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."

Abstract

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

List of Figures

List of Tables

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]:

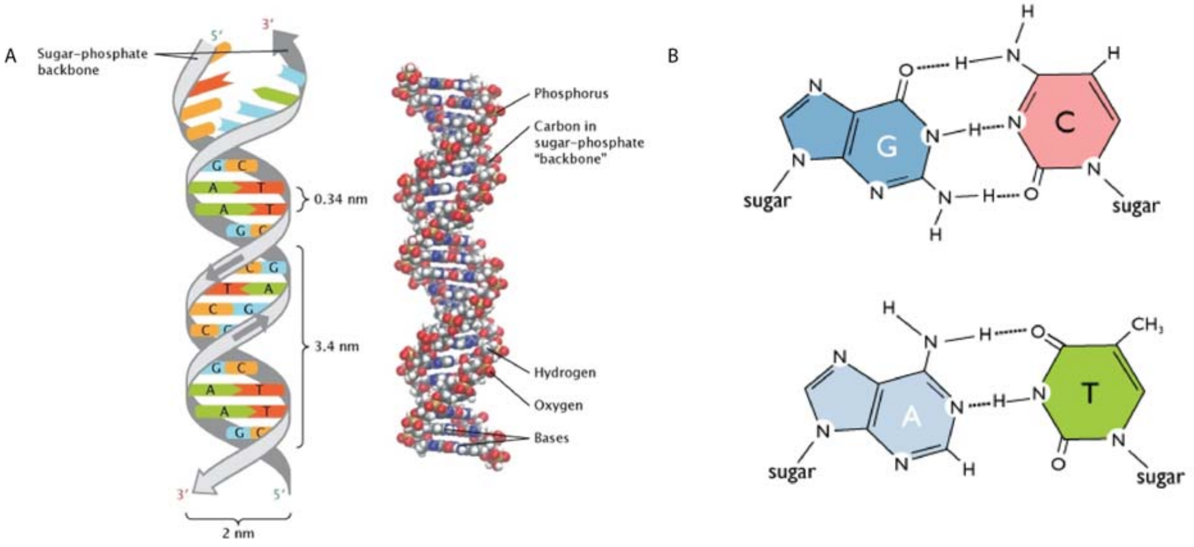
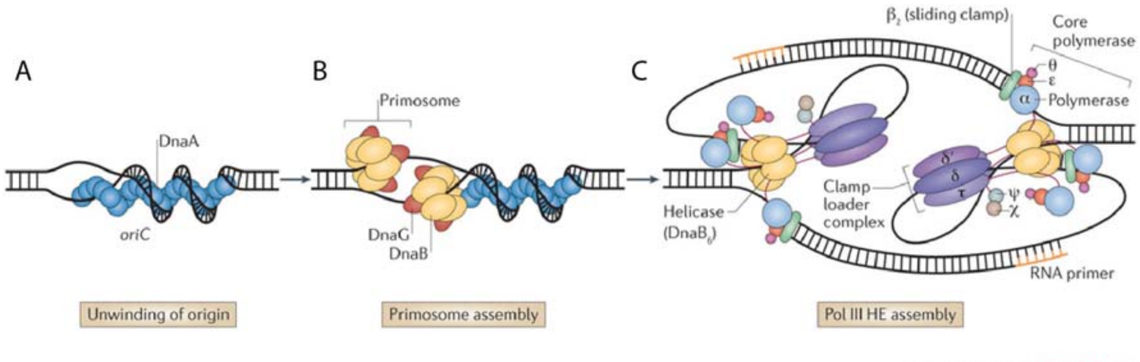


Figure 1.1: Structure of DNA
A - Structure of the DNA double helix. Reproduced from [23] in accordance with the publisher’s terms of use. B - Pairing of the 4 DNA bases: adenine forms hydrogen bonds with thymine, guanine bonds with cytosine. From [24].

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



Nature Reviews | Microbiology

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 ($\beta2$ clamp) and a clamp loader. The ring-shaped $\beta2$ 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 $\beta2$ 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 $\beta2$ 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

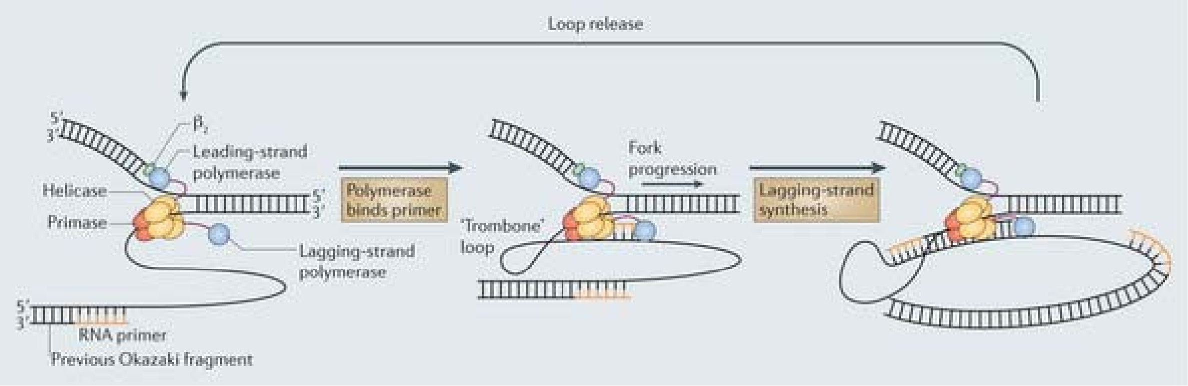


Figure 1.3: Lagging strand DNA synthesis in *E. coli*
Model of lagging strand DNA synthesis coordination: lagging strand polymerase action is thought to generate a "Trombone" loop to accommodate both polymerases and their opposing movement at the replication fork. Reproduced from [37] with permission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

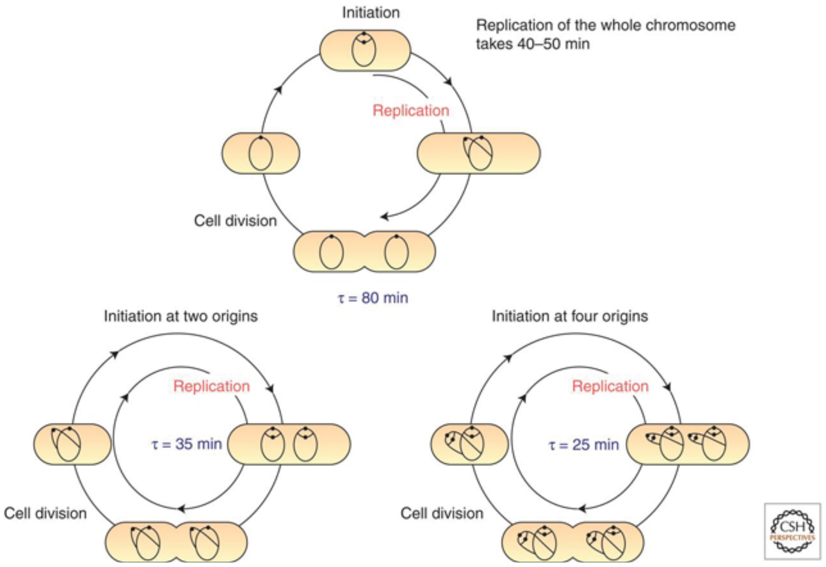


Figure 1.4: Overlapping replication cycles in *E. coli*
Depending on growth conditions population doubling time in *E. coli* can be shorter than the time required to replicate the chromosome due to re-initiation before cell division. Reproduced from [53] with permission from the publisher. Copyright (2013) Cold Spring Harbor Laboratory Press

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]).

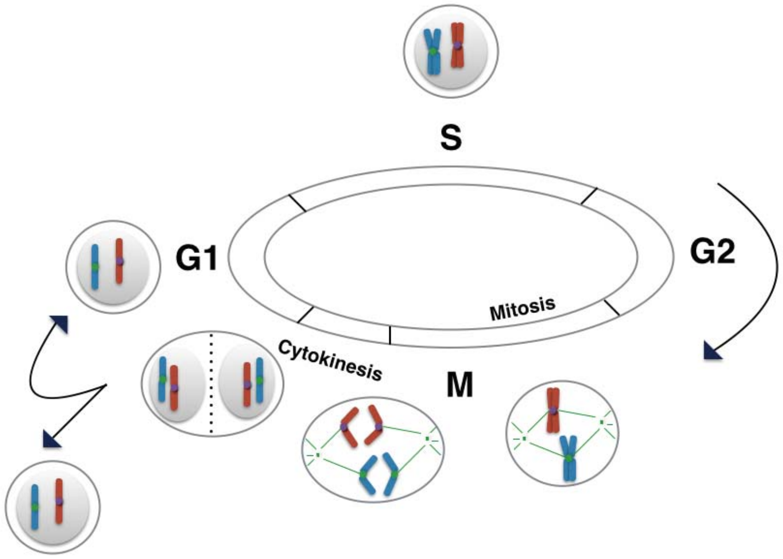


Figure 1.5: The eukaryotic cell cycle
A schematic of the eukaryotic cell cycle. S-phase (replication) and M-phase (mitosis followed by cytokinesis) alternate and are separated by two growth phases (G1 and G2).

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, epigentic 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 epigentic 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[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[101–104] and *Xenopus* DNA seems to have little or no sequence requirement for replication initiation *in vitro*[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[88, 106–112].

Replication Licensing Preventing re-replication is a two step process in eukaryotes[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[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 [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[117]. In vertebrates, ORC is potentially targeted to origins by HMGA1a which contains the AT-hook motif[84]. In humans, CDC6 is recruited to ORC by MCM8 which interacts with ORC2 and CDC6[118]. Another MCM family member, MCM9, binds Cdt1 directly and promotes recruitment of the MCM2-7 replicative helicase complex[119–130]. Loading of the MCM2-7 is stimulated by ORC and CDC6 ATPases activities[131, 132]. The MCM complex is loaded directly onto DNA and forms a double hexamer[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[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 [114, 139, 140]. Activation of the helicase requires the

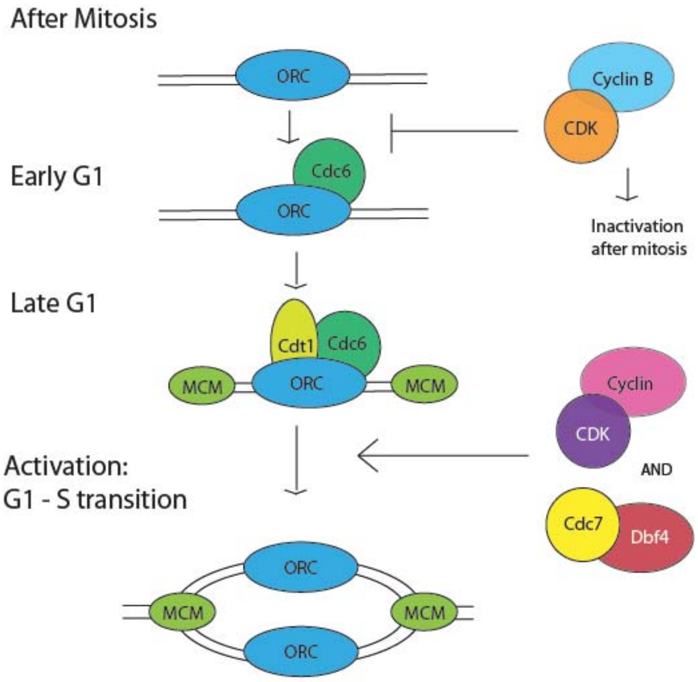


Figure 1.6: Licensing of eukaryotic origins of replication
A schematic detailing the ordered assembly of licensing factors on eukaryotic origins starting with the Origin Recognition Complex (ORC). Until the end of mitosis licensing is prevented by actions of cyclin-CDKs. After anaphase the licensing 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$^{\mathrm{Cdc4}}$ [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$^{\mathrm{Cdc4}}$-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

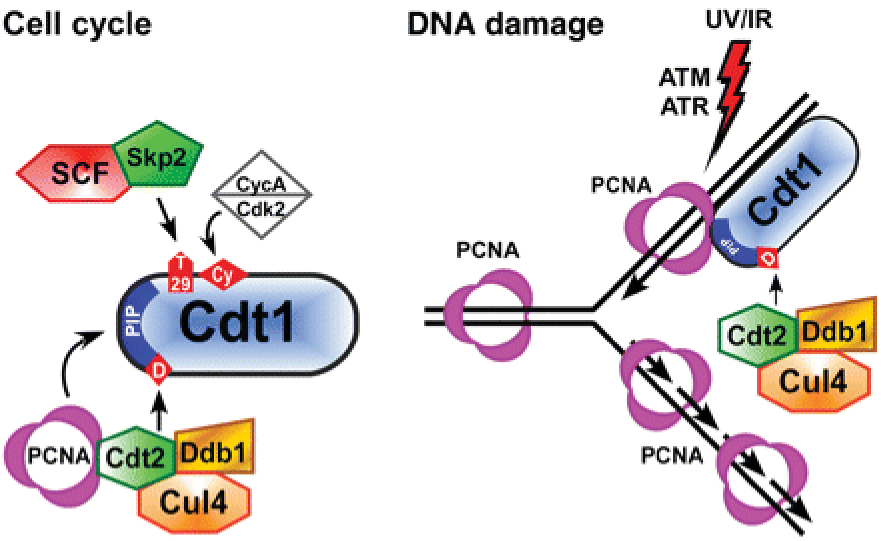


Figure 1.7: Degradation of Cdt1 during the cell cycle and in response to DNA damage During S phase, Ctd1 is targeted for degradation by the SCF–Skp2 and by the PCNA–Cul4–Ddb1–Cdt2 pathways. Cyclin-CDK activity results in the phosphorylation of Cdt1, which in turn allows recruitment of SCF–Skp2, an E3 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$^{\mathrm{Cdh1}}$, 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-

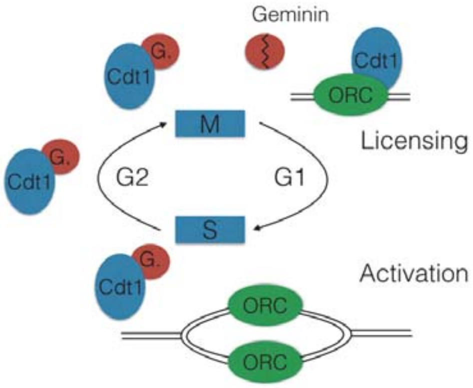


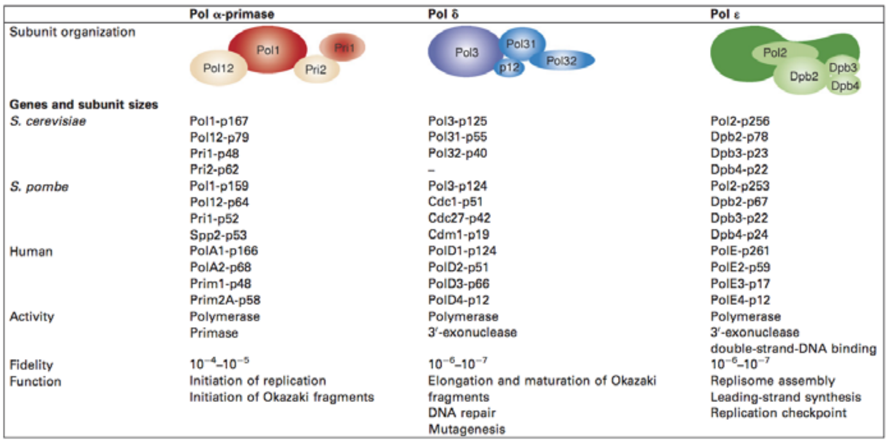
Figure 1.8: Regulation of Cdt1 by association with Geminin
Geminin regulates Cdt1 by binding and sequestering it during S and G2-phase. Its degradation after mitosis allows Cdt1 participation in origin licensing in early G1 phase.

chiometry of the Geminin-Cdt1 complex. A "permissive" heterotrimer is thought to promote Cdt1-mediated Mcm2–7 loading in G1, while an "inhibitory" heterohexamer is likely sequestering Cdt1[244–246].

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



	Pol α-primase	Pol δ	Pol ε
Subunit organization	Pol1 Pri1 Pol12 Pri2	Pol3 Pol31 p12 Pol32	Pol2 Dpb2 Dpb3 Dpb4
Genes and subunit sizes			
S. cerevisiae	Pol1-p167 Pol12-p79 Pri1-p48 Pri2-p62	Pol3-p125 Pol31-p55 Pol32-p40 –	Pol2-p256 Dpb2-p78 Dpb3-p23 Dpb4-p22
S. pombe	Pol1-p159 Pol12-p64 Pri1-p52 Spp2-p53	Pol3-p124 Cdc1-p51 Cdc27-p42 Cdm1-p19	Pol2-p253 Dpb2-p67 Dpb3-p22 Dpb4-p24
Human	PolA1-p166 PolA2-p68 Prim1-p48 Prim2A-p58	PolD1-p124 PolD2-p51 PolD3-p66 PolD4-p12	PolE-p261 PolE2-p59 PolE3-p17 PolE4-p12
Activity	Polymerase Primase	Polymerase 3′-exonuclease	Polymerase 3′-exonuclease double-strand-DNA binding
Fidelity	10^{-4}–10^{-5}	10^{-6}–10^{-7}	10^{-6}–10^{-7}
Function	Initiation of replication Initiation of Okazaki fragments	Elongation and maturation of Okazaki fragments DNA repair Mutagenesis	Replisome assembly Leading-strand synthesis Replication checkpoint

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' exonucleolytic 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 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 heterotetramer 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 subunit, 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

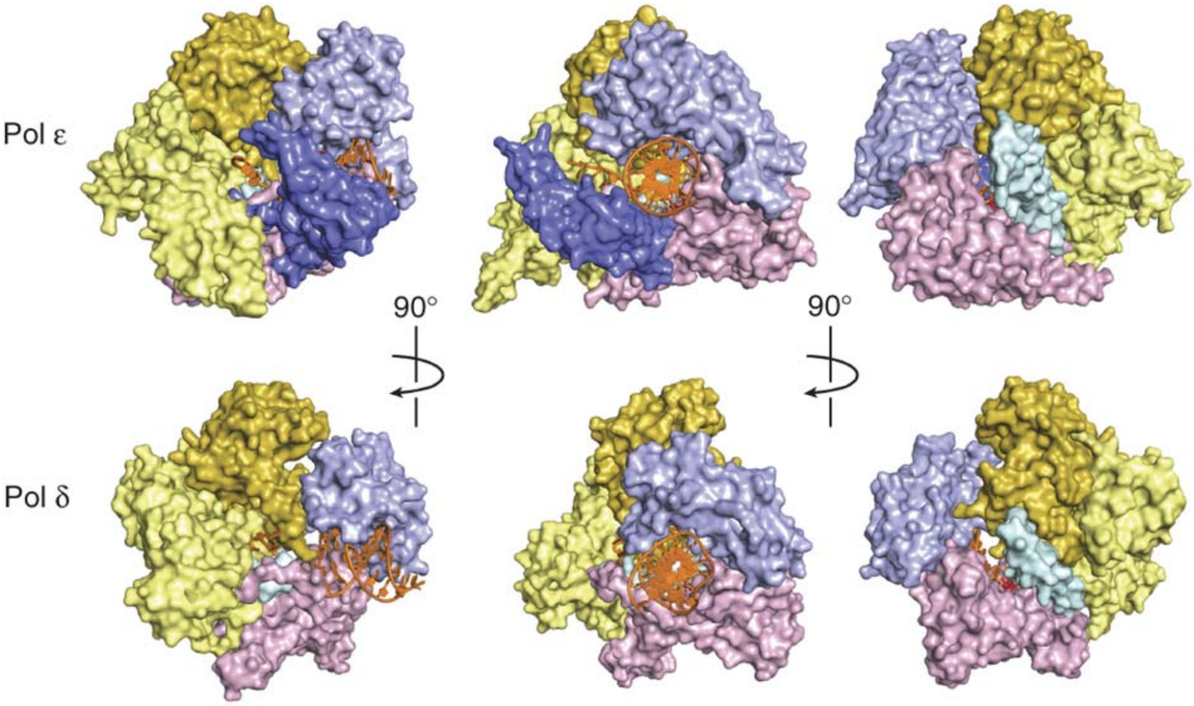


Figure 1.9: Structure of DNA polymerase δ and DNA polymerase ε
Surface representations of Pol δ (PDB 3IAY20) and Pol ε: The fingers (cyan), palm (pink), thumb (light blue), exonuclease (gold) and N-terminal (light yellow) domains are shown in three orientations. In the case of Polε, the additional, newly discovered P domain (dark blue) is shown. Reproduced from [296] with permission from the publisher.

fragments it is nonetheless crucial for cell survival[295].

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].

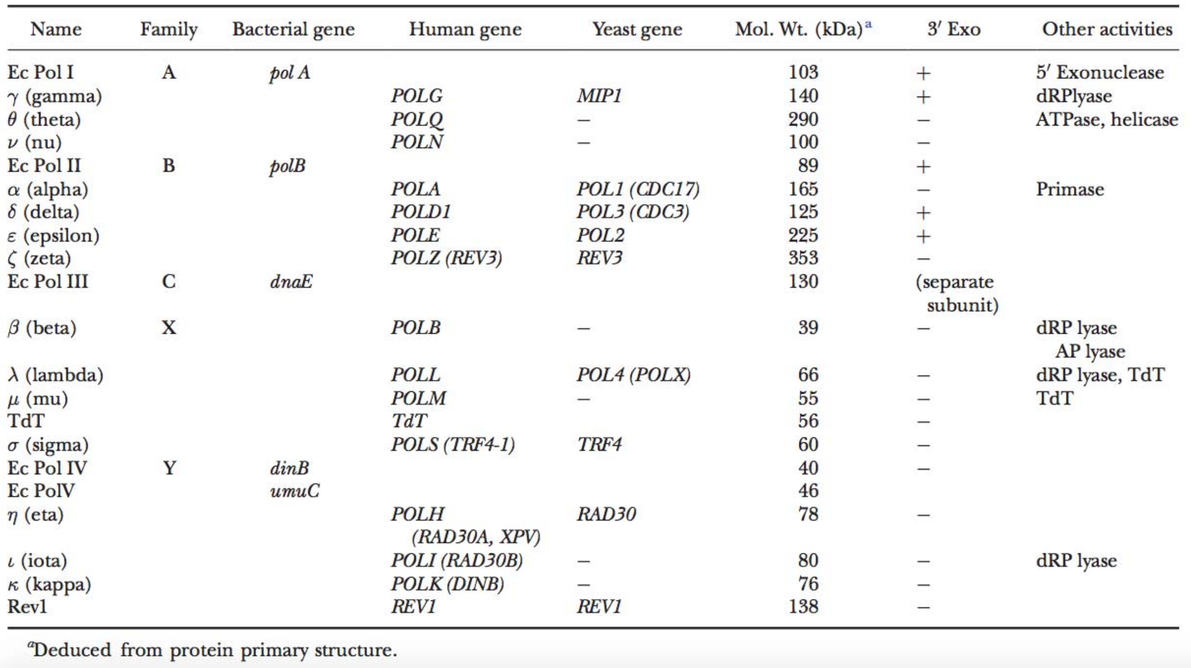
Considering the evidence for Pol δ as the lagging strand polymerase, this would place Pol ε on the leading strand. However, this does not elucidate how much Pol δ contributes to leading strand replication. In fact, Pol δ could well replicate the vast majority of leading strand, which is also supported 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.

1.1.1.4 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 catalyze 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].



Name	Family	Bacterial gene	Human gene	Yeast gene	Mol. Wt. (kDa)[a]	3′ Exo	Other activities
Ec Pol I	A	*pol A*			103	+	5′ Exonuclease
γ (gamma)			*POLG*	*MIP1*	140	+	dRPlyase
θ (theta)			*POLQ*	–	290	–	ATPase, helicase
ν (nu)			*POLN*	–	100	–	
Ec Pol II	B	*polB*			89	+	
α (alpha)			*POLA*	*POL1 (CDC17)*	165	–	Primase
δ (delta)			*POLD1*	*POL3 (CDC3)*	125	+	
ε (epsilon)			*POLE*	*POL2*	225	+	
ζ (zeta)			*POLZ (REV3)*	*REV3*	353	–	
Ec Pol III	C	*dnaE*			130	(separate subunit)	
β (beta)	X		*POLB*	–	39	–	dRP lyase AP lyase
λ (lambda)			*POLL*	*POL4 (POLX)*	66	–	dRP lyase, TdT
μ (mu)			*POLM*	–	55	–	TdT
TdT			*TdT*		56	–	
σ (sigma)			*POLS (TRF4-1)*	*TRF4*	60	–	
Ec Pol IV	Y	*dinB*			40	–	
Ec PolV		*umuC*			46		
η (eta)			*POLH (RAD30A, XPV)*	*RAD30*	78	–	
ι (iota)			*POLI (RAD30B)*	–	80	–	dRP lyase
κ (kappa)			*POLK (DINB)*	–	76	–	
Rev1			*REV1*	*REV1*	138	–	

[a]Deduced from protein primary structure.

Table 1.2: Families of DNA polymerases
Reproduced from [324] with permission from the publisher.

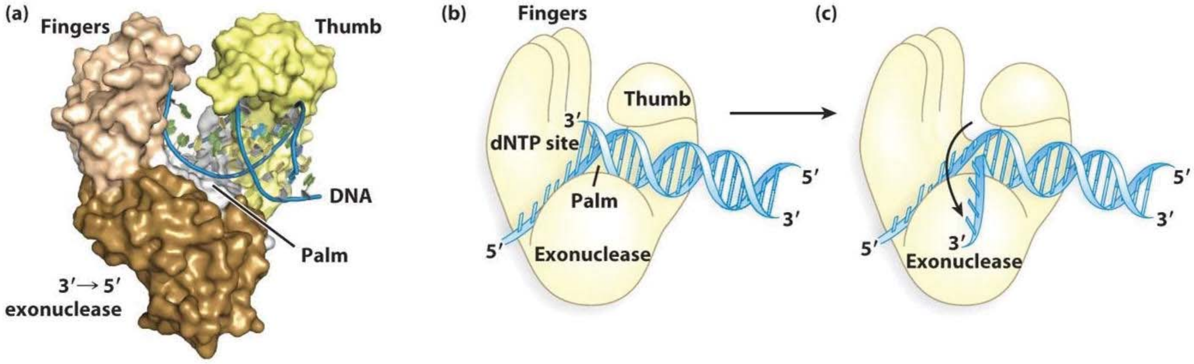


Figure 1.10: Structure and representation of replicative DNA polymerases
A - Surface crystal structure of a DNA polymerase complexed with DNA. B & C - Cartoon representation of the DNA polymerase structure in polymerization mode (B) and proofreading mode (C). Figure reproduced from [325]. Used by permission of the publisher.

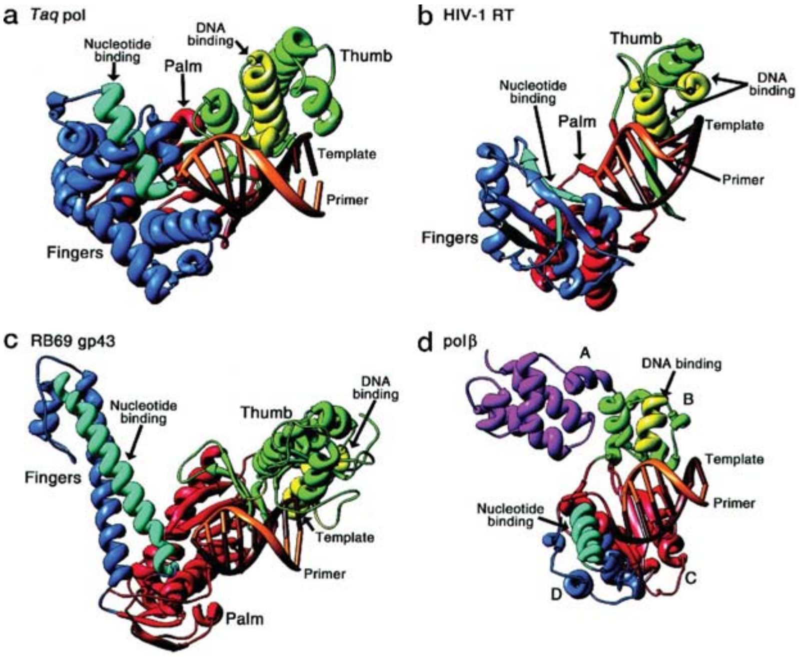


Figure 1.11: Comparison of primer-template DNA bound to four DNA polymerases. A - Taq DNA polymerase bound to DNA (co-crystal structure); B - the binary complex of HIV-1 RT and DNA (co-crystal structure); C - the model of DNA bound to RB69 gp43 (homology model); D - the ternary complex of rat pol β with 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

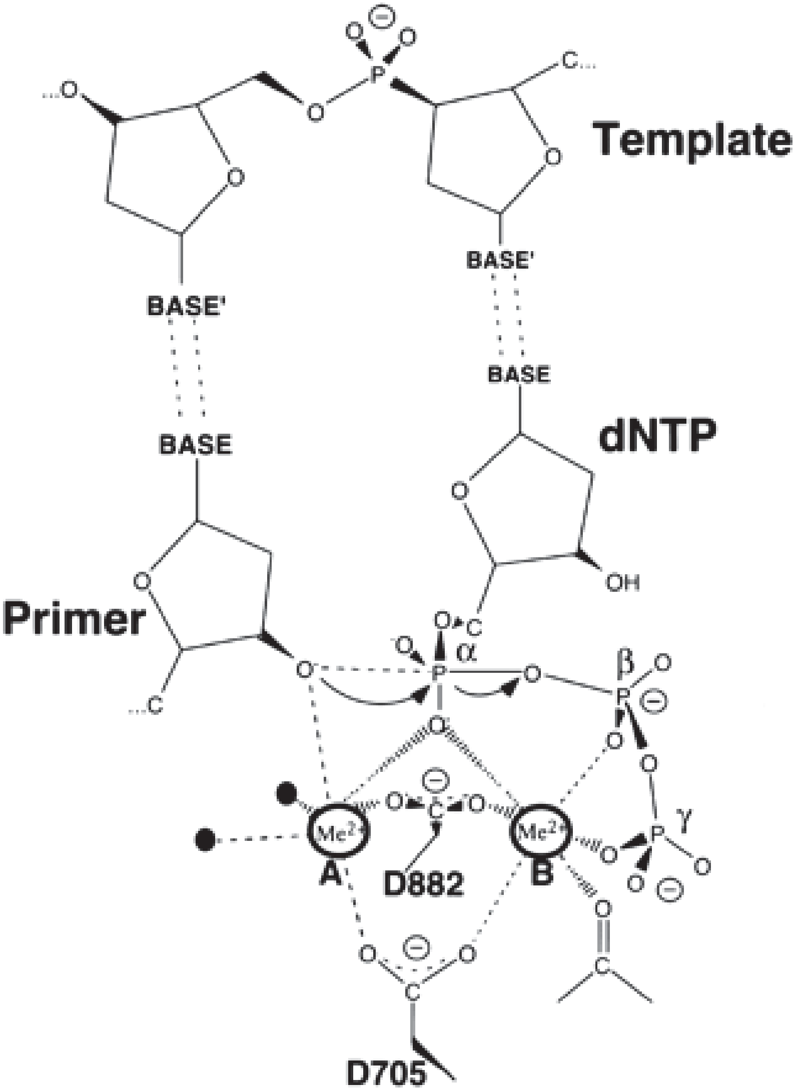
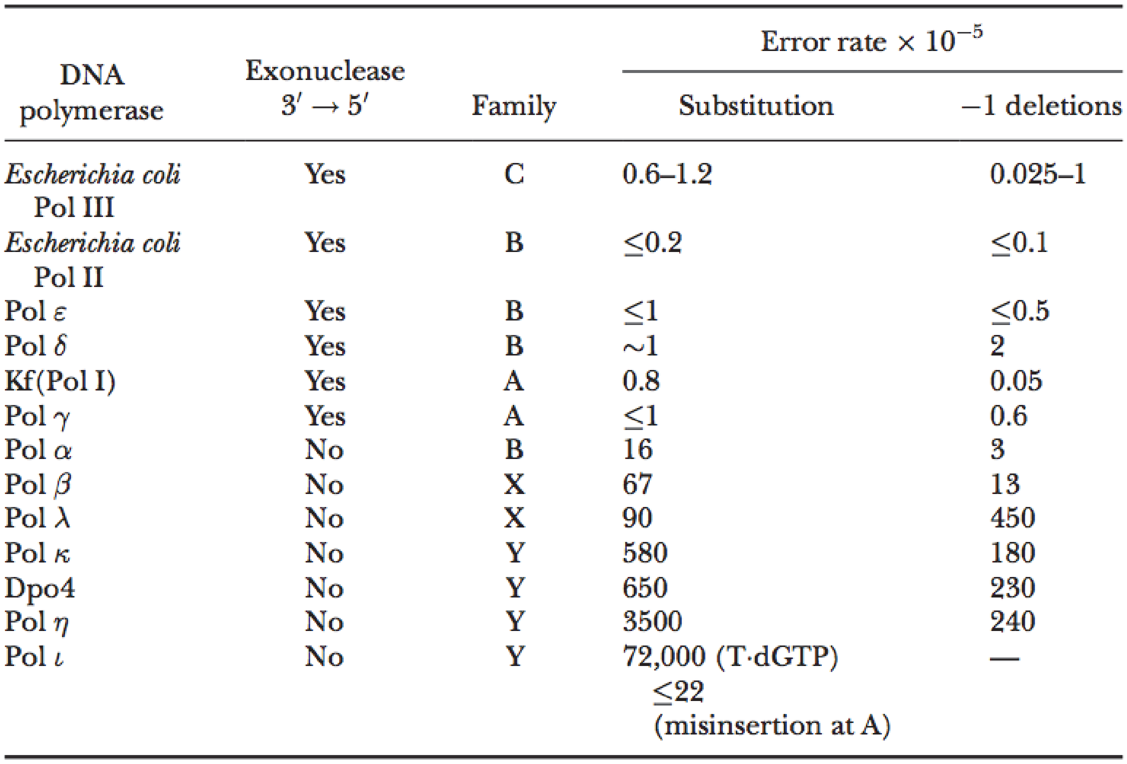


Figure 1.12: Mechanism of DNA polymerization
Polymerisation uses a two-metal ion mechanism that is thought to stabilize the resulting penta-coordinated transition state. (The two essential aspartates are annotated with the *E. coli* DNA polymerase I numbers.) Figure reproduced from [323] in accordance with the publisher's copyright permission policy.



DNA polymerase	Exonuclease $3' \rightarrow 5'$	Family	Error rate $\times 10^{-5}$	
			Substitution	-1 deletions
Escherichia coli Pol III	Yes	C	0.6–1.2	0.025–1
Escherichia coli Pol II	Yes	B	≤ 0.2	≤ 0.1
Pol ε	Yes	B	≤ 1	≤ 0.5
Pol δ	Yes	B	~ 1	2
Kf(Pol I)	Yes	A	0.8	0.05
Pol γ	Yes	A	≤ 1	0.6
Pol α	No	B	16	3
Pol β	No	X	67	13
Pol λ	No	X	90	450
Pol κ	No	Y	580	180
Dpo4	No	Y	650	230
Pol η	No	Y	3500	240
Pol ι	No	Y	72,000 (T·dGTP) ≤ 22 (misinsertion at A)	—

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

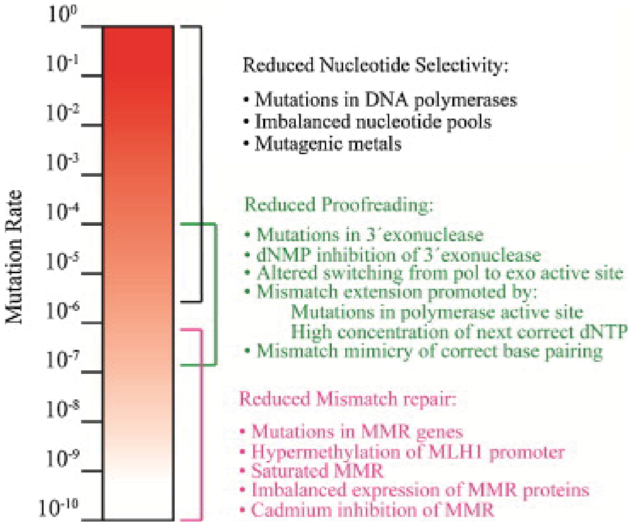


Figure 1.13: Replication fidelity
Nucleotide selectivity, exonuclease activity (proofreading) and mismatch repair all contribute to DNA replication fidelity in series to different degrees. The brackets indicate the magnitude of the contribution the different processes can make and examples of defects 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 coworkers 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 solfataricus,* the extended β-hairpin is found in the exonuclease domains of all other B-type polymerases

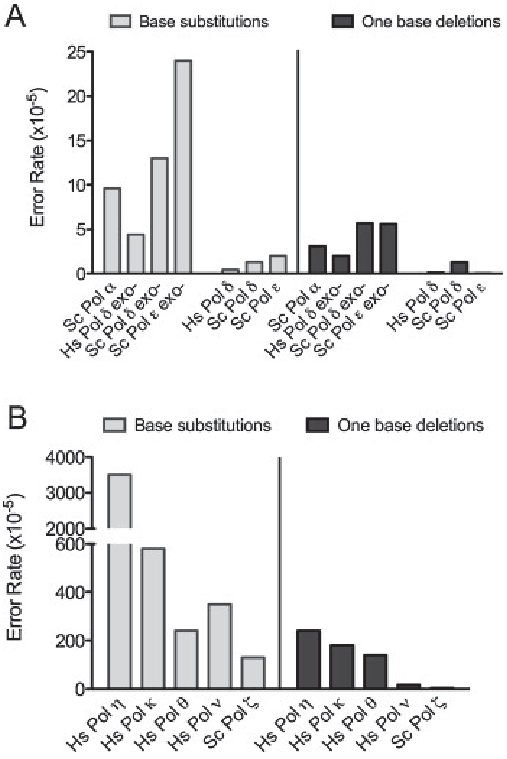


Figure 1.14: Fidelity of different DNA polymerases
"Eukaryotic DNA polymerase error rates for single base mutations. (A) Error rates for *Homo sapiens* Pol δ and *S. cerevisiae* polymerases α, δ and ε, for single base substitutions (BS; light grey bars) and one base deletions (dark grey). (B) Error rates 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 premutagenic 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 350-450nm 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 O^6-methylguanine back to guanine by the O^6-alkylguanine DNA alkyltransferases (also known as MGMT)[374]. O^6-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 S_N2-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 O^4-methylthymine in addition to O^6-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-synethesises 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

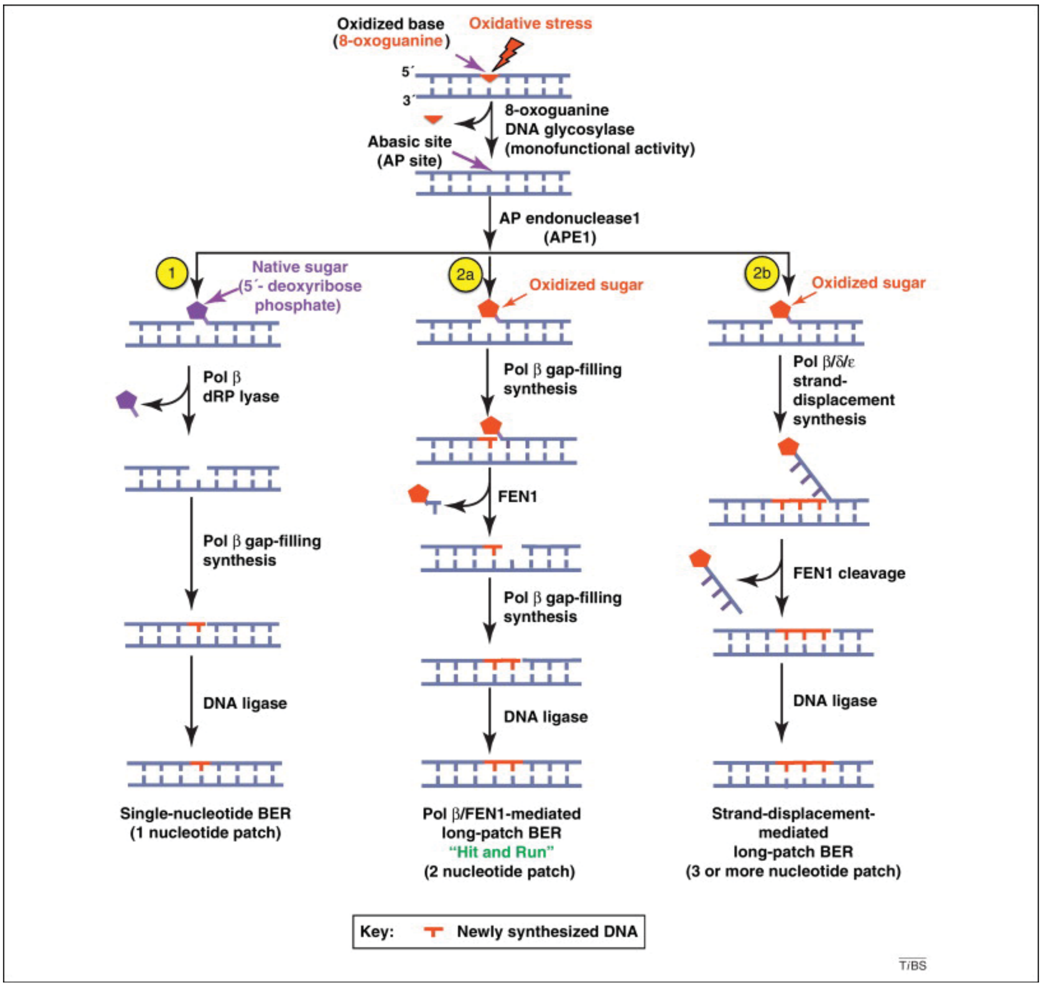


Figure 1.15: Base excision repair (BER) of oxidized DNA base lesions

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 basic (AP) site. AP endonuclease 1 (APE1) subsequently incises at the 5' side of the AP site sugar leaving either a native or oxidised sugar phospahte. 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 a 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

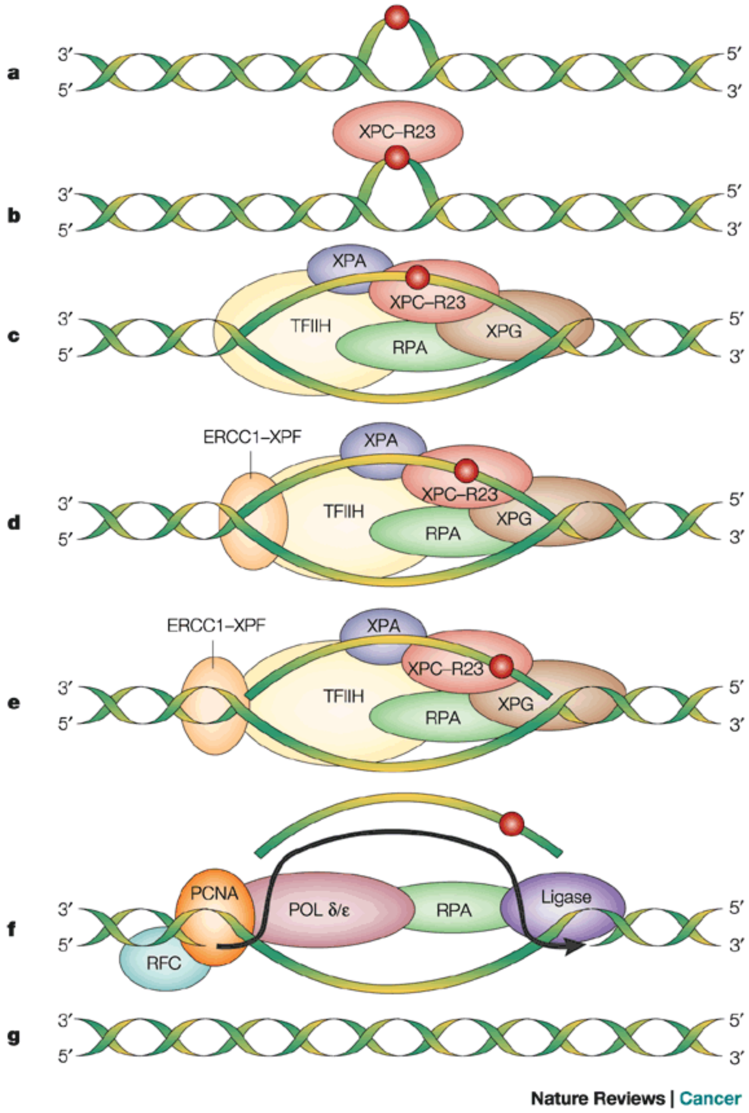


Figure 1.16: Nucleotide excision repair (NER)

A - Nucleotide excision repair (NER) repairs damaged DNA bases that distorts 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.

base[403, 404]. In prokaryotes, this critical distinction between strands is made using methylation marks, whereas in eukaryotes the process remains unclear[405]. Mismatch repair has been extensively studied in *E. coli*[406] 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[407, 408] and specifically nicks the nascent strand to create an initiation site[402, 409]. The first step in MMR is the recognition and binding of this type of lesion by the MutS dimer[410], 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[411] generating ssDNA which is swiftly covered by SSB. Depending on the relative position of the mismatch to the nick different exonucleases excise the DNA[410]. The resulting gap is repaired by the DNA polymerase holoenzyme and DNA ligase[402] and, lastly, a deoxyadenosine methylase methylates the daughter strand.

Eukaryotic MMR is very similar, but not completely understood[401, 402]. Like prokaryotic MMR it shows substrate specificity, bidirectionally 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[405]. 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[412–415]. The eukaryotic equivalents of MutS are formed by Msh2 and Msh3 (MutSα), which recognises base-base mismatches and 1-2base indels, and by Msh2 and Msh6 (MutSβ), which recognises larger INDELs[412, 416–419]. Both are ATPases and involved in recognition of mismatches[401]. Mlh1 heterodimerises to form MutL homologs[401]: with Pms2 to form MutLα, Pms1 to form MutLβ and MLH3 to form MutLγ[413, 420–422]. 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[423, 424], as well as Msh6 and Msh3[425–428], it has been proposed that PCNA recruits MutSα and MutSβ to newly replicated DNA to monitor newly synthesised DNA for mismatches[429, 430]. Evidence from *S. cerevisiae* has identified Exo1 as the only exonucleasece definitively involved . It can also bind Msh2 and Mlh1 [431–436] and has been shown to catalyse 5’ directed mismtach excision in the presence of MutS and RPA[437, 438]. However, considering that *exo1* null yeast and mice show only weak mutator phenotypes[434, 439], there are likely other important exonucleases[410].

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[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[467], and Polη contributes to DNA synthesis during HR[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)[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[471], and the spindle assembly checkpoint, which does prevent mitosis until the mitotic spindle has been assembled and all chromosomes are properly attached[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[473, 474], during S phase (intra-S checkpoint), which slows down S phase progression and promotes alternative replication mechanisms such as TLS[475], and at the metaphase/anaphase transition in M phase (G2/M checkpoint), which prevents division of damaged chromatids in the budding yeast *S. cerevisiae*[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[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

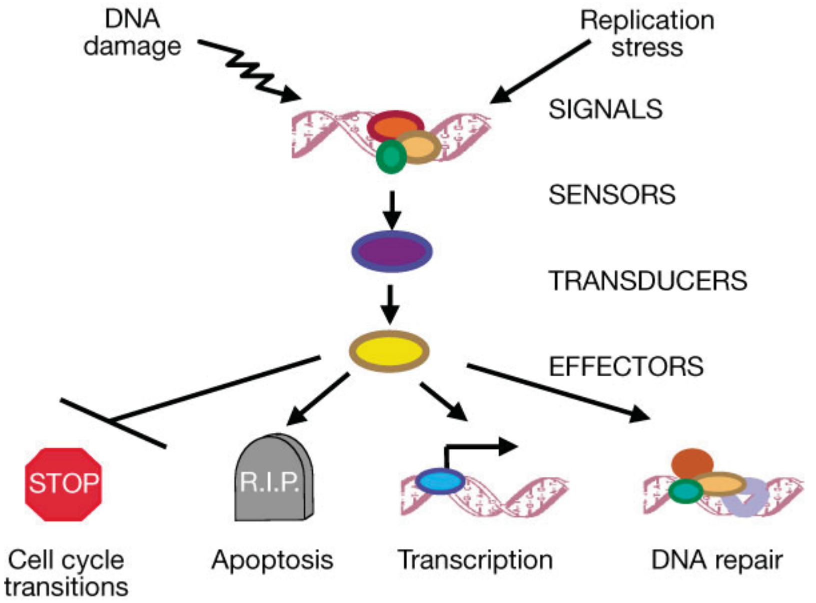


Figure 1.17: A general outline of the DNA damage signal transduction pathway
Arrows represent activating events and perpendicular ends represent an inhibitory event. The stop sign depicts cell-cycle arrest and a tombstone signifies apoptosis. DNA damage-induced transcription is represented by the helix with the arrow, while the helix with oval shaped subunit representations depicts damage-induced DNA-repair. For simplicity, the network of interating pathways is instead outlined as a linear sequence of events. Reproduced from [482] with permission from the publisher.

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 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.

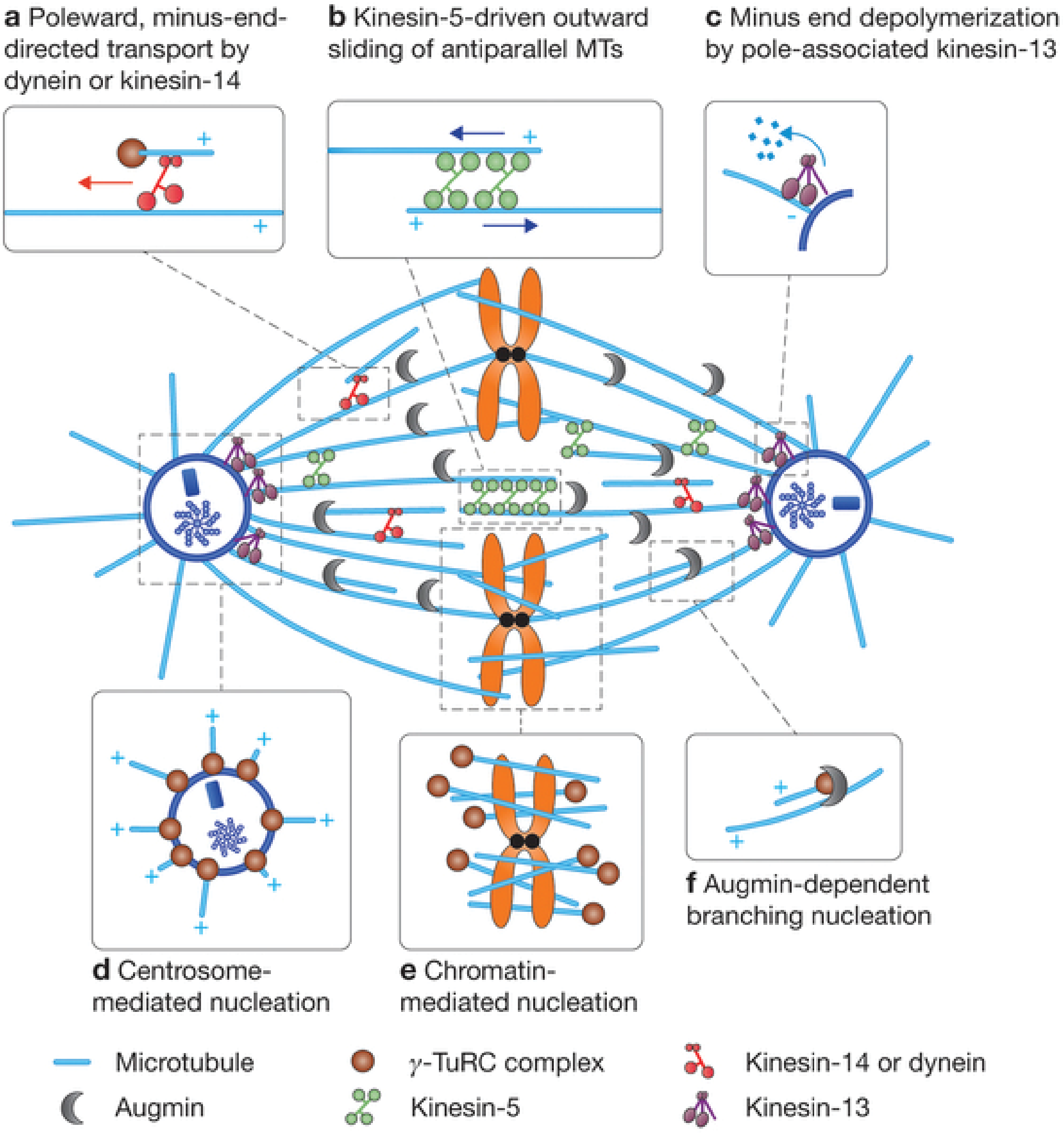

a Poleward, minus-end-directed transport by dynein or kinesin-14
b Kinesin-5-driven outward sliding of antiparallel MTs
c Minus end depolymerization by pole-associated kinesin-13
+
+
+
+
-
+
+
+
+
+
+
+
+
+
+
d Centrosome-mediated nucleation
e Chromatin-mediated nucleation
f Augmin-dependent branching nucleation
Microtubule
Augmin
γ-TuRC complex
Kinesin-5
Kinesin-14 or dynein
Kinesin-13

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.

1.2.1.1 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-occuring 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)

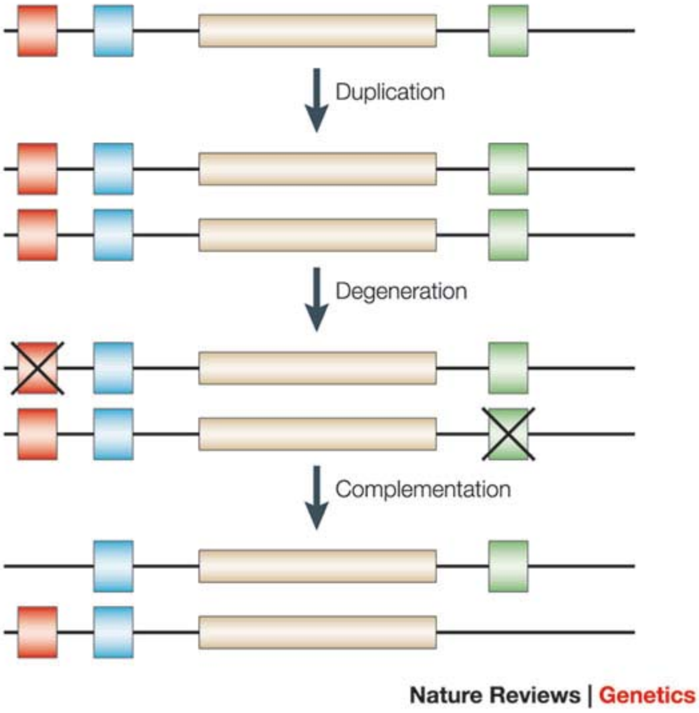
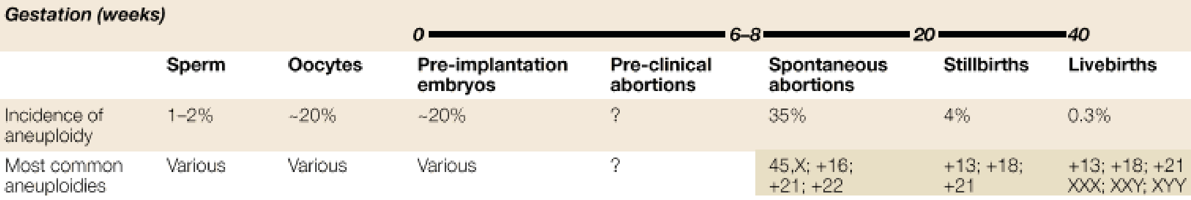


Figure 1.19: Gene duplications: the duplication-degeneration (DDC) model
The duplication-degeneration (DDC) model relies on complementary degenerative changes in two duplicated genes in a way that the two together retain the original function. The coloured boxes represent cis regulatory elements, but mutations in other functional elements such as a protein domain or splice site is possible. Reproduced from [485] with permission from the publisher.

genes[485]. 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



Gestation (weeks)			0		6–8	20	40
	Sperm	Oocytes	Pre-implantation embryos	Pre-clinical abortions	Spontaneous abortions	Stillbirths	Livebirths
Incidence of aneuploidy	1–2%	~20%	~20%	?	35%	4%	0.3%
Most common aneuploidies	Various	Various	Various	?	45,X; +16; +21; +22	+13; +18; +21	+13; +18; +21 XXX; XXY; XYY

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 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].

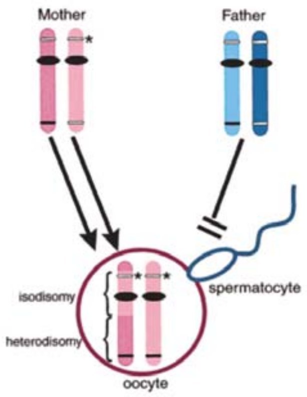
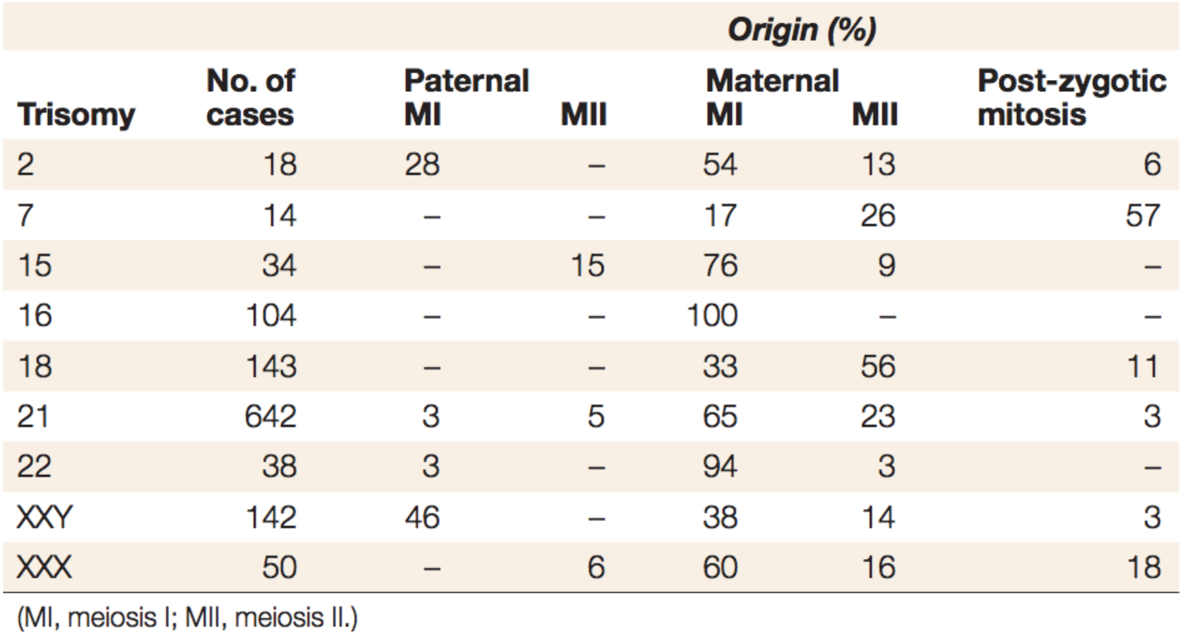


Figure 1.20: Uniparental Disomy - A special case of aneuploidy
An example of uniparental disomy: a non-disjunction event in maternal meiosis leads to the transmission of both copies of a particular chromosome pair to the gamete, which is then fertilised by a spermatocyte that is nullisomic for the pair 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 pubisher.

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 ore 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



		Origin (%)				
Trisomy	No. of cases	Paternal MI	Paternal MII	Maternal MI	Maternal MII	Post-zygotic mitosis
2	18	28	–	54	13	6
7	14	–	–	17	26	57
15	34	–	15	76	9	–
16	104	–	–	100	–	–
18	143	–	–	33	56	11
21	642	3	5	65	23	3
22	38	3	–	94	3	–
XXY	142	46	–	38	14	3
XXX	50	–	6	60	16	18

(MI, meiosis I; MII, meiosis II.)

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

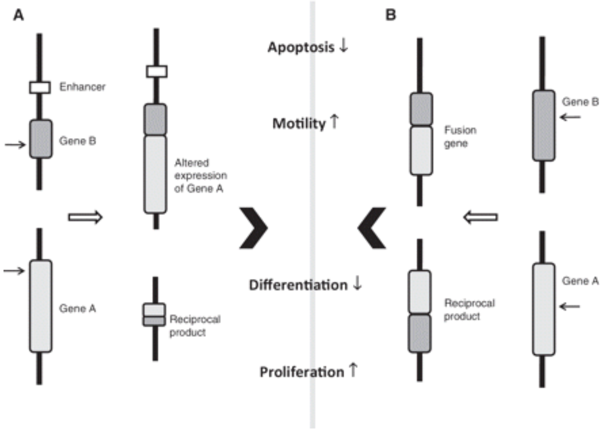


Figure 1.21: Consequences of chromosomal translocations
Chromosomal translocations can result in the placement of genes near different regulatory elements (A) or in aberrant gene fusions (B). Reproduced from [534] with permission from the publisher.

70s was identified to be a translocation involving the long arm of Chr22 and the long arm of Chr9 to form what is now commonly called the Philadelphia chromosome[530, 531]. In 1982, it was determined that *ABL1* was translocated in the process[532] and now it is clear that this particular translocation causes the fusion of two genes, *BCR* and *ABL*, to form an aberrant chimeric *BCR-ABL* which as a constitutively active tyrosine kinase promotes uncontrolled cellular proliferation and cancer partly through signaling through the oncogene *RAS*[533]. This and other clinically relevant translocations sparked investigation of these types of mutations: translocations are usually the result of reciprocal swapping of chromosome arms from heterologous chromosomes following a DNA DSB[534]. They can have severe consequences as the above example suggests: deregulation of key cellular proteins by generating aberrant gene fusions or placement of a gene under different transcriptional control causing aberrant gene expression[534](Fig. 1.21). While the exact mechanism of chromosomal translocations is still under investigations, there is evidence that *AID* and the *RAG* complex, proteins that cause DSBs critical for V(D)J recombination in immune cells, are involved. Cryptic *RAG* target sites have been identified elsewhere in the genome, which could explain the fact that in many known cases the IgH locus on chromosome 14 is involved in a chromosomal translocation[534]. However, since expression of the *RAG* complex is restricted to distinct types of immune cells, they cannot account for all chromosomal translocations, and other mechanisms such as BIR have been implicated in the generation of chromosomal translocations[534].

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, tetraplexes 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-

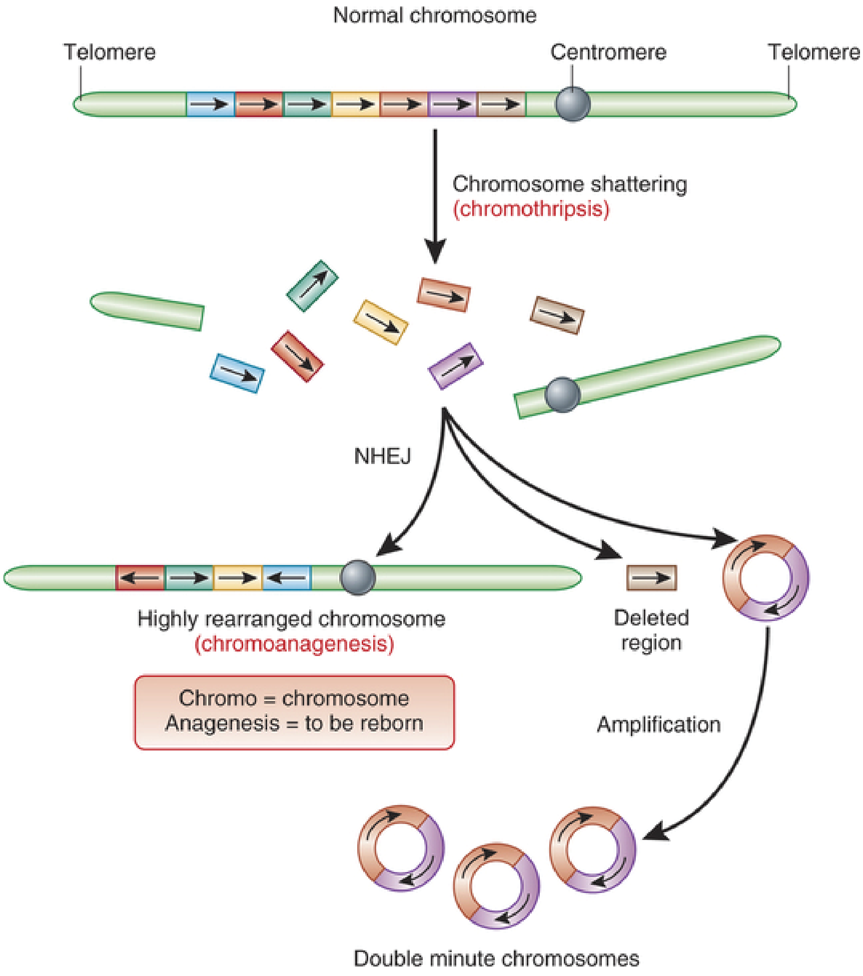


Figure 1.22: Chromothripsis
Chromothripsis is the shattering of one or more chromosomes, leading to the simultaneous generation of many double strand breaks, most of which are repaired by NHEJ in a manner leading to chromoanagenesis: the generation of a highly rearranged chromosome. Broken DNA fragments can also circularise to generate double minute chromosomes which are often amplified. Reproduced from [535] with permission from the publisher.

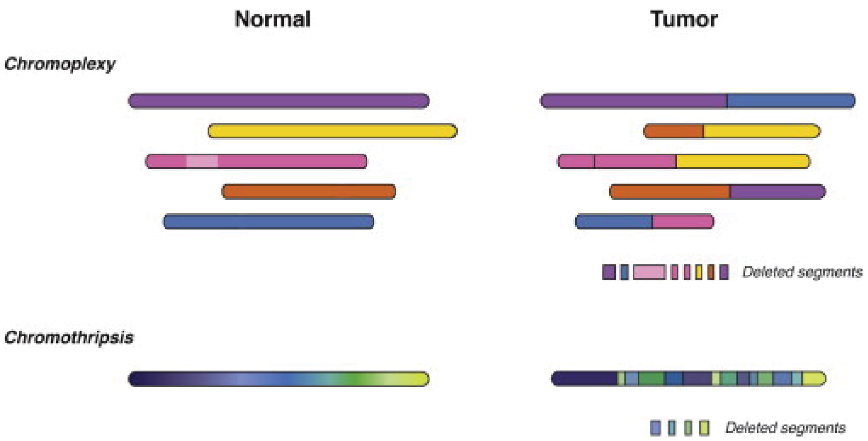


Figure 1.23: Chromoplexy and Chromothripsis
Schematic representations of genomic rearrangements found in Chromoplexy (top) and Chromothripsis (bottom). Reproduced from [536] with the permission of the publisher.

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-

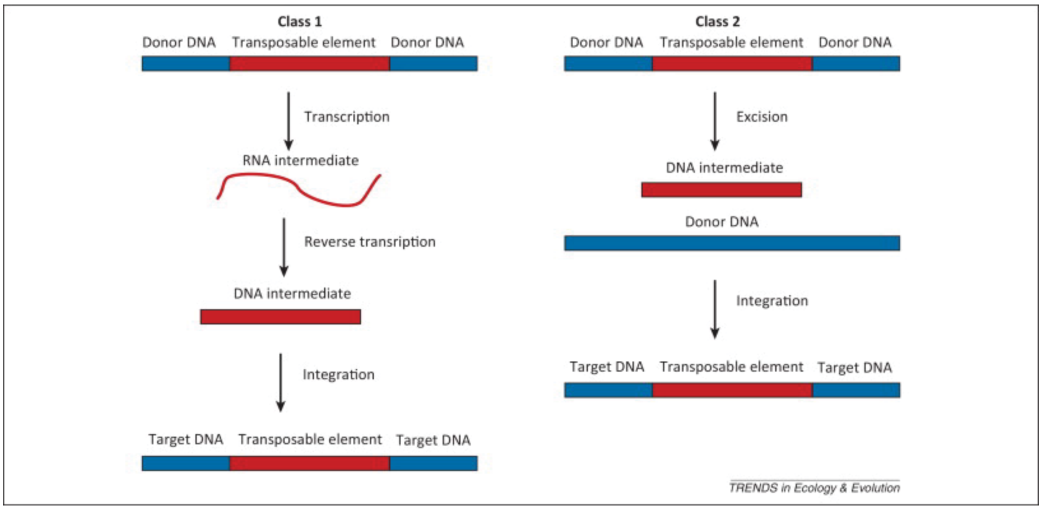


Figure 1.24: Classes of DNA transposons
There are two types of transposable elements: retrotransposons (Class I) and DNA transposons (Class II). Class I move via an RNA intermediate (“copy and paste”), while the latter excise themselves from the DNA (“cut and paste”).
Reproduced from [548] with permission 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 contributer 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].

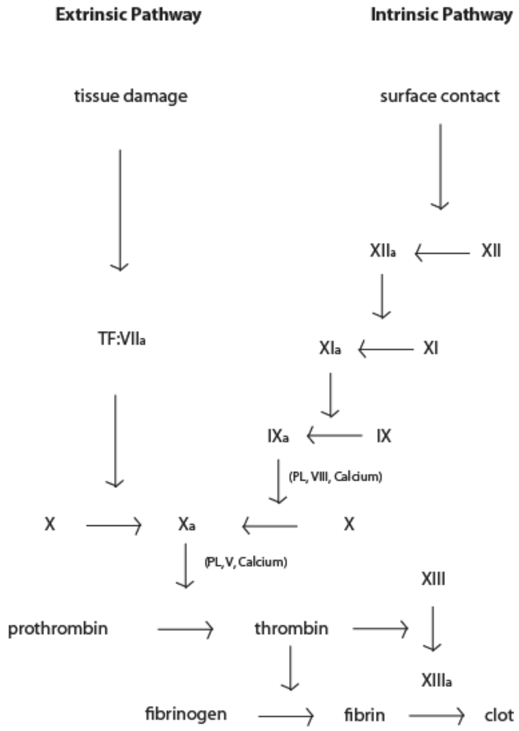


Figure 1.25: Blood clotting cascade
Schematic representation of the blood-clotting cascade. Many of the involved factors are serine proteases which - by cleaving - activate another downstream serine protease. The amplification inherent in signal transduction cascades allows a small stimulus to generate 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 *Agrobacterim 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[563], the DNA sequence of the gene codes for the sequence of amino acids[564]. 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[565, 566]. 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(Fig: 1.27). 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 *FBN1* gene causing Marfan syndrome, an autosomal dominant disease affecting the connective tissue[567]. Fibrillin-1, encoded by *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 TGFβ. Point mutations in *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

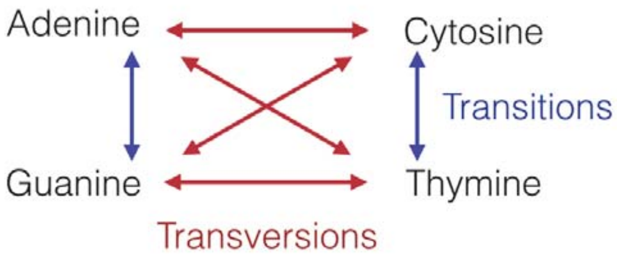


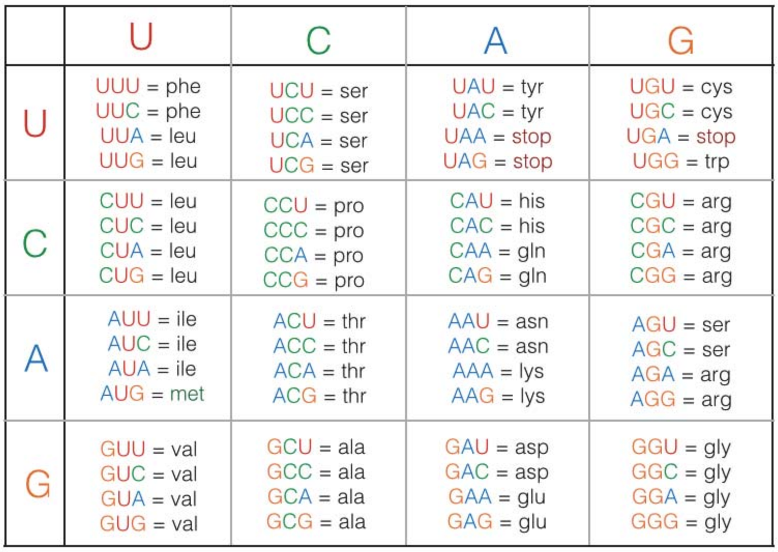
Figure 1.26: Transitions and Transversions

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



	U	C	A	G

U	UUU = phe
	UUC = phe
	UUA = leu
	UUG = leu
		UCU = ser
		UCC = ser
		UCA = ser
		UCG = ser
			UAU = tyr
			UAC = tyr
			UAA = stop
			UAG = stop
				UGU = cys
				UGC = cys
				UGA = stop
				UGG = trp

C	CUU = leu
	CUC = leu
	CUA = leu
	CUG = leu
		CCU = pro
		CCC = pro
		CCA = pro
		CCG = pro
			CAU = his
			CAC = his
			CAA = gln
			CAG = gln
				CGU = arg
				CGC = arg
				CGA = arg
				CGG = arg

A	AUU = ile
	AUC = ile
	AUA = ile
	AUG = met
		ACU = thr
		ACC = thr
		ACA = thr
		ACG = thr
			AAU = asn
			AAC = asn
			AAA = lys
			AAG = lys
				AGU = ser
				AGC = ser
				AGA = arg
				AGG = arg

G	GUU = val
	GUC = val
	GUA = val
	GUG = val
		GCU = ala
		GCC = ala
		GCA = ala
		GCG = ala
			GAU = asp
			GAC = asp
			GAA = glu
			GAG = glu
				GGU = gly
				GGC = gly
				GGA = gly
				GGG = gly

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

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

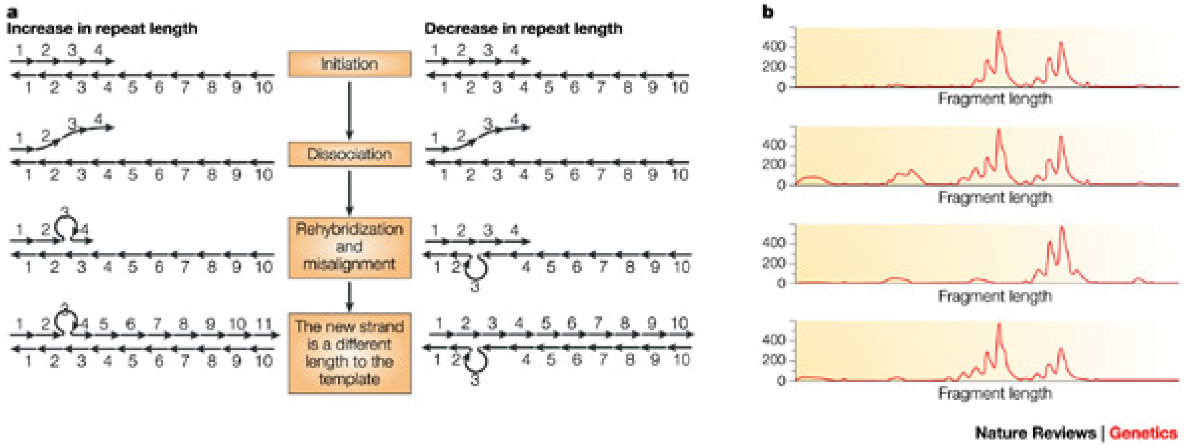

a
Increase in repeat length
Decrease in repeat length
Initiation
Dissociation
Rehybridization and misalignment
The new strand is a different length to the template
b
400
200
0
Fragment length
Nature Reviews | Genetics

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

1931[579]. 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-

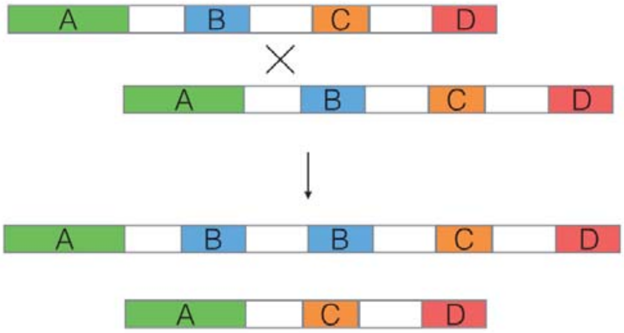


Figure 1.29: Unequal crossovers result in chromosome rearrangements
Crossovers sometimes occur between similar but not equivalent regions of the genome leading to an unequal exchange of DNA between chromosomes.

tion 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 O_2-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 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 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 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 photo-pyrimidine 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.

1.4.2 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 karykinetic 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

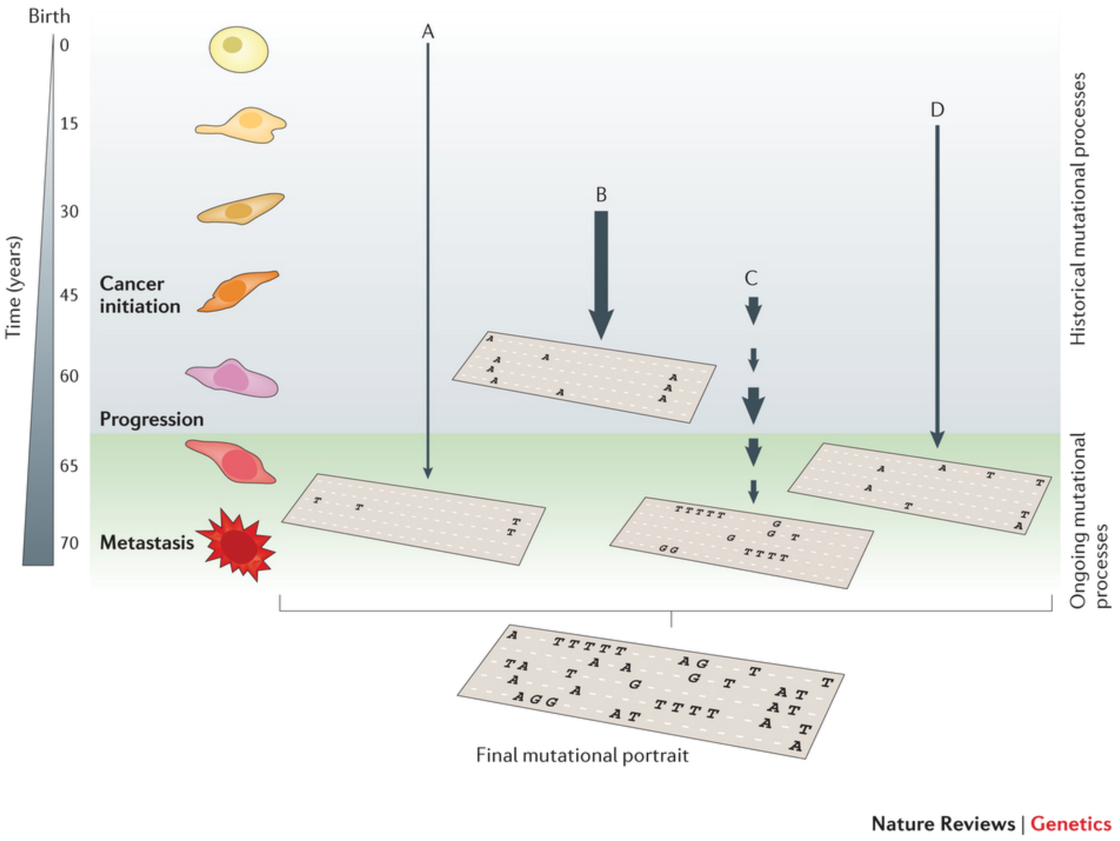

Birth
0
15
30
45
60
65
70
Time (years)
Cancer initiation
Progression
Metastasis
A
B
C
D
Historical mutational processes
Ongoing mutational processes
Final mutational portrait
Nature Reviews | Genetics

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 highy which is reflected in the final mutational portrait. Reproduced from [700] with permission from the publisher.

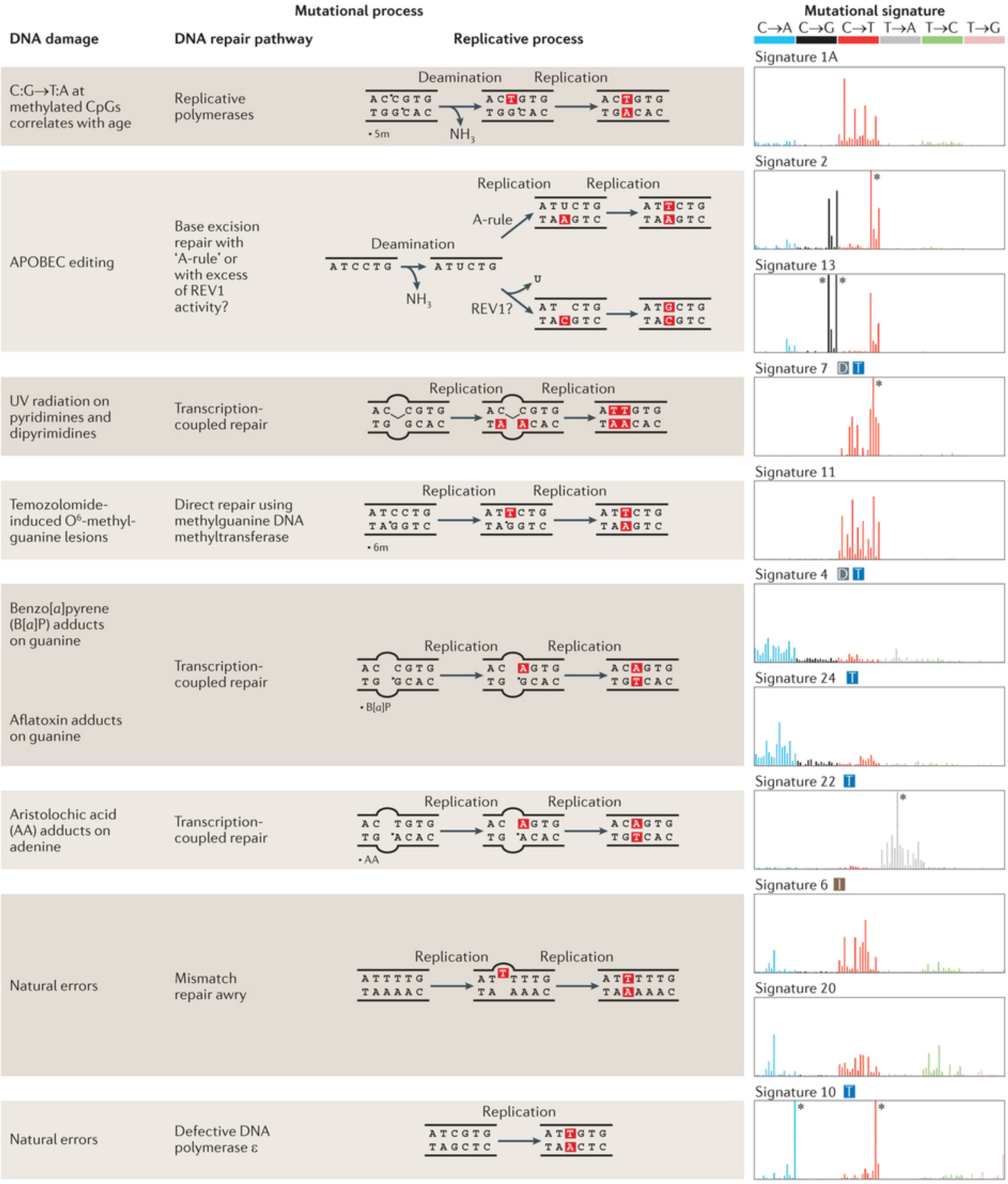
tumorigenesis are vital exercises, demonstrated by the fact that identification of potent carcinogens can be used in public health campaigns to drastically curb exposure to the substance and reduce cancer incidence[619], the ability of health care professionals to screen for predisposing mutations and 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 $\varepsilon^{\text{Exo}-}$ mice developing mainly intestinal tumours with 50% survival of ~16 months and Pol $\delta^{\text{Exo}-}$ mice exhibiting primarily thymic lymphomas with 50% survival of ~ 6 months. Considering that the error rates and specificities of Pol $\varepsilon^{\text{Exo}-}$ and Pol $\delta^{\text{Exo}-}$ 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


Mutational process
DNA damage
DNA repair pathway
Replicative process
Mutational signature
C→A C→G C→T T→A T→C T→G
C:G→T:A at methylated CpGs correlates with age
Replicative polymerases
Deamination
Replication
5m
NH3
Signature 1A
APOBEC editing
Base excision repair with 'A-rule' or with excess of REV1 activity?
A-rule
REV1?
U
Signature 2
Signature 13
UV radiation on pyridimines and dipyrimidines
Transcription-coupled repair
Signature 7 D T
Temozolomide-induced O6-methyl-guanine lesions
Direct repair using methylguanine DNA methyltransferase
6m
Signature 11
Benzo[a]pyrene (B[a]P) adducts on guanine
Aflatoxin adducts on guanine
B[a]P
Signature 4 D T
Signature 24 T
Aristolochic acid (AA) adducts on adenine
AA
Signature 22 T
Natural errors
Mismatch repair awry
Signature 6 I
Signature 20
Defective DNA polymerase ε
Signature 10 T
*

Nature Reviews | Genetics

Figure 1.31: Summary of known mutational signatures
Reproduced from [700] with permission from the publisher.

