. The number of parallel lines are as follows: n=18 (POL2, \textit{rev3}\textDelta, \textit{pol2-S312F}, \textit{pol2-P301R}), n=17 (\textit{pol2-A480V}, \textit{pol2-S312F rev3}\textDelta), n=16 (\textit{pol2-P301R rev3}\textDelta) and n=15 (\textit{pol2-A480V rev3}\textDelta). Significance of the difference in number of SNVs was determined by two-tailed, unpaired t-Tests. ** P \leq 0.01; *** P \leq 0.001
4.4 Observed mutagenesis in *pol2-P301R* strains is not due to increased participation of Polζ in DNA replication

Figure 4.4: Mutational patterns observed in *pol2-P301R* cells and *pol2-P301R rev3Δ* cells are highly similar. All single nucleotide variants identified in the *pol2-P301R* cells and *pol2-P301R rev3Δ* cells displayed in Fig. 4.3 are displayed in their trinucleotide context. Relative contribution to the total of single nucleotide variants is given. No adjustment for *S. cerevisiae* genome-wide trinucleotide frequencies is made.
4.5 Examining polymerase mutations in other organisms

The work in budding yeast *S. cerevisiae* has identified *pol2-P301R, pol2-S312F, pol2-L439V, pol2-M459K* and *pol3-S483N* as cancer mutations that likely increase the mutation rate, thus promoting tumour progression by increasing the probability of acquiring further cancer promoting mutations. To bring this evaluation of the candidate polymerase mutation full circle, mutations identified as likely tumor promoting in yeast cells will be introduced into mammalian systems, to show whether the assertions made in budding yeast hold and observing effects in multicellular organisms.

4.5.1 The Pole and Pold1 mutations in mouse models

In a collaboration with Ian Tomlinson’s group in Oxford, the Adams group started to construct germline Pole and Pold1 mutations in mice[768]. The Tomlinson lab provided me with two constructs: Pold1-S476N and Pole-L424V. The former is a conditional knock-in of the S476N mutation into the endogenous Pold1 gene by a loxP mediated introduction of a mutated exon 12 (Fig. 4.5-B). The latter introduces the L424V mutation similarly by a conditional knock-in of a mutated exon 13 of Pole (Fig. 4.5-A). The constructs were designed by Ian Tomlinson and use inverted loxP sites. Inversion of these sites results in an expression of a fluorescent marker and a switch from the unmodified to the mutated exon. Where regions of homology exist, the sequence of the exons were optimised by using synonymous codons (for instance exon 12-14), to decrease the likelihood of secondary DNA structures.

The constructs were introduced into JM8.F6 mouse ES cells (C57BL/6N strain) by Graham Duddy from the Sanger Institute ES Cell Mutagenesis Team. I checked 48 Neo-resistant clones each for proper integration of the constructs by using PCR across the homology arms. Five clones for the Pold1 construct and three clones for the Pole construct were identified. One of the Pold1-targeted clones was excluded for trisomy of chromosome 8. The clones for each construct were micro-injected by the Sanger mouse facility and 40-50 mouse embryos each were transferred for gestation. For both constructs chimeras were obtained and mated to generate non-chimeric progeny. The F1 progeny was genotyped by James Hewinson (Adams group). The mice carrying the conditional Pole-L424V mutation have been further crossed to Flp-deleter mice to make them conditional. These mice have been sent to Oxford for phenotyping and further experiments and their sperm cryopreserved should the line need reviving. The mice carrying the conditional Pold1-S476N strain were successfully generated after a second round of microinjection and are currently being crossed to Flp-deleter mice. These mice should give further indications about the nature of tumors that arise due to these
Constructs used to generate knock-in conditional mutants for A) *Pold1* S476N (mouse equivalent of human *POLE* p.L424V) and B) *Pole* L424V (mouse equivalent of *POLD1* S478N) in mouse embryonic stem cells. Inversion of loxP sites results in the expression of a fluorescent marker and the switch of the unmodified to a mutated exon. In the case of *Pold1* exon 12-14 the exons were optimised by synonymous mutations to decrease homology mediated secondary structures. FRT sites are also included allowing the excision of the PGK Neo cassette used for clone selection. The construct was designed by Ian Tomlinson (Oxford).
Polymerase mutations in mammalian systems and in combination with other mutations

two mutations: their prognosis, their organ tropism and any possible drug sensitivities.

4.5.2 Human POLE P286R mutant cell lines

Additional to the mice engineered to carry the two identified germline mutations from [768], I decided to also generate the POLE Pro286Arg mutant in a human cell line to confirm that mutation rate increases detected in S. cerevisiae can also be detected in human cell lines and that other characteristics like mutational patterns are conserved.

To that end, recent advances in gene editing techniques were chosen. To make the point mutation in human cells a CRISPR-Cas9D10A nickase-based system developed in the Jackson group was used to construct the plasmid for transfection[887]. This involves designing guide RNAs (gRNAs) that will target the mutated CAS9 enzyme to the POLE gene and introduce single-stranded breaks either side of the residue to be mutated. Furthermore, a 200bp long oligonucleotide (ssODN) is supplied that carries the genomic sequence around that locus with the designed mutation as well as additional mutations to prevent re-nicking from the Cas9D10A after recombination has taken place. gRNAs and the ssODN have been designed and cloned and are awaiting transfection into human cell lines (see Chapter 6.4 for sequences). The human cell line to use is currently being chosen. After genotyping a POLE P286R mutated human cell line will be used to assess the effects of this mutation on genomic integrity.

4.6 Summary

In this chapter, a selection of candidate pol2 mutations, shown in Chapter 3 to lead to varying increases in mutation number, were assessed for their behaviour in combination with other mutations. Inactivating mismatch repair in these haploid strains lead to either minature colonies, in case of the weakest mutator pol2-A480V, or to synthetic lethality in case of stronger mutators. This is concordant with an hypothesis of a lethal mutation rate leading to extinction. Experiments combining intermediate mutators with mismatch repair deficiency or repeating these experiments in diploid backgrounds are logical next steps.

Since, intriguingly, most mutator pol2 mutations described in this work lead to mutation accumulation in excess of what results due to the mutations in the pol2-4 background (mutations of two critical residues in the ExoI motif to alanine), which are reported to abrogate the catalytic activity of the exonuclease domain. Combination of these mutations to alanine with strong mutator candidate pol2 mutations results in noticeable decreases of mutation accumulation, but mutation numbers are still larger than those accumulated by pol2-4 strains.
To investigate whether a bulk of the mutagenesis could be due to the increased involvement of a translesion polymerase like Polζ, double mutants were generated and increases in mutation accumulation observed for all double mutants, though deletion of the catalytic subunit of Polζ, itself, leads to a reduction in accumulated mutations. To determine a possible source of the additional mutagenesis mutational patterns were plotted and no difference between the pattern of mutations in the polymerase ε mutants versus the mutations in the Polε/Polζ double mutant was observed. This suggests that, if anything, Polζ limits the mutagenesis due to the pol2 mutations.

Having identified a subset of DNA polymerase mutations that lead to an increased accumulation of mutations in budding yeast, it is crucial to demonstrate that these findings are relevant in a mammalian system. To that end, in collaboration with the Sanger Mouse Facility and Ian Tomlinson’s group in Oxford, mice with conditional knock-in mutations for Pole-L424V and Pold1-S476N were created. To take this work back into a human system, we are also in the process of creating a POLE P286R mutated human cell line.

**Evaluating hypotheses**

**Aims:**

- Polymerase mutations and mismatch repair deficiency act synergistically when present in the same cell.

  *When combining those mutations in haploid cells, most lead to synthetic lethality. The only surviving colonies we obtained - pol2-A480V msh2Δ - will need to be assessed for how both mutator phenotypes interact.*

- The mutagenesis observed in mutant strains is independent of Polε-exonuclease activity.

  *Combining polymerase mutations with the exonuclease-deactivating missense mutations in the same gene, leads to a reduction in mutagenesis for strong mutator polymerases suggesting that they are not independent of the exonuclease catalytic site.*

- The mutagenesis observed in mutant strains may be due to increased involvement of a translesion polymerase in DNA replication.

  *Contrary to what one would expect if this hypothesis were true, cells deficient for the translesion polymerase Polζ display more mutagenesis, not less. Thus, Polζ is not the source of the mutagenesis observed, but other polymerases have not been ruled out.*
• Candidate polymerase mutations identified as mutators in *Saccharomyces cerevisiae* will be mutators in mammalian systems.

_Mice carrying conditional knock-in mutations for the reported human germline varaints, POLE L424V and POLD1 S478N, were generated. A POLE P286R mutated human cell line is also currently being made._
Chapter 5

Discussion and future directions

5.1 Whole-genome sequencing as a flexible tool to address problems in cell biology

In this thesis, I have used whole-genome sequencing of model organisms to address questions in cell biology in DNA repair and replication. In the first part of this work, we have successfully used whole-genome sequencing to identify suppressor mutations in synthetic viability screens. In this type of experiment a selectable phenotype, usually due to a mutation, is alleviated by a second mutation. This allows inferences about a relationship between the two mutations and between the second mutation and the phenotype. These types of genetic interactions can be more informative than synthetic lethality, which sometimes arises due to the inactivation of two important, but unrelated pathways. While this type of screen has been utilised to uncover genetic interactions for decades, the identification of the secondary suppressor mutation is often labrious and time-consuming. The work in this thesis has shown that whole-genome sequencing can be utilised to correctly identify a suppressor mutation and that follow-up of these suppressors can yield relevant biological insight[801]. Currently, Dr. Fabio Puddu and I are validating suppressor mutations identified in a third *Saccharomyces cerevisiae* suppressor screen looking at proteins involved in replication stress response. Exploiting recent advances in culturing haploid mouse cells has also allowed us to extend this work to mammalian systems and demonstrate that the technique works to identify known suppressors to 6-thioguanine sensitivity as a proof-of-principle[1124]. Currently, we are extending this work to identify mutations alleviating the sensitivity to other chemicals of interest.

Suppressors identified in yeast usually arose without the need for mutagenesis, which com-
plicates the identification of suppressor mutations. However, it is known that the genetic background influences the pattern of spontaneously arising mutations, which may influence and limit the kind of suppressor that can be identified. Such differences in mutation patterns are the key interest in a nation-wide multi-institute project that the Jackson and Adams group are involved in (COMSIG). In an attempt to understand mutational processes in budding yeast, our group is identifying mutational patterns caused by deletion of any yeast gene.

As such we have propagated the *S. cerevisiae* gene deletion collection for a defined period of time and are sequencing strains to identify mutations acquired in that time frame. By adapting the analysis protocol I have developed for budding yeast genetic screens, I was able to identify acquired mutations in such mutation accumulation experiments. The acquired mutations will uncover mutational patterns and will generate a dataset from which mutational signatures can be extracted as it has been successfully demonstrated for human cancers [761–764]. It is expected that a catalogue correlating genetic defects and mutation patterns will in the future assist with elucidating the history and aetiology of cancer samples.

### 5.2 Polymerase mutations as drivers of mutagenesis

As part of the overarching effort to identify patterns of mutations associated with the loss of particular genes, I focused my attention on DNA polymerases. Tasked with duplicating the entire genome, they are prime candidates for sources of mutagenesis and recent work has identified mutations in DNA polymerase δ and ε as factors predisposing to familial colorectal cancer[768]. Further work identified more mutations in these DNA polymerases[768, 809, 810], but failed to identify which mutations had an impact on mutagenesis and which did not.

To examine the global effects of mutated DNA polymerases on genome stability, I used mutation accumulation experiments, propagating yeast strains carrying DNA polymerase mutations for a fixed amount of time. Whole-genome sequencing every sample at the beginning and end of the experiment allowed us to obtain lists of mutations accumulated by each sample. The design of the experiment was aided by similar experiments carried out in the group of Alain Nicolas, whose work with different budding yeast mutant provided information on the numbers of mutations expected in a wild-type strain[835]. Having used both mutation accumulation experiments and classic genetic reporter assays, I find that there is some agreement between these two methods, similar to what has been observed by Alain Nicolas[835]. However, whole-genome sequencing provides less variable data and more information, making it the better choice for this work.
5.2 Polymerase mutations as drivers of mutagenesis

Bioinformatic predictions and frequencies of mutations in the COSMIC dataset were used to prioritise three mutations in the human replicative polymerases, POLE S297F, POLE P286R and POLE V411L, and mutation accumulation experiments in Saccharomyces cerevisiae showed significant mutation increases for two of these mutations, pol2-S312F and pol2-P301R, which correspond to the human POLE S297F and POLE P286R, respectively. Other polymerase mutations conferring increases in mutation numbers are pol2-L439V, pol2-M459K and pol3-S483N.

Intriguingly, while the POLE V411L variant is the most commonly observed mutation, among the mutations studied in this thesis, in sequenced cancers, the budding yeast equivalent resulted in only a small, 1.4-fold increase over wild-type in diploid cells. As a comparison, the second most common mutation - POLE P286R (pol2-P301R) - resulted in a 27-fold increase over wild-type. This could either be due to a difference between the yeast and human version of POLE V411L or suggest that mutation frequency in cancers are not necessarily predictors of the severity of the resulting phenotype. It is also possible that the POLE V411L mutation promotes tumourigenesis by a manner other than mutation rate increases.

Exonuclease deficient strains, which are expected to show mutation rate increases, were included as a reference. The pol3-01 and pol2-4 alleles result in mutations of two acidic amino acids, involved in metal ion coordination, affecting proofreading, but not the polymerase activity of the encoded proteins. Considering that candidate polymerase mutations are located in the exonuclease domain and should affect the exonuclease activity, we expected mutation number increases to fall between wild-type and pol3-01 or pol2-4 strains. Surprisingly, the increases observed for pol3-01 and pol2-4 heterozygous diploid cells, were only 1.7- and 3.3-fold over wild type, respectively, meaning that every mutator strain identified in this work showed mutation increases exceeding those observed in the corresponding exouclease deficient strain.

Recently, some of these findings have been validated by work from another group using classical reporter gene assays on strains carrying pol2-P301R or pol2-4 mutations[888]. Here, mutation rate estimates for pol2-P301R also exceeded those from pol2-4 strains.

How these polymerase mutations exert their mutagenic potential is a question that remains open. One possibility is that the pol3-01 and pol2-4 alleles may not be truly proofreading deficient. While this would explain how other mutations in the exonuclease domain could be more deleterious, it is unlikely since these alleles have been studied extensively in vivo and in
Another possibility is that instead of a reduced exonuclease activity, these polymerase mutant strains actually have a hyperactive exonuclease, leading to removal of correctly paired nucleotides and idling. This could also explain why mutating the catalytic residues of the POL2 exonuclease domain to alanine alleviated the mutator phenotype of strong mutators as for example pol2-P301R. However, it is possible that combining polymerase mutations with mutations in the exonuclease catalytic residues results in structural changes not present in the initial mutant protein. Thus, I have not excluded the possibility that the reduction in the mutator phenotype severity is due to the loss of exonuclease catalytic activity, specifically. Indeed, a polymerase that excises correctly paired nuclotides would lead to decreased processivity. In such a case mutagenesis could arise from a less accurate DNA polymerase having increased access to the replication fork to compensate for the less processive replicative polymerase.

As recent work has placed Polζ in the replisome[885, 886] and indicates it can take over for Polε in cases of destabilizing mutations in its subunits[883, 884], the catalytic subunit of Polζ, REV3, was a natural target for our work to identify the source of mutagenesis in polymerase mutant strains. Unlike the dpb2 mutagenesis, which seems to be REV3-dependent, the pol2-P301R-dependent mutagenesis is potentiated in the absence of Rev3. Thus, it seems that, if anything, Rev3 is protective against pol2-P301R-dependent mutagenesis rather than introducing mutations. One hypothesis could be, that the missense mutation in the exonuclease domain does not just affect proofreading accuracy, but causes hyperactivity which leads to processivity decreases. A less processive, stalling polymerase could then be switched out more often in the replisome for Polζ, which as recent research indicates also plays a role in the replication of undamaged DNA, thus decreasing the access of the mutated polymerase epsilon to DNA during replication. If the mutagenesis is not entirely due to decreased proofreading - as indicated by mutating the exonuclease catalytic residues in mutated polymerase genes - and if it is not due to Polζ, the question remains which process introduces these mutations. It is possible that yet another polymerase is responsible, but it could also be that they are due to a loss of fidelity in the polymerase active site of Polε. The exonuclease and polymerase active site are on the same polypeptide and it is possible that a point mutation in one domain also affects the activity of the other. To test whether this occurs, I propose generating a mutated Polε, that carries the P301R mutation as well as mutations inactivating the catalytic activity of the polymerase domain. Since mutations in the catalytic residues of POL2 have been reported lethal[279], it is likely that this construct would also be lethal in haploid yeast cells, which is
why this work would be carried out in heterozygous diploids. If successful, these experiments could determine whether misincorporation, rather than deficient proofreading, by Polε causes the increased mutagenesis in Polε mutants.

Beyond this, it is known that the composition and concentration of the dNTP pool is correlated with mutation rates[889]. While difficult to explain it is conceivable that an altered polymerase can lead to imbalances in nucleotide pools, which are known to affect the accuracy of DNA replication. In fact, recent work has shown that mutagenesis due to DNA polymerase mutations in the polymerase domain of the protein depends on DUN1, which is known to stimulate ribonucleotide reductase (RNR) activity, which in turn is responsible for precise regulation of dNTP pools[890–892]. In this model, defective polymerases lead to an accumulation of incomplete replication intermediates, which in turn leads to checkpoint activation[892]. Checkpoint activation increases dNTP levels via an activation of Dun1. At these increased dNTP levels, a mutated DNA polymerase will more readily extend the incomplete termini and likely make more misinsertions. While a lot of this work has focused on mutations in the polymerase domain of DNA polymerases, it will be interesting to explore if a similar mechanism operates in our exonuclease domain mutated strains. Considering that dNTP pool levels seem to correlate with polymerase mutator severity, targeting dNTP pools could be a target for therapy of polymerase-mutated cancers[891].

To test the interactions between the mutator phenotypes caused by mismatch repair (MMR) deficiency and that caused by polymerase mutations, we tried to obtain double mutants to examine how mutation numbers and two distinctive mutational patterns interact. By doing so, we found that pol2-P301R, pol2-S312F and pol2-M459K were lethal in combination with a deletion in MSH2, a key mismatch repair player. Similar results have been obtained for simultaneous loss of mismatch repair and exonuclease activity by others in haploid yeast and mice[338, 769, 893, 894]. In yeast, recent work has pointed to a threshold of mutation rates that are acceptable: any higher mutation rate results in replication error-induced extinction (EEX)[895, 896].

Interestingly, it has been reported that the phenotype of exonuclease domain mutations found in cancer depends MMR deficiency[897] and that mutation of MSH2 and MSH6 mutations in addition to exonuclease polymerase mutations is a common event[806]. In fact, there are known cases of children with inherited biallelic mismatch repair deficiency that acquired early somatic driver mutations in DNA polymerase ε or δ[898]. Of those polymerase mutations identified in the children’s brain tumors one, POLE S297F, is included in this work as
well. Its yeast equivalent, pol2-S312F, is however lethal with msh2Δ in haploid cells. This discrepancy could either be due to the MMR status of the cancer cell, or to the fact that the yeast cells were haploid and did not have a wild-type DNA polymerase. Equally, it is possible that the cancer cells have acquired suppressor mutations that allow for this otherwise lethal combination.

The acquired mutations in polymerase mutant strains were used to visualise trinucleotide mutational patterns in the strongest mutators, pol3-S483N and pol2-P301R, and compare them to the wild type. While the former shows a pattern fairly similar to the wild-type, the latter shows three distinct peaks among other more subtle differences to the wild-type. Mutational signature extraction predicts three signatures in the data and signatures obtained by different algorithms produce similar, but subtly different results. While similarities between the human signature 10 and the signature extracted from yeast polymerase ε mutants can be seen, there are striking differences between the two as well. That being said, COSMIC signatures are extracted from an amalgamation of patterns in vastly more mutational data. While the 8815 mutations used for signature extraction in this work are sufficient to extract signatures, the human cancers COSMIC signatures are derived from are based on hundreds of thousands of mutations. That and inherent difference between human and yeast genomes and mutational processes can account for these differences. Similar differences can be seen between mutational patterns in the yeast strains and the human cancer samples reported by Shibrot et al. [899]: while the POLE-Pro286Arg mutated samples do show similarly low relative levels of C>G>G:C mutations and high relative levels of C>G>A:T mutations, the human samples show very low contributions of T:A>A:T mutations, while in the yeast strains they contribute approximately a quarter of all mutations to the overall mutational pattern. They report that TCG→TTG and TCT→TAT mutations account for >50% of the mutations found. However, in the yeast samples this is not the case: the TCT→TAT is one of the most common changes, but it accounts for less than 10% of all changes, while the TCG→TTG mutation is not common in the mutated yeast cells. Again, this could be due to differences in observed mutation numbers, inherent biological differences between human and yeast or the fact that the human cancers with 100 mutations per Mb will have acquired more mutations which can further contribute to the mutation pattern. Further work will show whether these differences hold true when mutation accumulation experiments are performed with mammalian cell lines.

Due to the limitations of variant calling algorithms when it comes to analyse repetitive sequences, acquired mutations were only identified from non-repetitive sequences. However,
repetitive regions can provide valuable information such as copy number estimates for rDNA repeats.

In this thesis, I detailed our approach to estimate rDNA repeat number from whole-genome sequencing data. Validation using strains of known rDNA copy number shows that our approach estimates copy number as accurately if not better than Southern blotting. While we could not identify striking deviations from copy numbers in wild-type strains, this technique could be used to identify as of yet unknown regulators of rDNA copy number in budding yeast.

5.3 Future directions

While this work provides initial insights into the effects of a collection of DNA polymerase mutants, further work will address how exactly these mutations cause mutation rate increases. For example it will be interesting to explore the difference between haploid, heterozygous and homozygous diploid mutants. Furthermore, it will be important to measure protein levels in heterozygous diploid cells to determine whether different mutation rates result from different ratios of wild-type versus mutated proteins. Additionally, the production of recombinant polymerases will allow us to determine whether these have defect in polymerisation, exonuclease activity or processivity in vitro.

To further validate the synthetic lethality between polymerase mutants and mismatch repair deficiency, I am planning to confirm it using plasmid eviction of a MSH2-carrying plasmid from a strain carrying both a polymerase mutation and a genomic deletion of MSH2. To define a possible mutation load threshold, combinations of other polymerase mutations, such as the pol2-L439V, and mismatch repair deficiency will be attempted in haploid cells. Considering their severe mutator phenotypes in single mutant cells, haploid double mutants of pol2-P301R and pol3-S483N may be inviable.

Indeed, if lethality arises from the increased mutation rate leading to an increased chance of mutating essential genes, mutation combinations that are lethal in a haploid background, could be viable in a diploid background. For instance, homozygous msh2Δ combined with heterozygous pol2-S312F could be viable, which would be in agreement with the existence of cancer cells with this genotype. Thus, using plasmid eviction I will attempt to construct diploid cells that are mismatch repair deficient and carry one of the polymerase mutations.

Considering our success with suppressor screens, we are also considering a screen to identify mutations that suppress or enhance the strong mutation rates of some polymerase mutations. The former could identify the manner in which these mutations operate, the latter could
identify targets for a synthetic lethality approach to treatment of affected patients. Suppressors for the synthetic lethality of mismatch repair deficiency and some polymerase mutations could also be aimed for.

While much of my work on mutation rates and signature has focused on polymerase mutants, it is entirely possible to extend this work to virtually any budding yeast mutant. In fact, as part of the COMSIG consortium, we will screen the entire yeast gene deletion collection for genes likely to regulate rDNA copy number maintenance. Beyond that we are on track to identify mutator phenotypes and mutational signatures for a wide array of nuclear gene deletions.

In summary, there are many questions we are looking forward to answer relating to the mutagenic potential of mutated DNA polymerases. We expect to uncover more answers by exploring genetic interactions with other mutations and examining key DNA polymerase mutations in mammalian systems. Additionally, our work will investigate the mutagenic potential of hundreds of genes and their associated mutational signatures. Hopefully, as we accumulate more information about how mutated proteins or their absence shapes a cell’s genome we will learn more about fundamental biological processes and the contributions such altered proteins can make to a cancer genome.
Chapter 6

Materials and Methods

This chapter provides further details of the materials and methods used in this work. Many of the methods used are described elsewhere in Chapters 2-4. To avoid repetition this chapter contains only additional materials and methods used during this Thesis.

