Understanding the biological processes generating the mutational spectra observed in genomes

Shuvro Prokash Nandi

Division of Cancer Genetics
School of Medicine
Cardiff University

A Thesis submitted to Cardiff University for
the degree of Doctor of Philosophy

Ph.D. 2018

Acknowledgments

Firstly, I would like to express my sincere gratitude to my supervisor, Professor Simon Reed for giving me the opportunity to undertake this Ph.D. and for his invaluable advice, guidance and unwavering support throughout. I would also like to thank my second supervisor Professor Jeremy Cheadle for supporting me on aspect of these Ph.D. studies.

I would like to thank members of the lab, particularly Dr. Patrick Van Eijk, without whom the bioinformatic analysis would be impossible, and whose contribution to this projects has been huge and should be fully acknowledged for his guidance, and invaluable discussions and ideas.

I am also grateful to be accompanied by Dr. Hamed Aula for his special kebab and Dr. Wenbin Wang for making fun around me. A special thanks to Felix Dobbs for delicious foods and posh British culture during the course of my studies.

I also wish to extend my thanks to former members of the lab, Dr Mark Bennett, Dr Katie Evans, Dr James Powell and Dr Richard Webster for their help and support during the early years of my PhD.

I would like to thank Trish, our lab attendant, for tidying up after me and sharing my time in the lab like a mom.

Special thanks to Commonwealth Scholarship Commission for providing me all the financial support to study in UK.

Lastly but more importantly I would like to acknowledge the close support of my parents, family and friends over the past four years.

DECLARATION

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed …… (candidate) Date ……

STATEMENT 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of Ph.D.

Signed …… (candidate) Date ……

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed …… (candidate) Date ……

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed …… (candidate) Date ……

STATEMENT 4: PREVIOUSLY APPROVED BAR ON ACCESS

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loans after expiry of a bar on access previously approved by the Academic Standards & Quality Committee.

Signed …… (candidate) Date ……

Summary

Maintaining genome stability is essential for life. Since DNA is constantly exposed to the deleterious effects of both the internal and external cellular environment, mechanisms have evolved to sense and repair the consequent genetic damages within the chromatin environment. Repair of UV-induced DNA damage requires chromatin remodeling. How repair of this damage is organised and initiated remains largely unknown. Previous work demonstrated in yeast cells that Global Genome Nucleotide Excision Repair (GG-NER) in chromatin is organized into domains in relation to open reading frames. In this thesis, by examining DNA damage-induced changes in the linear structure of nucleosomes at these sites, I show how chromatin remodeling is initiated during GG-NER. In undamaged cells, I found that the GG-NER chromatin-remodeling complex occupies chromatin and establishes the nucleosome structure at genomic locations, now referred to as GG-NER complex binding sites (GCBS's). These sites are frequently located at genomic boundaries that delineate chromosomally interacting domains (CIDs), which represent regions of higher-order nucleosome-nucleosome interaction. Repair in chromatin is initiated from these sites by the GG-NER complex-dependent disruption of dynamic nucleosomes that flank GCBS's, demonstrating the importance of this mechanism to the efficient removal of DNA damage by NER. I then studied how this affects the pattern of mutations acquired in the genome, establishing a novel workflow to catalogue the acquired mutations in yeast cells treated with or without UV radiation. Additionally, the use of NMF for *de novo* extraction of mutational signatures from these mutational catalogues, successfully decomposed the biological processes of mutagen exposure and DNA repair deficiencies. I showed that the genomic features that are important for repair organisation determine the location and types of mutations within genome. These studies may explain how novel cancer genes involved in chromatin modification drive tumorigenesis.

Abbreviations

A - Adenine base

AA - Aromatic amines

AID - activation-induced deaminase

AGT - O6-alkylguanine-DNA alkyl-transferase

ALL - Acute lymphoblastic leukaemia

AML - Acute myeloid leukaemia

AP site - Apurinic or apyrimidinic site

APOBEC - Apolipoprotein B mRNA editing enzyme catalytic polypeptide

ARR - Access-repair-restore

ATM and ATR - Ataxia telangiectasia mutated and. ATM and RAD3-related

bp - Base pairs

BER - Base excision repair

BSA - Bovine serum albumin

BRCA1 and BRCA2 - BReast CAncer gene

C - Cytosine base

CID - Chromosomally interacting domains (Also known as Micro-C boundaries)

ChIP - Chromatin immunoprecipitation

Chip - DNA microarrays

ChIP-Seq - Chromatin immunoprecipitation and sequencing

CLL - Chronic lymphocytic leukaemia

CPD - Cyclobutane pyrimidine dimer

CS - Cockayne Syndrome

CSB - Cockayne Syndrome group B

DBs - Double base substitution signatures

DDR - DNA-damage response

DNA - Deoxyribonucleic Acid

dRP - Deoxyribose phosphate

dNTP - Deoxy-nucleoside triphosphate

DSB - Double strand break

DSBR - Double strand break

EDTA - Ethylenediaminetetraacetic acid

EXO1 - exonuclease 1

FA - Formaldehyde

G- Guanine base

Gcn5 - Histone acetyl-transferase in yeast

GG-NER - Global genome nucleotide excision repair

GCBSs - GG-NER Complex Binding Sites

GRF - General regulatory factor

HAT - Histone acetyltransferase

HR - Homologous recombination

Htz1 – Histone variant H2A.Z

HNPCC - Hereditary nonpolyposis colorectal cancer

Indels - Insertions and deletions

ICGC - International Cancer Genome Consortium

IDs - Indel signatures.

IARC - International Agency for Research on Cancer

ICLs – Inter-strand crosslinks

IN - Input samples

IP - Immunoprecipitated samples

IR - Ionizing radiation

LM-PCR - Ligation-mediated polymerase chain reaction

MMR - Mismatch repair

MNase-Seq - Micrococcal nuclease digestion and sequencing

MUTYH - mutY DNA glycosylase gene

NA - N-nitrosamines

NER - Nucleotide excision repair

NGS - Next generation sequencing

NHEJ - Non-homologous end joining

NMF - Nonnegative matrix factorization

RON - Reactive Nitrogen species

NFR - Nucleosome free regions

ORF - open reading frames

PARP1 - Poly [ADP-ribose] polymerase 1

PAH - Polycyclic aromatic hydrocarbons

PBS - Phosphate-buffered saline

PCAWG - Pan-cancer analysis of whole genomes

PCNA - Proliferating cell nuclear antigen

PCR - Polymerase chain reaction

PMSF - Phenylmethylsulphonyl fluoride

POLE – DNA polymerase Epsilon

TOP1 - DNA topoisomerase 1

PTMs - Post-translational modifications

qPCR - Quantitative real-time polymerase chain reaction

RNA - Ribonucleic Acid

rNTPs - Ribo-nucleoside triphosphate

ROS - Reactive oxygen species

RSS - Residual sum of squire

SSBs - Single base substitutions signature

SDS - Sodium dodecyl sulphate

SGD - *Saccharomyces* Genome Database

SNVs - single nucleotide variations

SPN - Strongly positioned nucleosomes

SSB - Single strand break

SSBR - Single strand break repair

SWI/SNF - SWItch/Sucrose Non-Fermentable, nucleosome remodeling complex

T-Thymine base

TAE - Tris-acetate-EDTA

TBST - Tris-buffered saline tween

TC-NER - transcription-coupled nucleotide excision repair

TCGA - The Cancer Genome Atlas

TE - Tris-EDTA buffer

TES - Transcription end sites

TLS - Translesion synthesis

TSS - Transcription start sites

U - Uracil base

UV - Ultraviolet light

UVA - Ultraviolet light A

UVB - Ultraviolet light B

UVC - Ultraviolet light C

$UV^{S}S$ - UV-Sensitive Syndrome

WES - Whole-exome sequences

WGS - Whole genome sequencing

WSS – Within-sum of squares

WTSI - Wellcome Trust Sanger Institute

WT – Wild-type

YPD - Yeast Extract Peptone Dextrose media

YLE - Yeast lytic enzymes

XR-seq - Excision repair sequencing

XP- Xeroderma Pigmentosum

XRCC - X-Ray Repair Cross Complementing

3D-DIP-Chip - DNA damage detection by DNA immunoprecipitation on microarray chips

5mC - 5-methyl cytosine

5-hmC - 5-hydroxymethyl cytosine

6-4PP - pyrimidine(6-4)pyrimidone photoproduct

8-oxoG - 7,8 dihydro-8-oxoguanine

4-NQO - 4-Nitroquinoline 1-oxide

Contents

Chapter I

Introduction

Contents

1.1 DNA and Chromatin

DNA is comprised of two strands that coil around each other in antiparallel to form a double helical macromolecule. The basic unit of DNA is the nucleotide, which is composed of one of four nitrogen-containing bases (adenine (A), thymine (T), cytosine (C), or guanine (G)), a deoxyribose sugar, and a phosphate group. According to base pairing rules, hydrogen bonds exist between A & T and C & G to form the double-stranded DNA structure. In eukaryotes, DNA wraps around histone proteins, forming nucleosomes and thus creating the so-called "beads on a string" structure. The next level of organisation is thought to involve a string of nucleosomes folded into a 30nm diameter chromatin fibre. These fibres then undergo further folding into even higher-order structures forming chromatin and chromosomes (Figure 1.1).

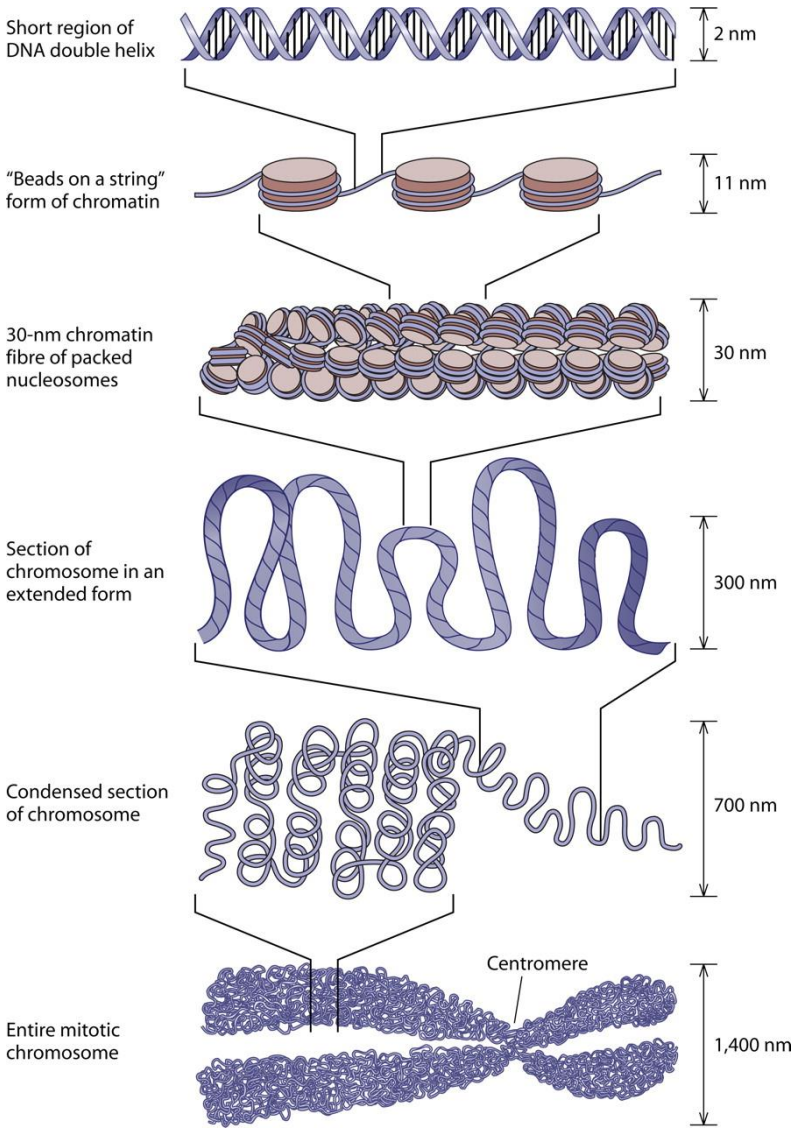


Figure 1.1: Chromosomes are composed of DNA tightly-wound around histones in a hierarchical folding pattern. Figure reproduced from (Jansen and Verstrepen 2011).

1.2 DNA Damage, DNA Repair and Mutation

DNA is the molecule of heredity, and it is highly prone to damage to its structure within the chromatin environment, due to the deleterious effects of continuous normal cellular metabolic processes as well as to genotoxic stresses such as ultraviolet (UV) radiation or chemical damage from the environment (Friedberg et al. 2005). Thousands of lesions occur daily in the DNA of each of our cells. DNA damage can cause disruption of cell division and altered gene regulation, while defective DNA repair can introduce DNA mutations that may alter the genetic information within the cell (Polo and Almouzni 2015). Therefore, repair of damaged DNA is fundamental to genome stability.

In the following section, I will describe the biological processes that can induce damages in genomic DNA.

1.2.1 Biological processes of DNA damage formation

As mentioned in the previous section, DNA can be damaged by either exogenous or endogenous factors (Figure 1.2), which can be of different origin. **Exogenous** factors, such as environmental, physical and chemical agents, damage DNA by directly interacting with DNA molecules (Sancar et al. 2004). Most of the exogenous factors, such as UV and ionizing radiation (IR), alkylating agents, and crosslinking agents can damage DNA by altering the structure of the DNA. Examples include dimerization of bases (Sinha and Häder 2002), formation of bulky adducts in DNA (Roos and Kaina 2006) or DNA strand breaks (Mehta and Haber 2014). On the other hand, most of the **endogenous** factors damage DNA by hydrolytic and oxidative reactions with water and reactive oxygen species, respectively, that are naturally present within cells (Lindahl 1993). This unavoidable tendency for the interaction of DNA with its adjacent surrounding molecules, can fuel the development of hereditary diseases and the formation of sporadic cancers (De Bont and van Larebeke 2004; Hoeijmakers 2009; Valavanidis et al. 2009; Tubbs and Nussenzweig 2017). Damaged DNA does not always result in the formation of mutated DNA; the damage is usually repaired by cellular DNA repair mechanisms leaving the genome unaffected in terms of its informational content. However, mutations (i.e. permanent, heritable changes in DNA sequence) may arise when repair is inaccurate or when a replication fork passes through a damaged site (Paulovich et al. 1997; Lindahl and Wood 1999). Most DNA (and RNA) polymerases will stall at such sites, but in a damaged cell, specialised translesion polymerases may read through damaged sites, possibly incorporating incorrect nucleotides as they do so. Thus, the cell completes DNA

replication, but at the cost of generation of mutations and the possible loss of genetic information (Waters et al. 2009; Sale et al. 2012).

In short, the DNA damage resulting from exposure to endogenous and environmental mutagens becomes a substrate for specific DNA repair pathways, which collectively determine the formation of mutations in the genome as discussed in the subsequent section.

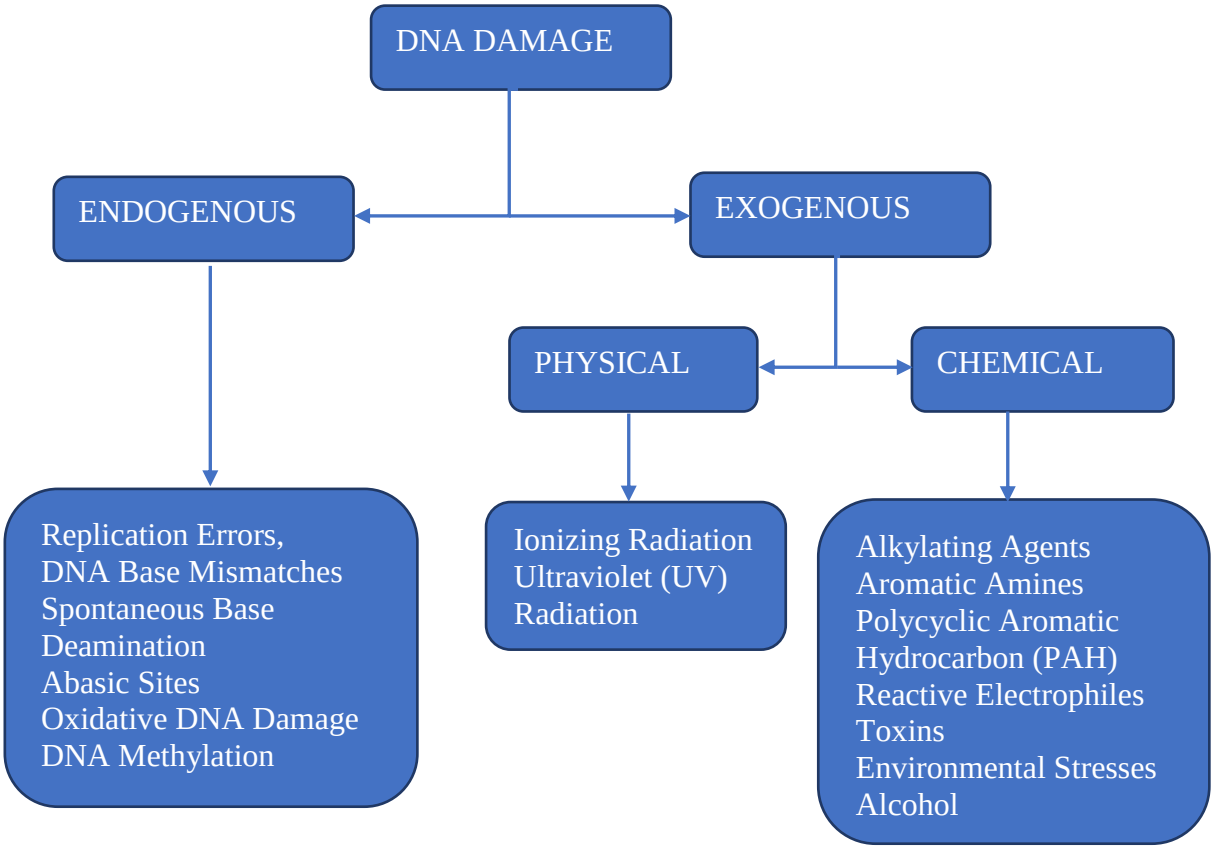


Figure 1.2: The different types of DNA damaging agents. Both endogenous and exogenous mutagens can damage DNA.

1.2.1.1 Endogenous DNA damages

1.2.1.1.1 Replication errors & DNA base mismatches

DNA replication is an essential biological process to ensure the accurate propagation of genetic information. Replication itself does not damage DNA, but it can result in the generation of mutations in the cells (Pray 2008). During replication, high fidelity DNA polymerases ensure the structural and biochemical stability of the genome and maintain genomic stability by ensuring the insertion of a correct complementary deoxynucleotide opposite the template base (Kunkel 2004; Swan et al. 2009). However, base substitutions and single base insertion or deletion errors still occurs at a rate of 10^{-6} to 10^{-8} per cell, per division in yeast (Kunkel 2004). Even though replication is a highly accurate process,

it is however, continuously affected by many types of DNA damages (Sale et al. 2012). Additionally, a specific family of DNA polymerases can tolerate certain types of damage, and is therefore able to replicate through damaged DNA, but sometimes with the cost of generating somatic mutations (Sale et al. 2012). This type of damage tolerance pathway will be described later.

Replication errors also accumulate from strand slippage events at repetitive sequences, causing insertions and deletions of nucleotides that can potentially change the DNA reading frame (Chatterjee et al. 2013). It has been reported that some specific regions in the human genome called 'hotspots' (such as micro- or mini-satellites) are more susceptible to replication error than others (Viguera et al. 2001).

Replicative polymerases can sometimes incorrectly incorporate uracil into the DNA, or end up with compromised fidelity because of the alterations of the relative and absolute concentrations of dNTPs and rNTPs pools, which are the substrates for replicative polymerases, within the cell's environment (Andersen et al. 2005; Kumar et al. 2010; Clausen et al. 2013). These incorrectly paired/incorporated nucleotides that escape proofreading and mismatch repair, may become fixed as mutations in the next round of DNA replication, and are a major source of spontaneous mutagenesis.

1.2.1.1.2 Deamination of DNA Bases

Spontaneous base deamination is one of the major sources of DNA damage within cells, which results in the hydrolytic removal of amine groups from the nitrogen bases. As a result, 5-methyl cytosine (5mC), 5-hydroxymethyl cytosine (5-hmC), cytosine (C), adenine (A) and guanine (G) spontaneously lose their exocyclic amine to become thymine (T), 5-hydroxymethyl uracil (5-hmU), uracil (U), hypoxanthine and xanthine respectively (Figure 1.3A and Figure 1.3B) (Duncan and Miller 1980; Lindahl 1993; Pfeifer et al. 2013). For example, the deamination of cytosine at the native C:G base pairing alters to a U:A base pair in the first round of replication, which in the next round of replication results in a CG:TA mutation. The rate at which these deamination processes occur depends on the types of bases involved. For example, the 5-mC and cytosine both get deaminated frequently, but 5-mC is deaminated more frequently than cytosine (Lindahl 1979; Shen et al. 1994). Consequently, the C>T transition at the CpG island accounts for one-third of the point mutations and is a major source of human disease (De Bont and van Larebeke 2004; Nabel et al. 2011).

In addition to spontaneous deamination of 5-mC, it can deaminate hydrolytically to thymine by the activity of both AID (activation-induced deaminase) and APOBEC (Apolipoprotein B mRNA editing enzyme catalytic polypeptide) family, resulting in C>T mutation (Morgan et al. 2004). The pattern of somatic hypermutations induced by the activity of AID is well-studied, and it predominantly deaminates cytosine that is flanked by 5' purines (Pham et al. 2003). The mutational pattern induced by the APOBEC family of enzymes depends on the subset of these enzymes that act on different sequence contexts. For example, APOBEC1, APOBEC3A and APOBEC3B can give raise to C>T or C>G substitutions at TpCpN trinucleotides (Taylor et al. 2013). In some cancer genomes (See later for details), it has been recently demonstrated that there is direct link between APOBEC deaminases and/or regions of somatic hypermutations (Nik-Zainal et al. 2012). Furthermore, adenine and guanine are spontaneously deaminated to hypoxanthine and xanthine respectively. During replication, hypoxanthine and xanthine preferentially pair with guanine and cytosine, resulting in T>C and C>T substitutions, although these events are rare (Lindahl 1993; Fernández et al. 2009). In addition to the endogenous deamination sources, exogenous exposure to UV radiation, intercalating agents, nitrous acid and sodium bisulfite can in general enhance base deamination rates in the DNA (Friedberg et al. 2005; Chatterjee and Walker 2017).

Taking together, the basic processes that result in or are responsible for the endogenous alterations to nucleotide and are potential sources of naturally occurring variants, and/or mutations. This suggests that the types and patterns of mutation induced by these processes also have a structure and organisation within the genome.

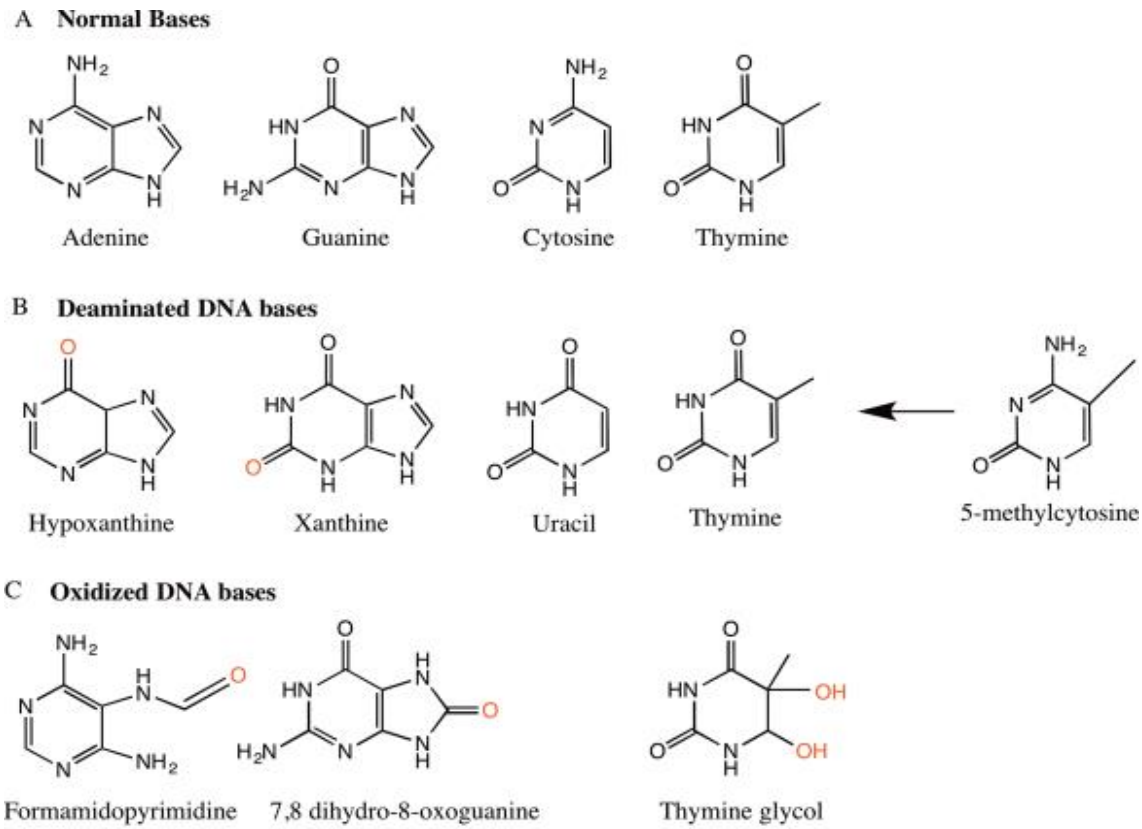


Figure 1.3: The basic chemical structure of common DNA base lesions. A: The normal bases: adenine (A), guanine (G), cytosine (C), and thymine (T). B: Base structure after deamination: hypoxanthine, xanthine, uracil, and thymine C: Base structure after oxidative damage: formamidopyrimidine derivative of adenine, 8-oxo-G, and thymine glycol. (Figure adapted and modified from (Chatterjee and Walker 2017).

1.2.1.1.3 Abasic Sites

Abasic or apurinic/apyrimidinic (AP) sites are one of the most frequent types of endogenous damage in DNA. AP sites can be formed by either spontaneous depurination, or as an intermediate during the base excision repair process. In this process, DNA glycosylase mediates hydrolytic cleavage at the N-glycosyl bond between the base and deoxyribose (Nakamura and Swenberg 1999). In human cells, the spontaneous depurination rate was estimated at 10,000 bases per cell, per day under normal physiological conditions; both extreme pH conditions and high temperatures positively impact their formation (Lindahl 1993). Because of their potential cytotoxicity and mutagenicity, these AP sites must be repaired efficiently. Most AP sites are effectively removed by AP endonucleases that cleave at their 5' end and allow the base excision repair (BER) pathway to repair them. Alternatively, AP sites can be bypassed by damage-tolerance processes and can generate substitutions, insertions, or even single strand break

(SSB) mutations (Nakamura and Swenberg 1999; Alseth et al. 2004; Chan et al. 2013). As different repair pathways are involved in removing AP damages, the final mutational spectra generally depends upon the repair background (Otterlei et al. 2000).

1.2.1.1.4 Oxidative DNA Damage

Oxidative DNA damage is formed as an inevitable consequence of cellular metabolism during respiration. Reactive Oxygen Species (ROS) such as superoxide radicals ($\bullet O_2^-$), hydrogen peroxide (H_2O_2), and hydroxyl radical (•OH) are capable of damaging DNA (Yu 1994). Similarly, reactive nitrogen species (RNS) are also damage DNA (Patel et al. 1999). One of the most biologically significant and well-studied oxidative base lesions formed from hydroxylation of the C-8 residue of guanine, is the saturated imidazole ring 7,8 dihydro-8-oxoguanine (8-oxoG) (Figure 1.3C). It has been shown that, 8-oxoG favours hydrogen bonding with adenine instead of cytosine, resulting in a higher prevalence of C:G>A:T transversion mutations during replication (Michaels et al. 1992). Additionally, because of the low oxidation potential of 8-oxoG, it can be further oxidized to other deleterious secondary DNA lesions, and thereby add to the overall mutational load within the genome (Cheng et al. 1992; Patel et al. 1999). Other ROS-induced DNA damages, such as thymine glycol or formamidopyrimidine (Figure 1.3C), either directly or indirectly, can also generate site-specific localised mutations (Basu et al. 1989; Smela et al. 2002; Bellon et al. 2009). ROS/RNS can also induce a variety of DNA lesions including the generation of AP sites, single or double strand breaks and deamination, which usually get repaired by BER, SSB repair (SSBR) pathways, or the double strand break repair (DSBR) pathways (Su 2006; Woodbine et al. 2011; Sallmyr and Tomkinson 2018). At low levels, ROS/RNS species also act as cellular messengers in redox signalling reactions, and affect important defined responses to invading pathogens by the immune system (Friedberg et al. 2005). However, excessive ROS/RNS species can cause a total of ~100 different oxidative base lesions, and are associated with the development of human diseases, such as cancer, Alzheimer's disease, Parkinson's disease, diabetes, and heart failure (Cooke et al. 2003; Pham-Huy et al. 2008; Reuter et al. 2010).

1.2.1.1.5 Methylated DNA bases

DNA methylation is a covalent post-replicative modification of genomic DNA by methyl transferases during normal methylation reactions, and it occurs mostly at the C5 position of the cytosine residues of CpG dinucleotides to form 5-mC (Holliday and Grigg 1993; Moore et al. 2013). 5mC is a mutable site, because it can undergo spontaneous deamination to form thymine. Methylation is important for controlling the regulation of

gene expression and for normal development, but methylation can also make DNA more susceptible to damages, and plays an important role in many types of cancer (Holliday and Grigg 1993). DNA methylation is associated with epigenetic modifications, and the interplay of these epigenetic modifications is crucial to regulate the functioning of the genome by changing chromatin architecture (Kulis and Esteller 2010). Some endogenously originated chemicals such as derivatives of N-nitroso compounds, bile salts, betaine and choline can also induce the formation of DNA methyl adducts inside the cells (Zhao et al. 1999). Methylated DNA damages can be repaired by either O6-methylguanine DNA methyltransferase by direct reversal, or by DNA glycosylases during BER (Wyatt and Pittman 2006; Cuozzo et al. 2007). Unrepaired or uncontrolled methylated DNA bases are a major source of spontaneous clustered DNA damage and the formation of strand breaks (Wyatt and Pittman 2006). Additionally, methylated DNA adducts such as O^6-methylguanine and it's derivative O^4-ethylthymine have been shown to induce mutations, producing G:C>A:T and T:A>C:G transition respectively (Loveless 1969; Dosanjh et al. 1991).

1.2.1.2 Exogenous DNA Damage

In addition to endogenous DNA damage, exogenous damaging agents, such as physical, chemical and biological agents, constantly disrupt the integrity of the DNA double helix. The International Agency for Research on Cancer (IARC) classify all the reported carcinogens associated with human cancers (World Health 2005). There are about 120 confirmed human carcinogens with an additional 400 probable/possible human carcinogens. Nevertheless, some common human cancers still have few (or no) identified causal agents (Cogliano et al. 2011). In the following section I will describe some of the mutational processes associated with exposure to physical, chemical and biological agents.

1.2.1.2.1 Exogenous Physical Agents

1.2.1.2.1.1 Ultraviolet (UV) Radiation

Some mutagens cannot be avoided completely, and one of the most significant, and best understood of these is ultraviolet (UV) radiation. Exposure to solar UV is both inevitable and a known carcinogen, with sequencing of the genomes of skin cancers revealing mutations indicative of DNA that has been damaged by UV light (Pacholczyk et al. 2016).

UV radiation generates damage within a DNA molecule by two pathways. In the most common or direct pathway, UV light predominantly damages DNA through the formation

of covalent bonds between adjacent pyrimidine bases forming, cyclobutane pyrimidine dimers (CPD), pyrimidine(6-4)pyrimidone photoproduct (6-4PP), and the Dewar photoproduct, which is an isomeric form of 6-4PP (Figure 1.4) (Douki and Cadet 2001; Ravanat et al. 2001; Sinha and Häder 2002; Mouret et al. 2006). These photolesions are assumed to cause UV specific mutations. In the indirect pathway, UV can induce oxidative stress mediated DNA damage through production of ROS by activation of some small molecules such as riboflavin and tryptophan (McCormick et al. 1976; Peak et al. 1984).

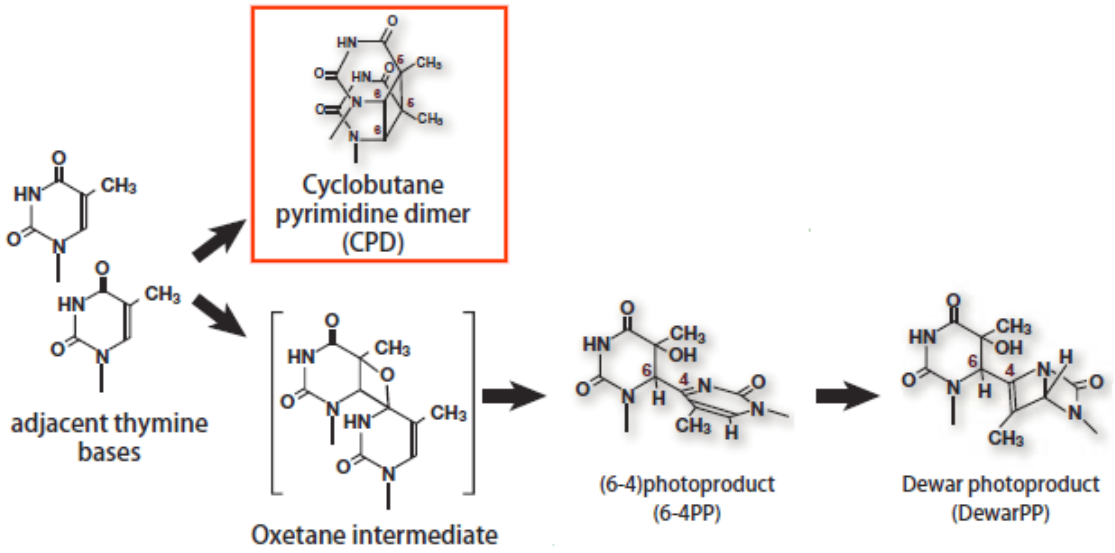


Figure 1.4: Main UV radiation-induced DNA base lesions. Representative CPD, shown here are cyclobutane thymine dimers. Representative (6 – 4)PP, shown here are derivatives of two thymine bases linked via C6 of one thymine base and C4 of the other thymine base. Figure adapted from https://www.cosmobio.com/contents/high_sensitivity_cpd_elisakit.html.

The extent to which UV light damages DNA depends on wavelength. UV is defined as electromagnetic radiation of a wavelength in the range 10 – 400 nm (Stark and Smith, 2006), with about 6% of the sun's output having wavelengths in this region (Schuch et al. 2009). However, wavelengths below 200nm are efficiently absorbed by atmospheric oxygen, whilst ozone completely and partially absorbs UV in the 200-280 nm and 280-340 nm range respectively (Freeman and Ryan 1990; Schuch et al. 2009). UV with wavelengths greater than 200 nm is subdivided into UVA (320-400 nm), UVB (280-320 nm) and UVC (200-280nm). Fortunately, very little UVC reaches the surface of the Earth, thanks to the protective effect of the ozone layer, for these shorter wavelengths (higher frequencies) are especially damaging (UVC induces approx. 10^2 and 10^5 times more

pyrimidine dimers in DNA than UVB and UVA respectively (Kuluncsics et al. 1999)). On the other hand, 0.3% of sunlight falling on the Earth's surface is UVB and 5% is UVA (Chatterjee and Walker 2017), each of which raise concerns for human health, since UVB is more damaging, but UVA is better at penetrating the superficial layers of skin (Mouret et al. 2006) and is less well filtered by sunscreens (Pfeifer et al. 2005).

UV irradiation is one of the most-studied mutagens, and can generate different types of mutations, which is a consequence of translesion DNA synthesis over UV-induced lesions. The most common UV-induced CPD generated at TT, TC, CT, and CC di-pyrimidines sites in a ratio of 68:16:13:3 respectively (Mitchell et al. 1992). But, the mutations it ultimately induces, also known as UV-induced mutational imprints, are predominantly composed of C>T, CC> TT and lower frequency of T>C, T>A, C>A, T>G single base substitutions with a few indirect oxidative type mutations like G>T (Ikehata 2018). Additionally, it can also induce 5-NTC-3> TTT mutations (Ikehata and Ono 2011). Furthermore, some thymine deletions at TT sites are also reported recently in melanoma cancer (Alexandrov et al. 2018). The biological process behind some of these different types of mutations associated with melanoma cancer are known but the majority are still remaining unknown.

Most of the CPDs get efficiently handled by different types of repair or DNA damage tolerance pathways such as, photoreactivation by photolyase, NER, BER, dimer bypass by error free and error prone TLS polymerases, recombinational repair or cell cycle check points (Friedberg et al. 2005). It also has been reported that TLS polymerase preferentially inserts an A opposite to the CPD providing supporting evidence for the 'A-rule' model of bypassing the most frequent UV-induced TT damages without generating mutations (Strauss 1991). However, the predominance of C>T substitution mutations at the UV-induced mutational imprint is due to deamination of cytosine, or methylated cytosine, at cytosine containing CPD's, followed by error free or error prone TLS (Figure 1.5). Additionally, T>C and T>A substitution mutations may be due to low frequencies of misincorporation by translesion polymerases of T and G opposite thymines in pyrimidine dimers rather than the more frequent and non-mutagenic A (Wang 2001). The final outcome of UV-induced mutational imprints may vary between different species, depending on absence or presence of deamination of methylation level of cytosine bases (Wang 2001).

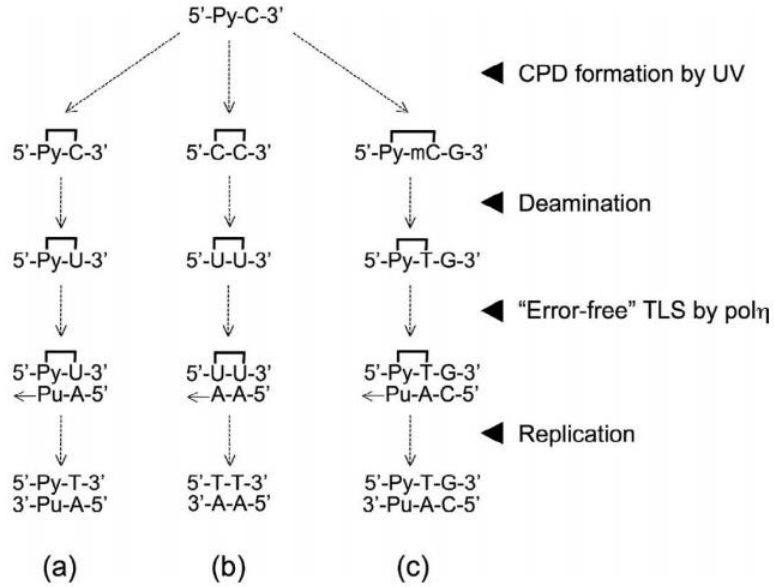


Figure 1.5: The most common mechanism of UV-induced mutation formation. (a) The usual pathway for C>T mutations via the deamination-mediated "error-free" TLS. (b) A specific pathway generation of CC>TT substitutions. (c) The pathway for C>T mutations via deamination of 5-mC. The CPD dimers are connected by bracket. Py, pyrimidine; Pu, purine; mC, methylcytosine (Figure adapted from (Ikehata and Ono 2011)).

As mentioned above, a pyrimidine dimer blocks DNA replication and transcription through the damaged site. Consequently, a single CPD can kill a bacterial cell that is deficient in all DNA repair pathways (Friedberg et al. 2005). However, a wild-type bacterium can survive a UV dose sufficient to generate thousands of CPDs, whilst a few hours exposure to sunlight is likely to induce tens of thousands of CPDs in the epidermal cells of unprotected human skin (Freeman and Ryan 1990; Snellman et al. 1992). This indicates that, under normal conditions, the repair machinery is highly efficient in eradicating UV-induced DNA damages.

1.2.1.2.1.2 Ionizing Radiation (IR)

Another highly lethal DNA damage is the group of DSBs that can be induced by ionizing radiation. Exposure to this type of radiation occurs during radiation therapy, commonly used for the treatment of certain cancers, where its ability to induce DSBs kills tumour cells. At the same time, IR is a potent carcinogen because of its ability to damage DNA. IR, is composed of alpha, beta, gamma, neutrons, and X-rays, being produced from diverse sources ranging from rocks, soil, and radon, to cosmic radiation and medical devices (Wood et al. 1990). Cumulatively, IR can damage the DNA either directly, by breaking the sugar-phosphate backbone of DNA or by indirect means, such as radiolysis

of the surrounding water to generate a cluster of highly reactive hydroxyl radicals (•OH) (Friedberg et al. 2005). Major lesions include 8-oxo-guanine, thymine glycol, and formamidopyrimidines. Apart from causing base lesions, IR also causes DNA double and single strand breaks with unique features, such as an excess of deletions and of an exceedingly rare type of rearrangement (Behjati et al. 2016). DSBs generated by IR are the most lethal form of DNA damage and are repaired via either homologous recombination (HR) or nonhomologous end-joining (NHEJ) pathways (Takata et al. 1998; Khanna and Jackson 2001).

1.2.1.2.2 Exogenous Chemical Agents

1.2.1.2.2.1 Alkylating and Crosslinking Agents

The major sources of alkylating agents are tobacco smoke, dietary components, biomass burning or industrial processing and chemotherapeutic agents such as Temozolamide or platinum-based drugs (Figure 1.6) (Pourquier 2011). Alkylating agents, including methyl methane sulfonate (MMS), ethyl methane sulfonate (EMS) and methyl nitrosourea (MNU) (Figure 1.6A); frequently interact with DNA to generate mutagenic lesions (Chatterjee and Walker 2017). Mostly, direct damage reversal, BER, NER and ICLs (Inter-strand crosslinks) repair are the main DNA repair pathways that respond to alkylated base damage (Wyatt and Pittman 2006). A specific type of mutation was identified in tumours from patients previously treated with Temozolamide, an alkylating agent. This mutations are enriched for C>T substitutions on guanine bases. A strong transcriptional strand-bias is also present in this type of mutation. In some cancer patients treated with platinum drugs, a high number of both C>T and T>A types of substitutions were found. Furthermore, tobacco smoking or chewing has also been associated with C>A type mutations. The NER and BER pathways are known to be involved in these mutational mechanisms.

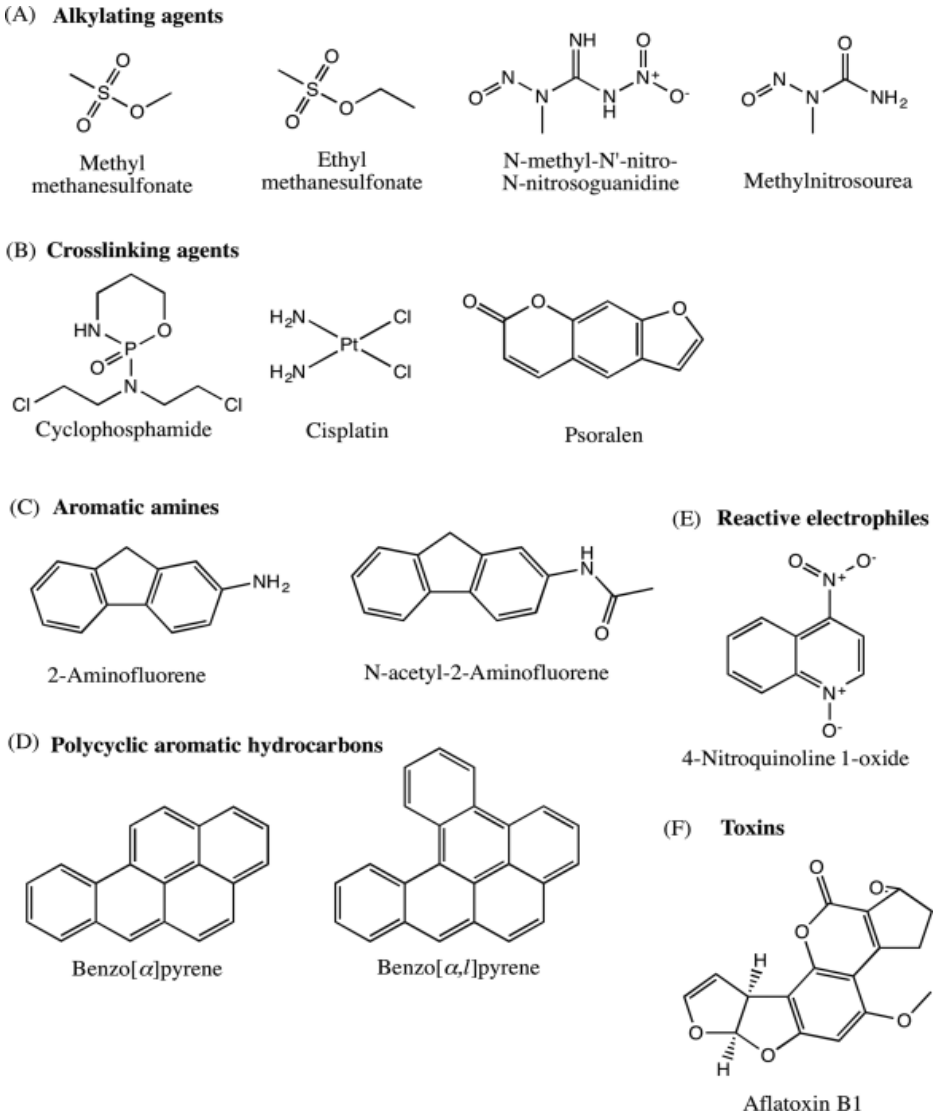


Figure 1.6: Most common chemical DNA damaging agents. A: Alkylating agents: MMS, EMS, MNNG, and MNU. B: Crosslinking agents: Cyclophosphamide, cisplatin and psoralen. C: Aromatic amines: AF and AAF. D: Polycyclic aromatic hydrocarbons: benzo(a)pyrene and dibenzo[a,l]pyrene. E: Reactive electrophiles: 4-NQO. **F**: Toxins: Afaltoxin B1 (Figure adapted from (Chatterjee and Walker 2017)).

1.2.1.2.2.2 Aromatic Amines and Polycyclic Aromatic Hydrocarbons

The major source of aromatic amines (AA) and polycyclic aromatic hydrocarbons (PAH) (Figure 1.6 C and D) are cigarette smoke, fuel, coal, industrial dyes, automobile exhaust, charred food and incomplete combustion of organic matter and fossil fuels (Stewart et al. 2010). Upon activation by the P450 monooxygenase system, AA and PAH are converted into the carcinogenic agents that either attack the C8 position of guanine or form adducts with DNA (Shimada and Fujii-Kuriyama 2004). C8-guanine lesions formed by AA are known to form persistent lesions that ultimately give rise to base substitutions and frame

shift mutations (Sproviero et al. 2014). Usually, the excision repair pathways such as NER and BER repair these types of DNA lesions if they are not bypassed by TLS polymerases (Jha et al. 2016).

1.2.1.2.2.3 Reactive Electrophiles

Reactive electrophiles, such as 4-nitroquinoline 1-oxide (4-NQO) are known mutagenic and carcinogenic agents (Figure 1.6 E). 4-NQO is converted into 4-hydroxyaminoquinoline 1-oxide upon metabolic activation and induces DNA adducts with C8 or N2 of guanine, and N6 of adenine, as well as causing oxidative stress that results in the 8-hydroxyguanine lesion, all of which significantly adds to the strand breakage events found in oral carcinogenesis (Brüsehafer et al. 2016).

These genotoxic bulky DNA adducts are repaired by the NER pathway in a similar manner as during the repair of UV-induced damages (Bastien et al. 2007). Other potent carcinogenic electrophiles, such as N-nitrosamines (NA), are by-products of tobacco smoke and are also encountered by humans in certain preserved meats. N-nitrosamines have been implicated in the development of esophageal, stomach, and nasopharyngeal cancers (Herrmann et al. 2015). Tobacco-specific nitrosamines can generate bulky DNA adducts that are substrates for NER and the extensive NA exposure might saturate the DNA repair capabilities, resulting in the formation of mutation patterns equivalent to those linked to NER defects described in the literature (Peterson 2010).

1.2.1.2.2.4 Toxins

Naturally, toxins can be found in contaminated cereals, tree nuts, oilseeds, spices, milk, and milk products (Bhat et al. 2010). The mostly studied toxins, such as aflatoxin B1 (Figure 1.6F) produced by fungi, are a potent carcinogen. After activation by the P-450 complex, aflatoxin B1 forms adducts with N7 of guanine which weakens the glycosidic bond resulting in depurination (Smela et al. 2001). Recent evidence shows that a specific mutational pattern was shown to be associated with aflatoxin exposure combined with other exposures. This mutational type exhibits a very strong transcriptional strand bias for C>A mutations, indicating the presence of guanine damage that is being repaired by TC-NER (Alexandrov et al. 2018).

Another toxin, aristolochic acid causes upper urinary tract infection, resulting from ingestion of the Aristolochia plant, predominant in specific parts of Asia (Poon et al. 2013). A strong correlation of the relevant mutations has been reported recently with

exposure of aristolochic acid that exhibits a predominance of A>T mutations (Huang et al. 2017).

1.2.1.2.2.5 Environmental Stresses

The catalogue of mutations found in human tumours can be altered by environmental stress regulation and stress-induced mutagenesis. Oxidative stress from the environment, extreme heat or cold and hypoxic conditions can cause DNA damage in both prokaryotes and eukaryotes (Halliwell and Gutteridge 2015). The mutational catalogue resulting from environmental stress also provides insight into the mechanisms involved in the process of evolution by natural selection. (Maharjan and Ferenci 2014; Kantidze et al. 2016). In some cancer cells, it has been reported that the stresses from the environment cause mutagenesis at trinucleotide repeats which are usually repaired by alternative-nonhomologous end joining (alt-NHEJ) components—XRCC1 and PARP1 (Chatterjee et al. 2016). Stress from food additives or preservatives such as benzoate and sorbate salts are also known to cause DNA damage (Piper and Piper 2017). Some of the daily used biological products, such as butyl paraben or bisphenol A found in cosmetics, have also been associated with the formation of DNA damage and alteration in cell cycle regulation (Pfeifer et al. 2015).

1.2.1.2.2.6 Alcohol

Chromosome analysis and DNA sequencing also recently helped to explain the genetic damage caused by acetaldehyde, a harmful chemical produced when the body processes alcohol and increases the risk of developing 7 types of cancer, including common types like breast and bowel cancer (Garaycoechea et al. 2018).

1.3 Biological processes of DNA damage repair

The previous sections focussed on the most common biological processes that can damage DNA. In this section, I describe the different types of DNA repair processes (Figure 1.7), which maintain the integrity of the genome after repairing damaged DNA. DNA repair pathways can also be categorised by the different types of DNA lesions that they are operating on. Firstly, repair of the damaged DNA base by either the direct reversal or BER pathways. Secondly, repair of multiple bulky base damages that may be repaired by either NER or MMR pathways. Finally, repair of DNA breaks by either SSBR or DSBR pathways. In addition, there is another type of specialised damage tolerance process, which operates during replication that bypasses the damaged sites, and allows replication to take place. This process is called TLS.

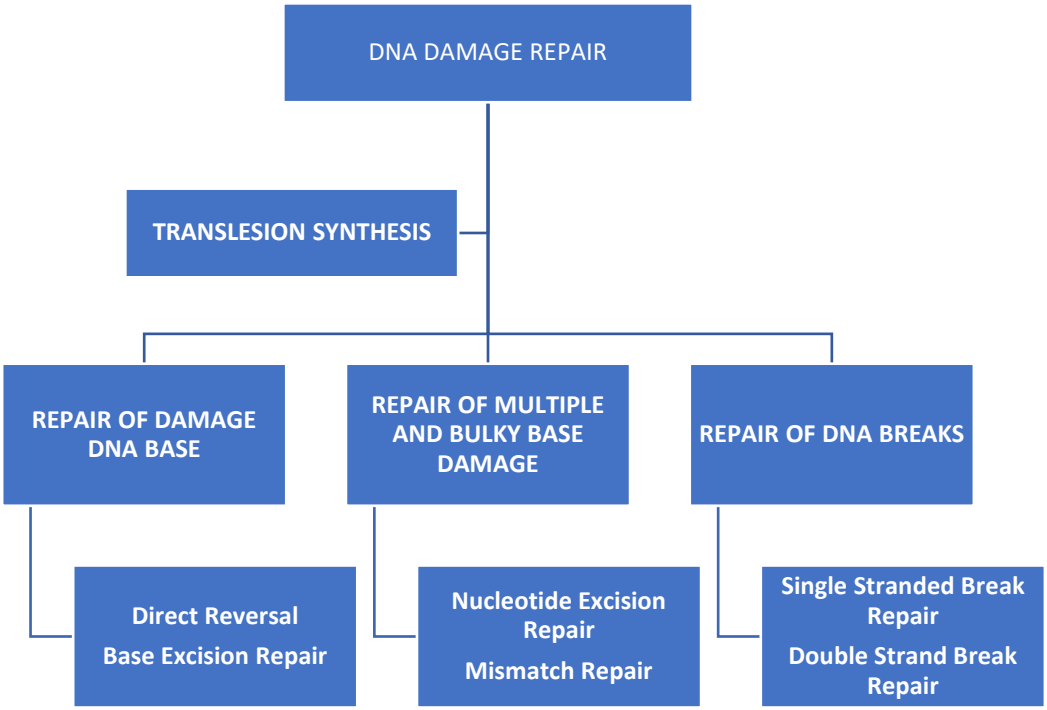


Figure 1.7: Damaged DNA can be removed by variety of repair processes. Translesion synthesis also known as post replicative repair, because of feeling of gaps those left during replication. The word repair in this case refer to the gaps, instead of original lesions itself.

1.3.1 Repair of base DNA damage

1.3.1.1 Direct reversal of DNA damage

DNA lesions such as UV photolesions can be reversed by photolyase-mediated photoreactivation in certain organisms (Friedberg et al. 2005). Indeed, experiments using UV-induced DNA damage in certain organisms, including yeast cells, must be carried out in the dark to prevent this process from directly reversing the damage. Reversal of alkylated bases is mediated by either the O^6-alkylguanine-DNA alkyltransferase (AGT) or the AlkB-related α-ketoglutarate-dependent dioxygenases (AlkB) (Duncan et al. 2002; Mishina et al. 2006). These repair enzymes have been shown to be hyper-active in certain types of cancers, causing resistance to treatment with alkylating agents (Pegg 2011). Compounds that inactivate these repair enzymes can be used in combination with therapeutic alkylating agents to circumvent resistance to cancer chemotherapy (Rabik et al. 2006). However, lack of AGT expression is associated with certain groups of cancers (Pegg 2011). In addition, a family of AGT homologs, inhibit the AGT enzyme by directing the repair of bulky alkyl damage to the NER pathway (Tubbs et al. 2009). Taken together, these repair processes play important roles in preventing the mutagenic potential of alkylating DNA damages, and maintain the integrity of the genome.

1.3.1.2 Base Excision Repair

The base damage produced by ionizing radiation, deamination, and by oxidizing agents are repaired predominantly by the base excision repair (BER) pathway (Sancar et al. 2004; Krokan and Bjørås 2013), which is distinct from NER because BER recognises those damages that are not perceived as significant distortions to the DNA helix (Wang and Vasquez 2014). However, some fraction of lesions that are repairable by base excision repair may also be dealt with by NER, but there is no backup repair system for NER (Swanson et al. 1999). In the BER pathway, after DNA damage induction, chromatin remodeling at the DNA damage site is followed by lesion recognition by a DNA glycosylase (Odell et al. 2013) which recognises and excises a damaged base from undistorted helices. An abasic site created from the monofunctional glycosylases gets committed to the short-patch-repair pathway, while the bifunctional glycosylases initiate the long-patch repair pathway of BER (Dianov and Hübscher 2013) (Figure 1.8)

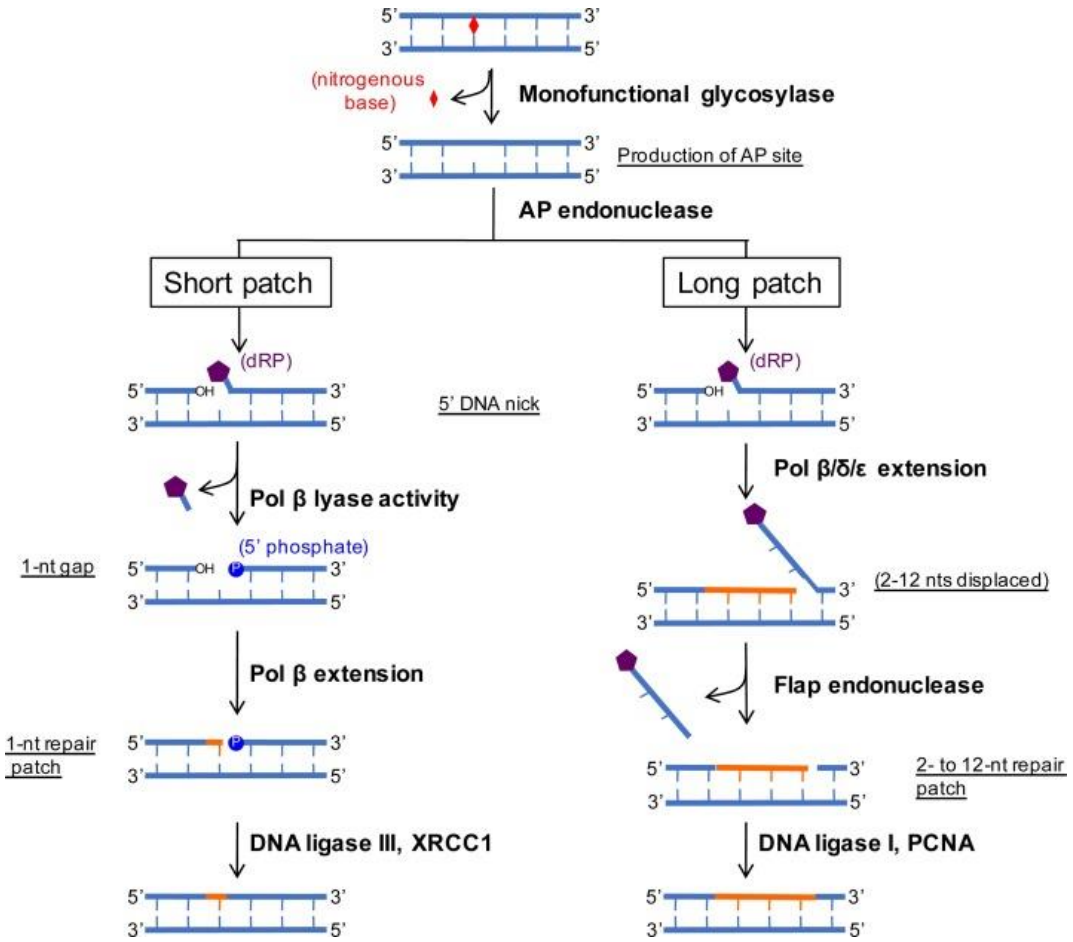


Figure 1.8: The simple illustrative diagram of base excision repair pathway. A non-helix distorting damage is spotted and processed by a glycosylase to creates an abasic site that is recognized by an AP endonuclease which generate a 5′ nick. This substrate is then handled by either short patch or long patch sub-pathways. In short patch repair sub-

pathway, the deoxyribonucleotide phosphate lyase activity of pol β to remove the dRP moiety and also filled the gap by extension of 1bp, followed by ligation of DNA by the DNA ligase III/XRCC1 complex. In the long patch repair sub-pathway, the gap felling DNA synthesis is proceeds by pol β, δ, ε to extend ≥2 nucleotides. The displaced 2-12 nucleotides DNA fragment is excised by Flap endonuclease followed via ligation by DNA ligase I and proliferating cell nuclear antigen (PCNA).(Figure adapted from (Meas et al. 2017)).

Defects in the BER pathway leads to cancer predisposition. Deletion mutations in BER genes have been shown to result in a higher mutation rate in a variety of organisms, implying that loss of BER could contribute to the development of cancer. Inherited mutations in the DNA glycosylase MUTYH are also known to increase susceptibility to colon cancer (Cheadle et al. 2003; Cheadle and Sampson 2003; Pilati et al. 2017; Viel et al. 2017). Indeed, somatic mutations acquired in the Pol β gene, involved in BER, have been found in 30% of human cancers, and some of these mutations lead to transformation when expressed in mouse cells (Starcevic et al. 2004). All of which adds to the significance of this repair pathway in the maintenance of global genome stability.

1.3.2 Repair of multiple and bulky base damage

1.3.2.1 Nucleotide excision repair

Nucleotide excision repair (NER) is the main pathway to remove bulky DNA lesions formed by UV light such as cyclobutane–pyrimidine dimers (CPDs) and 6–4 pyrimidine–pyrimidone photoproducts (6–4PPs), environmental mutagens or by chemotherapeutic agents such as platinum drugs (Friedberg 2001; Schärer 2013; Marteijn et al. 2014). Deficiencies in NER are associated with several different human syndromes: Xeroderma Pigmentosum (XP); which is associated with a predisposition to skin cancers; Cockayne Syndrome (CS); rare UV-Sensitive Syndrome ($UV^{S}S$); premature aging and Cerebro-Oculo-Facio-Skeletal syndrome (Friedberg et al. 2005; Schärer 2013; Marteijn et al. 2014)**.** However, like the BER pathway, NER also contributes to the instability mechanisms involved in triplet repeat disorders **(**McMurray 2010).

In both prokaryotes and eukaryotes, NER represents one of the most important repair systems. NER is characterised by damage recognition by a damage recognition protein complex, followed by dual incision of the damaged DNA strand on both sides of the lesion, resulting in the removal of the damage in an oligonucleotide fragment. Finally, a DNA polymerase is recruited by PCNA (Proliferating Cell Nuclear Antigen) to fill the

resulting gap using the undamaged DNA strand as a template. After this, a DNA ligase seals the nick to restore the DNA, an error-free process (Marteijn et al. 2014). There are two major branches of NER. The rapid acting transcription coupled repair pathway (TC-NER), that operates on the transcribed strand of actively transcribing genes, and the slower acting global genome repair pathway (GG-NER), that operates in all non-transcribed regions in the DNA (Friedberg et al. 2005). These pathways utilise the same basic set of proteins, with a subset involved in the early steps of DNA damage recognition being unique to each. GG-NER is initiated by the recognition of damage-induced DNA helix distortions by a damage recognition complex, and TC-NER is initiated by stalling of RNA polymerase II (RNA Pol II) at a lesion during transcription (Friedberg et al. 2005).

Genetic predisposition to cancer in somatic cells can be inherited because of the malfunction of genes required for the normal processing of DNA damage by all the major DNA repair pathways including nucleotide excision repair (NER) (de Boer and Hoeijmakers 2000; Sugasawa 2008; Broustas and Lieberman 2014). A wide variety of human genetic diseases are caused by mutations in genes that encode these DNA repair pathways. Defective NER has been clearly documented in the hereditary cancer-prone disease *xeroderma pigmentosum* (XP) and this is the primary cellular phenotype of this autosomal recessive disease. This demonstrated the importance of NER as a fundamental mechanism for protecting the functional integrity of the human genome (de Boer and Hoeijmakers 2000), providing important support for the somatic mutation theory of cancer.

Mutations in the NER-gene ERCC2 were previously associated with a specific mutational profile (Kim et al. 2016), suggesting the link between mutational profile and deficiency of NER pathways. The mutational profile analysis of TC-NER predominantly focused on transcription strand bias analysis. The transcriptional strand bias of mutation is due to both TC-NER and an excess of DNA damage induced in untranscribed compared to the transcribed strands of genes (Haradhvala et al. 2016). Both processes, however, result in more mutations occurring in the untranscribed compared to the transcribed strands of genes. The mechanism(s) underlying this amplification of transcriptional strand bias is still unknown and appears to be mutation-type specific.

In eukaryotes, the genome is organised into nuclear domains that have distinct chromatin structures and functions: highly repetitive sequences, centromeres, telomeres, non-coding sequences, inactive genes, RNA polymerase-I, II and -III and transcribed genes (Dixon

et al. 2012). Like DNA transcription and replication, NER is affected by the structure of chromatin, and it is plausible that the kinetics of DNA repair varies among domains, suggesting that the mutation rate may also differ within nuclear domains (Smerdon 1991). To initiate NER, chromatin remodeling, mediated by specific NER components, enables the NER machinery to efficiently recognise and repair the DNA lesion (Schärer 2013; Marteijn et al. 2014; Yu et al. 2016).

1.3.2.2 Mismatch repair

The post-replicative mismatch repair (MMR) pathway recognises and repairs mis-incorporated bases, as well as erroneous indels that arise during DNA replication and DNA recombination repair activity. Typical substrates for the MMR pathway are base mismatches that have arisen during replication and the insertion-deletion loops that occur within repetitive DNA sequences that have resulted from strand slippage events (Friedberg et al. 2005). In this process, the mis-incorporated damages are recognised by MSH2 and MSH6. A short section of the DNA around the mismatches is excised by the MMR endonuclease PMS2 and exonuclease 1 (EXO1), followed by the gap felling via coordinated activity by PCNA and DNA Pol δ (Figure 1.9).

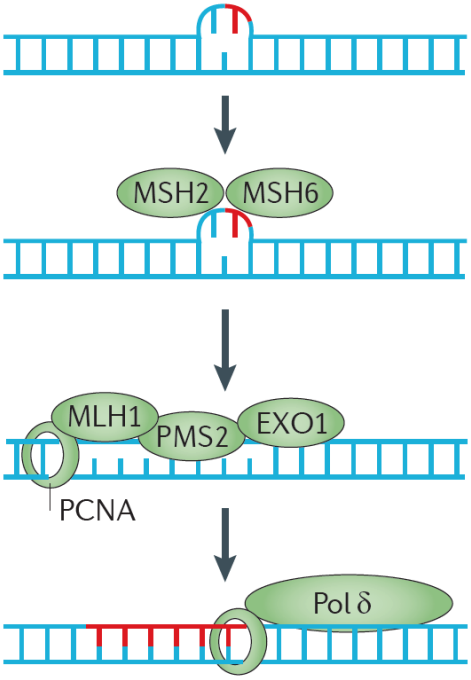


Figure 1.9: A simple illustrative diagram of the mismatch repair process. MMR correct the DNA mismatches generated during replication. (Figure adapted from (Helleday et al. 2014).

MMR is a major contributor to genome stability, where it improves replication fidelity by at least 100-fold (Kunkel and Erie 2005). Hence, defects in the MMR pathway increase the spontaneous mutation rate (Hsieh and Yamane 2008). Somatic mutations in MMR-related proteins also affect genomic stability and result in in the formation of microsatellite instability (Meier et al. 2018). Recently, the association between different numbers of substitutions and indel mutations with defective MMR have been reported in cell culture-based study (Zou et al. 2018). Importantly, the association of germline mutations in the MMR genes results in hereditary nonpolyposis colorectal cancer (HNPCC), or Lynch syndrome, which represents a risk factor for colon, ovarian and other types of cancer (Wu et al. 1999). In addition, chromatin structure and genome organisation also have been shown to regulate repair by providing access of damaged DNA to MMR components (Pan et al. 2014).

1.3.3 Repair of DNA breaks

1.3.3.1 Single Stranded Break Repair

Single Stranded Breaks (SSBs) are usually accompanied by single nucleotide gaps, resulting in discontinuities in one of the strands of the DNA double helix. SSBs are often generated by endogenous oxidative damage to the DNA, from abasic sites generated during BER, or from erroneous activity of the DNA topoisomerase 1 (TOP1) enzyme (Demple and DeMott 2002; Hegde et al. 2008). Most SSBs are repaired by a rapid global Single Stranded Break Repair process by four basic steps: detection of SSBs, processing of DNA ends, filling the gap in the DNA and ligation of the nick in the DNA (Caldecott 2008).

If SSBs are left unrepaired they pose a serious threat to genome stability and cell survival. Sometimes SSBs interrupt DNA replication, which can result in the formation of double strand breaks (DSBs) (Kuzminov 2001). Mutations in the SSBs repair complex are found frequently in certain human genetic disorders, such as spinocerebellar ataxia with axonal neuropathy 1 and ataxia-oculomotor apraxia. These patients often lack chromosomal instability and cancer predisposition (Caldecott 2008), demonstrating the importance of DNA repair mechanisms to other aspects of human health other than cancer.

1.3.3.2 Double Strand Breaks Repair

Extremely toxic DNA Double Strand Breaks (DSBs) are generated by both exogenous and endogenous DNA insults such as chemical and physical DNA damaging agents (Khanna and Jackson 2001). Defective repair of DSB can produce mutations and cause

many human syndromes, neurodegenerative diseases, immunodeficiency and cancer, and all are associated with defective repair of these DNA lesions (Varon et al. 1998; Khanna and Jackson 2001). Two main pathways, Homologous Recombination (HR) and Non-Homologous End-Joining (NHEJ) are responsible for repairing DNA DSBs (Shrivastav et al. 2008).

HR is a complex multistage repair process and only operates when a double-stranded copy of the broken sequence is available in the genome to act as a template. The pathway may induce small scale mutations and chromosomal aberrations such as tandem duplications (Pfeiffer et al. 2000). On the other hand, DSBs can also be repaired by NHEJ, which simply splices together two broken DNA ends, resulting from IR or other types of cleavage of the DNA. This process is often mediated by microhomology patches at ends, which have been created as a result of end-resection. This process will give rise to mutations and is therefore error-prone (Pfeiffer et al. 2000).

Both HR and NHEJ repair mechanisms will leave their own characteristic imprint of activity in the genome. For example, a strong association with both somatic and germline mutations in two important breast cancer genes BRCA1 and BRCA2 has been reported recently (Zámborszky et al. 2017). Similarly, substitution mutations are predominant in pancreatic cancer patients treated with platinum drugs (Alexandrov et al. 2018). Strong correlation is found between indel and those exhibits microhomology mediated overlaps at deletion boundaries, indicating that HR is mainly responsible for these types of mutations. By contrast, shorter or no microhomology at deletion boundaries indicating the NHEJ is associated with this mutational pattern found in cancer (Alexandrov et al. 2018). These indel patterns are characteristic of DNA double strand break repair and indicate that at least two distinct forms of end-joining mechanism are operative in human cancer (Ceccaldi et al. 2016).

1.3.4 Translesion synthesis

Usually, replicative polymerases are highly accurate, and do not tolerate damaged bases in the DNA. A damaged DNA template can lead them to stall and block replication fork progression. The fate of the stalled replication forks can be decided by two different mechanisms: switching of the templates following the blockage of the polymerase by the damage and restarting the replication process by switching the DNA polymerases. This allows the cell to continue replication with the added risk of introducing mutations (Wang 2001). The latter process called translesion synthesis (TLS) is carried out by highly

conserved TLS polymerases, which have a relatively lower fidelity than replicative DNA polymerases (Sale 2013). This prevents the introduction of DSB at stalled replication forks at the expense of the formation of single base substitutions.

In addition to their traditional DNA damage bypass functions, TLS polymerases are now known to play a role in other cellular pathways such as ICL repair and can play a role in the BER and NER pathways, by synthesising new DNA after the excision step with the potential for the increase of mutational burden (Knobel and Marti 2011). If incorrect nucleotides were incorporated by TLS polymerases, they would become mutations in the next round of replication, which can drive tumorigenesis and cause disease (Goodman 2002). Although an inherited deficiency in Pol η results in a marked predisposition to skin cancer, spontaneously arising mutations of the TLS polymerases have not emerged as a common event in human cancer (Lange et al. 2011). Recently it was reported that particular types of mutations are associated with Pol η activity. Nonetheless, polymorphisms in the TLS polymerases and dysregulation of their expression have been linked to cancer (Curtin 2012; Sale 2013). Additionally, most common TLS DNA polymerase η mutational profiles have been reported in a variety of different types of cancer (Rogozin et al. 2018). Taken together, mutational processes induced by DNA polymerases result in tumour cell development as a result of tolerating DNA damage. Some of these enzymes could be targets for future therapeutic interventions.

1.4 Mutagenesis

Damaged DNA is different from mutated DNA. DNA damage often persists and factors that contribute to the persistence of these DNA damages include the initial levels of damage induced, repair efficiency, chromatin structure of the genome and phase of cell cycle when the damage becomes encountered by replication enzymes (Wang 2001). Apart from spontaneous mutations, these persistent lesions may be bypassed via damage tolerance pathways at the cost of generating mutations. Recent advances in genomic technologies, including the use of microarrays and high throughput sequencing, have opened new opportunities for genome-wide surveys of DNA damage levels, repair efficiency, and mutation rates. DNA damage-induced mutations result from either the lack of DNA repair acting on the lesion or an error-prone repair attempt. Consequently, mutations are a direct biological output of DNA damage and repair events that often translate to the very same processes underlying genetic alterations that cause diseases like cancer. The contextual analysis of damage-induced mutations will provide a useful

method to gain insight into lesion formation and repair dynamics on a whole-genome scale in both model systems and clinical cancer samples.

1.5 DNA Damage Response

Survival of the organism depends on the maintenance of gnome stability. In response to both exogenous and endogenous DNA damaging agents described in previous sections, to maintain genome stability, a series of coordinated events also known as the DNA-damage response (DDR) that sense DNA damage, stop cell cycle progression, signal its presence, modulate transcription, promote subsequent repair and induce apoptosis if the damage load is too severe (Figure 1.10).

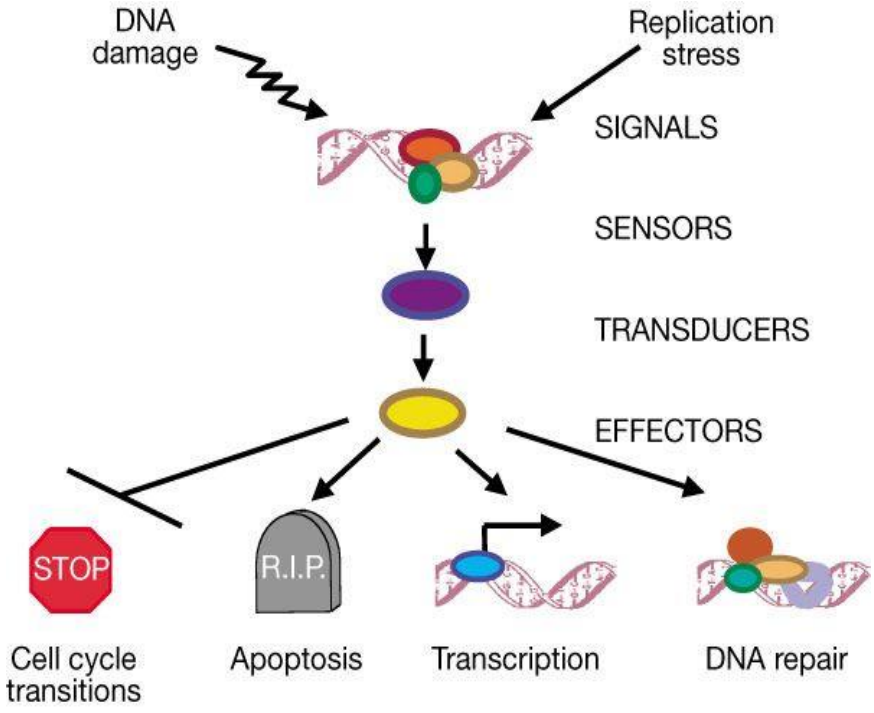


Figure 1.10: A contemporary view of the general outline of the DDR. Which is a programmed response to the effects of damage which affects regulation of replication (cell cycle response), gene expression (transcriptional regulation), genome stability (DNA repair pathways) and cell survival (apoptosis). The interactive network pathways are also mentioned, such as signals, sensors, transducers and effectors (Figure adapted from (Zhou and Elledge 2000)).

In general, all the pathways in DDR encounter a similar set of highly regulated events. These include detection of DNA damage, recruitment of repair factors at these damage sites and subsequent removal of these damages (Lord and Ashworth 2012b). Thereby, cells have developed a number of DNA damage detection and repair mechanisms for repair of various types of damages that occurs to the DNA by various DNA damaging

processes. In this coordinated DDR process, specific DNA repair pathway detect and repair specific types of lesion (Ciccia and Elledge 2010). For example, error form replication, spontaneous base deaminating, alkylating agents exposure results in mismatch or abasic sites within genome which frequently get repaired by MMR or BER processes (Figure 1.11). Additionally, exposure to UV irradiation or chemotherapeutic agents can cause helix distorting lesions or bulky adducts within DNA which preferentially repaired by NER process (Figure 1.11). In addition to various repair processes, cells have also developed DNA damage tolerances processes known as translesion synthesis. In this processes, various low fidelity DNA polymerases allowed cells to proceed replication, after bypassing the damaged DNA, but with generation of mutations.

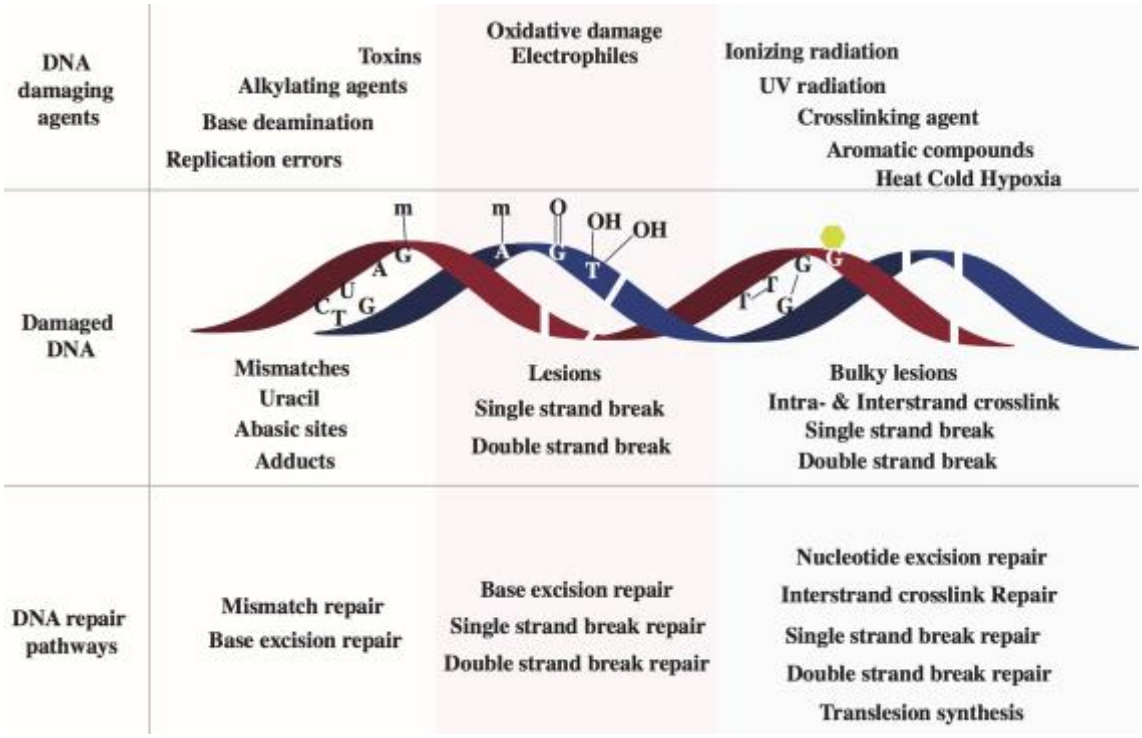


Figure 1.11: Schematic of different DNA lesions-induced DNA repair process for maintaining genome stability. The upper panel shows the various exogenous and endogenous DNA damaging agents. The middle panel represent corresponding DNA lesions induced within genome. And the lower panel depicts the subsequent repair pathways for removal of specific damages. TLS is included as post replicative repair, because of feeling of gaps those left during replication. The word repair in this context refer to the gaps, instead of original lesions itself. (Figure adapted from (Chatterjee and Walker 2017).

Assembly of the DDR factors at the damage sites is highly coordinated process (Harper and Elledge 2007). In the process, DNA damage sensors detect chromatin-associated DNA damage signals, which ultimately determine the physiological response of the cell to DNA damage (Lazzaro et al. 2009). Consequently, chromatin remodeling and histone modifiers are important determinant of the DDR response (Van Attikum and Gasser 2009; Chatterjee and Walker 2017), because accessibility of DNA into chromatin by the specific DDR and repair factors depends on modification of histone as mentioned in access-repair-restore (ARR) model (Smerdon 1991). Disruption or dysregulation of these DDR pathways are associated with many types of human diseases including cancer (Lord and Ashworth 2012a). Additionally, mutations affecting the DDR network components are the cause of several cancer predisposition syndromes including ATM/ATR (Ciccia and Elledge 2010). Therefore, determining how DNA damage in chromatin is detected, efficiently repaired, the chromatin restored and how these events are organised in the genome, is fundamental to understanding the mechanisms that underpin the relationship between genome stability and human health.

1.6 Genomic Instability and Cancer

In this section, I will outline aspects of our current understanding of cancer and mutagenesis. A key hallmark of cancer is instability of the genome (Hanahan and Weinberg 2011). Cancer is a clonal disease and arises because of the acquisition of somatic mutations in the genome of cells during the lifetime of the individual. Only a small minority of somatic mutations are considered to be driver mutations because this class of genetic changes occur in certain critical genes that are involved in the processes that maintain the stability of the genome, and when mutated, they confer a growth advantage that can be considered causative in the process of transforming a normal cell into a malignant one. The vast majority of mutations are considered to be passenger mutations; genetic changes that have accumulated as a result of genomic instability induced either as a result of cellular transformation or malignancy, or as bystander mutations. Bystander mutations arise from the same processes that create the driver mutations, but accrue in genes or regions of the genome that do not cause the phenotypic changes that result the emergence of the cancer. Although cancer is a clonal disease, heterogeneity is displayed in the mutational load of cancer cells due to the acquisition of late stage mutations. Thus, a cancer genome can be thought of as containing the record of mutagenic processes that have occurred before and after the acquisition of a neoplastic transformation by the progenitor cells of the tumour. In principle, through mitotic cell

division, the DNA sequence of cancer cells harbours a set of somatic mutations acquired as a result of different mutagenic processes that occurred within the cellular environment. This collection of somatic mutations, also known as the catalogue of somatic mutations, is distinct from germline mutations, which are inherited genetic variation, which can predispose individuals to disease, by driving their cells towards the early development of cancer (Stratton et al. 2009). The genomes of all cancer cells carry somatic mutations, and the pattern of these mutations, that is their type and distribution in the genome, reflect the cumulative effects of biological mutational processes, such as the induction of DNA damage and the efficiency of the repair processes operative between the fertilized egg and the formation of the cancer cell, as shown in Figure 1.12 (Stratton et al. 2009). As noted, somatic mutations occur in the cells while the cells are phenotypically normal, reflecting both the intrinsic mutagenic processes of normal cell growth and division and environmental or lifestyle exposures to genotoxic agents. During the development of cancer, DNA repair processes may contribute to the mutational burden and confer a mutator phenotype, which is initiated after the acquisition of driver mutations, while passenger mutations are also carried forward without providing any clonal growth advantage. However, these passenger mutations that are much more numerous, are caused by the same mutagenic processes that generate the driver mutations. Their prevalence provides sufficient statistical power, making it possible to determine the mutational mechanisms involved in the development of an individual patient's cancer.

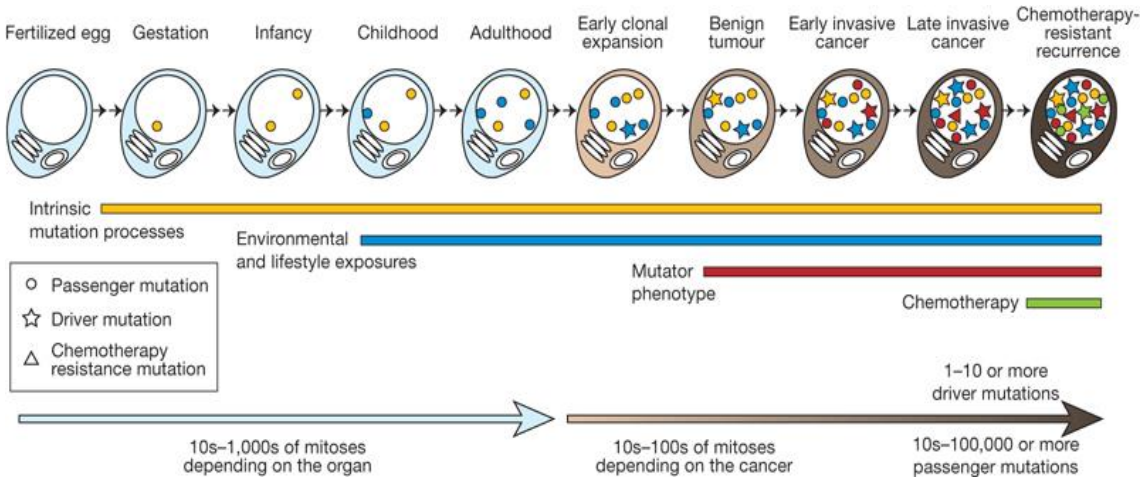


Figure 1.12: Process of acquisition of somatic mutations in cancer genomes showing the timing of the somatic mutations acquired by the cancer cell and the biological processes that contribute to the formation of those mutations (Figure adapted from (Stratton et al. 2009).

1.6.1 The landscape of somatic mutations found in cancer genomes

The identification of a catalogue of somatic mutations found in cancer genome studies is made possible by next generation sequencing (NGS) techniques using either whole genome sequences (WGS) or whole-exome sequences (WES) (Nik-Zainal 2014). In a cancer cell genome, there are different classes of somatic mutations observed (Stratton et al. 2009; Nik-Zainal et al. 2012). These include DNA base substitutions; insertions or deletions of small or large segments of DNA; rearrangements, in which DNA has been broken and then re-joined to a DNA segment from elsewhere in the genome; copy number changes by either gene amplification or complete absence of a DNA sequence from the cancer genome as shown in Figure 1.13 and represented as a circos plot.

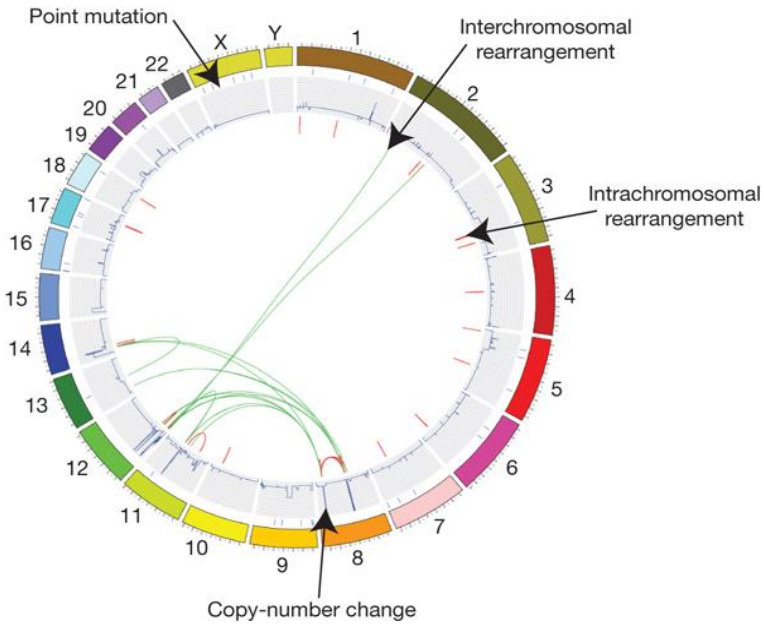


Figure 1.13: CIRCOS plot of a small-cell lung cancer patient. This plot depicts a complete genome of a lung cancer cell with the chromosomes indicated around the outside. A catalogue of several classes of somatic mutations are described, such as point mutations, copy number changes and rearrangement within the genome. Arrows indicate the type of somatic mutations present in this individual genome (Stratton et al. 2009).

When this catalogue of somatic mutations was mapped according to their genomic coordinates and their inter-mutational distance, it was obvious that the distribution of mutations differs within different types of cancer. For example, in the case of acute myeloid leukaemia (AML) (Figure 1.14, upper panel) or Acute lymphoblastic leukaemia (ALL) (Figure 1.14, lower panel), there is a difference in the distribution of various types of base substitution mutations over the background load of mutation. In the case of AML this background load of mutations is dominated by C>T types of mutations, whereas for

ALL these background loads are dominated by both C>T and C>G types of mutations, indicating that there is a structure and pattern to the distribution and types of mutations identified in different cancer cells.

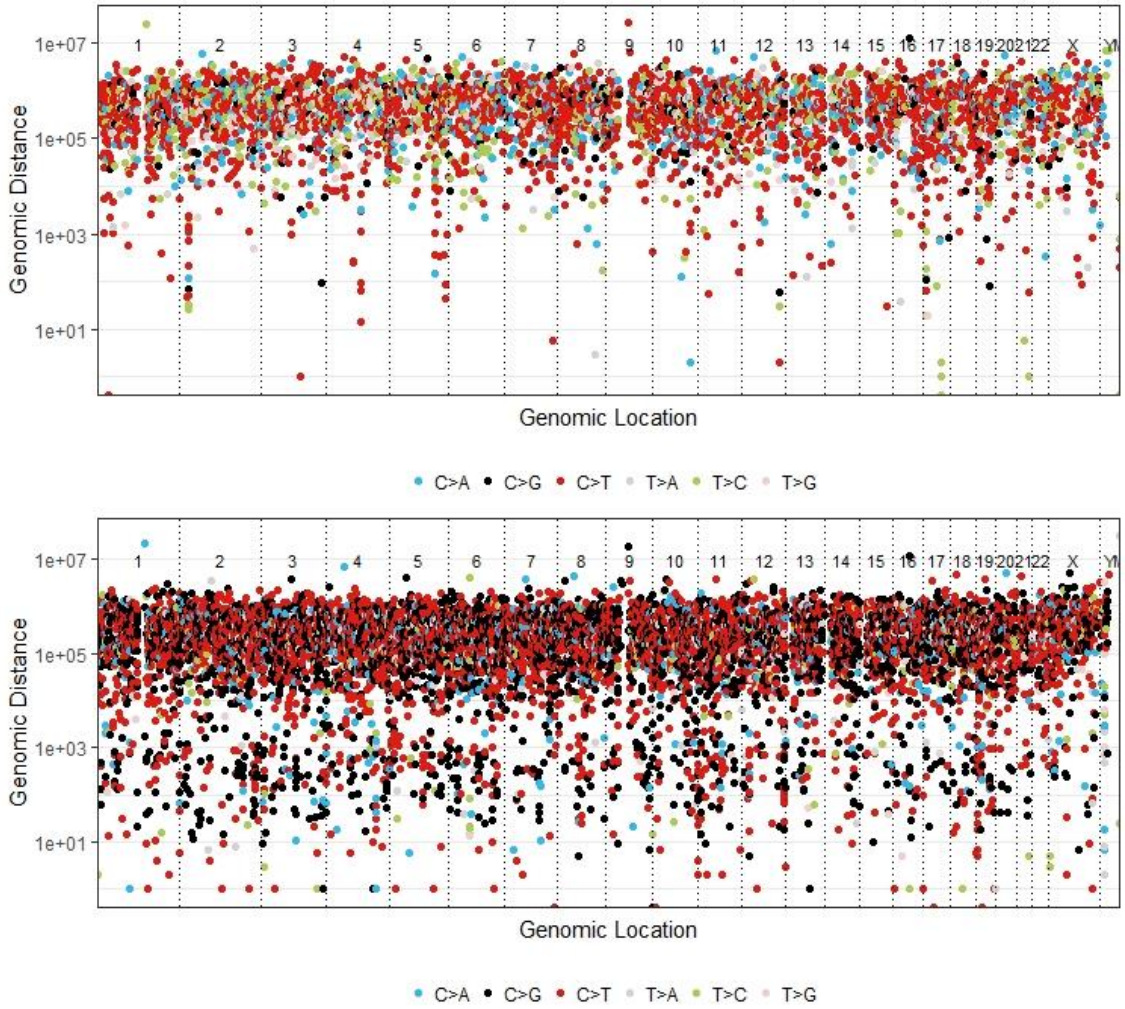


Figure 1.14: Rainfall plot of AML (upper panel) and ALL (lower panel) cancer sample in which each dot represents a single somatic mutation ordered on the horizontal axis according to its genomic coordinates in the human genome. The vertical axis, on a log scale, denotes the distance between each somatic substitution, mutation from the previous mutation. Different types of substitution mutations are colour coded. The somatic mutation data for both AML and ALL were downloaded from ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl/.

Interestingly, some cancer genomes also showed regions of somatic hypermutation, in which the mutational load is higher in a particular region within the genome (Nik-Zainal et al. 2012). This phenomenon of somatic hypermutation, is referred to as “Kataegis”, after the Greek for thunderstorm. This phenomenon, is common in several types of cancer such as breast, pancreas, lung, liver, medulloblastomas, CLL, B-cell lymphomas, acute

lymphoblastic cancers, whereas AML did not display any evidence of Kataegis (Alexandrov et al. 2013). Figure 1.15 displayed an example of Kataegis in a cancer genome. Patient with ID PD4107a showed regions of somatic hypermutations for C>T types of mutation over the background load of mutation (Figure 1.15A). Rainfall plot for patient ID PD4103a displayed Kataegis occurring at multiple loci through the genome for both C>T and C>G types of mutation (Figure 1.15B), while rainfall plot of patient ID PD4085a, shows no Kataegis Figure 1.15C.

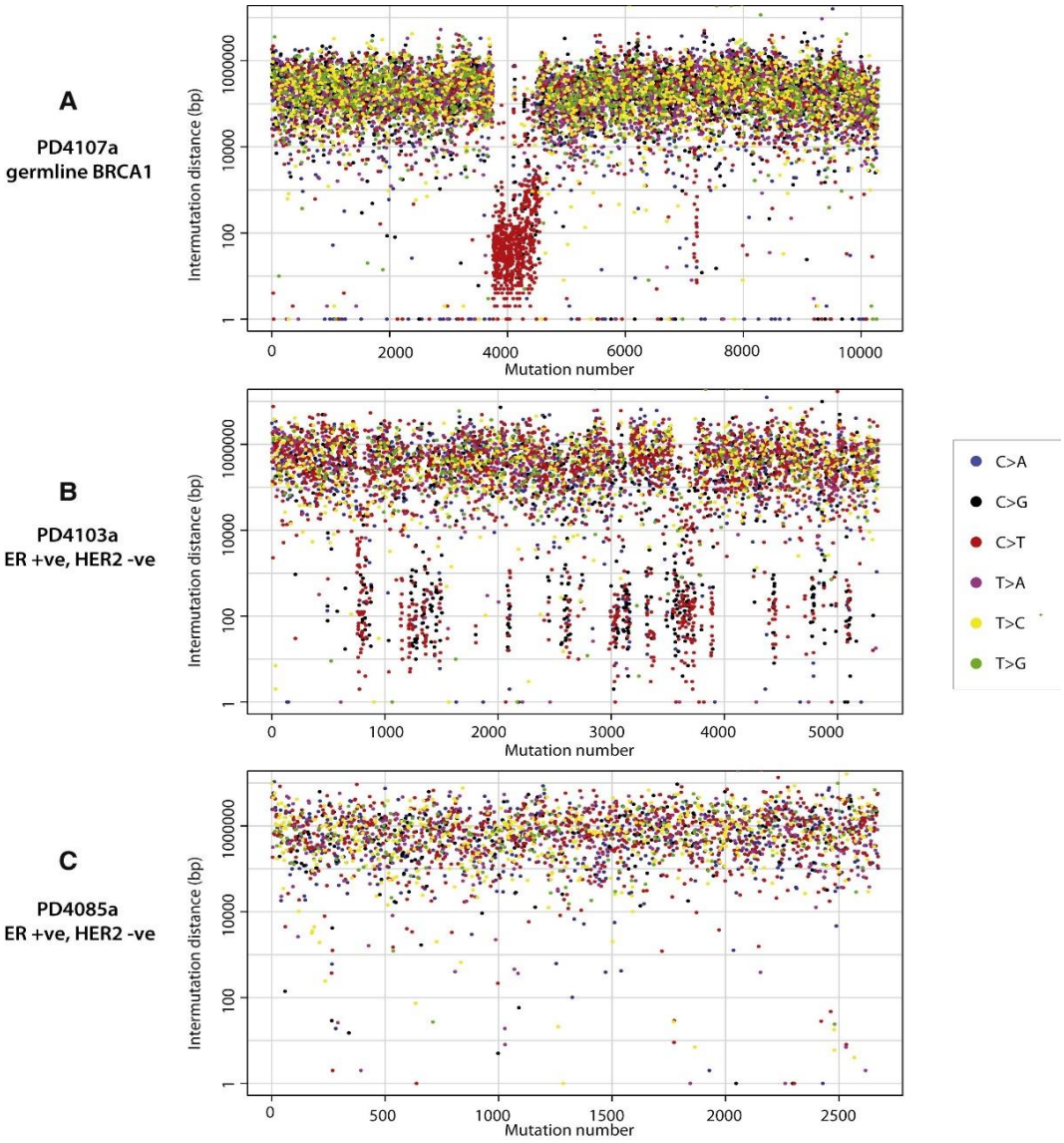


Figure 1.15: Rainfall plots of the mutational catalogue of three breast cancer patients showing regional clustering of substitution mutations. (Figure adapted and modified from (Nik-Zainal et al. 2012)).

