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]:

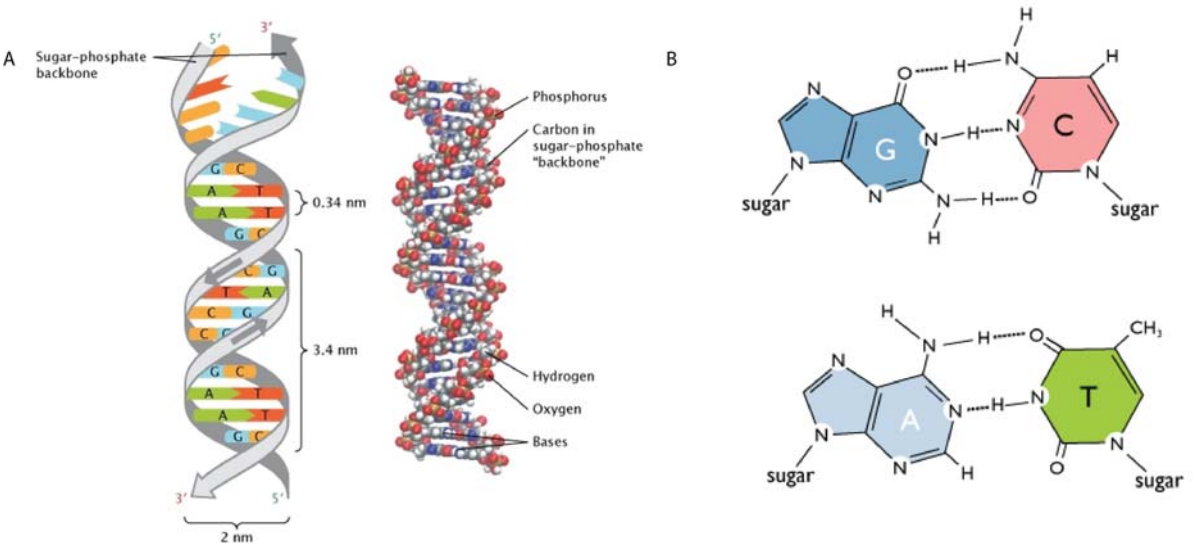


Figure 1.1: Structure of DNA
A - Structure of the DNA double helix. Reproduced from [23] in accordance with the publisher's terms of use. B - Pairing of the 4 DNA bases: adenine forms hydrogen bonds with thymine, guanine bonds with cytosine. From [24].

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

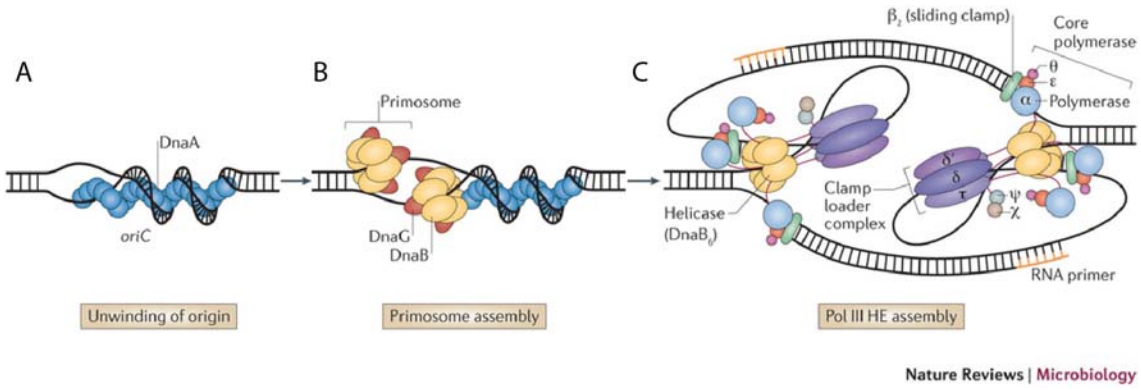


Figure 1.2: Replication initiation in *E. coli*
A - Unwinding of the DNA origin by oriC; B - Loading of the DNA helicases and DNA primases; C - Assembly of the replisome — Reproduced from [37] with permission from the publisher.

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

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

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

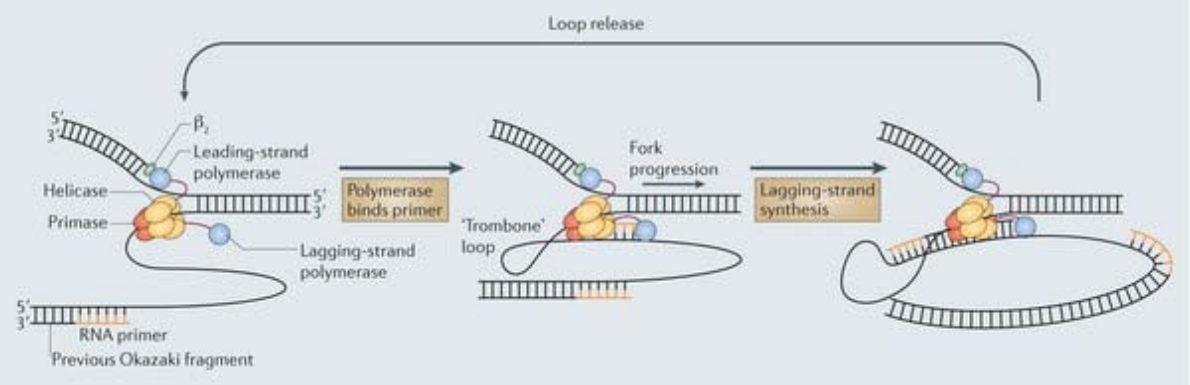


Figure 1.3: Lagging strand DNA synthesis in *E. coli*
Model of lagging strand DNA synthesis coordination: lagging strand polymerase action is thought to generate a "Trombone" loop to accommodate both polymerases and their opposing movement at the replication fork. Reproduced from [37] with permission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

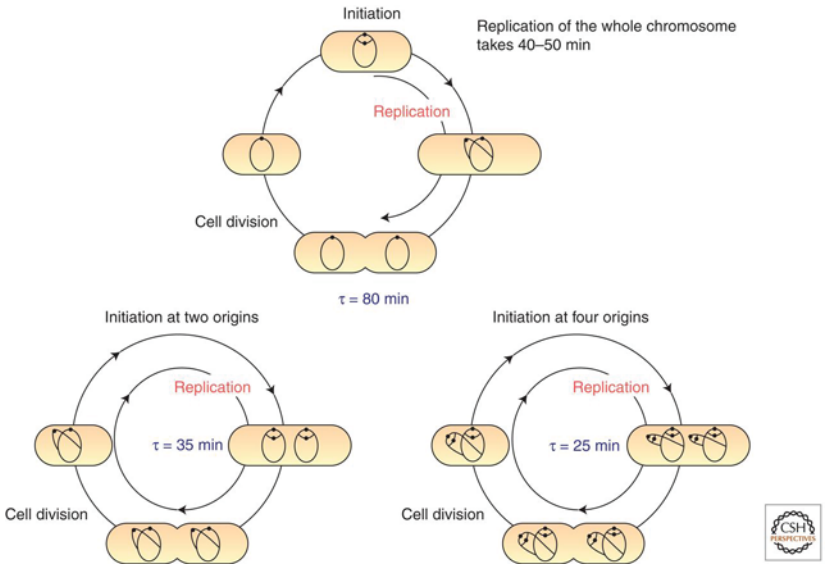


Figure 1.4: Overlapping replication cycles in *E. coli*
Depending on growth conditions population doubling time in *E. coli* can be shorter than the time required to replicate the chromosome due to re-initiation before cell division. Reproduced from [53] with permission from the publisher. Copyright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

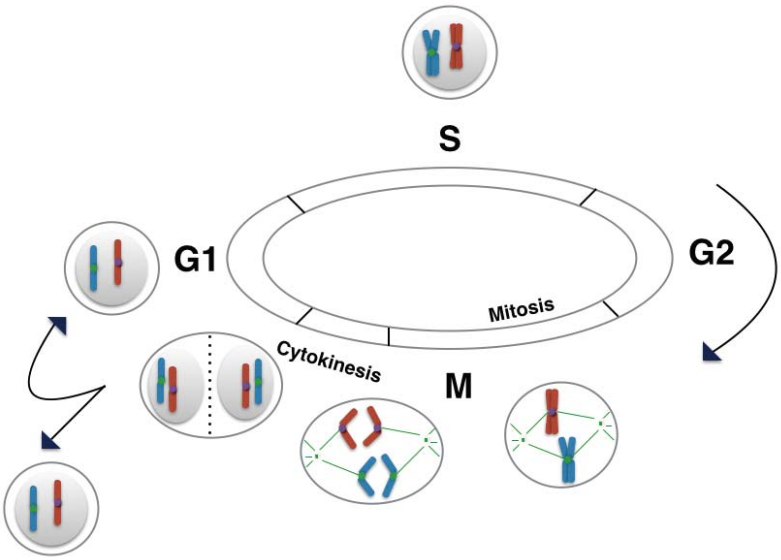


Figure 1.5: The eukaryotic cell cycle
A schematic of the eukaryotic cell cycle. S-phase (replication) and M-phase (mitosis followed by cytokinesis) alternate and are separated by two growth phases (G1 and G2).