6.1 Growth Medium

6.1.1 *Escherichia coli* Growth Media

**LB**

- LB mix (FM) 200g
- NaOH (10M) a few drops
- H₂O up to 8L

**LB-Amp**

- LB mix (FM) 200g
- NaOH (10M) a few drops
- 800μl Ampicillin (50mg/ml)
- H₂O up to 8L
Materials and Methods

LB-Amp-Agar

- LB mix 125g
- NaOH (10M) a few drops
- $500\mu l$ Ampicillin (50mg/ml)
- H$_2$O up to 5L
- Agar per 1L bottle 14g

6.1.2 *Saccharomyces cerevisiae* Growth Media

**YPD/ rich medium**

- Yeast extract 10g
- Peptone 20g
- H$_2$O up to 1000ml

  pH 5.4-5.7

  10x Glucose (20% w/v solution) final conc. 2%

**YPD-Agar**

- YPD medium

  Agar 2%

**Water-agar**  Made by autoclaving 500ml bottle filled with H$_2$O 300ml Agar 8g

**YNB (10X)**  Final concentration is 0.17%. Made by dissolving 8.5g in 500ml water. Filter sterilised and stored at 4°C.

**Ammonium sulphate (100X)**  Final concentration is 5g/L. Made by dissolving 5g in 500ml H$_2$O and sterilising. Stored at 4°C.
6.1 Growth Medium

**Monosodium glutamate (MSG; 100X)** Final concentration is 1g/L. Prepared by dissolving 50g in 500ml and filter sterilise. Stored at 4° C.

**Amino acids Mixture (25X)**

- L-Arginine 1.25g (f.c.: 50mg/L)
- L-Aspartate 2.00g (f.c.: 80mg/L)
- L-Isoleucine 1.25g (f.c.: 50mg/L)
- L-Methionine 0.5g (f.c.: 20mg/L)
- L-Phenylalanine 1.25g (f.c.: 50mg/L)
- L-Threonine 2.5g (f.c.: 100mg/L)
- L-Tyrosine 1.25g (f.c.: 50mg/L)
- L-Valine 3.5g (f.c.: 140mg/L)

Prepared by covering the powdered amino acids with 20ml ethanol ON at RT, then dissolved by adding 980ml H₂O. Stored at 4° C.

**Amino acid bases (100X)** (Adenine, Histidine, Leucine, Lysine, Tryptophan, Uracil) Final concentration is 100mg/L. Prepared by dissolving 5g in 500ml and filter sterilisation (Uracil is sterilised with ethanol). Stored at 4° C.

**SD (Synthetic-dropout)**

- 10X YNB (DIFCO) Solution 40ml
- 25X Amino acid Mixture
- 16ml 100X MSG or Ammonium sulphate 4ml
- 10X Glucose 40ml
- 100X Adenine 4ml

---

*1SD plates containing G418 use MSG*
Materials and Methods

100X Histidine 4ml
100X Tryptophan 4ml
100X Uracil 4ml
100X Leucine 4ml
100X Lysine 4ml

Other chemicals (G418, Thialysine) as needed

H₂O up to 400ml

This media is filter sterilised. Bases (e.g. uracil, histidine) are omitted according to experimental requirements to generate the required auxotrophic marker selection. Glucose can be substituted with other sugars as needed. SD-Agar is made by substituting the water with a bottle of melted water-agar, followed by pouring into petri dishes.

**FOA medium**  This is used to counter-select the \( UR A3 \) marker. \( UR A3^+ \) cells die on FOA medium, while \( U r a^- \) cells survive. The solution is added (after filter sterilisation once FOA has dissolved) to a sterile bottle of 200ml H₂O and 8g agar.

10X YNB (DIFCO) Solution 40ml
100X Ammonium Sulphate 4ml
25X Amino acid Mixture 16ml

100X Histidine 4ml
100X Tryptophan 4ml
100X Uracil 2ml
100X Leucine 4ml
100X Lysine 4ml
100X Adenine 4ml
100X Lysine 4ml

FOA 400mg

H₂O up to 200ml
6.2 Other solutions

**VB medium**  This is used to starve yeast cells to induce them to undergo meiosis/sporulation.

- NaAC anhydrous 8.2g
- KCl 1.9g
- MgSO\(_4\) 0.35g
- NaCl 1.2g
- Agar 15g
- H\(_2\)O up to 1L

### 6.2 Other solutions

Most solutions were prepared by the staff of the Gurdon Institute as follows.

**EDTA (0.5M, pH 8.0)**

- EDTA (Fisher) 372.2g
- NAOH pellets 100g
- 10M NaOH to pH
- H\(_2\)O up to 2L

**Sodium Acetate (3M pH 5.2)**

- Sodium Acetate (anhydrous, Fisher) 492.18g
- Glacial Acetic Acid ~200ml (enough for pH 5.2)
- H\(_2\)O up to 2L

**Sodium Chloride (5M)**

- NaCl (Fisher) 584.4g
- H\(_2\)O up to 2L
Sodium dodecyl sulphate, SDS (20%)

SDS (Melford) 800g
H₂O up to 4L

TAE (50X)

Tris 1210g
Glacial acetic acid 285.5ml
EDTA 0.5M pH8.0
H₂O up to 5L

TBE (10X)

Tris (Melford) 540g
Orthoboric Acid (Fisher) 275g
EDTA (0.5M pH 8.0) 200ml
H₂O up to 5L

TE (pH 8.0)

1M Tris pH8.0
EDTA 0.5M pH8.0
H₂O up to 2L

Tris (1M, pH 6.8)

Tris (Melford) 242.2g
Conc. HCl to pH ~160ml
H₂O up to 2L
Tris (1M, pH 7.4)

Tris (Melford) 242.2g

Conc. HCl to pH ~146ml

H₂O up to 2L

Tris (1M, pH 7.5)

Tris (Melford) 242.2g

Conc. HCl to pH ~142ml

H₂O up to 2L

Tris (1M, pH 8.0)

Tris (Melford) 242.2g

Conc. HCl to pH ~96ml

H₂O up to 2L

Tris (1M, pH 8.8)

Tris (Melford) 242.2g

Conc. HCl to pH ~36ml

H₂O up to 2L

6.3 Microbial Strains

6.3.1 Escherichia coli strains

One Shot® TOP10  F- mcrA Δ( mrr-hsdRMS-mcrBC) Φ80lacZΔM15 Δ lacX74 recA1 araD139 Δ( araleu)7697 galU galK rpsL (StrR) endA1 nupG

This chemically competent strain for plasmid construction was purchased from Invitrogen (Cat# C404010).
**XL1-Blue**  recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F’ proAB lacIqZΔM15 Tn10 (Tetr)]

This chemically competent strain used for plasmid construction was made in-house.

**MAX Efficiency® Stbl2™**  F- mcrA Δ(mcrBC-hsdRMS-mrr) recA1 endA1lon gyrA96 thi supE44 relA1 λ- Δ(lac-proAB)

This chemically competent strain was used for plasmid construction with unstable inserts and was purchased from Invitrogen (Cat# 10268019).

### 6.3.2 *Saccharomyces cerevisiae* strains

<table>
<thead>
<tr>
<th>Name</th>
<th>Genotype</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>K699</td>
<td>MATa ade2-1 trp1-1 leu2-3,112 his3-11,15 ura3 can1-100</td>
<td>Kim Nasmyth</td>
</tr>
<tr>
<td>K700</td>
<td>MATα ade2-1 trp1-1 leu2-3,112 his3-11,15 ura3 can1-100</td>
<td>Kim Nasmyth</td>
</tr>
<tr>
<td>YMH8</td>
<td>(K699) pol2::URA3-POL2</td>
<td>This work</td>
</tr>
<tr>
<td>YMH9</td>
<td>(YMH8)(K700)</td>
<td>This work</td>
</tr>
<tr>
<td>YMH10</td>
<td>(K699) pol3::URA3-pol3-S483N</td>
<td>This work</td>
</tr>
<tr>
<td>YMH11</td>
<td>(YMH10)(K700)</td>
<td>This work</td>
</tr>
<tr>
<td>YMH12</td>
<td>(K699) pol2::URA3-pol2-L439V</td>
<td>This work</td>
</tr>
<tr>
<td>YMH13</td>
<td>(YMH12)(K700)</td>
<td>This work</td>
</tr>
<tr>
<td>YMH14</td>
<td>(K699) pol2::URA3-pol2-V426L</td>
<td>This work</td>
</tr>
<tr>
<td>YMH15</td>
<td>(YMH14)(K700)</td>
<td>This work</td>
</tr>
<tr>
<td>YMH16</td>
<td>(K699) pol2::URA3-pol2-S312F</td>
<td>This work</td>
</tr>
<tr>
<td>YMH17</td>
<td>(YMH16)(K700)</td>
<td>This work</td>
</tr>
<tr>
<td>YMH18</td>
<td>(K699) pol2::URA3-pol2-P301R</td>
<td>This work</td>
</tr>
<tr>
<td>YMH19</td>
<td>(YMH18)(K700)</td>
<td>This work</td>
</tr>
<tr>
<td>YMH20</td>
<td>(K699) pol2::URA3-pol2-D290V</td>
<td>This work</td>
</tr>
<tr>
<td>YMH21</td>
<td>(YMH20)(K700)</td>
<td>This work</td>
</tr>
<tr>
<td>YMH22</td>
<td>(K699) pol2::URA3-pol2-M459K</td>
<td>This work</td>
</tr>
<tr>
<td>YMH23</td>
<td>(YMH22)(K700)</td>
<td>This work</td>
</tr>
<tr>
<td>YMH24</td>
<td>(K699) pol2::URA3-pol2-Q468R</td>
<td>This work</td>
</tr>
<tr>
<td>YMH25</td>
<td>(YMH24)(K700)</td>
<td>This work</td>
</tr>
<tr>
<td>Name</td>
<td>Genotype</td>
<td>Reference</td>
</tr>
<tr>
<td>-----------------</td>
<td>----------------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>YMH26</td>
<td>(K699) pol2::URA3-pol2-A480V</td>
<td>This work</td>
</tr>
<tr>
<td>YMH27</td>
<td>(YM H26)(K700)</td>
<td>This work</td>
</tr>
<tr>
<td>YMH28</td>
<td>(K699) pol2::URA3-pol2-4</td>
<td>This work</td>
</tr>
<tr>
<td>YMH29</td>
<td>(YM H28)(K700)</td>
<td>This work</td>
</tr>
<tr>
<td>YMH30</td>
<td>(K699) pol3::URA3-POL3</td>
<td>This work</td>
</tr>
<tr>
<td>YMH31</td>
<td>(K700)pol3::URA3-POL3</td>
<td>This work</td>
</tr>
<tr>
<td>YMH32</td>
<td>(K699) pol3::URA3-pol3-01</td>
<td>This work</td>
</tr>
<tr>
<td>YMH33</td>
<td>(K700) pol3::URA3-pol3-01</td>
<td>This work</td>
</tr>
<tr>
<td>YMH34</td>
<td>(K699) pol3::URA3-pol3-R316C</td>
<td>This work</td>
</tr>
<tr>
<td>YMH35</td>
<td>(K700) pol3::URA3-pol3-R316C</td>
<td>This work</td>
</tr>
<tr>
<td>YMH36</td>
<td>(K699) pol3::URA3-pol3-P332L</td>
<td>This work</td>
</tr>
<tr>
<td>YMH37</td>
<td>(K700) pol3::URA3-pol3-P332L</td>
<td>This work</td>
</tr>
<tr>
<td>YMH38</td>
<td>(K699) pol3::URA3-pol3-S375R</td>
<td>This work</td>
</tr>
<tr>
<td>YMH39</td>
<td>(K700) pol3::URA3-pol3-S375R</td>
<td>This work</td>
</tr>
<tr>
<td>YMH40</td>
<td>(K699) pol3::URA3-pol3-V397M</td>
<td>This work</td>
</tr>
<tr>
<td>YMH41</td>
<td>(K700) pol3::URA3-pol3-V397M</td>
<td>This work</td>
</tr>
<tr>
<td>YMH42</td>
<td>(K699) rev3::KanMX</td>
<td>This work</td>
</tr>
<tr>
<td>YMH43</td>
<td>(K699) pol2::URA3-pol2-A480V rev3::KanMX</td>
<td>This work</td>
</tr>
<tr>
<td>YMH44</td>
<td>(K699) pol2::URA3-pol2-P301R rev3::KanMX</td>
<td>This work</td>
</tr>
<tr>
<td>YMH46</td>
<td>(K699) pol2::URA3-pol2-S312F rev3::KanMX</td>
<td>This work</td>
</tr>
<tr>
<td>YMH52</td>
<td>(K700) pol2::URA3-pol2-S312F</td>
<td>This work</td>
</tr>
<tr>
<td>YMH53</td>
<td>(K700) pol2::URA3-pol2-P301R</td>
<td>This work</td>
</tr>
<tr>
<td>YMH54</td>
<td>(K699) msh2::KanMX</td>
<td>This work</td>
</tr>
<tr>
<td>YMH56</td>
<td>(K699) pol2::URA3-pol2-A480V msh2::KanMX</td>
<td>This work</td>
</tr>
<tr>
<td>YMH58</td>
<td>(K699) pol2::URA3-pol2-D290A-E292A-A480V</td>
<td>This work</td>
</tr>
<tr>
<td>YMH60</td>
<td>(K699) pol2::URA3-pol2-D290A-E292A-M459K</td>
<td>This work</td>
</tr>
<tr>
<td>YMH62</td>
<td>(K699) pol2::URA3-pol2-D290A-E292A-P301R</td>
<td>This work</td>
</tr>
<tr>
<td>YMH64</td>
<td>(K699) pol2::URA3-pol2-D290A-E292A-S312F</td>
<td>This work</td>
</tr>
<tr>
<td>YMH66</td>
<td>(K700) pol2::URA3-pol2-A480V</td>
<td>This work</td>
</tr>
<tr>
<td>YMH67</td>
<td>(K700) pol2::URA3-pol2-M459K</td>
<td>This work</td>
</tr>
<tr>
<td>YMH68</td>
<td>(YM H30)(K700)</td>
<td>This work</td>
</tr>
<tr>
<td>YMH69</td>
<td>(YM H36)(K700)</td>
<td>This work</td>
</tr>
<tr>
<td>YMH70</td>
<td>(YM H38)(K700)</td>
<td>This work</td>
</tr>
<tr>
<td>Name</td>
<td>Genotype</td>
<td>Reference</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------------------------------------------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>YMH71</td>
<td>(K700) pol3::URA3-pol3-01</td>
<td>This work</td>
</tr>
<tr>
<td>YMH72</td>
<td>(YMH34)(K700)</td>
<td>This work</td>
</tr>
<tr>
<td>YMH73-75</td>
<td>(K699) pol2::URA3-pol2-A480V msh2::KanMX</td>
<td>This work</td>
</tr>
<tr>
<td>YMH78</td>
<td>(K699)pol3::URA3-pol3-S483N(K700)pol2::URA3-pol2-P301R</td>
<td>This work</td>
</tr>
<tr>
<td>YMH81</td>
<td>(K699)(K700)pol2::URA3-pol2-P301R/pol2::URA3-pol2-P301R</td>
<td>This work</td>
</tr>
<tr>
<td>YMH82</td>
<td>(K699)(K700)pol2::URA3-pol2-M459K/POL2 msh2::KanMX/MSH2</td>
<td>This work</td>
</tr>
<tr>
<td>YMH83</td>
<td>(K699)(K700)pol2::URA3-pol2-S312F/POL2 msh2::KanMX/MSH2</td>
<td>This work</td>
</tr>
<tr>
<td>YMH84</td>
<td>(K699)(K700)pol2::URA3-pol2-P301R/POL2 msh2::KanMX/MSH2</td>
<td>This work</td>
</tr>
<tr>
<td>NOY408-1b</td>
<td>MATa ade2-1 ura3-1 his3-11 trp1-1 leu2-3,112 can1-100</td>
<td>Nogi et al. 1991</td>
</tr>
<tr>
<td>YSI101</td>
<td>(NOY408-1b) except fob1Δ::LEU2 (~150 rDNA copies)</td>
<td>[866]</td>
</tr>
<tr>
<td>YSI102</td>
<td>(NOY408-1b) except ~20 rDNA copies</td>
<td>[866]</td>
</tr>
<tr>
<td>YSI103</td>
<td>(NOY408-1b) except ~40 rDNA copies</td>
<td>[866]</td>
</tr>
<tr>
<td>YSI104</td>
<td>(NOY408-1b) except ~80 rDNA copies</td>
<td>[866]</td>
</tr>
<tr>
<td>YSI105</td>
<td>(NOY408-1b) except ~110 rDNA copies</td>
<td>[866]</td>
</tr>
</tbody>
</table>

### 6.4 Oligonucleotides

**Oligonucleotides to generate POL2 mutants by site directed mutagenesis**

<table>
<thead>
<tr>
<th>Oligonucleotide</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>pol2_S312F_SDM_fw</td>
<td>TAGATCATAATAAATGATGATTTTTTTATATGATCGATGGGGAAGG</td>
</tr>
<tr>
<td>pol2_S312F_SDM_rv</td>
<td>CCTTCCCCATCGATCATATAAAAATCATCATTATTGATCTA</td>
</tr>
<tr>
<td>pol2_V426L_SDM_fw</td>
<td>ACATGGATTGTGGCTGGGTGAAGCGTGATTCTATATTACC</td>
</tr>
<tr>
<td>pol2_V426L_SDM_rv</td>
<td>GGTAAA TAAGAA TCACGCTTCAAGCAAAACGAAACAAATCCATGT</td>
</tr>
<tr>
<td>pol2_P301R_SDM_fw</td>
<td>GCAAGGCGCCCTTTAAATTCGGGAGTCCGCGGATGATCAAATT</td>
</tr>
<tr>
<td>pol2_P301R_SDM_rv</td>
<td>ATTTGATCTACGGCGGAATCCCGGAATTTAAGGCCGCTTG</td>
</tr>
<tr>
<td>pol2_D290V_SDM_fw</td>
<td>ACCCTGTGGTAAATGCCATTGTTGATAGAAACCACGAAAGCCGCC</td>
</tr>
<tr>
<td>pol2_D290V_SDM_rv</td>
<td></td>
</tr>
</tbody>
</table>
6.4 Oligonucleotides

GGCGGCTTCGTGTTTCTATAACAAATGCCATTACCACACAGGGT
pol2_M459K_SDM_fw
TTGAACTGGAATCCCGAAATTAAAG
pol2_M459K_SDM_rv
TTTTCAATTCGACCGCTCTTTAATTCCGGATCCAGTTCAA
pol2_Q468R_SDM_fw
CGTATGCATTGAAAAGCCAACGCCACCTTTCCGAAATTTCTGT
pol2_Q468R_SDM_rv
ACAGAATATTCGGAAGGTCGCCGGCTTTTCAAATGCATACG
pol2_A480V_SDM_fw
ATTCTGTTTCCGATGCAGTCTACGTATTACCTTTACATGAA
pol2_A480V_SDM_rv
TTCATGTAAGGTAAATACGTAACGACTGCATCGGAAACAGAAAT
Pol2Faiat-SDM-fw
CCCTGTGGTAAATGGCAATTT
Pol2Faiat-SDM-rv
TTTAAGGCCGCTTCGGGTTGCTATAGCAACACGAAGGCCGCCTTTAAA
Oligonucleotides to generate POL3 mutants by site directed mutagenesis

pol3_FAIAC_SDM_fw
TGCGTATCATGTCCTTTGATATCGAGTGCTGGTAGGATTTGG
pol3_FAIAC_SDM_rv
CCAATCTTACCAGCACAATCGATATCAAAAGGACATGATACGCA
pol3_V397M_SDM_fw
TCATCAAGTTGATCCTGATATGATCATTTGGTATAATAACTAC
pol3_V397M_SDM_rv
GTAGTATTATAACCAATGATCATCAGGATCAAACCTTTGATGAA
pol3_R316C_SDM_fw
GGTCTCAATACAGCTCCAATTGATATCATGTCCTTTGATATCGA
pol3_R316C_SDM_rv
TCGATATCAGGACATGATAACAAATGAGCTGTATGAGACC
pol3_S483N_SDM_fw
CCTACAGTTGAATGCAATGCAATGCAGCACTTTTTAGGTGAACA
pol3_S483N_SDM_rv
TGGTACCTAAAAGAGTGCACATTGACTGCATTCAACGTTAGG

Pol2Faiat-SDM-fw
Pol2Faiat-SDM-rv

Materials and Methods

**Oligonucleotides to check polymerase mutant generation**

Sc-pol3_out_fw GAAGAGCATGACCTGTCATCATTC
pRS306_out_rv GACCATGATTACGCAAGCTCG
pol3_sq1 TACCAAAAGGAAATTTCCG
pol3_sq2 GTCATCCAAATGCCAAAGT
pol3_sq3 ACTACAAATTTTTGATATCCC
Pol2_promoter_rv: 5’GATCCATATTTGACACAGGCTGTT
pRS306_fw: 5’GGCGGACAGGTATCCGGTAAG

**Oligonucleotides to delete MSH2**

MSH2-F1 TTAATCTGTGACCTAACAATTAATGAAGATGCTCTAGTTATATATAATTA
MSH2-R1 TATCTATAGTTTCTAAGATGCTCTAGTTATATATAATTA
MSH2.3 TAAAGCCAATGAATTTACG
MSH2.4 TTTCCAGTGGTCTAGGACC

**Oligonucleotides to delete REV3**

REV3-F1 ATACAAATTAATGTTTGTGGGCGAATTTAATTTGGAAA CGGATCCCGGCTGTTAATTA
REV3-R1 ATAACTACTCATTTTTGCGAGACATATCTGTTAGTTAATTA
REV3.3 ACTGTGTAGAGAAAAAGAC
REV3.4 AATGTGTGGGAGAACCTATACG
6.5 Solutions

**Oligonucleotides to check the integration of polymerase mutant constructs into MEFs**

EHom1\_1fw GCTTGGGTGATGTTTGCTCGTCTGTTAAA
EHom1\_1rv CCGCGCTGTTCCTCCTTCTTCTCATCCT
EHom2\_1fw GCGGCA TGGACGAGCTGTACAAGTGA TTA
EHom2\_1rv CCGCGCTGTTCCTCCTTCTTCTCATCCT
D1Hom1\_1fw AGAGAATTGCTGAGAAAGGGAGGAGAC
D1Hom1\_1rv CCGCGCTGTTCCTCCTTCTTCTCATCCT
D1Hom2\_1fw CCGCGA TAA TGAGCCTGAAGGAGAC
D1Hom2\_1rv CCGCGCTGTTCCTCCTTCTTCTCATCCT

**Oligonucleotides to generate the POLE P286R mutation in human cells**

sgRNA-1 5’-ACCG-ATCTGGTCTGTCTCAGCA TC-3’
and 5’-AAAC-GA TGCTGAGACAGACCAGA T-3’

sgRNA-2 5’-ACCG-TCGATGGCCAGGTGACAGG-3’
and 5’-AAAC-CCTGCTCACCTGGCCCATCGA-3’

ssODN (all mutations in lower case) CAAGGTCCCCATCCCAGGAGCTTACTTCCAGAAG-
gCACCTGCTCACCTGGCCAT CGATCATGTAGGAAATCATCATAATCTGGTCTGTCTCAGCAT-
CAcGAAAATTTGAG GGGCAGGTCTTGGCTCAATGCAAATGCAGCAACCAGCAGGCT-
CTGTGGGGA CAAAA TAAGCA TAAAGCCAAGCTCTAAACTCCCCA

6.5 Solutions

**TE (1X)** Tris-HCl pH 7.4 10mM
   EDTA 1mM

**TAE(1X)** Tris-Acetate pH 8.0 40mM
   EDTA 10mM

**Gel Loading Dye, Purple(6X)** Purchased from NEB (Cat# B7024S).

**HyperLadder™ 1kb** Purchased from BIOLINE (Cat# BIO-33026).
6.6 Protocols

**DNA restriction** DNA is digested with appropriate restriction enzymes according to specification of the supplier (New England Biolabs).

**DNA ligation** Fragments of DNA were ligated using the Quick Ligation™ Kit according to the specifications of the supplier (New England Biolabs).

**Agarose gel electrophoresis** DNA to be run on the gel is mixed with 1/6 Volume of 6X Gel Loading Dye and loaded onto an agarose gel (0.6-2%) containing 5μg/ml Ethidium bromide. A molecular marker (Hyperladder 1kb) is also loaded for size measurements. The gel is run in 1X TAE buffer and DNA is visualised under UV-light (260nm).