Regions of Kataegis contain thousands of mutations commonly found in the vicinity of genomic rearrangements, suggesting mutational patterns are not random and there exists structure and organisation to their induction and distribution throughout the genome. The

underlying causes of Kataegis are proposed to be due to the function of the APOBEC family enzyme activities in some cancer types, while the AID functions are in other cancer types. The biological processes underlying these phenomena are yet to be determined.

1.6.2 Mutational strand asymmetry is observed in cancer genomes.

A recent study based on genome-wide mutational pattern analysis also revealed that, some cancer types showed significant stand biases in the distribution of mutations towards either the transcribed or non-transcribed strand within the cancer genome. This observation uncovered, UV- smoking- and liver-cancer associated mutations exhibited transcriptional strand asymmetries, whereas POLE and APOBEC associated mutations displayed replicative strand asymmetries. It was noted that, post replicative repair processes, such as DNA mismatch repair (MMR), balances these asymmetries for replication errors (Haradhvala et al. 2016). Classification of mutations found in cancer genomes according to their strand asymmetry provides important evidence for the biological impact for both the types of DNA damage, such as UV or tobacco smoke, and the DNA repair processes involved in their removal, such as NER or MMR. This observation also provides supporting evidence that mutations induced in cancer genomes are not caused by purely random events, and that there are multiple factors affecting their introduction and distribution throughout the genome.

1.6.3 Somatic mutations and mutational signatures

The concept of mutational signatures was first reported in 2012, by Alexandrov and colleagues. To decipher mutational signatures from a catalogue of somatic mutations found in cancer genomes, Dr Alexandrov used a novel mathematical algorithm called ‘SigProfiler’, which is based on a nonnegative matrix factorization (NMF) process. This system has been commonly used to extract biologically meaningful components from complex biological data sets (Lee and Seung 1999). The mutational signatures were extracted from a complex matrix generated from the mutation count matrix and by providing a sequence context to each type of the single base substitution mutation measured (Alexandrov et al. 2013b). The detailed protocol of mutational signature extraction used in this study is provided in Material and Methods section of Chapter IV. For single base substitution signatures, the profile of each signature is displayed as the contribution of each of the possible 96 trinucleotide contexts, which is composed of the 6 possible types of single base substitution (C>A, C>G, C>T, T>A, T>C and T>G) with the 4 possible types of 5' adjacent bases and 4 types of 3' adjacent bases (6*4*4=96). This

same conceptual framework of matrix factorization can be applied to examining indels or rearrangement signatures as well.

Historically, cancer research has focused on the discovery of driver mutations, primarily due to the population association studies required to attribute genetic linkage to the disease. Driver mutations represent the key somatic mutations that are causally implicated in oncogenesis. There are only relatively few driver mutations that can confer selective clonal growth advantages during the evolution of disease (Stratton et al. 2009; Alexandrov 2014). However, a cancer genome contains many more genetic changes than a mere handful of driver mutations. Each cancer can bear many thousands of passenger mutations that will not be causative of cancer development, but that are nevertheless a rich source of historical information about the processes involved in generating the genetic changes that can occur in the cell (Nik-Zainal et al. 2012; Alexandrov et al. 2013b). The passenger mutations do not result in positive selection, but because of the clonal nature of cancer, these passenger mutations are represented and serve as a record of the mutational processes that have occurred throughout the development of a cancer in an individual (Stratton et al. 2009; Nik-Zainal et al. 2012). Each mutational process leaves a characteristic imprint on the cancer genome, which is defined as a mutational signature.

The mutational signatures, which are eventually derived from the matrix factorisation process, are representative of the pattern of mutations that accrue in the genomes of tumours caused by the combined effect of both DNA damage induction and the imperfect DNA repair processes that occur in normal cells during the life time of an individual. These mutational signatures are the end-result of what is extracted from the catalogue of somatic mutations found in cancer genomes. This mutational catalogue is uncovered via tumour DNA sequencing and is the outcome of one or more mutational processes that have been operative throughout or at certain times during the development of that tumour over the life of the patient. Each mutational process is defined by the DNA damage process and operative DNA repair pathways, which each leave a characteristic imprint within the cancer genome.

By studying ~7000 cancer genomes from 30 different cancer types from the Cancer Genome Atlas (TCGA), initially the Wellcome Trust Sanger Institute (WTSI) uncovered the presence of more than 20 signatures of processes that cause mutations in DNA. Later, by analyses of 10,952 exomes and 1,048 genomes across 40 distinct types of human cancer from the TCGA and the International Cancer Genome Consortium (ICGC), the WTSI has categorized 30 reference mutational signatures in the COSMIC database. This

analysis showed that, the same signature can contribute to a number of different cancer types. It also showed that all the cancers contain two or more signatures, reflecting the variety of processes that work together during the development of cancer as illustrated in figure 1.16. However, from this figure we can see that the frequency and distribution of mutational signatures is non-random, which suggests a structure and organization with respect to the distribution of mutational signatures within cancer genomes.

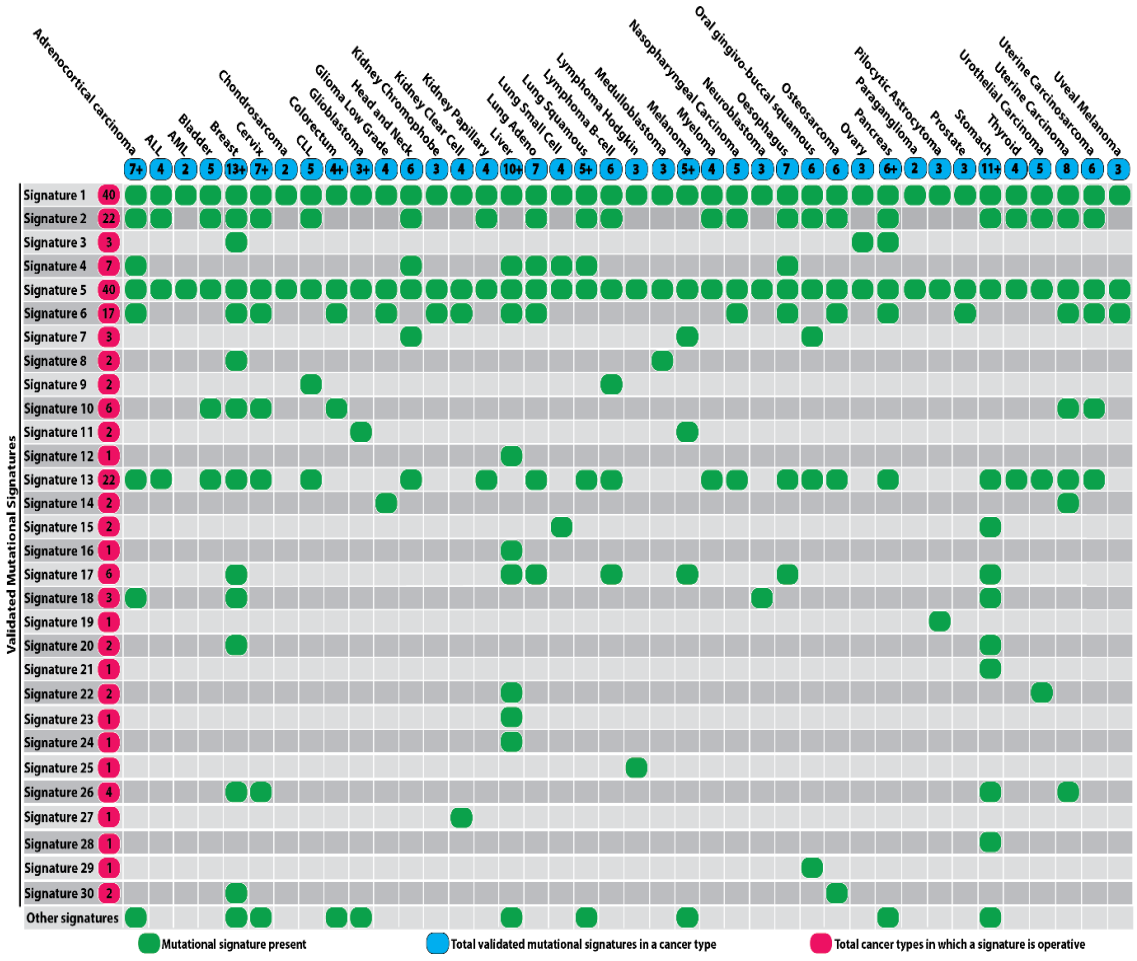


Figure 1.16: Frequency and distribution of mutational signatures across a variety of human cancer types (Figure adapted from, Catalogue of Somatic Mutations in Cancer (COSMIC) website).

Recently, the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) network working group reported around 49 Single-Base Substitution (SBS) signatures, together with 17 indel signatures and 11 dinucleotide (tandem) mutation signatures after studying ~23,000 tumour genomes across 71 different cancer types (Alexandrov et al. 2018). By examining these mutational signatures, in association with the cancer from which they originated and where known, their probable aetiology, this revealed a great deal more information about the mutagenic processes involved in carcinogenesis. Around one third

of these 49 SBS signatures have been attributed to endogenous mutagenic processes occurring during the normal metabolism of the cell. These include the activity of APOBEC family of deaminases and the deamination of 5m-Cytosine, the defective DNA repair processes including MMR, HR and BER, as well as the defective function of DNA polymerases. A further 14 are attributed to environmental exposures such as UV, PAH from tobacco smoking, aldehyde from alcohol consumption, aflatoxin B1, aristolochic acids or chemotherapeutic agents such as platinum drugs, temozolomide or azathioprine. The causes of almost half of the mutational signatures, however, are as yet not identified (Petljak and Alexandrov 2016; Letouzé et al. 2017; Alexandrov et al. 2018). This is currently a major focus of research involving global consortia in an effort to help us to understand comprehensively the fundamental causes of cancer.

In the initial cancer genome study, it was found that in melanomas and squamous skin carcinomas, the single base substitution mutation signature referred to as SBS7, was predominantly composed of C>T at C<u>C</u>N and T<u>C</u>N trinucleotides (the mutated base is underlined), together with many fewer T>N mutations. The likely biological process behind this signature was caused by UV-induced dipyrimidine formation, and subsequent translesion synthesis by error-prone polymerases, since this mutational mechanism has been previously described (Ikehata and Ono 2011). With the largest data set currently available, there are now 4 sub-classes of SBS7, including SBS7a,b,c and d. SBS7a characterised by predominance of C>T at T<u>C</u>N; similarly, SBS7b described by C>T at C<u>C</u>N and to a minor extent at T<u>C</u>N; SBS7c and SBS7d, which are characterised by predominance of T>A at N<u>T</u>T and T>C at N<u>T</u>T respectively. This study suggested that, same exposure can result in multiple mutational signatures (Alexandrov et al. 2018). Furthermore, the complexity of the mutational processes operative in some cancers and the inherent challenges in extracting their mutational signatures pose a problem for these studies. For example, the mutational catalogue from lung cancer genomes should contain the combined activity of ~60 different types of carcinogen that mutate DNA. Each of these chemicals may have their own unique mutational signature. Nevertheless, it has been mentioned that SBS4 is associated with tobacco smoking, indicating that the combined activity of multiple carcinogen exposures can generate a single mutational signature. On the other hand, MMR deficiency is associated with multiple mutational signatures (Alexandrov et al. 2013b; Alexandrov et al. 2018). Furthermore, most but not all mutagens induce mutations in tumours. Biological carcinogens that do not directly damage DNA, but rather accelerate cell divisions, thereby leaving less opportunity for

cells to repair DNA damage or errors induced during replication (Singer and Grunberger 2012).

Therefore, understanding the biological basis of these signatures is fundamental for cancer genome studies. One approach to achieving this is to extract mutational signatures from a model systems with known exposures to mutagens and/or defective DNA repair pathways. Matching of mutational signatures found extracted from these model systems with signatures from naturally occurring cancer genomes may provide clues to the causes that drive tumourigenesis in different types of cancer. These approaches applied to mutational signatures derived from thousands of human tumours will provide substantial insight into the DNA damage and repair processes that underlie the acquisition of somatic mutations across the spectrum of human cancer. In summary, the frequency and distribution of mutational signatures is non-random, suggesting that certain features and organization of the genome in cells influences the mutational end-points that are observed in tumours.

1.6.4 Novel Cancer Genes

It is well established that DNA modifications have an influence on mutagenesis. For example, spontaneous deamination of methylated cytosine is only one of many mutational processes caused by epigenetic phenomena. However, whole genome sequencing studies of cancer genomes have also revealed novel cancer genes, that were not identified by genome-wide association (GWAS) studies. This may have been due to insufficient power of the sample size to reach statistical significance for more subtle genetic effects. Many of these suggested novel cancer-causing genes are involved in chromatin modification, chromatin remodeling and DNA repair (Kandoth et al. 2013a; Papaemmanuil et al. 2013). For instance, the landscape of somatic mutations derived from the 560 breast cancer whole genome study uncovered ~93 driver genes involved in the development of breast cancer, many of those are involved in chromatin modification and remodeling as well as DNA repair (Nik-Zainal et al. 2016). Mutational signature analysis from WCG or WES provided critical insights into DNA repair defects and exposure to mutagenic processes during cancer progression. It is well established that chromatin and the epigenomic context influences DNA damage and repair pathway choices (Wang et al. 2007a; Wang et al. 2007b). Recently, some exome sequencing data implicates mutations in DNA repair, chromatin modifier or chromatin remodeling genes, which impose higher risk for multiple myeloma, ovarian clear cell carcinoma, B-cell lymphomas and Kabuki syndrome (Jones et al. 2010; Ng et al. 2010; Lunning and Green 2015; Waller et al. 2018). But how

these novel cancer genes involved in chromatin remodeling or modification drive tumorigenesis is currently unknown.

1.7 Structure and organisation of genome-wide DNA damage and repair

In this section I will describe the recent advances in our understanding of the structure and organisation of genome-wide DNA damage and repair in the context of both DNA and chromatin. I will focus particularly on how UV-induced DNA damages are formed and efficiently repaired by the NER pathway within the cellular environment. Collectively, these observations will demonstrate the importance of understanding how genetic damage is formed and efficiently repaired within the cells.

Several decades of research into the fundamental biochemical mechanisms of DNA repair have revealed the basic enzymatic functions of the multiple pathways that evolved in cells to recognize, remove and correct a bewildering variety of lesions that frequently occur in the genomes of cells (Friedberg et al. 1995). Within these DNA repair pathways, one of the major ones is known as nucleotide excision repair (NER), and a great deal is known about its fundamental molecular mechanism (Friedberg 2003; Marteijn et al. 2014). Nucleotide excision repair (NER) acts on a spectrum of DNA damage that have the common property of distorting the DNA double helix to some degree, and this feature is thought to be important for the recognition stage of repair. Over thirty polypeptides are involved in the basic NER reaction. Two damage-recognition pathways exist (Figure 1.17): the transcription coupled repair pathway (TC-NER) that operates on the transcribed strands of transcribing genes and involves RNA polymerase II in damage recognition; and the global genome repair pathway (GG-NER) that operates on all DNA, including non-transcribing and repressed regions of the genome, involving a unique subset of proteins in the early stages of DNA damage recognition (Fousteri and Mullenders 2008). Following the initial stages of DNA damage detection, these two pathways converge and utilise the same DNA repair proteins to complete the later stages of repair. The majority of yeast NER genes have well-conserved structural and/or functional human homologues, and the main features of both the GG-NER and TC-NER pathways are evolutionarily conserved (Hoeijmakers 1993; Hoeijmakers 1994).

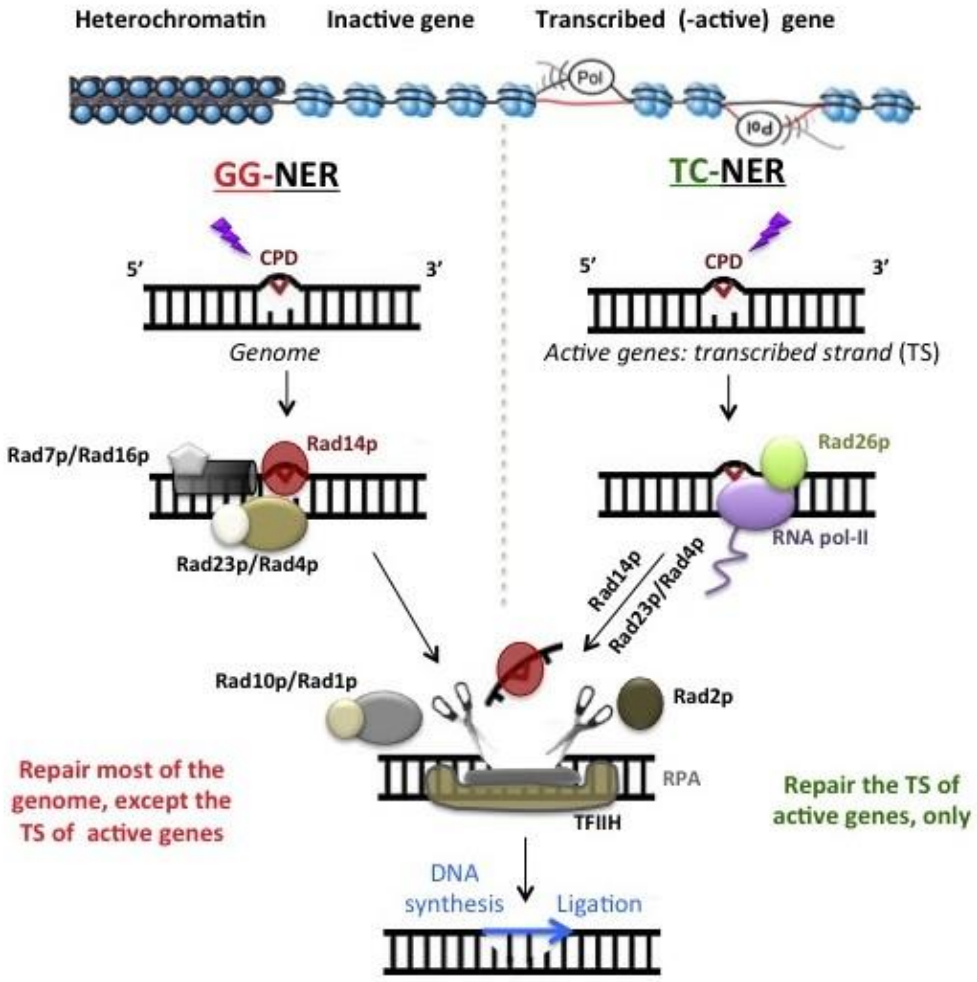


Figure 1.17: NER in yeast removes UV induced DNA damages. It is performed by a large multi-enzymatic complex made of more than 30 proteins, that repairs DNA via distinct steps: recognition of the lesion, incision of the damaged DNA strand upstream and downstream of the lesion, excision of the resulting ~30 nucleotide of DNA fragment containing the lesion, filling the gap by DNA synthesis, and ligation of the newly synthesized patch (Figure adapted from http://www.conconilab.ca/projects).

In more recent years, these studies have been expanded to show how the NER process operates throughout the entire genome. Novel approaches for detecting and representing DNA damage and repair throughout whole genomes have helped to explain the distribution of genome-wide DNA damage and the relative rates of repair observed throughout the genome. Microarray and next generation sequencing-based techniques have been developed to address how DNA damage is induced and repaired within the genomic context. For example, our laboratory developed a genome-wide DNA repair assay based on ChIP-Chip (Teng et al. 2011; Powell et al. 2015). This method relies upon the affinity capture of UV-induced DNA damage and its separation from undamaged

regions of the genome. The genetic damage can be identified by hybridisation of fluorescently labelled DNA to whole-genome DNA microarrays, revealing a map of both the levels and locations of the damage, and its heterogeneous distribution throughout the genome (Figure 1.18A) (Teng et al. 2011; Powell et al. 2015). Repeating this process at different times after induction of damage permits the calculation of relative DNA repair rates at individual sites throughout the genome, which also reveals a heterogeneous distribution (Figure 1.18B) (Teng et al. 2011; Powell et al. 2015). In addition to plotting of the data in relation to the linear genome as shown, it is possible to collate and display these genomic DNA repair rates as composite plots around ORFs, or indeed any other genomic feature (Powell et al. 2015), revealing the pattern of DNA repair in relation to these genomic features, as shown in Figure 1.18C.

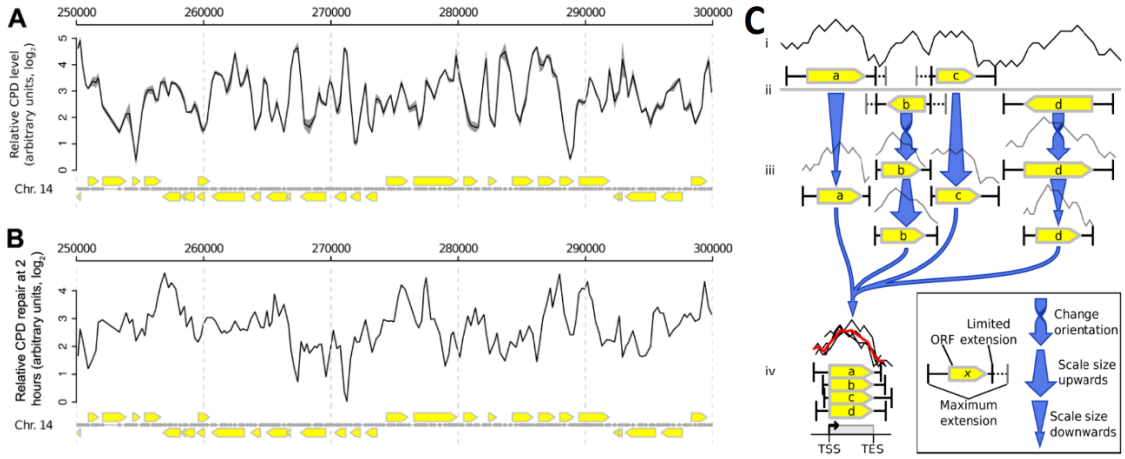


Figure 1.18: Genome-wide UV-induced DNA damage is heterogeneously distributed and DNA repair is organised around ORF structure. (A) A linear genome plot is shown here highlighting positions 250,000 to 300,000 of yeast chromosome 14 depicting the results of the 3D-DIP-Chip from wild-type cells. The gray dots on the linear plot indicate the position of the probes on the microarray. The yellow arrows indicate the ORFs and their direction of transcription. The results from the 3D-DIP-Chip experiment are shown as the average arbitrary log2 ratios of IP over Input for three independent biological experiments. (B) CPD repair rate displayed in a linear genome plot. (C) This cartoon depicts the process of generating composite plots using Sandcastle (Bennett et al. 2015). A collection of genomic positions corresponding to a genomic feature (e.g. TSS or transcription factor binding site) are taken and the data in a window surrounding that site are compiled into a composite profile from which a trend line is generated. Enhanced rates of repair in ORFs are due to TC-NER of actively transcribed genes. (Figure adapted and modified from (Yu et al. 2016).

Comparison of genome-wide DNA repair rates between wild-type and various mutant strains defective in DNA repair or repair-related chromatin remodeling can then be studied to determine the effect of the mutation on the normal distribution of relative repair rates throughout the genome. Figure 1.19 (black line) demonstrates the wild-type pattern of DNA repair rates after induction of DNA damage following UV irradiation, revealing an even rate of repair in intergenic regions and increased rates in the ORFs, where TC-NER will contribute to overall repair rates along with GG-NER (Hanawalt and Spivak 2008). In response to UV, deletion of the GG-NER factor Rad16, an ATP-dependent SWI/SNF superfamily chromatin remodeler, the histone modifier GCN5 (a histone acetyl transferase) or the histone variant (Htz1) significantly alter the distribution of wild-type repair rates throughout the genome. This observation suggests that, defects in the mechanism of GG-NER, histone modification or nucleosome components, might also alter the pattern of mutations induced throughout the genome. *These findings led us to pose key questions that represent the central aims of this thesis: How is NER organised within the genome, and to what extent do changes in the distribution of DNA repair rates caused by defects in this organisation affect the pattern of UV-induced genomic mutations?*

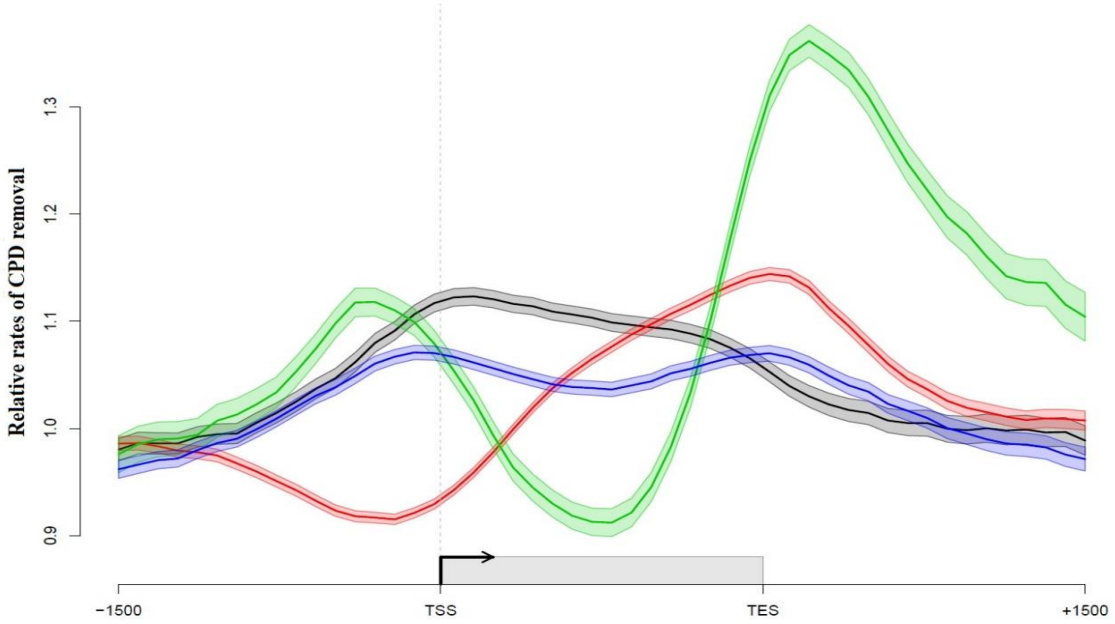


Figure 1.19: Relative rates of CPD repair around ORF structures. Solid lines show the mean of CPD repair rates in wild-type (n = 3, black line), *rad16* (n = 2, red line), *gcn5* (n = 2, green) and *htz1* (n=2, blue line) mutant cells. Shaded areas indicate the standard deviation, with CPD levels plotted as arbitrary units on the y-axis. (Data used here for

plotting was generated by previous colleague in our laboratory, and ORF plot was made using Sandcastle (Bennett et al. 2015).

In the nucleus, DNA is packaged into the nucleoprotein complex of chromatin. At present, how NER operates on naked DNA is well understood, but detailed knowledge of how it operates in chromatin is still emerging (Adam et al. 2015). To study how DNA damage is recognised and removed from DNA packaged into chromatin, our laboratory decided to examine the genome-wide locations of GG-NER factor chromatin binding both before and after UV irradiation using the ChIP-Chip and ChIP-seq techniques. These studies have recently culminated in the demonstration that GG-NER is organised and initiated from specific genomic locations (Yu et al. 2016; van Eijk et al. 2018). However, these recent reports were predicated on earlier studies that identified a protein complex of Rad7, Rad16 and Abf1 from the yeast *Saccharomyces cerevisiae*. This complex is now referred to as the GG-NER complex. Our lab showed that efficient GG-NER requires Abf1 to be bound to specific DNA binding sites (Reed et al. 1999). These can be found at hundreds of locations throughout the yeast genome (Yu et al. 2009). The Rad16 protein is a member of the SWI/SNF super-family of chromatin remodeling factors. Proteins in this super-family contain conserved ATPase motifs and are subunits of protein complexes with chromatin-remodeling activity (Flaus and Owen-Hughes 2011). Since Rad16 operates on repressed and non-transcribed regions of the genome during GG-NER, it has long been assumed that its role might involve chromatin remodeling (Verhage et al. 1994), conceivably, to improve access to damaged DNA. Rad16 also contains a C3HC4 type RING domain, which is an important motif in ubiquitin E3 ligase proteins. Our laboratory reported that the GG-NER complex also has E3 ubiquitin ligase activity involving the Cul3 and Elc1 proteins, explaining the results reported by other researchers (Gillette et al. 2006). Rad7 is part of an E3 ligase complex that ubiquitinates Rad4 a core yeast NER protein that binds to damaged DNA (Gillette et al. 2006). This ubiquitination of Rad4 in response to UV irradiation specifically regulates NER via a pathway that requires de novo protein synthesis and it directly influences NER and UV survival. Importantly, it is established that Rad7 and Rad16 exist in a complex within the cell and that deletion strains of either component showed identical phenotypes (Verhage et al. 1994; Reed et al. 1998).

Initial studies examining events at a single genetic locus, showed that the GG-NER complex promotes UV-induced chromatin remodeling necessary for DNA repair by recruiting the histone acetyl-transferase (HAT) Gcn5 onto chromatin. This promotes

increased histone H3 acetylation levels in the locale that in turn alter chromatin structure (Yu et al. 2011). Later using whole genome ChIP-Chip assays, this study also showed that chromatin occupancy of the histone acetyl-transferase Gcn5 is controlled by the GG-NER complex in response to UV damage, which regulates histone H3 acetylation and chromatin structure, thereby promoting efficient DNA repair of UV-induced lesions (Yu et al. 2016). During these studies, our lab also reported that the presence of histone H2A variant, Htz1 (H2A.Z) in nucleosomes has a positive effect in promoting efficient NER in yeast (Yu et al. 2013). Htz1 is well-known to enhance the occupancy of the histone acetyltransferase Gcn5 on chromatin and promotes histone H3 acetylation after UV irradiation. It was reported that this series of events results in increased binding of one of the core NER factors, Rad14, to damaged DNA. Cells lacking Htz1 exhibit both increased UV sensitivity and defective removal of UV-induced DNA damage in the Htz1-containing nucleosomes located at the repressed *MFA2* promoter, but not in the HMRa locus where Htz1 is absent. Based on all the results obtained from these studies, the following model of how GG-NER is organised and functions in response to UV was proposed (Figure 1.20).

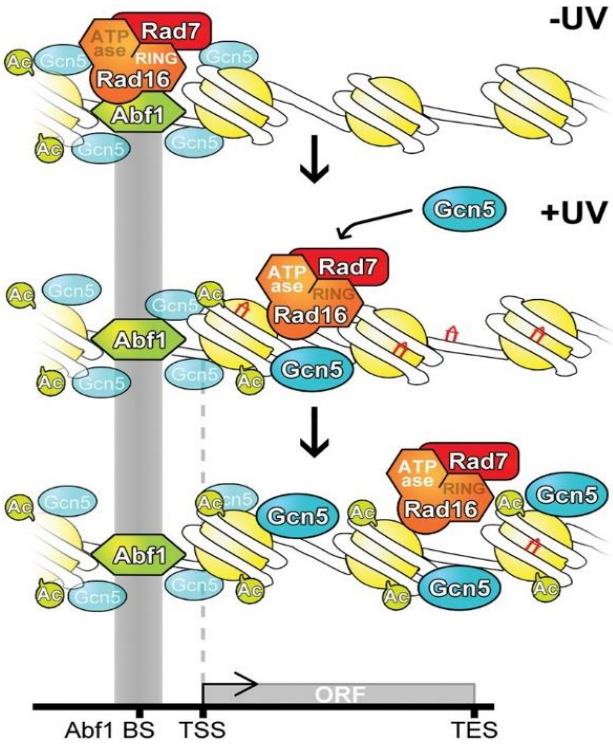


Figure 1.20: Model to illustrate how GG-NER is organized in the yeast genome. (Top panel) In undamaged cells, the GG-NER complex is located at multiple Abf1 binding sites predominantly in the promoter regions of genes. This occupancy is dependent on the RING domain of the Rad16 protein. The enrichment of GG-NER-independent basal

levels of Gcn5 can be detected at these sites. (Middle panel) In response to UV irradiation, the GG-NER complex dissociates from the Abf1 component at Abf1 binding sites. This process depends on the activity of the ATPase domain in Rad16. Concomitantly, the HAT Gcn5 is recruited onto the chromatin with its increased levels and distribution dependent on the Rad7-Rad16 GG-NER complex. (Bottom panel) During this process, histone H3 acetylation is increased over a domain defined by the redistribution of the Rad7-Rad16 proteins from Abf1 binding sites. This mechanism drives the chromatin remodeling necessary for the efficient repair of UV damage (Figure adapter from (Yu et al. 2016)).

Recently, using improved ChIP-seq techniques, the precise location of the GG-NER complex binding sites at nucleotide resolution was reported (van Eijk et al. 2018). The significance of this to my thesis is that measuring GG-NER complex binding at nucleotide resolution makes it possible to map nucleosome structure in relation to these binding sites. My contribution to this study involved the genome-wide mapping of nucleosomes in yeast cells, both before and after UV irradiation, to identify the GG-NER-dependent changes in the linear nucleosome structure in response to UV damage. The importance of this becomes clear when examining the precise position of UV-induced mutations in the genome, which are also mapped at nucleotide resolution. The key findings of the report showed that GG-NER is organized and initiated from a specific subset of novel genomic features now referred to as GG-NER Complex Binding Sites (GCBSs). In response to UV damage, the GG-NER complex remodels dynamic (Htz1-containing) nucleosomes immediately adjacent to GCBS's, to promote efficient DNA repair in the surrounding regions. This remodeling permits redistribution of GG-NER complex components to neighbouring regions of chromatin, where they promote the efficient repair of DNA damage. This remodeling process is dependent on the function of the GG-NER complex, since defects in its key functional domains results in a significant reduction of DNA repair rates in the surrounding regions of the genome. Remarkably, it was found that GCBSs frequently map precisely to the boundaries of newly-identified chromosomally interacting domains (CIDs) in the genome (Hsieh et al. 2015). These boundaries define domains of higher-order nucleosome-nucleosome interaction (Hsieh et al. 2015; Hsieh et al. 2016). It was suggested that organizing GG-NER into these higher-order chromatin domains, likely reduces the genomic search space for the detection of DNA damage, ensuring the efficient recognition and repair of DNA lesions throughout the genome. The loss of this higher-order genomic organization within chromatin, due to defective GG-NER-dependent chromatin remodeling, and damage-induced histone modifications,

causes major disruption to the distribution of relative repair rates in the genome. It is conceivable that such alterations might also affect the distribution of mutational patterns found throughout the genome, and this will be investigated in detail in this thesis.

1.8 Genome-wide mutational patterns and their relationship to genomic structure

Understanding the underlying causes contributing to the formation of mutational signatures and mutational patterns derived from whole-genome mutation data requires prior knowledge of genome-wide DNA damage and repair dynamics (Alexandrov et al. 2013b; Poon et al. 2014).

The mutational patterns associated with DNA damaging agents have been examined by studying reporter genes or plasmid-based assays following exposure to mutagens, such as UV light (Pfeifer et al. 2005). These studies identified the factors, including the position of DNA damage, the ability of NER to remove the lesions, as well as the mutagenic or error-free bypass of lesions by DNA damage tolerance mechanisms, that ultimately give rise to a diagnostic, UV damage-induced mutational pattern. More precisely, the sequence specific pattern of UV-induced CPD or 6-4PP (Sage 1993), the error free bypass of T-T CPD or 6-4PP by pol eta (Johnson et al. 2000), the possible elevated levels of deamination of adducted cytidines (Burger et al. 2003), and the error-prone bypass of T-C or C-C CPD (Ikehata and Ono 2011), together with the low frequencies of misincorporation by translesion polymerases of T and G opposite thymines in pyrimidine dimers, rather than the more frequent and non-mutagenic A (Wang 2001; Ikehata and Ono 2011), all these factors combine in a complex way to result in a UV-induced mutational pattern with the following hierarchy C>T, T>C and T>A. Comparison of these mutational patterns with the mutations identified in tumour suppressor genes such as TP53 in human cancers provided the initial evidence for the causative role of sunlight in skin cancer (Pfeifer et al. 2005). Large-scale, whole-genome and exome sequence analysis of mutations found in human melanomas demonstrate that exposure to UV light is the primary cause of mutations found in these tumours (Alexandrov et al. 2013). Further sequencing of additional human cancer genomes has identified additional context-specific mutational patterns consistent with exposure of cells to other DNA damaging agents. For example, the C>A substitution caused by tobacco exposure found in lung cancers of smokers (Alexandrov et al. 2013).

Initial investigations utilizing next-generation sequencing technology showed that the mutational rate across a variety of human tumour genomes, as well as the human germline, is heterogeneous, suggesting the different aspects of chromosome structure and

function, such as difference in replication timing and transcription activity affecting strand asymmetry may be playing an important role in the pattern of mutations induced (Stamatoyannopoulos et al. 2009; Hodgkinson et al. 2012). Studies showed that during the cell cycle, late replicating regions have higher mutation densities, and a strong negative correlation exists between mutation densities with the levels of transcription in these areas (Lawrence et al. 2013). The current explanation for these regional differences includes an increased error rate in late replicating regions, and the activity of the TC-NER pathway to remove damages in these areas. These suggestions are supported by several studies that show a higher rate of UV, tobacco and aristolochic acid-induced mutations occur on the non-transcribed strands of genes in melanoma, lung and urinary tract cancer respectively (Pleasance et al. 2010b; Alexandrov et al. 2013; Hoang et al. 2013).

Chromatin structure itself also appears to have an effect on the distribution of regional mutational density in cancer genomes. Regions of heterochromatin showed higher mutational density than euchromatic regions, and it was suggested that DNA repair factor accessibility to open chromatin is likely to be an underlying cause (Schuster-Böckler and Lehner 2012). Other reports showed that a lower rate of somatic mutations in cancer genomes, specifically in accessible regulatory DNA, is the result of an intact global genome repair pathway (Polak et al. 2014). Furthermore, a high level of UV-induced mutation was associated with repressive histone modifications across the genomes of tumours from GG-NER deficient patients (Zheng et al. 2014).

In addition to sequence context specificity of DNA damaging agents, regional differences in the DNA repair processes, presumably due to the genomic architecture within chromatin, can also influence the mutation levels throughout the genome. Two studies involving human cells, found that the molecular machinery that initiates gene transcription, prevents repair proteins from accessing DNA, resulting in increased mutation rates at sites of promoter or transcription-factor binding (Perera et al. 2016; Sabarinathan et al. 2016) (Figure 1.21).

Taken together, these studies indicate that the structure and organisation of the DNA repair mechanisms in the genome may influence the mutational patterns generated within them.

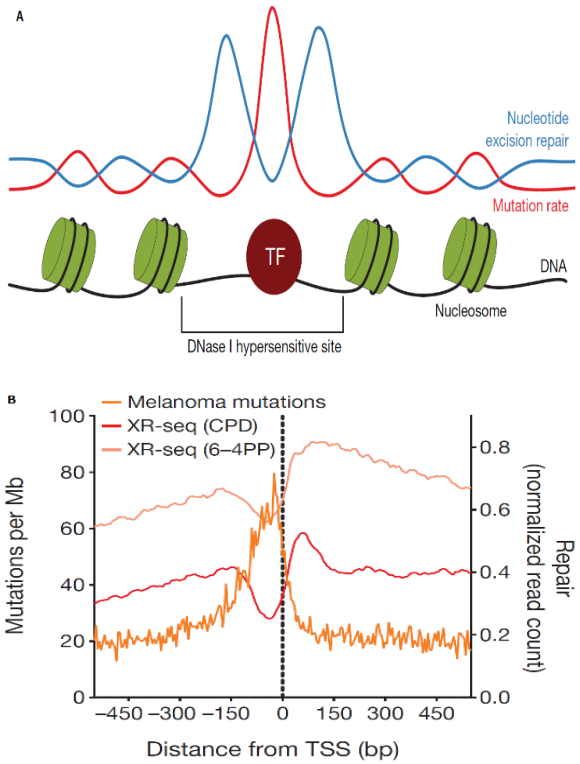


Figure 1.21: Model showing the mutation rate and repair rate in TFBS and nucleosome sites (A), Average melanoma mutation and XR-seq profiles for CPD and 6–4PP in normal skin fibroblast (B) showing mutation rate inversely correlate with DNA repair rate at transcription start sites. (Figures adapted from (Perera et al. 2016; Sabarinathan et al. 2016)).

Although these comparative analyses of the mutational patterns in cancer genomes in relation to genome structure and organisation have advanced our understanding of the significance of genomic DNA damage induction and its repair, this type of analysis has its limitations. The major drawback is the disparity between the cell types used to generate the datasets available for comparison. These can range from tumour cells from patients to study mutations, through to established cell lines to investigate DNA repair rates. Therefore, I chose to examine these events using yeast as a model organism, where DNA damage and repair events can be compared in isogenic strains in a highly controlled fashion, so that a clearer understanding of the relationship between DNA damage, repair and mutation can be revealed.

1.9 Yeast as a model organism to study genome stability

Saccharomyces cerevisiae (commonly known as baker's yeast), has been used extensively as a model organism for studying cellular processes in evolutionarily distant species, including humans (Botstein et al. 1997). *S. cerevisiae* was the first eukaryotic organism

whose complete genome sequence was announced in 1996. Since then, yeast has continued to be a pioneer organism that has facilitated the advancement of 'functional genomics' and 'systems biology' approaches (Botstein and Fink 2011). In the field of genomics, yeast has been used as a model organism to study the function of processes that affect humans including aging and cancer (Millot et al. 2012; Guaragnella et al. 2014; Cazzanelli et al. 2018). More importantly, DNA damage and repair play important role in the development of cancer and yeast has been used as a model for understanding how DNA damage and repair is organised within the genome (Sancar et al. 2004; Kunkel and Erie 2005; Reed 2011; Teng et al. 2011; Yu et al. 2011). It has also been suggested that yeast can be used as a potential isogenic model organism for understanding the biological process of mutational signature observed in human cancers (Alexandrov et al. 2013b). Our laboratory has developed methods for measuring genome-wide DNA damage and repair using yeast as a model organism (Teng et al. 2011; Bennetft et al. 2015; Powell et al. 2015) and used these to determine how these events are organised with the genome (Yu et al. 2016). In this thesis, I will extend these studies to examine how they impinge on the pattern of mutations induced in the genome. The availability of a well-annotated reference genome and ability to accumulate mutations over generations, ensures the capture of the mutagenic processes active in the cell. It is also possible to clonally expand the yeast cells from a single cell, which is the basis for the clonality found in cancer genomes (Larrea et al. 2010; Lujan et al. 2014; Serero et al. 2014; Segovia et al. 2015).

1.10 Aims of the current study

In this study, I will examine how DNA repair is structured and organised within the yeast genome in relation to the nucleosome structure before and in response to UV damage. To do this, I will use MNase-seq for mapping genome-wide nucleosome structure, and its changes in response to DNA damages, in both wild-type and GG-NER deficient yeast cells. Correlating these changes in nucleosome remodeling in response to UV-damages and comparing them with genome wide DNA damage and repair rates will explain how UV-induced DNA damages are efficiently repaired by the GG-NER pathway in chromatin.

I will then establish a workflow for the study of somatic mutational patterns induced in isogenic yeast strains. Here, I will examine events in wild-type, NER defective, chromatin modifier defective or histone variant defective yeast strain, to study their effect on the mutational patterns induced in response to UV induced DNA damages.

After complete genome sequencing of these yeast cells exposed to UV damage, or defective in the various DNA repair or chromatin modifier pathways, I will examine the mutational patterns and signatures induced using the algorithms employed in the cancer genome sequencing projects reported recently (Nik-Zainal et al. 2012; Alexandrov et al. 2013b). Comparing these with the mutational signatures described from analysis of cancer genomes will provide novel insight into the biological processes that generate mutational signatures observed in cancer genomes.

Our genome-wide DNA repair studies suggest that altering the distribution of DNA repair rates in the genome might also alter the type and distribution of genome-wide mutational patterns, which are linked to several diseases.

Comparing the distribution of relative repair rates with the distribution of mutation from cells with the same genetic background will enable us to establish the precise relationship between the processes of DNA damage, repair and mutation.

These studies will determine how GG-NER remodels chromatin for efficient repair in response to DNA damage. How defects in chromatin remodeling alter DNA repair rates in the genome, and how this affects the distribution of mutational patterns observe will explain how genomic instability gives rise to specific mutational signatures that drive tumorigenesis. This will provide insight into the causes of the mutational spectra observed in cancer genomes.

Chapter II

Material and Methods

Contents

This chapter provides details about the methodologies developed during this study for measuring genome-wide binding of chromatin associated factors, nucleosome mapping and mutagenesis following UV irradiation using *Saccharomyces cerevisiae* (baker's yeast) as model organism. These include basic microbiology techniques such as culturing, propagation and maintenance of yeast along with the routine techniques employed for UV treatments, DNA and chromatin isolation, fragmentation with sonication and quantitative real-time PCR (RT-PCR). Detailed methodology related to (i) ChIP-Seq: Chromatin immunoprecipitation and sequencing, (ii) MNase-Seq: Micrococcal nuclease digestion and sequencing and other assays used for subsequent chapters are also outlined here. The DNA sequencing was performed by using Wales Gene Park sequencing facilities available within the Cardiff University campus.

Originally, our laboratory developed ChIP-Chip (Teng et al. 2011) or 3D-DIP-Chip (Powell et al. 2015) for measurement of gneome-wide DNA damage and repair. Same techiniques also used to measure the chromatin accupency of DNA binding proteins. Additionally, this protocol was adopted to detect cisplatin and oxaliplatin induced DNA damage in human cells as well (Powell et al. 2015). Here I describe, how I modified the ChIP-Chip protocol to move it to support next-generation sequencing (NGS) downstream processing, referred to as ChIP-Seq, using similar techniques as described in our previous work for affinity capturing of the DNA binding proteins with target specific antibodies. However, instead of hybridizing on a microarry, I then sequenced the immunoprecipitaed (IP) and corresponding input (IN) samples after capturing the DNA bound proteins. In a similar way, for measuring the gneome-wide nucleosme occupency, I adopt the MNase-Seq technique. The library preparation for ChIP-Seq or MNase-Seq experiment are also mentined here. These libraries were sequenced on Ion-Proton platform by our colleagues at the Wales Gene Park Sequencing facility. After getting raw sequence data back, I checked the quality of each sequence file and the subsequent bioinformatic analysis using these NGS data. Thus, in the relevant chapter, I will cover the specific yeast strains, computational and bioinformatic methods used for data analysis. The chemical and reagents used for these routine experiments are mentioned in Appendix I.

2.1 Yeast cell culture

All the yeast strain used in this study, were taken from glycerol stocks, were first streaked either onto yeast extract peptone dextrose (YPD) plates or in a small portion of YPD liquid culture media. This allowed cells to recover before proceeding with the protocols mentioned below.

Yeast cells were grown in YPD liquid medium at 30^{o}C with 180 rpm in an INFORS HT multitron incubation shaker. For most experimental purposes, cells were grown to log-phase which is equivalent to OD ~0.6 and cell count is ~ $2x10^7$ cells/mL. The cell growth was monitored by two different methods. First, 1 mL of the cell cultures was taken out and used to measure the optical density at 600 nm with a Jenway 6300 Visible Spectrophotometer, blanked against YPD liquid media without cells. Second, cell density was counted using Hawksley improved cell counting chamber. Each strain was checked for their generation time and set up to a dilution for overnight growth to get mid log-phase cells the next day. UV irradiation was performed next morning, and cells were stored depending on purpose of study. For mutational studies, the yeast cells were grown on YPD agar plate at 30°C in a LEEC compact incubator until colonies had formed (typically 2-3 days). Plates were stored at 4°C for weeks. For long term storage, cells from plates were stored in glycerol and frozen at -80°C.

2.2 UV irradiation

For UV irradiation, log-phase cells were collected by centrifugation and resuspended in ice-cold 1x PBS to $2x10^7$ cells/mL as described previously (Yu, 2011, Yu 2016). Using 10 J/m^2s^1 UV-C (254 nm) for 10 seconds, cells were irradiated with 100 J/m^2 . The cell suspension was kept in the dark to prevent photoreactivation and cells were spun down and resuspended in fresh YPD media and incubated fixed amount of time (depending on the experiment) at 30°C to allow repair to take place. The following steps were followed:

1. Overnight grown log-phase yeast cells in YPD media (density of 2~$4x10^7$ cells/mL) were collected by centrifugation at 6,000 rpm for 5 minutes using a Sorvall Evolution RC Superspeed Centrifuge (Thermo Scientific).

2. Yeast cells were then washed once with ice-cold 1x PBS and resuspended in cold PBS to the same starting cell density of 2~$4x10^7$ cells/mL. Fifty mL of cells were poored into a 150 mm diameter Pyrex dish for UV irradiation.

3. Using 10 J/m^2s^1 for 10 seconds, cells were irradiated with 100 J/m^2 UV-C (254 nm). The intensity of the UV light was measured by digital UVX Radiometer (UVP's) using UVX-25 sensor which is sensitive for 254 nm UV-C light.

4. After UV irradiation, cells were collected by centrifugation, followed by resuspension in YPD media for repair to take place or resuspended in YPD media for downstream applications including crosslinking by formaldehyde if necessary. All unirradiated control cultures were processed similarly and subsequent processes were carried out in the dark to maintain identical experimental conditions.

Note: For the mutagenesis study, the first UV treatment was performed on cells suspensions in PBS, whereas the following exposures were performed on plate. The details protocol will be mentioned in next chapter.

2.3 Crosslinking

Preparation of chromatin for ChIP experiments requires cross-linking of protein-DNA complexes to assess the genomic occupancy or our protein of interest. After the indicated repair time in YPD, cells were treated with formaldehyde to a final concentration of ~1% to crosslink protein-DNA complexes and incubated at room temperature shaking at 180 rpm for 15 minutes. The crosslinking reaction was quenched by adding 125 mM final concentration of Glycine and incubated at room temperature with 180 rpm shaking for 5 minutes. The cells were harvested by centrifugation and washed twice with ice-cold 1x PBS. The final wash was performed using cold FA/SDS buffer (see the Appendix I) and the cells were transferred to a 2 mL Eppendorf tube. After collecting the cells by centrifugation at 6,000 rpm for 5 minutes, they were snap-frozen in liquid nitrogen and stored at -80°C for long term storage.

Note: For accumulation of mutations from whole genome sequences (WGS), cells were snap-frozen in liquid nitrogen without crosslinking. Following UV treatment, the log-phase cells washed three times with 1x PBS and stored.

2.4 Preparation of yeast chromatin

To investigate the protein-DNA or protein-protein interactions using ChIP-Seq, it is necessary to prepare chromatin from cells. Formaldehyde cross-linked cells were lysed using glass beads and vortexing. DNA was fragmented using sonication with a Bioruptor to a size of approximately 300bp. The maximum number of cells used for chromatin preparation was $1x10^8$ cells/sample from haploid yeast strains. All the steps performed

during chromatin purifications are carried out at 4°C and on ice. The following steps were followed:

1. UV irradiated, and formaldehyde cross-linked cells were re-suspended in 500 µL FA/SDS (+PMSF) in a 2 mL microcentrifuge tube.

2. Yeast cells were lysed mechanically by adding 500 µL of 425-600 µm glass beads (Sigma-Aldrich) to each sample and vortexing on a Disruptor Genie® vortex mixer for 10 minutes at 4°C in dark.

3. The chromatin was separated from the beads by puncturing a hole in the 2 mL microcentrifuge tube with a hot needle prior to the placement of the microcentrifuge tube in a 15 mL Falcon tube.

4. The lysate (~500 µL) was collected in the 15 mL Falcon tube by centrifugation at 2,000 rpm for 2 minutes at 4°C, followed by washing the glass beads with 500 µL of FA/SAS(+PMSF) buffer with further centrifugation at 2,000 rpm for 2 minutes at 4°C.

5. The cell lysate was transferred to a 2 mL microcentrifuge tube and centrifuged at 12,000 rpm for 20 minutes at 4°C in a microfuge (Beckman Coulter, 22R centrifuge) to remove any soluble proteins not cross-linked.

6. After removing the supernatant, the pellet was resuspended in 1 mL FA/SDS (+PMSF) buffer and sonicated using a BioRuptor sonicator (Diagenode) with power set at the 'high' for 8 to 10 cycles of 30 seconds on and 30 seconds off at 4°C (as described in section 2.5).

7. Following sonication, a further centrifugation was performed at 13,000 rpm for 10 minutes at 4°C and the supernatant was collected and centrifuged again at 13,000 rpm for 20 minutes at 4°C (both using Beckman-Coulter Microfuge 22R). Supernatant (Chromatin) was collected into fresh 1.5 mL microcentrifuge tubes.

8. The protein content was quantified using the Bradford assay (Bio-Rad). The supernatant containing the chromatin was snap frozen in liquid nitrogen and stored at -80°C for subsequent analysis.

2.5 Yeast chromatin fragmentation by sonication

Chromatin was fragmented using a Bioruptor® sonicator (Diagenode) to get desired fragment for ChIP-Seq. Four 2 mL round bottom tubes, each containing 300 µL of chromatin samples, were placed in a 2 mL microtube unit placed in a cooling water bath. The power at Bioruptor® machine set to high and sonication was conducted at 4°C cool

water tank for 30 seconds ON and 30 seconds OFF for each cycle. For microarray hybridization an average fragment size of 500 bp was required but in the case of NGS shorter fragments were desired. Sonication condition employed were as follows:

Yeast Chromatin - 12 cycles of 30s ON, 30s OFF

Following sonication, the 2 mL microcentrifuge tubes were centrifuged at 12,000 rpm in a bench top centrifuge for 20 minutes at 4°C. Finally, the supernatant was transferred to a fresh 1.5 mL microcentrifuge tube. The fragmented chromatin was snap frozen with liquid nitrogen and stored at -80 °C. The length of the fragmented products was confirmed by agarose gel electrophoresis after purifying the DNA and using the FastRuler low range DNA ladder (Fermentas) as reference.

2.6 Chromatin immunoprecipitation (ChIP)

Chromatin prepared as in section 2.5 was subjected to immunoprecipitation using specific antibodies raised against the protein of interest and magnetic Dynabeads. The antibody and Dynabeads serve to pull down or immunoprecipitate the DNA fragments where the protein of interest has bound (an immunoprecipitated or IP sample). In addition, a control input (IN) sample was obtained without immunoprecipitation. Following this, both IP and IN samples undergo cross-link reversal, pronase and RNase digestion before DNA purification. The following steps were followed:

1. For each sample to be immunoprecipitated, 50 µL of Dynabeads™ Protein G (Invitrogen) was taken into a 1.5 mL microcentrifuge tube and washed 3 times with 500 µL PBS-(0.1% BSA) (see Appendix I). For example, for 6 samples to be immunoprecipitated, 6x50 = 300 µL of Dynabeads™ Protein G was used.

2. After washing, the Dynabeads™ Protein G was resuspended in 100 µL PBS (0.1%BSA) per sample before the addition of a specific antibody. An antibody titration experiment was performed beforehand to determine the amount of antibody to add per sample for optimal immunoprecipitation.

3. The mixture of Dynabeads and antibody was incubated at 30°C, for 30 minutes at 1,300 rpm in an Eppendorf thermomixer. At this stage the antibody should attach to the Dynabeads.

4. The Dynabeads were collected using a DynaMag-2 Magnet (Invitrogen) held against the tube and washed 3 times with 500 µL PBS (0.1% BSA) to get rid of unbound antibody. Between washing steps, after adding 500 µL PBS (0.1% BSA), Dynabeads were resuspended by placing the tube into a rotor and place the tube back into DynaMag-2

Magnet. After the final wash, the Dynabeads were re-suspended in 50 µL of PBS (0.1% BSA) per sample. From this Dynabead master mix, 50 µL of beads was added to 100 µL of sonicated chromatin (~2-3µg) . In addition, 30 µL 10 x PBS-BSA (10 mg/mL) was added and the final volume was adjusted to 300 µL in total with PBS. This was incubated at 21°C at 1,300 rpm for 3 hours in an Eppendorf Thermomixer.

5. After the incubation, the ChIP reaction mixture was collected using a DynaMag-2 Magnet (Invitrogen) and the supernatant were removed. The beads were washed with 500 µL FA/SDS buffer as described before. This was followed by a series of washes: 2 washes with 1 mL FA/SDS +NaCl buffer; 1 wash with 500 µL LiCl buffer and finally with 500 µL cold TE (see Appendix I).

6. After the final wash, the DNA was eluted off the Dynabeads with 125 µL of Pronase buffer (see Appendix I) at 65°C, at 900 rpm for 20 minutes in an Eppendorf Thermomixer. Following this the supernatant was transferred to a fresh 1.5 mL microcentrifuge tube and 6.25 µL of Pronase (20mg/mL, Roche) was added before incubation overnight at 65°C in a water bath.

7. For the input (IN) samples, 50 µL of sonicated chromatin was adjusted to 100 µL with TE buffer, before the addition of 25 µL 5 x Pronase Buffer. As like the IP samples, the input samples all had 6.25 µL of Pronase added and were incubated overnight at 65°C in a water bath (Clifton).

8. The following day, the IP and IN samples were all treated with 2 µL of 10 mg/mL RNase A (Fisher Scientific) and incubated for 1 hour at 37°C. After the incubation, the DNA was purified using Invitrogen PureLink PCR purification kit by following the manufacturer's instructions and eluted into 50 µL elution buffer. From this stage the IP and IN DNA could either be used in a qPCR (to quality check for IP) or used as the starting point for genome-wide analysis.

2.7 Quantitative real-time PCR (qPCR)

The qPCR is performed using the iTaq™ Universal SYBR® Green Supermix (Bio-Rad) and CFX Connect™ Real-Time PCR Detection System (Bio-Rad). The *MFA2* gene targeted primers were used for this experiment in Yeast. The *MFA2* promoter primer sequences are:

Forward: 5´ - AAAGCAGCATGTTTTCATTTGAAACA - 3´ and

Reverse: 5´ - TATGGGCGTCCTATGCATGCAC - 3´.

The SYBR green dye fluorescence technology-based method used in this assay is both qualitative and quantitative. SYBR green fluorescence increases up to a 1,000-fold while bound to dsDNA compared to its unbound fluorescence. Fluorescence was measured during each PCR cycle of amplification.

During each qPCR experiment a set of standards were included, where one input (IN) sample was serially diluted 10-fold to provide a range of dilutions from 10^{-1} to 10^{-6}. These standards were used as a reference for relative DNA quantification between samples. For quantification of yeast samples, IN samples were diluted 1,000-fold and the immunoprecipitated (IP) samples diluted 5-fold. All dilutions were conducted in water. 5 µL of these diluted samples were added to a 5 µL mix of SYBR green and primer pair to achieve a final primer concentration of 500 nM, and a final volume in each well of 10 µL. qPCR reactions were performed in Hard-Shell® 96 well PCR plates (Bio-Rad). Each PCR reaction was performed in triplicate including the standards. Plates were sealed with plastic film, mixed briefly by vortexing and centrifuged at 2,000rpm for 2 minutes. For this studies the qPCR reaction was performed under the following conditions:

1. 95°C for 3 minutes
2. 95°C for 15 seconds
3. 57°C for 20 seconds
4. 57°C for 20 seconds – followed by optical image
5. Go to step 2 x 45 times
6. 95°C for 1 minute
7. Melt curve from 65°C to 95°C, increment 0.05 °C

For 0.05 – followed by optical image

End.

2.8 Preparation of yeast nucleosomal DNA for MNase-Seq

Micrococcal nuclease **(**MNase) digestion and sequencing adapted for nucleosome position mapping as described previously (Kent et al. 2011). Briefly:

1. $8x10^8$ log-phase yeast cells per samples were resuspended in 250 µL of 1 M sorbitol solution. To create spheroplasts, 200 µL of freshly prepared YLE buffer (see Appendix I, 1M sorbitol, 11.25 mM 2-mercaptoethanol, and 22.5 mg/mL yeast lytic enzyme) was added.

2. Cells were incubated at 22 °C for 3 minutes to remove cell walls, and then collected by a pulse spin at 12,000 RCF in a microcentrifuge.

3. The pellet was washed gently in 1 M sorbitol and re-centrifuged as above but with the tube rotated 180° relative to the previous spin.

4. Spheroplasts were resuspended into 400 µL of digestion buffer containing 1 M sorbitol, 50 mM NaCl, 10 mM Tris-HCl (pH7.5), 5 mM MgCl2, 1 mM CaCl2, 1 mM 2-mercaptoehanol, 0.5 mM spermidine, 0.075 % NP40 and transferred to a 1.5 mL Eppendorf tube. Following resuspension into digestion buffer, 4 µL of 30 U/µL MNase (Thermo Scientific) was added and cells were incubated at 37 °C for 10 minutes.

5. The cell suspension was then centrifuged at 14,000 RCF for 5 seconds and supernatant was quickly transferred to a fresh tube containing 40 µL of STOP solution (5 % SDS, 250 mM EDTA) and mixed well to terminate the reaction.

6. 10 µL of 10 mg/mL RNase A (Thermo Scientific) was added and incubated at 37 °C for 1 hour for complete removal of any RNA released. Next, 20 µL of pronase (20mg/mL) was added to each sample and the samples were incubated at 65°C overnight to reverse the protein-DNA cross-link and digest all DNA bound protein.

7) Two phenol/chloroform 1:1 (v/v) and one chloroform extraction were performed as mention in section and the aqueous layer (~450 µL) was transferred into to fresh tube each time.

8. DNA was precipitated by adding 45 µL of 3M Sodium Acetate (NaAc) (Invitrogen™) followed by 1 mL of chilled absolute ethanol (Fisher scientific). DNA pellets were collected by centrifugation (how fast, how long) and air dried after washing the pellet with 70% freshly prepared ethanol (Fisher scientific).

9. Finally, 100 µL of TE buffer was added to resuspend the DNA, and then further purified by PureLink™ PCR Purification Kit (Invitrogen), eluted into 50 µL elution buffer. The quality of the MNase prep. was check by running the purified nucleosomal DNA in 1.5% agarose gel.

2.9 Ion-Proton library preparation for ChIP-Seq and MNase-Seq

This library preparation protocol was adapted from the Life-Technologies Ion ChIP-Seq Library Preparation on the Ion Proton^{TM} System, Publication Number 4473623 Revision B. With minor modification to the blunt-ending, DNA purification and amplification as described below. I use ≤10 ng of DNA from ChIP or MNase-prep, quantified by using the Qubit® 2.0 Fluorometer. In the case of IP samples, when samples are below the detection limit of the Qubit®, then all the purified DNA (40-50 µL) was used after PureLink PCR purification (Invitrogen). The following steps were followed:

PreCR repair: 40 µL or ≤10ng of IP samples and 10 ng of Input samples diluted into 40 µL with ddH_2O per sample was taken in a 1.5 mL Eppendorf LoBind® Tube and 10 µL the PreCR repair mixture (NEB) was added. This reaction mixture was incubated for 20 minutes at 37°C to repair any UV induced damages. DNA was purified with magnetic beads using 1.8x sample volume of beads (CleanNA) according to the Life-Technologies ChIP-Seq protocol. It is important to use freshly (same day) prepared 70% ethanol for washing the beads for each purification. The DNA sample was eluted into 50 µL low TE.

End-repair & blunt ending: T4 DNA polymerase (NEB) was used for end-repair and blunt ending. 50 µL eluted DNA from previous section was combined with 0.2 µL of T4 DNA polymerase (3,000 U/µL) in 1x NEB buffer 2 with 1.0 µL dNTPs (10mM) in an end-volume of 120 µL. This reaction was mixed well by pipetting and incubated for 30 minutes at room temperature. Again, the DNA was purified using 1.8x sample volume of beads (CleanNA) according to the Life-Technologies ChIP-Seq protocol. This time the end-repaired DNA was eluted into 40 µL low TE. (optional: the DNA can be stored at 4°C at this point)

Ligation: Ligation mixtures were prepared using 40 µL eluted DNA, T4 DNA ligase (NEB, 400 U/µL) and the universal P1 adapter and A barcoded adapter (Eurofins) and 10x Ligase buffer in an end-volume of 100 µL according to the LifeTechnologies ChIP-Seq protocol. Care was taken to prevent cross contaminating between samples with the barcoded adapters. Ligation was performed for 30 minutes at room temperature. Adapter ligated DNA was purified using 1.5x sample volume of beads (CleanNA) according to the Life-Technologies ChIP-Seq protocol and eluted in 40 µL low TE (optional: the DNA can be stored at 4°C at this point).

Note: this step was critical, because longer incubations introduce adapter-adapter concatemers that get amplified during the next step. The P1 and other barcoded adapters (Eurofins) used for these thesis are attached in Appendix I, Table A1.1.

Nick repair & amplification: Q5 High Fidelity polymerase (NEB) was used to perform the nick repair reaction and amplification steps. These steps were performed in an end-volume of 100 µL. 40 µL purified DNA from previous steps was combined with 2 units of Q5 HF polymerase and 1 µL forward and 1 µL reverse amplification primers (20mM) (Eurofins), 20 µL 5x Q5 reaction buffer, 2.5 µL dNTPs (10 mM), made up to 100 µL using ddH_2O and mixed by pipetting up and down.

The reaction mixture tubes were placed into a thermal cycler by following the PCR cycling program:

Step	T	Time
Nick Repair	72°C	20min
Denature	95°C	5min
Denature	97°C	15sec
Anneal	60°C	15sec
Extend	72°C	60sec
		18 cycles
Hold	70°C	5min
Hold	4°C	-

It is possible to optimise the number of PCR cycles if the input DNA is much lower than 1 ng. The amplified DNA was purified with magnetic beads using 1.5x sample volume of beads (CleanNA) and eluted in 40 µL low TE. 2 µL of this purified DNA was used for analysis on a TapeStation 2200 (Agilent Technologies) before size selection.

Size-selection and DNA purification using SPRI beads: Magnetic SPRI beads (CleanNA) were used according to the Life-Technologies ChIP-Seq protocol to size-select the library prep. The first round of purification, a 0.7x sample volume of SPRI beads was used to selectively capture DNA >350bp, keeping the library DNA in the supernatant. During the second step 80 µL of SPRI beads were used (approx. 0.5x of sample volume), binding all DNA >160bp to the beads. Finally, the library DNA was eluted in 25 µL low TE.

Quantification & quality control: The quality of the prepared library was checked before and after size selection by running the samples on a High Sensitivity tape on the TapeStation 2200 (Agilent Technologies). DNA concentration was quantified by using a 2 µL sample on the Qubit and converted into nM using:

$$[nM] = \frac{\left[\frac{ng}{\mu L}\right]}{660 \times DNA} \times 1 \cdot 10^6$$

A diluted 100pM of pooled libraries were prepared and sent out for sequencing. The emulsion PCR on the Ion Chef requires 75pM. At this stage samples were submitted for

sequencing to our colleagues at the Wales Gene Park. The pooled library and individual DNA preps can be frozen at this point for long term storage.

2.10 Preparation of yeast genomic DNA

For yeast DNA extraction, firstly the cells were treated with Yeast Lytic Enzyme (YLE, MP BIOMEDICALS) with final concentration 1 mg/mL to create spheroplasts, which could then by lysed with a lysis buffer containing sodium dodecyl sulphate (SDS). Following RNase-A and Pronase treatment, to digest the RNA and protein respectively, the DNA was extracted using phenol/chloroform and precipitated using ethanol. The maximum number of cells used for genomic DNA isolation mentioned here was $5x10^9$ cells/sample (250 mL $2x10^7$ cells/mL) and this yields on average 300-500 μg of genomic DNA from haploid yeast strains. Depending of further application, the purified DNA are used for sequencing experiments or kept at -20°C for long term storage.

The following steps were followed:

1. Cells in PBS for each sample (either UV treated or untreated), were collected by centrifugation at 5,000 rpm for 5 minutes and transferred to a 15 mL falcon tube. Following removal of supernatant, cells were washed once with 5 mL Sorbitol-TE solution (see Appendix I).

2. Cells were re-suspended in a freshly prepared 5 mL Sorbitol-TE-YLE solution (see Appendix I) and mixed well by shaking. Cells were incubated for either 30 minutes at 37°C with occasional shaking or alternatively incubated overnight at 4°C in the dark with rotation (Labnet Revolver™). The production of spheroplasts can be monitored using a light microscope. For this 5 μL of cell suspension and 5 μL of 5% SDS were mixed and investigated under a microscope. There will be no or few intact yeast cells due to the addition of SDS, compared to spheroplasts not treated with SDS.

3. Spheroplasts were then gently centrifuged at 4,000 rpm for 5 minutes (Beckman-Coulter Microfuge 22R) at 4°C and re-suspended in 5 mL of lysis buffer/PBS 1:1(v/v) solution (see Appendix I). 300 μL 10 mg/mL of RNase-A (Sigma, reconstituted from dry powder using manufacturer protocol) was added to each sample, mixed well by vortexing and incubated at 37°C for 1 hour with occasional shaking. After this 200 μL Pronase (Roche, 20 mg/mL in TE buffer, prepared from dry powder according to manufacturer protocol) was added to each sample and incubated at 37°C incubator for 1 hour and then at 65°C water bath (Clifton) for 1 hour, with occasional shaking. A clear solution at this stage indicates complete lysis of the cells and successful deproteination.

4. An equal volume of phenol/chloroform 1:1 (v/v) was added, mixed well and centrifuged at 10,000 rpm for 10 minutes with Avanti™ JA-20 rotor (Beckman Coulter).

At this stage two phases formed, with cell debris and denatured proteins forming the interphase between two pages. The nucleic acids were in the aqueous upper phase and were carefully transferred to a new 15 mL polypropylene tube without breaking the interphase.

5. To ensure complete deproteinization a second phenol/chloroform 1:1 (v/v) and third chloroform/isoamylalcohol (24:1) extractions were performed as mentioned in the first extraction. The absence of protein precipitate at the interphase was indicative of complete deproteinization. Finally, the aqueous phase was transferred to a new 15 mL falcon tube.
6. DNA was precipitated by adding double volumes of chilled absolute ethanol (-20°C) to each sample with gentle shaking by inverting the tube, prior to storing the samples at -20°C overnight or -80°C for 30 minutes.

7. DNA pellets were collected by centrifugation at 10,000 rpm for 10 minutes at 4°C. The pellets were air dried and then re-suspended in 1 mL TE buffer before reprecipitation with addition of 1 volume chilled isopropanol (1 mL). The samples were gently shaken until the DNA became visible in the solution. The DNA precipitate was removed with the end of a pipette tip, squeezed dry against the tube wall, and transferred to a fresh 1.5 mL centrifuge tube containing 500 μL TE. Alternatively, the DNA precipitate can be collected by centrifugation at 12,000 rpm for 10 minutes at room temperature and re-suspended in 500 μL TE.
8. After the DNA was fully dissolved in the TE, the quality of the DNA was checked using non-denaturing agarose gel electrophoresis. In addition, the concentration of the DNA was measured using Qubit 2 Fluorometer (Invitrogen).
9. The DNA was kept at 4°C for short-term or at -20°C for longer term storage.

2.11 Library preparation for Illumina Mi-Seq and Hi-Seq sequencing

For ChIP-Seq and MNase-Seq the Wales Gene Park Ion Proton™ System facilities were used (Life Technologies) after preparing the libraries as described in the previous sections.

For whole genome sequencing, the Illumina Mi-Seq and Hi-Seq platforms were used. The library preparation and yeast whole genome sequencing for Illumina Mi-Seq or Hi-Seq were performed by our colleagues at the Wales Gene Park sequencing facilities. Briefly, 300 ng of genomic DNA of each sample was sheared in an end-volume of 55 μL to obtain 200bp fragments using the Covaris® ME220 Focused-ultrasonicator™. The sheared samples were then cleaned using a 1.8x Agencourt AMPure® Bead clean up (Beckman

Coulter®). The sheared samples then underwent library construction using the NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (E7645, New England Biolabs®). This process involved end-repair, adaptor ligation, size-selection using AMPure® XP beads PCR enrichment (5 cycles) of the adaptor-ligated DNA using NEBNext® Mulitplex Oligos for Illumina® (Dual Index primers set 1 & 2) (E7780 & E7600) followed by a clean-up of the PCR product using AMPure® XP beads. The libraries were validated using the Agilent 2100 Bioanalyser and a high-sensitivity kit (Agilent Technologies) to ascertain the insert size, and the Qubit® (Life Technologies) was used to perform the fluorometric quantitation. Following validation, the libraries were normalized to 4 nM, pooled together and clustered on the cBot™2 following the manufacturer's recommendations. The pool was then sequenced using either 75-base paired-end (2x75bp PE) or 150-base paired-end (2x150bp PE) dual index read format on the Illumina HiSeq® 4000 and Illumina MiSeq® system respectively, according to the manufacturer's instructions.

2.12 Quality control of the raw sequence data using FastQC

After getting raw sequence data back, I first checked the quality of the sequences for all the ChIP-Seq, MNase-Seq and whole genome sequence (WGS) data. The raw fastq sequence files were checked for the quality of the sequence reads using FastQC; the quality control tool for high throughput sequence data (Andrews 2010). FastQC can use fastq or bam files to report on read quality. The FastQC output is a HTML file that summarizes the findings. Low quality data were re-sequenced. All the results of these quality-controlled data are attached with this thesis as an e-Appendix file.

2.13 Data analysis

The NGS data were aligned with the BWA mem algorithm using sacCer3 as reference genome (Li and Durbin 2010). DNA occupency sites were determined by the MACS2 peak calling algorithm using the input sequence data to normalise the IP data. The details subsequent analysis for ChIP-Seq and MNase-Seq data will be provided in the Capter III. And the bioinfomratic analysis for WGS data will be provided in the Chapter IV.

Note: For bioionformatic data analysis, I took help from Dr. Patrick van Eijk, post doctoral research associate in our laboratory.

Chapter III

Global-Genome Nucleotide Excision Repair is initiated from a novel class of genomic features

Contents

3.1 Background

The rates at which lesions are removed by DNA repair can vary widely throughout the genome with important implications for genomic stability. Previous studies from our laboratory described genomic tools for the analysis of genome-wide DNA damage and repair rates. After measuring the distribution of nucleotide excision repair (NER) rates for UV-induced DNA damages throughout the budding yeast genome, these studies revealed that, in normal cells, genomic repair rates display a distinctive pattern, indicating that this DNA repair pathway is highly organised within the genome. After comparing the genome-wide DNA repair rates in wild-type (WT) cells and cells defective in the chromatin remodeling function of the GG-NER process (*rad16* mutant), UV-induced histone modification (*gcn5* mutant), or the damage-induced exchange of the histone variant HTZ1 (*htz1* mutant), it was noted that these mutant strains significantly alter the distribution of NER rates through the genome.

Previously, a complex of proteins was purified from yeast cells that are uniquely required for global genome nucleotide excision repair (GG-NER) in this simple eukaryote (Reed et al. 1999). The complex, known as the GG-NER complex, is comprised of the SWI/SNF superfamily member, Rad16, the Rad7 protein and the yeast general regulatory factor (GRF), Abf1. Recently, it has been reported that GG-NER is organised into domains around ORF structure, thus promoting efficient repair in the surrounding regions of the genome (Yu et al. 2016). This report showed that loss of GG-NER function severely affected repair rates around the promoter regions of genes containing Abf1 binding sites. Based on ChIP-Chip experiments, there are ~4,000 Abf1 binding sites found across the yeast genome (Zentner et al. 2015; Yu et al. 2016). It is well established that the GRF Abf1 exists in excess over the other GG-NER components (Rad7 and Rad16) in the cell, and has a wide range of different functions outside of GG-NER (Yarragudi et al. 2007; Schlecht et al. 2008; Ganapathi et al. 2010; Zhang et al. 2012). Consequently, in order to examine GG-NER function in relation to nucleosome structure, the linear arrangement of nucleosomes in the genome, a refined list of Abf1 binding sites was required to identify a more precisely measured set of GG-NER complex binding sites. This is necessary in order to permit the accurate mapping of nucleosome positions in relation to the GG-NER complex binding sites. To do this, Abf1 ChIP-seq experiments were performed to map the genome-wide occupancy of Abf1 in chromatin at nucleotide resolution, as a precursor to identifying the precise location of GG-NER complex binding sites. Distinctive genomic features associated with Abf1 binding, include Abf1 consensus binding

sequences (n = 1,752), and genome-wide Nucleosome Free Regions (NFR, n = 6,589), with which Abf1 occupancy is frequently associated, as described previously (Yarragudi et al. 2007; Hartley and Madhani 2009; Ganapathi et al. 2010; Ozonov and van Nimwegen 2013). Performing these experiments will enable the precise mapping of GG-NER complex binding sites in relation to these features.

It is well established that, efficient repair of UV-induced DNA damage requires chromatin remodeling. Previous studies in our laboratory, demonstrated that the GG-NER complex regulates UV-induced histone H3 acetylation, by controlling chromatin occupancy of the histone acetyl transferase, Gcn5 on chromatin (Teng et al. 2008). This UV-induced hyperacetylation of histones promotes an open chromatin conformation required for efficient repair of DNA damage (Yu et al. 2005; Teng et al. 2008). Additionally, it also revealed that, the histone H2A variant, HTZ1 (H2A.Z), in nucleosomes has a positive function in promoting efficient NER in yeast. HTZ1 inherently enhances the occupancy of the histone acetyltransferase Gcn5 on chromatin to promote histone H3 acetylation after UV irradiation (Yu et al. 2013). Consequently, this results in increased binding of an important NER recognition factor encoded by Rad14, to damaged DNA. A broad range of research also has revealed a role for histone modification and histone variant exchange in a variety of DNA repair pathways, including NER (Adam et al. 2015; Polo 2015; Polo and Almouzni 2015). However, at this stage, the precise details of how these modifications enhance repair of damage from chromatin and its effect on the distribution of mutations remains to be determined.

In this chapter, I describe the experiments performed to determine how chromatin is remodeled during GG-NER, to permit efficient repair of UV-induced DNA damage in yeast cells. These include experiments to map the precise genomic location of GG-NER complex binding in chromatin, that were used to establish a novel class of genomic features that are now referred to as GG-NER complex binding sites (GCBS's). Based on previous studies by my laboratory colleagues, I reasoned that the binding of the GG-NER complex to these sites located throughout the genome, might establish the chromatin structure at the level of the linear organisation of nucleosomes. I wanted to determine how the alteration of this structure in response to DNA damage promoted efficient repair of DNA damage in the genome. To investigate this, I performed MNase-seq (Jiang and Pugh 2009; Zhang and Pugh 2011) experiments, and adapted them to map nucleosomes across the entire genome to reveal their physical organisation, and to measure changes in their structure in response to UV-induced DNA damage. Changes to the physical

organisation of nucleosomes is known to control the accessibility of DNA to certain DNA binding proteins including transcription factors, replication factors and DNA repair complexes, thereby regulating these activities within the cell. In this chapter, I report the linear organisation of nucleosome structure in the genome in relation to the chromatin occupancy of the GG-NER complex. These studies helped to identify the boundaries of repair domains and to demonstrate the importance of GCBS's for initiating the efficient repair of the genome from these origins of GG-NER.

The work described in this chapter established the genomic locations and functional importance of GCBS's, a novel class of genomic feature, which frequently map to the boundaries of the newly-identified chromosomally interacting domains (CIDs) in the genome (Hsieh et al. 2015). These domains define regions of higher-order nucleosome-nucleosome interaction within the genome (Hsieh et al. 2015; Hsieh et al. 2016). Our observations show that repair of DNA damage by GG-NER is organised and initiated from GCBS's, and that the GG-NER complex remodels dynamic nucleosomes located immediately adjacent to this novel class of genomic features, to promote efficient DNA repair in the genome.

3.2 Material and methods

3.2.1 Yeast strains used in this study

The yeast strain used for this study, their genotype and the assay used are mentioned in table 3.1. To delete the RAD16 gene from the HA-Htz1 epitope-tagged W303-1B strain, a *RAD16::HIS3* disruption construct residing on pUC18 was used. The pUC18 *RAD16::HIS3* (Reed et al. 1998) was digested using EcoRI and BamHI and used for transforming yeast. The lithium-acetate transformation was used and selected for successful genomic integration of the disruption construct on His- selection plates. From the transformation plate 12 individual colonies were re-streaked on fresh media for single colony PCR and confirmation of UV phenotype. Successful clones were stored as glycerol stocks and used for the detection of genome-wide H2A.Z occupancy using ChIP-seq in the absence of RAD16.

Table 3.1: Yeast strains and their respective genotype used in this study

Yeast strain	Genotype	Assay
BY4742 (Wild-type)	MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0	Abf1 CHIP-Seq MNase-Seq
rad16	W303-1B *RAD16Δ::HIS3*	MNase-Seq
HA-HTZ1	W303-1B HA-HTZ1::KanMX	CHIP-Seq
HA-HTZ1 *rad16*	W303-1B HA-HTZ1::KanMX *RAD16::HIS3*	CHIP-Seq

3.2.2 Experimental overview

The full details regarding yeast cell culture, UV irradiation and crosslinking, chromatin preparation, immunoprecipitation, micrococcal nuclease digestion, DNA extraction and Ion- proton library preparation for ChIP-Seq/Mnase-Seq techniques are mentioned in the Material and Methods Chapter II (also see appendix II Figure A2.1 – A2.4). For Abf1 ChIP-Seq experiments 2 µg of Abf1 antibody (Abf1 (y-90): sc-25755, Santa Cruz Biotechnology) was used during immunoprecipitation. For Htz1 ChIP-Seq assay, 2 µg of anti-HA tag monoclonal antibody (Millipore, Cat. # 05-904) was used. The optimum antibody concentrations was determined by antibody titration assays using qPCR. Following Ion-proton library preparation, the DNA sequencing was performed by using Wales Gene Park sequencing facilities available within the Cardiff University campus.

3.2.3 Data analysis & access

NGS data was aligned to the reference genome (sacCer3) and processed for downstream analysis to detect peaks (MACS2), map nucleosomes (DANPOS) or NFRs (HOMER). Subsequent annotation data was assigned to relevant features using the ChIPpeakAnno package (Zhu et al. 2010), which was also used for the calculation of overlaps and drawing of Venn diagrams presented in this chapter. Full details of all bioinformatics data analysis are included in the next section. The Micro-C boundary positions data was obtained from the supplementary data accompanying the Hsieh et al. Manuscript (Hsieh et al. 2015). A list of genome-wide NFRs was obtained from (Yadon et al. 2010). The list of Abf1 consensus motifs can be obtained using the multiple EM for motif elicitation (MEME) (Bailey et al. 2006) with find individual motif occurrences (FIMO) algorithm (Grant et al. 2011). The data described in this chapter was submitted to ArrayExpress (https://www.ebi.ac.uk/fg/annotare/) and can be retrieved using accession code E-MTAB-6569.

3.2.4 Data processing

The trimmed Fastq files were aligned to the sacCer3 reference genome using the BWA-MEM 0.7.12-r1039 (Li and Durbin 2010) and piped into samtools 0.1.17 (r973:277) (Li and Durbin 2009) to convert the output to a sorted BAM file. These BAM files were used as input for downstream processing using MACS2 and DANPOS (see next section).

3.2.5 Peak detection of ChIP-seq data using MACS2

In order to perform peak detection with MACS2 of two biological replicates, we merged the sorted bam files (input and IP) using samtools as suggested by the MACS2 developer notes before calling peaks. Running MACS2, we set the genome size to $12x10^6$ bp, used a bandwidth of 100 bp, allowed for a peak fold-change between 1 and 100 and set the regions that are checked around the peak positions to calculate the maximum local lambda to 2,000 and 100,000 in order to capture the bias from a long-range effects like an open chromatin domains.

```
macs2 callpeak -t IP_merge.bam -c IN_merge.bam -f BAM -g
1.2e7 -n merge -B --bw 100 -q 0.05 -m 1 100 --slocal 2000 -
-llocal 100000
```

MACS2 outputs the normalised input and IP traces as bedgraph files and the peak and summit position as tab-delimited data that were loaded in IGB (Freese et al. 2016) for inspection. The peaks called for Abf1 binding in this output are included in e-appendix

and were used for the annotation and overlap calculations used to characterise the GCBS's described in the results section.

3.2.6 Mapping Nucleosomes derived from MNase-seq data using DANPOS

Aligned MNase-seq data, in sorted bam file format, was submitted to DANPOS (Chen et al. 2013) for mapping of nucleosome positions. Submitting multiple datasets to DANPOS allows for fold-change normalisation that compensates for different read-depth or coverage between datasets. Using the MNase-seq data from wild-type and *Rad16* deleted cells from non-irradiated (-UV) and 0 minutes and 30 minutes post-UV samples, DANPOS successfully and consistently maps ~65,000 nucleosomes for each dataset. DANPOS outputs the statistical information on position, fuzziness and occupancy in a spread-sheet format and produces a wig-file that contains the genome-wide trace of the nucleosome positions. These wig files were accessed through IGB (Freese et al. 2016) for viewing and generating snapshots as described in the Results section. This output was also uploaded to SeqPlots (Stempor and Ahringer 2016) for plotting (see following section). All composite plots described in the Results sections used the data described here.

3.2.7 Visualisation of genome-wide ChIP-seq, MNase-seq and ChIP-chip data

Genomic intervals of binding sites, NFRs, motifs or ORFs were uploaded to SeqPlots as canonical BED files. Genome-wide traces of continuous data such as those of MNase-seq and ChIP-seq data were uploaded as wig or bigWig (.bw) files. To convert ChIP-chip and 3D-DIP-chip data into a format that was amenable for plotting the output from Sandcastle (Bennett et al. 2015) was converted in the following way. Using the 'writeCCT' script from Sandcastle we exported the array data in a tab-delimited format. For compatibility, the chromosome names were converted to roman numerals using standard command line operations using Perl. This rudimentary BED file can now be converted to a wig file using the UCSC BedToWig script (http://genomewiki.ucsc.edu/index.php/File:BedToWig.sh) setting the span to the size of the probes on the array (58bp). Finally, the wig-file was manually converted to a bigWig file using the UCSC wigToBigWig script (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/wigToBigWig) using a file containing the chromosome sizes from sacCer3 in the process. Some features overlapped and had to be manually corrected and the data that was loaded into SeqPlots. All figures presented in the Results section were generated using SeqPlots, exported and processed using Adobe Photoshop CS2 (Willmore 2006).

3.2.8 Defining GG-NER complex binding sites

The Abf1 binding sites as detected by MACS2 were loaded in the R-statistical environment and annotated using the ChIPpeakAnno R-package (Zhu et al. 2010). BED files containing the coordinates of the Abf1 binding sites, genome-wide Nucleosome Free Regions (NFRs) (Yadon et al. 2010) and Abf1 consensus motifs (Bailey and Elkan 1994; Khan et al. 2017) were loaded using the BED2RangedData function. makeVennDiagram was used to generate the Venn diagram as shown in Figure 3.1. Using annotatePeakInBatch, ensembl annotation data was used to map the GCBS's to the nearest genomic features such as TSS or Gene.

The MACS2 output that detected 4,026 Abf1 binding sites from two biological ChIP-seq repeats, was used in conjunction with genome-wide NFRs (n = 6,589) (Yadon et al. 2010) and Abf1 consensus motifs (n = 1,752) to generate a three-way Venn diagram to find the genomic positions where these features overlap using ChIPpeakAnno (Zhu et al. 2010). This initial attempt to define GCBSs, did not include Rad16 binding sites identified from ChIP-chip data (n = 1,652) because the resolution of microarray data is such that the size of genomic intervals detected as peaks is at least 1 order of magnitude bigger than the features under investigation here. Therefore, multiple Abf1 binding sites, NFRs or motifs would overlap with a single Rad16 binding site, eliminating the high-resolution information obtained from the Abf1 ChIP-seq data (Figure 3.1) shows the resulting three-way Venn diagram.

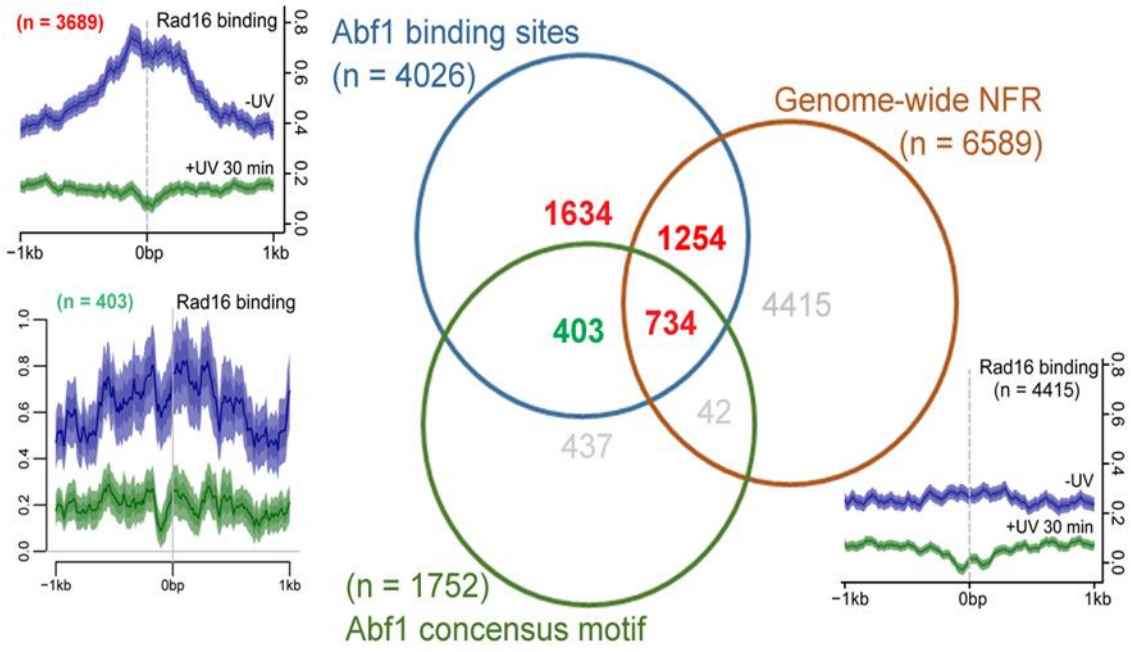


Figure 3.1: Venn diagram to classify Abf1 binding sites into categories based on the overlap between NFRs (n = 6589) and/or Abf1 consensus motifs (n = 1752) allowing no gap between the features. Inserts show Rad16 enrichment using ChIP-chip data. The GG-NER complex occupancy at the subcategories is highlighted in red.

The list of 4 subcategories of Abf1 binding sites (highlighted in red and green) were used to plot Rad16 occupancy from our previous ChIP-chip data (Yu et al. 2016). Of all classes of Abf1 binding sites, the 403 sites (highlighted in green), did not show any significant enrichment of Rad16. Therefore, this group of Abf1 binding sites that contain a motif but are not positioned at NFRs, do not qualify as GCBSs and are likely genomic positions where Abf1 executes its other functions not related to repair. The other 3 groups of Abf1 binding sites are individually (data not shown) and collectively enriched for Rad16 as shown by the insert in Figure 3.1 displaying the Rad16 occupancy at the combined 3689 positions. As a negative control we also plotted the Rad16 occupancy at the 4415 NFRs that do not contain Abf1 binding sites or motifs. These positions do not qualify as a GCBS by these criteria and we therefore find no enrichment for Rad16 at these sites as expected. This selects the first high-level set of GCBS's.

3.2.9 Annotating GG-NER complex binding sites using ChIPpeakAnno

As described in the methods sections, the R-package ChIPpeakAnno (Zhu et al. 2010) was used to annotate the set of 3,689 GCBS's selected based on the presence and/or absence of NFRs and motif sequence. These GCBS positions are enriched for the upstream and overlapStart annotation when considering their position relative to ORFs. This is in line with the previous findings from our laboratory (Yu et al. 2016). The downstream, overlapEnd, inside or includeFeature were less frequently present in this group of GCBSs. Figure 3.2 shows a pie chart of the distribution of the different orientations of GCBSs in relation to gene structure.

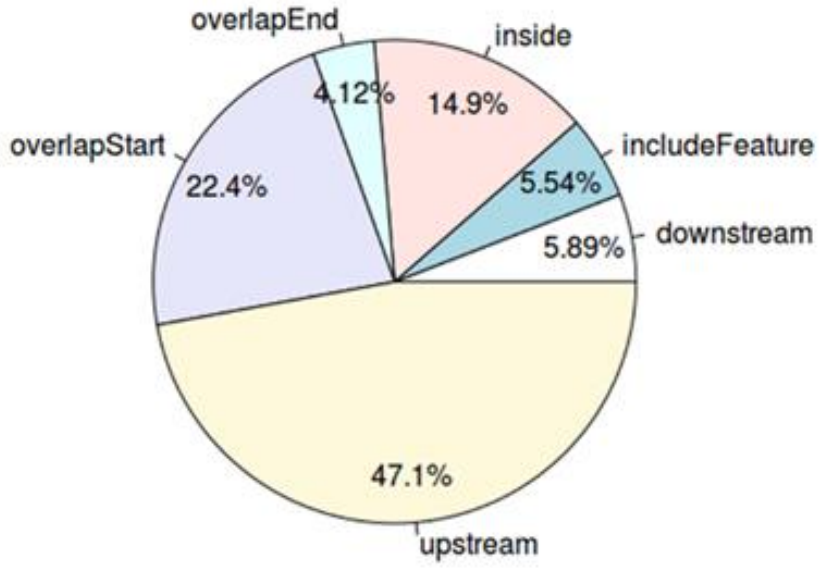


Figure 3.2: First-order GCBS's are annotated according to their location in relation to nearest gene. Categories of annotation refer to upstream of genes (overlapStart, upstream), inside of genes (inside), genes inside the binding site (includeFeature) and downstream of genes (overlapEnd, downstream).

3.2.10 K-mean clustering and heatmap plotting

Analysing the aggregate nucleosome positions around GCBS's (n = 2,664) in the context of gene structure revealed that the +1 nucleosome sits almost exactly over the TSS in this representation. This atypical conformation could be unique to these positions or be an artefact of composite plotting. In order to reveal if the orientation between the TSS, NFR and Abf1 binding sites are uniquely positioned at a certain class of GCBS's, we selected those GCBS's that overlap with an NFR. From the total list of GCBS's we obtained a subset of 1,985 sites that map to an NFR. The nucleosome structure around the remaining GCBS's will still reveal a nucleosome depleted region but are less pronounced and are not detected as NFR as a consequence (data not shown). With no NFR detected at these sites, it is impossible to generate nucleosome maps with the NFR as a frame of reference. Therefore, we selected the set of unique NFR positions and accompanying genes annotated to these positions to obtain a list of TSS and NFR positions to perform this analysis (n = 1,887). Next, we generated a heatmap of nucleosome occupancy around the GCBS/NFR positions aligned at the corresponding TSS or centred around the NFR using Seqplots (Stempor and Ahringer 2016). This data was then imported into the R statistical environment and we used the NbClust R-package to assess the K-mean clusters (Charrad et al. 2012). The intra-cluster variation was calculated by the within-cluster sum of

squares (WSS) using the wssplot function. The optimal number of clusters is reached when this parameter is minimised (Figure 3.3). The NbClust package further evaluates the optimal number of clusters for K-means clustering using another 25 criteria. Based on these findings and visual inspection of the heatmaps, we selected 13 clusters for K-means calculations aligned at TSS's and 4 clusters for the NFR centred data. This resulted in the heatmaps shown in the result section (Figure 3.7) (note that in A, 2 clusters contain only 1 trace, leaving 11 visible clusters on the heatmap).

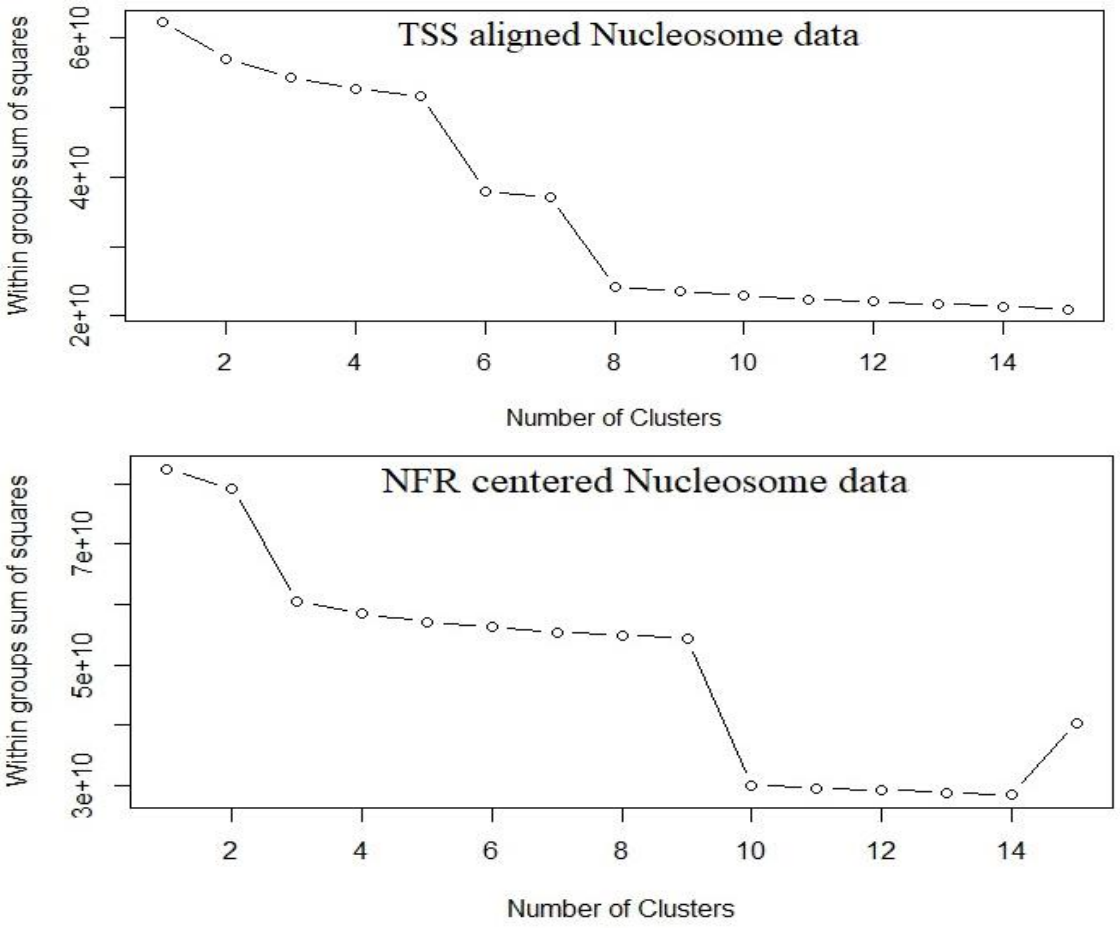


Figure 3.3: K-means cluster optimisation using NbClust. Shown here are the number of criteria (y-axis) that calculate the best-fit for the chosen number of clusters (x-axis) of the (A) TSS aligned and (B) NFR centred nucleosome data.