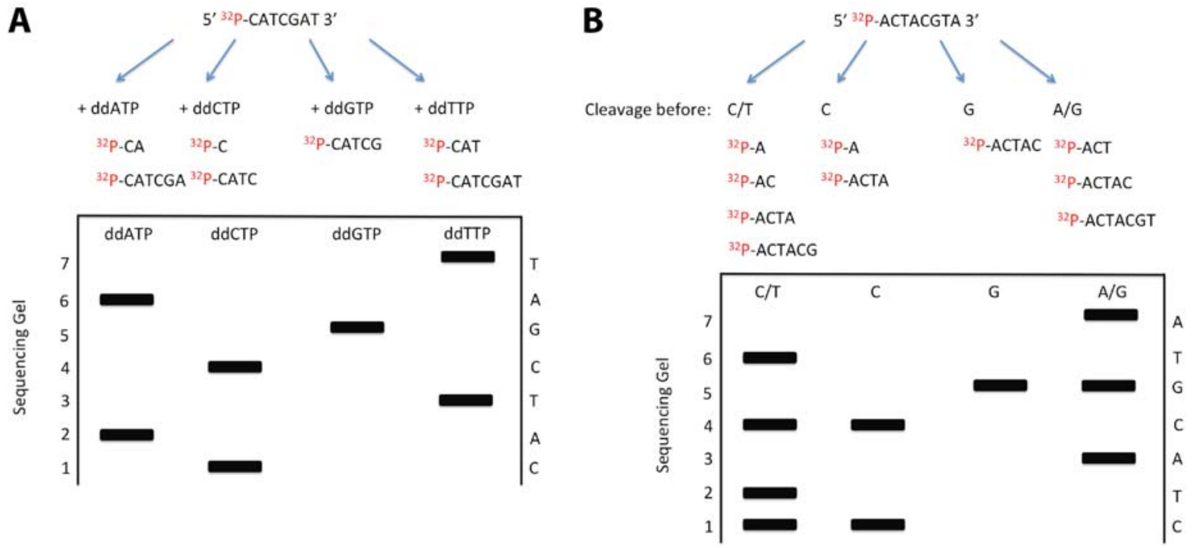


Figure 1.32: Early sequencing techniques: Gilbert and Sanger
Schematics to illustrate the principles of early DNA sequencing strategies. **A|** Dideoxy “Sanger” Sequencing: This type of sequencing required a DNA template, a primer a DNA polymerase, all four standard dNTPs and one of the di-deoxy NTPs, 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 spearate 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 nucelotide 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 flourescent 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

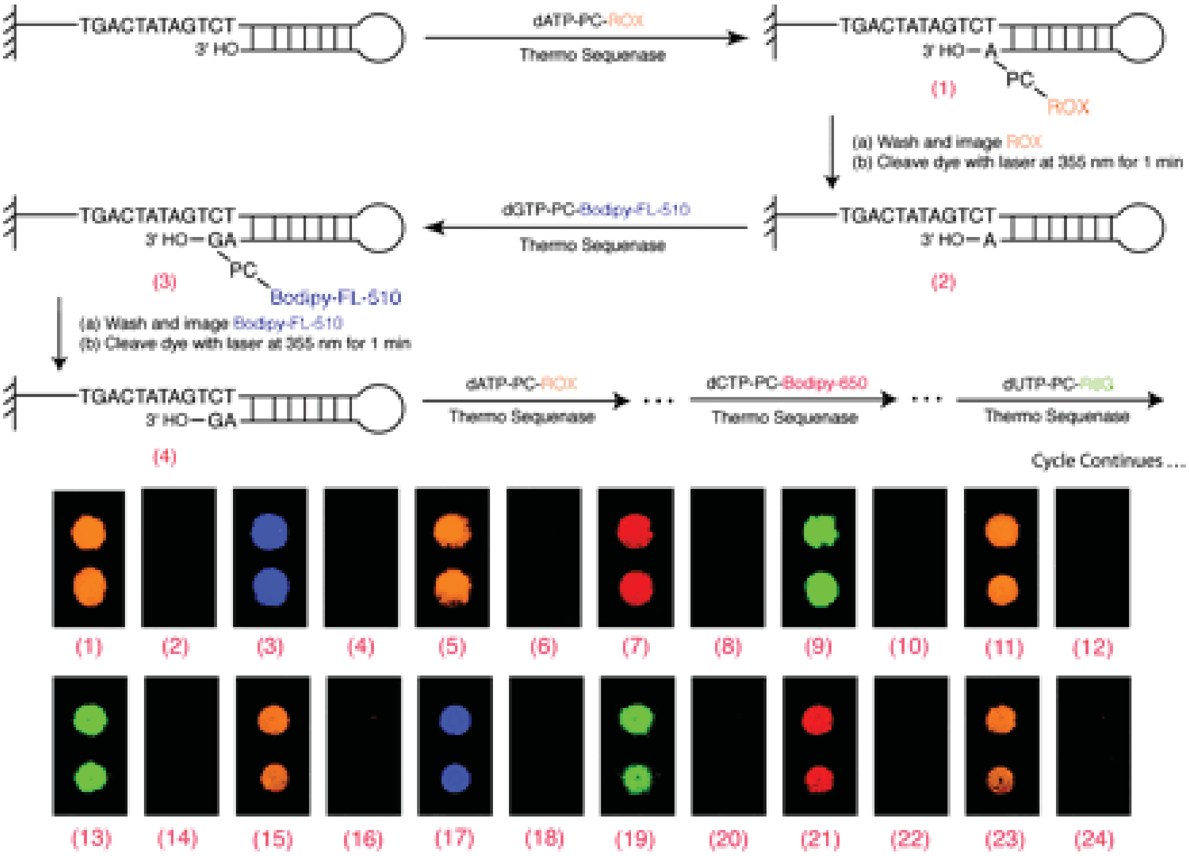


Figure 1.33: Sequencing by synthesis

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.

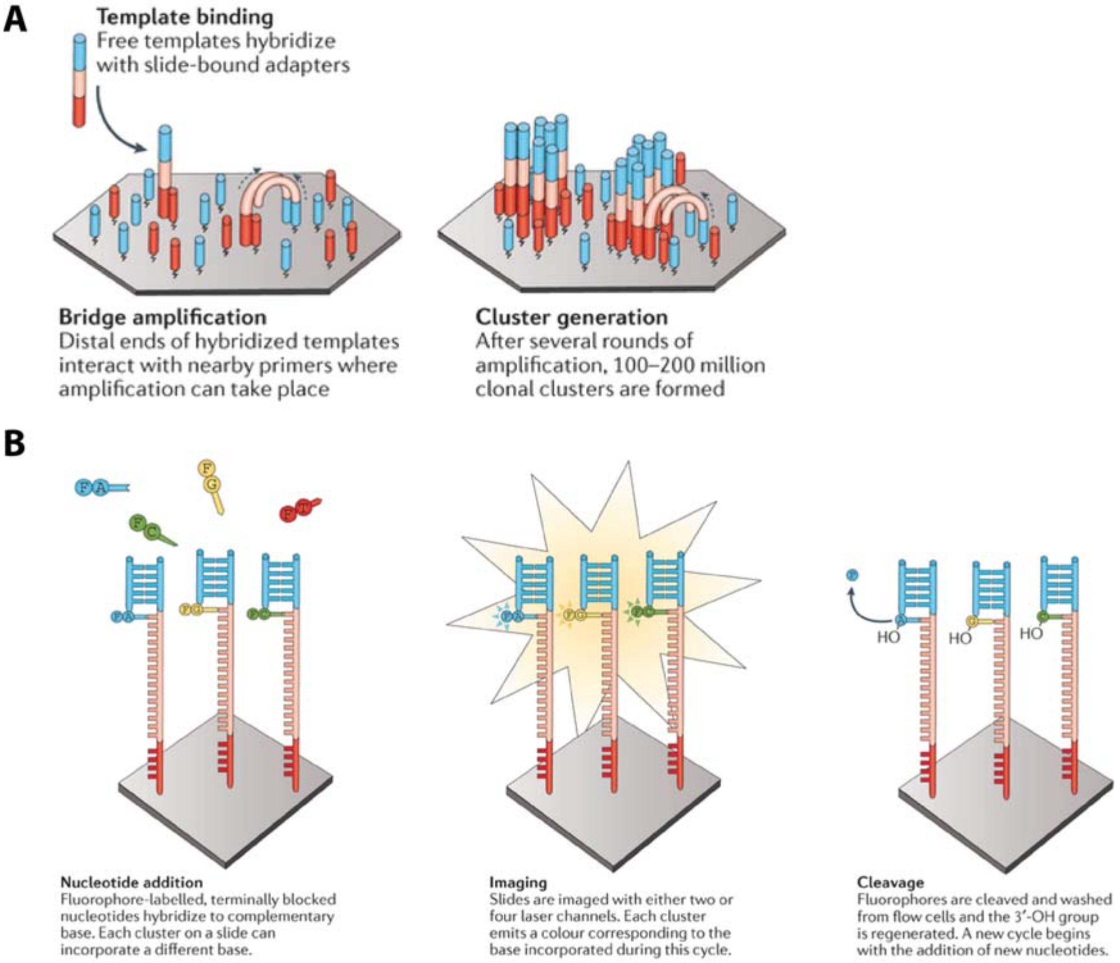

A
Template binding
Free templates hybridize with slide-bound adapters
Bridge amplification
Distal ends of hybridized templates interact with nearby primers where amplification can take place
Cluster generation
After several rounds of amplification, 100–200 million clonal clusters are formed
B
F A
F G
F C
F T
HO
HO
HO
Nucleotide addition
Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.
Imaging
Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.
Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.

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 looses 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.

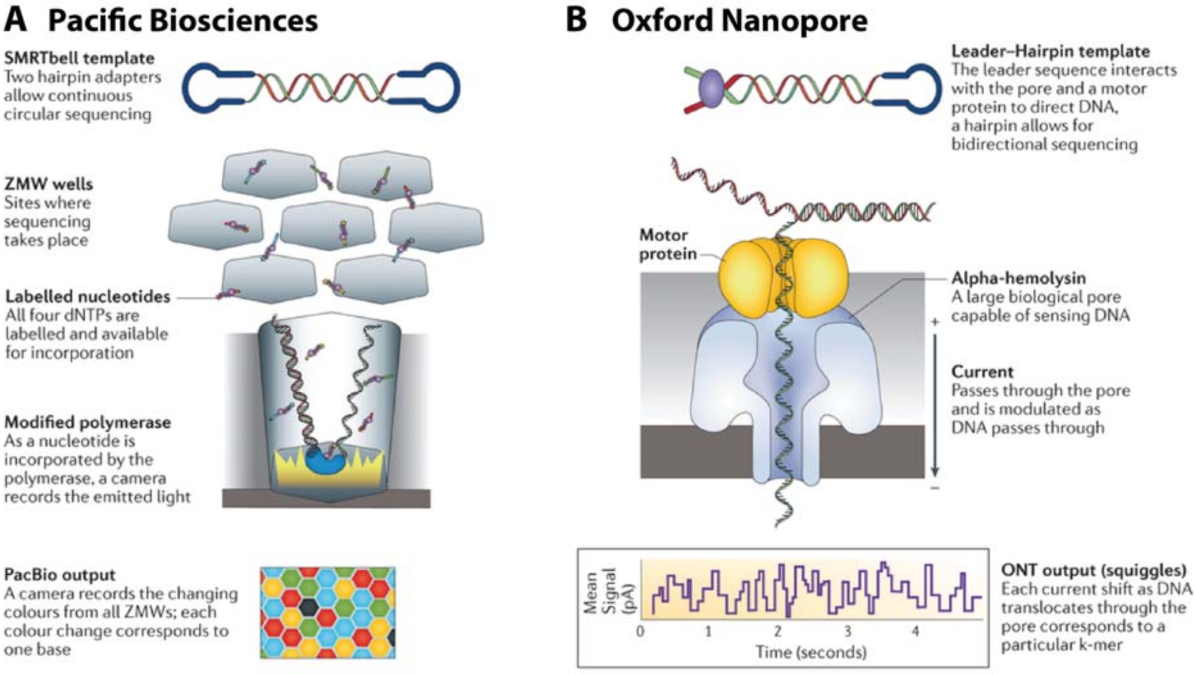

A Pacific Biosciences
SMRTbell template
Two hairpin adapters allow continuous circular sequencing
ZMW wells
Sites where sequencing takes place
Labelled nucleotides
All four dNTPs are labelled and available for incorporation
Modified polymerase
As a nucleotide is incorporated by the polymerase, a camera records the emitted light
PacBio output
A camera records the changing colours from all ZMWs; each colour change corresponds to one base
B Oxford Nanopore
Leader–Hairpin template
The leader sequence interacts with the pore and a motor protein to direct DNA, a hairpin allows for bidirectional sequencing
Motor protein
Alpha-hemolysin
A large biological pore capable of sensing DNA
+
Current
Passes through the pore and is modulated as DNA passes through
−
Mean Signal (pA)
0
1
2
3
4
Time (seconds)
ONT output (squiggles)
Each current shift as DNA translocates through the pore corresponds to a particular k-mer

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 MATa 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 MATa 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 ~5μm 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

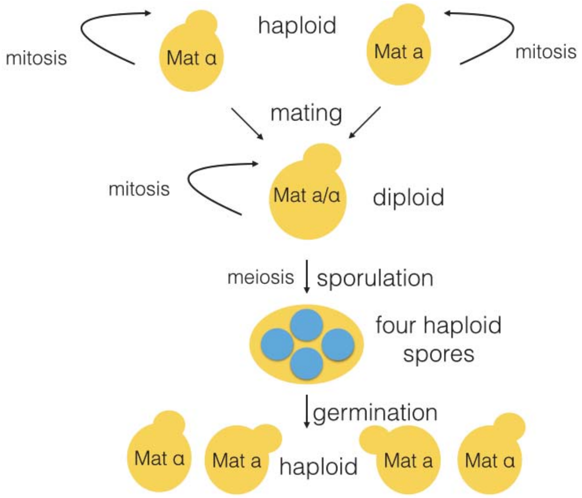


Figure 1.36: Life cycle of the budding yeast *Saccharomyces cerevisiae*
Haploid and diploid yeast cells can reproduce by mitosis. A haploid Mat a and a haploid Mat α can mate and form a diploid. A diploid cell can undergo meiosis and generate four haploid spores.

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 flourescent 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)[795] for which 4958 homologs can be identified in humans[796]. Most budding yeast genes do not contain any intron (only ~4% do)[780] 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[780]. 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. Chromosome 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[797]. 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)[780]. 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)[798], initiation of transcription (Nobel Prize in 2006)[799] and eukaryotic telomere structure (Nobel Prize in 2009)[800], 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.

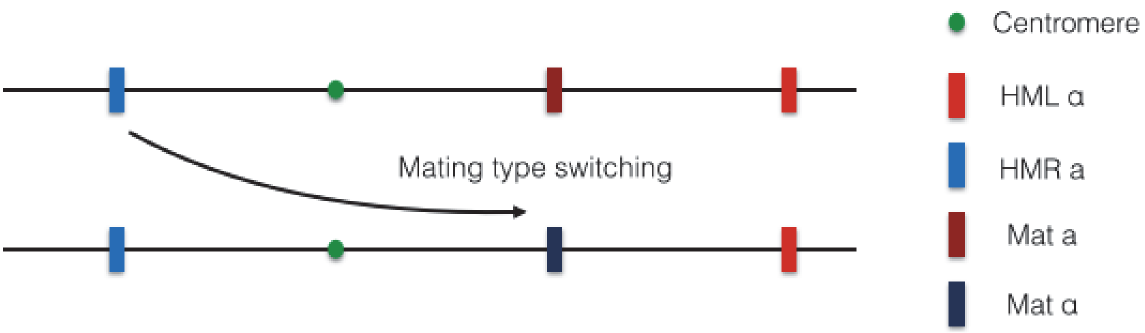
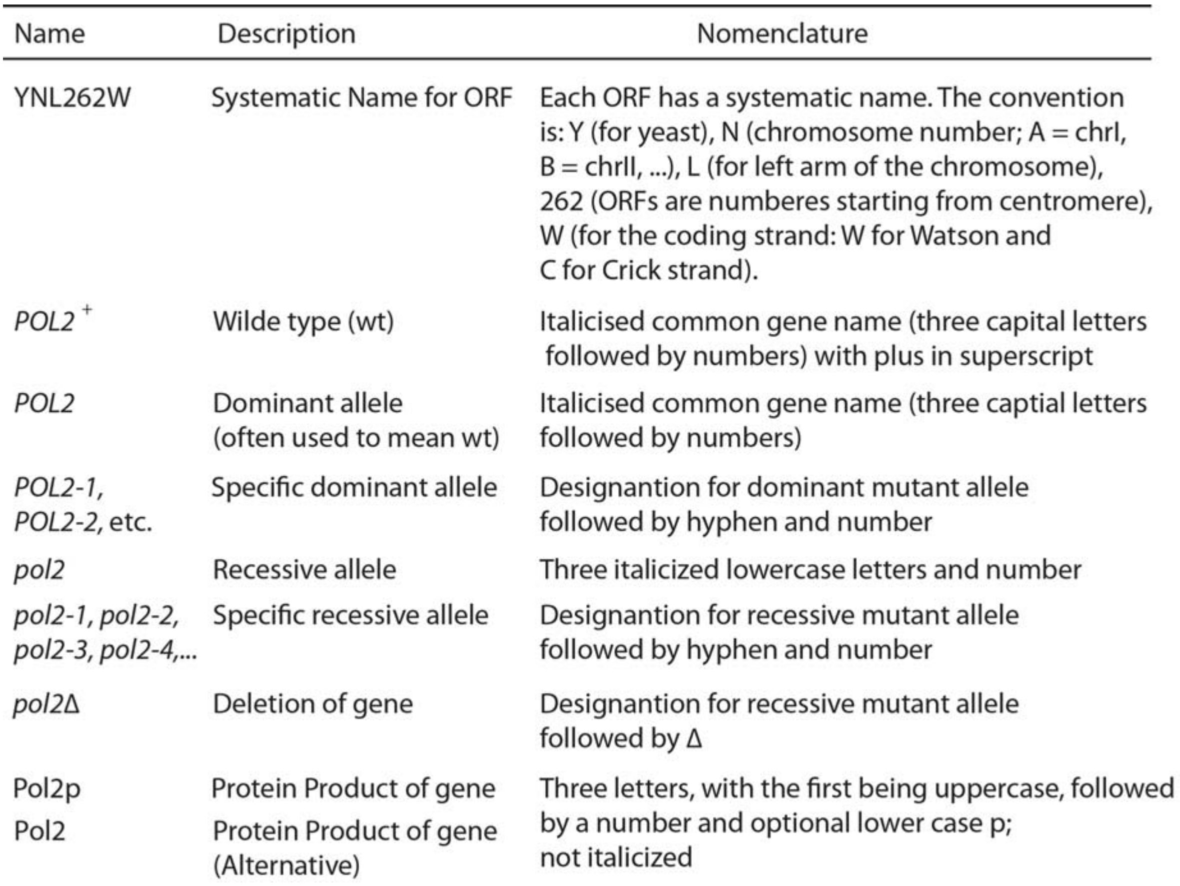
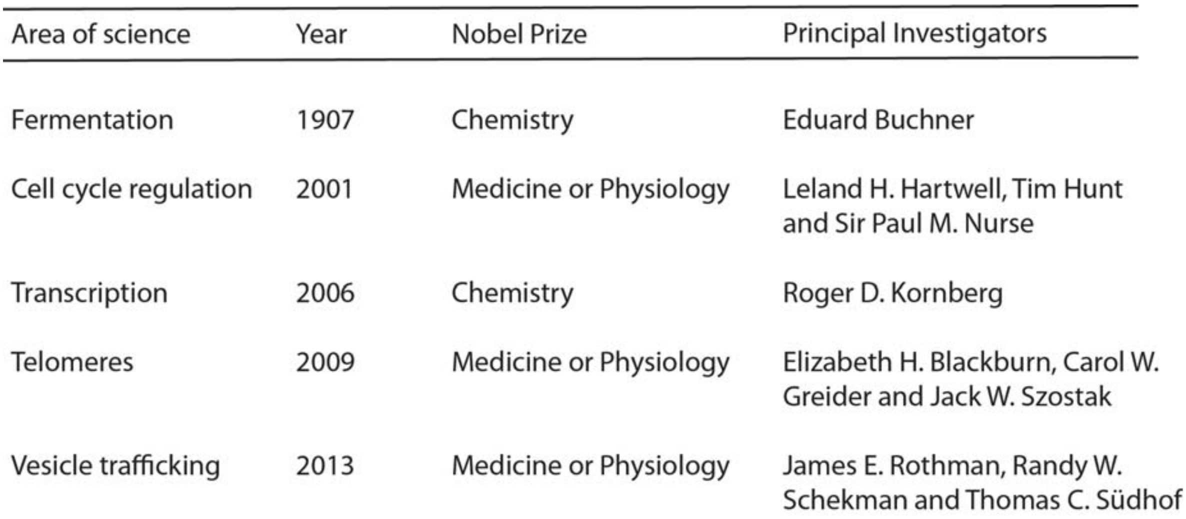


Figure 1.37: The budding yeast mating type locus
The mating type is determined by the genetic information contained within the mating type locus on Chromosome III. Yeast cells also contain inactive copies of the genetic information for both mating types in silent mating type cassettes at the ends of 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.



Name	Description	Nomenclature
YNL262W	Systematic Name for ORF	Each ORF has a systematic name. The convention is: Y (for yeast), N (chromosome number; A = chrI, B = chrII, ...), 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).
POL2^{+}	Wilde type (wt)	Italicised common gene name (three capital letters followed by numbers) with plus in superscript
POL2	Dominant allele (often used to mean wt)	Italicised common gene name (three captial letters followed by numbers)
POL2-1, *POL2-2*, etc.	Specific dominant allele	Designantion for dominant mutant allele followed by hyphen and number
pol2	Recessive allele	Three italicized lowercase letters and number
pol2-1, *pol2-2*, *pol2-3*, *pol2-4*,...	Specific recessive allele	Designantion for recessive mutant allele followed by hyphen and number
pol2Δ	Deletion of gene	Designantion for recessive mutant allele followed by Δ
Pol2p	Protein Product of gene	Three letters, with the first being uppercase, followed by a number and optional lower case p; not italicized
Pol2	Protein Product of gene (Alternative)	

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



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

Table 1.7: A selection of Nobel Prizes awarded for work using *S. cerevisiae* as a model organism.

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 tumuor 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

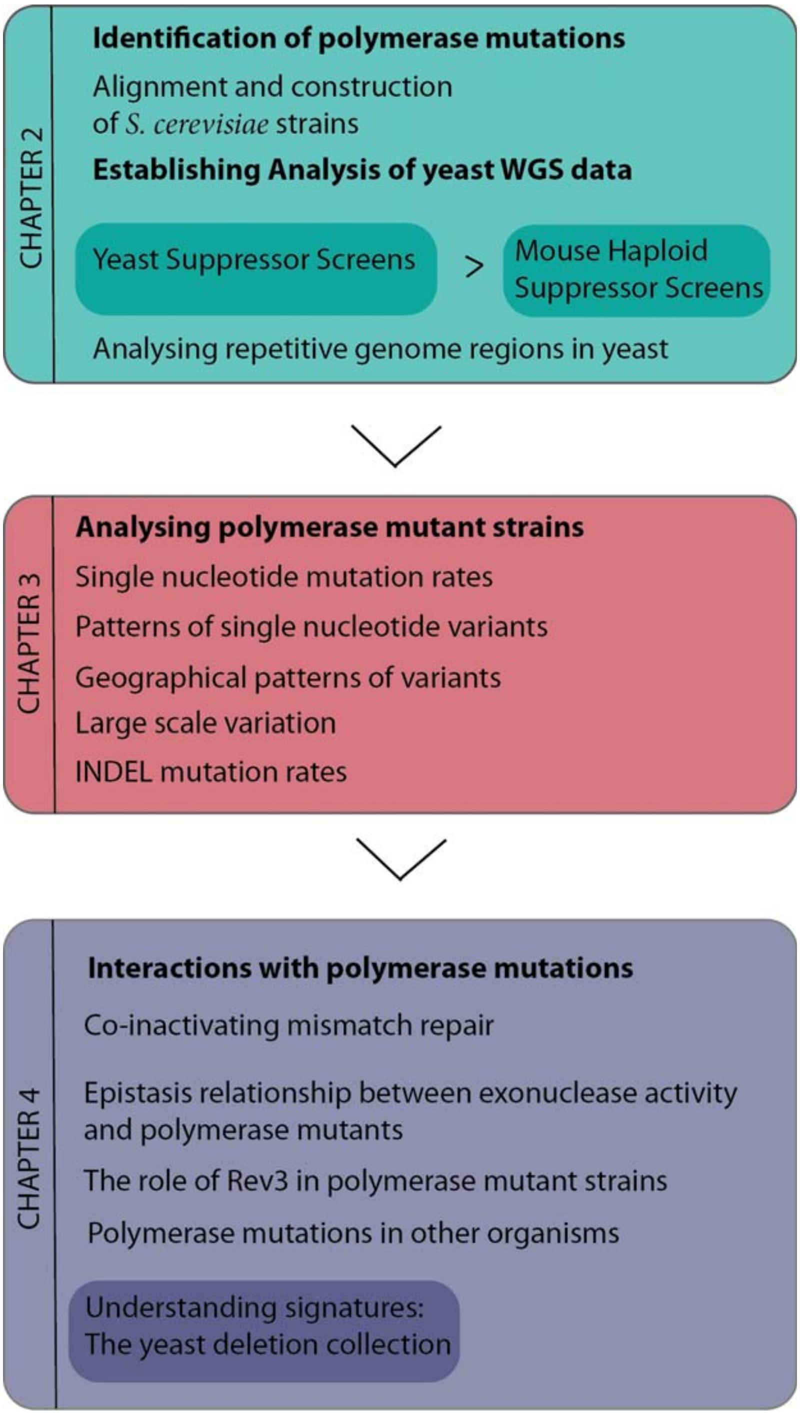

CHAPTER 2
Identification of polymerase mutations
Alignment and construction of *S. cerevisiae* strains
Establishing Analysis of yeast WGS data
Yeast Suppressor Screens
>
Mouse Haploid Suppressor Screens
Analysing repetitive genome regions in yeast
CHAPTER 3
Analysing polymerase mutant strains
Single nucleotide mutation rates
Patterns of single nucleotide variants
Geographical patterns of variants
Large scale variation
INDEL mutation rates
CHAPTER 4
Interactions with polymerase mutations
Co-inactivating mismatch repair
Epistasis relationship between exonuclease activity and polymerase mutants
The role of Rev3 in polymerase mutant strains
Polymerase mutations in other organisms
Understanding signatures: The yeast deletion collection

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 *POLE* coding for the catalytic subunit of Polε[259]. At the same time, pedigree-sequencing of families with a history of colorectal cancer identified two predisposing germline variants *POLE* L424V and *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 Polε and Polδ in yeast cause a base substitution phenotype of varying severity. Mutations affecting the catalytic residues of the proofreading domain of *POL3* (*pol3-01*) cause a mutator phenotype with increased base substitution and frameshift mutations[338]. Similar mutations in Polε (*pol2-4*) reduce proofreading activity about 100-fold *in vitro*, while leaving polymerase activity at wild-type levels[312]. *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 (Pol$\varepsilon^{\text{exo-/exo-}}$; the mouse equivalent of the yeast *pol2-4* mutation) showed a predisposition to cancer, while Pol$\varepsilon^{\text{exo-/+}}$were virtually indistinguishable from wild-type in this respect[769]. Spontaneous mutations were more frequent in Pol$\varepsilon^{\text{exo-}}$ mice than in Pol$\delta^{\text{exo-}}$ mice, in contrast to the budding yeast, where the *pol3-01* mutation causes a higher mutational frequency than *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. *Atp1a1* and *Hprt* in mice versus *URA3*, *CAN1* and *SUP4-o* in yeast), further confirming the need for improved methods to assess mutation rate increases. Mice deficient for both Polε and Polδ proofreading activity

Mutation	Assay gene	Fold change to wt	Publication
pol3-01	*his7-2*	240	[338]
pol3-01	*URA3* [a]	130	[338]
pol3-01	*URA3* [a]	52	[282]
pol3-01	*lys2::InsLD*	0.6	[308]
pol3-01	*his7-2*	74	[308]
pol3-01	*his7-2*	630	[282]
pol3-01	*CAN1*	110	[308]
pol3-01	*SUP4-o* [b]	32-106	[303]
pol3-01	*trp1–289*	100	[282]
pol3-01	*lys2::InsE* [c]	26 - 188	[807]
pol2-4	*CAN1*	5	[312]
pol2-4	*ade5-1*	43	[312]
pol2-4	*URA3* [a]	15	[282]
pol2-4	*his7-2*	24	[312]
pol2-4	*his7-2*	63	[282]
pol2-4	*leu2-1*	18	[312]
pol2-4	*hom3-10*	9	[312]
pol2-4	*his1-7*	31	[312]
pol2-4	*SUP4-o* [b]	2.9	[303]
pol2-4	*trp1–289*	3.9	[282]
pol2-4	*lys2::InsE* [c]	1.2 - 6	[807]
pol2-16	*URA3* [a]	1.6	[282]
pol2-16	*his7–2*	1.4	[282]
pol2-16	*trp1–289*	1.9	[282]

Table 2.1: Polymerase exonuclease domain mutations in *S. cerevisiae*

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[285]. [a]Forward mutation of *URA3*. [b]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*. [c]lys2::InsE alleles contain various sizes of dA homonucleotide runs. For similar experiments, see [338, 339, 808].

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 Altlas (TCGA), a project to catalogue genetic mutations responsible 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 identified 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 assembled 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 recurrence 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)

Gene	AA change	Chr	Pos(37)	Pos(38)	REF	ALT	
POLD1	p.Arg311Cys	19	50905959	50402702	C	T	[810]
POLD1	p.Gly426Ser	19	50909472	50406215	G	A	[768]
POLD1	p.Pro327Leu	19	50906319	50403062	C	T	[768]
POLD1	p.Ser370Arg	19	50906449	50403192	C	A	[768]
POLD1	p.Ser478Asn	19	50909713	50406456	G	A	[768]
POLD1	p.Val392Met	19	50906786	50403529	G	A	[810]
POLE	p.Met444Lys	12	133250189	132673603	A	T	[809]
POLE	p.Ala456Pro	12	133249857	132673271	C	G	[810]
POLE	p.Ala465Val	12	133249829	132673243	G	A	[809]
POLE	p.Arg446Gln	12	133250183	132673597	C	T	[810]
POLE	p.Asp275Val	12	133253217	132676631	T	A	[810]
POLE	p.Gln453Arg	12	133250162	132673576	T	C	[809]
POLE	p.Leu424Val	12	133250250	132673664	G	C	[768]
POLE	p.Pro286Arg	12	133253184	132676598	G	C	[810]
POLE	p.Pro436Arg	12	133250213	132673627	G	C	[809]
POLE	p.Ser297Phe	12	133253151	132676565	G	A	[810]
POLE	p.Val411Leu	12	133250289	132673703	C	A	[810]

Table 2.2: Genomic locations of mutations in DNA polymerases in different human genome assemblies

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.

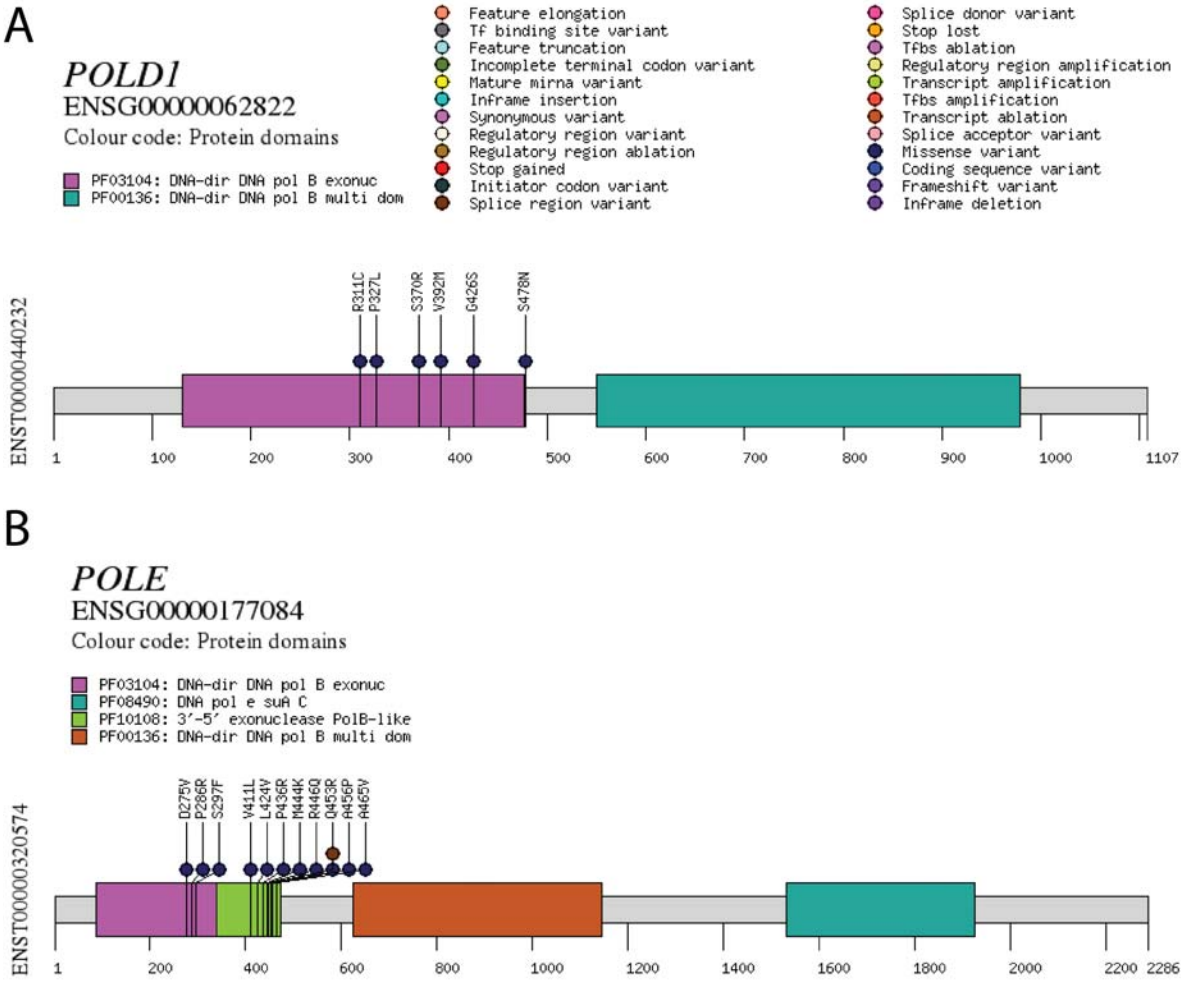


Figure 2.2: Locations of DNA polymerase mutations within the proteins
The locations of the mutations within the protein with reference to the domain structure is given. Plot was generated by Dr. Carla Daniela Robles Espinoza using a custom written script.

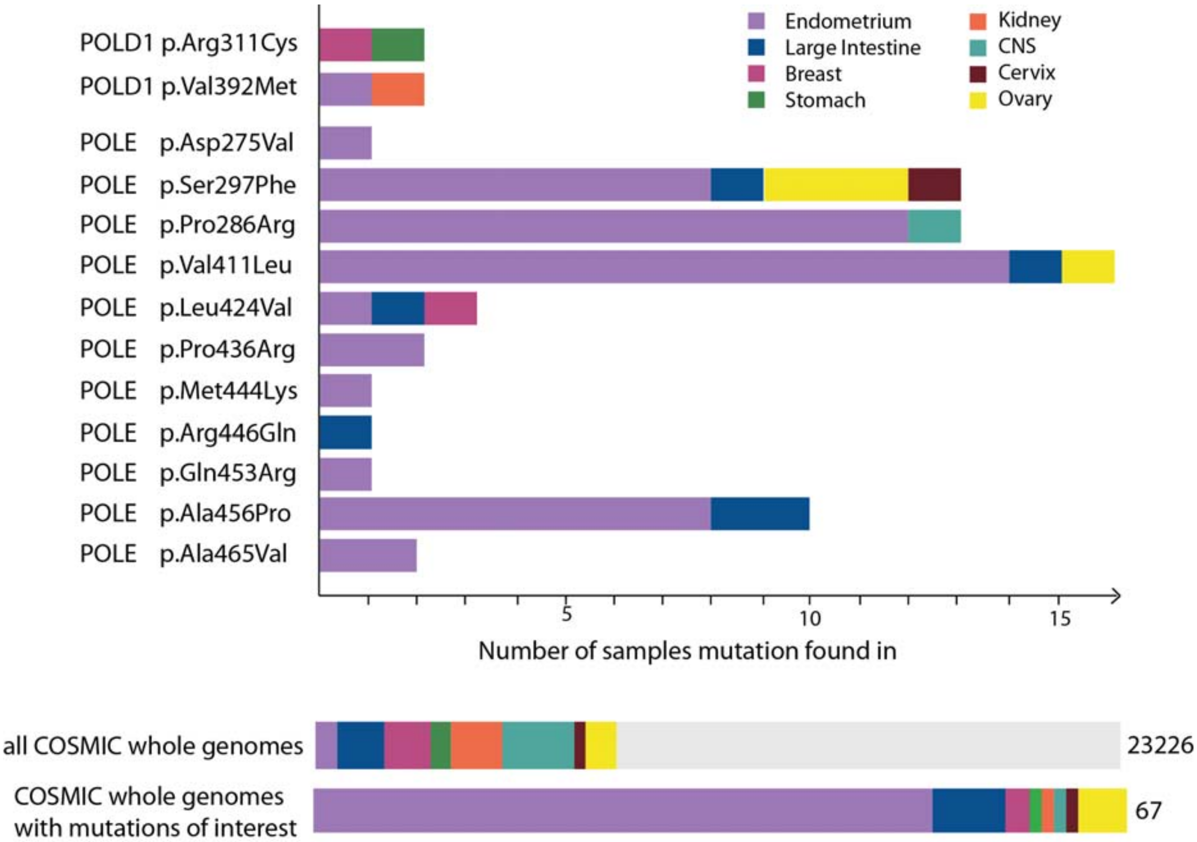


Figure 2.3: Prevalence of polymerase mutations of interest in COSMIC
The list of DNA polymerase mutations were cross-referenced with the COSMIC whole-genome data (v74)[812]. Recurrence of mutations in the whole dataset is displayed with information about the tissue of origin. For comparison, the composition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

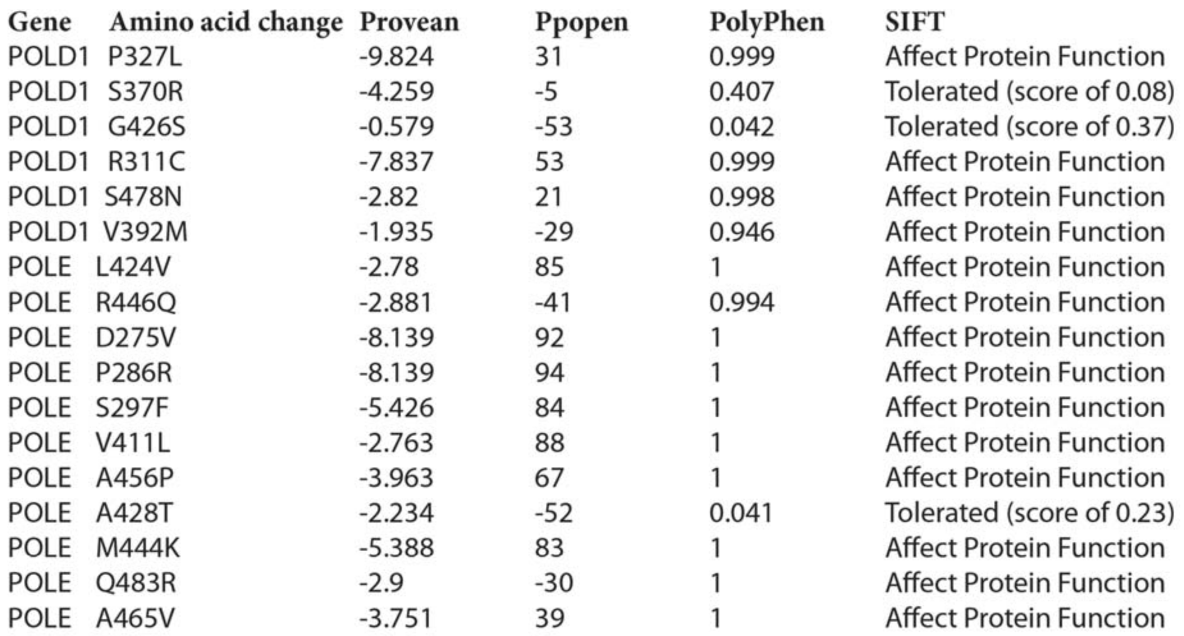
is a vast database of somatic changes observed in human cancer samples[812]. To assess the prevalence of these mutations, I accessed their curation of 22,690 whole cancer genomes and analysed mutation recurrence and tumour origin(Fig. 2.3). Recurrence indicates that DNA polymerase mutations to prioritise 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-

Gene	AA change	dbSNP	1000Genomes	500Exomes & CGP
POLD1	p.Arg311Cys	rs201010746 T=0.00001 (ExAC)	T=0.0002/1	rs201010746
POLD1	p.Gly426Ser	-	-	lowQual
POLD1	p.Pro327Leu	rs397514633 (OMIM)	-	-
POLD1	p.Ser370Arg	-	-	-
POLD1	p.Ser478Asn	rs397514632 (OMIM)	-	-
POLD1	p.VAL392Met	rs778843530 A=0.000008 (ExAC)	-	-
POLE	p.Met444Lys	-	-	-
POLE	p.Ala456Pro	-	-	-
POLE	p.Ala465Val	-	-	-
POLE	p.Arg446Gln	rs151273553 T=0.0003 (ExAC)	-	-
POLE	p.Asp275Val	-	-	-
POLE	p.Gln453Arg	-	-	-
POLE	p.Leu424Val	rs483352909 A=0.000008 (ExAC)	-	-
POLE	p.Pro286Arg	-	-	-
POLE	p.Pro436Arg	-	-	-
POLE	p.Ser297Phe	-	-	-
POLE	p.Val411Leu	-	-	-

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 dbSNP 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.



Gene	Amino acid change	Provean	Ppopen	PolyPhen	SIFT
POLD1	P327L	-9.824	31	0.999	Affect Protein Function
POLD1	S370R	-4.259	-5	0.407	Tolerated (score of 0.08)
POLD1	G426S	-0.579	-53	0.042	Tolerated (score of 0.37)
POLD1	R311C	-7.837	53	0.999	Affect Protein Function
POLD1	S478N	-2.82	21	0.998	Affect Protein Function
POLD1	V392M	-1.935	-29	0.946	Affect Protein Function
POLE	L424V	-2.78	85	1	Affect Protein Function
POLE	R446Q	-2.881	-41	0.994	Affect Protein Function
POLE	D275V	-8.139	92	1	Affect Protein Function
POLE	P286R	-8.139	94	1	Affect Protein Function
POLE	S297F	-5.426	84	1	Affect Protein Function
POLE	V411L	-2.763	88	1	Affect Protein Function
POLE	A456P	-3.963	67	1	Affect Protein Function
POLE	A428T	-2.234	-52	0.041	Tolerated (score of 0.23)
POLE	M444K	-5.388	83	1	Affect Protein Function
POLE	Q483R	-2.9	-30	1	Affect Protein Function
POLE	A465V	-3.751	39	1	Affect Protein Function

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”.

ware that employ strategies from evolutionary sequence comparisons to structure-based predictions: PROVEAN/SIFT [815–819], Poly-phen2 [820–823], PredictProtein(PPopen) [824], Mechismo [825] and Mutation Taster [826]. When considering all the scores for one mutation combined, *POLE* S297F, *POLE* P286R, *POLE* V411L and *POLE* A456P score as highly damaging to protein function across different software tools.

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

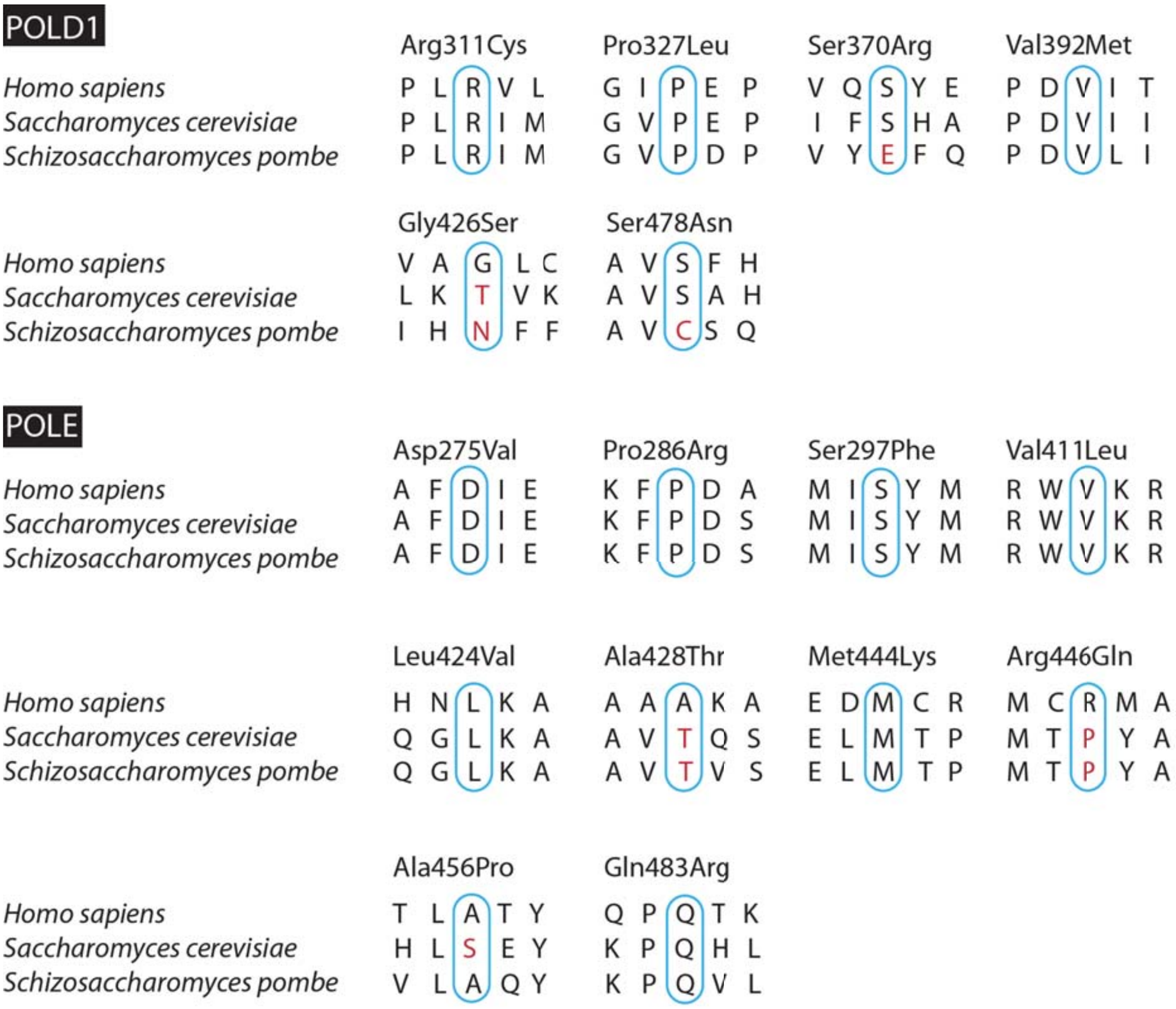


Figure 2.4: Alignment of polymerase residues of interest to the yeast proteins Sequences were aligned using Clustal Omega version 1.2.1[827–829]. Sequences used for alignment (uniprot ID in parenthesis): *Homo sapiens* POLE (Q07864), *Saccharomyces cerevisiae POL2* (P21951), *Homo sapiens POLD1* (P28340), *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.

Human variant	Conserved	*S. cerevisiae* variant
POLD1 p.Arg311Cys	Yes	*pol3 p.Arg3116Cys*
POLD1 p.Gly426Ser	No, (T)	-
POLD1 p.Pro327Leu	Yes	*pol3 p.Pro322Leu*
POLD1 p.Ser370Arg	Yes	*pol3 p.Ser375Arg*
POLD1 p.Ser478Asn	Yes	*pol3 p.Ser483Asn*
POLE p.Met444Lys	Yes	*pol2 p.Met459Lys*
POLE p.Ala456Pro	No, (S)	-
POLE p.Ala428Thr	No, (T)	-
POLE p.Ala465Val	Yes	*pol2 p.Ala480Val*
POLE p.Arg446Gln	No, (P)	-
POLE p.Asp275Val	Yes	*pol2 p.Asp290Val*
POLE p.Gln453Arg	Yes	*pol2 p.Gln468Arg*
POLE p.Leu424Val	Yes	*pol2 p.Leu439Val*
POLE p.Pro286Arg	Yes	*pol2 p.Pro301Arg*
POLE p.Ser297Phe	Yes	*pol2 p.Ser312Phe*
POLE p.Val411Leu	Yes	*pol2 p.Val426Leu*

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.

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^r) or canavanine (Can^r) 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 frameshifts (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-

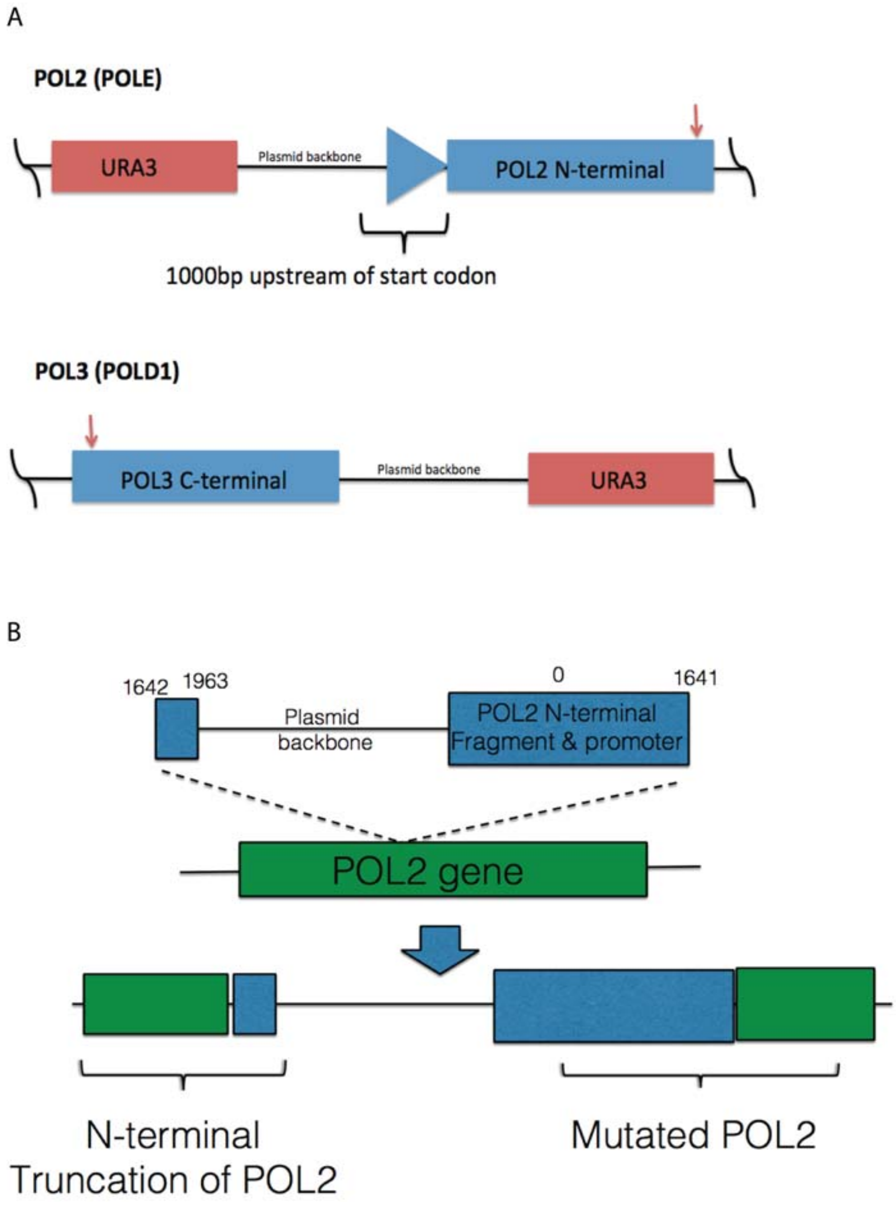
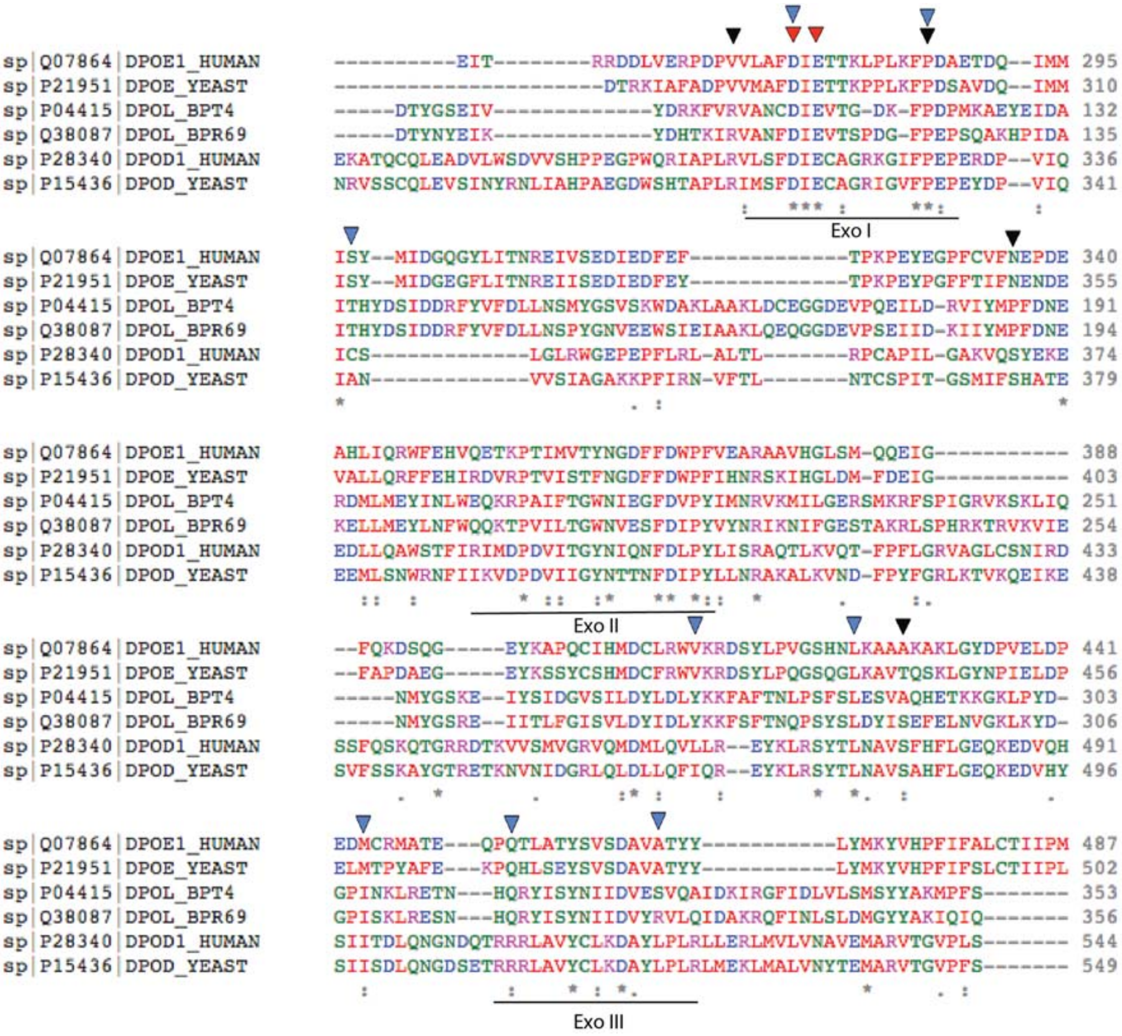


Figure 2.5: Rationale for plasmid construction

A| Two different types of vectors were designed - one for *POL2* and on 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 linerarisation. The red arrows denote approximate sites for linerarisation, 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.



```
sp|Q07864|DPOE1_HUMAN      ----------EIT--------RRDDLVERPDPVVLAFDIETTKLPLKFPDAETDQ--IMM 295
sp|P21951|DPOE_YEAST       ----------------------DTRKIAFADPVVMAFDIETTKPPLKFPDSAVDQ--IMM 310
sp|P04415|DPOL_BPT4        -----DTYGSEIV-------------YDRKFVRVANCDIEVTG-DK-FPDPMKAEYEIDA 132
sp|Q38087|DPOL_BPR69       -----DTYNYEIK-------------YDHTKIRVANFDIEVTSPDG-FPEPSQAKHPIDA 135
sp|P28340|DPOD1_HUMAN      EKATQCQLEADVLWSDVVSHPPEGPWQRIAPLRVLSFDIECAGRKGIFPEPERDP--VIQ 336
sp|P15436|DPOD_YEAST       NRVSSCQLEVSINYRNLIAHPAEGDWSHTAPLRIMSFDIECAGRIGVFPEPEYDP--VIQ 341
                                                            :   *** :     **:       :

sp|Q07864|DPOE1_HUMAN      ISY--MIDGQGYLITNREIVSEDIEDFEF-------------TPKPEYEGPFCVFNEPDE 340
sp|P21951|DPOE_YEAST       ISY--MIDGEGFLITNREIISEDIEDFEY-------------TPKPEYPGFFTIFNENDE 355
sp|P04415|DPOL_BPT4        ITHYDSIDDRFYVFDLLNSMYGSVSKWDAKLAAKLDCEGGDEVPQEILD-RVIYMPFDNE 191
sp|Q38087|DPOL_BPR69       ITHYDSIDDRFYVFDLLNSPYGNVEEWSIEIAAKLQEQGGDEVPSEIID-KIIYMPFDNE 194
sp|P28340|DPOD1_HUMAN      ICS-------------LGLRWGEPEPFLRL-ALTL-------RPCAPIL-GAKVQSYEKE 374
sp|P15436|DPOD_YEAST       IAN-------------VVSIAGAKKPFIRN-VFTL-------NTCSPIT-GSMIFSHATE 379
                           *                       . :                                *

sp|Q07864|DPOE1_HUMAN      AHLIQRWFEHVQETKPTIMVTYNGDFFDWPFVEARAAVHGLSM-QQEIG----------- 388
sp|P21951|DPOE_YEAST       VALLQRFFEHIRDVRPTVISTFNGDFFDWPFIHNRSKIHGLDM-FDEIG----------- 403
sp|P04415|DPOL_BPT4        RDMLMEYINLWEQKRPAIFTGWNIEGFDVPYIMNRVKMILGERSMKRFSPIGRVKSKLIQ 251
sp|Q38087|DPOL_BPR69       KELLMEYLNFWQQKTPVILTGWNVESFDIPYVYNRIKNIFGESTAKRLSPHRKTRVKVIE 254
sp|P28340|DPOD1_HUMAN      EDLLQAWSTFIRIMDPDVITGYNIQNFDLPYLISRAQTLKVQT-FPFLGRVAGLCSNIRD 433
sp|P15436|DPOD_YEAST       EEMLSNWRNFIIKVDPDVIIGYNTTNFDIPYLLNRAKALKVND-FPYFGRLKTVKQEIKE 438
                             ::  :        * ::  :*   ** *::  *      .     :.

sp|Q07864|DPOE1_HUMAN      --FQKDSQG-----EYKAPQCIHMDCLRWVKRDSYLPVGSHNLKAAAKAKLGYDPVELDP 441
sp|P21951|DPOE_YEAST       --FAPDAEG-----EYKSSYCSHMDCFRWVKRDSYLPQGSQGLKAVTQSKLGYNPIELDP 456
sp|P04415|DPOL_BPT4        -----NMYGSKE--IYSIDGVSILDYLDLYKKFAFTNLPSFSLESVAQHETKKGKLPYD- 303
sp|Q38087|DPOL_BPR69       -----NMYGSRE--IITLFGISVLDYIDLYKKFSFTNQPSYSLDYISEFELNVGKLKYD- 306
sp|P28340|DPOD1_HUMAN      SSFQSKQTGRRDTKVVSMVGRVQMDMLQVLLR--EYKLRSYTLNAVSFHFLGEQKEDVQH 491
sp|P15436|DPOD_YEAST       SVFSSKAYGTRETKNVNIDGRLQLDLLQFIQR--EYKLRSYTLNAVSAHFLGEQKEDVHY 496
                                .  *       .      :* :    :       *  *.  :           .

sp|Q07864|DPOE1_HUMAN      EDMCRMATE---QPQTLATYSVSDAVATYY-----------LYMKYVHPFIFALCTIIPM 487
sp|P21951|DPOE_YEAST       ELMTPYAFE---KPQHLSEYSVSDAVATYY-----------LYMKYVHPFIFSLCTIIPL 502
sp|P04415|DPOL_BPT4        GPINKLRETN---HQRYISYNIIDVESVQAIDKIRGFIDLVLSMSYYAKMPFS------- 353
sp|Q38087|DPOL_BPR69       GPISKLRESN---HQRYISYNIIDVYRVLQIDAKRQFINLSLDMGYYAKIQIQ------- 356
sp|P28340|DPOD1_HUMAN      SIITDLQNGNDQTRRRLAVYCLKDAYLPLRLLERLMVLVNAVEMARVTGVPLS------- 544
sp|P15436|DPOD_YEAST       SIISDLQNGDSETRRRLAVYCLKDAYLPLRLMEKLMALVNYTEMARVTGVPFS------- 549
                             :           :    * : *.                  *     . :
```

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): *Homo sapiens POLE* (Q07864), *Saccharomyces cerevisiae* POL2 (P21951), *Homo sapiens POLD1* (P28340), *Saccharomyces cerevisiae* POL3 (P15436), *Enterobacteria phage T4* 43 (P04415), *Enterobacteria phage RB69* 43 (Q38087), *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 *POLD1* are highlighted by black triangles, those in *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 Can^r 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-

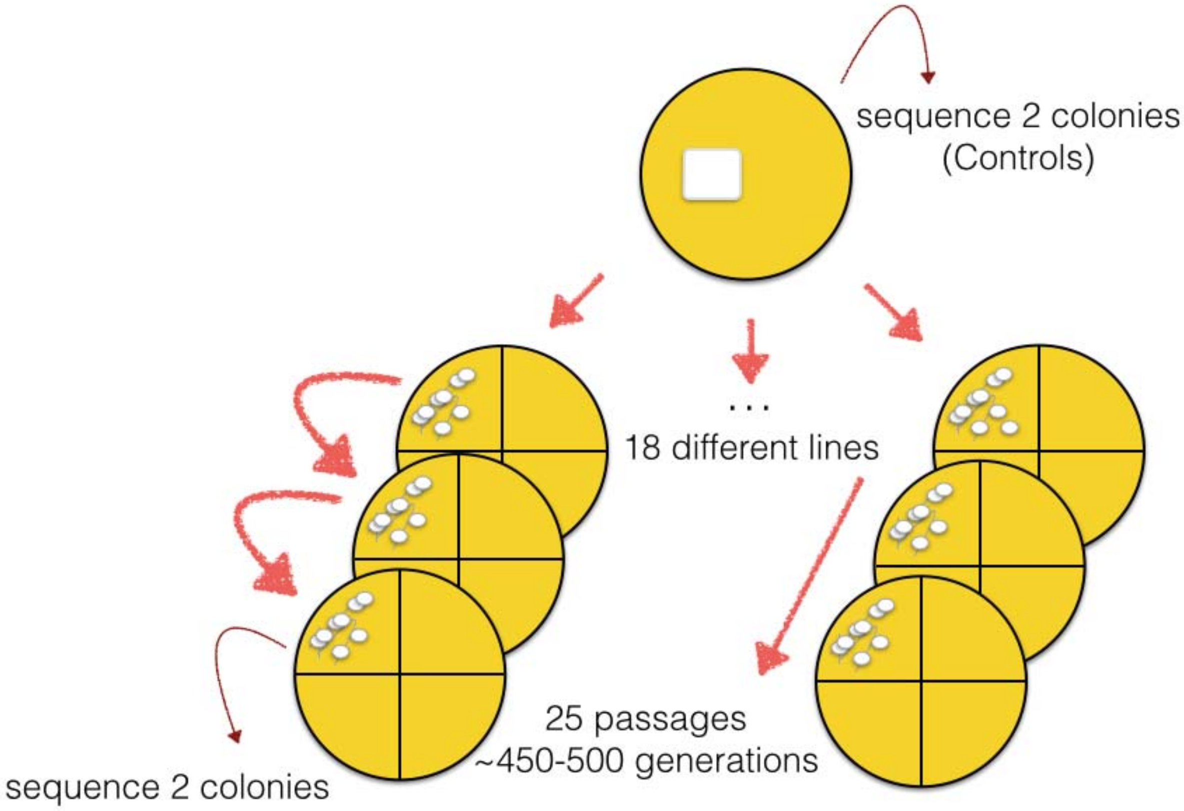


Figure 2.7: Mutation accumulation experiment: manual propagation of mutated *S. cerevisiae* strains
Experimental strategy: the heterozygous diploid polymerase mutant strains (all derived from the same wild-type W303 strain) were patched onto YPAD. From each patch 18 different parallel mutation accumulation 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 Resequencing 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

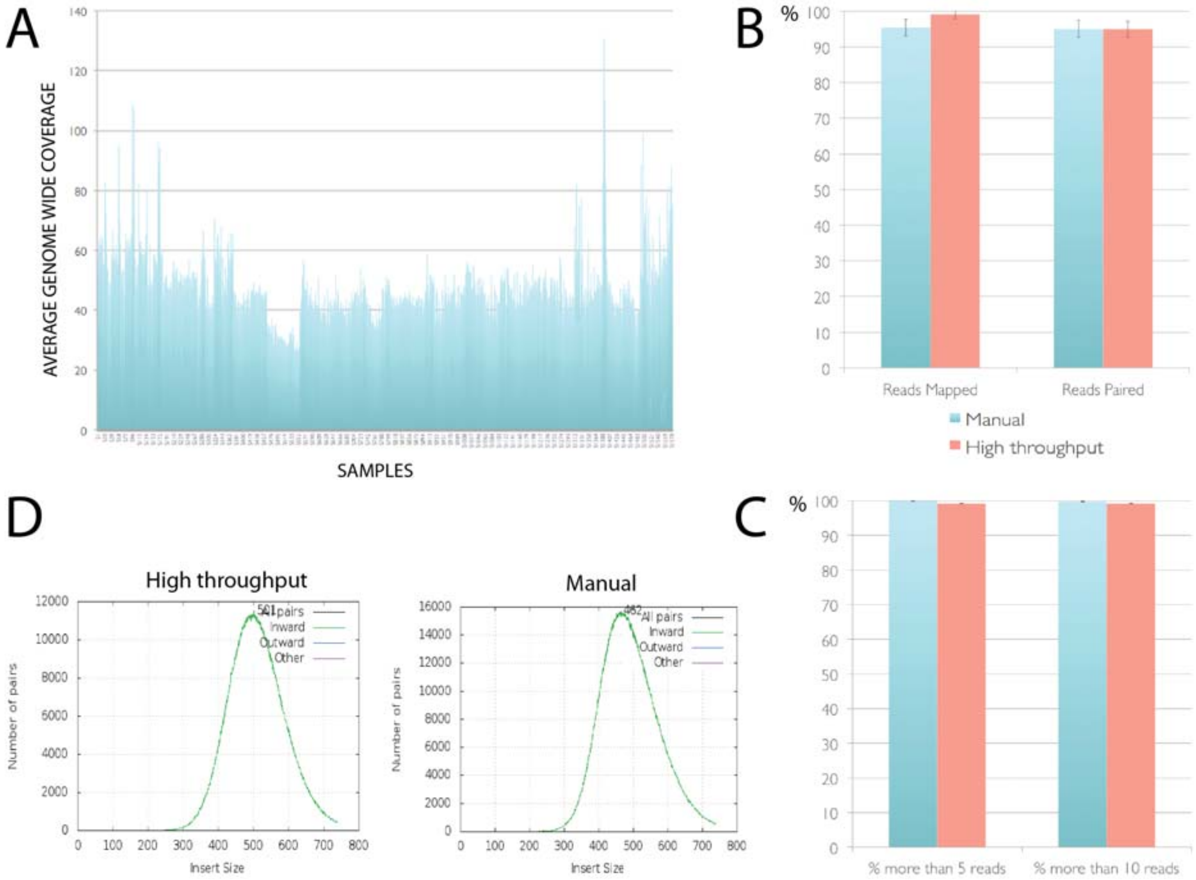


Figure 2.8: DNA extracted using a high-throughput protocol produces high quality sequencing data

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.

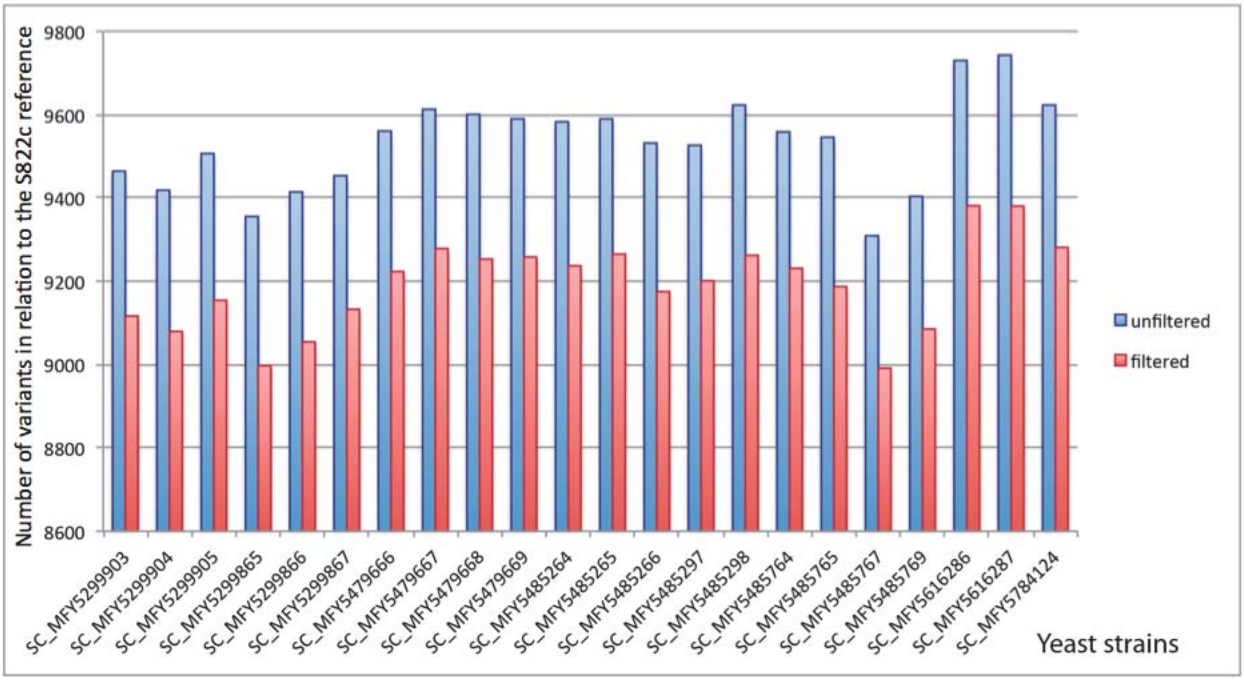


Figure 2.9: The number of variants in W303 strains compared to the S288c reference genome Aligned sequencing data from 22 *S. cerevisiae* strains of the W303 background strains was used to identify the number of background mutations to be expected when sequencing W303 *S. cerevisiae* strains. The samples 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.1.1.1) 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: 0.33×10^{-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

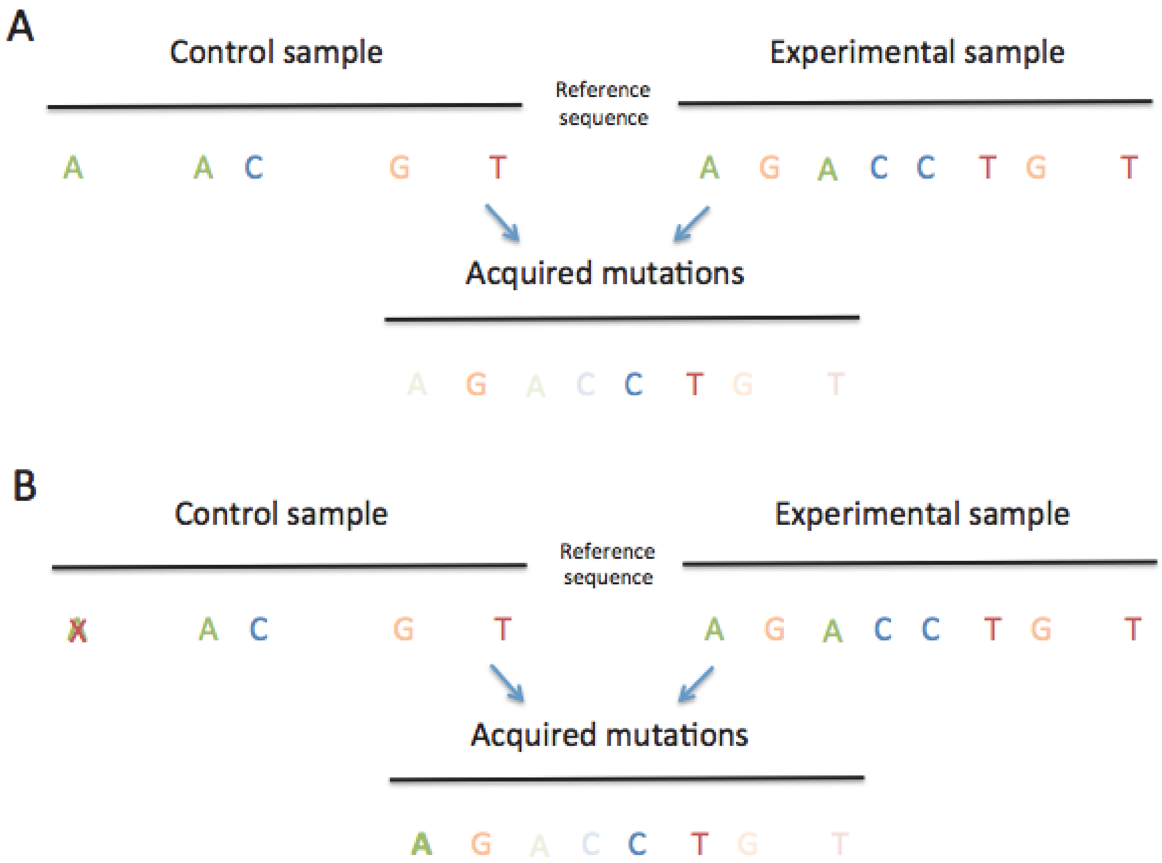


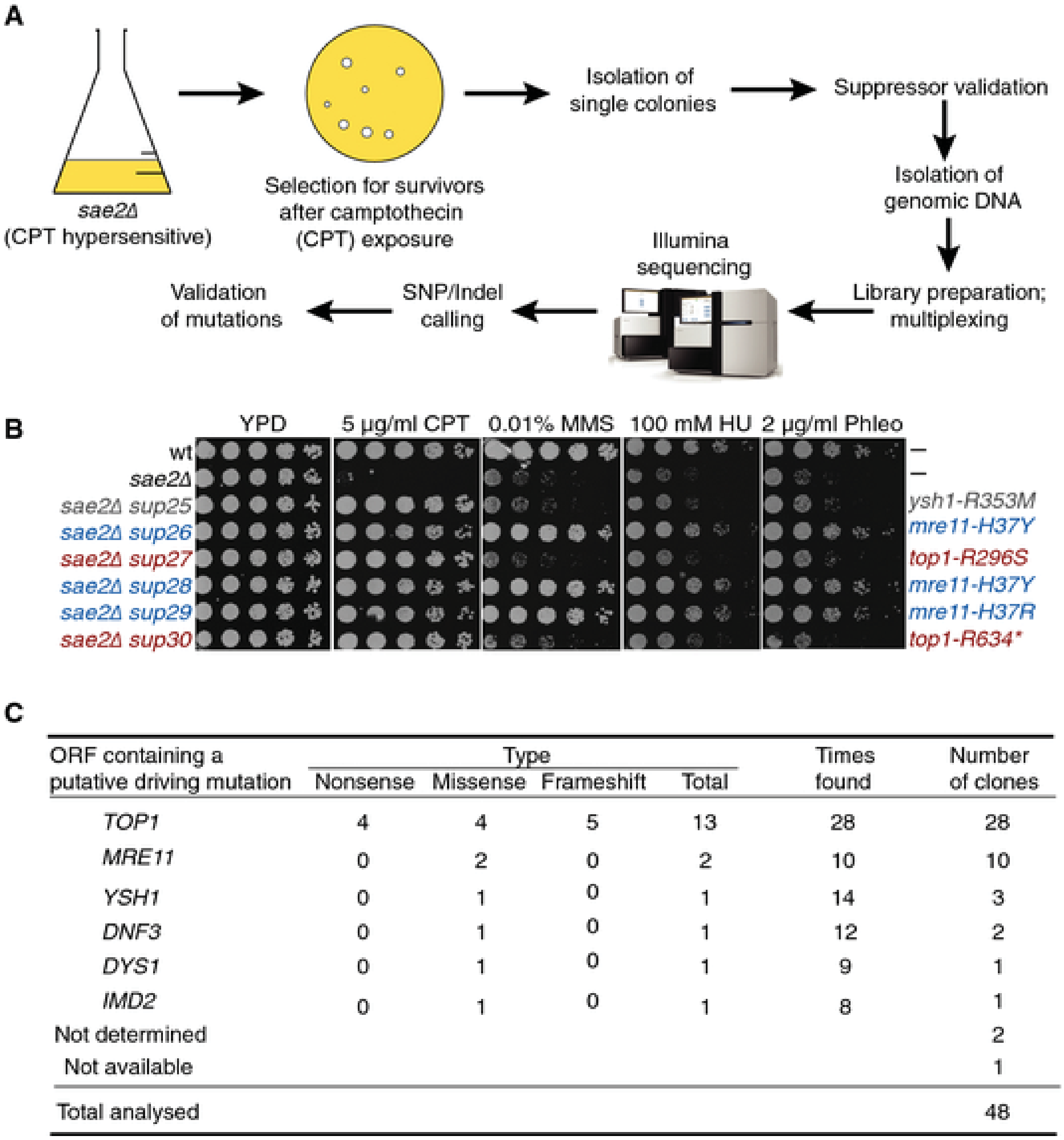
Figure 2.10: Experimental strategy to identify acquired mutations
A| In most sequencing experiments performed in this work, single nucleotide variants and small INDELs have been called with respect to a reference genome for both a control (pre-treatment) and an experimental (post-treatment) sample. The 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-occuring 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 *sae2*Δ cells - Dr. Tobias Oelschlägel performed a synthetic viability genomic screening identifying *sae2*Δ 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



ORF containing a putative driving mutation	Type				Times found	Number of clones
	Nonsense	Missense	Frameshift	Total		
TOP1	4	4	5	13	28	28
MRE11	0	2	0	2	10	10
YSH1	0	1	0	1	14	3
DNF3	0	1	0	1	12	2
DYS1	0	1	0	1	9	1
IMD2	0	1	0	1	8	1
Not determined						2
Not available						1
Total analysed						48

Figure 2.11: Sequencing analysis identifies mutations capable of suppressing *sae2*Δ 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 sae2Δ cells and sequenced 16 suppressors of tof1Δ 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.