**Plasmid extraction from Escherichia coli** Plasmids were extracted from *E. coli* grown overnight in the appropriate culture medium using the QIAprep Spin Miniprep Kit (QIAGEN) as directed.

**DNA extraction from agarose gels** After gel electrophoresis a small slice of agarose, containing the DNA to be purified, is excised from the gel, weighed and the DNA is extracted using the QIAquick® Gel Extraction Kit (QIAGEN). A small aliquot is run on an agarose gel to assess the quality and efficiency of purification.

**DNA precipitation** 1/16 volume of KAc 3M pH 5.0 and 1 volume of Isopropanol (propan-2ol) are added to the DNA solution. Samples are spun for 10’ at top centrifuge speed at RT and the supernatant is discarded. The pellet is washed with 1ml of 70% (-20°C) EtOH and the pellet is dried. The pellet is resuspended in 10-30μl TE buffer or water.

**PCR (Polymerase chain reaction)** PCR uses DNA as a template to amplify a target DNA fragment. Two oligonucleotides, flanking the fragment, acting as primers for the polymerase are required. The DNA polymerases used are Taq (qiagen 201203), Phusion (NEB #M0530L), The reaction mix contains:

- Template DNA 25-100ng (depending on whether it is plasmid or genomic)
- Oligonucleotides 20pmol each
- 10X DNA polymerase buffer 5μl
- dNTPs (2mM each) 5μl
- DNA polymerase 2units
Reactions are carried out in cycler machines from and consist of the following steps:

1. First denaturation 2’ @94°C
2. Denaturation 1’ @94°C
3. Annealing 1’ @Tm-5°C
4. Extension 1’ per kb of target fragment size + 2’ @72°C
5. Repeat steps 2-4 for 25-30 cycles
6. Final extension 10’ @72°C

The Tm is the lower melting temperature of the two oligonucleotides. All parameters can be adjusted depending on the DNA template, the purpose of the PCR and the DNA polymerase. For instance, for a yeast colony PCR (a diagnostic PCR where whole yeast cells are added to the reaction mix skipping the DNA extraction step), Step 1 should be increased to 7’ to allow breaking of the cells and liberation of genomic DNA.

**Site-directed mutagenesis** Perform using Agilent Technologies QuickChange Lightning Kit (#210519-5) according to the manufacturers’ instructions. Primers are designed according to the manufacturers’ instructions (see Chapter 6.4 and 6.4).

**Escherichia coli transformation** Chemically competent cells are transformed with DNA according to the manufacturers’ protocols.

**Saccharomyces cerevisiae transformation** The strain to be transformed is grown up in 50μl of the appropriate medium until the culture has reached a concentration between 5x10⁶ and 1x10⁷ cells/ml. The cells are pelleted and washed with 25ml of sterile water. Cells are resuspended in 500μl of water of which 100μl are used for a transformation. Cells are pelleted again and resuspended in 360μl transformation mix (33% PEG-4000, 0.1M LiAc, 0.27mg/ml salmon-sperm DNA) and an appropriate amount of transforming DNA is added. The suspension is incubated at 42°C for 5’ (plasmid transformation) - 40’ (a transformation requiring an integration event). Cells are pelleted and washed with sterile water, resuspended in 200μl water and plated on selective medium. Should the selection require some time for gene expression (for instance resistance to G418) cells are suspended in rich medium and grown for 2hours at 30°C before plating.

**Saccharomyces cerevisiae ONE-STEP gene deletion and tagging[900]** To generate a transformation cassette that features the selectable marker flanked by two regions of homology
suitable oligonucleotides are designed and ordered. The transformation cassette is amplified by PCR Mix preparation:

- 5μl F1 Oligonucleotide
- 5μl R1 Oligonucleotide
- 50μl 2mM dNTPs
- 50μl 10x Taq/Dynazime Buffer
- 5μl pFA6 template plasmid (1:20 QIA)
- 382.5μl H2O
- 2.5μl Taq/Dynazime

The solution is mixed and 100μl are aliquoted in each tube.

Program:
- 2' @94°C
- 1' @94°C
- 1' @45°C
- 4' @72°C - 5 cycles
- 1' @94°C
- 1' @52°C
- 4' @72°C - 30 cycles
- 10' @72°C

The PCR product is purified with Gel Cleanup Kit (Eppendorf)/Gel Cleanup System (Promega) without band extraction and resuspended in 30μl H2O. The yield is checked on a gel and 1-2μg (usually 5-6μl) is transformed into yeast cells using standard transformation protocols. Plates are replicated at least once and at least 8 single colonies are isolated to check integration of the cassette. Deletion is checked by colony PCR (and subsequently perhaps by Western blotting): a small amount of cells is placed in a PCR tube. The following mix is prepared and 50μl of it is aliquoted into each PCR tube:

- 1μl FOR Oligonucleotide
- 1μl REV Oligonucleotide
- 5μl 2mM dNTPs
- 5μl 10X Taq/Dynazime Buffer
- 7.5μl H2O
- 0.5μl Taq/Dynazime

The PCR is run with the following programme:
- 7’ @94°C
- 30” @94°C
6.6 Protocols

30” @50/42°C (ca 5° below lower melting temperature)
4’ @72°C - 45 cycles
10’ @72°C
PCR products are checked on an agarose gel.

**Saccharomyces cerevisiae gDNA extraction**  Collection - Cells were collected by pelleting 50ml of yeast culture (107 cells/ml, 3000rpm, 2min). Cells were washed with 1ml 0.9M sorbitol 0.1M EDTA, the supernatant is discarded and cells are frozen for storage.

Extraction - Cells are resuspended in 400μl 0.9M sorbitol 0.1M EDTA 14mM β-mercaptoethanol with 100μl of 4-5mg/ml zymolase and incubated at 37°C 30-45 minutes, 850rpm shaking. Cells are centrifuged for 30” at 13,000rpm, the supernatant is removed and cells are resuspended in 400μl of 1Z TE (pH8) with 90μl of the following freshly prepared solution: 1.5 ml of EDTA ph8.5 + 0.6 ml TRIS base 2M + 0.6 ml SDS 10%. The solution with the cells is gently mixed and incubated for 30min at 65°C, shaking 850rpm. 80μl Potassium Acetate 5M is added and cells are incubated 60min on ice. Cells are spun 15min at 13,000rpm at 4°C, the supernatant is decanted into a new tube, 500-1000μl 100% ethanol (EtOH) kept at -20°C is added and the liquids are mixed by inverting. Samples are left 30min at -80°C or at -20°C overnight to precipitate the DNA. Tubes are centrifuged 5min at 13,000rpm at 4°C, the supernatant is discarded and the pellet is washed with 1ml of chilled 70% EtOH (centrifuged 5min, 13,000rpm). The supernatant is removed, the DNA pellet is allowed to dry and the DNA is resuspended in 500μl 1X TE. Once resuspended, 5μl RNaseA is added and incubated for 30min at 37°C. Green phenol/chloroform tubes are pulsed down and the DNA solution is added. 500μl of phenol/chloroform is added, the solutions are mixed by vortexing and the tubes are centrifuged for 5min at maximum speed. The layer of liquid above the gel phase is moved to fresh tubes, 0.5ml isopropanol is added and the liquids are mixed by inverting. Samples are centrifuged 15min at 13,000rpm, the supernatant is discarded and the pellet is washed with 1ml 70% EtOH and the DNA is allowed to dry, then resuspended in 50μl 1X TE.

**Saccharomyces cerevisiae high-throughput gDNA extraction**  Cells were grown in 2ml 96-well plates 1.5ml YPD at 30°C shaking for 48 hours. Then, plates were spun down at 4000rpm for 5’ and the supernatant removed. Cells were resuspended in 500μl of:

- 22.5ml 2M sorbitol
- 10ml 0.5M EDTA
- 50μl 14mM \(\beta\)-mercaptoethanol
5ml RNase A (stock 10mg/ml)
12.5ml H2O
200 – 250mg zymoliase
and incubated for 2 hours at 37°C with shaking, followed by spinning down and removal of the supernatant. Cells were resuspended in 200μl of:
16ml ATL buffer (qiagen)
2ml proteinase k (qiagen)
2ml RNAse A (stock 10mg/ml)
and incubated at 56°C with shaking for 24 hours. The plate was placed on the robot (CAS1820 by Corbett Robotics) which carried out the following steps in a 96-well format.

1 | 400μl of buffer AL mixed 50:50 (qiagen 19075) with 100% ethanol was added to samples using filter tips (qiagen 990610).
2 | After mixing the total volume (600μl) was loaded onto a capture plate (qiagen 950901) and vacuum applied at 70kPa for 2’30”.
3 | The capture plate was washed twice with 600μl of buffer DXW (qiagen 950154) and once with 600μl of buffer DWF (qiagen 950163).
4 | The vaccum was applied at 30kPa to remove remaining liquid.
5 | 100μl of buffer E (qiagen 950172) was added, incubated for 30” and vaccum applied for 5’ at 50kPa to elute samples into the elution plate (qiagen 990602).

The eluted samples were then transfered into a 96-well plate for sequencing.

**Mating *Saccharomyces cerevisiae***

There are two different options for mating two haploid cells to generate a diploid.

1 | Two small quantaties of yeast cells are mixed on a YPD plate and incubated at 30°C. The next day a small quantity of yeast is suspended in 50μl H2O and a drop of 30μl is placed on a new YPD plate and the plate is tilted to spread the cells thinly. Under the dissection microscope roughly 10 diploids are identified (which at this stage appear in a “dumbbell” shape) and placed to an empty space on the plate.

2 | Small amounts of strains to be mated are inoculated in 5ml YPAD and incubated static overnight at 30C. Cells are resuspended and diluted 1:2000 in ddH2O. 100μl are plated on a YPAD plate and incubated overnight at 30°C. The next morning approximately 10 of the small colonies are picked and spread on a new YPAD plate. Diploid colonies are generally bigger, thus are picked first.

In both cases colonies are checked for ploidy by FACS and/or sporification.
6.7 Automated serial propagation platform

The evolving populations are maintained on top of agar surfaces in a home-made evolution chamber that controls moisture, light and temperature. Cells are kept in a 1536-well plate format on the agar surface and every fourth position is left empty. At each transfer, evolving populations are i) pinned onto the next evolution plate ii) pinned onto a scanning plate. The plate is scanned to track colony growth. Copies of evolution plates are stored as a frozen record at regular intervals. At the end of the timespan the final evolution plate is deconvoluted and populations are preserved in 96-well plates filled with 30% glycerol. The platform has since been published here [836].

6.8 Illumina sequencing

1-3 μg of extracted DAN was then supplied to the Sequencing Facility at the Wellcome Trust Sanger Institute, who sheared the DNA to 100-1,000 bp by using a Covaris E210 or LE220 (Covaris, Woburn, MA, USA) and size-selected fragments (350-450 bp) with magnetic beads (Ampure XP; Beckman Coulter). Illumina paired-end DNA library preparation were prepared by the Sanger, samples indexed and multiplexed. The DNA was sequenced on the Illumina HiSeq2500 generating 100bp paired-end reads which were aligned by the Sanger to the *S. cerevisiae* S288c assembly (R64-1-1/EF4) from Saccharomyces Genome Database (obtained from the Ensembl genome browser) using BWA[901], currently considered one of the most efficient alignment tools[860], and PCR duplicates were marked by using Picard ‘MarkDuplicates’[902](see B.1.1.2).

6.9 Sequencing analysis

For parameters of all programmes used for sequencing analysis used see Appendix B.1.1.2 and for scripts written in the course of this work see https://github.com/mareikeherzog/thesis-scripts.

6.9.1 Quality control of DNA sequencing

Extracted DNA was tested for total volume, concentration and total amount by the sequencing facility of the Wellcome Trust Sanger Institute using gel electrophoresis and the Quant-iT™ PicoGreen® dsDNA Assay Kit (ThermoFisher Scientific). The quality of the sequencing data
post-alignment was assessed using SAMTools stats (1.1+htslib-1.1), plot-bamstats, bamcheck and plot-bamcheck[903].

6.9.2 Alignment of sequencing reads to the reference genome

Fastq files were aligned to the relevant reference genome using BWA[901] and PCR duplicates marked using Picard MarkDuplicates[902] by the Wellcome Trust Sanger Institute. Where required files were realigned to a different reference genome using the same tools.

6.9.3 Variant Calling of SNPs and INDELs, Annotation and Filtering

Variant Calling was carried out using SAMTools mpileup[903], BCFtools call[903] and Scalpel[853]. Variants were annotated with Variant Effect Predictor[904] and vcf files were processed using BCFtools, VCFtools[905], BEDTools[906] and custom scripts.

6.9.4 Extracting mutational signatures

To extract mutational signatures, the SomaticSignatures[763] R package was used. A custom script was used to format all mutations from all strains into the required input format. An R wrapper script written by Kim Wong was used to run the different functionalities of the SomaticSignatures package in sequence following their methodology[907]. The number of signatures was set to 2-8. The normalizeMotifs function was used to normalize to whole genome trinucleotide frequencies. Signatures were also extracted using EMu[764].

6.9.5 Scripts written for this work

To analyse sequencing data software detailed in the previous sections was used (see 6.9.6 and B.1.1.2 for commands and parameters). However, some analysis steps required the use of scripts written specifically for that particular analysis. Scripts used to generate data detailed in this thesis are described below. The code for these scripts is stored in an online repository (https://github.com/mareikeherzog/thesis-scripts)

**av_cov_bait_regions.pl**  A programme that takes a bam file as an input and calls a file that contains list of genomic regions. The script then uses samtools mpileup output to work out the coverage across all the bases within those regions and returns the average coverage. Written for mouse, but can be adapted to other organisms. Used in WES experiments to check the average coverage across regions covered by the baits.
bam_stats_table.pl  A script turn the output from samtools stats in a table with key QC metrics to quickly check for substandard sequencing data.

bamtofastq.pl  A script that generates commands like “bam2fastq -o reads#.fastq 13791_2#1.bam” for samples of interest.

budding_yeast_gene_name_conversion.pl  A script that takes a list of S. cerevisiae systematic gene names and returns their standard name and a description. This basic operation has been re-purposed for other scripts that handle S. cerevisiae vcf files. The equivalent for S. pombe has also been written (fission_yeast_gene_name_conversion.pl).

consequence_display.pl  This script will go through a vcf file and count the consequences of the mutations that were called. If a mutation is associated with more than one consequence the one deemed more sever will be displayed. (Severity is indicated by the order of consequences in the array e.g. a gained stop codon is judged more severed than an inframe deletion). Other variations of this script have been written to deal with multi-sample vcfs, distinguish between SNVs and INDELs or categorise mutations as ’coding’, ’intronic’, ’regulatory’ and ’non-coding’.

coverage_of_gene_mouse.pl  This script can be used to get the coverage across all exons of a specific gene for mouse WES bam files. A variation to do the same for human sequencing data has also been written.

filter_bait_regions.pl  This is used in the analysis of mouse WES or targeted exon sequencing experiments. The script takes a vcf file (variant calling from the experiment in question) and a bed file that contains the genomic locations of the regions of interest (in a standard WES experiment that would be a file containing the location of all mouse exons). The script then removes all variants from the vcf file that do not fall within a region of interest.

gt-filter.pl  These custom filters for vcf-annotate allow filtering of vcf files on three metrics. Genotypes set to . for samples with DP < 10, Genotypes set to . for samples with GQ < 95 and a minimum value of MQ>30 is required. Written with the help of Dr. Thomas Keane and Shane McCarthy.
intersect_vcf_mutlists.pl  This script was used to check whether all mutations introduced into simulated genomes were actually found by the simulated sequencing and subsequent analysis and are present in vcf files or whether mutations found in the analysis were present in the mutation lists.

mask-hets.pl  This custom filters for vcf-annotate will set genotypes to . for all mutations that are heterozygous e.g. 0/1, 1/2, etc. mask-homs.pl to remove homozygous mutations has been written, too.

merge_bams_samtools.pl  A script that takes a list of bam file locations and, if the sample names of two successive bam files are the same, merges them into one bam file.

rDNA_cnv_estimate.pl  A programme that will estimate the copy number for rDNA repeats.

raindrop_plot_distances_morechr.pl  A script that takes a vcf file of mutations and outputs the distances between mutations in a way that they can be plotted with gnuplot to make a raindrop plot.

remove_shared_variants.pl  A script that takes a multi-sample vcf file and removes mutations that occur in more than one sample. A variant to only remove mutations present in all samples has been written.

samtools.stats.cov.pl  A script that takes samtools stats output and computes how many nucleotides have a coverage less than 5 or a coverage less than 10.

subset_loop_no_conversion.pl  A script that takes a multi-sample vcf file as input and utilizes the vcf-subset command to separate the vcf file into its samples.

ty-realign.sh  A script that takes a list of bam files, locates the corresponding fastq files and realigns them to the Ty custom reference genome.

vcf_stats_table_all.pl  A script that will take the output of vcf-stats and output a table with the information such as INDEL_Count, SNV_Count, Transitions, Transversions, C>T, A>G, A>T, C>G, G>T, A>C as well as different lengths of small INDELs.
**6.9 Sequencing analysis**

**vcf_to_gene_list.pl** A script that turns a vcf file into a table of mutations that affect genes. The information printed is: the type of mutation (SNV or INDEL), the chromosome, the position, the gene (its systematic and common name and a description), the consequence of the mutation (e.g. frameshift mutation), the number of homozygous and heterozygous mutations found across all samples in the vcf file, the names of samples carrying the mutation.

**6.9.6 Step-by-step workflow of variant analysis**

After quality control and alignment to a reference genome, analysis to extract variants present in samples that are not present in controls was carried out with the following steps and commands (see also B.1.1.2 for command parameters and 6.9.5 used):

**Step1:** Variant calling was performed against a reference genome

- S. cerevisiae: samtools mpileup -f Saccharomyces_cerevisiae.EF4.69.dna_sm.toplevel.fa -g -t DP,DV -C50 -pm3 -F0.2 -d10000 sample.bam | bcftools call -vm -f GQ > sample.vcf
- Mouse: samtools mpileup -f GRCm38_68.fa -g -t DP,DV -C50 -pm3 -F0.2 -d10000 sample.bam | bcftools call -vm -f GQ > sample.vcf
- optional (INDELs only): scalpel --somatic --normal control.bam --tumor sample.bam --bed WES_regions.bed --ref genome.fa

**Step2:** Ensembl variant effect predictor (VEP) was run on the vcf files

- variant_effect_predictor.pl --species saccharomyces_cerevisiae|mus_musculus -i sample.vcf --format vcf -o sample.vep.txt --force_overwrite --database
- vcf2consequences_vep -v sample.vcf -i sample.vep.txt > sample.csqs.vcf

**Step3:** The vcf files were checked for expected mutations (e.g. check for deletions, polymerase mutations or other expected mutations that should be present)

**Step4:** Filtering

- for mouse WES: perl filter_bait_regions.pl -i sample.csqs.vcf > sample.ex.vcf
- for scalpel generated vcf files: cat sample.somatic.vcf | vcf-annotate -f gt-filter.pl > sample.filt.vcf
Materials and Methods

- `bcftools norm -f Saccharomyces_cerevisiae.EF4.69.dna_sm.toplevel.fa|GRCm38_68.fa
  sample.csqs.vcf > sample.norm.vcf`

- `cat sample.norm.vcf | vcf-annotate -H -f +/-q=30/Q=50/SnpGap=7 > sample.annotate.vcf`

- `cat sample.annotate.vcf | vcf-annotate -f gt-filter.pl > sample.gq.vcf`

- Optional for haploid samples: `cat sample.gq.vcf | vcf-annotate -f mask-hets.pl > sample.hets.vcf`

**Step 5:** Files were subjected to vcf-subset to remove variants that did not pass filters

- `subset_loop_no_conversion.pl` -> carries out the following command in a loop: `vcf-subset -c sample_name sample.vcf -e > sample.sub.vcf`

**Step 6:** Sample files were intersected to remove any variants not acquired in the course of the experiment

- `cat sample.sub.vcf | vcf-sort > sample.sort.vcf`; `bgzip -f sample.sort.vcf`; `tabix -f -p vcf sample.sort.vcf.gz`

- `vcf-isec -f -a -c sample.vcf.gz control1.vcf.gz control2.vcf.gz (...) > sample.isec1.vcf`

(commands for mouse samples also include files with variants obtained from sequencing mice of the same background)

- If scalpel was also used: `bedtools intersect -header -a sample.vcf -b sample.somatic.filt.vcf
  > sample.merged.vcf`; `cat sample.vcf | grep "#" -v | grep "INDEL" -v >> sample.merged.vcf`

(these commands retain all post-filtering and intersection SNVs identified by samtools mpileup and those INDELs identified by both variant callers).

- If there are replicates for a sample (e.g. post-propagation polymerase strains had two colonies from the same line sequenced): `vcf-isec -f -a sample1.isec1.vcf.gz sample2.isec1.vcf.gz
  > sample.merge.vcf`

**Step 7:** All sample files were merged from one experiment into one vcf file

- For `x` in sort.*.vcf.gz, do list=$list'echo "$x"'; list=$list’ ’; done

- `vcf-merge $list 2>/dev/null > experiment_merge.vcf`
**Step 8 (optional):** Variants present in all samples or variants present in more than one sample were removed

- perl remove_shared_variants.pl -i Experiment_merge.vcf > merge_unique.vcf

**Step 9:** Specific outputs were produced depending on the experiment

- Output number of SNVs and INDELs: for x in *.merge.vcf; do n=$(echo $x | sed 's/.merge.vcf//g'); m=$(cat $x | grep "#" -v | wc -l); o=$(cat $x | grep "#" -v | grep "INDEL" -v | wc -l); p=$(cat $x | grep "#" -v | grep "INDEL" | wc -l); printf "$n	$m	$o	$p\n" >> Mutations.results.txt; done

- Calculate mean number of mutations and the standard deviation of numbers in the Mutations.results.txt file ($2 for total number of mutations, $3 for SNVs and $4 for INDELs): for x in $sample_names; do t=`cat Mutations.results.txt | grep $x | awk '{sum+=+$2} END { print (sum/NR)}'`; s=`cat Mutations.results.txt | grep $x | awk '{sum+=+$2; array[NR]=$2} END {for(x=1;x<=NR;x++) {sumsq+=((array[x]-(sum/NR))^2);}} print sqrt(sumsq/NR)}``; printf "$x	$t	$s\n"; printf "$x	$t	$s\n" >> Mutations.results.txt; done

- Output mutation patterns (e.g. transitions versus transversions): perl vcf_stats_table_all.pl experiment_merge.vcf > stat_table.txt

- Output a list of mutation affecting genes: perl vcf_to_gene_list.pl -i experiment_merge.vcf > genelist.txt
References


[13] The White House: Office of the Press Secretary. Remarks Made by the President, Prime Minister Tony Blair of England (via satellite), Dr. Francis Collins, Director of the National Human Genome Research Institute, and Dr. Craig Venter, President and Chief Scientific Officer, Celera Genomics Corporation, on the Completion of the First Survey of the Entire Human Genome Project, 2000.