3.2.11 NFR detection using HOMER

The detection of NFR was guided by the HOMER documentation (found at http://homer.ucsd.edu/homer/ngs/peaks.html). First, optimisation of the nucleosome detection of this data was achieved to match the >60,000 nucleosomes detected using DANPOS (see nucleosome mapping section). Next, the nucleosomes detected by

HOMER was used to perform the NFR detection using the -nfr tag on two replicate MNase-seq datasets.

To enable nucleosome detection using HOMER the aligned data (.bam format) was converted into a tagDirectory using the makeTagDirectory function as follows:

```
makeTagDirectory /Data/Experiment -format sam
/DATA/Experiment/WTU_bwa.sorted.bam
```

This generates a TagDirectory folder in /Data/Experiment that HOMER can use for nucleosome detection. Next, the findPeaks function was used to detect nucleosomes in wildtype untreated MNase-seq data:

```
findPeaks /Data/Experiment/TagDirectory -style histone -size
180 -minDist 10 -fdr 0.22 -F 0 -L 0 -C 0 -o auto
```

The false discovery rate was fine-tuned to achieve a similar nucleosome detection compared to DANPOS, which maps between 60,000 to 70,000 nucleosomes. After this optimisation, running the algorithm can be repeated with the inclusion of the -nfr tag to detect NFRs. This analysis was applied on both MNase-seq datasets and detected 9,500 and 10,000 NFRs, respectively. Finally, we combined the bed-files and selected only those that are in common between these using bedtools (Quinlan and Hall 2010) to calculate the intersection:

```
bedtools intersect -a WT_Nucl-UV.NFR_1.bed -b WT_Nucl-
UV.NFR_2.bed > WTU_NFR_merged.bed
```

This process also flattens the bed file at positions where HOMER calls multiple NFRs at a single genomic location contains, resulting in a merged bed file that contains 5556 NFR positions. This NFR detection results in ~1,100 fewer positions being mapped compared to literature (Yadon et al. 2010). To test whether this affected the GCBS annotation performed earlier, the same three-way Venn diagram was generated as shown in Figure 3.1 with the detected NFRs and find very similar numbers of overlap. The majority of the ~1,100 NFRs that we fail to detect using HOMER, are contained within the 4,415 class of NFRs that do not map to an Abf1 binding site or consensus sequence (Figure 3.1). Moreover, this group of ~1,100 NFRs are enriched for those NFRs that reside inside ORFs that are generally difficult to detect. Taken together, the detected NFRs do not alter the GCBS classification performed earlier using literature annotated NFRs (Yadon et al. 2010).

3.2.12 Sorting GCBS's by NFR size

From Abf1 chromatin occupancy around GCBS's it appears that, occupancy is slightly assymmetric, with a shoulder to the left-side of the GCBS, upstream relative to the nearest ORF. In order to investigate whether this feature of the data represents Abf1 binding at distinct confirmations both at the -1 nucleosome and the NFR or relates to the occupancy of Abf1 at broader NFRs, the GCBS positions were sorted by NFR size and the nucleosome and Abf1 ChIP-seq data plotted as a heatmap around these positions. To do this, the NFRs that overlap with our list of GCBS's was selected. The list of Abf1 binding sites (n = 4,026) overlaps with 1,985 genome-wide NFRs (Yadon et al. 2010) in a pairwise analysis (Figure 3.1). During the overlap calculations and annotation using ChIPpeakAnno (Zhu et al. 2010), duplicate NFR or gene entries are introduced when 2 NFRs overlap with a single Abf1 binding site, or when the same gene gets annotated to 2 different Abf1 binding sites (i.e. 1 upstream and 1 downstream). Therefore, in order to retrieve a list of unique NFRs and TSSs the duplicates that this process can generate was eliminated, resulting in a list of 1766 unique genes and accompanying NFRs. The heatmap representing nucleosome occupancy at these positions was generated using SeqPlots (Stempor and Ahringer 2016) and imported this into the R statistical environment to sort the data by NFR size.

3.2.13 Micro-C boundary data processing and plotting

Datasets from the Micro-C XL experiments were retrieved from the ENA repository (Study PRJNA336566). From the double cross-linked data available the 3% FA and 3mM DSG for 40-minute dataset (SRR4000672) were used for plotting Figure 3.8. To achieve this, HiC pro was used to align the data, build the contact maps, normalise the data and QC, following the instructions of the authors (https://github.com/nservant/HiC-Pro and enclosed documentation). Next, the HiC-plotter was used to visualise the Micro-C XL data in conjunction with the MNase-seq and ChIP-seq data (Akdemir and Chin 2015).

3.3 Results

3.3.1 Identification of changes to the genome-wide linear arrangement of nucleosomes in response to UV damage

In order to determine how chromatin is remodeled in response to DNA damage, first the organisation of nucleosome structure throughout the genome in undamaged cells was investigated. The physical arrangement of nucleosomes can be thought of as a structured array of nucleosome units distributed throughout the linear genome. Within a population of cells, the precise translational setting of a nucleosome within its unit, in any given cell, may vary, centring at a favoured site, which is commonly referred to as its nucleosome position. A single nucleosome position, and its change in response to environmental conditions, can be characterised by a combination of three parameters that define it. Firstly, the nucleosome position itself, secondly, its occupancy and finally, its fuzziness, with the latter term meaning the degree of freedom that a nucleosome has, to take up its unitary position within in a population of cells. This degree of freedom is high when a fuzzy nucleosome takes up a wider range of positions in a cell population, and vice versa for low fuzziness nucleosomes. In addition to describing the position and fuzziness score of a nucleosome unit, it is also possible to measure its occupancy, which is defined by its peak height as shown in Figure 3.4A. This refers to the frequency that the nucleosome unit is occupied by nucleosomes within the population of cells. Trans-acting factors can alter nucleosome structure by changing the position and/or fuzziness of a nucleosome, as well as affecting the nucleosome occupancy at any given position in response to environmental changes. Consequently, MNase-seq technique was used to map nucleosomes and measure alterations to their structure in wildtype cells before and after exposure to UV irradiation using a bioinformatics pipeline known as DANPOS (Chen et al. 2013). This software was specifically developed for determining genomic changes in nucleosome position, fuzziness and occupancy in cells under different environmental conditions . Using this pipeline, >60,000 nucleosome positions were consistently mapped with high accuracy (see Materials and Methods for full details). In response to UV irradiation, changes to nucleosome occupancy at various positions across the genome are readily observed when the mapped nucleosome traces are plotted in a linear fashion. A representative 8 Kbp section of yeast chromosome-I is shown in Figure 3.4A (note that occupancies with, or without UV exposure of cells are indicated in grey and black, respectively). The aggregated changes in nucleosome occupancy, fuzziness and position for all >60,000 nucleosomes are summarised in Figure 3.4B to D. It has been reported

that certain types of DNA damage cause a significant loss of total nucleosomes from chromatin (Hauer et al. 2017). However, no global change in nucleosome occupancy levels was observed throughout the yeast genome in WT cells treated with UV radiation (Figure 3.4B). A small increase in the frequency of low-occupancy nucleosomes (<350 normalised reads) was noticeable immediately after UV irradiation (Figure 3.4B, red line). Genome-wide nucleosome fuzziness on the other hand, is altered to a greater extent in response to UV irradiation. For example, an increase in the frequency of fuzzy nucleosomes was detected, with a reciprocal decrease in the frequency of low-fuzziness nucleosomes both at 0 and 30 minutes after UV irradiation (Figure 3.4C). Finally, the genomic position of nucleosomes was plotted as the inter-nucleosomal distance, or nucleosome spacing, in base pairs. As expected, the average spacing for all nucleosomes, as defined by the length of linker and nucleosomal DNA, is enriched for distances of between 160 to 180 base pairs, as shown in Figure 3.4D. As a result of UV irradiation, a small loss of nucleosomes with this spacing is observed, with a reciprocal gain in more closely spaced nucleosomes also apparent (i.e. those with <140bp spacing). These observations reveal the extent of the UV-induced alteration of the linear nucleosome structure throughout the entire genome, showing that chromatin is remodeled at only a sub-set of nucleosomes, via discrete local changes in certain of them. Since UV-induced lesions are essentially distributed uniformly throughout the genome, this observations suggests that repair of damage may be initiated through nucleosome remodeling at specific sites in the linear genome in response to UV irradiation.

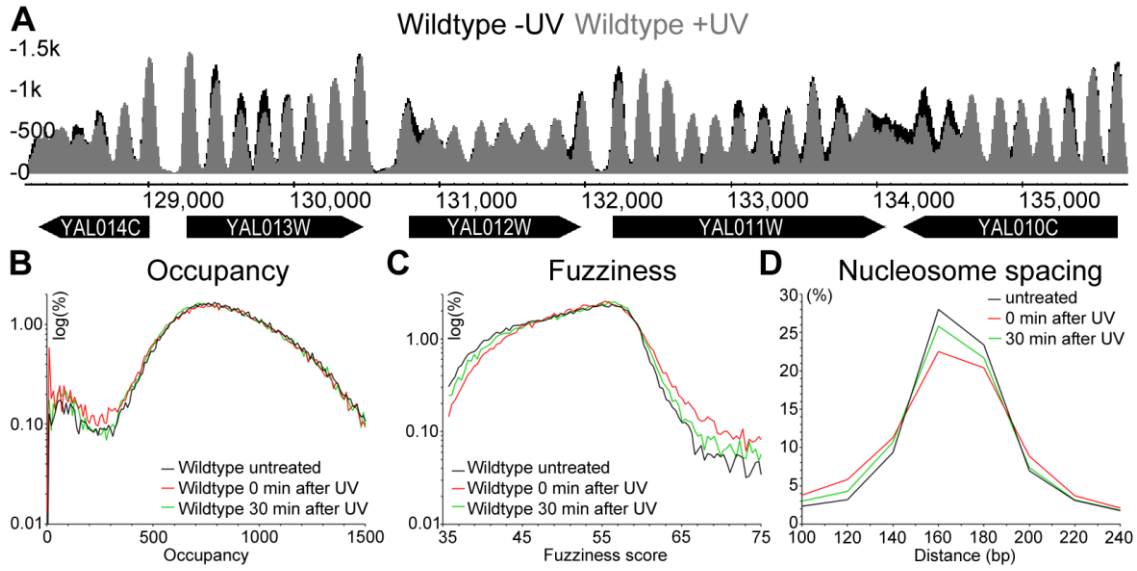


Figure 3.4: UV-induced changes to the genome-wide nucleosome landscape. A) Represented here are the nucleosome traces of wild-type cells before (black) and after

UV irradiation (grey) in an 8 Kbp region on chromosome I (128,000 to 136,000). The genes and their systematic names are indicated by the black arrows underneath the traces. The y-axis on the left indicates the relative read-counts that define the nucleosome peaks in this region. B) Genome-wide changes to wild-type nucleosome occupancy (peak height) in response to UV irradiation are quantified here. The distribution of relative occupancy (in reads) of all >60,000 nucleosomes as a log-scale of percentage is shown here. C) As B but now quantifying the degree of freedom a nucleosome has to occupy its unitary position, expressed as Fuzziness of all nucleosomes in response to UV irradiation. D) As B and C but now quantifying the change in the distribution of nucleosome spacing, reflecting the position of nucleosomes in the linear genome, expressed in base pairs for all nucleosomes after UV irradiation.

3.3.2 The canonical nucleosome structure observed at all transcription start sites is maintained after UV irradiation of cells

Figure 3.4 identified the changes in nucleosome structure, as described by three parameters, following exposure of cells to UV irradiation. Next, we wanted to determine the genomic location of these changes with respect to genomic features. Therefore, first we investigated whether the UV-induced changes in nucleosome structure described above can be seen when these events are examined and nucleosomes plotted at all transcription start sites (TSS's) (Xu et al. 2009). After examining the nucleosome structure in relation to all 5,171 TSS's, the structure around these genomic features remained unaltered after UV irradiation (Figure 3.5A). This result demonstrates that no gross UV-induced changes to the nucleosome landscape occur in the context of this well-established genomic feature of nucleosome organisation. Disruption of nucleosome structure at this feature has been previously described for mutants defective in certain essential SWI/SNF ATP-dependent chromatin remodelers, including CHD1, ISW1 or INO80 (van Bakel et al. 2013), which regulate gene expression by controlling nucleosome structure and occupancy at these sites. Indeed, in yeast, nucleosome sliding, which shifts the translational setting of the nucleosome, altering its position, is a well-known mechanism to control gene expression (van Bakel et al. 2013). Part of the cellular response to DNA damage controls the gene expression of various DNA-damage responsive genes via this mechanism. Therefore, nucleosome sliding at a single UV-responsive gene locus was investigated, by plotting nucleosomes at the DNA damage inducible locus, RAD51 (Shinohara et al. 1992), as shown in Figure 3.5B. As expected, these data demonstrate that nucleosome sliding can be detected at this locus after DNA

damage induction. The absence of linear nucleosome sliding when all TSSs are examined in aggregate, however, indicates that this mechanism of nucleosome remodeling does not occur globally throughout the genome in response to UV damage.

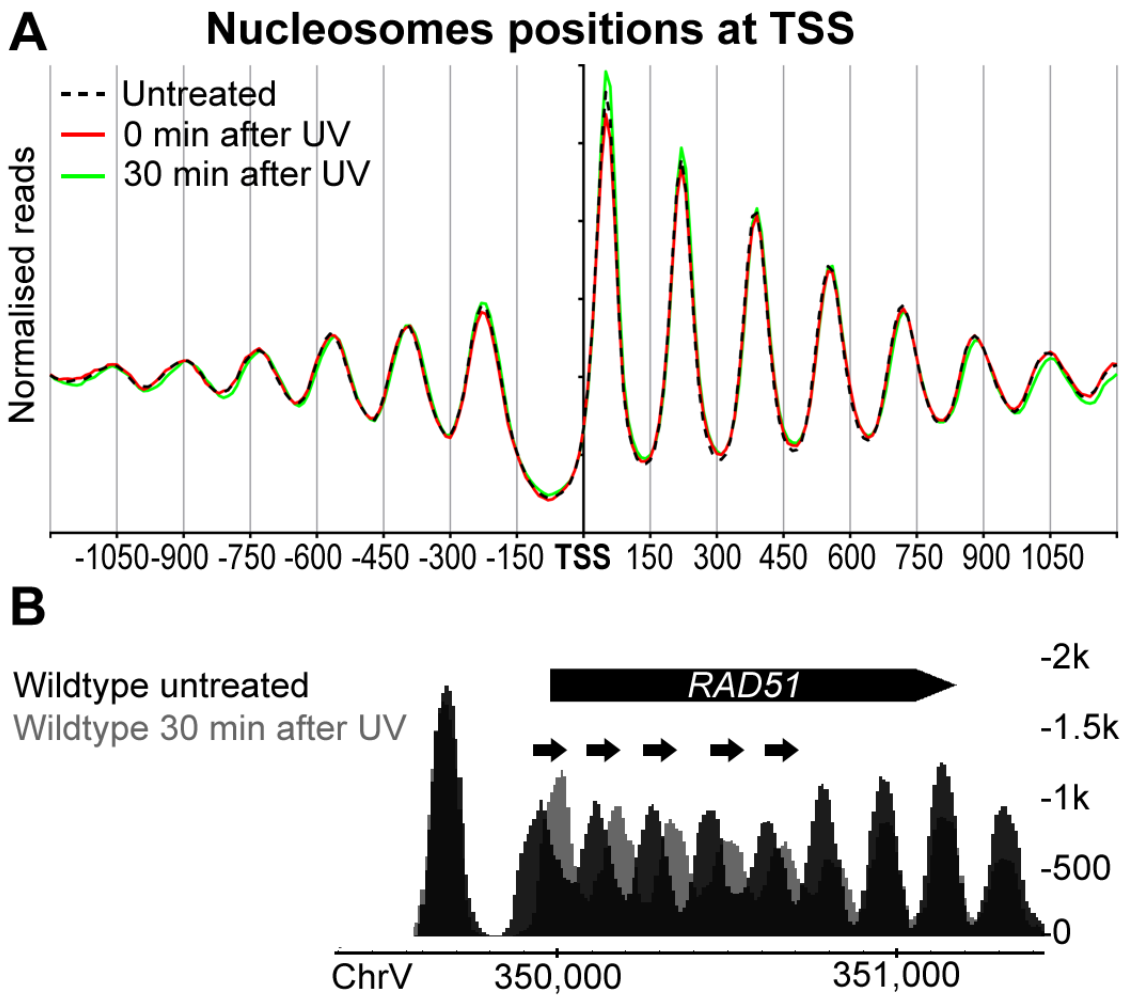


Figure 3.5: Nucleosome occupancy around all TSS in yeast does not change in response to UV irradiation. A) Composite plot of nucleosomes positions relative to all TSS (n = 5,171). Genome-wide MNase-seq data was used to aggregate nucleosome positioning in relation to TSS positions in wild-type cells before and after UV irradiation. B) UV-induced changes in nucleosome positions at the DNA damage inducible gene *RAD51*, shows sliding of nucleosomes after irradiation.

3.3.3 UV-induced nucleosome remodeling occurs at nucleosomes positioned immediately adjacent to GG-NER complex binding sites

Figure 3.4 revealed the genome-wide changes in nucleosome structure following exposure of cells to UV. However, these changes are likely due to the collective effect of a variety of mechanisms in addition to that of DNA repair by GG-NER. For example, UV-induced changes to gene expression as part of the DNA damage response are also likely to cause alterations to nucleosome structure. Therefore, I reasoned that UV-induced, GG-NER complex-dependent nucleosome remodeling might occur in relation to a novel class of genomic feature defined by the locations of GG-NER complex binding throughout the geneome. Recently, it has been reported that GG-NER is organised into domains around ORF structure, thus promoting efficient repair (Yu et al. 2016). This report showed that loss of GG-NER function severely affected repair rates around the promoter regions of genes containing Abf1 binding sites. Abf1 is a component of the GG-NER complex and is a known general regulatory factor (GRF) in yeast. Abf1 ChIP-seq experiments were performed to map the genome-wide occupancy of Abf1 in chromatin at nucleotide resolution as a precursor to identifying the precise location of GG-NER complex binding. In agreement with previously published Abf1 ChIP-chip data, and other reports (Zentner et al. 2015; Yu et al. 2016), ~4,000 Abf1 binding sites were detected by MACS2 (Zhang et al. 2008) (methods and e-appendix). It is well established that the GRF Abf1 exists in excess over the other GG-NER components (Rad7 and Rad16) in the cell and has a wide range of different functions outside of GG-NER (Yarragudi et al. 2007; Schlecht et al. 2008; Ganapathi et al. 2010; Zhang et al. 2012). Therefore, in order to examine nucleosome structure in relation to GG-NER function, a refined list of ~4000 Abf1 binding sites was established to identify a novel set of GG-NER complex binding sites (e-appendix). To do this, distinctive genomic features associated with Abf1 binding were examined, including Abf1 consensus sequences (n = 1,752), and genome-wide NFRs (n = 6,589), with which Abf1 occupancy is frequently associated as described in Figure 3.6A and Figure 3.1 (Yarragudi et al. 2007; Hartley and Madhani 2009; Ganapathi et al. 2010; Ozonov and van Nimwegen 2013). This enabled me to categorise Abf1 binding sites according to these features. To subclassify the Abf1 sites associated with GG-NER, previously published Rad16 ChIP-chip genome-wide occupancy data (Yu et al. 2016) was used to identify the set of Abf1 sites enriched for GG-NER complex binding. This yielded ~3,600 Abf1 binding sites that are enriched for Rad16 (Figure 3.6A). The majority of these genomic positions (~70%) are located in promoter regions upstream of genes making it possible to examine the surrounding nucleosome structure

(Figure 3.6A, Figure 3.2). This list is used to define a novel class of genomic features, which we now refer to as GG-NER complex binding sites (GCBS's, n = 2664). Full details of the analysis can be found in the Materials and Methods section. Annotating the data in this way now enables us to map nucleosomes directly at GCBS's, or at GCBS's in relation to ORF structure. The strand information of the nearest annotated gene orientates the data in such a way that the ORFs are positioned downstream (i.e. to the right) of the GCBS's.

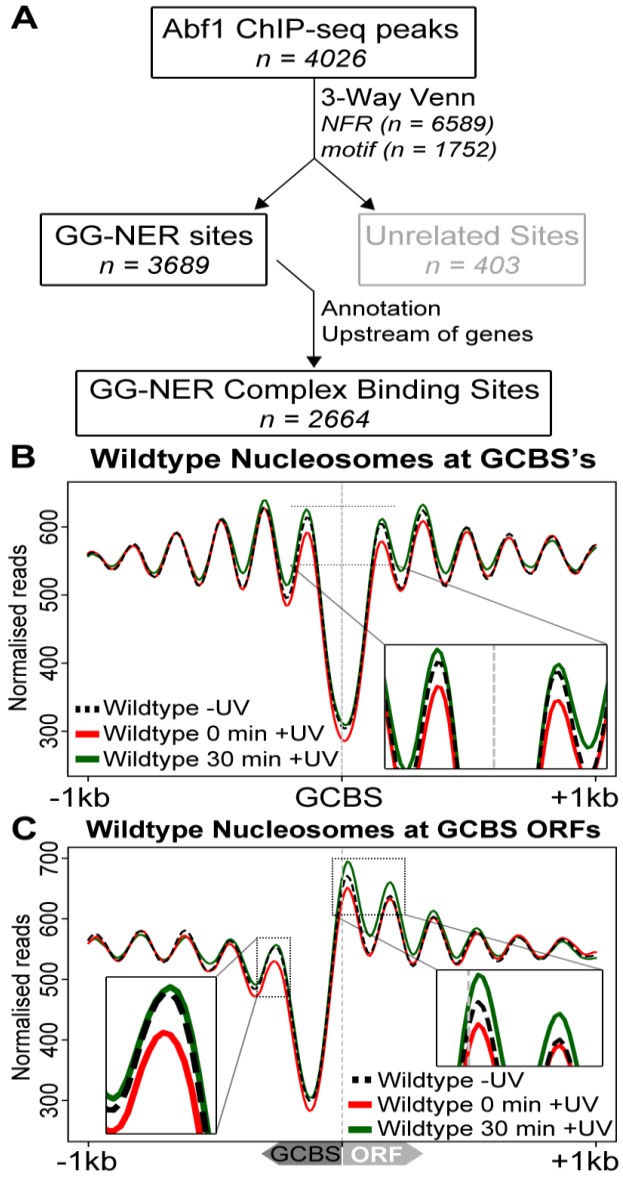


Figure 3.6: Identification of the genomic list of GCBS's and the nucleosome occupancy in relation to these sites. A) Flow chart to illustrate the bioinformatics analysis performed to identify genome-wide GCBS's by refining and filtering the list of Abf1 ChIP-seq peaks using NFR positions, motif sites and annotation information. B) MNase-seq data of wild-type cells was used to plot cumulative nucleosome positions around GCBS's (n = 2,664)

in the absence of UV irradiation and at different intervals after UV irradiation, displaying regularly spaced nucleosome arrays at these genomic locations. The x-axis denotes the 2 Kbp regions surrounding the GCBS's while the y-axis indicates nucleosome occupancy as measured by normalised reads. C) Nucleosome occupancy in wild-type cells before and after UV damage. MNase-seq data of untreated and UV-treated cells is shown as cumulative graphs around GCBS's in relation to ORF structure. The inserts highlight the nucleosome remodeling at the -1 position (on the left) and the remodeling at positions +1 and +2 (on the right).

Using this list, the composite plots of nucleosome positions directly at GCBS's (Figure 3.6B) and at GCBS-adjacent ORFs (Figure 3.6C) were examined. This reveals the position of an NFR at these locations, flanked by an array of positioned nucleosomes as others have reported previously for Abf1 binding sites (Lai and Pugh 2017). In Figure 3.6C, it is evident that this class of GCBS-adjacent nucleosomes is located directly over the position of the TSS. Typically, the +1 nucleosome is positioned further into the ORF when nucleosomes are mapped to all TSSs (Figure 3.5A). To determine whether this novel subset of TSS-positioned nucleosomes is unique to this class of GG-NER complex binding sites, K-means clustering of the individual nucleosome traces was performed to identify a subclass of genomic positions that uniquely contain a +1 nucleosome at the TSS, or whether this is a common feature amongst these genes. Interestingly, this analysis finds 13 clusters that all display a different distance between the TSS and the +1 nucleosome (Figure 3.7A). To represent the nucleosome structure independently of GCBS position, the centre of the coordinates of the NFR at these GCBS's was used. Plotting the MNase-seq data in this orientation uniformly aligns the nucleosome arrays, revealing the presence of only four clusters Figure 3.7B). Therefore, the +1 nucleosome position at the TSS as shown in Figure 3.6C is explained by the averaging of different nucleosome traces that exhibit a highly variable distance between the TSS and the +1-nucleosome position. This demonstrates that this is not a typical feature of these GCBS-associated regions, but simply reflects the variable distance between TSS and NFR in this class.

Next, we examined the effect of UV irradiation on nucleosome structure at these positions. In wild-type cells, loss of nucleosome occupancy at GCBS-adjacent +1 and -1 nucleosomes can be discerned immediately after UV irradiation (Figure 3.6B & C, red line). Following 30 minutes of repair time, nucleosome occupancy is restored to pre-damage levels (Figure 3.6B, green line), with evidence of increased nucleosome

occupancy at the +1 and +2 positions (Figure 3.6C, green line). These experiments reveal the precise genomic location of UV-induced remodeled nucleosomes in relation to GCBS's.

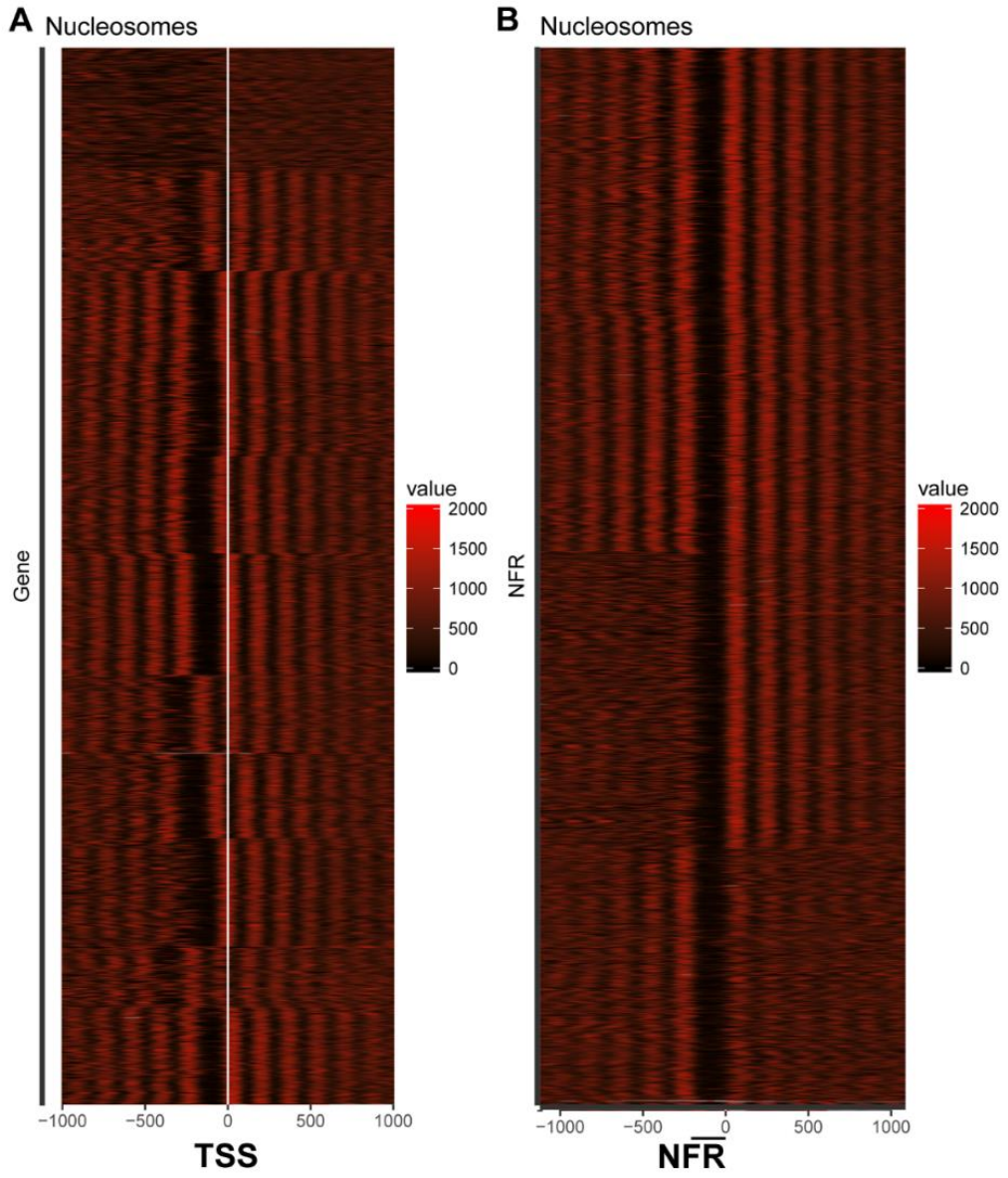


Figure 3.7: The relationship of TSS to NFR distance and the +1 nucleosome at genes downstream of GCBS's is displayed here. Heatmaps were generated using SeqPlots and transferred to the statistical R-environment for further analysis. A) Using K-means clustering analysis, the traces were grouped into 13 clusters of similar structure. The heatmap displays the relative nucleosome density around the TSS, highlighted by the white line in the middle of the figure. This includes nucleosomes 1 Kbp up- and downstream. The intensity of the heatmap is proportional to the normalised read-depth indicated in the figure. B) The same nucleosome data as displayed in A) was used, but now aligned at the accompanying NFR. K-means clustering of the data in this orientation

resulted in the identification of 4 clusters. The intensity of the heatmap as a function of normalised reads is indicated in the figure. The bar at the bottom indicates the orientation of the NFR at these genomic positions.

3.3.4 GG-NER complex binding sites are located at the boundary regions of specific Chromosomally Interacting Domains.

Figure 3.6 revealed UV-induced nucleosome remodeling in only a small subset of GCBS-adjacent nucleosomes. Since GG-NER operates throughout the genome, it should be considered how so few localized changes in nucleosome structure might contribute to chromatin remodeling in a wider context throughout the genome. The standard MNase-seq method only reveals changes in the linear arrangement of nucleosomes (Figure 3.4A, MNase-seq technique). Therefore, investigation of the genomic locations of GCBS's in relation to domains of higher-order chromatin structure is necessary to determine how these events are organised in relation to higher order chromatin structure. Recent advances in methods such as 3C and the related HiC, have led to the introduction of a chromatin capture method called Micro-C (Hsieh et al. 2015; Hsieh et al. 2016). This technique measures higher-order nucleosome-nucleosome interactions in chromatin. Micro-C follows the same principles as other 3C methods but uses MNase instead of restriction enzymes to digest cross-linked chromatin. This allows the detection of distal nucleosome-nucleosome interactions that have recently led to the discovery of chromosomally interacting domains (CIDs, n = 3061) at nucleosome resolution for the first time in yeast (Hsieh et al. 2015). These authors reported that boundary sites that demarcate CIDs are often found upstream of highly expressed genes and are enriched for nucleosome pairs that flank NFRs, in a similar fashion to the features that were observed in relation to GCBS's. Therefore, the relationship between the genomic locations of the GCBS's identified above, and the boundary sites of these newly described CIDs was examined. To this end, the genomic positions of the CID boundaries was retrieved from published data (Hsieh et al. 2015), and the overlap between these positions and the GCBS's was calculated. Remarkably, the GCBS's mapped predominantly to CID boundary positions, with around 50% being located *precisely* at these sites, as shown in Figure 3.8A. To examine the significance of this observation, a similar number of random genomic positions was taken, and overlap was calculated for the boundaries and GCBS's with these randomly chosen sites. This revealed only two boundaries overlapping at these positions compared with over 1,200 GCBS's found at CID boundaries (Figure 3.8A). Figure 3.8B shows a representation of the newly discovered CID chromatin landscape in

relation to the linear setting of nucleosomes (Figure 3.8D), and also shows the position of two GCBS's at CID boundaries, exemplified by the binding of Abf1 (Figure 3.8C). This confirms that GCBS's colocalise precisely at the boundary positions of a specific sub-set of CID boundaries, and therefore occupy sites in the genome that demarcate regions of higher-order chromatin structure. This suggests that GG-NER may be organised and initiated from these specific CID boundary regions, to which the GG-NER complex is bound in the absence of DNA damage.

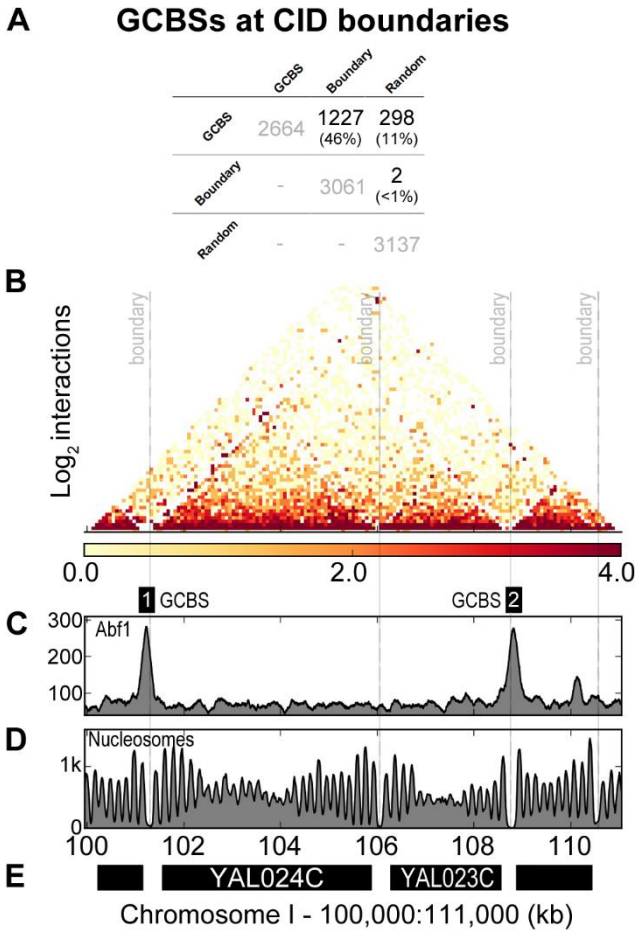


Figure 3.8 – GCBS's are located at the boundaries of Chromosomally Interacting Domains. A) Overlap calculations identified the number and identity of GCBS's (n = 2,664) at CID boundaries (n = 3,061) and at random sites (n = 3,137). The percentage of GCBS's in each subcategory is indicated between brackets. B) Micro-C data (Hsieh et al. 2016) was used to plot nucleosome-nucleosome interactions in a 11 Kbp window on chromosome I. The grey dashed lines indicate 4 boundary positions documented in the literature (Hsieh et al. 2015). The intensity of the heatmap is a measure for the normalised interactions indicated beneath the panel. C) Abf1 ChIP-seq data is plotted here to highlight two GCBS's in this region of the genome labelled as GCBS 1 and 2. D) The

nucleosome landscape is presented here by plotting MNase-seq data at this genomic location. E) Indicated in black bars are the genes located within this region of the genome. The labels on the x-axis highlight the genomic coordinates in Kbp. The y-axis on each panel indicates peak height as normalised reads.

3.3.5 GCBS-adjacent nucleosome remodeling in response to UV damage is dependent on the GG-NER complex

So far, the results reveal the genomic location of UV-induced nucleosome remodeling in relation to GCBS's in wild-type cells (Figure 3.6B & C). In order to determine whether this remodeling is dependent on the GG-NER complex, similar experiments were conducted in GG-NER defective, RAD16 deleted cells. To do this, nucleosomes were first mapped in untreated *rad16* mutants and compared them to the wild-type pattern (Figure 3.9A, grey line, Figure 3.10A). A reduced nucleosome occupancy was observed at the positions immediately adjacent to the GCBS's in these GG-NER defective cells. This demonstrates that the GG-NER complex is necessary for establishing the normal nucleosome structure adjacent to these locations in undamaged wildtype cells (Compare black line with grey line, Figure 3.9A & Figure 3.10A).

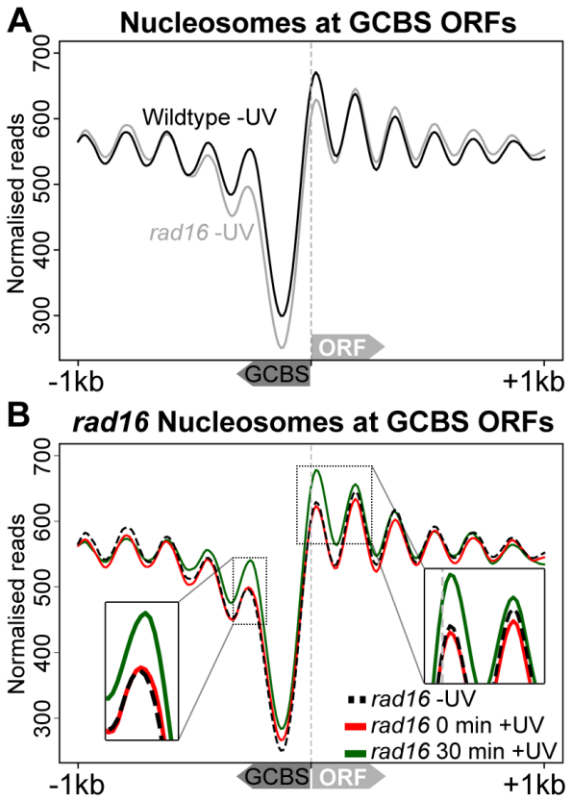


Figure 3.9: GG-NER complex adjacent nucleosomes are established and remodeled following UV irradiation in a Rad16-dependent fashion. A) MNase-seq data of wild-type and *rad16* mutant cells was used to plot cumulative nucleosome positions around GCBS's

(n = 2,664) in the absence of UV irradiation. The annotation of the nearest gene was used to infer strand information to align these genomic positions according to gene orientation as indicated by the arrows on the x-axis depicting the relative direction the GCBS and ORF. The x-axis denotes 2 Kbp regions surrounding the GCBS's, while the y-axis indicates nucleosome occupancy as measured by normalised reads. B) As described in A but showing UV-induced changes to nucleosome positions around GCBS's and accompanying ORFs in *rad16* mutated, GG-NER defective cells. Next, UV treated RAD16 deleted cells showed no loss of nucleosome occupancy at these positions (Figure 3.9B, Figure 3.10B). This shows that the nucleosome remodeling observed at these sites in wild-type cells (Figure 3.6B & C) is dependent on the GG-NER complex.

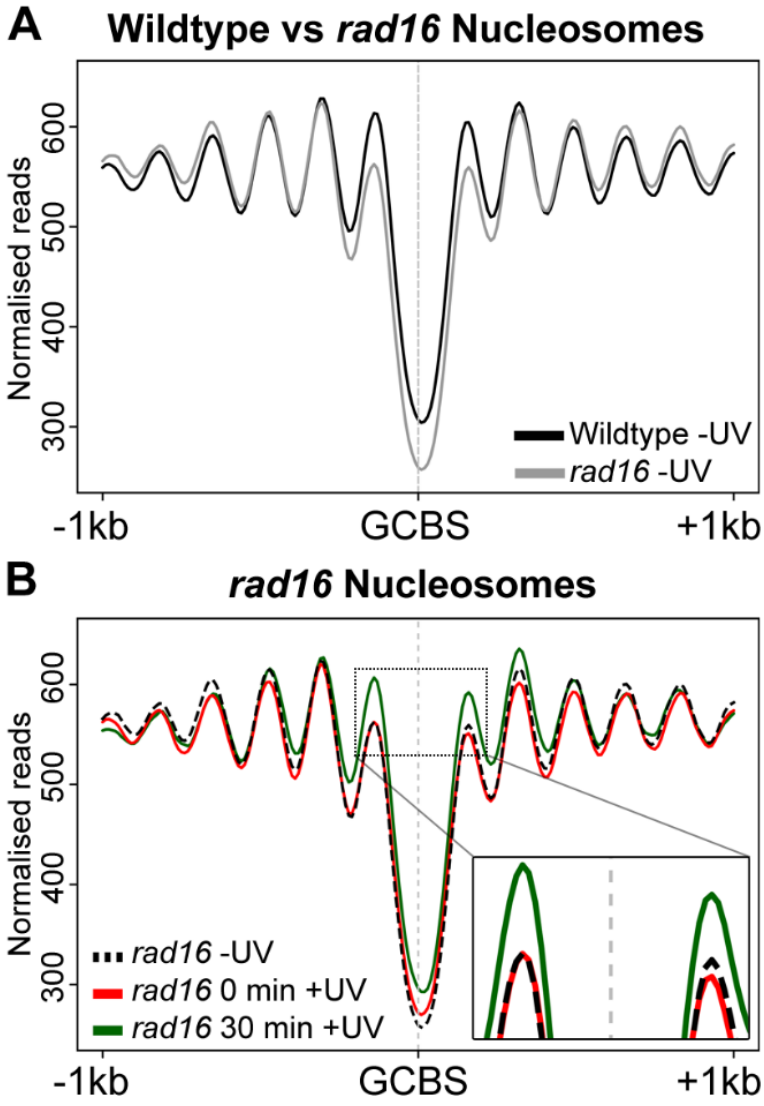


Figure 3.10: Nucleosome occupancy and UV-induced chromatin remodeling at GCBS's is GG-NER dependent. A) Composite plots centred at GCBS positions are shown here depicting the nucleosome data from both wild-type and *RAD16* deleted GG-NER

defective cells. B) As described in A, plotting MNase-seq data from *RAD16* deleted GG-NER defective cells both before and after UV irradiation.

To confirm that GG-NER dependent remodeling of nucleosomes is specific to GCBS sites, the nucleosomes at all TSSs in the RAD16 deleted strain was analysed and no UV-induced changes at these sites were observed (Figure 3.11). In fact, nucleosomes were found to accumulate at these sites 30 minutes after UV irradiation in GG-NER defective cells (Figure 1.9B, green line, Figure 3.10B, green line). Collectively, these results indicate that the nucleosome remodeling process observed at the GCBS-adjacent nucleosomes is a process that initiates the chromatin remodeling required for GG-NER (Weber et al. 2014).

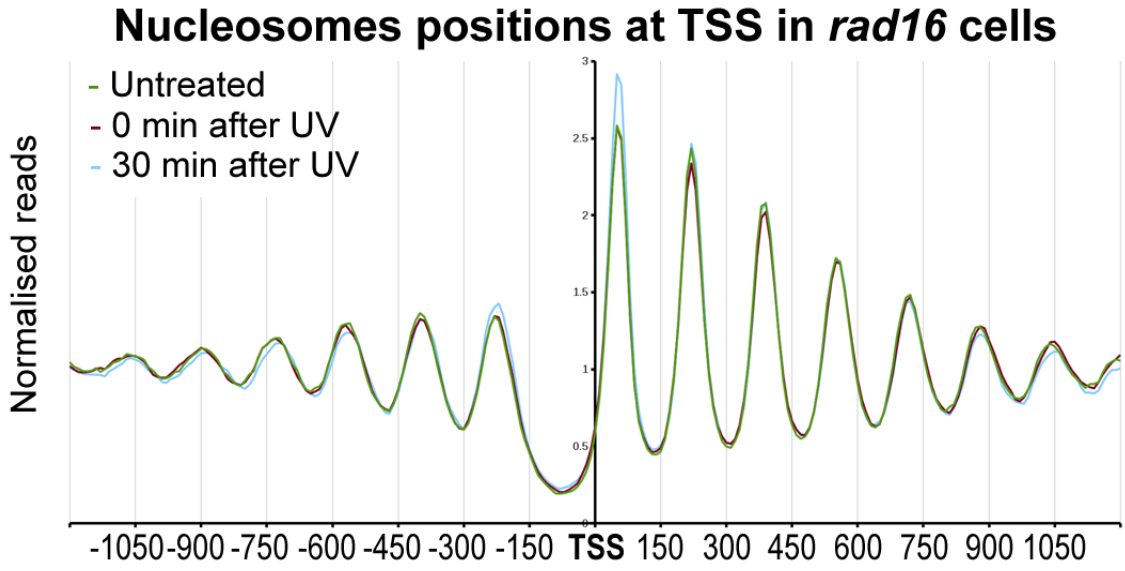


Figure 3.11: Nucleosome occupancy around all TSS in RAD16 deleted yeast does not change in response to UV irradiation. Composite plot of nucleosomes positions relative to all TSS (n = 5,171). Genome-wide MNase-seq data was used to aggregate nucleosome positioning in relation to TSS positions in wild-type cells before and after UV irradiation.

3.3.6 GCBS-adjacent nucleosomes are histone H2A.Z-containing barrier structures that are remodeled by the GG-NER complex in response to UV damage

The UV-induced loss of nucleosome occupancy observed in wild-type cells is consistent with histone exchange events that occur at dynamic nucleosomes, as described by others in the context of gene transcription (van Bakel et al. 2013). Dynamic nucleosomes are often associated with the functional response of the cell to environmental change or stress (Lai and Pugh 2017). This physical organisation of the chromatin controls the accessibility of binding proteins to the DNA in chromatin, such as transcription factors,

thus regulating their activity in the cell. Such nucleosomes contain the histone variant H2A.Z and have been described previously as 'barrier nucleosomes' that signifies their highly dynamic nature (Weber et al. 2014). As such, they represent nodes in the genome; gate-like structures that must be modified to permit proper functioning of events occurring at such locations. A role for histone variants in DNA repair has been noted in both NER and other repair mechanisms (Adam et al. 2015). Indeed, in yeast, our laboratory previously reported that histone H2A.Z is involved in NER (Yu et al. 2013). Therefore, I examined the occupancy of histone H2A.Z at nucleosomes adjacent to GCBS's. To study this, ChIP-seq experiments using HA-tagged H2A.Z were performed to map the positions of genome-wide H2A.Z-containing nucleosomes. After measuring the change in their occupancy in response to UV irradiation ~16,000 H2A.Z containing nucleosomes were detected. Initial analysis of the genome-wide distribution of histone H2A.Z confirmed the presence of this histone variant predominantly at nucleosomes flanking NFRs upstream of genes. These display an asymmetric pattern of binding as described previously in the literature (Guillemette et al. 2005; Raisner et al. 2005; Albert et al. 2007; Weber et al. 2014). After examining the H2A.Z occupancy in GCBS-adjacent nucleosomes, the presence of H2A.Z containing nucleosomes was observed at both the +1 and -1 positions (Figure 3.12A). Compared with the loss of overall nucleosome occupancy described earlier (Figure 3.6C), in the case of H2A.Z containing histones, occupancy is lost uniquely from the +1 nucleosome position in response to UV irradiation (Figure 3.12A, red line). In addition, after 60 minutes of repair time, H2A.Z occupancy returns to its pre-damage level (Figure 3.12A, green line), which is consistent with the wild-type recovery of nucleosome occupancy shown in Figure 3.6C. These differences may reflect variations in the type and timing of the histone eviction/exchange events occurring in the +1 and -1 nucleosomes during repair. Collectively, these data demonstrate that nucleosome remodeling occurs at promoter NFRs adjacent to GG-NER complex binding sites. In response to UV irradiation, histone eviction or exchange occurs at these nucleosomes.

In order to test the GG-NER complex-dependence of H2A.Z loss at these sites, the H2A.Z ChIP-seq experiment was repeated in cells deleted for RAD16. It was observed that, in the absence of DNA damage, histone H2A.Z occupies GCBS-adjacent nucleosomes in *rad16* mutant cells in a similar fashion to that observed in wildtype cells (compare Figure 3.12A black dashed line with 9B black dashed line, see also Figure 3.13). However, in response to UV, no loss of histone H2A.Z occupancy from the +1 nucleosome can be

detected in the GG-NER defective RAD16 deleted cells (Figure 3.12B, red line, Figure 3.13B, red line).

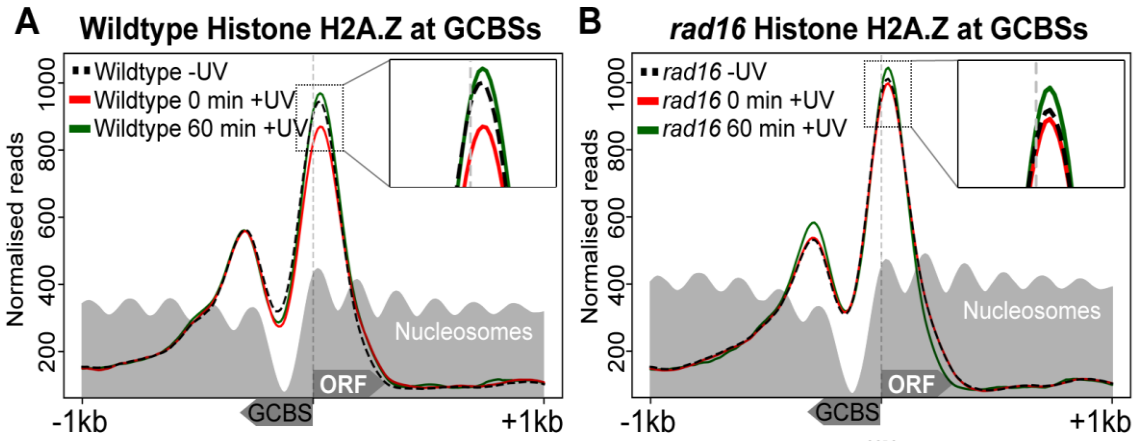


Figure 3.12: UV-induced loss of H2A.Z occupancy requires GG-NER complex-dependent nucleosome remodeling around GCBS's. A) The UV-induced change to H2A.Z occupancy in wild-type cells around GCBS-associated TSS's is shown here using H2A.Z ChIP-seq data, prior to UV irradiation and 0 or 60 minutes after UV damage. The light grey trace represents the nucleosome positioning in the absence of DNA damage retrieved from the data shown in Figure 3.9. The insert highlights the UV-induced changes to H2A.Z occupancy at the +1 position. B) As described in A, but now representing the H2A.Z occupancy at GCBS-bound promoter regions in GG-NER defective RAD16 deleted cells.

Consistent with the findings described for nucleosome occupancy observed in wild-type cells (Figure 3.6B & C), 60 minutes after UV irradiation, H2A.Z occupancy also accumulates to levels higher than those observed prior to UV irradiation in wild-type cells (Figure 3.12B, green line, Figure 3.13B, green line). Absence of H2A.Z loss from these sites in a GG-NER defective mutant, confirms a role for the GG-NER complex in this process. These data demonstrate that histone loss at H2A.Z-containing nucleosomes, adjacent to GCBS's, is driven by the GG-NER complex to alter chromatin structure during the initial stages of GG-NER in response to UV damage.

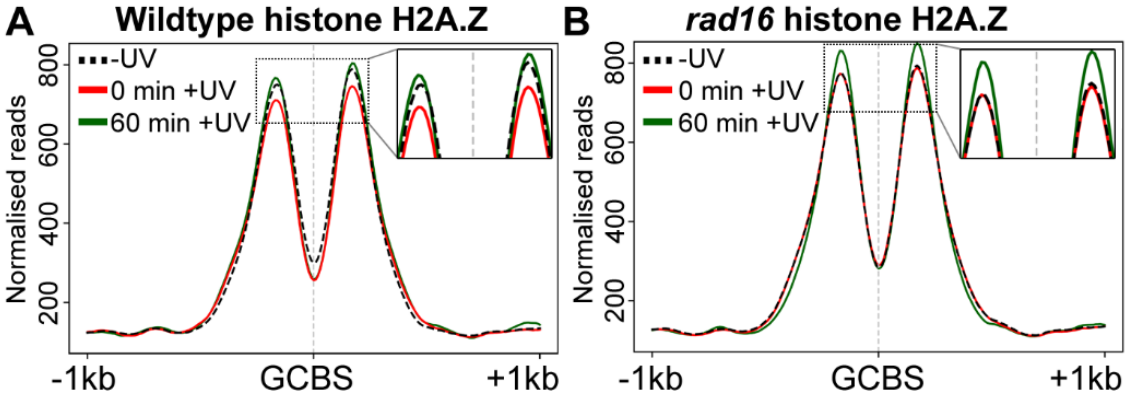


Figure 3.13: GCBS-adjacent nucleosomes contain histone H2A.Z that are remodeled in response to UV irradiation. A) H2A.Z occupies nucleosomes adjacent to GCBS's (n = 2664), plotted here from ChIP-seq data measuring H2A.Z before and 0 or 60 minutes after UV irradiation. The x-axis represents the GCBS's and 1 Kbp either side of these positions. H2A.Z occupancy is quantified as normalised ChIP-seq reads on the y-axis. B) As described in A, but plotting H2A.Z nucleosomes in *RAD16* deleted cells in response to UV irradiation.

3.3.7 Chromatin remodeling during GG-NER is initiated from GCBS's that define origins of repair within the genome

Finally, to determine the significance of the nucleosome remodelling mechanism described above to the repair of UV damage, genome-wide DNA repair rates were examined in relation to this novel class of genomic features. To this end, the relative rates of DNA repair described previously (Yu et al. 2016) were plotted in relation to GCBS's in order to establish how the function of the GG-NER complex promotes the efficient repair of UV-induced DNA damage. It is also possible, to map relative rates of DNA repair in relation to nucleosome positions at GCBS's. In wild-type cells, relative repair rates are high across a 2 Kbp window surrounding GCBS's (Figure 3.14A, and Figure 3.15C), with variation in the rates observed in relation to nucleosome positions, a phenomenon recently reported by others (Mao et al. 2016). However, in the absence of the GG-NER complex, relative CPD repair rates in the vicinity of GCBS's are severely reduced, as shown in RAD16 deleted cells (Figure 3.14A, grey line; Figure 3.15C).

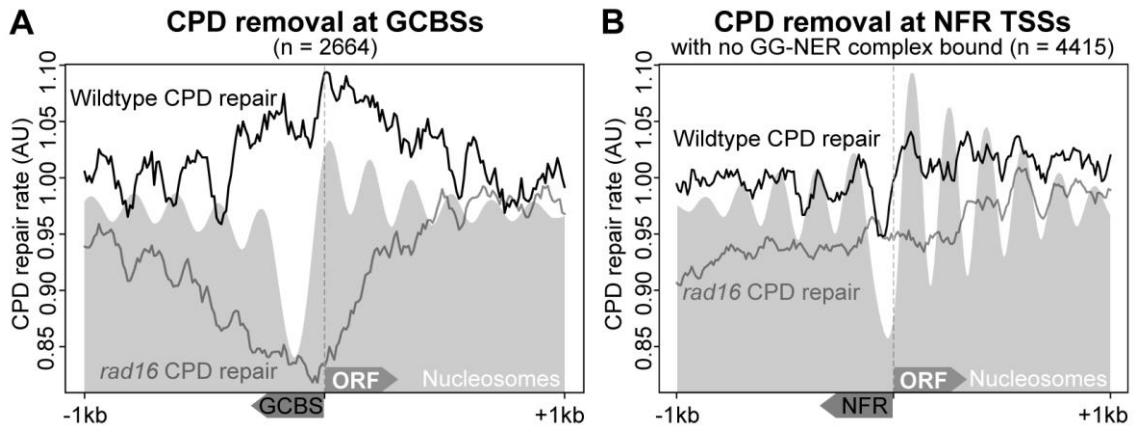


Figure 3.14: GG-NER complex binding at a subset NFRs organises repair in chromatin but this is not a general feature of NFRs. A) Relative CPD repair rates are plotted around GCBS's for both wild-type and *rad16* GG-NER defective mutant cells in relation to the nucleosome landscape, indicated as the grey shaded area. The x-axis indicates the orientation of both the GCBS and the ORF in relation to TSSs. B) As described in A, but here plotting the relative repair rates at non-GCBS-associated NFRs (n = 4,415, see Figure 3.1), orienting the data in relation to the nearest gene aligning at the TSS with the NFR positioned upstream. The x-axis indicates regions 1 Kbp up- and downstream from these positions. The grey shaded area represents the nucleosome data at these positions.

The data demonstrates that although relative rates of repair are most severely affected at GCBS's in GG-NER defective mutants, the effect on repair extends well beyond the location of remodelled nucleosomes that are immediately adjacent to the GCBS's. This suggests that the role of the GG-NER complex in chromatin remodeling likely extends beyond the local alteration of GCBS-adjacent nucleosome structure in the linear genome. It is conceivable that this may involve the disruption of higher-order nucleosome interactions of the type that comprise the CIDs, the boundaries of which are frequently occupied by the GG-NER complex as described earlier (Figure 3.8). For comparison, an in-silico control containing a set of genomic NFRs, that are *not* associated with GG-NER complex binding were also examined. As expected, no GG-NER-dependent nucleosome remodeling was observed at these sites (see Figure 3.1, n = 4415). (Figure 3.15A & B). Importantly, the relative rates of GG-NER at these sites in RAD16 deleted cells are not affected in the same way as those observed at GCBS's. The relative repair rate at these NFRs is reduced, but similarly distributed in both RAD16 deleted cells and wild-type cells (Figure 3.15 D). These observations confirm that the mechanism of repair organised and initiated from GCBS's, is not simply a common feature of *all* NFRs, but rather is specifically dependent on the occupancy and function of the GG-NER complex being

present at these sites. In conclusion, these studies show that GCBS's are novel genomic features that represent sites from which GG-NER is initiated following the remodeling of adjacent nucleosomes at these locations in response to UV damage.

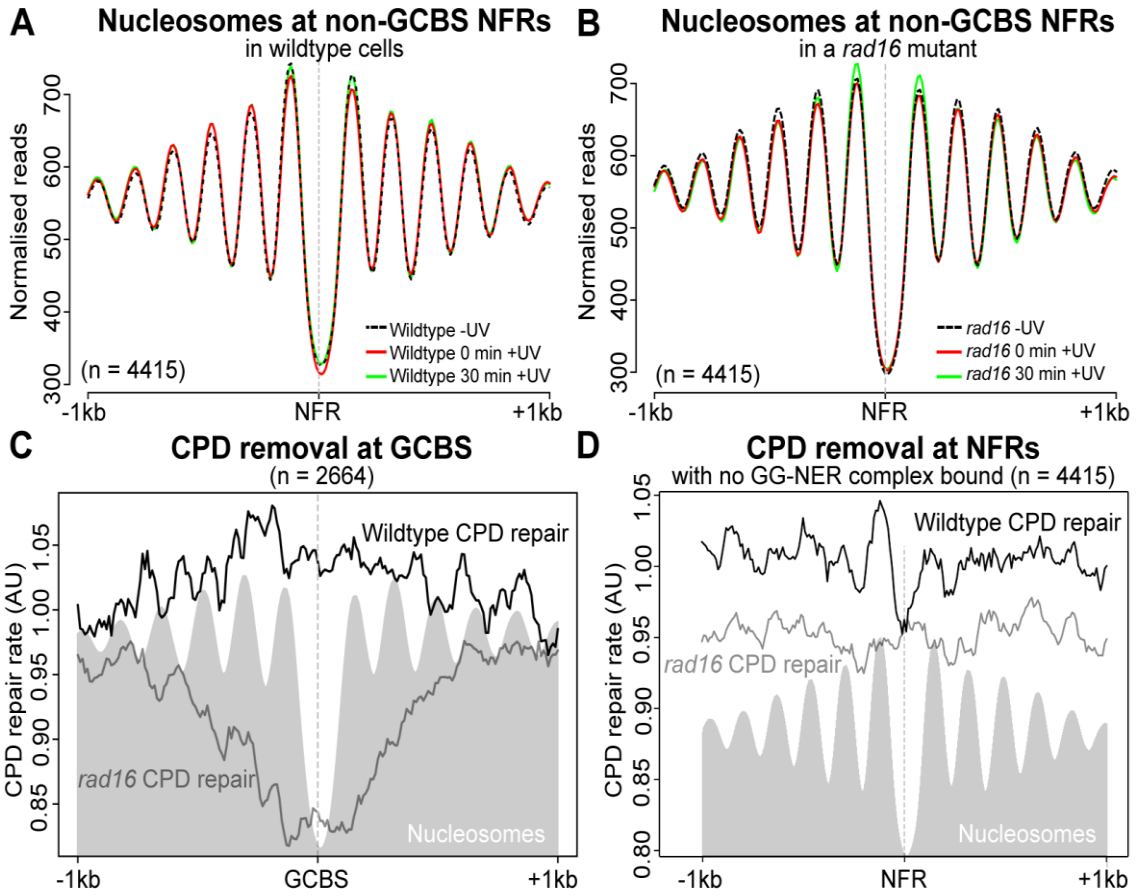


Figure 3.15: UV-induced nucleosome remodeling and repair require chromatin binding of the GG-NER complex and are not common features of NFRs. The NFR positions (n = 4415) that do *not* overlap with an Abf1 binding sites or Abf1 consensus motif (Figure 3.1), were used to plot the nucleosome data as composite plots in (A) wild-type and (B) *rad16* deleted GG-NER defective cells. On the x-axis the NFR position and 1 Kbp regions up- and downstream are displayed, while the y-axis depicts the normalised reads to indicate the nucleosome occupancy at these genomic locations. C) Relative repair rates from wild-type and GG-NER deficient cells were plotted at GCBS's to show the effect on repair as described previously. The nucleosome landscape is presented as the grey shaded area, showing the repair rates in the context of nucleosome positions. The GCBS positions, including 1 Kbp on either side are presented here. The CPD repair rates are expressed as arbitrary units. D) As described in C, but now using a set of genome-wide NFRs (n = 4415) to which *no* GG-NER complex is bound. Nucleosome and repair data of wild-type and *RAD16* deleted cells was plotted at these genomic features.

3.4 Summary

In cells, maintaining the integrity of the genome is essential for life. Since DNA is constantly exposed to the deleterious effects of both the internal and external cellular environment, mechanisms have evolved to sense and repair the consequent genetic damage. The ability to efficiently detect and repair the presence of DNA damage that is packaged into chromatin is of paramount importance, and defects in the process are associated with a variety of diseases including cancer.

The core findings of this chapter reveal that chromatin remodeling during repair of DNA damage by the nucleotide excision repair pathway (NER) is initiated from specific sites of GG-NER complex binding at the boundary sites of CIDs, which are genomic regions of higher-order nucleosome-nucleosome interaction. This result demonstrates that in undamaged cells, the complex occupies these sites and is bounded by nucleosomes containing the histone variant H2A.Z. In response to DNA damage, these boundary nucleosomes are remodelled in a GG-NER complex-dependent fashion, and this enables the Rad7 and Rad16 components of the complex to redistribute to more distal sites within the CID (van Eijk et al. 2018). Finally, it also demonstrates the importance of this mechanism to the efficient removal of DNA damage by NER, providing insight into how defects in chromatin remodeling might drive mutagenesis in cells.

NER recognises and repairs a broad range of lesions, including those induced by UV light and a variety of chemical carcinogens. Two sub-pathways of NER exist that differ in their mechanism of initiating damage recognition. During transcription-coupled repair (TC-NER), recognition is initiated by the stalling of RNA polymerase II as it encounters the damaged DNA. This couples repair of DNA damage to the process of transcription and also establishes how this process is organised within the genome. This coupling results in an efficient mechanism for removing genetic damage and restoring gene expression to damaged transcribed DNA strands, an important function of the DNA damage response. Stalling of RNA pol II subsequently recruits NER factors that function in later stages of the NER process. These factors are common to later stage repair events in the global genome repair pathway (GG-NER). However, less is known about how repair of DNA damage is initiated and organised in the GG-NER sub-pathway, which repairs all non-transcribed regions of the genome. In yeast, this pathway relies upon a protein complex that is unique for the function of GG-NER. Using genomic techniques, recently our laboratory showed that GG-NER is organised into domains related to the promoter regions of open reading frames (Yu et al. 2016). It was demonstrated that efficient DNA

repair around these sites depends on the GG-NER complex regulating the histone acetylation status of nucleosomes in the vicinity, which alters the chromatin structure. However, until now, it was not known how chromatin remodeling is initiated during GG-NER.

To tackle this problem, genome-wide nucleosome maps were generated to analyse UV-induced changes to the nucleosome landscape. The genomic distribution of changes to the three core nucleosome parameters were examined that quantify occupancy, fuzziness and position, to identify the subset of nucleosomes that are altered in response to UV-irradiation. These findings demonstrated that chromatin remodeling at this level occurs predominantly through dispersed local changes to nucleosome occupancy and fuzziness. However, nucleosome sliding in the context of gene expression can also be detected at known DNA-damage responsive genes, in line with previously published data (Lai and Pugh 2017). This data shows that the remodeling of positioned nucleosomes adjacent to GCBS's at many hundreds of genomic features in aggregate does not occur via nucleosome sliding. These results are consistent with previous biochemical observations (Yu et al. 2004; Yu et al. 2009). The complex can translocate along DNA *in vitro* through the activity of the SWI/SNF and helicase domains of Rad16, but it cannot slide nucleosomes *in vitro* (Yu et al. 2009). I suggest that this same mechanism also drives the nucleosome remodeling events described in this chapter. Furthermore, the GG-NER complex binding sites identified in this study, are not simply regions of repair initiation, but they are also locations of UV-induced histone remodeling, involving nucleosomes containing the histone variant H2A.Z. My results show that in undamaged cells these H2A.Z-containing nucleosomes represent barriers; gate-like structures that constrain and sequester the GG-NER complex at these genomic positions. DNA repair may be initiated by structural rearrangement of these barrier sites, allowing the GG-NER complex to redistribute from its initial binding locations in undamaged cells. The UV-induced loss of H2A.Z-containing nucleosomes essentially relieves the barrier effect, permitting the GG-NER complex to redistribute. Intriguingly, this process might serve to concurrently restrict RNA pol II transcription that is known to require H2A.Z-containing nucleosomes for efficient gene transcription initiation (Weber et al. 2014). Therefore, this mechanism may contribute to the inhibition of bulk transcription in response to DNA damage, while at the same time driving the efficient search for DNA damage by the GG-NER complex. Shut-down and restoration of normal gene expression is an established hallmark in

maintaining the stability of the genome in response to DNA damage (Ciccia and Elledge 2010).

Higher order chromatin structure in yeast has been identified following the introduction of methods that map distal nucleosome-nucleosome interactions, forming structural units that are classified as CIDs (Hsieh et al. 2015; Hsieh et al. 2016). These structures typically encompass 1 to 5 genes, and range in size from a few kilobases up to 10 Kbp. My results described in this chapter showed that ~50% of GCBS's can be found precisely at the boundaries between these genomic features. Conceivably, these nucleosome-nucleosome interactions contained within CIDs may represent higher-order levels of nucleosome structure that may also be remodeled during GG-NER. We will further investigate this notion in greater detail in the future. In line with previous studies in our laboratory, the DNA translocase activity of the GG-NER complex could induce the remodeling of higher order chromatin structure, similar to the loop-extrusion model suggested for CTCF-Cohesin complexes in higher eukaryotes (Sanborn et al. 2015). In this model, two CTCF-Cohesin complexes bind to the chromatin and extrude DNA through the cohesin ring structure until they encounter a CTCF binding site (Sanborn et al. 2015). The CTCF and cohesin factors reside at the base or boundary of these loop structures, which may be analogous to the boundary positions to which the GCBS complex binds in the yeast genome. Although the loop-extrusion model has not been demonstrated in yeast, and the lack of a yeast homolog for CTCF, excludes the possibility of a direct parallel mechanism. However, this study also suggests that the redistribution of the GG-NER complex, by virtue of the DNA translocase activity of Rad16 could act as a wedge to disrupt the higher-order contacts that exist in the DNA loops that make up the CIDs. Future research aims to investigate the remodeling mechanism of higher-order chromatin structure using the micro-C methodology.

To conclude this chapter, this study demonstrates that in undamaged cells, DNA repair complexes are positioned at hundreds of boundary regions that define the presence of CIDs; genomic domains of higher order nucleosome-nucleosome interactions. Suggesting that this arrangement might represent origins of DNA repair initiation that promote the efficient repair of DNA damage in chromatin. Initiating chromatin remodeling from defined origins could effectively reduce the search space for DNA damage recognition, by compartmentalising the genome into functional modular chromatin structures that can be rapidly remodeled and efficiently repaired. Therefore,

characteristic structural features of CIDs emerge when the genome is organised in this way – this ensures the rapid search and repair of genetic damage in chromatin.

The occupancy of the GG-NER complex upstream of genes in undamaged cells, shows that it is an inherent component of chromatin, as well as playing a role in repairing its structure in response to damage. The previous work in our laboratory showed that, in response to UV, the GG-NER complex regulates acetylation status, by controlling Gcn5 occupancy within these binding sites and promotes efficient repair (Yu et al. 2016). Another study also showed that, the histone variant H2AZ (Htz1), inherently enhances the occupancy of the histone acetyltransferase Gcn5 on chromatin to promote histone H3 acetylation after UV irradiation for efficient repair (Yu et al. 2013). Importantly, deletion of either Rad16/Rad7 (component of GG-NER, chromatin remodeler), Gcn5 (chromatin modifier), Htz1 (histone variant) can alter the normal pattern and distribution of WT DNA repair rates throughout the genome, raising the possibility that defective DNA repair, chromatin modification or faulty histone variant exchange, might also affect the distribution of mutations acquired within the genome. Determining whether this is indeed the case will likely help to explain how novel classes of cancer-causing genes, which are involved in modifying chromatin structure, drive tumorigenesis. This question will now be addressed in the following chapters.

Chapter IV

Measuring acquired mutations from UV damaged yeast cells: establishing a workflow

Contents

4.1 Background

In the previous chapter I undertook experiments to identify the genomic location of nucleosomes that were remodelled by the GG-NER complex in response to UV damage. These studies revealed a novel class of genomic features now described as GCBSs. These sites represent origins of GG-NER; positions from which GG-NER is initiated in response to DNA damage. The genomic distribution of relative NER rates was significantly affected by defects in either the GG-NER complex, or factors affecting the modification of chromatin. In the current chapter I undertake experiments to examine the genomic distribution and type of mutations acquired in the genomes of normal or mutation-prone cells, either damaged or undamaged by UV irradiation. My aim is to understand how the structure and organisation of the genome modulates the patterns of mutations acquired in cells. Recent efforts have been aimed at unravelling the structure and organisation of DNA repair in response to DNA damage, as well as the induction of mutations in chromatin. These findings have provided important insight into the fundamental mechanism that contributes to maintaining genome stability (Yu et al. 2011; Adar et al. 2016; Mao et al. 2016; Yu et al. 2016; Mao et al. 2017b). The genome-wide analysis of DNA damage, repair and mutagenesis has begun to address the mechanism of mutational heterogeneity observed within the yeast cells (Wyrick and Roberts 2015; Mao et al. 2017a). Both endogenous and exogenous DNA damage results in a heterogenous mutational pattern throughout the genome that can be a result of either the lack of organised DNA repair, or error-prone repair (Lawrence et al. 2013; Tubbs and Nussenzweig 2017). Therefore, it has been proposed that mutations that contribute to the development of cancer and other human disease are the biological output of both DNA damage induction and inefficient repair. Both of these biological processes operate within the genome in an extraordinary variety of different genomic contexts in order to maintain genomic integrity (Yu et al. 2016; van Eijk et al. 2018). Therefore, appropriate analysis of the genome-wide organisation of DNA damage, repair and mutation has been shown to be a useful process to gain insight into the mechanisms behind the stability of the genome and its impact on human health.

Detection of acquired somatic mutations in response to either endogenous/exogenous processes or a combination of both from whole genome sequence (WGS) data is one of the major challenges for accurate interpretation and validation of the mutational catalogues obtained from cancer genomes (Zou et al. 2018). Next generation sequencing (NGS) offers a powerful tool to investigate genome-wide variations such as small but

frequent single nucleotide variations (SNVs), insertion and deletions (Indels) and large but less frequent rearrangements, that accumulate within the genome during the life time of an individual. However, analysis of these genetic variations for comparative study using different genetic backgrounds, such as cell line-based models that contain tumour mutations or sequencing artefacts, often make it difficult to produce a reliable call on a given somatic mutation (Shi et al. 2018; Xu 2018). Distinguishing mutations from natural SNVs is further hampered by the absence of appropriate control data derived from a matching germline sample (Ding et al. 2010). A variety of experimental model systems have been developed for studying endogenous or environmental mutagenic processes that attempt to mimic the mutational processes observed in cancer genomes.

A recently reported software called 'IsoMut' was designed for fast and accurate mutation detection from WGS of multiple isogenic samples (Pipek et al. 2017), providing a unique opportunity to detect acquired mutations in a controlled system. Tumour samples contain acquired somatic mutations as well as germline variations (Alexandrov et al. 2015). Subtraction of the germline variation identified in normal tissue from the SNVs detected in the cancer genome will result in a final catalogue of somatic variants exclusive to the cancer of a patient. Mapping these acquired mutations according to their trinucleotide sequence-context allows for the generation of a mutational profile that is unique for each individual cancer. Recently, a non-negative matrix factorisation (NMF) algorithm was employed for the extraction of mutational signatures from these mutational profiles (Alexandrov et al. 2013b). Interestingly, the contribution and distribution of these mutational signatures differs between different cancer types, suggesting that various types of biological processes are operative during the development of cancer. The biological mechanisms underlying these novel mutational processes are, however, in large part unknown.

Of the many sources of mutations, replication of DNA can introduce variants due to the misincorporation of nucleotides that occurs predominantly during the later phases of the cell cycle (Waters and Walker 2006). This suggests that early and late replicating regions of the genome might have different mutational loads, which has been documented in literature (Sima and Gilbert 2014). Therefore, studying mutation induction in relation to replication timing will help to explain the heterogenous distribution of mutations within the genome. The alteration of the genetic landscape can also be associated with mutations induced by transcriptional activity within the cell (Kim and Jinks-Robertson 2012). In this context, transcriptional strand bias correlates well with regional heterogeneity in the

distribution of mutations found in cancer genomes (Haradhvala et al. 2016). Taken together, genomic context, accessibility of the repair factors within chromatin, replication timing and transcriptional activity within genome combine to modulate the genome-wide distribution of mutations.

One source of exogenous DNA damage is ultra-violet (UV) irradiation. DNA is a main target for UV damages and the most abundant mutagenic and cytotoxic DNA damages caused by UV light are cyclobutane pyrimidine dimers (CPDs), 6–4 Photoproducts (6-4PPs) and their Dewar valence isomers (Ikehata and Ono 2011). To handle these UV-induced damages, cells have developed several repair and tolerance mechanisms. NER, BER, TLS and post-replicative MMR all play important roles in the cellular response to these UV-induced lesions through the action of DNA damage detection mechanisms, repair and lesion bypass (Sinha and Häder 2002; Ikehata and Ono 2011). The concerted action of these pathways results in the repair and accurate replication of DNA. However, high levels of DNA damage or compromised repair activity can result in unrepaired lesions that are tolerated by lesion bypass. Persisting lesions and lesion bypass together constitute the source of mutagenesis.

Since, repair rates are affected by genomic context (Mao et al. 2016; Yu et al. 2016; van Eijk et al. 2018), the pattern of acquired mutations in relation to different genomic features such as, open reading frames (ORF), transcription start sites (TSS) and transcription end sites (TES), will be considered here in order to observe the enrichment or depletion of mutations at these genomic features. Nucleosomes provide the first level of DNA compaction and chromatin structure covering 75-90% of the genome (Kornberg and Lorch 1999). These structural units play an important role in regulating gene expression (Li and Reinberg 2011), DNA damage distribution (Mao et al. 2017b) and organisation of DNA repair (Reed 2011). It is well established that regional differences in chromatin environment, such as euchromatic to heterochromatic regions influence gene expression, DNA repair and DNA replication (Groth et al. 2007). Chromatin accessibility of these higher order chromatin structures is modulated by covalent modification of histone proteins, histone variant exchange and other DNA binding proteins. Therefore, the effect of strongly positioned nucleosomes (SPN), nucleosomes that occupy the same translational positioned in every cells, and nucleosome free regions (NFR) on the distribution of mutations will be examined. Similarly, plotting the mutation data in relation to dense and open chromatin will inform on the distribution of mutations in relation to these features of the chromatin environment. We hypothesize that the

organisation of repair within chromatin might influence the distribution of acquired mutations. The previous chapter describes the GG-NER-dependent nucleosome remodelling required for efficient repair at GCBSs. In short, we observed lower relative rates of repair in these genomic sites in GG-NER deficient cells after UV exposure (van Eijk et al. 2018), indicating that both nucleosome remodelling and GG-NER is organised and initiated from these sites. I will determine the significance of this on the pattern of acquired mutations found in the genome of yeast cells.

One way to understand the impact that the structure and organisation of DNA repair has on the process of mutagenesis, is to allow mutations to accumulate through many cell divisions with or without exposure to DNA damages, and to sequence the genome to identify the types, number and locations of the mutations that arise. The distribution of mutations may vary in response to both exposure to DNA damage, and in relation to genomic context. This has the potential to reveal the relationship between relative repair rate, replication timing, DNA accessibility and a range of other parameters with mutagenesis after exposure of cells to DNA damage. To study how the distribution of mutations alters either in response to DNA damage, defects in DNA repair, or a combination of both a workflow is required for studying the genome-wide mutational distribution in both wild-type and DNA repair deficient cells. For this we use yeast as a model organism and UV-induced DNA damage as a model lesion and known mutagen. To establish this method, wild-type and MMR-defective yeast cells were subjected to a series of cell passages to generate several hundreds of generations that are expected to accumulate mutations due to endogenous DNA damage with each generation (Lujan et al. 2014). Importantly, for this study I also introduce UV radiation exposure to induce DNA damage to examine the accumulation of mutations due to exogenous DNA damage. In the context of this study, the MMR deficient background was included because MMR deficiency leads to a well-known mutator phenotype (Lujan et al. 2014; Meier et al. 2014; Meier et al. 2018) and is one of the most studied DNA repair defects known to be associated with human cancers (Alexandrov et al. 2018).

I chose the baker's yeast, *Saccharomyces cerevisiae* as model system for studying mutagenesis, since previous work in our laboratory resulted in the development of tools and methods for measuring genome-wide DNA damage and repair (Teng et al. 2011; Bennett et al. 2015; Powell et al. 2015). This provides a unique opportunity to compare the distribution of UV-induced mutations with DNA repair rate data from the same isogenic background. In addition, *S. cerevisiae* has a well-annotated reference genome

and mutation accumulation protocols are available. Some limitations of interspecies differences when comparing yeast data to human models and cancer have to be recognized. However, many of the DNA repair pathways are conserved and most mutational processes are common across species. Moreover, it is straight-forward to clonally expand yeast cells from a single cell. This clonal expansion can be considered as a model for the clonal nature of events that occur in cancer genomes (Larrea et al. 2010; Lujan et al. 2014; Serero et al. 2014; Segovia et al. 2015).

Additionally, I will compare the experimentally controlled biological processes that induce the mutational patterns in yeast with the mutational signatures found in human cancer genomes. This will determine the relationship between the mutational signatures found in yeast with the mutational signatures described from sequencing cancer genomes.

After extracting the mutational catalogues, I will compare the distribution of DNA repair rates with the mutational patterns to determine how rates of repair affect the distribution of mutational patterns. Finally, I will plot the landscape of UV-induced mutations in wild-type and MMR defective genomes and compare them with the pan cancer analysis of whole genome (PCAWG) mutational signatures, to identify which of the cancer genome mutational signatures they most closely resemble.

4.2 Material and Method

To establish the workflow for accumulating and mapping acquired mutations, I start with wild-type BY4742 yeast cells and the *msh2* mutant cells defective for MMR. Msh2, human homolog for MusS, protein binds to DNA mismatches; form complex with Msh3 and Msh6 that bind to the DNA mismatches to initiate the mismatch repair process. As the DNA damaging agent, I used UV which is well-characterised genotoxic agents and form bulky adducts in DNA and often associated with different types of skin cancers. DNA damage induced by UV is mostly repaired by the NER and to a lesser extent the MMR processes.

4.2.1 Propagation of cells for accumulation of mutations with or without UV irradiation

The experimental strategy for accumulation of mutation in wild-type and MMR defective yeast cells with or without exposure to UV-induced DNA damage is described here. As per figure 4.1, yeast cells were subjected to 30 bottleneck passages by re-streaking single colonies on YPD plate for each strain with occasional UV light exposure. To start, the relevant yeast strains were retrieved from an -80°C glycerol stock, streaked onto YPD plates and incubated at 30°C. Next, a single colony was taken and inoculated into YPD liquid media and grown overnight to log-phase. Growth was monitored by measuring OD_{600} and cell counting. Log-phase growth is reached at about OD ~0.6 or a cell count of ~ $2x10^7$ cells/mL. These log-phase cells were subjected to 100 J/m^2 UV irradiation in 1× PBS using 10 J/m^2s^1 UV-C (254 nm) for 10 seconds. The details of UV irradiation are mentioned in Material and Methods chapter (Chapter II). Un-irradiated cells were kept aside as a control. Both UV irradiated, and un-irradiated cells were then streaked onto YPD agar plates to obtain single colonies (referred to as passage 1). Then 10 clones/colonies were taken with an aseptic loop from each YPD plate and streaked onto a new YPD plate with 4 clones per plate. At the same time, the cells from passage 1 were stored in glycerol at -80°C. For each clone, further propagations were carried out every 2 or 3 days (e.g. every Monday, Wednesday and Friday (Figure 4.1)) until passage 30. Every third passage the cells were treated with UV by exposing the plate to 10 J/m^2 of UV-C to mimic a process similar to intermittent exposure to UV irradiation as humans can experience from the environment. Cells from passage 1, 4, 7, 10, 13, 16, 19, 22, 25, 28 and 30 were stored in glycerol stocks at -80°C. DNA was extracted (Appendix Figure A3.1) as mentioned in the chromosomal DNA extraction section in Material and Methods chapter (Chapter II) from passage 1 and passage 30 for subsequent WGS.

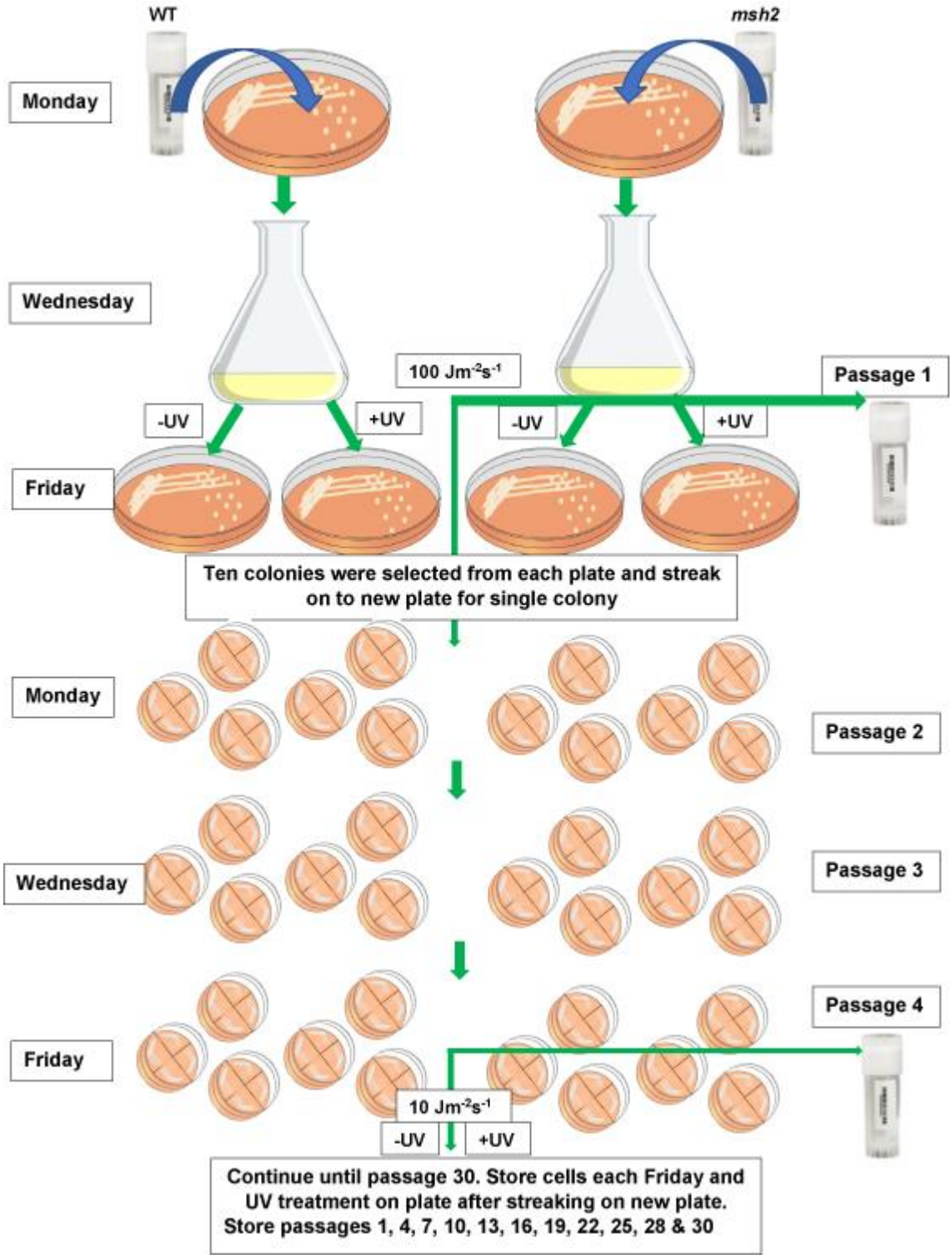


Figure 4.1: Study layout for accumulation of mutations using yeast as an isogenic model organism. Both wild-type and *msh2* cells were subjected to 30 single cell bottleneck passage on solid media with or without occasional exposure to UV damage. DNA was extracted and purified from passage 1 and passage 30 samples. Sequencing library prepared and then send out for sequencing using Illumina sequencing platform.

This strategy allows for the accumulation of mutations over 1,100 generations of yeast cells with or without exposure of UV light. All the strain of *S. cerevisiae* used for this study are haploid and descended from the S288C yeast strain: MATα; *his3D1; leu2D0; lys2D0; ura3D0*; BY4742 (Y10000). S288C is the strain used for generation of the *S. cerevisiae* reference genome referred to as sacCer3 (Engel et al. 2013). To acquire mutations, 10 clones from both wild-type and MMR deficient cells were propagated from passage 1 till passage 30. To accumulate UV-induced mutations, the same process was followed with occasional exposure of UV irradiation. Genomic DNA was extracted from a single passage 1 (P1 control) clone and ten passage 30 clones. Subsequently, a sequencing library was prepared by following the Illumina library preparation protocol and genomic DNA was sequenced using the Illumina sequencing platform (Details are mentioned in chapter 2). Whole genome sequence data was obtained by Illumina paired end sequencing with read sizes of 75 and 150 bases in two sequencing batches, Mi-Seq (150bp) and Hi-Seq (75bp).

4.2.2 Quality Checking, Alignment, Sorting and Indexing with Reference Genome

As mentioned in the previous chapter, after sequencing, all the raw paired-end fastq files were checked with FastQC (Babraham Bioinformatics) for the quality of the sequenced reads. Raw data files with low coverage were resequenced to get the minimum coverage level (>20x). The raw paired-end reads were then aligned by Burrows-Wheeler Alignment Tool (BWA, version 0.7.5a-r405) (Li and Durbin 2009) using sacCer3 (*S. cerevisiae* S288C) as a reference genome. As part of the alignment pipeline, the resulting SAM files were sorted and converted to BAM files, PCR duplicates were removed from the BAM files and indexing was performed using samtools (Li et al. 2009), Finally, the samtools 'flagstat' algorithm was used to extract information on the alignment.

The pipeline for alignment, SAM to BAM conversion, sorting, PCR duplicate removal and indexing is shown below:

```
[bwa mem -t 7 -M -R
"@RG\tID:F068_M_021\tPL:ILLUMINA\tPU:0\tLB:F068_M_021\tSM:F
068_M_021" /home/user/bwa-0.7.15/sacCer3.fa
F068_M_021_R1.fastq F068_M_021_R2.fastq | \
samtools view -Shu - | \
samtools sort - F068_M_021
samtools rmdup F068_M_021.bam > F068_M_021.bam
samtools index F068_M_021.bam
samtools flagstat F068_M_021.bam > F068_M_021.b.stat]
```

Aligned BAM files were checked for the correct genotypes of the strain used for these studies before calling mutations using Integrated Genomics Browser (IGB, Freese et al. 2016) (Figure 4.2).

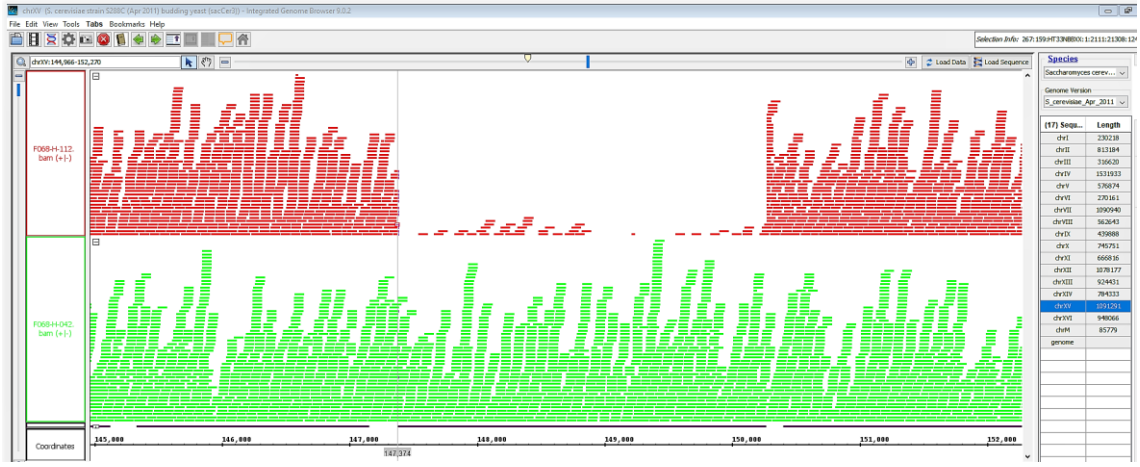


Figure 4.2: Snapshot of a short section of chromosome XV showing the position of *msh2* (YOL090W) deletion in yeast genome (track in red) and same positing in the wild-type genome (green track).

4.2.3 Collecting the catalogue of somatic mutations using IsoMut

After the generation of BAM files, I used 'IsoMut' (Pipek et al. 2017) for variant calling, which is a fast and accurate mutation detection tools for whole genome sequences of multiple isogenic samples. In brief, after applying a base quality filter [30], data from all samples were compared at each genomic position and filtered using optimised parameters of minimum mutated allele frequency [0.2], minimum coverage of the mutated sample [5] and minimum reference allele frequency of all the other samples [0.93]. The raw output from IsoMut was used to filter mutations called using a probability-based quality 'S score' calculated from the mutated sample and one other sample with the lowest reference allele frequency. The code used by IsoMut is available for download online (https://github.com/riblidezso/isomut). After running IsoMut, two output files are generated, one containing SNVs and one containing Indels. Both files contain information about the position within the genome of each variant, the mutated and altered allele, the coverage at the variant positions, the S score of the mutations and other parameters (all the code for running IsoMut used for this study and output results are attached in the e-Appendix file). After loading the variant information containing files into the 'R' statistical environment (Team 2014), I next generate the 'tuning curve' which uses the 'S score' reported by IsoMut to determine a threshold value to minimise false positives or background mutation detection using the control samples. The fine-tuning step is strongly

encouraged and yields better results than using the predefined filtering parameters only (personal communication, Dr. David Szuts). This two-step detection and filtering method allows the user to select less strict values for the 'sample_mut_freq_min' and 'sample_cov_min' filters when running 'IsoMut' to reduce the detection limit followed by further filtering the mutation detection based on the S score. For evaluation, the presence of mutations called by IsoMut were confirmed in the raw sequence data using a genome viewer or browser such as the integrated genomics browser (IGB) (Freese et al. 2016).

4.2.4 Subtracting background mutations to generate the final catalogue of acquired mutations

The 'tuning curve' was used to set a cut-off point, based on S score of control sample to cancel out background mutations or genetic variance from acquired mutations obtained during the experimental time course. The aim is to set the threshold such that the data from the first passage clones will have zero to 1 mutation per genome (recommendation based on (Pipek et al. 2017)). In these samples no unique, treatment-induced mutations should be present, thus the score can be tuned by minimizing the number of detected mutations in these samples, while maintaining satisfyingly high numbers in samples that underwent mutagenic treatment. The tuning procedure can be individually carried out for SNVs and Indels to achieve optimal results for all types of mutations. The value of the S score is related to the probability of false positive mutation. The score is calculated as the negative logarithm of Fisher's exact test p value on the two 'noisiest' samples (Pipek et al. 2017). Thus, a higher score means higher confidence mutation call. Please note that, the score itself has no clear physical probabilistic interpretation but is proved to be efficient for this purpose. The code for generating tuning curve are attached as an e-Appendix. At the end of this process I end up with a high quality, filtered and unique list of mutations from each genetic background of yeast cells considered in this study.

4.2.5 Mapping somatic mutations according to different genomic features

After generating a catalogue of the acquired mutations, both substitution mutations and Indels are plotted according to their types, load and location within the yeast genome. In case of Indels mutation, the mutational catalogue was filtered using R package `'dplyr'` based on types of Indels such as insertion or deletions, subtypes (A, T, G or C) of insertion or deletions and 1, 2, 3, 4, ≥5 bp microhomology or repeat mediated Indels. The number of Indels is then plotted as a bar chart according to their frequency within different types

of genomic background. For a detailed description of output of Indels, please see the e-Appendix file.

The substitution mutations, which are the main focus of this study, are plotted using the 'MutationalPatterns' Bioconductor R package (Blokzijl et al. 2018). First of all, to evaluate the enrichment and depletion, of substitutions, the observed and expected mutations were plotted around genomic features that exist as linear structures of DNA from *Saccharomyces* Genome Database (SGD) such as ORF, TSS, TES and random sites. About 3,100 yeast genome-wide random sites with 400pb window were used as a control. Secondly, chromatin associated genomic features such as NFR, SPN, and Micro-C boundaries, GCBSs with NFR associated TSS, non GCBS with NFR associated TSS were also used to test the effect of chromatin environment on mutational distribution. The strongly positioned nucleosomes data in the yeast genome (~10, 000, with the nucleosome score > 5) was obtained from (Brogaard et al. 2012) supplemental file for comparison with NFR. The ratio between observed and expected mutations was used to calculate the statistical significance using a one-sided binomial test as mentioned in the 'MutationalPatterns' package (Blokzijl et al. 2018). These genomic features files are imported as bed files and first converted into 'GRanges' objects. Next, the total number of observed mutations that fall into these genomic ranges was counted. The mutational density (in mutations per bp) was estimated by calculating the total mutation load divided by the total number of bases surveyed. The surveyed bases are the positions in the genome, which have enough high-quality reads to call a mutation. Most of the sequences used here are mapped ≥ 98%, so for simplicity, the complete sacCer3 ref genome was used as a surveyed region for each sample. To test for significant enrichment or depletion, the $\log_2$ ratio between observed and expected was calculated.

Third, the mutations were plotted according to the histone acetylation status that was used as a measure for open versus closed chromatin. High and low levels of histone H3-acetylation regions were generated using previously reported microarray data from our laboratory (Yu et al. 2005). The probe information from the microarray was used as a window for generating genome-wide histone acetylation status (data attached as an e-Appendix). Next, this data was converted into a 'GRanges' object for detection of enrichment or depletion of mutation, within high or low acetylated regions of whole yeast genome, using 'MutationalPatterns'.

Fourth, the mutations were plotted based on early and late replication timing information. Early and late replicative strand information for the yeast genome was generated by following the Repli-seq analysis protocol (Marchal et al. 2018) using data from (Muller et al. 2014). The data used here for replication timing is attached as an e-Appendix. The resulting replication timing data was loaded into R for replicative timing bias analysis using 'MutationalPatterns' package.

Fifth, the mutations were plotted according to the transcriptional strand for strand bias analysis. The transcriptional strand information was used after loading SGD transcript database as a 'TxDb' objects `'TxDb.Scerevisiae.UCSC.sacCer3.sgdGene'` into the R statistical package using the algorithm as mentioned in 'MutationalPatterns' package.

Six, the rainfall plot was generated, to observe the genome-wide distribution of the mutational pattern using inter-mutational distance.