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

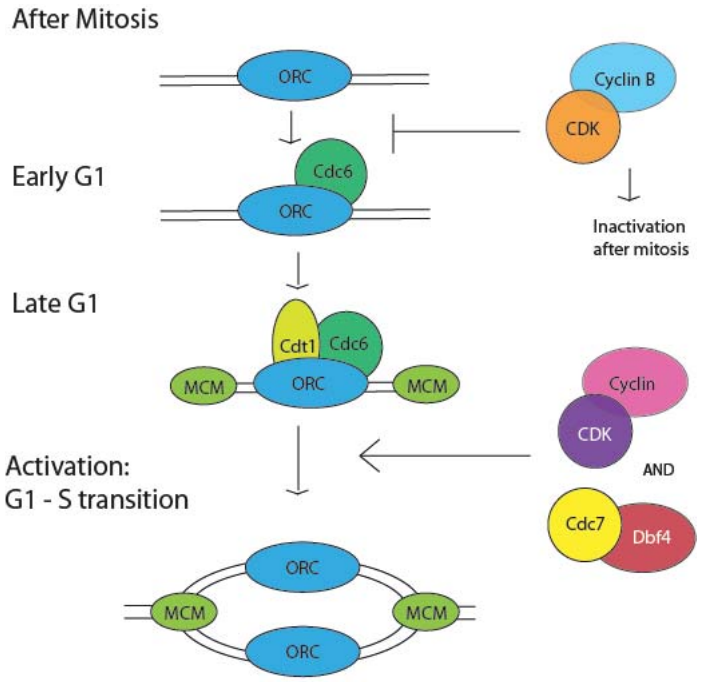


Figure 1.6: Licensing of eukaryotic origins of replication
A schematic detailing the ordered assembly of licensing factors on eukaryotic origins starting with the Origin Recognition Complex (ORC). Until the end of mitosis licensing is prevented by actions of cyclin-CDKs. After anaphase the licensing 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^{Cdc4} [170–173]. Additionally, Cdc6 expression is blocked due to CDK regulation of the transcription factor Swi5 and CDK phosphorylation and subsequent binding of Cdc6 blocks its licensing activities [114, 174]. Upon completion of mitosis, CDK activity is lowered in two key ways: the mitotic cyclin Clb2 is degraded by the 26S proteasome, and the CDK inhibitor Sic1 inhibits G1-CDK activity [175–178]. Degradation of Clb2 releases Cdc6 to participate in licensing again [174]. The phosphatase Cdc14 is also involved in promoting preRC assembly. Among other things it removes CDK-dependent phosphorylation from Swi5, which can then activate expression of Cdc6 and Sic1 [179, 180]. Cdc14 dephosphorylates Sic1, which protects it from SCF^{Cdc4}-mediated degradation [179]. The CDK inhibitor Sic1 is also a key barrier to firing origins too early [178, 181, 182]. As the cell cycle progresses, phosphorylation of Sic1 by the CDK complexes Cln-Cdc28 and Clb-Cdc28 targets it for ubiquitin-mediated degradation [183–185].

Negative regulation of licensing factors To prevent re-licensing of origins after initiation of replication, many of the licensing factors described above are subject to negative regulation often in multiple ways. This is likely to prevent reassembly of a preRC complex, subsequent reloading of the helicase and thus re-replication. Interfering with this negative regulation has been shown to be able to cause re-replication in many cases[114, 169, 186].

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

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

Cdt1 overexpression causes re-replication making Cdt1 regulation a key part of licensing regulation[214]. Targeting Cdt1 for degradation is conserved in higher eukaryotes including *Caenorhabditis elegans, Drosophila, Xenopus*, and mammals[114, 215]. There are multiple mechanisms to degrade Cdt1, highlighting the importance of this process[75, 114, 216] (Fig. 1.7). One CDK-dependent mechanism involves the SCF–Skp2 E3 ubiquitin ligase complex to mark Cdt1 for degradation[206, 217–221]. In human cells, Cdk2 and Cdk4 bind Cdt1 and phosphorylate it, thereby recruiting the E3 ligase to mark Cdt1 for degradation during S and G2 phase. While this pathway has been observed in human cells, it is not conserved in other metazoans suggesting it could be an evolutionarily recent addition[221]. Because impairment of this pathway still leads to Cdt1 degradation, another mechanism for Cdt1 degradation was identified involving the Cul4–Ddb1–Cdt2 complex and it has been demonstrated to be essential for Cdt1 degradation from *S. pombe* to metazoans[114, 203, 219, 221–228]. Degradation

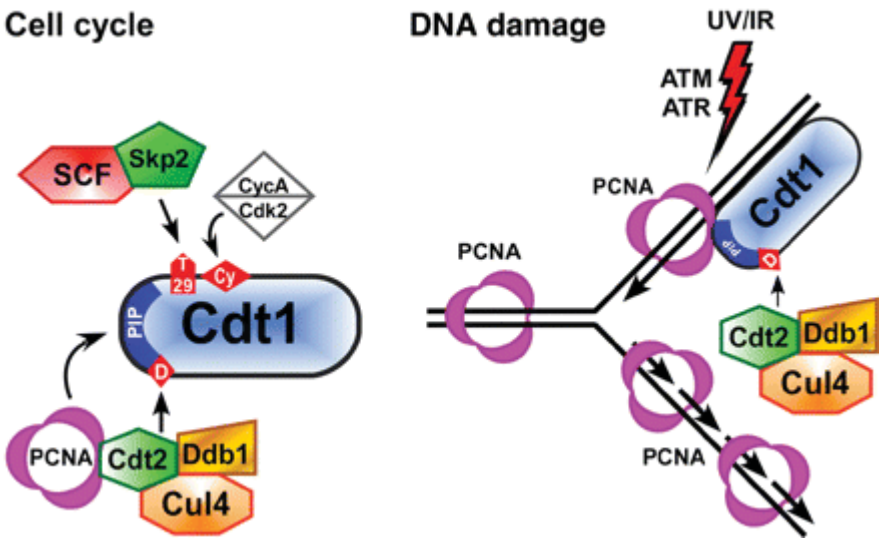


Figure 1.7: Degradation of Cdt1 during the cell cycle and in response to DNA damage During S phase, Ctd1 is targeted for degradation by the SCF–Skp2 and by the PCNA–Cul4–Ddb1–Cdt2 pathways. Cyclin-CDK activity results in the phosphorylation of Cdt1, which in turn allows recruitment of SCF–Skp2, an E3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

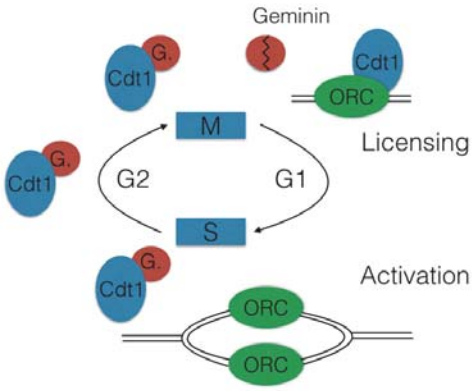


Figure 1.8: Regulation of Cdt1 by association with Geminin
Geminin regulates Cdt1 by binding and sequestering it during S and G2-phase. Its degradation after mitosis allows Cdt1 participation in origin licensing in early G1 phase.

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

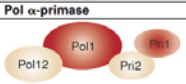
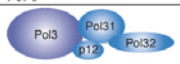
	Pol α-primase	Pol δ	Pol ε
Subunit organization			
Genes and subunit sizes			
S. cerevisiae	Pol1-p167	Pol3-p125	Pol2-p256
	Pol12-p79	Pol31-p55	Dpb2-p78
	Pri1-p48	Pol32-p40	Dpb3-p23
	Pri2-p62	–	Dpb4-p22
S. pombe	Pol1-p159	Pol3-p124	Pol2-p253
	Pol12-p64	Cdc1-p51	Dpb2-p67
	Pri1-p52	Cdc27-p42	Dpb3-p22
	Spp2-p53	Cdm1-p19	Dpb4-p24
Human	PolA1-p166	PolD1-p124	PolE-p261
	PolA2-p68	PolD2-p51	PolE2-p59
	Prim1-p48	PolD3-p66	PolE3-p17
	Prim2A-p58	PolD4-p12	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 sub-

unit, that stabilise its interactions with other subunits, cause an increased mutation rate[259]. In fact, evidence suggests, that the presence of mutated Dpb2 protein in the cell does not only affect the intrinsic fidelity of Pol ε, but also promotes the increased participation of DNA polymerase zeta (Pol ζ; the catalytic subunit encoded by *REV3* in *S. cerevisiae*), an error-prone polymerase, in DNA replication[288]. The inter-origin distance can be long in eukaryotes: in budding and fission yeast it is on average 38kb[289] and can be much longer in higher eukaryotes[259]. The eukaryotic replicative polymerases Pol δ and Pol ε have been found to be comparable in their high processivity in the presence of the sliding clamp PCNA[290, 291], reviewed in [259]. However, while Pol δ processivity requires its interaction with PCNA, Pol ε seems to be highly processive even without PCNA[292]. Pol ε shows high affinity for DNA, but a low affinity for PCNA; in contrast, Pol δ shows the opposite affinities for those two binding partners[291]. Recently, a structure for POLE has shed some light on this phenomenon: Pol ε has an extra domain (P domain) close to the DNA, allowing it to encircle the nascent double-stranded DNA, likely decreasing it "falling off" the DNA(Fig. 1.9). Lagging strand replication, like in *E. coli*, requires more steps to be initiated. Replication has to be initiated several times by primase and the primer elongated by Pol α[78]. Pol α is then switched to Pol δ which elongates the growing DNA strand until it encounters the previously synthesized Okazaki fragment[78]. Subsequently, the discontinuously synthesized fragments of DNA are joined up in a process called "Okazaki fragment maturation"[78]. This process needs to be highly accurate to avoid insertions and deletions (INDELs) and efficient so that none of the nearly 100,000 nicks in the DNA, generated during one budding yeast S phase, remain[293]. The former would alter the genetic information and the latter, if unrepaired and then replicated, would result in a double-strand break (DSB) in the DNA. And while DSBs can be repaired, a small number of lesions can overwhelm the repair system and cause cell death[294]. The nicks between fragments are processed by Pol δ and Rad27(FEN1) and ligated by DNA ligase I[293]. Pol δ displaces 2-3 nucleotides of any RNA or DNA of the next fragment that it meets[293]. Rad27, a 5' flap endonuclease, efficiently processes the nick. If it is absent or not functioning at an optimal level, Pol δ idles (it backs up using its exonuclease activity)[293]. This process is thought to keep the length of displaced downstream nucleotides to a minimum[293]. This behaviour was not observed for Pol ε consistent with Pol δ as the lagging strand polymerase. At some point Pol δ will switch from idling to strand displacement[293]. In these cases, displaced DNA will be single-stranded and coated by RPA, which makes it an inefficient target for FEN1, especially if the DNA forms secondary structures[295]. As demonstrated in yeast, in these cases the essential Dna2 nuclease/helicase will cleave these flaps[295]. While it is thought to be the less common path to process Okazaki