References


References


[569] A. Turner, J. Beales, S. Faure, R. P. Dunford, and D. A. Laurie. The pseudo-
response regulator Ppd-H1 provides adaptation to photoperiod in barley. *Science*,


of mammalian DNA repair and the DNA damage checkpoints. *Annu Rev Biochem*,

[572] A. A. Morley and D. R. Turner. The contribution of exogenous and endogenous muta-

[573] R. De Bont and N. van Larebeke. Endogenous DNA damage in humans: a review of


[577] E. Viguera, D. Canceill, and S. D. Ehrlich. Replication slippage involves DNA poly-


[579] H. B. Creighton and B. McClintock. A correlation of cytological and genetical crossing-


References


X. S. Puente, M. Pinyol, V. Quesada, L. Conde, G. R. Ordonez, N. Villamor, G. Es-
caramis, P. Jares, S. Bea, M. Gonzalez-Diaz, L. Bassaganyas, T. Baumann, M. Juan,
M. Lopez-Guerra, D. Colomer, J. M. Tubio, C. Lopez, A. Navarro, C. Tornador,
M. Aymerich, M. Rozman, J. M. Hernandez, D. A. Puente, J. M. Freije, G. Velasco,
A. Gutierrez-Fernandez, D. Costa, A. Carrio, S. Guijarro, A. Enjuanes, L. Hernane-
dez, J. Yague, P. Nicolas, C. M. Romeo-Casabona, H. Himmelbauer, E. Castillo, J. C.
Dohm, S. de Sanjose, M. A. Piris, E. de Alava, J. San Miguel, R. Royo, J. L. Gelpi,
D. Torrents, M. Orozco, D. G. Pisano, A. Valencia, R. Guigo, M. Bayes, S. Heath,
P. J. Campbell, I. Gut, A. Lopez-Guillermo, X. Estivill, E. Montserrat, C. Lopez-Otin,
and E. Campo. Whole-genome sequencing identifies recurrent mutations in chronic

T. J. Pugh, S. D. Weeraratne, T. C. Archer, D. A. Pomeranz Krummel, D. Auclair,
J. Bochicchio, M. O. Carneiro, S. L. Carter, K. Cibulskis, R. L. Erlich, H. Greulich,
M. S. Lawrence, N. J. Lennon, A. McKenna, J. Meldrim, A. H. Ramos, M. G. Ross,
C. Russ, E. Shefler, A. Sivachenko, B. Sogoloff, P. Stojanov, P. Tamayo, J. P. Mesirov,
V. Amani, N. Teider, S. Sengupta, J. P. Francois, P. A. Northcott, M. D. Taylor, F. Yu,
S. M. Pfister, T. M. Roberts, M. Meyerson, S. L. Pomeroy, and Y. J. Cho. Medul-
loblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature*,

T. J. Pugh, O. Morozova, E. F. Attiyeh, S. Asgharzadeh, J. S. Wei, D. Auclair, S. L.
Carter, K. Cibulskis, M. Hanna, A. Kiezun, J. Kim, M. S. Lawrence, L. Lichenstein,
A. McKenna, C. S. Pedamallu, A. H. Ramos, E. Shefler, A. Sivachenko, C. Sougnez,
C. Stewart, A. Ally, I. Birol, R. Chiu, R. D. Corbett, M. Hirst, S. D. Jackman, B. Kamoh,
A. H. Khodabakshi, M. Krzywinski, A. Lo, R. A. Moore, K. L. Mungall, J. Qian,
A. Tam, N. Thiessen, Y. Zhao, K. A. Cole, M. Diamond, S. J. Diskin, Y. P. Mosse,
A. C. Wood, L. Ji, R. Spoto, T. Badgett, W. B. London, Y. Moyer, J. M. Gastier-Foster,
M. A. Smith, J. M. Guidry Auvil, D. S. Gerhard, M. D. Hogarty, S. J. Jones, E. S.
Lander, S. B. Gabriel, G. Getz, R. C. Seeger, J. Khan, M. A. Marra, M. Meyerson, and
84, 2013.

V. Quesada, L. Conde, N. Villamor, G. R. Ordonez, P. Jares, L. Bassaganyas, A. J. Rams-
say, S. Bea, M. Pinyol, A. Martinez-Trillos, M. Lopez-Guerra, D. Colomer, A. Navarro,


References


References


References


References


Appendix A

List of Abbreviations

APC/C  Anaphase promoting complex/Cyclosome
APOBEC  apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
BAM  Binary sequence alignment and mapping
BER  Base excision repair
BIR  Break-induced replication
bp  Base pairs
CDK  Cyclin-dependent kinase
ChIP  chromatin immunoprecipitation
CNV  Copy number variation
CPD  Cyclobutane pyrimidine dimer
CPT  Camptothecin D-loop Displacement loop
DDC  Duplication–degeneration–complementation model
DDT  DNA damage tolerance
DNA  Deoxyribonucleic acid
dNTP  Deoxynucleoside triphosphate
DSB  Double strand break
DSBR | Classical double-strand break repair  
--- | ---  
dsDNA | Double-stranded DNA  
EtBR | Ethidium bromide  
EtOH | Ethanol  
f.c. | Final concentration  
FISH | Fluorescence in situ hybridization  
gDNA | Genomic DNA  
GG-NER | Global genome-wide nucleotide excision repair  
GRCh37 | Genome Reference Consortium human genome (build 37)  
HR | Homologous recombination  
HU | Hydroxurea  
IARC | International Agency for Research on Cancer  
INDEL | Small insertion/deletion  
IR | Ionising radiation  
kb | Kilobase pairs  
LOF | loss-of-function  
LP-BER | Long patch base excision repair  
LTR | Long terminal repeats  
MMEJ | Microhomology-mediated end joining  
MMR | DNA mismatch repair  
MMS | Methyl methanesulfonate  
NER | Nucleotide excision repair  
NGS | Next-generation sequencing
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHEJ</td>
<td>Non-homologous end joining</td>
</tr>
<tr>
<td>NIR</td>
<td>Non-ionising radiation</td>
</tr>
<tr>
<td>NMD</td>
<td>Nonsense-mediated decay</td>
</tr>
<tr>
<td>NMF</td>
<td>Nonnegative matrix factorization</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PEG</td>
<td>Polyethylene Glycol Pol Polymerase</td>
</tr>
<tr>
<td>Phleo</td>
<td>Phleomycin</td>
</tr>
<tr>
<td>ORF</td>
<td>Open Reading Frame</td>
</tr>
<tr>
<td>RFC</td>
<td>Replication factor C</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>RNS</td>
<td>Reactive nitrogen species</td>
</tr>
<tr>
<td>ROS</td>
<td>Reactive oxygen species</td>
</tr>
<tr>
<td>rpm</td>
<td>Revolutions per minute</td>
</tr>
<tr>
<td>RT</td>
<td>Room Temperature</td>
</tr>
<tr>
<td>SAC</td>
<td>Spindle assembly checkpoint</td>
</tr>
<tr>
<td>SDSA</td>
<td>Synthesis-dependent strand annealing</td>
</tr>
<tr>
<td>SGA</td>
<td>Synthetic Gene Array</td>
</tr>
<tr>
<td>SGD</td>
<td>Saccharomyces Genome Database</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>SNV</td>
<td>Single nucleotide variant</td>
</tr>
<tr>
<td>SSA</td>
<td>single-strand annealing</td>
</tr>
<tr>
<td>ssDNA</td>
<td>Single-stranded DNA</td>
</tr>
<tr>
<td>SV</td>
<td>Structural Variant</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>TC-NER</td>
<td>Transcription-coupled nucleotide excision repair</td>
</tr>
<tr>
<td>TCGA</td>
<td>The Cancer Genome Atlas</td>
</tr>
<tr>
<td>TE</td>
<td>Transposable Element</td>
</tr>
<tr>
<td>TLS</td>
<td>Translesion synthesis</td>
</tr>
<tr>
<td>Tm</td>
<td>Melting temperature (e.g. for oligonucleotides)</td>
</tr>
<tr>
<td>Tris</td>
<td>Tris(hydroxymethyl)aminomethane</td>
</tr>
<tr>
<td>UPD</td>
<td>Uniparental disomy</td>
</tr>
<tr>
<td>UV</td>
<td>Ultraviolet</td>
</tr>
<tr>
<td>UV-A</td>
<td>Ultraviolet A</td>
</tr>
<tr>
<td>UV-B</td>
<td>Ultraviolet B</td>
</tr>
<tr>
<td>UV-C</td>
<td>Ultraviolet C</td>
</tr>
<tr>
<td>VEP</td>
<td>Variant Effect Predictor</td>
</tr>
<tr>
<td>WES</td>
<td>Whole-exome sequencing</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole-genome sequencing</td>
</tr>
<tr>
<td>YNB</td>
<td>Yeast Nitrogen Base</td>
</tr>
<tr>
<td>YPD</td>
<td>Yeast Extract - Peptone - Dextrose</td>
</tr>
</tbody>
</table>
Appendix B

Supplementary Tables, Electronic Files and Articles Published

B.1 Supplementary figures, tables and notes

B.1.1 Software tools and parameters used

B.1.1.1 Software tools and parameters used for simulated genomes and capillary sequencing analysis

<table>
<thead>
<tr>
<th>Step</th>
<th>Software/Tool</th>
<th>Command</th>
<th>Command</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABI sequence alignment</td>
<td>BWA[908]</td>
<td>bwasw</td>
<td></td>
</tr>
<tr>
<td>Variant Calling of ABI files</td>
<td>SAMTools [903]</td>
<td>mpileup</td>
<td>-u</td>
</tr>
<tr>
<td>Variant Calling of ABI files</td>
<td>BCFtools[903]</td>
<td>view</td>
<td>-c -v</td>
</tr>
<tr>
<td>Filtering of ABI vcf files</td>
<td>VCFtools [905]</td>
<td>vcf-annotate</td>
<td>-f +/d=2/D=5</td>
</tr>
<tr>
<td>Generate INDEL Set</td>
<td>pIRS[840]</td>
<td>pirs diploid</td>
<td>-a 3 -v 0 -d 0.000075</td>
</tr>
<tr>
<td>Generate Control Set</td>
<td>pIRS[840]</td>
<td>pirs diploid</td>
<td>-s 0.0001</td>
</tr>
<tr>
<td>Generate Mutated Set</td>
<td>pIRS[840]</td>
<td>pirs diploid</td>
<td>-s 0.000025</td>
</tr>
<tr>
<td>Simulate sequencing</td>
<td>pIRS[840]</td>
<td>pirs simulate</td>
<td>-x 40(20,30,50) -m 450</td>
</tr>
<tr>
<td>Alignment</td>
<td>BWA[908]</td>
<td>v0.6.2</td>
<td>-q 15</td>
</tr>
<tr>
<td>Variant Calling</td>
<td>SAMTools [903]</td>
<td>mpileup</td>
<td>-g-tDP,DV-C50-PM3-F0.2-d10000</td>
</tr>
<tr>
<td>Variant Calling</td>
<td>BCFtools[903]</td>
<td>call</td>
<td>-vm -f GQ</td>
</tr>
<tr>
<td>Intersecting Variants</td>
<td>BEDtools[906]</td>
<td>intersect</td>
<td>-a -b -v</td>
</tr>
<tr>
<td>Visualising variants</td>
<td>IGV[909, 910]</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### B.1.1.2 Software tools and parameters used for sequencing analysis of *S. cerevisiae*

<table>
<thead>
<tr>
<th>Step</th>
<th>Software/Tool</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read alignment</td>
<td>BWA[901]</td>
<td><code>bwa aln -l 32 -t 6 -f</code></td>
</tr>
<tr>
<td>Read alignment</td>
<td>BWA[901]</td>
<td><code>bwa sampe -P -a 742</code></td>
</tr>
<tr>
<td>Mark PCR Duplicates</td>
<td>Picard MarkDuplicates[902]</td>
<td>-</td>
</tr>
<tr>
<td>Variant Calling</td>
<td>SAMTools mpileup[903]</td>
<td><code>-g-t DP,DV-C50-pm3-F0.2-d10000</code></td>
</tr>
<tr>
<td>Variant Calling</td>
<td>BCFtools call[903]</td>
<td><code>-vm -f GQ</code></td>
</tr>
<tr>
<td>Variant Annotation</td>
<td>Variant Effect Predictor[904]</td>
<td><code>-species saccharomyces_cerevisiae</code></td>
</tr>
<tr>
<td>Normalising INDELs</td>
<td>BCFtools norm[903]</td>
<td>-</td>
</tr>
<tr>
<td>Variant Filtering</td>
<td>VCFtools vcf-annotate[905]</td>
<td><code>-H -f +/q=30/Q=50/SnpGap=7</code></td>
</tr>
<tr>
<td>Variant Filtering</td>
<td>VCFtools vcf-annotate[905]</td>
<td>custom written filters (see B.1.4)</td>
</tr>
<tr>
<td>Subsetting samples</td>
<td>VCFtools vcf-subset[905]</td>
<td><code>-e</code></td>
</tr>
<tr>
<td>Sorting vcf files</td>
<td>VCFtools vcf-sort[905]</td>
<td>-</td>
</tr>
<tr>
<td>Intersecting vcf files</td>
<td>VCFtools vcf-iseq[905]</td>
<td><code>-f -a -c</code></td>
</tr>
<tr>
<td>Merging vcf files</td>
<td>VCFtools vcf-merge[905]</td>
<td>-</td>
</tr>
</tbody>
</table>
B.1.2 Strains used in mutation accumulation (MA) experiments

B.1.2.1 Manual propagation of strains heterozygous diploid for candidate polymerase mutations

<table>
<thead>
<tr>
<th>Yeast strain</th>
<th>polymerase mutation</th>
<th>ploidy &amp; genotype</th>
<th>parallel lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>YMH9/YMH68</td>
<td>wild-type</td>
<td>diploid</td>
<td>72</td>
</tr>
<tr>
<td>YMH29</td>
<td>pol2-4</td>
<td>heterozygous diploid</td>
<td>54</td>
</tr>
<tr>
<td>YMH27</td>
<td>pol2-A480V</td>
<td>heterozygous diploid</td>
<td>18</td>
</tr>
<tr>
<td>YMH21</td>
<td>pol2-D290V</td>
<td>heterozygous diploid</td>
<td>18</td>
</tr>
<tr>
<td>YMH13</td>
<td>pol2-L439V</td>
<td>heterozygous diploid</td>
<td>18</td>
</tr>
<tr>
<td>YMH23</td>
<td>pol2-M459K</td>
<td>heterozygous diploid</td>
<td>18</td>
</tr>
<tr>
<td>YMH19</td>
<td>pol2-P301R</td>
<td>heterozygous diploid</td>
<td>18</td>
</tr>
<tr>
<td>YMH25</td>
<td>pol2-Q468R</td>
<td>heterozygous diploid</td>
<td>18</td>
</tr>
<tr>
<td>YMH17</td>
<td>pol2-S312F</td>
<td>heterozygous diploid</td>
<td>18</td>
</tr>
<tr>
<td>YMH15</td>
<td>pol2-V426L</td>
<td>heterozygous diploid</td>
<td>18</td>
</tr>
<tr>
<td>YMH71</td>
<td>pol3-01</td>
<td>heterozygous diploid</td>
<td>18</td>
</tr>
<tr>
<td>YMH69</td>
<td>pol3-P332L</td>
<td>heterozygous diploid</td>
<td>18</td>
</tr>
<tr>
<td>YMH72</td>
<td>pol3-R316C</td>
<td>heterozygous diploid</td>
<td>18</td>
</tr>
<tr>
<td>YMH70</td>
<td>pol3-S375R</td>
<td>heterozygous diploid</td>
<td>18</td>
</tr>
</tbody>
</table>

B.1.2.2 Automated propagation of strains haploid and heterozygous diploid for candidate polymerase mutations

Table of strains included in the population bottleneck mutation accumulation experiment. Both heterozygous diploid (Het.) mutant strains and haploid mutant strains were propagated.
B.1.3 6-Thioguanine suppressor screen of haploid mouse cells

Bait locations for the exon-capture experiment (6Thioguanine haploid mouse cell suppressor screen)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chr</th>
<th>Location</th>
<th>No of exons</th>
<th>Mean coverage (fold)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dnmt1</td>
<td>9</td>
<td>20907206-20959888</td>
<td>39</td>
<td>604.7</td>
</tr>
<tr>
<td>Hprt</td>
<td>X</td>
<td>52988137-53021659</td>
<td>9</td>
<td>317.2</td>
</tr>
<tr>
<td>Mlh1</td>
<td>9</td>
<td>111228228-111271791</td>
<td>19</td>
<td>527.5</td>
</tr>
<tr>
<td>Mlh3</td>
<td>12</td>
<td>85234529-85270591</td>
<td>12</td>
<td>528.6</td>
</tr>
<tr>
<td>Msh2</td>
<td>17</td>
<td>87672330-87723713</td>
<td>16</td>
<td>566.9</td>
</tr>
<tr>
<td>Msh3</td>
<td>13</td>
<td>92211872-92355003</td>
<td>24</td>
<td>497.3</td>
</tr>
<tr>
<td>Msh4</td>
<td>3</td>
<td>153857149-153906138</td>
<td>20</td>
<td>511.5</td>
</tr>
<tr>
<td>Msh5</td>
<td>17</td>
<td>35028605-35046745</td>
<td>24</td>
<td>560.8</td>
</tr>
<tr>
<td>Msh6</td>
<td>17</td>
<td>87975050-87990883</td>
<td>10</td>
<td>572</td>
</tr>
<tr>
<td>Pms1</td>
<td>1</td>
<td>53189187-53297018</td>
<td>13</td>
<td>488.4</td>
</tr>
<tr>
<td>Pms2</td>
<td>5</td>
<td>143909964-143933968</td>
<td>15</td>
<td>541.4</td>
</tr>
<tr>
<td>Setd2</td>
<td>9</td>
<td>110532597-110618633</td>
<td>21</td>
<td>577.7</td>
</tr>
</tbody>
</table>

B.1.4 Custom filters for DNA sequencing Filters

The custom quality filters on any variant with a sequencing depth of less than 10 reads and a genotype quality if less than 25.
B.2 Electronic files of supplementary information

The remaining supplementary information has been placed in the Cambridge research repository Apollo as these are large files that do not need to be printed. Here included is the name under which they can be found and a short description of the data they contain. The DOI links under which they can be viewed are https://doi.org/10.17863/CAM.7296 (the mouse synthetic lethality screen) and https://doi.org/10.17863/CAM.7299 (the polymerase mutation project). Supplementary files for the Puddu, et al. (2015) publication [801] can be found with the journal article online.

B.2.1 Supplementary files for the mouse synthetic lethality screens

The sequencing data generated in the course of this project is available for download in the European Nucleotide Archive (PRJEB4302, PRJEB5755, PRJEB12638).

B.2.1.1 6TG_mouse_Sup1.xlsx

This file includes two tables. Table 1 includes all homozygous mutations identified through whole-exome sequencing of the first 7 suppressor clones we submitted for sequencing. Table 2 includes all mutations of the clones in which no mutation in Hprt could be identified.

B.2.1.2 6TG_mouse_Sup2.xlsx

This file includes four tables. Table 1 includes all homozygous mutations affecting Dnmt1, Hprt, Mlh1, Msh2, Msh6 and Pms2 genes identified on the targeted exon-capture experiment performed on 189 clones. Table 2 includes all heterozygous mutations. Table 3 includes PROVEAN and SIFT predictions for identified mutations. Table 4 summarizes the potential causative mutation for all suppressor screens with references when identified mutations were previously described.

B.2.1.3 6TG_mouse_Sup3.xlsx

This file includes three tables. Table 1 includes all homozygous mutations identified in 66 suppressor clones (23 orphan clones plus 43 clones with identified mutations). Table 2 includes all heterozygous mutations identified in the same clones. Table 3 contains all mutations identified in the 23 orphan clones.
B.2.1.4  6TG_mouse_Sup4.xlsx

This file includes three tables. Table 1 describes the bait regions used in the exon capture experiment. Table 2 includes the average coverage of targeted sequences in the exon-capture sequencing experiment. Table 3 includes DNA sequencing coverage for the whole-exome sequencing experiments.

B.2.2  Supplementary files for the mouse synthetic lethality screens

B.2.2.1  MA_SampleNames.pdf

This file lists all the samples used in manual propagation experiments and their corresponding sample name in the sequencing data files.

B.2.2.2  S1-3.experiment_merge.vcf

This file contains all acquired mutations across Set 1-3 (all pol2 mutants and pol3-S483N plus control samples) of the manual mutation accumulation experiments.

B.2.2.3  S4.experiment_merge.vcf

This file contains all acquired mutations across Set 4 (all remaining pol3 strains plus control samples) of the manual mutation accumulation experiments.

B.2.2.4  S5.experiment_merge.vcf

This file contains all acquired mutations across Set 5 (used for the figures in Chapter 4.3 and 4.4) of the manual mutation accumulation experiments.

B.3  Articles published during my PhD

During the course of this work, I was part of several publications, two of which are published or accepted for publication, one of which is in review and three of which are in preparation. In this appendix, published or accepted publications are listed and a short summary of the work as well as a description of my contribution is included. The articles can be found at the end of the dissertation.

In this work synthetic viability screening was used in budding yeast to identify mutations that can suppress the DNA sensitivity phenotype that results from the loss of Sae2, a protein involved in DNA repair. These suppressor mutations all affected specific residues in the Mre11 protein which is also involved in DNA repair. Further analysis revealed that the mutated Mre11 protein has a decreased affinity to ssDNA suggesting that in wild type cells Sae2 is required to remove Mre11 from the damaged DNA site in the course of the repair. My main contribution to this work is the analysis of whole genome sequencing data of 48 suppressor colonies under the supervision of Thomas Keane, leading to the identification of the \textit{mre11-H37R} and \textit{mre11-H37Y} mutations.

Genome-wide genetic screening with chemically-mutagenized haploid embryonic stem cells Josep Forment, Mareike Herzog, Julia Coates, Tomasz Konopka, Bianca Gapp, Sebastian Nijman, David Adams, Thomas Keane and Stephen Jackson. Nature Chemical Biology [Accepted 24th Aug 16]

This is a proof-of-principle work showing that synthetic viability screening in haploid, mouse embryonic stem cells is feasible. All known genes whose inactivation leads to suppression were identified in this work. This work demonstrates that causative mutations can be identified, that synthetic viability screens can map essential domains of a protein and that causative mutations can be identified even if mutagenesis generated more “passanger” mutations to sift through. This work is a demonstration of the feasibility of classical genetic screenings in mammalian cells and provides a new, powerful tool to explore mammalian genetic interactions. My contribution to this work is the analysis of all sequencing data of DNA from resistant clones and the identification of all critical mutations identified in this work.

Chromatin determinants impart camptothecin hypersensitivity in the absence of the Tof1/Csm3 replication pausing complex Fabio Puddu, Mareike Herzog, Nicola Geisler, Vincenzo Costanzo, Steve Jackson. Nucleic Acids Research [Submitted]

In budding yeast the absence of the Tof1/Csm3 complex causes hypersensitivity to camptothecin. Using a synthetic viability approach, we have identified that disruption of Sir-dependent heterochromatin by inactivation of histone H4-K16 deacetylation can suppress this...
sensitivity in *tof1Δ* and wild-type cells. My main contribution to this work is the analysis of all suppressor colonies whole genome sequencing and identification of inactivating mutations in the genes *SIR3* and *SIR4*, as well as analysis of ChIP-Seq data together with Fabio Puddu.
Synthetic viability genomic screening defines Sae2 function in DNA repair


Abstract

DNA double-strand break (DSB) repair by homologous recombination (HR) requires 3’ single-stranded DNA (ssDNA) generation by 5’ DNA-end resection. During meiosis, yeast Sae2 cooperates with the nuclease Mre11 to remove covalently bound Spo11 from DSB termini, allowing resection and HR to ensue. Mitotic roles of Sae2 and Mre11 nuclease have remained enigmatic, however, since cells lacking these display modest resection defects but marked DNA damage hypersensitivities. By combining classic genetic suppressor screening with high-throughput DNA sequencing, we identify Mre11 mutations that strongly suppress DNA damage sensitivities of sae2Δ cells. By assessing the impacts of these mutations at the cellular, biochemical and structural levels, we propose that, in addition to promoting resection, a crucial role for Sae2 and Mre11 nuclease activity in mitotic DSB repair is to facilitate the removal of Mre11 from ssDNA associated with DSB ends. Thus, without Sae2 or Mre11 nuclease activity, Mre11 bound to partly processed DSBs impairs strand invasion and HR.

Keywords: Mre11; Sae2; suppressor screening; synthetic viability; whole-genome sequencing

Subject Categories: DNA Replication, Repair & Recombination

DOI: 10.15252/embj.201590973 | Received 8 January 2015 | Revised 16 March 2015 | Accepted 2 April 2015 | Published online 21 April 2015


Introduction

The DSB is the most cytotoxic form of DNA damage, with ineffective DSB repair leading to mutations, chromosomal rearrangements and genome instability that can yield cancer, neurodegenerative disease, immunodeficiency and/or infertility (Jackson & Bartek, 2009). DSBs arise from ionising radiation and radiomimetic drugs and are generated when replication forks encounter single-stranded DNA breaks or other DNA lesions, including DNA alkylation adducts and sites of abortive topoisomerase activity. DSBs are also physiological intermediates in meiotic recombination, being introduced during meiotic prophase I by the topoisomerase II-type enzyme Spo11 that becomes covalently linked to the 5’ end of each side of the DSB (Keeney et al., 1997). The two main DSB repair pathways are non-homologous end-joining (NHEJ) and homologous recombination (Lisby et al., 2004; Symington & Gautier, 2011). In NHEJ, DNA ends need little or no processing before being ligated (Daley et al., 2005). By contrast, HR requires DNA-end resection, a process involving degradation of the 5’ ends of the break, yielding 3’ single-stranded DNA (ssDNA) tails that mediate HR via pairing with and invading the sister chromatid, which provides the repair template.