Finally, the substitutions were plotted as 96 trinucleotide profiles and 192 trinucleotide profile in relation to the transcribed or non-transcribed strand information (originally used for mutational signature analysis (Alexandrov et al. 2013b), see the next section).

4.2.6 NMF for *de novo* extraction of mutational signatures

All the substitutions within the catalogue of somatic mutations were used to extract the mutational signatures. For this analysis, a 96-trinucleotide matrix was generated after pooling samples of the same genotype and treatment. Then NMF was used for *de novo* extraction of mutational signatures from the 96-trinucleotide mutational profile matrix. This is the same approach originally used by Dr. Ludmil Alexandrov for deciphering mutational signatures from cancer genomes (Alexandrov et al. 2013b). NMF is a set of algorithms in multivariate data analysis and linear algebra where a matrix V is factorised into (usually) two matrices W and H, with the property that all three matrices have no negative elements (Figure 4.3). This non-negativity makes the resulting matrices easier to inspect.

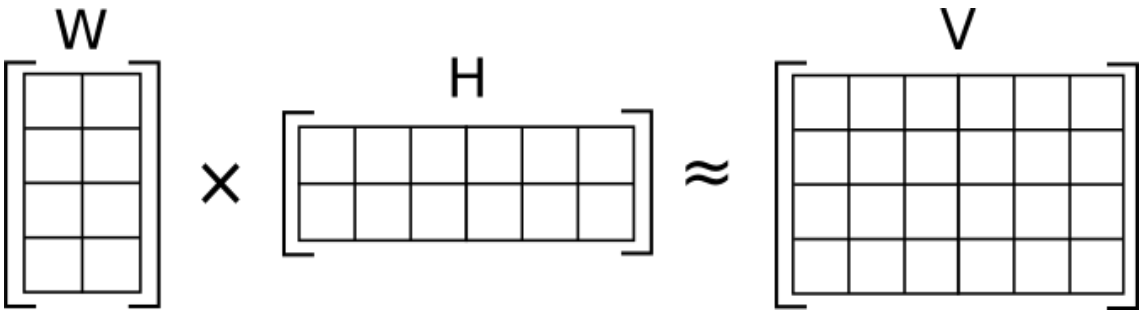


Figure 4.3: Illustration of approximate non-negative matrix factorization (NMF) in which the product matrix V is represented by the two smaller matrices W and H, which, when multiplied, the resulting matrix will be approximately reconstructed V matrix.

For examples, the product matrix, V can be a [a×b] non-negative matrix with N > 0 a factorisation rank. Non-negative Matrix Factorization (NMF) consists in finding an approximation V ≈ WxH, where, W and H are [a x N] and [N x b] non-negative matrices, respectively. In practice, the factorization rank N is often chosen such that N << min(a, b). The objective for this is to infer low-dimensional structure from high-dimensional omics data to summarise and extract information regarding complex biological processes contained in V into N factors: the columns of W. Choosing this factorisation rank (N) is the critical parameter, which is the number of mutational signatures in this case. It can be possible to estimate the optimal factorisation rank using the NMF Bioconductor package in R. As suggested by the developer of the NMF package in R, a common way of deciding on N is to try different values, compute some quality measure of the results, and choose the best value according to this quality criteria. Several approaches have then been proposed to choose the optimal value of N. For example, (Brunet et al. 2004) proposed to take the first value of N for which the cophenetic coefficient starts decreasing, (Hutchins et al. 2008) suggested to choose the first value where the residual sum of squire (RSS) curve presents an inflection point, and (Frigyesi and Höglund 2008) considered the smallest value at which the decrease in the RSS is lower than the decrease of the RSS obtained from random data. Here we follow mainly both the 'Brunet 2004' and 'Frigyesi 2008' approaches for deciding on N. At the same time a consensus heatmap was generated for checking the clustering of the samples based on factorisation rank or N.

4.2.7 Cosine Similarity and Reconstruction of a Mutational profile

After generating the 96 trinucleotides mutation count matrixes from the catalogue of somatic mutations, two separate approaches were followed in trying to find the optimum contribution of or similarity to the known mutational signatures observed in repertoires of cancer genomes. Firstly, the Cosine Similarity between two count matrixes can be used

to express the level of similarity between the mutational profile of a sample and the 49 known PCAWG mutational signatures (Alexandrov et al. 2018). Secondly, the Reconstruction of the mutational profile of each sample using the known PCAWG signatures can inform on what biological processes contributed to the pattern observed. The 49 know PCAWG single base substitution signatures profile is attached as an e-Appendix.

The cosine of two non-zero vectors, such as A and B, can be measured by using the dot product formula.

$$\mathbf{A.B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos\theta$$

The cosine similarity cos(θ), for this A and B vectors will be the measure that calculate the cosine distances of the angle between them, can be represented by the following formula.

$$similarity = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where A_i and B_i constituent of the A and B vectors. Because the elements of A and B are nonnegative, the cosine similarity has a range between 0 and 1. When the cosine similarity is 1 between two matrixes, these matrixes are the same. In contrast, when the similarity is 0, the matrixes are independent. Cosine similarities were calculated based on the pooled samples' 96-trinucleotide matrix with the recently reported 49 PCAWG mutational signatures (Alexandrov et al. 2018).

To reconstruct the samples' mutational profile matrix, the contribution of any set of matrices (known signatures) to the mutational profile matrices of a sample can be quantified using a non-negative least-squares (NNLS) optimisation (the equation below), where the weights are not allowed to become negative:

$\text{Min}_H \|W \cdot H - V\|^2$

where H is the weight or exposure matrix and is non-negative, W is the features matrix and V is the original (or product) matrix (Figure 4.3). This approach is unique and useful for this study to find out the contribution of known mutational signature matrices to small cohorts, or individual samples. This Euclidian norm of residual minimisation approaches is used for reconstruction of the mutational profile of a single sample using already known

or *de novo* extracted mutational signatures. The 'MutationalPatterns' R/Bioconductor package intergrades this algorithm form another R package called 'pracma' in R.

Again, the cosine similarity of the original mutational profile and the reconstructed mutational profile indicates how well the mutational profile of each sample can be reconstructed using existing signatures. This information can be used to explain the cause of sample's mutational profile. A lower cosine similarity (close to 0) between original and reconstructed profile indicates that the causes of the sample's mutational profile cannot be explained or fully reconstructed by the existing signature.

4.3 Results

4.3.1 Establishing the catalogue of acquired mutations from isogenic yeast strains

In order to establish a workflow for the detection of induced mutations, I used wild-type and mismatch repair deficient yeast strains and UV radiation as a DNA damaging agent and known mutagen. This experimental design allows me to study the contribution to mutagenesis of both endogenous and exogenous DNA damage, as well as account for defects in DNA repair. Mismatch deficient cells have a well-known mutator phenotype and act as a positive control for mutation in this context. Both wild-type and MMR deficient *msh2* cells were processed as described in Figure 4.1 (see Materials and Methods), generating 4 individual treatments representing 3 biological processes. The treatments are represented by the samples of wild-type cells (i) untreated and (ii) treated with UV irradiation, and the *msh2* mutant (iii) untreated and (iv) UV treated. The biological processes are represented by (a) endogenous processes inducing mutagenesis after 30 passages in all untreated samples, (b) UV-induced mutagenesis in the UV-treated samples and (c) MMR defective mutagenesis during this experimental time period (the *msh2* mutant samples). After NGS the resulting sequencing reads were processed according to the pipeline described in the methods section of this chapter, generating an output from IsoMut variant calling that generates two files. One containing the substitution mutations, and the other containing short Indels. To account for a level of background mutation detection, the passage 1 (P1 control) and P30 clones for both wild-type and *msh2* mutant yeast cells were used to select a cut-off aiming for 0 to 1 mutations detected in the early founder passage (P1). The background was subtracted after plotting the IsoMut output data into the R statistical language as a tuning curve (Figure 4.4). This tuning curve depicts the cumulative mutations below a certain S score on the Y-axis and their respective S scores on X-axis. This tuning curve reveals at what S score the threshold of 0 to 1 mutation per genome is reached in the P1 control samples (Figure 4.4). In the case of the wild-type and *msh2* mutant data, I set the threshold at an S score of ≥ 3.2 in order to filter out the vast majority of any background mutations. This procedure using the tuning curve filtration method was performed for all experimental data sets described in this chapter and the results are attached as an Appendix figure A3.2-A3.4.

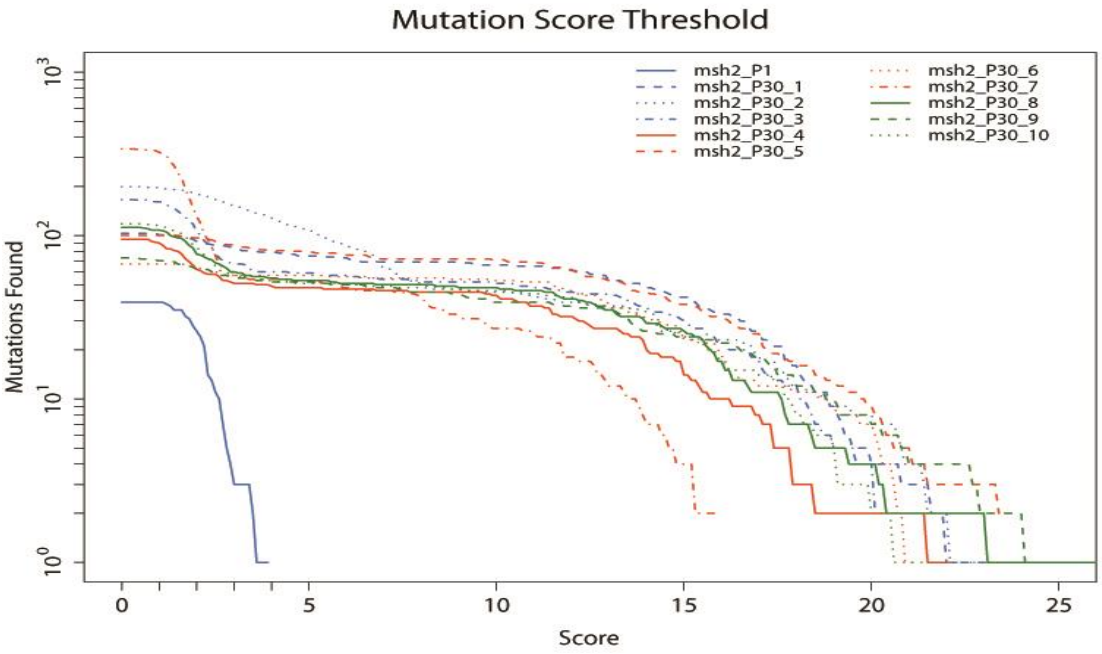


Figure 4.4: Tuning curve showing the cumulative distribution of mutations for *msh2* mutant cells as detected by IsoMut. The threshold score was determined based on the control P1 samples. Setting up a cut off value of score 4 (based on P1 control) will filter out all the background parental clone mutations and result in unique sub-clonal substations.

At the end of this analysis, the output file contains the acquired catalogue of mutations for both substitutions and indels. An overview of the number of substitutions and indel mutations accumulated in the wild-type and *msh2* mutant cells are provided in table 4.1.

Table 4.1: Number of SNV and short indels detected in the wild-type and *msh2* mutant yeast cells with or without UV damage.

Treatment*	Passage	n	Total SNV	SNV Mean	Total Indels	Indels Mean
BY4742_WT	Starting Clone	1	1	1	0	0
	End Clone	10	128	12.8	13	1.3
BY4742_WT_UV	Starting Clone	1	2	2	1	1
	End Clone	10	642	64.2	35	3.5
msh2	Starting Clone	1	0	0	1	1
	End Clone	10	668	66.8	474	47.4
msh2_UV	Starting Clone	1	3	3	1	1
	End Clone	10	1499	149.9	587	58.7

Abbreviations: WT, wild-type; m*sh2*, mismatch repair defective cells; UV, Ultraviolet; n=number of clones. Independent mutations in the starting clone represent false positives of the number of detections. SNVs = Single Nucleotide Variations. InDels = Insertions and deletions. * All the yeast strains used in this experiment are haploid and alpha mating type.

After 30 passages, the mutational load in wild-type cells is 128 from a total of 10 clones after ~1,100 generations. The calculated spontaneous base substitution rate is $9.6x10^{-10}$ $bp^{-1}cell\text{-}cycle^{-1}$, which is similar to a previously reported mutation rate in budding yeast ($3.6x10^{-10}$ $bp^{-1}cell\text{-}cycle^{-1}$) (Serero et al. 2014). The two- or three-fold increase could be due to the different strains used in this study and/or the detection by IsoMut and filtration used here. Under the same propagation conditions, the mutational load in *msh2* cells is 668 from 10 clones. Which is about 5-fold higher than the isogenic wild-type mutational load, indicating that the majority of these mutations result from unrepaired replication errors. In *MSH2* deleted cells, the mutation rate is $4.8x10^{-9}$ $bp^{-1}cell\text{-}cycle^{-1}$, which is lower than previously reported for *MSH2* deleted cells ($1.6x10^{-8}$ $bp^{-1}cell\text{-}cycle^{-1}$) (Lujan et al. 2014). In response to UV irradiation, the substitution rate in wild-type cells ($5.0x10^{-9}$ $bp^{-1}cell\text{-}cycle^{-1}$) increases more than 5-fold compared to the spontaneous mutation rate observed in untreated cells. As expected, these results confirm that UV irradiation is a strong mutagen (Ikehata and Ono 2011). In *msh2* mutant cells irradiated with UV, the mutation rate is more than 2-fold higher compared to untreated *msh2* cells resulting in $1.1x10^{-8}$ mutations $bp^{-1}cell\text{-}cycle^{-1}$. The contribution of both UV irradiation and the MMR defect in these cells can be observed by the 2-fold higher mutation rate in UV-treated *msh2* cells as compared to their wild-type counterparts. Additionally, in response to UV damage in *msh2* cells the mutational load is even higher at ~12-fold compared to wild-type undamaged cells. This variation in mutational load and type of substitutions can be explained by the fact that lack of MMR itself has an effect on the mutational load even without exposure to DNA damaging agents. Note that these mutation frequencies were calculated based on substitution mutations, not considering indels. Taken together, these initial findings demonstrate that the substitutions induced by two well-known mutagens can be accurately detected with this novel work flow through the combined use of propagation of yeast cells, UV irradiation as a mutagen and IsoMut for mutation detection.

Considering indel mutations, we find that these events are relatively rare in wild-type cells, even in response to UV damage, indicating that endogenous DNA damage

replication errors are not frequently generating indels. In the case of MMR defective cells, the indels are detected more frequently after passage 30, even without DNA damage. Importantly, I found that the indels generated in this mutant occur at a similar order of magnitude to that found in a similar study (Meier et al. 2018) using *C. elegans* as a model organism. Overall, deletions are more dominant than insertions when MMR is absent. Both A and T base insertions and deletions are higher in both UV irradiated and unirradiated cells (Figure 4.5). No significant change in the total number of indels was observed after UV damage in *msh2* mutant cells. However, higher relative rate of microhomology or repeat mediated insertions are observed in response to UV damage (Table 4.1 and Figure 4.5).

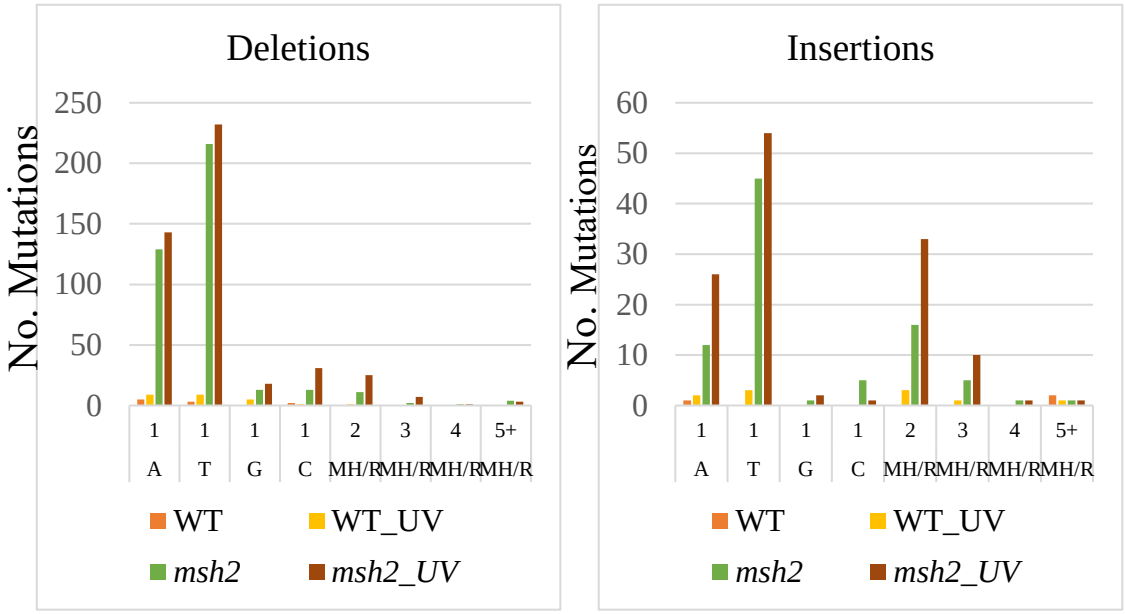


Figure 4.5: The total number of short Indels in wild-type and MMR defective (*msh2* mutant) cells. MH/R = microhomology or repeat mediated Indels. Y-axis represent the total number of mutations.

These results indicate that the biological processes of mutagenesis are comparable between different organisms. The following sections will focus on the substitution mutations and their genome-wide distributions.

4.3.2 The distribution of base substitution mutations in relation to genomic features

Mutational heterogeneity and the dispersed distribution of mutations within the genome is frequently observed in human diseases like cancer (Lawrence et al. 2013). Moreover, features of the linear genome such as genes and transcription factor binding sites (TFBSs) have been identified as hotspots for mutations induced during cancer development (Perera et al. 2016; Sabarinathan et al. 2016). Similarly, replication timing and transcriptional

activity have been shown to correlate with the presence of mutations found in cancer (Haradhvala et al. 2016). Interestingly, similar findings for the cellular repair capacity at different genomic features suggest that heterogeneous repair rates can influence the distribution of mutations within genome (Wyrick and Roberts 2015; Adar et al. 2016).

All the possible single base substitutions found in higher eukaryotes can be denoted as C>A, C>G, C>T, T>A, T>C, T>G, G>T, G>C, G>A, A>T, A>G and A>C. Because of the complementarity between DNA bases and the inability to identify whether the substitution originated on the DNA leading or lagging strand, these 12 substitutions can instead be represented as either of the following 6 substitutions:

C>A, C>G, C>T, T>A, T>C, T>G

or

G>T, G>C, G>A, A>T, A>G, A>C

For consistency, I am going to use the following annotation [C>A, C>G, C>T, T>A, T>C, T>G], for the 6 types of base substitution analysed throughout this thesis.

Recent studies showed that repair of UV-induced damages is modulated by the linear genetic context (Mao et al. 2016; Yu et al. 2016), therefore, I obtained a list of genomic features for ORFs, TSS and TESs to calculate the ratio of observed and expected mutations at these features using the substitutions obtained for the four experimental groups described earlier. I used SGD (Cherry et al. 1998) as a resource for identifying and collating these genomic features. To examine the significance of the difference between observed and expected mutations, a similar number of random genomic positions were examined as a control group. Using this information, the number of mutations in relation to the above described features can now be plotted (Figure 4.6 top panel) and the difference between the observed and expected number of base substitutions is expressed as the $\log_2$ ratio between these two (Figure 4.6 bottom panel). Counting mutations across all ORFs shows that in wild-type cells there are over 100 mutations found at these sites (Figure 4.6 top ORF panel). This number is consistent with the expected mutational load based on the overall mutation frequency throughout the genome. As a result, the $\log_2$ ratio of observed over expected is effectively zero, and demonstrates no difference in the levels of mutations in relation to ORFs compared with the genome overall (Figure 4.6 bottom ORF panel).

As observed previously (Table 4.1), UV irradiation induces mutations in the genome. This result is confirmed here as we observe a UV-induced increase in mutations found in ORFs, TSS's and TES's and of course at randomly selected sites (Figure 4.6). We noted that, there are slightly *fewer* UV-induced mutations than expected in ORFs. This is confirmed by the $\log_2$ ratio of observed over expected mutations, which is negative, but this difference did not reach statistical significance (Figure 4.6 bottom ORF panel).

Similar observations are made for the *msh2* mutant: (i) UV irradiation results in increased levels of mutations at ORFs and (ii) slightly fewer mutations are detected at ORFs compared to the expected number. However, the absence of MMR also increases the overall mutational load at these genomic features in the absence of UV damage (Figure 4.6 top ORF panel) and the effect of MMR and UV-irradiation on mutation induction is cumulative. Overall, there are no significant differences of the expected versus observed frequency of mutations when the data is plotted over all ORFs in the yeast genome, indicating that the induction of mutations in ORFs occurs at a similar frequency to the genome overall.

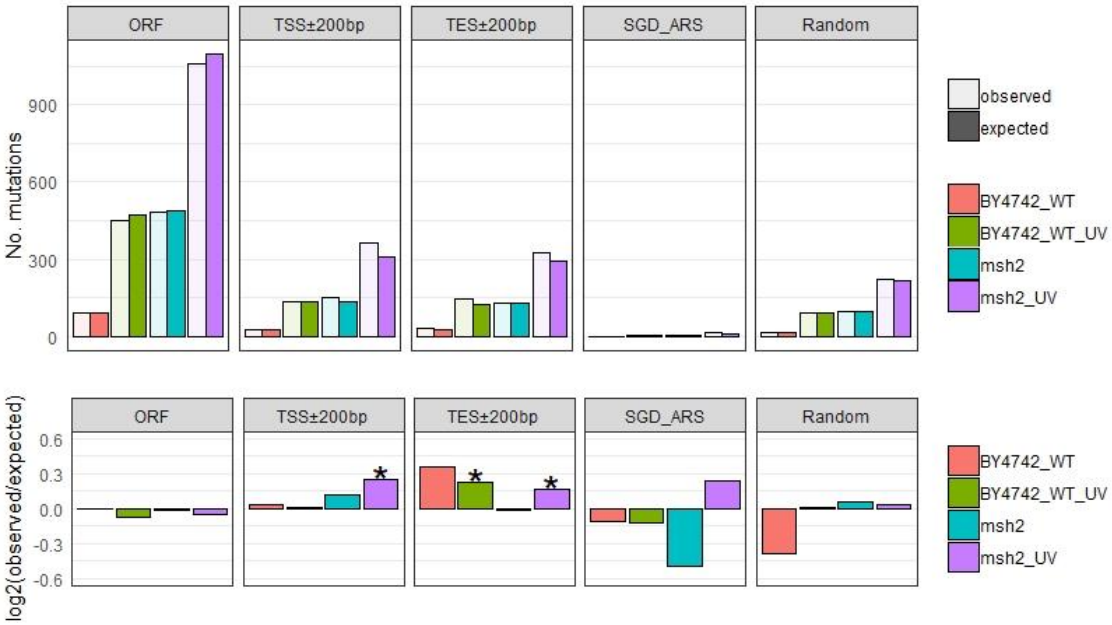


Figure 4.6: Enrichment and depletion of point mutations in relation to linear genomic features such as ORF, TSS, TES, ARS and random genomic sites for both wild-type and MMR defective yeast cells either with, or without UV damage. The $\log_2$ ratio of the number of observed and expected point mutations indicates the effect size of the enrichment (above the midline) or depletion (below the midline) in each region. Asterisks indicate enrichments or depletions of mutations that show a statistically significant difference from the expected number ($P < 0.05$, one-sided binomial test).

After scaling and orienting the genomic positions at TSS and TES, and including 200 bp flanking regions around these sites, the mutational load was again quantified and compared to the expected number of mutations in relation to these sites. At TSSs in wild-type cells, no significant mutational bias is observed, even after exposure to UV damage. However, at TESs significant enrichment in mutational load is observed after UV exposure, as indicated by an asterisk (Figure 4.6). It is possible now to compare these findings with the relative repair rates measured for UV-induced DNA damage (Figure 4.7). The results show that in wild-type cells, the higher relative repair rate around TSSs results in the expected number of mutations being induced at these sites compared to the genome overall. However, at TESs a higher mutational load than expected is observed and this correlates with a significant reduction in the relative rates of repair observed at the 3´ ends of the genes. In both TSSs, and TESs, significantly higher levels of mutations are observed than expected in UV damaged MMR-defective cells (Figure 4.6 top TSS and TES panel). This difference is dependent on UV damage, as no significant difference is observed in untreated cells. This demonstrates that UV-induced mutations are enriched at both TSSs and TESs within MMR defective yeast cells (Figure 4.6 bottom TSS panel). At ARSs, the origins of replication in yeast, fewer than expected mutations are observed in untreated MMR defective cells, but then higher levels of mutations are detected after UV damage, but not significantly so. The random sites included in this analysis confirm that there are no significant differences in mutational load in any of the cases examined (Figure 4.6 top and bottom random panel).

Collectively, we observe that mutation induction around gene structure in response to UV damage is *significantly* enriched only at TES in wild-type cells, and at both in TSS and TES positions in MMR defective cells. In contrast, a minor depletion of UV-induced mutations is detected within ORFs, however this difference is not significant. Therefore, the *faster* relative repair rates of UV-induced DNA damage observed in the ORF's of wild-type cells results in the expected numbers of mutations found within ORFs as compared to the genome overall (Figure 4.7). These results demonstrate that the distribution of mutations within the genome varies in relation to genomic structures.

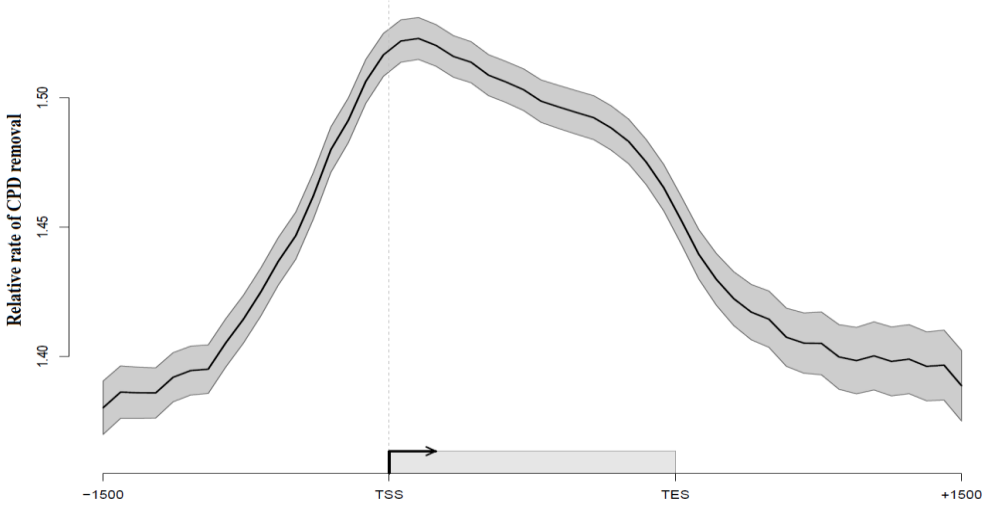


Figure 4.7: Relative rates of CPD repair in relation to ORF. Solid lines show the mean of relative CPD repair rates in wild-type (n = 3, black line) cells. Shaded areas indicate the standard deviation; with relative rates of CPD repair plotted as arbitrary units on the y-axis. (Data used here for plotting was generated by previous colleague in our laboratory, and ORF plot was made using Sandcastle (Bennett et al. 2015; Yu et al. 2016)). It is important to note that the assay (Teng et al. 2011; Powell et al. 2015) used measures the DNA damage and repair in cells on *both* strands of the DNA molecule, therefore representing the combined activity of the GG-NER and TC-NER pathways. In wild-type cells the relative repair rate reveals a uniform distribution in intergenic regions with enhanced rates of repair throughout the ORFs, which reflects the combined activity of GG-NER and TC-NER operating in these regions of the genome (Hu et al. 2015). Two points of inflection in the relative repair rates are observed at both TSS and TES regions of genes. These represent genomic regions where the relative rates of repair show a significant change.

4.3.3 Mutation induction in relation to other structural genomic features

DNA is wrapped around histone proteins, which form the basic unit of chromatin called the nucleosome (Lai and Pugh 2017). Within the linear structure of chromatin, repair is modulated by strongly positioned nucleosomes (SPN), which reportedly impairs accessibility of the repair complex to DNA lesions (Nag and Smerdon 2009; Rodriguez et al. 2015). In addition to the linear nucleosome structure, higher order chromatin structure such as Micro-C boundaries that define chromosomally interaction domains (CID's), also exists in yeast (Hsieh et al. 2015), and affect the relative repair rates (van Eijk et al. 2018).

Therefore, I decided to measure the mutation induction around nucleosome positions such as SPNs or nucleosome free regions (NFR), that are an important component of the linear chromatin structure (Lai and Pugh 2017), as well as higher order chromatin organisation, around Micro-C boundaries. Active regulatory regions within chromatin are typically histone-free and enable the binding of transcription-initiation factors and other regulatory proteins. Some well-known regulatory factor binding sites in yeast includes autonomously replicating sequence binding factor 1 (Abf1), DNA enhancer binding protein 1 (Reb1) and GG-NER complex binding sites (GCBSs). Abf1 is a general regulatory factor (GRF) in yeast, which displays multiple functions, for example, transcription, replication as well as GG-NER (Yu et al. 2009). On the other hand, the Reb1 protein, mainly involves transcription regulation, binding predominantly to the promoter and enhancer regions of rRNA transcription sites (Kasinathan et al. 2014). As mentioned in the previous chapter, GCBSs are the sites in the genome where GG-NER is initiated from (van Eijk et al. 2018). Two studies (Perera et al. 2016; Sabarinathan et al. 2016) recently showed that accumulation of mutations detected in melanoma cancers correlate with TF binding, or DNaseI hypersensitive sites (Perera et al. 2016; Sabarinathan et al. 2016). This work established that these regions of high mutational loads correspond to sites of slower repair as determined by XR-seq, another genome-wide DNA damage and repair assay (Adar et al. 2016). As a result, mutations accumulate at these sites. This higher mutational count could be explained by binding of transcription factors to promoters, which might impair NER and result in increased levels of mutations at active promoters. All of this work signifies the importance of repair organisation in the context of the linear genome and higher order chromatin structure. Building on these findings, here I analyse how genomic features that are important for repair organisation might help to explain the heterogenous distribution of mutations observed in cancer genomes (Salk et al. 2010).

After plotting the observed and expected mutation frequencies in relation to the linear structure of chromatin, we find no significant enrichment or depletion of mutations in relation to genome-wide NFRs, in both wild-type and *msh2* deleted cells, either in the presence or absence of UV damage (Figure 4.8, lower panel, NFR). When events are examined in relation to SPNs, similar observations are made. However, in *msh2* cells, higher mutational loads than expected are detected at SPN, and these become significantly enriched after UV exposure. This result demonstrates that strongly positioned nucleosomes in the genome do affect the stability of the genome causing higher levels of

mutation at this subset of nucleosomes in MMR defective cells treated and this becomes statistically significant with cells are exposed to UV radiation (Figure 4.8, lower panel, SPN). This result demonstrates that in wild type cells, the mismatch repair mechanism reduces the mutational load at strongly positioned nucleosomes, particularly after exposure to UV radiation. Importantly, observing mutational events in the context of all micro-C sites that exist at the boundaries of higher order chromatin structures known as CIDs, in this context significantly higher levels of mutations than expected are observed in wild-type UV damaged cells compared to undamaged cells (Figure 4.8, top panel, Micro-C Boundaries). This result shows that CID boundary regions are susceptible to the accumulation of UV-induced mutations in wild-type cells. Furthermore, in *msh2* cells significant enrichment of mutations is observed both before and after UV damage (Figure 4.8, top panel, Micro-C Boundaries). This result indicates that both defects in MMR and UV-induced DNA damage contributes to the accumulation of mutations at these boundaries of higher order chromatin structure. This suggests that Micro-C boundaries in the genome represent important sites affecting genome stability and could be sites from which DNA repair by a number of different DNA repair pathways is organised.

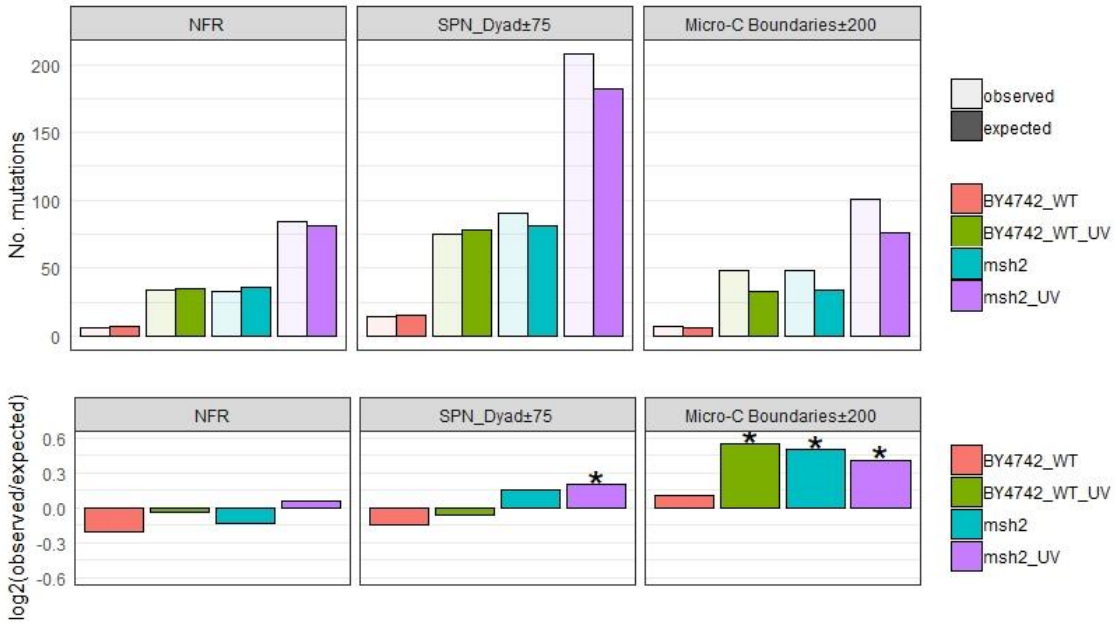


Figure 4.8: Enrichment and depletion of substitution mutations in chromatin associated genomic features such as NFR, SPN_Dyad±75bp, Micro-C boundaries ±200bp for both wild-type and MMR defective yeast cells with or without UV damage. The $\log_2$ ratio of the number of observed and expected point mutations indicates the effect size of the enrichment or depletion in each region. Asterisks indicate significant enrichments or depletions ($P < 0.05$, one-sided binomial test).

Finally, mutational loads when plotted in relation to Abf1, Reb1 and also GCBS binding motifs were examined. All these sites are occupied by two well-known general regulatory factors ABF1 and REB1, with GCBSs being the sites at which the Abf1-containing GG-NER complex occupies. In wild-type cells no significant mutational bias is observed around Abf1 or GCBSs, either before or after UV irradiation (Figure 4.9), indicating that mutational loads at these locations is similar to that expected throughout the genome. At Reb1 motifs, fewer mutations are observed than expected in the absence of UV damage, although this is not significant, but following UV damage, a significantly higher number of mutations than expected is observed at Reb1 binding sites (Figure 4.9), suggesting that these sites are particularly susceptible to damage-induced mutations. It is not clear why this is the case at this point. The opposite scenario is observed in the case of the *msh2* mutant; significant UV induced mutational bias is observed around Abf1 and GCBS summits, while no significant bias is detected around the Reb1 motif. These results demonstrate that MMR protects against the accumulation of UV-induced mutations at these sites in wild type cells, again suggesting the importance of these sites to the organisation of MMR in the genome.

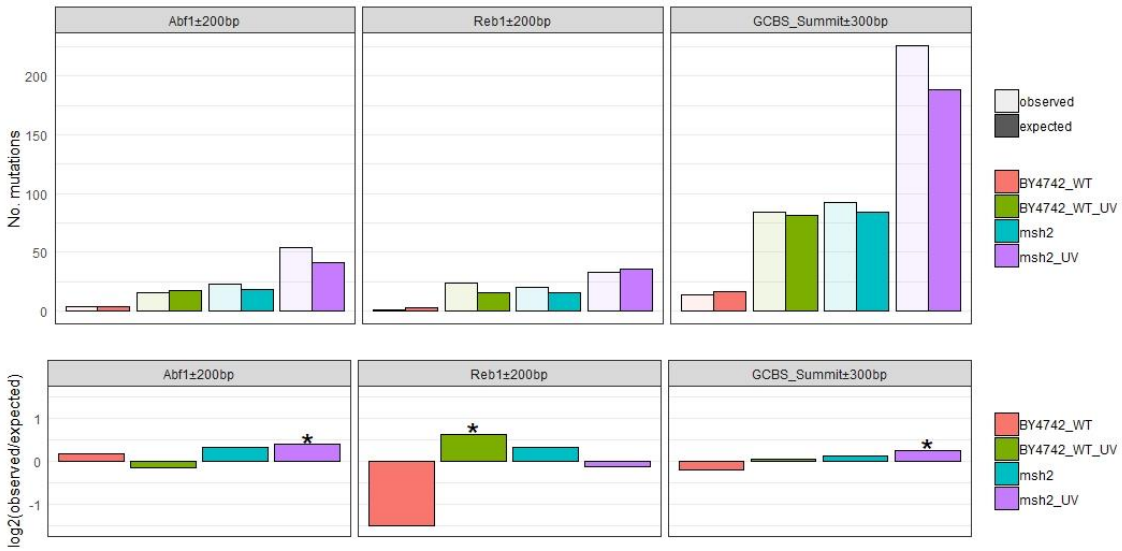


Figure 4.9: Enrichment and depletion of somatic point mutations in chromatin associated genomic features including Abf1_BS±200bp, Reb1_BS±200bp and GCBS_Summits±300 bp for both wild-type and MMR defective yeast cells, with or without UV damage. The $\log_2$ ratio of the number of observed and expected point mutations indicates the effect size of the enrichment or depletion in each region. Asterisks indicate significant enrichments or depletions ($P < 0.05$, one-sided binomial test).

To summarise, in response to DNA damage, the distribution of mutations varies with linear chromatin structure from nucleosome-depleted regions, to strongly positioned nucleosomes. Higher mutational density at Micro-C boundaries in both wild-type and *msh2* cells indicates the deficiency of organised repair at these sites affects the location of mutational hotspots in the genome. In addition to nucleosome structures, regulatory protein binding sites also alter UV-induced mutagenesis within the yeast genome, demonstrating the complex structure and organisation of biological processes that determine the UV-induced mutagenesis throughout the yeast genome.

4.3.4 Analysis of the different types of base substitution observed in cells

To study the total amount of base substitutions and their relative contribution to individual samples for the experiments performed, the contribution of the various types of base substitution are plotted according to the samples examined, either treated with or without UV damage. In total, 2,941 base substitutions were detected from both wild-type and *msh2* mutant cells, with or without exposure to UV damage. Overall, the *total* contribution of mutations is dominated by C>T followed by T>C, T>A, C>A, C>G and T>G types of base substitutions. The *relative* contribution of base substitutions to each of the different samples is shown in Figure 4.10.

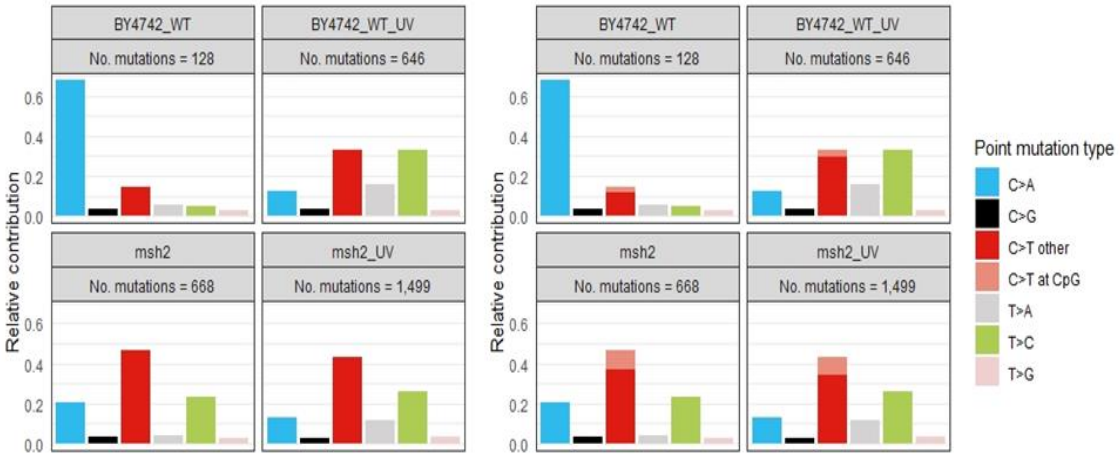


Figure 4.10: Relative contribution of each mutation type in the base substitution catalogues with sample types (left panel). Relative contribution of each mutation type in the base substitution catalogues with sample types and C>T at CpG island (right panel). Total number of mutations in each sample group is also noted above each bar chart.

The mutational pattern of untreated wild-type cells shows a predominant contribution of the C>A type of mutation. In contrast, under similar conditions, C>T mutations are more prevalent in untreated *msh2* cells. The characteristic pattern of substitution types found in

msh2 cells are C>T and T>C followed by C>A substitution (Figure 4.10). This is in line with a recently published study that validates the concept of mutational signatures using a human cell line or the *C. elegans* model organism (Meier et al. 2018; Zou et al. 2018). At the same time, characteristic predominance of C>T and T>C transitions followed by T>A and C>A transversion mutations are found in both wild-type and *msh2* cell lines exposed to UV radiation (Figure 4.10). This is due to the combined effect of UV damage and lack of MMR (Figure 4.10).

In addition, the mutation spectrum with distinction between C>T at CpG sites and other sites is also shown. As can be seen C>T mutations at CpG sites are relatively infrequent in comparison to overall C>T mutations. This is likely caused by a lower level of methylation of cytosine at these sites in yeast. This phenomenon is known to be one of the major contributors to UV-induced C>T type of mutations observed at mCpG sites in human cancer (Poulos et al. 2017) (Figure 4.10, right panel).

4.3.5 GCBSs are significantly enriched for certain types of UV-induced mutation in MMR defective cells

In Chapter 3, we showed that UV damage-induced nucleosome remodelling by the GG-NER complex is initiated from hundreds of genomic locations where the complex is bound. These sites exist predominantly at the nucleosome free regions (NFR) of gene promoters and are referred to as GG-NER complex binding sites (GCBS's). The results revealed that distinct functions of the GG-NER complex contribute differentially to the establishment of GG-NER complex occupancy at GCBS's in the absence of damage, and to its redistribution following UV irradiation. This study also showed that, these sites are frequently located at genomic boundaries or domains that delineate chromosomally interacting domains (CIDs), also known as Micro-C boundaries (Hsieh et al. 2015; van Eijk et al. 2018). However, not all NFR regions are GCBSs. NFRs can also be identified at other regions of the genome that are not bound by the GG-NER complex, and therefore these sites can be used as a negative control. Since repair is not organised in relation to these positions (van Eijk et al. 2018), it is conceivable that mutations may be enriched at these sites, and depleted GCBSs. Counting the total number of mutations in wild-type cells at these two genomic locations, reveals that none of the different types of base substitutions are significantly enriched at GCBSs, either before or after UV damage. This result is consistent with results reported in Figure 4.9 earlier in this chapter. Significantly higher levels of four types of UV-induced base substitutions are observed in MMR defective cells (Figure 4.11). This observation is also consistent with earlier results,

confirming the importance of the MMR pathway to specifically protecting GCBSs against the accumulation of UV-induced mutations. This could indicate that GG-NER and MMR function co-operatively at GCBSs.

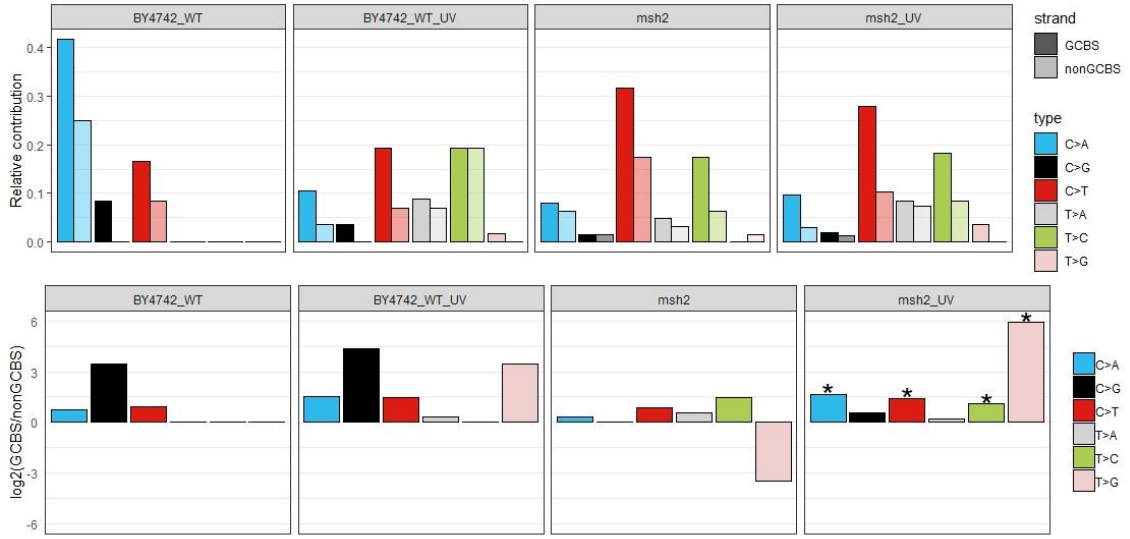


Figure 4.11: Relative contribution of the 6 base substitution types at GCBSs (dark shaded) and non-GCBSs NFRs (light shaded) sites. Log_2 ratio of the number of mutations on GCBS and non-GCBS NFR regions per indicated base substitution for each of the samples listed is shown in the lower section of this figure. The log_2 ratio indicates the extent of the bias, and asterisks (*) indicate significant regional asymmetries ($P<0.05$, one-sided binomial test).

4.3.6 Distribution of mutations in relation to 'open' and 'closed' chromatin based on histone H3 acetylation status

Organisation of chromatin plays an important role in the regional mutational density in human cancer (Schuster-Böckler and Lehner 2012). The mutational density is lower at positions in the chromatin the are marked as 'open', while a higher mutational load is associated with heterochromatic marks that signify 'closed' or compact chromatin (Polak et al. 2014). It is suggested that open chromatin confirmation is less likely to acquire mutations due to increased accessibility of the chromatin to the DNA repair factors leading to more efficient repair of DNA damage. It is well established that, UV-induced histone H3 acetylation (H3-Ac) is required for efficient GG-NER after damage induction (Yu et al. 2016). As described above for mutations detected in cancer, histone H3 acetylation and chromatin density might also affect the mutation distribution within chromatin. To study this, we classified high and low acetylated regions from previous ChIP-chip analysis (Yu et al .2016). We considered the distribution of mutations in high

and low acetylated regions within the genome as a surrogate marker for open and closed chromatin respectively. Plotting the mutational load according to these features showed that endogenous mutations in untreated wild-type cells are not distributed significantly differently. However, in response to DNA damage and/or deletion of *MSH2*, the distribution of various different types of base substitutions is significantly biased towards low acetylated (or closed) regions of the genome (Figure 4.12).

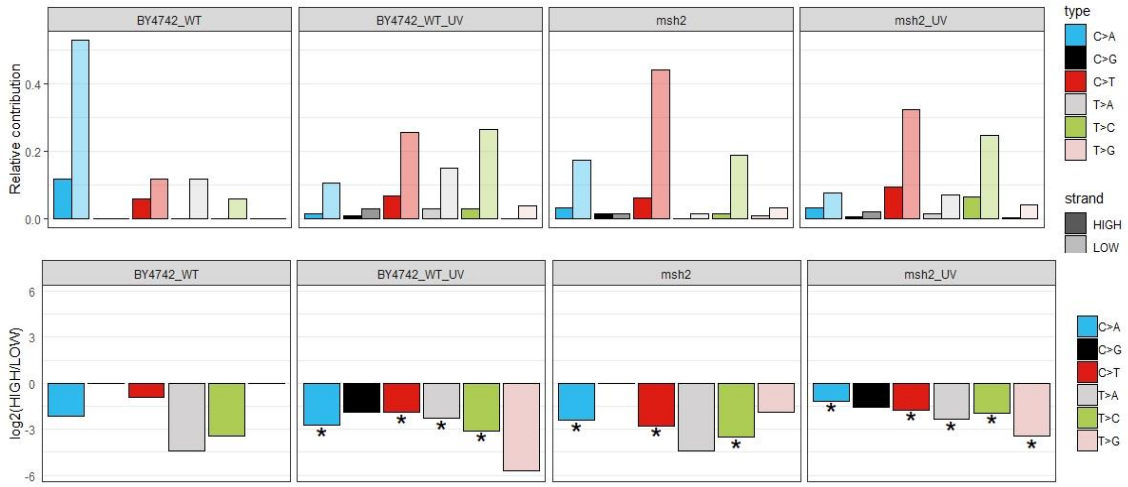


Figure 4.12: Significant mutational distribution bias towards high and low acetylated regions within yeast genome. Log_2 ratio of the number of mutations on the high and low acetylated regions per indicated base substitution of each samples shown in lower section of this figure. The log_2 ratio indicates the effect of the bias and asterisks (*) indicate significant acetylation region asymmetries ($P<0.05$, one-sided binomial test).

4.3.7 UV damage and defective MMR causes increased mutations in late replicating regions of the genome

Several studies showed that the heterogeneity in the distribution of mutations observed in human and cell-based assays, is due to differences in the early and late replicating regions of the genome (Woo and Li 2012). Higher rates of mutation in late replicating regions of the genome have also been observed in several cancers (Donley and Thayer 2013), indicating that replication timing has an important influence on distribution of mutations within genome. The involvement of replication-associated mechanisms in mutational heterogeneity can be evaluated by testing for a mutational bias towards early or late replicating regions in the genome. Repli-seq experiments make it possible to precisely map the genome-wide replication timing after release from alpha factor arrest during the yeast cell cycle (Muller et al. 2014). We used the datasets from these experiments to annotate the yeast genome, and assign early versus late replicating regions at a resolution

of 5 kb. Plotting the UV damage repair rate data at the centre of late replicating regions within the yeast genome, shows that in both wild-type (black line) and GG-NER defective (red line) yeast cells, the relative repair rate is mostly affected during late regions within yeast genome (Figure 4.13). This observation suggests that replication timing might also modulate the genome wide distribution of mutations.

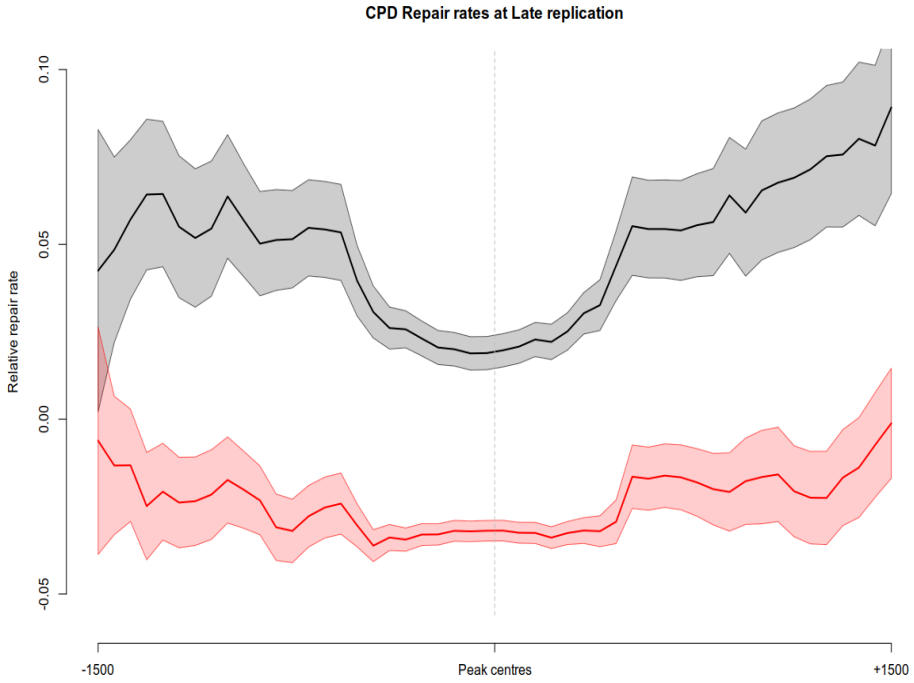


Figure 4.13: Relative CPD repair rate from at the centres of the late replicating regions within yeast genome. Solid lines show the mean of CPD repair rates in wild-type (n = 3, black line), Rad16 cells (n = 2, red line). Shaded areas indicate the standard deviation, with CPD levels plotted as arbitrary units on the y-axis. The repair rate data plotted here was obtained from (Yu et al. 2016) and the replication timing data was obtained from (Muller et al. 2014) (see the Material and Methods for details).

Mapping the mutational distribution in the early and late replicating regions within the yeast genome showed no significant bias in mutational distribution in untreated wild-type cells (Figure 4.14). However, in response to UV DNA damage, there is a striking change in the relative contribution of mutational types, and most of the mutational load displays a significant bias towards late replicating regions within the yeast genome (Figure 4.14). This observation suggests that, in line with observation made in several cancers, including melanoma (Donley and Thayer 2013), that higher levels of mutations are observed in late replicating regions of the genome for multiple different types of base substitution.

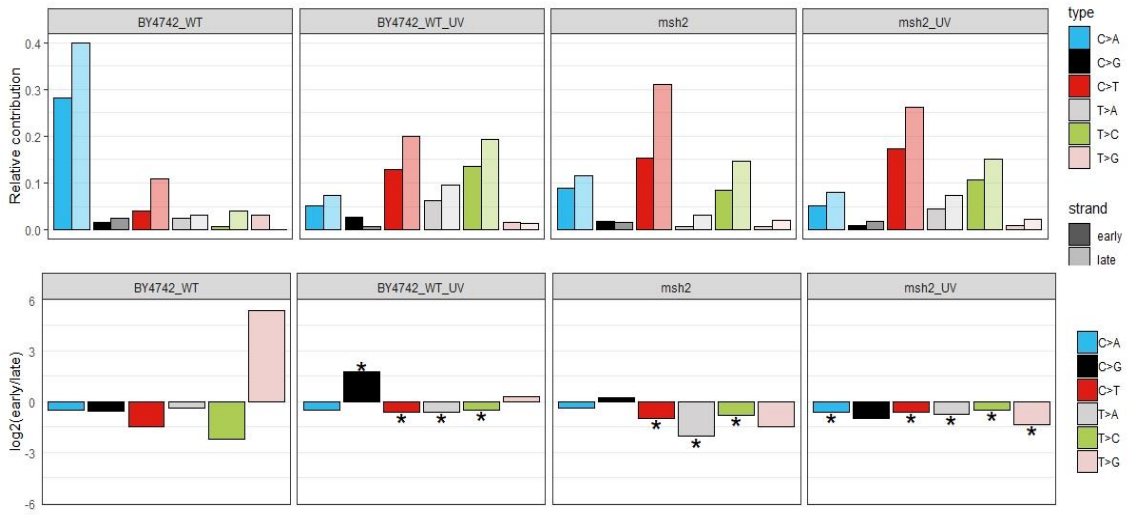


Figure 4.14: Mutational distribution with replication timing information. The upper panel shows the relative contribution of each nucleotide changes is subdivided into early, (dark shades), and late (light shades). Log_2 ratio of the number of mutations on the early and late replicative regions per indicated base substitution of each samples shown in lower section of this figure. The log_2 ratio indicates the effect size of the bias and asterisks (*) indicate significant replicative timing asymmetries ($P<0.05$, one-sided binomial test).

In the case of *msh2* mutant yeast cells, the relative distribution of mutation types is similar in both undamaged and damaged cells (Figure 4.14). Although the absolute mutational load is higher in UV-damaged MMR deficient cells, both UV damaged and undamaged *msh2* mutant cells show a significant mutational bias towards late replicating regions (Figure 4.14). This observation confirms that the significant mutational distribution towards late replicative regions within yeast genome are associated with error-prone repair and post replicative MMR deficiency (Lang and Murray 2011; Lujan et al. 2014).

4.3.8 Distribution of mutations in relation to transcriptional strand bias

In cancer genomes, somatic mutations exhibit transcriptional strand asymmetry in their density along the genome (Lawrence et al. 2013; Haradhvala et al. 2016). The repair rates in response to DNA damage in wild-type cells are also known to vary, with higher rates of repair observed over open reading frames (Figure 4.7). This is a result of the combined activity of both TC-NER and GG-NER operating in these regions (Yu et al. 2016). To determine whether the mutational catalogue obtained in this study shows a transcriptional strand bias, we examined events in both UV damaged and undamaged wild-type and *msh2* mutant yeast cells (Figure 4.15).

In untreated wild-type cells, there is no significant biases in the various types of substitution mutation between the transcribed and untranscribed strands, except for C>A type substitutions that show a significant bias towards the transcribed strand. However, in UV treated wild type cells there is a general shift towards a mutational bias in the non-transcribed strand, with a significant bias found in the case of C>T and T>C types of mutations (Figure 4.15). This observation is consistent with the combined activity of both GG-NER and TC-NER rapidly removing damage from the transcribed strands of genes, resulting in fewer mutations in this strand. Surprisingly, T>A types of mutations are significantly enriched in the transcribed strand of genes (Figure 4.15). This novel observation could indicate that the frequent UV-induced TT CPD has different repair kinetics during TC-NER and GG-NER for this particular type of lesion in actively transcribed genes (Figure 4.15). Alternatively, it might be the result of a different type of UV-induced lesion that is either over-represented, or difficult to repair in transcribed strands.

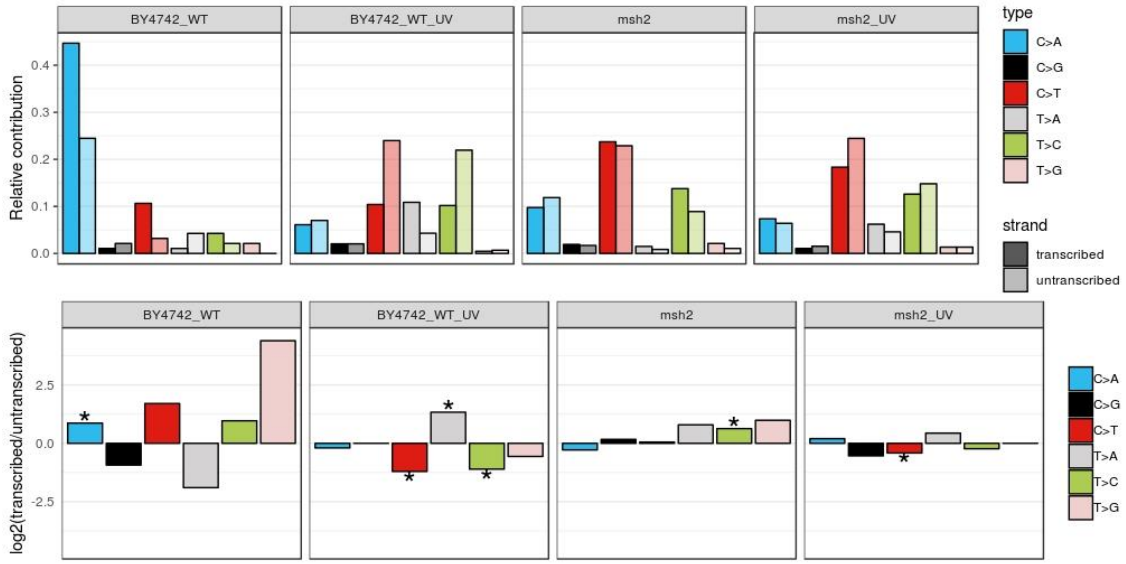


Figure 4.15: Mutational distribution with transcriptional strand information. The upper panel shows the relative contribution of each nucleotide changes is subdivided into either transcribed (dark shades) and untranscribed (light shades). Log_2 ratio of the number of mutations on the transcribed and untranscribed strand per indicated base substitution of each samples shown in lower section of this figure. The log_2 ratio indicates the effect of the bias and asterisks (*) indicate significant transcriptional strand asymmetries ($P<0.05$, two-sided binomial test or Poisson test).

In the case of *MSH2* deleted cells, the relative contribution of mutation types is very similar in both undamaged and UV-damaged cells with little evidence of transcriptional

strand asymmetry. As might be expected, defective post-replicative MMR does not result in mutational strand asymmetry with respect to transcriptional status. Some strand bias is observed in the case of UV-induced C>T and unirradiated T>C mutations. However, it is impossible to properly distinguish between UV-induced and replication-induced C>T and T>C mutations at this level of analysis. It is more likely that the higher level of these two mutation types is induced during replication, because they do not display a consistent strand asymmetry as would be expected if they were induced by UV damage.

4.3.9 Mutational spectrum analysis

In order to study the biological processes of genomic instability observed in cancer, it is important to understand the mutational density, distribution and spectrum of mutations (Pleasance et al. 2010a). In cancer genomes, mutational heterogeneity is observed in the mutational load along the cancer genome as well as in the mutation spectra found within different tumours (Lawrence et al. 2013). To study the spectrum of substitution mutations, the types of mutations can be plotted according to the linear chromosomes as rainfall plots and the 96 trinucleotides mutation context plot.

4.3.10 Rainfall plots

To study the genome-wide distribution of acquired substitutions, rainfall plots were generated to visualise the mutational types and the genome-wide inter-mutational distance. This plot was originally used to understand the localised hypermutations observed in breast and certain other cancer types (Nik-Zainal et al. 2012), also called "Kataegis". The rainfall plots of wild-type cells, without or with exposure to UV lesions are plotted in figure 4.16 and figure 4.17, respectively. The mutations appear evenly distributed throughout the genome. In unirradiated wild-type cells this distribution is dominated by C>A mutations (Figure 4.16), as previously noted (Figure 4.10). This might be due to endogenous damages during cell replication, and is consistent with findings from human cell model-based analysis (Zou et al. 2018). However, in UV-damaged cells, the distribution of mutations is dominated by both C>T and T>C mutations followed by T>A and C>A types of substitutions (Figure 4.17), indicating the effect of UV damage on the mutational type and providing evidence for the UV-induced mutational processes (Pfeifer et al. 2005; Ikehata and Ono 2011).

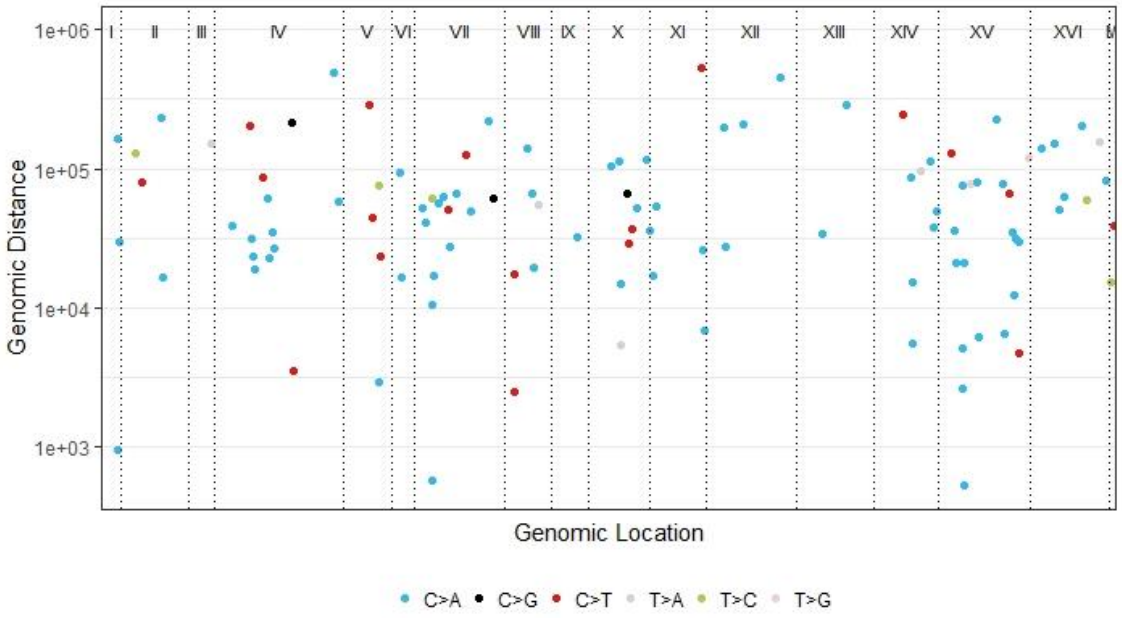


Figure 4.16: Rainfall plot of wild-type cells showing the genomic location of mutation with their inter-mutational distance. The coloured dots represent 6 possible types of substitutions.

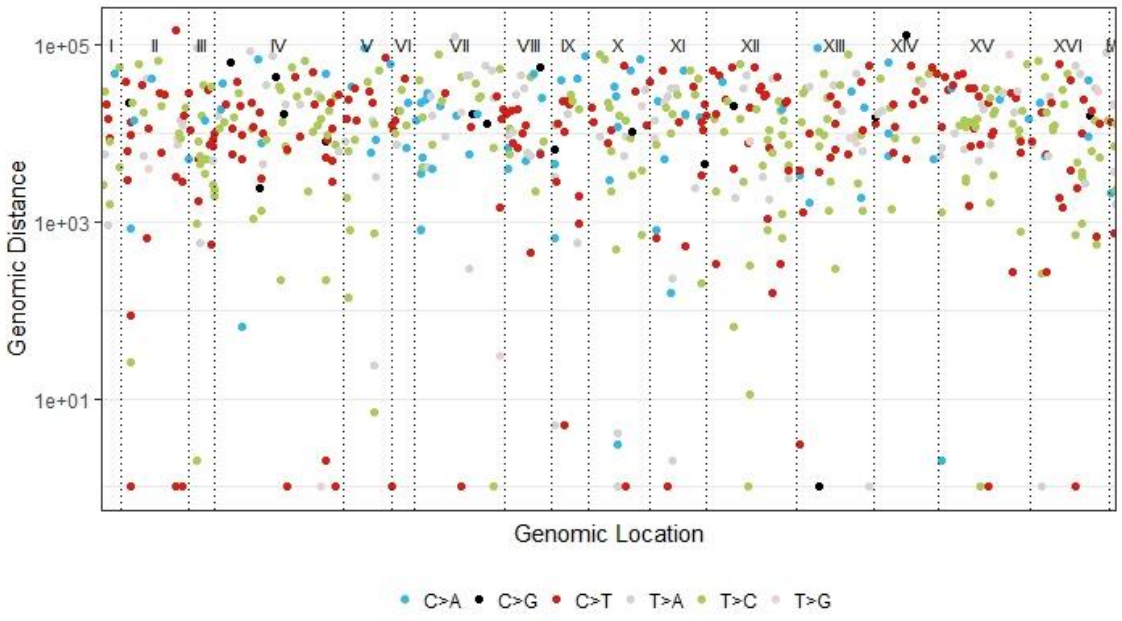


Figure 4.17: Rainfall plot of wild-type UV-exposed cells showing the genomic location of mutation with their inter-mutational distance. The coloured dots represent the 6 possible types of substitutions.

The rainfall plots of *msh2* cells, without or with exposure to UV lesions are plotted in figure 4.18 and figure 4.19, respectively. Comparing figure 4.16 (rainfall plot of

undamaged wild-type) and 4.18 (rainfall plot of undamaged *msh2* mutant) depicts the effect of MMR deficiency on the mutational distribution. Like wild-type cells, there is an apparently even distribution of mutations in *msh2* mutants, with no evidence of Kataegis. However, the mutational load is strikingly different from wild-type cells even without exposure to UV damage. The mutational pattern of *msh2* cells is dominated by C>T and T>C types of transitions, followed by the C>A transversion. This result is in line with the previously examined MMR defective *C. elegans* or human cell line-based mutational patterns studied previously (Lujan et al. 2014; Meier et al. 2014; Meier et al. 2018; Zou et al. 2018), suggesting that the main mutational processes active in MMR defective cells is likely conserved between different organisms.

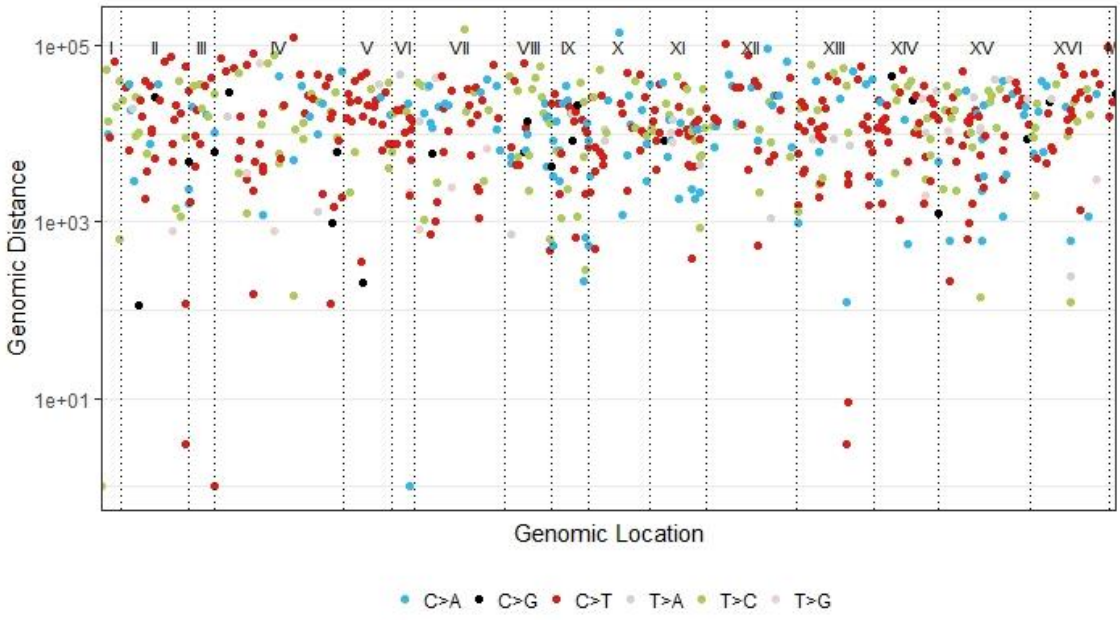


Figure 4.18: Rainfall plot of *msh2* cells showing the genomic location of mutation within yeast genome with their international distance. The coloured dots represent 6 possible types of substitutions.

The rainfall plots of *msh2* cells after exposure to UV radiation indicate that the mutational pattern is also dominated by C>T, T>C types of substitutions, followed by T>A and also C>A types of transversion. The overall mutational load is higher compared to that detected in undamaged cells, indicating the additional effect of UV damages on the induction and distribution of mutations (Figure 4.19).

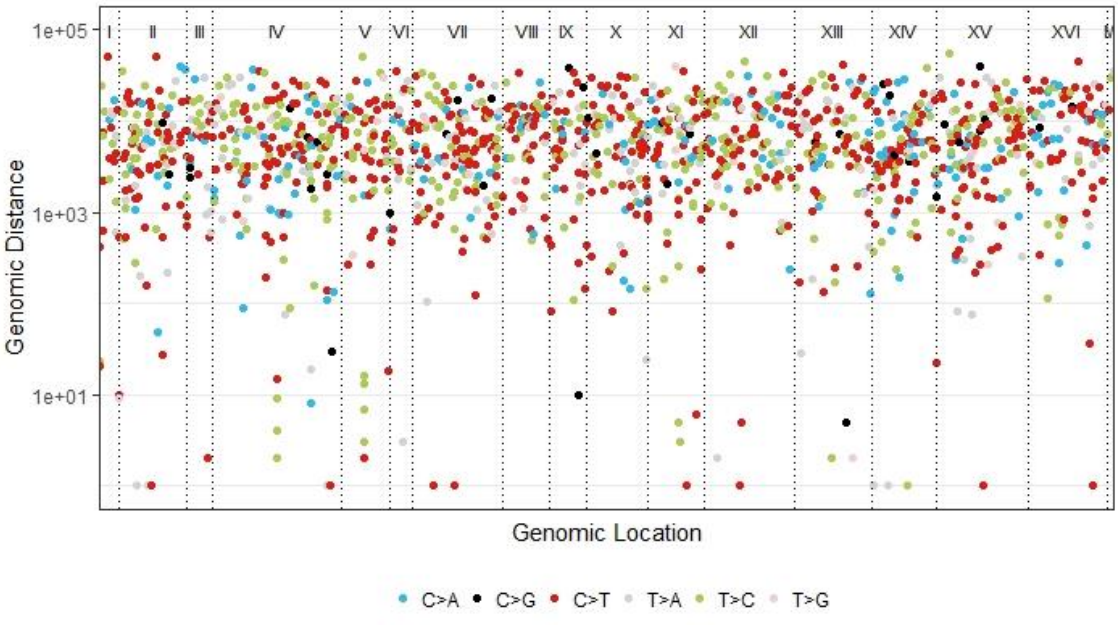


Figure 4.19: Rainfall plot of *msh2* cells treated with UV radiation showing the genomic location of mutations within the yeast genome, with their inter-mutational distance and the types of mutation also illustrated.

4.3.11 Substitution mutation spectra as illustrated by the 96 trinucleotide mutation subtypes

From the study of whole genome cancer data, mutational signatures, were extracted from the catalogue of somatic mutations that, strikingly, revealed the mutational processes that give rise to mutation during the development of cancer (Alexandrov et al. 2013b). The biological processes of some of these mutational signatures, that are informative about the cumulative processes of DNA damage and repair, are known, but most remain unknown (Alexandrov et al. 2018). Performing controlled biological experiments will in time reveal causes of the cryptic mutational signatures extracted from human cancer genomes, and provide clues about the biological processes of DNA damage and repair that generated them in the first place (Alexandrov et al. 2013b). Signatures derived from controlled experiments can be compared with existing unknown cancer signatures using quantitative tools such as a cosine similarity test. The cosine similarity expresses the similarity between two large vectors. In order to test the extent of cosine similarity with the recently reported PCAWG mutational signatures, the 96 trinucleotide mutational profile output is required from both our wild-type and MMR defective *msh2* mutant yeast cells. This novel way of representing substitution mutations is achieved by considering the bases immediately 5´ and 3´ to each mutated base, creating a sequence context. This

generates a 96 tri-nucleotide mutation context or profile (6 types of substitution ∗ 4 possible types of 5' base ∗ 4 possible types of 3' base) that can be quantitatively analysed (Alexandrov et al. 2013b). Mapping all the possible substitutions as their 96 trinucleotide mutational profile in this way, reveals both the types, and the context of the mutations studied here. Plotting substitution mutations in this way reveals that the types of substitutions vary greatly between different sample types and depends on the repair processes, as well as the nature of mutagen exposure (Figure 4.20). In agreement with our previous findings, changes in the relative number of substitution type, now in their trinucleotide context (as opposed to the 6 base substitution types at a single nucleotide position), can be observed. In wild-type cells, major changes are observed in C>A substitutions at N<u>C</u>N (mutated base underlined) sites along with fewer other types of substitutions (Figure 4.20). In response to UV-induced damages, these patterns change towards C>T @ T<u>C</u>N, C>T @ C<u>C</u>N followed by T>A @ T<u>T</u>A or A<u>T</u>A and T>C @ T<u>T</u>N, CTA & G<u>T</u>A (mutated bases underlined).

In the case of MMR deficient cells, the mutational context around C>T and C>A are dominated by N<u>C</u>N and T>C at N<u>T</u>N trinucleotides. In response to UV irradiation, the MMR deficient cells also display UV-induced mutational patterns observed in wild-type cells in addition to the characteristic MMR substitution patterns. These changes in mutation pattern are due to the combined effect of both UV damage and defective MMR.

Plotting the substitution mutations according to the 96 possible trinucleotide context, has important advantages. It provides the possibility to distinguish between similar types of mutation such as C>T or T>C substitutions in wild-type and *msh2* mutants cells, are generated in different sequence contexts (Figure 4.20). Importantly, this representation allows the use of NMF that enables the distinction between the underlying signatures that make up the mutational profile as displayed in Figure 4.20 (see next section for further details). These so-called meta-signatures have the potential to reveal information about the processes that contributed to the formation of the mutational catalogues observed in these cells.

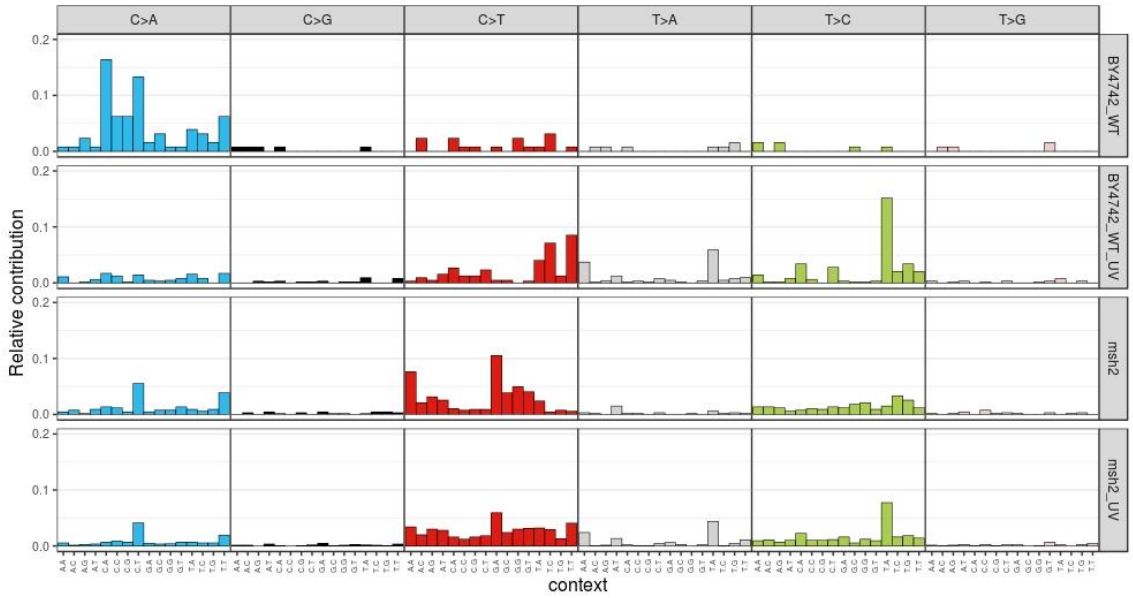


Figure 4.20: Profiles of 96 trinucleotide mutation contexts in wild-type and MMR defective yeast cells, with or without exposure to UV damage. The types of mutations are in different colours. The X-axis indicates the 5´ and 3´ base context of the mutated base. The relative contribution of each of the six types of base substitutions are displayed on Y-axis.

4.3.12 The 96 trinucleotide mutational profile of base substitutions derived from yeast cells shows similarity to the PCAWG mutational signature profiles

In an attempt to extract the mutational signatures derived from the biological processes active and generating mutations in our experimental system (e.g. mutagen exposure, repair deficiency, replication function, etc), we calculated the cosine similarity between the 96 trinucleotides mutational profile of the samples described here and compared them to the PCAWG signature profile. At a glance, figure 4.21 illustrates the heatmap representation of the cosine similarity of mutational profiles between wild-type and *msh2* mutant samples with or without UV damages, and the recently reported mutational signatures obtained from cancer genomes. Strikingly, none of the mutational patterns match perfectly to the published mutational signatures that were extracted from human cancer genomes. This is probably due to the different genetic background between yeast and human cells, as wells as differences in the biological processes occurring in the two different biological systems used. Under investigation here are only the mutations induced during the experimental time period described, exclusively reflecting the mutational pattern induced by a highly controlled set of biological processes. Cases of sporadic cancer, on the other hand, are highly heterogeneous, and caused by a bewildering

array of complex exposures to mutagens and fluctuations in DNA repair capacity that are unique to individuals. The controlled biological processes in this study are: (i) endogenous processes during replication (wild-type), (ii) endogenous and UV damage induced mutational processes (wild-type UV treatment), (iii) endogenous and MMR deficiency (*msh2*) and (iv) endogenous, MMR deficiency and UV damage induced mutational process (msh2 UV treatment).

In the case of wild-type yeast cells, the mutational profile shows high similarity with PCAWG single base substitution signatures(SBSs), SBS4 and SBS38 followed by SBS18, SBS20, SBS8, SBS14, SBS24, SBS29, SBS35, SBS36 and SBS40 (Figure 4.21). These signatures were attributed to various types of endogenous and indirect effect of exogenous oxidative damaging agents such as tobacco smoke, reactive oxygen species, indirect effect of UV light, MMR deficiency and so on (Pfeifer et al. 2002; De Bont and van Larebeke 2004; Alexandrov et al. 2018). This indicates that the mutational process during cell division induced by endogenous oxidative species introduced by cellular respiration is active in our system. This finding correlates well with the recently published mutational signature validation study, based on a HAP1 cell model (Zou et al. 2018).

In wild-type cells after exposure to UV damages, the resulting mutational profile displays similarity with PCAWG signatures, SBS2, SBS3, SBS5. SBS7a & SBSb, SBS11, SBS12, SBS40 and SBS41 (Figure 4.21). Most of these signatures are attributed to activity of the APOBEC family of cytidine deaminases (which convert cytosine bases (C) to uracil (U), result in C>T mutations), UV light exposure, BER and some unknown processes (SBS12 and SBS41) (Alexandrov et al. 2018). These observation reveal that the UV-induced mutational pattern described here, has only low similarity with UV-related PCAWG signatures found in melanoma cancer genomes. We anticipated that the exposure to UV light in malignant melanoma and UV irradiated yeast cells would be more similar. However, this is not the case, and this could be due to the absence of cytosine methylation in yeast at CpG islands, as noted in figure 4.10. This unique mutational pattern in this controlled experimental system is the final output of the yeast cells' repair capacity in response to UV damage over time that is likely different from that which occurs in patient melanoma cells. In addition, *msh2* cells display a mutational pattern that has the highest similarity with PCAWG signature SBS44 followed by SBS26, SBS21, SBS20, SBS15, SBS14, and SBS6 in the mutational pattern (Figure 4.21), These have been reported in the journal of 'The Repertoire of Mutational Signatures in Human Cancer', and is proposed to be associated with MMR deficiency in cancer. This confirms

that the biological processes generating mutational signatures due to MMR defects observed in cancer genomes are conserved between human and yeast cells.

After UV irradiation, the mutational profile of *msh2* mutant cells, displayed a pattern similar to the combined pattern of UV damaged wild-type cells and *MSH2* deleted cells, indicating the combined effect of defective MMR and UV damages. This result is mirrored when the cosine similarity between all 4 biological samples and the 49 PCAWG signatures are represented as a clustered heatmap as shown below (Figure 4.21).

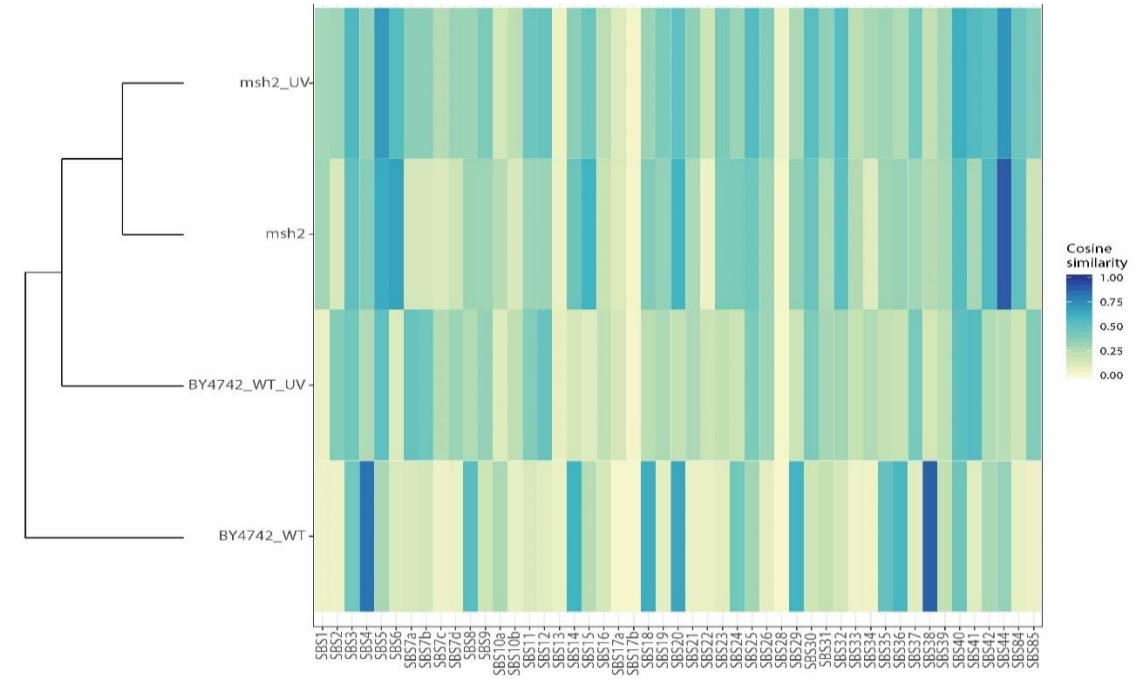


Figure 4.21: Heatmap shows the cosine similarity between the mutational profile of both UV damaged or undamaged wild-type and MMR deficient cells with PCAWG signature profile.

Collectively, the cosine similarity between the mutational profile of experimentally induced substitutions in a controlled biological system and the PCAWG mutational signatures derived from cancer genomes, provide a first proof of principle that this workflow is able to generate meaningful mutation signature data and the generation of meta-signatures that allow the analysis of controlled mutation induction. In the future, this could provide important insight into the processes that give rise to the mutational signatures found in human cancers.

4.3.13 Identification of the PCAWG signatures that can most accurately reconstruct the 96 mutational profile signatures observed within the experimental samples

An alternative way to illustrate the relationship between the experimentally generated mutational profiles to the known and published mutational signatures, is to calculate the contribution of the known (e.g. PCAWG), or user-specified (*de novo* extracted) mutational signatures to the mutational profiles from the individual samples and then to derive a set of minimal PCAWG signatures that can sufficiently reproduce the mutational profile detected in our biological system. This can be achieved by using a protocol called non-negative least square (NNLS) from the 'pracma' package, which is integrated within the MutationalPatterns package written in R (Blokzijl et al. 2018). This result allows the study of the minimal set of previously identified mutational signatures in wild-type and DNA repair defective cells and helps establish the extent of overlap between the investigated molecular processes underlying the mutational signatures in our yeast system, and the ones that are active and gave rise to the cancer genome mutational signatures.

Interestingly, the mutational catalogue of substitutions derived from yeast cells can be reconstructed predominantly by PCAWG SBS38, SBS4 and SBS20. The PCAWG SBS38 is attributed to indirect effect of UV light, SBS4 is attributed to tobacco smoking and SBS20 is attributed to concurrent DNA polymerase Delta-1 (the gene that encode the catalytic subunit of Pol δ) mutations and mismatch repair deficiency (Alexandrov et al. 2018). The aetiology proposed for the damage induced by these signatures is also quite similar to the types of damage induced by endogenous processes during cellular respiration. Therefore, the mutational signatures generated by these biological processes, contributed mostly to reconstruct the mutational profile of unirradiated wild-type cells. In response to UV damage, the landscape of mutations in wild-type cells can be reconstructed best by a combination of SBS41, SBS12 and SBS7b (Figure 4.22). The proposed aetiology of PCAWG SBS7b is UV light exposure, whereas the causes of SBS41 and SBS12 signatures are unknown at this point (Alexandrov et al. 2018). Additionally, the mutational landscape of unirradiated *msh2* can be reconstructed with a predominant contribution of SBS44. On the other hand, mutations derived from irradiated *msh2* (*msh2_UV)* can be reconstructed with contributions of SBS44, SBS41, SBS12 and SBS7b. The proposed aetiology of cryptic SBS44 is defective DNA mismatch repair (Alexandrov et al. 2018), which is consistent with the results of the experiments described here.

Importantly, the minimal set of PCAWG signatures that contribute more than 10% to the reconstructed profile of UV exposed *msh2* mutants, are a combination of those required for the reconstruction of the UV exposed wild type and *msh2* untreated mutant cells (Figure 4.22). This confirms the additive nature of both mutagenic processes active in the yeast strains studied here. However, note that not all PCAWG signatures, that originally showed similarity to the *de novo* extracted mutation profiles (Figure 4.21), are required to reconstruct the samples mutational profile. This is because PCAWG mutational signatures are not independent and actually showed some cosine similarities amongst each other (Figure A3.5 in Appendix). More importantly, these findings underpin the strength of the comparative analysis between PCAWG signatures and mutational profiles derived from defined mutagenic processes in an *in vivo* model-based workflow.

Figure 4.22: The optimal relative contribution of PCAWG signatures to reconstruct the mutational profiles of the wild-type and *msh2* mutant samples with or without UV damage. The PCAWG signatures with ≥ 10% contribution in at least one of the experimental samples are plotted.

Using the relative contributions shown in Figure 4.22, this information can now be used to generate a reconstruction profile that mimics the original mutational profile for each of the 4 experimental conditions described in this study. The results of this analysis are shown in Figure 4.23. In the top panel the original mutation profile of UV-treated wild-type cells is shown, while the middle panel depicts the reconstructed profile using contributions from the minimum of the known, and mostly similar PCAWG signatures. The difference between these profiles is expressed as the $\log_2$ of their ratio, and displayed at the bottom of Figure 4.23. This difference calculation identifies that certain C>T and T>C substitution are either over- or under-represented in the reconstructed profile. Comparison of the original with the reconstructed mutational profile reveals which

trinucleotide peaks cannot be reconstructed with the given signatures (Figure 4.23) and provides important insight into the mutational mechanisms active in the system studied, that cannot be accounted for by the signatures from PCAWG. It is possible that these difference are the result of the different genetics of our model system, or due to the use of UV-C as the mutagen, as opposed to UV-A and UV-B exposure typical for skin cancers (Pfeifer et al. 2005).

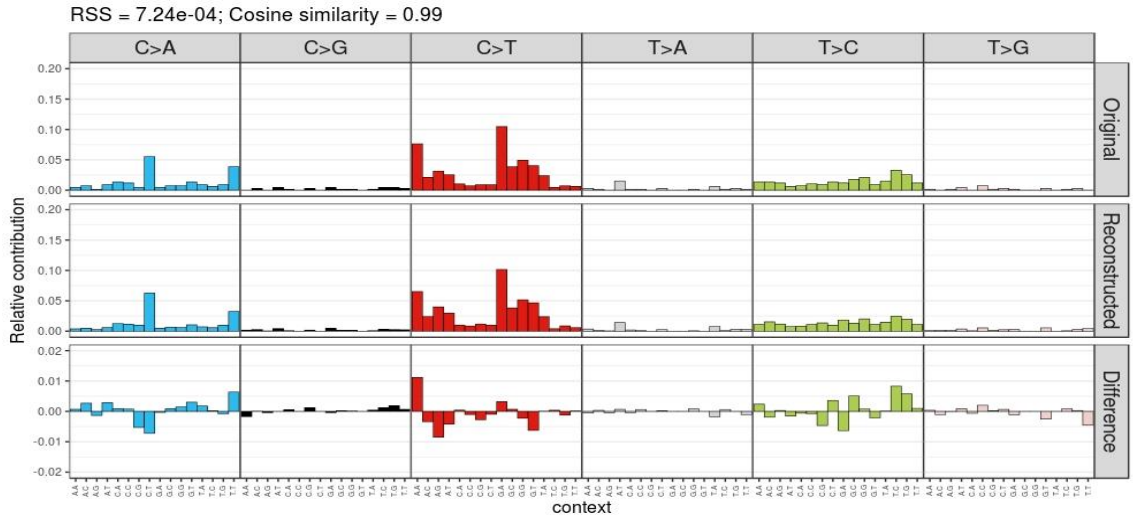


Figure 4.23: Relative contribution of each of the 96 trinucleotide changes to the original mutational profile (upper panel) and the reconstructed mutational profile (middle panel), and the difference between these profiles (lower panel) for the wild-type after UV damage. The residual sum of squares (RSS) and the cosine similarity between the original and the reconstructed mutational profile are indicated.

To test how well each experimental samples' mutational profile can be reconstructed by the PCAWG mutational signatures, the cosine similarity was calculated between the original and the reconstructed mutational profile for all the samples. A low similarity between the original and the reconstructed profile indicates that the analysed mutational profile cannot be fully explained by the provided PCAWG signatures. The mutational profiles of most samples cannot be reconstructed precisely with the PCAWG signatures ($\alpha < 0.95$, Figure 4.24), while the wild-type can be reconstructed with high accuracy ($\alpha > 0.95$, Figure 4.24). As mentioned above, this suggests that unassessed mutational processes in our model system might underlie the observed catalogue of somatic mutations, which cannot be faithfully reproduced with the PCAWG signatures.

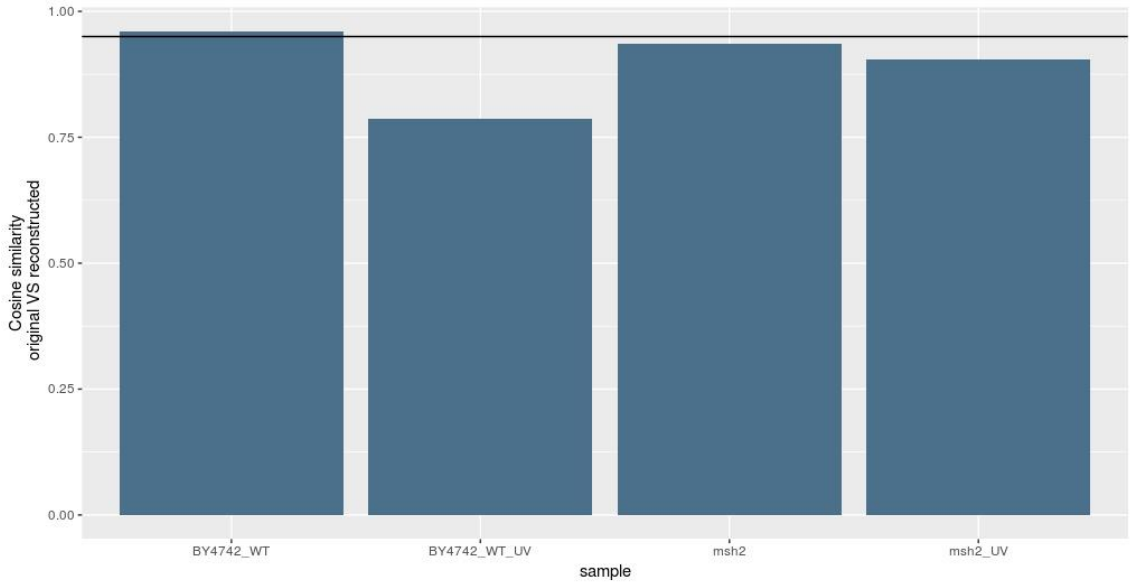


Figure 4.24: Cosine similarity between the original mutational profile and the reconstructed mutational profile based on the optimal linear contribution of all 49 PCAWG signatures. The line indicates the threshold of cosine similarity = 0.95.

4.3.14 *De Novo* Mutational Signature extraction from the controlled yeast-based system using NMF

The mutational signatures obtained from cancer genomes are derived from the pattern of mutations that arises because of the combined effect of DNA damage and repair processes that occur in the cells of patients during the life-time of the individual. These signatures represent the mutational processes and are characterised by their contribution to the 96 tri-nucleotide mutation contexts. Alexandrov and colleagues (2012) initially used the mutation count matrix and then applied the non-negative matrix factorisation (NMF) algorithm to extract mutational signatures from the catalogue of somatic mutations found in cancer genomes. The fundamental principle of this approach is to extract the minimal set of components that can most accurately regenerate the original heterogeneous mutational profile obtained from different cancers. These components or signatures are representative of the biological processes active in a cohort of samples. In case of PCAWG, this cohort consisted of >20,000 cancer genomes of most cancer types.

For this experimental model system, the controlled biological processes are:

- endogenous processes of DNA damage and repair in normal cells,
- the exogenous process of DNA damage (UV in this case) and
- the defect in repair (MMR deficiency)

Using a model organism approach for extracting mutational signatures from the mutational catalogue of a controlled biological experimental system will be invaluable to identify the underlying biological processes represented in the mutational signatures found in cancer genomes that remain currently unknown. The recent development of numerous software tools paved the way for a wide application of NMF in the study of mutational profiles and the underlying signatures (Alexandrov et al. 2013b; Gehring et al. 2015; Blokzijl et al. 2018). The most critical parameter in NMF is the factorisation rank N, which indicates the number of mutational signatures that can be extracted from the mutational count matrix. The best way to determine the optimum factorisation rank is by an itterative process of trial and error, that is, try different values of N and pick the one performing best for the application at hand (Gaujoux and Seoighe 2010). In this process a number of statistical and arithmetic parameters are calculated to evaluate the successful factorisation process.

To choose the factorisation rank, a dimension reduction approach was used, in which, the number of signatures, or factorisation rank, should be lower than the number of samples in the mutation matrix. We used 'MutationalPattern' (Blokzijl et al. 2018) and 'NMF' (Gaujoux and Seoighe 2010), both Bioconductor, packages in the R statistical environment, to estimate the best factorisation rank. The most common approach is to choose the smallest rank at which, the cophenetic correlation coefficient starts decreasing, as mentioned previously (Brunet et al. 2004) or alternatively, consider the smallest value at which the decrease in the RSS (Residual Sum of Squares) is lower than the decrease of the RSS obtained from random data (Frigyesi and Höglund 2008).