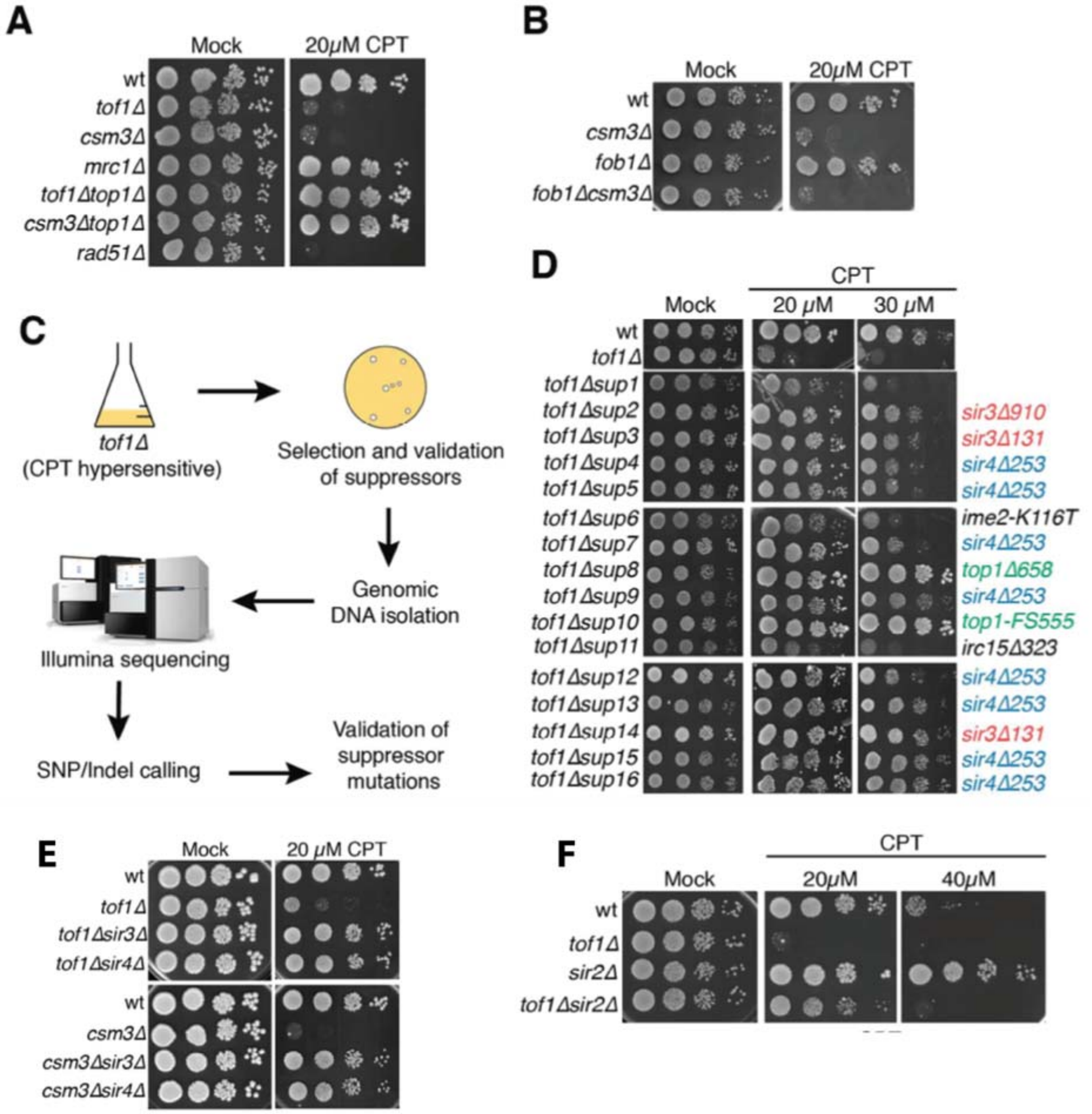


Figure 2.12: Mutations in *SIR3* and *SIR4* identified as the cause for the hypersensitivity of *tof1*Δ cells to camptothecin

(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 *tof1*Δ strains. **(E)** sir3 and sir4 deletions suppress the hypersensitivity of *tof1*Δ cells. **(F)** Deletion of SIR2 (encoding the third member of complexes containing Sir3p and Sir4p) also suppresses the hypersensitivity of *tof1*Δ 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.

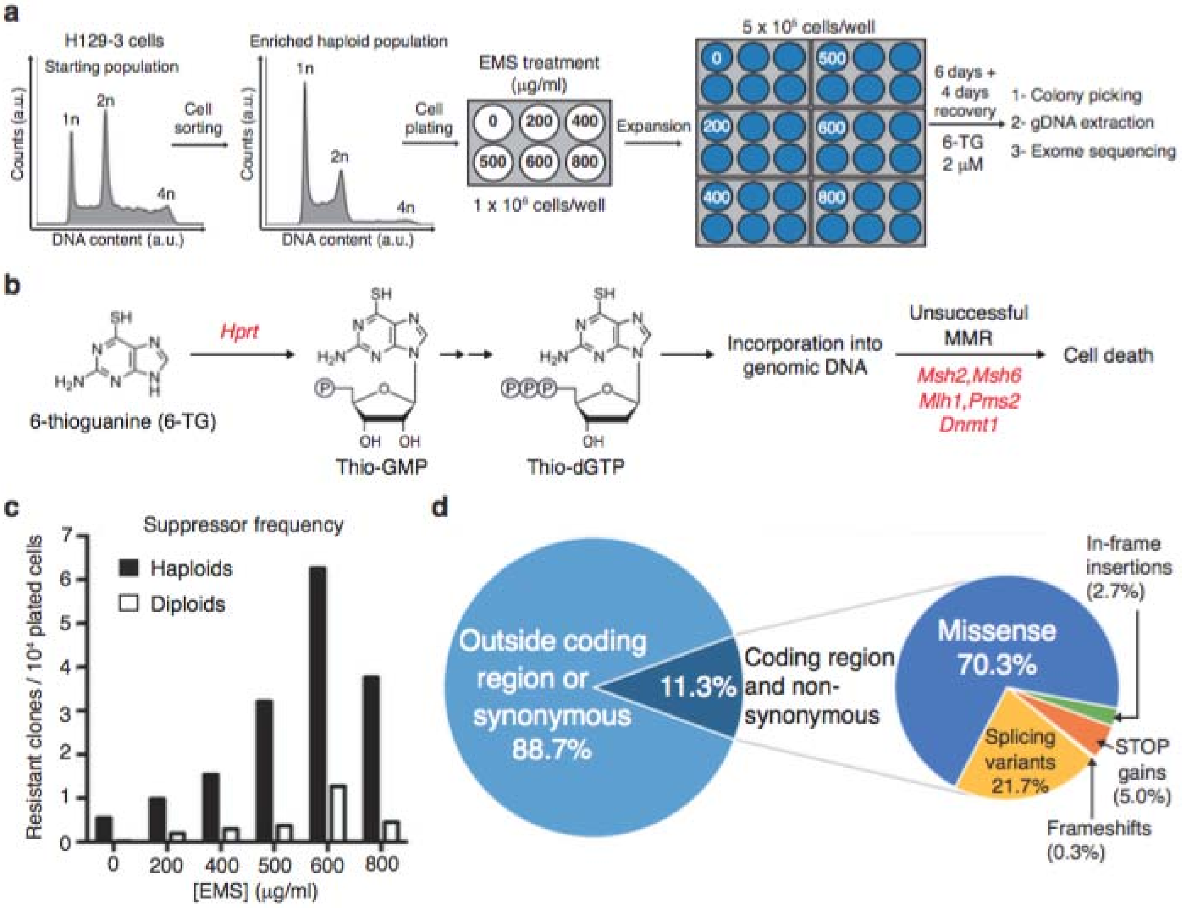
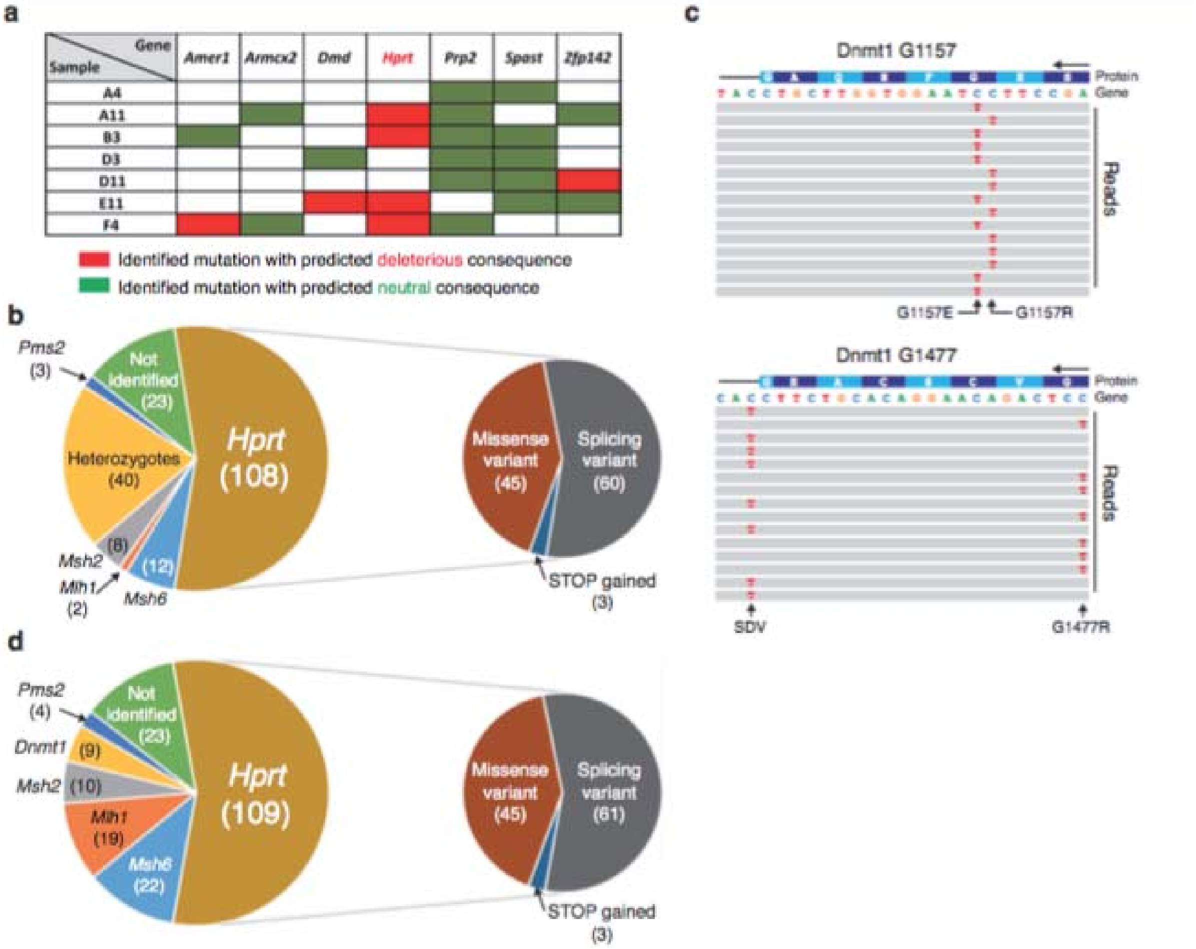


Figure 2.13: Generation of mutagenized libraries
(a) Experimental workflow. (b) Schematic of 6-TG metabolism and genotoxicity. Inactivating mutations in the genes highlighted in red have been shown to confer resistance to 6-TG. (c) Number of suppressors recovered at increasing concentrations of ethyl methanesulfonate (EMS) treatment. (d) Mutation consequences identified by whole-exome sequencing of 7 suppressor clones.

2.4.4 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-tioguanine (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



Sample \ Gene	*Amer1*	*Armcx2*	*Dmd*	*Hprt*	*Prp2*	*Spast*	*Zfp142*
A4					🟩	🟩	
A11		🟩		🟥	🟩		🟩
B3	🟩			🟥	🟩	🟩	
D3			🟩		🟩	🟩	
D11					🟩	🟩	🟥
E11			🟥	🟥		🟩	🟩
F4	🟥	🟩		🟥	🟩		

🟥 Identified mutation with predicted deleterious consequence
🟩 Identified mutation with predicted neutral consequence

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 suppres-

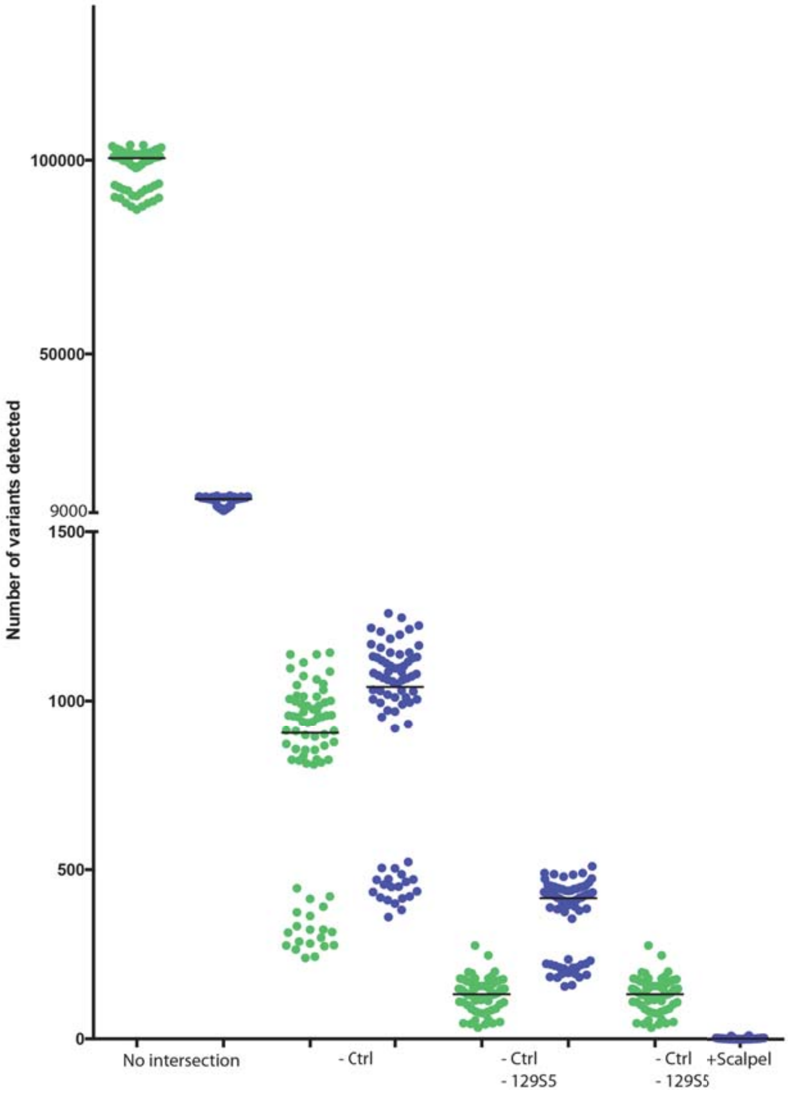
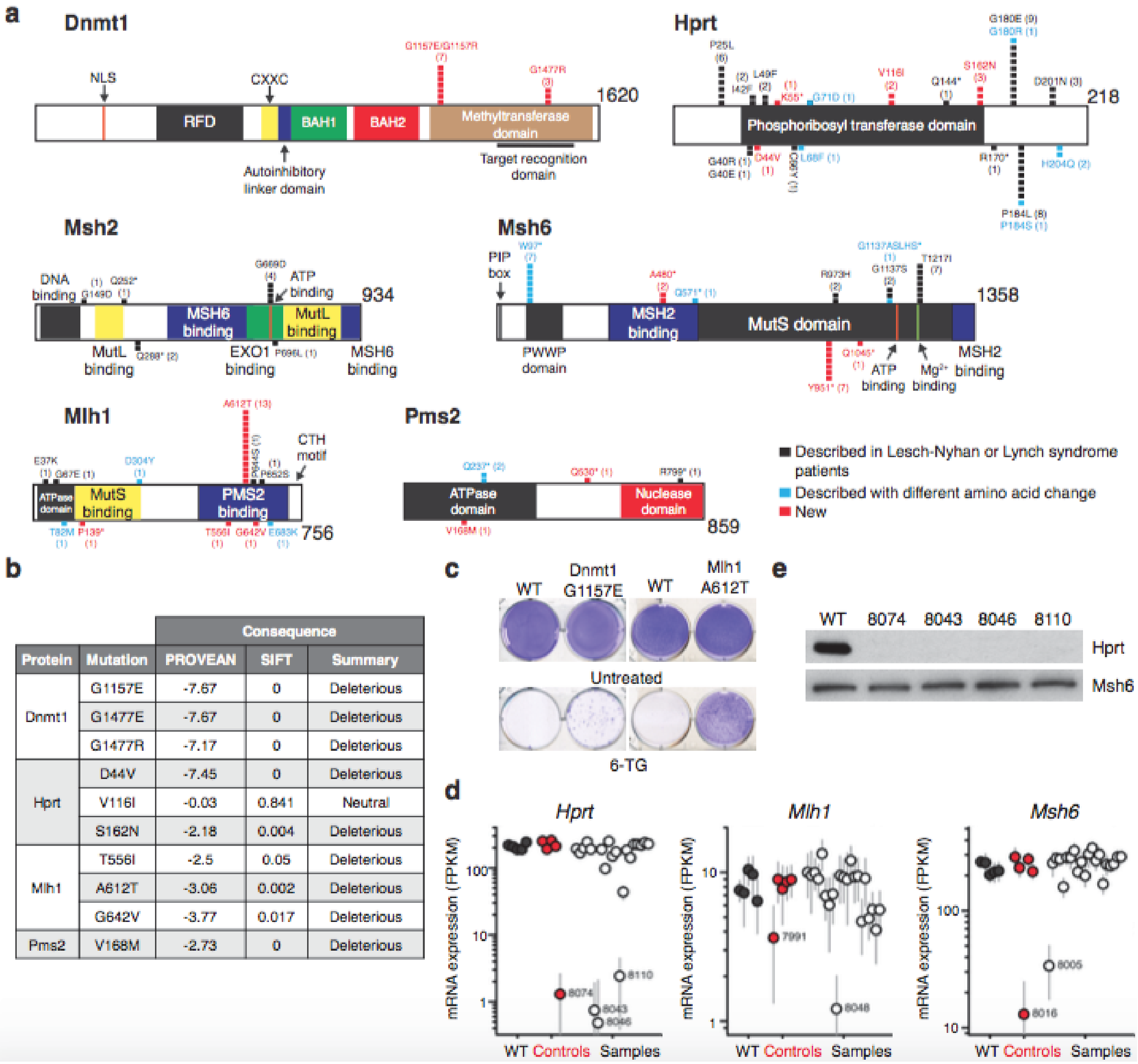


Figure 2.15: Using multiple controls and multiple variant callers to enrich for high confidence variants
Effects of using more than a sequenced control sample to clear samples of background mutations and using more than one INDEL caller to enrich for high confidence INDEL calls using 74 WES mouse samples. 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].



		Consequence		
Protein	Mutation	PROVEAN	SIFT	Summary
Dnmt1	G1157E	-7.67	0	Deleterious
	G1477E	-7.67	0	Deleterious
	G1477R	-7.17	0	Deleterious
Hprt	D44V	-7.45	0	Deleterious
	V116I	-0.03	0.841	Neutral
	S162N	-2.18	0.004	Deleterious
Mlh1	T556I	-2.5	0.05	Deleterious
	A612T	-3.06	0.002	Deleterious
	G642V	-3.77	0.017	Deleterious
Pms2	V168M	-2.73	0	Deleterious

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

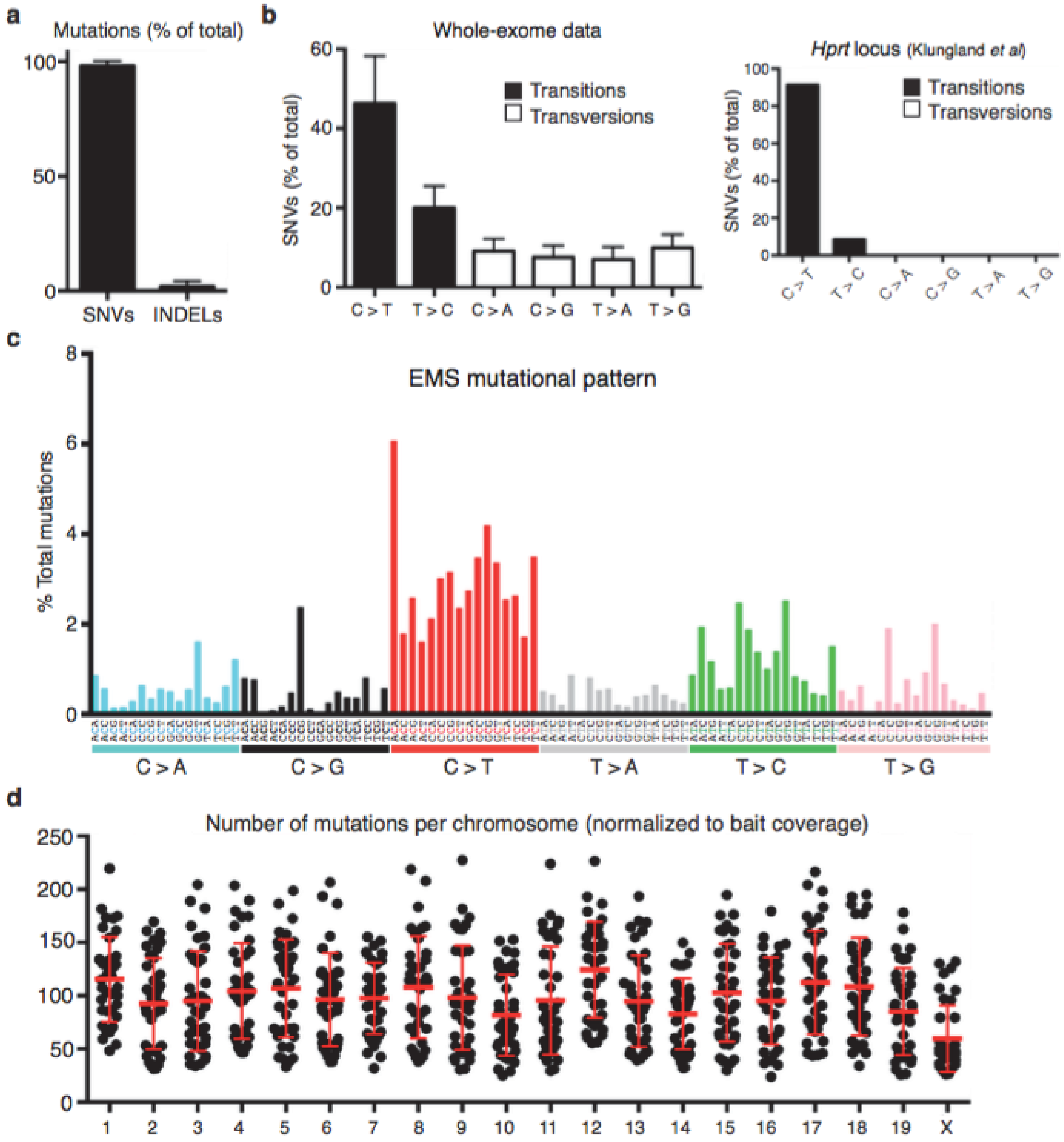


Figure 2.17: EMS mutagenic action
(a) Distribution of mutation types identified by whole-exome sequencing of 66 suppressor clones. SNV, single-nucleotide variant. INDEL, insertion or deletion. Only homozygous mutations were considered. (b) Distribution of identified SNVs. (c) EMS mutational pattern. (d) Number of mutations per chromosome in sequenced clones. Mutation numbers (both homozygous and heterozygous) were normalized to exon bait coverage.

(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

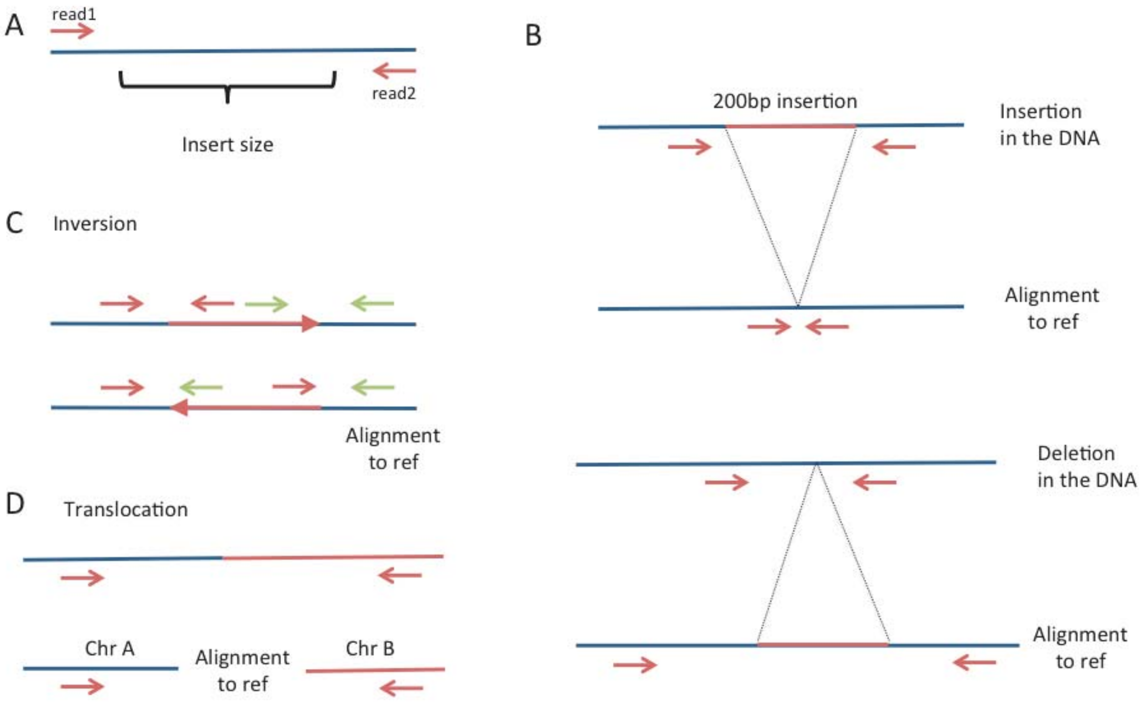


Figure 2.18: Relationship between read pairs and structural variants
A | Schematic of Illumina paired-end sequencing: a fragment of DNA is sequenced from both sides inwards for 150bp (may vary depending on sequencing machine) leaving a fragment in the middle unsequenced. Its size depends on the library 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 inversions 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

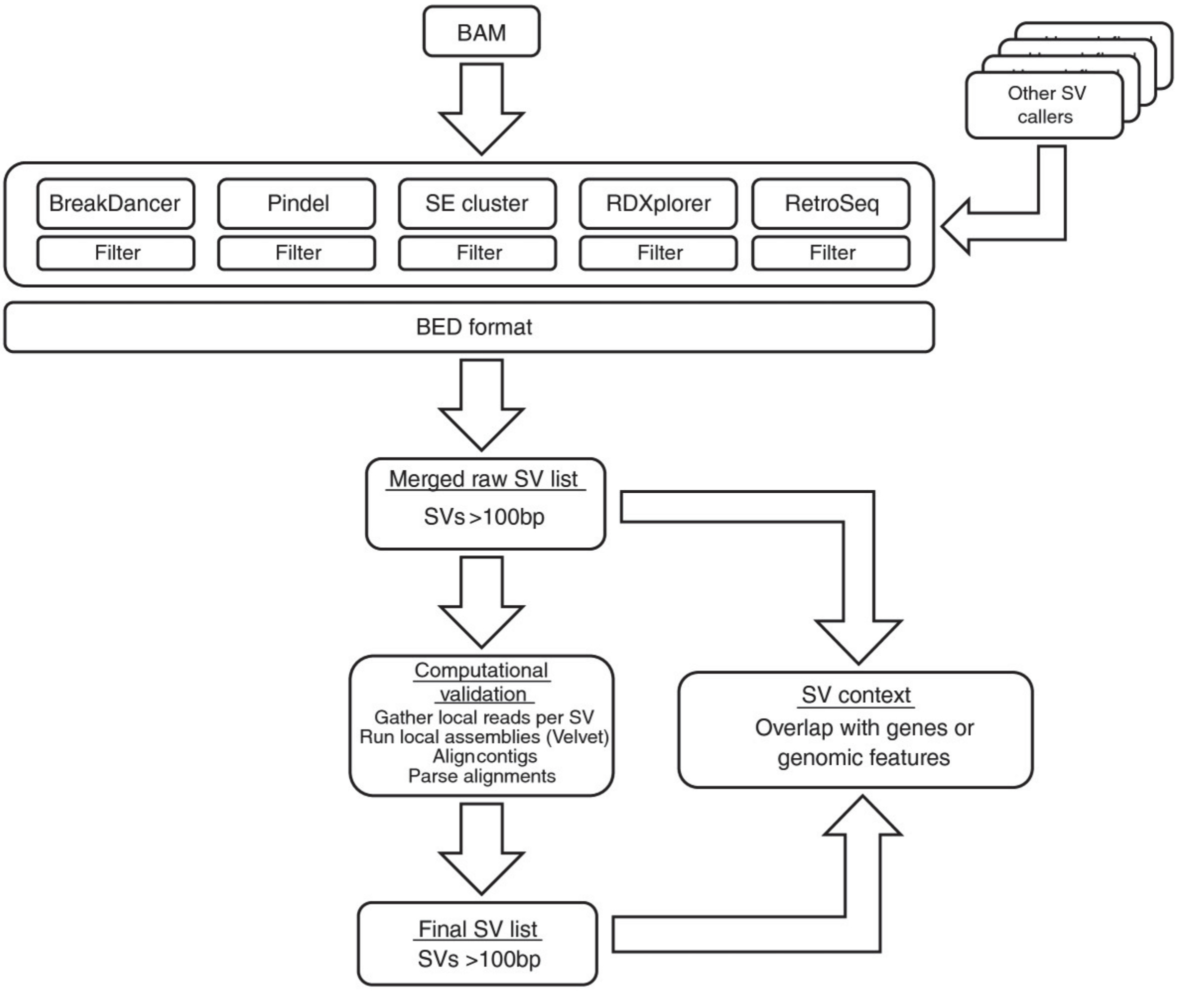


Figure 2.19: An overview of the SVMerge pipeline
"SVMerge uses a suite of software tools to detect structural variants (SVs) from mapped reads. The calls are filtered, merged and then validated computationally by local *de novo* assembly. The output is in BED format, allowing for easy downstream analysis 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.

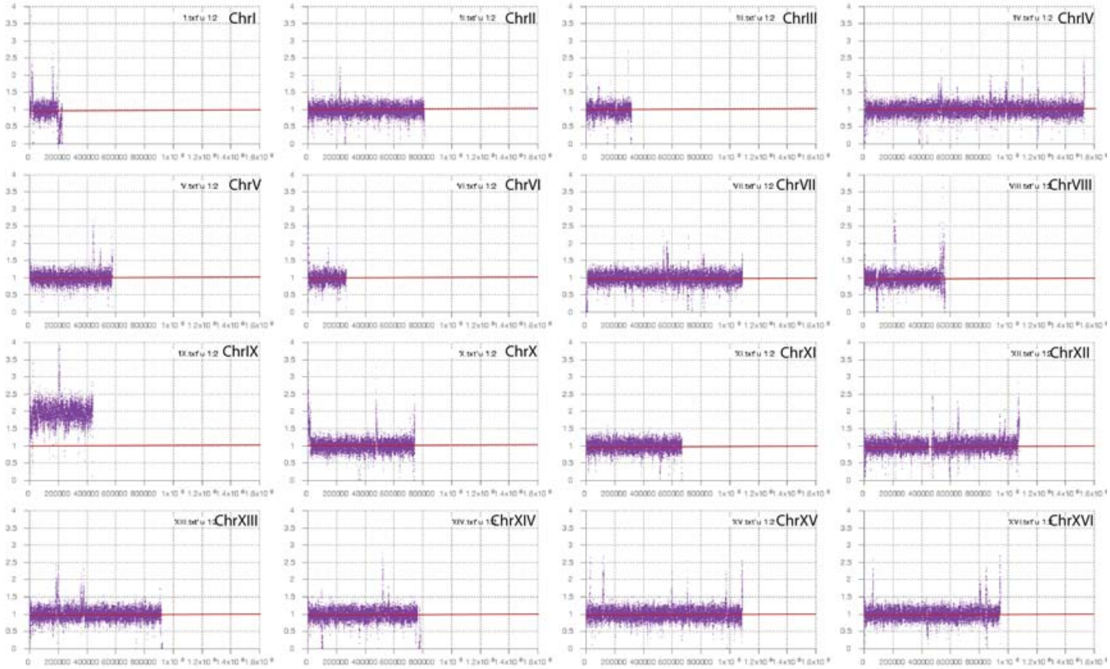


Figure 2.20: Visualising aneuploidy in budding yeast

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.

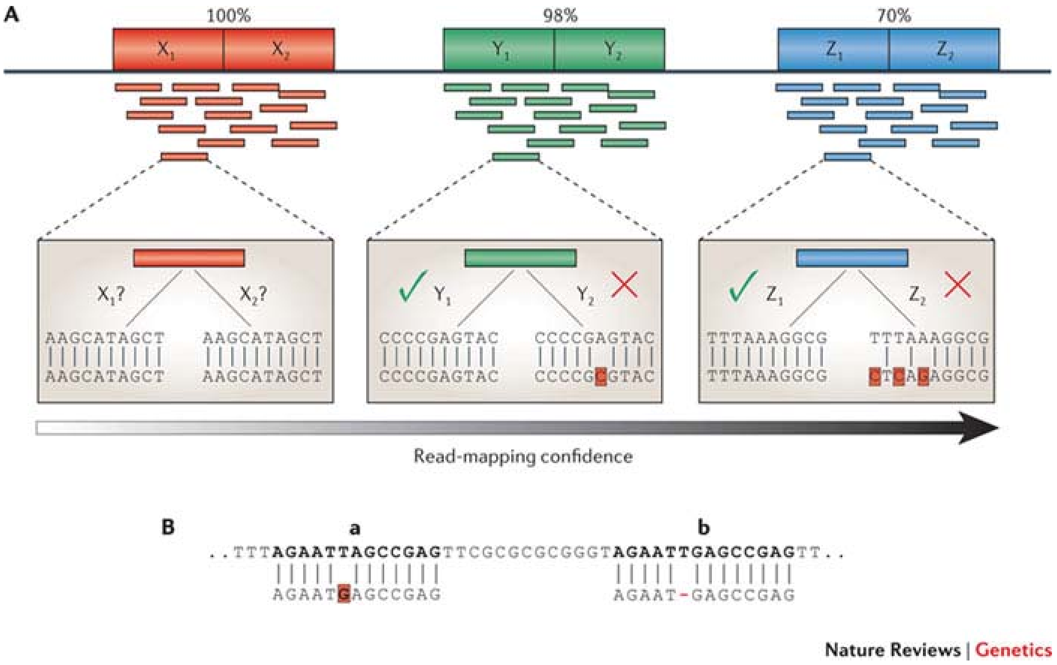
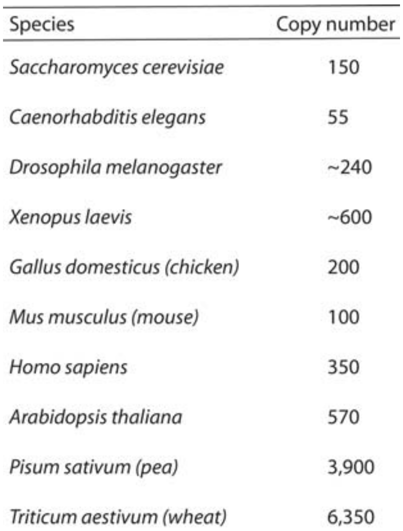


Figure 2.21: Ambiguities in read mapping

"**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.



Species	Copy number
Saccharomyces cerevisiae	150
Caenorhabditis elegans	55
Drosophila melanogaster	~240
Xenopus laevis	~600
Gallus domesticus (chicken)	200
Mus musculus (mouse)	100
Homo sapiens	350
Arabidopsis thaliana	570
Pisum sativum (pea)	3,900
Triticum aestivum (wheat)	6,350

Table 2.6: Haploid copy number of rDNA repeats across Eukaryotic species
Selection of rDNA repeats observed in different eukaryotic species [865].

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 demand 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

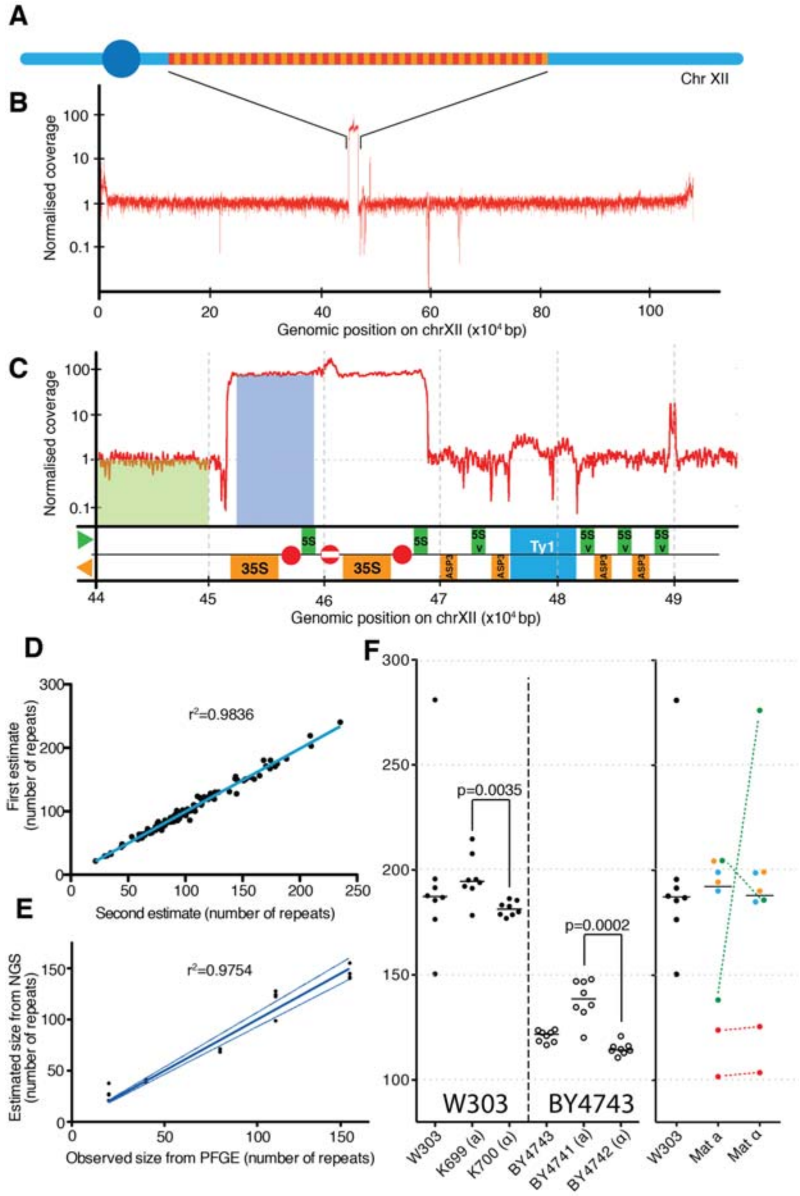


Figure 2.22: Next-generation sequencing data can be used to estimate rDNA copy number reliably

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 MATα. 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).

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

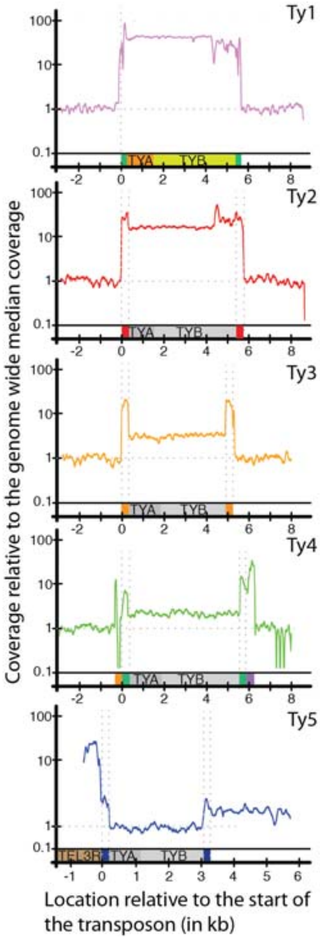


Figure 2.23: Next-generation sequencing data could also be used to assess Ty element copy number

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 references 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[873], it relies entirely on the fact that the human repeat is invariable (a TTAGGG tandem repeat). In contrast, the *S. cerevisiae* telomere repeat is degenerate with the consensus sequence $G_{2–3}(TG)_{1–6}$[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 *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 *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.

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 *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 identfied 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 5.5×10^{-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

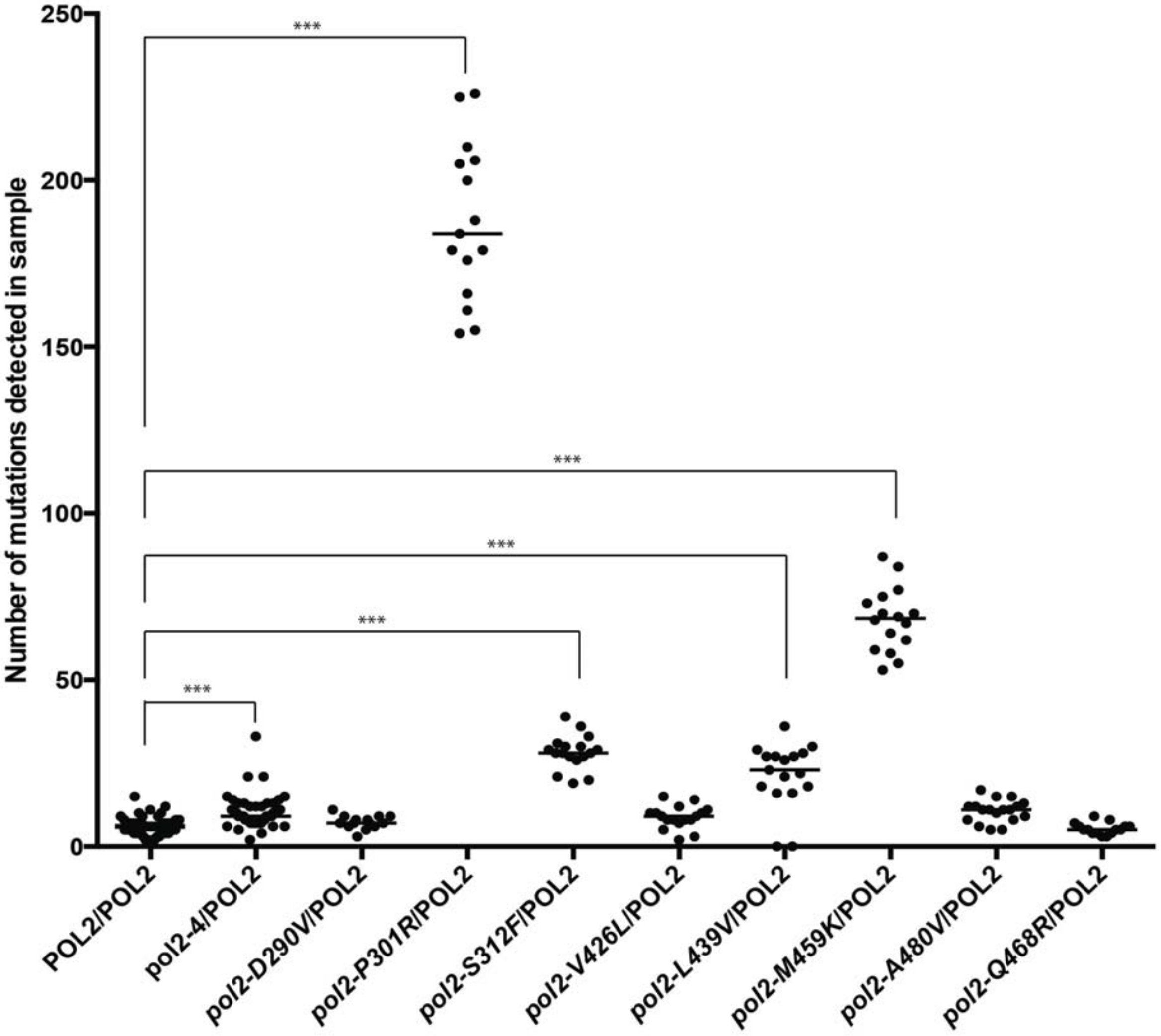


Figure 3.1: Number of single-nucleotide variants per sample in *pol2* mutant strains *Pol2* mutant strains were whole-genome sequenced before and after a three-month propagation on rich medium. The number of acquired single-nucleotide variants was determined for each parallel line that was propagated. 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

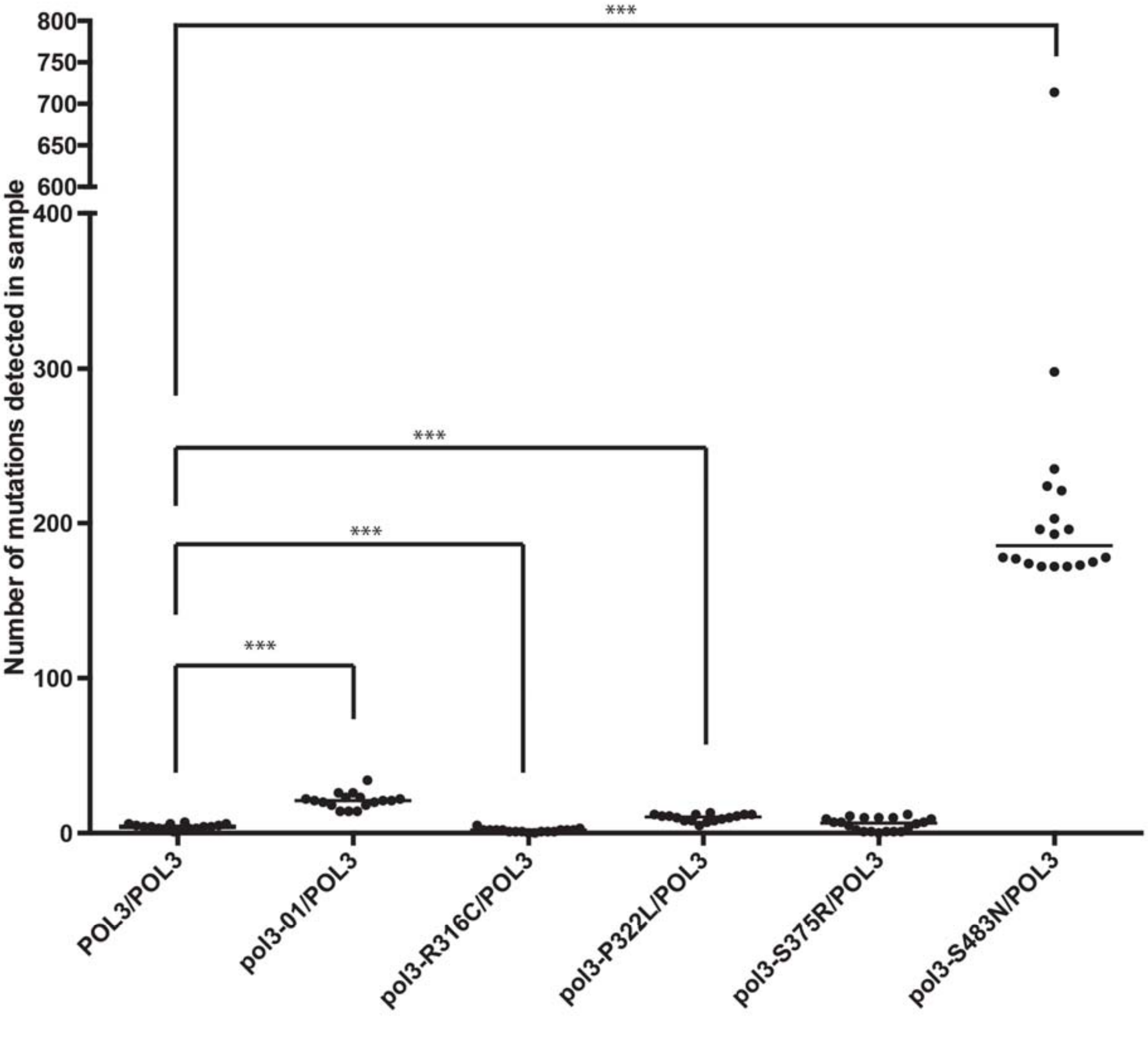


Figure 3.2: Number of single-nucleotide variants per sample in *pol3* mutant strains *Pol3* mutant strains were whole-genome sequenced before and after a three-month propagation on rich medium. The number of acquired single-nucleotide variants was determined for each parallel line that was propagated. The number of samples for each strain are as follows: n=65 (*POL*3), 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.

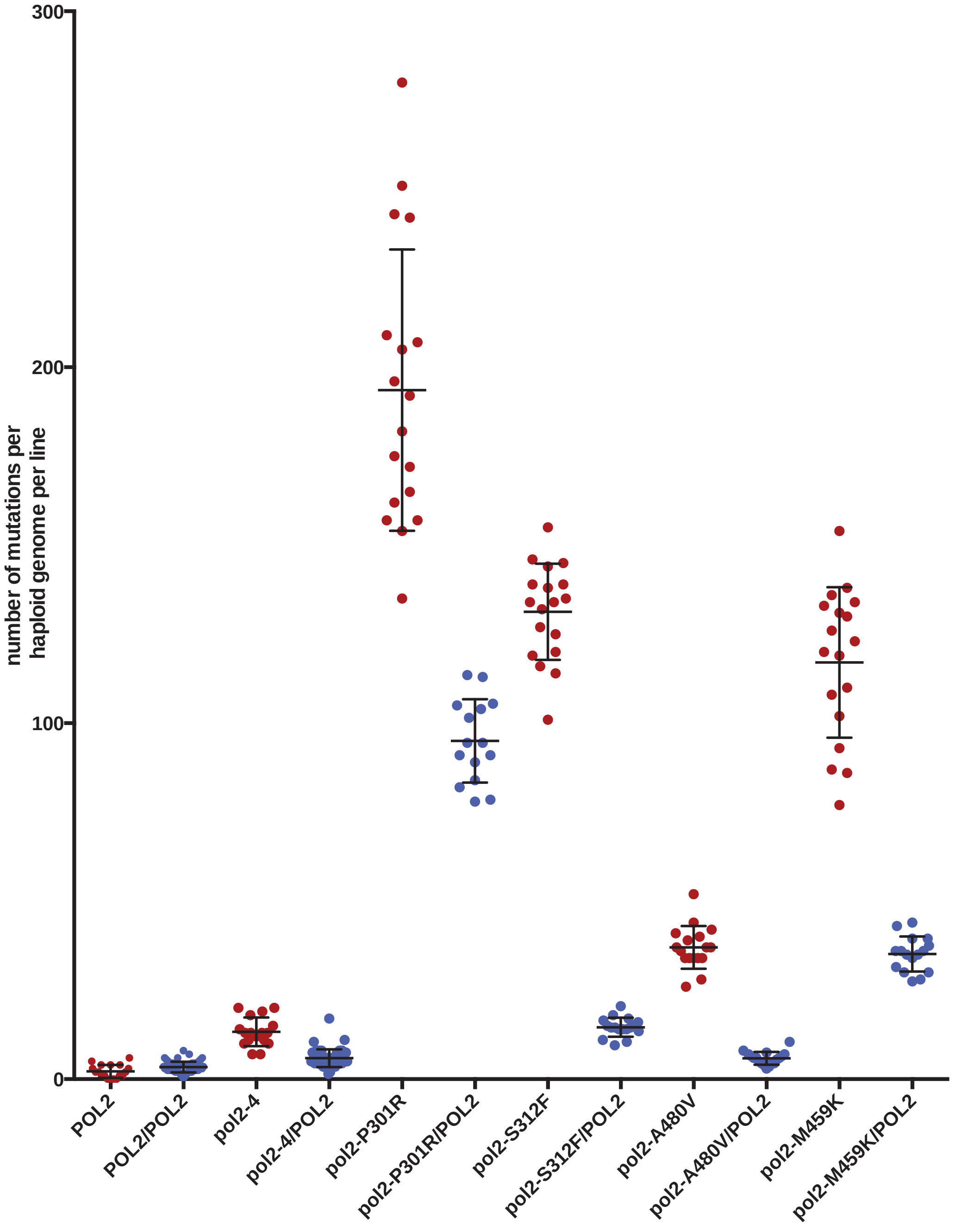
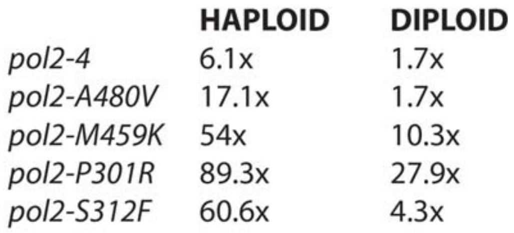


Figure 3.3: Number of single-nucleotide variants per line per haploid genome for selected haploid and heterozygous diploid *pol2* mutant strains

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.



	HAPLOID	**DIPLOID**
pol2-4	6.1x	1.7x
pol2-A480V	17.1x	1.7x
pol2-M459K	54x	10.3x
pol2-P301R	89.3x	27.9x
pol2-S312F	60.6x	4.3x

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.

different effect on mutation numbers.

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).

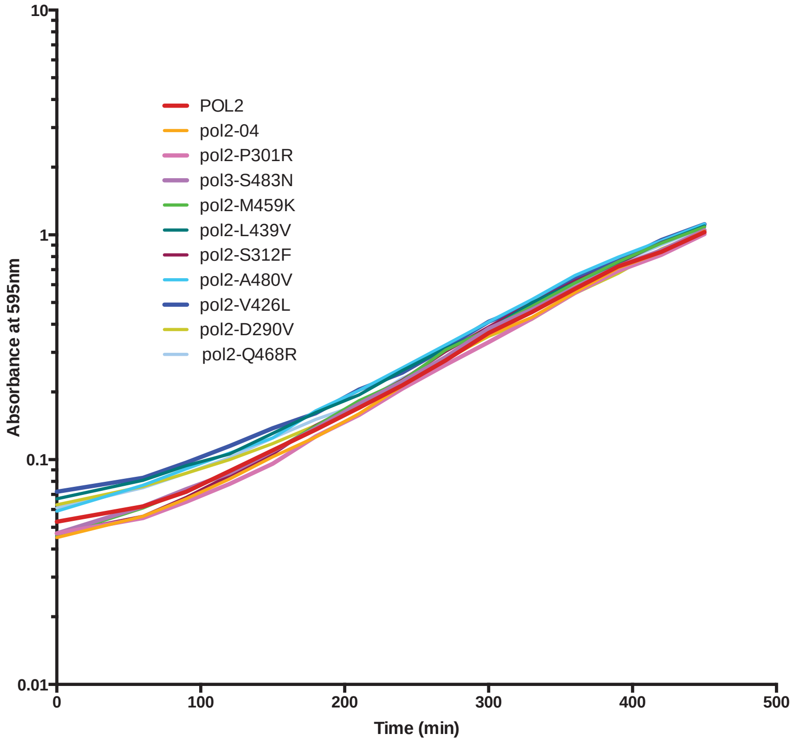
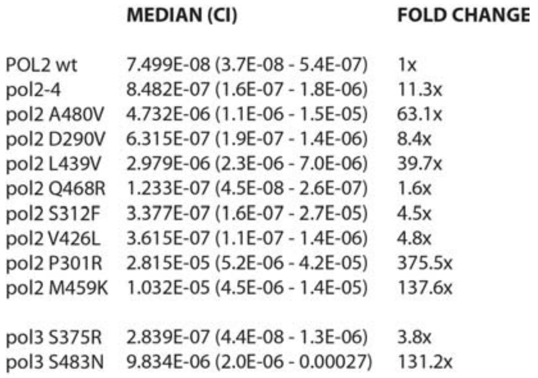


Figure 3.4: Growth of *S. cerevisiae* mutant strains in rich medium

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.



	MEDIAN (CI)	FOLD CHANGE
POL2 wt	7.499E-08 (3.7E-08 - 5.4E-07)	1x
pol2-4	8.482E-07 (1.6E-07 - 1.8E-06)	11.3x
pol2 A480V	4.732E-06 (1.1E-06 - 1.5E-05)	63.1x
pol2 D290V	6.315E-07 (1.9E-07 - 1.4E-06)	8.4x
pol2 L439V	2.979E-06 (2.3E-06 - 7.0E-06)	39.7x
pol2 Q468R	1.233E-07 (4.5E-08 - 2.6E-07)	1.6x
pol2 S312F	3.377E-07 (1.6E-07 - 2.7E-05)	4.5x
pol2 V426L	3.615E-07 (1.1E-07 - 1.4E-06)	4.8x
pol2 P301R	2.815E-05 (5.2E-06 - 4.2E-05)	375.5x
pol2 M459K	1.032E-05 (4.5E-06 - 1.4E-05)	137.6x
pol3 S375R	2.839E-07 (4.4E-08 - 1.3E-06)	3.8x
pol3 S483N	9.834E-06 (2.0E-06 - 0.00027)	131.2x

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.

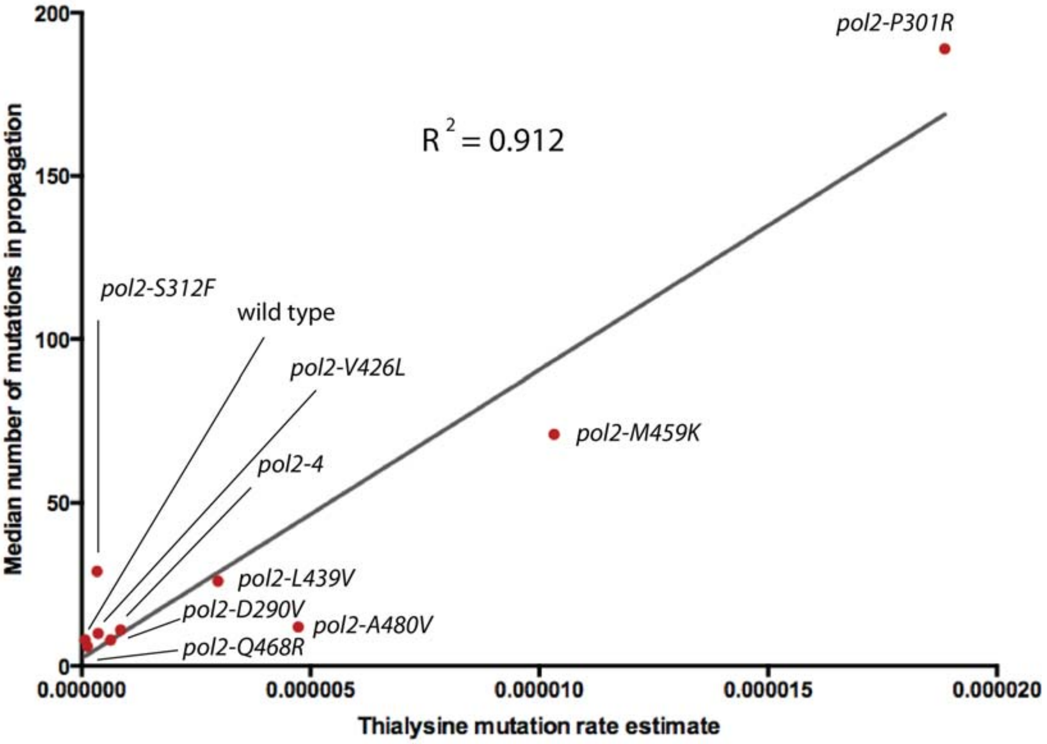


Figure 3.5: Correlation of mutation rate estimates and mutation accrual for *pol2* mutant strains For all *pol2* mutant strains and a wild type control, mutation rate estimates for haploid mutants obtained using thialysine resistance were plotted against the number of mutations detected after propagating 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]

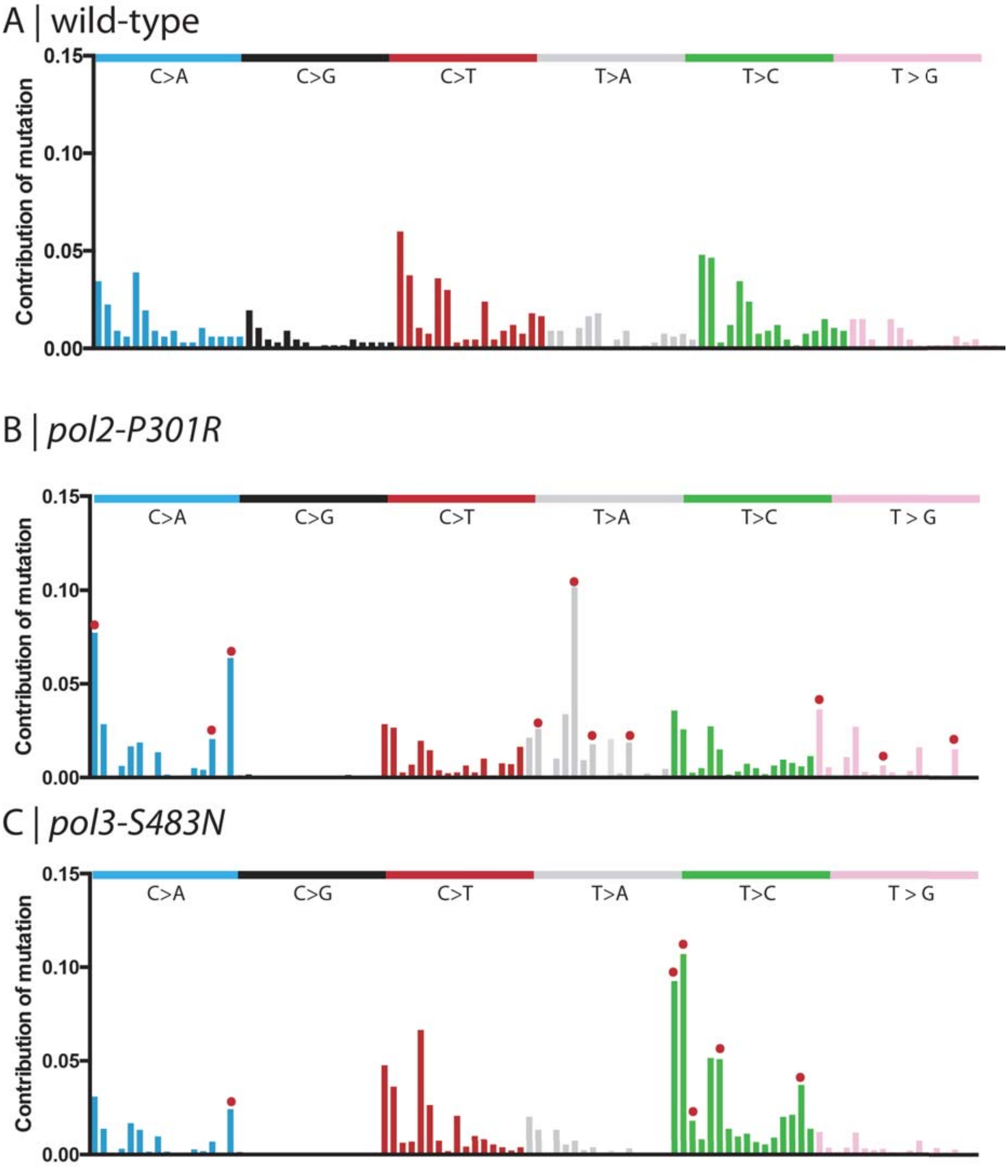


Figure 3.6: Single-nucleotide variant patterns

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 χ^2 are marked by a “red dot”.

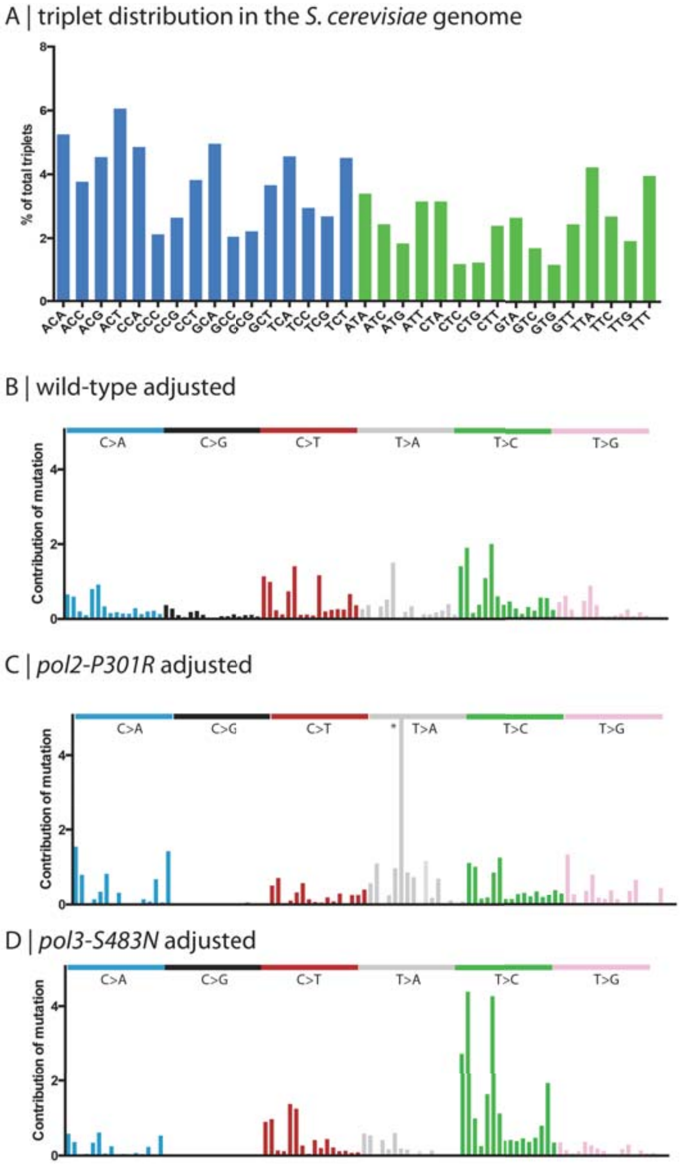


Figure 3.7: Single-nucleotide variant patterns adjusted to frequencies of trinucleotides in *S. cerevisiae*

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.

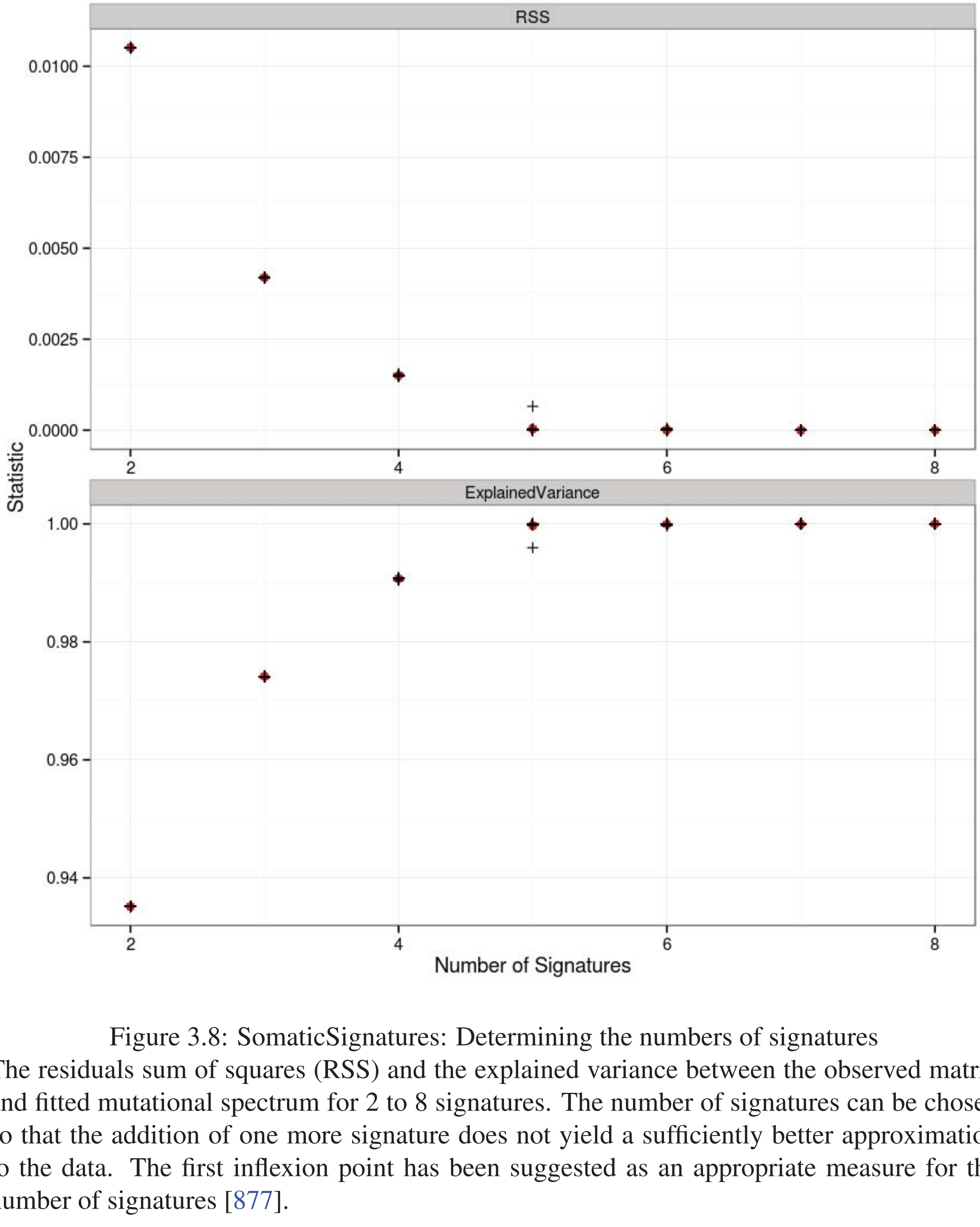


Figure 3.8: SomaticSignatures: Determining the numbers of signatures
The residuals sum of squares (RSS) and the explained variance between the observed matrix and fitted mutational spectrum for 2 to 8 signatures. The number of signatures can be chosen so that the addition of one more signature does not 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 improvements 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 indicates that two signatures explain nearl 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>TAT also preferred in *pol2-P301R* samples), while Signature 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 approximately 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 mutations 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 identified 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

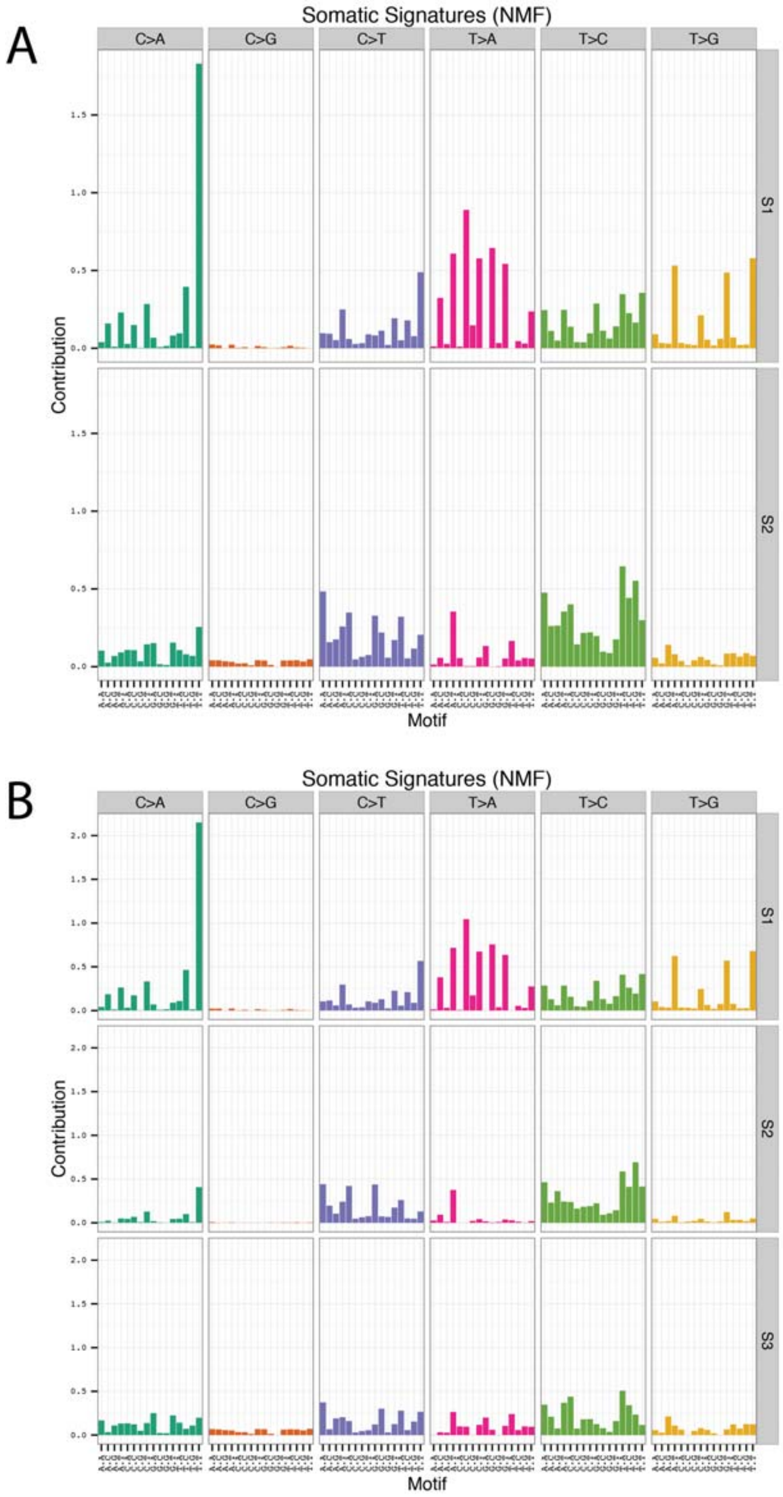


Figure 3.9: 2-3 signatures are determined using Non-negative matrix factorization
Two and three signatures were extracted from mutation data of all mutator strains and the wild type combined using SomaticSignatures with NMF. The signatures extracted are displayed in the 96 trinucleotide-change channel format indicating each mutation's contribution to the overall pattern. **A|** Two signatures extracted from mutator strains. **B|** Three signatures extracted from mutator strains.

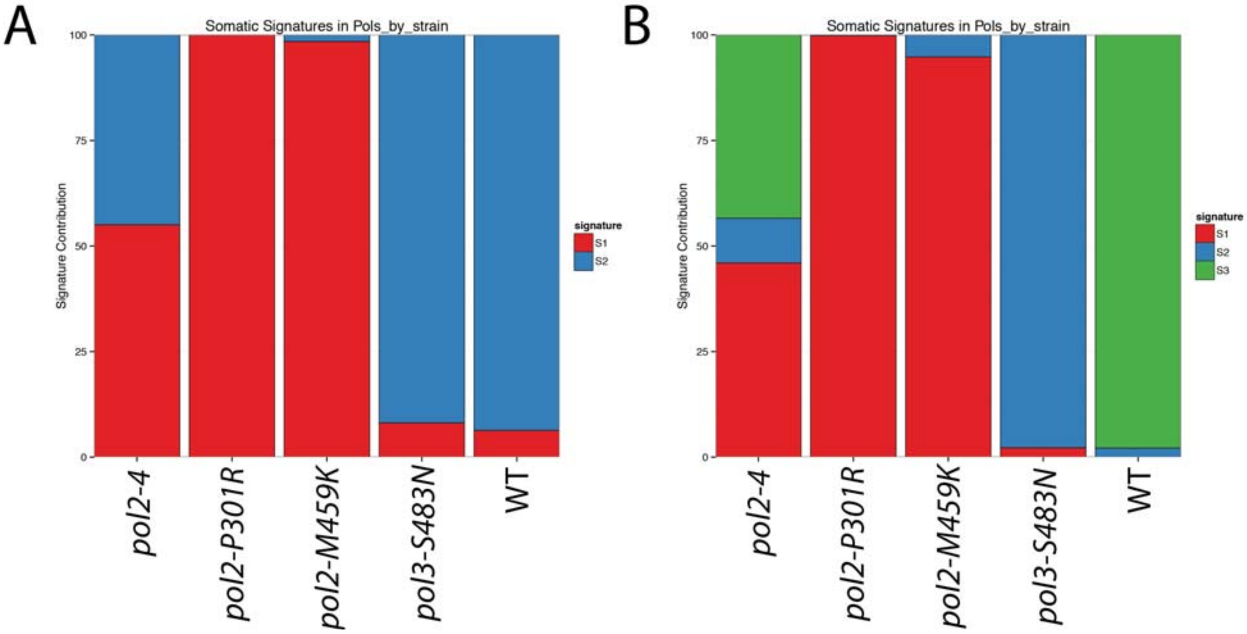


Figure 3.10: Contribution of the signatures to the variant pattern
Two and three signatures were extracted from mutation data of all mutator strains and the wild type combined using SomaticSignatures with NMF. The contribution of each of the two and three signatures to the mutational landscape of each 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

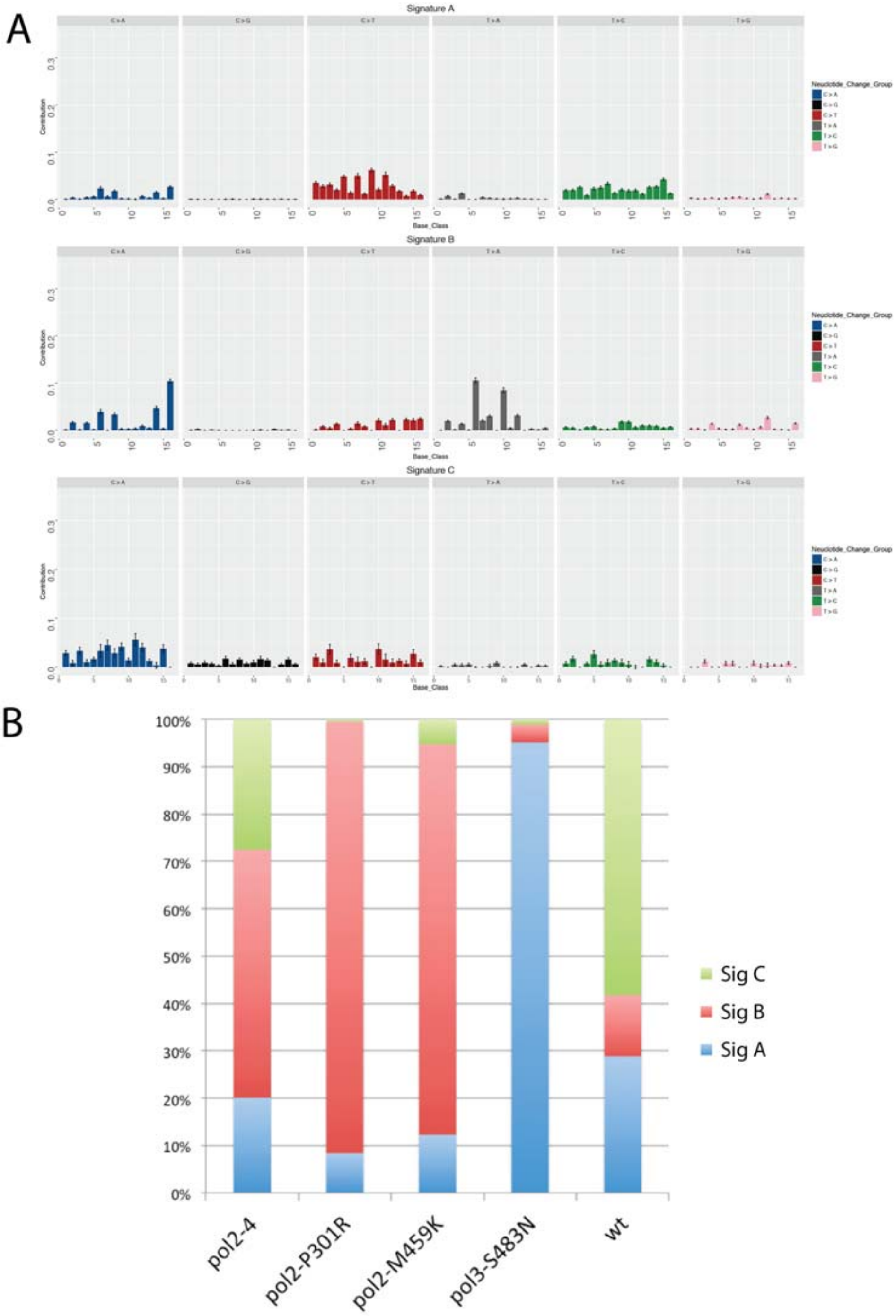


Figure 3.11: EMu: Validating Signature Analysis
Signature analysis was also performed using EMu. Using model selection criteria, such as the Bayesian information criterion (BIC), EMu determined that a model with three signatures has the strongest statistical support. **A|** Signatures displayed in their trinucleotide mutation pattern. **B|** Contribution of signatures displayed in **A** to the total mutations observed in mutator strains.
]

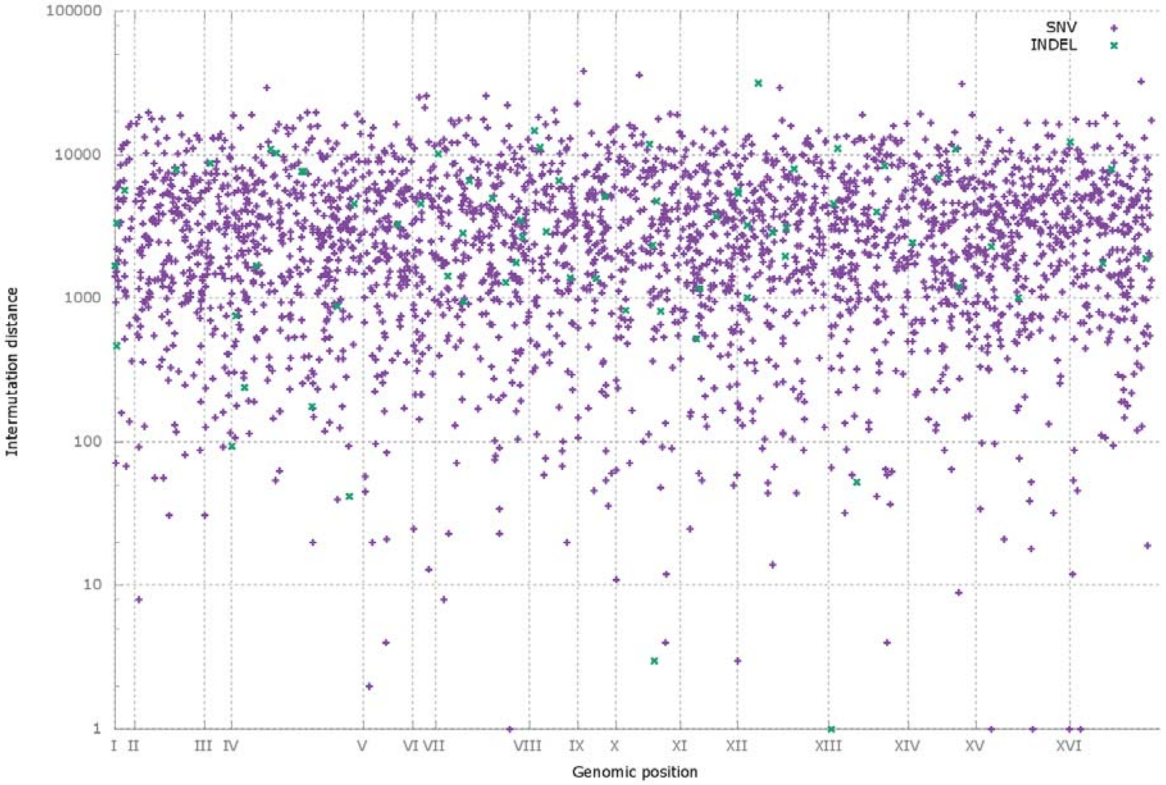


Figure 3.12: No observed clustering of mutations acquired by *pol2-P301R* strains
Mutations acquired across 15 parallel lines of propagated *pol2-P301R* strains were pooled, sorted and inter-mutation distances were determined and plotted.