Reflecting the above requirements, cells defective in resection components display HR defects and hypersensitivity to various DNA-damaging agents. This is well illustrated by Saccharomyces cerevisiae cells harbouring defects in the Mre11–Rad50–Xrs2 (MRX) complex, which binds and juxtaposes the two ends of a DSB (Williams et al., 2008) and, through Mre11 catalytic functions, provides nuclease activities involved in DSB processing (Furuse et al., 1998; Williams et al., 2008; Stracker & Petrini, 2011). Once a clean, partially resected 5’ end has been generated, the enzymes Exo1 and Sgs1/Dna2 are then thought to act, generating extensive ssDNA regions needed for effective HR (Mimitou & Symington, 2008; Zhu et al., 2008). Notably, while Mre11 nuclease activity is essential in meiosis to remove Spo11 and promote 5’ end resection, in mitotic cells, resection is only somewhat delayed in the absence of Mre11 and almost unaffected by mre11-nd (nuclease-dead) mutations (Ivanov et al., 1994; Moreau et al., 1999), indicating the existence of MRX-nuclease-independent routes for ssDNA generation.

Another protein linked to resection is S. cerevisiae Sae2, the functional homolog of human CtIP (Sartori et al., 2007; You et al., 2009). Despite lacking obvious catalytic domains, Sae2 and CtIP have been reported to displayendonuclease activity in vitro (Lengsfeld et al., 2007; Makharashvili et al., 2014; Wang et al., 2014), and their functions are tightly regulated by cell cycle- and DNA damage-dependent phosphorylations (Baroni et al., 2004; Huertas et al., 2008; Huertas & Jackson, 2009; Barton et al., 2014). In many ways, Sae2 appears to...
function together with MRX in DSB repair. For instance, mre11-nd as well as mre11S and rad50S hypomorphic alleles phenocopy Sae2 deletion (sae2Δ) in meiosis, yielding unprocessed Spo11-DNA complexes (Keeney & Kleckner, 1995; Nairz & Klein, 1997; Prinz et al., 1997). Furthermore, recent findings have indicated that Sae2 stimulates Mre11 endonuclease activity to promote resection, particularly at protein-bound DSB ends (Cannavo & Cejka, 2014). Also, both sae2Δ and mre11-nd mutations cause hypersensitivity towards the anti-cancer drug camptothecin (Deng et al., 2005), which yields DSBs that are repaired by HR. Nevertheless, key differences between MRX and Sae2 exist, since sae2Δ leads to persistence of MRX at DNA damage sites (Lisby et al., 2004) and hyperactivation of the MRX-associated Tel1 protein kinase (Usui et al., 2001), the homolog of human ATM, while MRX inactivation abrogates Tel1 function (Fukunaga et al., 2011). These findings, together with sae2Δ and mre11-nd cells displaying only mild resection defects (Clerici et al., 2005), highlight how Sae2 functions in HR cannot be readily explained by it simply cooperating with MRX to enhance resection.

As reported below, by combining classic genetic screening for suppressor mutants with whole-genome sequencing to determine their genotype, we are led to a model that resolves apparent para-

Supervised authors (SVGS; Fig 1A). To do this, we took cultures of a yeast strain (bearing a full deletion of the SAE2 gene) and subjected them to next-generation Illumina sequencing. We then used bioinformatics tools (see Materials and Methods) to identify mutations altering DSB chains.

Results

SVGS identifies Mre11 mutations as sae2Δ suppressors

To gain insights into why yeast cells lacking Sae2 are hypersensitive to DNA-damaging agents, we performed synthetic viability genomic screening (SVGS; Fig 1A). To do this, we took cultures of a sae2Δ yeast strain (bearing a full deletion of the SAE2 locus) and plated them on YPD plates supplemented with camptothecin, which stabilises DNA topoisomerase I cleavage complexes and yields replication-dependent DSBs that are repaired by Sae2-dependent HR (Deng et al., 2005) (Fig 1A). Thus, we isolated 48 mutants surviving camptothecin treatment that spontaneously arose in the population analysed. In addition to verifying that all indeed contained the SAE2 gene deletion yet were camptothecin resistant, subsequent analyses revealed that 10 clones were also largely or fully suppressed for sae2Δ hypersensitivity to the DNA-alkylating agent methyl methanesulphonate (MMS), the replication inhibitor hydroxyurea (HU), the DSB-generating agent phleomycin and ultraviolet light (Supplementary Fig S1).

To identify mutations causing these suppression phenotypes, genomic DNA from the 48 clones was isolated and analysed by next-generation Illumina sequencing. We then used bioinformatics tools (see Materials and Methods) to identify mutations altering open reading frames within the reference S. cerevisiae genome (Fig 1A). This revealed that 24 clones displaying camptothecin resistance but retaining sae2Δ hypersensitivity towards other DNA-damaging agents possessed TOP1 mutations (Fig 1B and C), thereby providing proof-of-principle for the SVGS methodology (TOP1 is a non-essential gene that encodes DNA topoisomerase I, the camptothecin target). Strikingly, of the remaining clones, 10 contained one or other of two different mutations in a single MRE11 codon, resulting in amino acid residue His37 being replaced by either Arg or Tyr (mre11-H37R and mre11-H37Y, respectively; Fig 1B and C and Supplementary Fig S1; note that TOP1 and MRE11 mutations are mutually exclusive). While some remaining clones contained additional potential suppressor mutations worthy of further examination, these were only resistant to camptothecin. Because of their broader phenotypes and undefined mechanism of action, we focused on characterising the MRE11 sae2Δ suppressor (mre11SUPsae2Δ) alleles.

mre11SUPsae2Δ alleles suppress many sae2Δ phenotypes

Mre11 His37 lies within a functionally undefined but structurally evolutionarily conserved α-helical region, and the residue is well conserved among quite divergent fungal species (Fig 2A). As anticipated from previous studies, deleting MRE11 did not suppress the DNA damage hypersensitivities of sae2Δ cells, revealing that mre11-H37R and mre11-H37Y were not behaving as null mutations (unpublished observation). In line with this, the mre11-H37R and mre11-H37Y alleles did not destabilise Mre11, producing proteins that were expressed at equivalent levels to the wild-type protein (Fig 2B). Nevertheless, expression of wild-type Mre11 reconstituted the mre11H37RΔ sae2Δ strains to camptothecin, and to a lesser extent to MMS (Fig 2C), indicating that mre11-H37R and mre11-H37Y were fully or partially recessive for the camptothecin and MMS resistance phenotypes, respectively. Furthermore, this established that expression of wild-type Mre11 is toxic to sae2Δmre11H37RΔ cells upon camptothecin treatment. Importantly, independent introduction of mre11-H37RΔ and mre11-H37YΔ alleles in a sae2Δ strain confirmed that each conferred suppression of sae2Δ hypersensitivity to various DNA-damaging agents (Fig 2D). The mre11-H37RΔ and mre11-H37YΔ alleles also suppressed camptothecin hypersensitivity caused by mutations in Sae2 that prevent its Mec1/Tel1-dependent (sae2-MT) or CDK-dependent (sae2-S267A) phosphorylation (Baroni et al., 2004; Huertas et al., 2008) (Fig 2E and F). By contrast, no suppression of sae2Δ camptothecin hypersensitivity was observed by mutating His37 to Ala (mre11-H37AΔ; Fig 2G), suggesting that the effects of the mre11H37RΔΔ alleles were not mediated by the abrogation of a specific function of His37 but more likely reflected functional alteration through introducing bulky amino acid side chains.

mre11SUPsae2Δ alleles do not suppress all sae2Δ phenotypes

In the absence of Sae2, cells display heightened DNA damage signal-

1510 The EMBO Journal Vol 34 | No 11 | 2015 © 2015 The Authors

Published online: April 21, 2015

The EMBO Journal

Sae2 function in DNA repair Fabio Puddu et al
Notably, however, mre11-H37R rescued the hypersensitivity of sae2Δ cells to etoposide, which produces DSBs bearing 5' DNA ends bound to Top2 (Supplementary Fig S2B; deletion of ERG6 was used to increase permeability of the plasma membrane to etoposide), suggesting that significant differences must exist between the repair of meiotic and etoposide-induced DSBs.

Next, we examined the effects of mre11SUPsae2Δ alleles on Sae2-dependent DSB repair by single-strand annealing (SSA), using a system wherein a chromosomal locus contains an HO endonuclease cleavage site flanked by two direct sequence repeats. In this system, HO induction produces a DSB that is then resected until two complementary sequences become exposed and anneal, resulting in repair by a process that deletes the region between the repeats (Fishman-Lobell et al., 1992; Vaze et al., 2002; Fig 3C). Despite displaying only mild resection defects (Clerici et al., 2006), we observed that sae2Δ cells were defective in SSA-mediated DSB repair and did not resume cell cycle progression after HO induction as fast as wild-type cells, in agreement with published work (Clerici et al., 2005). Notably, mre11-H37R did not alleviate these sae2Δ phenotypes (Fig 3D and E).

Finally, we examined the effect of the mre11-H37R mutation on telomere-associated functions of the MRX complex and Sae2. It has been established that simultaneous deletion of SGS1 and SAE2 results in synthetic lethality/sickness, possibly due to excessive telomere shortening (Mimitou & Symington, 2008; Hardy et al., 2014). To test whether mre11-H37R can alleviate this phenotype, we crossed a sae2Δmre11-H37R strain with a sgs1Δ strain. As shown in Supplementary Fig S2C, we were unable to recover either sgs1Δsae2Δ nor sgs1Δsae2Δmre11-H37R cells, implying that mre11-H37R cannot

**Figure 1.** SVGS identifies mutations suppressing sae2Δ DNA damage hypersensitivity.

A Outline of the screening approach that was used to identify suppressors of sae2Δ camptothecin (CPT) hypersensitivity.

B Validation of the suppression phenotypes; a subset (sup25–sup30) of the suppressors recovered from the screening is shown along with mutations identified in each clone.

C Summary of the results of the synthetic viability genomic screening (SVGS) for sae2Δ camptothecin (CPT) hypersensitivity. The ORF and the type of mutation are reported together with the number of times each ORF was found mutated and the number of clones in which each ORF was putatively driving the resistance.
sae2Δ DNA damage hypersensitivities but not sae2Δ meiotic phenotypes requiring Mre11-mediated Spo11 removal from recombination intermediates, nor mitotic SSA functions that have been attributed to Sae2-mediated DNA-end bridging (Clerici et al., 2005). Subsequent analyses revealed that suppression did not arise largely through channeling of DSBs towards NHEJ because the key NHEJ factor Yku70 was not required for mre11-H37R or mre11-H37Y to suppress the camptothecin sensitivity of a sae2Δ strain (Fig 3F). In addition, this analysis revealed that the previously reported suppression of sae2Δ-mediated DNA damage hypersensitivity by Ku loss (Mimitou & Symington, 2010; Foster et al., 2011) was considerably less effective than that caused by mre11-H37R or mre11-H37Y. Also, suppression of sae2Δ camptothecin hypersensitivity by mre11-H37R alleles did not require Exo1, indicating that in contrast to suppression of sae2Δ phenotypes by Ku loss (Mimitou & Symington, 2010), mre11-H37R and mre11-H37Y did not cause cells to become particularly reliant on Exo1 for DSB processing (Fig 3G). Further characterisations, focused on mre11-H37R, revealed that while not suppressing camptothecin hypersensitivity of an xrs2Δ strain (Fig 3H), it almost fully rescued the camptothecin hypersensitivity of a strain expressing the rad50S allele, which phenocopies sae2Δ by somehow preventing functional Sae2–MRX interactions that are required for Sae2 stimulation of Mre11 endonuclease activity (Keeney & Kleckner, 1995; Hopfner et al., 2000; Cannavo & Cejka, 2014; Fig 3I).

**H37R does not enhance Mre11 nuclease activity but impairs DNA binding**

To explore how mre11-H37R mutations might operate, we over-expressed and purified wild-type Mre11, Mre11 H37Y, and Mre11 H37A (Fig 4A and Supplementary Fig S2F) and then subjected these to biochemical analyses. All the proteins were expressed at similar levels and fractionated with equivalent profiles, suggesting that the Mre11 mutations did not grossly affect protein structure or stability. Since Sae2 promotes Mre11 nuclease functions, we initially speculated that sae2Δ suppression would be mediated by mre11-H37R alleles having intrinsically high, Sae2-independent nuclease activity. Surprisingly, this was not the case, with Mre11 H37R actually exhibiting lower nuclease activity than the wild-type protein (Fig 4B). Furthermore, by electrophoretic mobility shift assays, we found that the H37R mutation reduced Mre11 binding to double-stranded DNA.
Conversely, mutation of H37 to alanine, which does not result in a sae2Δ phenotype, did not negatively affect dsDNA-binding activity (Fig 4C) and only partially impaired ssDNA binding (Fig 4D).

Taken together with the fact that the lack of Sae2 only has minor effects on mitotic DSB resection (Clerici et al., 2005), the above results suggested that the sae2Δ suppressive effects of mre11SUPsae2Δ mutations were associated with weakened Mre11 DNA binding and

(dsDNA; Fig 4C) and abrogated Mre11 binding to ssDNA (Fig 4D). Conversely, mutation of H37 to alanine, which does not result in a supH37R phenotype, did not negatively affect dsDNA-binding activity (Fig 4C) and only partially impaired ssDNA binding (Fig 4D).
Figure 4. Mre11H37R is impaired biochemically, particularly at the level of ssDNA binding.

A Mre11 and Mre11H37R were purified to homogeneity from yeast cultures.

B 3’ exonuclease activity assay on Mre11 and Mre11H37R leading to release of a labelled single nucleotide, as indicated.

C, D Electrophoretic mobility shift assays on Mre11, Mre11H37R and Mre11H37A with dsDNA (C) or ssDNA (D).

E Quantification of mre11-H37R suppression of sae2Δ cell DNA damage hypersensitivity. Overnight grown cultures of the indicated strains were diluted and plated on medium containing the indicated doses of CPT. Colony growth was scored 3–6 days later. Averages and standard deviations are shown for each point.

F Intragenic suppression of CPT hypersensitivity of mre11-nd (mre11-H125N) by mre11-H37R. Overnight grown cultures of the indicated strains were treated as in (E). Dotted lines represent data from (E). Averages and standard deviations are shown for each point.

G Mre11 nuclease activity is not required for mre11-H37R-mediated suppression of sae2Δ CPT hypersensitivity. Overnight grown cultures of the indicated strains were treated as in (E). The dotted lines represent data from (E). Averages and standard deviations are shown for each point.

The EMBO Journal
Sae2 function in DNA repair
Fabio Puddu et al

Published online: April 21, 2015

Vol 34 | No 11 | 2015

© 2015 The Authors
were not linked to effects on resection or Mre11 nuclease activity. In line with this idea, by combining mutations in the same Mre11 polypeptide, we established that mre11-H37R substantially rescued camptothecin hypersensitivity caused by mutating the Mre11 active site residue His125 to Asn (Moreau et al., 2001; mre11-H125N; Fig 4E and Supplementary Fig S2F and G), which abrogates all Mre11 nuclease activities and prevents processing of DSBS when their 5’ ends are blocked (Moreau et al., 1999). Even sae2Δ mre11-H37R,H125N cells were resistant to camptothecin and MMS, indicating that Mre11-nuclease-mediated processing of DNA ends is not required for H37R-dependent suppression, nor for DNA repair in this Sa2-deficient setting (Fig 4G and Supplementary Fig S2G). Furthermore, while sae2Δ strains were more sensitive to camptothecin than mre11-H125N strains, the sensitivities of the corresponding strains carrying the mre11-H37R allele were comparable (compare curves 1 and 2 with 3 and 4 in Fig 4F) indicating that mre11-H37R suppresses not only the sae2Δ-induced lack of Mre11 nuclease activity, but also other nuclease-independent functions of Sa2. Nevertheless, mre11-H37R did not rescue the camptothecin hypersensitivity of sae2Δ cells to wild-type levels, suggesting that not all functions of Sa2 are suppressed by this Mre11 allele (Fig 4E and F).

**Identifying an Mre11 interface mediating sae2Δ suppression**

To gain further insights into how mre11ΔLOF/Pow2Δ alleles operate and relate this to the above functional and biochemical data, we screened for additional MRE1 mutations that could suppress camptothecin hypersensitivity caused by Sa2 loss. Thus, we propagated a plasmid carrying wild-type MRE1 in a mutagenic E. coli strain, thereby generating libraries of plasmids carrying mre11 mutations. We then introduced these libraries into a sae2Δmre11Δ strain and screened for transformants capable of growth in the presence of camptothecin (Fig 5A). Through plasmid retrieval, sequencing and functional verification, we identified 12 sae2Δ suppressors, nine carrying single mre11 point mutations and three being double mutants (Supplementary Fig S3A). One single mutant was mre11-H37R, equivalent to an initial spontaneously arising suppressor that we had identified. Among the other single mutations were mre11-P110L and mre11-L89V, both of which are located between Mre11 nuclease domains II and III, in a region with no strong secondary structure predictions (Fig 5B). Two of the three double mutants contained mre11-P110L combined with another mutation that was presumably not responsible for the resistance phenotype (because mre11-P110L acts as a suppressor on its own), whereas the third contained both mre11-Q70R and mre11-G193S. Subsequent studies, involving site-directed mutagenesis, demonstrated that effective sae2Δ suppression was mediated by mre11-Q70R, which alters a residue located in a highly conserved α-helical region (Fig 5C). Ensuing comparisons revealed that the mutations identified did not alter Mre11 protein levels (Supplementary Fig S3B) and that mre11-Q70R suppressed sae2Δ camptothecin hypersensitivity to similar extents as mre11-H37R and mre11-H37Y, whereas mre11-L89V and mre11-P110L were marginally weaker suppressors (Fig 5D).

To map the locations of the various mre11ΔLOF/Pow2Δ mutations within the Mre11 structure, we used the dimeric tertiary structure (Schiller et al., 2012) of the Schizosaccharomyces pombe Mre11 counterpart, Rad32, as a template to generate a molecular model of S. cerevisiae Mre11. The resulting structure had a near-native QMEAN score (0.705 vs 0.778; Benkert et al., 2008), indicating a reliable molecular model. Strikingly, ensuing analyses indicated that the mre11ΔLOF/Pow2Δ mutations clustered in a region of the protein structure distal from the nuclease catalytic site and adjacent to, but distinct from, the interface defined as mediating contacts with dsDNA in the Pyrococcus furiosus Mre11 crystal structure (Williams et al., 2008; Fig 5E; the predicted path of dsDNA is shown in black, while the mre11ΔLOF/Pow2Δ mutations and residues involved in nuclease catalysis are indicated in red and orange, respectively). Furthermore, this analysis indicated that H37 and Q70 are located close together, on two parallel α-helices and are both likely to be solvent exposed (Fig 5F). By contrast, the L89 side chain is predicted to be in the Mre11 hydrophobic core, although modelling suggested that the mre11-L89V mutation might alter the stability of the α-helix containing Q70. We noted that, in the context of the Mre11 dimer, H37 and Q70 are located in a hemicylindrical concave area directly below the position where dsDNA is likely to bind (Fig 5E right, shown by pink hemispheres). Furthermore, by specifically mutating other nearby residues to arginine, we found that the mre11-L77R mutation also strongly suppressed sae2Δ camptothecin hypersensitivity (Fig 5G). As discussed further below, while it is possible that certain mre11ΔLOF/Pow2Δ alleles somehow influence the established dsDNA-binding interface of Mre11, we speculate that mre11-H37R/Y and mre11-Q70R, and at least some of the other suppressors, act by perturbing interactions normally mediated between the Mre11 hemicylindrical concave region and ssDNA (modelled in Fig 5G and discussed further below). Consistent with this idea, we found that the Mre11L77R protein was markedly impaired in binding to ssDNA but not to dsDNA (Supplementary Figs S2E and S3C). However, because P110 lies in the ‘latching loop’ region of eukaryotic Mre11
that is likely to mediate contacts with Xrs2 (Schiller et al., 2012), sae2Δ suppression by this mutation might arise through altering such contacts. A recent report by L. Symington and colleagues reached similar conclusions (Chen et al., 2015).

Taken together, our findings suggested that, in addition to its established dsDNA-binding mode, Mre11 mediates distinct, additional functional contacts with DNA that, when disrupted, lead to suppression of sae2Δ phenotypes. Thus, we suggest that, during DSB processing, duplex DNA entering the Mre11 structure may become partially unwound, with the 5' terminal DNA impairs HR unless counteracted by the actions of Sae2 or weakened by mre11Δsaes2Δ alleles.

sae2Δ phenotypes reflect Mre11-bound DNA repair intermediates

A prediction arising from the above model is that Mre11 persistence and associated Tel1 hyperactivation in sae2Δ cells would be counteracted by mre11Δsaes2Δ mutations. To test this, we constructed yeast strains expressing wild-type Mre11 or Mre11H37R fused to yellow-fluorescent protein (YFP) and then used fluorescence microscopy to examine their recruitment and retention at sites of DNA damage induced by ionising radiation. In line with published work (Lisby et al., 2004), recruitment of wild-type Mre11 to DNA damage foci was more robust and persisted longer when Sae2 was absent (Fig 6A). Moreover, such Mre11 DNA damage persistence in sae2Δ cells was largely attenuated by mre11-H37R (Fig 6A; compare red and orange curves). By contrast, mre11-H37R had little or no effect on Mre11 recruitment and dissociation kinetics when Sae2 was present (compare dark and light blue curves). Importantly, we found that HR-mediated DSB repair was not required for H37R-induced suppression of Mre11-focus persistence in sae2Δ cells, as persistence and suppression still occurred in the absence of the key HR factor, Rad51 (Fig 6B). Also, in accord with our other observations, we found that the rad50Δ allele caused Mre11 DNA damage-focus persistence in a manner that was suppressed by the mre11-H37R mutation (Fig 6C).

Previous work has established that Mre11 persistence on DSB ends, induced by lack of Sae2, leads to enhanced and prolonged DNA damage-induced Tel1 activation, associated with Rad53 hyperphosphorylation (Usui et al., 2001; Lisby et al., 2004; Clerici et al., 2006; Fukunaga et al., 2011). Supporting our data indicating that, unlike wild-type Mre11, Mre11H37R is functionally released from DNA ends even in the absence of Sae2, we found that in a mec1Δ background (in which Tel1 is the only kinase activating Rad53; Sanchez et al., 1996), DNA damage-induced Rad53 hyperphosphorylation was suppressed by mre11-H37R (Fig 7A).

While we initially considered the possibility that persistent Tel1 hyperactivation might cause the DNA damage hypersensitivity of sae2Δ cells, we concluded that this was unlikely to be the case because Tel1 inactivation did not suppress sae2Δ DNA damage hypersensitivity phenotypes (Supplementary Fig S3D). Furthermore, Tel1 loss actually reduced the ability of mre11-H37R to suppress the
Figure 7. Tel1 participates in regulating Mre11 dynamics after DNA damage.

A mre11-H37R suppresses Tel1 hyperactivation induced by Mre11 IRIF persistence in sae2Δ cells.
B Deletion of Tel1 weakens the suppression of the sensitivity of a sae2Δ strain mediated by mre11-H37R.
C Deletion of Tel1 reduces the hyperaccumulation of Mre11 to IRIF and impairs the suppression of their persistence mediated by mre11-H37R (average and standard deviation from two independent experiments).
D mre11-H37R suppresses the sensitivity to CPT of a tel1Δ strain.
E Model for the role of MRX, Sae2 and Tel1 in response to DSBs.
discuss

Our data help resolve apparent paradoxes regarding Sae2 and MRX function by suggesting a revised model for how these and associated factors function in HR (Fig 7E). In this model, after being recruited to DSB sites and promoting Tel1 activation, resection and ensuing Mec1 activation, the MRX complex disengages from processed DNA termini in a manner promoted by Sae2 and facilitated by Tel1 and Mre11 nuclease activity. Sae2 is required to stimulate Mre11 nuclease activity (Cannavo & Čekja, 2014) and subsequently to promote MRX eviction from the DSB end. However, our data suggest that Sae2 can also promote MRX eviction in the absence of DNA-end processing, as mre11-h37r suppresses the phenotypes caused by sae2Δ and mre11-nd to essentially the same extent. Thus, according to our model, when Sae2 is absent, both the nuclease activities of Mre11 and MRX evi- cion are impaired. Under these circumstances, despite resection taking place—albeit with somewhat slower kinetics than in wild-type cells—MRX persists on ssDNA bearing the 3′ terminal OH, thereby delaying repair by HR. In cells containing the mre11-H37R mutation, however, weakened DNA binding together with Tel1 activity promotes MRX dissociation from DNA even in the absence of Sae2, thus allowing the nascent ssDNA terminus to effectively engage in the key HR events of strand invasion and DNA synthesis (Fig 7E). Nevertheless, it is conceivable that abrogation of pathological Tel1-mediated hyperactivation contributes to the resistance of sae2Δmre11-h37r cells to DNA-damaging agents. In this regard, we note that the site of one of the sae2Δ suppressors, P110, lies in the ‘latching loop’ region of eukaryotic Mre11 that is likely to mediate contacts with Xrs2 (Schiller et al., 2012), suggesting that, in this case, sae2Δ suppression might arise through weakening this interaction and dampening Tel1 activity.