In the factorisation estimation plots, each curve represents a summary measure over the range of ranks in the survey. The colours correspond to the type of data to which the measure is related, such as 'coefficient matrix', 'basis component matrix', 'best fit', or 'consensus matrix'. In order to avoid overfitting, it is recommended to run the same procedure on randomised data and choose the best fit by comparing with randomised data.

After plotting all the individual samples' mutational data as a 96 trinucleotide substitution matrix (10 clones from each sample, and 4 samples generated a [96x40] matrix), the factorisation rank estimation was performed (Figure 4.25) by choosing a rank between 2 and 10. The same process was performed after randomisation of the sample data for comparative analysis as described previously (Gaujoux and Seoighe 2010). The cophenetic correlation coefficient starts lowering at rank 3. By selecting the factorisation rank 3, this sets the number of signatures to be extracted. At the same time, the consensus

matrix of the estimated factorisation rank for 2 and 3 were generated to check the best fit for these data sets (Figure 4.26). Successful factorisation was confirmed by sample clustering of the same data. As shown in Figure 4.26, when a rank of 3 is chosen, the samples cluster according to the mutational process they represent, with both samples derived from *msh2* cells clustering together. When a rank of 2 is chosen, clustering performs more poorly, with untreated wild-type and *msh2* mutant data clustering together. Therefore, a rank of 3 was selected, and used in the following data analysis.

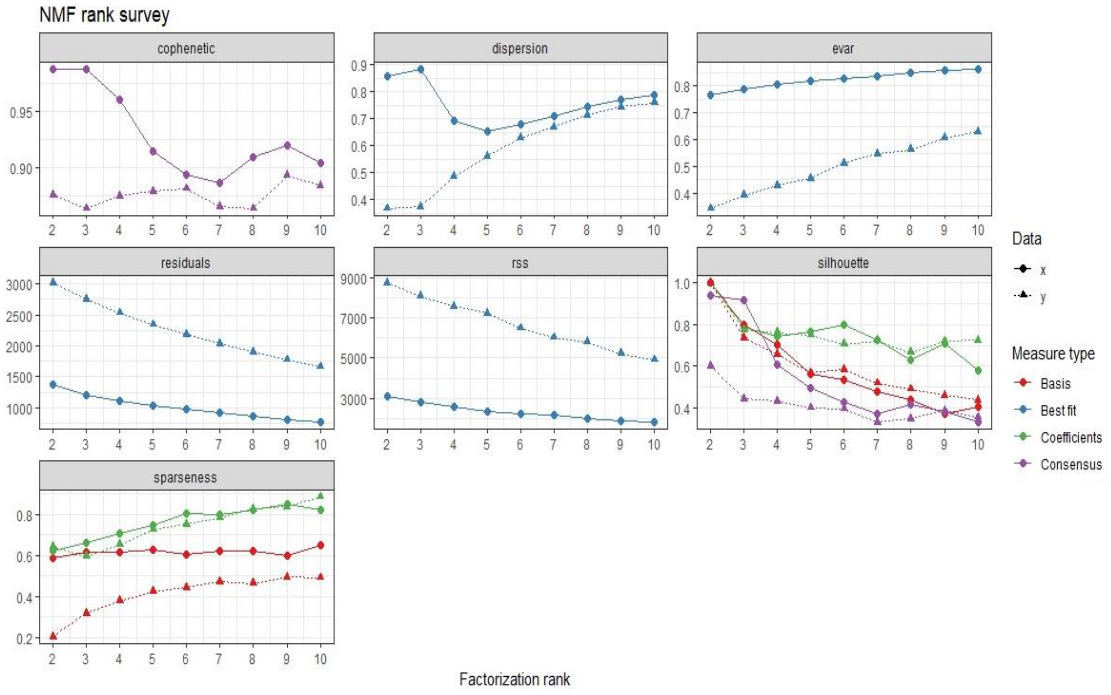


Figure 4.25: Factorisation rank survey. For the estimation of the rank, the above-mentioned quality measures were computed from 50 runs for each value of rank for both samples and randomised data. The estimation is based on Brunet's algorithm. The data marked as 'X', represents the experimental data set, and the data marked as 'Y' represents the same data following randomization.

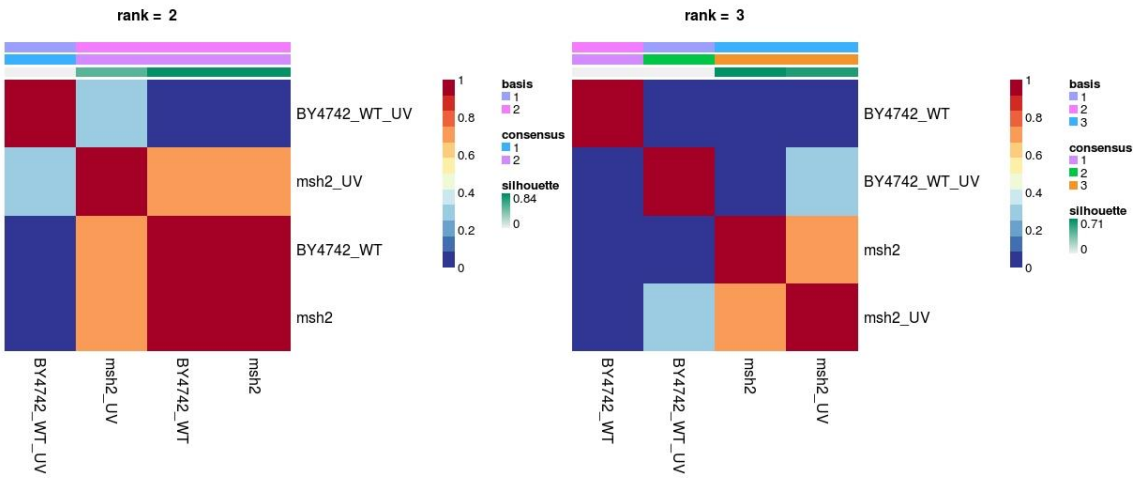


Figure 4.26: Estimation of the rank, consensus matrixes computed from 50 runs for each value of rank 2 and 3 after pooling the samples into groups.

4.3.15 *De novo* signatures extracted from the mutation profiles of yeast cells, extract signatures that correlate with mutagen exposure and DNA repair deficiency

Using a rank of 3 derived from the factorisation rank survey described above, we performed NMF to extract mutational signatures derived from the yeast model system. The resulting 96 trinucleotide mutational profile matrix of the 3 *de novo* extracted mutational signatures are plotted in Figure 4.27. These were named as Signatures A, B & C. The absolute and relative contribution of each of these mutational signatures in each sample was also plotted as a bar chart (Figure 4.28).

Signature A is characterised predominantly by C>A transversions at C<u>C</u>N and T<u>C</u>N trinucleotide context (the mutated base is underlined), with minimum contribution from other types of base substitutions. This signature is exclusively responsible for the mutational pattern found in unirradiated wild-type cells, with this signature contributing only weakly to unirradiated *msh2* cells. This indicates that these mutations are predominantly cause by the normal biological process of mutagenesis due to endogenous DNA damages and replication, in the presence of intact repair processes.

Signature B, on the other hand, is characterised predominantly by C>T at A<u>C</u>N and G<u>C</u>N trinucleotides, T>C at N<u>T</u>N trinucleotides with a few C>A mutations at N<u>C</u>N trinucleotides. This signature contributes to the two samples derived from MMR defective cells is specifically caused by defects in the function of MMR.

Finally, signature C is characterised predominantly by C>T at T<u>C</u>N trinucleotides and T>C at T<u>T</u>N trinucleotide with a small amount of T>A and C>A types of substitution.

Mutations derived from this signature contribute uniquely to samples treated with UV radiation, and they are caused by UV-induced mutagenesis (Alexandrov et al. 2018). Importantly, through the unbiased NMF, we again detect that mutations induced in the UV irradiated *msh2* mutant cells are a true reflection of the additive effect of two independent mutagenic processes (i.e. UV and MMR deficiency).

Taken together, these observation provide the first evidence that a mutagenesis study using the yeast model organism and pipeline developed here, can be used to study de novo extracte mutational signatures induced by controlled exposure to known mutagens or DNA repair defects. Moreover, NMF is able to accurately distinguish mutational signatures derived from multiple latent mutagenic processes active in a model system.

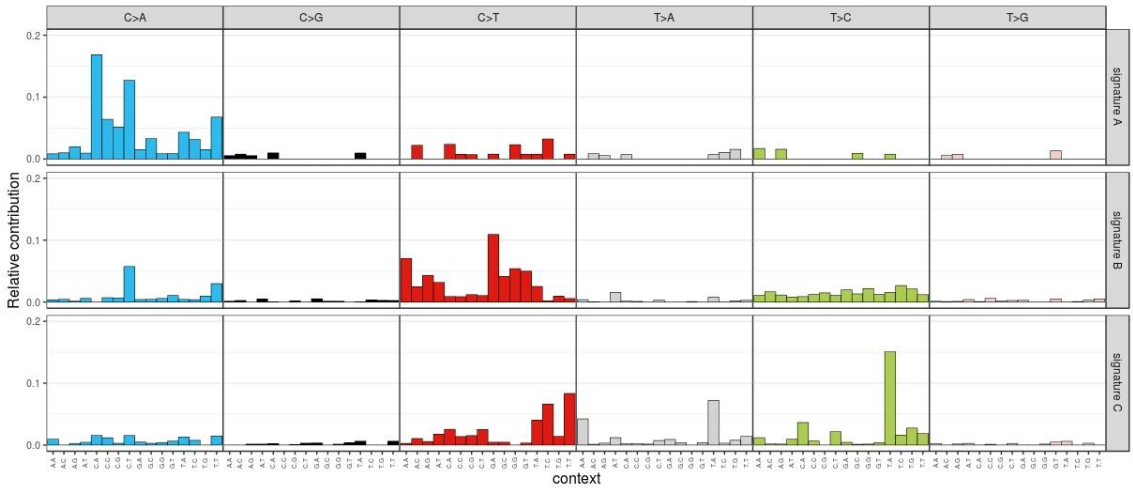


Figure 4.27: Relative contribution of indicated 96-tri-nucleotide changes to the three mutational signatures that were extracted *de novo* by NMF analysis of the acquired somatic mutational catalogue of the experimental model system.

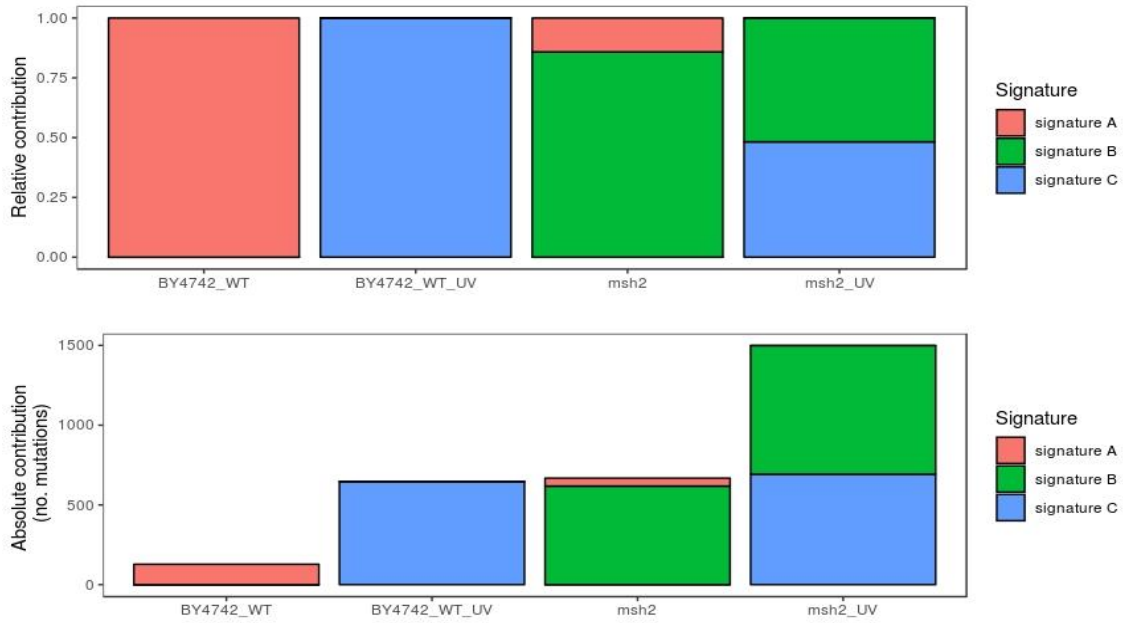


Figure 4.28: Relative and absolute contribution of de novo extracted mutational signatures A, B and C in wild-type and *msh2* cells both with or without UV damages.

Remarkably, hierarchical clustering that makes use of the relative contribution of each of the three extracted de novo mutational signatures to each experimental sample, detects two main clusters and one sub-cluster as shown in Figure 4.29. This result clearly demonstrate that, the three biological processes that were originally selected to conduct this experiment, can indeed be extracted using NMF and hierarchical clustering.

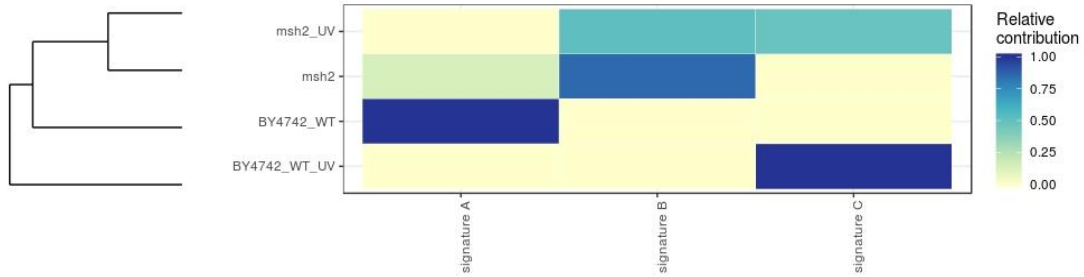


Figure 4.29: Heatmap showing relative contribution of de novo extracted signatures into individual samples. The samples are hierarchically clustered (average linkage) using the Euclidean distance between the vectors of relative contrition with the signatures.

In summary, the three investigated controlled mutagenic biological processes were successfully extracted using NMF and their contribution to the corresponding samples accurately represents the active biological processes. This includes both the DNA damage and DNA repair processes involved, which generates the mutational signatures identified. Future work exposing cells to different known and unknown environmental mutagens using such a workflow will be instrumental in describing the currently unknown causes

of the signatures that are associated with these environmental conditions. Similarly, this approach can be used to subject genetic mutants in various DNA repair pathways to this analysis in order to uncover how genetic defects in different repair pathways, possibly in combination, impinge on mutational outcome. Such studies will help uncover some of the causes of the currently unknown mutational signatures extracted from human cancer genomes.

4.3.16 The cosine similarity of the *de novo* extracted signatures with PCAWG signatures

In order to compare the novel signatures described above with the known PCAWG signatures, the cosine similarity test can apply here again. The cosine similarity between the *de novo* extracted mutational signatures, and the PCAWG signatures observed in cancer genomes will quantify how accurately the mutational signatures observed in cancer (Alexandrov et al. 2013b) can represent those signatures derived from a controlled biological experimental model, and *vice versa*. The cosine similarity of extracted mutational signatures from yeast cells with the PCAWG signatures are presented in figure 4.30.

From this cosine similarity heatmap, it is obvious that signature A is highly similar to PCAWG SBS signature 4 and 38 (cosine similarity $\varepsilon = 0.9$), which were attributed to tobacco smoking and an indirect effect of UV light, respectively. Additionally, relatively low similarity is also observed with other SBS signatures such as SBS8, SBS10, SBS14, SBS18, SBS20, SBS24, SBS29, SBS34, SBS36 and SBS40. As mentioned in previous sections, signature A is predominantly contributing to the mutations induced in unirradiated wild-type cells, indicating that, the active mutagenic processes operating during growth of yeast cells on agar plates generate a range of mutational signatures some of which are highly related to mutation signatures found in human cancers that are thought to be caused by polycyclic aromatic hydrocarbons found in cigarette smoke and the indirect effects of oxidative damage associated with exposure to UV light. Similar findings have also been reported using isogenic human cell-based models (Zou et al. 2018). This shows that the basic mutagenic processes linked to endogenously induced oxidative DNA damage and replication are closely related at this level of analysis, in line with previous findings (Tubbs and Nussenzweig 2017).

Signature B shows high similarity with PCAWG SBS 44 (cosine similarity $\varepsilon = 0.9$), followed by slightly lower similarity with SBS6, SBS14, SBS15 & SBS20. All these

PCAWG signatures are attributed to various forms of defective DNA mismatch repair. Again, confirming that defects in a conserved DNA repair pathway results in set of related mutation signatures in both human cancers and cell-based model organisms.

Signature C does not show high similarity with any of the PCAWG signatures, but displays moderate similarity with SBS7a, SBS7b, SBS7c & SBS7d along with SBS5, SBS40 and SBS41. The underlying biological processes of SBS 7a, b, c & d is due to exposure to solar UV light, but the aetiology of the rest of the mutational signatures are all unknown. This can be explained by the fact that, cancer types with mutational profiles that are similar to SBS type 7 suffer from exposure to predominantly UV-A and UV-B, and have an unknown genetic component that is unique to each patient. This heterogeneity does not apply in a model organism-based approach. However, it is likely that the UV-like signature we observed is likely caused specifically by the exposure to UV-C type radiation, as this is the main wavelength generated by germicidal lamps used in the laboratory. In contrast, the mutational signatures derived from cancer genomes have a high degree of similarity between them and cannot be studied in isolation as compared to the exposure and repair defect described here in yeast.

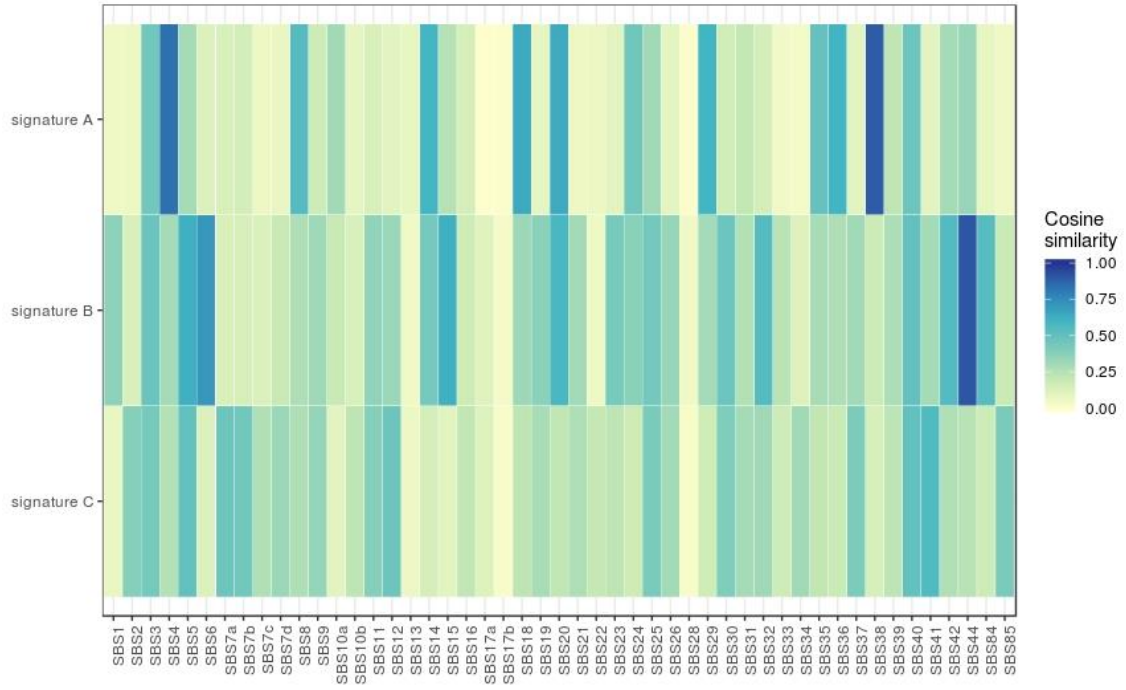


Figure 4.30: Heatmap of cosine similarities between the *de novo* extracted signature and PCAWG signatures.

4.4 Summary

In this chapter, a workflow is developed for acquiring the genome-wide distribution of mutations with or without UV damage and subsequent analysis of mutational types and pattern distribution for comparative analysis with relative repair rates using strains of *S. cerevisiae* as an isogenic model organism. The same workflow will be used to measure accumulation of mutations in various different strains, and subsequent analysis conducted to examine how the structure and organisation of NER (Chapter-V), and the chromatin factors that modulate NER (Chapter-VI), determine the mutational heterogeneity observed within genome.

This chapter describes the development and significance of a novel workflow for accruing and filtering a catalogue of acquired mutations derived from yeast cells. Additionally, I demonstrate that the use of NMF for *de novo* extraction of mutational signatures from these mutational catalogues, successfully decomposed the biological processes of mutagen exposure and DNA repair deficiencies. Likewise, the cosine similarity and reconstruction of mutational profiles with known PCAWG signatures provided additional confirmation for the conservation of the mechanisms of mutagenesis between yeast and human genomes.

Here I provide a proof-of-principle for the use of 'IsoMut' to accumulate unique heterogenous sub-clonal mutations from multiple isogenic yeast strains. The subsequent filtration using a tuning curve for passage 1 control sample mutations results in the selection of sub-clonal mutations that occurred exclusively during the ~1100 generations of yeast cell divisions. This way of accumulating mutations is important for understanding the biological processes operating within a controlled isogenic biological system. IsoMut uniquely exploits the isogenic nature of the samples by filtering out SNVs that are shared between different clones. This approach successfully detects the mutational pattern with or without UV exposure, allowing us to understand the biological process operative in a population of cells.

The core finding of this chapter reveals that the distribution of mutations within the genome follows a similar structure and organisation to that of the relative repair rates described previously. Building on these findings, here I showed that the genomic features that are important for repair organisation determine the location and types of mutations within genome. These observations may help to explain the extent of the mutational heterogeneity observed in cancer genomes.

Plotting the distribution of mutations in various genomic features showed that mutations are not randomly distributed within genome. Following UV damage, the distribution of mutations within the yeast genome is altered in the context of the genomic features that are defined by their differential repair rates. Interestingly, similar findings for the cellular repair capacity at different genomic features suggest that heterogeneous repair rates can influence the distribution of alkylating damage induced mutations within the yeast genome (Mao et al. 2017a).

Higher order chromatin structure modulates the mutational distribution within the genome especially at Micro-C boundaries. This observation is in line with the finding that mutations occur at regulatory regions such as TF binding sites that are commonly found in different types of cancer. In eukaryotes, several higher order chromatin structures have been described such as gene loops, enhancer-promoter loops, "topologically-associating domains"/ "chromosomally-interacting domains" (TADs/CIDs), and lamina-associated domains (LADs). These higher order chromatin structures each have different biological properties. For example, association of gene loops with the promoter and terminator determines directionally in yeast (O'Sullivan et al. 2004; Tan-Wong et al. 2012). Similarly, TADs/CIDs are associated with the functional regulatory domain in mammals (Symmons et al. 2014). Recently, several whole cancer genome sequencing studies showed that hotspots of mutations in similar types of higher order chromatin organisation are mostly associated with CTCF containing chromatin loops (Guo et al. 2018; Kaiser and Semple 2018). Future research aims to investigate, to what extent this higher order chromatin structures modulates the distribution of mutations, and also how repair is organised from these sites to maintain genome stability. Moreover, analysis described in this chapter has the potential to help determine whether the mutations at these positions in human cells are caused by genotoxin exposure and/or defects associated with how the repair mechanisms are organised in the genome.

In addition to higher order chromatin structure, well-known active regulatory regions such as TFBSs or GG-NER factor binding sites also showed variation in the mutational distribution depending on the types of TFBSs, following UV exposure. Similar results observed in two recent papers revealed that binding of transcription factors to its cognate binding sites might interfere with NER factors and thereby results in a higher density of mutations found at these sites in melanoma (Perera et al. 2016; Sabarinathan et al. 2016). This observation also suggests that the essential biological processes such as variation in transcription and DNA repair might impede each other's activities.

At the same time, it will be interesting to see whether gene expression levels in different mutant backgrounds has an effect on mutation induction. It is known that UV irradiation has an effect on gene expression. However, how this impinges on the mutational asymmetry observed in this experiments is not known. It is possible that the mutational strand asymmetry correlates with gene expression levels in response to UV damage and other environmental exposures.

Furthermore, sequencing of several intermediate clones sampled during the mutagenesis protocol will help to reveal the timing of mutagenic events following UV damage in various repair mutants. The connection between early repair kinetics (<3hrs) and the accumulation of mutations over several passages or generations is not known. It would be interesting to measure the mutation accumulation rate, expressed as allele frequency. At passage 30 the majority of mutations have an allele frequency of 1. Detecting mutations at early passages will reveal the kinetics of the allele frequency as measured by IsoMut. This is an important aspect of the somatic mutation theory of cancer.

The use of NMF to infer the mutational signatures or the biological processes operating in a set of complex omics datasets was successfully employed during the decomposition of active biological processes operating within these experimental datasets. The employment of factorization rank survey and consensus matrix for evaluation of a number of mutational signatures was successful for describing the active biological process operating within the yeast genome. The use of this concept can be anticipated in future translational research to harness the power of such whole genome analyses for diagnostic and personalised medicine methods to improve cancer therapy.

The study has significant implications for understanding the biological processes of genomic instability, as defined by the higher than normal mutation frequency, often observed in cancers. This study demonstrated the utility of whole genome sequencing in cell lines as a mutagenesis assay. We measured the mutagenic effect and defined the mutation spectrum caused by UV irradiation; a common genotoxic agent. Matching mutational signatures to DNA repair deficiency has a tremendous potential to stratify cancer therapy based on the relationship between these two phenomena and patient response to chemotherapy. This approach appears advantageous over genotyping marker genes, as mutational signatures provide a read-out for cellular repair deficiency associated with either genetic or epigenetic defects. Following on from our study, we expect that analysing DNA repair–defective model organisms and human cell lines, alone or in conjunction with defined genotoxic agents, will contribute to a more precise definition

and mechanistic understanding of the mutational signatures occurring in cancer genomes and will help to establish the aetiology of these signatures.

Chapter V

Defective NER alters the genome-wide pattern of mutations in response to UV damage

Contents

5.1 Background

Nucleotide excision repair (NER) is highly organised within the genome and over 30 gene products are involved in this biochemically complex process (Friedberg et al. 1995). At the same time, the genetic information contained within the DNA molecule is highly prone to damage to its structure due to both, the deleterious effects of normal cellular metabolic processes (endogenous) (Tubbs and Nussenzweig 2017) and to genotoxic stresses such as ultraviolet (UV) radiation or chemical damage from the environment (exogenous) (Friedberg et al. 2005). Thousands of lesions occur daily in the DNA of each of our cells (Tubbs and Nussenzweig 2017). Such DNA damage can cause disruption of cell division and altered gene regulation, while defective DNA repair can introduce DNA mutations that may alter the genetic information within the cell. Therefore, repair of damaged DNA is fundamental to genome stability, which when compromised is one of the major hallmarks of cancer. The malfunction of genes required for processing DNA damage by all major DNA repair pathways, including NER, can lead to cancer predisposition (Holmquist and Gao 1997). Defective NER has been clearly documented in the hereditary cancer-prone disease xeroderma pigmentosum (XP), as well as rare recessive syndromes such as Cockayne syndrome (CS) and Trichothiodystrophy (TTD) (de Boer and Hoeijmakers 2000). Recently, somatic mutations in core NER genes has been reported in urothelial tumours (Kim et al. 2016). This demonstrated the importance of NER as a fundamental mechanism for protecting the integrity of the genome (Hoeijmakers 2001), and provided important evidence for the somatic mutation theory of cancer.

NER removes many types of DNA damages induced by physical agents such as UV light (de Laat et al. 1999), chemical agents such as polycyclic aromatic hydrocarbons (PAH) in tobacco smoke (Hecht 1999) and chemotherapeutic agents such as platinum drugs (Reed 1998). Among them, well-known and well-studied, are UV-induced DNA photoproducts, such as cyclobutene pyrimidine dimers (CPD) and 6–4 photoproducts (6-4 PP), which get efficiently repaired by NER. Two sub-pathways of NER exist characterised by the initial damage recognition steps: the rapid acting transcription coupled repair pathway (TC-NER) that operates on the transcribed strands of actively transcribing genes and involves RNA polymerase II in the damage recognition step; and the slower acting global genome repair pathway (GG-NER) that operates on all DNA, including non-transcribed and repressed regions in the genome. This pathway involves a unique subset of proteins in the early stages of DNA damage recognition. Following the

initial stages of DNA damage detection, these two pathways converge and utilise the same set of DNA repair proteins. Most of the yeast NER genes have well conserved structural and/or functional human homologues, and the main features of both GG-NER and TC-NER pathways are evolutionarily conserved (Hoeijmakers 1993; Hoeijmakers 1994).

Initially, work on GG-NER in yeast cells showed that for efficient repair, a complex of proteins, also known as the GG-NER complex is involved in the early stages of this process. The GG-NER complex is comprised of Abf1, Rad16 and Rad7 in which Abf1 is bound to specific DNA binding sites (Reed et al. 1999), which can be found at hundreds of locations throughout the yeast genome (Yu et al. 2009). Abf1, is also known as a general regulatory factor in yeast and plays an important role in the efficient organisation of GG-NER throughout the genome (Yu et al. 2009). The Rad16 protein is a member of the SWI/SNF super-family of chromatin remodeling factors. Proteins in this super-family contain conserved ATPase motifs and are subunits of protein complexes with chromatin-remodelling activity (Flaus and Owen-Hughes 2011). Since Rad16 operates on repressed and non-transcribed regions of the genome during GG-NER, it has long been assumed that its role might involve chromatin remodeling (Verhage et al. 1994), conceivably, to improve access to damaged DNA. As described in Chapter 3, and in our recent publication (van Eijk et al. 2018), Rad16 is involved in the remodeling nucleosomes adjacent to GG-NER complex binding sites (GCBSs) for efficient repair in response to UV damage. Rad16 also contains a C3HC4 type RING domain, which is known as an E3 ubiquitin ligase domain. Previously reported studies showed that, the GG-NER complex also has E3 ubiquitin ligase activity involving the Cul3 and Elc1 proteins (Pintard et al. 2004; Willems et al. 2004; Gillette et al. 2006). Moreover, Rad7 is part of an E3 ligase complex that ubiquitinates Rad4; a core yeast NER protein which binds to damaged DNA (Gillette et al. 2006). This ubiquitination of Rad4 in response to UV specifically regulates NER via a pathway that requires *de novo* protein synthesis and directly influences NER and UV survival. Importantly, it is established that Rad7 and Rad16 exist in a complex within the cell and that deletion strains of either component have an identical phenotype (Verhage et al. 1994; Guzder et al. 1998; Reed et al. 1998).

Like Rad16, Rad26, a TC-NER factor in the yeast *S. cerevisiae,* is also a member of the SWI/SNF super-family of DNA-dependent ATPases involved in chromatin remodeling (van Gool et al. 1994; Osley et al. 2007) and is the homologue of the human Cockayne syndrome group B (CSB) gene (Guzder et al. 1996; Lee et al. 2001). Both Rad26 and CSB are involved in the preferential repair of UV lesions on the transcribed strand, and

in this process, they function together with the other components of NER (Teng and Waters 2000; Ghosh-Roy et al. 2013).

Rad4, on the other hand, is a protein that acts in the core NER pathway and is involved in removal of bulky lesions (Min and Pavletich 2007). Rad4 forms a heterodimer with Rad23, which is homologous to the human XPC-hHR23A/B complex, and is involved in damaged DNA recognition in human cells (Masutani et al. 1994; Jansen et al. 1998). It is well established that, in humans, NER deficiencies are associated with both Xeroderma Pigmentosum (XP) and Cockayne syndrome (Foury 1997; Friedberg et al. 2005). Similarly, many studies have reported correlations between SNVs in NER genes in a host of different cancers, predicting oncogenesis due to the presence of susceptibility alleles identified in NER genes (Kim et al. 2016; Phillips 2018).

Genome-wide DNA damage and repair mapping with high resolution methods using yeast as a model organism showed that in response to UV damage, wild-type DNA repair rates can be altered greatly by either defects in NER or factors that modulate NER efficiency (Teng and Waters 2000; Teng et al. 2011; Yu et al. 2016). Comparing repair rates in various different NER factor mutants revealed that the distribution of repair rate varies significantly among different mutants (Yu et al. 2016). Additionally, the UV sensitivity among NER defective mutants also varies. For example, *RAD4* deleted yeast cells are highly UV sensitive, followed by *RAD16*, *RAD7* and *RAD26* (Yu et al. 2011; Waters et al. 2012; Kong et al. 2016). The UV sensitivity of *rad16* and *rad7* mutants is identical, because they function as a complex. Taken together, in response to UV damage, variations in the repair capacity exist within the yeast model organism, depending on the genetic background for repair.

Genome-wide mutational pattern studies from whole cancer genomes showed that, mutations are heterogeneously distributed over the linear genomic structure, which could be caused by variation in genomic DNA repair capacity in the genome (Salk et al. 2010). However, the extent to which the variation in distribution of relative repair rates contributes to the mutational heterogeneity observed in cancer is not known (Lawrence et al. 2013). Additionally, recent reports have begun to measure and decipher the non-random nature of the mutational patterns that shape the somatic cancer genomes of different cancer types. The analysis of whole cancer genome sequencing data is helping to determine the causes of these mutational patterns, based on our current knowledge of DNA damage and repair mechanisms (Haradhvala et al. 2016). For example, in melanoma cancers, the mutational load is usually higher at positions of higher nucleosome

occupancy (heterochromatin) and in late replicating regions (Hodgkinson et al. 2012; Woo and Li 2012). Recent studies also suggest that, both of these chromosome features share lower levels of NER activity, suggesting that NER is potentially one of the key factors determining the location of UV induced mutations (Hodgkinson et al. 2012; Woo and Li 2012; Polak et al. 2014).

DNA is packaged inside the cell into a protein-DNA complex that is important for compaction and regulation of metabolic processes that require access to the DNA. This so-called chromatin structure is also affects DNA repair and was recognised as being fundamentally restrictive to the repair process (Smerdon 1991). Chromatin organisation due to histone modification can also have influence on the mutational rate in human cancer, suggesting differential accessibility of repair factors to the damages (Schuster-Böckler and Lehner 2012). Most recently, genomic DNA repair rates from cell culture-based studies have been correlated with the incidence of mutations in skin and other human cancers, suggesting that cancer-associated mutations occur in regions of the genome that are more difficult to repair (Sabarinathan et al. 2016). Recent evidence also suggests that, in human cells, binding of transcription factors at DNase I–hypersensitive sites in gene promoters results in lower levels of DNA repair and higher rates of mutation. This suggests that NER may also be organized in the human genome (Perera et al. 2016; Sabarinathan et al. 2016). Collectively, these studies demonstrate the importance of understanding the genomic organisation of DNA repair mechanisms and how they contribute the final landscape of mutational patterns. The majority of mutations are ultimately a result of failure to repair DNA in a timely fashion, therefore we need to know all the details of the many ways in which repair can break down to help understand mutational heterogeneity and outcome. It is anticipated that future translational research will harness the power of such whole-genome analyses for diagnostic and personalised medicine approaches to improve cancer therapy.

The whole-genome sequencing studies (The Cancer Genome Atlas network, 2012) have measured the genome-wide tumour-specific somatic mutation patterns, which report the entire spectrum of mutations accumulated during tumorigenesis within the cancers of individuals. These studies revealed the association of multiple mutational signatures, which are indicative of the mutational processes responsible for the mutations found in sporadic cancers, derived from the so-called normal population, across a range of human cancer types (Alexandrov et al. 2013). The association of some of the mutational signatures identified are known: *signatures caused by exposure to environmental*

mutagens, such as solar UV light, associated mainly with skin cancers, and PAHs contained in cigarette smoke, associated mainly with lung cancers; *signatures that are a result of defective DNA repair in one of the various repair pathways*, such as defective MMR, associated with colorectal cancer (Alexandrov et al. 2018). Therefore a mutational signature represents the combined effect of DNA damage and defective DNA repair (Zou et al. 2018). The spectrum of mutational signatures varies among different types of cancer, the same mutational signature can be present in multiple cancer types, whereas the same cancer genome can contain two or more mutational signatures (Alexandrov 2014). However, the causes of many of the mutation signatures identified to date, remain to be determined (Alexandrov et al. 2018).

Overall, NER plays an important role in maintaining genome stability by removing both endogenous and exogenous DNA helix distorting damages. NER is highly organised within the genome and maintains a uniform repair rate in response to both endogenous and exogenous damages. This repair rate can be affected by both defects in either NER factors or genomic features that modulate NER. These factors might alter the distribution of mutational patterns, which is a feature often associated with various cancers and age-related diseases. The causes of the heterogenous distribution of mutations in different types of cancers are not known. I hypothesised that variation in the distribution of DNA repair rates within the chromatin environment, due to changes in the efficiency of repair, might contribute to the distribution of mutational patterns observed in cancer genomes.

To examine this, in this chapter I will measure the distribution of mutations within the genome in response to UV damage in cells where several NER factors are defective causing defects in the GG-NER and TC-NER pathways. I will use yeast as a model organism to provide a controlled experimental system. Plotting the catalogue of mutations obtained from these experiments, in relation to genomic context, will allow us to determine the effects of changes in the rates of DNA repair on the distribution of mutational patterns observed in yeast cells.

In the previous chapter, I described the protocol and bioinformatics pipeline for measuring the distribution of mutations present in UV damaged wild-type and mutation-prone yeast cells. In this chapter, I am going to use the same approach to study how defective NER, in both the GG-NER and TC-NER sub-pathways, alters the distribution of mutations in the yeast genome. This will enable us to determine whether alterations in the distribution of relative repair rates in the genome affect the type and distribution of mutations.

Determining the effect of NER on the genomic pattern of UV-induced mutations

To disrupt the GG-NER and TC-NER pathways, I used *rad16* or *rad7* mutated yeast cells, in which the GG-NER sub-pathway is defective, and *rad26* yeast cells, in which the TC-NER sub-pathway is defective. *rad4* mutated yeast cells, in which the core NER pathway is defective were included as a control. I also included the mutational catalogue obtained from wild-type yeast cells described in the previous chapter for comparison. For *rad4* deleted cells, UV-treated results are not available, because *rad4* yeast cells are highly sensitive to UV light in comparison to certain other NER factors.

Plotting the mutational catalogue from undamaged wild-type and various NER deficient cells, around a variety of genomic features that affect genomic structure, will inform us whether the defects in either GG- or TC-NER, alter the distribution of mutations in and around different genomic features, both in terms of mutation types, and their pattern. Comparing this pattern with the UV-induced pattern will determine how the combined activity of DNA damage exposure and/or defective repair modulates the mutational distribution within genome.

These observations will demonstrate the importance of understanding how genetic damage is formed and repaired throughout the entire genome in response to UV radiation. And whether the repair factors, that are directly involved in nucleosome remodeling, alter the distribution of the mutational pattern observed.

5.2 Material and Methods

The yeast strains used for this study, their genotype and source are mentioned in Table 5.1.

Table 5.1: Yeast strains and their respective genotypes used for this chapter.

Strain ID*	Lab ID	Genotype	Source
Wild-type (BY4742)	247	Wild-type MATa his3delta1 leu2delta0 ura3delta0	Euroscarf
rad16	218	BY4742 MATa his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0 YBR114w::kanMX4	Euroscarf
rad4	272	BY4742 MATa his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0 YJR052w::kanMX4	Euroscarf
rad7	217	BY4742 MATa his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0 YJR052w::kanMX4	Euroscarf
rad26	224	W303 MAT alpha Δrad26	Reed Lab, CU

* All the yeast strains used in this experiment are haploid and alpha mating type.

The biological processes under investigation in this chapter, and the related yeast strains used for the study of NER deficiency on mutagenesis are illustrated in figure 5.1. Rad16 and Rad7 are involved in early stages of GG-NER, whereas Rad26 is involved in TC-NER. Rad4 is considered as a core NER factor and is highly sensitive to UV damages.

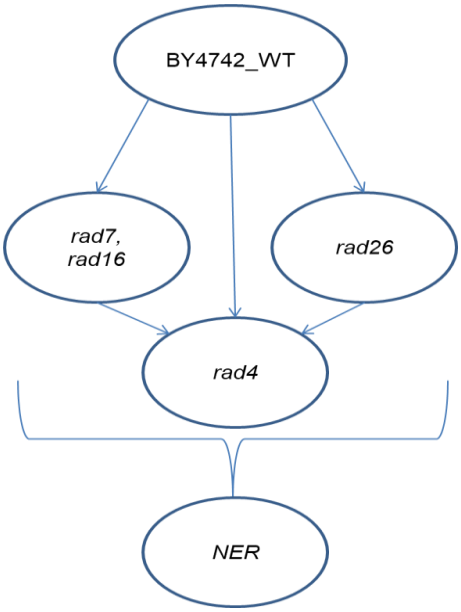


Figure 5.1: Graphical presentation of biological process of NER studied here in this chapter. Rad7 and Rad16 are involved in the initial stage of GG-NER and Rad26 is involved in early stages of TC-NER, the two sub-pathways of NER. Rad4 is involved in the core NER pathway. The BY4247 is the wild-type background control for all of these repair mutants.

For accumulation of mutations with or without UV irradiation, cells were propagated through ~1,100 generations from passage 1 to passage 30 as described in section 4.2.1 of the previous chapter. The materials and methods employed for growing the yeast strain used for this chapter and subsequent DNA extractions (Appendix-IV, Figure A4.1) were performed by following the protocols described in the Material and Methods Chapter-II. The sequencing library was prepared by following Illumina library preparation protocol and gDNA sequenced using the Illumina sequencing platform (Details are described in Chapter II). Whole genome sequence data was obtained using Illumina Hi-Seq paired-end sequencing chemistry, with read sizes of 75 bp. The raw paired-end sequences were processed using the same pipeline as mentioned in the previous chapter in section 4.2.2. The IsoMut bioinformatics pipeline was used to detect acquired mutations as described in section 4.2.3 of the previous chapter. All the computer coding used to run IsoMut is attached as an e-Appendix. The background mutations detected were subtracted after plotting the tuning curves of all samples as described in the previous chapter section 4.2.4, to determine the cut-off value based on the S score. The tuning curves were generated for both substitution mutations and indels. All the substitution mutations are attached as an e-Appendix. For mapping substitution mutations according to different genomic features, section 4.2.5 from previous chapter was followed. Mutational signatures and cosine similarity analysis was then performed using the bioinformatic workflow as described in section 4.2.6 and 4.2.7, respectively.

5.3 Results

5.3.1 Measuring the total number of acquired genomic mutations in cells

In order to examine the acquired total mutational catalogue from experimental samples, mutations from passage 1 were used as a control sample and filtered using the tuning curve (Appendix-IV, Figure A4.2-A4.8) to allow for only a very few false positive mutations (Table 5.2). Note: The total number of indels generated in this experimental setting in wild-type and NER defective yeast cells is negligible. This was expected, and therefore the subsequent analysis will focus on single nucleotide variations or base substitution mutations.

Table 5.2: Number of SNV and short indel mutations in wild-type and various NER defective yeast cells with or without UV damage.

Strains*	Passage	n	Total SNV	SNV Mean	Total Indels	InDels Mean
BY4742_WT	Starting Clone	1	1	1	0	0
	End Clone	10	128	12.8	13	1.3
BY4742_WT_UV	Starting Clone	1	2	2	1	1
	End Clone	10	642	64.2	35	3.5
rad16	Starting Clone	1	5	5	0	0
	End Clone	10	274	27.4	72	7.2
*rad16*_UV	Starting Clone	1	2	2	4	4
	End Clone	10	3136	313.6	78	7.8
rad7	Starting Clone	1	1	1	0	0
	End Clone	10	464	46.4	53	5.3
*rad7*_UV	Starting Clone	1	1	1	0	0
	End Clone	10	3515	351.5	157	15.7
rad26	Starting Clone	1	1	1	0	0
	End Clone	10	243	24.3	10	1
*rad26*_UV	Starting Clone	1	1	1	1	1
	End Clone	10	891	89.1	33	3.3
rad4	Starting Clone	1	2	2	0	0
	End Clone	10	853	85.3	6	0.6

Abbreviations: WT, wild-type; UV, Ultra Violet; n = number of clones. Independent mutations in the starting clone represent false positives of the number of detections. SNV = Single Nucleotide Variation. InDels = Insertions and deletions.

* All the yeast strains used in this experiment are haploid and alpha mating type.

5.3.2 Loss of NER increases total mutations in both undamaged and UV damaged cells

In the absence of UV damage, in comparison to wild-type cells, there was approximately a two-fold increase in base substitution mutations in TC-NER defective *rad26* cells. This number is about three to four-fold higher in GG-NER defective *rad16* and *rad7* cells (Table 5.1). However, the mutational load is about six-fold higher in the case of core NER defective *rad4* cells (Table 5.1). This indicates that the repair capacity of different NER mutants varies for the repair of endogenously induced DNA damage within cells. In the presence of UV damage, these mutational loads also vary between different NER deficient yeast strains compared to wild-type cells. A slight increase in mutational load is observed in UV treated TC-NER defective *rad26* cells (Table 5.1). The mutational load was about six-fold higher for both GG-NER defective *rad16* and *rad7* mutants (Table 5.1), confirming that at the level of UV-induced mutations, Rad7 and Rad16 play an important role in the maintenance of genome stability. Because of its high sensitivity to UV damage, the mutational profile for the NER defective *RAD4* deleted cells was not obtainable. These mutational outcomes correspond well with the previously studied DNA repair rates and UV sensitivities of these strains in our laboratory (Yu et al. 2016).

5.3.3 Mutational asymmetry caused by defective NER is observed around linear genomic features in response to DNA damage

To measure the distribution of the mutational load in relation to the linear arrangement of genomic features, I calculated the observed and expected levels of mutation at ORFs, TSSs, TESs and the ARS sequences obtained from the *Saccharomyces* Genome Database (SGD) (Figure 5.2, upper panel). The $\log_2$ ratio of observed over expected mutations from the experimental datasets at these genomic locations were plotted in the lower panel shown of Figure 5.2. The range for the total number of mutations detected at the different linear genomic features, is due to the variation in the total number for each of these SGD features present in the genome (Figure 5.2 upper panel).

Examining mutational loads in relation to ORFs in GG-NER defective *rad16* and *rad7* mutant cells reveals a small but significant depletion in the total number of observed

versus expected mutations in these regions of the genome. In fact, *rad16* cells showed significantly fewer mutations even without exposure to UV radiation (Figure 5.2, ORF panel). This result likely reflects, the competition between global genome and transcription-coupled NER pathways for lesions in this region of the genome. It is conceivable that loss of GG-NER results in enhanced TC-NER in this genomic context, resulting in fewer mutations observed in these regions, to a level below that which is expected. When mutation data is plotted around TSSs including a 200 bp flanking region on either side, no significant difference in the observed versus the expected mutation levels are seen in this context (Figure 5.2). However, when mutations are examined at TESs, significantly higher levels of UV-induced mutations than expected are observed in wild type cells. Also, in GG-NER defective *rad16* mutants, significantly higher levels of mutation are seen in the absence of UV damage, but this result is reversed in response to UV damage, where significantly fewer mutations than expected are observed at TES (Figure 5.2, TES panel). In contrast to wild-type cells, after UV irradiation GG-NER defective *rad7* and *rad16* mutant cells both showed significant enrichment of the mutational load in relation to ARS sites (Figure 5.2, ARS panel), This indicates that these sites are more susceptible to mutation induction when the GG-NER pathway is lost. No significant change in mutation levels are observed in TC-NER defective rad26 deleted cells, indicating that loss of transcription repair coupling does not significantly alter the levels of mutation observed at these sites. Finally, no significant change in mutation levels is observed when events are examined at randomly selected sites. Collectively, these results demonstrate that defective GG-NER results in an altered pattern and level of mutations observed at different genomic locations, resulting in both higher, and lower levels of mutation than expected, depending on the genomic location examined. We suggest that this result could be caused by changes in the relative rates of repair occurring at different locations in the genome. This is illustrated by observing the change in the genomic location of points of inflection in the repair rate curves for different mutant NER backgrounds (Figure 5.3). It appears that genomic locations where the relative rates of repair shift from high to low and *vice versa*, represent regions of genome instability, resulting in altered levels of mutation at these locations.

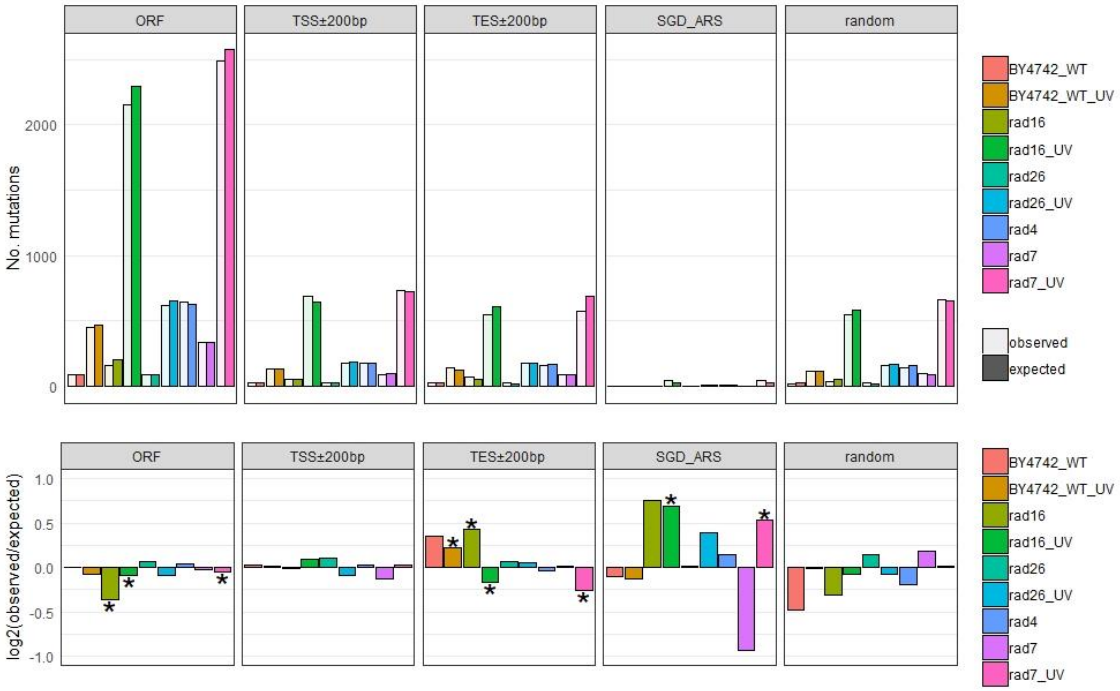


Figure 5.2: Enrichment and depletion of base substitutions in linear genomic features such as ORF, TSS, TES, ARS and random sites for both wild-type and various NER defective yeast cells with or without UV damage. The $\log_2$ ratio of the number of observed and expected point mutations indicates the effect size of the enrichment or depletion in each region. Asterisks indicate significant enrichments or depletions ($P < 0.05$, one-sided binomial test).

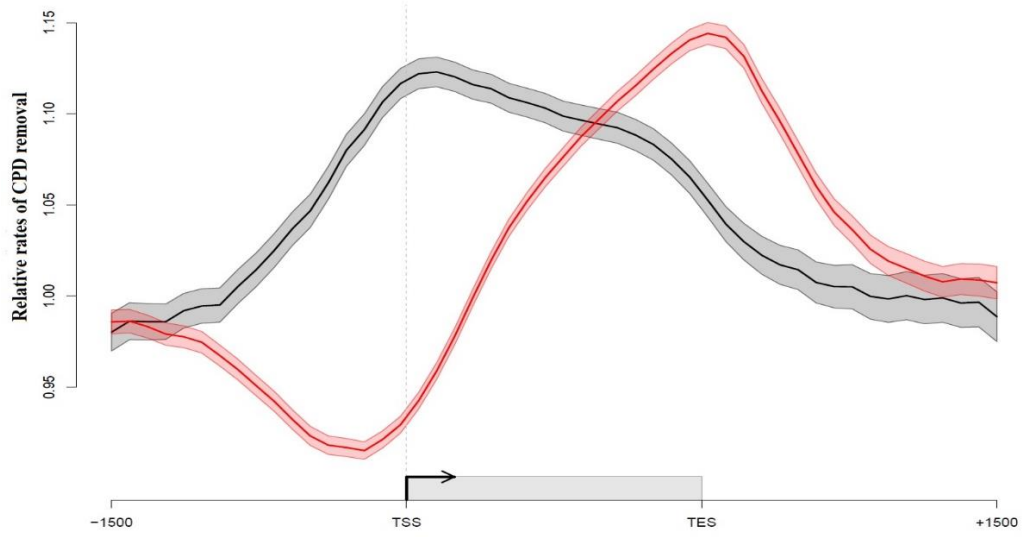


Figure 5.3: Relative rates of CPD repair around ORF structures. Solid lines show the mean of CPD repair rates in wild-type (n = 3, black line), and *rad16* cells (n = 2, red line), Shaded areas indicate the standard deviation, with CPD levels plotted as arbitrary units on the y-axis.

5.3.4 Distribution of mutational density in relation of chromatin structure

To investigate the mutational distribution around the linear chromatin structure in various NER deficient strains in response to UV damage, the observed and expected mutations were plotted in relation to nucleosome depleted NFRs, nucleosome containing SPNs, and also around the higher order structure Micro-C boundaries (Figure 5.4). Additionally, to investigate the binding effects of regulatory proteins to their DNA motifs, the observed and expected mutations were plotted around two of the most common yeast general regulatory factors (Abf1 and Reb1) binding sites, as well as around the GCBSs (Figure 5.5).

In wild-type cells, fewer mutations than expected are found around both NFRs and SPNs although the numbers are not statistically significant. This suggests that while repair is active, DNA lesions are efficiently repaired around these sites reducing the mutational load, and this holds even after UV damage (Figure 5.4, NFR and SPN panel). However, at micro-C boundaries, a statistically significant increase in UV-induced mutations is observed compared to that expected in wild type cells. This indicates that these locations are particularly susceptible to mutation induction especially following exposure to UV damage. Importantly, loss of GG-NER in *rad7* and *rad16* deleted cells results in significantly higher levels of UV-induced mutations than that expected at NFRs and Micro-C boundaries, which are regions that are related to GG-NER complex function. Interestingly, loss of TC-NER function at these locations also results in higher levels of mutation than expected after UV damage, and this is statistically significant at NFRs. Surprisingly, significantly fewer UV-induced mutations than expected were found at the sites of strongly positioned nucleosomes, which are predominantly found in the +1 position of open reading frames (Figure 5.4, SPN panel). It has been suggested that strongly positioned nucleosomes might obstruct the formation UV-induced DNA damage (Mao et al. 2016), but why this specific subset of nucleosomes shows unexpectedly lower levels of UV-induced mutation remains unknown.

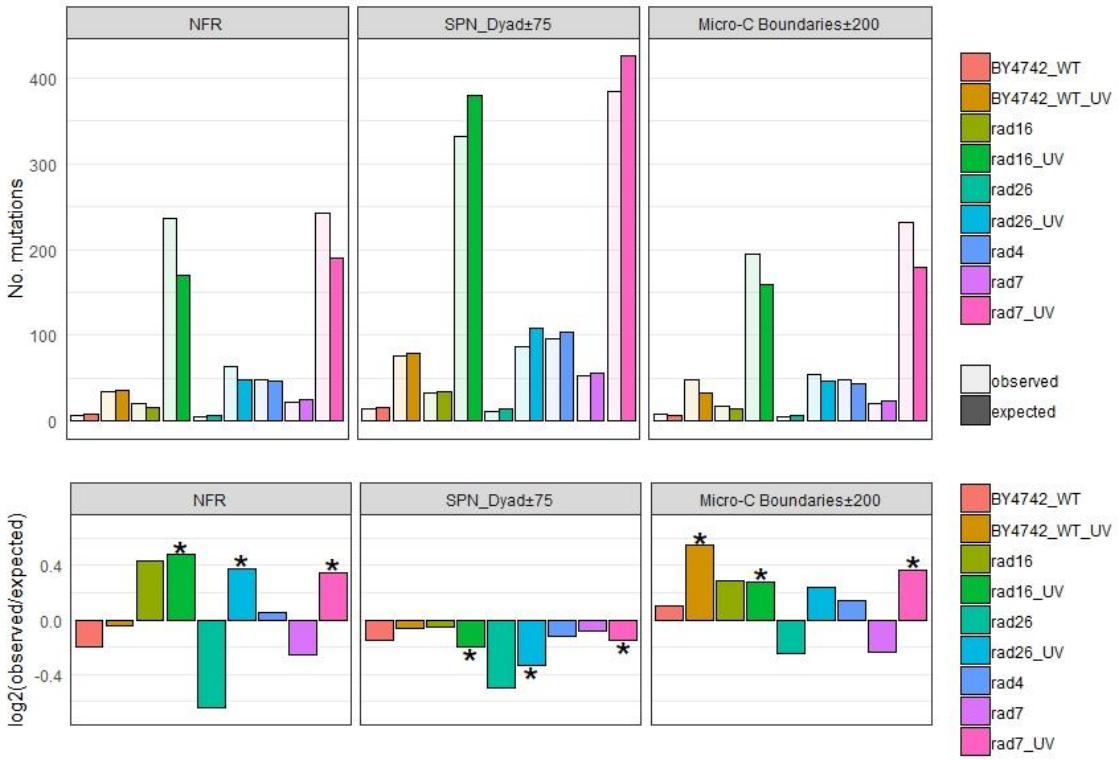


Figure 5.4: Enrichment and depletion of somatic substitutions in the primary structure of chromatin features such as NFR, SPN, and Micro-C boundaries for both wild-type and NER defective yeast cells, with or without UV damage. The $\log_2$ ratio of the number of observed and expected point mutations indicates the effect size of the enrichment or depletion in each region. Asterisks indicate significant enrichments or depletions for mutation ($P < 0.05$, one-sided binomial test).

Lastly, examination of the mutational load around Abf1, Reb1 and also GCBSs protein binding sites was examined. No significant enrichment or depletion of mutations is detected when the data is plotted around Abf11 binding sites (1644 motifs ± 150 bp either side) (Figure 5.5, Abf1 panel). This result could be due to the broad variety of other roles of Abf1 motifs in yeast cells, and the other activities of Abf1 factor binding at these sites, where it functions as a general regulatory factor. In contrast to Abf1, when the data is plotted around Reb 1 binding sites (1500 motif ± 150 bp either side), a significantly higher than expected level of mutation is observed in both wild-type and GG-NER defective cells in response to UV damage, but not in TC-NER defective cells, (Figure 5.5, Reb1 panel). This suggests that different classes of transcription factor binding to DNA might also affect the distribution of mutations within yeast genome, although the precise relationship between Abf1 and Reb1 binding sites is not well understood. Finally, examining mutation levels around GCBSs, the sites from which GG-NER is initiated in yeast cells, revealed that statistically significant higher numbers of UV-induced mutations

than expected were found at these sites in GG-NER defective *rad16* mutated cells compared to wild type cells. This demonstrates that defective GG-NER observed around these GCBSs can result in higher levels of mutation than expected. Surprisingly, in *rad7* deleted cells that are also GG-NER defective, higher than expected levels of mutation were not observed. It is not clear why this is the case, but if confirmed, this result would be the first indication of a different molecular phenotype (UV-induced mutation induction) for Rad7 and Rad16.

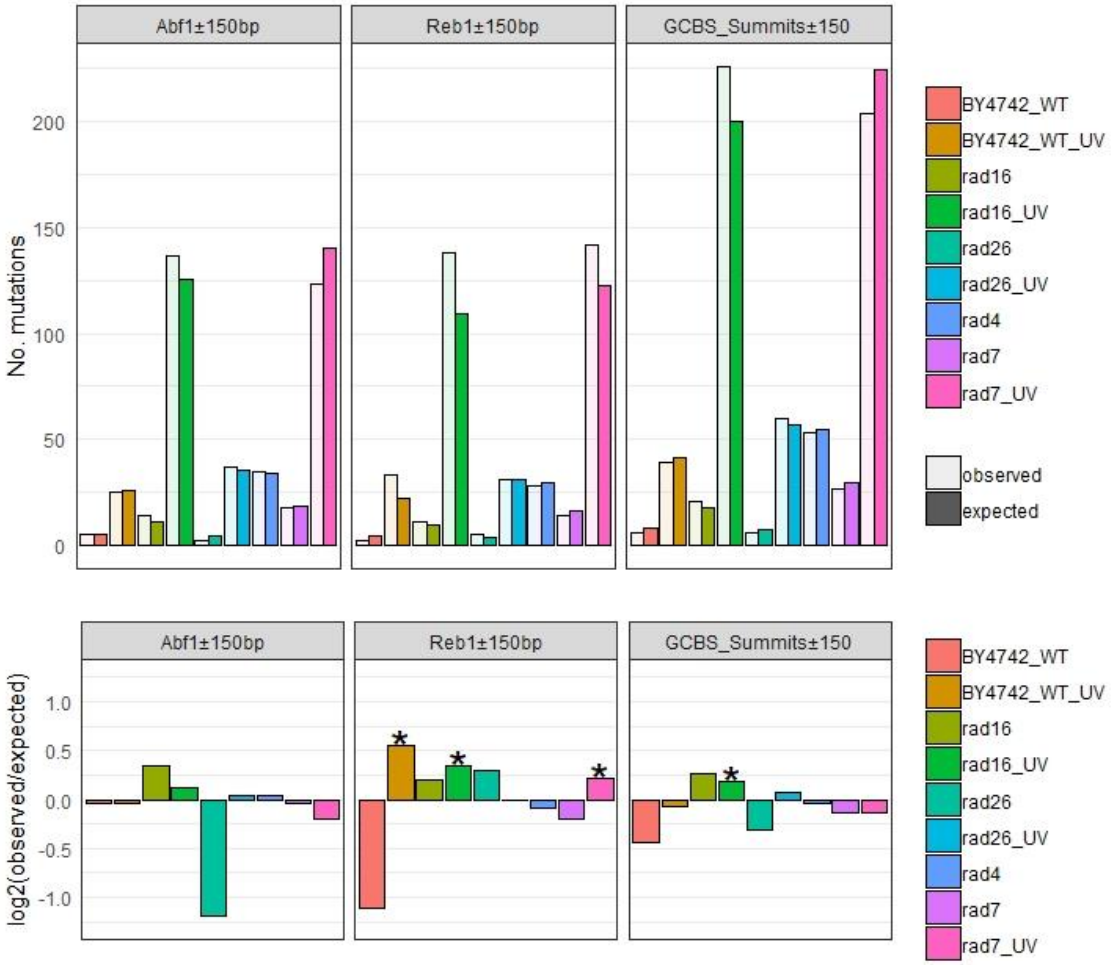


Figure 5.5: Enrichment and depletion of somatic substitutions in primary structure of chromatin features such as Abf1, Reb1 and GCBS summits for both wild-type and NER defective yeast cells with or without UV damage. The $\log_2$ ratio of the number of observed and expected point mutations indicates the effect size of the enrichment or depletion in each region. Asterisks indicate significant enrichments or depletions ($P < 0.05$, one-sided binomial test).

5.3.5 Substitution mutations suggests mutational heterogeneity

In order to investigate the mutational heterogeneity among the distribution of the 6 possible types of base substitutions, the relative contribution of the substitution mutation

spectra was plotted in Figure 5.6 for the strains used in this experiment, both with and without exposure to UV light.

In undamaged cells, the mutational load in wild-type cells are two or three-fold lower than the GG-NER or TC-NER deficient cells individually, but when the core repair factor for NER Rad4 is deleted the mutational load goes as much as ~7 fold higher (Figure 5.6) (See table 5.1 for total number of substitutions observed in various controlled experimental model background). This higher mutational load can be explained by the fact that Rad4 plays an essential role in the NER process, and it has much higher sensitivity to UV damages. Although the mutational load in undamged cells is similar between GG-NER (*rad16, rad7*) and TC-NER (*rad26*) deficient cells, there is a difference in the distribution of mutation types. This distinctive pattern of mutational distribution indicates, important differences in the two repair sub-pathways with respect to their effect on mutation. Overall *rad4, rad7* and *rad16* along with wild-type cells showed a similar pattern to the relative contribution of mutation types, but *rad26* shows a clearly different pattern in the relative contribution to the mutational load, indicating that TC-NER generates a distinctive pattern of mutations in the genome.

As expected, after UV exposure, the overall mutational load is higher for both GG-NER and TC-NER deficient cells, compared to both wild type cells and their undamaged counterparts. However, the GG-NER defective cells showed a higher level of mutational load compared to TC-NER defective cells, demonstrating the importance of GG-NER for repairing UV induced damages and preventing mutagenesis.

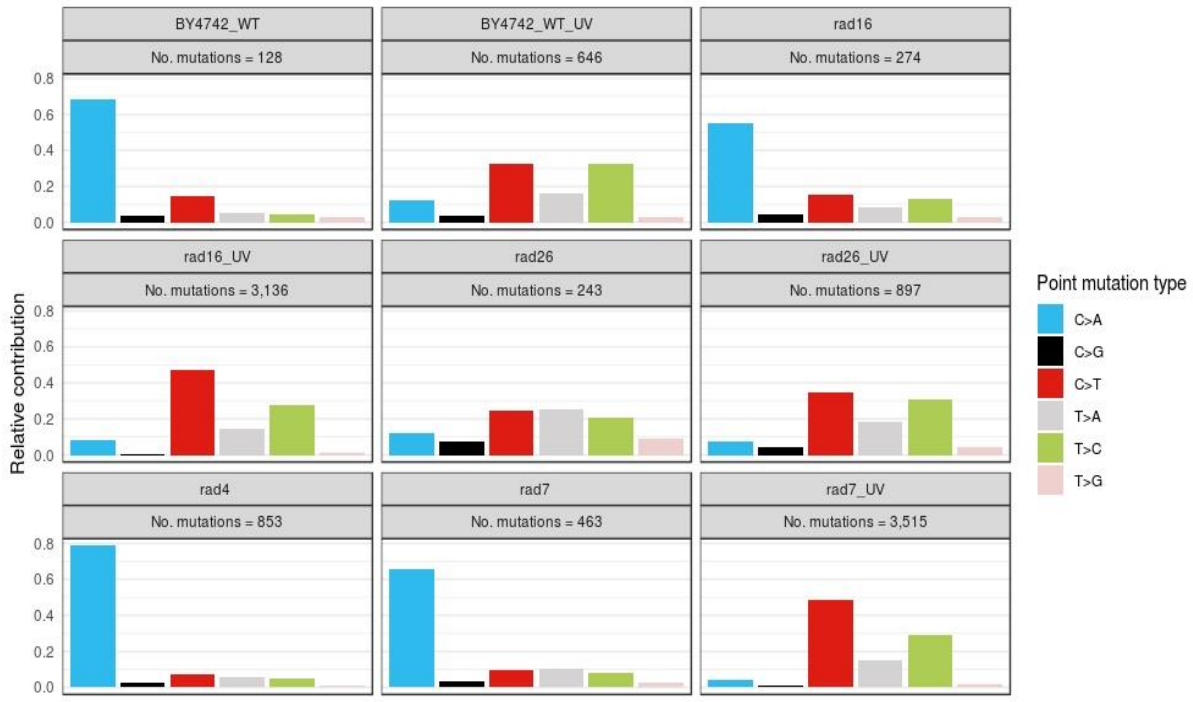


Figure 5.6: Relative contribution of each mutation type in the base substitution catalogue of each sample listed. The total number of substitutions observed after passage 30 in each category is also shown.

5.3.6 Distribution of the type and mutational load at GCBSs - the origins of GG-NER

As mentioned in results section of chapter III, the GG-NER complex occupies chromatin at NFR sites of a specific subset of gene promoters. This establishes the nucleosome structure at these genomic locations, which is refer to as GCBS's. These sites are frequently located at genomic boundaries that delineate CIDs, also known as Micro-C boundaries. Not all NFR regions are the sites for GCBSs. Consequently, I decided to explore the difference in the distribution of mutations at GCBSs associated NFR and other NFR.

To investigate the distribution of mutational types at the genomic NFR with GCBSs and NFR without GCBSs, the relative contribution of each mutation type was plotted over the GCBS-related NFRs (dark shaded bar) and non-GCBS-related NFRs (light shaded bar) (Figure 5.7, upper panel). Furthermore, the $\log_2$ ratio of the mutational count over these two sites were plotted to determine the bias in the distribution of mutations (Figure 5.7 lower panel). A negative value indicates a mutational bias towards non GCBS-containing NFRs, whereas a positive value indicates a mutational bias towards GCBS-containing NFRs.

As can be seen, in general all mutation types showed higher levels than expected at the GCBS-containing NFR sites, compared to non GCBS-containing NFRs. However, significantly higher levels of mutations than expected are detected in UV treated *RAD7* or *RAD16* deleted GG-NER deficient cells. In this case, two well-known, UV-induced base substitution types are observed, with significant C>T substitution seen in *rad16*, and both C>T and T>C substitutions in *rad7* cells enriched around GCBSs. This demonstrates how defective GG-NER can alter the distribution and type of mutations in the genome. The types of mutation found in TC-NER defective *rad26* deleted cells also showed a distinctive pattern of mutation induction at GCBS-containing NFRs, even in the absence of UV damage. Furthermore, following UV damage, the mutational distribution changes at these sites with a significant bias for T>A mutations at GCBS-containing NFRs, demonstrating that defects in TC-NER result in UV-induced T>A type substitutions around GCBS-containing NFRs (Figure 5.7, *rad26* panel). It is conceivable that this mutation type could be caused by the failure of rad26-deleted cells to recover RNA synthesis after UV exposure, as opposed to the failure to repair directly induced DNA damage by UV light, which does not typically induce T>A type substitutions. Failure to recover RNA synthesis is the primary molecular phenotype of *rad26* deleted cells, and this is also the case for Cockayne's syndrome B patients where defects occur in the *CSB* gene; the human homologue of yeast *Rad26*.

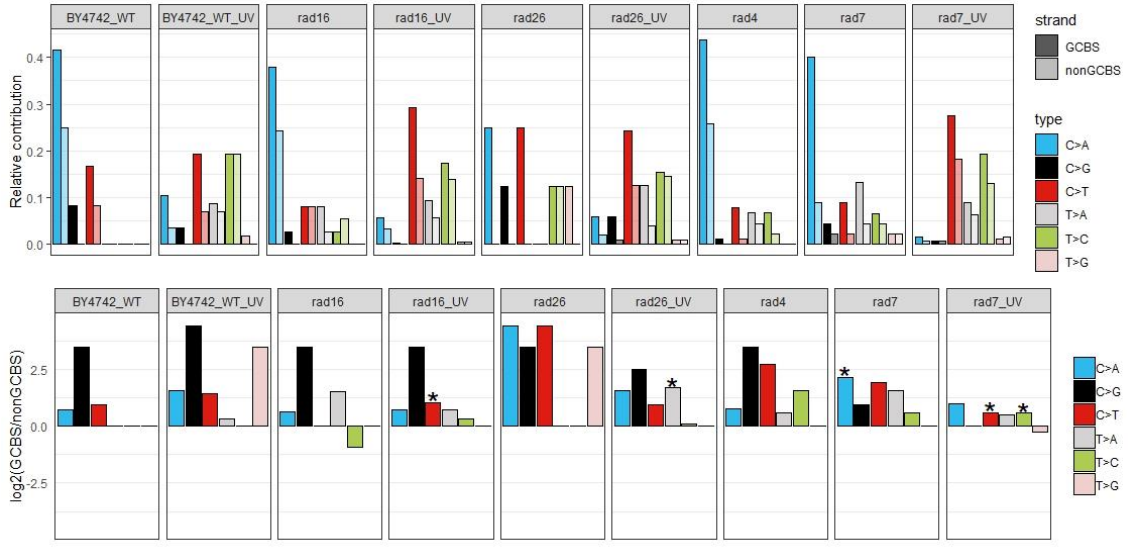


Figure 5.7: Enrichment and depletion of the distribution of mutational load around GCBSs NFR sites or non-GCBSs NFR sites. The relative contribution of 6 substitutions at GCBS containing NFR(dark shaded) and non-GCBS containing NFR (light shaded) for individual sample (upper panel). Log_2 ratio of the number of mutations on GCBS and non-GCBS NFR regions per indicated base substitution of each samples shown in lower

section of this figure. The log_2 ratio indicates the effect of the bias and asterisks (*) indicate significant region asymmetries ($P<0.05$, one-sided binomial test).

5.3.7 DNA damage and NER deficiency significantly increase the mutational load within yeast genomic regions with low accessibility

It has been demonstrated that histone acetylation alters the nucleosome structure. An open chromatin structure enables various biochemical activities to function more efficiently within the genome, including DNA repair. To study how the mutational pattern is distributed around regions of high and low acetylation within the yeast genome, the mutational catalogues were plotted around these regions. In figure 5.8, the upper panel bar plot shows the relative contribution of the 6 possible substitution types within high (dark shaded) and low (light shaded) acetylation status regions of the yeast genome. The lower panel shows the log_2 ratio for each pair of bars represented in the upper panel. The negative log_2 ratio indicates higher levels of mutation at regions of low-level acetylation, and positive log_2 ratio indicate enrichment of mutational load at highly acetylated regions within the yeast genome. Asterisks (*) indicate significant asymmetries for either high or low acetylated regions within yeast genome ($P<0.05$, one-sided binomial test).

As expected, higher levels of all types of base substitution mutations are observed in regions of low acetylation within genome (Figure 5.8). Significant enrichment of mutational distribution at low acetylated regions within yeast genome in response to UV damages correlates with inefficient repair at these heterochromatic regions within genome. In response to UV damage, all strains showed a significant mutational bias towards regions of low-level acetylated for C>A and T>A type substitutions as well as C>T, T>C mutations caused by UV-induced DNA damage. In the absence of DNA damage, all GG-NER defective cells also exhibited a bias for only C>A mutation towards low acetylated regions, indicating that endogenously induced C>A type mutations are also induced at less accessible regions for repair within yeast genome.

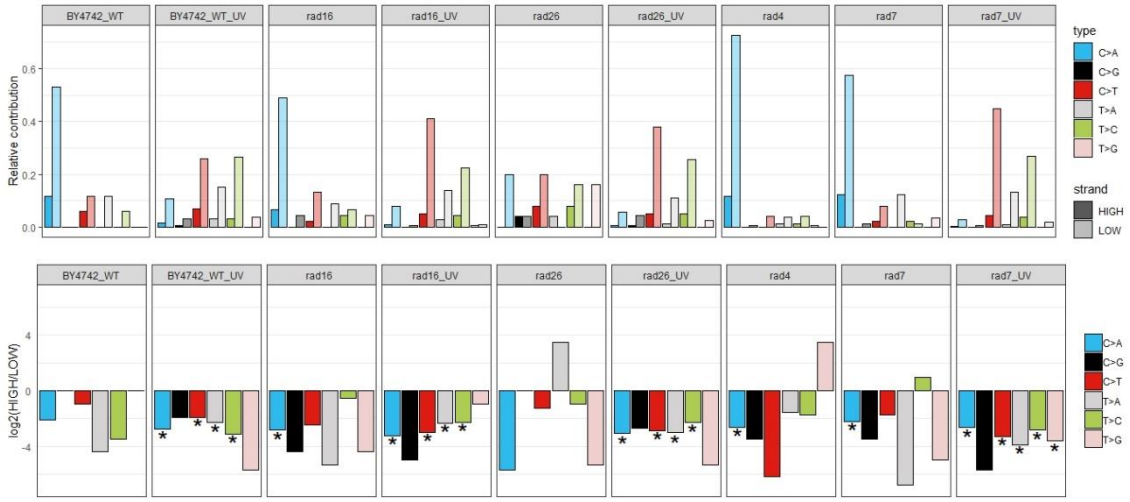


Figure 5.8: Mutational distribution bias towards high and low acetylated regions within wild-type and NER defective yeast genome. Log_2 ratio of the number of mutations on the high and low acetylated regions per indicated base substitution of each samples shown in lower section of this figure. The log2 ratio indicates the effect of the bias and asterisks (*) indicate significant acetylation region asymmetries ($P<0.05$, one-sided binomial test).

5.3.8 Replication timing and UV-induced mutagenesis in NER defective cells

The distribution of mutations varies depending on the origin of replication within the genome, from early replicating regions to late replicating regions. One study suggests that high fidelity template switching occurs during early stages of replication, whereas error-prone TLS occurs at the later stages where damage tolerance causes mutational biases observed in cancer cells (Lang and Murray 2011). Another study suggests that imbalance of dNTP pools at different stages of replication contributes to mutational bias towards late replicating regions (Pai and Kearsey 2017). Moreover, chromatin accessibility during early stages of replication also creates accessibility to repair enzymes, resulting in higher rates of repair (Adar et al. 2016). All of these studies suggest replication timing might affect the distribution of mutational pattern within the genome. To investigate whether the distribution of mutation types varies in relation to replication timing with or without UV damage, the relative contribution of each mutation type was plotted from wild-type and various NER defective yeast strains according to their early (dark shaded) and late (light shaded) replication status throughout the yeast genome (Figure, upper panel). The log_2 ratio of these early versus late mutational loads indicates the mutational bias with respect to replication time. Positive log_2 ratio indicates a bias towards mutations in early replicating regions and negative values indicate a bias during late replication (Figure 5.9, lower panel).

In wild-type undamaged cells no significant base substitution bias is observed. However, following UV damage, a significant mutational bias towards late replicating regions is observed for C>T, T>C and T>A mutations, with C>G type mutations enriched at early replicating sites, although the relative contribution of this to the total load is very low. This indicates that UV induced damages are more mutagenic at late replicating regions within yeast genome.

GG-NER does not appear to have a striking effect on the type and distribution of mutations found in early and late replicating regions of the genome. Strikingly, however, loss of TC-NER in undamaged *rad26*-deleted cells showed significant mutational bias toward early replicating regions for C>T, T>C and T>A types of mutations, suggesting lack of TC-NER at the early stages of replication, generates unexpectedly high levels of mutations at these sites in the yeast genome. In response to UV damage, this bias is essentially lost. This observation uncovers an unexpected connection between TC-NER and early and late replicating regions of the genome. Finally, in untreated *rad4* NER deleted cells, all types of mutations are biased towards at late replicating regions, but with significant bias for only C>A and T>A types. This may indicative of early replication-associated NER, due to increased accessible chromatin during replication (Adar et al. 2016). Collectively, these results show that replication timing and repair capacity play an important role in determining the distribution mutations in both endogenously and exogenously damaged cells.

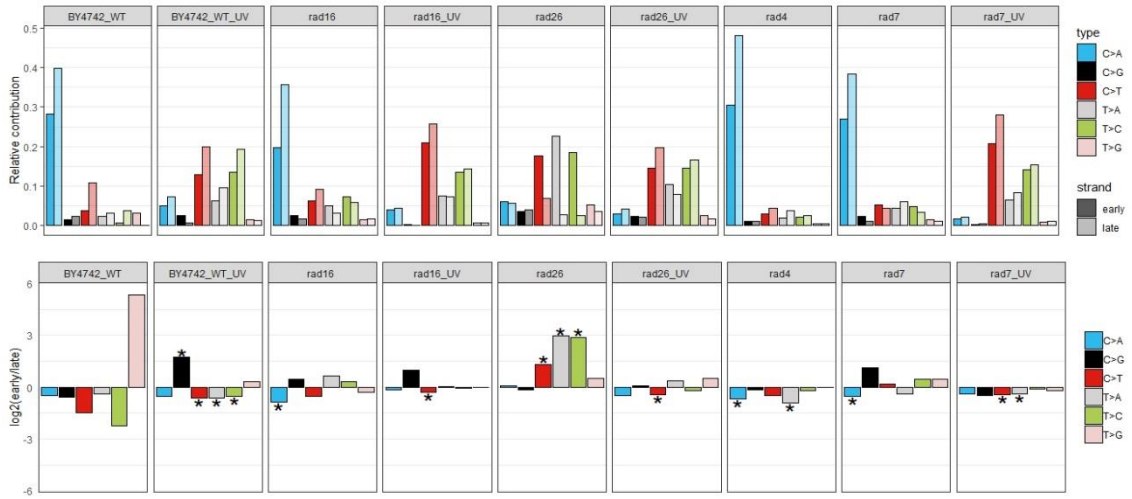


Figure 5.9: Distribution of mutations in early and late replicative regions within wild-type and NER defective yeast cells. The upper panel shows the relative contribution of each nucleotide changes and is subdivided into early (dark shades) and late (light shades) phases. Log_2 ratio of the number of mutations in the early and late replicative regions per

indicated base substitution of each samples shown in lower section of this figure. The log_2 ratio indicates the effect of the bias and asterisks (*) indicate significant replicative timing asymmetries ($P<0.05$, one-sided binomial test).

5.3.9 Transcriptional strand asymmetry observed in mutational distribution in TC-NER and GG-NER defective cells.

Transcriptional strand asymmetry in the distribution of mutation patterns is one of the frequently observed phenomena in skin, lung and liver cancers (Haradhvala et al. 2016). This transcription strand asymmetry is observed because UV-induced DNA damage occurs in one strand and is repaired at different rates by both TC-NER (faster repair rate) and GG-NER (slower repair rate) (Reed 2011). To investigate the transcriptional strand bias in the distribution of mutation patterns in GG-NER and TC-NER defective yeast cells, the relative distribution of mutational loads in transcribed (dark shaded) and untranscribed (light shaded) strands were plotted in figure 5.10 (upper panel). The log_2 ratio indicates the bias in their distribution either in the transcribed or in the untranscribed strand. The positive log_2 ratio indicates bias towards the transcribed strand and negative log_2 ratio indicates bias towards the untranscribed strand (Figure 5.10, lower panel). In general, significant transcriptional strand bias in the distribution of substitution mutations is observed predominantly in UV exposed cells compared to undamaged cells. Similarly, to that observed in wild-type cells, following UV damage, GG-NER defective *rad16* and *rad7* cells show a significant enrichment of C>A, C>T and T>C bias towards the untranscribed strand and T>A bias towards the transcribed strand. This mutational bias towards the untranscribed strand is due to lack of GG-NER operating on UV induced DNA damage particularly in the nontranscribed strand. The presence of active TC-NER in these cells removes damages from transcribed strand resulting in a mutational strand bias towards the untranscribed strand. No significant transcription strand bias is observed in *rad26* cells, demonstrating the significance of TC-NER in contributing to the transcriptional strand bias on the distribution of mutations observed in both wild-type and GG-NER defective cells following UV damage. This also demonstrate the importance of GG-NER function for maintaining the balance in the distribution mutations in TC-NER defective cells.

As noted in the previous chapter, significantly higher levels than expected of T>A type substitutions show a strand bias towards the transcribed strand, which is also observed in all other UV exposed cells *except* for *rad26* deleted cells. At this moment the cause of T>A types of mutations and their bias towards the transcribed strand is unknown, but it

appears to be a Rad26 dependent phenomenon. This striking feature needs to be analysed in greater details for T>A types of mutations for their bias towards transcribed strand.

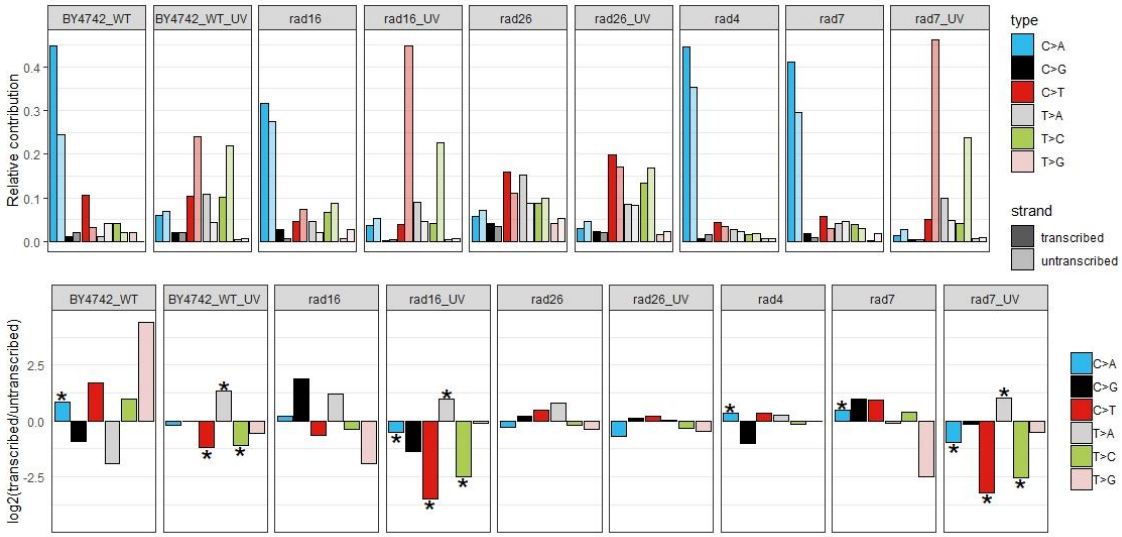


Figure 5.10: Distribution of substitution mutations with transcriptional strand information. The upper panel shows the relative contribution of each nucleotide change as a bar plot and is subdivided into both the transcribed (dark shades) and untranscribed (light shades) strands. Log_2 ratio of the number of mutations on the transcribed and untranscribed strand per indicated base substitution of each sample is shown in the lower section of this figure. The log_2 ratio indicates the effect of the bias and asterisks (*) indicates significant transcriptional strand asymmetries ($P<0.05$, two-sided binomial test or Poisson test).

5.3.10 Genome-wide distribution of acquired mutations in wild-type and GG-NER defective cells

In order to investigate the mutational hotspots and variation in the distribution of mutation types throughout the genome, rainfall plots were generated by plotting the genomic locations at which the substitution mutations were observed (x-axis) with their inter-mutational distance also plotted (y-axis). The rainfall plots of *rad7*, *rad26* and *rad4* are described in Figure 5.11 – 5.15. Without UV exposure, in GG-NER deficient *rad7* cells, the mutational pattern was dominated by C>A substitutions, along with a few other types of mutations. This observation is similar for *rad16* cells (Appendix-IV, Figure A4.9) as well as *rad4* cells (Figure 5.13), but not in *rad26* cells (Figure 5.12). In the previous chapter, the rainfall plot of wild-type cells, without UV damage also showed similar patterns (Chapter IV, Figure 4.16), except the mutational load is much higher in GG-NER deficient cells. This indicates that both wild-type and GG-NER deficient cells

predominantly generate C>A(G>T) substitution mutations globally within the yeast genome, without exposure to exogenous damage.

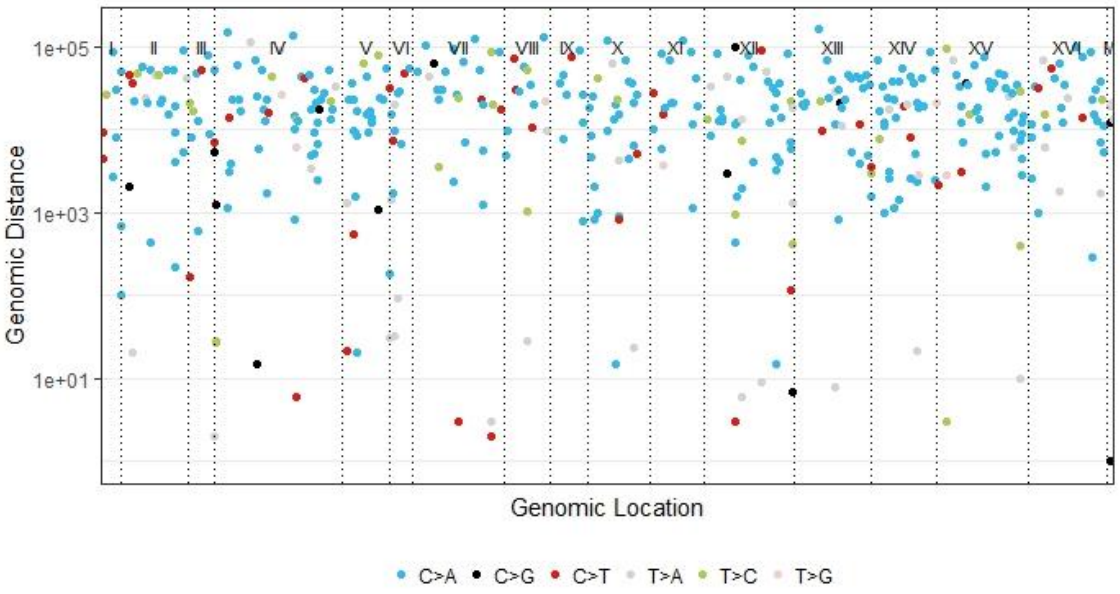


Figure 5.11: Rainfall plot of undamaged *rad7* cells showing the genomic location of mutation with their inter-mutational distance. The coloured dots represents the 6 possible types of base substitutions.

Following UV damage, in GG-NER defective cells, the mutation pattern is dominated by C>T, T>C, T>A and C>A substitutions, which is similar to that of wild-type cells following UV damage, except that the overall mutational load is higher in GG-NER defective *rad7* cells (Figure 5.12) and *rad16* cells (Appendix-IV, Figure A4.10). This indicates that a lack of GG-NER increases mutagenesis globally within the genome after UV damage induction. This observation also demonstrates the importance of GG-NER for maintaining genomic stability. In figure 5.12, a highly dense distribution of mutation types is observed throughout the genome, this is due to UV-induced CC>TT doublet mutations within GG-NER deficient cells, which was also observed in melanoma cancer (Pleasance et al. 2010a).

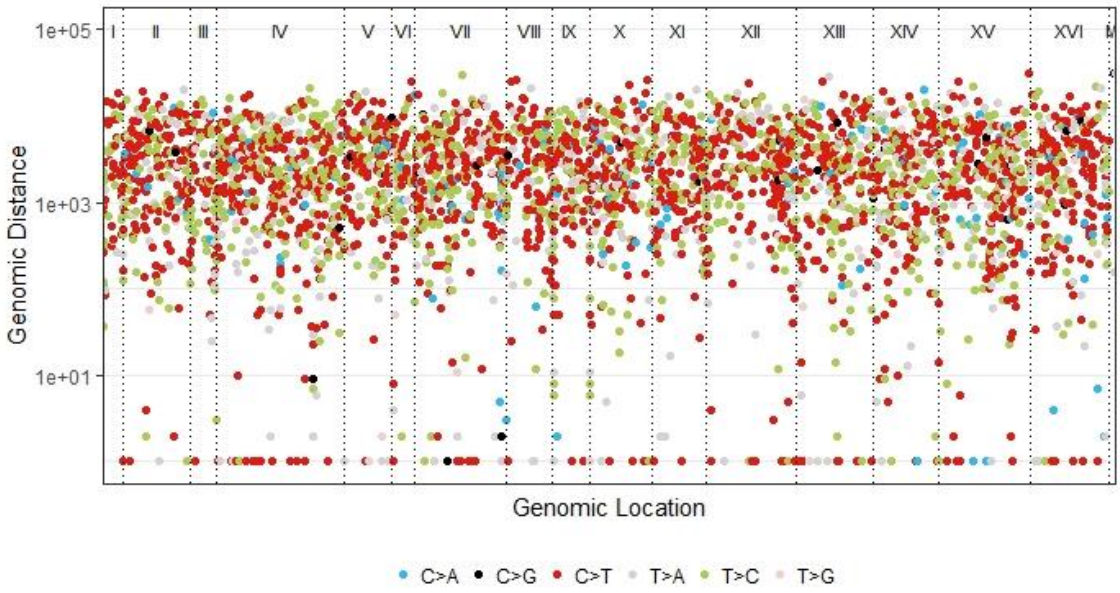


Figure 5.12: Rainfall plot of *rad7* cells after UV damage induction, showing the genomic location of mutation with their inter-mutational distance. The coloured dots represents 6 possible types of substitutions.

An exception was observed in *rad26* cells not exposed to UV damage, where the background mutational load is represented by all possible types of base substitution (Figure 5.13). This suggests that lack of TC-NER alters the wild-type and GG-NER-defective pattern of mutations without UV damage. This indicates that the biological mechanism of TC-NER changes the pattern of mutations observed in the genome. Although in *rad26* cells, without UV damage, the mutations showed a unique pattern in comparison to wild-type and other NER defective cells, following UV damage, this mutational distribution changed towards a predominantly UV-induced pattern of mutation types (Figure 5.14). The total mutational load after UV damage in *rad26* mutant cells was slightly higher than wild-type UV damaged cells, however, this load is significantly lower in comparison to the GG-NER deficient UV exposed cells (Table 5.1). These observations are in line with the known UV sensitivities of these strains. This also indicates the differences in the roles of Rad26 in TC-NER and Rad7/16 in GG-NER for maintaining genome stability. An interesting finding from this analysis is that both undamaged and UV-exposed rad26 mutant cells showed a higher amount of mitochondrial mutations, which probably linked to frailer to the recovery of RNA synthesis, mentioned in its human homolog Cockayne syndrome group B mutant cells (Cleaver et al. 2014, Scheibye-Knudsen et al. 2012). However, the types of mutations within the mitochondrial genome might explain the phenotype of Cockayne syndrome.

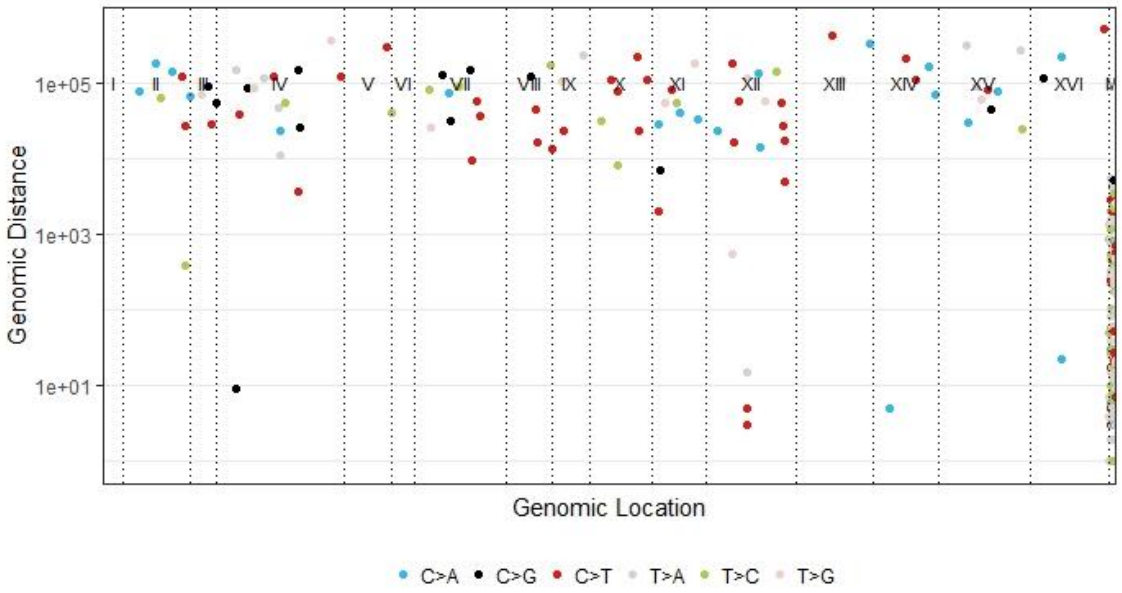


Figure 5.13: Rainfall plot of *rad26* mutant cells without UV damage, showing the genomic location of mutation with their inter-mutational distance. The colour dot represents 6 possible types of substitutions.

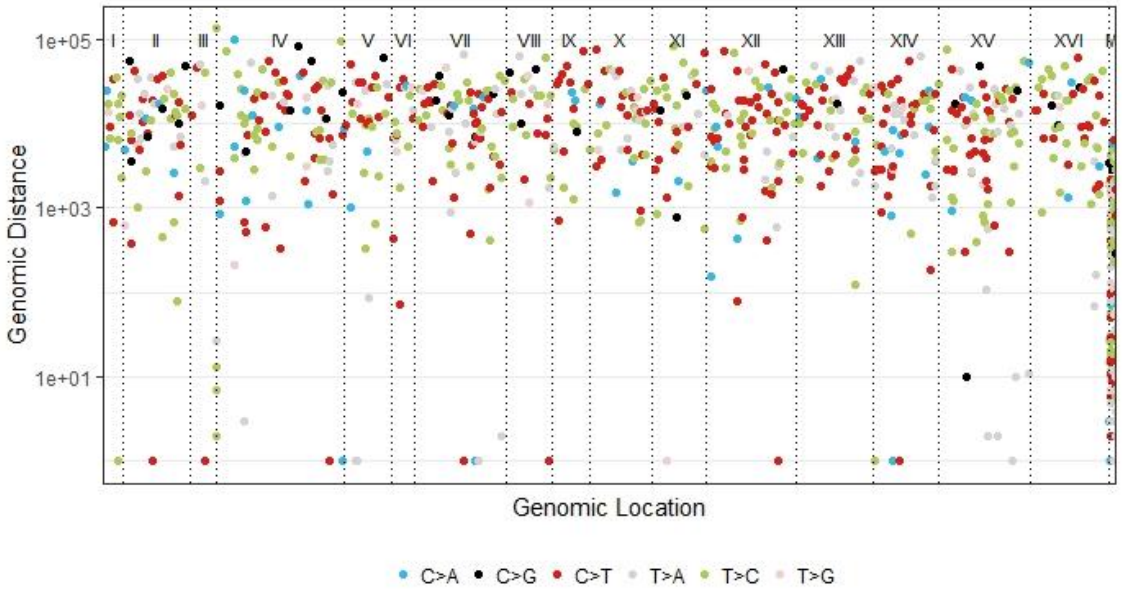


Figure 5.14: Rainfall plot of *rad26* mutant cells after UV damage induction, showing the genomic location of mutation with their inter-mutational distance. The colour dot represents 6 possible types of substitutions.