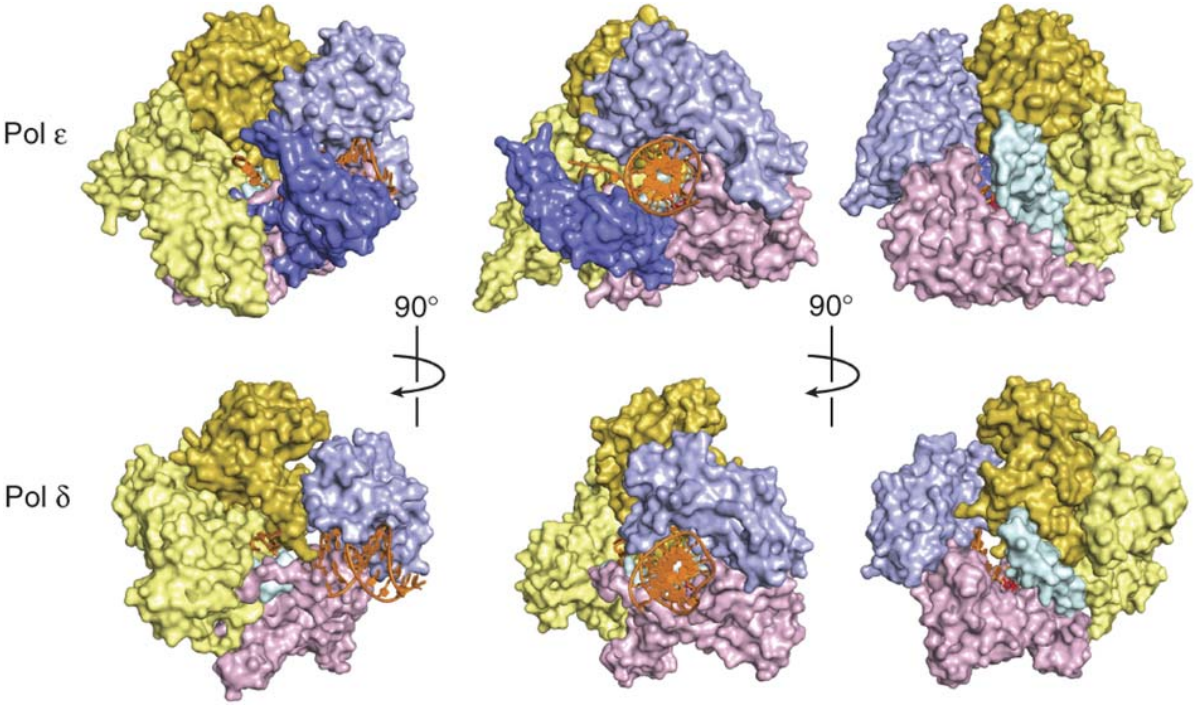


Figure 1.9: Structure of DNA polymerase δ and DNA polymerase ε
Surface representations of Pol δ (PDB 3IAY20) and Pol ε: The fingers (cyan), palm (pink), thumb (light blue), exonuclease (gold) and N-terminal (light yellow) domains are shown in three orientations. In the case of Polε, the additional, newly 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.

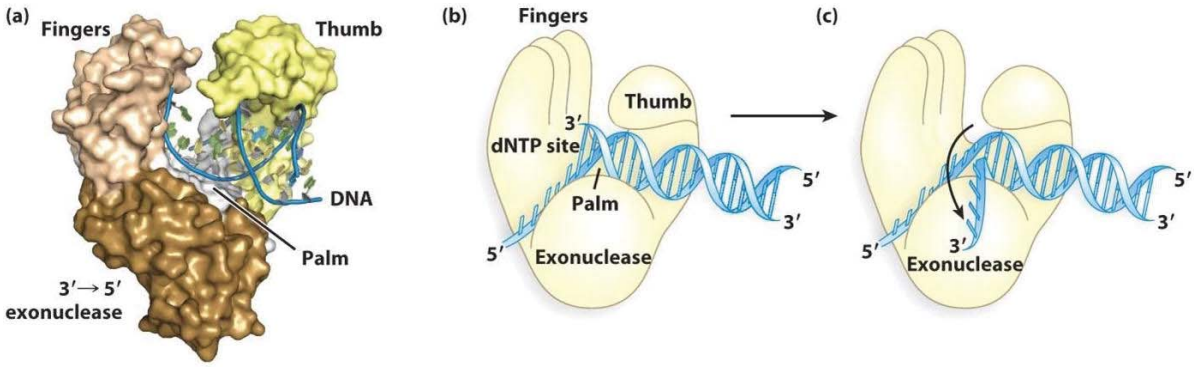


Figure 1.10: Structure and representation of replicative DNA polymerases
A - Surface crystal structure of a DNA polymerase complexed with DNA. B & C - Cartoon representation of the DNA polymerase structure in polymerization mode (B) and proofreading mode (C). Figure reproduced from [325]. Used by permission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

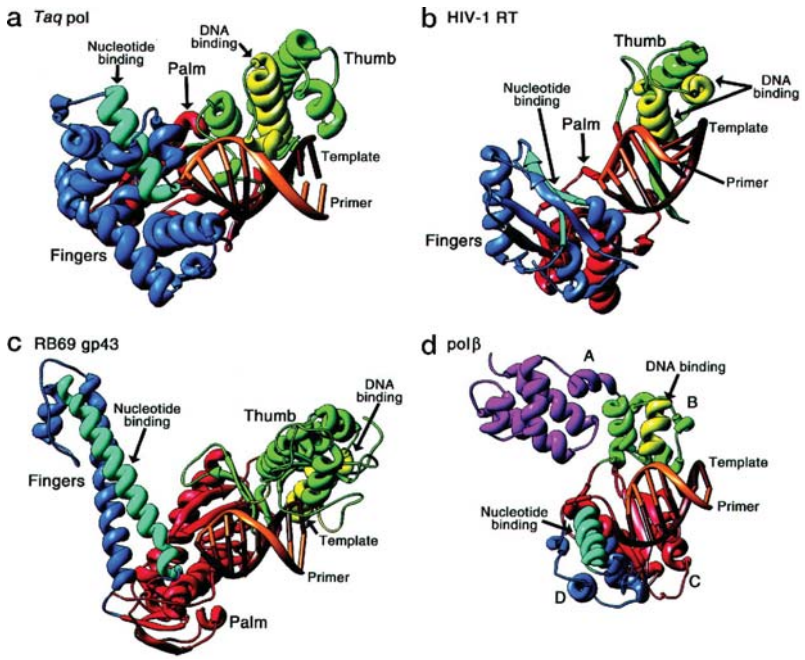


Figure 1.11: Comparison of primer-template DNA bound to four DNA polymerases.
A - Taq DNA polymerase bound to DNA (co-crystal structure); B - the binary complex of HIV-1 RT and DNA (co-crystal structure); C - the model of DNA bound to RB69 gp43 (homology model); D - the ternary complex of rat pol β with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

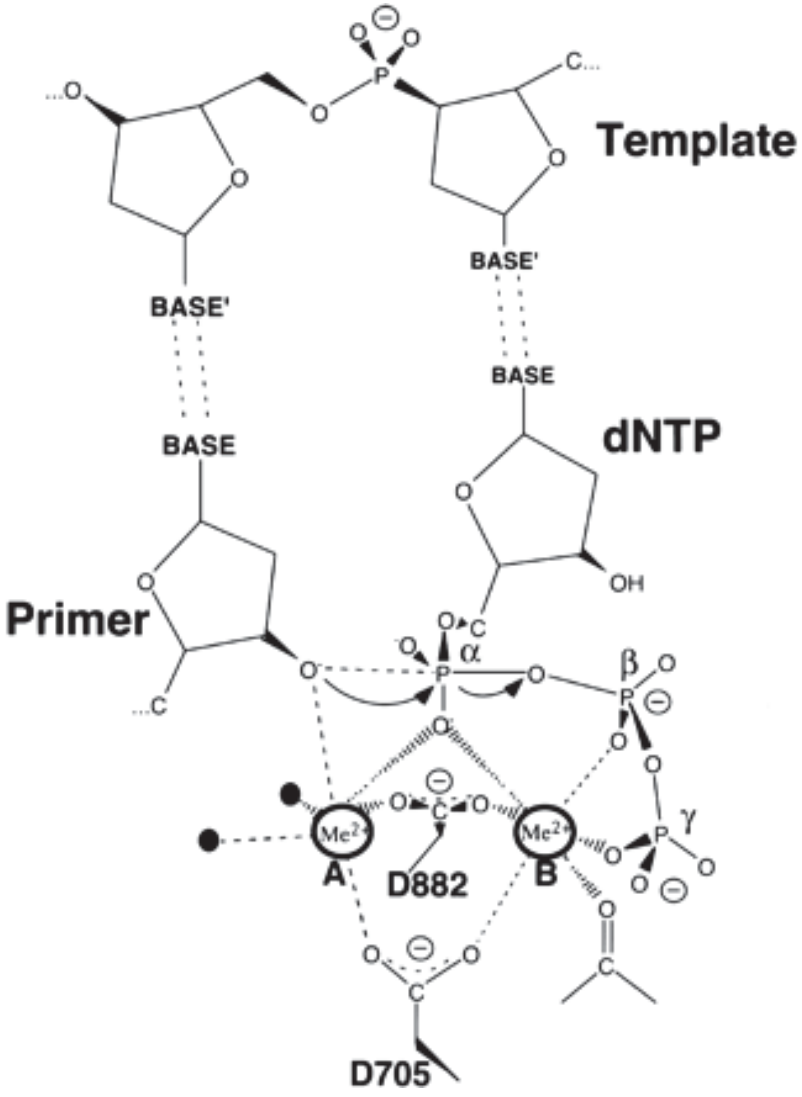


Figure 1.12: Mechanism of DNA polymerization
Polymerisation uses a two-metal ion mechanism that is thought to stabilize the resulting penta-coordinated transition state. (The two essential aspartates are annotated with the *E. coli* DNA polymerase I numbers.) Figure reproduced from [323] in accordance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