Our results also highlight how the camptothecin hypersensitivity of strains carrying a nuclease-defective version of Mre11 does not reflect defective Mre11-dependent DNA-end processing per se, but rather stems from stalling of MRX on DNA ends. We propose that this event delays or prevents HR, possibly by impairing the removal of 3′-bound Top1 as is suggested by the fact that in S. pombe, rad50Δ or mre11-nd alleles are partially defective in Top1 removal from damaged DNA (Hartsuiker et al., 2009). This interpretation also offers an explanation for the higher DNA damage hypersensitivity of sae2Δ cells compared to cells carrying mre11-H125N alleles: while sae2Δ cells are impaired in both Mre11 nuclease activity and Mre11 eviction—leading to MRX persistence at DNA damage sites and Tel1 hyperactivation—mre11-H125N cells are only impaired in Mre11 nuclease activity. Indeed, despite having no nuclease activity, the mre11-H125N mutation does not impair NHEJ, telomere maintenance, mating type switching or Mre11 interaction with Rad50/Xrs2 or interfere with the recruitment of the Mre11—Rad50—Xrs2 complex to foci at sites of DNA damage (Moreau et al., 1999; Lisby et al., 2004; Krogh et al., 2005). In addition, our model explains why the mre11-H37R mutation does not suppress meiotic defects of sae2Δ cells, because Sae2-stimulated Mre11 nuclelease activity is crucial for removing Spo11 from meiotic DSB 5′ termini. Finally, this model explains why mre11-H37R does not suppress the sae2Δ deficiency in DSB repair by SSA because the sae2Δ defect in SSA is suggested to stem from impaired bridging between the two ends of a DSB rather than from the persistence of MRX on DNA ends (Clerici et al., 2005; Andres et al., 2015; Davies et al., 2015). In this regard, we note that SSA does not require an extendable 3′-OH DNA terminus to proceed and so could ensue even in the presence of blocked 3′-OH DNA ends.

We have also found that the mre11-H37R mutation suppresses the DNA damage hypersensitivities of cells impaired in CDK- or Mec1/Tel1-mediated Sae2 phosphorylation. This suggests that such kinase-dependent control mechanisms—which may have evolved to ensure that HR only occurs after the DNA damage checkpoint has been triggered—also operate, at least in part, at the level of promoting MRX removal from partly processed DSBs. Accordingly, we found that TEL1 deletion causes moderate hypersensitivity to camptothecin that can be rescued by the mre11-H37R allele, implying that the same type of toxic repair intermediate is formed in sae2Δ and tell1Δ cells and that in each case, this can be rescued by MRX dissociation caused by mre11-H37R (Fig 7E). Supporting this idea, it has been previously shown that resection relies mainly on Exo1 in both tell1Δ and sae2Δ cells (Clerici et al., 2006; Manietro et al., 2007). We suggest that the comparatively mild hypersensitivity of tell1Δ strains to camptothecin is due to Tel1 loss allowing DSB repair intermediates to be channelled into a different pathway, in which Exo1-dependent resection (Manietro et al., 2007) leads to the activation of Mec1, which can then promote Sae2 phosphorylation and subsequent MRX removal (Fig 7E). The precise role of Tel1 in these events is not yet clear, although during the course of our analyses, we found that the deletion of TEL1 reduced the suppressive effects of mre11-H37R on sae2Δ DNA damage sensitivity and Mre11-focus persistence. This suggests that, in the absence of Sae2, Tel1 facilitates MRX eviction by mre11-H37R, possibly by phosphorylating the MRX complex itself.

Given the apparent strong evolutionary conservation of Sae2, the Mre11—Rad50—Xrs2 complex and their associated control mechanisms, it seems likely that the model we have proposed will also apply to other systems, including human cells. Indeed, we speculate the profound impacts of proteins such as mammalian CtIP and BRCA1 on HR may not only relate to their effects on resection but may also reflect them promoting access to ssDNA bearing 3′ termini so that HR can take place effectively. Finally, our data highlight the power of SVGS to identify genetic interactions—including those such that we have defined that rely on separation-of-function mutations rather than null ones—and also to inform on underlying biological and biochemical mechanisms. In addition to...
being of academic interest, such mechanisms are likely to operate in medical contexts, such as the evolution of therapy resistance in cancer.

Materials and Methods

Strain and plasmid construction

Yeast strains used in this work are derivatives of SK1 (meiotic phenotypes), YMV80 (SSA phenotypes) and haploid derivatives of W303 (all other phenotypes). All deletions were introduced by one-step gene disruption. pRS303-derived plasmids, carrying a wt or mutant MRE11 version, were integrated at the MRE11 locus in an mre11A::KanMX6 strain. Alternatively, the same strain was transformed with pRS416-derived plasmids containing wild-type or mutant MRE11 under the control of its natural promoter. Strains expressing mutated mre11-YFP were obtained in two steps: integration of a pRS306-based plasmid (pFP118.1) carrying a mutated version of Mre11 in a MRE11-YFP sae2Δ strain, followed by selection of those ‘pop-out’ events that suppressed camptothecin hypersensitivity of the starting strain. The presence of mutations was confirmed by sequencing. Full genotypes of the strains used in this study are described in Supplementary Table S1; plasmids are described in Supplementary Table S2.

Whole-genome paired-end DNA sequencing and data analysis

DNA (1–3 μg) was sheared to 100–1,000 bp by using a Covaris E210 or LE220 (Covaris, Woburn, MA, USA) and size-selected (350–450 bp) with magnetic beads (Ampure XP; Beckman Coulter). Sheared DNA was subjected to Illumina paired-end DNA library preparation and PCR-amplified for six cycles. Amplified libraries were sequenced with the HiSeq platform (Illumina) as paired-end 100 base reads according to the manufacturer’s protocol. A single sequencing library was created for each sample, and the sequencing coverage per sample is given in Supplementary Table S3. Sequencing reads from each lane were aligned to the S. cerevisiae S288c assembly (R64-1-1) from Saccharomyces Genome Database (obtained from the Ensembl genome browser) by using BWA (v0.5.9-r16) with the parameter ‘-q 15’. All lanes from the same library were then merged into a single BAM file with Picard tools, and PCR duplicates were marked by using Picard ‘MarkDuplicates’ (Li et al., 2009). All of the raw sequencing data are available from the ENA under accession ERP001366. SNPs and indels were identified by using the SAMTools (v0.1.19) mpileup function, which finds those ‘pop-out’ events that suppressed camptothecin hypersensitivity of the starting strain. The presence of mutations was confirmed by sequencing. Full genotypes of the strains used in this study are described in Supplementary Table S1; plasmids are described in Supplementary Table S2.

Molecular modelling

A monomeric molecular model of S. cerevisiae Mre11 was generated with the homology modelling program MODELLER (Sali & Blundell, 1993) v9.11, using multiple structures of Mre11 from S. pombe (PDB codes: 4FBW and 4FBK) and human (PDB code: 3T1I) as templates. A structural alignment of them was made with the program BATON (Sali & Blundell, 1990) and manually edited to remove unmatched regions. The quality of the model was found to be native-like as evaluated by MODELLER’s NDOPE (~1.2) and GA341 (1.0) metrics and the QMEAN server (Benkert et al, 2009) (http://swissmodel.expasy.org/qmean/) (0.705). The monomeric model was subsequently aligned on the dimeric assembly of the 4FBW template to generate a dimer, and the approximate position of DNA binding was determined by aligning the P. furiosus structure containing dsDNA (PDB code: 3DSC) with the dimeric model. All images were obtained using the PyMOL Molecular Graphics System.

Microscopy

Exponentially growing yeast strains carrying wild-type or mutant Mre11-YFP were treated with 40 Gy of ionising radiations with a Faxitron irradiator (CellRad). At regular intervals, samples were taken and fixed with 500 μl of Fixing Solution (4% paraformaldehyde, 3.4% sucrose). Cells were subsequently washed with wash solution (100 mM potassium phosphate pH 7.5, 1.2 M sorbitol) and mounted on glass slides. Images were taken at a DeltaVision microscope. All these experiments were carried out at 30°C.

In vitro assays

For the electrophoretic mobility shift assay (EMSA), a radiolabelled DNA substrate (5 nM) was incubated with the indicated amount of Mre11 or Mre11ΔH37R in 10 μl buffer (25 mM Tris–HCl, pH 7.5, 1 mM DTT, 100 μg/ml BSA, 150 mM KCl) at 30°C for 10 min. The reaction mixtures were resolved in a 10% polyacrylamide gel in TBE buffer (89 mM Tris–borate, pH 8.0, 2 mM EDTA). The gel was dried onto Whatman DE81 paper and then subjected to phosphorimaging analysis. For nuclelease assay, 1 mM MnCl2 was added to the reactions and the reaction mixtures were incubated at 30°C for 20 min and deproteinised by treatment with 0.5% SDS and 0.5 mg/ml proteasine K for 5 min at 37°C before analysis in a 10% polyacrylamide gel electrophoresis in TBE buffer.

Additional Materials and Methods can be found in the Supplementary Methods.

Supplementary information for this article is available online: http://emboj.embopress.org
Acknowledgements

We thank M.P. Longhese, R. Rothstein and J. Haber for providing strains and plasmids, Sir T. Blundell and T. Ochi for advice on structural biology and for providing comments to the manuscript. Research in the Jackson laboratory is funded by Cancer Research UK Programme Grant C6/A11224, the European Research Council and the European Community Seventh Framework Programme Grant Agreement No. HEALTH-F2-2010-259893 (DDResponse). Core funding is provided by CRUK (C6946/Al1492) and the Welcome Trust (WT0992069). SP receives his salary from the University of Cambridge, UK, supplemented by CRUK. TO, IC and FP were funded by Framework Programme Grant Agreement No. HEALTH-F2-2010-259893 (DDResponse). FP also received funding from EMBO (Fellowship ALTF 1287-2011). NG and IS are funded by the Wellcome Trust (201126/Z/13/Z). DJA and TMK were supported by Cancer Research UK and the Wellcome Trust (WT098051). PS and HN were supported by NIH grants ROI1E0007061 and K99ES0241-1, respectively.

Author contributions

The initial screening was conceived and designed by TO, EV, DJA and SPJ. Alignment of whole-genome sequencing data, variant calling and subsequent analysis was carried out by MH and TMK. Experiments for the characterisation of the mre11-h37r mutant were conceived by TO, IC, FP and SPJ, and were carried out by TO, FP, IC, NJG, EV and IS. Biochemical assays were designed by SPJ, PS and HN and carried out by HN. The identification of further mre11:-proneo mutants was designed by FP and SPJ and carried out by NJG. Modelling of S. cerevisiae Mre11 was performed by BO-M, and subsequent analyses were carried out by BO-M and FP. The manuscript was largely written by SPJ and FP, and was edited by all other authors.

Conflict of interest

The authors declare that they have no conflict of interest.

References


License: This is an open access article under the terms of the Creative Commons Attribution 4.0 License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
Genome-wide genetic screening with chemically-mutagenized haploid embryonic stem cells

Josep V. Forment¹,², Mareike Herzog¹,², Julia Coates¹, Tomasz Konopka³, Bianca V. Gapp³, Sebastian M. Nijman³,⁴, David J. Adams², Thomas M. Keane² and Stephen P. Jackson¹,²

¹The Wellcome Trust CRUK Gurdon Institute and Department of Biochemistry, University of Cambridge, Cambridge, UK
²The Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
³Ludwig Institute for Cancer Research Ltd. and Target Discovery Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK
⁴Research Center for Molecular Medicine of the Austrian Academy of Sciences (CeMM), Vienna, Austria

Authors for correspondence:
Josep V. Forment j.forment@gurdon.cam.ac.uk
Stephen P. Jackson s.jackson@gurdon.cam.ac.uk
Abstract

In model organisms, classical genetic screening via random mutagenesis has provided key insights into the molecular bases of genetic interactions, helping defining synthetic-lethality, -viability and drug-resistance mechanisms. The limited genetic tractability of diploid mammalian cells, however, has precluded this approach. Here, we demonstrate the feasibility of classical genetic screening in mammalian systems by using haploid cells, chemical mutagenesis and next-generation sequencing, providing a new tool to explore mammalian genetic interactions.
Classical genetic screens with mutagens have been extremely valuable in assigning functionality to genes in many model organisms\(^1\-^3\). Since most mutagenic agents yield random single-nucleotide variants (SNVs), clustering of mutations can provide valuable information on the functionality of protein domains and also define key amino acid residues\(^4\). The discovery of RNA interference (RNAi) allowed forward genetic screening in human cell cultures\(^4\) and, more recently, insertional mutagenesis in near-haploid human cancer cells\(^5\) and whole-genome CRISPR/Cas9 small-guide RNA (sgRNA) libraries have been used for this purpose\(^6\)-\(^8\). Although powerful, such loss-of-function (LOF) approaches miss phenotypes caused by separation-of-function or gain-of-function SNV mutations\(^9\),\(^10\), are less informative on protein function, and are not well suited to studying functions of essential genes.

Here, we describe the generation of SNV-mutagenized mammalian cell libraries, and establish their suitability to identify recessive suppressor mutations using resistance to the antimetabolite 6-thioguanine (6-TG) as a proof-of-principle.

Comprehensive libraries of homozygous SNV-containing mutant clones are not feasible to obtain in cells with diploid genomes. To circumvent this issue, we used H129-3 haploid mouse embryonic stem cells (mESCs)\(^11\) treated with varying doses of the DNA-alkylating agent ethylmethanesulfonate (EMS), a chemical inducer of SNVs\(^12\) (Fig. 1a, Supp. Fig. 1a). For comparison purposes, the same procedure was performed on diploid H129-3 mESCs (Supp. Fig. 1b). Haploid and diploid mutant libraries were then screened for suppressors of cellular sensitivity to the toxic nucleotide precursor 6-TG (Fig. 1b). Libraries of the EMS dose that produced more 6-TG resistant clones showed a near 6-fold difference between haploid and diploid cells (Supp. Fig. 1c), highlighting the increased accumulation of suppressor mutations in the haploid genetic background.

196 resistant clones were isolated from haploid libraries treated with 6-TG. To test the feasibility of identifying causative suppressor mutations, DNA from seven of these resistant clones and from control mESCs not treated with EMS was subjected to whole-exome sequencing. Homozygous SNVs and base insertions/deletions (INDELs) were identified (Fig. 1c), and only a small proportion of them affected coding sequences and were non-synonymous (Fig. 1d, Supp. Table 1). When
analyzing this subset, suppressor gene candidates were defined as those appearing mutated in multiple independent clones and harboring potential deleterious mutations (Supp. Table 1). Importantly, *Hprt*, the gene encoding hypoxanthine-guanine phosphoribosyltransferase, the sole 6-TG target\textsuperscript{13} (Fig. 1b), appeared mutated in five of the sequenced clones. Moreover, it was the only candidate suppressor gene carrying potentially deleterious mutations in all clones where mutational consequences could be assigned (Fig. 1e, Supp. Table 1). These results established that, without using any previous knowledge regarding the identity of suppressor loci, we identified *Hprt* as a top gene candidate after sequencing of very few clones.

In addition to mutations in the *Hprt* gene, inactivation of DNA mismatch repair (MMR) protein components Msh2, Msh6, Mlh1 and Pms2 has also been shown to confer resistance to 6-TG\textsuperscript{14}, as does mutations in DNA methyltransferase Dnmt1\textsuperscript{15}. In fact, the two whole-exome sequenced clones that did not carry mutations in *Hprt* presented nonsense mutations in *Msh6* and *Pms2* (Supp. Table 1, Supp. Fig. 1d). To analyze coverage of the mutant libraries, we subjected the 189 additional suppressor clones to targeted sequencing of the known suppressor genes (Fig. 1b). Importantly, deleterious mutations in most of these genes were identified in several independent resistant clones (Fig. 2a, Supp. Table 2). Thus, if the same non-targeted whole-exome sequence approach carried out in the initial analysis of seven suppressor clones would have been applied to all of them, *Hprt, Msh2, Msh6, Mlh1* and *Pms2* (as genes carrying independent homozygous deleterious mutations in different resistant clones) would have been identified as strong suppressor gene candidates, confirming the feasibility of the approach.

Interestingly, a subset of clones presented heterozygous deleterious mutations in known suppressor genes (Supp. Table 2). These could have arisen after diploidization of the original EMS-treated haploid population, or could have occurred in the small proportion of diploid H129-3 cells present during EMS treatment of the enriched haploid population (Fig. 1a). Regardless of their origin, deleterious heterozygous mutations could only generate 6-TG resistance if each would affect one allele of the gene, effectively inactivating both copies. Heterozygous mutations in
the Dnmt1 gene occurred in such close proximity that they could be analyzed from the same sequencing reads. No co-occurrence of heterozygous mutations in the same reads indicated that Dnmt1 mutant clones were compound heterozygotes (Fig. 2b). As these mutations all scored as potentially deleterious for Dnmt1 protein function (Supp. Table 2), it is likely that they are causative of the suppression to 6-TG sensitivity in these clones (see below). When deleterious heterozygous mutations were taken into account, Dnmt1 could also be included in the list of suppressor gene candidates (Fig. 2c).

Highlighting the applicability of the methodology to identify functionally important protein regions, missense and nonsense variants linked to clinically-relevant mutations in Hprt (causative of the inherited neurological disorder Lesch-Nyhan syndrome and its variants16) and in genes involved in DNA MMR (linked to the inherited colon cancer predisposition Lynch syndrome17) were effectively retrieved (Fig. 3a). Furthermore, and due to the mutational preferences of EMS (see below), mRNA splicing variant mutations potentially affecting total protein levels of Dnmt1, Hprt, Mlh1, Msh2 and Msh6 were also found (Supp. Table 2). These were particularly prevalent in Hprt (Fig. 3a), and a detailed analysis of them confirmed their deleterious consequence at the protein level (Supp. Figure 2). Production of aberrant mRNA splicing forms, with the subsequent reduction or absence of protein product, is thus an important consequence of the mutagenic action of EMS.

Non-described mutations in Dnmt1, Hprt, Mlh1, Msh6 and Pms2 were also identified, most of which with predicted deleterious effects on the protein product (Fig. 3b, Supp. Table 2). Newly identified A612T and G1157E mutations in Mlh1 and Dnmt1, respectively, were introduced de novo into wild-type mESCs by CRISPR/Cas9 gene editing (Supp. Fig. 3). We chose these mutations as they are missense mutations only identified in heterozygotes, and we wanted to test their ability to generate suppression when occurring in homozygosis. Importantly, H129-3 mESCs carrying engineered A612T Mlh1 or G1157E Dnmt1 mutations were resistant to 6-TG treatment to differing extents when compared to their wild type counterparts (Fig. 3c), showing their potential as causative mutations of the suppressor phenotype.
A small group of resistant clones (23) did not present mutations in any of the known suppressor genes (Fig. 2a,c). These “orphan” clones were subjected to whole-exome DNA sequencing and RNA sequencing. DNA sequencing of the unassigned suppressor clones and several control samples allowed an unprecedented description of EMS mutagenic action at the whole-exome level, confirming its preference in producing SNVs, and transitions rather than transversions (Supp. Fig. 4). Although whole-exome sequencing effectively retrieved causative mutations in all control samples resistant to 6-TG, no other obvious gene candidate could be identified from the remaining orphan suppressors (Supp. Table 3). RNA sequencing, however, revealed significantly reduced expression levels of Hprt, Mlh1 or Msh6 as potential causes of suppression in several such clones (Fig. 3d,e; Supp. Table 4). Further studies will be required to define whether epigenetic alterations or mutations in transcriptional regulatory sequences outside of exon regions, and hence not covered during DNA sequencing, could explain the nature of these orphan suppressor clones.

Collectively, our findings establish that classical genetic screening can be effectively performed in mammalian systems by combining the use of haploid cells, a chemical inducer of SNVs, and next-generation DNA and RNA sequencing techniques. Use of haploid cells when creating libraries of SNV mutants allowed identification of recessive suppressor point mutations, in contrast to diploid cell screening where only dominant mutations are effectively retrieved\(^{18}\). Furthermore, EMS induction of SNVs allowed generation of complex mutant libraries, thus increasing the probability of identification of suppressor loci compared to isolation of rare, spontaneous suppressor events\(^{19}\). Importantly, through screening for cellular resistance to 6-TG we identified point mutations in all described suppressor genes, showing high coverage capability. Moreover, as we have established for 6-TG suppressor loci, SNVs have value in delineating key residues required for protein function, thus helping to explain molecular mechanisms of suppression. SNV-based mutagenesis will also be a useful technique to investigate genetic interactions of essential genes, and we envisage the applicability of this approach into haploid cells of human origin\(^{20-22}\). Chemical mutagenesis of haploid cells, either alone or in combination with
LOF screens, thus has the potential to bring functional genomics in mammalian systems to a hitherto unachieved comprehensive level.

**Methods**

Methods and any associated references are available in the online version of the paper.
Figure legends

**Figure 1. Generation of mutagenized libraries.** (a) Experimental workflow. (b) Schematic of 6-TG metabolism and genotoxicity. Inactivating mutations in the genes highlighted in red have been shown to confer resistance to 6-TG. (c) Mutation types identified by whole-exome sequencing of 7 suppressor clones. (d) Consequences of identified mutations. (e) Genes harboring independent mutations in different clones. Mutations were assigned as deleterious or neutral according to PROVEAN and SIFT software (see Methods).

**Figure 2. Identification of suppressor mutations.** (a) Distribution of homozygous mutations identified in suppressor gene candidates; numbers of independent clones are in brackets and types of Hprt mutations are shown in detail. (b) Examples of sequencing reads obtained for heterozygous mutations affecting the Dnmt1 gene. SNVs causing missense mutations G1157E or G1157R (top panel) and G1477R or affecting the splicing donor sequence on intron 36 (bottom panel; see also Supp. Fig. 2), were never detected in the same sequencing read, indicating that they locate to different alleles. (c) Distribution of suppressor gene candidate mutations identified, including heterozygous deleterious mutations.

**Figure 3. Clinically-relevant and newly-identified suppressor mutations.** (a) Distribution of point mutations on Dnmt1, Hprt and MMR proteins; each square represents an independent clone. Asterisks (*) denote STOP-codon gains. (b) Predicted consequences of potential new suppressor mutations. Consequences were predicted as in Fig. 1e. (c) De novo introduction of new mutations Dnmt1 G1157E and Mlh1 A612T confers cellular resistance to 6-TG. (d) Hprt, Mlh1 and Msh6 mRNA expression levels (fragments per kilobase per million reads). Black dots indicate wild-type (WT) samples, red dots represent clones with already identified mutations (controls), and white dots represent samples for which no causative mutations were identified (see Supp. Table 2 for identifiers). Error bars represent uncertainties on expression estimates. (e) Reduced Hprt mRNA levels correspond to reduced protein production as detected by western blot.
References


Acknowledgements

We thank all S.P.J. lab members for discussion, and especially A. Blackford, F. Puddu and C. Schmidt for critical reading of the manuscript. We thank M. Leeb for H129-3 cells and advice on haploid ES cell culture conditions. We thank C.D. Robles-Espinoza for helping designing the array of baits for the exon-capture experiment, and J. Hewinson for technical support. Research in the S.P.J. laboratory is funded by Cancer Research UK (CRUK; programme grant C6/A11224), the European Research Council and the European Community Seventh Framework Programme (grant agreement no. HEALTH-F2-2010-259893; DDResponse). Core funding is provided by Cancer Research UK (C6946/A14492) and the WellcomeTrust (WT092096). S.P.J. receives salary from the University of Cambridge, supplemented by CRUK. J.V.F. is funded by Cancer Research UK programme grant C6/A11224 and the Ataxia Telangiectasia Society. J.C. is funded by Cancer Research UK programme grant C6/A11224. D.J.A. is supported by CRUK. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. [311166]. B.V.G. is supported by a Boehringer Ingelheim Fonds PhD fellowship.
Supplementary Figure legends

Supplementary Figure 1. Mutant library production controls and top candidate suppressor mutations identified. (a) Cellular toxicity to various EMS doses used to generate mutant libraries. (b) Cell cycle profile of haploid and diploid H129-3 mESCs. (c) EMS-mutagenized haploid and diploid mESC libraries were treated with 2 μM 6-TG for 6 days, and surviving cells were stained with crystal violet (left panel). Suppressor frequencies to 6-TG treatment of the different EMS-mutagenized libraries, represented as number of suppressor clones isolated per 10,000 plated cells (right panel). (d) Top candidate mutations conferring 6-TG resistance in the 7 suppressor clones sequenced (left panel). Asterisks (*) denote STOP-codon gains. SDV, splicing donor variant (see Supp. Fig. 2). Protein depletion in some clones was confirmed by western blotting (right panel).