An interesting observation from this rainfall plot analysis is that, in *rad4* cells, without UV damage, the mutational pattern was dominated by C>A types of substitution (Figure 5.15), suggesting that this core NER factor is also involved in maintaining endogenously

induced DNA helix distorting damages. The overall mutational load is also higher in the *rad4* cells and this interesting phenomenon, shows the importance of Rad4 protein in NER for maintaining genome stability.

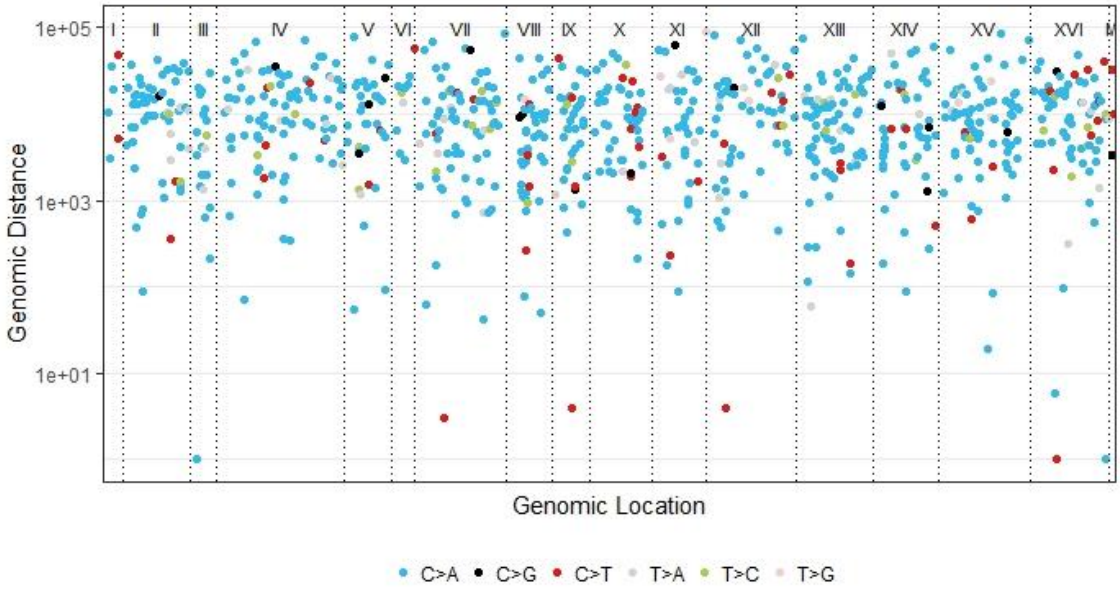


Figure 5.15: Rainfall plot of *rad4* mutant cells without UV damage, showing the genomic location of mutation with their inter-mutational distance. The colour dot represents 6 possible types of substitutions.

5.3.11 The 96 trinucleotide mutation profile indicates variation in mutational pattern between NER defective cells

All six classes of base substitutions, their flanking sequence context and their relative contributions to each sample, showed that the mutational pattern in untreated cells, varies between GG-NER and TC-NER defective cells (Figure 5.16). The 96 trinucleotides mutation profiles of wild-type, GG-NER deficient cells showed a similar pattern, with only subtle differences between them (Figure 5.16). However, TC-NER deficient cells showed a distinctive pattern in comparison to wild-type and GG-NER defective cells. Wild-type and GG-NER along with core NER showed a predominance of C>A mutations in the context of C<u>C</u>N and T<u>C</u>N (mutated base underlined), with only a few other types of substitutions detected. However, TC-NER defective mutants showed an even contribution of C>A along with C>T, T>A and T>C mutations, with fewer contributions from other classes of substitutions. These results indicate that defective TC-NER contributes to unique types of mutational profile. Looking into further details, core NER deficiency and wild-type cells showed a similar mutational pattern in comparison to GG-

NER deficiency. Additionally, *rad7* and *rad16* cells showed a high degree of similarity, consistent with their function in a complex during the early stages of GG-NER.

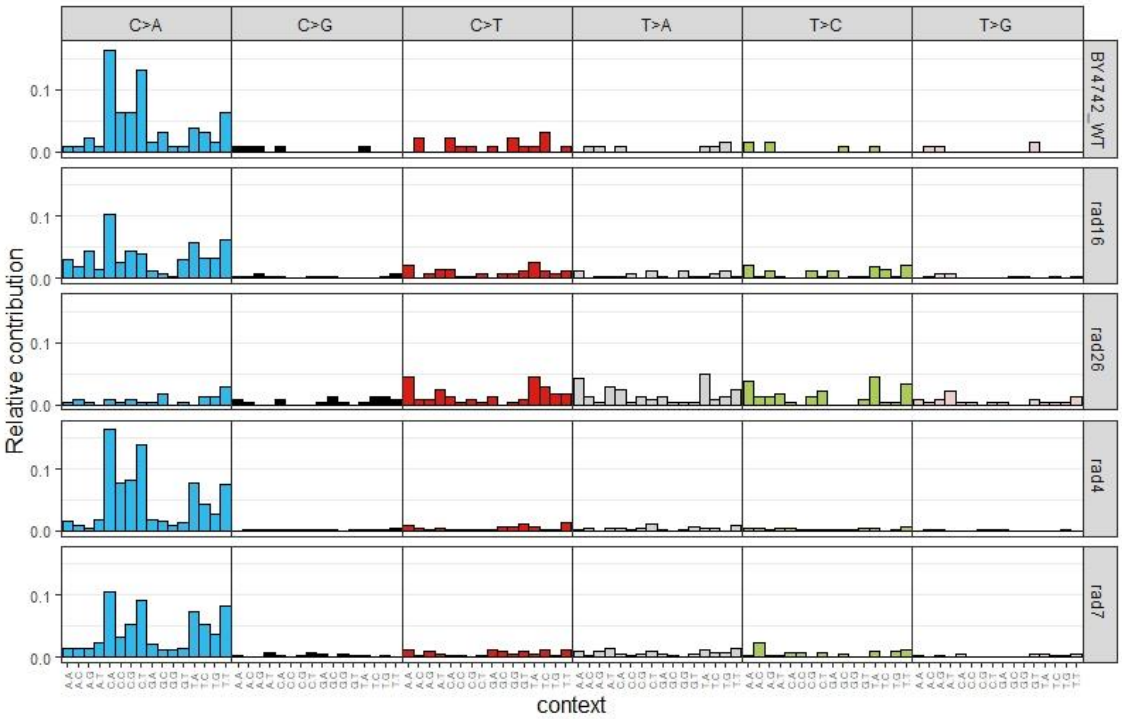


Figure 5.16: The relative contribution of 96 trinucleotides substitution mutation profile of wild-type and NER defective cells in the absence of exogeneous UV damage. The *rad7* and *rad16* mutant represent GG-NER deficiency and *rad26* mutant represent TC-NER deficiency, while *rad4* mutant represents as core NER defective cells.

Without UV damage, the GG-NER and TC-NER mutants showed distinctive patterns in their 96 trinucleotide mutation types. The *rad16* and *rad7* mutant cells shows higher similarity in their mutational pattern, compared to *rad26* mutant cell, reflecting the significant differences in the function of the GG-NER and TC-NER sub-pathways for dealing with the repair of endogenously produced DNA damages.

However, following UV damage, both GG-NER and TC-NER deficient cells show a similar mutational profile with subtle differences between them (Figure 5.17). The *rad16* and *rad7* mutants exhibited higher similarity in their mutation profiles to each other than to *rad26* mutant cells, with UV-induced predominance of C>T at C<u>C</u>N (mutated base underlined), T>C at T<u>T</u>N and T>A at T<u>T</u>A and A<u>T</u>A trinucleotides. In *rad26* cells, the significant differences observed around C>T at C<u>C</u>N (mutated base underlined) and relative predominance of T>A types of substitutions around N<u>T</u>N sites.

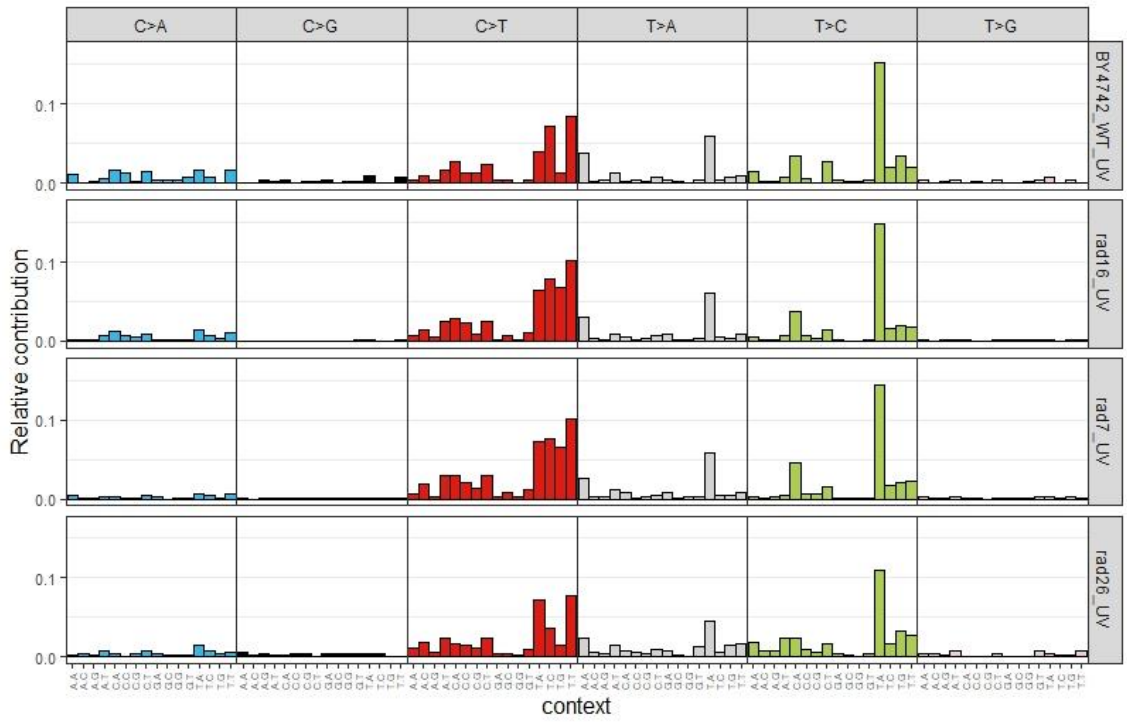


Figure 5.17: The relative contribution of 96 trinucleotide substitution mutations of wild-type, GG-NER defective *rad7*, *rad16* mutants and TC-NER deficient *rad26* mutant following exposure to UV damage.

5.3.12 The similarity of the 96 trinucleotide mutational profile to PCAWG signatures

The cosine similarity between each samples' 96 trinucleotide mutational profile and PCAWG signatures, reflects which cancer signatures showed the highest similarity to each samples' mutational profile. Figure 5.18 shows the heatmap representation of the similarity between each samples' mutational profile and PCAWG signature profile.

Strikingly, PCAWG single base substitution signatures (SBS)7a, SBS7b, SBS7c & SBS7d were mostly similar with UV-treated yeast strains, and this mutational signature is predicted to be due to exposure to UV light in malignant melanoma tumours (Alexandrov et al. 2018). The cosine similarity of wild-type and PCAWG mutational signatures were described in previous chapter. The core NER defective (*rad4* mutant) cells showed a similar pattern of mutation to wild-type cells, although the mutational loads was ~7 times higher in defective NER cells (Table 5.1). This shows that core NER defective cells generate a similar pattern of mutational profile from endogenous damage, differing only in the mutational load within the yeast genome. GG-NER deficient cells also generate similar mutational patterns as core NER deficient cells, with a higher similarity to SBS4 and SBS38 followed by SBS10a, SBS14, SBS18, SBS20, SBS24,

SBS29 SBS35, SBS36 and SBS40. The proposed aetiology of these signatures were various induced oxidative and/or replicative processes (Table A4.1 in Appendix-IV) (Alexandrov et al. 2018).

Interestingly, TC-NER deficiency showed a higher similarity with the PCAWG SBS5, SBS40 and SBS25 and a lower similarity with SBS41, SBS3, SBS8, SBS9, SBS12, SBS16 and SBS37. The probable association of most of these cryptic cancer signatures are unknown. This suggest that defective TC-NER might drive the biological processes responsible for generating these cryptic mutational signatures observed in cancer genomes, whose biological processes are currently unknown (Table A4.1 in Appendix-IV) (Alexandrov et al. 2018).

The hierarchical clustering of the samples shows closely related biological processes, for example *rad7* and *rad16* cluster together because of their biological role in the early stage of GG-NER involved in DNA damage recognition (Yu et al. 2016). There is stark difference in sample clustering between UV-treated and untreated cells. Remarkably, *rad26*, which represents defective TC-NER, clusters on its own and displays a unique mutational profile (Figure 5.18).

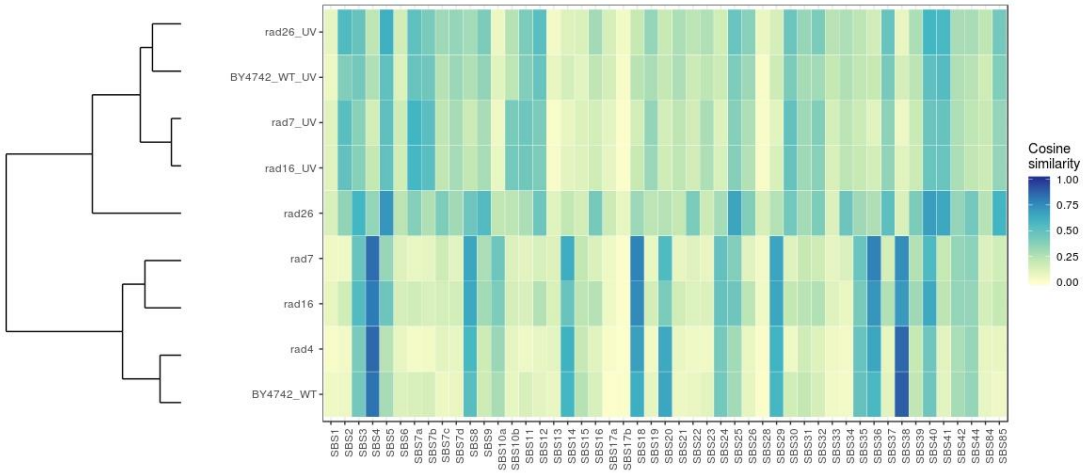


Figure 5.18: Heatmap of the cosine similarity between the mutational profile of individual samples with PCAWG signatures. The samples are hierarchically clustered (average linkage) using the Euclidean distance between the vectors of cosine similarities with the signatures.

5.3.13 Optimum contribution of PCAWG signatures to reconstruct individual samples for the 96 trinucleotides mutational profile

As described in the previous chapter, one way to investigate the contribution of the previously identified cryptic PCAWG mutational signatures in cells with altered DNA

damage and repair, is to reconstruct the individual samples' mutational profile using the existing mutational signatures. This will provide an indication of the molecular process that generate the samples' mutational profile (Blokzijl et al. 2018). Figure 5.19 shows that the mutational landscape of undamaged wild-type, *rad4*, *rad7* and *rad16* can be predominantly reconstructed with the contribution of PCAWG SBS4, SBS36 and SBS38. The proposed aetiology, for the SBS4 is 'exposure to tobacco smoke', for SBS36 it is 'defective base excision repair and MUTYH mutation' and for SBS38 is 'indirect effects of UV light'. All of these processes generate predominantly C>A (G>T) types of substitutions mutations within the genome by either directly or indirectly inducing DNA damage. This observation provides supporting evidence for endogenously produced oxidative type DNA damages involved in generating these mutational profiles. The mutational profile of *rad26* mutant cells showed unique characteristics, and can be reconstructed by those PCAWG SBSs whose origin is mostly unknown (Alexandrov et al. 2018) (Appendix-IV, Table A4.1). This suggest that the biological processes of these unknown cryptic cancer mutation signatures may be due, to some extent, to defective TC-NER processes. In response to UV damage, the landscape of wild-type, *rad26*, *rad7* and *rad16* can be reconstructed with contributions of SBS2, SBS7b, SBS12 and SBS41. The proposed aetiology of SBS2 is 'APOBEC activity', which is member of the family of cytidine deaminases, which convert the cytosine nucleotide to Uracil; SBS7b is 'UV exposure', whereas the aetiology of SBS12 and SBS41 are unknown (Table A4.1 in Appendix-IV). Not all PCAWG signatures that are similar to the *de novo* extracted signatures are required to reconstruct a mutational profile. This is because PCAWG mutational signatures are not independent (Appendix-III, Figure A3.5).

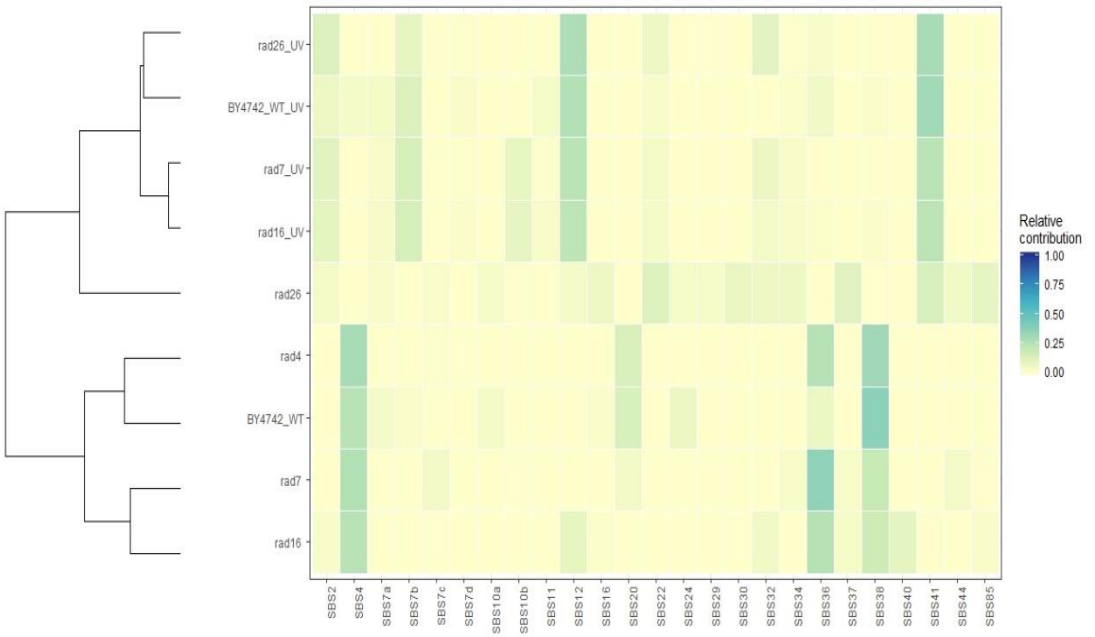


Figure 5.19: The optimal relative contribution of PCAWG signatures to reconstruct the mutational profiles of the wild-type, GG-NER and TC-NER experimental samples with or without UV exposure. The samples were hierarchically clustered (average linkage) using the Euclidean distance between the vectors of cosine similarities with the signatures. The PCAWG signatures with ≥ 10% contribution in at least one of the experimental samples were plotted.

To investigate whether each samples' mutational profile can be reconstructed by the provided mutational signatures, the cosine similarity was calculated between the original and the reconstructed mutational profile. The mutational profiles of most samples were not reconstructed very well with the PCAWG signatures ($\alpha < 0.95$, Figure 5.20), while some samples (*rad4, rad7,* and wild-type) can be reconstructed ($\alpha > 0.95$, Figure 5.20) with high confidence. A low similarity between the original and the reconstructed profile indicates that the analysed mutational profile cannot be fully explained by the provided PCAWG signatures, which suggests that additional, unassessed mutational processes might underlie the observed catalogue of somatic mutations.

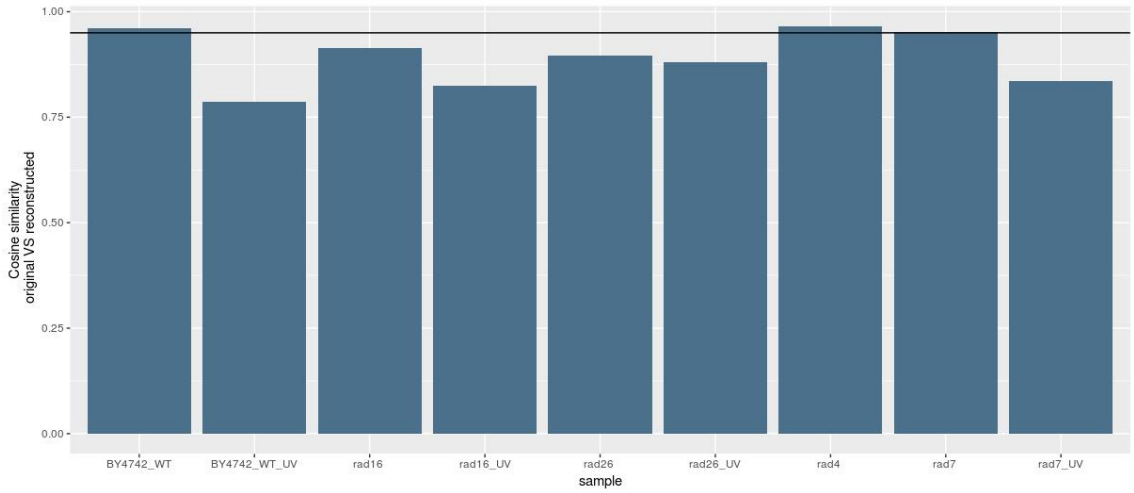


Figure 5.20: Cosine similarity between the original mutational profile and the reconstructed mutational profile based on the optimal linear contribution of all 49 PCAWG signatures. The line indicates the threshold of cosine similarity = 0.95.

5.3.14 *De novo* mutational signature extraction delineates the active biological processes

As mentioned in the previous chapter, the non-negative matrix factorization was used to extract mutational signatures from the catalogue of experimentally generated mutational profile. This signatures will represent the biological processes active in a cohort of samples. The individual conditioned tested in this chapter are represented by the samples of wild-type cells (i) untreated and (ii) treated with UV; the GG-NER mutants (*rad7/rad16*) (iii) untreated and (iv) treated with UV; the TC-NER mutant (*rad26*) (v) untreated or (vi) treated with UV and core NER mutant (*rad4*) (vii) untreated. The biological processes are represented by (a) normal cellular metabolic process-induced mutagenesis after 30 passage in all untreated samples, (b) UV-induced mutagenesis in the UV-treated samples after 30 passage and (c) the deficiency in NER sub-pathways induced mutagenesis during this experimental time period (*rad4*, *rad7*, *rad16*, *rad26*).

Following plotting the individual clones mutational catalogue as 96 trinucleotide substitution matrix (10 clone form each treatment and 9 treatment generated a [96x90] matrix). The factorization rank survey (Figure 5.21) and consensus heatmap clustering (Figure 5.22) was generated for rank between 2 to 8 to obtain best value of rank (N) (for details please see Chapter 4, section 4.14). Based on the factorization rank survey, the third rank showed the highest cophenetic coefficient. These observations suggest that three mutational signatures can be extracted with high confidence from experimental samples mutational profile under study.

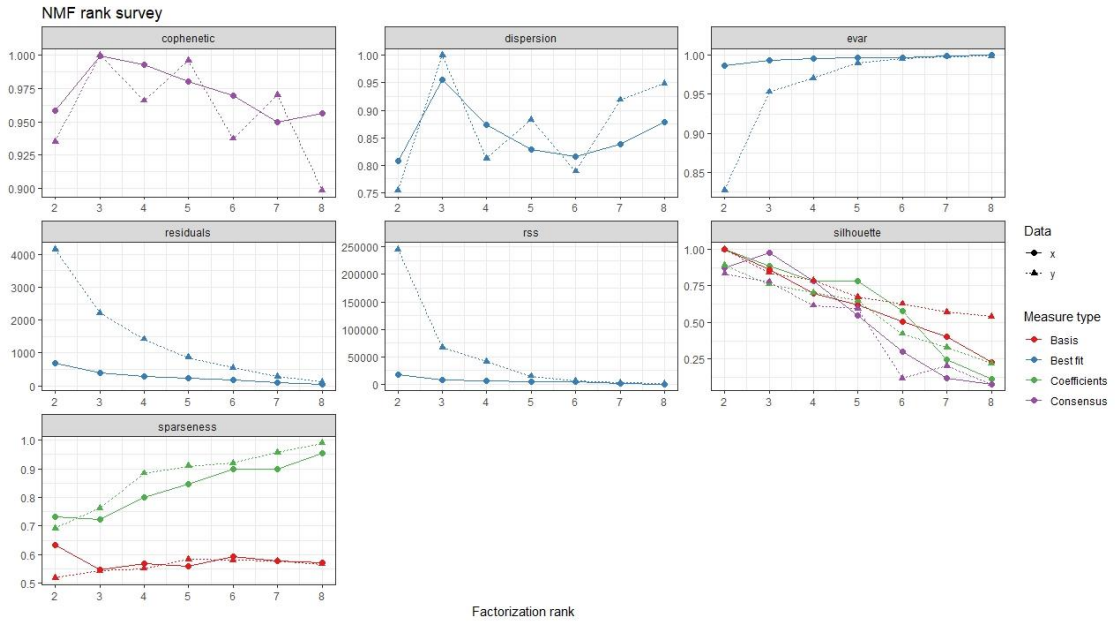


Figure 5.21: Factorization rank survey showing the rank or number of mutational signatures can be extracted from mutational catalogue matrix. For estimation of the rank, mentioned quality measures were computed from 50 runs for each value of rank in both samples and randomised data. The estimation is based on Brunet's algorithm. The data marked as 'X', represents the experimental data set, and the data marked as 'Y' represents the same data following randomization.

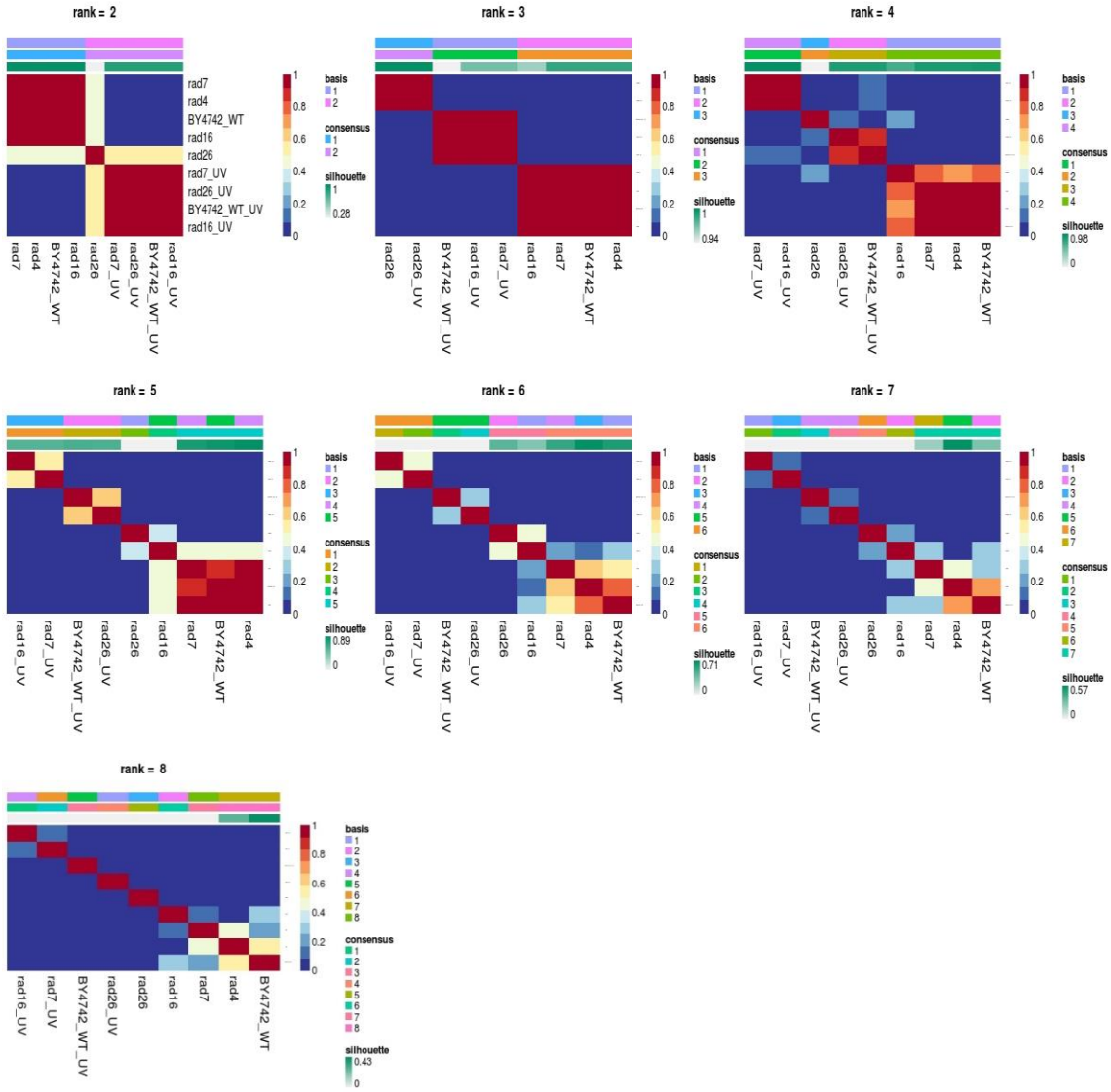


Figure 5.22: Heatmap of consensus matrix of possible factorization rank from 50 runs for each value of the rank mentioned in figure 5.21

Following the suggested best factorization rank, three mutational signatures were extracted from the samples 96 trinucleotide substitution matrix. These three signatures are name as signatures A, B and C. The resulting 96 trinucleotide mutational profile matrix of the 3 *de novo* extracted mutational signatures are plotted in Figure 5.23.

Signature A is characterised by predominant C>T at TCN trinucleotides (the mutated base is underlined) and T>C at TTN trinucleotides with small amount of T>A and C>A substitutions, indicative of UV induced mutagenesis, similar to that observed the previous chapter where it was designated signature C (UV signature, Cosine similarity > 0.95).

Signature B is characterised by C>A types of substitution at CCN and TCN trinucleotides with minimum contribution of other types of substitution mutations, which is similar to

signature A in the previous chapter (cosine similarity >0.95). This indicates that the oxidative damage by normal cellular metabolic processes in untreated cells.

Signatures C is characterised by a relatively even distribution of mutations across the 96 possible trinucleotide. Similar type of signatures were found in a recent repertoire of cancer genomes mutational signature analysis, in which SBS5 and SBS40 showed this flat or featureless phenomenon (Alexandrov et al. 2018). Interestingly, the biological process of these flat like signatures are unknown.

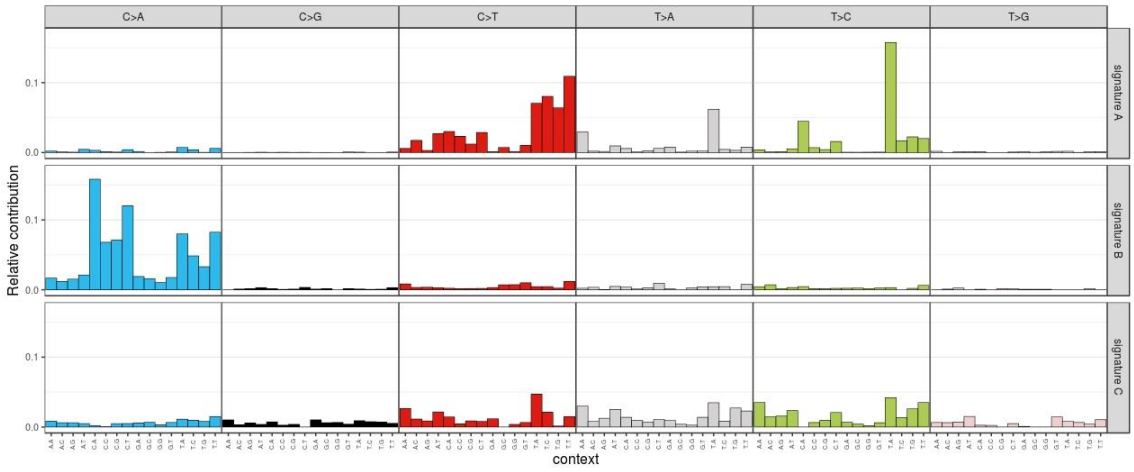


Figure 5.23: Relative contribution of indicated 96 trinucleotides mutations to the three mutational signatures that were extracted *de novo* by NMF analysis of the acquired mutational catalogue of the experimental model system. The wild-type and NER defectives cells' mutational profile with or without UV exposure were considered for *de novo* mutational signatures extraction.

NMF was also used to estimate the contribution of each *de novo* extracted mutational signatures to the experimental samples mutational catalogue. The result indicates that, multiple mutational processes contribute to each of the experimental samples, although in some case one process is dominant (Figure 5.24).

The experimentally generated mutational signature A, which is thought to be a UV induced signature from the Chapter-IV, section 4.14, contributes predominantly to the samples exposed to UV. This observation again confirms that the mutational signature A represent UV induced biological process.

Experimentally generated mutational signature B predominantly contributes to undamaged wild-type and GG-NER deficient cells. This suggest that similar biological processes are active in wild-type and GG-NER defective cells, for dealing with endogenous DNA damages, leading to subtle variations between the final mutational

load. Conversely, this observation demonstrates that different biological process (wild-type and GG-NER) can generate the same mutational pattern, although the total mutational loads differ between undamaged wild-type, *rad4*, *rad7*, *rad16* mutants.

Experimentally generated mutational signature C, which predominantly contributes to the TC-NER defective *rad26* mutant cells with fewer contribution to wild-type, *rad7* and undamaged *rad16* cells, suggesting the biological process of defective TC-NER.

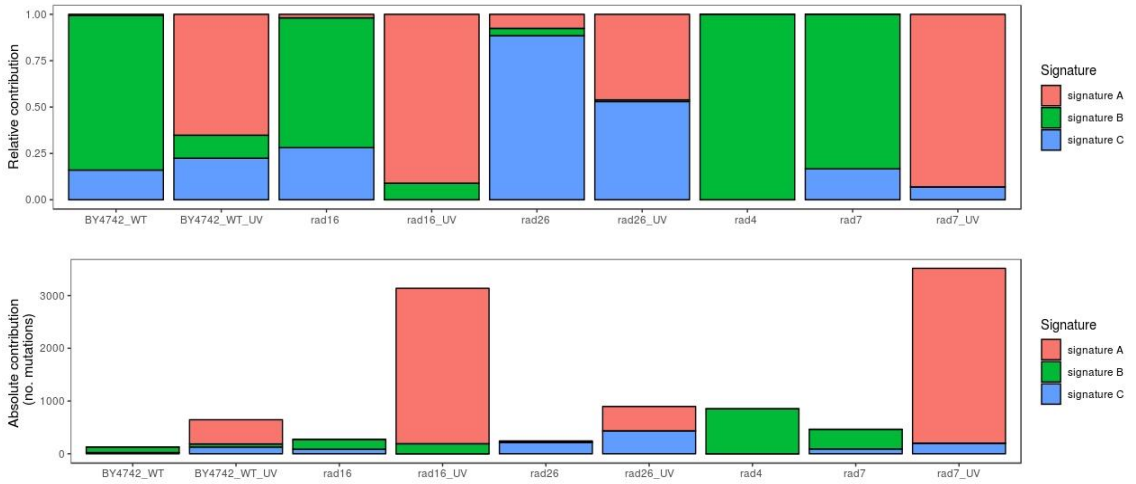


Figure 5.24: Relative and absolute contribution of each mutational signatures in each wild-type and NER defective yeast cells mutational profile with or without exposure to UV damage.

Remarkably, hierarchical clustering by using relative contribution of each of the three experimentally extracted signatures to each samples mutational catalogue as features shows two main clusters and one sub-cluster (Figure 5.25). This demonstrate that, three active biological processes were involved in this experiment depicted by NMF algorithm. These includes the oxidative damage during cellular metabolic process, UV induced DNA damage and TC-NER repair deficiency. Surprisingly, GG-NER does not generates any separate mutational process, although defect in GG-NER fasten both normal metabolic process- induced and UV-induced mutagenesis within yeast genome.

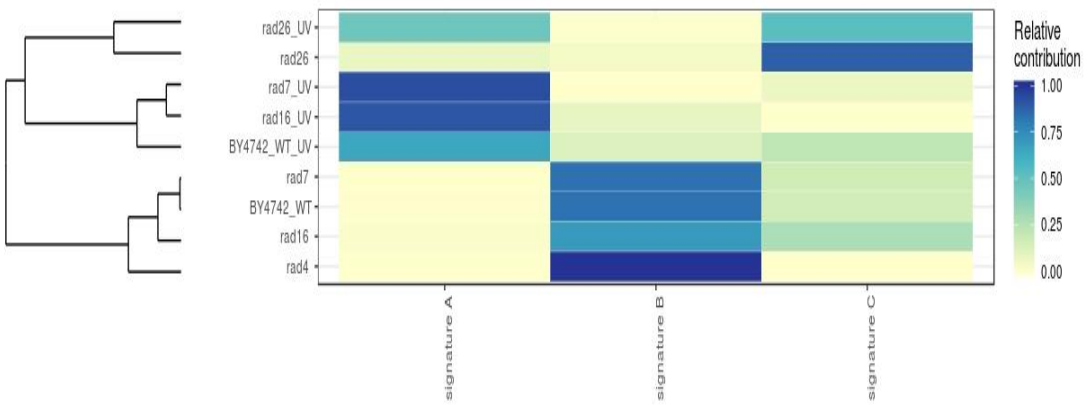


Figure 5.25: Heatmap showing relative contribution of de novo extracted signatures into individual samples. The samples are hierarchically clustered (average linkage) using the Euclidean distance between the vectors of relative contrition with the signatures.

5.3.15 The cosine similarity of the *de novo* extracted signatures with PCAWG signatures

The cosine similarity test was again applied in order to compare the *de novo* extracted signatures described above with the known PCWAG signatures. Experimentally generated mutational signature A (*UV signature, contribution mostly to UV-exposed sample and same as previous chapter signature C*), again displayed moderate similarity with PCAWG SBS7a & SBS7b, along with SBS2, SBS5, SBS40 and SBS41 (Figure 5.26). The underlying biological processes of SBS7a, & SBS7b were due to UV light, but the aetiology of the rest of the mutational signature are unknown. As mention in previous chapter, these moderate similarity with known PCAWG signatures associated with UV-exposure probably due to types of UV light used for this experiments and heterogeneity in melanoma cancer patients.

Experimentally generated mutational signature B is highly similar to PCAWG SBS4 and SBS38 (cosine similarity ~ 0.9), which were attributed to tobacco smoking and indirect effect of UV light respectively (Figure 5.26). Additionally, relatively low similarity also observed with other SBS signatures such as SBS8, SBS10, SBS14, SBS18, SBS20, SBS24, SBS29, SBS34, SBS36 and SBS40 (Figure 5.26). As mentioned in the previous section, signature B is predominantly contributing to unirradiated wild-type cells (*similar to signatures A in previous chapter*) and GG-NER defective cells, indicating that, the oxidative process during normal cellular metabolic activities. Similar findings were also mentioned using isogenic human cell-based models (Zou et al. 2018), in which the wild-type parental clone showed similar mutational pattern.

Experimentally generated mutational signature C, showed similarity mostly with PCAWG SBS5 & SBS40, along with SBS3, SBS12, SBS25 and SBS37 (Figure 5.26). The underlying biological processes of most of these PCAWG signatures are unknown. As mention in previous section, signature C is predominantly contributing to both unirradiated and irradiated TC-NER defective cells, these observations suggest that defective TC-NER could be the main aetiology of these mutational signatures observed in cancer, whose biological processes are not determined yet.

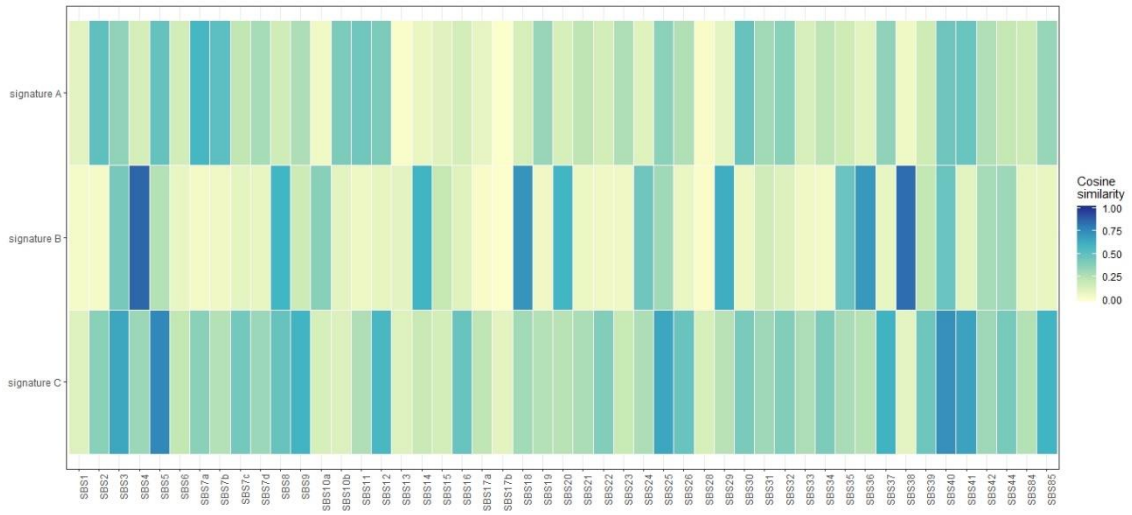


Figure 5.26: Heatmap of cosine similarities between the de novo extracted signature from wild-type and NER defective yeast cells with or without UV exposures and know PCAWG signatures from cancer gnomes.

5.4 Summary

In this chapter, the same approach described in Chapter IV was used to measure the catalogue of acquired mutations from both GG-NER (rad7/rad16 mutants) and TC-NER (rad26 mutants) defective yeast strains, with or without UV exposures over ~1100 generations. Using a similar process, the mutational profile of undamaged *rad4* mutant was examined as a control for core-NER deficiency, and wild-type yeast cells in the similar genetic background for comparative study is also included.

The key findings of this chapter reveals that the heterogenous distribution of mutational patterns induced by either UV exposure or normal cellular metabolic processes is determined by the alterations in the genome-wide relative repair rates in both wild-type and GG-NER deficient cells. Additionally, this study also indicates the presence of a novel mutational signature, caused by the biological process of TC-NER deficiency, which correlates well with those PCAWG signatures (SBS7 and SBS40), whose biological processes are currently unknown.

As mentioned previously, NER recognised a broad range of lesions including both oxidative damage induced by normal cellular metabolic processes or damage induced by variety of carcinogens such as UV light. The two sub-pathways of NER vary depending on how they detect damage in the first place. In TC-NER these damage recognition processes are facilitated by coupling the stalling of RNA pol II and subsequent recruitment of NER factors to these sites. Rad26 is believed to be involved in this early stages of TC-NER. The GG-NER complex factors Rad16 and Rad7 are also involved in early stages of the GG-NER process. However, less is known about what the mutational outcome when these factors are missing. To address this question, the genomic distribution of mutational patterns, with or without UV damages, were examined. The findings demonstrate that the mutational output varies significantly between the GG-NER and TC-NER defective cells. The additional observations for the core NER defective *rad4* mutant, with significantly higher but similar mutational patterns to that found in wild-type undamaged cells, demonstrate that defective NER increases the accumulation of mutations within the genome.

Overall these results show that the repair kinetics and differential DNA repair rates, as well as competition between GG-NER and TC-NER determine the mutational heterogeneity observed within various genomic features in the yeast genome. This study adds mechanistic insight, showing the negative correlation between genome-wide relative

repair rates and acquired mutational density, results in lower mutational load within the genome where the relative repair rate is high. At this moment, we don't know what causes the higher repair rates at 3´ end of a genes in GG-NER defective cell (Figure 5.3). At the same time, we currently don't know the pattern of repair rates following UV exposure in TC-NER defective cells. Future work will aim to analyse the TC-NER deficiency, as well for getting the complete picture of how NER operates within the chromatin environment, and how these processes determine the mutational patterns observed within the yeast genome.

The organisation of GG-NER in the genome is not simply a common feature associated with all NFRs. As mentioned in Chapter III, the relative rates at nonGCBS-containing NFRs in *rad16* mutant cells is affected differently, compared to GCBS-containing NFRs. The relative repair rates at non-GCBS NFRs in *rad16* mutant cells are also reduced, but not in the same way as in GCBSs in comparison to wild-type cells following UV exposure (Figure 3.15). The significant enrichments in the distribution of UV-induced substitutions in both *rad16* and *rad7* mutants around the GCBS-containing NFRs signify that the defect in GG-NER repair organisation around these sites results in increased numbers of observed mutations. This also shows that, structure and organisation of GG-NER is important to maintaining the genetic stability within yeast genome. This might explain the non-random distribution of mutational patterns reported in human melanoma cancer (Pleasance et al. 2010a). Additionally, significant mutational bias is observed around Micro-C boundaries, the regions of higher order nucleosome-nucleosome interactions, following UV irradiation. Most of the newly identified classes of GG-NER complex binding sites are located around these boundaries, and following UV damage, GG-NER complex-mediated nucleosome remodeling in the vicinity of the GCBSs is required for efficient repair. This demonstrates that defective GG-NER results in higher mutational load at these sites within yeast genome.

Transcriptional strand asymmetry in the distribution of mutations is one of the common cancer genome signatures (Haradhvala et al. 2016). This study provides insight into how various types of UV-induced mutations are distributed within transcribed and non-transcribed strands in yeast genome. My study also provides evidence that active TC-NER causes mutational strand asymmetry, also frequently observed in cancer genomes, particularly in melanoma. This study also shows that, the types of mutations vary between transcribed and untranscribed strand; indicating that, the initial UV-induced DNA damage recognition by either TC-NER or GG-NER sub-pathway might be the cause of

mutational strand asymmetry. Additionally, because of repair timing differences between transcribed and untranscribed strand as well as the different repair kinetics of various types of UV-induced damages by both of these sub-pathways (Adar et al. 2016), this might determine the particular patterns observed following UV exposure. Further study targeting strand specific repair kinetics analysis are needed to investigate this in more detail.

The unique mutational pattern generated by TC-NER deficiency in both undamaged and UV damaged cells resulting in novel mutational signatures, was characterised by even distribution of all possible types of base substitutions. In contrast, similar mutational patterns in wild-type, GG-NER deficiency and core NER deficiency indicate the similar biological process of mutational pattern generation occurs in both wild-type and core-NER deficient cells. However, the higher mutational load in core NER or GG-NER deficient cells demonstrates how repair defects increases the accumulation of mutations within the various genomic locations in yeast.

As described in chapter IV, NMF was successfully employed to decompose mutational signatures from a cohort of samples. The factorization rank survey and consensus heatmap test validated the existence of three mutational signatures. Two of them similar to those found in Chapter-IV and one novel signature. The biological process for this novel signature is defective TC-NER factor in yeast Rad26, which is the human homologue of CSB and involved in repair of both UV induced and oxidative DNA lesions in the nucleus as well as in mitochondria (Melis et al. 2013). CSB, a SWI/SNF ATPase-containing chromatin remodeling factor, in human cells plays crucial roles, not only in TC-NER, but also in RNA pol II transcription activities. Significant differences are observed in the mutagenesis mechanisms between the GG-NER and TC-NER sub-pathways when dealing with the repair of both endogenous and exogenously induced DNA damages. Failure to recover RNA synthesis is the primary molecular phenotype of rad26 mutants in yeast cells, and this is also the case for Cockayne's syndrome B patients, where defects occurs in the CSB gene; the human homologue of yeast Rad26.

To conclude this chapter, the structure and organisation of repair play important role in maintaining genome stability. UV is a strong mutagen. Even in wild-type cells, while the repair process is intact, UV-induced mutations are predominant in these cells. The distribution of UV-induced mutational pattern in wild-type cells depends on the structure and organisation of repair processes. Loss of organised repair, such as defective GG-NER and TC-NER, or core NER factors, alters and/or increases the accumulation and

distribution of mutations within yeast genome, even without exogenous genotoxic stress. This process results in similar or distinct mutational signatures detected within yeast genome, depending on the repair factors involved in the repair of the endogenously induced DNA damage. Following exogeneous UV-induced DNA damages, again defective GG-NER or TC-NER affects the wild type accumulation and distribution of UV induced mutations, resulting in distinct UV-induced mutational signatures within yeast genome.

UV-induced mutations are more difficult to repair at low-acetylated genomic regions, which are represent a more closed chromatin than high-acetylated regions. Recent studies suggest that UV induced damages are uniformly distributed in relation to linear genomic structure (Teng et al. 2011; Hu et al. 2017), However, the repair rate of these damages is modulated by the accessibility of the damaged DNA, which is determined by acetylation status of chromatin, as well as histone variant exchange to facilitate repair (Yu et al. 2011; Yu et al. 2016; Hu et al. 2017). This variation in the repair rate due to accessibility, might also determine the mutational distribution observed within cells. This question will be addressed in next chapter.

Chapter VI

Chromatin modification and variant exchange influences the genomic location of mutational distribution

Contents

6.1 Background

Genomic instability is an established hallmark of cancer and defective DNA repair is a major contributor to its cause. DNA repair pathways are integrated within a system-wide process known as the DNA Damage Response (DDR). In this system, DNA damage sensors detect chromatin-associated DNA damage signals, which ultimately determine the physiological response of the cell to DNA damage (Lazzaro et al. 2009). Therefore, determining how DNA damage in chromatin gets efficiently repaired, and how these events contribute to the mutational endpoint, is fundamental to understanding the mechanisms that underpin the relationship between genome stability and human health.

The compaction of chromatin plays an important role in determining the region of DNA where damages occurs (Freeman and Ryan 1990), the accessibility of the repair factors to the damaged DNA (Reed 2011; Yu et al. 2011) and therefore has the potential to influence the distribution of mutations. The target of UV damages are DNA molecules, which are wrapped in chromatin structure. The basic unit of chromatin is known as the nucleosome which is composed of a histone octamer enveloped by 147 bp of DNA and variable length linker DNA connecting the adjacent nucleosomes. In addition to the canonical histones H2A, H2B, H3 and H4, histone variants such as H2A.Z and H3.3 also exist in yeast (Gurard-Levin et al. 2014). It is known that, the physical arrangement of the nucleosomes in the genome provides an important framework that supports the ordered modification of histone tails and variant turnover (Polo 2015). The chemical modification of N-terminal tails of histone such as acetylation, methylation, phosphorylation, ubiquitination and indeed the exchange of histone variants within the octamer structure alters the physiochemical properties of the nucleosome and regulates chromatin-associated biological functions within the cells, including DNA replication, transcription and repair (Lai and Pugh 2017). The integrity of the genome is constantly insulted by genotoxic agents. Inefficient or inaccurate repair of these lesions induced, poses a serious threat to genome stability and can lead to cancer development (Nair et al. 2017). For efficient repair, cells have developed several ATP-dependent chromatin remodeling and post transcriptional modification of histone to modulate chromatin structure necessary for accessibility of the repair machinery to the damage embedded in the chromatin (Lans et al. 2012). Defects in such regulatory processes modulate the mutational endpoint and are also implicated in diseases associated with ageing, including cancer (Luijsterburg and van Attikum 2011).

NER is the sole mechanism for removing helix distorting bulky adducts from the genome, including those formed endogenously by oxidative or hydrolytic processes and exogenously by UV damage or chemotherapeutic drugs such as cisplatin or oxaliplatin. It is well established that the NER efficiency is affected by nucleosome structure both *in vivo* (Hara et al. 2000; Liu 2015) and *in vitro* (Nag and Smerdon 2009), suggesting that even the basic level of chromatin compaction constitutes a hindrance to repair. To study how repair is initiated in chromatin, it has been recently demonstrated that global genome nucleotide excision repair (GG-NER) in chromatin is organised into domains around open reading frames (Yu et al. 2016). To identify these domains and the DNA damage-induced changes in the linear structure of nucleosomes, the process of chromatin remodeling during repair is demonstrated (van Eijk et al. 2018). In undamaged cells, the GG-NER complex occupies chromatin at nucleosome free regions (NFR) of specific gene promoters. This establishes the nucleosome structure at these genomic locations, which is refer to as GG-NER complex binding sites (GCBS's). These sites are frequently located at genomic boundaries or domains that delineate chromosomally interacting domains (CIDs). These boundaries define domains of higher-order nucleosome-nucleosome interaction. The efficient repair of DNA damage in chromatin is initiated following disruption of H2A.Z-containing nucleosomes adjacent to GCBSs by the GG-NER complex. Likewise, histone variant exchange is important during cellular processes including repair. To resolve the inhibitory effect of nucleosome on repair proteins, variant turnover is facilitated by either ATP dependent chromatin remodelers or acetylation of histone (Lai and Pugh 2017). In response to UV damage, Htz1 also promotes NER in yeast by enhancing the occupancy of HAT Gcn5 on chromatin to promote ionone H3 acetylation and open chromatin structure necessary for efficient repair (Yu et al. 2013). Loss of histone variant showed increase UV sensitivity and alter wild-type relative repair rates within genome in response to UV DNA damage (Figure 6.1), suggesting this might alter the distribution of mutation pattern within yeast gnome after UV damage. Taken together, these studies lead to ask how mutations are distributed across chromosomes and what are the factors that govern where the disease-causing mutation is prone to occur?

Recent cancer genome study proposed that, the variation in the distribution of mutations in genomic regions is primarily due to differential DNA repair capacity within genome. Suggesting the variation in the accessibility of the repair factors to the DNA caused by compaction of chromatin. It has been well established that, accessibility of the DNA by repair factors within chromatin is altered as a result of acetylation of histone by HAT,

Gcn5. In response to UV damage histone modification was necessary (Yu et al. 2016; Hodges et al. 2018) which determine open chromatin structure for efficient repair. Because, compete loss of UV induced H3AC at K9/14 after deleting GCN5, alter the wild-type repair rates within yeast genome (Yu et al. 2016) when the data was plotted around ORF. Additionally, proteins that modify nucleosomes also often associated with mutational heterogeneity in human cancer, suggesting that differences in repair due to histone modification may be an important contributor to heterogenous mutation rates.

Recent effort to measure and decipher the non-random nature of the mutational pattern that shapes the somatic cancer genome of different cancer types. This include efforts to explain the causes of these mutational pattern based on our current knowledge of DNA damage and repair mechanism (Haradhvala et al. 2016). More recently, genomic DNA repair rates have been correlated with the incidence of mutations in skin and other cancers, suggesting the cancer associated mutations occurs at the genomic regions which are difficult to repair (Sabarinathan et al. 2016). Additionally, by studying 23,000 tumours of 71 cancer types (Alexandrov et al. 2018), it has been reported around 49 Single-Base Substitution (SBS) signatures, together with 17 Indels (insertion and deletion mutation) signatures and 11 dinucleotide (tandem) mutation signatures. The biological processes behind most mutational signatures are not determined yet. Significantly, these studies have also revealed novel cancer genes, many of which are involved in chromatin remodeling and modification. We speculate that tumorigenesis in these sporadic cancers may be driven by mutations in these chromatin modifiers that disrupt the normal landscape of genome-wide DNA repair rates. This subsequently alters the distribution of mutations in the genome. These observations demonstrate the importance of understanding how genetic damage is formed and repaired in chromatin, throughout the entire genome. Knowing the underlying causes that give rise to cancer will permit a more accurate assessment of the risk of developing the disease, and aid in selecting and developing appropriate treatments for individuals.

Mutagenesis study in yeast model organism will provide the evidence in our current understanding of how the distribution and repair of UV-induced DNA damage influence mutagenesis in human skin cancers. These findings will reveal the influence of chromatin on repair and mutagenesis of DNA lesions on a genome-wide scale. In cancer genome, mutation density (like damage density) are highly heterogenous. It would be interesting to determine to what extent these mechanisms regulate mutation rate in human cancers,

since it is known that certain histone modifications correlate with mutation density in sequenced cancer genomes (Schuster-Böckler and Lehner 2012).

In Chapter III, I showed that, chromatin remodeling during repair of DNA damage by NER is initiated from GCBSs at boundaries of higher order chromatin structure. In undamaged cells, the GG-NER complex occupied at sites within chromatin bounded by histone variant, H2A.Z. Following UV damage, these histone variant contain boundary nucleosomes remodeled by GG-NER complex dependent manner for efficient repair. It also established from our laboratory previous study that, in response to UV damage, GG-NER complex regulate histone acetylation around these GG-NER complex bounding sites for generating open chromatin structure, required for efficient repair (Yu et al. 2011). Follow up study showed that, loss of either histone modifier (*gcn5* mutant) or histone variant (*htz1* mutant) alter the genome-wide wild-type relative repair rate (Yu et al. 2016). In Chapter IV, I established a protocol and bioinformatic pipeline for measuring the distribution of mutations observed in wild-type and repair defective yeast cell. In this chapter, I am going to use the same approach for accumulating and analysing the genome-wide distribution of mutations in *htz1* and *gcn5* mutant with or without exposure to UV damage. In the previous chapter, I provide the details outcome of how defects in NER pathway alter the genome-wide distribution of mutations in and around different genomic features. In this chapter I am going to explore, whether loss of histone variant (*HTZ1*) or histone acetyltransferase (*GCN5*) alter the distribution of mutations within yeast genome with or without UV exposure. This will enable us to determine how the alteration of the wild-type relative repair rates, due to defect in chromatin modification or variant exchange processes, affect the distribution of mutations within yeast genome. This might reveal, how defect in chromatin modification or remodeling drive tumorigenesis in cancer genome.

6.2 Material and Methods

Yeast strain used for this study

The yeast strains used for this study and their respective genotype is mentions in Table 6.1.

Table 6.1: Yeast strain and their respective genotypes used for this chapter.

Strain ID*	Function in	Genotype	Source
BY4742	Wild-type function	Wild-type MATa his3delta1 leu2delta0 ura3delta0	Euroscarf
gcn5	Histone Modification	BY4742 gcn5Δ	Reed Lab, CU
htz1	Transcription Regulation & nucleosome remodeling	BY4742 MATa his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0 htz1::kanMX4	Euroscarf

* All the yeast strain used in this experiment are haploid and matting type alpha.

For accumulation of mutations with or without UV irradiation, the propagation of cells through ~1,100 generation from passage 1 to passage 30 was performed by following the protocol described in section 4.2.1 of the Chapter IV. The materials and methods employed for growing the yeast strain used for this chapter and subsequent DNA extraction were performed by following the protocols as mentioned in the Material and Methods section Chapter-II. The sequencing library was prepared by following Illumina library preparation protocol and sequenced using Illumina sequencing platform (Details are mentioned in Chapter II). Whole genome sequence data was obtained by using Illumina Hi-Seq paired-end sequencing chemistry with read sizes of 75 bp. The raw paired-end sequences were processes using the same ages pipeline as mentioned in Chapter IV, section 4.2.2. The bioinformatics pipeline followed to accumulate acquired mutations using IsoMut was employed by following section 4.2.3, Chapter IV. All the code used to run IsoMut are attached as eAppendix. The background mutations detected were subtracted after plotting the tuning curves of all samples (Appendix V, Figure A5.1 – A5.4) as described in Chapter IV, section 4.2.4 to determine the cut-off value based on S score. The tuning curves were generated for both substitution mutations and Indels. For mapping substitution mutations according to different genomic features section 4.2.5 from Chapter IV was followed. Mutational signatures and cosine similarity analysis was then performed using the bioinformatic workflow as described in Chapter IV, section 4.2.6 and 4.2.7, respectively.

6.3 Results

6.3.1 Loss of chromatin components *GCN5* or *HTZ1* does not alter the total mutational load in presence or absence of UV damage.

Using the pipeline described in Chapter IV, I extracted the catalogue of mutations from the datasets derived from wild-type, *gcn5* and *htz1* mutant cells. We sequenced genomic DNA from both untreated and UV-irradiated cells and processed the data accordingly. An overview of the number of substitutions and Indels mutation accumulated in the wild-type, *gcn5* and *htz1* mutant yeast cells are provided in table 6.2. Please note that the total number of indels generated during this experiment in wild-type, *gcn5* and *htz1* mutant yeast cells is negligible as expected. Therefore, I am going to focus on single nucleotide variations or substitution mutations for subsequent analysis.

Table 6.2: Number of SVNs and short Indels in the wild-type, *gcn5* and *htz1* mutant yeast cells with or without UV damage.

Treatment*	Passage	n	Total SNVs	SNV Mean	Total Indels	Indels Means
BY4742_WT	Starting Clone	1	1	1	0	0
	End Clone	10	128	12.8	13	1.3
BY4742_WT_UV	Starting Clone	1	2	2	1	1
	End Clone	10	642	64.2	35	3.5
gcn5	Starting Clone	1	1	1	3	3.0
	End Clone	10	99	9.9	11	1.1
*gcn5*_UV	Starting Clone	1	1	1	3	3
	End Clone	10	570	57.0	24	2.4
htz1	Starting Clone	1	1	1	2	2.0
	End Clone	10	121	12.1	9	0.9
*htz1*_UV	Starting Clone	1	1	1	1	1
	End Clone	10	769	76.9	30	3

Abbreviations: WT, wild-type; *gcn5*, histone acetyltransferase mutant; UV, Ultra-violate; n=number of clones. SNVs = Single Nucleotide Variations. Indels = Insertions and deletions. Independent mutations in the starting clone represent false positives of the number of detections.

In undamaged cells, following 30 passages, the mutational load in wild-type cells is 128 SNVs from 10 clones after ~1,100 generations. During the same propagation time, the

mutational load in *gcn5* cells and *htz1* is 99 and 121, respectively (Table 6.2). These numbers indicate a similar total mutation load, demonstrating that these mutant backgrounds do not display an inherent mutator phenotype, as expected. Following UV damage, the mutational load increases to 642, 570 and 769 for the wild-type, *gcn5* and *htz1* mutants, respectively, which is ~5-fold higher in comparison to the data from the undamaged samples. These results confirm that UV irradiation is a strong mutagen (Ikehata and Ono 2011). However, similar increase in mutational load is observed between the wild-type, *gcn5* and *htz1* mutant yeast cells, even after UV exposure. This result is consistence with their UV sensitivity (Yu et al. 2013). Collectively, these initial findings demonstrate that substitution mutations increase after exposure to UV irradiation and that deletion of *GCN5* or *HTZ1* does not alter the total mutational load. However, relative repair rate analysis showed that, in response to UV damage, both of these mutants alter the wild-type relative repair rates within the yeast genome (Figure 6.1). These findings lead us to ask the question whether the *gcn5* or *htz1* mutants alter the locations of mutations induced within yeast genome by altering the genome-wide relative repair rates.

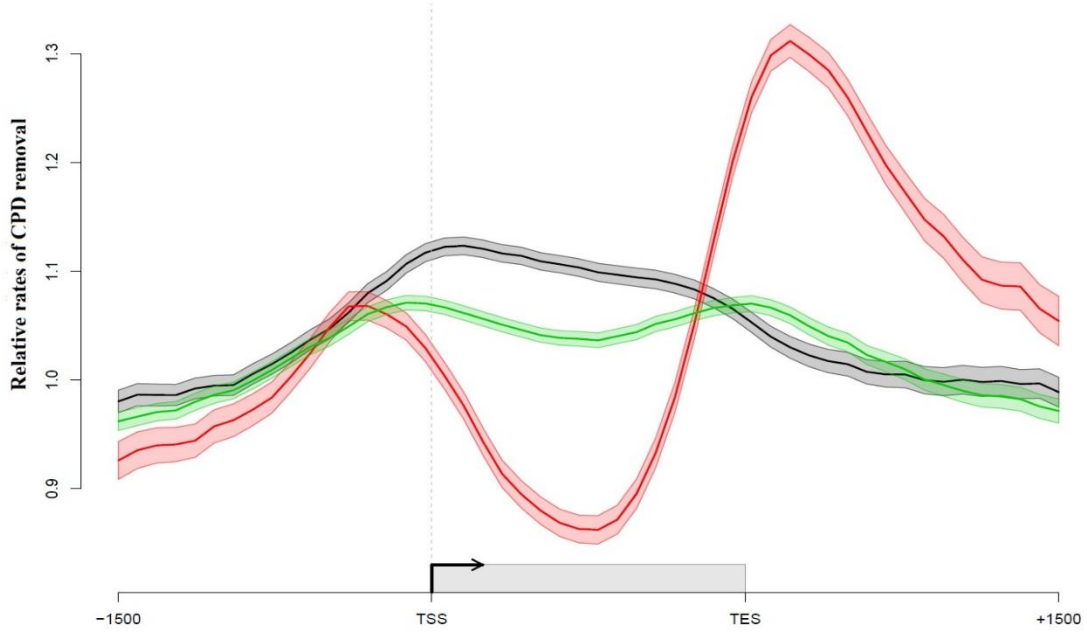


Figure 6.1: Relative rates of CPD repair around ORF structures. Solid lines show the mean of CPD repair rates in wild-type (n = 3, black), *gcn5* (n = 2, red) and *htz1* (n=2, green) mutant. Shaded areas indicate the standard deviation, with CPD levels plotted as arbitrary units on the y-axis. (Data used here for plotting was obtained from previous work by former colleagues in our laboratory. The composite ORF plot shown here, was made using Sandcastle (Bennett et al. 2015).

6.3.2 Distribution of mutations in relation to genomic features

To measure the distribution of mutations in relation to the linear arrangement of genomic structures, I calculated the observed mutations at all OFRs, TSSs, TESs, ARS from the *Saccharomyces* Genome Database (SGD) and compared them to the expected mutations based on mutation density and feature size (Figure 6.2, upper panel, as mentioned in chapter IV, section 4.3.2). The $\log_2$ ratio of observed over expected mutations from wild-type, *gcn5* and *htz1* mutant yeast cells were plotted for these genomic locations (Figure 6.2 lower panel). The variation in the total number of mutations in the different linear genomic structures are a function of the amount and size of these genomic, SGD features (Figure 6.2 upper panel).

As shown previously, the UV-induced accumulation of mutations is readily observed at all the features studied here. This is reflected also when considering the random sites. This set of negative control positions reveals the increase in UV-induced mutations but fails to enrich for mutations when comparing these sites to the number of expected mutations. Importantly, the number of UV-induced mutations observed at ORFs is consistently lower than expected in all backgrounds studied here (Figure 6.2, ORF panel). Interestingly, this difference is only significant in the *gcn5* mutant. The endogenous mutations induced around ORFs in the absence of UV irradiation are much lower (~5-fold as described) but show only minor differences between observed versus expected that are not significant.

Similarly, mutations detected around TSSs and TESs are only slightly different compared to the calculated, expected mutation load based on mutation frequency and cumulative feature size (Figure 6.2). Interestingly, I observe less mutations at TSSs in the *htz1* mutant cells than expected. It is important to note that the majority of H2A.Z containing nucleosomes are located at the +1 nucleosomes of most TSSs (Albert et al. 2007). Absence of this characteristic chromatin feature of TSSs alters the repair rates (Figure 6.1) and could also affect the mutation induction as a result. This makes sense in the context of the model for GG-NER chromatin remodeling we recently proposed (van Eijk et al. 2018). However, the depletion of mutations at TSS in these *htz1* mutant cells in not significant. It remains to be determined whether this difference in mutation induction at TSSs represent a biological outcome of the change in chromatin structure and repair rates in the absence of H2A.Z.

As described previously, all of the mutants analysed in this chapter have slightly higher than expected mutations at the TES. The characteristic structure of this portion of the genome makes it refractory to repair, thus resulting in a site that can accumulate mutations. It has to be noted that only the data for wild-type cells shows a significant difference based on the $\log_2$ ratio between the observed versus expected (Figure 6.2, TES panel).

Finally, the ARS positions, due to their low numbers (n = 352), are sites of low mutation load. Moreover, the observed mutations at these positions do not deviate substantially from the expected numbers (Figure 6.2, ARS panel). The $\log_2$ ratio for the *htz1* mutant data in the absence and presence of UV irradiation, however, shows a large depletion of mutations in this background. Importantly, these changes are not significant, potentially due to the lower numbers of mutations found at this small subsection of the genome.

Overall, no large bias in the mutational load is observed over the features represented in Figure 6.2. As explained before, in wild-type cells, significant bias in the observed mutational load in comparison to genome-wide expected mutation density is observed in and around TESs.

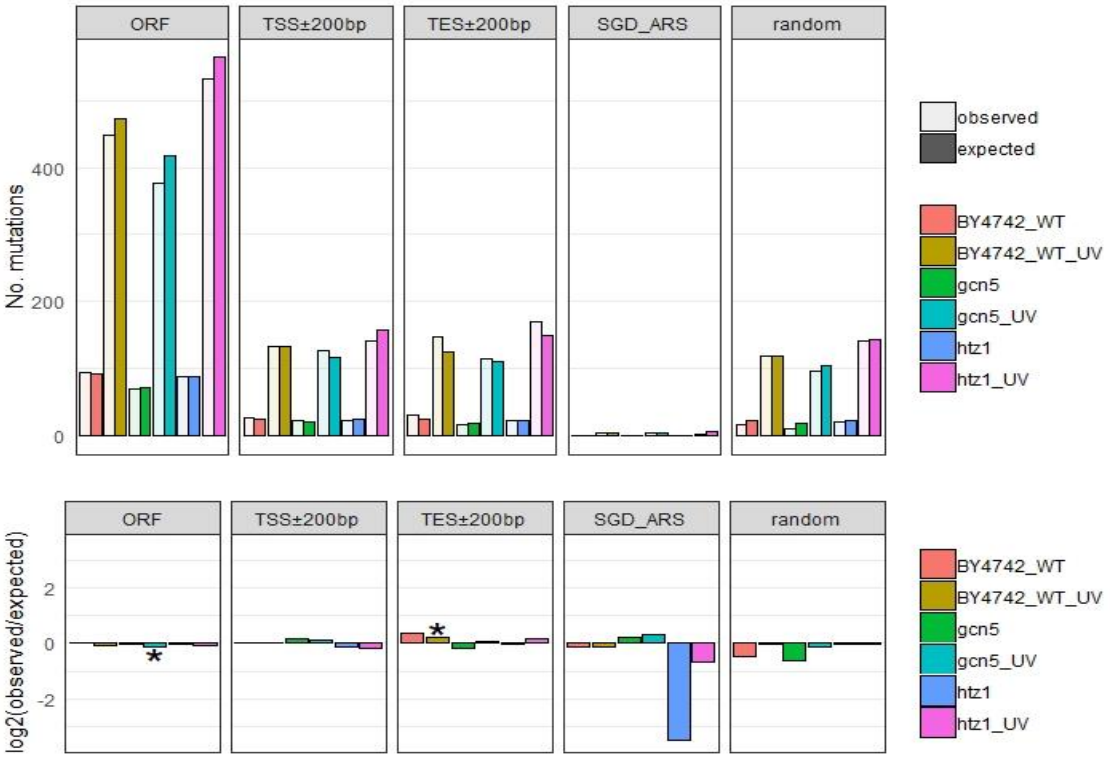


Figure 6.2: Enrichment and depletion of point mutations in linear genomic features such as ORF, TSS, and TES genomic sites for wild-type, *gcn5* and *htz1* mutant yeast cells with

or without UV damage. The $\log_2$ ratio of the number of observed and expected point mutations indicates the effect size of the enrichment or depletion in each region. Asterisks indicate significant enrichments or depletions ($P < 0.05$, one-sided binomial test).

When the data is plotted around linear chromatin structures, no significant bias in their observed mutational load is detected in comparison to genome-wide expected mutation density around NFRs (Figure 6.3, NFR panel). It is known that the majority of H2A.Z-containing nucleosomes are positioned next to NFRs (Albert et al. 2007) and that NFRs and H2A.Z are important regulatory components of chromatin that are intimately linked (Hartley and Madhani 2009; Weber et al. 2014). Therefore, it is conceivable that *HTZ1* deletion might influence mutation induction at NFRs. Indeed, in the absence of H2A.Z, mutations are depleted at NFRs, but this difference is not significant (figure 6.3, NFR panel).

On the other hand, around SPNs, detected mutations are consistently lower than expected and significantly so in the case of *gcn5* mutant (Figure 6.3, SPN panel). Surprisingly, this indicates that even though the nucleosome structure is inherently refractory to repair (Hara et al. 2000), the ~10,000 strongly positioned nucleosomes used in this analysis do not represent sites of hypermutation in the various genetic backgrounds tested here. Indeed, the mutation load is consistently *lower* than expected at these positions. It remains to be determined what the exact repair rate is at these nucleosomes in order to confirm whether the efficiency of repair can be correlated with mutation induction.

Finally, Micro-C boundaries were considered here. The cumulative mutations observed at these locations is compared to the expected mutation frequency at these sites. The resulting data reveal a persistent increase in UV-induced mutations over expected (Figure 6.3, Micro-C panel). Something about the structure of these genomic positions makes it so that mutations accumulate there. This difference in mutation load is high and significant in wild-type and *gcn5* mutant cells but smaller and not significant in *htz1* cells (Figure 6.3, Micro-C panel). Conversely, mutations induced in the absence of DNA damage do not accumulate at these sites as the difference is much smaller (wild-type) or even depleted in the case of *gcn5* and *htz1* mutant cells. This indicates that endogenous DNA damages and UV-induced DNA damage cause mutations differently when the absence of Gcn5 or Htz1 alters the chromatin and DNA repair rates. It is important to note that it is currently not possible to measure the repair of endogenous lesions by NER. Therefore, I cannot correlate the mutation induction in untreated cells described with the mutation accumulation or depletion detected here.

Figure 6.3: Enrichment and depletion of somatic point mutations in chromatin associated genomic features such as NFR, SPN_Dyad±75bp, Micro-C boundaries±200bp for wild-type, *gcn5* and *htz1* mutant yeast cells with or without UV damage. The $\log_2$ ratio of the number of observed and expected point mutations indicates the effect size of the enrichment or depletion in each region. Asterisks indicate significant enrichments or depletions ($P < 0.05$, one-sided binomial test).

Lastly, when the mutational density data of wild-type *htz1* and *gcn5* are plotted around Abf1, Reb1 and GG-NER complex binding sites, significant heterogeneity in their mutational distributions observed around these sites (Figure 6.4). The TFBSs for Abf1 (n = 1644) and Reb1 (n = 1511) were derived from the MEME-motif obtained from the Yeastract database (Teixeira et al. 2017). In contrast to some of the previous genomic features analysed, these TFBSs reveal stark differences in the distribution of mutations around them. Firstly, mutations tend to accumulate at Abf1 TFBSs in the absence of Gcn5 compared to wild-type cells (Figure 6.4, Abf1 panel). Mutations are enriched at these sites as shown by the higher amount of observed mutation over expected, both in the presence and absence of UV irradiation (Figure 6.4, Abf1 panel). In wild-type cells on the other hand, the mutations induced at these positions are close to the number expected based on mutation frequency and feature size. Interestingly, this accumulation of mutations at Abf1 TFBSs is only observed in the absence of UV irradiation in *htz1* mutant cells. After DNA damage induction, the mutations at Abf1 TFBSs is close to the expected number of mutations calculated in these cells.

In the case of Reb1 the picture is different. Reb1 is a GRF similar to Abf1, but with no known function in repair of UV-induced lesions. Interestingly, in wild-type cells the mutations that are detected within 200bp of these TFBSs are depleted in the absence of DNA damage but enriched over the number that is expected when considering UV-induced mutations (Figure 6.4, Reb1 panel). This differential is lost in both mutants, indicating that mutation induction around Reb1 binding sites is not affected in the absence of Gcn5 or H2A.Z (Figure 6.4).

Given that the organisation of repair is more accurately represented by GCBSs as opposed to TFBSs, I also tested whether mutations are enriched around these genomic features. In the absence of UV irradiation, the mutation load is as expected, with only minor, non-significant differences between observed and expected (Figure 6.4, GCBS panel). However, UV-induced mutations are enriched at these GCBSs in wild-type and *gcn5* mutant cells. Interestingly, the mutations observed at GCBSs in the *gcn5* mutant are much higher than those in wild-type cells (Figure 6.4, GCBS panel). Moreover, this difference is highly significant. In the absence of H2A.Z on the other hand, the difference is not observed, and mutations are detected at a level that is to be expected based on total mutation frequency. Taken together, these findings indicate that the altered repair rates observed in *GCN5* deleted cells might have an impact on mutation induction at GCBSs. This appears to not be the case for H2A.Z depletion.

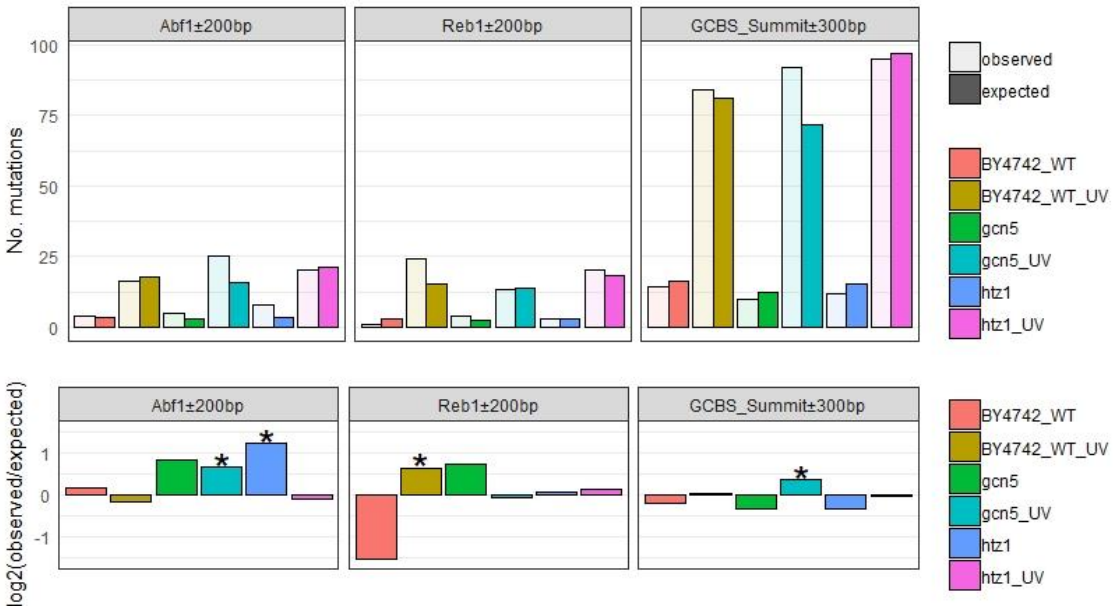


Figure 6.4: Enrichment and depletion of somatic point mutations in chromatin associated genomic features such as Abf1_BS±200bp, Reb1_BS±200bp and GCBS_Summits±300 bp for both wilt-type, *gcn5* and *htz1* mutant yeast cells with or without UV damage. The

log_2 ratio of the number of observed and expected point mutations indicates the effect size of the enrichment or depletion in each region. Asterisks indicate significant enrichments or depletions ($P < 0.05$, one-sided binomial test).

It is known that Gcn5 and H2A.Z are both important for repair of UV-induced DNA damage in a chromatin context (Yu et al. 2016; van Eijk et al. 2018). Similarly, genomic features such as the newly described GCBSs as well as Abf1 binding sites and NFRs are important feature of repair organisation. Therefore, I tried to determine here whether the mutation induction at these genomic position in the relevant mutation backgrounds would reveal changes in the mutation landscape induced by UV irradiation.

Collectively, significantly fewer observed versus expected mutations can be found at OFRs in *gcn5* mutant cells. Significant enrichment and depletion in the observed mutational load in comparison to genome-wide expected mutation density around SPN and Micro-C boundaries are prominent in case of wild-type and *gcn5* mutant. Differences in their genomic mutational distribution were also detected between wild-type, *gcn5* and *htz1* mutant yeast cells, when the observed and expected mutational density were plotted around Abf1, Reb1 or GCBSs binding sites. These results signify, that, although the overall mutational load does not vary between wild-type, *gcn5* and *htz1* mutant yeast cells, even after UV damage, the locations of mutational load varies between them.

6.3.3 The substitution mutational types in wild-type, *gcn5* and *htz1* mutant cells follows the similar pattern.

As mentioned in section 6.3.1, the total mutational load does not vary between wild-type, *gcn5* and *htz1* mutant yeast cells even after UV exposure. When the mutation types from undamaged and UV damaged experiments of wild-type, *gcn5* and *htz1* mutant yeast cells are plotted according to their relative contribution, few differences are observed in unirradiated samples (Figure 6.5, wild-type, *gcn5* and *htz1* mutant yeast cells). The predominant relative contribution of C>A substitution observed in wild-type cells is less prominent in *gcn5* and *htz1* mutants (Figure 6.5). Due to the lack of C>A substitution in these 2 mutant backgrounds, the relative contributions of the other types of substitution is higher than in wild-type cells. Strikingly, T>C types of mutations are relatively frequent in H2A.Z depleted cells (Figure 6.5, bottom middle panel). However, following UV exposure, the relative contribution of substitutions looks fairly similar between wild-type, *gcn5* and *htz1* mutant yeast cells (Figure 6.5, wild-type, *gcn5* and *htz1* mutant yeast cells,

UV exposed panel). This indicates that, following UV damage, the *gcn5* and *htz1* mutant yeast cells do not alter the wild-type substitution mutation rate and type.

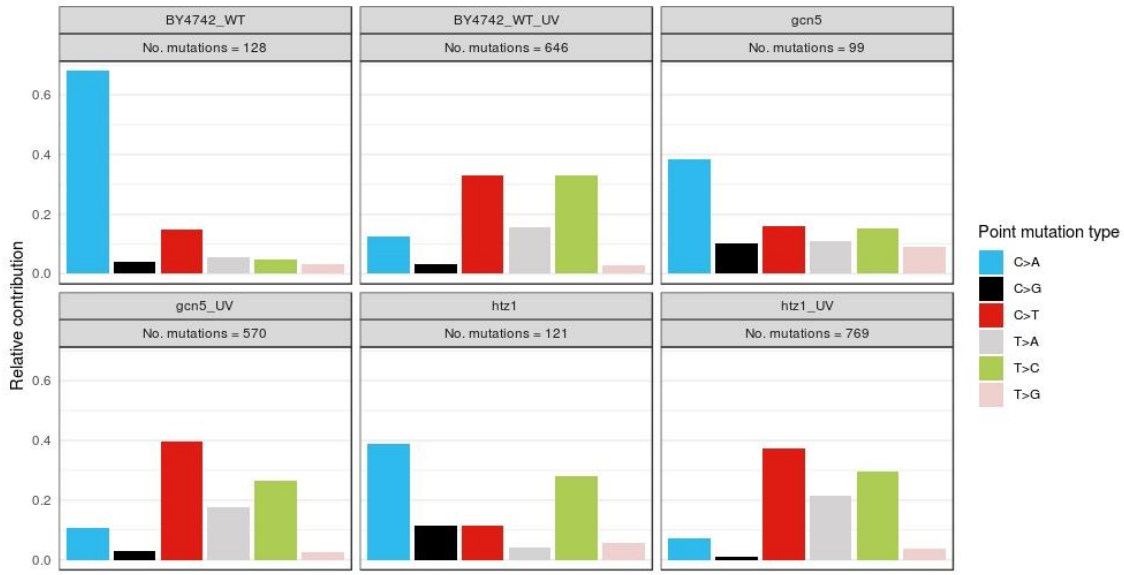


Figure 6.5: Relative contribution of each mutation type in the base substitution catalogues in wild-type, *gcn5* and *htz1* mutant cells with or without UV exposure. Total number of substitution mutations in each sample group is also mentioned.

6.3.4 Substitution mutations in wild-type, *gcn5* and *htz1* mutant cells are differentially enriched around GCBSs and non-GCBS-NFRs.

Nucleotide excision repair (NER) is organised from GCBSs. These sites are predominantly bounded by H2A.Z containing nucleosomes and are sites of UV-induced hyperacetylation by Gcn5 (Yu et al. 2016). To investigate, whether deletion of *GCN5* or *HTZ1* alters the distribution of substitution mutations around the GCBSs compared to non-GCBS NFRs; the relative contribution of substitutions is plotted at these positions in the upper panel of Figure 6.6. The $\log_2$ ratio of mutations at GCBSs versus non-GCBSs contacting NFR substitutions are plotted in Figure 6.6, lower panel.

As described previously, a subset of substitutions is enriched at GCBS in unirradiated cells (Figure 6.6, wild-type panel). Interestingly, the predominantly UV-induced mutations (T>C & T>C) are not enriched at GCBSs compared to non-GCBS NFRs in wild-type cells. This indicates that if damage induction is similar, repair efficiency is comparable between these two genomic features. In, *GCN5* deleted cells the different types of mutations detected are distributed differently compared to the unirradiated wild-type counterpart. In this instance C>A and C>T substitutions are enriched at GCBSs, whereas C>G and T>C mutations are higher at non-GCBS NFRs (Figure 6.6, gcn5 panel).

After UV irradiation the mutation distribution is completely altered. I now observe a significant enrichment of C>A and C>T mutations at GCBSs and non-significant enrichment of all other types of substitutions except T>G (Figure 6.6, gcn5 panel).This is the first preliminary finding that reveals a clear distinction between the distribution of mutations between wild-type and *gcn5* mutant cells, in line with the difference in repair rates we observed previously (Figure 6.1, (Yu et al. 2016)).

Absence of H2A.Z has a less severe effect on the distribution of mutation around the GCBS and non-GCBS NFR positions. Here, C>A mutations are enriched at GCBSs in the absence of DNA damage. Similarly, C>T mutation are enriched at GCBS's albeit not significantly (Figure 6.6, *htz1* panel). The higher rate of C>G mutants at GCBS's observed in wild-type cells in not detected here. Moreover, the distribution of UV-induced mutations in the absence of H2A.Z is only slightly different from that observed in wild-type cells (Figure 6.6, compare wild-type and *htz1* UV panels). In *htz1* mutant cells the C>A and C>T are enriched at GCBS's, with only the latter significantly. C>G types of substitution are similarly enriched at GCBS between wild-type and *htz1* mutant cells. Even though H2A.Z is an important component of chromatin and plays a role in organised DNA repair, the impact it has on repair does not translate directly to major alteration to the distribution of UV-induced mutations in this context.

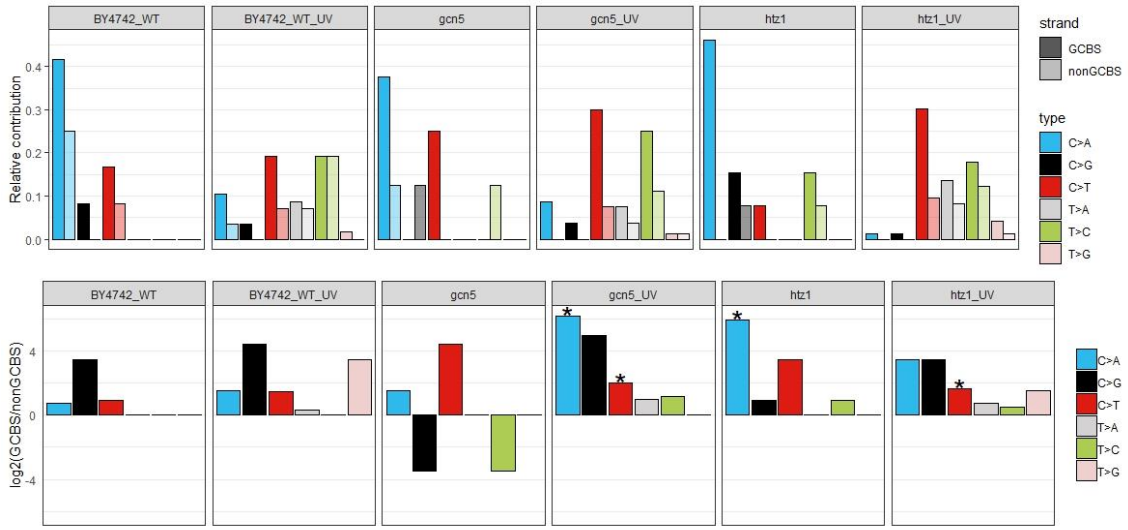


Figure 6.6: Relative contribution of 6 substitution types at GCBSs (dark shaded) and non-GCBSs NFRs (light shaded) sites in wild-type, *gcn5* and *htz1* mutant yeast cells with or with UV exposure. The $\log_2$ ration indicates the effect size of substitution mutation bias and asterisks (*) indicate significant bias ($P < 0.05$, one-sided binomial test).

6.3.5 Significant mutational bias observed in wild-type, *gcn5* and *htz1* mutant cells toward open chromatin.

It is well established that, chromatin remodeling is required for efficient repair of UV-induced DNA damage (Yu et al. 2011). Previous studies in our laboratory, demonstrated that the GG-NER complex regulates UV-induced histone H3 acetylation, by controlling chromatin occupancy of the histone acetyl transferase Gcn5 on chromatin (Teng et al. 2008). This UV-induced hyperacetylation of histones promotes an open chromatin conformation required for efficient repair of DNA damage (Yu et al. 2011; Yu et al. 2016). Additionally, it was also revealed that, the histone H2A variant, H2A.Z (*HTZ1*), in nucleosomes has a positive function in promoting efficient NER in yeast. H2A.Z inherently enhances the occupancy of the histone acetyltransferase Gcn5 on chromatin to promote histone H3 acetylation after UV irradiation (Yu et al. 2013). To investigate, whether loss of either *GCN5* or *HTZ1* alters the distribution of UV-induced mutations around the high and low acetylated regions within the yeast genome, we annotated the genome using previous histone H3 acetylation data. Binning the genome in this way allows the designation of genomic windows with enriched probes as High acetylated regions and those without as Low acetylated regions. This annotation delineates open versus closed chromatin. Indeed, in response to UV damage, a significant mutational bias is observed towards low acetylated, open chromatin regions within yeast genome for all mutation types (Figure 6.7).

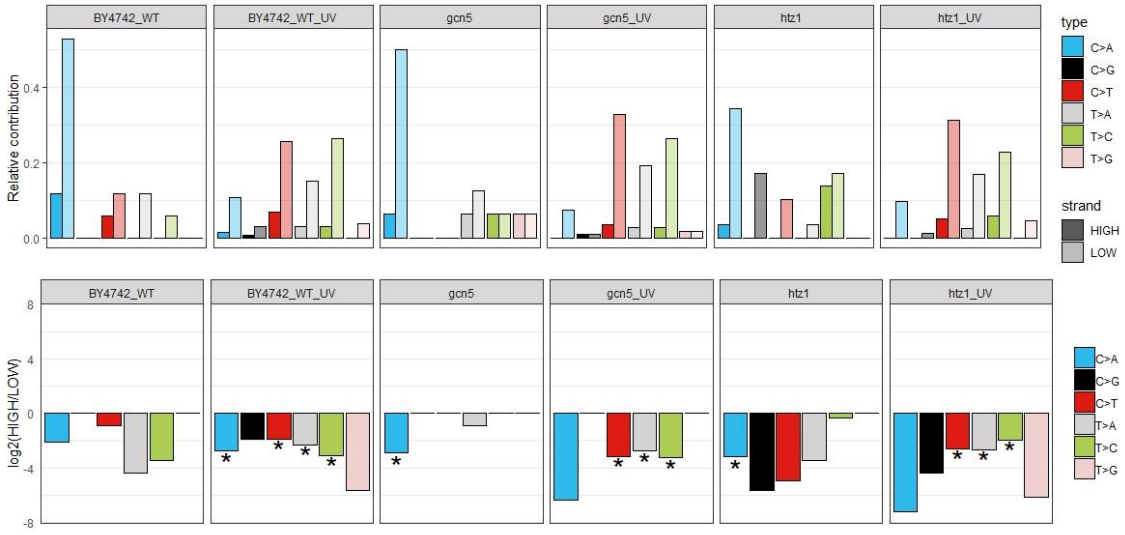


Figure 6.7: Significant mutational distribution bias towards high and low acetylated regions in wild-type, *gcn5* and *htz1* mutant yeast. Log_2 ratio of the number of mutations on the high and low acetylated regions per indicated base substitution of each samples

shown in lower section of this figure. The $\log_2$ ratio indicates the effect of the bias and asterisks (*) indicate significant acetylation region asymmetries ($P<0.05$, one-sided binomial test).

6.3.6 Distribution of mutations in wild-type, *gcn5* and *htz1* mutant cells depends on replication timing

Chromatin remodeling is required for efficient replication (Vincent et al. 2008). Importantly, replication-timing and chromatin structure have both been associated with mutagenesis (Stamatoyannopoulos et al. 2009; Schuster-Böckler and Lehner 2012). To investigate whether variation in distribution of substitution mutations can be observed in early and late replicating regions within wild-type, *gcn5* and *htz1* mutant yeast cells, the relative contrition of substitution mutation is plotted according to early (dark shaded) versus late (light shaded) replicating regions within yeast genome (Figure 6.8). To do this we obtained a list of 5kb windows from Repli-seq data, that annotate the entire yeast genome as either early or late replicating. Using this information, it becomes apparent that the relative distribution of substitutions varies between wild-type, *gcn5* and *htz1* mutant yeast within early versus late replicating regions (Figure 6.8). Mutations occur predominantly in late replication regions of the genome. In unirradiated wild-type cells T>G mutations are an exception and appear more frequently in early replication DNA (Figure 6.8, wild-type panel). As expected, UV-induced mutations in wild-type cells are enriched in late replication regions of the genome. However, the C>G type of mutation is found mostly in early replicating sites. Both *gcn5* and *htz1* mutants show no significant mutations in early or late regions of the genome in the absence of DNA damage (Figure 6.8). The UV-induced mutations in these mutants are fairly similar to the wild-type pattern of early versus late replicating DNA. The $\log_2$ ratio of early versus late mutations is altered but the relative contributions do not differ greatly (Figure 6.8, top panel). Interestingly, the C>G mutations are enriched in late replicating regions of the genome in *GCN5* deleted cells as compared to wild-type cells. The bias towards mutations accumulating in late replicating DNA in wild-type is maintain in both *gcn5* and *htz1* mutants, however the absolute difference does not pass the significance test (Figure 6.8). These data demonstrate that loss of either *GCN5* or *HTZ1* alters the distribution of substitutions in relation to replication timing in response to UV damage.

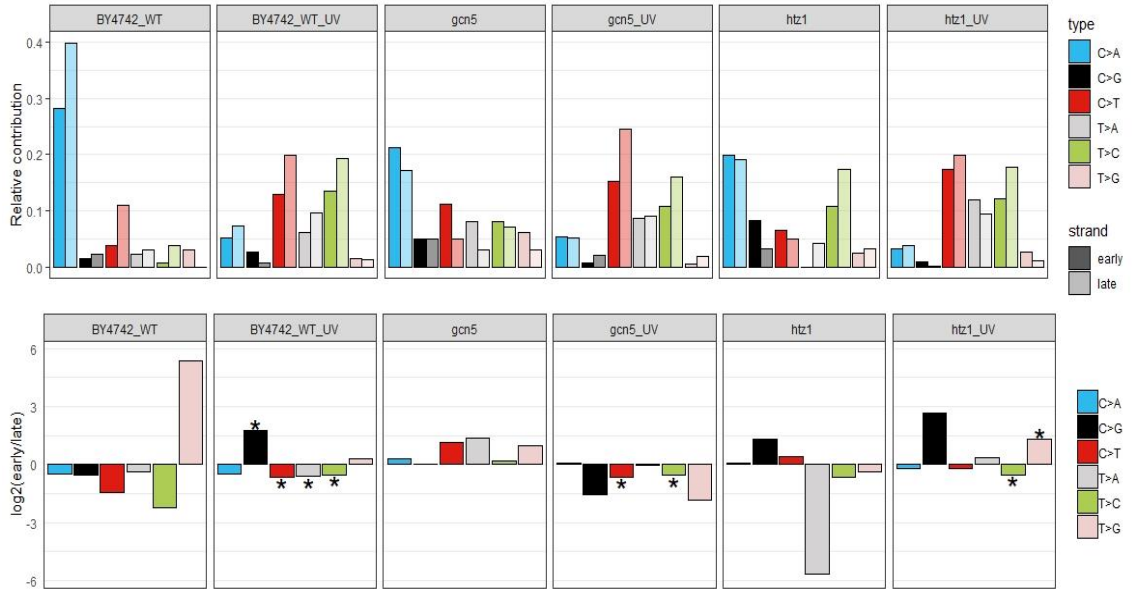


Figure 6.8: Mutational distribution with replication timing information. The upper panel shows the relative contribution of each nucleotide changes is subdivided into either early (dark shades) and late (light shades). Log_2 ratio of the number of mutations on the early and late replicative regions per indicated base substitution of each samples shown in lower section of this figure. The log_2 ratio indicates the effect of the bias and asterisks (*) indicate significant replicative timing asymmetries ($P<0.05$, one-sided binomial test).