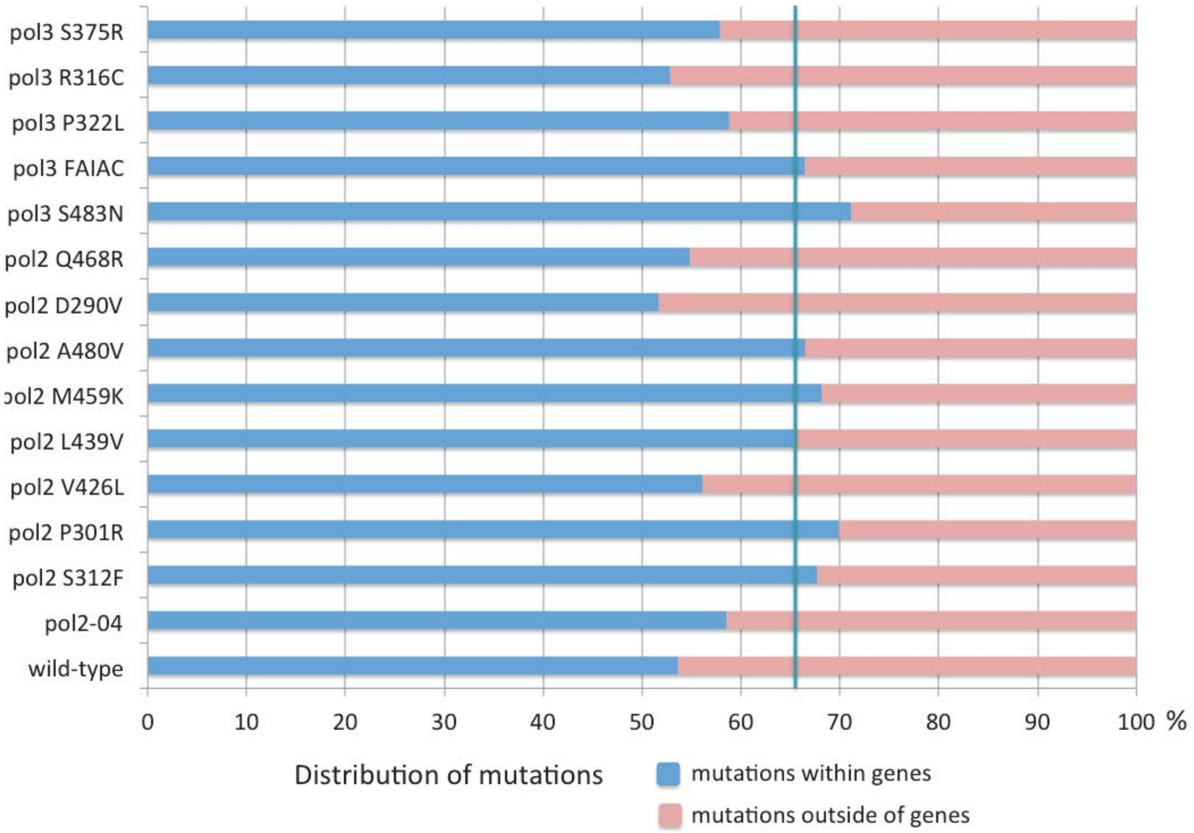


Figure 3.13: Mutations falling inside and outside of genes in heterozygous diploid polymerase mutant strains
Each mutation acquired during the three month propagation of the heterozygous diploid polymerase mutant strains was scored for their presence inside one of the 5133 verified *Saccharomyces cerevisiae* open reading frames (ORFs). Here the percentage of how many mutations fall inside a gene versus outside a gene for each strain is given. The vertical, teal coloured line at 65.5% represents the genome wide percentage of nucleotides that form ORFs.

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

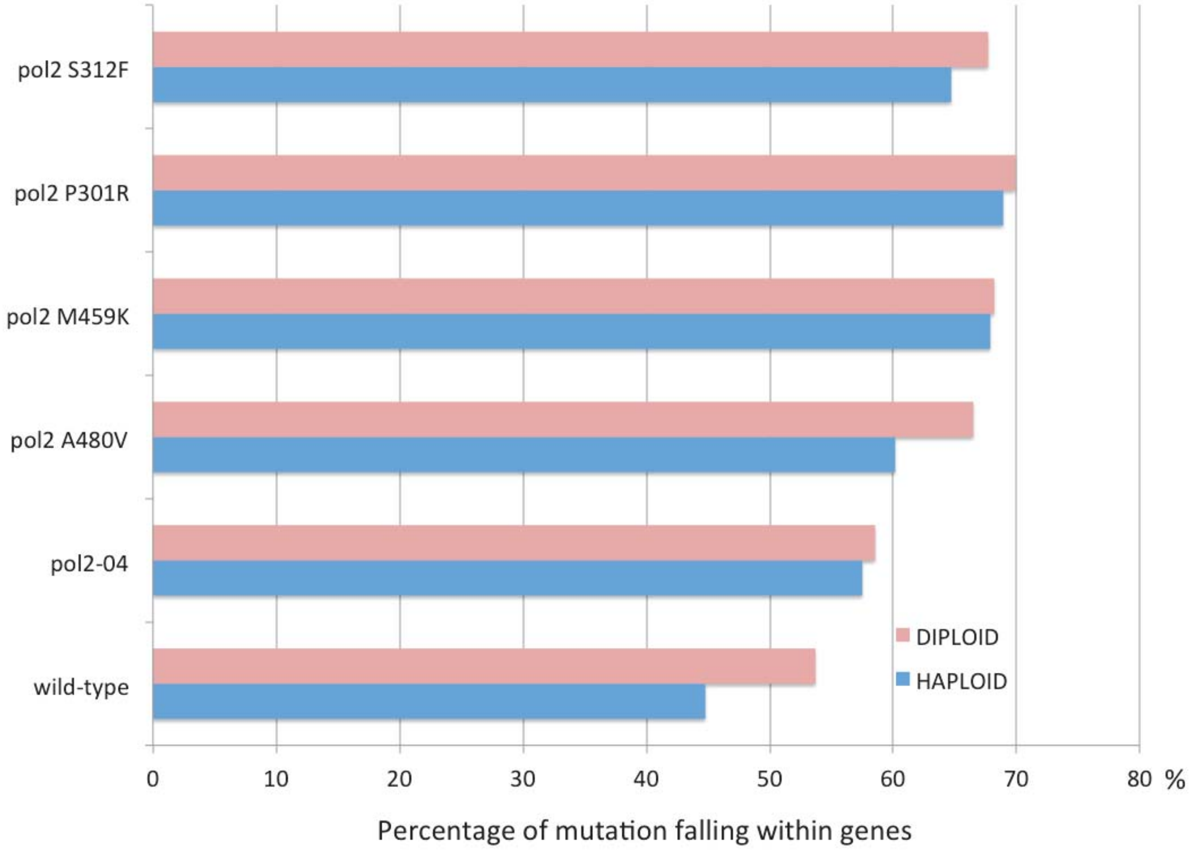


Figure 3.14: Percentage of mutations within genes in haploid strains
The percentage of mutations that fall within genes is shown for both haploid and heterozygous diploid strains for a selection of *pol2* mutants.

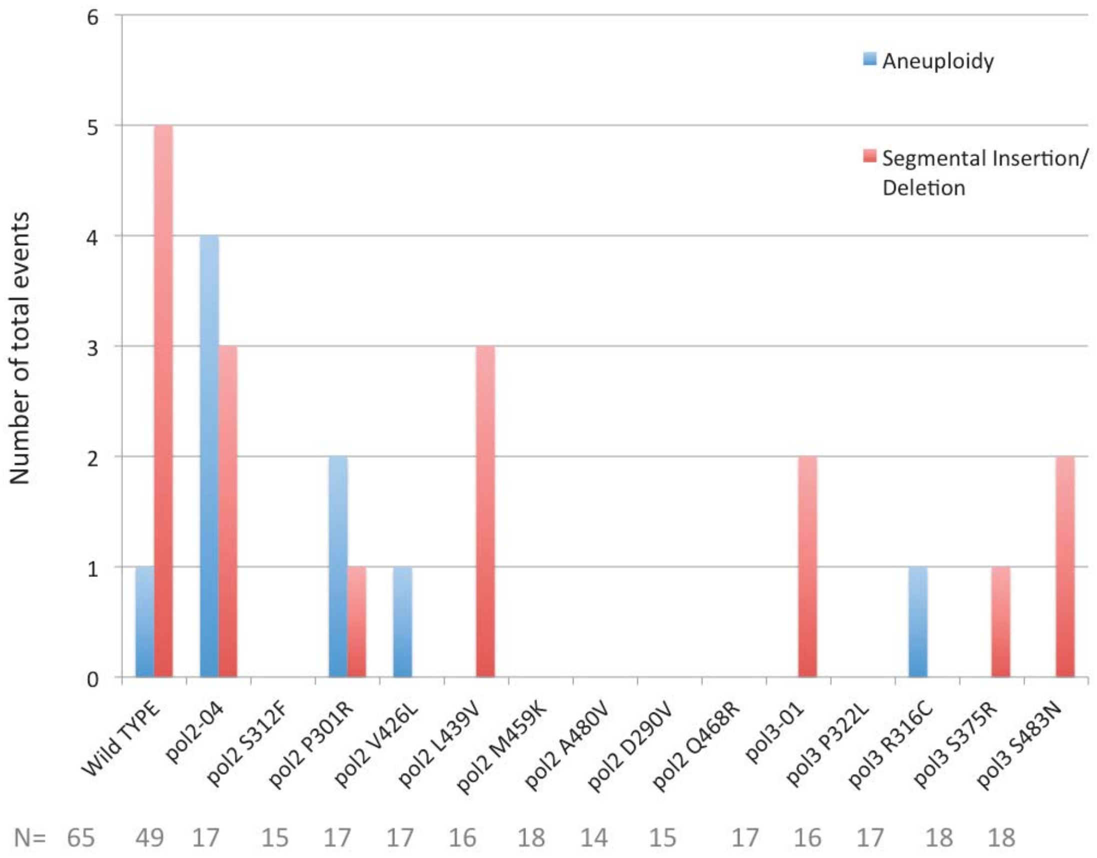


Figure 3.15: Number of total aneuploidy and segmental insertions/duplications identified The number of events of aneuploidy and large deletions/amplification was determined from the coverage data of all strains and counted. The number of samples available for each strain is listed below the figure legend.

replication.

3.5 Large-scale variation: aneuploidy, CNVs and rDNA copy number

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 ± 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 *PMS*1. 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 MutS$\alpha$and MutSβ), Mlh1 (MutL homolog) and Pms1 (which forms a heterodimer with Mlh1) leads to high numbers of SNVs and small INDELs.

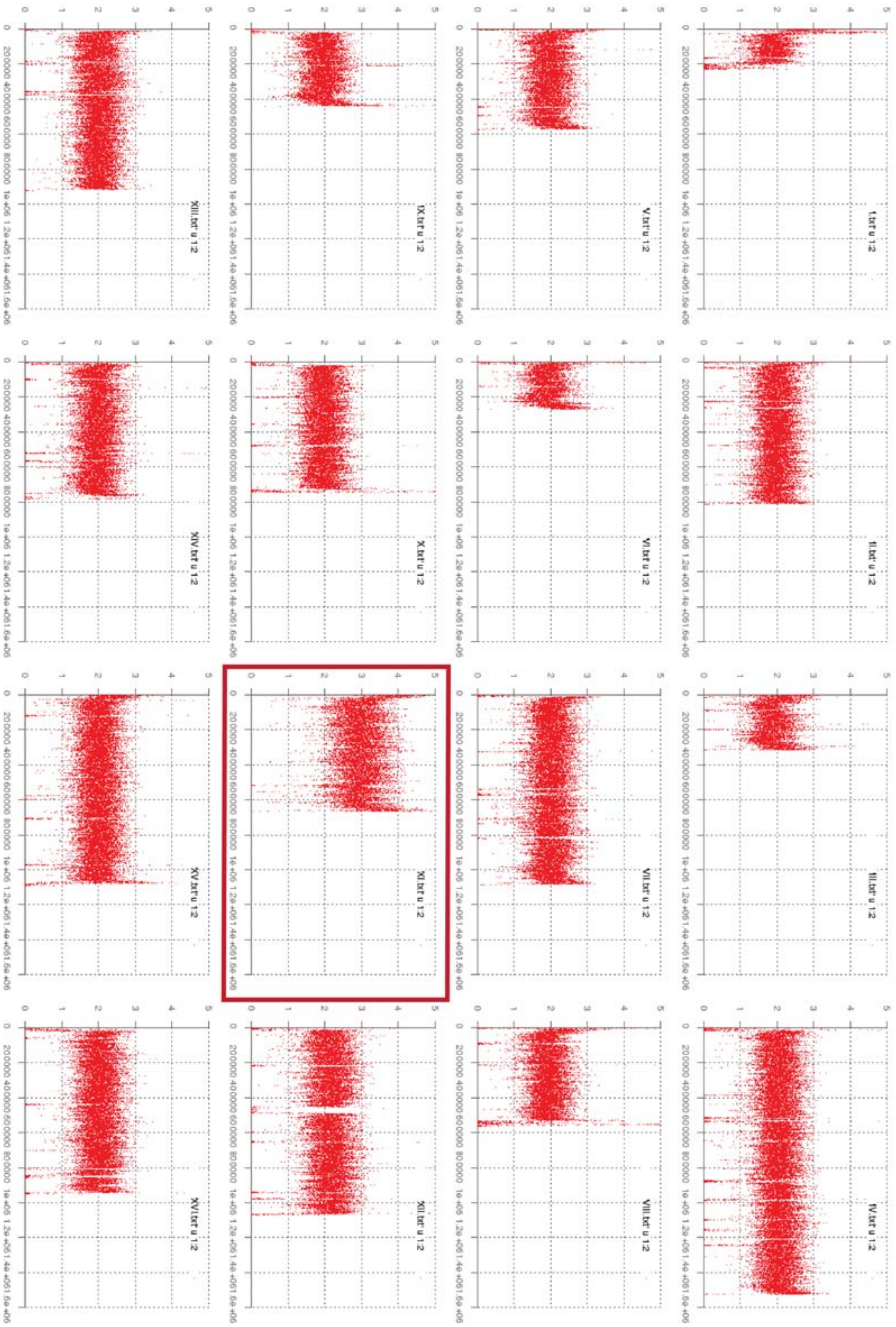


Figure 3.16: Example of aneuploidy in *S. cerevisiae*
This example shows the coverage profile of a *pol2-04* propagated heterozygous diploid strain with an aneuploidy for chromosome XI (3n). At repetitive regions the coverage drops to 0 (for instance see chromosome XII) due to mapping quality thresholds in the program.

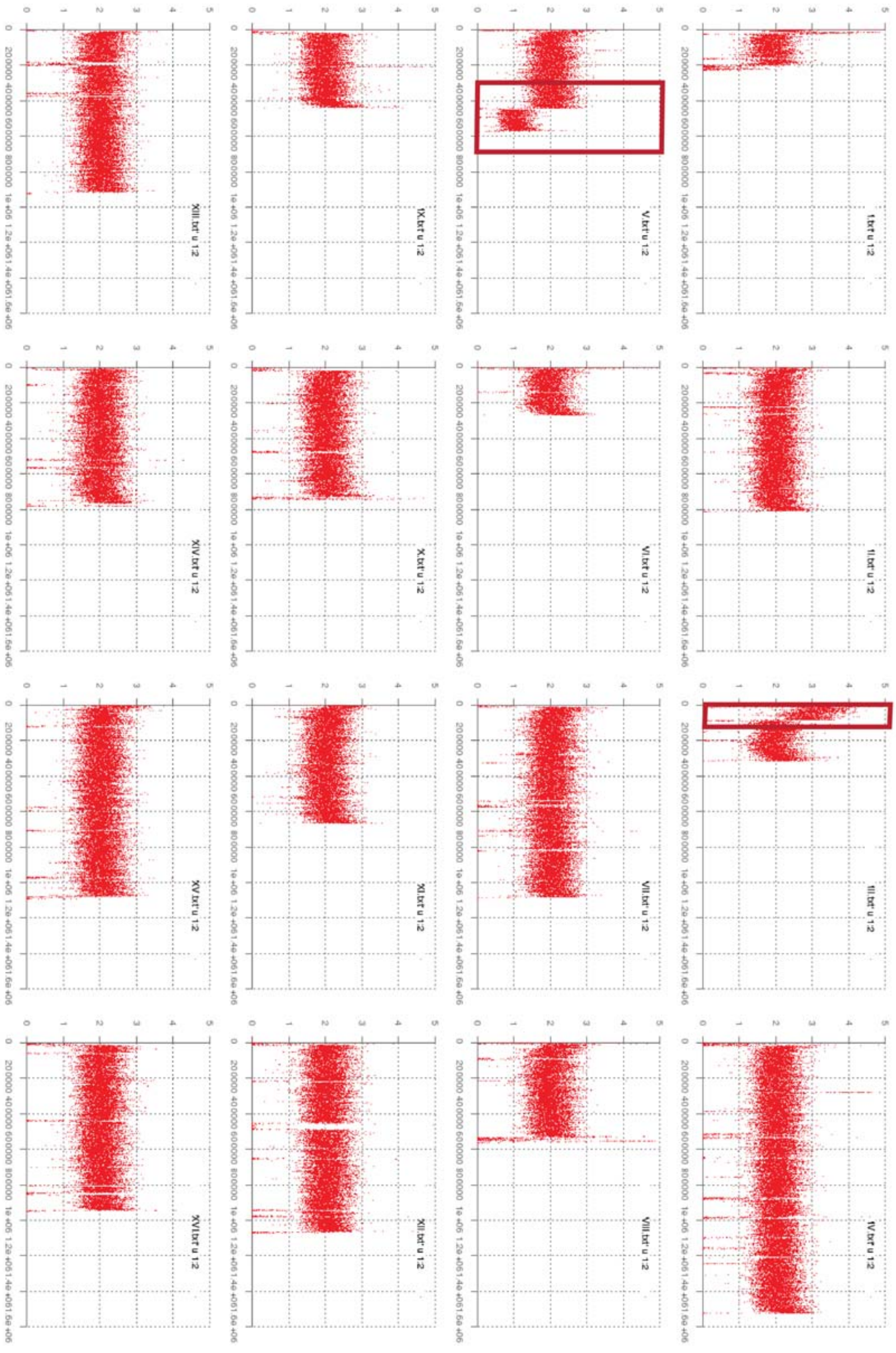


Figure 3.17: Example of segmental deletions and amplifications in *S. cerevisiae*
This example shows the coverage profile of a *pol3-01* propagated heterozygous diploid strain with a segmental deletion on chromosome V and an amplification of a region of chromosome III. At repetitive regions the coverage drops to 0 (for instance see chromosome XII) due to mapping quality thresholds in the program.

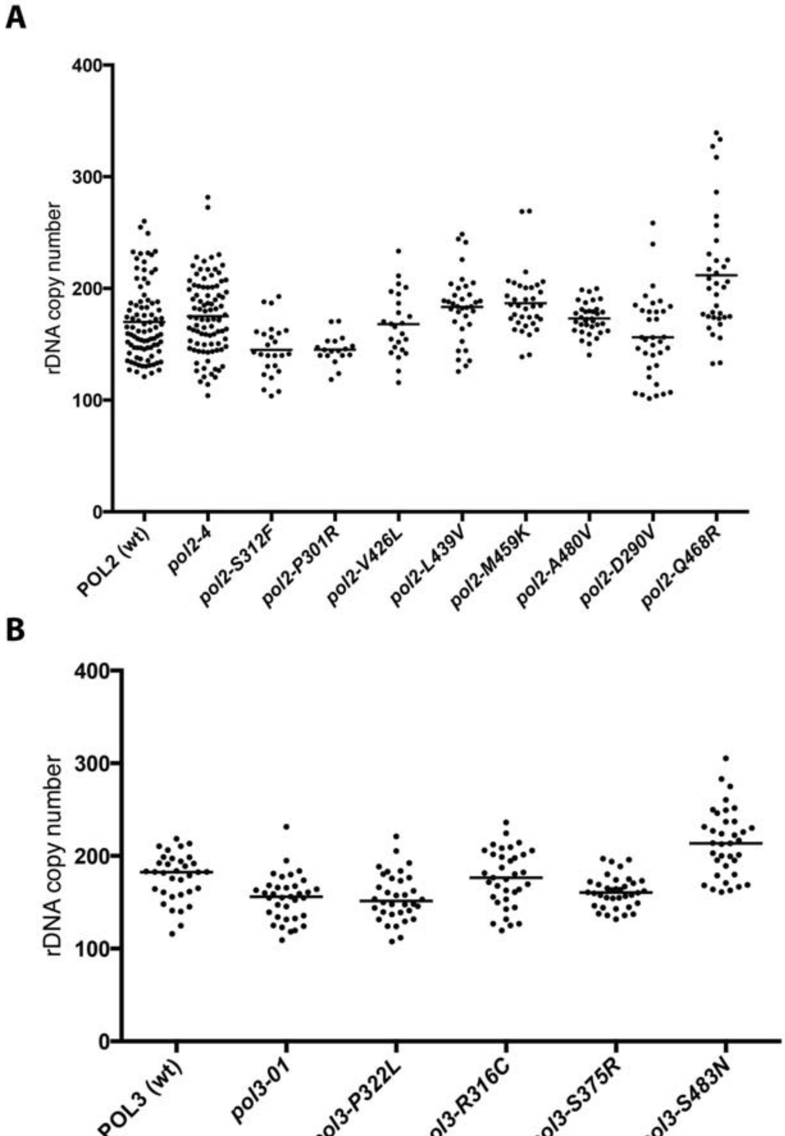


Figure 3.18: rDNA copy number changes in polymerase mutants

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.

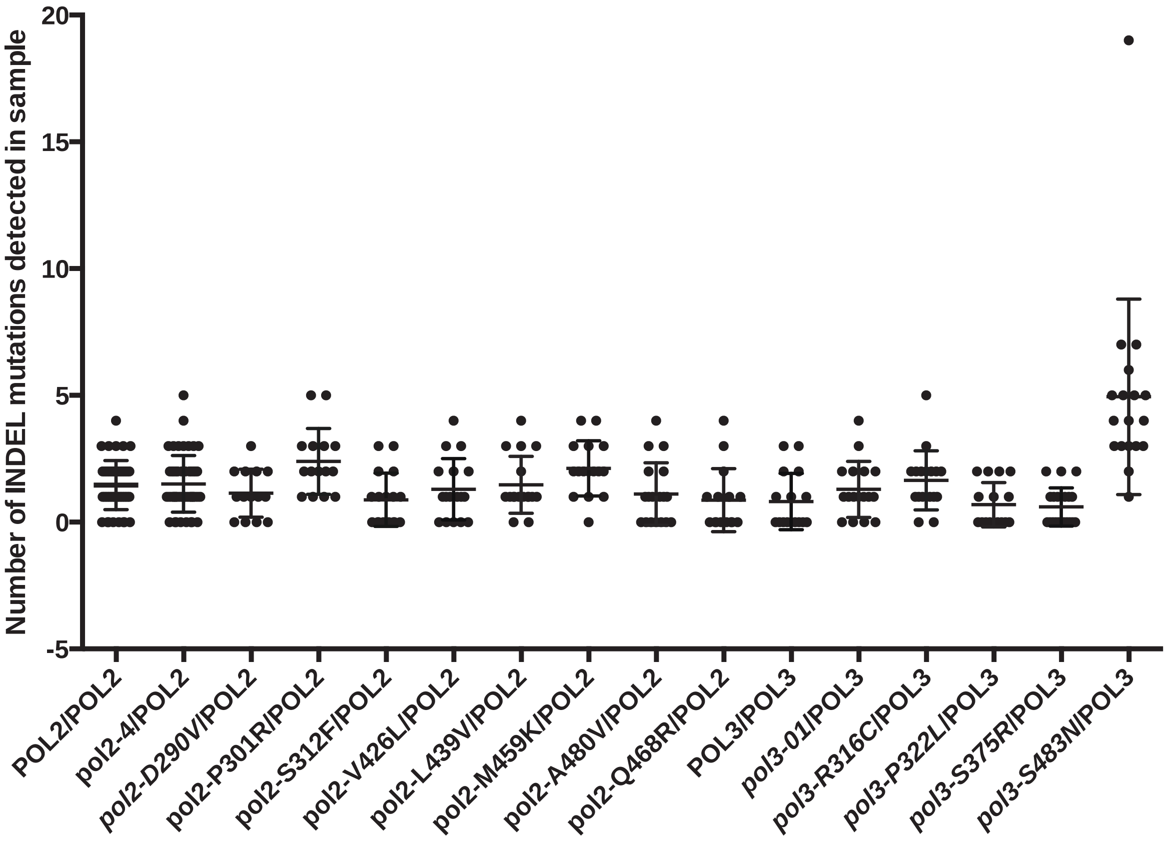


Figure 3.19: No increase in the number of INDELs detected per sample across strains
The number of INDELs per sample was determined from the heterozygous diploid polymerase mutant strains that were propagated for three months using single colony bottlenecks. The back bars indicate mean and standard deviation.

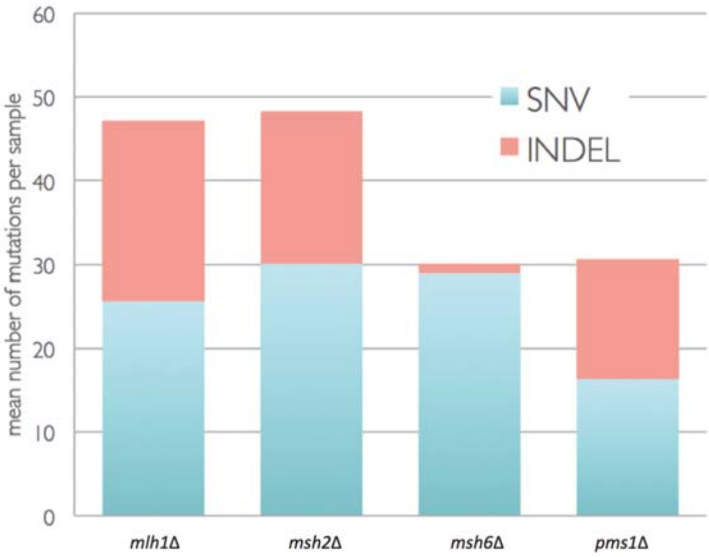


Figure 3.20: Mutation accrual in strains with mismatch repair deficiencies
Strains with deletions in mismatch repair genes were propagated for three months using population bottlenecks. Ten colonies each were sequenced and single-nucleotide variants as well as INDELs were identified. Results for strains carrying deletions in *MLH1, MSH2, MSH6* and *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 *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 *in vivo* using the budding yeast system. Strains carrying the mutations *pol2-P301R, pol2-S312F, pol2-L439V, pol2-M459K* or *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 identfied 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 Polε 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 *pol2-P301R, pol2-S312F, pol2-L439V, pol2-M459K* or *pol3-S483N* were all shown to significantly increase the number of acquired mutations when in heterozygosis. Strikingly, the *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 *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 *pol2-L439V* and *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 *pol2-P301R* mutation, resulting in similarly striking mutation increases as *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 acuired 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 MAT α strain deleted for *MSH2 (msh2*Δ) was created (see Chapter 6.6) and mated to haploid MATa cells: *pol2-P301R, pol2-S312F, pol2-A480V* and *pol2-M459K* (see Chapter 6.6). The strains now heterozygous diploid for *msh2*Δ and the polymerase mutation were kept on sporulation medium to undergo meiosis. Tetrads were issolated and dissected. After two days of growth tetrads were replicated onto selection plates selecting for the polymerase mutation (ura- plates), the *msh2*Δ mutation (G418 plates) and plates selecting for either mating type. This allowed confirmation that tetrads were haploid

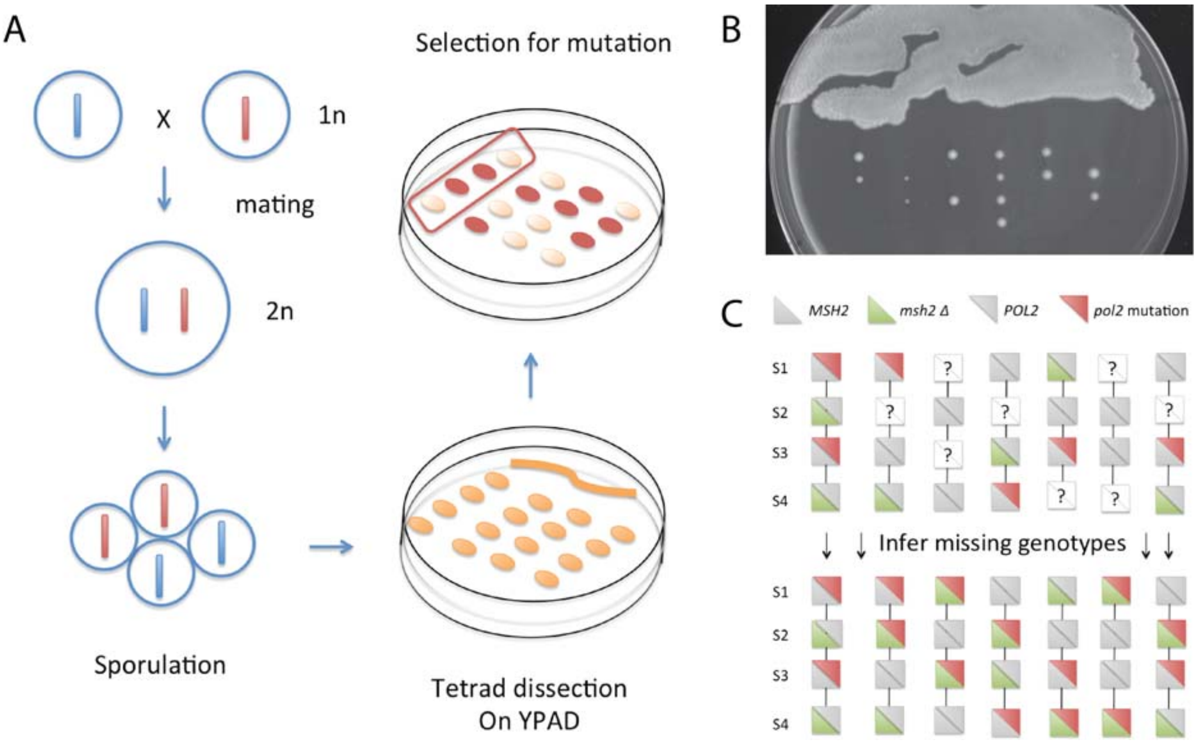


Figure 4.1: Tetrad dissection to generate double mutants and detect synthetic lethality **A|** For tetrad dissection, two haploid single mutant strains of opposite mating types are mated and the resulting diploid forced to undergo meiosis resulting in a tetrad. Alternatively, an already existent diploid 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 mutatant 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 *msh2*Δ 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 *msh2*Δ in a haploid background. *pol2-A480V msh2*Δ 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 *msh2*Δ 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 viablity of other polymerase mutants in combination with mismatch repair deficiency can be made. For instance *pol2-Q468R msh2*Δ double mutant should be viable, a *pol3-S483N msh2*Δ double mutant should not be and the viability of *pol2-04 msh2*Δ and *pol2-L439V msh2*Δ 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 deficienct 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-

polymerase mutant	tetrads dissected	double mutants expected	double mutants recovered
pol2-A480V	4	4	3*
pol2-M459K	8	10	0
pol2-P301R	10	14	0
pol2-S312F	15	14	0

Table 4.1: Synthetic lethality of polymerase mutants and mismatch repair deficiency
Diploid strains heterozygous for both *msh2*Δ 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.

tween 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-

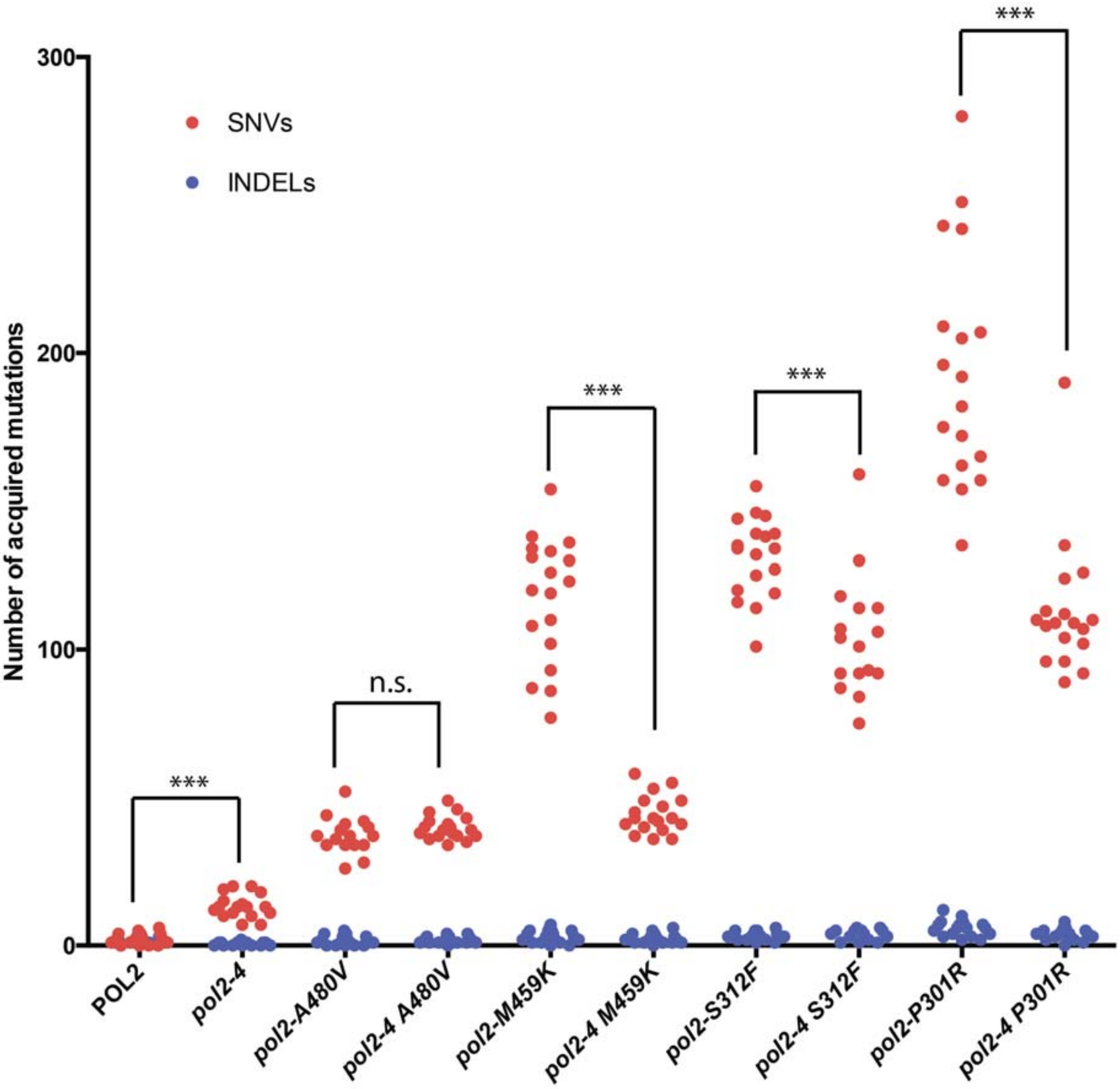


Figure 4.2: The mutagenesis observed in strong mutator strains is partially rescued by mutating critical residues in the exonuclease domain active site
Combinations of candidate *pol2* mutations and the *pol2-D290A,E292A* mutations in the same gene were achieved by site directed mutagenesis of the plasmids 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-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. $P > 0.05$; *** $P \leq 0.001$

MCM-GINS) complex, which is thought to bring the polymerase ε into the replisome. Polymerase ζ, recently found to also interact with polymerase δ subunits Pol31 and Pol32[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 Polζ, *REV3*, substantially decreases the mutator phenotype found in *dpb2-100* cells, but has little effect on the mutator phenotype observed in *pol2-4* cells[883].

To investigate whether this could partly explain the apparent divergence between *pol2-4* mutant strains and the other polymerase mutations, whose presence results in higher mutation numbers than exonuclease deficiency, a *rev3*Δ strain was constructed and double mutants of polymerase mutations with *rev3*Δ were generated by mating and tetrad 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 muttaion numbers observed in the *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 *dpb2-100* cells[883], where a double mutant with *rev3*Δ decreased mutation numbers, leading them to conclude that Polζ is responsible for a substantial part of the mutagenesis in *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 *pol2-A480V* strains with the corresponding *pol2-A480V rev3*Δ double mutant (a mean of 38.6 versus a mean of 49.1 mutations) and the *pol2-S312F* with the corresponding *pol2-S312F rev3*Δ cells (a mean of 134.6 versus a mean of 161.6 mutations), the mutation increase oberserved in *pol2-P301R* strains upon combination with a deletion of *REV3* is striking: *pol2-*P301R cells acquire a mean of 116.8 mutations, while *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 *pol2-P301R* cells and *pol2-P301R rev3*Δ cells were compared (Fig. 4.4). Even though the number of mutations acquired by the *pol2-P301R rev3*Δ cells is more than double that of the *pol2-P301R* cells, there are no striking differences between the two patterns.

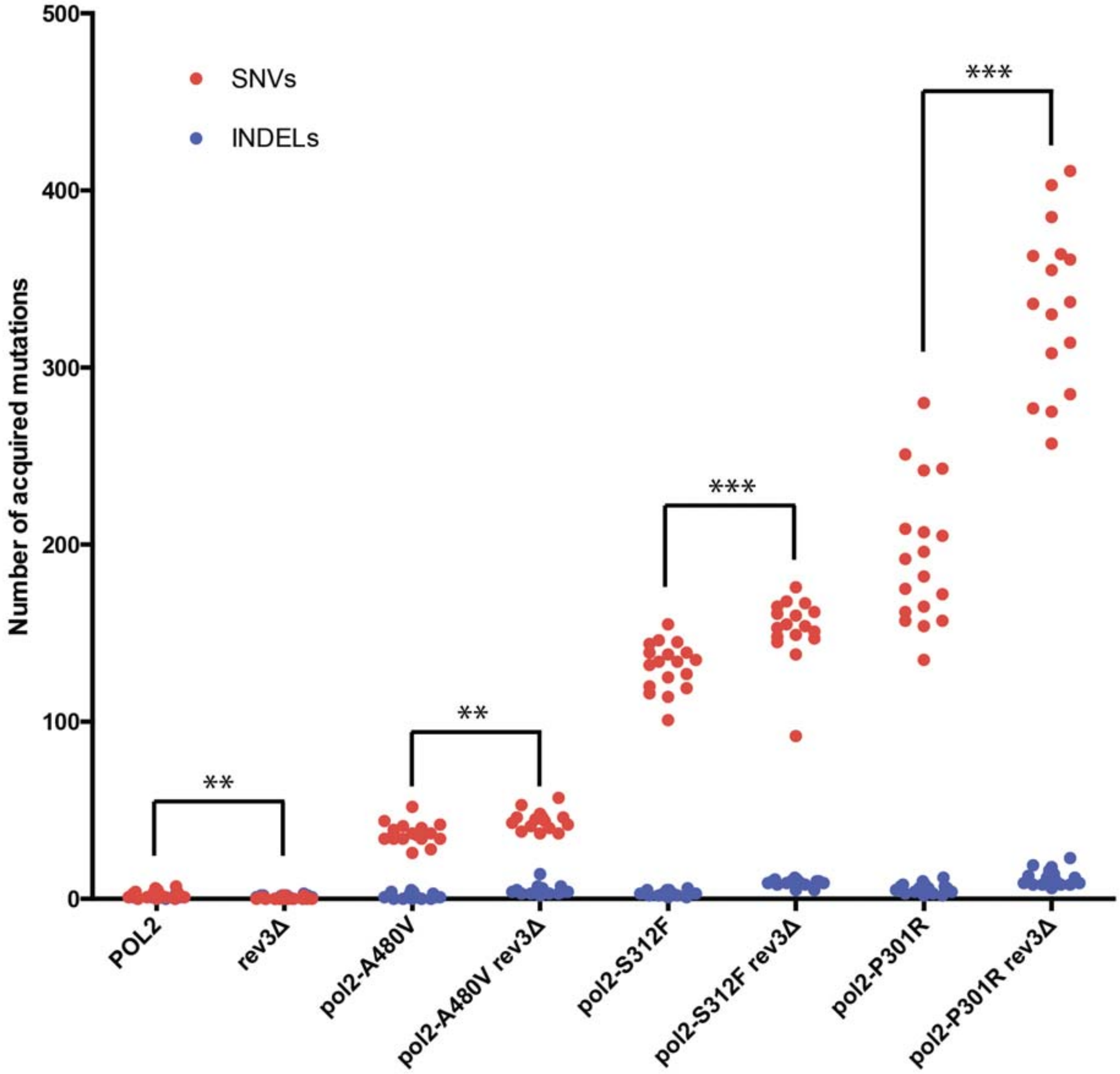


Figure 4.3: Synergystic effects on mutation number between *rev3*Δ and *pol2-P301R* Polymerase mutation and *rev3*Δ haploid double mutants were obtained by crossing and propagated alongside haploid single mutants for 13 passages using single colony bottlenecks. Samples were sequenced before and after 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, *rev3*Δ, *pol2-S312F*, *pol2-P301R*), n=17 (*pol2-A480V*, *pol2-S312F rev3*Δ), n=16 (*pol2-P301R rev3*Δ) and n=15 (*pol2-A480V rev3*Δ). Significance of the difference in number of SNVs was determined by two-tailed, unpaired t-Tests. ** $P \leq 0.01$; *** $P \leq 0.001$

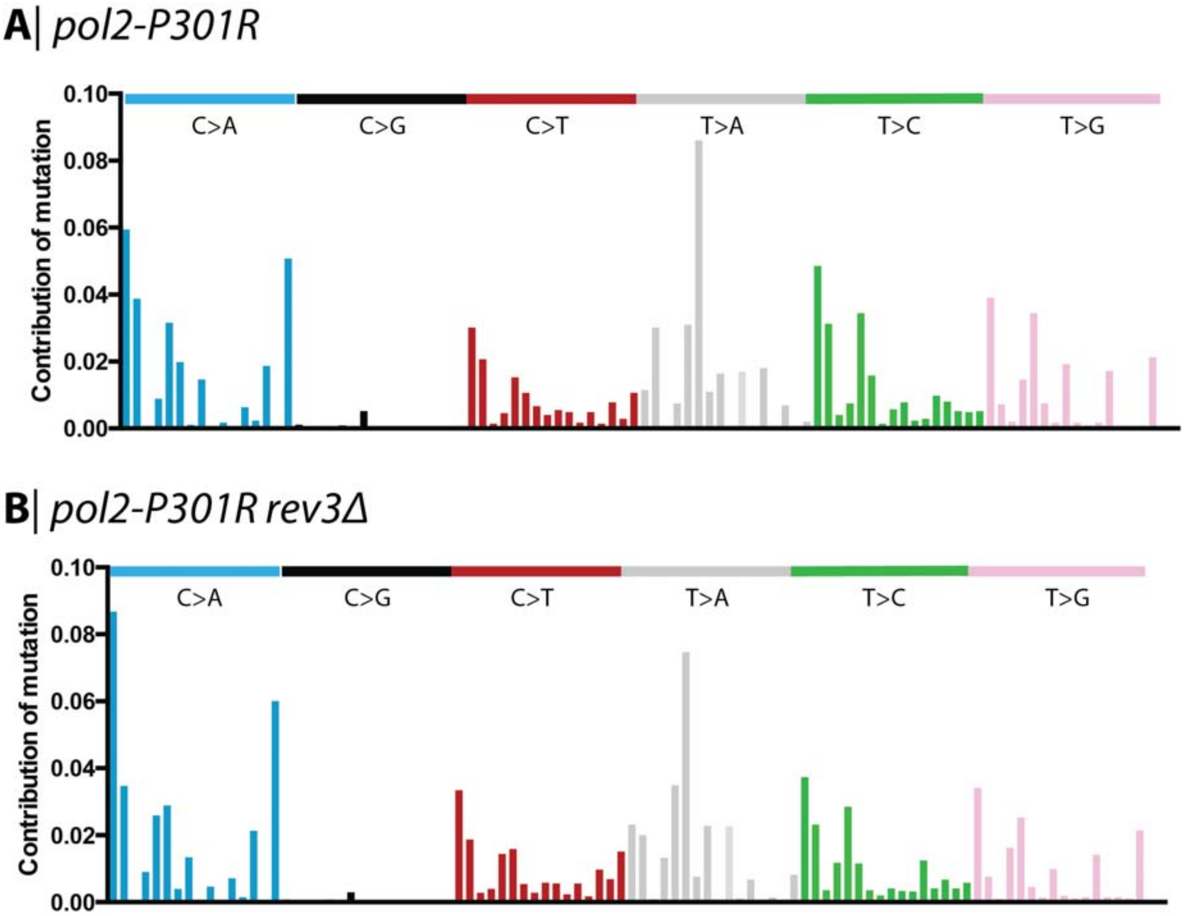


Figure 4.4: Mutational patterns observed in *pol2-P301R* cells and *pol2-P301R rev3*Δ cells are highly similar

All single nucleotide variants identified in the *pol2-P301R* cells and *pol2-P301R rev3*Δ cells displayed in Fig. 4.3 are displayed in their trinucleotide context. Relative contribution to the total of single nucelotide varaints is given. No adjustment for *S. cerevisiae* genome-wide trinucelotide 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 identfied 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 identfied. 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

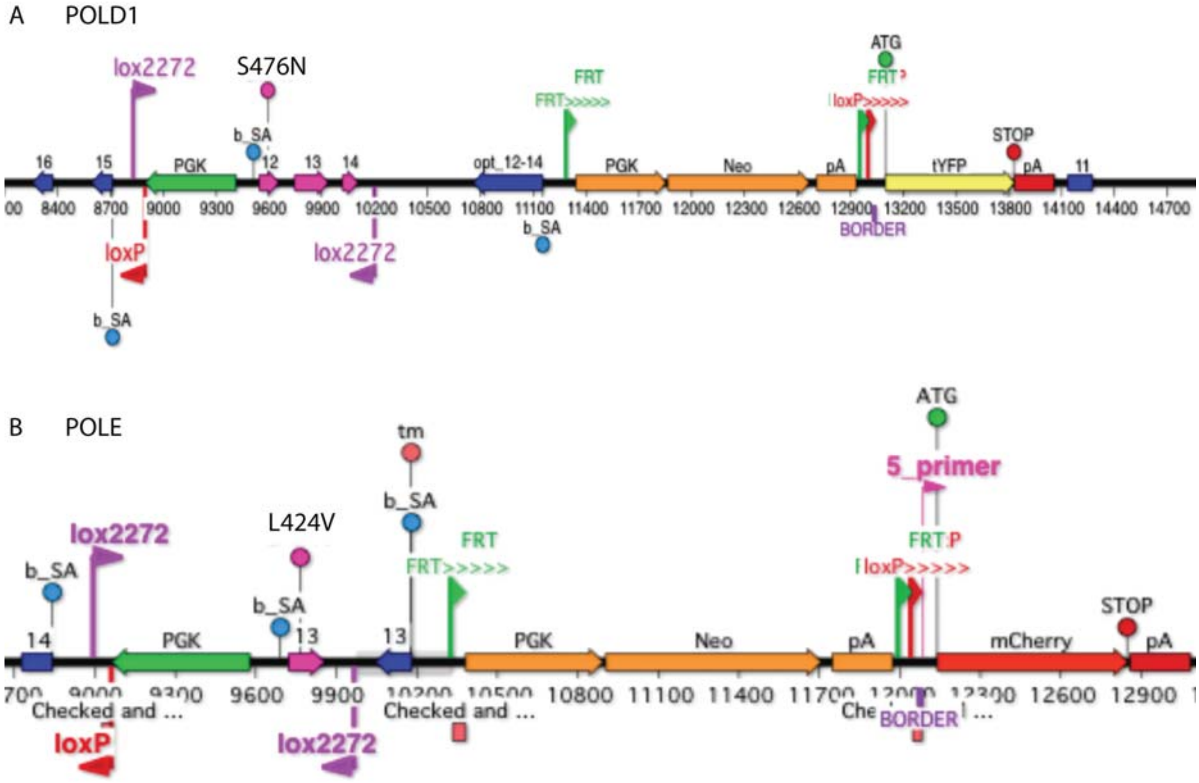


Figure 4.5: Constructs used for conditional knock-in mutations in mice
Constructs used to generate knock-in conditional mutants for **A|** *Pold1* S476N (mouse equivalent of human *POLE* p.L424V) and **B|** *Pole* L424V (mouse equivalent of *POLD1* S478N) in mouse mebryonic stem cells. Inversion of loxP 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 alowing the excision of the PGK Neo cassette used for clone selection. The construct was designed by Ian Tomlinson (Oxford).

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-Cas9^{D10A} 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 Cas9^{D10A} 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 polymarase 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 identfied 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 Polεexonuclease 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 complicates 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.

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 identfied 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* *vitro [282, 303, 312, 338, 347, 807]* .

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 Polεin 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 likley 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 Polεcauses 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 opperates in our exoncuelase 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 phenoype 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 childrens' brain tumors one, *POLE S297F*, is included in this work as well. Its yeast equivalent, *pol2-S312F*, is however lethal with *msh2*Δ 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 comapre 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 alogrithms 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 experimenst 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 identifed 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 willallow 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 *msh2*Δ 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 identfiy 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

H_2O up to 8L

LB-Amp

LB mix (FM) 200g

NaOH (10M) a few drops

800μl Ampicillin (50mg/ml)

H_2O up to 8L

LB-Amp-Agar

LB mix 125g

NaOH (10M) a few drops

500μl Ampicillin (50mg/ml)

H_2O up to 5L

Agar per 1L bottle 14g

6.1.2 *Saccharomyces cerevisiae* Growth Media

YPD/ rich medium

Yeast extract 10g

Peptone 20g

H_2O 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_2O 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_2O and sterilising. Stored at 4° C.

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 H_2O. 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 [1]

10X Glucose 40ml

100X Adenine 4ml

[1] SD plates containing G418 use MSG

100X Histidine 4ml

100X Tryptophan 4ml

100X Uracil 4ml

100X Leucine 4ml

100X Lysine 4ml

Other chemicals (G418, Thialysine) as needed

H_2O 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 *URA3* marker. *URA3*$^+$ cells die on FOA medium, while Ura$^-$ cells survive. The solution is added (after filter sterilisation once FOA has dissolved) to a sterile bottle of 200ml H_2O 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

H_2O up to 200ml

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_2O 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_2O 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_2O up to 2L

Sodium Chloride (5M)

NaCl (Fisher) 584.4g

H_2O up to 2L

Sodium dodecyl sulphate, SDS (20%)

SDS (Melford) 800g

H_2O up to 4L

TAE (50X)

Tris 1210g

Glacial acetic acid 285.5ml

EDTA 0.5M pH8.0

H_2O up to 5L

TBE (10X)

Tris (Melford) 540g

Orthoboric Acid (Fisher) 275g

EDTA (0.5M pH 8.0) 200ml

H_2O up to 5L

TE (pH 8.0)

1M Tris pH8.o

EDTA 0.5M pH8.0

H_2O up to 2L

Tris (1M, pH 6.8)

Tris (Melford) 242.2g

Conc. HCl to pH ~160ml

H_2O up to 2L

Tris (1M, pH 7.4)

Tris (Melford) 242.2g

Conc. HCl to pH ~146ml

H_2O up to 2L

Tris (1M, pH 7.5)

Tris (Melford) 242.2g

Conc. HCl to pH ~142ml

H_2O up to 2L

Tris (1M, pH 8.0)

Tris (Melford) 242.2g

Conc. HCl to pH ~96ml

H_2O up to 2L

Tris (1M, pH 8.8)

Tris (Melford) 242.2g

Conc. HCl to pH ~36ml

H_2O 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 lacI^{q}ZΔM15 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

Name	Genotype	Reference
K699	MATa ade2-1 trp1-1 leu2-3,112 his3-11,15 ura3 can1-100	Kim Nasmyth
K700	MATα ade2-1 trp1-1 leu2-3,112 his3-11,15 ura3 can1-100	Kim Nasmyth
YMH8	(K699) pol2::URA3-POL2	This work
YMH9	(YMH8)(K700)	This work
YMH10	(K699) pol3::URA3-pol3-S483N	This work
YMH11	(YMH10)(K700)	This work
YMH12	(K699) pol2::URA3-pol2-L439V	This work
YMH13	(YMH12)(K700)	This work
YMH14	(K699) pol2::URA3-pol2-V426L	This work
YMH15	(YMH14)(K700)	This work
YMH16	(K699) pol2::URA3-pol2-S312F	This work
YMH17	(YMH16)(K700)	This work
YMH18	(K699) pol2::URA3-pol2-P301R	This work
YMH19	(YMH18)(K700)	This work
YMH20	(K699) pol2::URA3-pol2-D290V	This work
YMH21	(YMH20)(K700)	This work
YMH22	(K699) pol2::URA3-pol2-M459K	This work
YMH23	(YMH22)(K700)	This work
YMH24	(K699) pol2::URA3-pol2-Q468R	This work
YMH25	(YMH24)(K700)	This work

Name	Genotype	Reference
YMH26	(K699) pol2::URA3-pol2-A480V	This work
YMH27	(YMH26)(K700)	This work
YMH28	(K699) pol2::URA3-pol2-4	This work
YMH29	(YMH28)(K700)	This work
YMH30	(K699) pol3::URA3-POL3	This work
YMH31	(K700)pol3::URA3-POL3	This work
YMH32	(K699) pol3::URA3-pol3-01	This work
YMH33	(K700) pol3::URA3-pol3-01	This work
YMH34	(K699) pol3::URA3-pol3-R316C	This work
YMH35	(K700) pol3::URA3-pol3-R316C	This work
YMH36	(K699) pol3::URA3-pol3-P332L	This work
YMH37	(K700) pol3::URA3-pol3-P332L	This work
YMH38	(K699) pol3::URA3-pol3-S375R	This work
YMH39	(K700) pol3::URA3-pol3-S375R	This work
YMH40	(K699) pol3::URA3-pol3-V397M	This work
YMH41	(K700) pol3::URA3-pol3-V397M	This work
YMH42	(K699) rev3::KanMX	This work
YMH43	(K699) pol2::URA3-pol2-A480V rev3::KanMX	This work
YMH44	(K699) pol2::URA3-pol2-P301R rev3::KanMX	This work
YMH46	(K699) pol2::URA3-pol2-S312F rev3::KanMX	This work
YMH52	(K700) pol2::URA3-pol2-S312F	This work
YMH53	(K700) pol2::URA3-pol2-P301R	This work
YMH54	(K699) msh2::KanMX	This work
YMH56	(K699) pol2::URA3-pol2-A480V msh2::KanMX	This work
YMH58	(K699) pol2::URA3-pol2-D290A-E292A-A480V	This work
YMH60	(K699) pol2::URA3-pol2-D290A-E292A-M459K	This work
YMH62	(K699) pol2::URA3-pol2-D290A-E292A-P301R	This work
YMH64	(K699) pol2::URA3-pol2-D290A-E292A-S312F	This work
YMH66	(K700) pol2::URA3-pol2-A480V	This work
YMH67	(K700) pol2::URA3-pol2-M459K	This work
YMH68	(YMH30)(K700)	This work
YMH69	(YMH36)(K700)	This work
YMH70	(YMH38)(K700)	This work

Name	Genotype	Reference
YMH71	(K700) pol3::URA3-pol3-01	This work
YMH72	(YMH34)(K700)	This work
YMH73-75	(K699) pol2::URA3-pol2-A480V msh2::KanMx	This work
YMH78	(K699)pol3::URA3-pol3-S483N(K700)pol2::URA3-pol2-P301R	This work
YMH81	(K699)(K700)pol2::URA3-pol2-P301R/pol2::URA3-pol2-P301R	This work
YMH82	(K699)(K700)pol2::URA3-pol2-M459K/POL2 msh2::KanMx/MSH2	This work
YMH83	(K699)(K700)pol2::URA3-pol2-S312F/POL2 msh2::KanMx/MSH2	This work
YMH84	(K699)(K700)pol2::URA3-pol2-P301R/POL2 msh2::KanMx/MSH2	This work
NOY408-1b	MATa ade2-1 ura3-1 his3-11 trp1-1 leu2-3,112 can1-100	Nogi et al. 1991
YSI101	(NOY408-1b) except fob1Δ::LEU2 (~150 rDNA copies)	[866]
YSI102	(NOY408-1b) except ~20 rDNA copies	[866]
YSI103	(NOY408-1b) except ~40 rDNA copies	[866]
YSI104	(NOY408-1b) except ~80 rDNA copies	[866]
YSI105	(NOY408-1b) except ~110 rDNA copies	[866]

6.4 Oligonucleotides

Oligonucleotides to generate POL2 mutants by site directed mutagenesis

pol2_S312F_SDM_fw

TAGATCAAATAATGATGATT**TTT**TATATGATCGATGGGGAAGG

pol2_S312F_SDM_rv

CCTTCCCCATCGATCATATA**AAA**AATCATCATTATTTGATCTA

pol2_V426L_SDM_fw

ACATGGATTGTTTCCGTTGG**CTG**AAGCGTGATTCTTATTTACC

pol2_V426L_SDM_rv

GGTAAATAAGAATCACGCTT**CAG**CCAACGGAAACAATCCATGT

pol2_P301R_SDM_fw

CGAAGCCGCCTTTAAAATTC**CGG**GATTCCGCCGTAGATCAAAT

pol2_P301R_SDM_rv

ATTTGATCTACGGCGGAATC**CCG**GAATTTTAAAGGCGGCTTCG

pol2_D290V_SDM_fw

ACCCTGTGGTAATGGCATTT**GTT**ATAGAAACCACGAAGCCGCC

pol2_D290V_SDM_rv

GGCGGCTTCGTGGTTTCTAT**AAC**AAATGCCATTACCACAGGGT

pol2_M459K_SDM_fw

TTGAACTGGATCCCGAATTA**AAG**ACGCCGTATGCATTTGAAAA

pol2_M459K_SDM_rv

TTTTCAAATGCATACGGCGT**CTT**TAATTCGGGATCCAGTTCAA

pol2_Q468R_SDM_fw

CGTATGCATTTGAAAAGCCA**CGG**CACCTTTCCGAATATTCTGT

pol2_Q468R_SDM_rv

ACAGAATATTCGGAAAGGTG**CCG**TGGCTTTTCAAATGCATACG

pol2_A480V_SDM_fw

ATTCTGTTTCCGATGCAGTC**GTT**ACGTATTACCTTTACATGAA

pol2_A480V_SDM_rv

TTCATGTAAAGGTAATACGT**AAC**GACTGCATCGGAAACAGAAT

Pol2Faiat-SDM-fw

CCCTGTGGTAATGGCATTT**GCT**ATA**GCA**ACCACGAAGCCGCCTTTAAA

Pol2Faiat-SDM-rv

TTTAAAGGCGGCTTCGGGT**TGC**TAT**AGC**AAATGCCATTACCACAGGG

Oligonucleotides to generate POL3 mutants by site directed mutagenesis

pol3_FAIAC_SDM_fw

TGCGTATCATGTCCTTT**GAT**ATC**GAG**TGTGCTGGTAGGATTGG

pol3_FAIAC_SDM_rv

CCAATCCTACCAGCACA**CTC**GAT**ATC**AAAGGACATGATACGCA

pol3_V397M_SDM_fw

TCATCAAAGTTGATCCTGAT**ATG**ATCATTGGTTATAATACTAC

pol3_V397M_SDM_rv

GTAGTATTATAACCAATGAT**CAT**ATCAGGATCAACTTTGATGA

pol3_R316C_SDM_fw

GGTCTCATACAGCTCCATTG**TGT**ATCATGTCCTTTGATATCGA

pol3_R316C_SDM_rv

TCGATATCAAAGGACATGAT**ACA**CAATGGAGCTGTATGAGACC

pol3_S483N_SDM_fw

CCTACACGTTGAATGCAGTC**AAT**GCGCACTTTTTAGGTGAACA

pol3_S483N_SDM_rv

TGTTCACCTAAAAAGTGCGC**ATT**GACTGCATTCAACGTGTAGG

pol3_P332L_SDM_fw
CTGGTAGGATTGGCGTCTTT**CTG**GAACCTGAATACGATCCCGT
pol3_P332L_SDM_rv
ACGGGATCGTATTCAGGTTC**CAG**AAAGACGCCAATCCTACCAG
pol3_S375R_SDM_fw
TAACAGGTTCAATGATTTTT**CGC**CACGCCACTGAAGAGGAAAT
pol3_S375R_SDM_rv
ATTTCCTCTTCAGTGGCGTG**GCG**AAAAATCATTGAACCTGTTA

Oligonucleotides to check polymerase mutant generation

Sc-pol3_out_fw GAAGAGCATGACCTGTCATCATTC
pRS306_out_rv GACCATGATTACGCCAAGCTCG
pol3_sq1 TACCAAAAGGAAAGTATTCG
pol3_sq2 GTCATCCAAATTGCCAACGT
pol3_sq3 ACTACAAATTTTGATATCCC
Pol2_promoter_rv: 5'GATCCATATTGCACACCAGAGCTGTT
pRS306_fw: 5'GGCGGACAGGTATCCGGTAAG

Oligonucleotides to delete MSH2

MSH2-F1 TTATCTGCTGACCTAACATCAAAATCCTCAGATTAAAAGT CGGATCCCCGGGTTAATTAA
MSH2-R1 TATCTATCGATTCTCACTTAAGATGTCGTTGTAATATTAA GAATTCGAGCTCGTTTAAAC
MSH2.3 TAAAGCCAATGAATTGGACG
MSH2.4 TTTCCAGTGGTCTAGAGACC

Oligonucleotides to delete REV3

REV3-F1 ATACAAAACTACAAGTTGTGGCGAAATAAAATGTTTGGAA CGGATCCCCGGGTTAATTAA
REV3-R1 ATAACTACTCATCATTTTGCGAGACATATCTGTGTCTAGA GAATTCGAGCTCGTTTAAAC
REV3.3 ACTGTTTAGAGAAAAGAAGC
REV3.4 AATGTGTGGGGAACTTATACG

Oligonucleotides to check the integration of polymerase mutant constructs into MEFs

EHom1_1fw GCTTGGGTGATGATGTTGGCTCCTGTAAA
EHom1_1rv CCGCGCTGTTCTCCTCTTCCTCATCTC
EHom2_1fw GCGGCATGGACGAGCTGTACAAGTGATTA
EHom2_1rv CCAGGACCTGCGGTAGTGGAAAGAGAAA
D1Hom1_1fw AGAGAATTGCTGAGAAAGGGGAGTGAGACA
D1Hom1_1rv CCGCGCTGTTCTCCTCTTCCTCATCTC
D1Hom2_1fw CCGCGATAATATGAGCCTGAAGGAGACCGT
D1Hom2_1rv TGGGTGGAGAAGGGCATCAGGAAGGAC

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

sgRNA-1 5'-ACCG-ATCTGGTCTGTCTCAGCATC-3'
and 5'-AAAC-GATGCTGAGACAGACCAGAT-3'
sgRNA-2 5'-ACCG-TCGATGGCCAGGTGAGCAGG-3'
and 5'-AAAC-CCTGCTCACCTGGCCATCGA-3'
ssODN (all mutations in lower case) CAAGGTCCCCATCCCAGGAGCTTACTTCCCAGAAGgCACCTGCTCACCTGGCCAT CGATCATGTAGGAAATCATCATAATCTGGTCTGTCTCAGCATCAcGAAAtTTGAG GGGCAGTTTGGTCGTCTCAATGTCAAATGCCAAAACCACAGGGTCCTGTGGGGA CAAAATAAGCATAAAGCCAAGCTCTAAACTCCCCA

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

dH_2O up to 50μl

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 Performed 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 5x10^6 and 1x10^7 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 l pFA6 template plasmid (1:20 QIA)

382.5μl l H2O

2.5μl l Taq/Dynazime

The solution is mixed and 100\mu l 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 30\mu l 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

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 0.9Msorbitol 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 zymoliase 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 resusupended 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 2hours 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 fiter 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 H_2O 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 ddH_2O. 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 forzen 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 aliged 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$^{\text{TM}}$ 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 signtures, the SomaticSignatures[763] R package was used. A customn 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 methodolgy[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.

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 alignent 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

- 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

Step5: 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

Step6: Sample files were intersected to remove any variants not aquired 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 varaints 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

Step7: 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

Step8(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

Step9: 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\t$m\t$o\t$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$t\t$s\n"; printf "$x\t$t\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

[1] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, 1953.

[2] M. H. Wilkins, A. R. Stokes, and H. R. Wilson. Molecular structure of deoxypentose nucleic acids. *Nature*, 171(4356):738–40, 1953.

[3] R. E. Franklin and R. G. Gosling. Molecular configuration in sodium thymonucleate. *Nature*, 171(4356):740–1, 1953.

[4] C. R. Darwin. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* John Murray, London, 5th edition edition, 1869.

[5] C. R. Darwin and A. R. Wallace. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *J Proc Linn Soc*, Zoology 3:45–62, 1858.

[6] G. Mendel. Versuche über Pflanzen-Hybriden. *Verhandlungen des Naturforschenden Vereines, Abhandlungen, Brünn*, pages 3–47, 1866.

[7] R. Dahm. Friedrich Miescher and the discovery of DNA. *Dev Biol*, 278(2):274–88, 2005.

[8] C. MacLeod. Oswald Theodore Avery, 1877-1955. *J Gen Microbiol*, 17(3):539–49, 1957. doi:10.1099/00221287-17-3-539.

[9] O. T. Avery. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *Journal of Experimental Medicine*, 79(2):137–158, 1944.

[10] A. D. Hershey. Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. *The Journal of General Physiology*, 36(1):39–56, 1952.

[11] M. O'Donnell, L. Langston, and B. Stillman. Principles and concepts of DNA replication in bacteria, archaea, and eukarya. *Cold Spring Harb Perspect Biol*, 5(7), 2013.

[12] N. E. Morton. Parameters of the human genome. *Proc Natl Acad Sci U S A*, 88(17):7474–6, 1991.

[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.

[14] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[15] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J.

Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, et al. The sequence of the human genome. *Science*, 291(5507):1304–51, 2001.

[16] J. Pellicer, M. F. Fay, and I. J. Leitch. The largest eukaryotic genome of them all? *Bot J Linn Soc*, 164(1):10–15, 2010.

[17] C. T. Friz. The biochemical composition of the free-living Amoebae Chaos chaos, Amoeba dubia and Amoeba proteus. *Comp Biochem Physiol*, 26(1):81–90, 1968.

[18] C. L. McGrath and L. A. Katz. Genome diversity in microbial eukaryotes. *Trends Ecol Evol*, 19(1):32–8, 2004.

[19] E. Chargaff, S. Zamenhof, and C. Green. Composition of human desoxypentose nucleic acid. *Nature*, 165(4202):756 – 757, 1950.

[20] J. D. Watson and F. H. Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361):964–7, 1953.

[21] *Nobel Lectures, Physiology or Medicine 1942-1962*. Elsevier Publishing Company, Amsterdam, 1964.

[22] M. Meselson and F. W. Stahl. The replication of DNA in Escherichia coli. *Proc Natl Acad Sci U S A*, 44(7):671–682, 1958.

[23] L. Pray. Discovery of DNA structure and function: Watson and Crick. *Nature Education*, 1(1):100, 2008.

[24] CyberBridge: Nucleotides and the double helix. url: http://cyberbridge.mcb.harvard.edu. Accessed 3rd June 2016.

[25] A. Johnson and M. O'Donnell. Cellular DNA replicases: components and dynamics at the replication fork. *Annu Rev Biochem*, 74:283–315, 2005.

[26] K. J. Marians. Prokaryotic DNA replication. *Annu Rev Biochem*, 61:673–719, 1992.

[27] M. O'Donnell, D. Jeruzalmi, and J. Kuriyan. Clamp loader structure predicts the architecture of DNA polymerase III holoenzyme and RFC. *Curr Biol*, 11(22):R935–R946, 2001.

[28] S. J. Benkovic, A. M. Valentine, and F. Salinas. Replisome-mediated DNA replication. *Annu Rev Biochem*, 70:181–208, 2001.

[29] C. S. McHenry. Chromosomal replicases as asymmetric dimers: studies of subunit arrangement and functional consequences. *Mol Microbiol*, 49(5):1157–1165, 2003.

[30] B. Grabowski and Z. Kelman. Archeal DNA replication: eukaryal proteins in a bacterial context. *Annu Rev Microbiol*, 57:487–516, 2003.

[31] W. F. McDonald, N. Klemperer, and P. Traktman. Characterization of a processive form of the vaccinia virus DNA polymerase. *Virology*, 234(1):168–75, 1997.

[32] T. R. Hernandez and I. R. Lehman. Functional interaction between the Herpes Simplex-1 DNA-Polymerase and Ul42 Protein. *J Biol Chem*, 265(19):11227–11232, 1990.

[33] F. R. Blattner. The Complete Genome Sequence of Escherichia coli K-12. *Science*, 277(5331):1453–1462, 1997.

[34] M. L. Mott and J. M. Berger. DNA replication initiation: mechanisms and regulation in bacteria. *Nat Rev Microbiol*, 5(5):343–54, 2007.

[35] K. E. Duderstadt, K. Chuang, and J. M. Berger. DNA stretching by bacterial initiators promotes replication origin opening. *Nature*, 478(7368):209–13, 2011.

[36] S. Zorman, H. Seitz, B. Sclavi, and T. R. Strick. Topological characterization of the DnaA-oriC complex using single-molecule nanomanipuation. *Nucleic Acids Res*, 40(15):7375–83, 2012.

[37] A. Robinson and A. M. van Oijen. Bacterial replication, transcription and translation: mechanistic insights from single-molecule biochemical studies. *Nat Rev Microbiol*, 11(5):303–15, 2013.

[38] N. Kresge, R. D. Simoni, and R. L. Hill. Arthur Kornberg's Discovery of DNA Polymerase I. *J Biol Chem*, 280(49), 2005.

[39] I. R. Lehman. Discovery of DNA polymerase. *J Biol Chem*, 278(37):34733–8, 2003.

[40] C. A. Wu, E. L. Zechner, J. A. Reems, C. S. McHenry, and K. J. Marians. Coordinated leading- and lagging-strand synthesis at the Escherichia coli DNA replication fork. V. Primase action regulates the cycle of Okazaki fragment synthesis. *J Biol Chem*, 267:4074–4083, 1992.

[41] J. P. Bouche, K. Zechel, and A. Kornberg. dnaG gene product, a rifampicin-resistant RNA polymerase, initiates the conversion of a single-stranded coliphage DNA to its duplex replicative form. *J Biol Chem*, 250(15):5995–6001, 1975.

[42] S. Wickner, M. Wright, and J. Hurwitz. Studies on In Vitro DNA Synthesis. * Purification of the dnaG Gene Product from Escherichia coli. *Proc Natl Acad Sci USA*, 70(5):1613–1618, 1973.

[43] X.-P. Kong, R. Onrust, M. O'Donnell, and J. Kuriyan. Three-dimensional structure of the beta subunit of E. coli DNA polymerase III holoenzyme: A sliding DNA clamp. *Cell*, 69(3):425–437, 1992.

[44] D. Jeruzalmi, O. Yurieva, Y. Zhao, M. Young, J. Stewart, M. Hingorani, M. O'Donnell, and J. Kuriyan. Mechanism of processivity clamp opening by the delta subunit wrench of the clamp loader Complex of E. coli DNA polymerase III. *Cell*, 106(4):417–428, 2001.

[45] H. Maki and A. Kornberg. The polymerase subunit of DNA polymerase III of Escherichia coli. II. Purification of the alpha subunit, devoid of nuclease activities. *J Biol Chem*, 260(24):12987–92, 1985.

[46] K. J. Marians. The interaction between helicase and primase sets the replication fork clock. *J Biol Chem*, 271(35):21398–21405, 1996.

[47] K. J. Marians. The Extreme C Terminus of Primase is Required for Interaction with DnaB at the Replication Fork. *J Biol Chem*, 271(35):21391–21397, 1996.

[48] N. K. Sinha, C. F. Morris, and B. M. Alberts. Efficient in vitro replication of double-stranded DNA templates by a purified T4 bacteriophage replication system. *J Biol Chem*, 255(9):4290–3, 1980.

[49] P. D. Chastain II, A. M. Makhov, N. G. Nossal, and J. Griffith. Architecture of the replication complex and DNA loops at the fork generated by the bacteriophage T4 proteins. *J Biol Chem*, 278(23):21276–85, 2003.

[50] B. M. Alberts, J. Barry, P. Bedinger, T. Formosa, C. V. Jongeneel, and K. N. Kreuzer. Studies on DNA Replication in the Bacteriophage T4 a–.gif System. *Cold Spring Harb Sym*, 47(0):655–668, 1983.

[51] T. Formosa, R. L. Burke, and B. M. Alberts. Affinity purification of bacteriophage T4 proteins essential for DNA replication and genetic recombination. *Proc Natl Acad Sci U S A*, 80(9):2442–6, 1983.

[52] K. Park, Z. Debyser, S. Tabor, C. C. Richardson, and J. D. Griffith. Formation of a DNA Loop at the Replication Fork Generated by Bacteriophage T7 Replication Proteins. *J Biol Chem*, 273(9):5260–5270, 1998.

[53] K. Skarstad and T. Katayama. Regulating DNA replication in bacteria. *Cold Spring Harb Perspect Biol*, 5(4):a012922, 2013.

[54] M. Manosas, M. M. Spiering, Z. Zhuang, S. J. Benkovic, and V. Croquette. Coupling DNA unwinding activity with primer synthesis in the bacteriophage T4 primosome. *Nat Chem Biol*, 5(12):904–12, 2009.

[55] S. Slater, S. Wold, M. Lu, E. Boye, K. Skarstad, and N. Kleckner. E. coli SeqA protein binds oriC in two different methyl-modulated reactions appropriate to its roles in DNA replication initiation and origin sequestration. *Cell*, 82(6):927–936, 1995.

[56] M. Lu. SeqA: A negative modulator of replication initiation in E. coli. *Cell*, 77(3):413–426, 1994.

[57] E. Boye, A. Lobner-Olesen, and K. Skarstad. Limiting DNA replication to once and only once. *EMBO Rep*, 1(6):479–83, 2000.

[58] *Les Prix Nobel. The Nobel Prizes 2001*. Nobel Foundation, Stockholm, 2002.

[59] K. Siddiqui, K. F. On, and J. F. Diffley. Regulating DNA replication in eukarya. *Cold Spring Harb Perspect Biol*, 5(9), 2013.

[60] J. M. Peters. The anaphase promoting complex/cyclosome: a machine designed to destroy. *Nat Rev Mol Cell Biol*, 7(9):644–56, 2006.

[61] D. Coudreuse and P. Nurse. Driving the cell cycle with a minimal CDK control network. *Nature*, 468(7327):1074–9, 2010.

[62] V. Fantl, G. Stamp, A. Andrews, I. Rosewell, and C. Dickson. Mice lacking cyclin D1 are small and show defects in eye and mammary gland development. *Genes Dev*, 9(19):2364–2372, 1995.

[63] P. Sicinski, J. L. Donaher, S. B. Parker, T. Li, A. Fazeli, H. Gardner, S. Z. Haslam, R. T. Bronson, S. J. Elledge, and R. A. Weinberg. Cyclin D1 provides a link between development and oncogenesis in the retina and breast. *Cell*, 82(4):621–630, 1995.

[64] P. Sicinski, J. L. Donaher, Y. Geng, S. B. Parker, H. Gardner, M. Y. Park, R. L. Robker, J. S. Richards, L. K. McGinnis, J. D. Biggers, J. J. Eppig, R. T. Bronson, S. J. Elledge, and R. A. Weinberg. Cyclin D2 is an FSH-responsive gene involved in gonadal cell proliferation and oncogenesis. *Nature*, 384(6608):470–4, 1996.

[65] E. Sicinska, I. Aifantis, L. Le Cam, W. Swat, C. Borowski, Q. Yu, A. A. Ferrando, S. D. Levin, Y. Geng, H. von Boehmer, and P. Sicinski. Requirement for cyclin D3 in lymphocyte development and T cell leukemias. *Cancer Cell*, 4(6):451–461, 2003.

[66] M. A. Ciemerych, A. M. Kenney, E. Sicinska, I. Kalaszczynska, R. T. Bronson, D. H. Rowitch, H. Gardner, and P. Sicinski. Development of mice expressing a single D-type cyclin. *Genes Dev*, 16(24):3277–89, 2002.

[67] K. Kozar, M. A. Ciemerych, V. I. Rebel, H. Shigematsu, A. Zagozdzon, E. Sicinska, Y. Geng, Q. Yu, S. Bhattacharya, R. T. Bronson, K. Akashi, and P. Sicinski. Mouse development and cell proliferation in the absence of D-cyclins. *Cell*, 118(4):477–91, 2004.

[68] M. Malumbres, R. Sotillo, D. Santamaria, J. Galan, A. Cerezo, S. Ortega, P. Dubus, and M. Barbacid. Mammalian cells cycle without the D-type cyclin-dependent kinases Cdk4 and Cdk6. *Cell*, 118(4):493–504, 2004.

[69] S. G. Rane, P. Dubus, R. V. Mettus, E. J. Galbreath, G. Boden, E. P. Reddy, and M. Barbacid. Loss of Cdk4 expression causes insulin-deficient diabetes and Cdk4 activation results in beta-islet cell hyperplasia. *Nat Genet*, 22(1):44–52, 1999.

[70] C. Berthet, E. Aleem, V. Coppola, L. Tessarollo, and P. Kaldis. Cdk2 Knockout Mice Are Viable. *Curr Biol*, 13(20):1775–1785, 2003.

[71] S. Ortega, I. Prieto, J. Odajima, A. Martin, P. Dubus, R. Sotillo, J. L. Barbero, M. Malumbres, and M. Barbacid. Cyclin-dependent kinase 2 is essential for meiosis but not for mitotic cell division in mice. *Nat Genet*, 35(1):25–31, 2003.

[72] Y. Geng, Q. Yu, E. Sicinska, M. Das, J. E. Schneider, S. Bhattacharya, W. M. Rideout, R. T. Bronson, H. Gardner, and P. Sicinski. Cyclin E Ablation in the Mouse. *Cell*, 114(4):431–443, 2003.

[73] T. Parisi, A. R. Beck, N. Rougier, T. McNeil, L. Lucian, Z. Werb, and B. Amati. Cyclins E1 and E2 are required for endoreplication in placental trophoblast giant cells. *EMBO J*, 22(18):4794–803, 2003.

[74] C. J. Sherr and J. M. Roberts. Living with or without cyclins and cyclin-dependent kinases. *Genes Dev*, 18(22):2699–711, 2004.

[75] S. S. Hook, J. J. Lin, and A. Dutta. Mechanisms to control rereplication and implications for cancer. *Curr Opin Cell Biol*, 19(6):663–71, 2007.

[76] J. G. Cook. Replication licensing and the DNA damage checkpoint. *Front Biosci*, 14(1):5013, 2009.

[77] B. M. Green, K. J. Finn, and J. J. Li. Loss of DNA replication control is a potent inducer of gene amplification. *Science*, 329(5994):943–6, 2010.

[78] P. Garg and P. M. Burgers. DNA polymerases that propagate the eukaryotic DNA replication fork. *Crit Rev Biochem Mol Biol*, 40(2):115–28, 2005.

[79] T. J. Kelly and G. W. Brown. Regulation of chromosome replication. *Annu Rev Biochem*, 69:829–80, 2000.

[80] D. Rhodes. Chromatin structure: The nucleosome core all wrapped up. *Nature*, 389(389):231–232, 1997.

[81] G. J. Filion, J. G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L. D. Ward, W. Brugman, I. J. de Castro, R. M. Kerkhoven, and B. Bussemaker, H. J.and van Steensel. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell*, 143(2):212–224, 2010.

[82] J. Hansen. Human mitotic chromosome structure: what happened to the 30-nm fibre? *EMBO J*, 31(7):1621–1623, 2012.

[83] K. Salma and P. T Benjamin. Gatekeepers of chromatin: Small metabolites elicit big changes in gene expression. *Trends Biochem Sci*, 37(11):477–483, 2012.

[84] C. Alabert and A. Groth. Chromatin replication and epigenome maintenance. *Nat Rev Mol Cell Bio*, 13:153–167, 2012.

[85] Y. Marahrens and B. Stillman. A yeast chromosomal origin of DNA replication defined by multiple functional elements. *Science*, 255(5046):817–823, 1992.

[86] J. V. Van Houten and C. S. Newlon. Mutational analysis of the consensus sequence of a replication origin from yeast chromosome III. *Mol Cell Biol*, 10(8):3917–3925, 1990.

[87] J. R. Broach, Y. Y. Li, J. Feldman, M. Jayaram, J. Abraham, K. A. Nasmyth, and J. B. Hicks. Localization and Sequence Analysis of Yeast Origins of DNA Replication. *Cold Spring Harb Sym*, 47(0):1165–1173, 1983.

[88] S. P. Bell and B. Stillman. ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature*, 357(6374):128–34, 1992.

[89] H. Rao, Y. Marahrens, and B. Stillman. Functional conservation of multiple elements in yeast chromosomal replicators. *Mol Cell Biol*, 14(11):7643–7651, 1994.

[90] J. F. Theis and C. S. Newlon. Domain B of ARS307 contains two functional elements and contributes to chromosomal replication origin function. *Mol Cell Biol*, 14(11):7652–7659, 1994.

[91] S. Lin and D. Kowalski. Functional equivalency and diversity of cis-acting elements among yeast replication origins. *Mol Cell Biol*, 17(9):5473–5484, 1997.

[92] H. Rao and B. Stillman. The origin recognition complex interacts with a bipartite DNA binding site within yeast replicators. *Proc Natl Acad Sci USA*, 92:2224–28, 1995.

[93] A. Rowley, J. H. Cocker, J. Harwood, and J. F. X. Diffley. Initiation complex assembly at budding yeast replication origins begins with the recognition of a bipartite sequence by limiting amounts of the initiator, ORC. *EMBO J*, 14:2631–2641, 1995.

[94] J. F. X. Diffley and B. Stillman. Purification of a yeast protein that binds to origins of DNA replication and a transcriptional silencer. *Proc Natl Acad Sci USA*, 85:2120–2124, 1988.

[95] M. Segurado, A. de Luis, and F. Antequera. Genome-wide distribution of DNA replication origins at A+T-rich islands in Schizosaccharomyces pombe. *EMBO Rep*, 4(11):1048–1053, 2003.

[96] M. Hayashi, Y. Katou, T. Itoh, A. Tazumi, Y. Yamada, T. Takahashi, T. Nakagawa, K. Shirahige, and H. Masukata. Genome-wide localization of pre-RC sites and identification of replication origins in fission yeast. *EMBO J*, 26(5):1327–39, 2007.

[97] J. C. Cadoret, F. Meisch, V. Hassan-Zadeh, I. Luyten, C. Guillet, L. Duret, H. Quesneville, and M. N. Prioleau. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc Natl Acad Sci U S A*, 105(41):15837–42, 2008.

[98] D. D. Dubey, J. Zhu, D. L. Carlson, K. Sharma, and J. A. Huberman. Three ARS elements contribute to the ura4 replication origin region in the fission yeast, Schizosaccharomyces pombe. *EMBO J*, 13(15):3638–3647, 1994.

[99] R. K. Clyne and T. J. Kelly. Genetic analysis of an ARS element from the fission yeast Schizosaccharomyces pombe. *EMBO J*, 14(24):6348–6357, 1995.

[100] D. D. Dubey, S. Kim, I. T. Todorov, and J. A. Huberman. Large, complex modular structure of a fission yeast DNA replication origin. *Curr Biol*, 6(4):467–473, 1996.

[101] M. L. DePamphilis. Replication origins in metazoan chromosomes: fact or fiction? *BioEssays*, 21(1):5–16, 1999.

[102] T. Kobayashi, T. Rein, and M. L. DePamphilis. Identification of Primary Initiation Sites for DNA replication in the Hamster Dihydrofolate Reductase Gene Initiation Zone. *Mol Cell Biol*, 18(6):3266–3277, 1998.

[103] P. Françon, D. Maiorano, and M. Méchali. Initiation of DNA replication in eukaryotes: questioning the origin. *FEBS Lett*, 452(1-2):201–213, 1999.

[104] P. A. Dijkwel, J. P. Vaughn, and J. L. Hamlin. Mapping of replication initiation sites in mammalian genomes by two-dimensional gel analysis: stabilization and enrichment of replication intermediates by isolation on the nuclear matrix. *Mol Cell Biol*, 11(8):3850–9, 1991.

[105] A. C. Spradling. ORC binding, gene amplification, and the nature of metazoan replication origins. *Genes Dev*, 13(20):2619–23, 1999.

[106] M. Muzi-Falconi and T. J. Kelly. Orp1, a member of the Cdc18/cdc6 family of S-phase regulators, is homologous to a component of the origin recognition complex. *Proc. Natl. Acad. Sci. USA*, 92(26):12475–12479, 1995.

[107] P. B. Carpenter and W. G. Dunphy. Identification of a novel 81-kDa component of the Xenopus Origin Recognition Complex. *J. Biol. Chem.*, 273:24891–24897, 1998.

[108] A. Rowles, J. P. J. Chong, L. Brown, M. Howell, G. I. Evan, and J. J. Blow. Interaction between the Origin Recognition Complex and the replication licensing systemin Xenopus. *Cell*, 87(2):287–296, 1996.