Supplementary Figure 2. Splicing mutants in the Hprt gene. (a) Types of splicing variant mutations identified in Hprt. Mutated positions are highlighted in bold, and followed by the changed base in brackets. Exonic sequences are in capital letters, intronic sequences in lower case. SDV, splicing donor variant. SAV, splicing acceptor variant. SRV, splicing region variant. (b) Position of splicing variant mutations in Hprt exon-intron junctions. (c) Hprt splicing variant mutations result in reduced Hprt protein levels as judged by western blot analysis.

Supplementary Figure 3. Knock-in generation of Dnmt1 G1157E and Mlh1 A612T mutant cell lines. (a) Upper panel. Position of small-guide RNAs (sgRNAs) designed to introduce the Dnmt1 G3662A mutation (nucleotide number based on cDNA sequence; amino acid G1157E mutation). Protospacer adjacent motif (PAM) sequences for each sgRNA are also depicted, and Cas9 nickase cutting sites marked with arrows. Lower panel. Dnmt1 sequence after gene editing. Mutations to abolish sgRNA binding, introduce the G1157E mutation and an EcoRI restriction site to allow screening, are in lower case and highlighted in pink. Right panel. EcoRI digestion of the PCR amplification of the region surrounding G3662 in wild-type (WT) and gene-edited cells. (b) Upper panel. Position of sgRNAs designed to introduce
the *Mlh1* G2101A mutation (nucleotide number of cDNA sequence; amino acid A612T mutation). PAM sequences are also depicted and Cas9 nickase cutting sites marked with arrows. Lower panel. *Mlh1* sequence after gene editing (annotations as in a). Right panel. *Eco*RI digestion of the PCR amplification of the region surrounding G2101 in WT and gene-edited cells.

**Supplementary Figure 4. EMS mutagenic action.** (a) Distribution of mutation types identified by whole-exome sequencing of 66 suppressor clones (23 orphan clones plus 43 clones with identified mutations). SNV, single-nucleotide variant. INDEL, insertion or deletion. Only homozygous mutations were considered. (b) Distribution of identified SNVs. (c) EMS mutational pattern. (d) Number of mutations per chromosome in sequenced clones. Mutation numbers (both homozygous and heterozygous) were normalized to exon bait coverage. (e) Heat map showing homogenous distribution of EMS-induced mutations in all chromosomes. Differences observed in the X chromosome could be accounted by its frequent loss in ES cells in culture (Robertson et al, *J Embryol Exp Morphol*, 74, 1983). *P* values were calculated by the Kruskal-Wallis test for multiple comparisons.
**Supplementary Table legends**

**Supplementary Table 1.** Homozygous mutations identified through whole-exome sequencing of 7 suppressor clones.

**Supplementary Table 2.** Homozygous mutations identified on the targeted exon-capture experiment performed on 189 suppressor clones. Heterozygous mutations affecting *Dnmt1, Hprt, Mlh1, Msh2, Msh6* and *Pms2* are also shown.

**Supplementary Table 3.** Homozygous mutations identified through whole-exome sequencing of 66 suppressor clones (23 orphan clones plus 43 clones with identified mutations). Heterozygous mutations affecting *Dnmt1, Hprt, Mlh1, Msh2, Msh6* and *Pms2* are also shown.

**Supplementary Table 4.** RNA sequencing data from 5 wild-type samples, 5 identified suppressor clones and 21 unidentified suppressor clones. Values represent fragments per kilobase per million reads.

**Supplementary Table 5.** DNA sequencing coverage for the whole-exome and targeted exon-capture experiments.
Forment et al, Figure 1
Forment et al, Figure 2
### Consequence

<table>
<thead>
<tr>
<th>Protein</th>
<th>Mutation</th>
<th>PROVEAN</th>
<th>SIFT</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dnmt1</td>
<td>G1157E</td>
<td>-7.67</td>
<td>0</td>
<td>Deleterious</td>
</tr>
<tr>
<td></td>
<td>G1147R</td>
<td>-7.17</td>
<td>0</td>
<td>Deleterious</td>
</tr>
<tr>
<td>Hprt</td>
<td>D44V</td>
<td>-7.45</td>
<td>0</td>
<td>Deleterious</td>
</tr>
<tr>
<td></td>
<td>V116I</td>
<td>-0.03</td>
<td>0.841</td>
<td>Neutral</td>
</tr>
<tr>
<td></td>
<td>S162N</td>
<td>-2.18</td>
<td>0.004</td>
<td>Deleterious</td>
</tr>
<tr>
<td>Mlh1</td>
<td>T556I</td>
<td>-2.50</td>
<td>0.05</td>
<td>Deleterious</td>
</tr>
<tr>
<td></td>
<td>A612T</td>
<td>-3.06</td>
<td>0.002</td>
<td>Deleterious</td>
</tr>
<tr>
<td></td>
<td>G642V</td>
<td>-3.77</td>
<td>0.017</td>
<td>Deleterious</td>
</tr>
<tr>
<td>Pms2</td>
<td>V168M</td>
<td>-2.73</td>
<td>0</td>
<td>Deleterious</td>
</tr>
</tbody>
</table>

**Forment et al, Figure 3**
Forment et al, Supplementary Figure 1
### Splicing variant

<table>
<thead>
<tr>
<th>Splicing variant</th>
<th>Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDV1</td>
<td>...GTG-g(a)tg...</td>
</tr>
<tr>
<td>SAV1</td>
<td>...cag(a)-ATT...</td>
</tr>
<tr>
<td>SDV2</td>
<td>...CAG-g(a)tt...</td>
</tr>
<tr>
<td>SAV2</td>
<td>...tag(a)-GAC...</td>
</tr>
<tr>
<td>SDV3</td>
<td>...TGT-g(a)ta...</td>
</tr>
<tr>
<td>SAV4</td>
<td>...tag(a)-AAT...</td>
</tr>
<tr>
<td>SDV4</td>
<td>...AAG-g(a)ta...</td>
</tr>
<tr>
<td>SAV5</td>
<td>...AAG-g(a)ta...</td>
</tr>
<tr>
<td>SRV5</td>
<td>...GAA-gtaag(t)...</td>
</tr>
<tr>
<td>SAV6</td>
<td>...cag(a)-CTT...</td>
</tr>
<tr>
<td>SDV7</td>
<td>...ACT-g(a)ta...</td>
</tr>
<tr>
<td>SDV8</td>
<td>...AAT-g(a)ta...</td>
</tr>
<tr>
<td>SRV8</td>
<td>...AAT-gtaag(a)...</td>
</tr>
</tbody>
</table>

Forment et al, Supplementary Figure 2
Forment et al, Supplementary Figure 3
Forment et al, Supplementary Figure 4
Some supplementary files may need to be viewed online via your Referee Centre at http://mc.manuscriptcentral.com/nar.

Chromatin determinants impart camptothecin hypersensitivity in the absence of the Tof1/Csm3 replication pausing complex

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Nucleic Acids Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>NAR-01713-M-2016</td>
</tr>
<tr>
<td>Manuscript Type</td>
<td>1 Standard Manuscript</td>
</tr>
<tr>
<td>Key Words:</td>
<td>Synthetic viability, Tof1, Fork Protection Complex, Camptothecin, SIR complex</td>
</tr>
</tbody>
</table>
Chromatin determinants impart camptothecin hypersensitivity in the absence of the Tof1/Csm3 replication pausing complex

Fabio Puddu¹,†, Mareike Herzog¹², Nicola J. Geisler¹, Vincenzo Costanzo³ and Stephen P. Jackson¹²,†

¹The Gurdon Institute and Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK
²The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK.
³IFOM (Fondazione Istituto FIRC di Oncologia Molecolare) via Adamello 16, 20139 Milan, Italy.

† to whom correspondence should be addressed:
Stephen P. Jackson, s.jackson@gurdon.cam.ac.uk
Correspondence may also be addressed to:
Fabio Puddu, f.puddu@gurdon.cam.ac.uk
Abstract

Camptothecin-induced Top1 locking on DNA generates a physical barrier to replication fork progression and creates topological stress. In *Saccharomyces cerevisiae*, absence of the Tof1/Csm3 complex causes camptothecin hypersensitivity by allowing replisome rotation, which converts impending topological stress to DNA catenation. By using a synthetic viability screening approach, we have discovered that inactivation of histone H4-K16 deacetylation suppresses much of the sensitivity of wild-type cells to camptothecin and the hypersensitivity of *tof1Δ* strains towards this agent. We show that disruption of Sir1-dependent heterochromatin that is established at silent mating-type loci and likely in other regions of the genome is sufficient to suppress camptothecin sensitivity in wild-type and *tof1Δ* cells. We have also found that the Tof1/Csm3 complex prevents loss of epigenetic silencing when this cannot be re-established by Sir1, and suggest a model in which DNA hypercatenation generated in the absence of the Tof1/Csm3 complex perturbs histone deposition.


**Introduction**

Separation of the two parental DNA strands during DNA replication creates positive supercoiling ahead of the replication fork. Such over-winding hinders replisome progression and must be removed for DNA replication to be completed. In *Saccharomyces cerevisiae*, the main DNA topoisomerase that relaxes positive supercoiling during DNA replication is Top1, a type IB topoisomerase (1, 2). Despite the importance of DNA uncoiling for replication, cells lacking Top1 can fully replicate their genome because replisomes, by rotating along their axes, can convert impending positive supercoiling into intertwines/concatenation between the two daughter DNA strands (3). The concatenation generated in this way is an obstacle to chromosome segregation and must be resolved by Top2, a type II topoisomerase, before the onset of mitosis (4). In contrast to Top1, Top2 is essential in yeast cells because a certain amount of concatenation is generated even in wild-type cells, possibly because Top1 cannot relieve topological stress between replisomes converging towards replication termination zones (5).

Consistent with this model, increased fork rotation has been observed when replication forks approach stable fork-pausing structures, such as centromeres, tRNA genes, inactive replication origins (6) and potentially retrotransposon long terminal repeats (LTRs) and transcriptionally repressed chromatin (7, 8).

To reduce the requirement for decatenation, replisome rotation is normally restricted by the Tof1/Csm3 complex (6), the yeast homolog of the mammalian Timeless/Tipin complex. Tof1 and Csm3 are also crucial for proper pausing of replication forks at the replication fork
barriers present in the tandem arrays that form the large ribosomal DNA locus (9). Independently of these functions, the Tof1/Csm3 complex also interacts with Mrc1 (10), which functions as an adaptor to transmit signals from the apical replication-checkpoint kinase Mec1 to the transducer kinase Rad53 during replication stress induced by nucleotide depletion (11). The fact that tof1Δ strains, similarly to mrc1Δ strains, show synergistic phenotypes in combination with loss of Rad9 – the other major checkpoint adaptor protein in S. cerevisiae – suggests that the Tof1/Csm3 complex recruits Mrc1 for the purpose of Rad53 activation (9, 12). In this regard, it is noteworthy that Mrc1 also has checkpoint-independent functions and can be recruited to replication forks independently of Tof1/Csm3 (11, 13, 14).

Despite the above findings, certain results have remained unexplained, and the exact role of the Tof1/Csm3 complex has remained elusive. For instance, tof1Δ and csm3Δ yeast strains were shown to be hypersensitive to high doses of camptothecin (15), a drug that induces DNA double-strand DNA breaks (DSBs) during S phase by trapping Top1 in a covalent complex with DNA. These strains, however, are not hypersensitive to other agents that induce DSBs, such as ionising radiation, or to drugs such as hydroxyurea that affect S phase progression (15), suggesting that the camptothecin hypersensitivity of tof1Δ and csm3Δ strains might arise through topologically stressed DNA structures generated by Top1 inhibition rather than from DNA damage per se (16, 17).

Materials and Methods

Yeast Strains and Plasmids. Yeast strains used for this work are haploid derivatives of W303 unless otherwise indicated, and are listed in Supplementary Table 1. All deletions were introduced by one-step gene disruption/tagging (18). Strains carrying histone H4 mutations were obtained by plasmid shuffling, transforming the strain JHY6 (hht1-hhf1Δ::KanMX6 hht2-hhf2Δ::HPH) with plasmids obtained by site-directed mutagenesis of plasmid pMR206 (HHT2-
HHF2; TRP1). Strains to detect transient loss of silencing events were prepared by introducing TOFI deletion in the diploid strain JRY9730 (sir1Δ::LEU2/SIR1 HMRα/HMRα-a2Δ::cre ura3Δ::PGPD-loxP-yEmRFP-TCYC1-kanMX-loxP-yEGFP-TADH1/ura3) (19) and by recovering the appropriate spores after sporulation.

Whole-genome paired-end DNA sequencing and data analysis was performed as previously described (20). All raw sequencing data are available from the European Nucleotide Archive (ENA) under the accession codes detailed in Supplementary Table 2. SNPs and indels were identified by using the SAMtools (v0.1.19) mpileup function, which finds putative variants and indels from alignments and assigns likelihoods, and BCFtools that performs the variant calling (21). The following parameters were used: for SAMtools (v0.1.19) mpileup -EDS -C50 -m2 -F0.0005 -d 10000’ and for BCFtools (v0.1.19) view ‘-p 0.99 -vcgN’. Functional consequences of the variants were produced by using the Ensembl VEP (22).

Drug sensitivity assays. Overnight-grown saturated cultures of the indicated strains were serially diluted (10 fold) in water. 10 μl drops of each dilution were the deposited on each plate. Images were scanned two to three days after plating and growth at 30°C.

Analysis of cell cycle progression. Exponentially growing cultures (30°C) were synchronised in G1 by addition of 5 μg/ml alpha factor for 2 hours. G1 synchronised cultures were then transferred to fresh YPD and released into S phase in the presence or in the absence of camptothecin and/or sirtinol. 45 minutes after the release, 20 μg/ml alpha factor was added to allow quantification of G1 cells by preventing re-entry into the cell cycle.

Mating and silencing assays. Mating assays were performed by using saturated cultures of the indicated strains. 10 fold serial dilutions of each culture were prepared and deposited on plates lacking amino acids previously seeded with tester strains 6122a and 6122alpha (HIS3 TRP1 LEU2 URA3 lys2). Growth ensued only after mating of the deposited strain with the
tester strain by mutual complementation of auxotrophies. Assays to detect transient loss of silencing events were performed as previously described (19).

**Analysis of ChIP-seq data**

ChIP-seq data were downloaded from the Sequence Read Archive (NCBI) using accession numbers specified in Supplementary Table 3. Reads were aligned using BWA-MEM. For each genomic position, coverage was calculated using bedtools genomcov and normalised using the genome-wide median of each sample. For each genomic position, the enrichment (E) was calculated as the ratio of the normalised coverages of IP and input samples. Every genomic position showing $E_{\text{str}2}>0.9$ and $E_{\text{str}3}>1.1$ and $E_{\text{str}2}>0.9$ and $E_{\text{GFP}}<1.1$ and $E_{\text{H}3}<0.75$ was exported to a bed file. These values were determined empirically and small adjustments did not substantially alter the final results. The bed file was queried with the coordinates of every annotated ORF to calculate the total number of positions in each ORF for which the above conditions are true. The final SIR score was obtained by dividing this number by the length of the ORF.

**Results**

To understand the roles of the Tof1/Csm3 complex during DNA replication, we investigated the basis for the camptothecin hypersensitivity of *TOF1* or *CSM3*-deleted cells. This hypersensitivity arises from the well-established trapping of Top1 in a covalent complex with DNA, as shown by the fact that it was rescued by *TOP1* deletion ([Figure 1A](#)). Notably, *mrclΔ* strains were not hypersensitive to camptothecin ([Figure 1A and (15)]), indicating that a defect in replication checkpoint activation does not explain the camptothecin hypersensitivity of *tof1Δ* or *csm3Δ* strains. Moreover, *tof1Δ/csm3Δ* sensitivity does not arise from issues connected to fork pausing at the replication fork barrier on ribosomal DNA, as pausing-
deficient \textit{fob1}Δ strains were not hypersensitive to camptothecin and \textit{FOB1} deletion did not alleviate the camptothecin hypersensitivity of a \textit{csm3}Δ strain (\textbf{Figure 1B}).

\textbf{\textit{SIR} gene mutations suppress camptothecin hypersensitivity of \textit{tof1}Δ/\textit{csm3}Δ cells}

To understand the origin of the hypersensitivity of \textit{tof1}Δ and \textit{csm3}Δ strains to camptothecin, we carried out a synthetic viability genomic screening (20) to identify mutations capable of suppressing such hypersensitivity (\textbf{Figure 1C}). We isolated sixteen resistant colonies and verified that they indeed displayed both resistance to camptothecin and absence of \textit{TOF1} (\textbf{Figure 1D} and \textbf{Supplementary Figure 1A}). We then sequenced their genomic DNAs to identify candidate mutations responsible for the suppression phenotype (Supplementary Table 1). Two of the sixteen strains – the most resistant ones – carried mutations that inactivated \textit{TOP1}, which encodes the drug target. Three strains carried either of two nonsense mutations that inactivated \textit{SIR3}, while eight of the remaining strains carried a nonsense mutation inactivating \textit{SIR4} (\textbf{Figure 1D}; premature stop codons are designated by a Δ following the position of the last amino acid residue encoded by the truncated gene).

Importantly, by directly introducing deletions of \textit{SIR3} and \textit{SIR4} in \textit{tof1}Δ and \textit{csm3}Δ strains, we verified that \textit{SIR3} or \textit{SIR4} inactivation mediated suppression of camptothecin hypersensitivity (\textbf{Figure 2A}). In the three remaining suppressor strains – the weakest suppressors – we could not identify any mutation responsible for the suppression. In one of these, no mutations were detected, while the other two carried point mutations in \textit{IME2} (Inducer of MEiosis, which is not expressed in exponentially growing cells) or \textit{IRC15}. However, ensuing studies established that neither \textit{IME2} nor \textit{IRC15} deletion suppressed the camptothecin hypersensitivity of \textit{tof1}Δ cells (\textbf{Supplementary Figure 1B and 1C}).

Sir3 and Sir4 form a ternary protein complex with the histone deacetylase catalytic subunit Sir2 (reviewed in (23)), with removal of any of the three subunits inactivating the
transcriptional silencing functions of the complex (24). Significantly, we established that, as for cells lacking Sir3 or Sir4, loss of Sir2 also alleviated the camptothecin hypersensitivity of tof1Δ cells (Figure 2B). Furthermore, by increasing the concentration of camptothecin, we found that deletion of SIR2, SIR3 and SIR4 also promoted camptothecin resistance in a wild-type yeast background (Figure 2B and supplementary Figure 1D). By contrast, SIR2 deletion did not alleviate the strong camptothecin hypersensitivity of a rad51Δ strain, which is defective in DSB repair (Figure 2C). These data thus indicated that the SIR complex affects camptothecin sensitivity only under specific genetic contexts, and that inactivating the SIR complex does not act as a general mediator of camptothecin sensitivity, for example by reducing Top1 activity, increasing cell permeability, or enhancing DNA DSB induction by camptothecin.

**Sir2 mediated deacetylation of H4-K16 imparts camptothecin sensitivity.**

To assess whether loss of the deacetylase activity of the Sir complex was responsible for the suppression of tof1Δ hypersensitivity to camptothecin, we used the small-molecule Sir2 inhibitor, sirtinol (25). This established that addition of 20µM sirtinol suppressed the camptothecin sensitivity of a tof1Δ strain and enhanced the resistance of a wild-type strain (Figure 2D and data not shown). While Sir2 homologs in higher eukayotes have been implicated in the deacetylation of proteins involved in DNA repair, such as PARP1, Ku70 and CtIP (26-28), the prime target for *S. cerevisiae* Sir2 is histone H4 lysine 16 (H4-K16), which is found in an acetylated state through much of the transcriptionally active yeast genome. In *S. cerevisiae*, deacetylation of this residue by Sir2 allows binding of Sir3, thus recruiting further Sir2 that removes acetylation marks from flanking H4-K16 residues, a process that is then propagated to produce a transcriptionally silent heterochromatic state (23). To explore whether the relevant target for Sir2 in relation to its effects on the camptothecin sensitivity of tof1Δ cells was H4-K16, we mutated this residue to glutamine (Q), a residue that mimics a
constitutively acetylated lysine and abrogates Sir3 binding (29). Strikingly, this \textit{hhf-K16Q}
mutation suppressed the camptothecin hypersensitivity of a \textit{tof1Δ} strain, and at higher doses
also reduced the camptothecin sensitivity of a wild-type strain (\textbf{Figure 2E}). Similarly, 
mutation of H4-K16 to glycine (G), which prevents binding by Sir3(29), strongly counteracted 
the camptothecin sensitivity of both \textit{tof1Δ} and wild-type cells. By contrast, mutating histone 
H4-K16 to non-acetylable arginine (R) produced much weaker suppression (\textbf{Figure 2E}). This 
finding was in agreement with published data showing that, despite encoding for a non-
acetylable residue and allowing increased Sir3 binding (29), the \textit{hhf-K16R} mutation actually 
reduces transcriptional silencing (30). Taken together, these results highlighted a correlation 
between loss of silencing and camptothecin resistance.

\textbf{An “acetylated H4-K16” template is responsible for camptothecin induced mitotic arrest.}

The above data supported a model in which the mechanism by which the SIR complex yields 
camptothecin sensitivity is via effects on H4-K16 deacetylation. In this regard, we reasoned 
that the SIR complex might impart camptothecin sensitivity by deacetylating newly 
incorporated histone H4 during DNA replication, or by it promoting a condensed chromatin 
template that impairs DNA replication. To discriminate between these two possibilities, we 
took advantage of the fact that camptothecin treatment of synchronised wild-type cells 
released from G1 into S-phase leads to a prolonged G2/M arrest (31). We first assessed the 
effect of \textit{TOF1} and \textit{CSM3} deletion on this particular phenotype by releasing synchronised 
wild-type, \textit{tof1Δ} and \textit{csm3Δ} cultures either in the presence or in the absence of camptothecin. 
As expected, wild-type cells treated with camptothecin did not delay bulk DNA replication 
compared to strains released in the absence of camptothecin but they did delay exit from the 
subsequent mitosis (\textbf{Figure 3A}). Significantly, compared to wild-type controls, cells deleted 
for \textit{TOF1} or \textit{CSM3} arrested for longer periods of time in G2/M following camptothecin
treatment (Figure 3B), a phenotype that correlated with persistence of the mitotic cyclin Clb2 (Figure 3C). Nevertheless, these cells eventually re-entered the cell cycle and continued proliferating, consistent with the fact that tof1Δ and csm3Δ strains were not killed by acute camptothecin treatment (Figure 3D; note that a repair defective rad51Δ strain was hypersensitive even to acute camptothecin treatment).

If Sir2 deacetylation activity during S phase promoted camptothecin sensitivity, one would expect that addition of sirtinol after the release from G1 would rescue the mitotic delay induced by camptothecin in tof1Δ cells. Conversely, if broad acetylation of the chromatin template was required to rescue the tof1Δ phenotype, sirtinol should lead to suppression only if tof1Δ cells were pre-grown in the presence of sirtinol. To discriminate between these two hypothesis, we grew hml and hmltof1Δ cells either in the presence or in the absence of sirtinol, and we then synchronised them in G1 by addition of alpha-factor (Figure 3E). We used a mutant hml background because sirtinol makes wild-type cells insensitive to alpha-factor by derepressing the HML/R (HM) loci (25, importantly, as shown in Supplementary Figure 2A, HML mutation did not affect camptothecin sensitivity). We then released the G1 synchronised cells into S phase in the presence of camptothecin alone, or in the presence of camptothecin plus sirtinol. Crucially, addition of sirtinol after the G1 release was not sufficient to rescue the mitotic delay of tof1Δ cells (Figure 3E and Supplementary Figure 2B). By contrast, pre-growing tof1Δ cells in the presence of sirtinol fully suppressed their mitotic delay, whether or not sirtinol was present during the subsequent camptothecin treatment (Figure 3E and Supplementary Figure 2B). Collectively, these findings supported a model in which much of the toxicity caused by camptothecin reflects replication-associated problems arising within chromatin regions containing de-acetylated H4-K16, with cells lacking Tof1 or Csm3 being particularly sensitive to this.
HM-like chromatin is responsible for *tof1Δ* strain hypersensitivity to camptothecin.