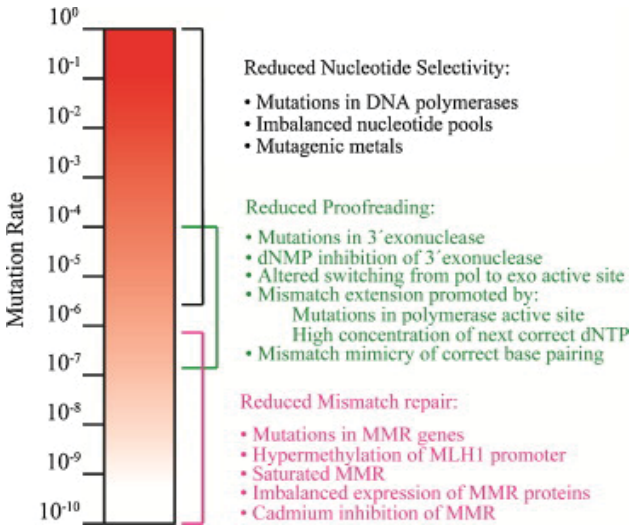


Figure 1.13: Replication fidelity
Nucleotide selectivity, exonuclease activity (proofreading) and mismatch repair all contribute to DNA replication fidelity in series to different degrees. The brackets indicate the magnitude of the contribution the different processes can make and examples of defects 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

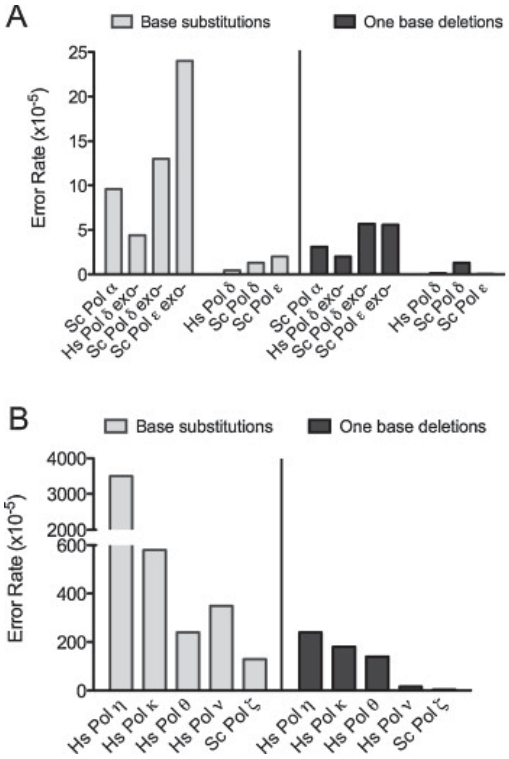


Figure 1.14: Fidelity of different DNA polymerases

"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

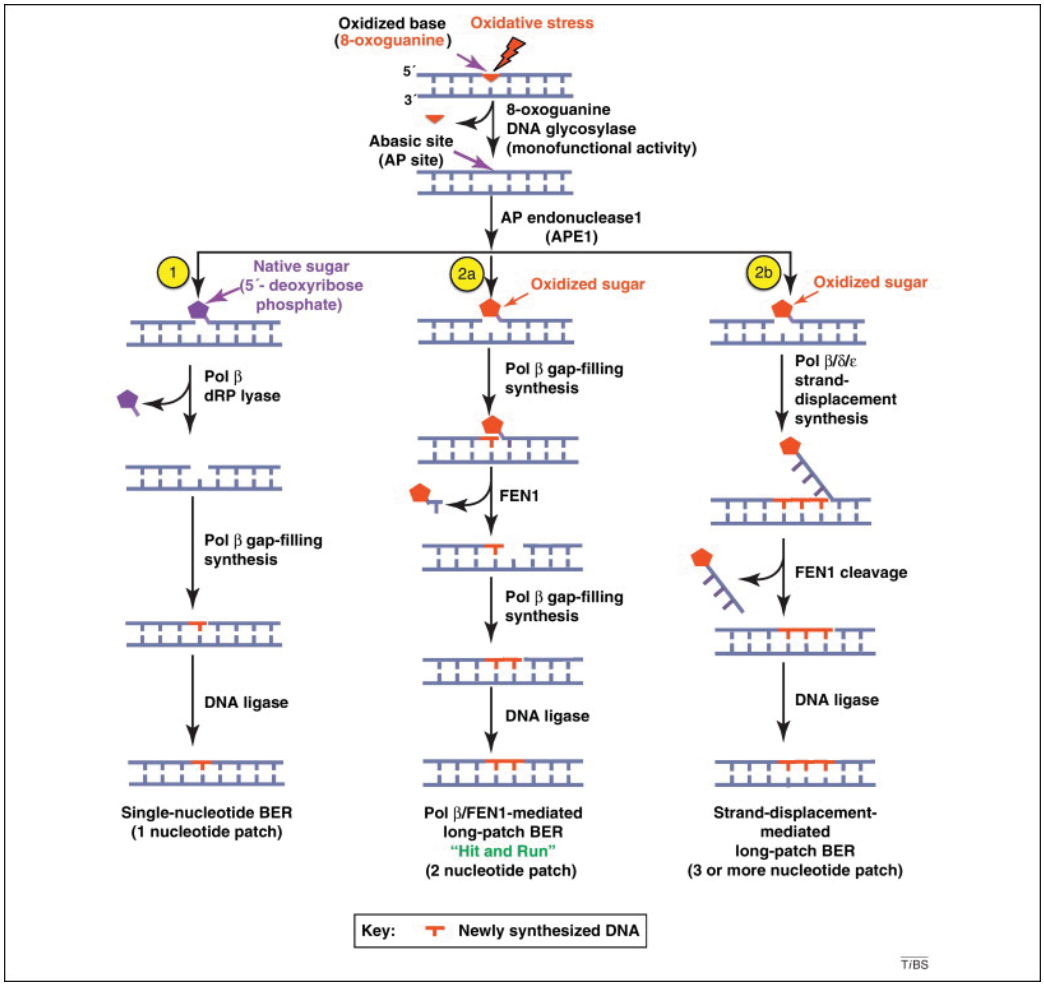


Figure 1.15: Base excision repair (BER) of oxidized DNA base lesions
BER demonstrated using 8-oxoguanine (8-oxoG) as an example. 8-oxoG is removed by the DNA glycosylase 8-oxoguanine DNA glycosylase (OGG1) leaving an basic (AP) site. AP endonuclease 1 (APE1) subsequently incises at the 5' side of the 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

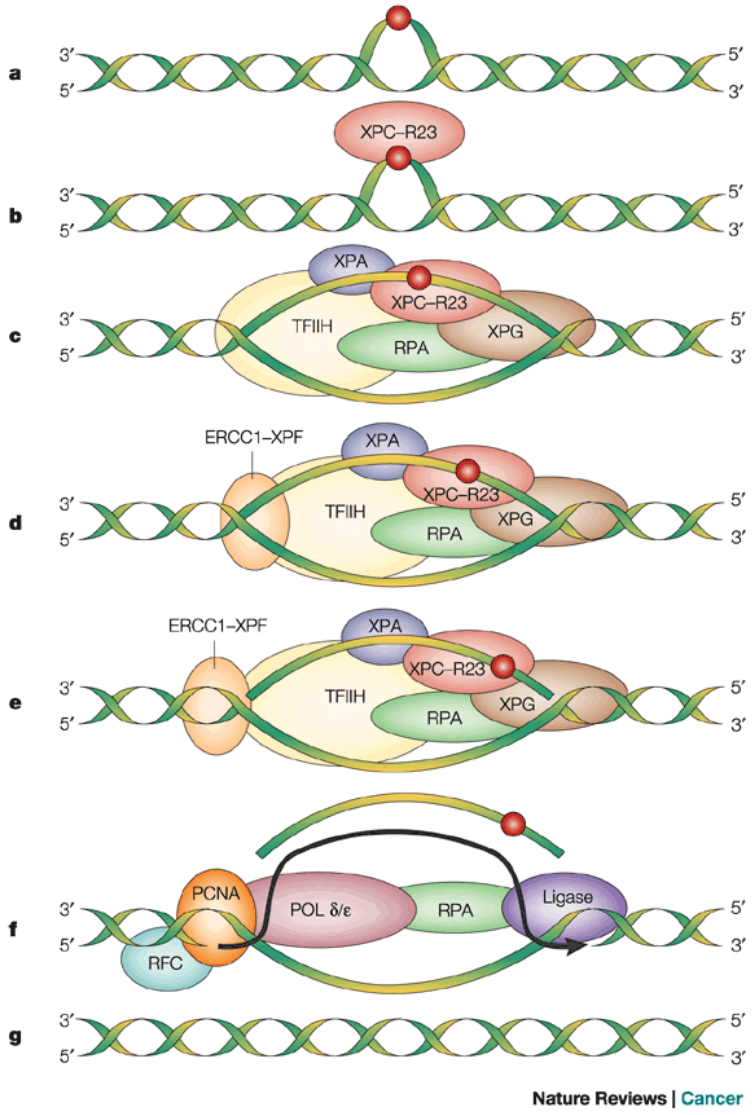


Figure 1.16: Nucleotide excision repair (NER)
A - Nucleotide excision repair (NER) repairs damaged DNA bases that distorts the DNA helix structure (such as photoproducts resulting from UV exposure). B - The damage is recognized by XPC (bound to HHRAD23B). C - Binding of the XPC-HHRAD23B heterodimer is 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

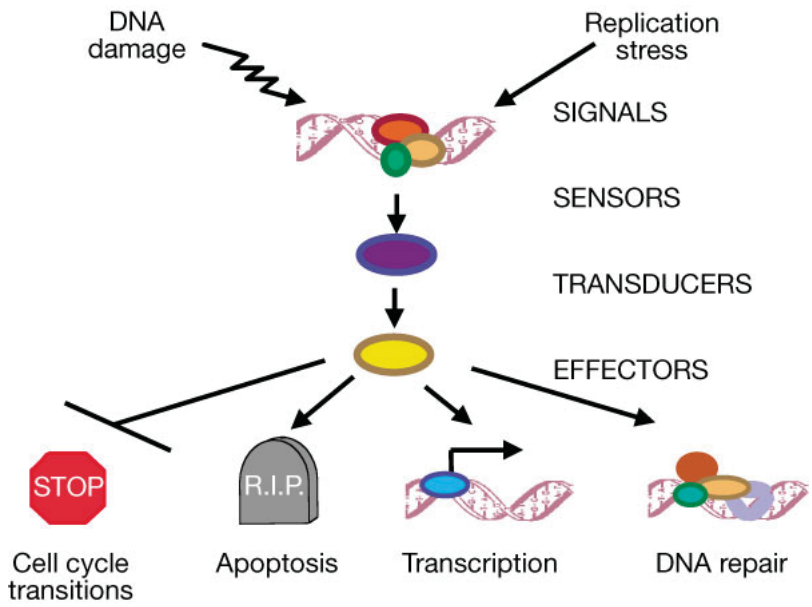


Figure 1.17: A general outline of the DNA damage signal transduction pathway
Arrows represent activating events and perpendicular ends represent an inhibitory event. The stop sign depicts cell-cycle arrest and a tombstone signifies apoptosis. DNA damage-induced transcription is represented by the helix 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.