[109] M. Gossen, D. T. Pak, S. K. Hansen, J. K. Acharya, and M. R. Botchan. A Drosophila homolog of the Yeast Origin Recognition Complex. *Science*, 270(5242):1674–1677, 1995.

[110] D. G. Quintana, Z. Hou, K. C. Thome, M. Hendricks, P. Saha, and A. Dutta. Identification of HsORC4, a member of the human Origin of Replication Recognition Complex. *J Biol Chem*, 272:28247–28251, 1997.

[111] K. Takahara, M. Bong, R. Brevard, R. L. Eddy, L. L. Haley, S. J. Sait, T. B. Shows, G. G. Hoffman, and D. S. Greenspan. Mouse and Human homologues of the Yeast Origin of Replication Recognition Complex subunit ORC2 and chromosomal localization of the cognate human gene ORC2L. *Genomics*, 31(1):119–122, 1996.

[112] K. A. Gavin, M. Hidaka, and B. Stillman. Conserved initiator proteins in Eukaryotes. *Science*, 270(5242):1667–71, 1996.

[113] S. P. Bell and A. Dutta. DNA replication in eukaryotic cells. *Annu Rev Biochem*, 71:333–74, 2002.

[114] E.E. Arias and J.C. Walter. Strength in numbers: preventing rereplication via multiple mechanisms in eukaryotic cells. *Genes Dev*, 21(5):497–51, 2007.

[115] L.S. Drury and J.F. Diffley. Factors affecting the diversity of DNA replication licensing control in eukaryotes. *Curr Biol*, 19(6):530–5, 2009.

[116] J. F. Diffley and J. H. Cocker. Protein-DNA interactions at a yeast replication origin. *Nature*, 357(6374):169–72, 1992.

[117] R.-Y. Chuang and T. J. Kelly. The fission yeast homologue of Orc4p binds to replication origin DNA via multiple AT-hooks. *Proc Natl Acad Sci USA*, 96(6):2656–2661, 1999.

[118] M. Volkening and I. Hoffmann. Involvement of human MCM8 in prereplication complex assembly by recruiting hcdc6 to chromatin. *Mol Cell Biol*, 25(4):1560–8, 2005.

[119] J. H. Cocker, S. Piatti, C. Santocanale, K. Nasmyth, and J. F. X. Diffley. An essential role for the Cdc6 protein in forming the pre-replicative complexes of budding yeast. *Nature*, 379:180–182, 1996.

[120] S. Chen, M.A. de Vries, and S.P. Bell. Orc6 is required for dynamic recruitment of Cdt1 during repeated Mcm2-7 loading. *Genes Dev*, 21(22):2897–907, 2007.

[121] J. C. W. Randell, J. L. B. Bowers, H. K. Rodríguez, and S. P. Bell. Sequential ATP hydrolysis by Cdc6 and ORC directs loading of the Mcm2-7 helicase. *Mol Cell*, 21(1):29–39, 2006.

[122] M. Rialland, F. Sola, and C. Santocanale. Essential role of human CDT1 in DNA replication and chromatin licensing. *J Cell Sci*, 115:1435–1440, 2002.

[123] H. Nishitani, Z. Lygerou, T. Nishimoto, and P. Nurse. The Cdt1 protein is required to license DNA for replication in fission yeast. *Nature*, 404:625–628, 2000.

[124] D. Maiorano, J. Moreau, and M. Méchali. XCDT1 is required for the assembly of pre-replicative complexes in Xenopus laevis. *Nature*, 404:622–625, 2000.

[125] S. Donovan, J. Harwood, L. S. Drury, and J. F. X. Diffley. Cdc6p-dependent loading of Mcm proteins onto pre-replicative chromatin in budding yeast. *Proc. Natl. Acad. Sci. USA*, 94:5611–5616, 1997.

[126] A. J. Whittaker, I. Royzman, and T. O. Orr-Weaver. Drosophila Double parked: a conserved, essential replication protein that colocalizes with the origin recognition complex and links DNA replication with mitosis and the down-regulation of S phase transcripts. *Genes Dev*, 14:1765–1776, 2000.

[127] P. J. Gillespie, A. Li, and J. J. Blow. Reconstitution of licensed replication origins on Xenopus sperm nuclei using purified proteins. *BMC Biochem*, 2:15, 2001.

[128] N. Mailand and J. F. X. Diffley. CDKs promote DNA replication origin licensing in human cells by protecting Cdc6 from APC/C-dependent proteolysis. *Cell*, 122(6):915–926, 2005.

[129] A. Svitin and I. Chesnokov. Study of DNA replication in Drosophila using cell free in vitro system. *Cell Cycle*, 9(4):815–9, 2010.

[130] A. Gambus, G. A. Khoudoli, R. C. Jones, and J. J. Blow. MCM2-7 form double hexamers at licensed origins in Xenopus egg extract. *J Biol Chem*, 286:11855–11864, 2011.

[131] J. L. Bowers, J. C. W. Randell, S. Chen, and S. P. Bell. ATP hydrolysis by ORC catalyzes reiterative Mcm2-7 assembly at a defined Origin of Replication. *Mol Cell*, 16(6):967–978, 2004.

[132] C. Speck, Z. Chen, H. Li, and B. Stillman. ATPase-dependent cooperative binding of ORC and Cdc6 to origin DNA. *Nat Struct Mol Biol*, 12:965–971, 2005.

[133] C. Evrin, P. Clarke, J. Zech, R. Lurz, J. Sun, S. Uhle, H. Li, B. Stillman, and C. Speck. A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. *Proc Natl Acad Sci USA*, 106(48):20240–20245, 2009.

[134] A. Rowles, S. Tada, and J. J. Blow. Changes in association of the Xenopus origin recognition complex with chromatin on licensing of replication origins. *J Cell Sci*, 112:2011– 2018, 1999.

[135] T. Seki and J. F. X. Diffley. Stepwise assembly of initiation proteins at budding yeast replication origins in vitro. *Proc Natl Acad Sci*, 97(26):14115–14120, 2000.

[136] D. Remus, F. Beuron, G. Tolun, J. D. Griffith, E. P. Morris, and J. F. Diffley. Concerted loading of Mcm2-7 double hexamers around DNA during DNA replication origin licensing. *Cell*, 139:719–730, 2009.

[137] A. Gambus, G. A. Khoudoli, R. C. Jones, and J. J. Blow. MCM2-7 form double hexamers at licensed origins in Xenopus egg extract. *J Biol Chem*, 286:11855–11864, 2011.

[138] D. Shechter, C. Y. Ying, and J. Gautier. DNA unwinding is an MCM complex-dependent and ATP hydrolysis-dependent process. *J Biol Chem*, 279:45586–45593, 2004.

[139] S.P. Bell and A. Dutta. DNA replication in eukaryotic cells. *Annu Rev Biochem*, 71:333–374, 2002.

[140] L. S. Drury and J. F. Diffley. Factors affecting the diversity of DNA replication licensing control in eukaryotes. *Curr Biol*, 19(6):530–5, 2009. doi: 10.1016/j.cub.2009.02.034.

[141] S. E. Moyer, P. W. Lewis, and M. R. Botchan. Isolation of the Cdc45/Mcm2-7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. *Proc Natl Acad Sci U S A*, 103(27):10236–41, 2006.

[142] T. Aparicio, E. Guillou, J. Coloma, G. Montoya, and J. Mendez. The human GINS complex associates with Cdc45 and MCM and is essential for DNA replication. *Nucleic Acids Res*, 37(7):2087–95, 2009.

[143] K. Bousset and J. F. Diffley. The Cdc7 protein kinase is required for origin firing during S phase. *Genes Dev*, 12(4):480–90, 1998.

[144] Y. Kamimura, H. Masumoto, A. Sugino, and H. Araki. Sld2, which interacts with Dpb11 in Saccharomyces cerevisiae, is required for chromosomal DNA replication. *Mol Cell Biol*, 18(10):6102–6109, 1998.

[145] Y. Kamimura, Y. S. Tak, A. Sugino, and H. Araki. Sld3, which interacts with Cdc45 (Sld4), functions for chromosomal DNA replication in Saccharomyces cerevisiae. *EMBO J*, 20(8):2097–107, 2001.

[146] L. Zou and B. Stillman. Formation of a preinitiation complex by S-phase cyclin CDK-dependent loading of Cdc45p onto chromatin. *Science*, 280(5363):593–6, 1998.

[147] H. Masumoto, S. Muramatsu, Y. Kamimura, and H. Araki. S-cdk-dependent phosphorylation of Sld2 essential for chromosomal DNA replication in budding yeast. *Nature*, 415(6872):651–5, 2002.

[148] S. Tanaka, T. Umemori, K. Hirai, S. Muramatsu, Y. Kamimura, and H. Araki. CDK-dependent phosphorylation of Sld2 and Sld3 initiates DNA replication in budding yeast. *Nature*, 445(7125):328–32, 2007.

[149] T. Tanaka, T. Umemori, S. Endo, S. Muramatsu, M. Kanemaki, Y. Kamimura, C. Obuse, and H. Araki. Sld7, an Sld3-associated protein required for efficient chromosomal DNA replication in budding yeast. *EMBO J*, 30(10):2019–30, 2011.

[150] P. Zegerman and J. F. X. Diffley. Phosphorylation of Sld2 and Sld3 by cyclin-dependent kinases promotes DNA replication in budding yeast. *Nature*, 445(7125):281–5, 2007.

[151] C. F. Hardy, O. Dryga, S. Seematter, P. M. Pahl, and R. A. Sclafani. mcm5/cdc46-bob1 bypasses the requirement for the S phase activator Cdc7p. *Proc Natl Acad Sci U S A*, 94(7):3151–5, 1997.

[152] U. P. Strausfeld, M. Howell, R. Rempel, J. L. Maller, T. Hunt, and J. J. Blow. Cip1 blocks the initiation of DNA replication in Xenopus extracts by inhibition of cyclin-dependent kinases. *Curr Biol*, 4(10):876–83, 1994.

[153] P. K. Jackson, S. Chevalier, M. Philippe, and M. W. Kirschner. Early events in DNA replication require cyclin E and are blocked by p21CIP1. *J Cell Biol*, 130(4):755–69, 1995.

[154] W. Jiang, D. McDonald, T. J. Hope, and T. Hunter. Mammalian Cdc7-Dbf4 protein kinase complex is essential for initiation of DNA replication. *EMBO J*, 18(20):5703–13, 1999.

[155] J. C. Walter. Evidence for sequential action of cdc7 and cdk2 protein kinases during initiation of DNA replication in Xenopus egg extracts. *J Biol Chem*, 275(50):39773–8, 2000.

[156] S. Waga and B. Stillman. The DNA replication fork in eukaryotic cells. *Annu Rev Biochem*, 67:721–51, 1998.

[157] L. I. Francis, J. C. Randell, T. J. Takara, L. Uchima, and S. P. Bell. Incorporation into the prereplicative complex activates the Mcm2-7 helicase for Cdc7-Dbf4 phosphorylation. *Genes Dev*, 23(5):643–54, 2009.

[158] Y. J. Sheu and B. Stillman. Cdc7-Dbf4 phosphorylates MCM proteins via a docking site-mediated mechanism to promote S phase progression. *Mol Cell*, 24(1):101–13, 2006.

[159] Y. J. Sheu and B. Stillman. The Dbf4-Cdc7 kinase promotes S phase by alleviating an inhibitory activity in Mcm4. *Nature*, 463(7277):113–7, 2010.

[160] S. Tanaka, R. Nakato, Y. Katou, K. Shirahige, and H. Araki. Origin association of Sld3, Sld7, and Cdc45 proteins is a key step for determination of origin-firing timing. *Curr Biol*, 21(24):2055–63, 2011.

[161] M. N. Sangrithi, J. A. Bernal, M. Madine, A. Philpott, J. Lee, W. G. Dunphy, and A. R. Venkitaraman. Initiation of DNA replication requires the RECQL4 protein mutated in Rothmund-Thomson syndrome. *Cell*, 121(6):887–98, 2005.

[162] A. Kumagai, A. Shevchenko, A. Shevchenko, and W. G. Dunphy. Treslin collaborates with TopBP1 in triggering the initiation of DNA replication. *Cell*, 140(3):349–59, 2010.

[163] L. Sanchez-Pulido, J. F. Diffley, and C. P. Ponting. Homology explains the functional similarities of Treslin/Ticrr and Sld3. *Curr Biol*, 20(12):R509–10, 2010.

[164] C. L. Sansam, N. M. Cruz, P. S. Danielian, A. Amsterdam, M. L. Lau, N. Hopkins, and J. A. Lees. A vertebrate gene, ticrr, is an essential checkpoint and replication regulator. *Genes Dev*, 24(2):183–94, 2010.

[165] D. Boos, L. Sanchez-Pulido, M. Rappas, L. H. Pearl, A. W. Oliver, C. P. Ponting, and J. F. Diffley. Regulation of DNA replication through Sld3-Dpb11 interaction is conserved from yeast to humans. *Curr Biol*, 21(13):1152–7, 2011.

[166] A. Ballabeni, R. Zamponi, G. Caprara, M. Melixetian, S. Bossi, L. Masiero, and K. Helin. Human CDT1 associates with CDC7 and recruits CDC45 to chromatin during S phase. *J Biol Chem*, 284(5):3028–36, 2009.

[167] Z. You and H. Masai. Cdt1 forms a complex with the minichromosome maintenance protein (MCM) and activates its helicase activity. *J Biol Chem*, 283(36):24469–77, 2008.

[168] O. M. Aparicio, D. M. Weinstein, and S. P. Bell. Components and dynamics of DNA replication complexes in S. cerevisiae: Redistribution of MCM proteins and Cdc45p during S phase. *Cell*, 91(1):59–69, 1997.

[169] J. J. Blow and A. Dutta. Preventing re-replication of chromosomal DNA. *Nat Rev Mol Cell Biol*, 6(6):476–86, 2005.

[170] L. S. Drury, G. Perkins, and J. F. Diffley. The cyclin-dependent kinase Cdc28p regulates distinct modes of Cdc6p proteolysis during the budding yeast cell cycle. *Curr Biol*, 10(5):231–40, 2000.

[171] S. Elsasser, Y. Chi, P. Yang, and J. L. Campbell. Phosphorylation controls timing of Cdc6p destruction: A biochemical analysis. *Mol Biol Cell*, 10(10):3263–3277, 1999.

[172] G. Perkins, L. S. Drury, and J. F. Diffley. Separate SCF(CDC4) recognition elements target Cdc6 for proteolysis in S phase and mitosis. *EMBO J*, 20(17):4836–45, 2001.

[173] L. S. Drury, G. Perkins, and J. F. Diffley. The Cdc4/34/53 pathway targets Cdc6p for proteolysis in budding yeast. *EMBO J*, 16(19):5966–76, 1997.

[174] S. Mimura, T. Seki, S. Tanaka, and J. F. Diffley. Phosphorylation-dependent binding of mitotic cyclins to Cdc6 contributes to DNA replication control. *Nature*, 431(7012):1118–23, 2004.

[175] M. Schwab, A. Schulze Lutum, and W. Seufert. Yeast Hct1 is a regulator of Clb2 cyclin proteolysis. *Cell*, 90(4):683–693, 1997.

[176] R. Visintin. CDC20 and CDH1: A family of substrate-specific activators of APC-dependent proteolysis. *Science*, 278(5337):460–463, 1997.

[177] T. T. Nugroho and M. D. Mendenhall. An inhibitor of yeast cyclin-dependent protein kinase plays an important role in ensuring the genomic integrity of daughter cells. *Mol Cell Biol*, 14(5):3320–3328, 1994.

[178] E. Schwob. The B-type cyclin kinase inhibitor p40SIC1 controls the G1 to S transition in S. cerevisiae. *Cell*, 79(2):233–244, 1994.

[179] R. Visintin, K. Craig, E. S. Hwang, S. Prinz, M. Tyers, and A. Amon. The phosphatase Cdc14 triggers mitotic exit by reversal of Cdk-dependent phosphorylation. *Mol Cell*, 2(6):709–718, 1998.

[180] D. Knapp, L. Bhoite, D. J. Stillman, and K. Nasmyth. The transcription factor Swi5 regulates expression of the cyclin kinase inhibitor p40SIC1. *Mol Cell Biol*, 16(10):5701–7, 1996.

[181] J. D. Donovan, J. H. Toyn, A. L. Johnson, and L. H. Johnston. P40SDB25, a putative CDK inhibitor, has a role in the M/G1 transition in Saccharomyces cerevisiae. *Genes Dev*, 8(14):1640–1653, 1994.

[182] B. L. Schneider, Q. H. Yang, and A. B. Futcher. Linkage of replication to start by the Cdk inhibitor Sic1. *Science*, 272(5261):560–562, 1996.

[183] R. M. R. Feldman, C. C. Correll, K. B. Kaplan, and R. J. Deshaies. A complex of Cdc4p, Skp1p, and Cdc53p/Cullin catalyzes ubiquitination of the phosphorylated CDK inhibitor Sic1p. *Cell*, 91(2):221–230, 1997.

[184] R. Verma, R. M. Feldman, and R. J. Deshaies. SIC1 is ubiquitinated in vitro by a pathway that requires CDC4, CDC34, and cyclin/CDK activities. *Mol Cell Biol*, 8(8):1427–1437, 1997.

[185] M. Koivomagi, E. Valk, R. Venta, A. Iofik, M. Lepiku, E. R. Balog, S. M. Rubin, D. O. Morgan, and M. Loog. Cascades of multisite phosphorylation control Sic1 destruction at the onset of S phase. *Nature*, 480(7375):128–31, 2011.

[186] G. Xouri, M. Dimaki, P. I. H. Bastiaens, and Z. Lygerou. Cdt1 interactions in the licensing process: A model for dynamic spatio-temporal control of licensing. *Cell Cycle*, 6(13):1549–1552, 2014.

[187] M. Weinreich, C. Liang, H. H. Chen, and B. Stillman. Binding of cyclin-dependent kinases to ORC and Cdc6p regulates the chromosome replication cycle. *Proc Natl Acad Sci U S A*, 98(20):11211–7, 2001.

[188] G. M. Wilmes, V. Archambault, R. J. Austin, M. D. Jacobson, S. P. Bell, and F. R. Cross. Interaction of the S-phase cyclin Clb5 with an "RXL" docking sequence in the initiator protein Orc6 provides an origin-localized replication control switch. *Genes Dev*, 18(9):981–91, 2004.

[189] V. Q. Nguyen, C. Co, and J. J. Li. Cyclin-dependent kinases prevent DNA re-replication through multiple mechanisms. *Nature*, 411(6841):1068–73, 2001.

[190] S. Chen and S. P. Bell. CDK prevents Mcm2-7 helicase loading by inhibiting Cdt1 interaction with Orc6. *Genes Dev*, 25(4):363–72, 2011.

[191] S. Chen, M. A. de Vries, and S. P. Bell. Orc6 is required for dynamic recruitment of Cdt1 during repeated Mcm2-7 loading. *Genes Dev*, 21(22):2897–907, 2007.

[192] V. Q. Nguyen, C. Co, K. Irie, and J. J. Li. Clb/Cdc28 kinases promote nuclear export of the replication initiator proteins Mcm2-7. *Curr Biol*, 10(4):195–205, 2000.

[193] V. Q. Nguyen, C. Co, and J. J. Li. Cyclin-dependent kinases prevent DNA re-replication through multiple mechanisms. *Nature*, 411(6841):1068–73, 2001.

[194] K. Labib, J. F. Diffley, and S. E. Kearsey. G1-phase and B-type cyclins exclude the DNA-replication factor Mcm4 from the nucleus. *Nat Cell Biol*, 1(7):415–22, 1999.

[195] M. E. Liku, V. Q. Nguyen, A. W. Rosales, K. Irie, and J. J. Li. CDK phosphorylation of a novel NLS-NES module distributed between two subunits of the Mcm2-7 complex prevents chromosomal rereplication. *Mol Biol Cell*, 16(10):5026–39, 2005.

[196] G. Oshiro, J. C. Owens, Y. Shellman, R. A. Sclafani, and J. J. Li. Cell cycle control of Cdc7p kinase activity through regulation of Dbf4p stability. *Mol Cell Biol*, 19(7):4888–96, 1999.

[197] M. Weinreich and B. Stillman. Cdc7p-Dbf4p kinase binds to chromatin during S phase and is regulated by both the APC and the RAD53 checkpoint pathway. *EMBO J*, 18(19):5334–46, 1999.

[198] M. F. Ferreira, C. Santocanale, L. S. Drury, and J. F. Diffley. Dbf4p, an essential S phase-promoting factor, is targeted for degradation by the anaphase-promoting complex. *Mol Cell Biol*, 20(1):242–8, 2000.

[199] P. V. Jallepalli, G. W. Brown, M. Muzi-Falconi, D. Tien, and T. J. Kelly. Regulation of the replication initiator protein p65cdc18 by CDK phosphorylation. *Genes Dev*, 11(21):2767–79, 1997.

[200] V. Gopalakrishnan, P. Simancek, C. Houchens, H. A. Snaith, M. G. Frattini, S. Sazer, and T. J. Kelly. Redundant control of rereplication in fission yeast. *Proc Natl Acad Sci U S A*, 98(23):13114–9, 2001.

[201] J. Hu and Y. Xiong. An evolutionarily conserved function of proliferating cell nuclear antigen for Cdt1 degradation by the Cul4-Ddb1 ubiquitin ligase in response to DNA damage. *J Biol Chem*, 281(7):3753–6, 2006.

[202] E. Ralph, E. Boye, and S. E. Kearsey. DNA damage induces Cdt1 proteolysis in fission yeast through a pathway dependent on Cdt2 and Ddb1. *EMBO Rep*, 7(11):1134–9, 2006.

[203] E. Guarino, M. E. Shepherd, I. Salguero, H. Hua, R. S. Deegan, and S. E. Kearsey. Cdt1 proteolysis is promoted by dual PIP degrons and is modulated by PCNA ubiquitylation. *Nucleic Acids Res*, 39(14):5978–90, 2011.

[204] P. Saha, J. Chen, K. C. Thome, S. J. Lawlis, Z. Hou, M. Hendricks, J. D. Parvin, and A. Dutta. Human CDC6/Cdc18 associates with Orc1 and Cyclin-cdk and is selectively eliminated from the nucleus at the onset of S phase. *Mol Cell Biol*, 18(5):2758–2767, 1998.

[205] B. O. Petersen, J. Lukas, C. S. Sorensen, J. Bartek, and K. Helin. Phosphorylation of mammalian CDC6 by cyclin A/CDK2 regulates its subcellular localization. *EMBO J*, 18(2):396–410, 1999.

[206] N. Sugimoto, Y. Tatsumi, T. Tsurumi, A. Matsukage, T. Kiyono, H. Nishitani, and M. Fujita. Cdt1 phosphorylation by cyclin A-dependent kinases negatively regulates its function without affecting geminin binding. *J Biol Chem*, 279(19):19691–7, 2004.

[207] J. Mendez and B. Stillman. Chromatin association of human origin recognition complex, cdc6, and minichromosome maintenance proteins during the cell cycle: assembly of prereplication complexes in late mitosis. *Mol Cell Biol*, 20(22):8602–12, 2000.

[208] A. S. Hemerly, S. G. Prasanth, K. Siddiqui, and B. Stillman. Orc1 controls centriole and centrosome copy number in human cells. *Science*, 323(5915):789–93, 2009.

[209] D. Coverley, C. Pelizon, S. Trewick, and R.A. Laskey. Chromatin-bound Cdc6 persists in S and G2 phases in human cells, while soluble Cdc6 is destroyed in a cyclin A-cdk2 dependent process. *J Cell Sci*, 113(11):1929–1938, 2000.

[210] C. Pelizon, M. A. Madine, P. Romanowski, and R. A. Laskey. Unphosphorylatable mutants of Cdc6 disrupt its nuclear export but still support DNA replication once per cell cycle. *Genes Dev*, 14(19):2526–33, 2000.

[211] L. M. Delmolino, P. Saha, and A. Dutta. Multiple mechanisms regulate subcellular localization of human CDC6. *J Biol Chem*, 276(29):26947–54, 2001.

[212] J. Kim, H. Feng, and E. T. Kipreos. C. elegans CUL-4 prevents rereplication by promoting the nuclear export of CDC-6 via a CKI-1-dependent pathway. *Curr Biol*, 17(11):966–72, 2007.

[213] C. Pelizon. Human replication protein Cdc6 is selectively cleaved by caspase 3 during apoptosis. *EMBO Rep*, 3(8):780–784, 2002.

[214] D. Maiorano, L. Krasinska, M. Lutzmann, and M. Mechali. Recombinant Cdt1 induces rereplication of G2 nuclei in Xenopus egg extracts. *Curr Biol*, 15(2):146–53, 2005.

[215] J. G. Cook. Replication licensing and the dna damage checkpoint. *Front Biosci*, 14(1):5013, 2009.

[216] M. Fujita. Cdt1 revisited: complex and tight regulation during the cell cycle and consequences of deregulation in mammalian cells. *Cell Div*, 1:22, 2006.

[217] X. Li, Q. Zhao, R. Liao, P. Sun, and X. Wu. The SCF(Skp2) ubiquitin ligase complex interacts with the human replication licensing factor Cdt1 and regulates Cdt1 degradation. *J Biol Chem*, 278(33):30854–8, 2003.

[218] E. Liu, X. Li, F. Yan, Q. Zhao, and X. Wu. Cyclin-dependent kinases phosphorylate human Cdt1 and induce its degradation. *J Biol Chem*, 279(17):17283–8, 2004.

[219] H. Nishitani, N. Sugimoto, V. Roukos, Y. Nakanishi, M. Saijo, C. Obuse, T. Tsurimoto, K. I. Nakayama, K. Nakayama, M. Fujita, Z. Lygerou, and T. Nishimoto. Two E3 ubiquitin ligases, SCF-skp2 and DDB1-Cul4, target human Cdt1 for proteolysis. *EMBO J*, 25(5):1126–36, 2006.

[220] T. Kondo, M. Kobayashi, J. Tanaka, A. Yokoyama, S. Suzuki, N. Kato, M. Onozawa, K. Chiba, S. Hashino, M. Imamura, Y. Minami, N. Minamino, and M. Asaka. Rapid degradation of Cdt1 upon UV-induced DNA damage is mediated by SCFSkp2 complex. *J Biol Chem*, 279(26):27315–27319, 2004.

[221] Y. Kim and E. T. Kipreos. The Caenorhabditis elegans replication licensing factor CDT-1 is targeted for degradation by the CUL-4/DDB-1 complex. *Mol Cell Biol*, 27(4):1394–406, 2007.

[222] D. Y. Takeda, J. D. Parvin, and A. Dutta. Degradation of Cdt1 during S phase is Skp2-independent and is required for efficient progression of mammalian cells through S phase. *J Biol Chem*, 280(24):23416–23, 2005.

[223] J. Hu, C. M. McCall, T. Ohta, and Y. Xiong. Targeted ubiquitination of CDT1 by the DDB1-CUL4A-ROC1 ligase in response to DNA damage. *Nat Cell Biol*, 6(10):1003–9, 2004.

[224] E. E. Arias and J. C. Walter. PCNA functions as a molecular platform to trigger Cdt1 destruction and prevent re-replication. *Nat Cell Biol*, 8(1):84–90, 2006.

[225] J. Jin, E. E. Arias, J. Chen, J. W. Harper, and J. C. Walter. A family of diverse Cul4-Ddb1-interacting proteins includes Cdt2, which is required for S phase destruction of the replication factor Cdt1. *Mol Cell*, 23(5):709–21, 2006.

[226] T. Senga, U. Sivaprasad, W. Zhu, J. H. Park, E. E. Arias, J. C. Walter, and A. Dutta. PCNA is a cofactor for Cdt1 degradation by CUL4/DDB1-mediated n-terminal ubiquitination. *J Biol Chem*, 281(10):6246–52, 2006.

[227] C. G. Havens and J. C. Walter. Docking of a specialized PIP box onto chromatin-bound PCNA creates a degron for the ubiquitin ligase CRL4Cdt2. *Mol Cell*, 35(1):93–104, 2009.

[228] A. Li and J. J. Blow. Cdt1 downregulation by proteolysis and geminin inhibition prevents DNA re-replication in Xenopus. *EMBO J*, 24(2):395–404, 2005.

[229] N. Sugimoto, I. Kitabayashi, S. Osano, Y. Tatsumi, T. Yugawa, M. Narisawa-Saito, A. Matsukage, T. Kiyono, and M. Fujita. Identification of novel human Cdt1-binding proteins by a proteomics approach: proteolytic regulation by APC/CCdh1. *Mol Biol Cell*, 19(3):1007–21, 2008.

[230] L. N. Truong and X. Wu. Prevention of DNA re-replication in eukaryotic cells. *J Mol Cell Biol*, 3(1):13–22, 2011.

[231] T. J. McGarry and M. W. Kirschner. Geminin, an inhibitor of DNA replication, is degraded during mitosis. *Cell*, 93(6):1043–1053, 1998.

[232] J. A. Wohlschlegel, B. T. Dwyer, S. K. Dhar, C. Cvetic, J. C. Walter, and A. Dutta. Inhibition of eukaryotic DNA replication by geminin binding to Cdt1. *Science*, 290(5500):2309–12, 2000.

[233] S. Tada, A. Li, D. Maiorano, M. Mechali, and J. J. Blow. Repression of origin assembly in metaphase depends on inhibition of RLF-B/Cdt1 by geminin. *Nat Cell Biol*, 3(2):107–13, 2001.

[234] C. Lee, B. Hong, J. M. Choi, Y. Kim, S. Watanabe, Y. Ishimi, T. Enomoto, S. Tada, Y. Kim, and Y. Cho. Structural basis for inhibition of the replication licensing factor Cdt1 by geminin. *Nature*, 430(7002):913–7, 2004.

[235] D. Maiorano, W. Rul, and M. Mechali. Cell cycle regulation of the licensing activity of cdt1 in xenopus laevis. *Exp Cell Res*, 295(1):138–49, 2004.

[236] K. Yanagi, T. Mizuno, Z. You, and F. Hanaoka. Mouse geminin inhibits not only Cdt1-MCM6 interactions but also a novel intrinsic Cdt1 DNA binding activity. *J Biol Chem*, 277(43):40871–80, 2002.

[237] I. S. Mihaylov, T. Kondo, L. Jones, S. Ryzhikov, J. Tanaka, J. Zheng, L. A. Higa, N. Minamino, L. Cooley, and H. Zhang. Control of DNA replication and chromosome ploidy by Geminin and Cyclin A. *Mol Cell Biol*, 22(6):1868–1880, 2002.

[238] W. Zhu, Y. Chen, and A. Dutta. Rereplication by depletion of geminin is seen regardless of p53 status and activates a G2/M checkpoint. *Mol Cell Biol*, 24:7140–7150, 2004.

[239] S. L. Kerns, S. J. Torke, J. M. Benjamin, and T. J. McGarry. Geminin prevents rereplication during Xenopus development. *J Biol Chem*, 282(8):5514–21, 2007.

[240] M. Melixetian, A. Ballabeni, L. Masiero, P. Gasparini, R. Zamponi, J. Bartek, J. Lukas, and K. Helin. Loss of Geminin induces rereplication in the presence of functional p53. *J Cell Biol*, 165(4):473–82, 2004.

[241] L. M. Quinn, A. Herr, T. J. McGarry, and H. Richardson. The Drosophila Geminin homolog: roles for Geminin in limiting DNA replication, in anaphase and in neurogenesis. *Genes Dev*, 15(20):2741–54, 2001.

[242] T. J. McGarry. Geminin deficiency causes a Chk1-dependent G2 arrest in Xenopus. *Mol Biol Cell*, 13(10):3662–71, 2002.

[243] H. Nishitani, S. Taraviras, Z. Lygerou, and T. Nishimoto. The human licensing factor for DNA replication Cdt1 accumulates in G1 and is destabilized after initiation of S-phase. *J Biol Chem*, 276(48):44905–11, 2001.

[244] M. Lutzmann, D. Maiorano, and M. Mechali. A Cdt1-geminin complex licenses chromatin for DNA replication and prevents rereplication during S phase in Xenopus. *EMBO J*, 25(24):5764–74, 2006.

[245] V. De Marco, P. J. Gillespie, A. Li, N. Karantzelis, E. Christodoulou, R. Klompmaker, S. van Gerwen, A. Fish, M. V. Petoukhov, M. S. Iliou, Z. Lygerou, R. H. Medema, J. J. Blow, D. I. Svergun, S. Taraviras, and A. Perrakis. Quaternary structure of the human Cdt1-Geminin complex regulates DNA replication licensing. *Proc Natl Acad Sci U S A*, 106(47):19807–12, 2009.

[246] C. Lee, B. Hong, J. M. Choi, Y. Kim, S. Watanabe, Y. Ishimi, T. Enomoto, S. Tada, Y. Kim, and Y. Cho. Structural basis for inhibition of the replication licensing factor Cdt1 by geminin. *Nature*, 430(7002):913–7, 2004.

[247] U. Hubscher, G. Maga, and S. Spadari. Eukaryotic DNA polymerases. *Annu Rev Biochem*, 71:133–63, 2002.

[248] P. M. J. Burgers. Eukaryotic DNA polymerases in DNA replication and DNA repair. *Chromosoma*, 107(4):218–227, 1998.

[249] S. L. Forsburg. Eukaryotic MCM proteins: Beyond replication initiation. *Microbiol Mol Biol Rev*, 68(1):109–131, 2004.

[250] I. R. Lehman and L. S. Kaguni. DNA polymerase alpha. *J Biol Chem*, 264(8):4265–8, 1989.

[251] M. S. Wold. Replication protein A: a heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism. *Annu Rev Biochem*, 66:61–92, 1997.

[252] E. Warbrick. The puzzle of PCNA's many partners. *BioEssays*, 22(11):997–1006, 2000.

[253] G. Maga and U. Hubscher. Proliferating cell nuclear antigen (PCNA): a dancer with many partners. *J Cell Sci*, 116(Pt 15):3051–60, 2003.

[254] C. M. Green. One ring to rule them all? Another cellular responsibility for PCNA. *Trends Mol Med*, 12(10):455–8, 2006.

[255] T. A. Kunkel and P. M. Burgers. Dividing the workload at a eukaryotic replication fork. *Trends Cell Biol*, 18(11):521–7, 2008.

[256] P. M. Burgers, E. V. Koonin, E. Bruford, L. Blanco, K. C. Burtis, M. F. Christman, W. C. Copeland, E. C. Friedberg, F. Hanaoka, D. C. Hinkle, C. W. Lawrence, M. Nakanishi, H. Ohmori, L. Prakash, S. Prakash, C. A. Reynaud, A. Sugino, T. Todo, Z. Wang, J. C. Weill, and R. Woodgate. Eukaryotic DNA polymerases: proposal for a revised nomenclature. *J Biol Chem*, 276(47):43487–90, 2001.

[257] R. C. Conaway and I. R. Lehman. A DNA primase activity associated with DNA polymerase alpha from Drosophila melanogaster embryos. *Proc Natl Acad Sci U S A*, 79(8):2523–7, 1982.

[258] R. C. Conaway and I. R. Lehman. Synthesis by the DNA primase of Drosophila melanogaster of a primer with a unique chain length. *Proc Natl Acad Sci U S A*, 79(15):4585–8, 1982.

[259] E. E. Henninger and Z. F. Pursell. DNA polymerase epsilon and its roles in genome stability. *IUBMB Life*, 66(5):339–51, 2014.

[260] T. Tsurimoto and B. Stillman. Replication factors required for SV40 DNA replication in vitro. II. Switching of DNA polymerase alpha and delta during initiation of leading and lagging strand synthesis. *J Biol Chem*, 266(3):1961–8, 1991.

[261] T. Tsurimoto and B. Stillman. Functions of replication factor C and proliferating-cell nuclear antigen: functional similarity of DNA polymerase accessory proteins from human cells and bacteriophage T4. *Proc Natl Acad Sci U S A*, 87(3):1023–1027, 1990.

[262] T. Tsurimoto, T. Melendy, and B. Stillman. Sequential initiation of lagging and leading strand synthesis by two different polymerase complexes at the SV40 DNA replication origin. *Nature*, 346(6284):534–9, 1990.

[263] S. Waga and B. Stillman. Anatomy of a DNA replication fork revealed by reconstitution of SV40 DNA replication in vitro. *Nature*, 369(6477):207–12, 1994.

[264] K. Fien, Y. S. Cho, J. K. Lee, S. Raychaudhuri, I. Tappin, and J. Hurwitz. Primer utilization by DNA polymerase alpha-primase is influenced by its interaction with Mcm10p. *J Biol Chem*, 279(16):16144–53, 2004.

[265] S. L. Sawyer, I. H. Cheng, W. Chai, and B. K. Tye. Mcm10 and Cdc45 cooperate in origin activation in Saccharomyces cerevisiae. *J Mol Biol*, 340(2):doi: 10.1016/j.jmb.2004.04.066, 2004.

[266] K. L. Collins and T. J. Kelly. Effects of T antigen and replication protein A on the initiation of DNA synthesis by DNA polymerase alpha-primase. *Mol Cell Biol*, 11(4):2108–2115, 1991.

[267] T. Melendy and B. Stillman. An interaction between replication protein-a and Sv40 T-antigen appears essential for primosome assembly during Sv40 DNA-replication. *J Biol Chem*, 268(5):3389–3395, 1993.

[268] K. J. Gerik, X. Li, A. Pautz, and P. M. J. Burgers. Characterization of the two small subunits of Saccharomyces cerevisiae DNA polymerase . *J Biol Chem*, 273(31):19747–19755, 1998.

[269] P. M. J. Burgers and K. J. Gerik. Structure and processivity of two forms of Saccharomyces cerevisiae DNA polymerase . *J Biol Chem*, 273(31):19756–19762, 1998.

[270] S. Zuo. Structure and activity associated with multiple forms of Schizosaccharomyces pombe DNA Polymerase delta. *J Biol Chem*, 275(7):5153–5162, 2000.

[271] V. N. Podust, L. S. Chang, R. Ott, G. L. Dianov, and E. Fanning. Reconstitution of human DNA polymerase delta using recombinant baculoviruses: the p12 subunit potentiates DNA polymerizing activity of the four-subunit enzyme. *J Biol Chem*, 277(6):3894–901, 2002.

[272] G. L. Moldovan, B. Pfander, and S. Jentsch. PCNA, the maestro of the replication fork. *Cell*, 129(4):665–79, 2007.

[273] D. J. Mozzherin. Architecture of the active DNA polymerase delta middle dot Proliferating Cell Nuclear Antigen middle dot Template-Primer Complex. *J Biol Chem*, 274(28):19862–19867, 1999.

[274] H. J. Einolf and F. P. Guengerich. Kinetic analysis of nucleotide incorporation by mammalian DNA polymerase delta. *J Biol Chem*, 275(21):16316–22, 2000.

[275] U. Wintersberger and E. Wintersberger. Studies on Deoxyribonucleic Acid polymerases from yeast. 1. Partial purification and properties of two DNA polymerases from mitochondria-free cell extracts. *Eur J Biochem*, 13(1):11–19, 1970.

[276] O. Chilkova, B. H. Jonsson, and E. Johansson. The quaternary structure of DNA polymerase epsilon from Saccharomyces cerevisiae. *J Biol Chem*, 278(16):14082–6, 2003.

[277] H. Pospiech and J. E. Syvaoja. DNA polymerase epsilon - more than a polymerase. *Scientific World J*, 3:87–104, 2003.

[278] T. Kesti, K. Flick, S. Keränen, J. E. Syväoja, and C. Wittenberg. DNA polymerase epsilon catalytic domains are dispensable for DNA replication, DNA repair, and cell viability. *Mol Cell*, 3(5):679–685, 1999.

[279] R. Dua, D. L. Levy, and J. L. Campbell. Analysis of the essential functions of the C-terminal protein/protein interaction domain of Saccharomyces cerevisiae pol epsilon and its unexpected ability to support growth in the absence of the DNA polymerase domain. *J Biol Chem*, 274(32):22283–22288, 1999.

[280] Y. I. Pavlov, P. V. Shcherbakova, and T. A. Kunkel. In vivo consequences of putative active site mutations in yeast DNA polymerases alpha, epsilon, delta, and zeta. *Genetics*, 159(1):47–64, 2001.

[281] W. Feng and G. D'Urso. Schizosaccharomyces pombe cells lacking the amino-terminal catalytic domains of DNA polymerase epsilon are viable but require the DNA damage checkpoint control. *Mol Cell Biol*, 21(14):4495–504, 2001.

[282] T. Ohya, Y. Kawasaki, S. Hiraga, S. Kanbara, K. Nakajo, N. Nakashima, A. Suzuki, and A. Sugino. The DNA polymerase domain of pol epsilon is required for rapid, efficient, and highly accurate chromosomal DNA replication, telomere length maintenance, and normal cell senescence in Saccharomyces cerevisiae. *J Biol Chem*, 277:28099–28108, 2002.

[283] R. Dua, S. Edwards, D. L. Levy, and J. L. Campbell. Subunit interactions within the Saccharomyces cerevisiae DNA polymerase epsilon (pol epsilon) complex. Demonstration of a dimeric pol epsilon. *J Biol Chem*, 275(37):28816–25, 2000.

[284] R. Dua, D. L. Levy, and J. L. Campbell. Role of the putative zinc finger domain of Saccharomyces cerevisiae DNA polymerase in DNA replication and the S/M checkpoint pathway. *J Biol Chem*, 273(45):30046–30055, 1998.

[285] R. Dua, D. L. Levy, and J. L. Campbell. Analysis of the essential functions of the C-terminal protein/protein interaction domain of Saccharomyces cerevisiae pol epsilon and its unexpected ability to support growth in the absence of the DNA polymerase domain. *J Biol Chem*, 274(32):22283–22288, 1999.

[286] T. A. Navas, Z. Zhou, and S. J. Elledge. DNA polymerase epsilon links the DNA replication machinery to the S phase checkpoint. *Cell*, 80(1):29–39, 1995.

[287] D. J. Netz, C. M. Stith, M. Stumpfig, G. Kopf, D. Vogel, H. M. Genau, J. L. Stodola, R. Lill, P. M. Burgers, and A. J. Pierik. Eukaryotic DNA polymerases require an iron-sulfur cluster for the formation of active complexes. *Nat Chem Biol*, 8(1):125–32, 2012.

[288] J. Kraszewska, M. Garbacz, P. Jonczyk, I. J. Fijalkowska, and M. Jaszczur. Defect of Dpb2p, a noncatalytic subunit of DNA polymerase varepsilon, promotes error prone replication of undamaged chromosomal DNA in Saccharomyces cerevisiae. *Mutat Res*, 737(1-2):34–42, 2012.

[289] W. Feng, D. Collingwood, M. E. Boeck, L. A. Fox, G. M. Alvino, W. L. Fangman, M. K. Raghuraman, and B. J. Brewer. Genomic mapping of single-stranded DNA in hydroxyurea-challenged yeasts identifies origins of replication. *Nat Cell Biol*, 8(2):148–55, 2006.

[290] P. M. Burgers. Saccharomyces cerevisiae replication factor C. II. Formation and activity of complexes with the proliferating cell nuclear antigen and with DNA polymerases delta and epsilon. *J Biol Chem*, 266(33):22698–706, 1991.

[291] O. Chilkova, P. Stenlund, I. Isoz, C. M. Stith, P. Grabowski, E. B. Lundstrom, P. M. Burgers, and E. Johansson. The eukaryotic leading and lagging strand DNA polymerases are loaded onto primer-ends via separate mechanisms but have comparable processivity in the presence of PCNA. *Nucleic Acids Res*, 35(19):6588–97, 2007.

[292] G. Chui and S. Linn. Further Characterization of HeLa DNA Polymerase epsilon. *J Biol Chem*, 270(14):7799–7808, 1995.

[293] P. Garg, C. M. Stith, N. Sabouri, E. Johansson, and P. M. Burgers. Idling by DNA polymerase delta maintains a ligatable nick during lagging-strand DNA replication. *Genes Dev*, 18(22):2764–73, 2004.

[294] M. A. Resnick. Similar responses to ionizing radiation of fungal and vertebrate cells and the importance of DNA double-strand breaks. *J Theor Biol*, 71(3):339–346, 1978.

[295] H. I. Kao, J. Veeraraghavan, P. Polaczek, J. L. Campbell, and R. A. Bambara. On the roles of Saccharomyces cerevisiae Dna2p and Flap endonuclease 1 in Okazaki fragment processing. *J Biol Chem*, 279(15):15014–24, 2004.

[296] M. Hogg, P. Osterman, G. O. Bylund, R. A. Ganai, E. B. Lundstrom, A. E. Sauer-Eriksson, and E. Johansson. Structural basis for processive DNA synthesis by yeast DNA polymerase varepsilon. *Nat Struct Mol Biol*, 21(1):49–55, 2014.

[297] D. Shore and A. Bianchi. Telomere length regulation: coupling DNA end processing to feedback regulation of telomerase. *EMBO J*, 28(16):2309–22, 2009.

[298] E. H. Blackburn, E. S. Epel, and J. Lin. Human telomere biology: A contributory and interactive factor in aging, disease risks, and protection. *Science*, 350(6265):1193–8, 2015.

[299] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74, 2011.

[300] S. A. Nick McElhinny, D. A. Gordenin, C. M. Stith, P. M. Burgers, and T. A. Kunkel. Division of labor at the eukaryotic replication fork. *Mol Cell*, 30(2):137–44, 2008.

[301] G. L. Moldovan, B. Pfander, and S. Jentsch. PCNA, the maestro of the replication fork. *Cell*, 129(4):665–79, 2007.

[302] P. V. Shcherbakova and Y. I. Pavlov. 3'->5' exonucleases of DNA polymerases epsilon and delta correct base analog induced DNA replication errors on opposite DNA strands in Saccharomyces cerevisiae. *Genetics*, 142(3):717–726, 1996.

[303] R. Karthikeyan, E. J. Vonarx, A. F. Straffon, M. Simon, G. Faye, and B. A. Kunz. Evidence from mutational specificity studies that yeast DNA polymerases delta and

epsilon replicate different DNA strands at an intracellular replication fork. *J Mol Biol*, 299(2):405–19, 2000.

[304] Y. I. Pavlov, C. S. Newlon, and T. A. Kunkel. Yeast origins establish a strand bias for replicational mutagenesis. *Mol Cell*, 10(1):207–213, 2002.

[305] Z. F. Pursell, I. Isoz, E. B. Lundstrom, E. Johansson, and T. A. Kunkel. Yeast DNA polymerase epsilon participates in leading-strand DNA replication. *Science*, 317(5834):127–30, 2007.

[306] S. J. Diede and D. E. Gottschling. Telomerase-mediated telomere addition in vivo requires DNA primase and DNA polymerases alpha and delta. *Cell*, 99(7):723–733, 1999.

[307] Y. H. Jin, P. Garg, C. M. Stith, H. Al-Refai, J. F. Sterling, L. J. Murray, T. A. Kunkel, M. A. Resnick, P. M. Burgers, and D. A. Gordenin. The multiple biological roles of the 3'–>5' exonuclease of Saccharomyces cerevisiae DNA polymerase delta require switching between the polymerase and exonuclease domains. *Mol Cell Biol*, 25(1):461–71, 2005.

[308] Y. H. Jin, R. Obert, P. M. Burgers, T. A. Kunkel, M. A. Resnick, and D. A. Gordenin. The 3'–>5' exonuclease of DNA polymerase delta can substitute for the 5' flap endonuclease Rad27/Fen1 in processing Okazaki fragments and preventing genome instability. *Proc Natl Acad Sci U S A*, 98(9):5122–7, 2001.

[309] M. E. Huang, B. Le Douarin, C. Henry, and F. Galibert. The Saccharomyces cerevisiae protein YJR043C (Pol32) interacts with the catalytic subunit of DNA polymerase alpha and is required for cell cycle progression in G2/M. *Mol Gen Genet*, 260(6):541–550, 1999.

[310] E. Johansson, P. Garg, and P. M. Burgers. The Pol32 subunit of DNA polymerase delta contains separable domains for processive replication and proliferating cell nuclear antigen (PCNA) binding. *J Biol Chem*, 279(3):1907–15, 2004.

[311] Y. I. Pavlov, C. Frahm, S. A. Nick McElhinny, A. Niimi, M. Suzuki, and T. A. Kunkel. Evidence that errors made by DNA polymerase alpha are corrected by DNA polymerase delta. *Curr Biol*, 16(2):202–7, 2006.

[312] A. Morrison, J. B. Bell, T. A. Kunkel, and A. Sugino. Eukaryotic DNA polymerase amino acid sequence required for 3'—>5' exonuclease activity. *Proc Natl Acad Sci U S A*, 88(21):9473–9477, 1991.

[313] M. Simon, L. Giot, and G. Faye. The 3' to 5' exonuclease activity located in the DNA-polymerase delta-subunit of Saccharomyces cerevisiae is required for accurate replication. *EMBO J*, 10(8):2165–2170, 1991.

[314] S. A. N. McElhinny, C. M. Stith, P. M. Burgers, and T. A. Kunkel. Inefficient proofreading and biased error rates during inaccurate DNA synthesis by a mutant derivative of Saccharomyces cerevisiae DNA polymerase delta. *J Biol Chem*, 282(4):2324–32, 2007.

[315] I. Miyabe, T. A. Kunkel, and A. M. Carr. The major roles of DNA polymerases epsilon and delta at the eukaryotic replication fork are evolutionarily conserved. *PLoS Genet*, 7(12):e1002407, 2011.

[316] T. Fukui, K. Yamauchi, T. Muroya, M. Akiyama, H. Maki, A. Sugino, and S. Waga. Distinct roles of DNA polymerases delta and epsilon at the replication fork in Xenopus egg extracts. *Genes Cells*, 9(3):179–191, 2004.

[317] S. Waga, T. Masuda, H. Takisawa, and A. Sugino. DNA polymerase epsilon is required for coordinated and efficient chromosomal DNA replication in Xenopus egg extracts. *Proc Natl Acad Sci U S A*, 98(9):4978–83, 2001.

[318] J. Fuss and S. Linn. Human DNA polymerase epsilon colocalizes with proliferating cell nuclear antigen and DNA replication late, but not early, in S phase. *J Biol Chem*, 277(10):8658–66, 2002.

[319] A. K. Rytkonen, M. Vaara, T. Nethanel, G. Kaufmann, R. Sormunen, E. Laara, H. P. Nasheuer, A. Rahmeh, M. Y. Lee, J. E. Syvaoja, and H. Pospiech. Distinctive activities of DNA polymerases during human DNA replication. *FEBS J*, 273(13):2984–3001, 2006.

[320] T. Zlotkin, G. Kaufmann, Y. Jiang, M. Y. Lee, L. Uitto, J. Syväoja, I. Dornreiter, E. Fanning, and T. Nethanel. DNA polymerase epsilon may be dispensable for SV40 - but not cellular - DNA replication. *EMBO J*, 15(9):2298–2305, 1996.

[321] F. B. Dean, P. Bullock, Y. Murakami, C. R. Wobbe, L. Weissbach, and J. Hurwitz. Simian virus 40 (SV40) DNA replication: SV40 large T antigen unwinds DNA containing the SV40 origin of replication. *Proc Natl Acad Sci U S A*, 84(1):16–20, 1987.

[322] S. Sengupta, F. van Deursen, G. de Piccoli, and K. Labib. Dpb2 integrates the leading-strand DNA polymerase into the eukaryotic replisome. *Curr Biol*, 23(7):543–52, 2013.

[323] S. S. Hook, J. J. Lin, and A. Dutta. Mechanisms to control rereplication and implications for cancer. *Curr Opin Cell Biol*, 19(6):663–71, 2007.

[324] K. Bebenek and T. A. Kunkel. Functions of DNA polymerases. *Adv Protein Chem*, 69:137–65, 2004.

[325] M. M. Cox, J. Doudna, and M. O'Donnell. *Molecular Biology: Principles and Practices*. WH Freeman and Company, 1st edition edition, Copyright 2012.

[326] W. Yang and R. Woodgate. What a difference a decade makes: insights into translesion DNA synthesis. *Proc Natl Acad Sci U S A*, 104(40):15591–8, 2007.

[327] P. J. Rothwell, V. Mitaksov, and G. Waksman. Motions of the fingers subdomain of klentaq1 are fast and not rate limiting: implications for the molecular basis of fidelity in DNA polymerases. *Mol Cell*, 19(3):345–55, 2005.

[328] A. K. Showalter and M. Tsai. A reexamination of the nucleotide incorporation fidelity of DNA polymerases. *Biochemistry*, 41(34):10571–10576, 2002.

[329] Y. C. Tsai and K. A. Johnson. A new paradigm for DNA polymerase specificity. *Biochemistry*, 45(32):9675–87, 2006.

[330] W. A. Beard, D. D. Shock, B. J. Vande Berg, and S. H. Wilson. Efficiency of correct nucleotide insertion governs DNA polymerase fidelity. *J Biol Chem*, 277(49):47393–8, 2002.

[331] J. W. Drake, B. Charlesworth, D. Charlesworth, and J. F. Crow. Rates of spontaneous mutation. *Genetics*, 148(4):1667–86, 1998.

[332] J. E. Sale. Translesion dna synthesis and mutagenesis in eukaryotes. *Cold Spring Harb Perspect Biol*, 5(3):a012708, 2013.

[333] L. A. Loeb and Jr. Monnat, R. J. DNA polymerases and human disease. *Nat Rev Genet*, 9(8):594–604, 2008.

[334] M. E. Arana and T. A. Kunkel. Mutator phenotypes due to DNA replication infidelity. *Semin Cancer Biol*, 20(5):304–11, 2010.

[335] J. E. Sale, A. R. Lehmann, and R. Woodgate. Y-family DNA polymerases and their role in tolerance of cellular DNA damage. *Nat Rev Mol Cell Biol*, 13(3):141–52, 2012.

[336] S. D. McCulloch and T. A. Kunkel. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res*, 18(1):148–61, 2008.

[337] R. R. Iyer, A. Pluciennik, V. Burdett, and P. L. Modrich. DNA mismatch repair: functions and mechanisms. *Chem Rev*, 106(2):302–23, 2006.

[338] A. Morrison, A. L. Johnson, L. H. Johnston, and A. Sugino. Pathway correcting DNA replication errors in Saccharomyces cerevisiae. *EMBO J*, 12(4):1467–73, 1993.

[339] A. Morrison and A. Sugino. The 3' -> 5' exonucleases of both DNA polymerases delta and epsilon participate in correcting errors of DNA replication in Saccharomyces cerevisiae. *Mol Gen Genet*, 242:289–296, 1994.

[340] A. Bernad, L. Blanco, J. Lázaro, G. Martín, and M. Salas. A conserved 3'->5' exonuclease active site in prokaryotic and eukaryotic DNA polymerases. *Cell*, 59(1):219–228, 1989.

[341] I. V. Shevelev and U. Hubscher. The 3'->5' exonucleases. *Nat Rev Mol Cell Biol*, 3(5):364–76, 2002.

[342] D. L. Ollis, P. Brick, R. Hamlin, N. G. Xuong, and T. A. Steitz. Structure of large fragment of Escherichia coli DNA polymerase I complexed with dTMP. *Nature*, 313(6005):762–766, 1985.

[343] L. Beese, V. Derbyshire, and T. Steitz. Structure of DNA polymerase I Klenow fragment bound to duplex DNA. *Science*, 260(5106):352–355, 1993.

[344] C. A. Brautigam and T. A. Steitz. Structural principles for the inhibition of the 3'-5' exonuclease activity of Escherichia coli DNA polymerase I by phosphorothioates. *J Mol Biol*, 277(2):363–377, 1998.

[345] P. S. Freemont, J. M. Friedman, L. S. Beese, M. R. Sanderson, and T. A. Steitz. Cocrystal structure of an editing complex of Klenow fragment with DNA. *Proc Natl Acad Sci U S A*, 85(23):8924–8, 1988.

[346] L. S. Beese and T. A. Steitz. Structural basis for the 3'-5' exonuclease activity of Escherichia coli DNA polymerase I: a two metal ion mechanism. *EMBO J*, 10:25–33, 1991.

[347] Y. H. Jin, R. Obert, P. M. Burgers, T. A. Kunkel, M. A. Resnick, and D. A. Gordenin. The 3'–>5' exonuclease of DNA polymerase delta can substitute for the 5' flap endonuclease Rad27/Fen1 in processing Okazaki fragments and preventing genome instability. *Proc Natl Acad Sci U S A*, 98(9):5122–7, 2001.

[348] C. M. Joyce and T. A. Steitz. Function and structure relationships in DNA polymerases. *Annu Rev Biochem*, 63:777–822, 1994.

[349] L. S. Beese, V. Derbyshire, and T. A. Steitz. Structure of DNA polymerase I Klenow fragment bound to duplex DNA. *Science*, 260(5106):352–5, 1993.

[350] L. A. Loeb and T. A. Kunkel. Fidelity of DNA synthesis. *Annu Rev Biochem*, 51:429–57, 1982.

[351] C. A. Dumstorf, A. B. Clark, Q. Lin, G. E. Kissling, T. Yuan, R. Kucherlapati, W. G. McGregor, and T. A. Kunkel. Participation of mouse DNA polymerase iota in strand-biased mutagenic bypass of UV photoproducts and suppression of skin cancer. *Proc Natl Acad Sci U S A*, 103(48):18083–8, 2006.

[352] K. Bebenek and T. A. Kunkel. Analyzing fidelity of DNA polymerases. *Methods Enzymol*, 262:217–32, 1995.

[353] M. Hogg, P. Aller, W. Konigsberg, S. S. Wallace, and S. Doublie. Structural and biochemical investigation of the role in proofreading of a beta hairpin loop found in the exonuclease domain of a replicative DNA polymerase of the B family. *J Biol Chem*, 282(2):1432–44, 2007.

[354] L. J. Reha-Krantz, L. A. Marquez, E. Elisseeva, R. P. Baker, L. B. Bloom, H. B. Dunford, and M. F. Goodman. The proofreading pathway of bacteriophage T4 DNA Polymerase. *J Biol Chem*, 273(36):22969–22976, 1998.

[355] S. A. Stocki, R. L. Nonay, and L. J. Reha-Krantz. Dynamics of bacteriophage T4 DNA polymerase function: identification of amino acid residues that affect switching between polymerase and 3' –> 5' exonuclease activities. *J Mol Biol*, 254(1):15–28, 1995.

[356] L. J. Reha-Krantz. Locations of amino acid substitutions in bacteriophage T4 tsL56 DNA polymerase predict an N-terminal exonuclease domain. *J Virol*, 63(11):4762–6, 1989.

[357] P. Wu, N. Nossal, and S. J. Benkovic. Kinetic characterization of a bacteriophage T4 antimutator DNA polymerase. *Biochemistry*, 37(42):14748–55, 1998.

[358] M. K. Swan, R. E. Johnson, L. Prakash, S. Prakash, and A. K. Aggarwal. Structural basis of high-fidelity DNA synthesis by yeast DNA polymerase delta. *Nat Struct Mol Biol*, 16(9):979–86, 2009.

[359] D. A. Korona, K. G. Lecompte, and Z. F. Pursell. The high fidelity and unique error signature of human DNA polymerase epsilon. *Nucleic Acids Res*, 39(5):1763–73, 2011.

[360] P. V. Shcherbakova, Y. I. Pavlov, O. Chilkova, I. B. Rogozin, E. Johansson, and T. A. Kunkel. Unique error signature of the four-subunit yeast DNA polymerase epsilon. *J Biol Chem*, 278(44):43770–80, 2003.

[361] A. Morrison and A. Sugino. The 3' -> 5' exonucleases of both DNA polymerases delta and epsilon participate in correcting errors of DNA replication in Saccharomyces cerevisiae. *Mol Gen Genet*, 242(3):289–296, 1994.

[362] J. M. Fortune, Y. I. Pavlov, C. M. Welch, E. Johansson, P. M. Burgers, and T. A. Kunkel. Saccharomyces cerevisiae DNA polymerase delta: high fidelity for base substitutions but lower fidelity for single- and multi-base deletions. *J Biol Chem*, 280(33):29980–7, 2005.

[363] T. A. Navas, Y. Sanchez, and S. J. Elledge. RAD9 and DNA polymerase epsilon form parallel sensory branches for transducing the DNA damage checkpoint signal in Saccharomyces cerevisiae. *Genes Dev*, 10(20):2632–43, 1996.

[364] A. Datta, J. L. Schmeits, N. S. Amin, P. J. Lau, K. Myung, and R. D. Kolodner. Checkpoint-dependent activation of mutagenic repair in Saccharomyces cerevisiae pol3-01 Mutants. *Mol Cell*, 6(3):593–603, 2000.

[365] M. F. Goodman. Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Annu Rev Biochem*, 71:17–50, 2002.

[366] L. Foulds. The natural history of cancer. *J Chronic Dis*, 8(1):2–37, 1958.

[367] E. Farber and R. Cameron. The sequential analysis of cancer development. 31:125–226, 1980.

[368] R. A. Weinberg. Oncogenes, antioncogenes, and the molecular bases of multistep carcinogenesis. *Cancer Res*, 49(14):3713–21, 1989.

[369] A. P. Eker, C. Quayle, I. Chaves, and G. T. van der Horst. DNA repair in mammalian cells: Direct DNA damage reversal: elegant solutions for nasty problems. *Cell Mol Life Sci*, 66(6):968–80, 2009.

[370] T. Carell, L. T. Burgdorf, L. M. Kundu, and M. Cichon. The mechanism of action of DNA photolyases. *Curr Opin Chem Biol*, 5(5):491–498, 2001.

[371] A. Sancar. Structure and function of DNA photolyase and cryptochrome blue-light photoreceptors. *Chem Rev*, 103(6):2203–37, 2003.

[372] J. I. Lucas-Lledo and M. Lynch. Evolution of mutation rates: phylogenomic analysis of the photolyase/cryptochrome family. *Mol Biol Evol*, 26(5):1143–53, 2009.

[373] Y. Mishina, E. M. Duguid, and C. He. Direct reversal of DNA alkylation damage. *Chem Rev*, 106(2):215–32, 2006.

[374] B. Kaina, M. Christmann, S. Naumann, and W. P. Roos. MGMT: key node in the battle against genotoxicity, carcinogenicity and apoptosis induced by alkylating agents. *DNA Repair (Amst)*, 6(8):1079–99, 2007.

[375] P. J. Abbott and R. Saffhill. DNA synthesis with methylated poly(dC-dG) templates. Evidence for a competitive nature to miscoding by. *Biochim Biophys Acta*, 562(1):51–61, 1979.

[376] I. Teo, B. Sedgwick, M. W. Kilpatrick, T. V. McCarthy, and T. Lindahl. The intracellular signal for induction of resistance to alkylating agents in E. coli. *Cell*, 45(2):315–24, 1986.

[377] B. Sedgwick, P. Robins, N. Totty, and T. Lindahl. Functional domains and methyl acceptor sites of the Escherichia coli ada protein. *J Biol Chem*, 263(9):4430–3, 1988.

[378] B. Demple, B. Sedgwick, P. Robins, N. Totty, M. D. Waterfield, and T. Lindahl. Active site and complete sequence of the suicidal methyltransferase that counters alkylation mutagenesis. *Proc Natl Acad Sci U S A*, 82(9):2688–92, 1985.

[379] D. M. Wilson III and V. A. Bohr. The mechanics of base excision repair, and its relationship to aging and disease. *DNA Repair (Amst)*, 6(4):544–59, 2007.

[380] A. B. Robertson, A. Klungland, T. Rognes, and I. Leiros. DNA repair in mammalian cells: Base excision repair: the long and short of it. *Cell Mol Life Sci*, 66(6):981–93, 2009.

[381] P. M. Girard and S. Boiteux. Repair of oxidized DNA bases in the yeast Saccharomyces cerevisiae. *Biochimie*, 79(9-10):559–566, 1997.

[382] X. Wu. Relationships between yeast Rad27 and Apn1 in response to apurinic/apyrimidinic (AP) sites in DNA. *Nucleic Acids Res*, 27(4):956–962, 1999.

[383] Z. Wang, X. Wu, and E. C. Friedberg. DNA repair synthesis during base excision repair in vitro is catalyzed by DNA polymerase epsilon and is influenced by DNA polymerases alpha and delta in Saccharomyces cerevisiae. *Mol Cell Biol*, 13(2):1051–1058, 1993.

[384] M. R. Kelley, Y. W. Kow, and David M. W. III. Disparity between DNA base excision repair in yeast and mammals. *Cancer Res*, 63(3):549–54, 2003.

[385] Y. Liu and S. H. Wilson. DNA base excision repair: a mechanism of trinucleotide repeat expansion. *Trends Biochem Sci*, 37(4):162–72, 2012.

[386] T. Nouspikel. DNA repair in mammalian cells: Nucleotide excision repair: variations on versatility. *Cell Mol Life Sci*, 66(6):994–1009, 2009.

[387] W. L. de Laat, N. G. Jaspers, and J. H. Hoeijmakers. Molecular mechanism of nucleotide excision repair. *Genes Dev*, 13(7):768–85, 1999.

[388] M. R. Stratton. Exploring the genomes of cancer cells: progress and promise. *Science*, 331(6024):1553–8, 2011.

[389] S. Tornaletti. DNA repair in mammalian cells: Transcription-coupled DNA repair: directing your effort where it's most needed. *Cell Mol Life Sci*, 66(6):1010–20, 2009.

[390] E. M. McNeil and D. W. Melton. DNA repair endonuclease ERCC1-XPF as a novel therapeutic target to overcome chemoresistance in cancer therapy. *Nucleic Acids Res*, 40(20):9990–10004, 2012.

[391] S. N. Guzder, Y. Habraken, P. Sung, L. Prakash, and S. Prakash. Reconstitution of yeast nucleotide excision repair with purified rad proteins, Replication Protein A, and transcription factor TFIIH. *J Biol Chem*, 270(22):12973–12976, 1995.

[392] J. Essers, A. F. Theil, C. Baldeyron, W. A. van Cappellen, A. B. Houtsmuller, R. Kanaar, and W. Vermeulen. Nuclear dynamics of PCNA in DNA replication and repair. *Mol Cell Biol*, 25(21):9350–9, 2005.

[393] X. Wu, E. Braithwaite, and Z. Wang. DNA ligation during excision repair in yeast cell-free extracts is specifically catalyzed by the CDC9 gene product. *Biochemistry*, 38(9):2628–35, 1999.

[394] J. Moser, H. Kool, I. Giakzidis, K. Caldecott, L. H. Mullenders, and M. I. Fousteri. Sealing of chromosomal DNA nicks during nucleotide excision repair requires XRCC1 and DNA ligase III alpha in a cell-cycle-specific manner. *Mol Cell*, 27(2):311–23, 2007.

[395] S. C. Shuck, E. A. Short, and J. J. Turchi. Eukaryotic nucleotide excision repair: from understanding mechanisms to influencing biology. *Cell Res*, 18(1):64–72, 2008.

[396] M. Fousteri and L. H. Mullenders. Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell Res*, 18(1):73–84, 2008.

[397] E. D. Pleasance, R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, I. Varela, M. L. Lin, G. R. Ordonez, G. R. Bignell, K. Ye, J. Alipaz, M. J. Bauer, D. Beare, A. Butler, R. J. Carter, L. Chen, A. J. Cox, S. Edkins, P. I. Kokko-Gonzales, N. A. Gormley, R. J. Grocock, C. D. Haudenschild, M. M. Hims, T. James, M. Jia, Z. Kingsbury, C. Leroy, J. Marshall, A. Menzies, L. J. Mudie, Z. Ning, T. Royce, O. B. Schulz-Trieglaff, A. Spiridou, L. A. Stebbings, L. Szajkowski, J. Teague, D. Williamson, L. Chin, M. T. Ross, P. J. Campbell, D. R. Bentley, P. A. Futreal, and M. R. Stratton. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–6, 2010.

[398] E. D. Pleasance, P. J. Stephens, S. O'Meara, D. J. McBride, A. Meynert, D. Jones, M. L. Lin, D. Beare, K. W. Lau, C. Greenman, I. Varela, S. Nik-Zainal, H. R. Davies, G. R. Ordonez, L. J. Mudie, C. Latimer, S. Edkins, L. Stebbings, L. Chen, M. Jia, C. Leroy, J. Marshall, A. Menzies, A. Butler, J. W. Teague, J. Mangion, Y. A. Sun, S. F. McLaughlin, H. E. Peckham, E. F. Tsung, G. L. Costa, C. C. Lee, J. D. Minna,

A. Gazdar, E. Birney, M. D. Rhodes, K. J. McKernan, M. R. Stratton, P. A. Futreal, and P. J. Campbell. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 463(7278):184–90, 2010.

[399] E. C. Friedberg. How nucleotide excision repair protects against cancer. *Nat Rev Cancer*, 1(1):22–33, 2001.

[400] R. D. Kolodner and G. T. Marsischky. Eukaryotic DNA mismatch repair. *Curr Opin Genetics Dev*, 9(1):89–96, 1999.

[401] T. A. Kunkel and D. A. Erie. DNA mismatch repair. *Annu Rev Biochem*, 74:681–710, 2005.

[402] P. Modrich and R. Lahue. Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Annu Rev Biochem*, 65:101–33, 1996.

[403] J. Jiricny. The multifaceted mismatch-repair system. *Nat Rev Mol Cell Biol*, 7(5):335–46, 2006.

[404] J. Peña-Diaz and J. Jiricny. Mammalian mismatch repair: error-free or error-prone? *Trends Biochem Sci*, 37(5):206–214, 2012.

[405] R. Fishel. Mismatch repair. *J Biol Chem*, 290(44):26395–403, 2015.

[406] R. S. Lahue, K. G. Au, and P. Modrich. DNA mismatch correction in a defined system. *Science*, 245(4914):160–4, 1989.

[407] J. Y. Lee, J. Chang, N. Joseph, R. Ghirlando, D. N. Rao, and W. Yang. MutH complexed with hemi- and unmethylated DNAs: coupling base recognition and DNA cleavage. *Mol Cell*, 20(1):155–66, 2005.

[408] C. Ban and W. Yang. Structural basis for MutH activation in E.coli mismatch repair and relationship of Muth to restriction endonucleases. *EMBO J*, 17(5):1526–34, 1998.

[409] M. S. Junop, G. Obmolova, K. Rausch, P. Hsieh, and W. Yang. Composite active site of an ABC ATPase: MutS uses ATP to verify mismatch recognition and authorize DNA repair. *Mol Cell*, 7(1):1–12, 2001.

[410] G. M. Li. Mechanisms and functions of DNA mismatch repair. *Cell Res*, 18(1):85–98, 2008.

[411] V. Dao and P. Modrich. Mismatch-, Muts-, Mutl-, and helicase II-dependent unwinding from the single-strand break of an incised heteroduplex. *J Biol Chem*, 273(15):9202–7, 1998.

[412] Drummond JT, Li GM, Longley MJ, and Modrich P. Isolation of an hMSH2-p160 heterodimer that restores DNA mismatch repair to tumor cells. *Science*, 268(5219):1909–1912, 1995.

[413] G. M. Li and P. Modrich. Restoration of mismatch repair to nuclear extracts of H6 colorectal tumor cells by a heterodimer of human Mutl homologs. *Proc Natl Acad Sci U S A*, 92(6):1950–4, 1995.

[414] T. A. Prolla, D. M. Christie, and R. M. Liskay. Dual requirement in yeast DNA mismatch repair for MLH1 and PMS1, two homologs of the bacterial mutL gene. *Mol Cell Biol*, 14(1):407–15, 1994.

[415] E. C. Friedberg, G. C. Walker, W. Siede, and R. A. Schultz. *DNA repair and mutagenesis*. ASM Press, Washington, 2nd edn edition, 2006.

[416] R. A. Reenan and R. D. Kolodner. Isolation and characterization of two Saccharomyces cerevisiae genes encoding homologs of the bacterial Hexa and Muts mismatch repair proteins. *Genetics*, 132(4):963–73, 1992.

[417] R. Fishel, M. K. Lescoe, M. R. Rao, N. G. Copeland, N. A. Jenkins, J. Garber, M. Kane, and R. Kolodner. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell*, 75(5):1027–38, 1993.

[418] F. Palombo, P. Gallinari, I. Iaccarino, T. Lettieri, M. Hughes, A. D'Arrigo, O. Truong, J. J. Hsuan, and J. Jiricny. GTBP, a 160-kilodalton protein essential for mismatch-binding activity in human cells. *Science*, 268(5219):1912–4, 1995.

[419] F. S. Leach, N. C. Nicolaides, N. Papadopoulos, B. Liu, J. Jen, R. Parsons, P. Peltomäki, P. Sistonen, L. A. Aaltonen, M. Nyström-Lahti, X. Y. Guan, J. Zhang, P. S. Meltzer, J. Yu, F. Kao, D. J. Chen, K. M. Cerosaletti, R. E. K. Fournier, S. Todd, T. Lewis, R. J. Leach, S. L. Naylor, J. Weissenbach, J. Mecklin, H. Järvinen, G. M. Petersen, S. R. Hamilton, J. Green, J. Jass, P. Watson, H. T. Lynch, J. M. Trent, A. de la Chapelle, K. W. Kinzler, and B. Vogelstein. Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell*, 75(6):1215–1225, 1993.

[420] C. E. Bronner, S. M. Baker, P. T. Morrison, G. Warren, L. G. Smith, M. K. Lescoe, M. Kane, C. Earabino, J. Lipford, A. Lindblom, P. Tannergard, R. J. Bollag, A. R. Godwin, A. C. Ward, M. Nordenskjold, R. Fishel, R. Kolodner, and R. M. Liskay. Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. *Nature*, 368(6468):258–61, 1994.

[421] N. C. Nicolaides, N. Papadopoulos, B. Liu, Y. F. Wei, K. C. Carter, S. M. Ruben, C. A. Rosen, W. A. Haseltine, R. D. Fleischmann, C. M. Fraser, M. D. Adams, J. C. Venter, M. G. Dunlop, S. R. Hamilton, G. M. Petersen, A. de la Chapelle, B. Vogelstein, and K. W. Kinzler. Mutations of two PMS homologues in hereditary nonpolyposis colon cancer. *Nature*, 371(6492):75–80, 1994.

[422] N. Papadopoulos, N. C. Nicolaides, Y. F. Wei, S. M. Ruben, K. C. Carter, C. A. Rosen, W. A. Haseltine, R. D. Fleischmann, C. M. Fraser, M. D. Adams, and et al. Mutation of a mutL homolog in hereditary colon cancer. *Science*, 263(5153):1625–9, 1994.

[423] L. Gu, Y. Hong, S. McCulloch, H. Watanabe, and G. M. Li. ATP-dependent interaction of human mismatch repair proteins and dual role of PCNA in mismatch repair. *Nucleic Acids Res*, 26(5):1173–8, 1998.

[424] A. Umar, A. B. Buermeyer, J. A. Simon, D. C. Thomas, A. B. Clark, R. M. Liskay, and T. A. Kunkel. Requirement for PCNA in DNA mismatch repair at a step preceding DNA resynthesis. *Cell*, 87(1):65–73, 1996.

[425] J. Bowers, P. T. Tran, A. Joshi, R. M. Liskay, and E. Alani. MSH-MLH complexes formed at a DNA mismatch are disrupted by the PCNA sliding clamp. *J Mol Biol*, 306(5):957–68, 2001.

[426] A. B. Clark, F. Valle, K. Drotschmann, R. K. Gary, and T. A. Kunkel. Functional interaction of proliferating cell nuclear antigen with MSH2-MSH6 and MSH2-MSH3 complexes. *J Biol Chem*, 275(47):36498–501, 2000.

[427] H. Flores-Rozas, D. Clark, and R. D. Kolodner. Proliferating cell nuclear antigen and Msh2p-Msh6p interact to form an active mispair recognition complex. *Nat Genet*, 26(3):375–8, 2000.

[428] H. E. Kleczkowska, G. Marra, T. Lettieri, and J. Jiricny. hMSH3 and hMSH6 interact with PCNA and colocalize with it to replication foci. *Genes Dev*, 15(6):724–36, 2001.

[429] P. J. Lau and R. D. Kolodner. Transfer of the MSH2.MSH6 complex from proliferating cell nuclear antigen to mispaired bases in DNA. *J Biol Chem*, 278(1):14–7, 2003.

[430] S. S. Shell, C. D. Putnam, and R. D. Kolodner. The N terminus of Saccharomyces cerevisiae Msh6 is an unstructured tether to PCNA. *Mol Cell*, 26(4):565–78, 2007.

[431] C. Schmutte, R. C. Marinescu, M. M. Sadoff, S. Guerrette, J. Overhauser, and R. Fishel. Human exonuclease I interacts with the mismatch repair protein hMSH2. *Cancer Res*, 58(20):4537–42, 1998.

[432] D. X. Tishkoff, N. S. Amin, C. S. Viars, K. C. Arden, and R. D. Kolodner. Identification of a human gene encoding a homologue of Saccharomyces cerevisiae EXO1, an exonuclease implicated in mismatch repair and recombination. *Cancer Res*, 58(22):5027–31, 1998.

[433] D. X. Tishkoff, A. L. Boerger, P. Bertrand, N. Filosi, G. M. Gaida, M. F. Kane, and R. D. Kolodner. Identification and characterization of Saccharomyces cerevisiae EXO1, a gene encoding an exonuclease that interacts with MSH2. *Proc Natl Acad Sci USA*, 94(14):7487–7492, 1997.

[434] N. S. Amin, M. N. Nguyen, S. Oh, and R. D. Kolodner. Exo1-dependent mutator mutations: model system for studying functional interactions in mismatch repair. *Mol Cell Biol*, 21(15):5142–55, 2001.

[435] F. C. Nielsen, A. C. Jager, A. Lutzen, J. R. Bundgaard, and L. J. Rasmussen. Characterization of human exonuclease 1 in complex with mismatch repair proteins, subcellular localization and association with PCNA. *Oncogene*, 23(7):1457–68, 2004.

[436] P. T. Tran, N. Erdeniz, L. S. Symington, and R. M. Liskay. EXO1-A multi-tasking eukaryotic nuclease. *DNA Repair (Amst)*, 3(12):1549–59, 2004.

[437] Y. Zhang, F. Yuan, S. R. Presnell, K. Tian, Y. Gao, A. E. Tomkinson, L. Gu, and G. M. Li. Reconstitution of 5'-directed human mismatch repair in a purified system. *Cell*, 122(5):693–705, 2005.

[438] J. Genschel and P. Modrich. Mechanism of 5'-directed excision in human mismatch repair. *Mol Cell*, 12(5):1077–1086, 2003.

[439] K. Wei, A. B. Clark, E. Wong, M. F. Kane, D. J. Mazur, T. Parris, N. K. Kolas, R. Russell, Jr. Hou, H., B. Kneitz, G. Yang, T. A. Kunkel, R. D. Kolodner, P. E. Cohen, and

W. Edelmann. Inactivation of Exonuclease 1 in mice results in DNA mismatch repair defects, increased cancer susceptibility, and male and female sterility. *Genes Dev*, 17(5):603–14, 2003. doi: 10.1101/gad.1060603.

[440] R. R. Tice and R. B. Setlow. DNA repair and replication in aging organisms and cells. *Handbook of the biology of aging*, pages 173–224, 1985.

[441] M. M. Vilenchik and A. G. Knudson. Endogenous DNA double-strand breaks: production, fidelity of repair, and induction of cancer. *Proc Natl Acad Sci U S A*, 100(22):12871–6, 2003.

[442] S. J. Boulton. DNA repair: Decision at the break point. *Nature*, 465(7296):301–2, 2010.

[443] D. Pang, S. Yoo, W. S. Dynan, M. Jung, and A. Dritschilo. Ku proteins join DNA fragments as shown by atomic force microscopy. *Cancer Res*, 57(8):1412–5, 1997.

[444] L. Chen, K. Trujillo, W. Ramos, P. Sung, and A. E. Tomkinson. Promotion of Dnl4-catalyzed DNA end-joining by the Rad50/Mre11/Xrs2 and Hdf1/Hdf2 complexes. *Mol Cell*, 8(5):1105–15, 2001.

[445] X. Wu, T. E. Wilson, and M. R. Lieber. A role for FEN-1 in nonhomologous DNA end joining: The order of strand annealing and nucleolytic processing events. *Proc Natl Acad Sci U S A*, 96(4):1303–1308, 1999.

[446] K. Lobachev, E. Vitriol, J. Stemple, M. A. Resnick, and K. Bloom. Chromosome fragmentation after induction of a double-strand break is an active process prevented by the RMX repair complex. *Curr Biol*, 14(23):2107–12, 2004.

[447] S. Moreau, J. R. Ferguson, and L. S. Symington. The nuclease activity of Mre11 is required for meiosis but not for mating type switching, end joining, or telomere maintenance. *Mol Cell Biol*, 19(1):556–566, 1999.

[448] J. San Filippo, P. Sung, and H. Klein. Mechanism of eukaryotic homologous recombination. *Annu Rev Biochem*, 77:229–57, 2008.

[449] S. Moreau, E. A. Morgan, and L. S. Symington. Overlapping functions of the Saccharomyces cerevisiae Mre11, Exo1 and Rad27 nucleases in DNA metabolism. *Genetics*, 159(4):1423–33, 2001.

[450] A. J. Rattray, C. B. McGill, B. K. Shafer, and J. N. Strathern. Fidelity of mitotic double-strand-break repair in Saccharomyces cerevisiae: a role for SAE2/COM1. *Genetics*, 158(1):109–22, 2001.

[451] F. Paques and J. E. Haber. Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev*, 63(2):349–404, 1999.

[452] A. Deem, A. Keszthelyi, T. Blackgrove, A. Vayl, B. Coffey, R. Mathur, A. Chabes, and A. Malkova. Break-induced replication is highly inaccurate. *PLoS Biol*, 9(2):e1000594, 2011.

[453] L. S. Waters, B. K. Minesinger, M. E. Wiltrout, S. D'Souza, R. V. Woodruff, and G. C. Walker. Eukaryotic translesion polymerases and their roles and regulation in DNA damage tolerance. *Microbiol Mol Biol Rev*, 73(1):134–54, 2009.

[454] J. E. Sale. Competition, collaboration and coordination–determining how cells bypass DNA damage. *J Cell Sci*, 125(Pt 7):1633–43, 2012.

[455] S. Sharma, C. M. Helchowski, and C. E. Canman. The roles of DNA polymerase zeta and the Y family DNA polymerases in promoting or preventing genome instability. *Mutat Res*, 743-744:97–110, 2013.

[456] L. C. Colis, P. Raychaudhury, and A. K. Basu. Mutational specificity of gamma-radiation-induced guanine-thymine and thymine-guanine intrastrand cross-links in mammalian cells and translesion synthesis past the guanine-thymine lesion by human DNA polymerase eta. *Biochemistry*, 47(31):8070–9, 2008.

[457] C. Masutani, R. Kusumoto, A. Yamada, N. Dohmae, M. Yokoi, M. Yuasa, M. Araki, S. Iwai, K. Takio, and F. Hanaoka. The XPV (xeroderma pigmentosum variant) gene encodes human DNA polymerase eta. *Nature*, 399(6737):700–4, 1999.

[458] R. E. Johnson. hRAD30 mutations in the variant form of Xeroderma Pigmentosum. *Science*, 285(5425):263–265, 1999.

[459] S. D. McCulloch, R. J. Kokoska, C. Masutani, S. Iwai, F. Hanaoka, and T. A. Kunkel. Preferential cis-syn thymine dimer bypass by DNA polymerase eta occurs with biased fidelity. *Nature*, 428(6978):97–100, 2004.

[460] R. E. Johnson. Fidelity of Human DNA Polymerase eta. *J Biol Chem*, 275(11):7447–7450, 2000.

[461] C. Masutani, R. Kusumoto, S. Iwai, and F. Hanaoka. Mechanisms of accurate translesion synthesis by human DNA polymerase eta. *EMBO J*, 19(12):3100–9, 2000.