The yeast genome contains three well-studied heterochromatic regions transcriptionally silenced by SIR proteins: the ribosomal DNA (rDNA) array, sub-telomeric regions and the cryptic mating-type loci (Figure 4A, 4B, and 4C). To establish whether loss of rDNA silencing mediated suppression of *tof1Δ* camptothecin hypersensitivity, we used a strain carrying a deletion of the entire rDNA locus complemented by a multi-copy plasmid containing the rDNA repeat unit (32). We found that deletion of the rDNA locus did not reduce *tof1Δ* hypersensitivity to camptothecin (Figure 4A), suggesting that this genomic region is not the prime target of SIR-mediated silencing that is lethal to *tof1Δ* cells exposed to camptothecin.

This notion was also supported by the fact that, while we observed suppression of camptothecin sensitivity with *sir2Δ, sir3Δ* or *sir4Δ*, silencing of the rDNA locus only requires Sir2, with *SIR4* deletion actually increasing rDNA silencing by delocalising Sir2 from telomeres (33).

To determine if loss of sub-telomeric silencing could rescue *tof1Δ* hypersensitivity to camptothecin, we employed a strain carrying a C-terminal truncation of Rap1 (*rap1Δ663*), the so-called *rap1-17* allele. This mutation completely disrupts transcriptional silencing at telomeres (telomere position effect) and partially affects silencing of the cryptic mating-type locus *HML* but not of that of *HMR* (34). While strains carrying the *rap1Δ663* allele grew slower than wild-type strains, presumably due to the role of Rap1 in regulating transcription of genes involved in ribosome formation and glycolysis (35, 36), they did not display altered sensitivity to camptothecin (Figure 4B). Notably, the *rap1Δ663* mutation also failed to suppress the camptothecin hypersensitivity of *tof1Δ* cells (Figure 4B), indicating that loss of telomere position effect does not promote survival in the presence of this drug.
At the cryptic mating type loci *HML* and *HMR*, silencing is established by replication origin recognition complex (ORC)-mediated recruitment of Sir1, which then attracts the SIR complex via an interaction with Sir4 (37, 38). Sir4 binding is also stabilised by an interaction with Rap1, which binds to its DNA consensus sequence located next to the ORC binding site (ACS, ARS consensus sequence, **Figure 4C**). For these reasons, deletion of *SIR1* results in partial loss of silencing at the cryptic mating-type loci, but does not affect telomeric or rDNA silencing (39). Strikingly, we found that *SIR1* deletion strongly alleviated the camptothecin hypersensitivity of a *tof1Δ* strain (**Figure 4C**). Similarly to what we had observed for *SIR2* deletion, disruption of *SIR1* also decreased the sensitivity of a wild-type strain to high levels of camptothecin but did not rescue the camptothecin hypersensitivity of a *rad51Δ* strain (**Figure 4C and Supplementary Figure 2C**). These data indicated that the de-acetylated H4-K16 bearing chromatin template that is toxic to *tof1Δ* and wild-type cells in the presence of camptothecin is generated in a Sir1-dependent manner, and were also consistent with our conclusions that camptothecin sensitivity is not mainly generated via the rDNA or telomeric loci.

**Tof1/Csm3 prevents loss of epigenetic information during DNA replication.**

Rather than being required to maintain epigenetic silencing, Sir1 re-establishes silent chromatin when it happens to be lost; and thus, within a population of *sir1Δ* cells, only a fraction has lost silencing (19, 24, 40). In spite of this, we noted that *SIR1* deletion suppressed *tof1Δ* hypersensitivity to essentially the same extent as *SIR2* deletion. To explain this apparent paradox, we hypothesised that *tof1Δ* cells might lose silencing more frequently than wild-type cells, and may thus require Sir1 to re-establish it. To test this hypothesis, we took advantage of the fact that co-expression of the two *HM* loci results in sterility (41) and used mating assays to measure the extent of silencing loss in wild-type, *sir1Δ, tof1Δ* and *sir1Δtof1Δ* strains.
This revealed that, under our experimental conditions, sir1Δ strains did not show a detectable mating defect, but deletion of TOF1 reduced the ability of sir1Δ cells to mate. (Figure 4D). Additionally, like sir2Δ cells, sir1Δtof1Δ cells failed to arrest in G1 in the presence of alpha-factor, despite the sir1Δ or tof1Δ single mutants being proficient in this assay (Figure 4E).

To directly assess loss of silencing at the HMR locus, we employed an experimental system designed to trap transient loss-of-silencing events (19). Briefly, we used a strain in which loss of silencing induces expression of the Cre recombinase integrated at the HMR locus. Cre then excises a fragment of DNA carrying genes encoding for red-fluorescent protein (RFP), expressed from the constitutive GPD promoter, and resistance to the antibiotic G418. This excision juxtaposes the GPD promoter to the gene coding for green-fluorescent protein (GFP), resulting in cells switching from red to green fluorescence (as well as from G418 sensitivity to resistance, Figure 4F). We grew cultures of wild-type, tof1Δ, sir1Δ, and tof1Δsir1Δ strains in the presence of G418 to prevent expansion of green clones and then plated them to obtain single colonies, which were scored for the presence of red/green sectors. The majority of colonies formed by wild-type and tof1Δ cells were either completely red or had very small sectors/dots of GFP signal, with tof1Δ colonies showing a larger proportion of the latter (Figure 4G). In agreement with previous results, most sir1Δ colonies had large green sectors and many of them were mainly or completely green, indicating prevalent loss of HMR silencing (19). Strikingly, all colonies of the double mutant sir1Δtof1Δ were completely green, highlighting extensive loss of HMR silencing (Figure 4G; we noticed that, while most colonies were fully green, small patches of red fluorescence could be detected in some of them, indicating that the cell that started the colony was originally red). Collectively, these findings supported a model in which replication in the absence of the Tof1/Csm3 complex strictly requires Sir1 for silencing maintenance, possibly because chromosome hyper-catenation created in the absence of Tof1 alters the dynamics of histone deposition and favours loss of HMR-like silencing.
Various SIR-bound genomic regions mediate camptothecin sensitivity.

To establish whether loss of H4-K16Ac or the associated leak of genetic information from HML and HMR was responsible for the suppression of tof1Δ phenotypes by SIR complex loss, we analysed the sensitivity of diploid tof1Δ/tof1Δ cells, which express simultaneously, at the MAT locus, the genetic information encoded by HMR and HML. If a leak of HML genetic information reduced the camptothecin hypersensitivity of MATa tof1Δ strains, one would expect a homozygous tof1Δ diploid strain to be less camptothecin sensitive than the corresponding haploid strain; however, this was not the case (Figure 5A). Moreover, the hypersensitivity of diploid tof1Δ/tof1Δ cells was also rescued by sirtinol, indicating that the loss of heterochromatin structure rather than leaked HM genetic information is responsible for suppression of camptothecin hypersensitivity (Figure 5A). However, when we then deleted the HML and HMR loci, we were surprised to observe that this did not rescue the camptothecin hypersensitivity of tof1Δ cells (Figure 5B), suggesting the existence of other genomic loci targeted by the Sir1/2/3/4 pathway.

To identify such genomic regions, we analysed a dataset of chromatin immunoprecipitation-sequencing (ChIP-seq) data for Sir2, Sir3, Sir4, GFP, acetylated histone H4-K16, and histone H3 (42, 43). In these datasets, we searched for genomic regions displaying increased binding of Sir2, Sir3 and Sir4 compared to neighbouring regions, even below the levels of statistical significance. We then removed any region that showed increased GFP binding to exclude ChIP bias towards highly expressed genes (43). We also removed any region where we could not observe a decrease in H4-K16 acetylation, the functional consequence Sir complex binding, or where such a decrease co-localised with loss of the H3 ChIP signal, suggesting depletion of nucleosomes. Strikingly, genomic regions identified in this manner co-localised with confirmed open reading frames (ORFs; three examples of which are shown in Figure 5C). We then defined a “SIR-binding score” for every ORF as the fraction of nucleotides for which the
above conditions held. While the majority of all ORFs had a null SIR score (indicative of no enrichment of SIR complex binding), we found that 111 of them showed an enrichment of Sir2/3/4 and concomitant loss of H4-K16 acetylation in at least 20% of their sequence (Supplementary Table 4). Of these 111 ORFs, 29 were localised in sub-telomeric regions or in regions proximal to the HM loci (Figure 5D, small grey dots), while the remaining 82 hits were positioned along chromosome lengths (Figure 5D, green dots). While the majority of the identified ORFs are expressed at high levels during exponential growth, high expression was not sufficient for a high SIR score (Supplementary Figure 2D based on data from (44)). Collectively, these findings highlighted how, in addition to functioning at its well-defined target loci, the SIR complex may also act at a variety of loci scattered throughout the genome, and suggested that these loci might also promote camptothecin toxicity in wild-type and tofΔ cells.

Recruitment of Sir1 at HM loci requires its interaction with the bromo-adjacent domain (BAH) region of Orc1. We therefore asked whether any of the loci we identified above was also positioned in proximity to a site bound by ORC. To do this, we calculated the distance between the centre of each ORF and the nearest ORC binding site (45). This analysis revealed that ~50% of SIR-positive ORFs were located less than 1.7 kbp from a site of ORC binding (Figure 5E). This distance is smaller than the median value of 7.7 kbp for all yeast ORF. We reasoned that, if ORC has a functional role in recruiting the SIR complex to these genomic loci, it should be possible to suppress the camptothecin hypersensitivity of tofΔ by preventing ORC-mediated recruitment of Sir1. In line with this hypothesis, we found that deleting the BAH domain of ORC1 suppressed tofΔ camptothecin hypersensitivity (Figure 5F; effects of ORC1 deletion could not be studied because it is an essential gene). These findings thus suggested that the chromatin substrates that become toxic to tofΔ cells exposed to camptothecin is at least partially formed in an ORC-dependent manner.
Discussion

We identified the Sir1, 2, 3 and 4 (SIR) genes as major mediators of the sensitivity of both wild-type and \textit{tof1Δ} cells to camptothecin. Furthermore, we established that, rather than merely reducing camptothecin action, deletion of SIR genes removes a factor that hinders cell proliferation in the presence of camptothecin in wild-type cells and that is particularly toxic to cells lacking the Tof1 replication pausing complex. Camptothecin promotes the accumulation of positive supercoiling during DNA replication by locking Topoisomerase 1 on DNA in a non-functional state (16, 17). Since Tof1/Csm3 restricts replisome rotation during DNA replication (6) and since the main force driving fork rotation is positive supercoiling, we hypothesize that an excess of positive supercoiling is the factor that is alleviated by deletion of SIR genes. Lack of Sir2, Sir3, or Sir4 leads to loss of histone H4 lysine 16 (H4-K16) deacetylation and subsequent impairment in heterochromatin formation. We have observed that inhibition of Sir2 deacetylase activity or mutation of H4-K16 to glutamine — a residue that mimics an acetylated lysine — also suppresses \textit{tof1Δ} camptothecin sensitivity. Importantly, this suppression is observed only if Sir2 activity is inhibited prior to camptothecin treatment, suggesting that a heterochromatic template becomes toxic to \textit{tof1Δ} cells when replicated in the presence of camptothecin.

Yeast genomes contain three well-characterized regions of transcriptionally silenced chromatin, namely the ribosomal DNA, sub-telomeric regions and the cryptic mating-type loci \textit{HML} and \textit{HMR}; and of these, only the cryptic mating type loci require Sir1 for their silencing (39). The fact that \textit{SIR1} deletion also suppresses the camptothecin sensitivity of \textit{tof1Δ} cells initially suggested to us that \textit{HML} and \textit{HMR} represent the chromatin templates that are toxic to \textit{tof1Δ} cells. However, we did not observe a reduction in \textit{tof1Δ} sensitivity to camptothecin by deleting \textit{HML} and \textit{HMR}, meaning that these two genomic loci alone are not responsible for the strong camptothecin sensitivity phenotype displayed by \textit{tof1Δ} cells.
By analysing publicly available ChIP-seq data, we identified various genomic loci that exhibit enhanced localisation of Sir2, Sir3 and Sir4 as well as H4-K16 under-acetylation. Notably, we found that these genomic loci co-localise with confirmed ORFs and are located closer to sites of ORC binding than the average yeast ORF. Indeed, we found that many of these sites co-localise with genomic loci that were previously shown to bind ORC despite not having origin activity (45). Importantly, some of the SIR-enriched loci also co-localise with sites of replication fork pausing and sites enriched in binding of Rrm3, a DNA helicase that relieves replication fork pauses (46, 47), suggesting that SIR-enriched loci are inherently difficult to replicate even in the absence of camptothecin. The fact that these ORFs are amongst the most highly expressed yeast genes and yet exhibit enhanced recruitment of the SIR silencing complex and signs of histone de-acetylation is enigmatic. One possibility is that strong transcription could prevent heterochromatin formation despite presence of the SIR complex. Indeed it has been shown that promoter strength affects the efficiency of silencing (48). By affecting transcription, camptothecin, could stimulate a temporary heterochromatinization of these genes, creating topological barriers to DNA replication. In this regard the hypersensitivity of tof1Δ cells to camptothecin might stem from the hyper-catenation that is generated when replication forks lacking Tof1/Csm3 approach barriers created by the Sir2/3/4 complex. In this regard, we note that increased catenation would likely require more time to be resolved, thereby potentially accounting for the M/G1 delay observed in tof1Δ cells following camptothecin treatment.

Lack of Sir1 does not lead automatically to loss of transcriptional silencing, but rather it removes the pathway required for its re-establishment after it is lost. We were thus initially surprised that SIR1 deletion also produced a strong suppression of camptothecin sensitivity, similar to that we observed with SIR2 deletion. To explain this unexpected result, we hypothesised that increased DNA catenation caused by loss of Tof1 might increase the frequency of silencing loss. Accordingly, we observed that sir1Δtof1Δ cells show phenotypes
that are consistent with loss of silencing at cryptic mating type loci. It is difficult to imagine how loss of Tof1 might lead to loss of silencing, but one possibility is that hyper-catenation of sister chromatids generated in the absence of Tof1 could transiently impair normal histone deposition/recycling, thereby promoting loss of parental heterochromatic marks. In regard to this, we note that yeast strains lacking histone chaperones Asf1 or CAF-1 also lose HML silencing in the absence of Sir1 (49, 50). Despite the threat to genome integrity, loss of nucleosomes on intertwined DNA strands might also represent a signal for stimulating Top2 decatenating activity, as it is suggested by the fact that nucleosome loss increases Top2 occupancy (51).

Inhibition of topoisomerase 1 is a widely used therapeutic strategy to selectively kill proliferating cancer cells, with camptothecin analogues being part of the standard-of-care provided by many cancer clinics worldwide. Various mechanisms of camptothecin resistance have been observed, ranging from overexpression of drug-efflux transporters that actively reduce intracellular drug concentration (52) to specific Top1 mutations that prevent its interaction with camptothecin (53, 54). Using yeast as a model system, we have found that inhibition of H4-K16 deacetylation by inactivation of the Sir2/3/4 complex represents an additional mechanism of camptothecin resistance. Further studies will be required to determine if this mechanism is evolutionary conserved and whether it plays a significant role in the emergence of resistance to camptothecin analogues in human cancers.

**Supplementary Data**

This manuscript contains two supplementary figures and four supplementary tables.

**Funding**

This work was supported by the European Molecular Biology Organization [ALTF1287-2011 to FP]; the Associazione Italiana per la Ricerca sul Cancro [to VC, 10932 to FP]; Cancer...
Research UK [C6/A11224, C6/A18796, C6946/A14492 (core funding)]; the European Research Council [DDREAM to SP], DNAMEREP to VC; the European Community Seventh Framework Programme [HEALTH-F2-2010-259893 (DDResponse)]; the Wellcome Trust [101126/Z/13/Z, 098051 to M.H. and WT092096 (core funding)]; the Armenian-Harvard Foundation [to VC]; and the Fondazione Telethon [to VC].

Conflict of interest statement. None declared

Acknowledgements

We thank Jasper Rine, Catherine A. Fox, Takehiko Kobayashi, Tony Kouzarides and Shelley Berger for the gift of strains and plasmids, all the members of the SP laboratory for helpful discussions, and Helena Santos-Rosa for advice with designing histone point mutations. The initial project was conceived by FP, VC and SP. Screening for suppressors and DNA extractions were carried out by FP and NJG. Analysis of whole genome sequencing data was carried out by MH and FP. Subsequent in vivo experiments and analyses of ChIP-seq data were designed and carried out by FP. The manuscript was largely written by FP and SP, with contributions made by all the other authors.

References


953–961.


**Figure Legends**

**Figure 1.** A synthetic viability screening to identify the cause for the hypersensitivity of tof1Δ cells to camptothecin

(A) Loss of Tof1 and Csm3 but not Mrc1 causes hypersensitivity to camptothecin in a Top1-dependent manner. (B) Loss of pausing at the replication fork barrier on rDNA does not cause camptothecin hypersensitivity. (C) Outline of the procedure for a synthetic viability screen.
(D) Synthetic viability screening identifies sir3 and sir4 alleles as suppressors of the camptothecin hypersensitivity of tof1Δ strains.

**Figure 2. Loss of the SIR complex suppresses camptothecin hypersensitivity of tof1Δ strains.**

(A) Deletion of SIR3 or SIR4 suppresses the hypersensitivity of tof1Δ cells to camptothecin. (B) Deletion of SIR2 also suppresses the hypersensitivity of tof1Δ cells to camptothecin and reduces the sensitivity of a wild-type strain. (C) Deletion of SIR2 cannot suppress the camptothecin hypersensitivity of a rad51Δ strain. (D) Inhibition of Sir2 deacetylase activity with sirtinol suppresses the hypersensitivity of tof1Δ cells to camptothecin. (E) Mutations that mimic a permanently acetylated H4-K16 (K16Q) or that remove the binding site for Sir3 (K16G) also suppress the sensitivity to camptothecin of wild-type and tof1Δ strains. Mutation K16R (non-acetylable residue) yields a less strong suppression, in line with reports that this mutation partially impairs silencing.

**Figure 3. An “acetylated H4-K16” template mediates sensitivity to camptothecin during DNA replication.**

(A) A wild-type strain released into S phase in the presence of camptothecin does not delay progression through S phase, but delays progression through the subsequent mitosis. (B) In the absence of Tof1 or Csm3, camptothecin treated cells remain arrested in G2/M for longer periods of time than wild-type cells. (C) tof1Δ and csm3Δ cells released into S phase in the presence of camptothecin delay destruction of the mitotic cyclin Clb2. (D) tof1Δ and csm3Δ cells are not hypersensitive to a pulse treatment with camptothecin. (E) tof1Δ cells and congenic wild-type cells were pre-grown either in the absence or in the presence of sirtinol. They were subsequently synchronised in G1 and released into S phase in the presence of camptothecin, either with or without sirtinol. Cell cycle progression was monitored by FACS
analysis. Quantification of G1 cells shows that sirtinol addition during camptothecin treatment
does not suppress the mitotic delay of \textit{tof1}Δ cells, while pre-growth in the presence of sirtinol
is sufficient to suppresses the camptothecin hypersensitivity phenotype of \textit{tof1}Δ cells.

Figure 4. Disruption of \textit{SIR1} suppresses camptothecin sensitivity in wild-type and \textit{tof1}Δ
cells.

(A) Deletion of the rDNA locus is not sufficient to suppress the hypersensitivity of \textit{tof1}Δ cells
to camptothecin. (B) A mutation in \textit{RAP1} that disrupts telomeric silencing does not suppress
the hypersensitivity of \textit{tof1}Δ cells to camptothecin. (C) Deletion of \textit{SIR1} suppresses
camptothecin sensitivity in wild-type and \textit{tof1}Δ cells. (D) \textit{sir1}Δ and \textit{tof1}Δ show a synergistic
defect in the ability to mate. (E) \textit{tof1}Δ\textit{sir1}Δ cells are unable to arrest in G1 after exposure to
alpha factor. (F) Outline of the genetic system used to detect loss of silencing events at the
\textit{HMR} locus: transient loss of silencing causes expression of the Cre recombinase and a switch
from RFP and KanMX expression to GFP expression. (G) Cells carrying the genetic reporter
described in (F) were grown in the presence of G418 to prevent expansion of the green clones,
and were then plated on YPD plates. A selection of representative colonies is shown. A
quantification of red, green and sectored colonies is shown on the right.

Figure 5. Disruption of ORC1-mediated binding of the SIR complex to highly transcribed
genes suppresses the hypersensitivity of \textit{tof1}Δ cells to camptothecin.

(A) Homozygous \textit{tof1}Δ/\textit{tof1}Δ diploid cells are as sensitive to camptothecin as \textit{tof1}Δ haploids
and their hypersensitivity can be rescued by sirtinol. (B) Deletion of \textit{HML} and \textit{HMR} cannot
suppress the camptothecin hypersensitivity of \textit{tof1}Δ strains. (C) Analysis of ChIP-seq data for
the indicated proteins. In green is the protein tested; in grey are the controls. Input samples
are shown in darker green/grey and immunoprecipitated samples are shown in lighter
green/grey. The position of each ORF is indicated by a black bar. (D) Identification of regions
bound by the SIR complex: for each ORF in the genome, a “SIR score” was calculated as the
fraction of the ORF for which both increased Sir2, Sir3, Sir4, and decreased H4-K16ac was
observable. ORFs were sorted based on their “SIR score”. Sub-telomeric ORFs and ORFs
proximal to HML and HMR are shown with small grey dots, while remaining ORFs are shown
with large green dots. (E) SIR-positive ORFs are on average located closer to sites of ORC
binding than ORFs in general. All yeast ORFs are shown in purple as a function of their
distance from the nearest site of ORC binding. SIR-positive ORFs (SIR score >0.2), are shown
in green. (F) Deletion of the BAH domain of ORC1 partially rescues the camptothecin
hypersensitivity of tof1Δ cells.
Figure 1

A

B

C

D

Figure 1

For Peer Review
Supplementary Figure 2

(A) wt, tof1Δ, tof1Δser3Δ, tof1Δser4Δ, csm3Δ, csm3Δser3Δ, csm3Δser4Δ

(B) wt, tof1Δ, ser2Δ, tof1Δser2Δ

(C) wt, red51Δ, ser2Δ, red51Δser2Δ

(D) wt, tof1Δ, mrc1Δ, tof1Δ, tof1Δshh-K16Q

(E) wt, tof1Δ, hsf-K16R, tof1Δshh-K16R, hsf-K16G, tof1Δshh-K16G
Figure 4
Supplementary Material

Supplementary Table 1: Yeast strains used in this study.

Supplementary Table 2: Whole-genome sequencing data.

Supplementary Table 3: SRA accession numbers for ChIP-seq data

Supplementary Table 4: ORFs with SIR score > 0.2

Supplementary Figure 1
(A) Suppressor strains recovered from the tof1Δ synthetic-viability screen are G418 resistant, indicating presence of the TOF1 deletion cassette. (B) Deletion of IRC15 does not suppress tof1Δ camptothecin hypersensitivity. (C) Deletion of IME2 does not suppress tof1Δ camptothecin hypersensitivity. (D) Deletion of SIR2, SIR3 or SIR4 increases resistance to camptothecin in a wild-type background.

Supplementary Figure 2
(A) Deletion of HML does not rescue tof1Δ hypersensitivity to camptothecin. (B) tof1Δ cells and congenic wild-type cells were pre-grown either in the absence or in the presence of sirtinol. They were subsequently synchronised in G1 and released into S phase in the presence of camptothecin, either with or without sirtinol. Cell cycle progression was monitored by FACS analysis. (C) Deletion of SIR1 does not rescue camptothecin hypersensitivity in rad51Δ cells. (D) The majority of SIR-positive ORFs are highly expressed genes, but high expression does not necessarily correlate with high SIR score.
Supplementary Figure 1

A

G418 (tof1Δ:KanMX6)

tof1Δ

tof1Δsup1

tof1Δsup2

tof1Δsup3

tof1Δsup4

tof1Δsup5

wt
tof1Δ

tof1Δsup6

tof1Δsup7

tof1Δsup8

tof1Δsup9

tof1Δsup10

tof1Δsup11

wt
tof1Δ

tof1Δsup12

tof1Δsup13

tof1Δsup14

tof1Δsup15

tof1Δsup16

wt
tof1Δ

B

Mock 20μM CPT

C

Mock 20μM CPT

D

Mock 40μM CPT

For PBA-40

sir2Δ

sir3Δ

sir4Δ