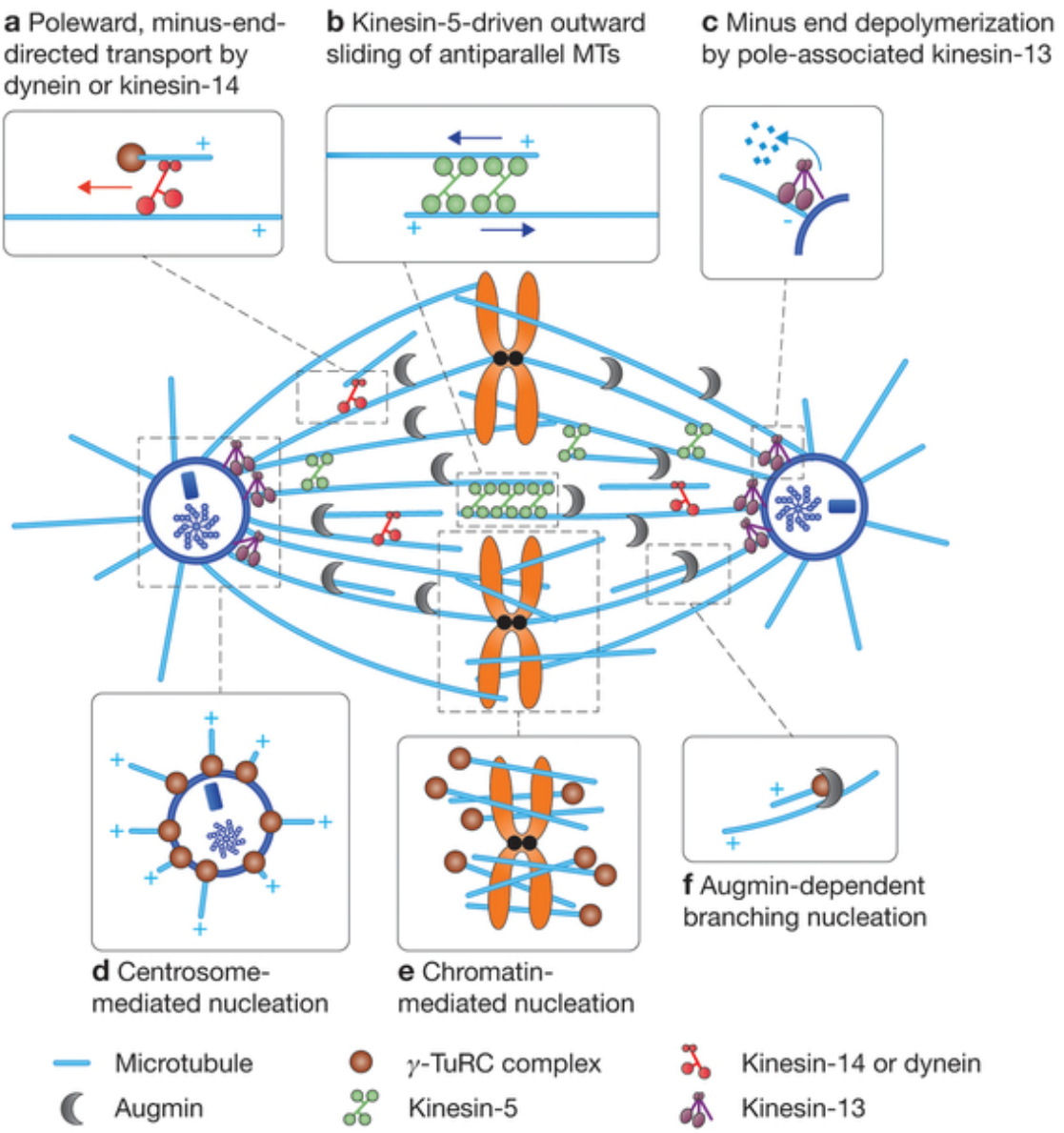


Figure 1.18: The mitotic spindle
The mitotic microtubule (MT) spindle assembles to separate the chromosomes. Spindle architecture depends on molecular kinesin and dynein motors. A - minus-directed motors can slide MTs poleward and contribute to MT-clustering at spindle poles. B - Kinesin-5 motors can 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)

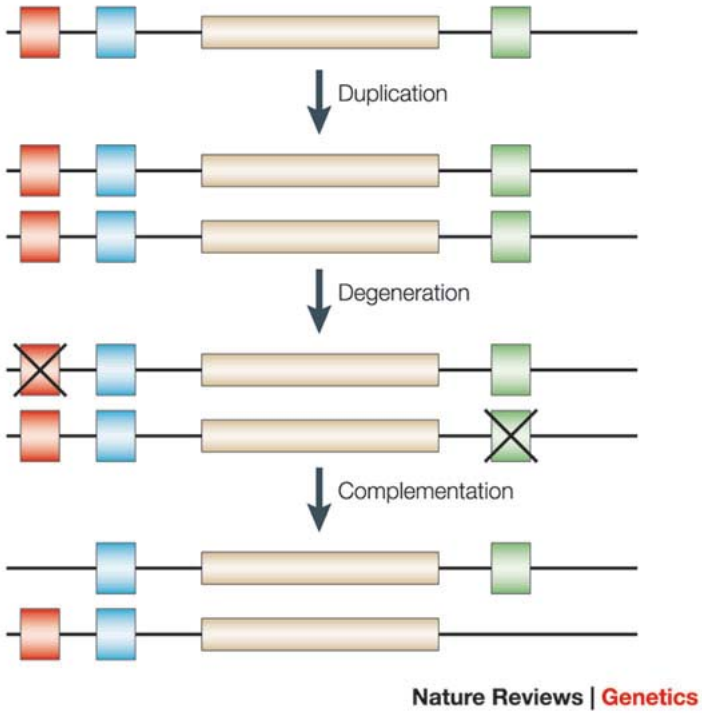


Figure 1.19: Gene duplications: the duplication-degeneration (DDC) model
The duplication-degeneration (DDC) model relies on complementary degenerative changes in two duplicated genes in a way that the two together retain the original function. The coloured boxes represent cis regulatory elements, but 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].

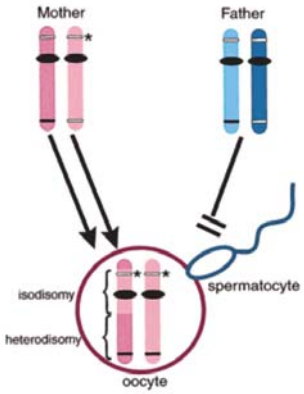


Figure 1.20: Uniparental Disomy - A special case of aneuploidy
An example of uniparental disomy: a non-disjunction event in maternal meiosis leads to the transmission of both copies of a particular chromosome pair to the gamete, which is then fertilised by a spermatocyte that is nullisomic for the pair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

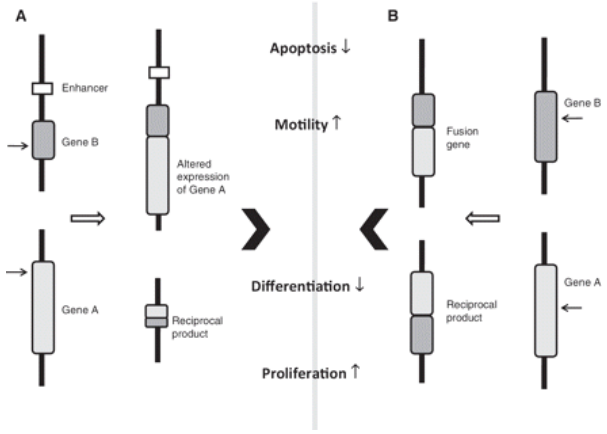


Figure 1.21: Consequences of chromosomal translocations
Chromosomal translocations can result in the placement of genes near different regulatory elements (A) or in aberrant gene fusions (B). Reproduced from [534] with permission from the publisher.

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-

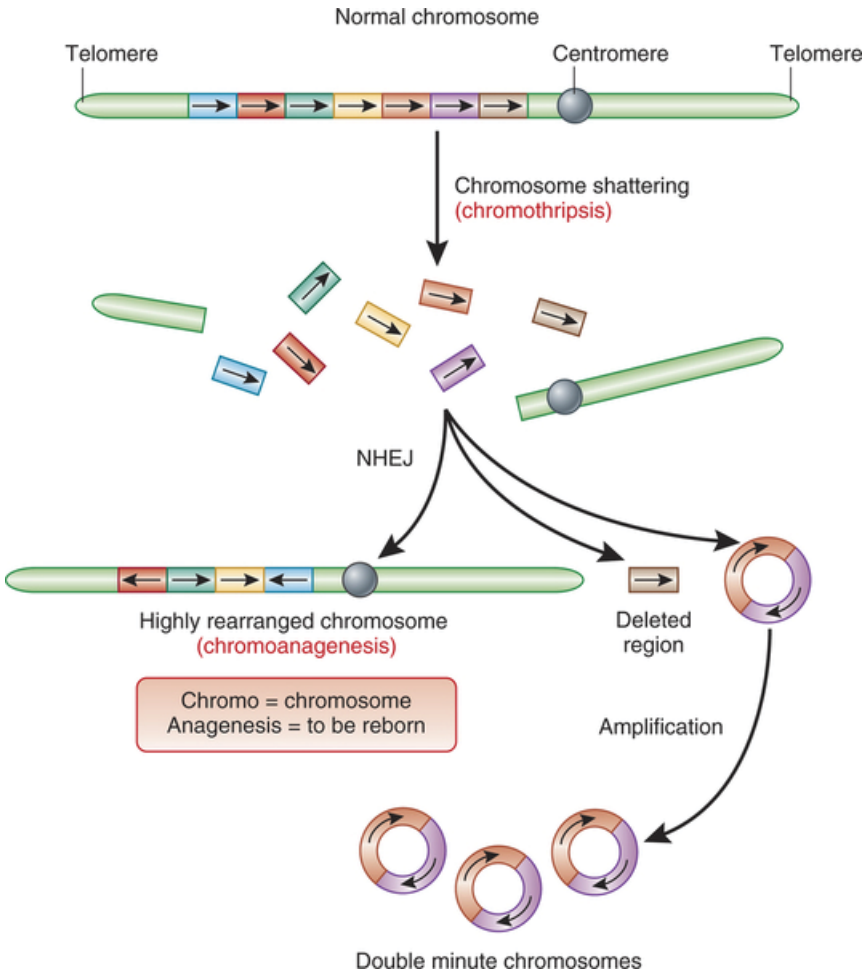


Figure 1.22: Chromothripsis

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.