[462] A. Vaisman, H. Ling, R. Woodgate, and W. Yang. Fidelity of Dpo4: effect of metal ions, nucleotide selection and pyrophosphorolysis. *EMBO J*, 24(17):2957–67, 2005.

[463] G. N. Gan, J. P. Wittschieben, B. O. Wittschieben, and R. D. Wood. DNA polymerase zeta (pol zeta) in higher eukaryotes. *Cell Res*, 18(1):174–83, 2008.

[464] M. R. Northam, P. Garg, D. M. Baitin, P. M. Burgers, and P. V. Shcherbakova. A novel function of DNA polymerase zeta regulated by PCNA. *EMBO J*, 25(18):4316–25, 2006.

[465] M. R. Northam, E. A. Moore, T. M. Mertz, S. K. Binz, C. M. Stith, E. I. Stepchenkova, K. L. Wendt, P. M. Burgers, and P. V. Shcherbakova. DNA polymerases zeta and Rev1 mediate error-prone bypass of non-B DNA structures. *Nucleic Acids Res*, 42(1):290–306, 2014.

[466] S. Prakash, R. E. Johnson, and L. Prakash. Eukaryotic translesion synthesis dna polymerases: specificity of structure and function. *Annu Rev Biochem*, 74:317–53, 2005.

[467] S. A. Nick McElhinny and D. A. Ramsden. Sibling rivalry: competition between Pol X family members in V(D)J recombination and general double strand break repair. *Immunol Rev*, 200:156–64, 2004.

[468] T. Kawamoto, K. Araki, E. Sonoda, Y. M. Yamashita, K. Harada, K. Kikuchi, C. Masutani, F. Hanaoka, K. Nozaki, N. Hashimoto, and S. Takeda. Dual roles for DNA polymerase eta in homologous DNA recombination and translesion DNA synthesis. *Mol Cell*, 20(5):793–9, 2005.

[469] M. J. McIlwraith, A. Vaisman, Y. Liu, E. Fanning, R. Woodgate, and S. C. West. Human DNA polymerase eta promotes DNA synthesis from strand invasion intermediates of homologous recombination. *Mol Cell*, 20(5):783–92, 2005.

[470] L. Hartwell and T. Weinert. Checkpoints: controls that ensure the order of cell cycle events. *Science*, 246(4930):629–634, 1989.

[471] T. A. Weinert, G. L. Kiser, and L. H. Hartwell. Mitotic checkpoint genes in budding yeast and the dependence of mitosis on DNA replication and repair. *Genes Dev*, 8(6):652–665, 1994.

[472] M. A. Hoyt. A new view of the spindle checkpoint. *J Cell Biol*, 154(5):909–11, 2001.

[473] W. Siede, A. S. Friedberg, I. Dianova, and E. C. Friedberg. Characterization of G1 checkpoint control in the yeast Saccharomyces cerevisiae following exposure to DNA-damaging agents. *Genetics*, 138(2):271–81, 1994.

[474] W. Siede, A. S. Friedberg, and E. C. Friedberg. RAD9-dependent G1 arrest defines a second checkpoint for damaged DNA in the cell cycle of Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, 90(17):7985–9, 1993.

[475] A. G. Paulovich and L. H. Hartwell. A checkpoint regulates the rate of progression through S phase in S. cerevisiae in Response to DNA damage. *Cell*, 82(5):841–847, 1995.

[476] T. Weinert and L. Hartwell. The RAD9 gene controls the cell cycle response to DNA damage in Saccharomyces cerevisiae. *Science*, 241(4863):317–322, 1988.

[477] C. J. Bakkenist and M. B. Kastan. DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. *Nature*, 421(6922):499–506, 2003.

[478] T. Sperka, J. Wang, and K. L. Rudolph. DNA damage checkpoints in stem cells, ageing and cancer. *Nat Rev Mol Cell Biol*, 13(9):579–90, 2012.

[479] I. A. Shaltiel, L. Krenning, W. Bruinsma, and R. H. Medema. The same, only different - DNA damage checkpoints and their reversal throughout the cell cycle. *J Cell Sci*, 128(4):607–20, 2015.

[480] H. Niida and M. Nakanishi. DNA damage checkpoints in mammals. *Mutagenesis*, 21(1):3–9, 2006.

[481] A. L. Gartel and A. L. Tyner. The role of the cyclin-dependent kinase inhibitor p21 in apoptosis. *Mol Cancer Ther*, 1(8):639–49, 2002.

[482] B. B. Zhou and S. J. Elledge. The DNA damage response: putting checkpoints in perspective. *Nature*, 408(6811):433–9, 2000.

[483] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, 5 edition, 2008.

[484] H. Wang, I. Brust-Mascher, and J. M. Scholey. Sliding filaments and mitotic spindle organization. *Nat Cell Biol*, 16(8):737–9, 2014.

[485] V. E. Prince and F. B. Pickett. Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet*, 3(11):827–37, 2002.

[486] R. V. Samonte and E. E. Eichler. Segmental duplications and the evolution of the primate genome. *Nat Rev Genet*, 3(1):65–72, 2002. doi: 10.1038/nrg705.

[487] H. R. Kobel and L. Du Pasquier. Genetics of polyploid Xenopus. *Trends Genet*, 2:310–315, 1986.

[488] P. Dehal and J. L. Boore. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*, 3(10):e314, 2005.

[489] Y. H. Nakanishi, H. Kato, and S. Utsumi. Polytene chromosomes in silk gland cells of the silkworm, Bombyx mori. *Experientia*, 25(4):384–5, 1969.

[490] Y. Suzuki, L. P. Gage, and D. D. Brown. The genes for silk fibroin in Bombyx mori. *J Mol Biol*, 70(3):637–649, 1972.

[491] J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler. Recent segmental duplications in the human genome. *Science*, 297(5583):1003–7, 2002.

[492] M. Long, E. Betran, K. Thornton, and W. Wang. The origin of new genes: glimpses from the young and old. *Nat Rev Genet*, 4(11):865–75, 2003.

[493] M. Lynch and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, 2000.

[494] M. K. Hughes and A. L. Hughes. Evolution of duplicate genes in a tetraploid animal, Xenopus laevis. *Mol Biol Evol*, 10(6):1360–9, 1993.

[495] S. Ohno. *Evolution by Gene Duplication*. Springer, New York, 1970.

[496] M. Kimura and T. Otha. On some principle governing molecular evolution. *Proc. Natl. Acad. Sci.*, 71:2848–2852, 1974.

[497] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge Univ. Press, Cambridge, 1983.

[498] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–45, 1999.

[499] M. Lynch and A. Force. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459–473, 2000.

[500] T. Boveri. *Zellenstudien II: Die Befruchtung und Teilung des Eies von Ascaris megalocephala*, volume 22. Zeit. Naturwiss., Jena, 1888.

[501] T. Boveri. Befruchtung. *Ergeb. Anat. Entwicklungsgesh.*, 1:386–485, 1892.

[502] T. Boveri. *Zellenstudien IV: Die Entwicklung dispermer Seeigeleier. Ein Beitrag zur Befruchtungslehre und zur Theorie des Kernes*, volume 43. Zeit. Naturwiss., Jena, 1907.

[503] T. Boveri. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *J Cell Sci*, 121 Suppl 1:1–84, 2008.

[504] F. Baltzer. Theodor Boveri. *Science*, 144(3620):809–815, 1964.

[505] T. Boveri. *Ergebnisse über die Konstitution der chromatischen Substanz des Zellkerns*. Gustav Fischer, Jena, 1904.

[506] T. Boveri. *The origin of malignant tumors*. Baillière, Tindall and Cox, London, 1929.

[507] A. J. F. Griffiths, J. H. Miller, D. T. Suzuki, R. C. Lewontin, and W. M. Gelbart. *An Introduction to Genetic Analysis*. W. H. Freeman, New York, 7th edition, 2000.

[508] T. Hassold and P. Hunt. To err (meiotically) is human: the genesis of human aneuploidy. *Nat Rev Genet*, 2(4):280–91, 2001.

[509] D. D. Sears, J. H. Hegemann, and P. Hieter. Meiotic recombination and segregation of human-derived artificial chromosomes in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, 89(12):5296–5300, 1992.

[510] D. A. Driscoll and S. Gross. Clinical practice. Prenatal screening for aneuploidy. *N Engl J Med*, 360(24):2556–62, 2009.

[511] Consortium International Wheat Genome Sequencing. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, 345(6194):1251788, 2014.

[512] T. Marcussen, S. R. Sandve, L. Heier, M. Spannagl, M. Pfeifer, Consortium International Wheat Genome Sequencing, K. S. Jakobsen, B. B. Wulff, B. Steuernagel, K. F. Mayer, and O. A. Olsen. Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, 345(6194):1250092, 2014.

[513] R. Riley and V. Chapman. Genetic Control of the Cytologically Diploid Behaviour of Hexaploid Wheat. *Nature*, 182(4637):713–715, 1958.

[514] E. Martinez-Perez, P. Shaw, and G. Moore. The Ph1 locus is needed to ensure specific somatic and meiotic centromere association. *Nature*, 411(6834):204–7, 2001.

[515] E. R. Sears. Cytogenetic studies with polyploid species of wheat. I. Chromosomal aberrations in the progeny of a haploid of Triticum vulgare. *Genetics*, 24(4):509–23, 1939.

[516] E. R. Sears. Nullisomics in Triticum vulgare. *Genetics*, 26:167–168, 1941.

[517] E. R. Sears. Cytogenetic studies with polyploid species of wheat. II. Chromosomal aberrations in the progeny of a haploid of Triticum vulgare. *Genetics*, 29:232–246, 1944.

[518] W. P. Robinson. Mechanisms leading to uniparental disomy and their clinical consequences. *BioEssays*, 22(5):452–459, 2000.

[519] J. Spence, R. Perciaccante, G. Greig, H. Willard, D. Ledbetter, J. Hejtmancik, and M. Pollack. Uniparental disomy as a mechanism of human genetic disease. *Am J Hum Genet*, 42:217–226, 1988.

[520] R. Voss, E. Ben-Simon, A. Avital, S. Godfrey, J. Zlotogora, J Dagan, Y Tikochinski, and J. Hillel. Isodisomy of Chromosome 7 in a patient with Cystic Fibrosis: Could uniparental disomy be common in humans? *Am J Hum Genet*, 45:373–380, 1989.

[521] E. Engel. Uniparental disomies in unselected populations. *Am J Hum Genet*, 63(4):962–6, 1998.

[522] R. D. Nicholls, J. H. Knoll, M. G. Butler, S. Karam, and M. Lalande. Genetic imprinting suggested by maternal heterodisomy in nondeletion Prader-Willi syndrome. *Nature*, 342(6247):281–5, 1989.

[523] J. Peters. The role of genomic imprinting in biology and disease: an expanding view. *Nat Rev Genet*, 15(8):517–30, 2014.

[524] A. Mertzanidou, L. Wilton, J. Cheng, C. Spits, E. Vanneste, Y. Moreau, J. R. Vermeesch, and K. Sermon. Microarray analysis reveals abnormal chromosomal complements in over 70% of 14 normally developing human embryos. *Hum Reprod*, 28(1):256–64, 2013.

[525] Y. B. Yurov, I. Y. Iourov, S. G. Vorsanova, T. Liehr, A. D. Kolotii, S. I. Kutsev, F. Pellestor, A. K. Beresheva, I. A. Demidova, V. S. Kravets, V. V. Monakhov, and I. V. Soloviev. Aneuploidy and confined chromosomal mosaicism in the developing human brain. *PLoS One*, 2(6):e558, 2007.

[526] T. H. Taylor, S. A. Gitlin, J. L. Patrick, J. L. Crain, J. M. Wilson, and D. K. Griffin. The origin, mechanisms, incidence and clinical consequences of chromosomal mosaicism in humans. *Hum Reprod Update*, 20(4):571–81, 2014.

[527] C. Lengauer, K. W. Kinzler, and B. Vogelstein. Genetic instabilities in human cancers. *Nature*, 396(6712):643–9, 1998.

[528] Z. Zhuang, W. S. Park, S. Pack, L. Schmidt, A. O. Vortmeyer, E. Pak, T. Pham, R. J. Weil, S. Candidus, I. A. Lubensky, W. M. Linehan, B. Zbar, and G. Weirich. Trisomy 7-harbouring non-random duplication of the mutant MET allele in hereditary papillary renal carcinomas. *Nat Genet*, 20(1):66–9, 1998.

[529] P. C. Nowell and D. A. Hungerford. A minute chromosome in human chronic granulocytic leukemia. *Science*, 132:1488–1501, 1960.

[530] J. D. Rowley. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. *Nature*, 243(5405):290–3, 1973.

[531] J. D. Rowley. Identification of a Translocation with Quinacrine Fluorescence in a Patient with Acute Leukemia. *Annales De Genetique*, 16(2):109–112, 1973.

[532] A. de Klein, A. G. van Kessel, G. Grosveld, C. R. Bartram, A. Hagemeijer, D. Bootsma, N. K. Spurr, N. Heisterkamp, J. Groffen, and J. R. Stephenson. A cellular oncogene is translocated to the Philadelphia chromosome in chronic myelocytic leukaemia. *Nature*, 300(5894):765–767, 1982.

[533] R. Kurzrock. Philadelphia Chromosome?Positive Leukemias: From Basic Mechanisms to Molecular Therapeutics. *Annals of Internal Medicine*, 138(10):819, 2003.

[534] M. Nambiar and S. C. Raghavan. How does DNA break during chromosomal translocations? *Nucleic Acids Res*, 39(14):5813–25, 2011.

[535] A. J. Holland and D. W. Cleveland. Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. *Nat Med*, 18(11):1630–8, 2012.

[536] M. M. Shen. Chromoplexy: a new category of complex rearrangements in the cancer genome. *Cancer Cell*, 23(5):567–9, 2013.

[537] E. I. McIvor, U. Polak, and M. Napierala. New insights into repeat instability: role of RNA*DNA hybrids. *RNA Biol*, 7(5):551–8, 2010.

[538] K. Kieburtz, M. MacDonald, C. Shih, A. Feigin, K. Steinberg, K. Bordwell, C. Zimmerman, J. Srinidhi, J. Sotack, J. Gusella, and I. Shoulson. Trinucleotide repeat length and progression of illness in Huntington's disease. *J Med Genet*, 31(11):872–4, 1994.

[539] A. Rosenblatt, B. V. Kumar, A. Mo, C. S. Welsh, R. L. Margolis, and C. A. Ross. Age, CAG repeat length, and clinical progression in Huntington's disease. *Mov Disord*, 27(2):272–6, 2012.

[540] B. McClintock. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A*, 36(6):344–355, 1950.

[541] T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, and A. H. Schulman. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*, 8(12):973–82, 2007.

[542] A. Miura, S. Yonebayashi, K. Watanabe, T. Toyama, H. Shimada, and T. Kakutani. Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. *Nature*, 411(6834):212–4, 2001.

[543] Jr. Kazazian, H. H., C. Wong, H. Youssoufian, A. F. Scott, D. G. Phillips, and S. E. Antonarakis. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, 332(6160):164–6, 1988.

[544] Y. Miki, I. Nishisho, A. Horii, Y. Miyoshi, J. Utsunomiya, K. W. Kinzler, B. Vogelstein, and Y. Nakamura. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res*, 52(3):643–5, 1992.

[545] J. V. Moran. Exon Shuffling by L1 Retrotransposition. *Science*, 283(5407):1530–1534, 1999.

[546] W. Gilbert. Why genes in pieces? *Nature*, 271:501, 1978.

[547] A. van Rijk and H. Bloemendal. Molecular mechanisms of exon shuffling: illegitimate recombination. *Genetica*, 118:245–249, 2003.

[548] J. A. Agren. Evolutionary transitions in individuality: insights from transposable elements. *Trends Ecol Evol*, 29(2):90–6, 2014.

[549] J. A. Kolkman and W. P. Stemmer. Directed evolution of proteins by exon shuffling. *Nat Biotechnol*, 19(5):423–8, 2001.

[550] L. Patthy. Modular exchange principles in proteins. *Curr Opin Struct Biol*, 1(3):351–361, 1991.

[551] M. Belfort and P. S. Perlman. Mechanisms of Intron Mobility. *J Biol Chem*, 270(51):30237–30240, 1995.

[552] L. Patthy. Introns and exons. *Curr Opin Struct Biol*, 4(3):383–392, 1994.

[553] L. Patthy. Intron-dependent evolution: Preferred types of exons and introns. *FEBS Lett*, 214(1):1–7, 1987.

[554] I. Chen and D. Dubnau. DNA uptake during bacterial transformation. *Nat Rev Microbiol*, 2(3):241–9, 2004.

[555] J. Lederberg and E. L. Tatum. Gene Recombination in Escherichia Coli. *Nature*, 158(4016):558–558, 1946.

[556] T. V. Matveeva and L. A. Lutova. Horizontal gene transfer from Agrobacterium to plants. *Front Plant Sci*, 5:326, 2014.

[557] T. Kyndt, D. Quispe, H. Zhai, R. Jarret, M. Ghislain, Q. Liu, G. Gheysen, and J. F. Kreuze. The genome of cultivated sweet potato contains Agrobacterium T-DNAs with expressed genes: An example of a naturally transgenic food crop. *Proc Natl Acad Sci U S A*, 112(18):5844–9, 2015.

[558] Y. M. Lo, T. K. Lau, L. Y. Chan, T. N. Leung, and A. M. Chang. Quantitative analysis of the bidirectional fetomaternal transfer of nucleated cells and plasma DNA. *Clin Chem*, 46(9):1301–9, 2000.

[559] D. W. Bianchi, G. K. Zickwolf, G. J. Weil, S. Sylvester, and M. A. DeMaria. Male fetal progenitor cells persist in maternal blood for as long as 27 years postpartum. *Proc Natl Acad Sci U S A*, 93(2):705–708, 1996.

[560] A. Crisp, C. Boschetti, M. Perry, A. Tunnacliffe, and G. Micklem. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biol*, 16:50, 2015.

[561] E. Freese. The difference between spontaneous and base-analogue induced mutations of Phage T4. *Proc Natl Acad Sci U S A*, 45(4):622–633, 1959.

[562] E. Freese. The specific mutagenic effect of base analogues on Phage T4. *J Mol Biol*, 1(2):87–105, 1959.

[563] G. W. Beadle and E. L. Tatum. Genetic control of biochemical reactions in Neurospora. *Proc Natl Acad Sci U S A*, 27(11):499–506, 1941.

[564] F. H. Crick. On protein synthesis. *Symp Soc Exp Biol*, 12:138–63, 1958.

[565] E. Regis. The Forgotten Code Cracker. *Scientific American*, 297(5):50–51, 2007.

[566] *Nobel Lectures, Physiology or Medicine 1963-1970.* Elsevier Publishing Company, Amsterdam, 1972.

[567] P. N. Robinson. The molecular genetics of Marfan syndrome and related microfibrillopathies. *J Med Genet*, 37(1):9–25, 2000.

[568] N. Nakamichi. Adaptation to the local environment by modifications of the photoperiod response in crops. *Plant Cell Physiol*, 56(4):594–604, 2015.

[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*, 310(5750):1031–4, 2005.

[570] A. S. Kondrashov and I. B. Rogozin. Context of deletions and insertions in human coding sequences. *Hum Mutat*, 23(2):177–85, 2004.

[571] A. Sancar, L. A. Lindsey-Boltz, K. Unsal-Kacmaz, and S. Linn. Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annu Rev Biochem*, 73:39–85, 2004.

[572] A. A. Morley and D. R. Turner. The contribution of exogenous and endogenous mutagens to in vivo mutations. *Mutat Res*, 428(1-2):11–5, 1999.

[573] R. De Bont and N. van Larebeke. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis*, 19(3):169–85, 2004.

[574] A. L. Jackson and L. A. Loeb. The contribution of endogenous sources of DNA damage to the multiple mutations in cancer. *Mutat Res*, 477(1-2):7–21, 2001.

[575] C. Bernstein, A. R. Prasad, V. Nfonsam, and H. Bernstein. *DNA Damage, DNA Repair and Cancer.* New Research Directions in DNA Repair. 2013.

[576] H. Ellegren. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet*, 5(6):435–45, 2004.

[577] E. Viguera, D. Canceill, and S. D. Ehrlich. Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J*, 20(10):2587–95, 2001.

[578] T. H. Morgan, A. H. Sturtevant, H. J. Muller, and C. B. Bridges. *The Mechanism of Mendelian heredity*. H. Holt and company, New York, 1915.

[579] H. B. Creighton and B. McClintock. A correlation of cytological and genetical crossing-over in Zea Mays. *Proc Natl Acad Sci U S A*, 17(8):492–497, 1931.

[580] S. I. Nagaoka, T. J. Hassold, and P. A. Hunt. Human aneuploidy: mechanisms and new insights into an age-old problem. *Nat Rev Genet*, 13(7):493–504, 2012.

[581] T. Chiang, R. M. Schultz, and M. A. Lampson. Meiotic origins of maternal age-related aneuploidy. *Biol Reprod*, 86(1):1–7, 2012.

[582] K. Ishiguro and Y. Watanabe. Chromosome cohesion in mitosis and meiosis. *J Cell Sci*, 120(Pt 3):367–9, 2007.

[583] W. D. Gilliland and R. S. Hawley. Cohesin and the maternal age effect. *Cell*, 123(3):371–3, 2005.

[584] T. Lindahl. Instability and decay of the primary structure of DNA. *Nature*, 362(6422):709–15, 1993.

[585] S. Obeid, N. Blatter, R. Kranaster, A. Schnur, K. Diederichs, W. Welte, and A. Marx. Replication through an abasic DNA lesion: structural basis for adenine selectivity. *EMBO J*, 29(10):1738–47, 2010.

[586] L. Haracska, I. Unk, R. E. Johnson, E. Johansson, P. M. Burgers, S. Prakash, and L. Prakash. Roles of yeast DNA polymerases delta and zeta and of Rev1 in the bypass of abasic sites. *Genes Dev*, 15(8):945–54, 2001.

[587] M. S. Cooke, M. D. Evans, M. Dizdaroglu, and J. Lunec. Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB J*, 17(10):1195–214, 2003.

[588] H. J. Helbock, K. B. Beckman, M. K. Shigenaga, P. B. Walter, A. A. Woodall, H. C. Yeo, and B. N. Ames. DNA oxidation matters: the HPLC-electrochemical detection assay of 8-oxo-deoxyguanosine and 8-oxo-guanine. *Proc Natl Acad Sci U S A*, 95(1):288–93, 1998.

[589] H. Wiseman and B. Halliwell. Damage to DNA by reactive oxygen and nitrogen species: role in inflammatory disease and progression to cancer. *Biochem J*, 313 (Pt 1):17–29, 1996.

[590] R. P. Patel, J. McAndrew, H. Sellak, C. R. White, H. Jo, B. A. Freeman, and V. M. Darley-Usmar. Biological aspects of reactive nitrogen species. *Biochem Biophys Acta*, 1411(2-3):385–400, 1999.

[591] M. D. Evans, M. Dizdaroglu, and M. S. Cooke. Oxidative DNA damage and disease: induction, repair and significance. *Mutat Res*, 567(1):1–61, 2004.

[592] M. Hori, T. Suzuki, N. Minakawa, A. Matsuda, H. Harashima, and H. Kamiya. Mutagenicity of secondary oxidation products of 8-oxo-7,8-dihydro-2'-deoxyguanosine 5'-triphosphate (8-hydroxy-2'- deoxyguanosine 5'-triphosphate). *Mutat Res*, 714(1-2):11–6, 2011.

[593] Q. Q. Wang, R. A. Begum, V. W. Day, and K. Bowman-James. Sulfur, oxygen, and nitrogen mustards: stability and reactivity. *Org Biomol Chem*, 10(44):8786–93, 2012.

[594] M. L. Michaels, C. Cruz, A. P. Grollman, and J. H. Miller. Evidence that MutY and MutM combine to prevent mutations by an oxidatively damaged form of guanine in DNA. *Proc Natl Acad Sci USA*, 89(15):7022–7025, 1992.

[595] T. Lindahl and B. Nyberg. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry*, 13(16):3405–10, 1974.

[596] M. Liu and D. G. Schatz. Balancing AID and DNA repair during somatic hypermutation. *Trends Immunol*, 30(4):173–81, 2009.

[597] S. G. Conticello. The AID/APOBEC family of nucleic acid mutators. *Genome Biol*, 9(6):229, 2008.

[598] H. D. Morgan, W. Dean, H. A. Coker, W. Reik, and S. K. Petersen-Mahrt. Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming. *J Biol Chem*, 279(50):52353–60, 2004.

[599] D. Ratel, J. L. Ravanat, F. Berger, and D. Wion. N6-methyladenine: the other methylated base of DNA. *BioEssays*, 28(3):309–15, 2006.

[600] M. Ehrlich, M. A. Gama-Sosa, L. H. Huang, R. M. Midgett, K. C. Kuo, R. A. McCune, and C. Gehrke. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res*, 10(8):2709–21, 1982.

[601] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–22, 2009.

[602] P. A. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*, 13(7):484–92, 2012.

[603] J. C. Shen, W. M. Rideout III, and P. A. Jones. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res*, 22(6):972–6, 1994.

[604] G. P. Pfeifer, S. Kadam, and S. G. Jin. 5-hydroxymethylcytosine and its potential roles in development and cancer. *Epigenetics Chromatin*, 6(1):10, 2013.

[605] J. R. Fernandez, B. Byrne, and B. L. Firestein. Phylogenetic analysis and molecular evolution of guanine deaminases: from guanine to dendrites. *J Mol Evol*, 68(3):227–35, 2009.

[606] V. J. Cogliano, R. Baan, K. Straif, Y. Grosse, B. Lauby-Secretan, F. El Ghissassi, V. Bouvard, L. Benbrahim-Tallaa, N. Guha, C. Freeman, L. Galichet, and C. P. Wild. Preventable exposures associated with human cancers. *J Natl Cancer Inst*, 103(24):1827–39, 2011.

[607] L. G. Hernandez, H. van Steeg, M. Luijten, and J. van Benthem. Mechanisms of non-genotoxic carcinogens and importance of a weight of evidence approach. *Mutat Res*, 682(2-3):94–109, 2009.

[608] J. McCann, E. Choi, E. Yamasaki, and B. N. Ames. Detection of carcinogens as mutagens in the Salmonella/microsome test: assay of 300 chemicals. *Proc Natl Acad Sci U S A*, 72(12):5135–9, 1975.

[609] R. Doll and A. B. Hill. Smoking and carcinoma of the lung; preliminary report. *Br Med J*, 2(4682):739–48, 1950.

[610] H. Witschi. A short history of lung cancer. *Toxicol Sci*, 64(1):4–6, 2001.

[611] E. L. Wynder and E. A. Graham. Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma; a study of 684 proved cases. *J Am Med Assoc*, 143(4):329–36, 1950.

[612] International Agency for Research on Cancer. Tobacco Smoke and Involuntary Smoking. *Tech. Rep. 83, IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, 2004. URL "http://monographs.iarc.fr/ENG/Monographs/vol83/mono83-6C.pdf".

[613] Centers for Disease Control, Prevention (US); National Center for Chronic Disease Prevention, Health Promotion (US); Office on Smoking, and Health (US). How tobacco smoke causes disease: The biology and behavioral basis for smoking-attributable disease: A report of the surgeon general. chapter 5: Cancer. 2010. URL "http://www.ncbi.nlm.nih.gov/books/NBK53010/".

[614] O. C. Ifegwu and C. Anyakora. Polycyclic Aromatic Hydrocarbons: Part I. Exposure. *Adv Clin Chem*, 72:277–304, 2015.

[615] J. Cook and E. L. Kennaway. Chemical compounds as carcinogenic agents: First Supplementary Report: L iterature of 1937. *Cancer Res*, 33:50–97, 1938.

[616] W. Levin, A. W. Wood, H. Yagi, P. M. Dansette, D. M. Jerina, and A. H. Conney. Carcinogenicity of benzo[a]pyrene 4,5-, 7,8-, and 9,10-oxides on mouse skin. *Proc Natl Acad Sci U S A*, 73(1):243–7, 1976.

[617] J. R. Brown and J. L. Thornton. Percivall Pott (1714-1788) and chimney sweepers' cancer of the scrotum. *Br J Ind Med*, 14(1):68–70, 1957.

[618] International Agency for Research on Cancer. A review of human carcinogens: Chemical agents and related occupations. *Tech. Rep. 100F, IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, 2012. URL "http://monographs.iarc.fr/ENG/Monographs/vol100F/mono100F-21.pdf".

[619] H. T. Butlin. Cancer of the scrotum in chimney sweeps and others. II. Why foreign sweeps do not suffer from scrotal cancer. *Br Med J*, 2(1-6), 1892.

[620] M. C. Poirier. Chemical-induced DNA damage and human cancer risk. *Nat Rev Cancer*, 4(8):630–7, 2004.

[621] J. H. Kim, K. H. Stansbury, N. J. Walker, M. A. Trush, P. T. Strickland, and T. R. Sutter. Metabolism of benzo[a]pyrene and benzo[a]pyrene-7,8-diol by human cytochrome P450 1B1. *Carcinogenesis*, 19(10):1847–53, 1998.

[622] D. E. Volk, V. Thiviyanathan, J. S. Rice, B. A. Luxon, J. H. Shah, H. Yagi, J. M. Sayer, H. J. Yeh, D. M. Jerina, and D. G. Gorenstein. Solution structure of a cis-opened (10R)-N6-deoxyadenosine adduct of (9S,10R)-9,10-epoxy-7,8,9,10-tetrahydrobenzo[a]pyrene in a DNA duplex. *Biochemistry*, 42(6):1410–20, 2003.

[623] D. Vesley. *Ionizing and Nonionizing Radiation*, pages 65–74. Springer, 1999.

[624] G. P. Pfeifer, Y. H. You, and A. Besaratinia. Mutations induced by ultraviolet light. *Mutat Res*, 571(1-2):19–31, 2005.

[625] J. F. Ward. *DNA Damage Produced by Ionizing Radiation in Mammalian Cells: Identities, Mechanisms of Formation, and Reparability*. Progress in Nucleic Acid Research and Molecular Biology, Volume 35, ELSEVIER, 1988.

[626] T. K. Kim, T. Kim, T. Y. Kim, W. G. Lee, and J. Yim. Chemotherapeutic dna-damaging drugs activate interferon regulatory factor-7 by the mitogen-activated protein kinase kinase-4-c-jun nh2-terminal kinase pathway. *Cancer Res*, 60(5):1153–6, 2000.

[627] C. Young. Solar ultraviolet radiation and skin cancer. *Occup Med (Lond)*, 59(82-8), 2009.

[628] C. Campbell, A. G. Quinn, B. Angus, P. M. Farr, and J. L. Rees. Wavelength specific patterns of p53 induction in human skin following exposure to UV radiation. *Cancer Res*, 53(12):2697–9, 1993.

[629] F. R. de Gruijl. Skin cancer and solar UV radiation. *Eur J Cancer*, 35(14):2003–9, 1999.

[630] E. M. Witkin. Ultraviolet-induced mutation and DNA repair. *Annu Rev Microbiol*, 23:487–514, 1969.

[631] D. E. Brash. Sunlight and the onset of skin cancer. *Trends Genet*, 13(10):410–4, 1997.

[632] H. Ikehata and T. Ono. The mechanisms of UV mutagenesis. *J Radiat Res*, 52(2):115–25, 2011.

[633] D. M. Parkin, D. Mesher, and P. Sasieni. 13. Cancers attributable to solar (ultraviolet) radiation exposure in the UK in 2010. *Brit J Cancer*, 105:S66–S69, 2011.

[634] S. Gandini, F. Sera, M. S. Cattaruzza, P. Pasquini, O. Picconi, P. Boyle, and C. F. Melchi. Meta-analysis of risk factors for cutaneous melanoma: II. Sun exposure. *Eur J Cancer*, 41(1):45–60, 2005.

[635] F. El Ghissassi, R. Baan, K. Straif, Y. Grosse, B. Secretan, B. Bouvard, L. Benbrahim-Tallaa, N. Guha, C. Freeman, L. Galichet, V Cogliano, and WHO International Agency for Research on Cancer Monograph Working Group. A review of human carcinogens. Part D: radiation. *Lancet Oncol*, 10(8):751–2, 2009.

[636] Sunbeds (Regulation) Act 2010, United Kingdom, 2010.

[637] H. W. Lim, W. D. James, D. S. Rigel, M. E. Maloney, J. M. Spencer, and R. Bhushan. Adverse effects of ultraviolet radiation from the use of indoor tanning equipment: time to ban the tan. *J Am Acad Dermatol*, 64(4):e51–60, 2011.

[638] G. Tweedale. Asbestos and its lethal legacy. *Nat Rev Cancer*, 2(4):311–5, 2002.

[639] G. Tweedale and P. Hansen. Protecting the workers: the medical board and the asbestos industry, 1930s-1960s. *Med Hist*, 42(4):439–57, 1998.

[640] J. E. Alleman and B. T. Mossman. Asbestos Revisited. *Scientific American*, 277(1):70–75, 1997.

[641] G. Liu, P. Cheresh, and D. W. Kamp. Molecular basis of asbestos-induced lung disease. *Annu Rev Pathol*, 8:161–87, 2013.

[642] K. Luus. Asbestos: mining exposure, health effects and policy implications. *Mcgill J Med*, 10(2):121–6, 2007.

[643] P. Boffetta. Epidemiology of environmental and occupational cancer. *Oncogene*, 23(38):6392–403, 2004.

[644] M. L. Newhouse, G. Berry, and J. C. Wagner. Mortality of factory workers in east London 1933-80. *Br J Ind Med*, 42(1):4–11, 1985.

[645] Institute of Medicine (US) Committee on Asbestos: Selected Health Effects. *Asbestos: Selected Cancers*. National Academy of Sciences, 2006.

[646] D. W. Kamp and S. A. Weitzman. The molecular basis of asbestos induced lung injury. *Thorax*, 54(7):638–52, 1999.

[647] F. Drablos, E. Feyzi, P. A. Aas, C. B. Vaagbo, B. Kavli, M. S. Bratlie, J. Pena-Diaz, M. Otterlei, G. Slupphaug, and H. E. Krokan. Alkylation damage in DNA and RNA–repair mechanisms and medical significance. *DNA Repair (Amst)*, 3(11):1389–407, 2004.

[648] R. Goldman and P. G. Shields. Food mutagens. *The Journal of nutrition*, 133(Suppl 3):965S–973S, 2003.

[649] Stephen S. Hecht. DNA adduct formation from tobacco-specific N-nitrosamines. *Mutat Res Fund Mol Mech Mut*, 424(1-2):127–142, 1999.

[650] B. P. Engelward. A chemical and genetic approach together define the biological consequences of 3-methyladenine lesions in the mammalian genome. *J Biol Chem*, 273(9):5412–5418, 1998.

[651] S. Cruet-Hennequart, M. T. Glynn, L. S. Murillo, S. Coyne, and M. P. Carty. Enhanced DNA-PK-mediated RPA2 hyperphosphorylation in DNA polymerase eta-deficient human cells treated with cisplatin and oxaliplatin. *DNA Repair (Amst)*, 7(4):582–96, 2008.

[652] O. S. Platt. Hydroxyurea for the treatment of sickle cell anemia. *N Engl J Med*, 358(13):1362–9, 2008.

[653] L.P. Wakelin. Polyfunctional DNA intercalating agents. *Medicinal research reviews*, 6:275–340, 1986.

[654] J. de Boer and J. H. Hoeijmakers. Nucleotide excision repair and human syndromes. *Carcinogenesis*, 21(3):453–60, 2000.

[655] M. Swift, D. Morrell, E. Cromartie, A. R. Chamberlin, M. H. Skolnick, and D. T. Bishop. The incidence and gene frequency of ataxia-telangiectasia in the United States. *Am J Hum Genet*, 39(5):573–83, 1986.

[656] Y. Shiloh and M. B. Kastan. ATM: Genome stability, neuronal development, and cancer cross paths. 83:209–254, 2001.

[657] K. A. Bernstein, S. Gangloff, and R. Rothstein. The RecQ DNA helicases in DNA repair. *Annu Rev Genet*, 44:393–417, 2010.

[658] D. von Hansemann. Über asymmetrische Zelltheilung in epithel Krebsen und deren biologische Bedeutung. *Virchow's Arch. Path. Anat.*, 119:299, 1890.

[659] R. Schimke, R. Kaufman, F. Alt, and R. Kellems. Gene amplification and drug resistance in cultured murine cells. *Science*, 202(4372):1051–1055, 1978.

[660] W. S. Sutton. The chromosomes in heredity. *Biological Bulletin*, 4:231–251, 1903.

[661] C. M. Croce. Oncogenes and cancer. *N Engl J Med*, 358(5):502–11, 2008.

[662] C. O. Nordling. A new theory on cancer-inducing mechanism. *Br J Cancer*, 7(1):68–72, 1953.

[663] P. Armitage and R. Doll. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer*, 8(1):1–12, 1954.

[664] P. Armitage and R. Doll. A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *Br J Cancer*, 11(2):161–9, 1957.

[665] K. W. Kinzler and B. Vogelstein. Lessons from hereditary colorectal cancer. *Cell*, 87(2):159–170, 1996.

[666] E. R. Fearon and B. Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767, 1990.

[667] P. Rous. A transmissible avian neoplasm. (Sarcoma of the common fowl.). *Journal of Experimental Medicine*, 12(5):696–705, 1910.

[668] R.J. Huebner and G.J. Todaro. Oncogenes of RNA tumor viruses as determinants of cancer. *Proc Natl Acad Sci USA*, 64(3):1087–94, 1969.

[669] P. Rous. Transmission of a malignant new growth by means of a cell-free filtrate. *JAMA-J Am Med Assoc*, 250(11):1445, 1983.

[670] D. Stehelin, H. E. Varmus, J. M. Bishop, and P. K. Vogt. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, 260(5547):170–173, 1976.

[671] D. H. Spector, H. E. Varmus, and J. M. Bishop. Nucleotide sequences related to the transforming gene of avian sarcoma virus are present in DNA of uninfected vertebrates. *Proc Natl Acad Sci U S A*, 75(9):4102–6, 1978.

[672] A. D. Levinson, H. Oppermann, L. Levintow, H. E. Varmus, and J. M. Bishop. Evidence that the transforming gene of avian sarcoma virus encodes a protein kinase associated with a phosphoprotein. *Cell*, 15(2):561–572, 1978.

[673] M. S. Collett and R. L. Erikson. Protein kinase activity associated with the avian sarcoma virus src gene product. *Proc Natl Acad Sci U S A*, 75(4):2021–4, 1978.

[674] W. Eckhart, M. A. Hutchinson, and T. Hunter. An activity phosphorylating tyrosine in polyoma T antigen immunoprecipitates. *Cell*, 18(4):925–933, 1979.

[675] O. N. Witte, A. Dasgupta, and D. Baltimore. Abelson murine leukaemia virus protein is phosphorylated in vitro to form phosphotyrosine. *Nature*, 283(5750):826–831, 1980.

[676] T. Hunter and B. M. Sefton. Transforming gene product of Rous sarcoma virus phosphorylates tyrosine. *Proc Natl Acad Sci U S A*, 77(3):1311–1315, 1980.

[677] H. Ushiro and S. Cohen. Identification of phosphotyrosine as a product of Epidermal Growth Factor-Activated Protein-Kinase in a-431 cell-membranes. *J Biol Chem*, 255(18):8363–8365, 1980.

[678] M. D. Waterfield, G. T. Scrace, N. Whittle, P. Stroobant, A. Johnsson, Å. Wasteson, B. Westermark, C. Heldin, J. S. Huang, and T. F. Deuel. Platelet-derived growth factor is structurally related to the putative transforming protein p28sis of simian sarcoma virus. *Nature*, 304(5921):35–39, 1983.

[679] R. F. Doolittle, M. W. Hunkapiller, L. E. Hood, S. G. Devare, K. C. Robbins, S. A. Aaronson, and H. N. Antoniades. Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science*, 221(4607):275–277, 1983.

[680] J. Downward, Y. Yarden, E. Mayes, G. Scrace, N. Totty, P. Stockwell, A. Ullrich, J. Schlessinger, and M. D. Waterfield. Close similarity of epidermal growth factor receptor and v-erb-B oncogene protein sequences. *Nature*, 307(5951):521–527, 1984.

[681] C. Shih and R. A. Weinberg. Isolation of a transforming sequence from a human bladder carcinoma cell line. *Cell*, 29(1):161–169, 1982.

[682] M. Goldfarb, K. Shimizu, M. Perucho, and M. Wigler. Isolation and preliminary characterization of a human transforming gene from T24 bladder carcinoma cells. *Nature*, 296(5856):404–409, 1982.

[683] S. Pulciani, E. Santos, A. V. Lauver, L. K. Long, K. C. Robbins, and M. Barbacid. Oncogenes in human tumor cell lines: molecular cloning of a transforming gene from human bladder carcinoma cells. *Proc Natl Acad Sci U S A*, 79(9):2845–9, 1982.

[684] L. F. Parada, C. J. Tabin, C. Shih, and R. A. Weinberg. Human EJ bladder carcinoma oncogene is homologue of Harvey sarcoma virus ras gene. *Nature*, 297(5866):474–478, 1982.

[685] E. Santos, S. R. Tronick, S. A. Aaronson, S. Pulciani, and M. Barbacid. T24 human bladder carcinoma oncogene is an activated form of the normal human homologue of BALB- and Harvey-MSV transforming genes. *Nature*, 298(5872):343–347, 1982.

[686] C. J. Der, T. G. Krontiris, and G. M. Cooper. Transforming genes of human bladder and lung carcinoma cell lines are homologous to the ras genes of Harvey and Kirsten sarcoma viruses. *Proc Natl Acad Sci U S A*, 79(11):3637–40, 1982.

[687] C. J. Tabin, S. M. Bradley, C. I. Bargmann, R. A. Weinberg, A. G. Papageorge, E. M. Scolnick, R. Dhar, D. R. Lowy, and E. H. Chang. Mechanism of activation of a human oncogene. *Nature*, 300(5888):143–149, 1982.

[688] E. P. Reddy, R. K. Reynolds, E. Santos, and M. Barbacid. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature*, 300(5888):149–152, 1982.

[689] E. Taparowsky, Y. Suard, O. Fasano, K. Shimizu, M. Goldfarb, and M. Wigler. Activation of the T24 bladder carcinoma transforming gene is linked to a single amino acid change. *Nature*, 300(5894):762–765, 1982.

[690] H. Harris. Cell fusion and the analysis of malignancy. *Proc R Soc Lond B Biol Sci*, 179:1–20, 1971.

[691] D. E. Comings. A general theory of carcinogenesis. *Proc Natl Acad Sci U S A*, 70(12):3324–8, 1973.

[692] W. K. Cavenee, T. P. Dryja, R. A. Phillips, W. F. Benedict, R. Godbout, B. L. Gallie, A. L. Murphree, L. C. Strong, and R. L. White. Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature*, 305(5937):779–784, 1983.

[693] S. Baker, E. Fearon, J. Nigro, Hamilton, A. Preisinger, J. Jessup, P. vanTuinen, D. Ledbetter, D. Barker, Y. Nakamura, R. White, and B. Vogelstein. Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science*, 244(4901):217–221, 1989.

[694] S. H. Friend, R. Bernards, S. Rogelj, R. A. Weinberg, J. M. Rapaport, D. M. Albert, and T. P. Dryja. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature*, 323(6089):643–6, 1986.

[695] W. H. Lee, R. Bookstein, F. Hong, L. J. Young, J. Y. Shew, and E. Y. Lee. Human retinoblastoma susceptibility gene: cloning, identification, and sequence. *Science*, 235(4794):1394–9, 1987.

[696] H. J. Huang, J. K. Yee, J. Y. Shew, P. L. Chen, R. Bookstein, T. Friedmann, E. Y. Lee, and W. H. Lee. Suppression of the neoplastic phenotype by replacement of the RB gene in human cancer cells. *Science*, 242(4885):1563–6, 1988.

[697] C. A. Finlay, P. W. Hinds, and A. J. Levine. The p53 proto-oncogene can act as a suppressor of transformation. *Cell*, 57(7):1083–1093, 1989.

[698] S. Baker, S. Markowitz, E. Fearon, J. Willson, and B. Vogelstein. Suppression of human colorectal carcinoma cell growth by wild-type p53. *Science*, 249(4971):912–915, 1990.

[699] E. K. Yim and J. S. Park. The role of HPV E6 and E7 oncoproteins in HPV-associated cervical carcinogenesis. *Cancer Res Treat*, 37(6):319–24, 2005.

[700] T. Helleday, S. Eshtad, and S. Nik-Zainal. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet*, 15(9):585–98, 2014.

[701] B. D. Howard and I. Tessman. Identification of the altered bases in mutated single-stranded DNA. II. In vivo mutagenesis by 5-bromodeoxyuridine and 2-aminopurine. *J Mol Biol*, 9:364–71, 1964.

[702] R. B. Setlow and W. L. Carrier. Pyrimidine dimers in ultraviolet-irradiated DNA's. *J Mol Biol*, 17(1):237–54, 1966.

[703] D. E. Brash, J. A. Rudolph, J. A. Simon, A. Lin, G. J. McKenna, H. P. Baden, A. J. Halperin, and J. Ponten. A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. *Proc Natl Acad Sci U S A*, 88(22):10124–8, 1991.

[704] M. Ozturk. p53 mutation in hepatocellular carcinoma after aflatoxin exposure. *Lancet*, 338(8779):1356–9, 1991.

[705] B. Bressac, M. Kew, J. Wands, and M. Ozturk. Selective G to T mutations of p53 gene in hepatocellular carcinoma from southern Africa. *Nature*, 350(6317):429–31, 1991.

[706] B. Vogelstein and K. W. Kinzler. Carcinogens leave fingerprints. *Nature*, 355(6357):209–10, 1992.

[707] M. S. Greenblatt, W. P. Bennett, M. Hollstein, and C. C. Harris. Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. *Cancer Res*, 54(18):4855–78, 1994.

[708] M. Hollstein, M. Hergenhahn, Q. Yang, H. Bartsch, Z. Q. Wang, and P. Hainaut. New approaches to understanding p53 gene tumor mutation spectra. *Mutat Res*, 431(2):199–209, 1999.

[709] M. Hollstein, D. Sidransky, B. Vogelstein, and C. Harris. p53 mutations in human cancers. *Science*, 253(5015):49–53, 1991.

[710] P. Stephens, S. Edkins, H. Davies, C. Greenman, C. Cox, C. Hunter, G. Bignell, J. Teague, R. Smith, C. Stevens, S. O'Meara, A. Parker, P. Tarpey, T. Avis, A. Barthorpe, L. Brackenbury, G. Buck, A. Butler, J. Clements, J. Cole, E. Dicks, K. Edwards, S. Forbes, M. Gorton, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, D. Jones, V. Kosmidou, R. Laman, R. Lugg, A. Menzies, J. Perry, R. Petty, K. Raine, R. Shepherd, A. Small, H. Solomon, Y. Stephens, C. Tofts, J. Varian, A. Webb, S. West, S. Widaa, A. Yates, F. Brasseur, C. S. Cooper, A. M. Flanagan, A. Green, M. Knowles, S. Y. Leung, L. H. Looijenga, B. Malkowicz, M. A. Pierotti, B. Teh, S. T. Yuen, A. G. Nicholson, S. Lakhani, D. F. Easton, B. L. Weber, M. R. Stratton, P. A. Futreal, and R. Wooster. A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat Genet*, 37(6):590–2, 2005.

[711] C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y. E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M. H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, and M. R. Stratton. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–8, 2007.

[712] A. F. Rubin and P. Green. Mutation patterns in cancer genomes. *Proc Natl Acad Sci U S A*, 106(51):21766–70, 2009.

[713] Consortium International Cancer Genome, T. J. Hudson, W. Anderson, A. Artez, A. D. Barker, C. Bell, R. R. Bernabe, M. K. Bhan, F. Calvo, I. Eerola, D. S. Gerhard, A. Guttmacher, M. Guyer, F. M. Hemsley, J. L. Jennings, D. Kerr, P. Klatt, P. Kolar, J. Kusada, D. P. Lane, F. Laplace, L. Youyong, G. Nettekoven, B. Ozenberger, J. Peterson, T. S. Rao, J. Remacle, A. J. Schafer, T. Shibata, M. R. Stratton, J. G. Vockley, K. Watanabe, H. Yang, M. M. Yuen, B. M. Knoppers, M. Bobrow, A. Cambon-Thomsen, L. G. Dressler, S. O. Dyke, Y. Joly, K. Kato, K. L. Kennedy, P. Nicolas, M. J. Parker, E. Rial-Sebbag, C. M. Romeo-Casabona, K. M. Shaw, S. Wallace, G. L. Wiesner, N. Zeps, P. Lichter, A. V. Biankin, C. Chabannon, L. Chin, B. Clement, E. de Alava, F. Degos, M. L. Ferguson, P. Geary, D. N. Hayes, T. J. Hudson, A. L. Johns,

A. Kasprzyk, H. Nakagawa, R. Penny, M. A. Piris, R. Sarin, A. Scarpa, T. Shibata, M. van de Vijver, P. A. Futreal, H. Aburatani, M. Bayes, D. D. Botwell, P. J. Campbell, X. Estivill, D. S. Gerhard, S. M. Grimmond, I. Gut, M. Hirst, C. Lopez-Otin, P. Majumder, M. Marra, J. D. McPherson, H. Nakagawa, Z. Ning, X. S. Puente, Y. Ruan, T. Shibata, M. R. Stratton, H. G. Stunnenberg, H. Swerdlow, V. E. Velculescu, R. K. Wilson, H. H. Xue, L. Yang, P. T. Spellman, G. D. Bader, P. C. Boutros, P. J. Campbell, et al. International network of cancer genome projects. *Nature*, 464(7291):993–8, 2010.

[714] N. Agrawal, M. J. Frederick, C. R. Pickering, C. Bettegowda, K. Chang, R. J. Li, C. Fakhry, T. X. Xie, J. Zhang, J. Wang, N. Zhang, A. K. El-Naggar, S. A. Jasser, J. N. Weinstein, L. Trevino, J. A. Drummond, D. M. Muzny, Y. Wu, L. D. Wood, R. H. Hruban, W. H. Westra, W. M. Koch, J. A. Califano, R. A. Gibbs, D. Sidransky, B. Vogelstein, V. E. Velculescu, N. Papadopoulos, D. A. Wheeler, K. W. Kinzler, and J. N. Myers. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science*, 333(6046):1154–7, 2011.

[715] S. C. Baca, D. Prandi, M. S. Lawrence, J. M. Mosquera, A. Romanel, Y. Drier, K. Park, N. Kitabayashi, T. Y. MacDonald, M. Ghandi, E. Van Allen, G. V. Kryukov, A. Sboner, J. P. Theurillat, T. D. Soong, E. Nickerson, D. Auclair, A. Tewari, H. Beltran, R. C. Onofrio, G. Boysen, C. Guiducci, C. E. Barbieri, K. Cibulskis, A. Sivachenko, S. L. Carter, G. Saksena, D. Voet, A. H. Ramos, W. Winckler, M. Cipicchio, K. Ardlie, P. W. Kantoff, M. F. Berger, S. B. Gabriel, T. R. Golub, M. Meyerson, E. S. Lander, O. Elemento, G. Getz, F. Demichelis, M. A. Rubin, and L. A. Garraway. Punctuated evolution of prostate cancer genomes. *Cell*, 153(3):666–77, 2013.

[716] M. F. Berger, M. S. Lawrence, F. Demichelis, Y. Drier, K. Cibulskis, A. Y. Sivachenko, A. Sboner, R. Esgueva, D. Pflueger, C. Sougnez, R. Onofrio, S. L. Carter, K. Park, L. Habegger, L. Ambrogio, T. Fennell, M. Parkin, G. Saksena, D. Voet, A. H. Ramos, T. J. Pugh, J. Wilkinson, S. Fisher, W. Winckler, S. Mahan, K. Ardlie, J. Baldwin, J. W. Simons, N. Kitabayashi, T. Y. MacDonald, P. W. Kantoff, L. Chin, S. B. Gabriel, M. B. Gerstein, T. R. Golub, M. Meyerson, A. Tewari, E. S. Lander, G. Getz, M. A. Rubin, and L. A. Garraway. The genomic complexity of primary human prostate cancer. *Nature*, 470(7333):214–20, 2011.

[717] M. F. Berger, E. Hodis, T. P. Heffernan, Y. L. Deribe, M. S. Lawrence, A. Protopopov, E. Ivanova, I. R. Watson, E. Nickerson, P. Ghosh, H. Zhang, R. Zeid, X. Ren, K. Cibul-

skis, A. Y. Sivachenko, N. Wagle, A. Sucker, C. Sougnez, R. Onofrio, L. Ambrogio, D. Auclair, T. Fennell, S. L. Carter, Y. Drier, P. Stojanov, M. A. Singer, D. Voet, R. Jing, G. Saksena, J. Barretina, A. H. Ramos, T. J. Pugh, N. Stransky, M. Parkin, W. Winckler, S. Mahan, K. Ardlie, J. Baldwin, J. Wargo, D. Schadendorf, M. Meyerson, S. B. Gabriel, T. R. Golub, S. N. Wagner, E. S. Lander, G. Getz, L. Chin, and L. A. Garraway. Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature*, 485(7399):502–6, 2012.

[718] Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–7, 2012.

[719] Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43–9, 2013.

[720] M. A. Chapman, M. S. Lawrence, J. J. Keats, K. Cibulskis, C. Sougnez, A. C. Schinzel, C. L. Harview, J. P. Brunet, G. J. Ahmann, M. Adli, K. C. Anderson, K. G. Ardlie, D. Auclair, A. Baker, P. L. Bergsagel, B. E. Bernstein, Y. Drier, R. Fonseca, S. B. Gabriel, C. C. Hofmeister, S. Jagannath, A. J. Jakubowiak, A. Krishnan, J. Levy, T. Liefeld, S. Lonial, S. Mahan, B. Mfuko, S. Monti, L. M. Perkins, R. Onofrio, T. J. Pugh, S. V. Rajkumar, A. H. Ramos, D. S. Siegel, A. Sivachenko, A. K. Stewart, S. Trudel, R. Vij, D. Voet, W. Winckler, T. Zimmerman, J. Carpten, J. Trent, W. C. Hahn, L. A. Garraway, M. Meyerson, E. S. Lander, G. Getz, and T. R. Golub. Initial genome sequencing and analysis of multiple myeloma. *Nature*, 471(7339):467–72, 2011.

[721] K. De Keersmaecker, Z. K. Atak, N. Li, C. Vicente, S. Patchett, T. Girardi, V. Gianfelici, E. Geerdens, E. Clappier, M. Porcu, I. Lahortiga, R. Luca, J. Yan, G. Hulselmans, H. Vranckx, R. Vandepoel, B. Sweron, K. Jacobs, N. Mentens, I. Wlodarska, B. Cauwelier, J. Cloos, J. Soulier, A. Uyttebroeck, C. Bagni, B. A. Hassan, P. Vandenberghe, A. W. Johnson, S. Aerts, and J. Cools. Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. *Nat Genet*, 45(2):186–90, 2013.

[722] L. Ding, G. Getz, D. A. Wheeler, E. R. Mardis, M. D. McLellan, K. Cibulskis, C. Sougnez, H. Greulich, D. M. Muzny, M. B. Morgan, L. Fulton, R. S. Fulton, Q. Zhang, M. C. Wendl, M. S. Lawrence, D. E. Larson, K. Chen, D. J. Dooling, A. Sabo, A. C. Hawes, H. Shen, S. N. Jhangiani, L. R. Lewis, O. Hall, Y. Zhu, T. Mathew, Y. Ren,

J. Yao, S. E. Scherer, K. Clerc, G. A. Metcalf, B. Ng, A. Milosavljevic, M. L. Gonzalez-Garay, J. R. Osborne, R. Meyer, X. Shi, Y. Tang, D. C. Koboldt, L. Lin, R. Abbott, T. L. Miner, C. Pohl, G. Fewell, C. Haipek, H. Schmidt, B. H. Dunford-Shore, A. Kraja, S. D. Crosby, C. S. Sawyer, T. Vickery, S. Sander, J. Robinson, W. Winckler, J. Baldwin, L. R. Chirieac, A. Dutt, T. Fennell, M. Hanna, B. E. Johnson, R. C. Onofrio, R. K. Thomas, G. Tonon, B. A. Weir, X. Zhao, L. Ziaugra, M. C. Zody, T. Giordano, M. B. Orringer, J. A. Roth, M. R. Spitz, II Wistuba, B. Ozenberger, P. J. Good, A. C. Chang, D. G. Beer, M. A. Watson, M. Ladanyi, S. Broderick, A. Yoshizawa, W. D. Travis, W. Pao, M. A. Province, G. M. Weinstock, H. E. Varmus, S. B. Gabriel, E. S. Lander, R. A. Gibbs, M. Meyerson, and R. K. Wilson. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455(7216):1069–75, 2008.

[723] A. M. Dulak, P. Stojanov, S. Peng, M. S. Lawrence, C. Fox, C. Stewart, S. Bandla, Y. Imamura, S. E. Schumacher, E. Shefler, A. McKenna, S. L. Carter, K. Cibulskis, A. Sivachenko, G. Saksena, D. Voet, A. H. Ramos, D. Auclair, K. Thompson, C. Sougnez, R. C. Onofrio, C. Guiducci, R. Beroukhim, Z. Zhou, L. Lin, J. Lin, R. Reddy, A. Chang, R. Landrenau, A. Pennathur, S. Ogino, J. D. Luketich, T. R. Golub, S. B. Gabriel, E. S. Lander, D. G. Beer, T. E. Godfrey, G. Getz, and A. J. Bass. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat Genet*, 45(5):478–86, 2013.

[724] A. Fujimoto, Y. Totoki, T. Abe, K. A. Boroevich, F. Hosoda, H. H. Nguyen, M. Aoki, N. Hosono, M. Kubo, F. Miya, Y. Arai, H. Takahashi, T. Shirakihara, M. Nagasaki, T. Shibuya, K. Nakano, K. Watanabe-Makino, H. Tanaka, H. Nakamura, J. Kusuda, H. Ojima, K. Shimada, T. Okusaka, M. Ueno, Y. Shigekawa, Y. Kawakami, K. Arihiro, H. Ohdan, K. Gotoh, O. Ishikawa, S. Ariizumi, M. Yamamoto, T. Yamada, K. Chayama, T. Kosuge, H. Yamaue, N. Kamatani, S. Miyano, H. Nakagama, Y. Nakamura, T. Tsunoda, T. Shibata, and H. Nakagawa. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat Genet*, 44(7):760–4, 2012.

[725] C. S. Grasso, Y. M. Wu, D. R. Robinson, X. Cao, S. M. Dhanasekaran, A. P. Khan, M. J. Quist, X. Jing, R. J. Lonigro, J. C. Brenner, I. A. Asangani, B. Ateeq, S. Y. Chun, J. Siddiqui, L. Sam, M. Anstett, R. Mehra, J. R. Prensner, N. Palanisamy, G. A. Ryslik, F. Vandin, B. J. Raphael, L. P. Kunju, D. R. Rhodes, K. J. Pienta, A. M. Chinnaiyan, and S. A. Tomlins. The mutational landscape of lethal castration-resistant prostate cancer. *Nature*, 487(7406):239–43, 2012.

[726] G. Guo, Y. Gui, S. Gao, A. Tang, X. Hu, Y. Huang, W. Jia, Z. Li, M. He, L. Sun, P. Song, X. Sun, X. Zhao, S. Yang, C. Liang, S. Wan, F. Zhou, C. Chen, J. Zhu, X. Li, M. Jian, L. Zhou, R. Ye, P. Huang, J. Chen, T. Jiang, X. Liu, Y. Wang, J. Zou, Z. Jiang, R. Wu, S. Wu, F. Fan, Z. Zhang, L. Liu, R. Yang, X. Liu, H. Wu, W. Yin, X. Zhao, Y. Liu, H. Peng, B. Jiang, Q. Feng, C. Li, J. Xie, J. Lu, K. Kristiansen, Y. Li, X. Zhang, S. Li, J. Wang, H. Yang, Z. Cai, and J. Wang. Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. *Nat Genet*, 44(1):17–9, 2012.

[727] E. Hodis, I. R. Watson, G. V. Kryukov, S. T. Arold, M. Imielinski, J. P. Theurillat, E. Nickerson, D. Auclair, L. Li, C. Place, D. Dicara, A. H. Ramos, M. S. Lawrence, K. Cibulskis, A. Sivachenko, D. Voet, G. Saksena, N. Stransky, R. C. Onofrio, W. Winckler, K. Ardlie, N. Wagle, J. Wargo, K. Chong, D. L. Morton, K. Stemke-Hale, G. Chen, M. Noble, M. Meyerson, J. E. Ladbury, M. A. Davies, J. E. Gershenwald, S. N. Wagner, D. S. Hoon, D. Schadendorf, E. S. Lander, S. B. Gabriel, G. Getz, L. A. Garraway, and L. Chin. A landscape of driver mutations in melanoma. *Cell*, 150(2):251–63, 2012.

[728] L. Holmfeldt, L. Wei, E. Diaz-Flores, M. Walsh, J. Zhang, L. Ding, D. Payne-Turner, M. Churchman, A. Andersson, S. C. Chen, K. McCastlain, J. Becksfort, J. Ma, G. Wu, S. N. Patel, S. L. Heatley, L. A. Phillips, G. Song, J. Easton, M. Parker, X. Chen, M. Rusch, K. Boggs, B. Vadodaria, E. Hedlund, C. Drenberg, S. Baker, D. Pei, C. Cheng, R. Huether, C. Lu, R. S. Fulton, L. L. Fulton, Y. Tabib, D. J. Dooling, K. Ochoa, M. Minden, I. D. Lewis, L. B. To, P. Marlton, A. W. Roberts, G. Raca, W. Stock, G. Neale, H. G. Drexler, R. A. Dickins, D. W. Ellison, S. A. Shurtleff, C. H. Pui, R. C. Ribeiro, M. Devidas, A. J. Carroll, N. A. Heerema, B. Wood, M. J. Borowitz, J. M. Gastier-Foster, S. C. Raimondi, E. R. Mardis, R. K. Wilson, J. R. Downing, S. P. Hunger, M. L. Loh, and C. G. Mullighan. The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nat Genet*, 45(3):242–52, 2013.

[729] F. W. Huang, E. Hodis, M. J. Xu, G. V. Kryukov, L. Chin, and L. A. Garraway. Highly recurrent TERT promoter mutations in human melanoma. *Science*, 339(6122):957–9, 2013.

[730] M. Imielinski, A. H. Berger, P. S. Hammerman, B. Hernandez, T. J. Pugh, E. Hodis, J. Cho, J. Suh, M. Capelletti, A. Sivachenko, C. Sougnez, D. Auclair, M. S. Lawrence, P. Stojanov, K. Cibulskis, K. Choi, L. de Waal, T. Sharifnia, A. Brooks, H. Greulich,

S. Banerji, T. Zander, D. Seidel, F. Leenders, S. Ansen, C. Ludwig, W. Engel-Riedel, E. Stoelben, J. Wolf, C. Goparju, K. Thompson, W. Winckler, D. Kwiatkowski, B. E. Johnson, P. A. Janne, V. A. Miller, W. Pao, W. D. Travis, H. I. Pass, S. B. Gabriel, E. S. Lander, R. K. Thomas, L. A. Garraway, G. Getz, and M. Meyerson. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150(6):1107–20, 2012.

[731] Y. Jiao, C. Shi, B. H. Edil, R. F. de Wilde, D. S. Klimstra, A. Maitra, R. D. Schulick, L. H. Tang, C. L. Wolfgang, M. A. Choti, V. E. Velculescu, Jr. Diaz, L. A., B. Vogelstein, K. W. Kinzler, R. H. Hruban, and N. Papadopoulos. DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. *Science*, 331(6021):1199–203, 2011.

[732] S. Jones, T. L. Wang, M. Shih Ie, T. L. Mao, K. Nakayama, R. Roden, R. Glas, D. Slamon, L. A. Diaz Jr., B. Vogelstein, K. W. Kinzler, V. E. Velculescu, and N. Papadopoulos. Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science*, 330(6001):228–31, 2010.

[733] D. T. Jones, N. Jager, M. Kool, T. Zichner, B. Hutter, M. Sultan, Y. J. Cho, T. J. Pugh, V. Hovestadt, A. M. Stutz, T. Rausch, H. J. Warnatz, M. Ryzhova, S. Bender, D. Sturm, S. Pleier, H. Cin, E. Pfaff, L. Sieber, A. Wittmann, M. Remke, H. Witt, S. Hutter, T. Tzaridis, J. Weischenfeldt, B. Raeder, M. Avci, V. Amstislavskiy, M. Zapatka, U. D. Weber, Q. Wang, B. Lasitschka, C. C. Bartholomae, M. Schmidt, C. von Kalle, V. Ast, C. Lawerenz, J. Eils, R. Kabbe, V. Benes, P. van Sluis, J. Koster, R. Volckmann, D. Shih, M. J. Betts, R. B. Russell, S. Coco, G. P. Tonini, U. Schuller, V. Hans, N. Graf, Y. J. Kim, C. Monoranu, W. Roggendorf, A. Unterberg, C. Herold-Mende, T. Milde, A. E. Kulozik, A. von Deimling, O. Witt, E. Maass, J. Rossler, M. Ebinger, M. U. Schuhmann, M. C. Fruhwald, M. Hasselblatt, N. Jabado, S. Rutkowski, A. O. von Bueren, D. Williamson, S. C. Clifford, M. G. McCabe, V. P. Collins, S. Wolf, S. Wiemann, H. Lehrach, B. Brors, W. Scheurlen, J. Felsberg, G. Reifenberger, P. A. Northcott, M. D. Taylor, M. Meyerson, S. L. Pomeroy, M. L. Yaspo, J. O. Korbel, A. Korshunov, R. Eils, S. M. Pfister, and P. Lichter. Dissecting the genomic complexity underlying medulloblastoma. *Nature*, 488(7409):100–5, 2012.

[734] Z. Kan, H. Zheng, X. Liu, S. Li, T. D. Barber, Z. Gong, H. Gao, K. Hao, M. D. Willard, J. Xu, R. Hauptschein, P. A. Rejto, J. Fernandez, G. Wang, Q. Zhang, B. Wang, R. Chen, J. Wang, N. P. Lee, W. Zhou, Z. Lin, Z. Peng, K. Yi, S. Chen, L. Li, X. Fan, J. Yang,

R. Ye, J. Ju, K. Wang, H. Estrella, S. Deng, P. Wei, M. Qiu, I. H. Wulur, J. Liu, M. E. Ehsani, C. Zhang, A. Loboda, W. K. Sung, A. Aggarwal, R. T. Poon, S. T. Fan, J. Wang, J. Hardwick, C. Reinhard, H. Dai, Y. Li, J. M. Luk, and M. Mao. Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. *Genome Res*, 23(9):1422–33, 2013.

[735] C. Love, Z. Sun, D. Jima, G. Li, J. Zhang, R. Miles, K. L. Richards, C. H. Dunphy, W. W. Choi, G. Srivastava, P. L. Lugar, D. A. Rizzieri, A. S. Lagoo, L. Bernal-Mizrachi, K. P. Mann, C. R. Flowers, K. N. Naresh, A. M. Evens, A. Chadburn, L. I. Gordon, M. B. Czader, J. I. Gill, E. D. Hsi, A. Greenough, A. B. Moffitt, M. McKinney, A. Banerjee, V. Grubor, S. Levy, D. B. Dunson, and S. S. Dave. The genetic landscape of mutations in Burkitt lymphoma. *Nat Genet*, 44(12):1321–5, 2012.

[736] R. D. Morin, M. Mendez-Lago, A. J. Mungall, R. Goya, K. L. Mungall, R. D. Corbett, N. A. Johnson, T. M. Severson, R. Chiu, M. Field, S. Jackman, M. Krzywinski, D. W. Scott, D. L. Trinh, J. Tamura-Wells, S. Li, M. R. Firme, S. Rogic, M. Griffith, S. Chan, O. Yakovenko, I. M. Meyer, E. Y. Zhao, D. Smailus, M. Moksa, S. Chittaranjan, L. Rimsza, A. Brooks-Wilson, J. J. Spinelli, S. Ben-Neriah, B. Meissner, B. Woolcock, M. Boyle, H. McDonald, A. Tam, Y. Zhao, A. Delaney, T. Zeng, K. Tse, Y. Butterfield, I. Birol, R. Holt, J. Schein, D. E. Horsman, R. Moore, S. J. Jones, J. M. Connors, M. Hirst, R. D. Gascoyne, and M. A. Marra. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature*, 476(7360):298–303, 2011.

[737] N. Nagarajan, D. Bertrand, A. M. Hillmer, Z. J. Zang, F. Yao, P. E. Jacques, A. S. Teo, I. Cutcutache, Z. Zhang, W. H. Lee, Y. Y. Sia, S. Gao, P. N. Ariyaratne, A. Ho, X. Y. Woo, L. Veeravali, C. K. Ong, N. Deng, K. V. Desai, C. C. Khor, M. L. Hibberd, A. Shahab, J. Rao, M. Wu, M. Teh, F. Zhu, S. Y. Chin, B. Pang, J. B. So, G. Bourque, R. Soong, W. K. Sung, B. Tean Teh, S. Rozen, X. Ruan, K. G. Yeoh, P. B. Tan, and Y. Ruan. Whole-genome reconstruction and mutational signatures in gastric cancer. *Genome Biol*, 13(12):R115, 2012.

[738] S. Nik-Zainal, L. B. Alexandrov, D. C. Wedge, P. Van Loo, C. D. Greenman, K. Raine, D. Jones, J. Hinton, J. Marshall, L. A. Stebbings, A. Menzies, S. Martin, K. Leung, L. Chen, C. Leroy, M. Ramakrishna, R. Rance, K. W. Lau, L. J. Mudie, I. Varela, D. J. McBride, G. R. Bignell, S. L. Cooke, A. Shlien, J. Gamble, I. Whitmore, M. Maddison, P. S. Tarpey, H. R. Davies, E. Papaemmanuil, P. J. Stephens, S. McLaren, A. P. Butler, J. W. Teague, G. Jonsson, J. E. Garber, D. Silver, P. Miron, A. Fatima, S. Boy-

ault, A. Langerod, A. Tutt, J. W. Martens, S. A. Aparicio, A. Borg, A. V. Salomon, G. Thomas, A. L. Borresen-Dale, A. L. Richardson, M. S. Neuberger, P. A. Futreal, P. J. Campbell, M. R. Stratton, and Consortium Breast Cancer Working Group of the International Cancer Genome. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–93, 2012.

[739] D. W. Parsons, S. Jones, X. Zhang, J. C. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, I. M. Siu, G. L. Gallia, A. Olivi, R. McLendon, B. A. Rasheed, S. Keir, T. Nikolskaya, Y. Nikolsky, D. A. Busam, H. Tekleab, Jr. Diaz, L. A., J. Hartigan, D. R. Smith, R. L. Strausberg, S. K. Marie, S. M. Shinjo, H. Yan, G. J. Riggins, D. D. Bigner, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu, and K. W. Kinzler. An integrated genomic analysis of human glioblastoma multiforme. *Science*, 321(5897):1807–12, 2008.

[740] M. Peifer, L. Fernandez-Cuesta, M. L. Sos, J. George, D. Seidel, L. H. Kasper, D. Plenker, F. Leenders, R. Sun, T. Zander, R. Menon, M. Koker, I. Dahmen, C. Muller, V. Di Cerbo, H. U. Schildhaus, J. Altmuller, I. Baessmann, C. Becker, B. de Wilde, J. Vandesompele, D. Bohm, S. Ansen, F. Gabler, I. Wilkening, S. Heynck, J. M. Heuckmann, X. Lu, S. L. Carter, K. Cibulskis, S. Banerji, G. Getz, K. S. Park, D. Rauh, C. Grutter, M. Fischer, L. Pasqualucci, G. Wright, Z. Wainer, P. Russell, I. Petersen, Y. Chen, E. Stoelben, C. Ludwig, P. Schnabel, H. Hoffmann, T. Muley, M. Brockmann, W. Engel-Riedel, L. A. Muscarella, V. M. Fazio, H. Groen, W. Timens, H. Sietsma, E. Thunnissen, E. Smit, D. A. Heideman, P. J. Snijders, F. Cappuzzo, C. Ligorio, S. Damiani, J. Field, S. Solberg, O. T. Brustugun, M. Lund-Iversen, J. Sanger, J. H. Clement, A. Soltermann, H. Moch, W. Weder, B. Solomon, J. C. Soria, P. Validire, B. Besse, E. Brambilla, C. Brambilla, S. Lantuejoul, P. Lorimier, P. M. Schneider, M. Hallek, W. Pao, M. Meyerson, J. Sage, J. Shendure, R. Schneider, R. Buttner, J. Wolf, P. Nurnberg, S. Perner, L. C. Heukamp, P. K. Brindle, S. Haas, and R. K. Thomas. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nat Genet*, 44(10):1104–10, 2012.

[741] S. Pena-Llopis, S. Vega-Rubin-de Celis, A. Liao, N. Leng, A. Pavia-Jimenez, S. Wang, T. Yamasaki, L. Zhrebker, S. Sivanand, P. Spence, L. Kinch, T. Hambuch, S. Jain, Y. Lotan, V. Margulis, A. I. Sagalowsky, P. B. Summerour, W. Kabbani, S. W. Wong, N. Grishin, M. Laurent, X. J. Xie, C. D. Haudenschild, M. T. Ross, D. R. Bentley, P. Kapur, and J. Brugarolas. BAP1 loss defines a new class of renal cell carcinoma. *Nat Genet*, 44(7):751–9, 2012.

[742] X. S. Puente, M. Pinyol, V. Quesada, L. Conde, G. R. Ordonez, N. Villamor, G. Escaramis, 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. Hernandez, 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, M. Gut, P. Klatt, J. Marshall, K. Raine, L. A. Stebbings, P. A. Futreal, M. R. Stratton, 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 lymphocytic leukaemia. *Nature*, 475(7354):101–5, 2011.

[743] 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, G. R. Crabtree, A. G. Kautzman, S. B. Gabriel, G. Getz, N. Jager, D. T. Jones, P. Lichter, S. M. Pfister, T. M. Roberts, M. Meyerson, S. L. Pomeroy, and Y. J. Cho. Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature*, 488(7409):106–10, 2012.

[744] 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. Sposto, 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 J. M. Maris. The genetic landscape of high-risk neuroblastoma. *Nat Genet*, 45(3):279–84, 2013.

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

T. Baumann, M. Aymerich, M. Rozman, J. Delgado, E. Gine, J. M. Hernandez, M. Gonzalez-Diaz, D. A. Puente, G. Velasco, J. M. Freije, J. M. Tubio, R. Royo, J. L. Gelpi, M. Orozco, D. G. Pisano, J. Zamora, M. Vazquez, A. Valencia, H. Himmelbauer, M. Bayes, S. Heath, M. Gut, I. Gut, X. Estivill, A. Lopez-Guillermo, X. S. Puente, E. Campo, and C. Lopez-Otin. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet*, 44(1):47–52, 2012.

[746] T. Rausch, D. T. Jones, M. Zapatka, A. M. Stutz, T. Zichner, J. Weischenfeldt, N. Jager, M. Remke, D. Shih, P. A. Northcott, E. Pfaff, J. Tica, Q. Wang, L. Massimi, H. Witt, S. Bender, S. Pleier, H. Cin, C. Hawkins, C. Beck, A. von Deimling, V. Hans, B. Brors, R. Eils, W. Scheurlen, J. Blake, V. Benes, A. E. Kulozik, O. Witt, D. Martin, C. Zhang, R. Porat, D. M. Merino, J. Wasserman, N. Jabado, A. Fontebasso, L. Bullinger, F. G. Rucker, K. Dohner, H. Dohner, J. Koster, J. J. Molenaar, R. Versteeg, M. Kool, U. Tabori, D. Malkin, A. Korshunov, M. D. Taylor, P. Lichter, S. M. Pfister, and J. O. Korbel. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell*, 148(1-2):59–71, 2012.

[747] G. Robinson, M. Parker, T. A. Kranenburg, C. Lu, X. Chen, L. Ding, T. N. Phoenix, E. Hedlund, L. Wei, X. Zhu, N. Chalhoub, S. J. Baker, R. Huether, R. Kriwacki, N. Curley, R. Thiruvenkatam, J. Wang, G. Wu, M. Rusch, X. Hong, J. Becksfort, P. Gupta, J. Ma, J. Easton, B. Vadodaria, A. Onar-Thomas, T. Lin, S. Li, S. Pounds, S. Paugh, D. Zhao, D. Kawauchi, M. F. Roussel, D. Finkelstein, D. W. Ellison, C. C. Lau, E. Bouffet, T. Hassall, S. Gururangan, R. Cohn, R. S. Fulton, L. L. Fulton, D. J. Dooling, K. Ochoa, A. Gajjar, E. R. Mardis, R. K. Wilson, J. R. Downing, J. Zhang, and R. J. Gilbertson. Novel mutations target distinct subgroups of medulloblastoma. *Nature*, 488(7409):43–8, 2012.

[748] C. M. Rudin, S. Durinck, E. W. Stawiski, J. T. Poirier, Z. Modrusan, D. S. Shames, E. A. Bergbower, Y. Guan, J. Shin, J. Guillory, C. S. Rivers, C. K. Foo, D. Bhatt, J. Stinson, F. Gnad, P. M. Haverty, R. Gentleman, S. Chaudhuri, V. Janakiraman, B. S. Jaiswal, C. Parikh, W. Yuan, Z. Zhang, H. Koeppen, T. D. Wu, H. M. Stern, R. L. Yauch, K. E. Huffman, D. D. Paskulin, P. B. Illei, M. Varella-Garcia, A. F. Gazdar, F. J. de Sauvage, R. Bourgon, J. D. Minna, M. V. Brock, and S. Seshagiri. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat Genet*, 44(10):1111–6, 2012.

[749] M. Sausen, R. J. Leary, S. Jones, J. Wu, C. P. Reynolds, X. Liu, A. Blackford, G. Parmigiani, Jr. Diaz, L. A., N. Papadopoulos, B. Vogelstein, K. W. Kinzler, V. E. Velculescu, and M. D. Hogarty. Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. *Nat Genet*, 45(1):12–7, 2013.

[750] J. S. Seo, Y. S. Ju, W. C. Lee, J. Y. Shin, J. K. Lee, T. Bleazard, J. Lee, Y. J. Jung, J. O. Kim, J. Y. Shin, S. B. Yu, J. Kim, E. R. Lee, C. H. Kang, I. K. Park, H. Rhee, S. H. Lee, J. I. Kim, J. H. Kang, and Y. T. Kim. The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res*, 22(11):2109–19, 2012.

[751] S. Seshagiri, E. W. Stawiski, S. Durinck, Z. Modrusan, E. E. Storm, C. B. Conboy, S. Chaudhuri, Y. Guan, V. Janakiraman, B. S. Jaiswal, J. Guillory, C. Ha, G. J. Dijkgraaf, J. Stinson, F. Gnad, M. A. Huntley, J. D. Degenhardt, P. M. Haverty, R. Bourgon, W. Wang, H. Koeppen, R. Gentleman, T. K. Starr, Z. Zhang, D. A. Largaespada, T. D. Wu, and F. J. de Sauvage. Recurrent R-spondin fusions in colon cancer. *Nature*, 488(7413):660–4, 2012.

[752] S. P. Shah, A. Roth, R. Goya, A. Oloumi, G. Ha, Y. Zhao, G. Turashvili, J. Ding, K. Tse, G. Haffari, A. Bashashati, L. M. Prentice, J. Khattra, A. Burleigh, D. Yap, V. Bernard, A. McPherson, K. Shumansky, A. Crisan, R. Giuliany, A. Heravi-Moussavi, J. Rosner, D. Lai, I. Birol, R. Varhol, A. Tam, N. Dhalla, T. Zeng, K. Ma, S. K. Chan, M. Griffith, A. Moradian, S. W. Cheng, G. B. Morin, P. Watson, K. Gelmon, S. Chia, S. F. Chin, C. Curtis, O. M. Rueda, P. D. Pharoah, S. Damaraju, J. Mackey, K. Hoon, T. Harkins, V. Tadigotla, M. Sigaroudinia, P. Gascard, T. Tlsty, J. F. Costello, I. M. Meyer, C. J. Eaves, W. W. Wasserman, S. Jones, D. Huntsman, M. Hirst, C. Caldas, M. A. Marra, and S. Aparicio. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403):395–9, 2012.

[753] M. S. Stark, S. L. Woods, M. G. Gartside, V. F. Bonazzi, K. Dutton-Regester, L. G. Aoude, D. Chow, C. Sereduk, N. M. Niemi, N. Tang, J. J. Ellis, J. Reid, V. Zismann, S. Tyagi, D. Muzny, I. Newsham, Y. Wu, J. M. Palmer, T. Pollak, D. Youngkin, B. R. Brooks, C. Lanagan, C. W. Schmidt, B. Kobe, J. P. MacKeigan, H. Yin, K. M. Brown, R. Gibbs, J. Trent, and N. K. Hayward. Frequent somatic mutations in MAP3K5 and MAP3K9 in metastatic melanoma identified by exome sequencing. *Nat Genet*, 44(2):165–9, 2012.

[754] P. J. Stephens, P. S. Tarpey, H. Davies, P. Van Loo, C. Greenman, D. C. Wedge, S. Nik-Zainal, S. Martin, I. Varela, G. R. Bignell, L. R. Yates, E. Papaemmanuil, D. Beare,

A. Butler, A. Cheverton, J. Gamble, J. Hinton, M. Jia, A. Jayakumar, D. Jones, C. Latimer, K. W. Lau, S. McLaren, D. J. McBride, A. Menzies, L. Mudie, K. Raine, R. Rad, M. S. Chapman, J. Teague, D. Easton, A. Langerod, Consortium Oslo Breast Cancer, M. T. Lee, C. Y. Shen, B. T. Tee, B. W. Huimin, A. Broeks, A. C. Vargas, G. Turashvili, J. Martens, A. Fatima, P. Miron, S. F. Chin, G. Thomas, S. Boyault, O. Mariani, S. R. Lakhani, M. van de Vijver, L. van 't Veer, J. Foekens, C. Desmedt, C. Sotiriou, A. Tutt, C. Caldas, J. S. Reis-Filho, S. A. Aparicio, A. V. Salomon, A. L. Borresen-Dale, A. L. Richardson, P. J. Campbell, P. A. Futreal, and M. R. Stratton. The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486(7403):400–4, 2012.

[755] N. Stransky, A. M. Egloff, A. D. Tward, A. D. Kostic, K. Cibulskis, A. Sivachenko, G. V. Kryukov, M. S. Lawrence, C. Sougnez, A. McKenna, E. Shefler, A. H. Ramos, P. Stojanov, S. L. Carter, D. Voet, M. L. Cortes, D. Auclair, M. F. Berger, G. Saksena, C. Guiducci, R. C. Onofrio, M. Parkin, M. Romkes, J. L. Weissfeld, R. R. Seethala, L. Wang, C. Rangel-Escareno, J. C. Fernandez-Lopez, A. Hidalgo-Miranda, J. Melendez-Zajgla, W. Winckler, K. Ardlie, S. B. Gabriel, M. Meyerson, E. S. Lander, G. Getz, T. R. Golub, L. A. Garraway, and J. R. Grandis. The mutational landscape of head and neck squamous cell carcinoma. *Science*, 333(6046):1157–60, 2011.

[756] K. Wang, J. Kan, S. T. Yuen, S. T. Shi, K. M. Chu, S. Law, T. L. Chan, Z. Kan, A. S. Chan, W. Y. Tsui, S. P. Lee, S. L. Ho, A. K. Chan, G. H. Cheng, P. C. Roberts, P. A. Rejto, N. W. Gibson, D. J. Pocalyko, M. Mao, J. Xu, and S. Y. Leung. Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nat Genet*, 43(12):1219–23, 2011.