6.3.7 Transcriptional strand asymmetry observed in the distribution of substitution mutations wild-type, *gcn5* and *htz1* mutant yeast

To investigate, whether the distribution of substitution mutations varies with transcriptional activity in wild-type, *gcn5* and *htz1* mutant yeast, the relative contribution of substitutions mutations on transcribed (dark shaded) and untranscribed (light shaded) strand is plotted in Figure 6.9. The strand-bias observed for the substitutions detected in wild-type cells, are not observed in the two mutant backgrounds. However, as with wild-type cells, a significant bias for UV-induced C>T and T>C types of mutations towards the untranscribed strand is observed for *gcn5* and *htz1* mutant yeast cells (Figure 6.9). Even though Gcn5 and H2A.Z play a role in transcription and repair organisation, their absence does not appear to have an impact on the biased accumulation of mutations in the untranscribed strand. It is possible that the intact TC-NER pathway is unaffected by the deletion of *GCN5* and *HTZ1* and therefore no strand-bias is observed. Conversely, the effect on repair in these mutants is predominantly through defective GG-NER.

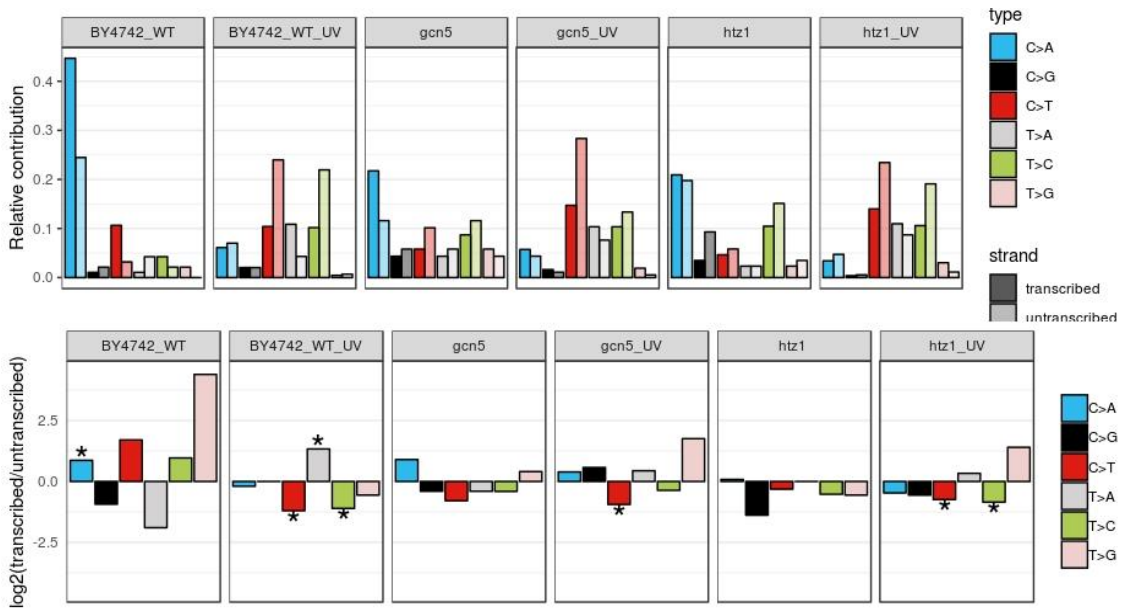


Figure 6.9: Mutational profile with transcriptional strand bias in wild-type, *gcn5* and *htz1* mutant yeast cells with or without UV exposure. The relative contribution of each trinucleotide change subdivided into the fraction of trinucleotide changes present on the transcribed (dark shades) and untranscribed strand (light shades). Log_2 ratio of the number of mutations on the transcribed and untranscribed strand per indicated base substitution for each sample type (lower panel). Asterisks indicate significant transcriptional strand asymmetries ($P < 0.05$, two-sided Poisson test).

6.3.8 Mutation spectrum and similarity with PCAWG signatures

To study the spectrum of mutations observed in linear orientation of the chromosomes, rainfall plots of the mutation data were generated for wild-type, *gcn5* and *htz1* mutant yeast to compare the mutations distribution on a genome-wide level. No significant changes were observed between them (Appendix-V, A5.5-A5.8). The UV-induced mutational pattern is detected in wild-type, *gcn5* and *htz1* mutant yeast, as demonstrated by the increased level of C>T and T>C types of substitutions. The rainfall plots of wild-type cells with or without UV exposure were plotted in chapter IV (Figure 4.16 and 4.17, respectively).

The 96 trinucleotides mutation profiles of wild-type, *gcn5* and *htz1* mutant yeast with or without exposure to UV damage are plotted in Figure 6.10. The wild-type, *gcn5* and *htz1* mutant yeast cells show similar patterns of trinucleotide context with subtle differences between them. In undamaged wild-type, *gcn5* and *htz1* mutant yeast cells this pattern is dominated by C>A types of substitution (Figure 6.10, without UV treatment panel). Importantly, the trinucleotide context of the C>A type of mutations is different between

these strains. Following UV damage, again *gcn5* and *htz1* mutant yeast cells display the characteristic UV-induced predominance of C>T at TCN, T>C at TCN (the mutated base underlined) substitution, which is similar to the wild-type pattern in this 96 trinucleotides mutation type context (Figure 6.10, UV exposed panel).

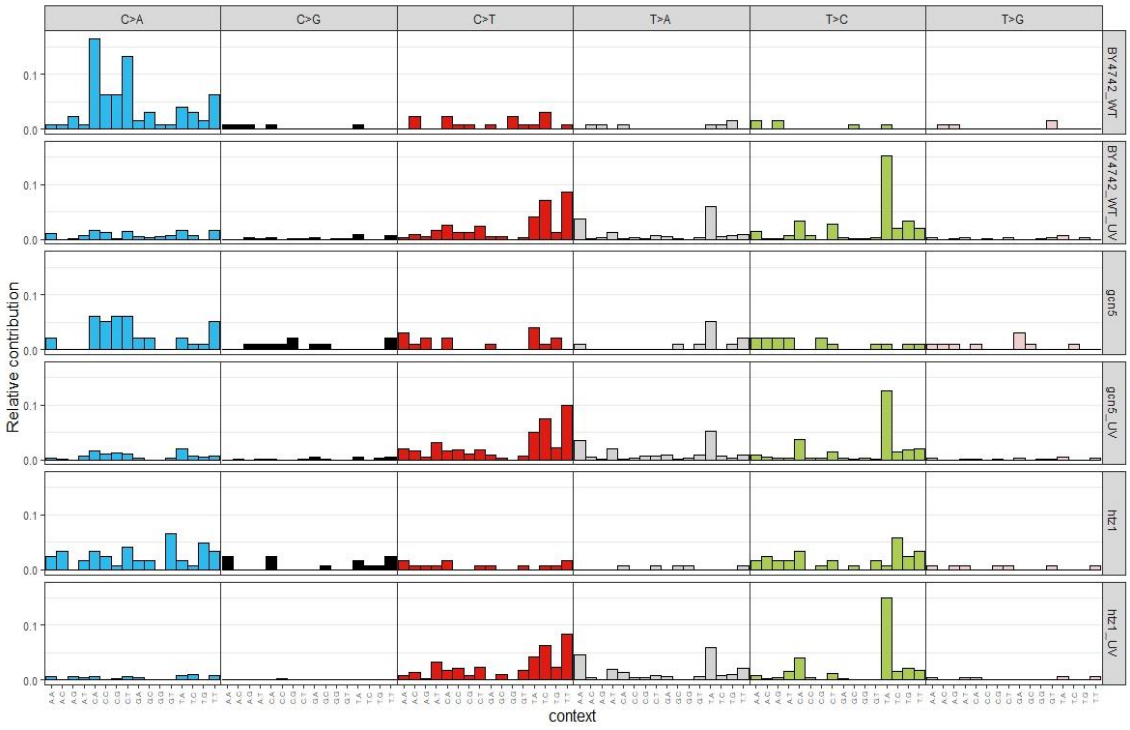


Figure 6.10: Retailed comparison of mutational profile showing 96 trinucleotide substitution types and their relative contribution to wild-type, *htz1* and *gcn5* mutant cells with or without UV exposure.

A similar result is obtained when the cosine similarity between the mutational profile of the individual samples and the PCAWG signatures was calculated (Figure 6.11). Undamaged, *gcn5* and *htz1* mutant yeast cells present a mutational profile that is highly similar to the same PCAWG signatures that I describe for wild-type cells in chapter IV (section 4.3.13, Figure 4.21). Likewise, following UV damage, the mutational profile of the *gcn5* and *htz1* mutant yeast cells reveal a similar profile to the wild-type pattern. Similar PCAWG signatures have high cosine similarity scores compared to those observed to match the wild-type pattern of mutations induced by UV damaged. Clustering, based on Euclidean distance between the vectors of cosine similarities with the signatures also groups undamaged wild-type, *gcn5* and *htz1* mutant yeast together, while UV damaged wild-type, *gcn5* and *htz1* mutant yeast cells also cluster together (Figure 6.11). This result demonstrates that the mutational patterns observed between

wild-type, *gcn5* and *htz1* mutant yeast cells map onto the PCAWG signature with a high degree of similarity.

Figure 6.11: Heatmap shows the cosine similarity between the mutational profile of wild-type, *htz1* and *gcn5* mutant cells with or without UV exposure with COSMIC signatures profile. The samples are hierarchically clustered (average linkage) using the Euclidean distance between the vectors of cosine similarities with the signatures.

Interestingly, reconstruction of the undamaged wild-type, *gcn5* and *htz1* mutant yeast cells' mutational profile using PCAWG signatures showed a distinctive pattern (Figure 6.12). The mutation profile reconstruction of undamaged *htz1* mutant cells is mostly made up of PCAWG SBS37, SBS29 and SBS14 signatures (Figure 6.12). The proposed aetiology for SBS37 is unknown, whereas SBS29 is associated with tobacco chewing and SBS14 is associated with concurrent POLE mutations and mismatch repair deficiency (Alexandrov et al. 2018). However, reconstruction of the mutational profile of *gcn5* undamaged cells requires those PCAWG signatures as noted for wild-type undamaged cells (Chapter IV, section 4.3.13). Clustering the mutational profiles of these mutant also reveals that the wild-type and *gcn5* undamaged profiles cluster together, whereas *htz1* clusters separately. Following UV damage, the *htz1* and *gcn5* mutational profiles can be reconstituted with PCAWG signatures similar to those used for the reconstruction of the wild-type UV-induced mutational profile (Chapter IV, section 4.3.13). Again, this phenomenon is visible from their clustering analyses as well (Figure 6.12). Collectively, differences between the mutational profile of undamaged wild-type, *htz1* and *gcn5* cells were observed and the reconstruction of these profiles required different PCAWG signatures. Therefore, these observations suggest that loss of *HTZ1* or *GCN5*, alters the

normal distribution of mutations within yeast cells and that the mutagenic process active in these mutants is different as they require different PCAWG signatures in their reconstruction. However, after UV damage, the mutation pattern is dominated by UV-induced mutations such as T>C and T>C resulting in a very similar distribution of mutations that can be reconstructed with similar PCAWG signatures.

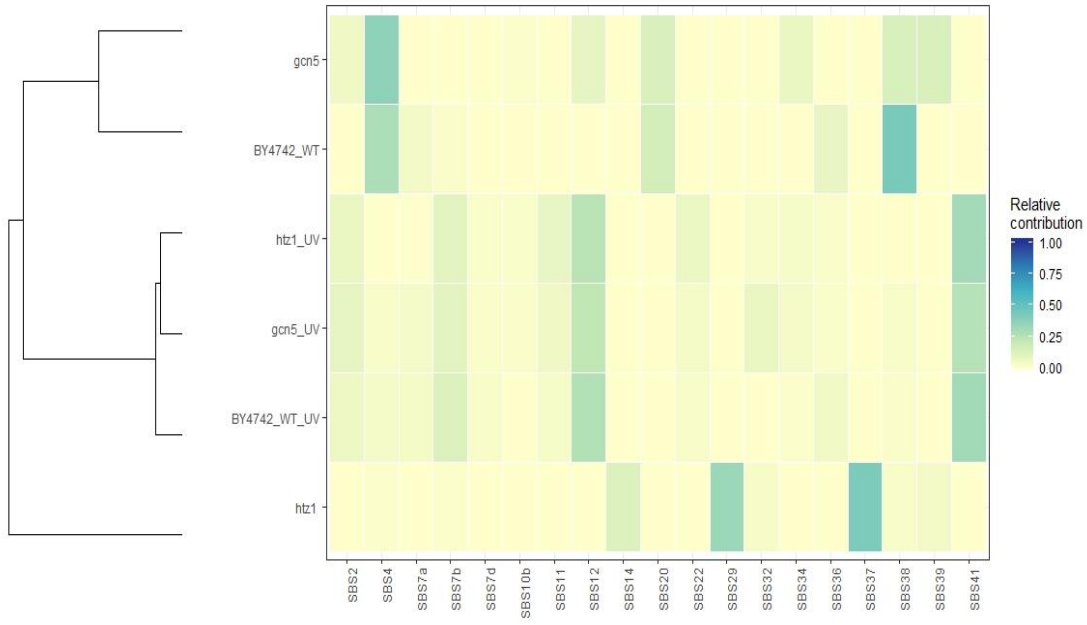


Figure 6.12: The optimal relative contribution of COSMIC signatures to reconstruct the mutational profiles of the wild-type, *gcn5* and *htz1* mutant yeast cells with or without UV damage. The samples are hierarchically clustered (average linkage) using the Euclidean distance between the vectors of cosine similarities with the signatures.

6.3.9 *De novo* mutational signature extraction using NMF

As explained in Chapter IV, the NMF rank estimation graph and consensus heatmap can be used to estimate the best rank to perform NMF with. Performing this analysis (Figure 6.13 & 6.14) on the datasets discussed in this chapter shows that two mutational signatures can be extracted from the data.

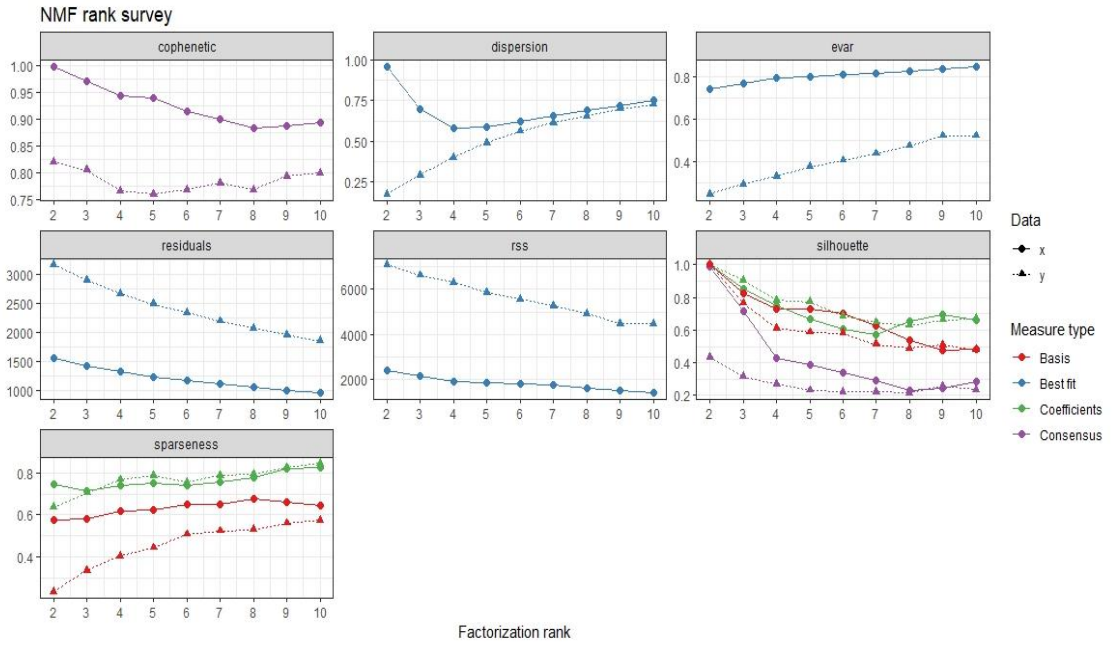


Figure 6.13: Estimation of the rank from wild-type, *htz1* and *gcn5* mutant cells mutational catalogue. Quality measures computed from 50 runs for each value of N. The estimation is based on Brunet's algorithm. The data mark as 'x' represents the experimental data set and the data mark as 'y' represents after randomization of the same data.

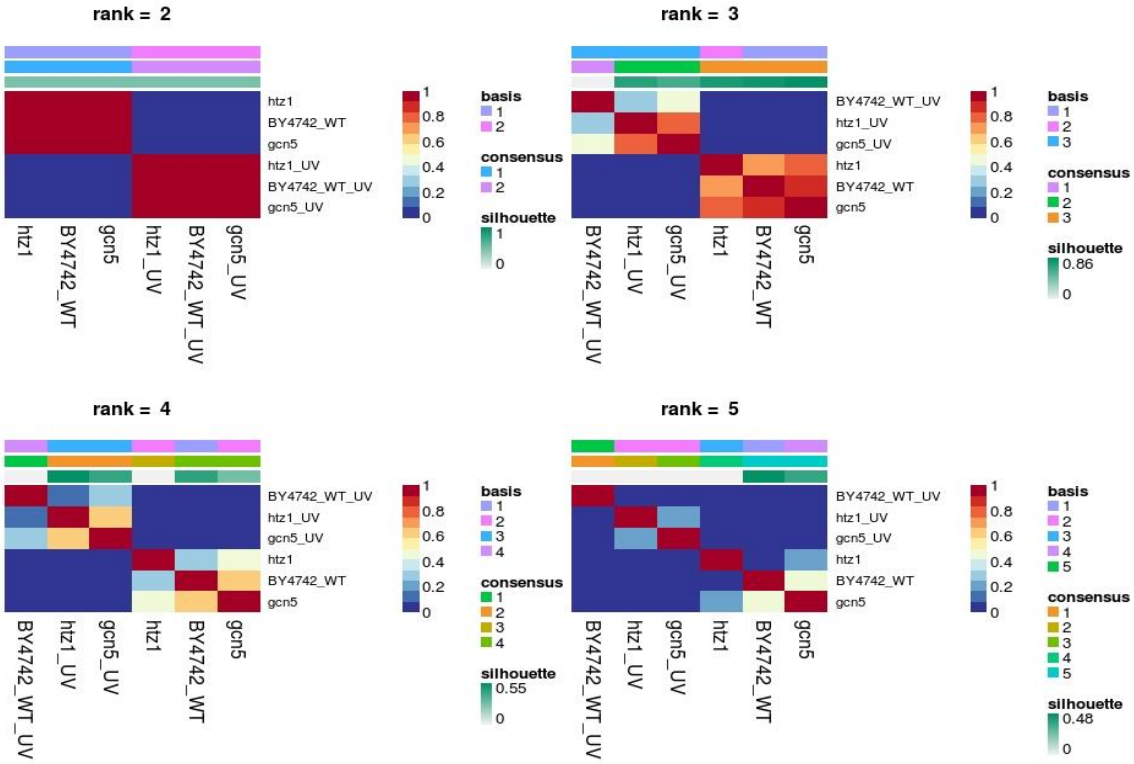


Figure 6.14: Estimation of the rank after grouping the individual samples together as wild-type, *htz1* and *gcn5* mutant cells mutational catalogue. Consensus matrices

computed from 50 runs for each value of rank 2 to 5. This is because after pooling the individual samples based on treatment (6 treatment, max clustering for the rank 2 to 5).

The two NMF decomposed signatures derived from wild-type, *htz1* and |*gcn5* mutant profiles are plotted in Figure 6.15. The relative contribution of these *de novo* extracted signatures A and B to the wild-type, *htz1* and *gcn5* mutational profile is plotted in Figure 6.15.

Signature A is characterised predominantly by C>A transversions at CCN and TCN trinucleotide context (the mutated base is underlined) with minimal contribution of the other types of substitutions. This signature is exclusively responsible for the mutational pattern noted previously in unirradiated wild-type cells (Chapter IV, section 4.3.13) and undamaged *htz1* and *gcn5* mutant cells. This indicates that these mutations are predominantly cause by the normal cellular metabolic activities. Similar processes are indeed active in *htz1* and *gcn5* mutant cells and this confirms that NMF can extract the signature attached to these endogenous processes.

Secondly, signature B is characterised predominantly by C>T at TCN trinucleotides and T>C at TTN trinucleotide with a small amount of T>A and C>A types of substitution. Mutations that are part of this signature contribute to all UV-treated cells and represent the biological process of UV-induced mutagenesis (Alexandrov et al. 2018).

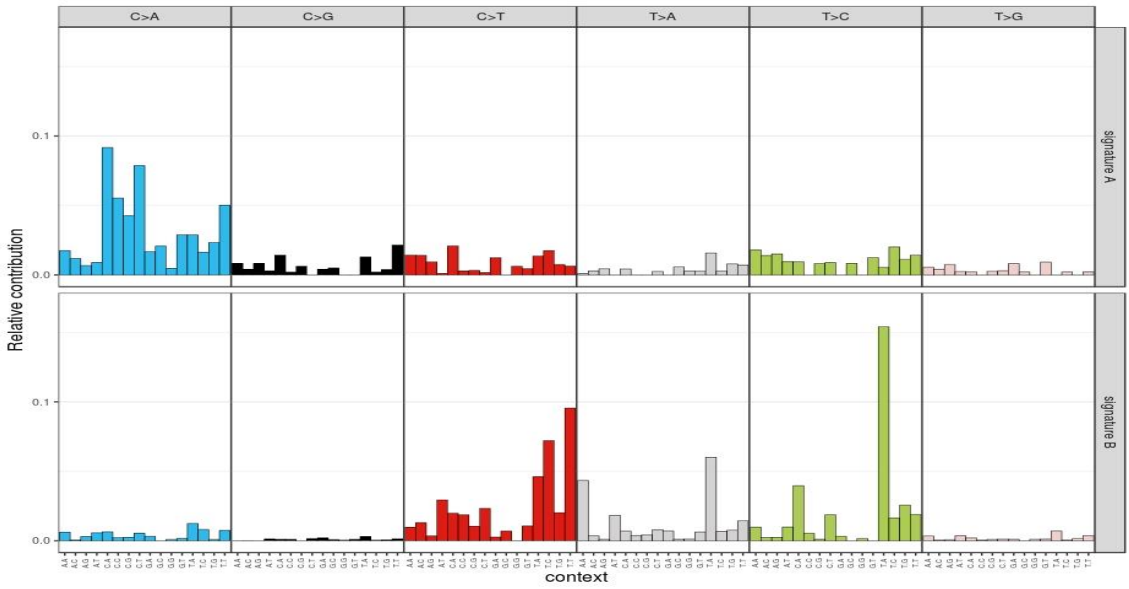


Figure 6.15: Relative contribution of indicated 96-trinucleotides changes to the two mutational signatures that were extracted *de novo* by NMF analysis of the acquired somatic mutational catalogue of the experimental model system. The wild-type, *htz1* and

gcn5 mutant with or without exposure to UV damages cells' mutational catalogue were considered for this analysis.

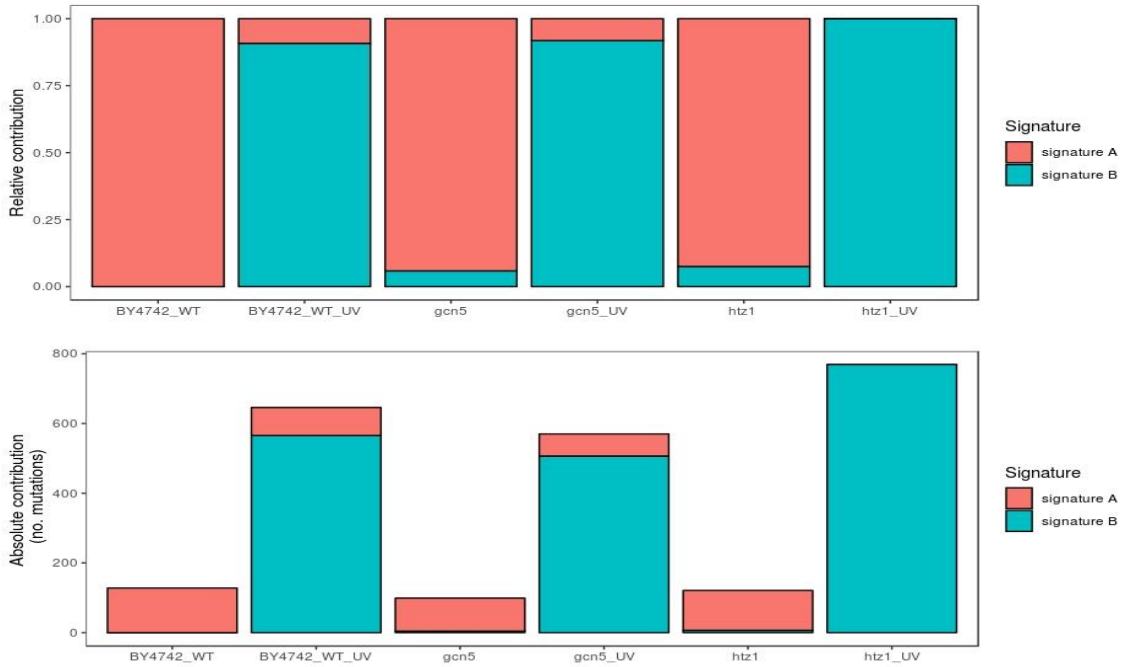


Figure 6.16: Relative and absolute contribution of each mutational signatures in wild-type, *htz1* and *gcn5* mutant cells mutational profile with or without UV exposure.

Interestingly, hierarchical clustering that makes use of the relative contribution of each of the two extracted signatures to each experimental sample, reveals two main clusters (Figure 6.17). This result indicates that, the two biological processes that were originally involved in this experiment can be extracted using NMF and clustering.

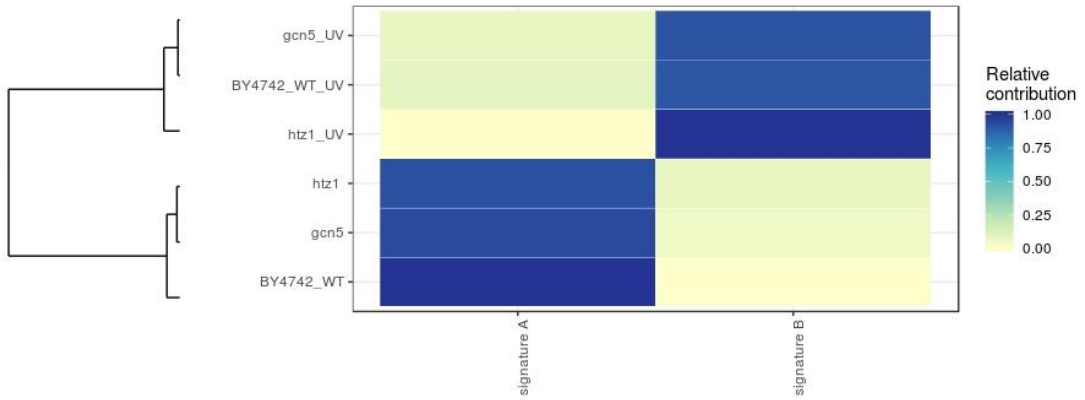


Figure 6.17: Heatmap showing relative contribution of each mutational signatures in each sample. The samples are hierarchically clustered (average linkage) using the Euclidean distance between the vectors of cosine similarities with the signatures.

The cosine similarity between these two mutational signatures (signature A and signature B) and PCAWG signatures is shown in Figure 6.18. A similar pattern is observed as during the analysis of the wild-type mutational spectrum with or without exposure to UV damage (Chapter IV, section 4.3.16, Figure 4.30, signatures A and C in that case). A high cosine similarity is also observed here between signature A described in this chapter and signature A described in chapter IV. A similar observation can be made for Signature B in this chapter and signature C in chapter IV.

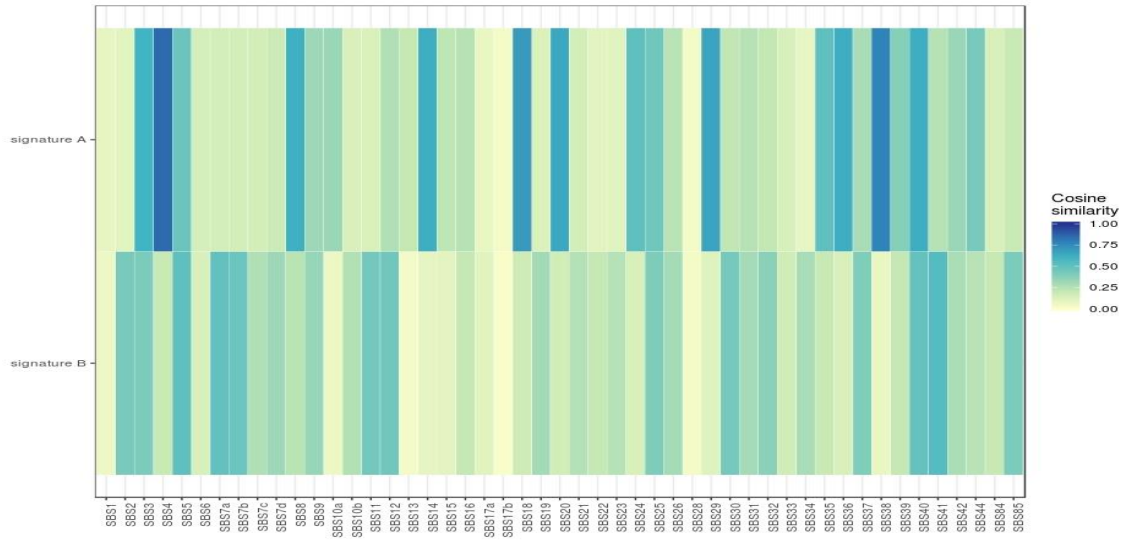


Figure 6.18: Heatmap of cosine similarities between the *de novo* extracted signatures from samples and COSMIC signatures.

In summary, the two signatures extracted using NMF from the mutational profile of wild-type, *htz1* and *gcn5* mutant cells and their contribution to the corresponding samples was able to extract the signatures that relate to the active biological processes. This includes both normal cellular metabolic process during respiration and the exposure to UV light. This result indicates that loss of *GCN5* or *HTZ1* does not generate unique mutational signatures. However, they alter the UV induced distribution of mutations at various genomic locations as observed from plotting the mutational load around various genomic features as well higher order chromatin structures.

6.4 Summary

In this chapter, I employed the same workflow developed in chapter-IV for acquiring the genome-wide distribution of mutations, with or without UV damage and subsequent analysis of the distribution of mutational types and patterns for comparative analysis with relative repair rates using *S. cerevisiae* as an isogenic model organism.

The key findings of this chapter are that the loss of chromatin modifier, or histone variant exchange capability, does not alter the mutational levels compared with that seen in wild-type cells, either with or without UV damage. However, deficiency of these processes alter both type and distribution of mutations within genome, importantly from those regions where it functions as part of the mechanism of GG-NER. Importantly, the distribution of the mutational load around various genomic features provides important evidence of histone modification or variant exchange regarding to DNA repair and mutagenesis.

In chapter III and in our recent publication we reported that the GG-NER process is organised from GCBSs, predominantly at the Micro-C boundaries, many of which are flanked by H2A.Z-containing nucleosomes (van Eijk et al. 2018). These results demonstrated that, in undamaged cells, the GG-NER complex occupies GCBSs within yeast genome, bounded by H2A.Z-containing nucleosomes. Following UV damage, these H2A.Z-containing boundary nucleosomes are remodeled in a GG-NER complex-dependent fashion for efficient repair. Another study showed that the majority of H2A.Z-containing nucleosomes are located at the +1 nucleosome of most TSS (Albert et al. 2007). The results in this chapter demonstrate that following UV damage, loss of H2A.Z results in an even distribution of mutations within the yeast genome, in comparison to wild type cells. This result also indicates that, the presence of this histone variant might contribute to the mutational hotspot around these genomic sites, such as higher-order Micro-C boundaries. One recent study showed that, histone variant (H3.3) is required for IgV gene diversification by generating regions of somatic-hypermutation generated by AID (Romanello et al. 2016). Another study suggested that, H3.3 is a ‘target marker’ for somatic hypermutations (Aida et al. 2013). Our result thus provide a possible mechanism for how the loss of a histone variant does not alter the total mutational load, but does alter the location of acquired mutations within the genome.

Following UV damage, results demonstrate that loss of GCN5 also results in higher mutational load around those sites from which GG-NER is organised. These observations

provide evidence of a requirement for Gcn5 in the efficient repair of UV-induced damages (Yu et al. 2011). The significant enrichment of mutational load following UV-exposure in *gcn5* mutant cells, notably at Micro-C boundaries, Abf1 binding sites and GCBSs, provide evidence that Gcn5 mediated acetylation is required around these sites for efficient repair.

Moreover, the relative contribution of base substitution mutations in early and late replicating regions also varies depending on both exposure to UV damage and capacity to acetylate histones, histone variant exchange, or a combination of these. Histone modifications have been implicated in many DNA-depended processes including DNA replication, transcription and repair. Dynamic changes in histone acetylation regulates the origin of replication in yeast cells (Unnikrishnan et al. 2010). Furthermore, our studies also showed that acetylation is required for efficient repair of UV induced damages by NER (Yu et al. 2005; Reed 2011). Like histone modification, histone variant exchange also alters the dynamic structure of chromatin. Loss of histone variant affects the replication and also the repair processes (Henikoff and Smith 2015). My study reveals that deletion of *GCN5* or *HTZ1* also alters the wild-type substitution mutation pattern within early vs late replicating regions in yeast gnome.

No novel mutation signatures were found in *gcn5* and *htz1* mutant cells. This suggests that the biological processes of histone modification by Gcn5 or histone variant exchange does not alter the acquired mutational pattern in defective cells, even after UV exposure. Although, loss of these biological processes does alter the location of these mutation within yeast genome, especially around higher-order chromatin structure at Micro-C boundaries.

Future study targeting these Micro-C boundaries will provide mechanistic insight into how factors that affect the process of GG-NER, such as loss of HAT activity or even histone variant exchange deficiency, drive tumorigenesis by, altering the mutational distribution around these higher-order chromatin structures. It may be that enrichment of mutations at these sites activate proto-oncogenes, (Flavahan et al. 2016; Hnisz et al. 2016) or inactivate tumour suppressor genes that drive tumorigenesis (Jia et al. 2017).

Chapter VII

General Discussion

In cells maintaining genomic stability is essential for life. Genomic instability, as defined by the higher than normal rates of mutations within the genome, is an established hallmark in human diseases such as cancer. Since DNA is constantly exposed to the deleterious effects of both the internal and external cellular environment, mechanisms have evolved to sense and repair the consequent genetic damages within the cellular chromatin environment. Defects in such processes result in mutations in the genetic code that are associated with a variety of diseases including cancer. Therefore, understanding the mechanisms that cause the mutational spectra in genomes will have a significant impact on individual risk and indeed cancer prevention, as well as having implications for the development of stratified medicine in the treatment of cancer.

The main achievement of this research described in this thesis has been to advance our understanding of how the organisation of biological processes that repair DNA damage contribute to the generation of mutational spectra induced in genome. The thesis begins by defining the genomic locations of GG-NER complex binding and demonstrating that remodelled H2A.Z-containing nucleosomes flanking GCBSs is necessary for efficient repair of UV-induced damage. Next, this study also demonstrates how defects in the DNA repair process alter the genomic distribution of mutation types, load and locations within yeast genome, following exposure to the same mutagen. Additionally, I explained, how factors, that are known to be involved in chromatin modifications to drive the DNA repair process, alter the type and genomic locations of mutations in response to UV damage.

NER is the sole mechanism for repairing helix distorting bulky adducts, including those formed endogenously by oxidative or hydrolytic processes (Melis et al. 2013) and exogenously by UV light (de Laat et al. 1999) or chemical agents such as polycyclic aromatic hydrocarbons (PAH) in tobacco smoke (Hecht 1999) and platinum drugs (Reed 1998). Among them, well-known and well-studied, are UV-induced DNA photoproducts, such as cyclobutene pyrimidine dimers (CPD) and 6–4 photoproducts (6-4 PP), which get efficiently repaired by NER. Two sub-pathways of NER exist characterised by the initial damage recognition steps: the rapid acting transcription coupled repair pathway (TC-NER) that operates on the transcribed strands of actively transcribing genes and involves RNA polymerase II in the damage recognition step; and the slower acting global genome repair pathway (GG-NER) that operates on all DNA, including non-transcribed and repressed regions in the genome. This pathway involves a unique subset of proteins in the early stages of DNA damage recognition known as GG-NER complex in yeast. Following the initial stages of DNA damage detection, these two pathways converge and utilise the

same set of DNA repair proteins. Most of the yeast NER genes have well conserved structural and/or functional human homologues, and the main features of both GG-NER and TC-NER pathways are evolutionarily conserved (Hoeijmakers 1994). Initially, work on GG-NER in yeast cells showed that for efficient repair, a complex of proteins, also known as the GG-NER complex is involved in the early stages of this process. The GG-NER complex is comprised of Abf1, Rad16 and Rad7 in which Abf1 is bound to specific DNA binding sites (Reed et al. 1998; Reed et al. 1999), which can be found at hundreds of locations throughout the yeast genome (Yu et al. 2009). Importantly, it is established that Rad7 and Rad16 exist in a complex within the cell and that deletion strains of either component have an identical phenotype (Verhage et al. 1994; Reed et al. 1998). Both Rad16 and Rad26 (a TC-NER factor in the yeast *S. cerevisiae)* are member of the SWI/SNF super-family of DNA-dependent ATPases involved in chromatin remodelling (van Gool et al. 1994; Osley et al. 2007). The homologue of Rad26 is the human Cockayne syndrome group B (CSB) gene (Guzder et al. 1996; Lee et al. 2001). Both Rad26 and CSB are involved in the preferential repair of UV lesions on the transcribed strand, and in this process, they function together with the other components of NER (Teng and Waters 2000; Ghosh-Roy et al. 2013). Rad4, is a core NER pathway and is involved in removal of bulky lesions (Min and Pavletich 2007). Rad4 forms a heterodimer with Rad23, which is homologous to the human XPC-hHR23A/B complex, and is involved in damaged DNA recognition in human cells (Masutani et al. 1994; Jansen et al. 1998). It is well established that, in humans, NER deficiencies are associated with both Xeroderma Pigmentosum (XP) and Cockayne syndrome (Foury 1997; Friedberg et al. 2005). To study how repair is initiated in chromatin, it has been recently demonstrated that global genome nucleotide excision repair (GG-NER) in chromatin is organised into domains around open reading frames (Yu et al. 2016). This study showed that, GG-NER complex regulates the histone acetyltransferase, Gcn5 at this sites for promoting efficient repair of UV induced lesions. Likewise, following UV damage, Htz1 also promote NER in yeast by enhancing the occupancy of HAT Gcn5 on chromatin to promote histone H3 acetylation and open chromatin structure necessary for efficient repair (Yu et al. 2013). Loss of GG-NER complex or histone modifier (GCN5 or HTZ1), showed increase UV sensitivity and alter wild-type relative repair rates within genome in response to UV DNA damage (Figure 1.19).

The whole-genome sequencing studies (The Cancer Genome Atlas network, 2012) have measured genome-wide tumour-specific somatic mutation patterns, which report the

entire spectrum of mutations accumulated during tumourigenesis within the cancers of individuals. These studies revealed the association of multiple mutational signatures, which are indicative of the mutational processes responsible for the mutations found in sporadic cancers, derived from the so-called normal population, across a range of human cancer types (Alexandrov et al. 2018). The causes of some of the mutational signatures identified are known and fall broadly into two groups. Firstly, signatures caused by exposure to environmental mutagens, such as solar UV light, associated mainly with skin cancers, and polycyclic aromatic hydrocarbons such as cigarette smoke, associated mainly with lung cancers. Secondly, these signatures can be the result of defective DNA repair in one of the various repair pathways. However, the causes of many of the mutation signatures identified remain to be determined (Kandoth et al. 2013b; Alexandrov et al. 2018). Significantly, these studies have also revealed novel cancer genes, many of which are involved in chromatin remodelling and modification. We speculate that tumourigenesis in these sporadic cancers may be driven by mutations in these chromatin modifiers that disrupt the normal landscape of genome-wide DNA repair rates. This subsequently alters the distribution of mutations in the genome. These observations demonstrate the importance of understanding how genetic damage is formed and repaired in chromatin, throughout the entire genome. Knowing the underlying causes that give rise to cancer will permit a more accurate assessment of the risk of developing the disease, and aid in selecting and developing appropriate treatments for individuals. DNA repair pathways are integrated within a system-wide process known as the DNA Damage Response (DDR). In this system, DNA damage sensors detect chromatin-associated DNA damage signals, which ultimately determine the physiological response of the cell to DNA damage (Lazzaro et al. 2009). Therefore, determining how DNA damage in chromatin is detected, efficiently repaired, the chromatin restored and how these events are organised in the genome, is fundamental to understanding the mechanisms that underpin the relationship between genome stability and human health. To study this our laboratory initially developed genomic tools for measuring genome-wide DNA damage and repair using microarrays (Teng et al. 2011; Powell et al. 2015). Originally developed for DNA repair studies in yeast, the technique has now been adapted and is functional for studies in human cells. So far, these study generated genome wide DNA damage and repair profiles for cells treated with the well-known mutagens UV light and cisplatin, a drug used extensively in chemotherapy of patients with solid tumours.

Nucleosome remodeling is essential to understand how repair is organised within genome. To address this question, MNase-Seq experiments were performed to determine how nucleosomes remodelled around those sites from where repair is organised. Additionally, Abf1 and Htz1 ChIP-Seq experiments were performed to measure the genomic location of these factors. This study revealed that chromatin remodeling during repair of DNA damage by NER is initiated from specific sites of GG-NER complex binding located at the boundary of higher-order chromatin structures known as Micro-C boundary or CID boundaries. This result demonstrated that, in undamaged cells, The GG-NER complex binds to these Micro-C boundaries, and is flanked by the barrier histone variant, H2A.Z-containing nucleosomes. Following UV damage, these boundary nucleosomes are remodelled in a GG-NER complex dependent fashion to facilitate repair initiated from these sites, to more distal sites between these Micro-C boundaries. This study also demonstrates the importance of this mechanism to the efficient removal of DNA damage by NER, providing insight into how defects in chromatin remodeling might drive mutagenesis in cells.

The genome-wide nucleosome maps were generated to analyse UV-induced changes to the nucleosome landscape. The genomic distribution of changes to the three core nucleosome parameters were examined that quantify *occupancy*, *fuzziness* and *position*, to identify the subset of nucleosomes that are altered in response to UV-irradiation. These findings demonstrated that chromatin remodeling at this level occurs predominantly through dispersed local changes to nucleosome occupancy and fuzziness. Furthermore, the GG-NER complex binding sites identified in this study, are not simply regions of repair initiation, but they are also locations of UV-induced histone remodeling, involving nucleosomes containing the histone variant H2A.Z. My results show that in undamaged cells these H2A.Z-containing nucleosomes represent barriers; gate-like structures that constrain and sequester the GG-NER complex at these genomic positions. DNA repair may be initiated by structural rearrangement of these barrier sites, allowing the GG-NER complex to redistribute from its initial binding locations in undamaged cells. The UV-induced loss of H2A.Z-containing nucleosomes essentially relieves the barrier effect, permitting the GG-NER complex to redistribute. Intriguingly, this process might serve to concurrently restrict RNA pol II transcription that is known to require H2A.Z-containing nucleosomes for efficient gene transcription initiation (Weber et al. 2014). Therefore, this mechanism may contribute to the inhibition of bulk transcription in response to DNA damage, while at the same time driving the efficient search for DNA damage by the GG-

NER complex. Shut-down and restoration of normal gene expression is an established hallmark in maintaining the stability of the genome in response to DNA damage (Ciccia and Elledge 2010).

Higher-order chromatin structure in yeast has been identified following the introduction of methods that map distal nucleosome-nucleosome interactions, forming structural units that are classified as CIDs (Hsieh et al. 2015; Hsieh et al. 2016). These structures typically encompass 1 to 5 genes, and range in size from a few kilobases up to 10 Kbp. My results described in this chapter showed that ~50% of GCBS's can be found precisely at the boundaries between these genomic features. Conceivably, these nucleosome-nucleosome interactions contained within CIDs may represent higher-order levels of nucleosome structure that may also be remodelled during GG-NER. We will further investigate this notion in greater detail in the future. In line with previous studies in our laboratory, the DNA translocase activity of the GG-NER complex could induce the remodeling of higher order chromatin structure, similar to the loop-extrusion model suggested for CTCF-Cohesin complexes in higher eukaryotes (Sanborn et al. 2015). In this model, two CTCF-Cohesin complexes bind to the chromatin and extrude DNA through the cohesin ring structure until they encounter a CTCF binding site (Sanborn et al. 2015). The CTCF and cohesin factors reside at the base or boundary of these loop structures, which may be analogous to the boundary positions to which the GCBS complex binds in the yeast genome. Although the loop-extrusion model has not been demonstrated in yeast, and the lack of a yeast homolog for CTCF, excludes the possibility of a direct parallel mechanism. However, this study also suggests that the redistribution of the GG-NER complex, by virtue of the DNA translocase activity of Rad16 could act as a wedge to disrupt the higher-order contacts that exist in the DNA loops that make up the CIDs. Future research aims to investigate the remodeling mechanism of higher-order chromatin structure using the micro-C methodology.

Cancer is a disease of genomic instability and is driven by somatic mutations acquired within the individual. The mutational history that gives rise to a tumour is recorded in that cancer genome because of the clonal nature of cancer. The mutational spectrum measured in a cancer genome is derived from a combination of extragenic and intragenic factors. In the recent cancer genome study, the mutational pattern/signatures were extracted from whole cancer genomes or whole exome sequencing data available publicly in different types of databases. When these mutational patterns are plotted according to their types, rainfall plots, inter-mutational distances, strand-biases and the frequency of

mutational signatures it displayed the non-random distribution of mutational patterns, suggesting a structure and organization within the cancer genome. Our laboratory revealed that, DNA damage and repair are also not randomly distributed, and there is a structure and organisation of genome-wide DNA damage and repair rates within the cell's chromatin environment. Importantly, these studies showed that, deletion of either Rad16/Rad7 (GG-NER or chromatin remodeling), Gcn5 (chromatin modifier), Htz1 (histone variant) can alter the pattern and distribution of wild type relative DNA repair rates throughout the genome, raising the possibility that defective DNA repair, chromatin modification or faulty histone variant exchange, might also affect the distribution of mutations acquired within the genome. Determining whether this is indeed the case will likely help to explain how novel classes of cancer-causing genes, which are involved in modifying chromatin structure, drive tumorigenesis.

To address this, a novel workflow was developed for accruing and filtering a catalogue of acquired mutations derived from yeast cells. My results demonstrate that the use of NMF for *de novo* extraction of mutational signatures from these mutational catalogues successfully decomposed the biological processes of mutagen exposure and DNA repair deficiency. Additionally, the cosine similarity and reconstruction of mutational profiles with known PCAWG signatures provided additional confirmation for the conservation of the mechanisms of mutagenesis between the yeast and human genomes. Here, I provide a proof-of-principle for the use of 'IsoMut' (Pipek et al. 2017) to accumulate unique heterogenous sub-clonal mutations from multiple isogenic samples. The subsequent filtration using a tuning curve for passage 1 control sample mutations results in the selection of sub-clonal mutations that occurred exclusively during the ~1100 generations of yeast cells division. This way of accumulating mutations is important for understanding the biological processes operating within a controlled isogenic biological system. IsoMut uniquely exploits the isogenic nature of the samples by filtering out SNVs that are shared between different clones. This approach successfully detects the mutational pattern with or without UV exposure allowing us to understand the biological process of mutations operative in a population of cells.

The genomic distribution of mutations were examined in realtion to genomic features for comparison with the distribution of relative repair rates in response to UV damage. This result reveals that the distribution of mutations within the genome is related to the structure and organisation of the relative repair rates described previously (Figure 1.19).

My results showed that, the expected number of UV-induced mutational distributions over ORFs results from the combined activity of GG-NER and TC-NER observed previously. This result showed that, in wild types cells, TES is more mutation prone than TSS, which is in line with our previously observed repair rate data. A similar pattern is also observed in melanoma cancer genome with higher prevalence of somatic substitutions at 3′ end, compared to 5′ end of genes (Pleasance et al. 2010a). This is possibly the consequence of expression-related repair processes, which are less well-organised at 3′ end of the transcript and hence the mutation prevalence is high. Additionally, significant depletions of mutations observed around TES in GG-NER defective cells also in line with the higher rate of repair in the 3' end of the genes observed previously (Yu et al. 2016). At this moment, we don't know what causes the higher relative repair rates at 3′ end of a genes in GG-NER defective cell. At the same time, we don't know the cause of the TC-NER defective relative repair rate pattern observed following UV exposure. Future work aims to analyse the relative repair rate of TC-NER deficiency, as well for getting the complete picture of how NER operates within the chromatin environment and how these processes determine the mutational patterns within yeast genome. However, the opposite phenomenon is observed in histone modifier defective cells; lower relative repair rates over ORF results in lower observed mutations. Although, this observation is in line with the known UV sensitivity of these strains, it is not easily explained with respect to the distribution of relative repair rate. This remains to be determined. Whether other types of chromatin modifier might be involved in the repair of UV-induced damage in Gcn5 mutants over ORF, which results in lower than expected mutations following UV damages.

Building on these findings, here I showed that the genomic features that are important for repair organisation determine the location and types of mutations within genome. This might be a main reason for mutational heterogeneity observed in cancer. Plotting the distribution of mutations in various genomic features showed that mutations are not randomly distributed within genome. Following UV damage, the distribution of mutations within the yeast genome is altered in the context of the genomic features that are defined by their differential repair rates. Interestingly, similar findings for the cellular repair capacity at different genomic features suggest that heterogeneous repair rates can influence the distribution of alkylating damage induced mutations within yeast genome (Mao et al. 2017a).

In chapter III, and our recent publication (van Eijk et al. 2018), we reported that, GG-NER is organised from a subset of Abf1 containing NFR, predominantly found at Micro-C boundary and bound by histone variant, H2A.Z, which is also positioned next to NFR (Albert et al. 2007). My data showed that, following UV exposure, the distribution of mutations within NFRs is influenced by loss of NER activity, and by the presence of H2A.Z histone variant in NFR-adjacent nucleosomes within the yeast genome. This result confirms that organisation of GG-NER determines the location of mutations within genome. Additionally this observation showed that in comparison to NFR, SPN (~10,000, data obtained from (Brogaard et al. 2012)), has a protective effect with respect to induction of mutations. This is in line with the initial lower amount of UV-induced damages observed (Mao et al. 2016). It remains to be determined whether, variation in repair rate, or the capacity of these nucleosomes to reduce DNA damage levels and therefore prevent subsequent accumulation of mutations at these sites.

As mentioned, GG-NER is organised from GCBSs, many of which are found at Micro-C boundaries, bordered by H2A.Z-containing nucleosomes. Importantly, my results described in this thesis showed that, following UV damage, lack of organised repair from these sites results in higher mutational load. One exception is observed in case of *htz1* mutant. This data suggests that, loss of nucleosome remodeling capacity by GG-NER complex, or loss of histone modification around GG-NER complex surrounding nucleosomes and presence of histone variant, contributes to the mutational load around these higher order Micro-C boundaries. This result also suggests that the structure and organisation of GG-NER around its origin of repair might contribute to the mutational heterogeneity observed in cancer. In line with these observations, one recent study also showed that, histone variant (H3.3) is required for IgV gene diversification by generating regions to somatic-hypermutation generated by AID (Romanello et al. 2016). Another study suggested that, H3.3 is a 'target marker' for somatic hypermutations (Aida et al. 2013). This observations thus provide a mechanistic insight about involvement of histone variant might determine the mutation hotspot in cancer genome.

In eukaryotes, several higher order chromatin organisation has been described such as gene loops, enhancer-promoter loops, "topologically-associating domains"/ "chromosomally-interacting domains" (TADs/CIDs), and lamina-associated domains (LADs). These higher order organisation is implicated in biological activities. For example, association of gene loop with promoter and terminator directionally in yeast (O'Sullivan et al. 2004; Tan-Wong et al. 2012), TADs/CIDs are associated with the

functional regulatory domain in mammals (Symmons et al. 2014). Most recently several whole cancer genome sequencing study also showed the hotspots of mutations in similar types of higher order chromatin organisation mostly associated with CTCF containing chromatin loops (Guo et al. 2018; Kaiser and Semple 2018). Future research aims to investigate, to what extent this higher order chromatin structure modulates the distribution of mutations, and also how repair is organised from these sites to maintain genome stability. Future study targeting these Micro-C techniques, will provide mechanistic insight of how factors that affect GG-NER processes such as loss of HAT or even histone variant exchange deficiency drive tumorigenesis by altering mutational distribution around these higher-order chromatin structures.

In addition to higher order chromatin structure, my result showed variation in the mutational distribution depending on the types of TFBSs, following UV exposure. Similar results observed in two recent papers revealed that binding of transcription factors to its cognate binding sites might interfere with NER factors and thereby results in a higher density of mutations found at these sites in melanoma (Perera et al. 2016; Sabarinathan et al. 2016). However, the observations of my results demonstrate that the essential biological processes such as variation in transcription and DNA repair might impede each other's activities and this process depends on types of regulatory factors and therefore its function within genome. Both Abf1 and Reb1 are GRFs in yeast, however their functions within yeast genomes are not fully characterised yet.

Organisation of repair is more accurately represented by GCBSs as opposed to TFBSs. Therefore, I used a list of GCBS summits and their flanking regions. These positions were derived from the Abf1 ChIP-Seq peak summits calculated by MCCS2 as described in our recent publication (van Eijk et al. 2018). The summits were used here, because the width of the Abf1 ChIP-Seq peaks can vary up to 1 order of magnitude. This way the width about each positioned is fixed to the area-of-effect observed for repair surrounding this positions. Mutational distributions in my result confirm that GG-NER complex and histone modification around GCBS summits is required for efficient repair. Additionally, histone variant also might responsible for mutational bias around this sites. These sites are predominantly bounded by H2A.Z containing nucleosomes and are sites for UV induced hyperacetylation around these domains (Yu et al. 2016). My results showed that alterations of substitution mutation as around GCBSs depends on GG-NER or TC-NER factor. Similarly, the mutational distribution in defective histone modification and histone

variant exchange provide the proof of involvement of Gcn5 and variant exchange during GG-NER process from this sites.

The transcriptions strand bias analysis following UV damages revels the mutational strand asymmetry between transcribed and untranscribed strand. A recent study targeting several cancer genome showed that both replicative and transcriptional asymmetries are widespread across cancer (Haradhvala et al. 2016) suggesting the involvement of both DNA damage and repair processes. My data for transcriptional strand asymmetry reveals, following UV damage, TC-NER activity is the main reason for transcriptional strand asymmetry. Additionally, this study also find a novel T>A types of mutational bias for transcribed strand probably due to competitive activity of TC-NER and GG-NER on UV induced lesions.

Following DNA damage, shut-down and restoration of normal gene expression is an established hallmark in maintaining genome stability (Ciccia and Elledge 2010). However, failure of recovery RNA synthesis after UV damage is observed in TC-NER defective *rad26* mutant yeast, which is the human homolog of CSB. Both CS and XP patients showed this phenomenon (Mayne and Lehmann 1982). It will be interesting to see whether gene expression levels in different mutant backgrounds has an effect on mutation induction. It is known that UV irradiation has an effect on gene expression. However, how this impinges on the mutational asymmetry observed in these experiments is not known. It is possible that the mutational strand asymmetry correlates with gene expression levels in response to UV damage.

Furthermore, sequencing of several intermediates clones will help to follow the timing of mutagenesis following UV damage in various repair mutants. The connection between early repair kinetics (<3hrs) and the accumulation of mutations over several passage or generations is not known. It would be interesting to measure the mutation accumulation rate, expressed as allele frequency. At passage 30 the majority of mutations have an allele frequency of 1. Detecting mutations at early passages will reveal the kinetics of the allele frequency as measured by IsoMut. This is an important aspect of the somatic mutation theory of cancer.

The results of this thesis also showed that, UV-induced mutations are comparatively difficult to repair at low acetylated regions which is marked for closed chromatin than high acetylated regions. Previous study showed that, UV induced damages are uniformly distributed in their linear genomic structure (Teng et al. 2011; Hu et al. 2017), however

the repair rate of these damages modulated by accessibility of the damaged DNA which is determined by acetylation status of chromatin to facilitate repair (Yu et al. 2011; Yu et al. 2016; Hu et al. 2017), which might result in heterogenous distribution of mutations within genome, often associated with human cancer (Schuster-Böckler and Lehner 2012). The recent access-repair-restore model showed that, the organisation of chromatin pose barrier to repair. Following DNA damage, chromatin remodelling is necessary for transcription as well as for efficient repair (Polo and Almouzni 2015). Faster repair at open chromatin in compare to close chromatin might determine the mutational load**.** This result showed that, mutational bias towards low acetylated regions within yeast genome following UV damage. Suggesting, UV induced mutations are prone to occurs at low accessible regions. This result also confirm that both UV induced substitutions and replication error-prone mutations are predominating at later stage of replication. This observation suggest, repair is less efficient at later stage of replication and result in higher mutational rate at late replicating regions. Similarly, late replicating regions are also regions of heterochromatic or low DNA accessible regions as mentioned in the previous section. Strikingly, in undamaged cells, enrichment of substitutions mutations at early replicating regions in *rad26* TC-NER mutant. At this moment, we don't know the reason behind this. I suggest that, this might be due to lack of RNA synthesis recovery following UV damage and replication around these site. This notion need further details investigations.

Plotting the UV induced relative repair rate around the centre of the late replicating regions within yeast genome displayed relatively lower repair rate in GG-NER defective *rad16* mutant cells in comparison to wild-type cells. Additionally, this data also reveals that, the repair rate affected mostly at the centre of the late replicating regions in both wild-type and *rad16* mutant cells and gradual increase in relative repair rates either sides of these late replicating regions, which is regions of early replicating area. Providing evidence of lower repair rate result in the higher mutations density towards late replicating regions. Future work targeting synchronise cells repair rate, replicating timing and mutagenesis study might able to tackle this heterogeneity in the distribution of substitution mutations frequently also observed in various types of cancer (Haradhvala et al. 2016).

The use of NMF to infer the mutational signatures or the biological processes operating in a set of complex omics data was successfully employed during the decomposition of active biological processes operating within these experimental datasets. The

employment of factorization rank survey and consensus matrix for evaluation of number of mutational signatures was successful for describing the active biological process operating within the yeast genome. This idea can be anticipated in future translational research to harness the power of such whole genome analysis for diagnostic and personalise medicine approaches to improve cancer therapy. My results showed that a novel mutational signatures (Signature C, in chapter V) is associated with TC-NER activity, which showed high similarity with the recently reported PCAWG mutational signatures SBS5 and SBS40. The biological processes of these cryptic signatures is unknown. The biological function of yeast Rad26 is homologous to human CSB as both are involved in repair of oxidative DNA lesions in the nucleus as well as in mitochondria (Melis et al. 2013). CSB, a SWI/SNF ATPase containing chromatin remodelling factor, in human cells, plays a crucial role, not only in onset of TC-NER, but also in RNA pol II transcription activities. In future, this striking role of Rad26 need to be addressed in more precise detail in relation to relative repair data in order to understand how these mechanisms contribute to the unique mutational pattern generated by TC-NER activity.

The study has significant implications for understanding the biological processes of genomic instability, frequently observed in cancer. This study demonstrated the utility of whole genome sequencing in cell lines as a mutagenesis assay. I measured the mutagenic effect and defined the mutation spectrum caused by UV irradiation; a common genotoxic agent. Matching mutational signatures to DNA repair deficiency has a tremendous potential to stratify cancer therapy tailored to DNA repair deficiency. This approach appears advantageous over genotyping marker genes, as mutational signatures provide a read-out for cellular repair deficiency associated with either genetic or epigenetic defects. Following on from our study, we expect that analysing DNA repair–defective model organisms and human cell lines, alone or in conjunction with exposure to defined genotoxic agents, will contribute to a more precise definition of mutational signatures occurring in cancer genomes and will help to establish the aetiology of these signatures.

In conclusion, the structure and organisation of repair plays an important role in maintaining genome stability. This study demonstrates that in undamaged cells, DNA repair complexes are positioned at hundreds of boundary regions that define the presence of CIDs; genomic domains of higher order nucleosome-nucleosome interactions. Suggesting that this arrangement might represent origins of DNA repair initiation that promote the efficient repair of DNA damage in chromatin. Initiating chromatin remodeling from defined origins could effectively reduce the search space for DNA

damage recognition by compartmentalising the genome into functional modular chromatin structures that can be rapidly remodelled and efficiently repaired. Therefore, characteristic structural features of CIDs emerge when the genome is organised in this way – this ensures the rapid search and repair of genetic damage in chromatin. Additionally, the distribution of the UV-induced mutational pattern in wild-type cells depends on the structure and organisation of repair processes. Loss of organised repair, such as defective GG-NER and TC-NER or factors that modulate these repair processes (Gcn5, Htz1), alters the normal distribution of the mutational pattern. Therefore, understanding of the structure and organisation of DNA damage and repair underlying the development of mutational spectra observed in human disease has enormous implications for prevention and therapy.

Appendix I

Chemical and reagents of liquid and solid media

1. Growth media

YPD (1.0 L)

10.0 g Bacto Yeast Extract
20.0 g Bacto Peptone
20.0 g Glucose
Made up to 1.0 L with H_2O. Autoclave at 125.0°C for 15.0 minutes.

YPD Agar (1.0 L)

10.0 g Bacto Yeast Extract
20.0 g Bacto Peptone
20.0 g Glucose
12.5 g Bacto Agar
Made up to 1.0 L with H_2O. Mixed well and autoclave at 125.0°C for 15.0 minutes.
Cool down to 55°C, pour ~25.0 ml per Petri dish and store at 4.0°C.

2. Stock Solutions

PBS (1.0 L)

8.0 g NaCL
0.2 g KCl
1.8 g $Na2HPO_4.2H\ H_20$
0.24 g KH_2PO_4
800.0 ml H_2O
(137.0 mM NaCl, 2.7 mM KCl, 10.0 mM $Na_2HPO_4.2H_2O$, 2.0 mM KH_2PO_4)
Adjust the pH to 7.4. Add H_2O to 1.0 L. Autoclave at 125.0°C for 15.0 minutes.

1.0 M Tris

121.1 g Tris base
800.0 ml H_2O
Adjust the pH to the desired value (7.6 or 8.0) by adding concentrated HCl.
Add H_2O to make 1.0 L.

0.5 M EDTA (pH 8.0)

186.1 g $EDTA.Na_2.2H_20$
800.0 ml H_2O
Adjust the pH with NaOH to pH 8.0.
Add approximately 100.0 ml H_2O to make 1.0 L. Autoclave at 125.0°C for 15.0 minutes.

10 x Tris-EDTA (TE) Buffer (400 ml)

40.0 ml 1.0 M Tris-HCl (pH 7.5)
08.0 ml 0.5 M EDTA (pH 8.0)
352.0 ml of H_20

Sorbitol TE (1.0 L) (Kept in 4.0°C)

165.0 g Sorbitol
100.0 ml Tris . HCl 1.0M (pH 8.0)
200.0 ml 0.5 M EDTA
(0.9 M sorbitol, 0.1 M Tris-HCl [pH 8.0], 0.1 M EDTA)

Add 500 ml H_20 to dissolve the sorbitol. Adjust the final volume to 1.0 L.

3.0 M Sodium acetate (pH 5.2) (400.0 ml)

163.24 g Sodium acetate.$3H_20$

300.0 ml H_2O

Adjust the pH to 5.2 with acetic acid. Add H_2O to make 400.0 ml

Use filter to sterilise.

5.0 M NaCl (400.0 ml)

Dissolve 116.9 g of NaCl in 350.0 ml of H_20.

Adjust the volume to 400.0 ml with H_20.

Sterilise by autoclaving.

10% SDS (1.0 L)

Dissolve 100.0 g of SDS in 800.0 ml of distilled H_20.

Add distilled H_20 to make a total volume of 1.0 L.

DNA Lysis Buffer (1.0 L)

240.0 g Urea

11.69 g NaCl

5.0 g CDTA

5.0 g SDS

100.0 ml 1.0 M Tris-HCl (pH 8.0)

(4.0 M urea, 200.0 mM NaCl, 100.0 mM Tris-HCl [pH 8.0], 10.0 mM CDTA, 0.5% SDS)

Add 700.0 ml of H_2O to dissolve the chemicals and then adjust the final volume to 1.0 L.

3. Solutions for gel electrophoresis

50x TAE (1.0 L)

242.0 g Tris base

136.0 g Sodium Acetate.$3H_20$

200.0 ml 0.5 M EDTA

500.0 ml H_2O

Adjust to pH 7.2 with acetic acid.

Add H_2O to make 1.0 L.

4. Solution for ChIP

FA/SDS Buffer + PMSF

50.0 mM HEPES-KOH (pH 7.5)

150.0 mM NaCl

1.0 mM EDTA

1.0 % Triton-X 100

0.1 % Sodium Deoxycholate

0.1 % SDS

1.0 mM PMSF (Add just before use)

PMSF (Phenylmethylsulfonyl fluoride)

1.75 g of the serine protease inhibitor PMSF is dissolved in 100.0 ml 100% ethanol to make 100.0 mM PMSF in 100.0 ml. Stored at 4.0°C.

5x Pronase Buffer (100.0 ml)
12.5 ml 1.0 M Tris (pH 7.5)
5.0 ml 0.5 M EDTA
25.0 ml 10% SDS
Add H_2O to make 100.0 ml.
(125.0 mM Tris pH 7.5, 25.0 mM EDTA, 2.5% SDS)
1x Pronase Buffer
25.0 mM Tris pH 7.5, 5.0 mM EDTA, 0.5% SDS
LiCl Buffer (500.0 ml)
5.0 ml 1.0 M Tris-Cl (pH 8.0)
25.0 ml 5.0 M LiCl
1.0 ml 0.5 M EDTA
2.5 ml NP40
25.0 ml 10% Sodium Deoxycholate
(10.0 mM Tris-Cl (pH 8.0), 250.0 mM LiCl, 1.0 mM ETDA, 0.5% NP40, 0.5% Sodium Deoxycholate)
Add H_2O to make 500.0 ml.

5. Solutions for Nucleosomal DNA preparation.
2.5 mL YLE buffer (10 reactions)
1.25 ml 2.0 M Sorbitol
2.0 µl 14.3M β-Mercaptoethanol
56.25 mg YLE Powder (22.5 mg/ml final concentration)
Make up to 2.5 mL with water. Use immediately.

6. Adapter Annealing Buffer Composition (1X)

10 mM Tris, pH 7.5 - 8.0
50 mM NaCl
1 mM EDTA
Stock : 100x TE in which 1X TE should be (10mM Tris pH 7.5, 1mM EDTA) and 5 M NaCl.

Table A1.1: The adapter sequences used for ChIP-Seq and MNase-Seq library preparation.

Adapter Names	Sequences
Ion P1 adapter	5'—CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGAT–3'
ABC1	5'-ATCGTTACCTTAGCTGAGTCGGAGACACGC-3'
ABC2	5'-ATCGTTCTCCTTACTGAGTCGGAGACACGC-3'
ABC3	5'-ATCGAATCCTCTTCTGAGTCGGAGACACGC-3'
ABC4	5'-ATCGATCTTGGTACTGAGTCGGAGACACGC-3'
ABC5	5'-ATCGTTCCTTCTGCTGAGTCGGAGACACGC-3'
ABC6	5'-ATCGAACTTGCAGCTGAGTCGGAGACACGC-3'
ABC7	5'-ATCGAATCACGAACTGAGTCGGAGACACGC-3'
ABC8	5'-ATCGTTCCGCTCACTGAGTCGGAGACACGC-3'
ABC9	5'-ATCGTTCTCCTTACTGAGTCGGAGACACGC-3'
ABC10	5'-ATCGTTCGGTCAGCTGAGTCGGAGACACGC-3'
ABC11	5'-ATCGATTCGAGGACTGAGTCGGAGACACGC-3'
ABC12	5'-ATCGAACCACCTACTGAGTCGGAGACACGC-3'
ABC13	5'-ATCGTCCGTTAGACTGAGTCGGAGACACGC-3'
ABC14	5'-ATCGACACTCCAACTGAGTCGGAGACACGC-3'
ABC15	5'-ATCGACCTCTAGACTGAGTCGGAGACACGC-3'
ABC16	5'-ATCGTCATCCAGACTGAGTCGGAGACACGC-3'
ABC17	5'-ATCGACGAATAGACTGAGTCGGAGACACGC-3'
ABC18	5'-ATCGCAATTGCCTCTGAGTCGGAGACACGC-3'
ABC19	5'-ATCGTCCGACTAACTGAGTCGGAGACACGC-3'
ABC20	5'-ATCGATGGATCTGCTGAGTCGGAGACACGC-3'
Modified Adapter Strand	
Ion P1 Adapter Modified	5'-ATCACCGACTGCCCATAGAGAGGAAAGCGGAGGCGTAGTG*T*T-3'
ABC_Modified_1	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGCTAAGGTAACGAT–3'
ABC_Modified_2	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGTAAGGAGAACGAT–3'
ABC_Modified_3	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGAAGAGGATTCGAT–3'
ABC_Modified_4	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGTACCAAGATCGAT–3'
ABC_Modified_5	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGCAGAAGGAACGAT–3'
ABC_Modified_6	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGCTGCAAGTTCGAT–3'
ABC_Modified_7	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGTTCGTGATTCGAT–3'
ABC_Modified_8	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGTTCCGATAACGAT–3'
ABC_Modified_9	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGTGAGCGGAACGAT–3'
ABC_Modified_10	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGCTGACCGAACGAT–3'
ABC_Modified_11	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGTCCTCGAATCGAT–3'
ABC_Modified_12	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGTAGGTGGTTCGAT–3'
ABC_Modified_13	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGTCTAACGGACGAT–3'
ABC_Modified_14	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGTTGGAGTGTCGAT–3'
ABC_Modified_15	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGTCTAGAGGTCGAT–3'
ABC_Modified_16	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGTCTGGATGACGAT–3'
ABC_Modified_17	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGTCTATTCGTCGAT–3'
ABC_Modified_18	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGAGGCAATTGCGAT–3'
ABC_Modified_19	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGTTAGTCGGACGAT–3'
ABC_Modified_20	5'—CCATCTCATCCCT*G*CGTGTCTCCGACTCAGCAGATCCATCGAT–3'
Primers for amplification	
Forward	5-CCATCTCATCCCTGCGTGTCTC-3
Reverse	5-AACCACTACGCCTCCGCTTTCC-3
*ABC = Adapter Barcoded	# Red color indicate barcoded regions.

Appendix II

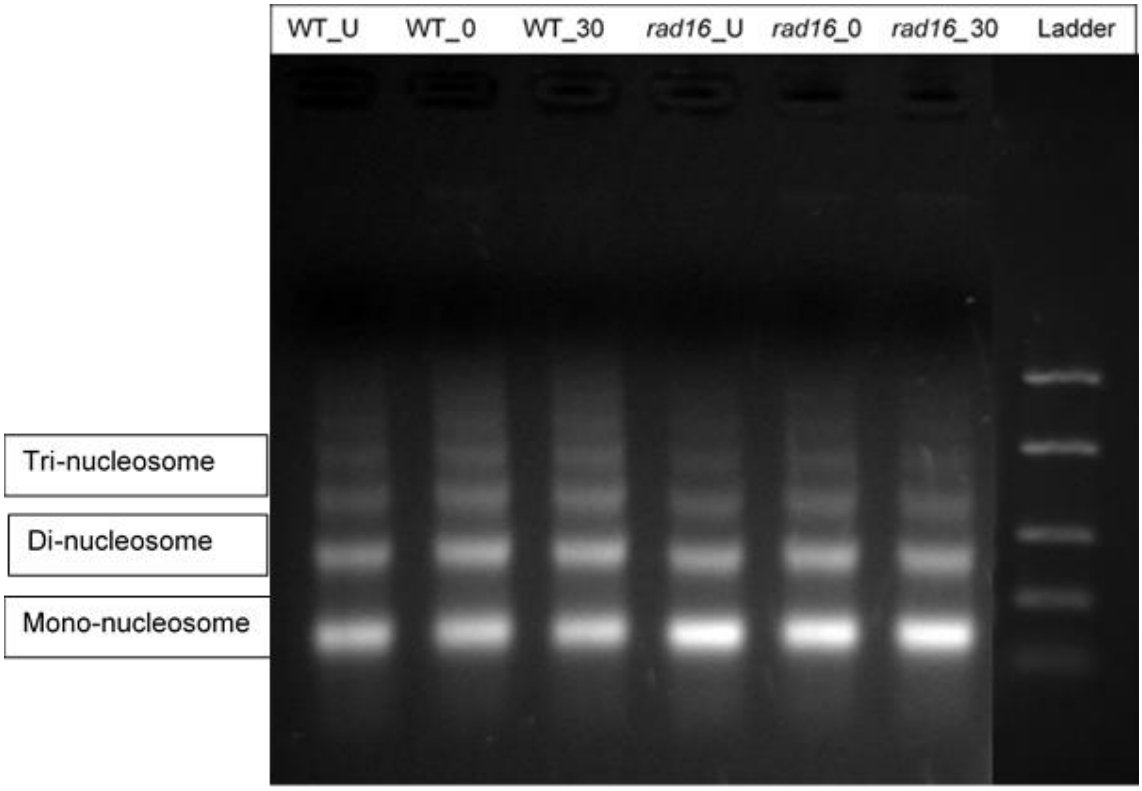


Figure A2.1: Agarose gel electrophoresis demonstrating the nucleosome DNA preparation in wild-type (left three lane) and *rad16* mutant (right three lanes) cells showing the mono-, di-, tri- nucleosomes fragments. FastRuler low range DNA ladder (ThermoFisher) was used to confirm nucleosomal sizes .

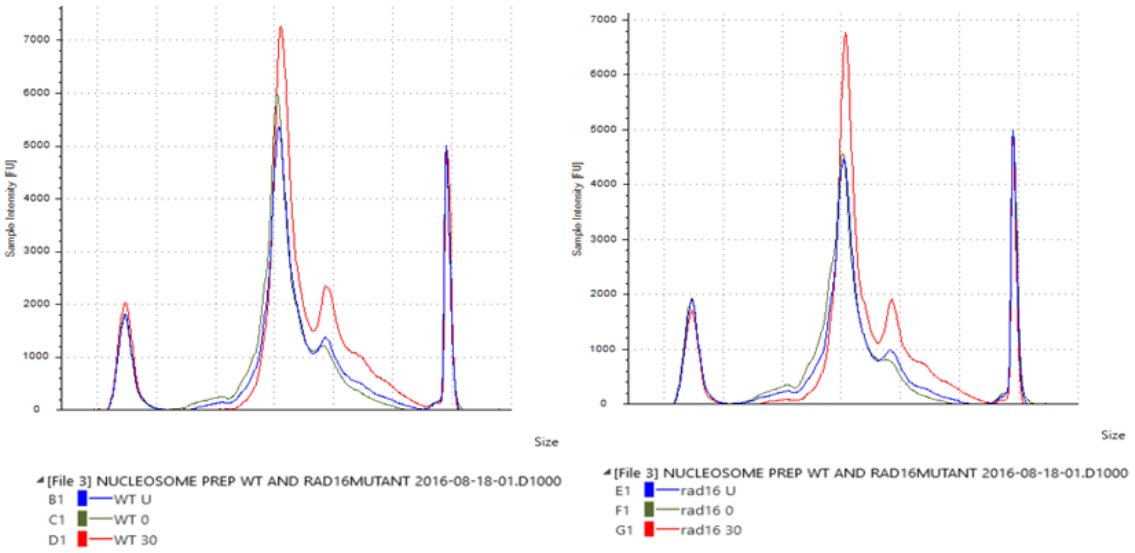
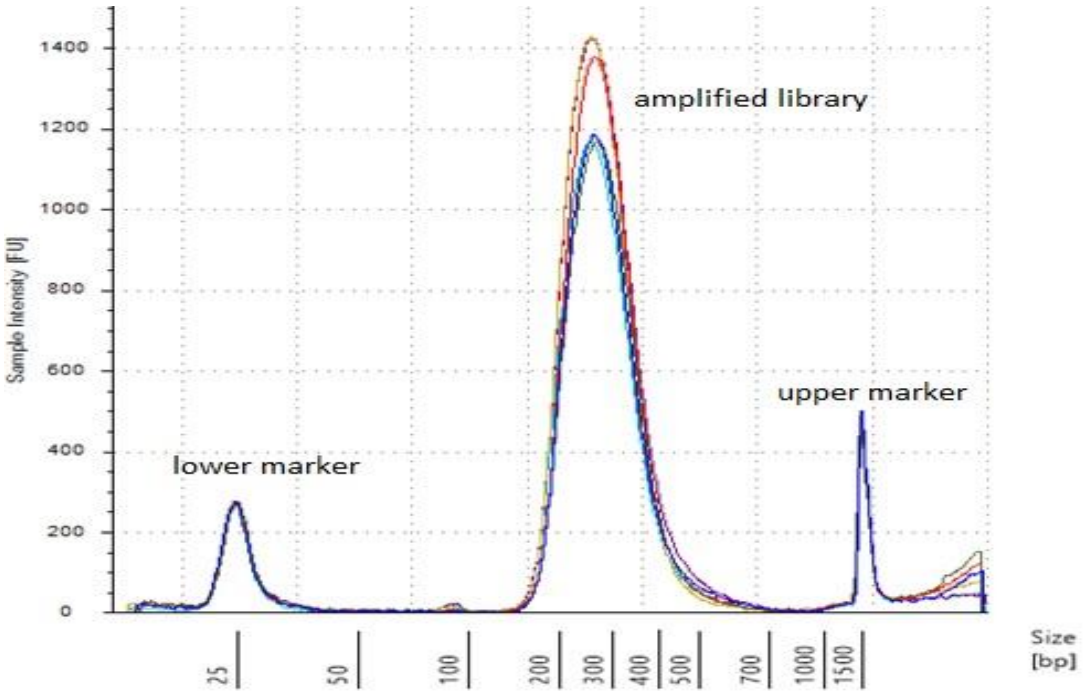


Figure A2.2: Bioanalyzer track showing successful nucleosomal library preparation for MNase-Seq in wild-type (left panel) and *rad16* mutant (right panel) cells.



FigureA2.3: Bioanalyzer track showing successful library preparation for ChIP-Seq of Abf1 in wild type cells (colour track represent different IP and IN samples).

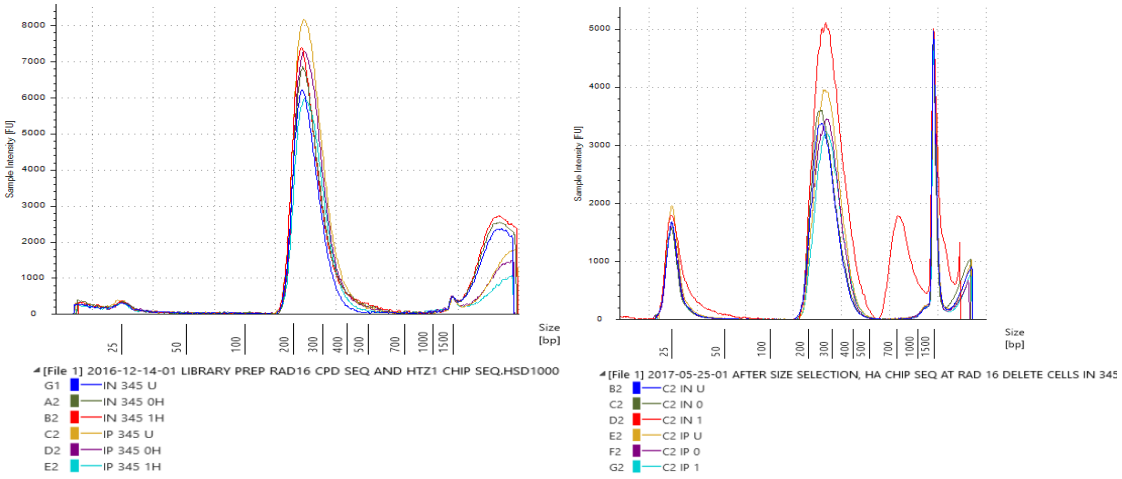


Figure A2.4: Bioanalyzer track showing successful library preparation for ChIP-Seq for Htz1 and in wild type (left panel) and *rad16* mutant cells (right panel)

Appendix III

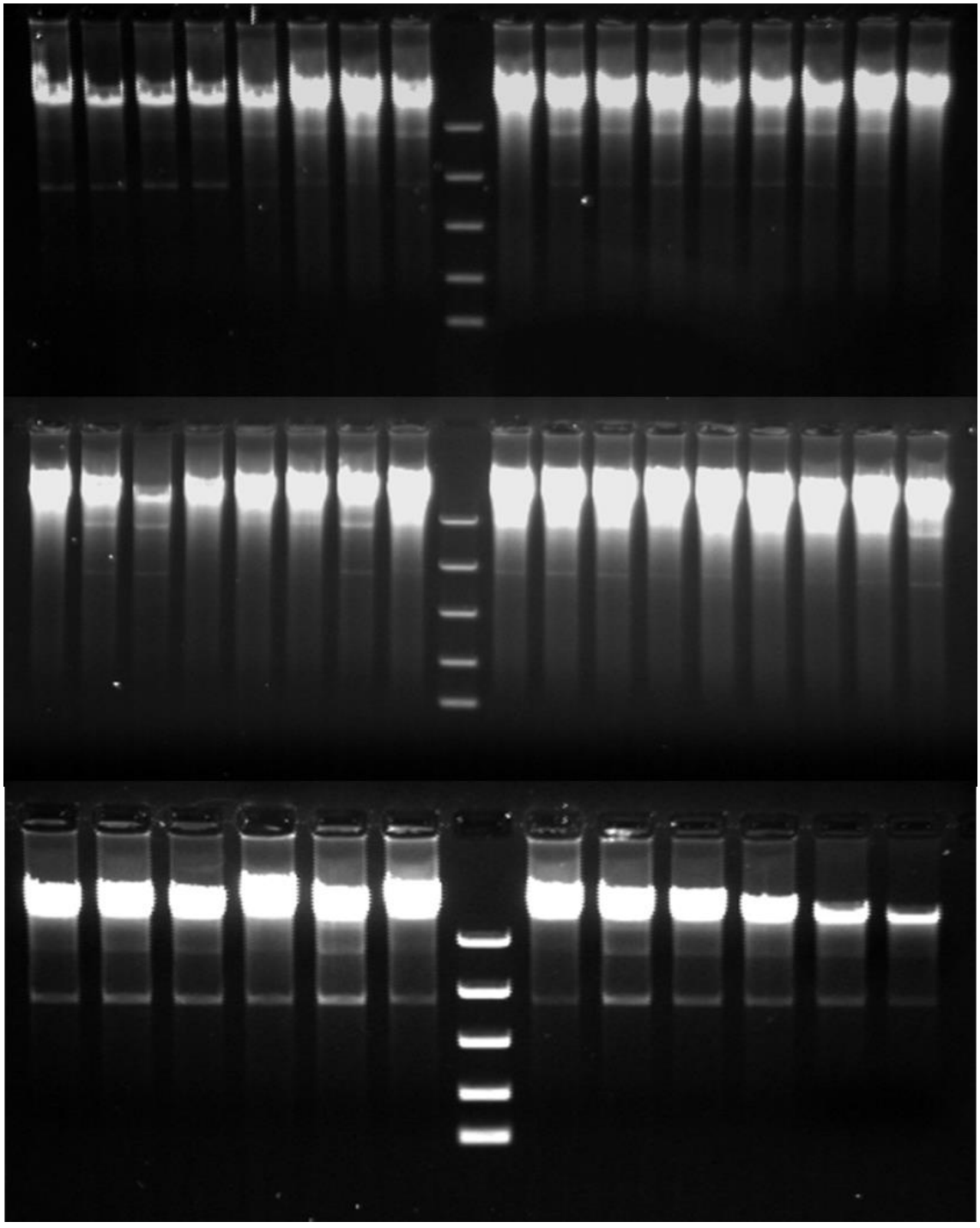


Figure A3.1: Extracted yeast genomic DNA examined by agarose gel electrophoresis. The 2-micron plasmids are also visible. The marker used here was high range FastRuler DNA ladders (ThermoFisher Scientific, Catalog #SM1123)

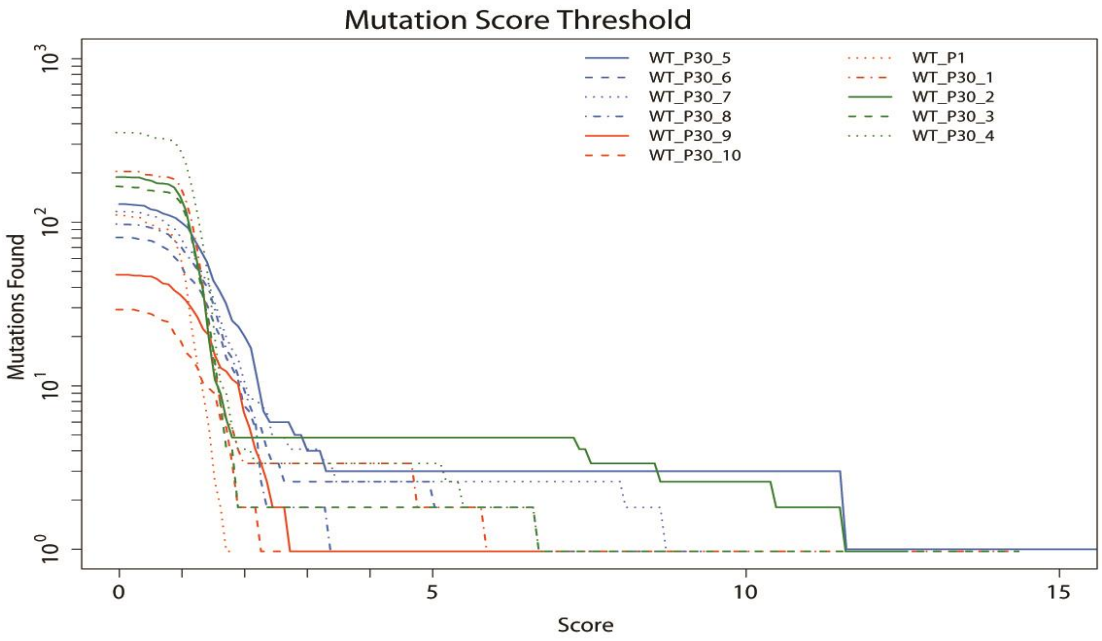


Figure A3.2: Tuning curve of wild-type cells showing the cumulative distribution of mutations as detected by IsoMut.

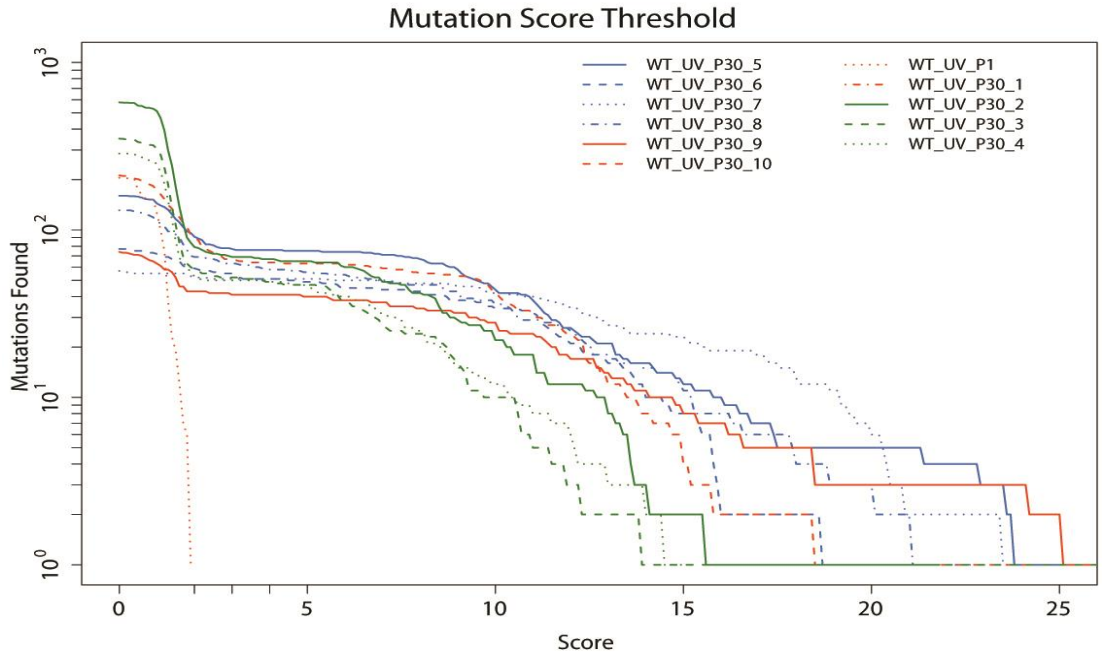


Figure A3.3: Tuning curve of UV-exposed wild-type cells showing the cumulative distribution of mutations as detected by IsoMut.

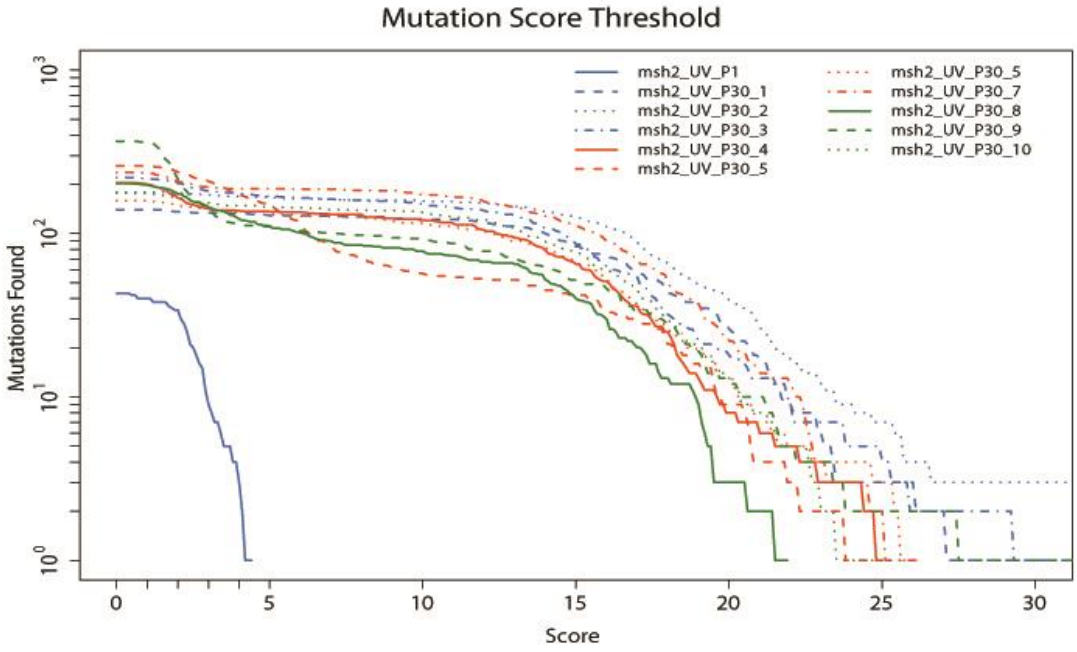


Figure A3.4: Tuning curve of UV-exposed *msh2* mutant cells showing the cumulative distribution of mutations as detected by IsoMut.

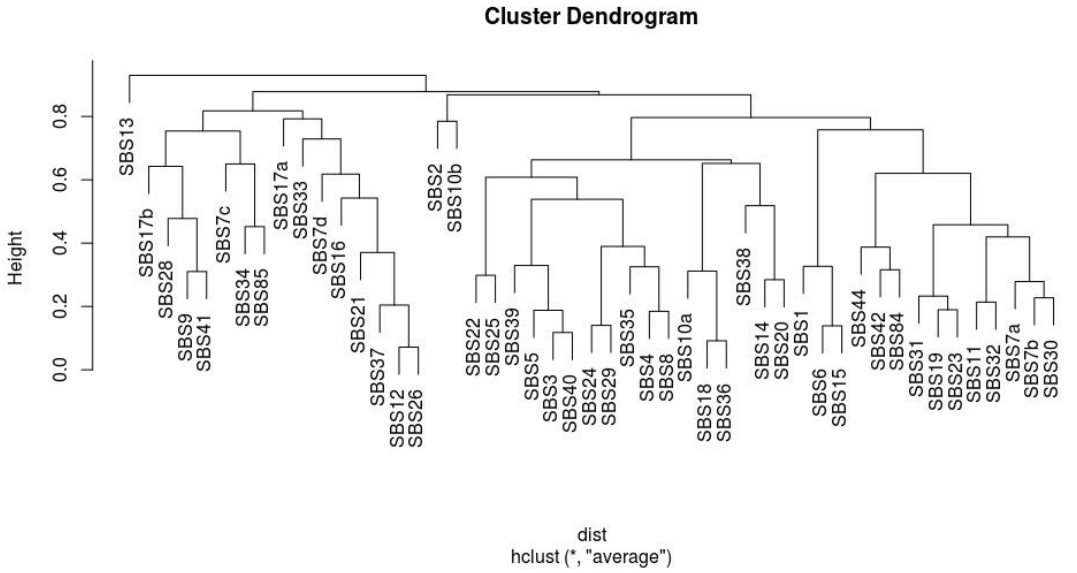


Figure A3.5: The hierarchical clustering of 49 PCAWG signatures based on the average linkage similarity.

Appendix IV

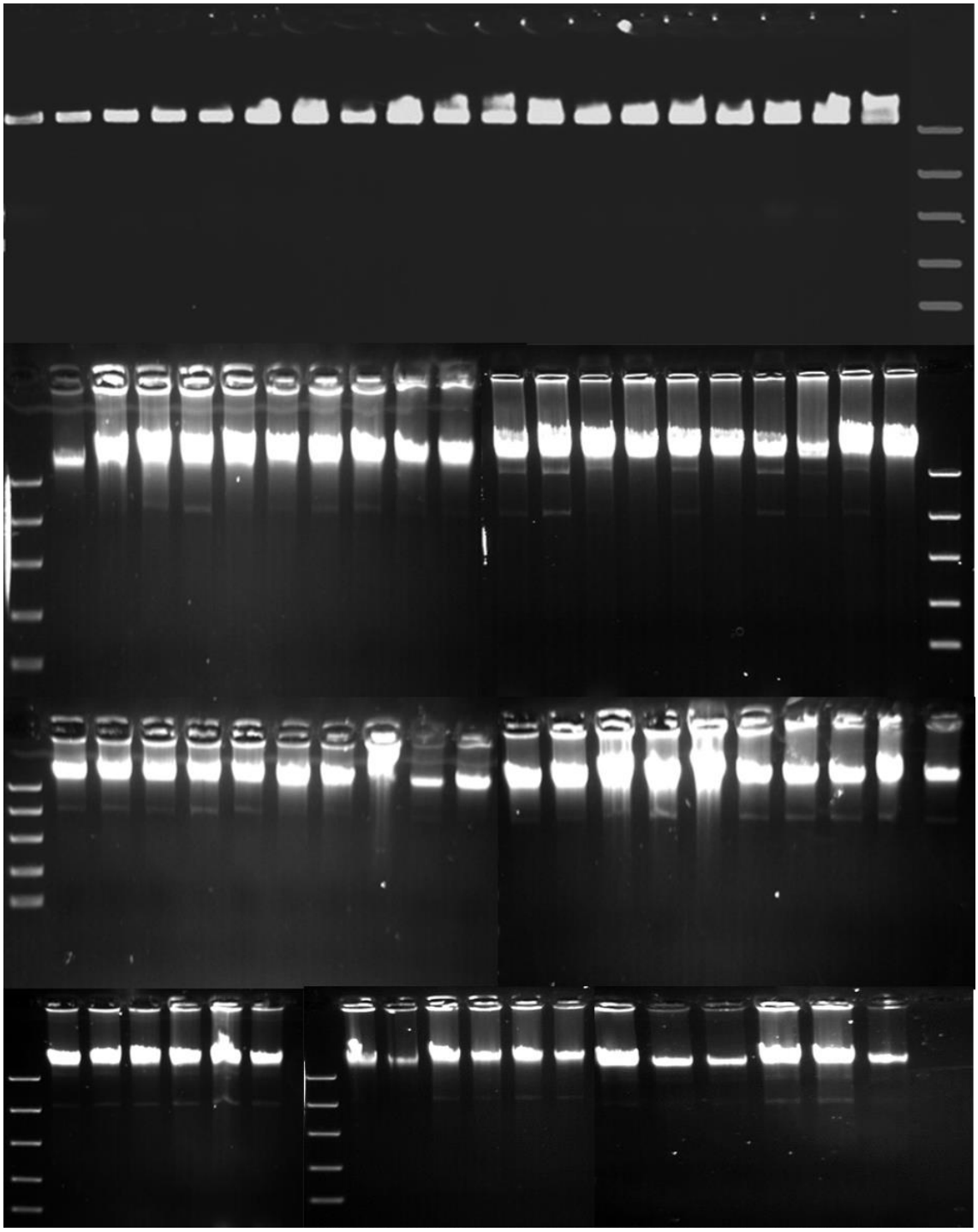


Figure A4.1: Extracted 78 yeast DNA samples (used for chapter V) examined by 1% agarose gel electrophoresis. DNA was extracted by yeast genomic DNA extraction protocol mentioned in chapter II. The marker used here was high range FastRuler DNA ladders (ThermoFisher Scientific, Catalog #SM1123).