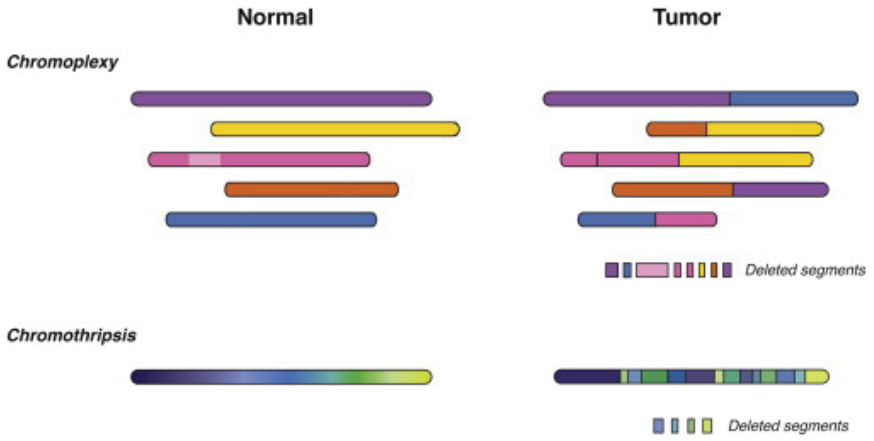


Figure 1.23: Chromoplexy and Chromothripsis
Schematic representations of genomic rearrangements found in Chromoplexy (top) and Chromothripsis (bottom). Reproduced from [536] with the permission of the publisher.

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-

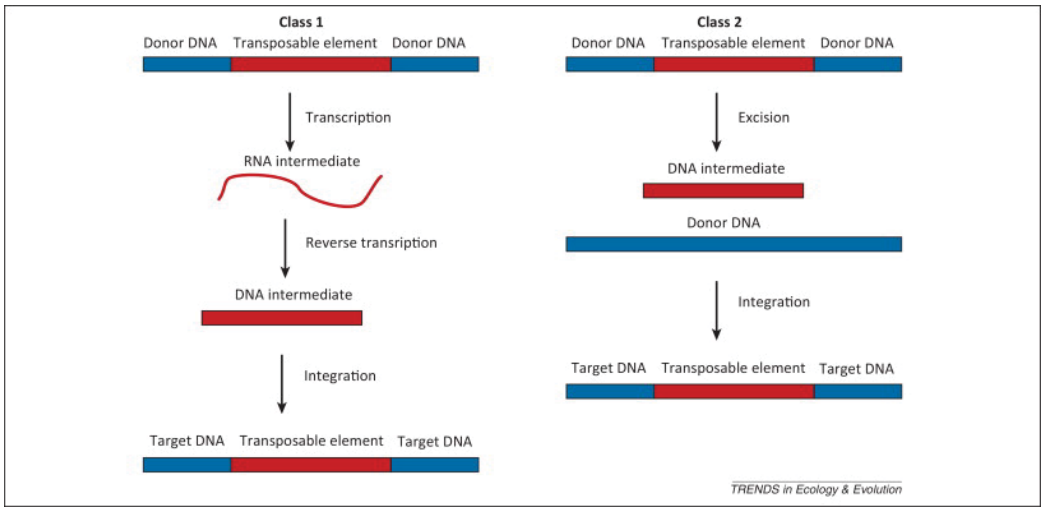


Figure 1.24: Classes of DNA transposons
There are two types of transposable elements: retrotransposons (Class I) and DNA transposons (Class II). Class I move via an RNA intermediate (“copy and paste”), while the latter excise themselves from the DNA (“cut and paste”).
Reproduced from [548] with permission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

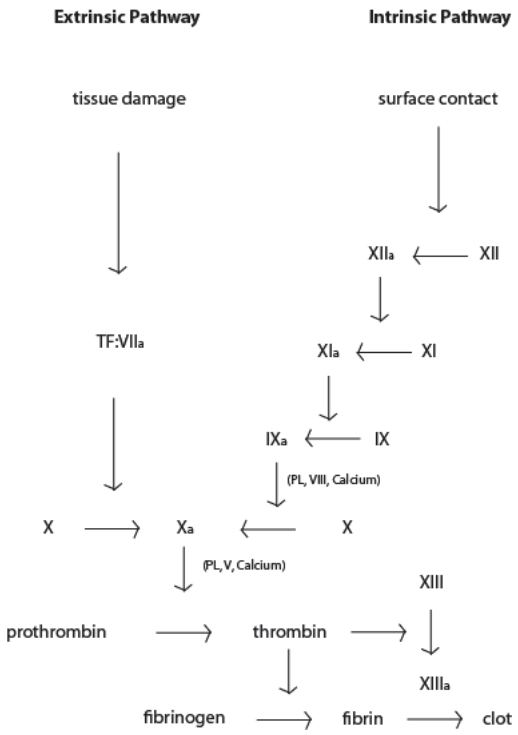


Figure 1.25: Blood clotting cascade

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

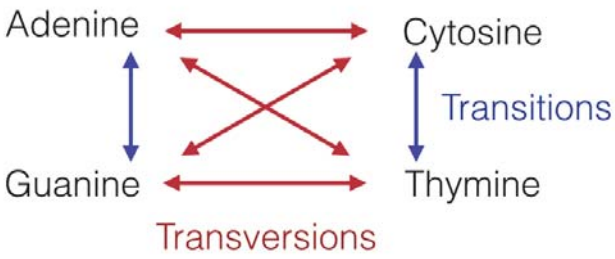


Figure 1.26: Transitions and Transversions

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

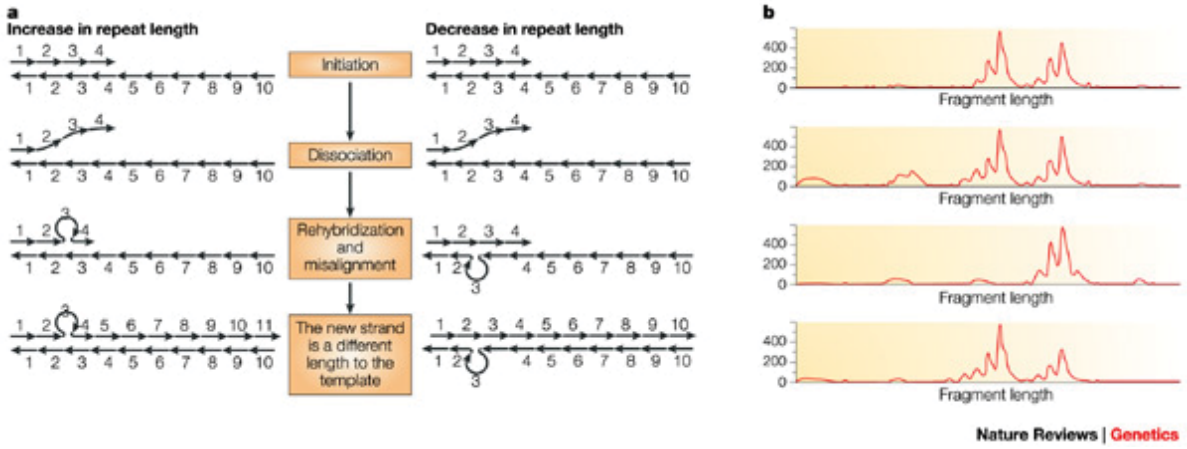


Figure 1.28: Replication slippage
Replication slippage involves the denaturation of the nascent strand from the template followed by missalignment during rehybridization. This leads to the new strand having a different length than the template and is especially common in repetitive regions of the genome. 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-

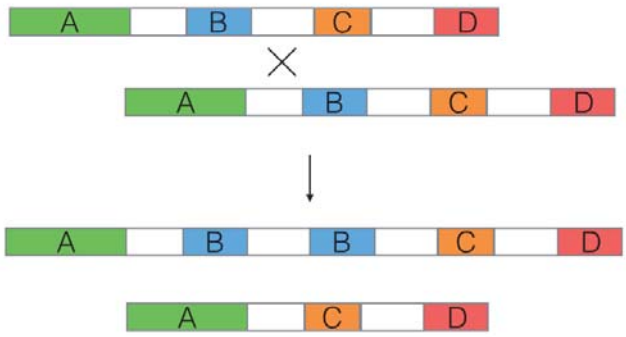


Figure 1.29: Unequal crossovers result in chromosome rearrangements
Crossovers sometimes occur between similar but not equivalent regions of the genome leading to an unequal exchange of DNA between chromosomes.