[757] J. Wu, Y. Jiao, M. Dal Molin, A. Maitra, R. F. de Wilde, L. D. Wood, J. R. Eshleman, M. G. Goggins, C. L. Wolfgang, M. I. Canto, R. D. Schulick, B. H. Edil, M. A. Choti, V. Adsay, D. S. Klimstra, G. J. Offerhaus, A. P. Klein, L. Kopelovich, H. Carter, R. Karchin, P. J. Allen, C. M. Schmidt, Y. Naito, Jr. Diaz, L. A., K. W. Kinzler, N. Papadopoulos, R. H. Hruban, and B. Vogelstein. Whole-exome sequencing of neoplastic cysts of the pancreas reveals recurrent mutations in components of ubiquitin-dependent pathways. *Proc Natl Acad Sci U S A*, 108(52):21188–93, 2011.

[758] Z. J. Zang, I. Cutcutache, S. L. Poon, S. L. Zhang, J. R. McPherson, J. Tao, V. Rajasegaran, H. L. Heng, N. Deng, A. Gan, K. H. Lim, C. K. Ong, D. Huang, S. Y. Chin, I. B. Tan, C. C. Ng, W. Yu, Y. Wu, M. Lee, J. Wu, D. Poh, W. K. Wan, S. Y. Rha, J. So, M. Salto-Tellez, K. G. Yeoh, W. K. Wong, Y. J. Zhu, P. A. Futreal, B. Pang,

Y. Ruan, A. M. Hillmer, D. Bertrand, N. Nagarajan, S. Rozen, B. T. Teh, and P. Tan. Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. *Nat Genet*, 44(5):570–4, 2012.

[759] J. Zhang, L. Ding, L. Holmfeldt, G. Wu, S. L. Heatley, D. Payne-Turner, J. Easton, X. Chen, J. Wang, M. Rusch, C. Lu, S. C. Chen, L. Wei, J. R. Collins-Underwood, J. Ma, K. G. Roberts, S. B. Pounds, A. Ulyanov, J. Becksfort, P. Gupta, R. Huether, R. W. Kriwacki, M. Parker, D. J. McGoldrick, D. Zhao, D. Alford, S. Espy, K. C. Bobba, G. Song, D. Pei, C. Cheng, S. Roberts, M. I. Barbato, D. Campana, E. Coustan-Smith, S. A. Shurtleff, S. C. Raimondi, M. Kleppe, J. Cools, K. A. Shimano, M. L. Hermiston, S. Doulatov, K. Eppert, E. Laurenti, F. Notta, J. E. Dick, G. Basso, S. P. Hunger, M. L. Loh, M. Devidas, B. Wood, S. Winter, K. P. Dunsmore, R. S. Fulton, L. L. Fulton, X. Hong, C. C. Harris, D. J. Dooling, K. Ochoa, K. J. Johnson, J. C. Obenauer, W. E. Evans, C. H. Pui, C. W. Naeve, T. J. Ley, E. R. Mardis, R. K. Wilson, J. R. Downing, and C. G. Mullighan. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature*, 481(7380):157–63.

[760] J. Zhang, G. Wu, C. P. Miller, R. G. Tatevossian, J. D. Dalton, B. Tang, W. Orisme, C. Punchihewa, M. Parker, I. Qaddoumi, F. A. Boop, C. Lu, C. Kandoth, L. Ding, R. Lee, R. Huether, X. Chen, E. Hedlund, P. Nagahawatte, M. Rusch, K. Boggs, J. Cheng, J. Becksfort, J. Ma, G. Song, Y. Li, L. Wei, J. Wang, S. Shurtleff, J. Easton, D. Zhao, R. S. Fulton, L. L. Fulton, D. J. Dooling, B. Vadodaria, H. L. Mulder, C. Tang, K. Ochoa, C. G. Mullighan, A. Gajjar, R. Kriwacki, D. Sheer, R. J. Gilbertson, E. R. Mardis, R. K. Wilson, J. R. Downing, S. J. Baker, D. W. Ellison, and Project St. Jude Children's Research Hospital-Washington University Pediatric Cancer Genome. Whole-genome sequencing identifies genetic alterations in pediatric low-grade gliomas. *Nat Genet*, 45(6):602–12, 2013.

[761] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A. L. Borresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjord, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jager, D. T. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. Lopez-Otin, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. Tutt, R. Valdes-Mas, M. M. van Buuren, L. van 't Veer,

A. Vincent-Salomon, N. Waddell, L. R. Yates, Initiative Australian Pancreatic Cancer Genome, Icgc Breast Cancer Consortium, Icgc Mmml-Seq Consortium, Icgc PedBrain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, and M. R. Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–21, 2013.

[762] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*, 3(1):246–59, 2013.

[763] J. S. Gehring, B. Fischer, M. Lawrence, and W. Huber. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics*, 31(22):3673–5, 2015.

[764] A. Fischer, C. J. Illingworth, P. J. Campbell, and V. Mustonen. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol*, 14(4):R39, 2013.

[765] S. Saadatmand, J. R. Vos, M. J. Hooning, J. C. Oosterwijk, L. B. Koppert, G. H. de Bock, M. G. Ausems, C. J. van Asperen, C. M. Aalfs, E. B. Gomez Garcia, H. Meijers-Heijboer, N. Hoogerbrugge, M. Piek, C. Seynaeve, C. Verhoef, M. Rookus, M. M. Tilanus-Linthorst, Breast Hereditary, and Netherlands Ovarian Cancer Research Group. Relevance and efficacy of breast cancer screening in BRCA1 and BRCA2 mutation carriers above 60 years: a national cohort study. *Int J Cancer*, 135(12):2940–9, 2014.

[766] D. J. Slamon, B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde, T. Fleming, W. Eiermann, J. Wolter, M. Pegram, J. Baselga, and L. Norton. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N Engl J Med*, 344(11):783–92, 2001.

[767] B. J. Druker, M. Talpaz, D. J. Resta, B. Peng, E. Buchdunger, J. M. Ford, N. B. Lydon, H. Kantarjian, R. Capdeville, S. Ohno-Jones, and C. L. Sawyers. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N Engl J Med*, 344(14):1031–7, 2001.

[768] C. Palles, J. B. Cazier, K. M. Howarth, E. Domingo, A. M. Jones, P. Broderick, Z. Kemp, S. L. Spain, E. Guarino, I. Salguero, A. Sherborne, D. Chubb, L. G. Carvajal-

Carmona, Y. Ma, K. Kaur, S. Dobbins, E. Barclay, M. Gorman, L. Martin, M. B. Kovac, S. Humphray, Corgi Consortium, W. G. S. Consortium, A. Lucassen, C. C. Holmes, D. Bentley, P. Donnelly, J. Taylor, C. Petridis, R. Roylance, E. J. Sawyer, D. J. Kerr, S. Clark, J. Grimes, S. E. Kearsey, H. J. Thomas, G. McVean, R. S. Houlston, and I. Tomlinson. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet*, 45(2):136–44, 2013.

[769] T. M. Albertson, M. Ogawa, J. M. Bugni, L. E. Hays, Y. Chen, Y. Wang, P. M. Treuting, J. A. Heddle, R. E. Goldsby, and B. D. Preston. DNA polymerase epsilon and delta proofreading suppress discrete mutator and cancer phenotypes in mice. *Proc Natl Acad Sci U S A*, 106(40):17101–4, 2009.

[770] R. E. Goldsby, N. A. Lawrence, L. E. Hays, E. A. Olmsted, X. Chen, M. Singh, and B. D. Preston. Defective DNA polymerase-delta proofreading causes cancer susceptibility in mice. *Nat Med*, 7(6):638–9, 2001.

[771] R. E. Goldsby, L. E. Hays, X. Chen, E. A. Olmsted, W. B. Slayton, G. J. Spangrude, and B. D. Preston. High incidence of epithelial cancers in mice deficient for DNA polymerase delta proofreading. *Proc Natl Acad Sci U S A*, 99(24):15560–5, 2002.

[772] G. M. Church and W. Gilbert. Genomic sequencing. *Proc Natl Acad Sci U S A*, 81(7):1991–5, 1984.

[773] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*, 74(12):5463–7, 1977.

[774] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. Comparison of next-generation sequencing systems. *J Biomed Biotechnol*, 2012:251364, 2012.

[775] T. S. Seo, X. Bai, D. H. Kim, Q. Meng, S. Shi, H. Ruparel, Z. Li, N. J. Turro, and J. Ju. Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proc Natl Acad Sci U S A*, 102(17):5926–31, 2005.

[776] S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, 17(6):333–51, 2016.

[777] E. R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends Genet*, 24(3):133–41, 2008.

[778] C. W. Fuller, L. R. Middendorf, S. A. Benner, G. M. Church, T. Harris, X. Huang, S. B. Jovanovich, J. R. Nelson, J. A. Schloss, D. C. Schwartz, and D. V. Vezenov. The challenges of sequencing by synthesis. *Nat Biotechnol*, 27(11):1013–23, 2009.

[779] Large-scale genome sequencing and analysis centers (LSAC): The cost of sequencing a human genome. url: https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome. accessed 10 july 2016.

[780] A. A. Duina, M. E. Miller, and J. B. Keeney. Budding yeast for budding geneticists: A primer on the Saccharomyces cerevisiae Model System. *Genetics*, 197(1):33–48, 2014.

[781] R. K. Mortimer and J. R. Johnston. Genealogy of principal strains of the yeast genetic stock center. *Genetics*, 113(1):35–43, 1986.

[782] A. Hinnen, J. B. Hicks, and G. R. Fink. Transformation of yeast. *Proc Natl Acad Sci U S A*, 75(4):1929–33, 1978.

[783] A. Baudin, O. Ozier-Kalogeropoulos, A. Denouel, F. Lacroute, and C. Cullin. A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. *Nucleic Acids Res*, 21(14):3329–30, 1993.

[784] C. B. Brachmann, A. Davies, G. J. Cost, E. Caputo, J. Li, P. Hieter, and J. D. Boeke. Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast*, 14(2):115–32, 1998.

[785] G. Giaever and C. Nislow. The yeast deletion collection: a decade of functional genomics. *Genetics*, 197(2):451–65, 2014.

[786] G. Prelich. Suppression mechanisms: themes from variations. *Trends Genet*, 15(7):261–6, 1999.

[787] W.-K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O'Shea. Global analysis of protein localization in budding yeast. *Nature*, 425:686–691, 2003.

[788] A. H. Tong and C. Boone. Synthetic genetic array analysis in Saccharomyces cerevisiae. *Methods Mol Biol*, 313:171–192, 2006.

[789] W. P. Tansey. Yeast Chromatin Immunoprecipitation (ChIP) Assay. *CSH Protoc*, 2007:pdb prot4642, 2007.

[790] C. T. Chien, P. L. Bartel, R. Sternglanz, and S. Fields. The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc Natl Acad Sci U S A*, 88(21):9578–9582, 1991.

[791] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 274(5287):546, 563–7, 1996.

[792] K. H. Wolfe and D. C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–713, 1997.

[793] M. Kellis, B. W. Birren, and E. S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature*, 428(6983):617–24, 2004.

[794] K. H. Wolfe. Origin of the yeast whole-genome duplication. *PLoS Biol*, 13(8):e1002221, 2015.

[795] YeastMine: List of verified ORFs. URL: "http://yeastmine.yeastgenome.org/".

[796] HumanMine. URL: "http://www.humanmine.org/humanmine/portal.do".

[797] J. E. Haber. Mating-type genes and MAT switching in Saccharomyces cerevisiae. *Genetics*, 191(1):33–64, 2012.

[798] R. W. Schekman. Nobel Lecture: Genetic and Biochemical Dissection of the Secretory Pathway. 2014.

[799] *From Les Prix Nobel. The Nobel Prizes 2009.* Nobel Foundation, Stockholm, 2010.

[800] *From Les Prix Nobel. The Nobel Prizes 2006.* Nobel Foundation, Stockholm, 2007.

[801] F. Puddu, T. Oelschlaegel, I. Guerini, N. J. Geisler, H. Niu, M. Herzog, I. Salguero, B. Ochoa-Montano, E. Vire, P. Sung, D. J. Adams, T. M. Keane, and S. P. Jackson. Synthetic viability genomic screening defines Sae2 function in dna repair. *EMBO J*, 34(11):1509–22, 2015.

[802] J. V. Forment, M. Herzog, J. Coates, T. Konopka, B. V. Gapp, S. M. Nijman, D. J. Adams, T. M. Keane, and S. P. Jackson. Genome-wide genetic screening with

chemically-mutagenized haploid embryonic stem cells. *Nat Chem Biol*, 13(1):12–14, 20174. Accepted Aug 2016.

[803] A. de la Chapelle. Genetic predisposition to colorectal cancer. *Nat Rev Cancer*, 4(10):769–80, 2004.

[804] H. T. Lynch and A. de la Chapelle. Hereditary colorectal cancer. *N Engl J Med*, 348(10):919–32, 2003.

[805] S. Briggs and I. Tomlinson. Germline and somatic polymerase epsilon and delta mutations define a new class of hypermutated colorectal and endometrial cancers. *J Pathol*, 230(2):148–53, 2013.

[806] E. Heitzer and I. Tomlinson. Replicative DNA polymerase mutations in cancer. *Curr Opin Genet Dev*, 24:107–13, 2014.

[807] H. T. Tran, J. D. Keen, M. Kricker, M. A. Resnick, and D. A. Gordenin. Hypermutability of homonucleotide runs in mismatch repair and DNA polymerase proofreading yeast mutants. *Mol Cell Biol*, 17(5):2859–2865, 1997.

[808] M. Strand, T. A. Prolla, R. M. Liskay, and T. D. Petes. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature*, 365(6443):274–6, 1993.

[809] Network Cancer Genome Atlas Research, C. Kandoth, N. Schultz, A. D. Cherniack, R. Akbani, Y. Liu, H. Shen, A. G. Robertson, I. Pashtan, R. Shen, C. C. Benz, C. Yau, P. W. Laird, L. Ding, W. Zhang, G. B. Mills, R. Kucherlapati, E. R. Mardis, and D. A. Levine. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73, 2013.

[810] D. N. Church, S. E. Briggs, C. Palles, E. Domingo, S. J. Kearsey, J. M. Grimes, M. Gorman, L. Martin, K. M. Howarth, S. V. Hodgson, Nsecg Collaborators, K. Kaur, J. Taylor, and I. P. Tomlinson. DNA polymerase epsilon and delta exonuclease domain mutations in endometrial cancer. *Hum Mol Genet*, 22(14):2820–8, 2013.

[811] NCBI Genome Remapping Service. URL: http://www.ncbi.nlm.nih.gov/genome/tools/remap.

[812] S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, C. Y. Kok, M. Jia, T. De, J. W. Teague, M. R. Stratton, U. McDermott, and P. J. Campbell. COSMIC: exploring the world's knowledge of

somatic mutations in human cancer. *Nucleic Acids Res*, 43(Database issue):D805–11, 2015.

[813] The Single Nucleotide Polymorphism Database (dbSNP). URL: http://www.ncbi.nlm.nih.gov/projects/SNP/.

[814] Genomes Project Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[815] P. Kumar, S. Henikoff, and P. C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073–1082, 2009.

[816] P. C. Ng and S. Henikoff. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet*, 7:61–80, 2006.

[817] P. C. Ng and S. Henikoff. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31(13):3812–3814, 2003. doi: 10.1093/nar/gkg509.

[818] P. C. Ng and S. Henikoff. Accounting for human polymorphisms predicted to affect protein function. *Genome Res*, 12(3):436–46, 2002.

[819] P. C. Ng and S. Henikoff. Predicting deleterious amino acid substitutions. *Genome Res*, 11(5):863–74, 2001.

[820] V. Ramensky, P. Bork, and S. Sunyaev. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*, 30(17):3894–900, 2002.

[821] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248–9, 2010.

[822] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet*, Chapter 7:Unit7 20, 2013.

[823] S. Sunyaev, V. Ramensky, I. Koch, W. Lathe III, A. S. Kondrashov, and P. Bork. Prediction of deleterious human alleles. *Hum Mol Genet*, 10(6):591–7, 2001.

[824] G. Yachdav, E. Kloppmann, L. Kajan, M. Hecht, T. Goldberg, T. Hamp, P. Honigschmid, A. Schafferhans, M. Roos, M. Bernhofer, L. Richter, H. Ashkenazy, M. Punta, A. Schlessinger, Y. Bromberg, R. Schneider, G. Vriend, C. Sander, N. Ben-Tal, and B. Rost. PredictProtein-an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res*, 42(W1):W337–W343, 2014.

[825] M. J. Betts, Q. Lu, Y. Jiang, A. Drusko, O. Wichmann, M. Utz, I. A. Valtierra-Gutierrez, M. Schlesner, N. Jaeger, D. T. Jones, S. Pfister, P. Lichter, R. Eils, R. Siebert, P. Bork, G. Apic, A. C. Gavin, and R. B. Russell. Mechismo: predicting the mechanistic impact of mutations and modifications on molecular interactions. *Nucleic Acids Res*, 43(2):e10, 2015.

[826] J. M. Schwarz, D. N. Cooper, M. Schuelke, and D. Seelow. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods*, 11(4):361–2, 2014.

[827] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*, 7:539, 2011.

[828] M. Goujon, H. McWilliam, W. Li, F. Valentin, S. Squizzato, J. Paern, and R. Lopez. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res*, 38(Web Server issue):W695–9, 2010.

[829] H. McWilliam, W. Li, M. Uludag, S. Squizzato, Y. M. Park, N. Buso, A. P. Cowley, and R. Lopez. Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res*, 41(Web Server issue):W597–600, 2013.

[830] M. S. Williamson, J. C. Game, and S. Fogel. Meiotic gene conversion mutants in Saccharomyces cerevisiae. I. Isolation and characterization of pms1-1 and pms1-2. *Genetics*, 110(4):609–46, 1985.

[831] H. Sychrova and M. R. Chevallier. Cloning and sequencing of the Saccharomyces cerevisiae gene LYP1 coding for a lysine-specific permease. *Yeast*, 9(7):771–82, 1993.

[832] G. F. Crouse. Mutagenesis assays in yeast. *Methods*, 22(2):116–9, 2000.

[833] C. Chen and R. D. Kolodner. Gross chromosomal rearrangements in Saccharomyces cerevisiae replication and recombination defective mutants. *Nat Genet*, 23(1):81–5, 1999.

[834] J. M. Sheltzer, H. M. Blank, S. J. Pfau, Y. Tange, B. M. George, T. J. Humpton, I. L. Brito, Y. Hiraoka, O. Niwa, and A. Amon. Aneuploidy drives genomic instability in yeast. *Science*, 333(6045):1026–30, 2011.

[835] A. Serero, C. Jubin, S. Loeillet, P. Legoix-Ne, and A. G. Nicolas. Mutational landscape of yeast mutator strains. *Proc Natl Acad Sci U S A*, 111(5):1897–902, 2014.

[836] M. Zackrisson, J. Hallin, L. G. Ottosson, P. Dahl, E. Fernandez-Parada, E. Landstrom, L. Fernandez-Ricaud, P. Kaferle, A. Skyman, S. Stenberg, S. Omholt, U. Petrovic, J. Warringer, and A. Blomberg. Scan-o-matic: High-resolution microbial phenomics at a massive scale. *G3 (Bethesda)*, 2016.

[837] M. Ralser, H. Kuhl, M. Ralser, M. Werber, H. Lehrach, M. Breitenbach, and B. Timmermann. The Saccharomyces cerevisiae W303-K6001 cross-platform genome sequence: insights into ancestry and physiology of a laboratory mutt. *Open Biol*, 2(8):120093, 2012.

[838] Wellcome Trust Sanger Institute. Saccharomyces Genome Resequencing Project. URL: http://www.sanger.ac.uk/research/projects/genomeinformatics/sgrp.html.

[839] H. Thorvaldsdottir, J. T. Robinson, and J. P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*, 14(2):178–92, 2013.

[840] X. Hu, J. Yuan, Y. Shi, J. Lu, B. Liu, Z. Li, Y. Chen, D. Mu, H. Zhang, N. Li, Z. Yue, F. Bai, H. Li, and W. Fan. pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics*, 28(11):1533–5, 2012.

[841] M. Lynch, W. Sung, K. Morris, N. Coffey, C. R. Landry, E. B. Dopman, W. J. Dickinson, K. Okamoto, S. Kulkarni, D. L. Hartl, and W. K. Thomas. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A*, 105(27):9272–7, 2008.

[842] M. Boutros and J. Ahringer. The art and design of genetic screens: RNA interference. *Nat Rev Genet*, 9(7):554–66, 2008.

[843] J. E. Carette, C. P. Guimaraes, M. Varadarajan, A. S. Park, I. Wuethrich, A. Godarova, M. Kotecki, B. H. Cochran, E. Spooner, H. L. Ploegh, and T. R. Brummelkamp. Haploid genetic screens in human cells identify host factors used by pathogens. *Science*, 326(5957):1231–1235, 2009.

[844] H. Koike-Yusa, Y. Li, E. P. Tan, C. Velasco-Herrera Mdel, and K. Yusa. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat Biotechnol*, 32(3):267–73, 2014.

[845] O. Shalem, N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, T. S. Mikkelsen, D. Heckl, B. L. Ebert, D. E. Root, J. G. Doench, and F. Zhang. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, 343(6166):84–7, 2014.

[846] T. Wang, J. J. Wei, D. M. Sabatini, and E. S. Lander. Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, 343(6166):80–84, 2014.

[847] T. Rolef Ben-Shahar, S. Heeger, C. Lehane, P. East, H. Flynn, M. Skehel, and F. Uhlmann. Eco1-dependent cohesin acetylation during establishment of sister chromatid cohesion. *Science*, 321(5888):563–6, 2008.

[848] M. Leeb and A. Wutz. Derivation of haploid embryonic stem cells from mouse embryos. *Nature*, 479(7371):131–4, 2011.

[849] G. A. Lepage and M. Jones. Purinethiols as feedback inhibitors of purine synthesis in ascites tumor cells. *Cancer Research*, 21(5):642–649, 1961.

[850] T. M. Keane, L. Goodstadt, P. Danecek, M. A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, N. A. Furlotte, E. Eskin, C. Nellaker, H. Whitley, J. Cleak, D. Janowitz, P. Hernandez-Pliego, A. Edwards, T. G. Belgard, P. L. Oliver, R. E. McIntyre, A. Bhomra, J. Nicod, X. Gan, W. Yuan, L. van der Weyden, C. A. Steward, S. Bala, J. Stalker, R. Mott, R. Durbin, I. J. Jackson, A. Czechanski, J. A. Guerra-Assuncao, L. R. Donahue, L. G. Reinholdt, B. A. Payseur, C. P. Ponting, E. Birney, J. Flint, and D. J. Adams. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–94, 2011.

[851] J. O'Rawe, T. Jiang, G. Sun, Y. Wu, W. Wang, J. Hu, P. Bodily, L. Tian, H. Hakonarson, W. E. Johnson, Z. Wei, K. Wang, and G. J. Lyon. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med*, 5(3):28, 2013.

[852] J. M. Zook, B. Chapman, J. Wang, D. Mittelman, O. Hofmann, W. Hide, and M. Salit. Integrating human sequence data sets provides a resource of benchmark SNP and INDEL genotype calls. *Nat Biotechnol*, 32(3):246–51, 2014.

[853] G. Narzisi, J. A. O'Rawe, I. Iossifov, H. Fang, Y. H. Lee, Z. Wang, Y. Wu, G. J. Lyon, M. Wigler, and M. C. Schatz. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nat Methods*, 11(10):1033–6, 2014.

[854] P. F. Swann, T. R. Waters, D. C. Moulton, Y. Z. Xu, Q. Zheng, M. Edwards, and R. Mace. Role of postreplicative DNA mismatch repair in the cytotoxic action of thioguanine. *Science*, 273(5278):1109–11, 1996.

[855] G. Guo, W. Wang, and A. Bradley. Mismatch repair genes identified using genetic screens in Blm-deficient embryonic stem cells. *Nature*, 429(6994):891–895, 2004.

[856] H. A. Jinnah, L. De Gregorio, J. C. Harris, W. L. Nyhan, and J. P. O'Neill. The spectrum of inherited mutations causing HPRT deficiency: 75 new cases and a review of 196 previously reported cases. *Mutat Res*, 463(3):309–26, 2000.

[857] K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, S. D. McGrath, M. C. Wendl, Q. Zhang, D. P. Locke, X. Shi, R. S. Fulton, T. J. Ley, R. K. Wilson, L. Ding, and E. R. Mardis. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*, 6(9):677–81, 2009.

[858] R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*, 15(6):R84, 2014.

[859] K. Wong, T. M. Keane, J. Stalker, and D. J. Adams. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol*, 11(12):R128, 2010.

[860] T. J. Treangen and S. L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*, 13(1):36–46, 2012.

[861] P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla,

S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, et al. The b73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112–5, 2009.

[862] T. Kobayashi. Regulation of ribosomal RNA gene copy number and its role in modulating genome integrity and evolutionary adaptability in yeast. *Cell Mol Life Sci*, 68(8):1395–403, 2011.

[863] T. Kobayashi, D. J. Heck, M. Nomura, and T. Horiuchi. Expansion and contraction of ribosomal DNA repeats in Saccharomyces cerevisiae: requirement of replication fork blocking (Fob1) protein and the role of RNA polymerase I. *Genes Dev*, 12(24):3821–3830, 1998.

[864] Takehiko Kobayashi. Strategies to maintain the stability of the ribosomal RNA gene repeats. *Genes Genet Syst*, 81(3):155–161, 2006.

[865] E. O. Long and I. B. Dawid. Repeated genes in eukaryotes. *Annu Rev Biochem*, 49:727–64, 1980.

[866] S. Ide, T. Miyazaki, H. Maki, and T. Kobayashi. Abundance of ribosomal RNA gene copies maintains genome integrity. *Science*, 327(5966):693–6, 2010.

[867] Y. Takeuchi, T. Horiuchi, and T. Kobayashi. Transcription-dependent recombination and the role of fork collision in yeast rDNA. *Genes Dev*, 17(12):1497–506, 2003.

[868] B. J. Brewer, D. Lockshon, and W. L. Fangman. The arrest of replication forks in the rDNA of yeast occurs independently of transcription. *Cell*, 71(2):267–276, 1992.

[869] T. Kobayashi, M. Hidaka, M. Nishizawa, and T. Horiuchi. Identification of a site required for DNA replication fork blocking activity in the rRNA gene cluster in Saccharomyces cerevisiae. *Mol Gen Genet*, 233(3):355–62, 1992.

[870] J. M. Kim, S. Vanguri, J. D. Boeke, A. Gabriel, and D. F. Voytas. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed

by the complete Saccharomyces cerevisiae genome sequence. *Genome Res*, 8(5):464–78, 1998.

[871] J. Gafner and P. Philippsen. The yeast transposon Ty1 generates duplications of target DNA on insertion. *Nature*, 286(5771):414–418, 1980.

[872] D. J. Garfinkel, K. M. Nyswaner, K. M. Stefanisko, C. Chang, and S. P. Moore. Ty1 copy number dynamics in Saccharomyces. *Genetics*, 169(4):1845–57, 2005.

[873] Z. Ding, M. Mangino, A. Aviv, T. Spector, R. Durbin, and Uk K. Consortium. Estimating telomere length from whole genome sequence data. *Nucleic Acids Res*, 42(9):e75, 2014.

[874] K. Forstemann, M. Hoss, and J. Lingner. Telomerase-dependent repeat divergence at the 3' ends of yeast telomeres. *Nucleic Acids Res*, 28(14):2690–2694, 2000.

[875] Y. O. Zhu, M. L. Siegal, D. W. Hall, and D. A. Petrov. Precise estimates of mutation rate and spectrum in yeast. *Proc Natl Acad Sci U S A*, 111(22):E2310–8, 2014.

[876] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91, 1999.

[877] L. N. Hutchins, S. M. Murphy, P. Singh, and J. H. Graber. Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics*, 24(23):2684–2690, 2008.

[878] L. B. Alexandrov and M. R. Stratton. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev*, 24:52–60, 2014. doi: 10.1016/j.gde.2013.11.014.

[879] Cosmic: Signatures of mutational processes in human cancer. url: "http://cancer.sanger.ac.uk/cosmic/signatures". accessed 10 july 2016.

[880] A. G. Lada, A. Dhar, R. J. Boissy, M. Hirano, A. A. Rubel, I. B. Rogozin, and Y. I. Pavlov. AID/APOBEC cytosine deaminase induces genome-wide kataegis. *Biol Direct*, 7:47; discussion 47, 2012.

[881] G. Poulogiannis, I. M. Frayling, and M. J. Arends. DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome. *Histopathology*, 56(2):167–79, 2010.

[882] E. Shinbrot, E. E. Henninger, N. Weinhold, K. R. Covington, A. Y. Goksenin, N. Schultz, H. Chao, H. Doddapaneni, D. M. Muzny, R. A. Gibbs, C. Sander, Z. F. Pursell, and D. A. Wheeler. Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome Res*, 24(11):1740–50, 2014.

[883] M. Garbacz, H. Araki, K. Flis, A. Bebenek, A. E. Zawada, P. Jonczyk, K. Makiela-Dzbenska, and I. J. Fijalkowska. Fidelity consequences of the impaired interaction between DNA polymerase epsilon and the GINS complex. *DNA Repair (Amst)*, 29:23–35, 2015.

[884] J. Kraszewska, M. Garbacz, P. Jonczyk, I. J. Fijalkowska, and M. Jaszczur. Defect of Dpb2p, a noncatalytic subunit of DNA polymerase varepsilon, promotes error prone replication of undamaged chromosomal DNA in Saccharomyces cerevisiae. *Mutat Res*, 737(1-2):34–42, 2012.

[885] A. V. Makarova, J. L. Stodola, and P. M. Burgers. A four-subunit DNA polymerase zeta complex containing Pol delta accessory subunits is essential for PCNA-mediated mutagenesis. *Nucleic Acids Res*, 40(22):11618–26, 2012.

[886] R. E. Johnson, L. Prakash, and S. Prakash. Pol31 and Pol32 subunits of yeast DNA polymerase delta are also essential subunits of DNA polymerase zeta. *Proc Natl Acad Sci U S A*, 109(31):12455–60, 2012.

[887] T. W. Chiang, C. le Sage, D. Larrieu, M. Demir, and S. P. Jackson. CRISPR-Cas9(D10A) nickase-based genotypic and phenotypic screening to enhance genome editing. *Sci Rep*, 6:24356, 2016.

[888] D. P. Kane and P. V. Shcherbakova. A common cancer-associated DNA polymerase epsilon mutation causes an exceptionally strong mutator phenotype, indicating fidelity defects distinct from loss of proofreading. *Cancer Res*, 74(7):1895–901, 2014.

[889] D. Kumar, J. Viberg, A. K. Nilsson, and A. Chabes. Highly mutagenic and severely imbalanced dNTP pools can escape detection by the S-phase checkpoint. *Nucleic Acids Res*, 38(12):3975–83, 2010.

[890] C. D. Sohl, S. Ray, and J. B. Sweasy. Pools and Pols: Mechanism of a mutator phenotype. *Proc Natl Acad Sci U S A*, 112(19):5864–5, 2015.

[891] L. N. Williams, L. Marjavaara, G. M. Knowels, E. M. Schultz, E. J. Fox, A. Chabes, and A. J. Herr. dNTP pool levels modulate mutator phenotypes of error-prone DNA polymerase epsilon variants. *Proc Natl Acad Sci U S A*, 112(19):E2457–66, 2015.

[892] T. M. Mertz, S. Sharma, A. Chabes, and P. V. Shcherbakova. Colon cancer-associated mutator DNA polymerase delta variant causes expansion of dNTP pools increasing its own infidelity. *Proc Natl Acad Sci U S A*, 112(19):E2467–76, 2015.

[893] C. N. Greene and S. Jinks-Robertson. Spontaneous frameshift mutations in Saccharomyces cerevisiae: accumulation during DNA replication and removal by proofreading and mismatch repair activities. *Genetics*, 159(1):65–75, 2001.

[894] P. M. Treuting, T. M. Albertson, and B. D. Preston. Case series: acute tumor lysis syndrome in mutator mice with disseminated lymphoblastic lymphoma. *Toxicol Pathol*, 38(3):476–85, 2010.

[895] A. J. Herr, M. Ogawa, N. A. Lawrence, L. N. Williams, J. M. Eggington, M. Singh, R. A. Smith, and B. D. Preston. Mutator suppression and escape from replication error-induced extinction in yeast. *PLoS Genet*, 7(10):e1002282, 2011.

[896] L. N. Williams, A. J. Herr, and B. D. Preston. Emergence of DNA polymerase epsilon antimutators that escape error-induced extinction in yeast. *Genetics*, 193(3):751–70, 2013.

[897] A. A. Agbor, A. Y. Goksenin, K. G. LeCompte, S. H. Hans, and Z. F. Pursell. Human Pol epsilon-dependent replication errors and the influence of mismatch repair on their correction. *DNA Repair (Amst)*, 12(11):954–63, 2013.

[898] A. Shlien, B. B. Campbell, R. de Borja, L. B. Alexandrov, D. Merico, D. Wedge, P. Van Loo, P. S. Tarpey, P. Coupland, S. Behjati, A. Pollett, T. Lipman, A. Heidari, S. Deshmukh, N. Avitzur, B. Meier, M. Gerstung, Y. Hong, D. M. Merino, M. Ramakrishna, M. Remke, R. Arnold, G. B. Panigrahi, N. P. Thakkar, K. P. Hodel, E. E. Henninger, A. Y. Goksenin, D. Bakry, G. S. Charames, H. Druker, J. Lerner-Ellis, M. Mistry, R. Dvir, R. Grant, R. Elhasid, R. Farah, G. P. Taylor, P. C. Nathan, S. Alexander, S. Ben-Shachar, S. C. Ling, S. Gallinger, S. Constantini, P. Dirks, A. Huang, S. W. Scherer, R. G. Grundy, C. Durno, M. Aronson, A. Gartner, M. S. Meyn, M. D. Taylor, Z. F. Pursell, C. E. Pearson, D. Malkin, P. A. Futreal, M. R. Stratton, E. Bouffet,

C. Hawkins, P. J. Campbell, U. Tabori, and Biallelic Mismatch Repair Deficiency Consortium. Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. *Nat Genet*, 47(3):257–62, 2015.

[899] E. Shinbrot, E. E. Henninger, N. Weinhold, K. R. Covington, A. Y. Goksenin, N. Schultz, H. Chao, H. Doddapaneni, D. M. Muzny, R. A. Gibbs, C. Sander, Z. F. Pursell, and D. A. Wheeler. Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome Res*, 24(11):1740–50, 2014.

[900] M. S. Longtine, A. McKenzie, D. J. Demarini, N. G. Shah, A. Wach, A. Brachat, P. Philippsen, and J. R. Pringle. Additional modules for versatile and economical PCR-based gene deletion and modification in Saccharomyces cerevisiae. *Yeast (Chichester, England)*, 14:953–961, 1998.

[901] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009.

[902] Picard. URL: "http://broadinstitute.github.io/picard/".

[903] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25:2078–9, 2009.

[904] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, 26(16):2069–70, 2010.

[905] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and Group Genomes Project Analysis. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–8, 2011.

[906] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–2, 2010.

[907] J. Gehring. Inferring Somatic Signatures from Single Nucleotide Variant Calls. URL: http://www.bioconductor.org/packages/devel/bioc/vignettes/SomaticSignatures/inst/doc/SomaticSignatures-vignette.html; last accessed: 02/06/2016.

[908] H. Li and R. Durbin. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, 26(5):589–95, 2010. doi: 10.1093/bioinformatics/btp698.

[909] J. T. Robinson, H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nat Biotechnol*, 29(1):24–6, 2011.

[910] H. Thorvaldsdottir, J. T. Robinson, and J. P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*, 14(2):178–92, 2013.

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

NHEJ	Non-homologous end joining
NIR	Non-ionising radiation
NMD	Nonsense-mediated decay
NMF	Nonnegative matrix factorization
PCR	Polymerase chain reaction
PEG	Polyethylene Glycol Pol Polymerase
Phleo	Phleomycin
ORF	Open Reading Frame
RFC	Replication factor C
RNA	Ribonucleic acid
RNS	Reactive nitrogen species
ROS	Reactive oxygen species
rpm	Revolutions per minute
RT	Room Temperature
SAC	Spindle assembly checkpoint
SDSA	Synthesis-dependent strand annealing
SGA	Synthetic Gene Array
SGD	Saccharomyces Genome Database
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SSA	single-strand annealing
ssDNA	Single-stranded DNA
SV	Structural Variant

TC-NER	Transcription-coupled nucleotide excision repair
TCGA	The Cancer Genome Atlas
TE	Transposable Element
TLS	Translesion synthesis
Tm	Melting temperature (e.g. for oligonucleotides)
Tris	Tris(hydroxymethyl)aminomethane
UPD	Uniparental disomy
UV	Ultraviolet
UV-A	Ultraviolet A
UV-B	Ultraviolet B
UV-C	Ultraviolet C
VEP	Variant Effect Predictor
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
YNB	Yeast Nitrogen Base
YPD	Yeast Extract - Peptone - Dextrose

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

Step	Software/Tool	Command	Command
ABI sequence alignment	BWA[908]	bwasw	-
Variant Calling of ABI files	SAMTools [903]	mpileup	-u
Variant Calling of ABI files	BCFtools[903]	view	-c -v
Filtering of ABI vcf files	VCFtools [905]	vcf-annotate	-f +/d=2/D=5
Generate INDEL Set	pIRS[840]	pirs diploid	-a 3 -v 0 -d 0.000075
Generate Control Set	pIRS[840]	pirs diploid	-s 0.0001
Generate Mutated Set	pIRS[840]	pirs diploid	-s 0.000025
Simulate sequencing	pIRS[840]	pirs simulate	-x 40(20,30,50) -m 450
Alignment	BWA[908]	v0.6.2	-q 15
Variant Calling	SAMTools [903]	mpileup	-g-tDP,DV-C50-pm3-F0.2-d10000
Variant Calling	BCFtools[903]	call	-vm -f GQ
Intersecting Variants	BEDtools[906]	intersect	-a -b -v
Visualising variants	IGV[909, 910]	-	-

B.1.1.2 Software tools and parameters used for sequencing analysis of *S. cerevisiae*

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

B.1.2 Strains used in mutation accumulation (MA) experiments experiments

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

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

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.

Het.	Haploid	polymerase mutation	parallel lines
YMH9	YMH8	wild-type	28
YMH29	YMH28	*pol2-4*	28
YMH27	YMH26	*pol2-A480V*	18
YMH21	YMH20	*pol2-D290V*	18
YMH13	YMH12	*pol2-L439V*	18
YMH23	YMH22	*pol2-M459K*	18
YMH19	YMH18	*pol2-P301R*	18
YMH25	YMH24	*pol2-Q468R*	18
YMH17	YMH16	*pol2-S312F*	18
YMH15	YMH14	*pol2-V426L*	18
YMH11	YMH10	*pol3-S384N*	18

B.1.3 6-Thioguanine supressor screen of haploid mouse cells

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

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

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).fsdjakl

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 identfied.

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 identfied mutations. Table 4 summarizes the potential causative mutation for all suppressor screens with references when identfied 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 identfied in the same clones. Table 3 contains all mutations identfied 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 seqeunces 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.

Synthetic viability genomic screening defines Sae2 function in DNA repair. Fabio Puddu, Tobias Oelschlaegel, Ilaria Guerini, Nicola J Geisler, Hengyao Niu, Mareike Herzog, Israel Salguero, Bernardo Ochoa-Montaño, Emmanuelle Viré, Patrick Sung, David J Adams, Thomas M Keane, Stephen P Jackson. *EMBO J.* 2015 **34**(11):1509-22. doi: 10.15252/embj.201590973. PMID: 25899817

In this work synthetic viability screening was used in budding yeast do 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 *mre11-H37R* and *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 identfied 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 *tof1*Δ 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.

Published online: April 21, 2015

Article

TRANSPARENT PROCESS

OPEN ACCESS

THE EMBO JOURNAL

Synthetic viability genomic screening defines Sae2 function in DNA repair

Fabio Puddu[1,†], Tobias Oelschlaegel[1,†], Ilaria Guerini[1], Nicola J Geisler[1], Hengyao Niu[3], Mareike Herzog[1,2], Israel Salguero[1], Bernardo Ochoa-Montaño[1], Emmanuelle Viré[1], Patrick Sung[3], David J Adams[2], Thomas M Keane[2] & Stephen P Jackson[1,2,*]

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 *sae2*Δ 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
The EMBO Journal (2015) 34: 1509–1522

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 display endonuclease 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

[1] The Gurdon Institute and Department of Biochemistry, University of Cambridge, Cambridge, UK
[2] The Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
[3] Molecular Biophysics and Biochemistry, Yale University School of Medicine, New Haven, CT, USA
*Corresponding author. Tel: +44 1223 334088; E-mail: s.jackson@gurdon.cam.ac.uk
[†]These authors contributed equally to this work

Published online: April 21, 2015

function together with MRX in DSB repair. For instance, *mre11-nd* as well as *mre11S* and *rad50S* hypomorphic alleles phenocopy *SAE2* deletion (*sae2*Δ) 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 *sae2*Δ 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 *sae2*Δ 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 *sae2*Δ 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 paradoxes regarding Sae2 and MRX functions, namely the fact that while deletion of either *SAE2* or *MRE11* causes hypersensitivity to DNA-damaging agents, the resection defect of *sae2*Δ strains is negligible compared to that of *mre11*Δ cells, and lack of Sae2 causes an increase in Mre11 persistence at DSB ends rather than a loss. Our model invokes Mre11/MRX removal from DNA as a critical step in allowing HR to proceed effectively on a resected DNA template.

Results

SVGS identifies Mre11 mutations as *sae2*Δ 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 *sae2*Δ 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 *sae2*Δ 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).

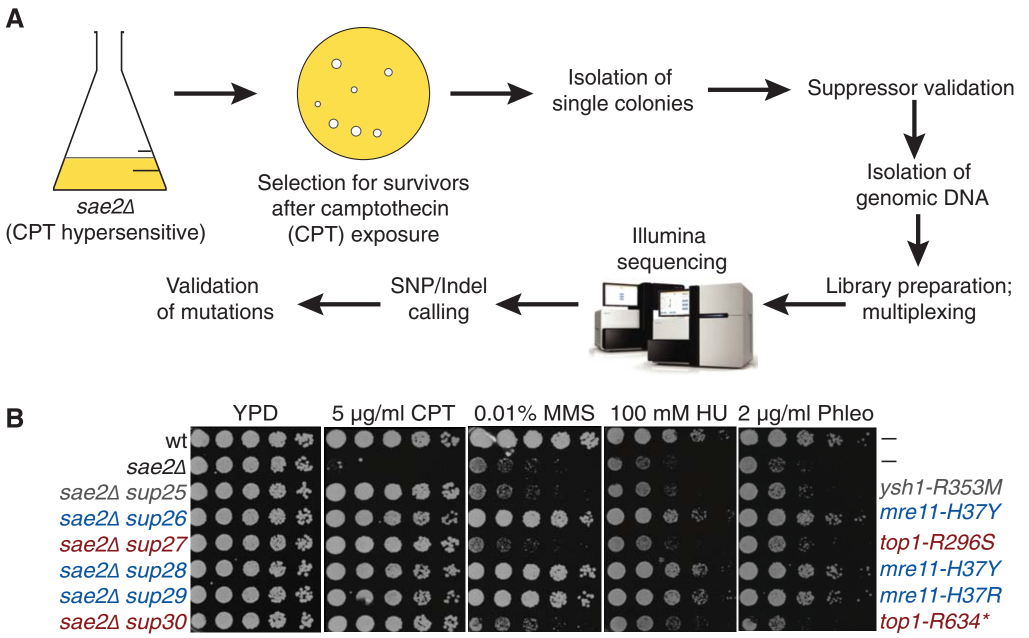
To identify mutations causing these suppression phenotypes, genomic DNA from the 48 clones was isolated and analysed by next-generation Illumina sequencing. We then used bioinformatics tools (see Materials and Methods) to identify mutations altering open reading frames within the reference *S. cerevisiae* genome (Fig 1A). This revealed that 24 clones displaying camptothecin resistance but retaining *sae2*Δ 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 sae2*Δ suppressor (*mre11*$^{SUPsae2\Delta}$) alleles.

mre11$^{SUPsae2\Delta}$ alleles suppress many *sae2*Δ 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 *sae2*Δ 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 resensitised the *mre11*$^{SUPsae2\Delta}$ *sae2*Δ 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 *sae2*Δ*mre11*$^{SUPsae2\Delta}$ cells upon camptothecin treatment. Importantly, independent introduction of *mre11-H37R* and *mre11-H37Y* alleles in a *sae2*Δ strain confirmed that each conferred suppression of *sae2*Δ hypersensitivity to various DNA-damaging agents (Fig 2D). The *mre11-H37R* and *mre11-H37Y* 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 *sae2*Δ camptothecin hypersensitivity was observed by mutating His37 to Ala (*mre11-H37A*; Fig 2G), suggesting that the effects of the *mre11*$^{SUPsae2\Delta}$ 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.

mre11$^{SUPsae2\Delta}$ alleles do not suppress all *sae2*Δ phenotypes

In the absence of Sae2, cells display heightened DNA damage signalling as measured by Rad53 hyperphosphorylation (Clerici *et al*, 2006). As we had found for the DNA damage hypersensitivities of *sae2*Δ cells, this read-out of Sae2 inactivity was also rescued by *mre11-H37R* (Fig 3A). By contrast, *mre11-H37R* did not suppress the sporulation defect of *sae2*Δ cells (unpublished observation). In line with this, *mre11-H37R* did not suppress impaired meiotic DSB processing caused by Sae2 deficiency, as reflected by aberrant accumulation of 5′-bound Spo11 repair intermediates within the *THR4* recombination hot spot (Goldway *et al*, 1993; Fig 3B; as shown in Supplementary Fig S2A, *mre11-H37R* did not itself cause meiotic defects when Sae2 was



C

ORF containing a putative driving mutation	Type				Times found	Number of clones
	Nonsense	Missense	Frameshift	Total		
TOP1	4	4	5	13	28	28
MRE11	0	2	0	2	10	10
YSH1	0	1	0	1	14	3
DNF3	0	1	0	1	12	2
DYS1	0	1	0	1	9	1
IMD2	0	1	0	1	8	1
Not determined						2
Not available						1
Total analysed						48

Figure 1. SVGS identifies mutations suppressing *sae2*Δ DNA damage hypersensitivity.

A Outline of the screening approach that was used to identify suppressors of *sae2*Δ 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 *sae2*Δ 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.

present). Notably, however, *mre11-H37R* rescued the hypersensitivity of *sae2*Δ 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 *mre11*$^{SUPsae2\Delta}$ 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 *sae2*Δ 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 *sae2*Δ 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 *sae2*Δ*mre11-H37R* strain with a *sgs1*Δ strain. As shown in Supplementary Fig S2C, we were unable to recover neither *sgs1*Δ*sae2*Δ nor *sgs1*Δ*sae2*Δ*mre11-H37R* cells, implying that *mre11-H37R* cannot

Published online: April 21, 2015

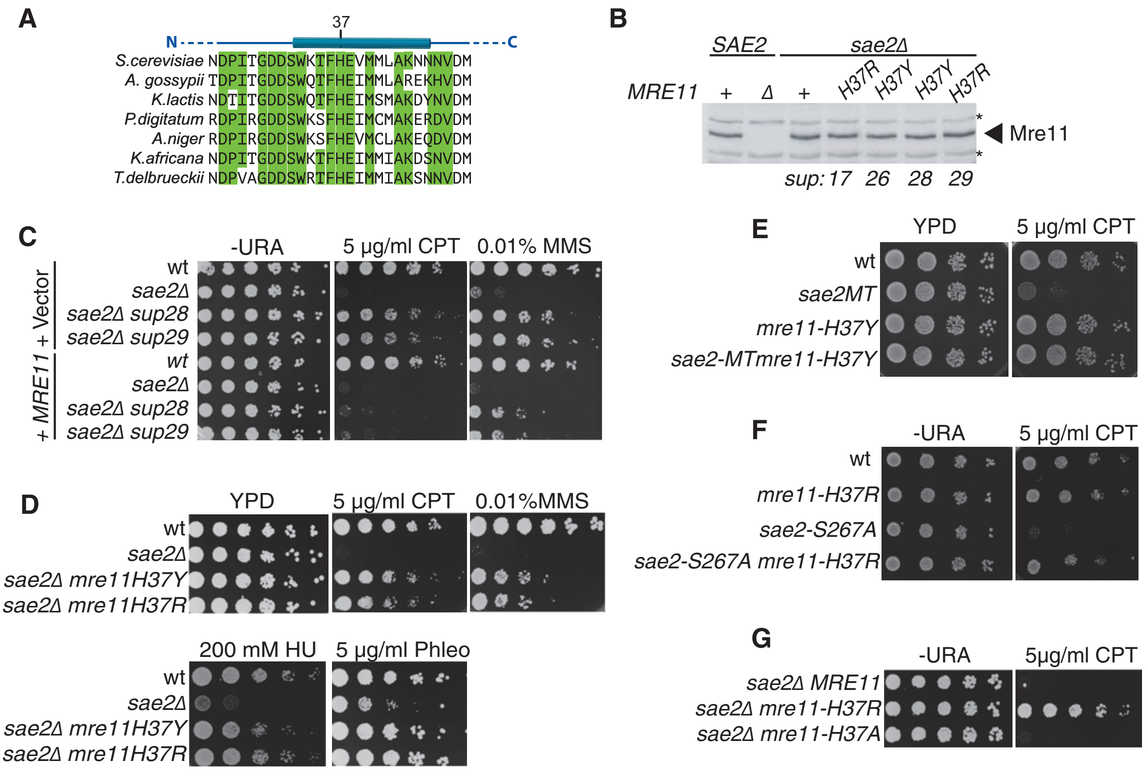


Figure 2. ***mre11-H37R* suppresses the CPT hypersensitivity of *sae2*Δ cells.**

- A Alignment of Mre11 region containing H37 in fungal species; secondary structure prediction is shown above.
- B Western blot with anti-Mre11 antibody on protein extracts prepared from the indicated strains shows that *mre11-H37R* and *mre11-H37Y* mutations do not alter Mre11 protein levels (* indicate cross-reacting proteins).
- C *sup28* and *sup29* suppression is rescued by expressing wild-type (wt) Mre11.
- D *mre11-H37R* and *mre11-H37Y* suppress *sae2*Δ DNA damage hypersensitivity.
- E, F *mre11-H37Y* suppresses DNA damage hypersensitivities of *sae2MT (sae2-2,5,6,8,9)* and *sae2-S267A* cells. CPT, camptothecin; Phleo, phleomycin.
- G *mre11-H37A* does not suppress *sae2*Δ.

suppress this phenotype. In agreement with this conclusion, the *mre11-H37R* mutation did not negatively affect Mre11-dependent telomere maintenance as demonstrated by Southern blot analysis (Supplementary Fig S2D).

Together, the above data revealed that *mre11*$^{SUPsae2\Delta}$ alleles suppressed *sae2*Δ DNA damage hypersensitivities but not *sae2*Δ 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 channelling 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 *sae2*Δ strain (Fig 3F). In addition, this analysis revealed that the previously reported suppression of *sae2*Δ-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 *sae2*Δ camptothecin hypersensitivity by *mre11*$^{SUPsae2\Delta}$ alleles did not require Exo1, indicating that in contrast to suppression of *sae2*Δ 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 *xrs2*Δ strain (Fig 3H), it almost fully rescued the camptothecin hypersensitivity of a strain expressing the *rad50S* allele, which phenocopies *sae2*Δ 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*$^{SUPsae2\Delta}$ mutations might operate, we overexpressed and purified wild-type Mre11, Mre11^{H37R} 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 *sae2*Δ suppression would be mediated by *mre11*$^{SUPsae2\Delta}$ 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

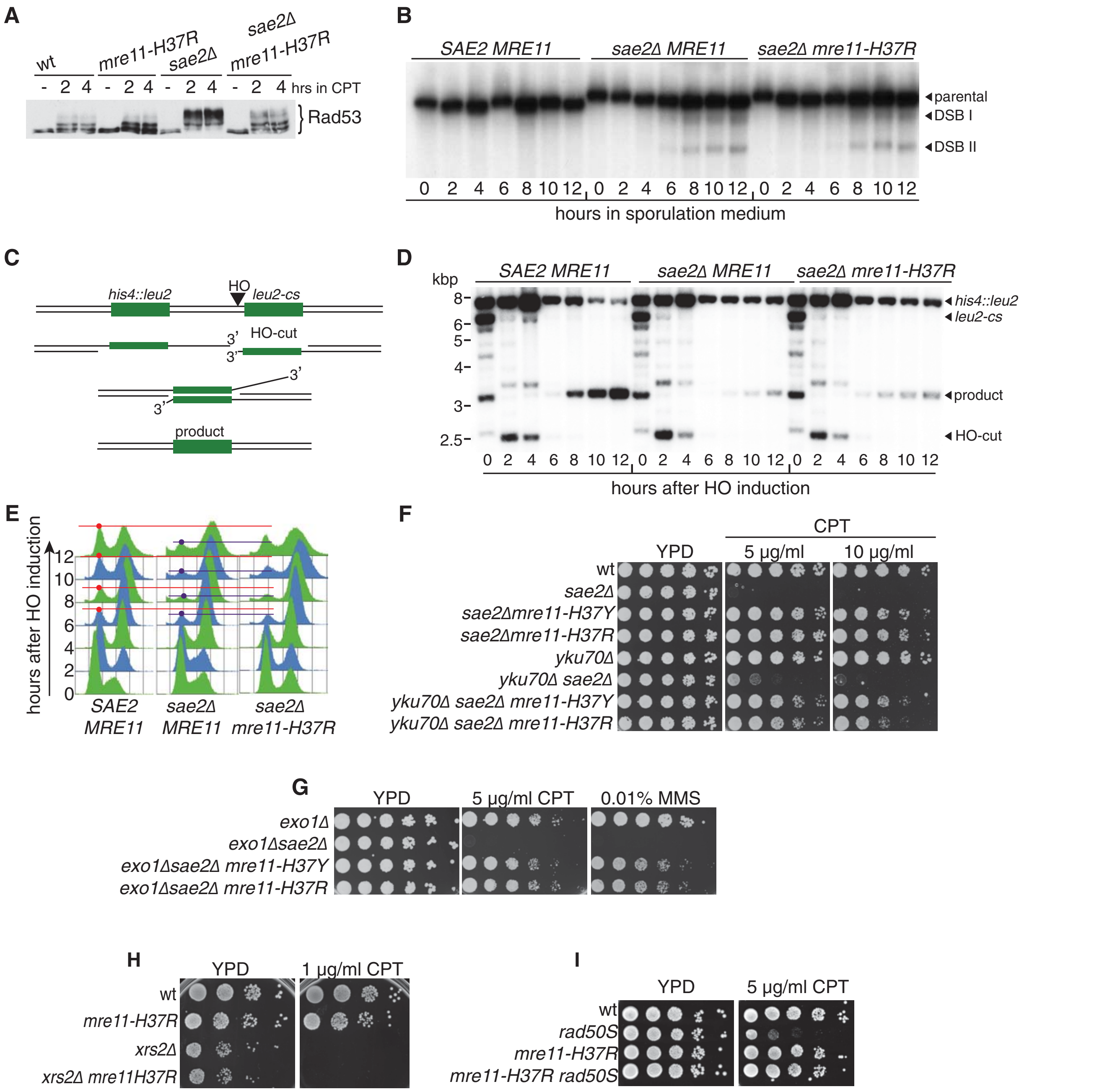


Figure 3. *mre11-H37R* suppresses some but not all *sae2*Δ phenotypes.

A *mre11-H37R* suppresses *sae2*Δ checkpoint hyperactivation.

B *mre11-H37R* does not rescue *sae2*Δ meiotic DSB processing defect.

C Outline of DSB repair by single-strand annealing (SSA).

D *mre11-H37R* does not rescue the SSA repair defect of *sae2*Δ strains.

E *mre11-H37R* does not rescue *sae2*Δ-dependent cell cycle arrest caused by DSB induction.

F, G Exo1 and Ku are not required for *mre11-H37R*-mediated suppression of *sae2*Δ hypersensitivity.

H *mre11-H37R* does not suppress *xrs2*Δ camptothecin (CPT) hypersensitivity.

I *mre11-H37R* suppresses *rad50S* CPT hypersensitivity.

(dsDNA; Fig 4C) and abrogated Mre11 binding to ssDNA (Fig 4D). Conversely, mutation of H37 to alanine, which does not result in a *sup*^{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 *sae2*Δ suppressive effects of *mre11*^{SUPsae2Δ} mutations were associated with weakened Mre11 DNA binding and

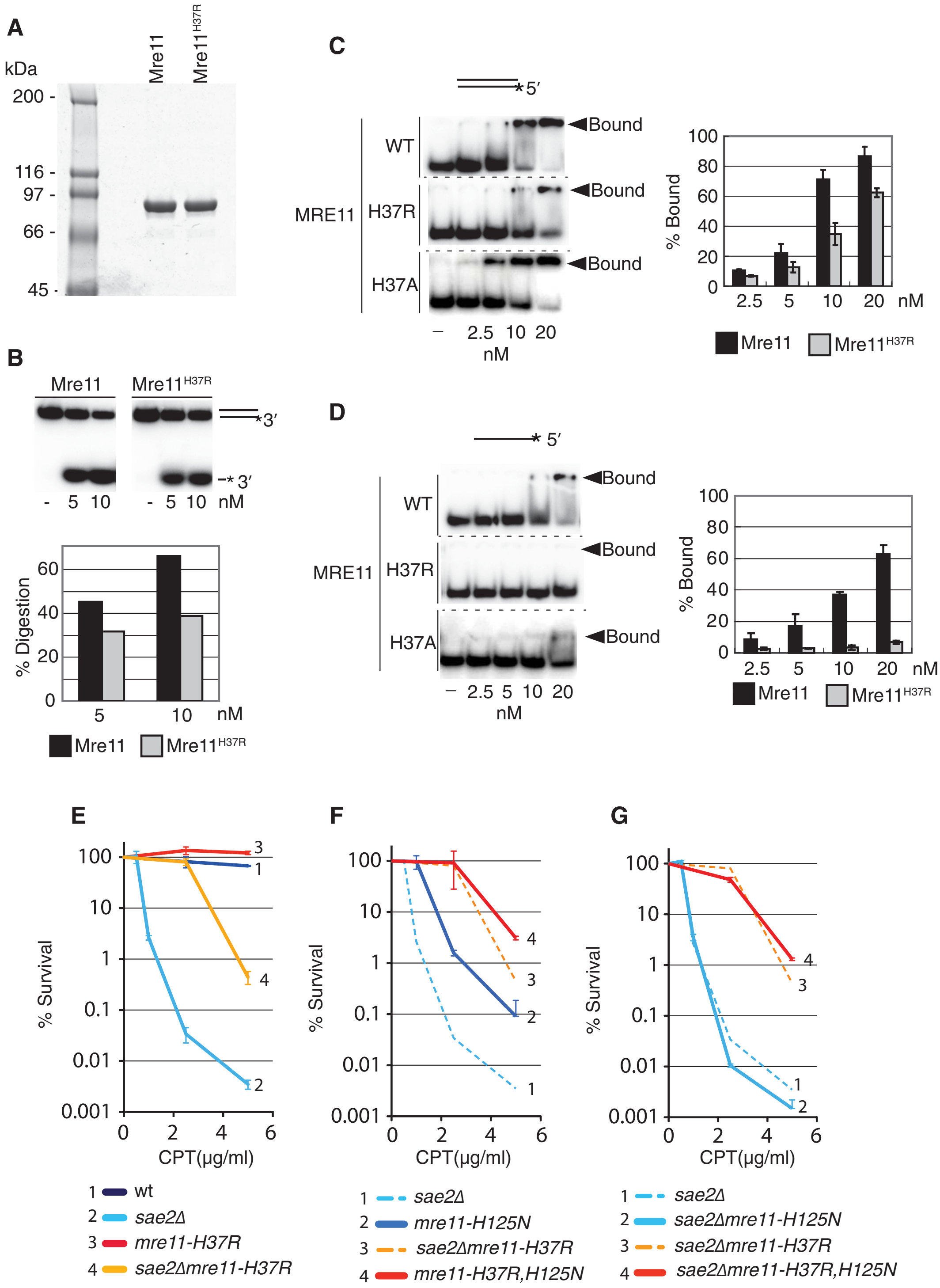


Figure 4. Mre11^{H37R} is impaired biochemically, particularly at the level of ssDNA binding.

A Mre11 and Mre11^{H37R} were purified to homogeneity from yeast cultures.

B 3′ exonuclease activity assay on Mre11 and Mre11^{H37R} leading to release of a labelled single nucleotide, as indicated.

C, D Electrophoretic mobility shift assays on Mre11, Mre11^{H37R} and Mre11^{H37A} 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.

Published online: April 21, 2015

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 *sae2*Δ *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 Sae2-deficient setting (Fig 4G and Supplementary Fig S2G). Furthermore, while *sae2*Δ 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 *sae2*Δ-induced lack of Mre11 nuclease activity, but also other nuclease-independent functions of Sae2. Nevertheless, *mre11-H37R* did not rescue the camptothecin hypersensitivity of *sae2*Δ cells to wild-type levels, suggesting that not all functions of Sae2 are suppressed by this *MRE11* allele (Fig 4E and F).

Identifying an Mre11 interface mediating *sae2*Δ suppression

To gain further insights into how *mre11*$^{SUPsae2\Delta}$ alleles operate and relate this to the above functional and biochemical data, we screened for additional *MRE11* mutations that could suppress camptothecin hypersensitivity caused by Sae2 loss. Thus, we propagated a plasmid carrying wild-type *MRE11* in a mutagenic *E. coli* strain, thereby generating libraries of plasmids carrying *mre11* mutations. We then introduced these libraries into a *sae2*Δ*mre11*Δ 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 *sae2*Δ 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 *sae2*Δ 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 *sae2*Δ 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 *mre11*$^{SUPsae2\Delta}$ 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 *mre11*$^{SUPsae2\Delta}$ 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 *mre11*$^{SUPsae2\Delta}$ 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 hemi-cylindrical 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 *sae2*Δ camptothecin hypersensitivity (Fig 5G). As discussed further below, while it is possible that certain *mre11*$^{SUPsae2\Delta}$ 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 hemi-cylindrical concave region and ssDNA (modelled in Fig 5G and discussed further below). Consistent with this idea, we found that the Mre11^{Q70R} 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

Figure 5. Identifying additional mutations in *MRE11* that mediate *sae2*Δ suppression.

A Outline of the plasmid mutagenesis approach to identify new *mre11*$^{SUPsae2\Delta}$ alleles. ^{LOF}: loss-of-function alleles. ^{SUP}: suppressor alleles.

B Mre11 with shaded boxes and blue shapes indicating phosphoesterase motifs and secondary structures, respectively; additional *mre11*$^{SUPsae2\Delta}$ mutations recovered from the screening are indicated.

C Fungal alignment and secondary structure prediction of the region of Mre11 containing Q70.

D *mre11-Q70R*, *mre11-L89V* and *mre11-P110L* alleles recovered from plasmid mutagenesis screening suppress *sae2*Δ hypersensitivity to camptothecin.

E Structural prediction of *S. cerevisiae* Mre11 residues 1–414, obtained by homology modelling using the corresponding *S. pombe* and human structures. The water-accessible surface of the two monomers is shown in different shades of blue. Red: residues whose mutation suppresses *sae2*Δ DNA damage hypersensitivity. Orange: residues whose mutation abrogates Mre11 nuclease activity.

F Model of Mre11 tertiary structure (residues 1–100). Residues are colour-coded as in (E).

G Top: *mre11-L77R* suppresses the DNA damage hypersensitivity of *sae2*Δ cells. Bottom: localisation of *mre11*$^{SUPsae2\Delta}$ suppressors on the molecular model of the Mre11 dimer. The two Mre11 monomers are shown in different shades of blue, and the proposed path of bound ssDNA is indicated by the orange filament.

H Model in which the two DNA filaments of the two DSB ends melt when binding to Mre11; the 5′ ends being channelled towards the active site and the 3′ end being channelled towards the Mre11$^{SUPsae2\Delta}$ region.

▶

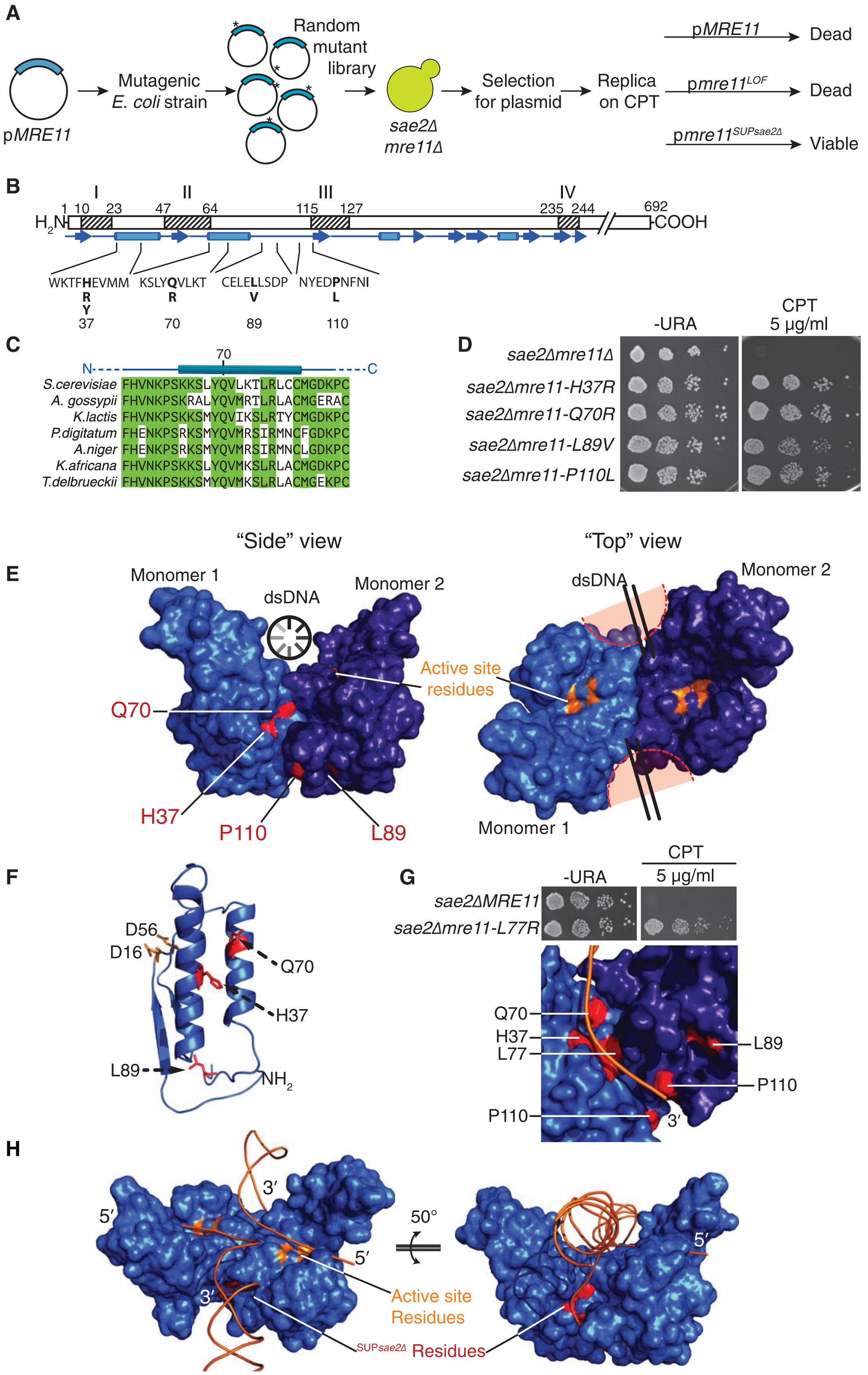

A
pMRE11
Mutagenic
E. coli strain
Random
mutant
library
sae2Δ
mre11Δ
Selection
for plasmid
Replica
on CPT
pMRE11
Dead
pmre11LOF
Dead
pmre11SUPsae2Δ
Viable
B
I
II
III
IV
1
10
23
47
64
115
127
235
244
692
H2N
COOH
WKTFHEVMM
R
Y
37
KSLYQVLKT
R
70
CELELLSDP
V
89
NYEDPNFNI
L
110
C
70
N
C
S.cerevisiae FHVNKPSKKSLYQVLKTLRLCCMGDKPC
A. gossypii FHVNKPSKRALYQVMRTLRLACMGERAC
K.lactis FHVNKPSKKSMYQVIKSLRTYCMGDKPC
P.digitatum FHENKPSRKSMYQVMRSIRMNCFGDKPC
A.niger FHENKPSRKSMYQVMRSIRMNCLGDKPC
K.africana FHVNKPSKKSLYQVMKSLRLACMGDKPC
T.delbrueckii FHVNKPSKKSMYQVMKSLRLACMGEKPC
D
-URA
CPT
5 μg/ml
sae2Δmre11Δ
sae2Δmre11-H37R
sae2Δmre11-Q70R
sae2Δmre11-L89V
sae2Δmre11-P110L
E
"Side" view
"Top" view
Monomer 1
dsDNA
Monomer 2
Active site
residues
Q70
H37
P110
L89
dsDNA
Monomer 2
Monomer 1
F
D56
D16
Q70
H37
L89
NH2
G
-URA
CPT
5 μg/ml
sae2ΔMRE11
sae2Δmre11-L77R
Q70
H37
L77
P110
L89
P110
3′
H
5′
3′
5′
3′
50°
5′
Active site
Residues
SUPsae2Δ Residues

Figure 5.

Published online: April 21, 2015

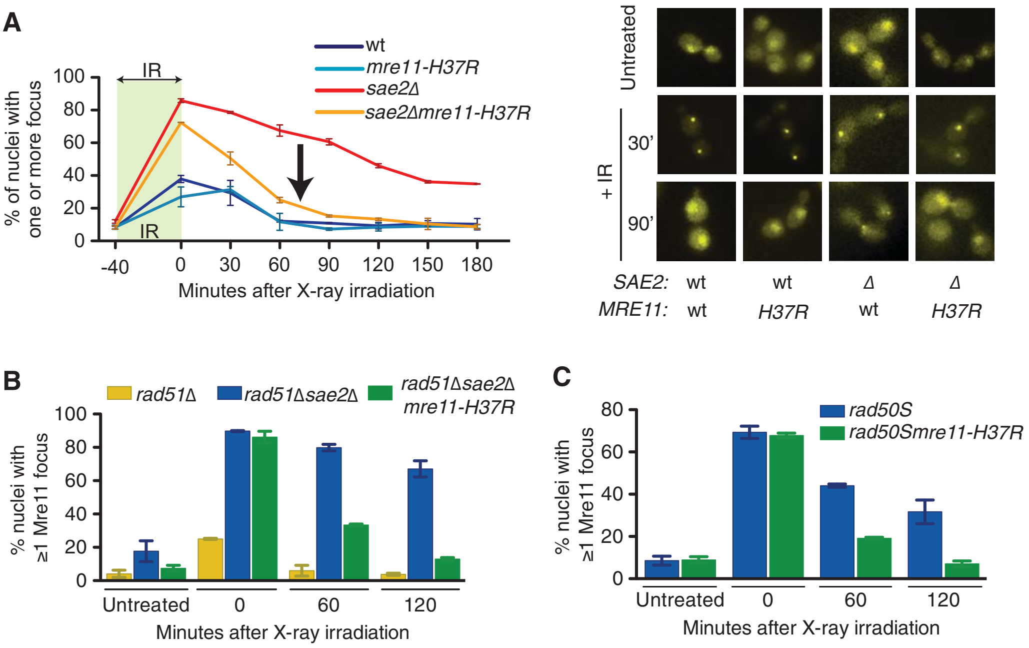


Figure 6. ***mre11$^{SUPsae2\Delta}$*** **alleles bypass the need for Sae2 to remove Mre11 from DSB ends.**

A IR-induced Mre11^{H37R} foci (IRIF) persist for shorter times than Mre11-wt IRIF in exponentially growing *sae2Δ* cells (average and standard deviations from two or more independent experiments).

B Effects of *sae2Δ* and *mre11-H37R* on Mre11 IRIF persistence still occur when Rad51 is absent, revealing that Mre11 IRIF persistence causes defective HR (average and standard deviation from two independent experiments).

C *mre11-H37R* suppresses Mre11 IRIF persistence in exponentially growing *rad50S* cells (average and standard deviation from two independent experiments).

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′ end being channelled towards the nuclease catalytic site and the resulting ssDNA—bearing the 3′ terminal OH—interacting with an adjacent Mre11 region that contains residues mutated in *mre11$^{SUPsae2\Delta}$* alleles (Fig 5G and H). In this regard, we note that Mre11 was recently shown in biochemical studies to promote local DNA unwinding (Cannon *et al*, 2013). Such a model would explain our biochemical findings, and would also explain our biological data if persistent Mre11 binding to the nascent 3′ terminal DNA impairs HR unless counteracted by the actions of Sae2 or weakened by *mre11$^{SUPsae2\Delta}$* 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 *mre11$^{SUPsae2\Delta}$* mutations. To test this, we constructed yeast strains expressing wild-type Mre11 or Mre11^{H37R} 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 *rad50S* 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, Mre11^{H37R} 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

Published online: April 21, 2015

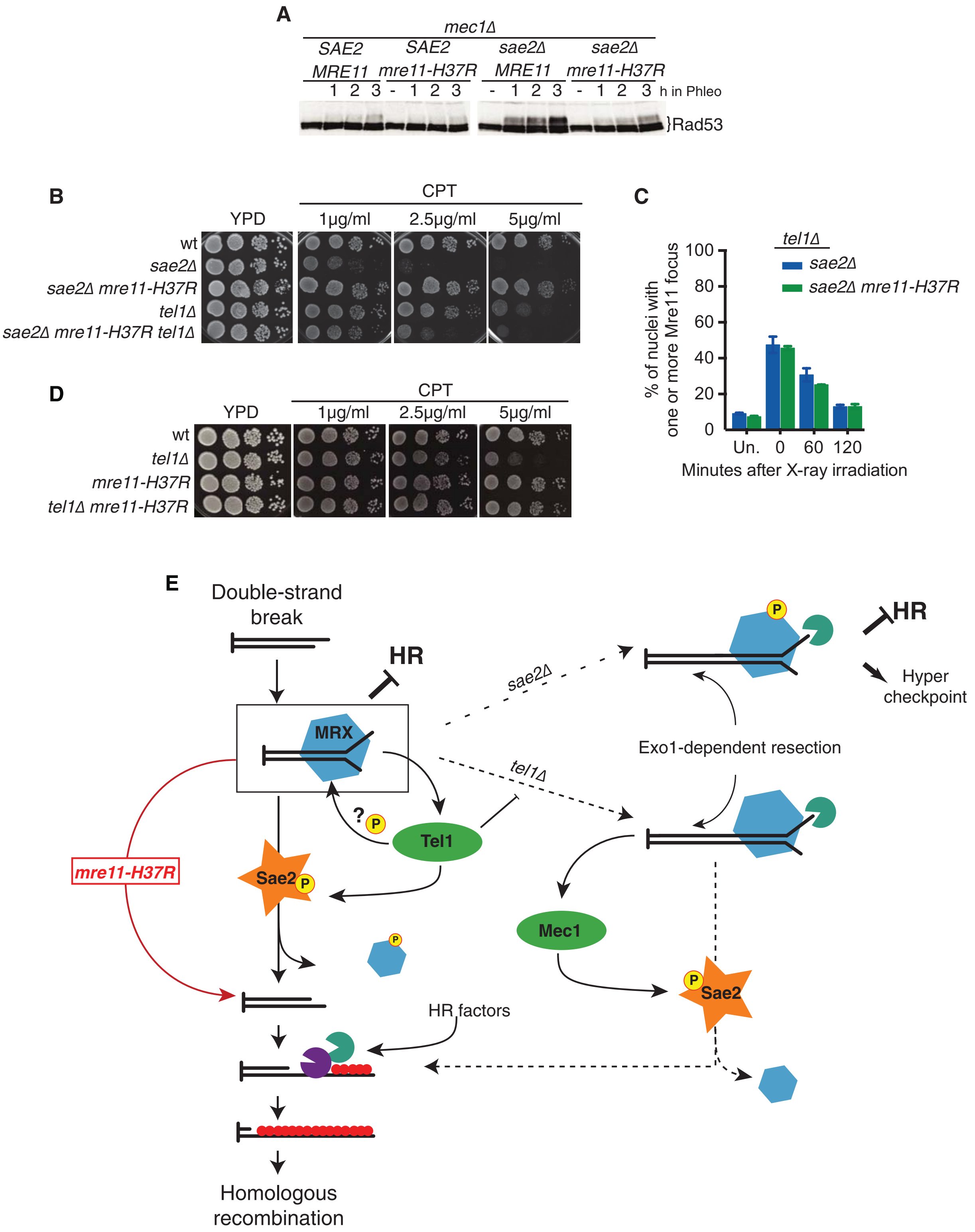


Figure 7. Tel1 participates in regulating Mre11 dynamics after DNA damage.

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.

Published online: April 21, 2015

camptothecin hypersensitivity of *sae2*Δ cells (Fig 7B). In accord with this, in the absence of Tel1, *mre11-H37R* no longer affected the dissociation kinetics of IR-induced Mre11 foci in *sae2*Δ cells (Fig 7C). Collectively, these data suggested that Tel1 functionally cooperates with Sae2 to promote the removal of Mre11 from DNA ends. In this regard, we noted that *mre11-H37R* suppressed the moderate camptothecin hypersensitivity of a *tel1*Δ strain (Fig 7D). We therefore propose that, while persistent DNA damage-induced Tel1 activation is certainly a key feature of *sae2*Δ cells, it is persistent binding of the MRX complex to nascent 3′ terminal DNA that causes toxicity in *sae2*Δ cells, likely through it delaying downstream HR events. Accordingly, mutations that reduce Mre11 ssDNA binding enhance the release of the Mre11 complex from DSB ends in the absence of Sae2, through events promoted by Tel1 (Fig 7E). In this model, Mre11 persistence at DNA damage sites is a cause, and not just a consequence, of impaired HR-mediated repair in sae2Δ cells.

Discussion

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 & Cejka, 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 *sae2*Δ 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 eviction 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 checkpoint hyperactivation contributes to the resistance of *sae2*Δ*mre11-H37R* cells to DNA-damaging agents. In this regard, we note that the site of one of the *sae2*Δ 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, *sae2*Δ 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, rad50S* 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 *sae2*Δ cells compared to cells carrying *mre11-H125N* alleles: while *sae2*Δ 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 *sae2*Δ cells, because Sae2-stimulated Mre11 nuclease activity is crucial for removing Spo11 from meiotic DBS 5′ termini. Finally, this model explains why *mre11-H37R* does not suppress the *sae2*Δ deficiency in DSB repair by SSA because the *sae2*Δ 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 *sae2*Δ and *tel1*Δ 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 *tel1*Δ and *sae2*Δ cells (Clerici *et al*, 2006; Mantiero *et al*, 2007). We suggest that the comparatively mild hypersensitivity of *tel1*Δ strains to camptothecin is due to Tel1 loss allowing DSB repair intermediates to be channelled into a different pathway, in which Exo1-dependent resection (Mantiero *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 *sae2*Δ 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

Published online: April 21, 2015

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 *mre11Δ::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-YPF 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 putative variants and indels from alignments and assigns likelihoods, and BCFtools that performs the variant calling (Li *et al*, 2009). The following parameters were used: for SAMtools (v0.1.19) mpileup -EDS -C50 -m2 -F0.0005 -d 10,000' and for BCFtools (v0.1.19) view '-p 0.99 -vcgN'. Functional consequences of the variants were produced by using the Ensembl VEP (McLaren *et al*, 2010).

MRE11 random mutagenesis

Plasmid pRS316 carrying *MRE11* coding sequence under the control of its natural promoter was transformed into mutagenic XL1-Red competent *E. coli* cells (Agilent Technologies) and propagated following the manufacturer's instructions. A plasmid library of ~3,000 independent random mutant clones was transformed into *mre11Δsae2Δ* cells, and transformants were screened for their ability to survive in the presence of camptothecin. Plasmids extracted from survivors loosing their camptothecin resistance after a passage on 5-fluoro-orotic acid (FOA) were sequenced and independently reintroduced in a *mre11Δsae2Δ* strain.