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

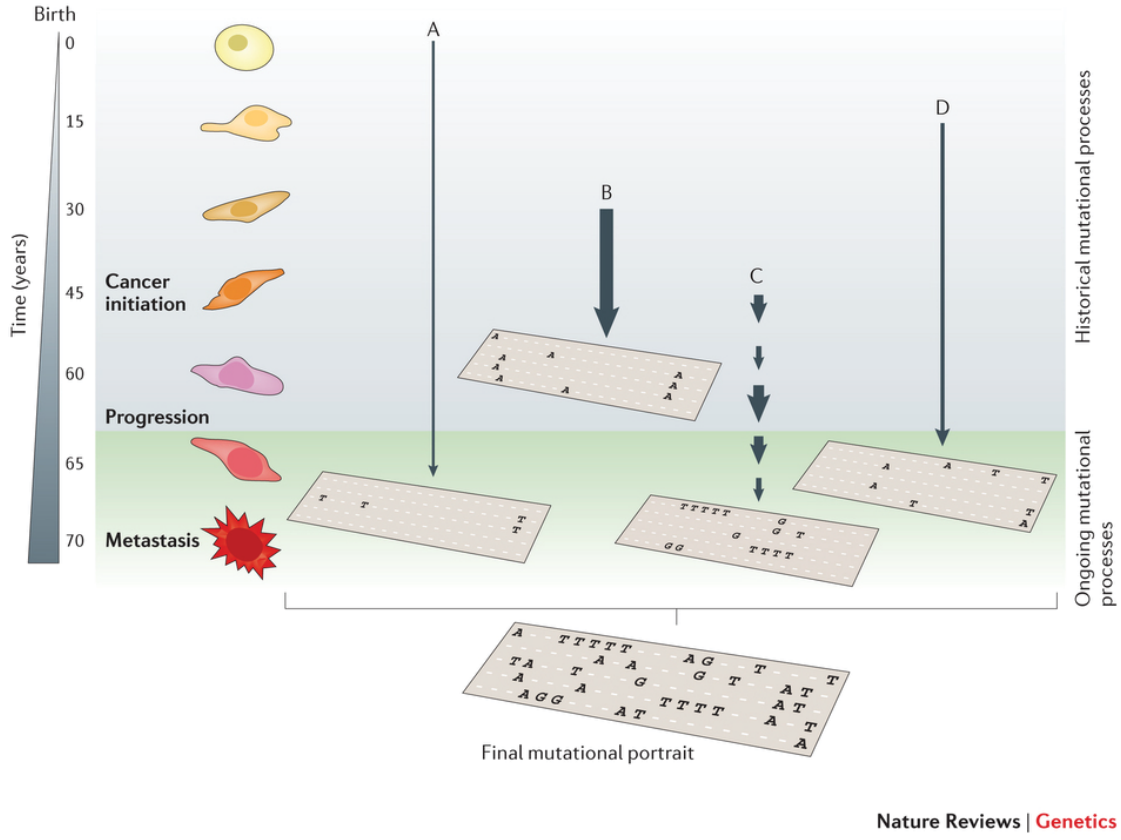


Figure 1.30: Mutational signatures leave their marks on the genome
A schematic of how different mutational processes leave a characteristic imprint on a genome. the mutational patterns generated, length of exposure and intensity of the mutagenic process can vary highy which is reflected in the final 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 $\times$ 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

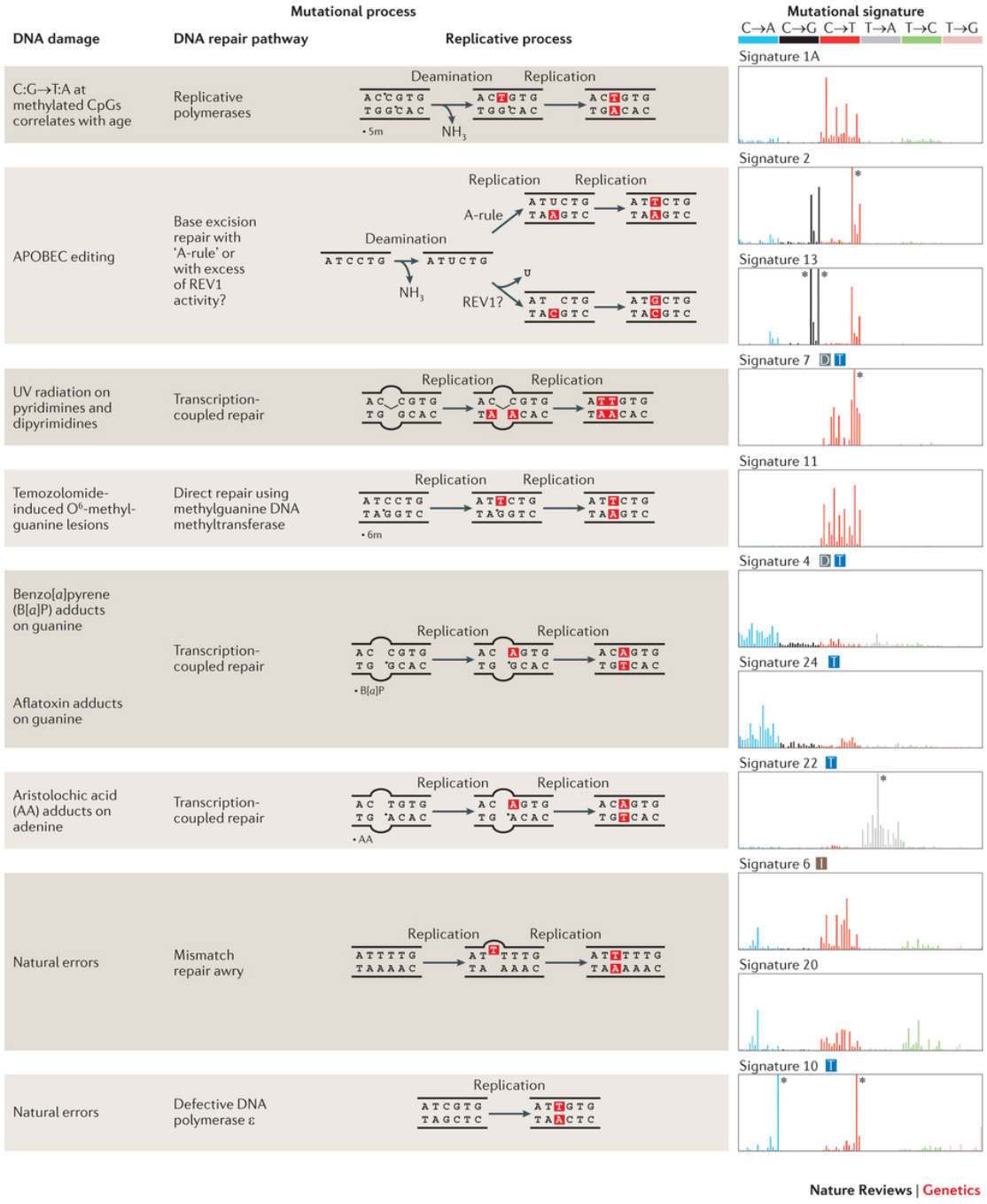


Figure 1.31: Summary of known mutational signatures
Reproduced from [700] with permission from the publisher.

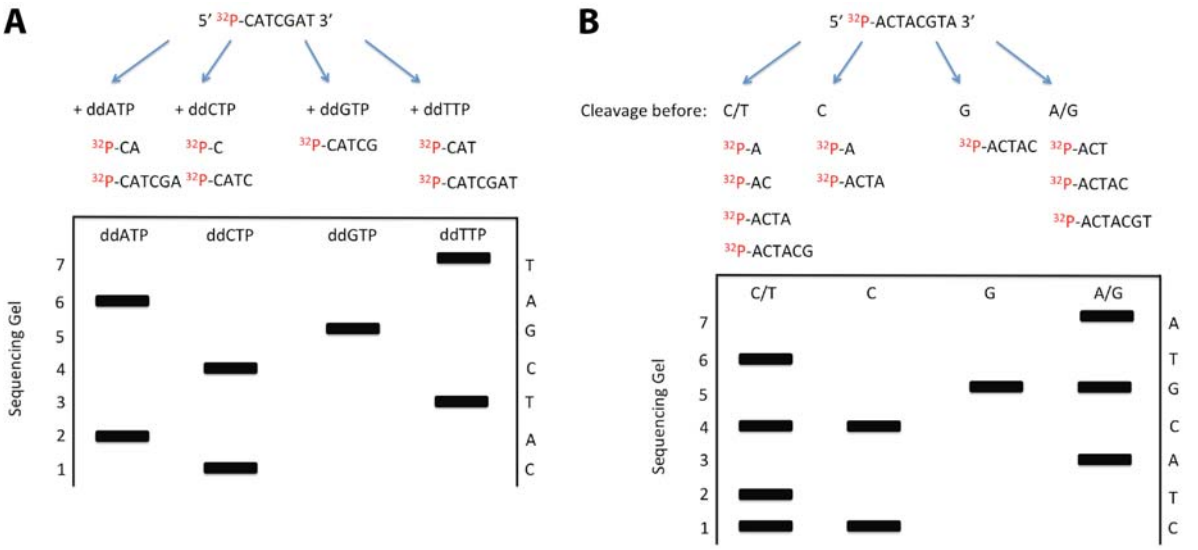


Figure 1.32: Early sequencing techniques: Gilbert and Sanger

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

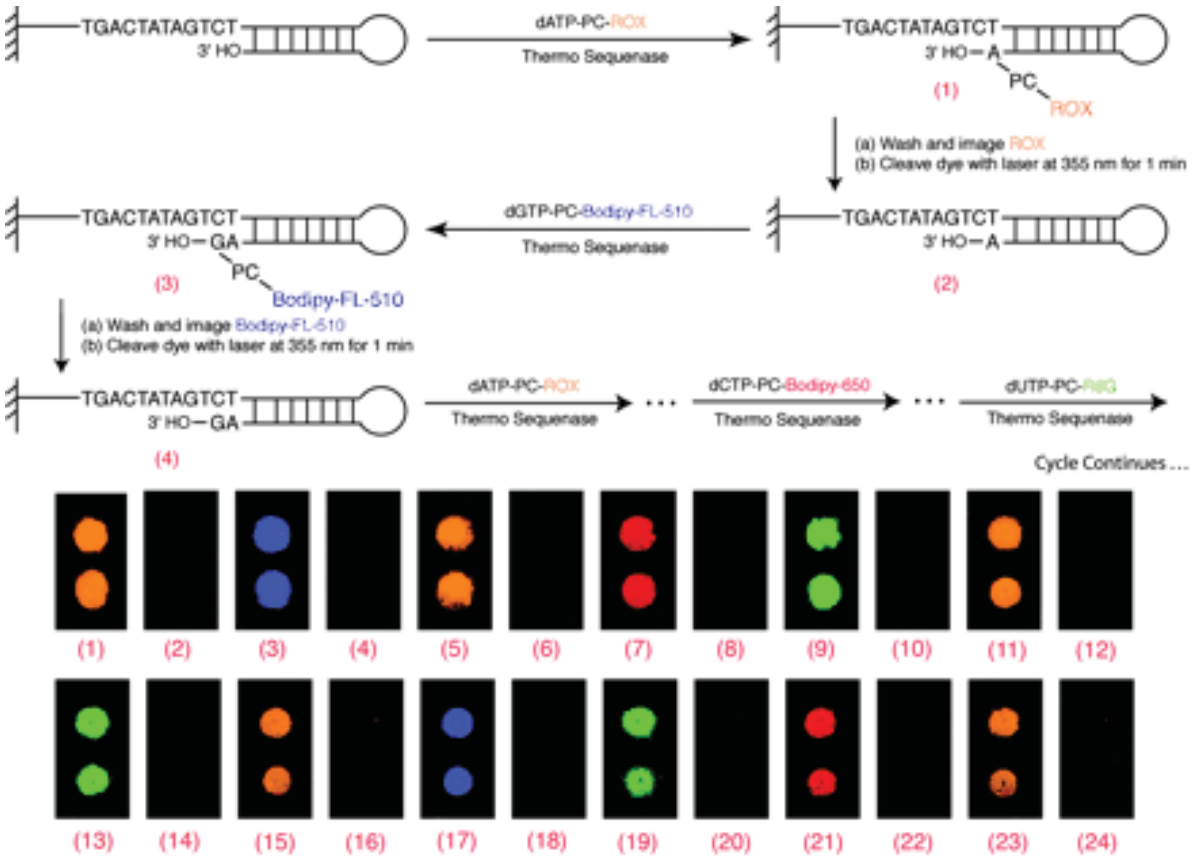


Figure 1.33: Sequencing by synthesis
Schematic of Sequencing by Synthesis (SBS) | 1) Incorporation of a fluorescent dATP-PC-ROX, after washing and imaging 2) the terminator is photo-cleaved. 3) Next, dGTP-PC-Bodipy-FL-510 is incorporated, excess nucleotides washed off and the fluorophore imaged. 4) This 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.