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 Mre11^{H37R} 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 nuclease assay, 1 mM $MnCl_2$ 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 proteinase 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/A14492) and the Wellcome Trust (WT092096). SPJ receives his salary from the University of Cambridge, UK, supplemented by CRUK. TO, IG 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 (101126/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 RO1ES007061 and K99ES021441, 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 *in vivo* characterisation of the *mre11-H37R* mutant were conceived by TO, IG, FP and SPJ, and were carried out by TO, FP, IG, NJG, EV and IS. Biochemical assays were designed by SPJ, PS and HN and carried out by HN. The identification of further *mre11*$^{supsae2\Delta}$ 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

Andres SN, Appel CD, Westmoreland JW, Williams JS, Nguyen Y, Robertson PD, Resnick MA, Williams RS (2015) Tetrameric Ctp1 coordinates DNA binding and DNA bridging in DNA double-strand-break repair. *Nat Struct Mol Biol* 22: 158 – 166

Baroni E, Viscardi V, Cartagena-Lirola H, Lucchini G, Longhese MP (2004) The functions of budding yeast Sae2 in the DNA damage response require Mec1- and Tel1-dependent phosphorylation. *Mol Cell Biol* 24: 4151 – 4165

Barton O, Naumann SC, Diemer-Biehs R, Künzel J, Steinlage M, Conrad S, Makharashvili N, Wang J, Feng L, Lopez BS, Paull TT, Chen J, Jeggo PA, Löbrich M (2014) Polo-like kinase 3 regulates CtIP during DNA double-strand break repair in G1. *J Cell Biol* 206: 877 – 894

Benkert P, Tosatto SCE, Schomburg D (2008) QMEAN: a comprehensive scoring function for model quality assessment. *Proteins* 71: 261 – 277

Benkert P, Künzli M, Schwede T (2009) QMEAN server for protein model quality estimation. *Nucleic Acids Res* 37: W510 – W514

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

Cannon B, Kuhnlein J, Yang S-H, Cheng A, Schindler D, Stark JM, Russell R, Paull TT (2013) Visualization of local DNA unwinding by Mre11/Rad50/Nbs1 using single-molecule FRET. *Proc Natl Acad Sci USA* 110: 18868 – 18873

Chen H, Donnianni RA, Handa N, Deng SK, Oh J, Timashev LA, Kowalczykowski SC, Symington LS (2015) Sae2 promotes DNA damage resistance by removing the Mre11–Rad50–Xrs2 complex from DNA and attenuating Rad53 signaling. *Proc Natl Acad Sci USA* 112: E1880 – E1887

Clerici M, Mantiero D, Lucchini G, Longhese MP (2005) The *Saccharomyces cerevisiae* Sae2 protein promotes resection and bridging of double strand break ends. *J Biol Chem* 280: 38631 – 38638

Clerici M, Mantiero D, Lucchini G, Longhese MP (2006) The *Saccharomyces cerevisiae* Sae2 protein negatively regulates DNA damage checkpoint signalling. *EMBO Rep* 7: 212 – 218

Daley JM, Palmbos PL, Wu D, Wilson TE (2005) Nonhomologous end joining in yeast. *Annu Rev Genet* 39: 431 – 451

Davies OR, Forment JV, Sun M, Belotserkovskaya R, Coates J, Galanty Y, Demir M, Morton CR, Rzechorzek NJ, Jackson SP, Pellegrini L (2015) CtIP tetramer assembly is required for DNA-end resection and repair. *Nat Struct Mol Biol* 22: 150 – 157

Deng C, Brown JA, You D, Brown JM (2005) Multiple endonucleases function to repair covalent topoisomerase I complexes in *Saccharomyces cerevisiae*. *Genetics* 170: 591 – 600

Fishman-Lobell J, Rudin N, Haber JE (1992) Two alternative pathways of double-strand break repair that are kinetically separable and independently modulated. *Mol Cell Biol* 12: 1292 – 1303

Foster SS, Balestrini A, Petrini JHJ (2011) Functional interplay of the Mre11 nuclease and Ku in the response to replication-associated DNA damage. *Mol Cell Biol* 31: 4379 – 4389

Fukunaga K, Kwon Y, Sung P, Sugimoto K (2011) Activation of protein kinase Tel1 through recognition of protein-bound DNA ends. *Mol Cell Biol* 31: 1959 – 1971

Furuse M, Nagase Y, Tsubouchi H, Murakami-Murofushi K, Shibata T, Ohta K (1998) Distinct roles of two separable in vitro activities of yeast Mre11 in mitotic and meiotic recombination. *EMBO J* 17: 6412 – 6425

Goldway M, Sherman A, Zenvirth D, Arbel T, Simchen G (1993) A short chromosomal region with major roles in yeast chromosome III meiotic disjunction, recombination and double strand breaks. *Genetics* 133: 159 – 169

Hardy J, Churikov D, Géli V, Simon M-N (2014) Sgs1 and Sae2 promote telomere replication by limiting accumulation of ssDNA. *Nat Commun* 5: 5004

Hartsuiker E, Neale MJ, Carr AM (2009) Distinct requirements for the Rad32Mre11 nuclease and Ctp1CtIP in the removal of covalently bound topoisomerase I and II from DNA. *Mol Cell* 33: 117 – 123

Hopfner KP, Karcher A, Shin DS, Craig L, Arthur LM, Carney JP, Tainer JA (2000) Structural biology of Rad50 ATPase: ATP-driven conformational control in DNA double-strand break repair and the ABC-ATPase superfamily. *Cell* 101: 789 – 800

Huertas P, Cortés-Ledesma F, Sartori AA, Aguilera A, Jackson SP (2008) CDK targets Sae2 to control DNA-end resection and homologous recombination. *Nature* 455: 689 – 692

Huertas P, Jackson SP (2009) Human CtIP mediates cell cycle control of DNA end resection and double strand break repair. *J Biol Chem* 284: 9558 – 9565

Ivanov EL, Sugawara N, White CI, Fabre F, Haber JE (1994) Mutations in XRS2 and RAD50 delay but do not prevent mating-type switching in *Saccharomyces cerevisiae*. *Mol Cell Biol* 14: 3414 – 3425

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

Keeney S, Kleckner N (1995) Covalent protein-DNA complexes at the 5′ strand termini of meiosis-specific double-strand breaks in yeast. *Proc Natl Acad Sci USA* 92: 11274 – 11278

Keeney S, Giroux CN, Kleckner N (1997) Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell* 88: 375–384

Krogh BO, Llorente B, Lam A, Symington LS (2005) Mutations in Mre11 phosphoesterase motif I that impair *Saccharomyces cerevisiae* Mre11-Rad50-Xrs2 complex stability in addition to nuclease activity. *Genetics* 171: 1561–1570

Lengsfeld BM, Rattray AJ, Bhaskara V, Ghirlando R, Paull TT (2007) Sae2 is an endonuclease that processes hairpin DNA cooperatively with the Mre11/Rad50/Xrs2 complex. *Mol Cell* 28: 638–651

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079

Lisby M, Barlow JH, Burgess RC, Rothstein R (2004) Choreography of the DNA damage response: spatiotemporal relationships among checkpoint and repair proteins. *Cell* 118: 699–713

Makharashvili N, Tubbs AT, Yang S-H, Wang H, Barton O, Zhou Y, Deshpande RA, Lee J-H, Lobrich M, Sleckman BP, Wu X, Paull TT (2014) Catalytic and noncatalytic roles of the CtIP endonuclease in double-strand break end resection. *Mol Cell* 54: 1022–1033

Mantiero D, Clerici M, Lucchini G, Longhese MP (2007) Dual role for *Saccharomyces cerevisiae* Tel1 in the checkpoint response to double-strand breaks. *EMBO Rep* 8: 380–387

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. *Bioinformatics* 26: 2069–2070

Mimitou EP, Symington LS (2008) Sae2, Exo1 and Sgs1 collaborate in DNA double-strand break processing. *Nature* 455: 770–774

Mimitou EP, Symington LS (2010) Ku prevents Exo1 and Sgs1-dependent resection of DNA ends in the absence of a functional MRX complex or Sae2. *EMBO J* 29: 3358–3369

Moreau S, Ferguson JR, Symington LS (1999) The nuclease activity of Mre11 is required for meiosis but not for mating type switching, end joining, or telomere maintenance. *Mol Cell Biol* 19: 556–566

Moreau S, Morgan EAA, Symington LSS (2001) Overlapping functions of the *Saccharomyces cerevisiae* mre11, exo1 and rad27 nucleases in DNA metabolism. *Genetics* 159: 1423

Nairz K, Klein F (1997) mre11S–-a yeast mutation that blocks double-strand-break processing and permits nonhomologous synapsis in meiosis. *Genes Dev* 11: 2272–2290

Prinz S, Amon A, Klein F (1997) Isolation of COM1, a new gene required to complete meiotic double-strand break-induced recombination in *Saccharomyces cerevisiae*. *Genetics* 146: 781–795

Sali A, Blundell TL (1990) Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol* 212: 403–428

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234: 779–815

Sanchez Y, Desany BA, Jones WJ, Liu Q, Wang B, Elledge SJ (1996) Regulation of RAD53 by the ATM-like kinases MEC1 and TEL1 in yeast cell cycle checkpoint pathways. *Science* 271: 357–360

Sartori AA, Lukas C, Coates J, Mistrik M, Fu S, Bartek J, Baer R, Lukas J, Jackson SP (2007) Human CtIP promotes DNA end resection. *Nature* 450: 509–514

Schiller CB, Lammens K, Guerini I, Coordes B, Feldmann H, Schlauderer F, Möckel C, Schele A, Strässer K, Jackson SP, Hopfner K-P (2012) Structure of Mre11-Nbs1 complex yields insights into ataxia-telangiectasia-like disease mutations and DNA damage signaling. *Nat Struct Mol Biol* 19: 693–700

Stracker TH, Petrini JHJ (2011) The MRE11 complex: starting from the ends. *Nat Rev Mol Cell Biol* 12: 90–103

Symington LS, Gautier J (2011) Double-strand break end resection and repair pathway choice. *Annu Rev Genet* 45: 247–271

Usui T, Ogawa H, Petrini JH (2001) A DNA damage response pathway controlled by Tel1 and the Mre11 complex. *Mol Cell* 7: 1255–1266

Vaze MB, Pellicioli A, Lee SE, Ira G, Liberi G, Arbel-Eden A, Foiani M, Haber JE (2002) Recovery from checkpoint-mediated arrest after repair of a double-strand break requires Srs2 helicase. *Mol Cell* 10: 373–385

Wang H, Li Y, Truong LN, Shi LZ, Hwang PY-H, He J, Do J, Cho MJ, Li H, Negrete A, Shiloach J, Berns MW, Shen B, Chen L, Wu X (2014) CtIP maintains stability at common fragile sites and inverted repeats by end resection-independent endonuclease activity. *Mol Cell* 54: 1012–1021

Williams RS, Moncalian G, Williams JS, Yamada Y, Limbo O, Shin DS, Groocock LM, Cahill D, Hitomi C, Guenther G, Moiani D, Carney JP, Russell P, Tainer JA (2008) Mre11 dimers coordinate DNA end bridging and nuclease processing in double-strand-break repair. *Cell* 135: 97–109

You Z, Shi LZ, Zhu Q, Wu P, Zhang Y-W, Basilio A, Tonnu N, Verma IM, Berns MW, Hunter T (2009) CtIP links DNA double-strand break sensing to resection. *Mol Cell* 36: 954–969

Zhu Z, Chung W-H, Shim EY, Lee SE, Ira G (2008) Sgs1 helicase and two nucleases Dna2 and Exo1 resect DNA double-strand break ends. *Cell* 134: 981–994

Genome-wide genetic screening with chemically-mutagenized haploid embryonic stem cells

Josep V. Forment[1,2], Mareike Herzog[1,2], Julia Coates[1], Tomasz Konopka[3], Bianca V. Gapp[3], Sebastian M. Nijman[3,4], David J. Adams[2], Thomas M. Keane[2] and Stephen P. Jackson[1,2]

[1]The Wellcome Trust CRUK Gurdon Institute and Department of Biochemistry, University of Cambridge, Cambridge, UK

[2]The Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

[3]Ludwig Institute for Cancer Research Ltd. and Target Discovery Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK

[4]Research 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[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[14], as does mutations in DNA methyltransferase Dnmt1[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 variants[16]) and in genes involved in DNA MMR (linked to the inherited colon cancer predisposition Lynch syndrome[17]) 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

1. Forsburg, S. L. The art and design of genetic screens: yeast. *Nat Rev Genet* **2,** 659–668 (2001).
2. St Johnston, D. The art and design of genetic screens: Drosophila melanogaster. *Nat Rev Genet* **3,** 176–188 (2002).
3. Jorgensen, E. M. & Mango, S. E. The art and design of genetic screens: Caenorhabditis elegans. *Nat Rev Genet* **3,** 356–369 (2002).
4. Boutros, M. & Ahringer, J. The art and design of genetic screens: RNA interference. *Nat Rev Genet* **9,** 554–566 (2008).
5. Carette, J. E. *et al.* Haploid genetic screens in human cells identify host factors used by pathogens. *Science* **326,** 1231–1235 (2009).
6. Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M. D. C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat Biotechnol* **32,** 267–273 (2014).
7. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343,** 84–87 (2014).
8. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343,** 80–84 (2014).
9. Rolef Ben-Shahar, T. *et al.* Eco1-dependent cohesin acetylation during establishment of sister chromatid cohesion. *Science* **321,** 563–566 (2008).
10. Puddu, F. *et al.* Synthetic viability genomic screening defines Sae2 function in DNA repair. *Embo J* **34,** 1509–1522 (2015).
11. Leeb, M. & Wutz, A. Derivation of haploid embryonic stem cells from mouse embryos. *Nature* **479,** 131–134 (2011).
12. Munroe, R. R. & Schimenti, J. J. Mutagenesis of mouse embryonic stem cells with ethylmethanesulfonate. *Methods Mol. Biol.* **530,** 131–138 (2009).
13. LePage, G. A. & Jones, M. Purinethiols as feedback inhibitors of purine synthesis in ascites tumor cells. *Cancer Res* **21,** 642–649 (1961).
14. Swann, P. F. *et al.* Role of postreplicative DNA mismatch repair in the cytotoxic action of thioguanine. *Science* **273,** 1109–1111 (1996).
15. Guo, G., Wang, W. & Bradley, A. Mismatch repair genes identified using genetic screens in Blm-deficient embryonic stem cells. *Nature* **429,** 891–895

(2004).

16. Jinnah, H. A., De Gregorio, L., Harris, J. C., Nyhan, W. L. & O'Neill, J. P. The spectrum of inherited mutations causing HPRT deficiency: 75 new cases and a review of 196 previously reported cases. *Mutat Res* **463,** 309–326 (2000).

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

18. Kasap, C., Elemento, O. & Kapoor, T. M. DrugTargetSeqR: a genomics- and CRISPR-Cas9-based method to analyze drug targets. *Nat Chem Biol* **10,** 626–628 (2014).

19. Smurnyy, Y. *et al.* DNA sequencing and CRISPR-Cas9 gene editing for target validation in mammalian cells. *Nat Chem Biol* **10,** 623–625 (2014).

20. Blomen, V. A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* **350,** 1092–1096 (2015).

21. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350,** 1096–1101 (2015).

22. Sagi, I. *et al.* Derivation and differentiation of haploid human embryonic stem cells. *Nature* (2016). doi:10.1038/nature17408

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 *Eco*RI restriction site to allow screening, are in lower case and highlighted in pink. *Right panel*. *Eco*RI 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

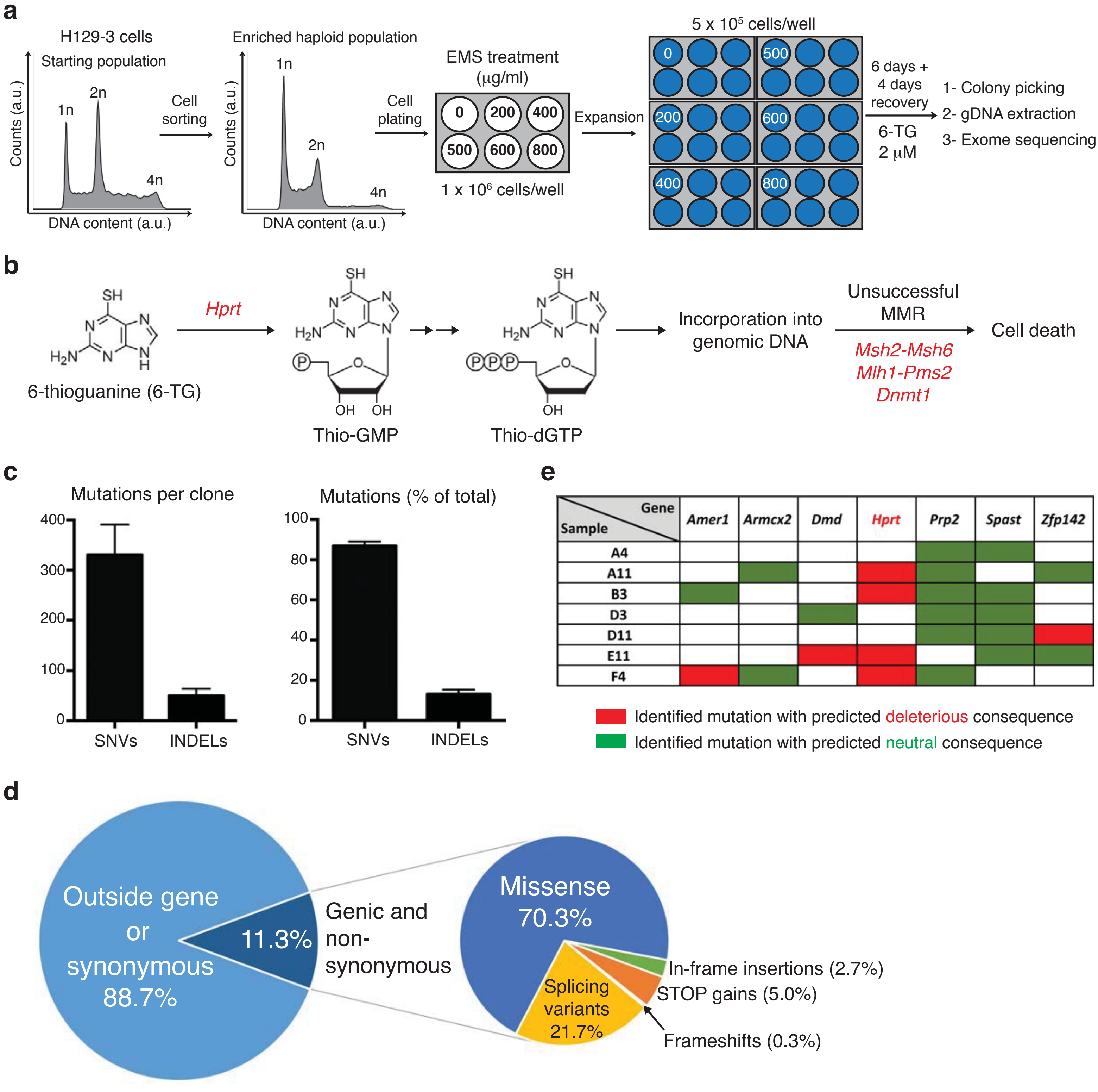
Supplementary Table 1. Homozygous mutations identified through whole-exome sequencing of 7 suppressor clones.

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.

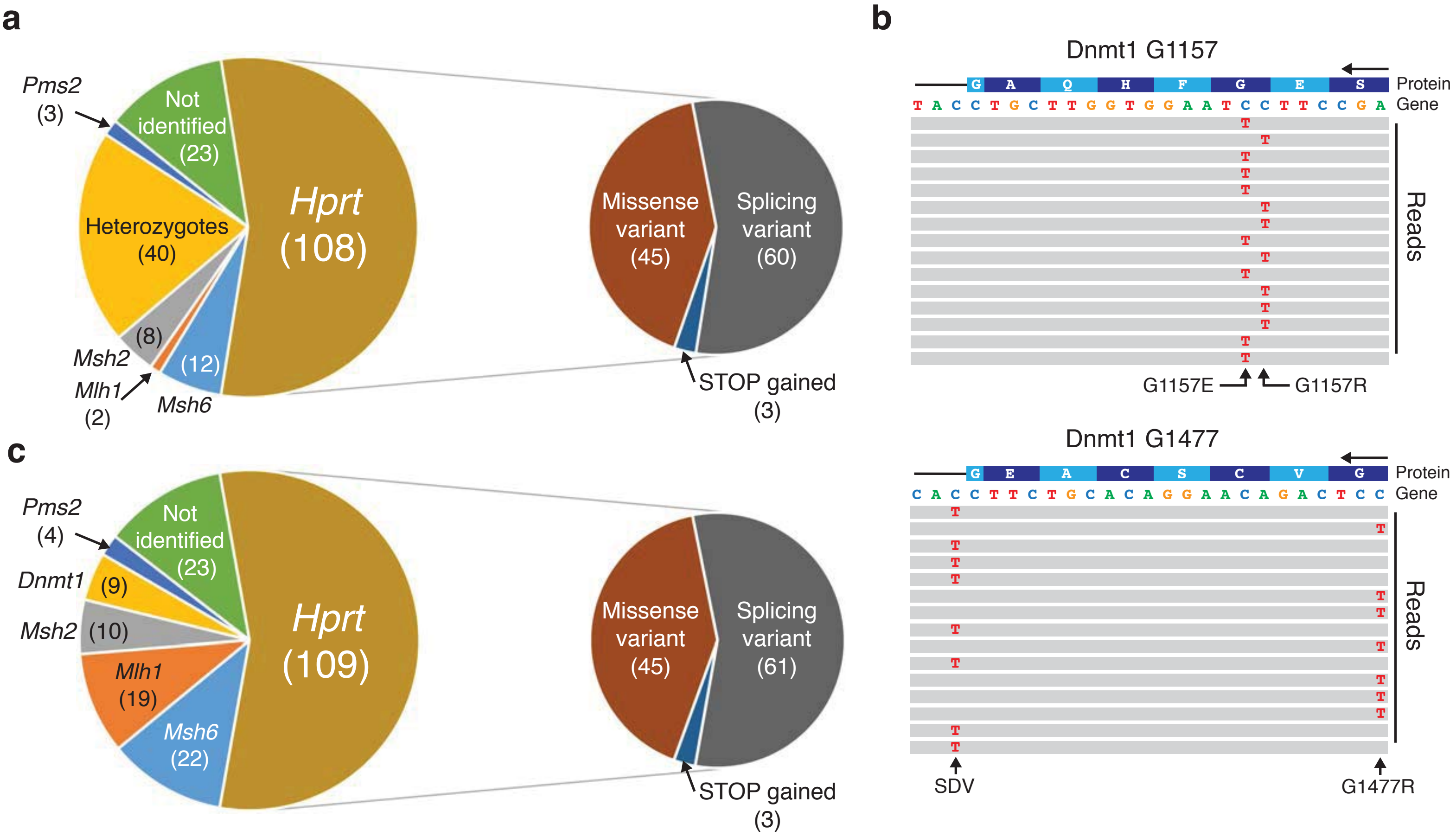


Sample \ Gene	Amer1	Armcx2	Dmd	Hprt	Prp2	Spast	Zfp142
A4					🟩	🟩	
A11		🟩		🟥	🟩		🟩
B3	🟩			🟥	🟩	🟩	
D3			🟩		🟩	🟩	
D11					🟩	🟩	🟥
E11			🟥	🟥		🟩	🟩
F4	🟥	🟩		🟥	🟩		

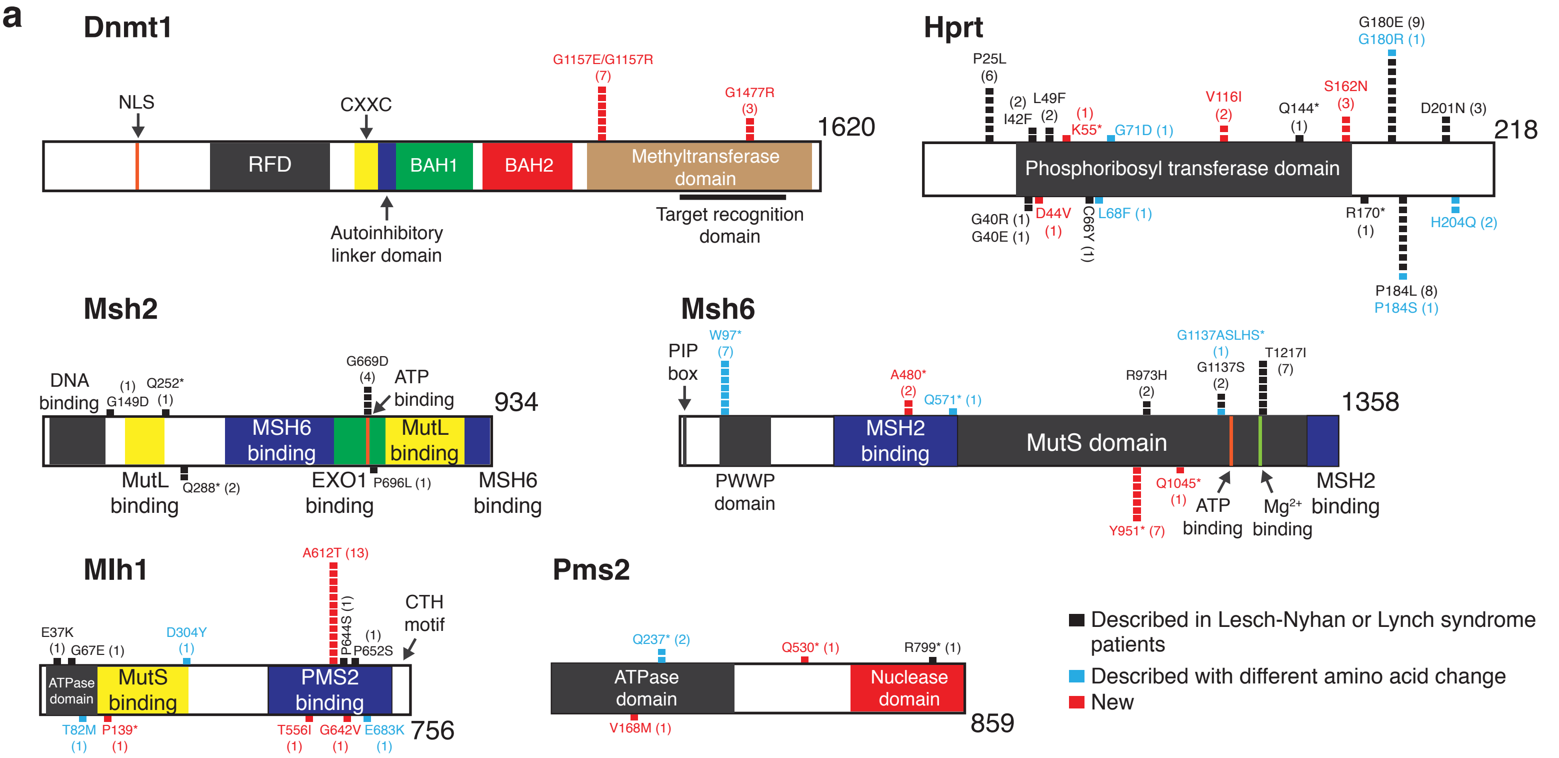
🟥 Identified mutation with predicted deleterious consequence

🟩 Identified mutation with predicted neutral consequence

Forment et al, Figure 1

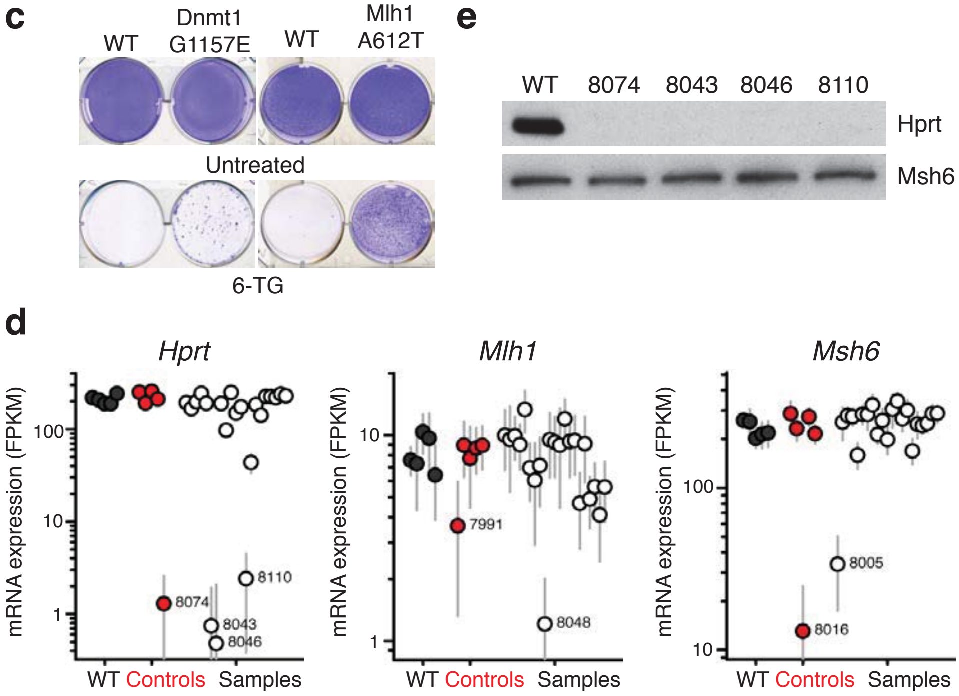


Forment et al, Figure 2

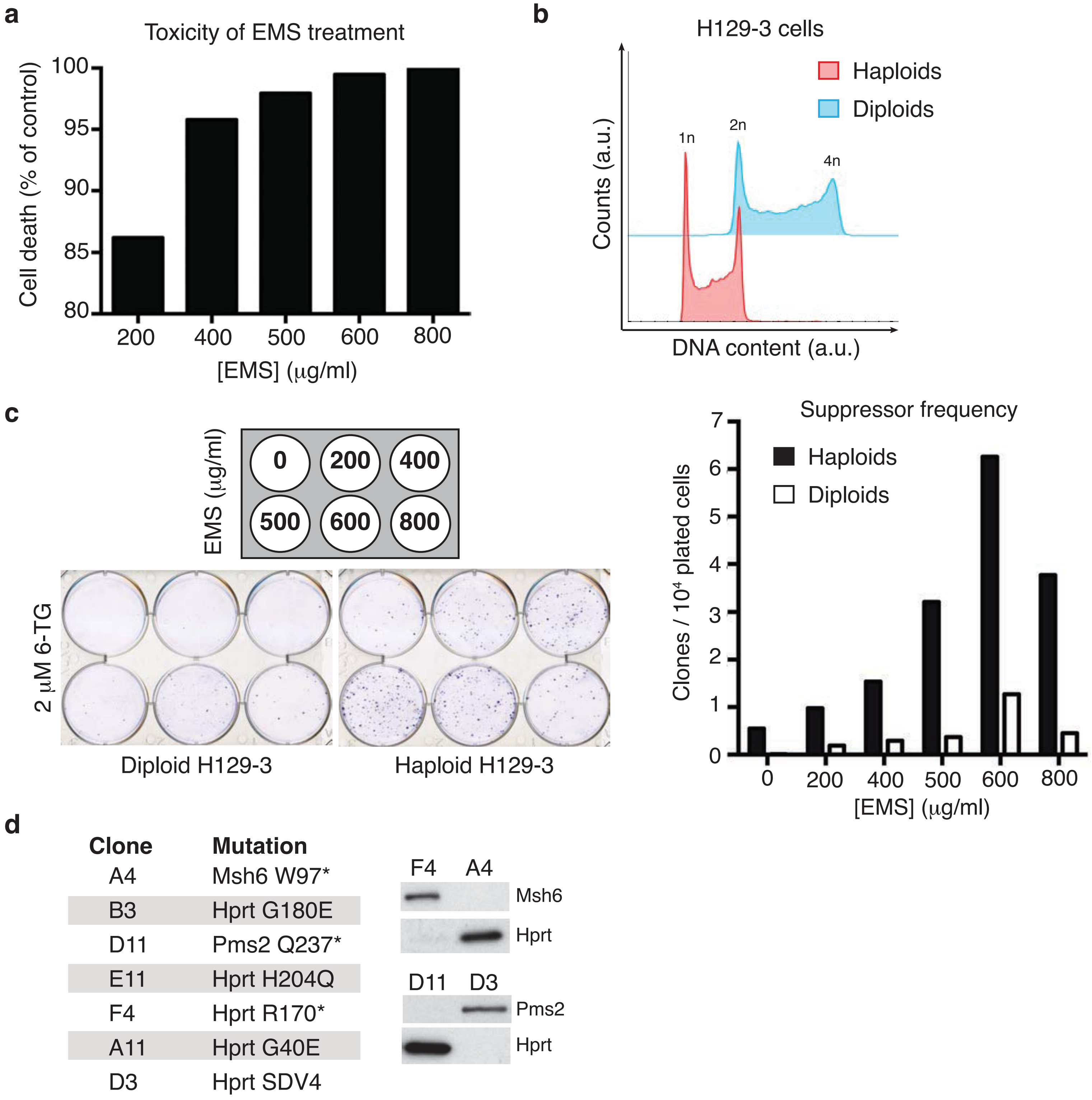


b

		Consequence		
Protein	Mutation	PROVEAN	SIFT	Summary
Dnmt1	G1157E	-7.67	0	Deleterious
	G1477E	-7.67	0	Deleterious
	G1477R	-7.17	0	Deleterious
Hprt	D44V	-7.45	0	Deleterious
	V116I	-0.03	0.841	Neutral
	S162N	-2.18	0.004	Deleterious
Mlh1	T556I	-2.5	0.05	Deleterious
	A612T	-3.06	0.002	Deleterious
	G642V	-3.77	0.017	Deleterious
Pms2	V168M	-2.73	0	Deleterious



Forment et al, Figure 3



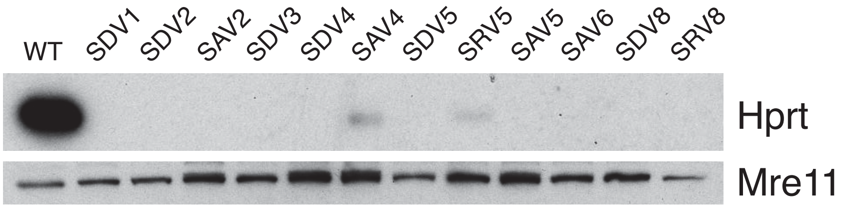
Clone	Mutation
A4	Msh6 W97*
B3	Hprt G180E
D11	Pms2 Q237*
E11	Hprt H204Q
F4	Hprt R170*
A11	Hprt G40E
D3	Hprt SDV4

Forment et al, Supplementary Figure 1

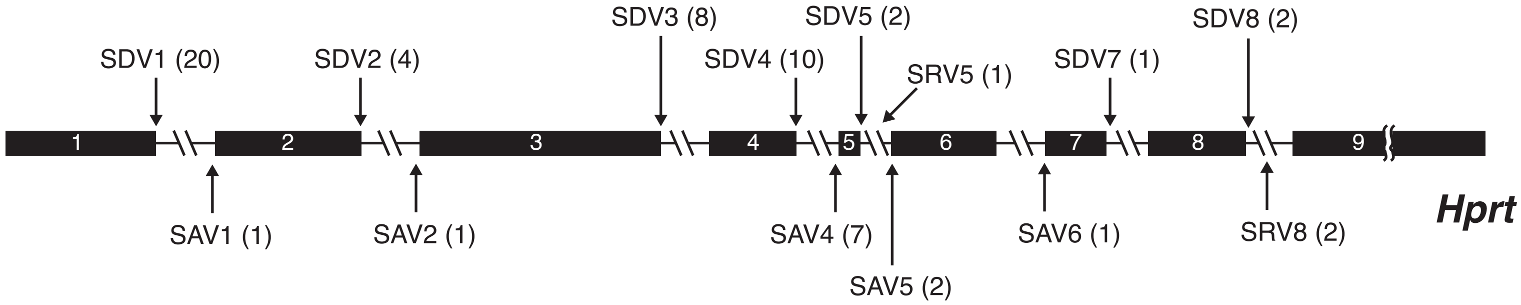
a

Splicing variant	Mutation
SDV1	...GTG-**g**(a)tg...
SAV1	...ca**g**(a)-ATT...
SDV2	...CAG-**g**(a)tt...
SAV2	...ta**g**(a)-GAC...
SDV3	...TGT-**g**(a)ta...
SDV4	...AAG-**g**(a)ta...
SAV4	...ta**g**(a)-AAT...
SDV5	...AAG-**g**(a)ta...
SRV5	...GAA-gtaa**g**(t)...
SAV5	...aa**g**(a)-GAT...
SAV6	...ca**g**(a)-CTT...
SDV7	...ACT-**g**(a)ta...
SDV8	...AAT-**g**(a)ta...
SRV8	...AAT-gtaa**g**(a)...

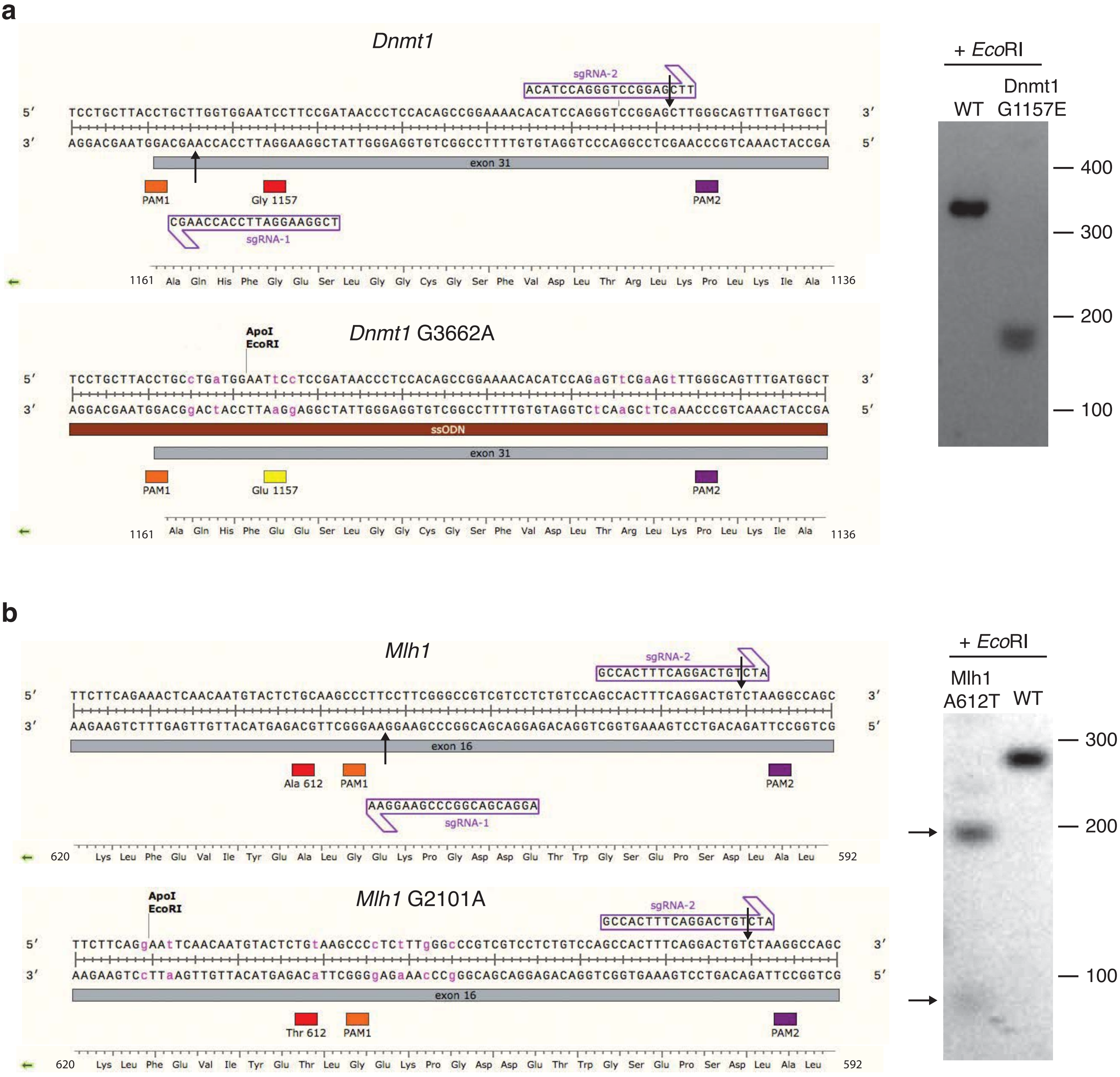
c



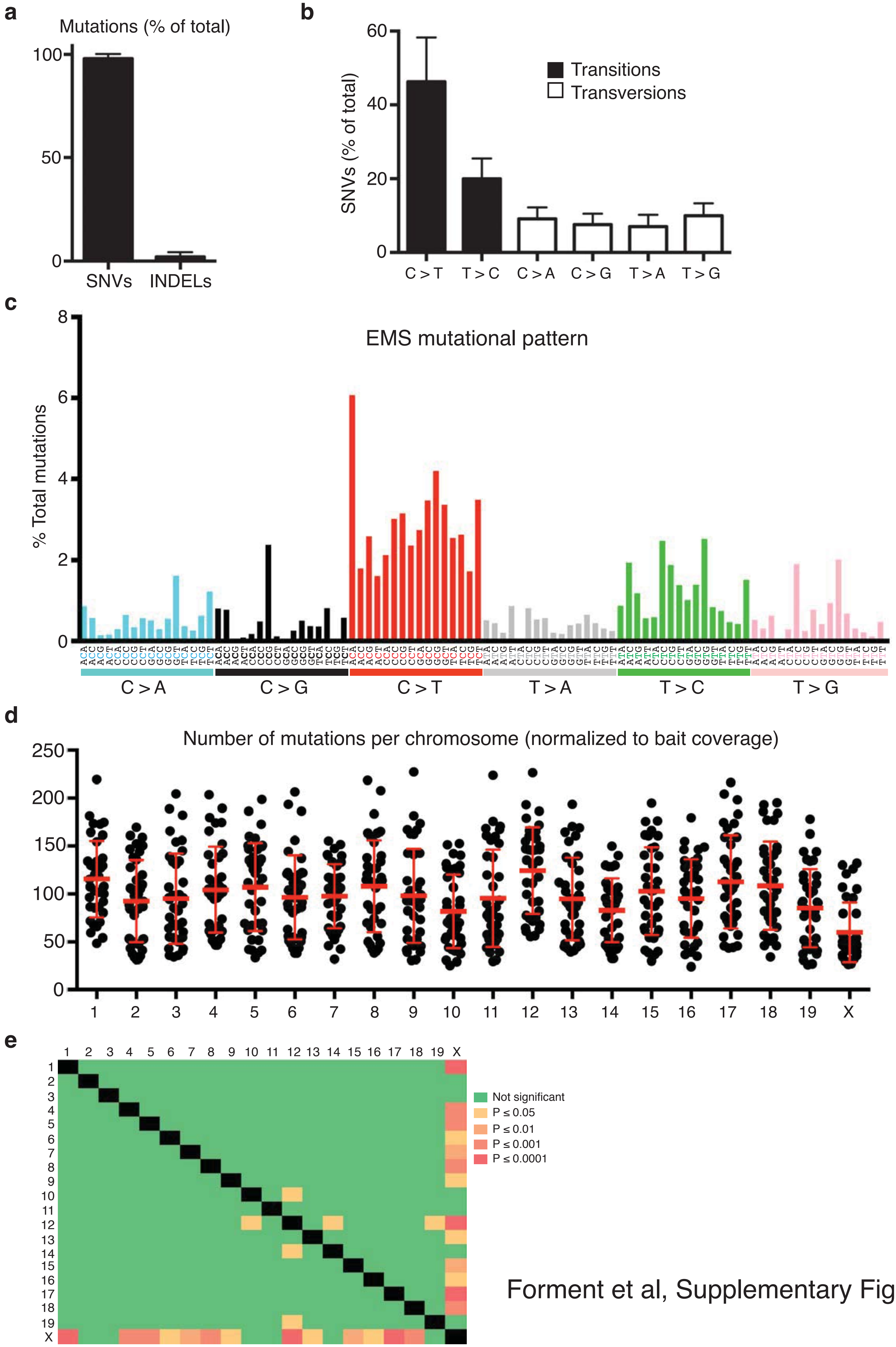
b



Forment et al, Supplementary Figure 2


a
Dnmt1
sgRNA-2
ACATCCAGGGTCCGGAGCTT
5′ TCCTGCTTACCTGCTTGGTGGAATCCTTCCGATAACCCTCCACAGCCGGAAAACACATCCAGGGTCCGGAGCTTGGGCAGTTTGATGGCT 3′
3′ AGGACGAATGGACGAACCACCTTAGGAAGGCTATTGGGAGGTGTCGGCCTTTTGTGTAGGTCCCAGGCCTCGAACCCGTCAAACTACCGA 5′
exon 31
PAM1
Gly 1157
PAM2
CGAACCACCTTAGGAAGGCT
sgRNA-1
1161 Ala Gln His Phe Gly Glu Ser Leu Gly Gly Cys Gly Ser Phe Val Asp Leu Thr Arg Leu Lys Pro Leu Lys Ile Ala 1136
ApoI
EcoRI
Dnmt1 G3662A
5′ TCCTGCTTACCTGCcTGaTGGAATtCcTCCGATAACCCTCCACAGCCGGAAAACACATCCAGaGTtCGaAGtTTGGGCAGTTTGATGGCT 3′
3′ AGGACGAATGGACGgACtACCTTAaGgAGGCTATTGGGAGGTGTCGGCCTTTTGTGTAGGTCtCAaGCtTCaAACCCGTCAAACTACCGA 5′
ssODN
exon 31
PAM1
Glu 1157
PAM2
1161 Ala Gln His Phe Glu Glu Ser Leu Gly Gly Cys Gly Ser Phe Val Asp Leu Thr Arg Leu Lys Pro Leu Lys Ile Ala 1136
+ EcoRI
WT
Dnmt1 G1157E
400
300
200
100
b
Mlh1
sgRNA-2
GCCACTTTCAGGACTGTCTA
5′ TTCTTCAGAAACTCAACAATGTACTCTGCAAGCCCTTCCTTCGGGCCGTCGTCCTCTGTCCAGCCACTTTCAGGACTGTCTAAGGCCAGC 3′
3′ AAGAAGTCTTTGAGTTGTTACATGAGACGTTCGGGAAGGAAGCCCGGCAGCAGGAGACAGGTCGGTGAAAGTCCTGACAGATTCCGGTCG 5′
exon 16
Ala 612
PAM1
PAM2
AAGGAAGCCCGGCAGCAGGA
sgRNA-1
620 Lys Leu Phe Glu Val Ile Tyr Glu Ala Leu Gly Glu Lys Pro Gly Asp Asp Glu Thr Trp Gly Ser Glu Pro Ser Asp Leu Ala Leu 592
ApoI
EcoRI
Mlh1 G2101A
sgRNA-2
GCCACTTTCAGGACTGTCTA
5′ TTCTTCAGgAAtTCAACAATGTACTCTGtAAGCCCcTCtTTgGGcCCGTCGTCCTCTGTCCAGCCACTTTCAGGACTGTCTAAGGCCAGC 3′
3′ AAGAAGTCcTTaAGTTGTTACATGAGACaTTCGGGgAGaAACcCCgGGCAGCAGGAGACAGGTCGGTGAAAGTCCTGACAGATTCCGGTCG 5′
exon 16
Thr 612
PAM1
PAM2
620 Lys Leu Phe Glu Val Ile Tyr Glu Thr Leu Gly Glu Lys Pro Gly Asp Asp Glu Thr Trp Gly Ser Glu Pro Ser Asp Leu Ala Leu 592
+ EcoRI
Mlh1 A612T
WT
300
200
100

Forment et al, Supplementary Figure 3



Forment et al, Supplementary Figure 4

OXFORD UNIVERSITY PRESS | Nucleic Acids Research

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

Journal:	*Nucleic Acids Research*
Manuscript ID	NAR-01713-M-2016
Manuscript Type:	1 Standard Manuscript
Key Words:	Synthetic viability, Tof1, Fork Protection Complex, Camptothecin, SIR complex

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

Fabio Puddu[1,†], Mareike Herzog[1,2], Nicola J. Geisler[1], Vincenzo Costanzo[3] and Stephen P. Jackson[1,2,†]

[1]The Gurdon Institute and Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK

[2]The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK.

[3]IFOM (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/catenation between the two daughter DNA strands (3). The catenation 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 catenation 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 *TOF1* deletion in the diploid strain JRY9730 (*sir1Δ::LEU2/SIR1 HMRa/HMRα-α2Δ::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 genomecov 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 $E_{sir2}>0.9$ and $E_{sir3}>1.1$ and $E_{sir2}>0.9$ and $E_{GFP}<1.1$ and $E_{H4\text{-}K16ac}<0.75$ and $E_{H3}>0.75$ 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, *mrc1Δ* 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 *fob1Δ* strains were not hypersensitive to camptothecin and *FOB1* deletion did not alleviate the camptothecin hypersensitivity of a *csm3Δ* strain **(Figure 1B)**.

SIR gene mutations suppress camptothecin hypersensitivity of *tof1Δ/csm3Δ* cells

To understand the origin of the hypersensitivity of *tof1Δ* and *csm3Δ* strains to camptothecin, we carried out a synthetic viability genomic screening (20) to identify mutations capable of suppressing such hypersensitivity **(Figure 1C)**. We isolated sixteen resistant colonies and verified that they indeed displayed both resistance to camptothecin and absence of *TOF1* **(Figure 1D and 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 *TOP1*, which encodes the drug target. Three strains carried either of two nonsense mutations that inactivated *SIR3*, while eight of the remaining strains carried a nonsense mutation inactivating *SIR4* **(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 *SIR3* and *SIR4* in *tof1Δ* and *csm3Δ* strains, we verified that *SIR3* or *SIR4* inactivation mediated suppression of camptothecin hypersensitivity **(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 *IME2* (Inducer of MEiosis, which is not expressed in exponentially growing cells) or *IRC15*. However, ensuing studies established that neither *IME2* nor *IRC15* deletion suppressed the camptothecin hypersensitivity of *tof1Δ* cells **(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 *hhf-K16Q* mutation suppressed the camptothecin hypersensitivity of a *tof1Δ* strain, and at higher doses also reduced the camptothecin sensitivity of a wild-type strain **(Figure 2E)**. Similarly, mutation of H4-K16 to glycine (G), which prevents binding by Sir3(29), strongly counteracted the camptothecin sensitivity of both *tof1Δ* and wild-type cells. By contrast, mutating histone H4-K16 to non-acetylable arginine (R) produced much weaker suppression **(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 *hhf-K16R* mutation actually reduces transcriptional silencing (30). Taken together, these results highlighted a correlation between loss of silencing and camptothecin resistance.

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 *TOF1* and *CSM3* deletion on this particular phenotype by releasing synchronised wild-type, *tof1Δ* and *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 **(Figure 3A)**. Significantly, compared to wild-type controls, cells deleted for *TOF1* or *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 *tof1Δ* 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 *tof1Δ* by preventing ORC-mediated recruitment of Sir1. In line with this hypothesis, we found that deleting the BAH domain of ORC1 suppressed *tof1Δ* 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 *tof1Δ* 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 *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 *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 *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 *HML* and *HMR*; and of these, only the cryptic mating type loci require Sir1 for their silencing (39). The fact that *SIR1* deletion also suppresses the camptothecin sensitivity of *tof1Δ* cells initially suggested to us that *HML* and *HMR* represent the chromatin templates that are toxic to *tof1Δ* cells. However, we did not observe a reduction in *tof1Δ* sensitivity to camptothecin by deleting *HML* and *HMR*, meaning that these two genomic loci alone are not responsible for the strong camptothecin sensitivity phenotype displayed by *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 SPJ, 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 Armenise-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 SPJ laboratory for helpful discussions, and Helena Santos-Rosa for advice with designing histone point mutations. The initial project was conceived by FP, VC and SPJ. 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 SPJ, with contributions made by all the other authors.

References

1. Schvartzman,J.B. and Stasiak,A. (2004) A topological view of the replicon. *EMBO reports*, **5**, 256–261.

2. Postow,L., Crisona,N.J., Peter,B.J., Hardy,C.D. and Cozzarelli,N.R. (2001) Topological challenges to DNA replication: conformations at the fork. *Proc Natl Acad Sci U S A*, **98**, 8219–8226.

3. Sundin,O. and Varshavsky,A. (1980) Terminal stages of SV40 DNA replication proceed via multiply intertwined catenated dimers. *Cell*, **21**, 103–114.

4. Baxter,J., Sen,N., Mart 'ı nez,V.L., De Carandini,M.E.M., Schvartzman,J.B., Diffley,J.F.X. and Aragón,L. (2011) Positive supercoiling of mitotic DNA drives decatenation by topoisomerase II in eukaryotes. *Science (New York, N.Y.)*, **331**, 1328–1332.

5. Fachinetti,D., Bermejo,R., Cocito,A., Minardi,S., Katou,Y., Kanoh,Y., Shirahige,K., Azvolinsky,A., Zakian,V.A. and Foiani,M. (2010) Replication termination at eukaryotic

chromosomes is mediated by Top2 and occurs at genomic loci containing pausing elements. *Molecular Cell*, **39**, 595–605.

6. Schalbetter,S.A., Mansoubi,S., Chambers,A.L., Downs,J.A. and Baxter,J. (2015) Fork rotation and DNA precatenation are restricted during DNA replication to prevent chromosomal instability. *Proc Natl Acad Sci U S A*, **112**, E4565–70.

7. Ivessa,A.S., Lenzmeier,B.A., Bessler,J.B., Goudsouzian,L.K., Schnakenberg,S.L. and Zakian,V.A. (2003) The Saccharomyces cerevisiae helicase Rrm3p facilitates replication past nonhistone protein-DNA complexes. *Molecular Cell*, **12**, 1525–1536.

8. Zaratiegui,M., Vaughn,M.W., Irvine,D.V., Goto,D., Watt,S., Bähler,J., Arcangioli,B. and Martienssen,R.A. (2011) CENP-B preserves genome integrity at replication forks paused by retrotransposon LTR. *Nature*, **469**, 112–115.

9. Tourrière,H., Versini,G., Cordón-Preciado,V., Alabert,C. and Pasero,P. (2005) Mrc1 and Tof1 promote replication fork progression and recovery independently of Rad53. *Molecular Cell*, **19**, 699–706.

10. Katou,Y., Kanoh,Y., Bando,M., Noguchi,H., Tanaka,H., Ashikari,T., Sugimoto,K. and Shirahige,K. (2003) S-phase checkpoint proteins Tof1 and Mrc1 form a stable replication-pausing complex. *Nature*, **424**, 1078–1083.

11. Alcasabas,A.A., Osborn,A.J., Bachant,J., Hu,F., Werler,P.J., Bousset,K., Furuya,K., Diffley,J.F.X., Carr,A.M. and Elledge,S.J. (2001) Mrc1 transduces signals of DNA replication stress to activate Rad53. *Nature cell biology*, **3**, 958–965.

12. Foss,E.J. (2001) Tof1p regulates DNA damage responses during S phase in Saccharomyces cerevisiae. *Genetics*, **157**, 567–577.

13. Osborn,A.J. and Elledge,S.J. (2003) Mrc1 is a replication fork component whose phosphorylation in response to DNA replication stress activates Rad53. *Genes & development*, **17**, 1755–1767.

14. Bando,M., Katou,Y., Komata,M., Tanaka,H., Itoh,T., Sutani,T. and Shirahige,K. (2009) Csm3, Tof1, and Mrc1 form a heterotrimeric mediator complex that associates with DNA replication forks. *The Journal of biological chemistry*, **284**, 34355–34365.

15. Redon,C., Pilch,D.R. and Bonner,W.M. (2006) Genetic analysis of Saccharomyces cerevisiae H2A serine 129 mutant suggests a functional relationship between H2A and the sister-chromatid cohesion partners Csm3-Tof1 for the repair of topoisomerase I-induced DNA damage. *Genetics*, **172**, 67–76.

16. Ray Chaudhuri,A., Hashimoto,Y., Herrador,R., Neelsen,K.J., Fachinetti,D., Bermejo,R., Cocito,A., Costanzo,V. and Lopes,M. (2012) Topoisomerase I poisoning results in PARP-mediated replication fork reversal. *Nature structural & molecular biology*, **19**, 417–423.

17. Koster,D.A., Palle,K., Bot,E.S.M., Bjornsti,M.-A. and Dekker,N.H. (2007) Antitumour drugs impede DNA uncoiling by topoisomerase I. *Nature*, **448**, 213–217.

18. Longtine,M.S., McKenzie,A., Demarini,D.J., Shah,N.G., Wach,A., Brachat,A., Philippsen,P. and Pringle,J.R. (1998) Additional modules for versatile and economical PCR-based gene deletion and modification in Saccharomyces cerevisiae. *Yeast (Chichester, England)*, **14**,

953–961.

19. Dodson,A.E. and Rine,J. (2015) Heritable capture of heterochromatin dynamics in *Saccharomyces cerevisiae*. *eLife*, **4**, 1–22.

20. Puddu,F., Oelschlaegel,T., Guerini,I., Geisler,N.J., Niu,H., Herzog,M., Salguero,I., Ochoa-Montaño,B., Viré,E., Sung,P., *et al.* (2015) Synthetic viability genomic screening defines Sae2 function in DNA repair. *The EMBO journal*, **34**, 1509–1522.

21. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, **25**, 2078–2079.

22. McLaren,W., Pritchard,B., Rios,D., Chen,Y., Flicek,P. and Cunningham,F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics (Oxford, England)*, **26**, 2069–2070.

23. Kueng,S., Oppikofer,M. and Gasser,S.M. (2013) SIR proteins and the assembly of silent chromatin in budding yeast. *Annual review of genetics*, **47**, 275–306.

24. Rine,J. and Herskowitz,I. (1987) Four genes responsible for a position effect on expression from HML and HMR in Saccharomyces cerevisiae. *Genetics*, **116**, 9–22.

25. Grozinger,C.M., Chao,E.D., Blackwell,H.E., Moazed,D. and Schreiber,S.L. (2001) Identification of a class of small molecule inhibitors of the sirtuin family of NAD-dependent deacetylases by phenotypic screening. *The Journal of biological chemistry*, **276**, 38837–38843.

26. Rajamohan,S.B., Pillai,V.B., Gupta,M., Sundaresan,N.R., Birukov,K.G., Samant,S., Hottiger,M.O. and Gupta,M.P. (2009) SIRT1 promotes cell survival under stress by deacetylation-dependent deactivation of poly(ADP-ribose) polymerase 1. *Molecular and cellular biology*, **29**, 4116–4129.

27. Jeong,J., Juhn,K., Lee,H., Kim,S.-H., Min,B.-H., Lee,K.-M., Cho,M.-H., Park,G.-H. and Lee,K.-H. (2007) SIRT1 promotes DNA repair activity and deacetylation of Ku70. *Exp. Mol. Med.*, **39**, 8–13.

28. Kaidi,A., Weinert,B.T., Choudhary,C. and Jackson,S.P. (2010) Human SIRT6 promotes DNA end resection through CtIP deacetylation. *Science (New York, N.Y.)*, **329**, 1348–1353.

29. Onishi,M., Liou,G.-G., Buchberger,J.R., Walz,T. and Moazed,D. (2007) Role of the conserved Sir3-BAH domain in nucleosome binding and silent chromatin assembly. *Molecular Cell*, **28**, 1015–1028.

30. Meijsing,S.H. and Ehrenhofer-Murray,A.E. (2001) The silencing complex SAS-I links histone acetylation to the assembly of repressed chromatin by CAF-I and Asf1 in Saccharomyces cerevisiae. *Genes Dev.*, **15**, 3169–3182.

31. Redon,C., Pilch,D.R., Rogakou,E.P., Orr,A.H., Lowndes,N.F. and Bonner,W.M. (2003) Yeast histone 2A serine 129 is essential for the efficient repair of checkpoint-blind DNA damage. *EMBO reports*, **4**, 678–684.

32. Wai,H.H., Vu,L., Oakes,M. and Nomura,M. (2000) Complete deletion of yeast chromosomal

rDNA repeats and integration of a new rDNA repeat: use of rDNA deletion strains for functional analysis of rDNA promoter elements in vivo. *Nucleic acids research*, **28**, 3524–3534.

33. Smith,J.S., Brachmann,C.B., Pillus,L. and Boeke,J.D. (1998) Distribution of a limited Sir2 protein pool regulates the strength of yeast rDNA silencing and is modulated by Sir4p. *Genetics*, **149**, 1205–1219.

34. Kyrion,G., Liu,K., Liu,C. and Lustig,A.J. (1993) RAP1 and telomere structure regulate telomere position effects in Saccharomyces cerevisiae. *Genes Dev.*, **7**, 1146–1159.

35. Vignais,M.L., Woudt,L.P., Wassenaar,G.M., Mager,W.H., Sentenac,A. and Planta,R.J. (1987) Specific binding of TUF factor to upstream activation sites of yeast ribosomal protein genes. *The EMBO journal*, **6**, 1451–1457.

36. Lieb,J.D., Liu,X., Botstein,D. and Brown,P.O. (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nature genetics*, **28**, 327–334.

37. Triolo,T. and Sternglanz,R. (1996) Role of interactions between the origin recognition complex and SIR1 in transcriptional silencing. *Nature*, **381**, 251–253.

38. Gardner,K.A., Rine,J. and Fox,C.A. (1999) A region of the Sir1 protein dedicated to recognition of a silencer and required for interaction with the Orc1 protein in saccharomyces cerevisiae. *Genetics*, **151**, 31–44.

39. Smith,J.S. and Boeke,J.D. (1997) An unusual form of transcriptional silencing in yeast ribosomal DNA. *Genes Dev.*, **11**, 241–254.

40. Sussel,L., Vannier,D. and Shore,D. (1993) Epigenetic switching of transcriptional states: cis- and trans-acting factors affecting establishment of silencing at the HMR locus in Saccharomyces cerevisiae. *Molecular and cellular biology*, **13**, 3919–3928.

41. Ivy,J.M., Klar,A.J. and Hicks,J.B. (1986) Cloning and characterization of four SIR genes of Saccharomyces cerevisiae. *Molecular and cellular biology*, **6**, 688–702.

42. Thurtle,D.M. and Rine,J. (2014) The molecular topography of silenced chromatin in Saccharomyces cerevisiae. *Genes Dev.*, **28**, 245–258.

43. Teytelman,L., Thurtle,D.M., Rine,J. and van Oudenaarden,A. (2013) Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc Natl Acad Sci U S A*, **110**, 18602–18607.

44. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, **320**, 1344–1349.

45. Shor,E., Warren,C.L., Tietjen,J., Hou,Z., Müller,U., Alborelli,I., Gohard,F.H., Yemm,A.I., Borisov,L., Broach,J.R., *et al.* (2009) The origin recognition complex interacts with a subset of metabolic genes tightly linked to origins of replication. *PLoS genetics*, **5**, e1000755.

46. Azvolinsky,A., Giresi,P.G., Lieb,J.D. and Zakian,V.A. (2009) Highly transcribed RNA polymerase II genes are impediments to replication fork progression in Saccharomyces cerevisiae. *Molecular Cell*, **34**, 722–734.

47. Mohanty,B.K., Bairwa,N.K. and Bastia,D. (2006) The Tof1p-Csm3p protein complex counteracts the Rrm3p helicase to control replication termination of Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, **103**, 897–902.

48. Ren,J., Wang,C.-L. and Sternglanz,R. (2010) Promoter strength influences the S phase requirement for establishment of silencing at the Saccharomyces cerevisiae silent mating type Loci. *Genetics*, **186**, 551–560.

49. Osada,S., Sutton,A., Muster,N., Brown,C.E., Yates,J.R., Sternglanz,R. and Workman,J.L. (2001) The yeast SAS (something about silencing) protein complex contains a MYST-type putative acetyltransferase and functions with chromatin assembly factor ASF1. *Genes Dev.*, **15**, 3155–3168.

50. Enomoto,S. and Berman,J. (1998) Chromatin assembly factor I contributes to the maintenance, but not the re-establishment, of silencing at the yeast silent mating loci. *Genes Dev.*, **12**, 219–232.

51. Sperling,A.S., Jeong,K.S., Kitada,T. and Grunstein,M. (2011) Topoisomerase II binds nucleosome-free DNA and acts redundantly with topoisomerase I to enhance recruitment of RNA Pol II in budding yeast. *Proc Natl Acad Sci U S A*, **108**, 12693–12698.

52. Maliepaard,M., van Gastelen,M.A., de Jong,L.A., Pluim,D., van Waardenburg,R.C., Ruevekamp-Helmers,M.C., Floot,B.G. and Schellens,J.H. (1999) Overexpression of the BCRP/MXR/ABCP gene in a topotecan-selected ovarian tumor cell line. *Cancer Res.*, **59**, 4559–4563.

53. Jensen,N.F., Agama,K., Roy,A., Smith,D.H., Pfister,T.D., Rømer,M.U., Zhang,H.-L., Doroshow,J.H., Knudsen,B.R., Stenvang,J., *et al.* (2016) Characterization of DNA topoisomerase I in three SN-38 resistant human colon cancer cell lines reveals a new pair of resistance-associated mutations. *J. Exp. Clin. Cancer Res.*, **35**, 56.

54. Chrencik,J.E., Staker,B.L., Burgin,A.B., Pourquier,P., Pommier,Y., Stewart,L. and Redinbo,M.R. (2004) Mechanisms of camptothecin resistance by human topoisomerase I mutations. **339**, 773–784.

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 *tof1Δ* cells, while pre-growth in the presence of sirtinol is sufficient to suppresses the camptothecin hypersensitivity phenotype of *tof1Δ* cells.

Figure 4. Disruption of *SIR1* suppresses camptothecin sensitivity in wild-type and *tof1Δ* cells.

(A) Deletion of the rDNA locus is not sufficient to suppress the hypersensitivity of *tof1Δ* cells to camptothecin. **(B)** A mutation in *RAP1* that disrupts telomeric silencing does not suppress the hypersensitivity of *tof1Δ* cells to camptothecin. **(C)** Deletion of *SIR1* suppresses camptothecin sensitivity in wild-type and *tof1Δ* cells. **(D)** *sir1Δ* and *tof1Δ* show a synergistic defect in the ability to mate. **(E)** *tof1Δsir1Δ* 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 *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 *tof1Δ* cells to camptothecin.

(A) Homozygous *tof1Δ/tof1Δ* diploid cells are as sensitive to camptothecin as *tof1Δ* haploids and their hypersensitivity can be rescued by sirtinol. **(B)** Deletion of *HML* and *HMR* cannot suppress the camptothecin hypersensitivity of *tof1Δ* 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 of 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

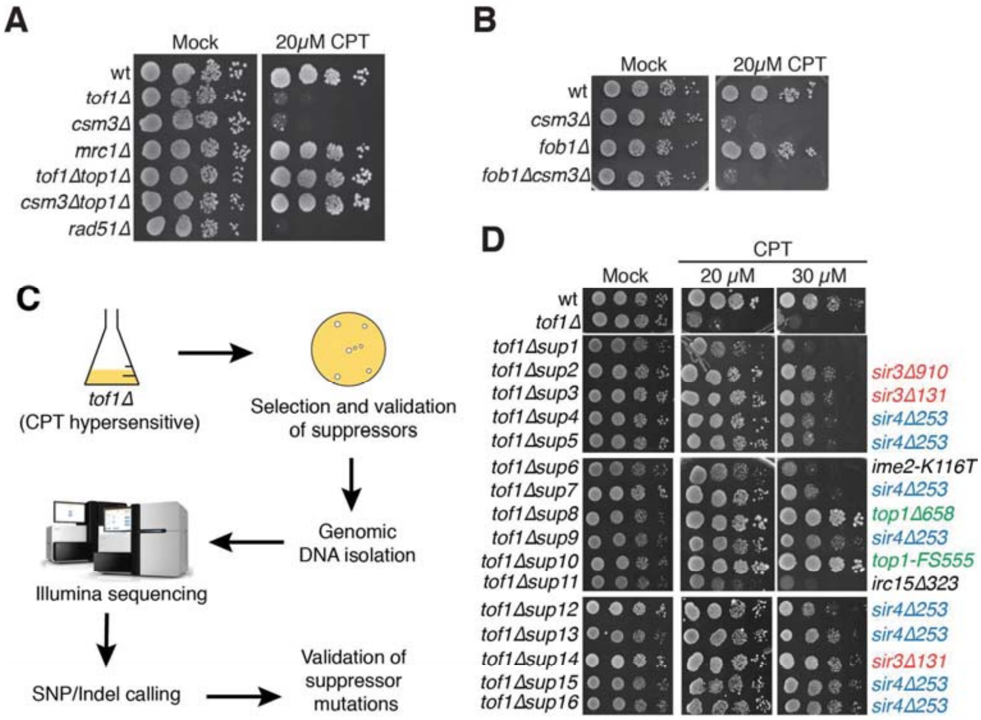


Figure 1

Figure 2

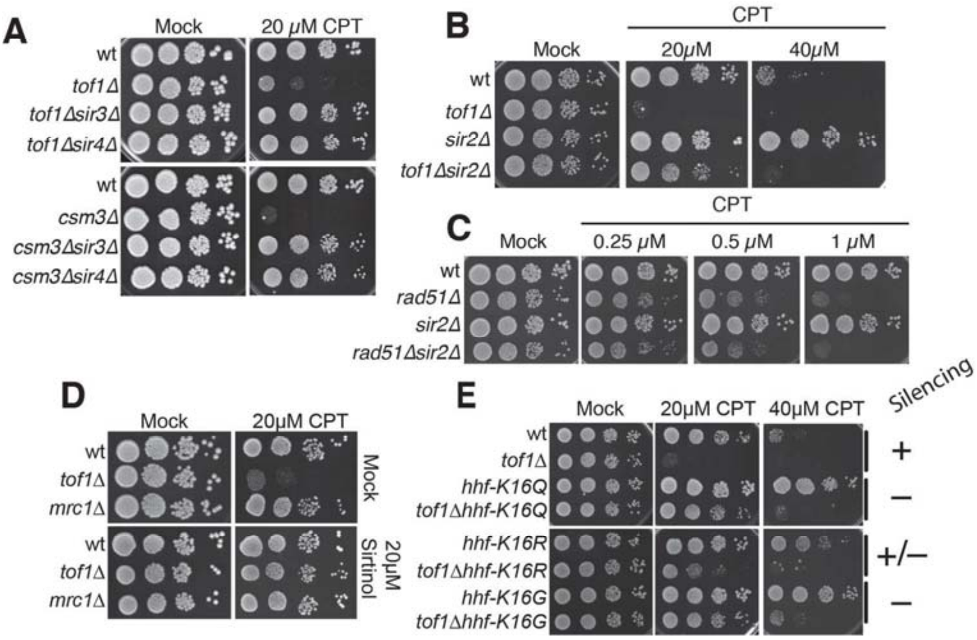


Figure 2

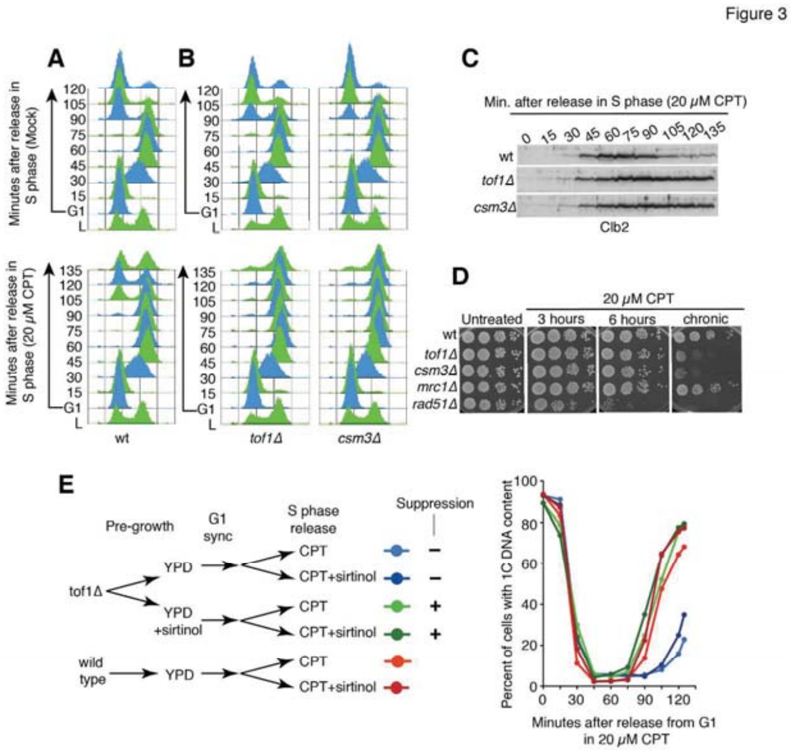


Figure 3
A
B
C
D
E
Minutes after release in S phase (Mock)
Minutes after release in S phase (20 μM CPT)
120
105
90
75
60
45
30
15
G1
L
135
wt
tof1Δ
csm3Δ
Min. after release in S phase (20 μM CPT)
0 15 30 45 60 75 90 105 120 135
Clb2
20 μM CPT
Untreated
3 hours
6 hours
chronic
mrc1Δ
rad51Δ
Pre-growth
G1 sync
S phase release
Suppression
tof1Δ
YPD
YPD +sirtinol
wild type
CPT
CPT+sirtinol
−
+
Percent of cells with 1C DNA content
Minutes after release from G1 in 20 μM CPT
0
20
40
60
80
100
30
60
90
120

Figure 3

Figure 4

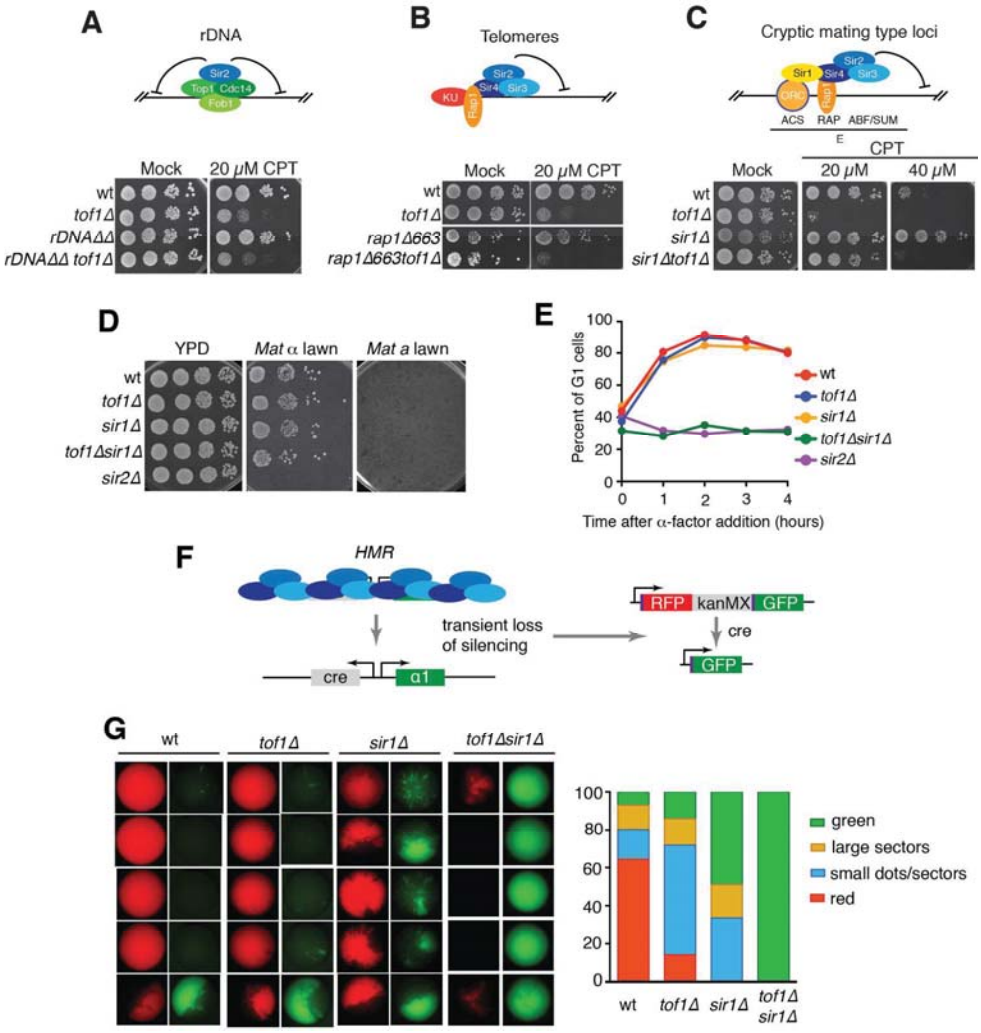


Figure 4

Figure 5

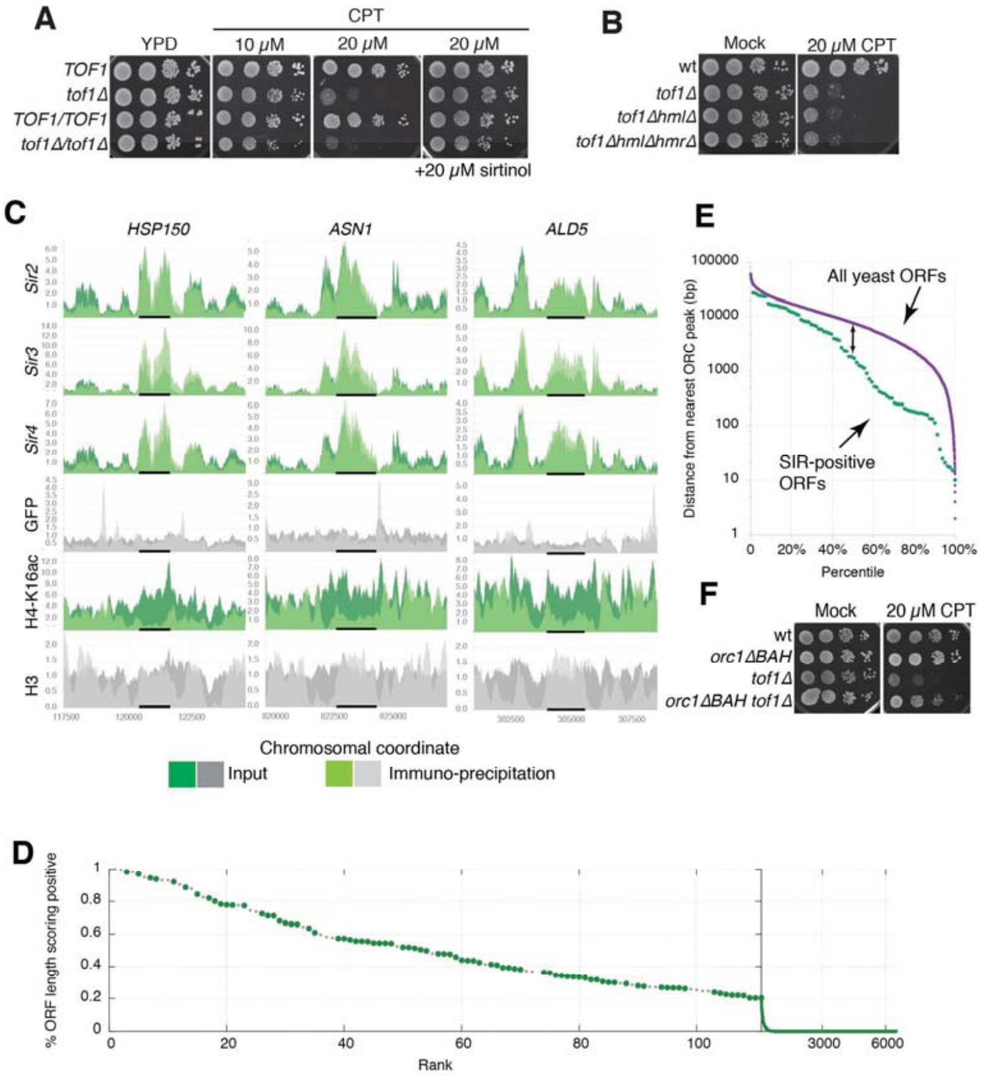


Figure 5

Supplementary Material

Supplementary Table 1: Yeast strains used in this study.

Supplementary Table 2: Whole-genome sequencing data.

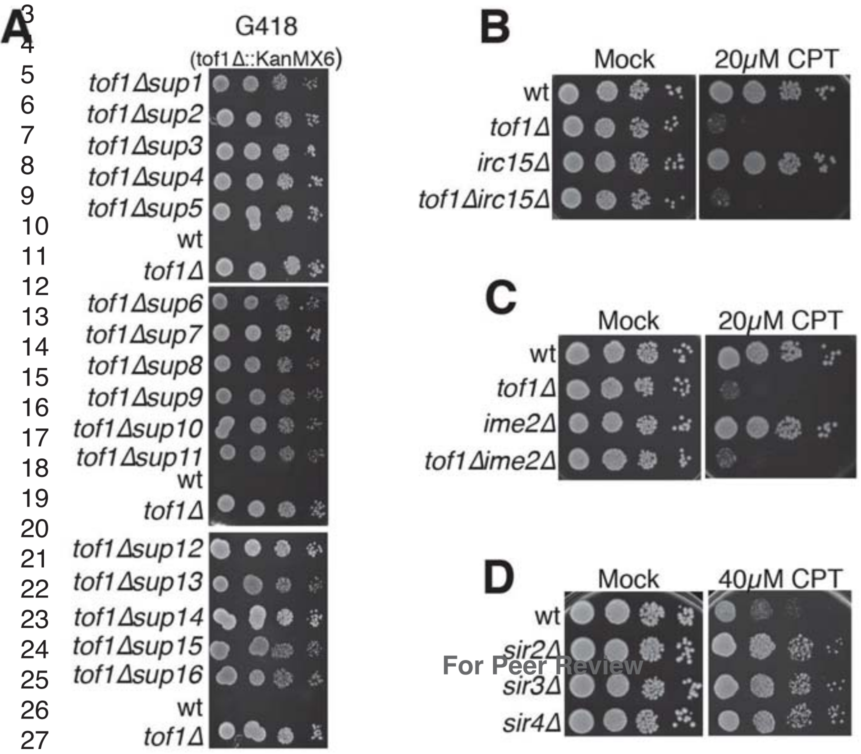
Supplementary Table 3: SRA accession numbers for ChIP-seq data

Supplementary Table 4: ORFs with SIR score > 0.2

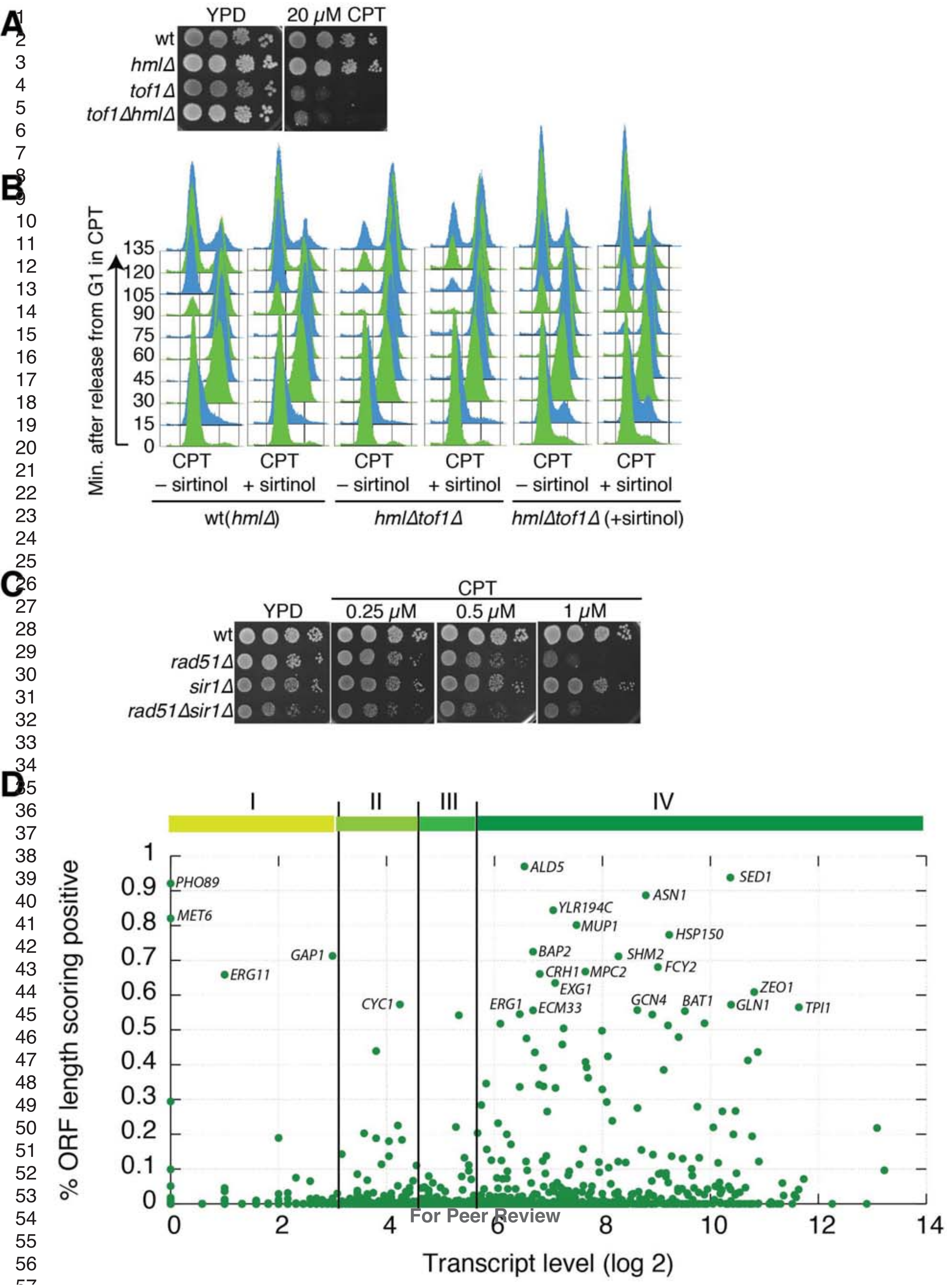
Supplementary Figure 1
(A) Suppressor strains recovered from the *tof1Δ* synthetic-viability screen are G418 resistant, indicating presence of the *TOF1* deletion cassette. **(B)** Deletion of *IRC15* does not suppress *tof1Δ* camptothecin hypersensitivity. **(C)** Deletion of *IME2* does not suppress *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



Supplementary Figure 2


A
YPD
20 μM CPT
wt
hmlΔ
tof1Δ
tof1ΔhmlΔ
B
Min. after release from G1 in CPT
CPT – sirtinol
CPT + sirtinol
CPT – sirtinol
CPT + sirtinol
CPT – sirtinol
CPT + sirtinol
wt(hmlΔ)
hmlΔtof1Δ
hmlΔtof1Δ (+sirtinol)
C
YPD
CPT
0.25 μM
0.5 μM
1 μM
wt
rad51Δ
sir1Δ
rad51Δsir1Δ
D
I
II
III
IV
% ORF length scoring positive
PHO89
MET6
ERG11
GAP1
CYC1
ALD5
YLR194C
MUP1
BAP2
CRH1
MPC2
EXG1
ERG1
ECM33
ASN1
HSP150
SHM2
FCY2
GCN4
BAT1
SED1
ZEO1
GLN1
TPI1
Transcript level (log 2)