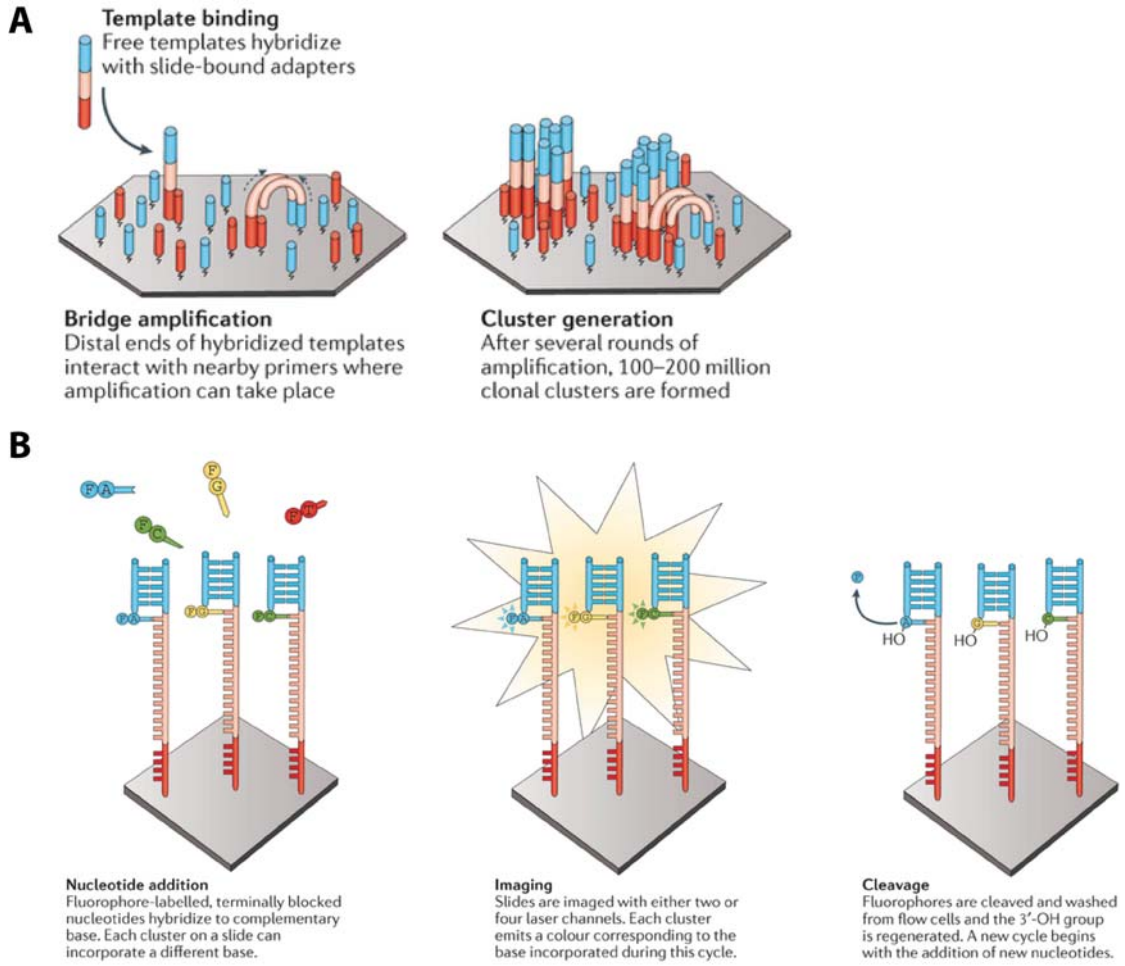


Figure 1.34: Solid-phase bridge amplification and sequencing by synthesis (Illumina)
A| In solid-phase bridge amplification, fragmented DNA is ligated to adapter sequences and bound to a primer immobilized on a solid support, such as a patterned flow cell. The free end can interact with other nearby 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.

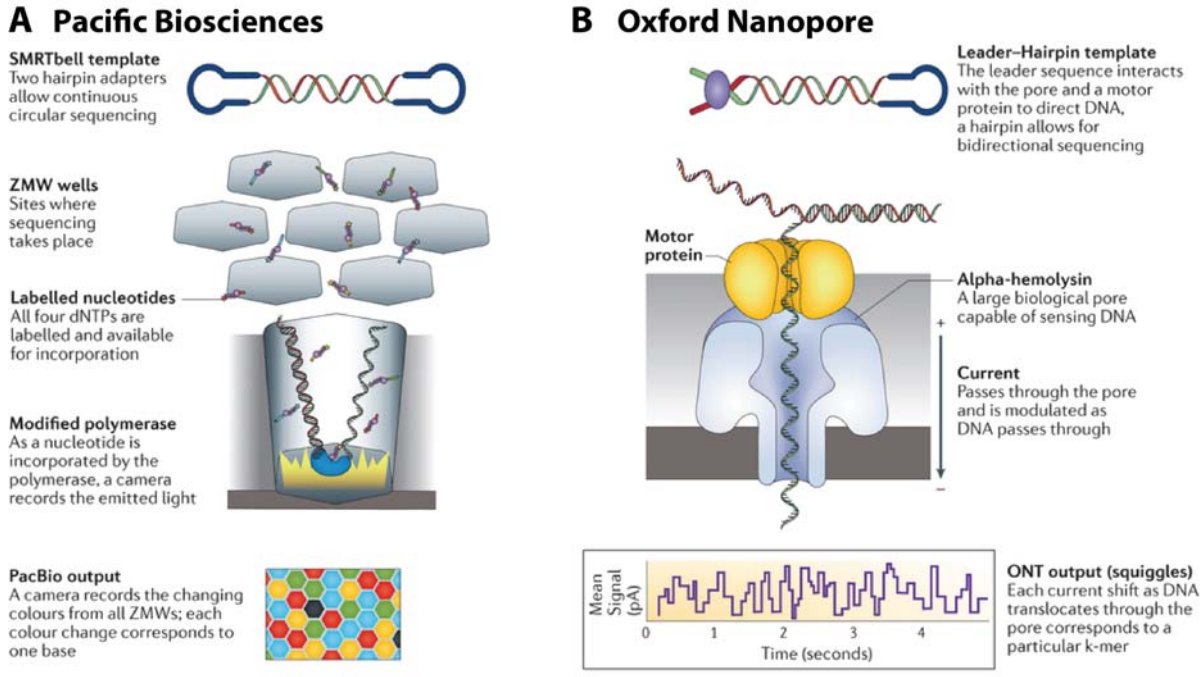


Figure 1.35: Third-generation Sequencing Techniques

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

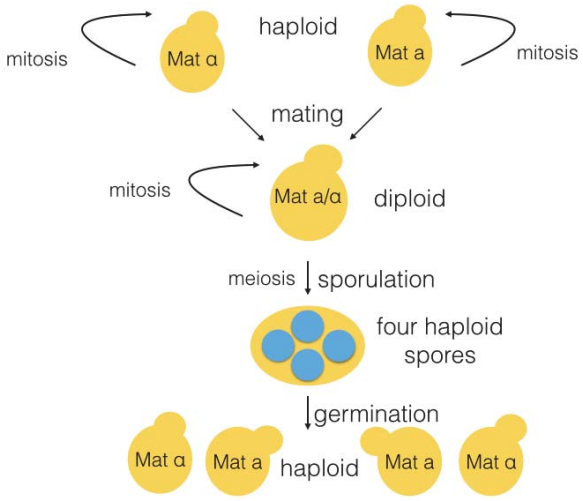


Figure 1.36: Life cycle of the budding yeast *Saccharomyces cerevisiae*
Haploid and diploid yeast cells can reproduce by mitosis. A haploid Mat a and a haploid Mat α can mate and form a diploid. A diploid cell can undergo meiosis and generate four haploid spores.

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.

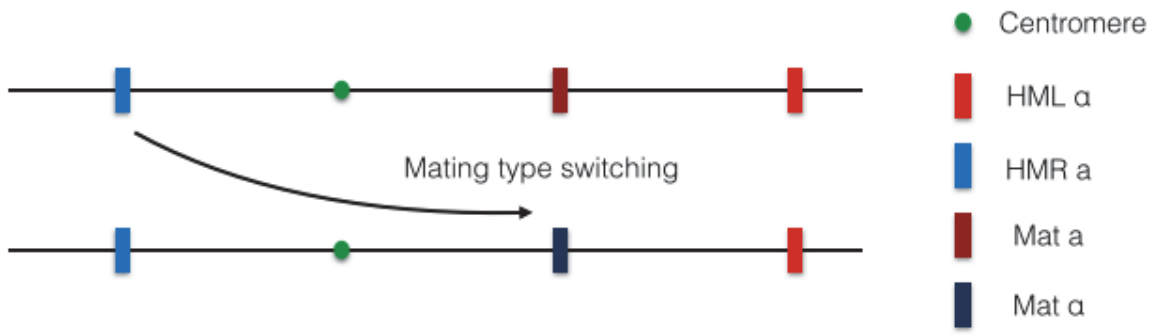


Figure 1.37: The budding yeast mating type locus

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
*pol2*Δ	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.