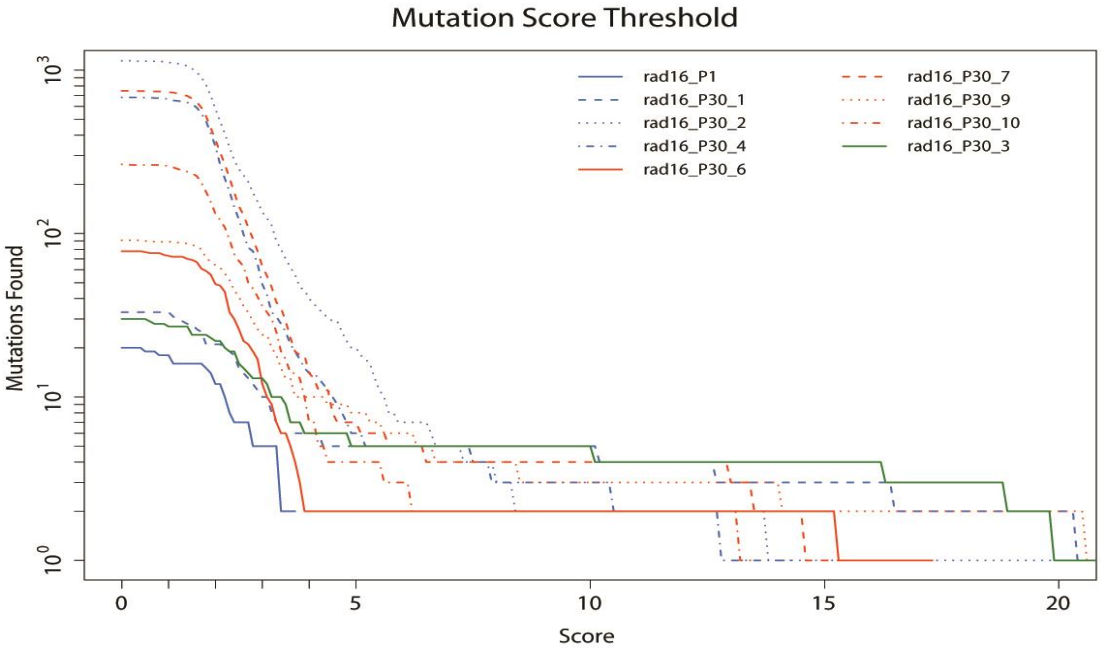


Figure A4.2: Tuning curve for *rad16* mutant cells substitution filtration

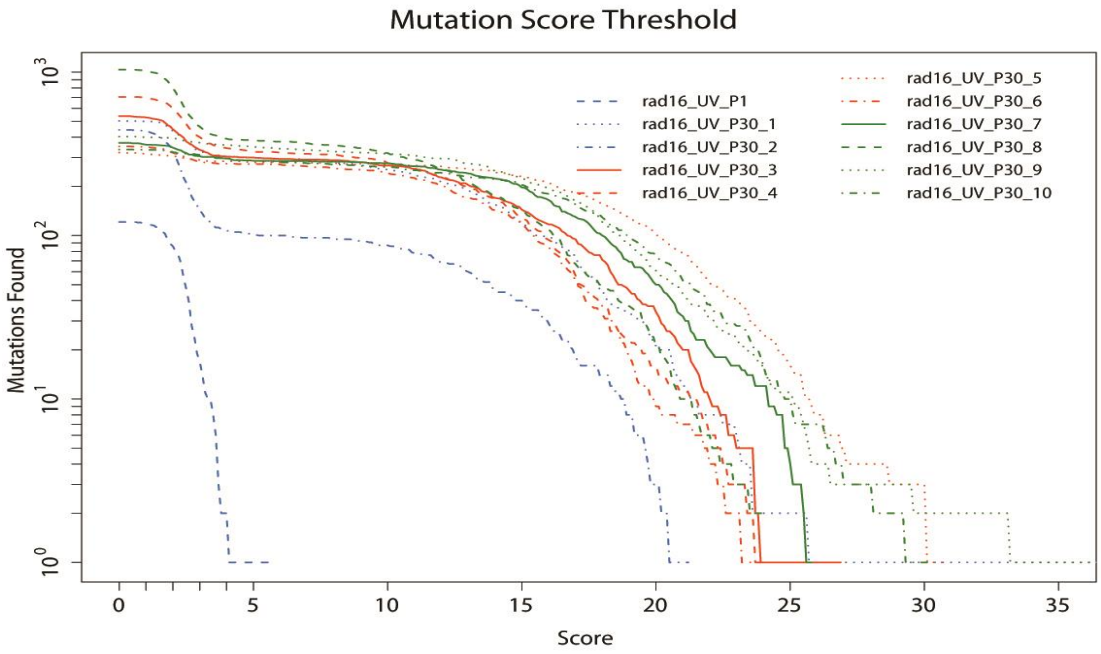


Figure A4.3: Tuning curve for UV-induced *rad16* mutant cells substitution filtration

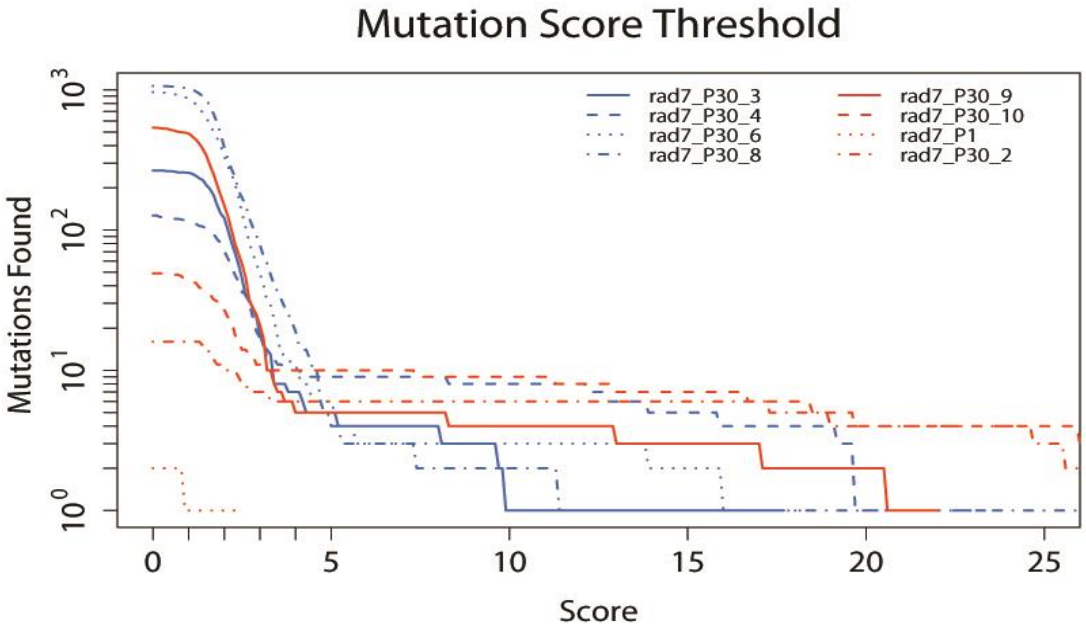


Figure A4.4: Tuning curve for *rad7* mutant cells substitution filtration

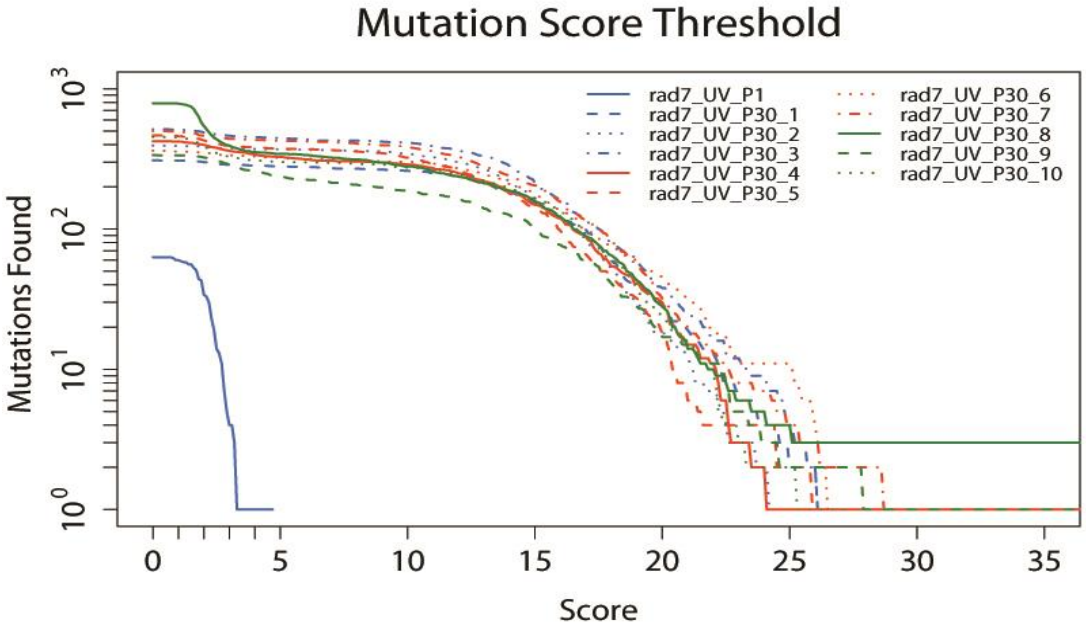


Figure A4.5: Tuning curve for UV-induced *rad7* mutant cells substitution filtration

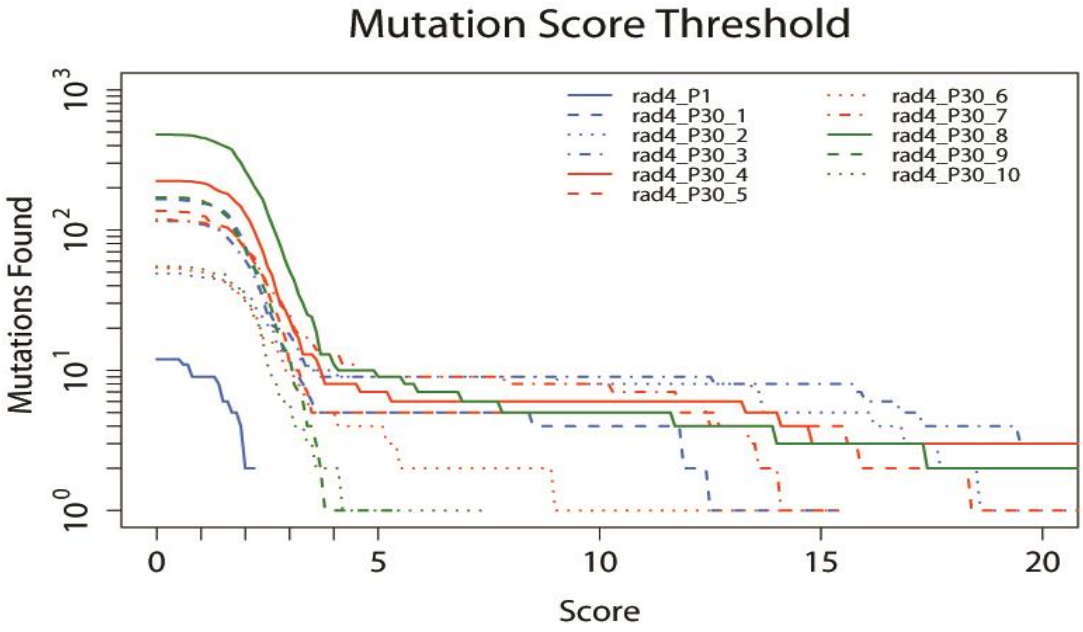


Figure A4.6: Tuning curve for *rad4* mutant cells substitution filtration

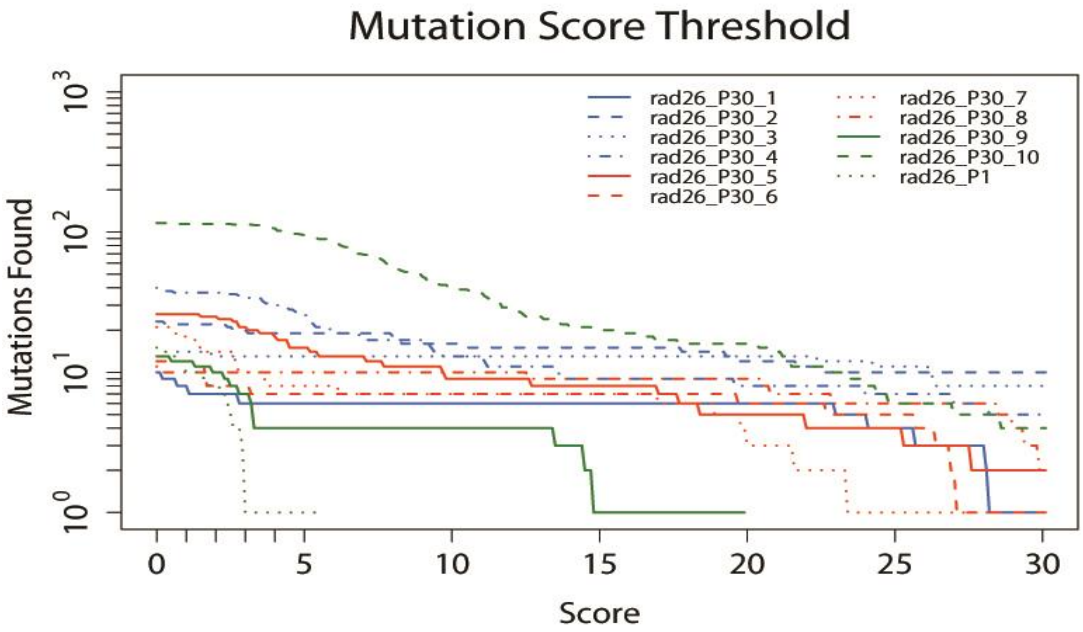


Figure A4.7: Tuning curve for *rad26* mutant cells substitution filtration

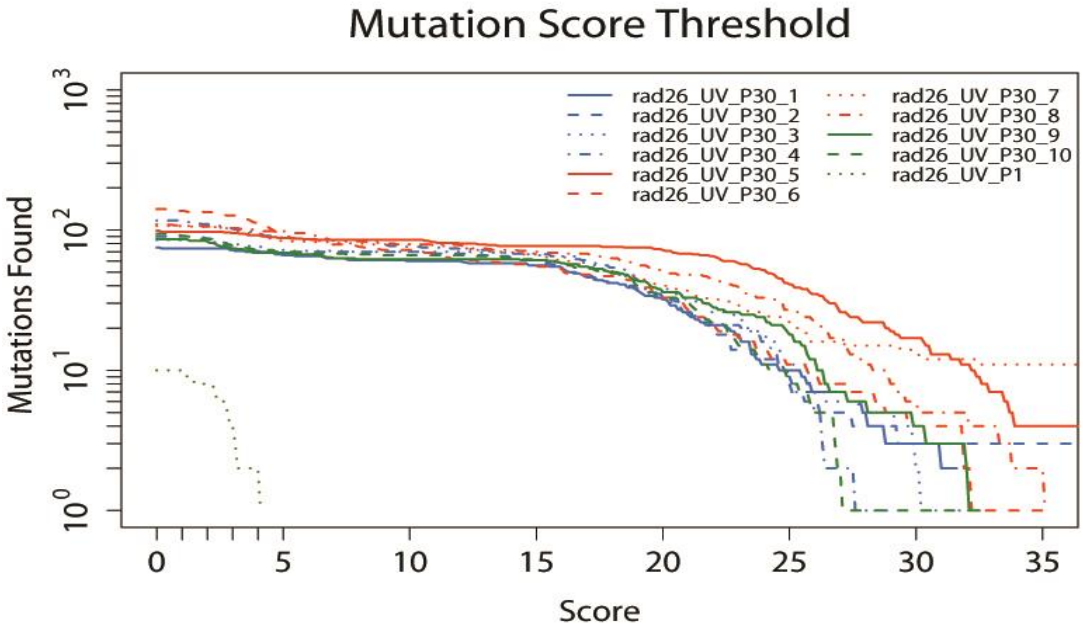


Figure A4.8: Tuning curve for UV-induced *rad26* mutant cells substitution filtration

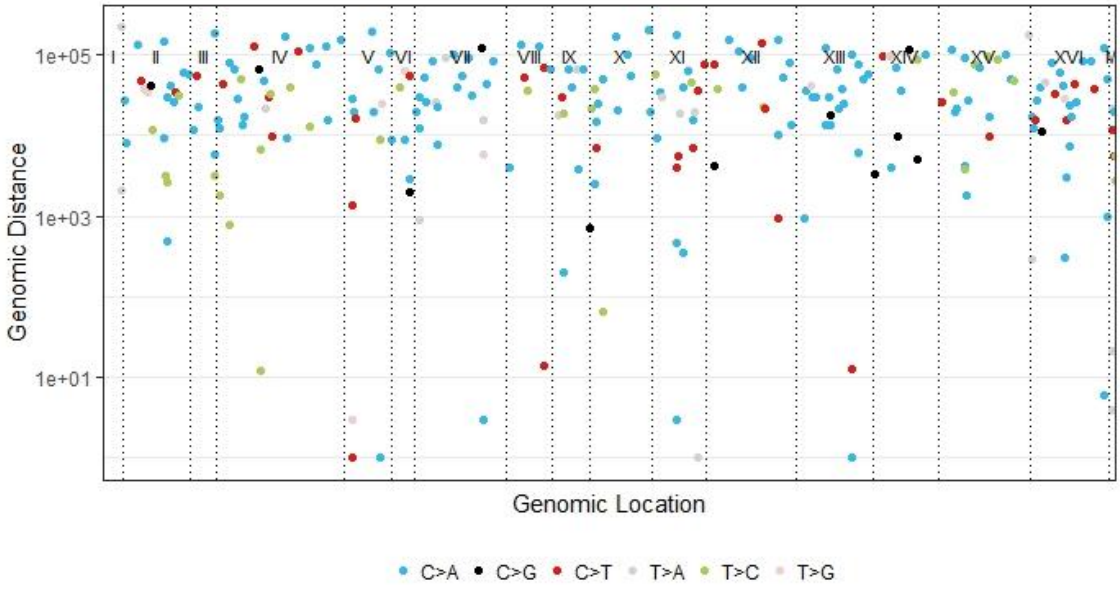


Figure A4.9: Rainfall plot of *rad16* mutant showing the genomic location of substitution mutations with inter-mutational distance

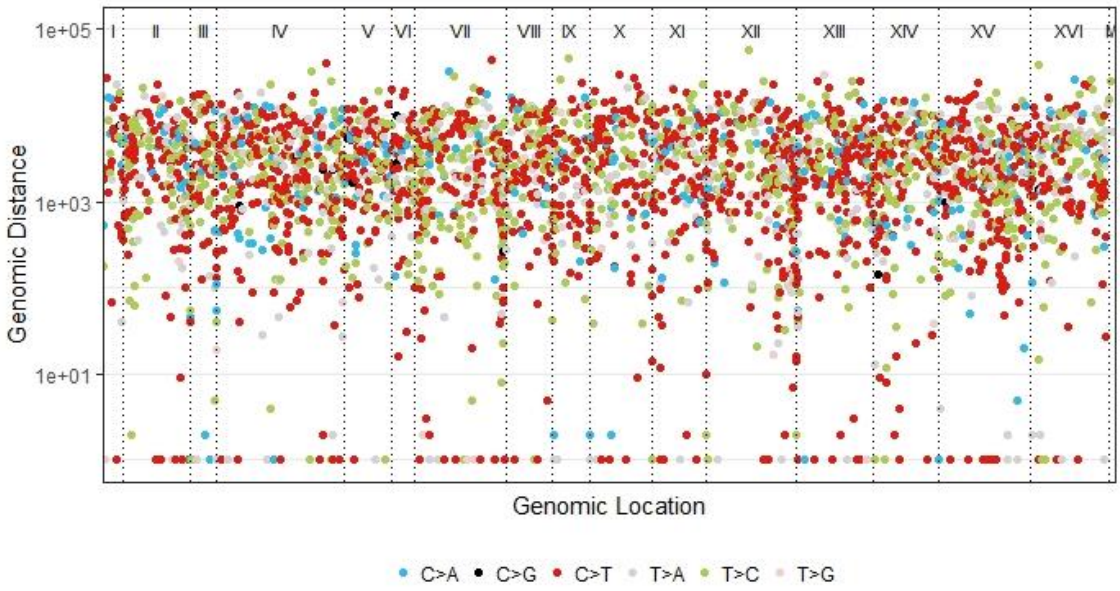


Figure A4.10: Rainfall plot of UV-induced *rad16* mutant showing the genomic location of substitution mutations with inter-mutational distance

Table A4.1: The Pan Cancer Analysis of Whole Genome (PCAWG) mutational signatures and their probable associations with biological processes of DNA damage and repair. This data obtained from (Alexandrov et al. 2018).

Signatures	**Proposed aetiology**
SBS1	Deamination of 5-methylcytosine
SBS2	APOBEC activity
SBS3	Defective HR DNA repair: BRCA1/2 mutation
SBS4	Tobacco smoking
SBS5	Unknown
SBS6	Defective DNA mismatch repair
SBS7a	Ultraviolet light exposure
SBS7b	Ultraviolet light exposure
SBS7c	Ultraviolet light exposure
SBS7d	Ultraviolet light exposure
SBS8	Unknown
SBS9	Polymerase η activity
SBS10a	POLE mutation
SBS10b	POLE mutation
SBS11	Temozolomide treatment

SBS12	Unknown
SBS13	APOBEC activity
SBS14	Concurrent POLE mutation and mismatch repair deficiency
SBS15	Defective DNA mismatch repair
SBS16	Unknown
SBS17a	Unknown
SBS17b	Unknown
SBS18	Reactive oxygen species
SBS19	Unknown
SBS20	Concurrent POLD1 mutation and mismatch repair deficiency
SBS21	Defective DNA mismatch repair
SBS22	Aristolochic acid exposure
SBS23	Unknown
SBS24	Aflatoxin exposure
SBS25	Chemotherapy
SBS26	Defective DNA mismatch repair
SBS27	Unknown
SBS28	Unknown
SBS29	Tobacco chewing
SBS30	Defective base excision repair: NTHL1 mutation
SBS31	Platinum treatment
SBS32	Azathioprine treatment
SBS33	Unknown
SBS34	Unknown
SBS35	Platinum treatment
SBS36	Defective base excision repair: MUTYH mutation
SBS37	Unknown
SBS38	Indirect effect of ultraviolet light
SBS39	Unknown
SBS40	Unknown
SBS41	Unknown
SBS42	Haloalkane exposure
SBS43	Unknown
SBS44	Defective DNA mismatch repair

SBS84	Unknown
SBS85	Unknown

Abbreviations: HR, Homologous Recombination; POBEC, apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like; BRCA 1, Breast cancer type 1(gene); POLE, DNA polymerase epsilon; POLD, DNA polymerase delta, NTHL1, nth like DNA glycosylase 1; MUTYH, mutY DNA glycosylase.

Appendix V

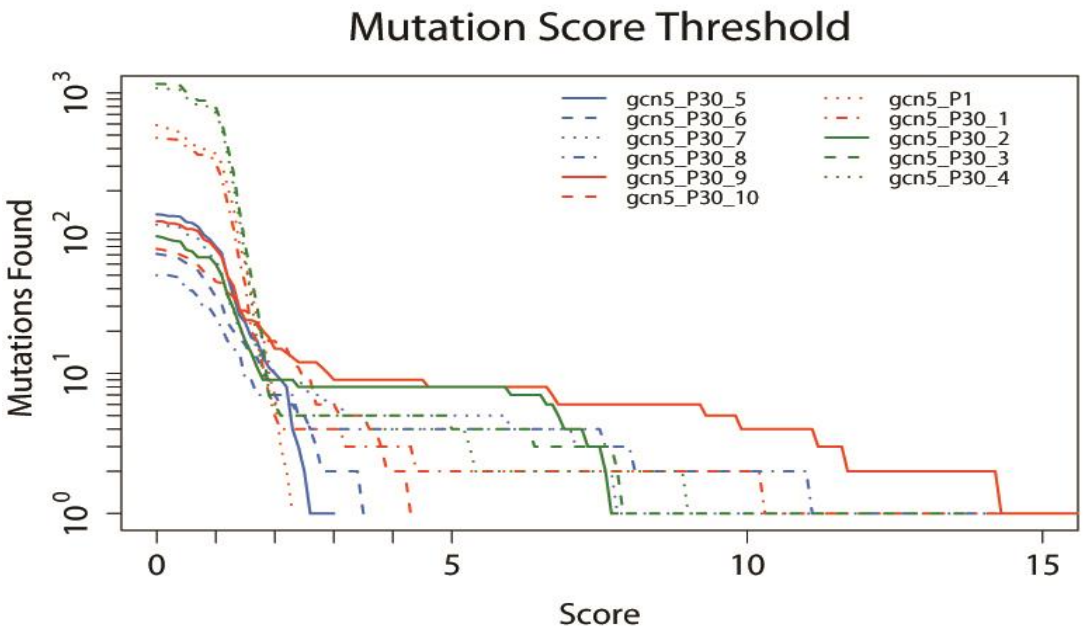


Figure A5.1: Tuning curve for *gcn5* mutant cells substitution filtration

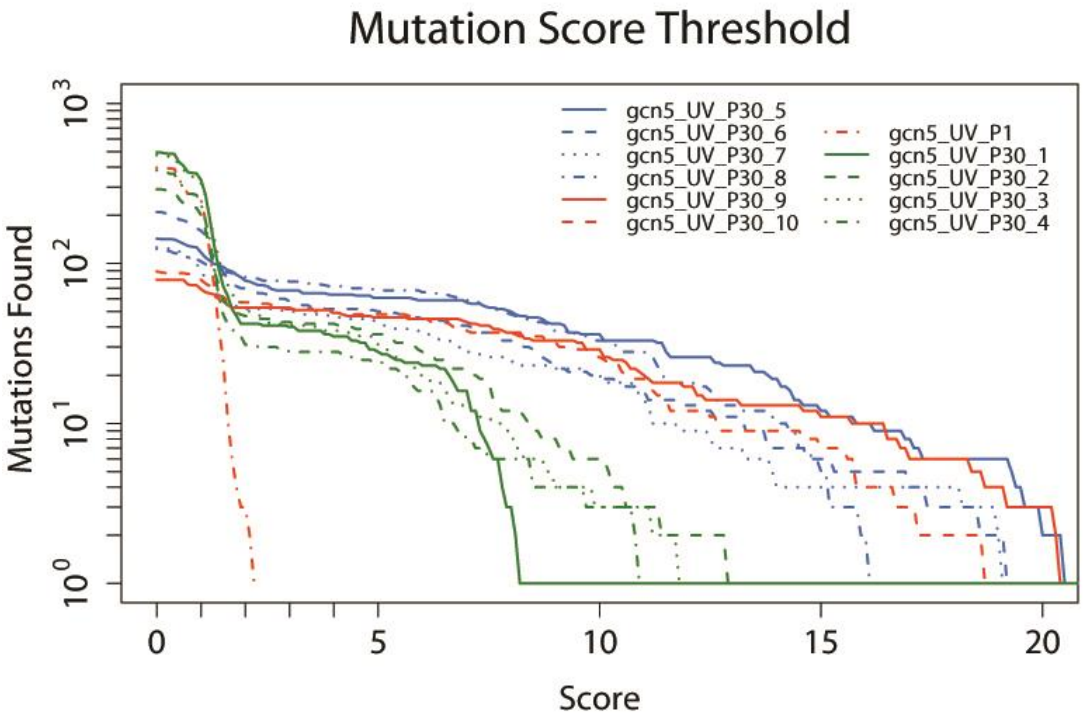


Figure A5.2: Tuning curve for UV-induced *gcn5* mutant cells substitution filtration

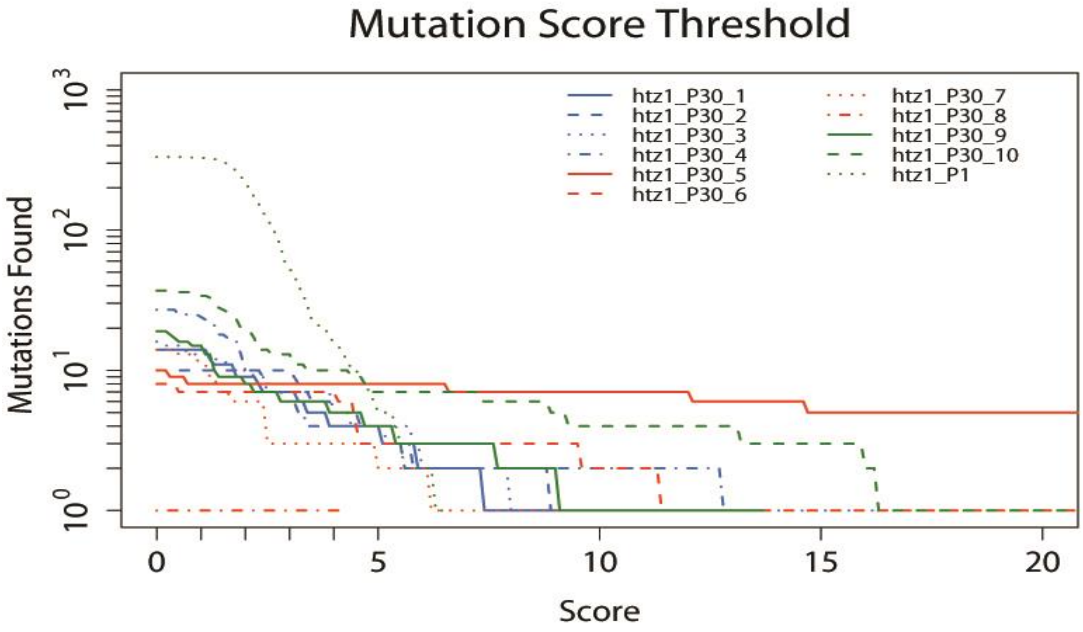


Figure A5.3: Tuning curve for *gcn5* mutant cells substitution filtration

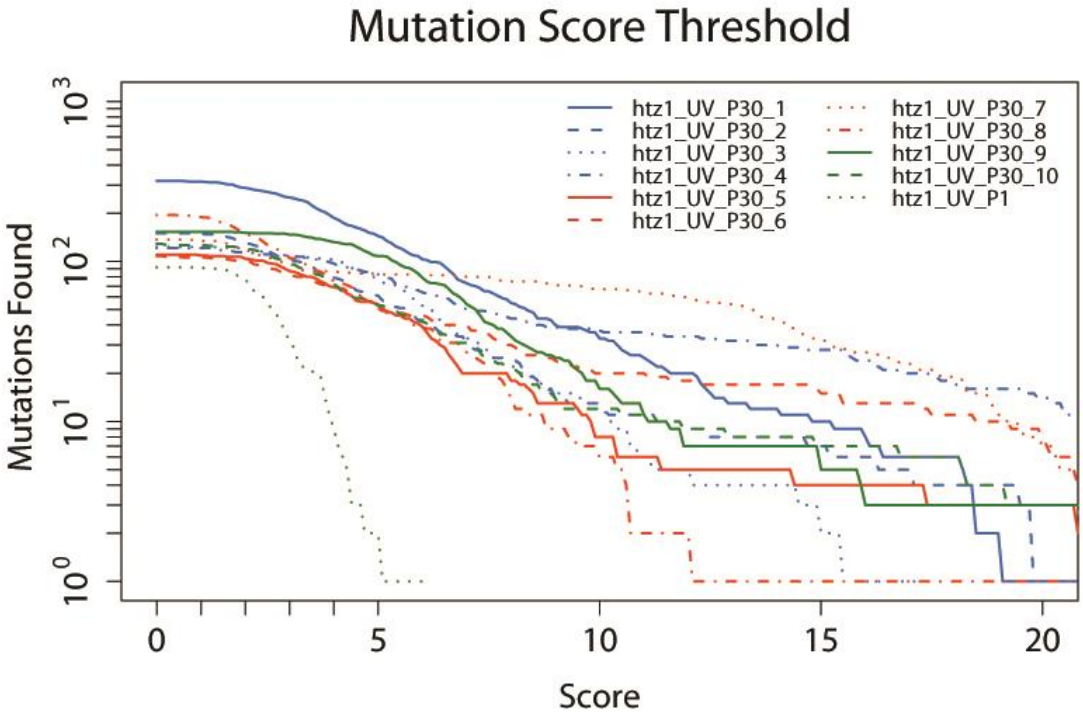


Figure A5.4: Tuning curve for UV-induced *gcn5* mutant cells substitution filtration

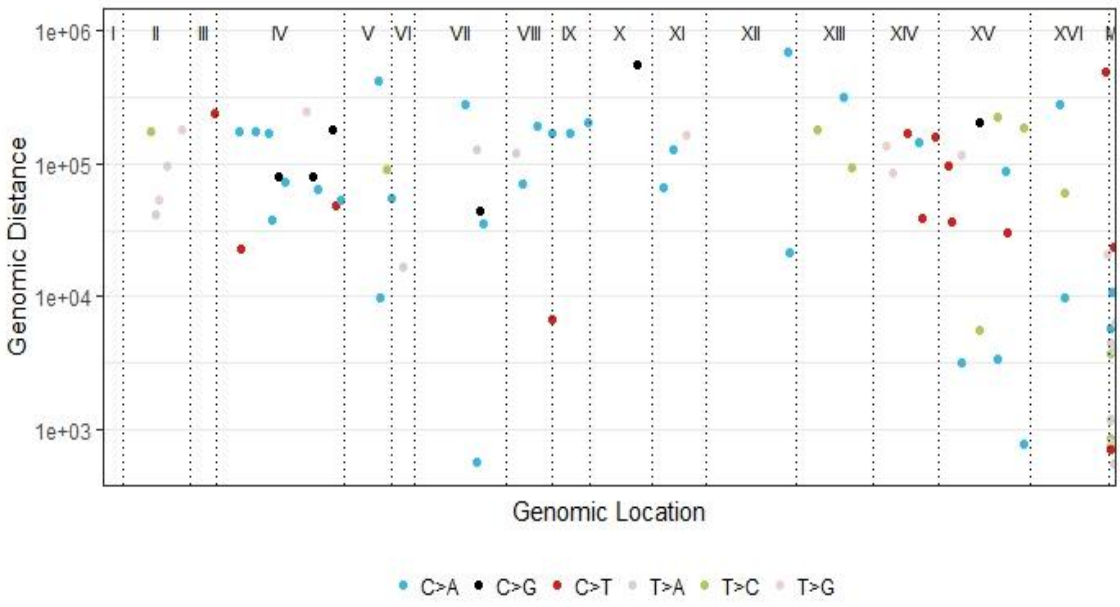


Figure A5.5: Rainfall plot of *gcn5* mutant showing the genomic location of substitution mutations with inter-mutational distance

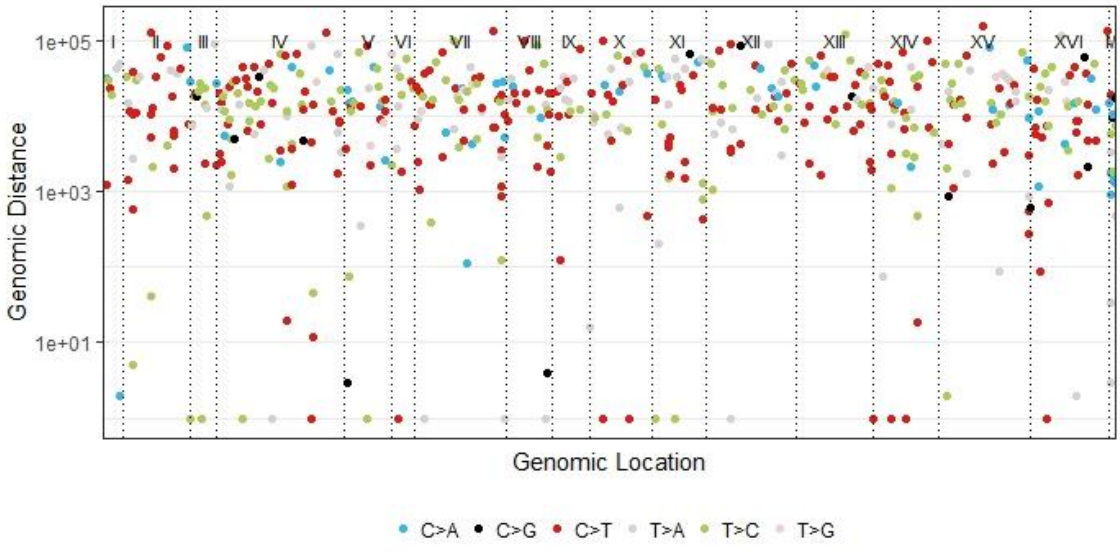


Figure A5.6: Rainfall plot of UV-induced *gcn5* mutant showing the genomic location of substitution mutations with inter-mutational distance

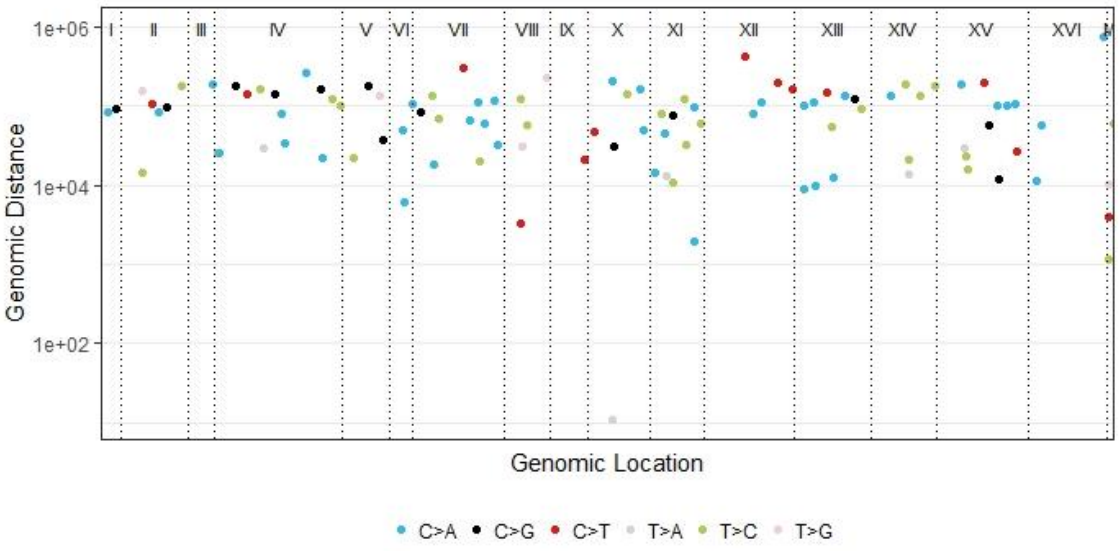


Figure A5.7: Rainfall plot of *htz1* mutant showing the genomic location of substitution mutations with inter-mutational distance

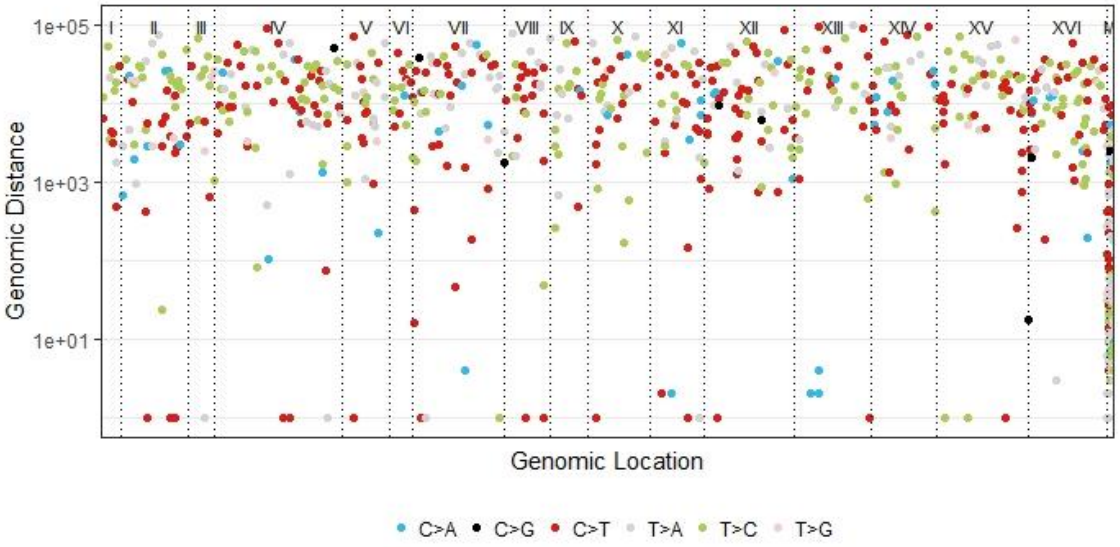


Figure A5.8: Rainfall plot of UV-induced *htz1* mutant showing the genomic location of substitution mutations with inter-mutational distance

References

Adam S, Dabina J, Polo SE. 2015. Chromatin plasticity in response to DNA damage: The shape of things to come. *DNA Repair* **32**: 120-126.

Adar S, Hu JC, Lieb JD, Sancar A. 2016. Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America* **113**: E2124-E2133.

Aida M, Hamad N, Stanlie A, Begum NA, Honjo T. 2013. Accumulation of the FACT complex, as well as histone H3. 3, serves as a target marker for somatic hypermutation. *Proceedings of the National Academy of Sciences*: 201305859.

Akdemir KC, Chin L. 2015. HiCPlotter integrates genomic data with interaction matrices. *Genome biology* **16**: 198.

Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF. 2007. Translational and rotational settings of H2A. Z nucleosomes across the Saccharomycescerevisiae genome. *Nature* **446**: 572.

Alexandrov L. 2014. Signatures of mutational processes in human cancer. *Mutagenesis* **29**: 499-499.

Alexandrov L, Kim J, Haradhvala NJ, Huang MN, Ng AWT, Boot A, Covington KR, Gordenin DA, Bergstrom E, Lopez-Bigas N. 2018. The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv*: 322859.

Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, Stratton MR. 2015. Clock-like mutational processes in human somatic cells. *Nature Genetics* **47**: 1402-+.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L. 2013. Signatures of mutational processes in human cancer. *Nature* **500**: 415-421.

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. 2013b. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports* **3**: 246-259.

Alseth I, Korvald H, Osman F, Seeberg E, Bjørås M. 2004. A general role of the DNA glycosylase Nth1 in the abasic sites cleavage step of base excision repair in Schizosaccharomyces pombe. *Nucleic acids research* **32**: 5119-5125.

Andersen S, Heine T, Sneve R, König I, Krokan HE, Epe B, Nilsen H. 2005. Incorporation of dUMP into DNA is a major source of spontaneous DNA damage, while excision of uracil is not required for cytotoxicity of fluoropyrimidines in mouse embryonic fibroblasts. *Carcinogenesis* **26**: 547-555.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in bipolymers.

Bailey TL, Williams N, Misleh C, Li WW. 2006. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic acids research* **34**: W369-W373.

Bastien N, Becerril C, Bespalov VA, Boiteux S, Boszko IP, Brash DE, Cadet J, Caney C, Cheung Jr KJ, Conconi A. 2007. *From DNA photolesions to mutations, skin cancer and cell death*. Royal Society of Chemistry.

Basu AK, Loechler EL, Leadon SA, Essigmann JM. 1989. Genetic effects of thymine glycol: site-specific mutagenesis and molecular modeling studies. *Proceedings of the National Academy of Sciences* **86**: 7677-7681.

Behjati S, Gundem G, Wedge DC, Roberts ND, Tarpey PS, Cooke SL, Van Loo P, Alexandrov LB, Ramakrishna M, Davies H et al. 2016. Mutational signatures of ionizing radiation in second malignancies. *Nat Commun* **7**: 12605.

Bellon S, Shikazono N, Cunniffe S, Lomax M, O'Neill P. 2009. Processing of thymine glycol in a clustered DNA damage site: mutagenic or cytotoxic. *Nucleic acids research* **37**: 4430-4440.

Bennett M, Evans KE, Yu SR, Teng YM, Webster RM, Powell J, Waters R, Reed SH. 2015. Sandcastle: software for revealing latent information in multiple experimental ChIP-chip datasets via a novel normalisation procedure. *Scientific Reports* **5**.

Bhat R, Rai RV, Karim AA. 2010. Mycotoxins in food and feed: present status and future concerns. *Comprehensive reviews in food science and food safety* **9**: 57-81.

Blokzijl F, Janssen R, van Boxtel R, Cuppen E. 2018. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med* **10**: 33.

Botstein D, Chervitz SA, Cherry M. 1997. Yeast as a model organism. *Science* **277**: 1259-1260.

Botstein D, Fink GR. 2011. Yeast: an experimental organism for 21st Century biology. *Genetics* **189**: 695-704.

Brogaard K, Xi L, Wang J-P, Widom J. 2012. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**: 496.

Broustas CG, Lieberman HB. 2014. DNA damage response genes and the development of cancer metastasis. *Radiation research* **181**: 111-130.

Brunet JP, Tamayo P, Golub TR, Mesirov JP. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 4164-4169.

Brüsehafer K, Manshian BB, Doherty AT, Zaïr ZM, Johnson GE, Doak SH, Jenkins GJ. 2016. The clastogenicity of 4NQO is cell-type dependent and linked to cytotoxicity, length of exposure and p53 proficiency. *Mutagenesis* **31**: 171-180.

Burger A, Fix D, Liu H, Hays J, Bockrath R. 2003. In vivo deamination of cytosine-containing cyclobutane pyrimidine dimers in E. coli: a feasible part of UV-mutagenesis. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **522**: 145-156.

Caldecott KW. 2008. Single-strand break repair and genetic disease. *Nature Reviews Genetics* **9**: 619.

Cazzanelli G, Pereira F, Alves S, Francisco R, Azevedo L, Dias Carvalho P, Almeida A, Côrte-Real M, Oliveira M, Lucas C. 2018. The Yeast Saccharomyces cerevisiae as a Model for Understanding RAS Proteins and Their Role in Human Tumorigenesis. *Cells* **7**: 14.

Ceccaldi R, Rondinelli B, D'Andrea AD. 2016. Repair Pathway Choices and Consequences at the Double-Strand Break. *Trends Cell Biol* **26**: 52-64.

Chan K, Resnick MA, Gordenin DA. 2013. The choice of nucleotide inserted opposite abasic sites formed within chromosomal DNA reveals the polymerase activities participating in translesion DNA synthesis. *DNA repair* **12**: 878-889.

Charrad M, Ghazzali N, Boiteau V, Niknafs A. 2012. NbClust Package: finding the relevant number of clusters in a dataset. *UseR! 2012*.

Chatterjee N, Lin Y, Yotnda P, Wilson JH. 2016. Environmental stress induces trinucleotide repeat mutagenesis in human cells by Alt-nonhomologous end joining repair. *Journal of molecular biology* **428**: 2978-2980.

Chatterjee N, Santillan BA, Wilson JH. 2013. Microsatellite repeats: canaries in the coalmine. In *Stress-Induced Mutagenesis*, pp. 119-150. Springer.

Chatterjee N, Walker GC. 2017. Mechanisms of DNA damage, repair, and mutagenesis. *Environmental and molecular mutagenesis*.

Cheadle JP, Dolwani S, Sampson JR. 2003. Inherited defects in the DNA glycosylase MYH cause multiple colorectal adenoma and carcinoma. *Carcinogenesis* **24**: 1281-1282.

Cheadle JP, Sampson JR. 2003. Exposing the MYtH about base excision repair and human inherited disease. *Human molecular genetics* **12**: R159-R165.

Chen KF, Xi YX, Pan XW, Li ZY, Kaestner K, Tyler J, Dent S, He XW, Li W. 2013. DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Research* **23**: 341-351.

Cheng KC, Cahill DS, Kasai H, Nishimura S, Loeb LA. 1992. 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G----T and A----C substitutions. *Journal of Biological Chemistry* **267**: 166-172.

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M. 1998. SGD: Saccharomyces genome database. *Nucleic acids research* **26**: 73-79.

Ciccia A, Elledge SJ. 2010. The DNA damage response: making it safe to play with knives. *Molecular cell* **40**: 179-204.

Clausen AR, Zhang S, Burgers PM, Lee MY, Kunkel TA. 2013. Ribonucleotide incorporation, proofreading and bypass by human DNA polymerase δ. *DNA repair* **12**: 121-127.

Cleaver JE, Brennan-Minnella AM, Swanson RA, Fong K-w, Chen J, Chou K-m, Chen Y-w, Revet I, Bezrookove V. 2014. Mitochondrial reactive oxygen species are scavenged by Cockayne syndrome B protein in human fibroblasts without nuclear DNA damage. *Proceedings of the National Academy of Sciences* **111**: 13487-13492.

Cogliano VJ, Baan R, Straif K, Grosse Y, Lauby-Secretan B, El Ghissassi F, Bouvard V, Benbrahim-Tallaa L, Guha N, Freeman C. 2011. Preventable exposures associated with human cancers. *Journal of the National Cancer Institute* **103**: 1827-1839.

Cooke MS, Evans MD, Dizdaroglu M, Lunec J. 2003. Oxidative DNA damage: mechanisms, mutation, and disease. *The FASEB Journal* **17**: 1195-1214.

Cuozzo C, Porcellini A, Angrisano T, Morano A, Lee B, Di Pardo A, Messina S, Iuliano R, Fusco A, Santillo MR. 2007. DNA damage, homology-directed repair, and DNA methylation. *PLoS genetics* **3**: e110.

Curtin NJ. 2012. DNA repair dysregulation from cancer driver to therapeutic target. *Nature Reviews Cancer* **12**: 801.

de Boer J, Hoeijmakers JHJ. 2000. Nucleotide excision repair and human syndromes. *Carcinogenesis* **21**: 453-460.

De Bont R, van Larebeke N. 2004. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19**: 169-185.

de Laat WL, Jaspers NGJ, Hoeijmakers JHJ. 1999. Molecular mechanism of nucleotide excision repair. *Genes & development* **13**: 768-785.

Demple B, DeMott MS. 2002. Dynamics and diversions in base excision DNA repair of oxidized abasic lesions. *Oncogene* **21**: 8926.

Dianov GL, Hübscher U. 2013. Mammalian base excision repair: the forgotten archangel. *Nucleic acids research* **41**: 3483-3490.

Ding L, Wendl MC, Koboldt DC, Mardis ER. 2010. Analysis of next-generation genomic data in cancer: accomplishments and challenges. *Human molecular genetics* **19**: R188-R196.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376.

Donley N, Thayer MJ. 2013. DNA replication timing, genome stability and cancer: late and/or delayed DNA replication timing is associated with increased genomic instability. *Semin Cancer Biol* **23**: 80-89.

Dosanjh MK, Galeros G, Goodman MF, Singer B. 1991. Kinetics of extension of O6-methylguanine paired with cytosine or thymine in defined oligonucleotide sequences. *Biochemistry* **30**: 11595-11599.

Douki T, Cadet J. 2001. Individual determination of the yield of the main UV-induced dimeric pyrimidine photoproducts in DNA suggests a high mutagenicity of CC photolesions. *Biochemistry* **40**: 2495-2501.

Duncan BK, Miller JH. 1980. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**: 560.

Duncan T, Trewick SC, Koivisto P, Bates PA, Lindahl T, Sedgwick B. 2002. Reversal of DNA alkylation damage by two human dioxygenases. *Proceedings of the National Academy of Sciences* **99**: 16660-16665.

Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, Dwight SS, Hitz BC, Karra K, Nash RS. 2013. The reference genome sequence of Saccharomyces cerevisiae: then and now. *G3: Genes, Genomes, Genetics*: g3-113.

Fernández JR, Byrne B, Firestein BL. 2009. Phylogenetic analysis and molecular evolution of guanine deaminases: from guanine to dendrites. *Journal of molecular evolution* **68**: 227-235.

Flaus A, Owen-Hughes T. 2011. Mechanisms for ATP-dependent chromatin remodelling: the means to the end. *The FEBS journal* **278**: 3579-3595.

Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, Suvà ML, Bernstein BE. 2016. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**: 110.

Foury F. 1997. Human genetic diseases: a cross-talk between man and yeast. *Gene* **195**: 1-10.

Fousteri M, Mullenders LH. 2008. Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell research* **18**: 73-84.

Freeman SE, Ryan SL. 1990. Wavelength dependence for UV-induced pyrimidine dimer formation in DNA of human peripheral blood lymphocytes. *Mutation Research/DNA Repair* **235**: 181-186.

Freese NH, Norris DC, Loraine AE. 2016. Integrated genome browser: visual analytics platform for genomics. *Bioinformatics* **32**: 2089-2095.

Friedberg EC. 2001. Nucleotide excision repair in eukaryotes. *e LS*.

Friedberg EC. 2003. DNA damage and repair. *Nature* **421**: 436-440.

Friedberg EC, Bardwell AJ, Bardwell L, Feaver WJ, Kornberg RD, Svejstrup JQ, Tomkinson AE, Wang Z. 1995. Nucleotide excision repair in the yeast Saccharomyces cerevisiae: its relationship to specialized mitotic recombination and RNA polymerase II basal transcription. *Philos Trans R Soc Lond B Biol Sci* **347**: 63-68.

Friedberg EC, Walker GC, Siede W, Wood RD. 2005. *DNA repair and mutagenesis*. American Society for Microbiology Press.

Frigyesi A, Höglund M. 2008. Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes. *Cancer informatics* **6**: CIN-S606.

Ganapathi M, Palumbo MJ, Ansari SA, He Q, Tsui K, Nislow C, Morse RH. 2010. Extensive role of the general regulatory factors, Abf1 and Rap1, in determining genome-wide chromatin structure in budding yeast. *Nucleic acids research* **39**: 2032-2044.

Garaycoechea JI, Crossan GP, Langevin F, Mulderrig L, Louzada S, Yang F, Guilbaud G, Park N, Roerink S, Nik-Zainal S. 2018. Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. *Nature* **553**: 171.

Gaujoux R, Seoighe C. 2010. A flexible R package for nonnegative matrix factorization. *BMC bioinformatics* **11**: 367.

Gehring JS, Fischer B, Lawrence M, Huber W. 2015. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**: 3673-3675.

Ghosh-Roy S, Das D, Chowdhury D, Smerdon MJ, Chaudhuri RN. 2013. Rad26, the transcription-coupled repair factor in yeast, is required for removal of stalled RNA polymerase-II following UV irradiation. *PloS one* **8**: e72090.

Gillette TG, Yu S, Zhou Z, Waters R, Johnston SA, Reed SH. 2006. Distinct functions of the ubiquitin–proteasome pathway influence nucleotide excision repair. *The EMBO journal* **25**: 2529-2538.

Goodman MF. 2002. Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Annual review of biochemistry* **71**: 17-50.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017-1018.

Groth A, Rocha W, Verreault A, Almouzni G. 2007. Chromatin challenges during DNA replication and repair. *Cell* **128**: 721-733.

Guaragnella N, Palermo V, Galli A, Moro L, Mazzoni C, Giannattasio S. 2014. The expanding role of yeast in cancer research and diagnosis: insights into the function of the oncosuppressors p53 and BRCA1/2. *FEMS yeast research* **14**: 2-16.

Guillemette B, Bataille AR, Gévry N, Adam M, Blanchette M, Robert F, Gaudreau L. 2005. Variant histone H2A. Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. *PLoS biology* **3**: e384.

Guo YA, Chang MM, Huang W, Ooi WF, Xing M, Tan P, Skanderup AJ. 2018. Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nature communications* **9**: 1520.

Gurard-Levin ZA, Quivy J-P, Almouzni G. 2014. Histone chaperones: assisting histone traffic and nucleosome dynamics. *Annual review of biochemistry* **83**: 487-517.

Guzder SN, Habraken Y, Sung P, Prakash L, Prakash S. 1996. RAD26, the yeast homolog of human Cockayne's syndrome group B gene, encodes a DNA-dependent ATPase. *Journal of Biological Chemistry* **271**: 18314-18317.

Halliwell B, Gutteridge JMC. 2015. *Free radicals in biology and medicine*. Oxford University Press, USA.

Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. *cell* **144**: 646-674.

Hanawalt PC, Spivak G. 2008. Transcription-coupled DNA repair: two decades of progress and surprises. *Nature reviews Molecular cell biology* **9**: 958.

Hara R, Mo J, Sancar A. 2000. DNA damage in the nucleosome core is refractory to repair by human excision nuclease. *Molecular and cellular biology* **20**: 9173-9181.

Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, Rheinbay E, Kim J, Maruvka YE, Braunstein LZ. 2016. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**: 538-549.

Harper JW, Elledge SJ. 2007. The DNA damage response: ten years after. *Molecular cell* **28**: 739-745.

Hartley PD, Madhani HD. 2009. Mechanisms that specify promoter nucleosome location and identity. *Cell* **137**: 445-458.

Hauer MH, Seeber A, Singh V, Thierry R, Sack R, Amitai A, Kryzhanovska M, Eglinger J, Holcman D, Owen-Hughes T. 2017. Histone degradation in response to DNA damage enhances chromatin dynamics and recombination rates. *Nature Structural and Molecular Biology* **24**: 99.

Hecht SS. 1999. Tobacco smoke carcinogens and lung cancer. *JNCI: Journal of the National Cancer Institute* **91**: 1194-1210.
Hegde ML, Hazra TK, Mitra S. 2008. Early steps in the DNA base excision/single-strand interruption repair pathway in mammalian cells. *Cell research* **18**: 27.
Helleday T, Eshtad S, Nik-Zainal S. 2014. Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics* **15**: 585-598.
Henikoff S, Smith MM. 2015. Histone variants and epigenetics. *Cold Spring Harbor perspectives in biology* **7**: a019364.
Herrmann SS, Granby K, Duedahl-Olesen L. 2015. Formation and mitigation of N-nitrosamines in nitrite preserved cooked sausages. *Food chemistry* **174**: 516-526.
Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, Goldmann J, Lajoie BR, Fan ZP, Sigova AA. 2016. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**: 1454-1458.
Hoang ML, Chen C-H, Sidorenko VS, He J, Dickman KG, Yun BH, Moriya M, Niknafs N, Douville C, Karchin R. 2013. Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Science translational medicine* **5**: 197ra102-197ra102.
Hodges AJ, Plummer DA, Wyrick JJ. 2018. NuA4 acetyltransferase is required for efficient nucleotide excision repair in yeast. *DNA repair*.
Hodgkinson A, Chen Y, Eyre-Walker A. 2012. The large-scale distribution of somatic mutations in cancer genomes. *Human mutation* **33**: 136-143.
Hoeijmakers JH. 1993. Nucleotide excision repair I: from E. coli to yeast. *Trends Genet* **9**: 173-177.
Hoeijmakers JH. 2001. Genome maintenance mechanisms for preventing cancer. *Nature* **411**: 366-374.
Hoeijmakers JHJ. 1994. Human nucleotide excision repair syndromes: molecular clues to unexpected intricacies. *European Journal of Cancer* **30**: 1912-1921.
Hoeijmakers JHJ. 2009. DNA damage, aging, and cancer. *New England Journal of Medicine* **361**: 1475-1485.
Holliday R, Grigg GW. 1993. DNA methylation and mutation. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **285**: 61-67.
Holmquist GP, Gao S. 1997. Somatic mutation theory, DNA repair rates, and the molecular epidemiology of p53 mutations. *Mutat Res* **386**: 69-101.
Hsieh P, Yamane K. 2008. DNA mismatch repair: molecular mechanism, cancer, and ageing. *Mechanisms of ageing and development* **129**: 391-407.
Hsieh T-HS, Fudenberg G, Goloborodko A, Rando OJ. 2016. Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome. *Nature methods* **13**: 1009.
Hsieh T-HS, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ. 2015. Mapping nucleosome resolution chromosome folding in yeast by micro-C. *Cell* **162**: 108-119.
Hu J, Adebali O, Adar S, Sancar A. 2017. Dynamic maps of UV damage formation and repair for the human genome. *Proc Natl Acad Sci U S A* **114**: 6758-6763.
Hu JC, Adar S, Selby CP, Lieb JD, Sancar A. 2015. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes & Development* **29**: 948-960.
Huang MN, Yu W, Teoh WW, Ardin M, Jusakul A, Ng AWT, Boot A, Abedi-Ardekani B, Villar S, Myint SS et al. 2017. Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Res* **27**: 1475-1486.
Hutchins LN, Murphy SM, Singh P, Graber JH. 2008. Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics* **24**: 2684-2690.

Ikehata H. 2018. Mechanistic considerations on the wavelength-dependent variations of UVR genotoxicity and mutagenesis in skin: the discrimination of UVA-signature from UV-signature mutation. *Photochemical & Photobiological Sciences*.

Ikehata H, Ono T. 2011. The mechanisms of UV mutagenesis. *Journal of radiation research* **52**: 115-125.

Jansen A, Verstrepen KJ. 2011. Nucleosome positioning in Saccharomyces cerevisiae. *Microbiology and Molecular Biology Reviews* **75**: 301-320.

Jansen LET, Verhage RA, Brouwer J. 1998. Preferential binding of yeast Rad4· Rad23 complex to damaged DNA. *Journal of Biological Chemistry* **273**: 33111-33114.

Jha V, Bian C, Xing G, Ling H. 2016. Structure and mechanism of error-free replication past the major benzo [a] pyrene adduct by human DNA polymerase κ. *Nucleic acids research* **44**: 4957-4967.

Jia R, Chai P, Zhang H, Fan X. 2017. Novel insights into chromosomal conformations in cancer. *Molecular cancer* **16**: 173.

Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews Genetics* **10**: 161.

Johnson RE, Washington MT, Prakash S, Prakash L. 2000. Fidelity of human DNA polymerase eta. *J Biol Chem* **275**: 7447-7450.

Jones S, Wang T-L, Shih I-M, Mao T-L, Nakayama K, Roden R, Glas R, Slamon D, Diaz LA, Vogelstein B. 2010. Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science*: 1196333.

Kaiser VB, Semple CA. 2018. Chromatin loop anchors are associated with genome instability in cancer and recombination hotspots in the germline. *Genome biology* **19**: 101.

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA. 2013a. Mutational landscape and significance across 12 major cancer types. *Nature* **502**: 333.

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA et al. 2013b. Mutational landscape and significance across 12 major cancer types. *Nature* **502**: 333-339.

Kantidze OL, Velichko AK, Luzhin AV, Razin SV. 2016. Heat stress-induced DNA damage. *Acta Naturae (англоязычная версия)* **8**.

Kasinathan S, Orsi GA, Zentner GE, Ahmad K, Henikoff S. 2014. High-resolution mapping of transcription factor binding sites on native chromatin. *Nature methods* **11**: 203.

Kent NA, Adams S, Moorhouse A, Paszkiewicz K. 2011. Chromatin particle spectrum analysis: a method for comparative chromatin structure analysis using paired-end mode next-generation DNA sequencing. *Nucleic Acids Res* **39**: e26.

Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, Bessy A, Cheneby J, Kulkarni SR, Tan G. 2017. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic acids research* **46**: D260-D266.

Khanna KK, Jackson SP. 2001. DNA double-strand breaks: signaling, repair and the cancer connection. *Nature genetics* **27**: 247.

Kim J, Mouw KW, Polak P, Braunstein LZ, Kamburov A, Tiao G, Kwiatkowski DJ, Rosenberg JE, Van Allen EM, D D'Andrea A. 2016. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature genetics* **48**: 600.

Kim N, Jinks-Robertson S. 2012. Transcription as a source of genome instability. *Nature Reviews Genetics* **13**: 204.

Knobel PA, Marti TM. 2011. Translesion DNA synthesis in the context of cancer research. *Cancer cell international* **11**: 39.

Kong M, Liu L, Chen X, Driscoll KI, Mao P, Böhm S, Kad NM, Watkins SC, Bernstein KA, Wyrick JJ. 2016. Single-molecule imaging reveals that Rad4 employs a dynamic DNA damage recognition process. *Molecular cell* **64**: 376-387.

Kornberg RD, Lorch Y. 1999. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**: 285-294.

Krokan HE, Bjørås M. 2013. Base excision repair. *Cold Spring Harbor perspectives in biology* **5**: a012583.

Kulis M, Esteller M. 2010. DNA methylation and cancer. In *Advances in genetics*, Vol 70, pp. 27-56. Elsevier.

Kuluncsics Z, Perdiz D, Brulay E, Muel B, Sage E. 1999. Wavelength dependence of ultraviolet-induced DNA damage distribution: involvement of direct or indirect mechanisms and possible artefacts. *Journal of Photochemistry and Photobiology B: Biology* **49**: 71-80.

Kumar D, Abdulovic AL, Viberg J, Nilsson AK, Kunkel TA, Chabes A. 2010. Mechanisms of mutagenesis in vivo due to imbalanced dNTP pools. *Nucleic acids research* **39**: 1360-1371.

Kunkel TA. 2004. DNA replication fidelity. *Journal of Biological Chemistry* **279**: 16895-16898.

Kunkel TA, Erie DA. 2005. DNA mismatch repair. *Annu Rev Biochem* **74**: 681-710.

Kuzminov A. 2001. Single-strand interruptions in replicating chromosomes cause double-strand breaks. *Proceedings of the National Academy of Sciences* **98**: 8241-8246.

Lai WKM, Pugh BF. 2017. Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nature Reviews Molecular Cell Biology* **18**: 548.

Lang GI, Murray AW. 2011. Mutation rates across budding yeast chromosome VI are correlated with replication timing. *Genome biology and evolution* **3**: 799-811.

Lange SS, Takata K-i, Wood RD. 2011. DNA polymerases and cancer. *Nature reviews cancer* **11**: 96-110.

Lans H, Marteijn JA, Vermeulen W. 2012. ATP-dependent chromatin remodeling in the DNA-damage response. *Epigenetics & chromatin* **5**: 4.

Larrea AA, Lujan SA, McElhinny SAN, Mieczkowski PA, Resnick MA, Gordenin DA, Kunkel TA. 2010. Genome-wide model for the normal eukaryotic DNA replication fork. *Proceedings of the National Academy of Sciences*: 201010178.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**: 214.

Lazzaro F, Giannattasio M, Puddu F, Granata M, Pellicioli A, Plevani P, Muzi-Falconi M. 2009. Checkpoint mechanisms at the intersection between DNA damage and repair. *DNA Repair (Amst)* **8**: 1055-1067.

Lee DD, Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**: 788-791.

Lee S-K, Yu S-L, Prakash L, Prakash S. 2001. Requirement for Yeast RAD26, a Homolog of the HumanCSB Gene, in Elongation by RNA Polymerase II. *Molecular and cellular biology* **21**: 8651-8656.

Letouzé E, Shinde J, Renault V, Couchy G, Blanc J-F, Tubacher E, Bayard Q, Bacq D, Meyer V, Semhoun J. 2017. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nature Communications* **8**: 1315.

Li G, Reinberg D. 2011. Chromatin higher-order structures and gene regulation. *Current opinion in genetics & development* **21**: 175-186.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**: 589-595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Lindahl T. 1979. DNA glycosylases, endonucleases for apurinic/apyrimidinic sites, and base excision-repair. In *Progress in nucleic acid research and molecular biology*, Vol 22, pp. 135-192. Elsevier.

Lindahl T. 1993. Instability and decay of the primary structure of DNA. *nature* **362**: 709.

Lindahl T, Wood RD. 1999. Quality control by DNA repair. *Science* **286**: 1897-1905.

Liu X. 2015. In vitro chromatin templates to study nucleotide excision repair. *DNA Repair (Amst)* **36**: 68-76.

Lord CJ, Ashworth A. 2012a. The DNA damage response and cancer therapy. *Nature* **481**: 287.

Lord CJ, Ashworth A. 2012b. The DNA damage response and cancer therapy. *Nature* **481**: 287-294.

Loveless A. 1969. Possible relevance of O–6 alkylation of deoxyguanosine to the mutagenicity and carcinogenicity of nitrosamines and nitrosamides. *Nature* **223**: 206.

Lujan SA, Clausen AR, Clark AB, MacAlpine HK, MacAlpine DM, Malc EP, Mieczkowski PA, Burkholder AB, Fargo DC, Gordenin DA. 2014. Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. *Genome research*: gr-178335.

Lunning MA, Green MR. 2015. Mutation of chromatin modifiers; an emerging hallmark of germinal center B-cell lymphomas. *Blood cancer journal* **5**: e361.

Maharjan R, Ferenci T. 2014. Mutational signatures indicative of environmental stress in bacteria. *Molecular biology and evolution* **32**: 380-391.

Mao P, Brown AJ, Malc EP, Mieczkowski PA, Smerdon MJ, Roberts SA, Wyrick JJ. 2017a. Genome-wide maps of alkylation damage, repair, and mutagenesis in yeast reveal mechanisms of mutational heterogeneity. *Genome Res*.

Mao P, Smerdon MJ, Roberts SA, Wyrick JJ. 2016. Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. *Proceedings of the National Academy of Sciences* **113**: 9057-9062.

Mao P, Wyrick JJ, Roberts SA, Smerdon MJ. 2017b. UV-Induced DNA Damage and Mutagenesis in Chromatin. *Photochem Photobiol* **93**: 216-228.

Marchal C, Sasaki T, Vera D, Wilson K, Sima J, Rivera-Mulia JC, Trevilla-Garcia C, Nogues C, Nafie E, Gilbert DM. 2018. Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. *Nat Protoc* **13**: 819-839.

Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JHJ. 2014. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nature reviews Molecular cell biology* **15**: 465.

Masutani C, Sugasawa K, Yanagisawa J, Sonoyama T, Ui M, Enomoto T, Takio K, Tanaka K, Van der Spek PJ, Bootsma D. 1994. Purification and cloning of a nucleotide excision repair complex involving the xeroderma pigmentosum group C protein and a human homologue of yeast RAD23. *The EMBO journal* **13**: 1831-1843.

Mayne LV, Lehmann AR. 1982. Failure of RNA synthesis to recover after UV irradiation: an early defect in cells from individuals with Cockayne's syndrome and xeroderma pigmentosum. *Cancer research* **42**: 1473-1478.

McCormick JP, Pachlatko JP, Eisenstark A. 1976. Characterization of a cell-lethal product from the photooxidation of tryptophan: hydrogen peroxide. *Science* **191**: 468-469.

McMurray CT. 2010. Mechanisms of trinucleotide repeat instability during human development. *Nature Reviews Genetics* **11**: 786.

Meas R, Wyrick JJ, Smerdon MJ. 2017. Nucleosomes regulate base excision repair in chromatin. *Mutation Research/Reviews in Mutation Research*.

Mehta A, Haber JE. 2014. Sources of DNA double-strand breaks and models of recombinational DNA repair. *Cold Spring Harbor perspectives in biology* **6**: a016428.

Meier B, Cooke SL, Weiss J, Bailly AP, Alexandrov LB, Marshall J, Raine K, Maddison M, Anderson E, Stratton MR. 2014. C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome research*: gr-175547.

Meier B, Volkova NV, Hong Y, Schofield P, Campbell PJ, Gerstung M, Gartner A. 2018. Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. *Genome research* **28**: 666-675.

Melis JP, van Steeg H, Luijten M. 2013. Oxidative DNA damage and nucleotide excision repair. *Antioxid Redox Signal* **18**: 2409-2419.

Michaels ML, Cruz C, Grollman AP, Miller JH. 1992. Evidence that MutY and MutM combine to prevent mutations by an oxidatively damaged form of guanine in DNA. *Proceedings of the National Academy of Sciences* **89**: 7022-7025.

Millot GA, Carvalho MA, Caputo SM, Vreeswijk MPG, Brown MA, Webb M, Rouleau E, Neuhausen SL, Hansen TvO, Galli A. 2012. A guide for functional analysis of BRCA1 variants of uncertain significance. *Human mutation* **33**: 1526-1537.

Min JH, Pavletich NP. 2007. Recognition of DNA damage by the Rad4 nucleotide excision repair protein. *Nature* **449**: 570-575.

Mishina Y, Duguid EM, He C. 2006. Direct reversal of DNA alkylation damage. *Chemical reviews* **106**: 215-232.

Mitchell DL, Jen J, Cleaver JE. 1992. Sequence specificity of cyclobutane pyrimidine dimers in DNA treated with solar (ultraviolet B) radiation. *Nucleic acids research* **20**: 225-229.

Moore LD, Le T, Fan G. 2013. DNA methylation and its basic function. *Neuropsychopharmacology* **38**: 23.

Morgan HD, Dean W, Coker HA, Reik W, Petersen-Mahrt SK. 2004. Activation-induced cytidine deaminase deaminates 5-Methylcytosine in DNA and is expressed in pluripotent tissues implications for epigenetic reprogramming. *Journal of Biological Chemistry* **279**: 52353-52360.

Mouret S, Baudouin C, Charveron M, Favier A, Cadet J, Douki T. 2006. Cyclobutane pyrimidine dimers are predominant DNA lesions in whole human skin exposed to UVA radiation. *Proceedings of the National Academy of Sciences* **103**: 13765-13770.

Muller CA, Hawkins M, Retkute R, Malla S, Wilson R, Blythe MJ, Nakato R, Komata M, Shirahige K, de Moura AP et al. 2014. The dynamics of genome replication using deep sequencing. *Nucleic Acids Res* **42**: e3.

Nabel CS, Manning SA, Kohli RM. 2011. The curious chemical biology of cytosine: deamination, methylation, and oxidation as modulators of genomic potential. *ACS chemical biology* **7**: 20-30.

Nag R, Smerdon MJ. 2009. Altering the chromatin landscape for nucleotide excision repair. *Mutation Research/Reviews in Mutation Research* **682**: 13-20.

Nair N, Shoaib M, Sørensen CS. 2017. Chromatin dynamics in genome stability: roles in suppressing endogenous DNA damage and facilitating DNA repair. *International journal of molecular sciences* **18**: 1486.

Nakamura J, Swenberg JA. 1999. Endogenous apurinic/apyrimidinic sites in genomic DNA of mammalian tissues. *Cancer research* **59**: 2522-2526.

Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC. 2010. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature genetics* **42**: 790.

Nik-Zainal S. 2014. Insights into cancer biology through next-generation sequencing. *Clin Med (Lond)* **14 Suppl 6**: s71-77.

Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA et al. 2012. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149**: 979-993.

Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I, Alexandrov LB, Martin S, Wedge DC et al. 2016. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**: 47-54.

O'Sullivan JM, Tan-Wong SM, Morillon A, Lee B, Coles J, Mellor J, Proudfoot NJ. 2004. Gene loops juxtapose promoters and terminators in yeast. *Nature genetics* **36**: 1014.

Odell ID, Wallace SS, Pederson DS. 2013. Rules of engagement for base excision repair in chromatin. *Journal of cellular physiology* **228**: 258-266.

Osley MA, Tsukuda T, Nickoloff JA. 2007. ATP-dependent chromatin remodeling factors and DNA damage repair. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **618**: 65-80.

Otterlei M, Kavli B, Standal R, Skjelbred C, Bharati S, Krokan HE. 2000. Repair of chromosomal abasic sites in vivo involves at least three different repair pathways. *The EMBO journal* **19**: 5542-5551.

Ozonov EA, van Nimwegen E. 2013. Nucleosome free regions in yeast promoters result from competitive binding of transcription factors that interact with chromatin modifiers. *PLoS computational biology* **9**: e1003181.

Pacholczyk M, Czernicki J, Ferenc T. 2016. The effect of solar ultraviolet radiation (UVR) on induction of skin cancers. *Medycyna pracy* **67**: 255-266.

Pai C-C, Kearsey SE. 2017. A Critical Balance: dNTPs and the Maintenance of Genome Stability. *Genes* **8**: 57.

Pan L, Penney J, Tsai L-H. 2014. Chromatin regulation of DNA damage repair and genome integrity in the central nervous system. *Journal of molecular biology* **426**: 3376-3388.

Papaemmanuil E, Gerstung M, Malcovati L, Tauro S, Gundem G, Van Loo P, Yoon CJ, Ellis P, Wedge DC, Pellagatti A. 2013. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*: blood-2013.

Patel RP, McAndrew J, Sellak H, White CR, Jo H, Freeman BA, Darley-Usmar VM. 1999. Biological aspects of reactive nitrogen species. *Biochimica et Biophysica Acta (BBA)-Bioenergetics* **1411**: 385-400.

Paulovich AG, Toczyski DP, Hartwell LH. 1997. When checkpoints fail. *Cell* **88**: 315-321.

Peak JG, Peak MJ, MacCoss M. 1984. DNA breakage caused by 334-nm ultraviolet light is enhanced by naturally occurring nucleic acid components and nucleotide coenzymes. *Photochemistry and photobiology* **39**: 713-716.

Pegg AE. 2011. Multifaceted roles of alkyltransferase and related proteins in DNA repair, DNA damage, resistance to chemotherapy, and research tools. *Chemical research in toxicology* **24**: 618-639.

Perera D, Poulos RC, Shah A, Beck D, Pimanda JE, Wong JWH. 2016. Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532**: 259-+.

Peterson LA. 2010. Formation, repair, and genotoxic properties of bulky DNA adducts formed from tobacco-specific nitrosamines. *Journal of nucleic acids* **2010**.

Petljak M, Alexandrov LB. 2016. Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis* **37**: 531-540.

Pfeifer D, Chung YM, Hu MCT. 2015. Effects of low-dose bisphenol A on DNA damage and proliferation of breast cells: the role of c-Myc. *Environmental health perspectives* **123**: 1271.

Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. 2002. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *oncogene* **21**: 7435.

Pfeifer GP, Kadam S, Jin S-G. 2013. 5-hydroxymethylcytosine and its potential roles in development and cancer. *Epigenetics & chromatin* **6**: 10.

Pfeifer GP, You Y-H, Besaratinia A. 2005. Mutations induced by ultraviolet light. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **571**: 19-31.

Pfeiffer P, Goedecke W, Obe G. 2000. Mechanisms of DNA double-strand break repair and their potential to induce chromosomal aberrations. *Mutagenesis* **15**: 289-302.

Pham P, Bransteitter R, Petruska J, Goodman MF. 2003. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* **424**: 103.

Pham-Huy LA, He H, Pham-Huy C. 2008. Free radicals, antioxidants in disease and health. *International journal of biomedical science: IJBS* **4**: 89.

Phillips DH. 2018. Mutational spectra and mutational signatures: Insights into cancer aetiology and mechanisms of DNA damage and repair. *DNA repair* **71**: 6-11.

Pilati C, Shinde J, Alexandrov LB, Assié G, André T, Hélias-Rodzewicz Z, Ducoudray R, Le Corre D, Zucman-Rossi J, Emile JF. 2017. Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *The Journal of pathology* **242**: 10-15.

Pipek O, Ribli D, Molnár J, Póti Á, Krzystanek M, Bodor A, Tusnady GE, Szallasi Z, Csabai I, Szüts D. 2017. Fast and accurate mutation detection in whole genome sequences of multiple isogenic samples with IsoMut. *BMC bioinformatics* **18**: 73.

Piper JD, Piper PW. 2017. Benzoate and sorbate salts: a systematic review of the potential hazards of these invaluable preservatives and the expanding spectrum of clinical uses for sodium benzoate. *Comprehensive Reviews in Food Science and Food Safety* **16**: 868-880.

Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordóñez GR, Bignell GR et al. 2010a. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**: 191-196.

Pleasance ED, Stephens PJ, O'meara S, McBride DJ, Meynert A, Jones D, Lin M-L, Beare D, Lau KW, Greenman C. 2010b. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**: 184.

Polak P, Lawrence MS, Haugen E, Stoletzki N, Stojanov P, Thurman RE, Garraway LA, Mirkin S, Getz G, Stamatoyannopoulos JA. 2014. Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nature biotechnology* **32**: 71.

Polo SE. 2015. Reshaping chromatin after DNA damage: the choreography of histone proteins. *Journal of molecular biology* **427**: 626-636.

Polo SE, Almouzni G. 2015. Chromatin dynamics after DNA damage: The legacy of the access-repair-restore model. *DNA Repair* **36**: 114-121.

Poon SL, McPherson JR, Tan P, Teh BT, Rozen SG. 2014. Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention. *Genome medicine* **6**: 24.

Poon SL, Pang S-T, McPherson JR, Yu W, Huang KK, Guan P, Weng W-H, Siew EY, Liu Y, Heng HL. 2013. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Science translational medicine* **5**: 197ra101-197ra101.

Poulos RC, Olivier J, Wong JWH. 2017. The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes. *Nucleic acids research* **45**: 7786-7795.

Pourquier P. 2011. Alkylating agents. *Bulletin du cancer* **98**: 1237-1251.

Powell JR, Bennett MR, Evans KE, Yu S, Webster RM, Waters R, Skinner N, Reed SH. 2015. 3D-DIP-Chip: a microarray-based method to measure genomic DNA damage. *Scientific Reports* **5**.

Pray L. 2008. DNA replication and causes of mutation. *Nature education* **1**: 214.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Rabik CA, Njoku MC, Dolan ME. 2006. Inactivation of O6-alkylguanine DNA alkyltransferase as a means to enhance chemotherapy. *Cancer treatment reviews* **32**: 261-276.

Raisner RM, Hartley PD, Meneghini MD, Bao MZ, Liu CL, Schreiber SL, Rando OJ, Madhani HD. 2005. Histone variant H2A. Z marks the 5′ ends of both active and inactive genes in euchromatin. *Cell* **123**: 233-248.

Ravanat J-L, Douki T, Cadet J. 2001. Direct and indirect effects of UV radiation on DNA and its components. *Journal of Photochemistry and Photobiology B: Biology* **63**: 88-102.

Reed E. 1998. Platinum-DNA adduct, nucleotide excision repair and platinum based anti-cancer chemotherapy. *Cancer treatment reviews* **24**: 331-344.

Reed SH. 2011. Nucleotide excision repair in chromatin: Damage removal at the drop of a HAT. *DNA Repair* **10**: 734-742.

Reed SH, Akiyama M, Stillman B, Friedberg EC. 1999. Yeast autonomously replicating sequence binding factor is involved in nucleotide excision repair. *Genes & development* **13**: 3052-3058.

Reed SH, You Z, Friedberg EC. 1998. The Yeast RAD7 and RAD16 Genes Are Required for Postincision Events during Nucleotide Excision Repair IN VITRO AND IN VIVO STUDIES WITHrad7 AND rad16 MUTANTS AND PURIFICATION OF A Rad7/Rad16-CONTAINING PROTEIN COMPLEX. *Journal of Biological Chemistry* **273**: 29481-29488.

Reuter S, Gupta SC, Chaturvedi MM, Aggarwal BB. 2010. Oxidative stress, inflammation, and cancer: how are they linked? *Free Radical Biology and Medicine* **49**: 1603-1616.

Rodriguez Y, Hinz JM, Smerdon MJ. 2015. Accessing DNA damage in chromatin: Preparing the chromatin landscape for base excision repair. *DNA repair* **32**: 113-119.

Rogozin IB, Goncearenco A, Lada AG, De S, Yurchenko V, Nudelman G, Panchenko AR, Cooper DN, Pavlov YI. 2018. DNA polymerase η mutational signatures are found in a variety of different types of cancer. *Cell Cycle* **17**: 348-355.

Romanello M, Schiavone D, Frey A, Sale JE. 2016. Histone H3.3 promotes IgV gene diversification by enhancing formation of AID-accessible single-stranded DNA. *Embo j* **35**: 1452-1464.

Roos WP, Kaina B. 2006. DNA damage-induced cell death by apoptosis. *Trends in molecular medicine* **12**: 440-450.

Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, Lopez-Bigas N. 2016. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**: 264-+.

Sage E. 1993. Distribution and repair of photolesions in DNA: genetic consequences and the role of sequence context. *Photochemistry and photobiology* **57**: 163-174.

Sale JE. 2013. Translesion DNA synthesis and mutagenesis in eukaryotes. *Cold Spring Harbor perspectives in biology* **5**: a012708.

Sale JE, Lehmann AR, Woodgate R. 2012. Y-family DNA polymerases and their role in tolerance of cellular DNA damage. *Nature reviews Molecular cell biology* **13**: 141.

Salk JJ, Fox EJ, Loeb LA. 2010. Mutational heterogeneity in human cancers: origin and consequences. *Annual Review of Pathological Mechanical Disease* **5**: 51-75.

Sallmyr A, Tomkinson AE. 2018. Repair of DNA double-strand breaks by mammalian alternative end-joining pathways. *Journal of Biological Chemistry*: jbc-TM117.

Sanborn AL, Rao SSP, Huang S-C, Durand NC, Huntley MH, Jewett AI, Bochkov ID, Chinnappan D, Cutkosky A, Li J. 2015. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences* **112**: E6456-E6465.

Sancar A, Lindsey-Boltz LA, Ünsal-Kaçmaz K, Linn S. 2004. Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annual review of biochemistry* **73**: 39-85.

Scheibye-Knudsen M, Ramamoorthy M, Sykora P, Maynard S, Lin P-C, Minor RK, Wilson DM, Cooper M, Spencer R, de Cabo R. 2012. Cockayne syndrome group B protein prevents the accumulation of damaged mitochondria by promoting mitochondrial autophagy. *Journal of Experimental Medicine* **209**: 855-869

Schlecht U, Erb I, Demougin P, Robine N, Borde V, van Nimwegen E, Nicolas A, Primig M. 2008. Genome-wide expression profiling, in vivo DNA binding analysis, and probabilistic motif prediction reveal novel Abf1 target genes during fermentation, respiration, and sporulation in yeast. *Molecular Biology of the Cell* **19**: 2193-2207.

Schuch AP, da Silva Galhardo R, de Lima-Bessa KM, Schuch NJ, Menck CFM. 2009. Development of a DNA-dosimeter system for monitoring the effects of solar-ultraviolet radiation. *Photochemical & Photobiological Sciences* **8**: 111-120.

Schuster-Böckler B, Lehner B. 2012. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *nature* **488**: 504.

Schärer OD. 2013. Nucleotide excision repair in eukaryotes. *Cold Spring Harbor perspectives in biology* **5**: a012609.

Segovia R, Tam AS, Stirling PC. 2015. Dissecting genetic and environmental mutation signatures with model organisms. *Trends in Genetics* **31**: 465-474.

Serero A, Jubin C, Loeillet S, Legoix-Né P, Nicolas AG. 2014. Mutational landscape of yeast mutator strains. *Proceedings of the National Academy of Sciences* **111**: 1897-1902.

Shen J-C, Rideout Iii WM, Jones PA. 1994. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic acids research* **22**: 972-976.

Shi W, Ng CKY, Lim RS, Jiang T, Kumar S, Li X, Wali VB, Piscuoglio S, Gerstein MB, Chagpar AB. 2018. Reliability of Whole-Exome Sequencing for Assessing Intratumor Genetic Heterogeneity. *bioRxiv*: 253195.

Shimada T, Fujii-Kuriyama Y. 2004. Metabolic activation of polycyclic aromatic hydrocarbons to carcinogens by cytochromes P450 1A1 and 1B1. *Cancer Sci* **95**: 1-6.

Shinohara A, Ogawa H, Ogawa T. 1992. Rad51 protein involved in repair and recombination in S. cerevisiae is a RecA-like protein. *Cell* **69**: 457-470.
Shrivastav M, De Haro LP, Nickoloff JA. 2008. Regulation of DNA double-strand break repair pathway choice. *Cell research* **18**: 134.
Sima J, Gilbert DM. 2014. Complex correlations: replication timing and mutational landscapes during cancer and genome evolution. *Curr Opin Genet Dev* **25**: 93-100.
Singer B, Grunberger D. 2012. *Molecular biology of mutagens and carcinogens*. Springer Science & Business Media.
Sinha RP, Häder D-P. 2002. UV-induced DNA damage and repair: a review. *Photochemical & Photobiological Sciences* **1**: 225-236.
Smela ME, Currier SS, Bailey EA, Essigmann JM. 2001. The chemistry and biology of aflatoxin B1: from mutational spectrometry to carcinogenesis. *Carcinogenesis* **22**: 535-545.
Smela ME, Hamm ML, Henderson PT, Harris CM, Harris TM, Essigmann JM. 2002. The aflatoxin B1 formamidopyrimidine adduct plays a major role in causing the types of mutations observed in human hepatocellular carcinoma. *Proceedings of the National Academy of Sciences* **99**: 6655-6660.
Smerdon MJ. 1991. DNA repair and the role of chromatin structure. *Current opinion in cell biology* **3**: 422-428.
Snellman E, Jansen CT, Lauharanta J, Kolari P. 1992. Solar ultraviolet (UV) radiation and UV doses received by patients during four-week climate therapy periods in the Canary Islands. *Photodermatology, photoimmunology & photomedicine* **9**: 40-43.
Sproviero M, Verwey AM, Rankin KM, Witham AA, Soldatov DV, Manderville RA, Fekry MI, Sturla SJ, Sharma P, Wetmore SD. 2014. Structural and biochemical impact of C8-aryl-guanine adducts within the NarI recognition DNA sequence: influence of aryl ring size on targeted and semi-targeted mutagenicity. *Nucleic Acids Res* **42**: 13405-13421.
Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nature genetics* **41**: 393.
Starcevic D, Dalal S, Sweasy JB. 2004. Is there a link between DNA polymerase beta and cancer? *Cell Cycle* **3**: 996-999.
Stempor P, Ahringer J. 2016. SeqPlots-Interactive software for exploratory data analyses, pattern discovery and visualization in genomics. *Wellcome open research* **1**.
Stewart PA, Coble JB, Vermeulen R, Schleiff P, Blair A, Lubin J, Attfield M, Silverman DT. 2010. The diesel exhaust in miners study: I. Overview of the exposure assessment process. *Ann Occup Hyg* **54**: 728-746.
Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. *Nature* **458**: 719-724.
Strauss BS. 1991. The 'A rule'of mutagen specificity: A consequence of DNA polymerase bypass of non-instructional lesions? *Bioessays* **13**: 79-84.
Su TT. 2006. Cellular responses to DNA damage: one signal, multiple choices. *Annu Rev Genet* **40**: 187-208.
Sugasawa K. 2008. Xeroderma pigmentosum genes: functions inside and outside DNA repair. *Carcinogenesis* **29**: 455-465.
Swan MK, Johnson RE, Prakash L, Prakash S, Aggarwal AK. 2009. Structural basis of high-fidelity DNA synthesis by yeast DNA polymerase δ. *Nature structural & molecular biology* **16**: 979.
Swanson RL, Morey NJ, Doetsch PW, Jinks-Robertson S. 1999. Overlapping specificities of base excision repair, nucleotide excision repair, recombination, and translesion

synthesis pathways for DNA base damage in Saccharomyces cerevisiae. *Molecular and cellular biology* **19**: 2929-2935.

Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, Ettwiller L, Spitz F. 2014. Functional and topological characteristics of mammalian regulatory domains. *Genome research*.

Takata M, Sasaki MS, Sonoda E, Morrison C, Hashimoto M, Utsumi H, Yamaguchi-Iwai Y, Shinohara A, Takeda S. 1998. Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. *The EMBO journal* **17**: 5497-5508.

Tan-Wong SM, Zaugg JB, Camblong J, Xu Z, Zhang DW, Mischo HE, Ansari AZ, Luscombe NM, Steinmetz LM, Proudfoot NJ. 2012. Gene loops enhance transcriptional directionality. *Science* **338**: 671-675.

Taylor BJ, Nik-Zainal S, Wu YL, Stebbings LA, Raine K, Campbell PJ, Rada C, Stratton MR, Neuberger MS. 2013. DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *Elife* **2**: e00534.

Team RC. 2014. R: A language and environment for statistical computing.

Teixeira MC, Monteiro PT, Palma M, Costa C, Godinho CP, Pais P, Cavalheiro M, Antunes M, Lemos A, Pedreira T. 2017. YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. *Nucleic acids research* **46**: D348-D353.

Teng Y, Liu H, Gill HW, Yu Y, Waters R, Reed SH. 2008. Saccharomyces cerevisiae Rad16 mediates ultraviolet-dependent histone H3 acetylation required for efficient global genome nucleotide-excision repair. *EMBO Rep* **9**: 97-102.

Teng Y, Waters R. 2000. Excision repair at the level of the nucleotide in the upstream control region, the coding sequence and in the region where transcription terminates of the Saccharomyces cerevisiae MFA2 gene and the role of RAD26. *Nucleic acids research* **28**: 1114-1119.

Teng YM, Bennett M, Evans KE, Zhuang-Jackson H, Higgs A, Reed SH, Waters R. 2011. A novel method for the genome-wide high resolution analysis of DNA damage. *Nucleic Acids Research* **39**.

Tubbs A, Nussenzweig A. 2017. Endogenous DNA damage as a source of genomic instability in cancer. *Cell* **168**: 644-656.

Tubbs JL, Latypov V, Kanugula S, Butt A, Melikishvili M, Kraehenbuehl R, Fleck O, Marriott A, Watson AJ, Verbeek B. 2009. Flipping of alkylated DNA damage bridges base and nucleotide excision repair. *Nature* **459**: 808.

Unnikrishnan A, Gafken PR, Tsukiyama T. 2010. Dynamic changes in histone acetylation regulate origins of DNA replication. *Nature structural & molecular biology* **17**: 430.

Valavanidis A, Vlachogianni T, Fiotakis C. 2009. 8-hydroxy-2′-deoxyguanosine (8-OHdG): a critical biomarker of oxidative stress and carcinogenesis. *Journal of environmental science and health Part C* **27**: 120-139.

Van Attikum H, Gasser SM. 2009. Crosstalk between histone modifications during the DNA damage response. *Trends in cell biology* **19**: 207-217.

van Bakel H, Tsui K, Gebbia M, Mnaimneh S, Hughes TR, Nislow C. 2013. A compendium of nucleosome and transcript profiles reveals determinants of chromatin architecture and transcription. *PLoS genetics* **9**: e1003479.

van Eijk P, Nandi SP, Yu S, Bennett M, Leadbitter M, Teng Y, Reed S. 2018. Nucleosome remodelling at origins of Global Genome-Nucleotide Excision Repair occurs at the boundaries of higher-order chromatin structure. *bioRxiv*: 283747.

van Gool AJ, Verhage R, Swagemakers SM, van de Putte P, Brouwer J, Troelstra C, Bootsma D, Hoeijmakers JH. 1994. RAD26, the functional S. cerevisiae homolog of the Cockayne syndrome B gene ERCC6. *The EMBO Journal* **13**: 5361-5369.

Varon R, Vissinga C, Platzer M, Cerosaletti KM, Chrzanowska KH, Saar K, Beckmann G, Seemanová E, Cooper PR, Nowak NJ. 1998. Nibrin, a novel DNA double-strand break repair protein, is mutated in Nijmegen breakage syndrome. *Cell* **93**: 467-476.

Verhage R, Zeeman AM, de Groot N, Gleig F, Bang DD, van de Putte P, Brouwer J. 1994. The RAD7 and RAD16 genes, which are essential for pyrimidine dimer removal from the silent mating type loci, are also required for repair of the nontranscribed strand of an active gene in Saccharomyces cerevisiae. *Molecular and Cellular Biology* **14**: 6135-6142.

Viel A, Bruselles A, Meccia E, Fornasarig M, Quaia M, Canzonieri V, Policicchio E, Urso ED, Agostini M, Genuardi M. 2017. A specific mutational signature associated with DNA 8-oxoguanine persistence in MUTYH-defective colorectal cancer. *EBioMedicine* **20**: 39-49.

Viguera E, Canceill D, Ehrlich SD. 2001. Replication slippage involves DNA polymerase pausing and dissociation. *The EMBO journal* **20**: 2587-2595.

Vincent JA, Kwong TJ, Tsukiyama T. 2008. ATP-dependent chromatin remodeling shapes the DNA replication landscape. *Nature structural & molecular biology* **15**: 477.

Waller RG, Darlington TM, Wei X, Madsen MJ, Thomas A, Curtin K, Coon H, Rajamanickam V, Musinsky J, Jayabalan D. 2018. Novel pedigree analysis implicates DNA repair and chromatin remodeling in multiple myeloma risk. *PLoS genetics* **14**: e1007111.

Wang G, Vasquez KM. 2014. Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. *DNA repair* **19**: 143-151.

Wang GG, Allis CD, Chi P. 2007a. Chromatin remodeling and cancer, part I: covalent histone modifications. *Trends in Molecular Medicine* **13**: 363-372.

Wang GG, Allis CD, Chi P. 2007b. Chromatin remodeling and cancer, part II: ATP-dependent chromatin remodeling. *Trends in Molecular Medicine* **13**: 373-380.

Wang Z. 2001. DNA damage-induced mutagenesis : a novel target for cancer prevention. *Mol Interv* **1**: 269-281.

Waters LS, Minesinger BK, Wiltrout ME, D'Souza S, Woodruff RV, Walker GC. 2009. Eukaryotic translesion polymerases and their roles and regulation in DNA damage tolerance. *Microbiology and Molecular Biology Reviews* **73**: 134-154.

Waters LS, Walker GC. 2006. The critical mutagenic translesion DNA polymerase Rev1 is highly expressed during G2/M phase rather than S phase. *Proceedings of the National Academy of Sciences* **103**: 8971-8976.

Waters R, Evans K, Bennett M, Yu S, Reed S. 2012. Nucleotide excision repair in cellular chromatin: studies with yeast from nucleotide to gene to genome. *International journal of molecular sciences* **13**: 11141-11164.

Weber CM, Ramachandran S, Henikoff S. 2014. Nucleosomes are context-specific, H2A. Z-modulated barriers to RNA polymerase. *Molecular cell* **53**: 819-830.

Willmore B. 2006. *Adobe Photoshop CS2 studio techniques*. Adobe Press.

Woo YH, Li W-H. 2012. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nature communications* **3**: 1004.

Wood ML, Dizdaroglu M, Gajewski E, Essigmann JM. 1990. Mechanistic studies of ionizing radiation and oxidative mutagenesis: genetic effects of a single 8-hydroxyguanine (7-hydro-8-oxoguanine) residue inserted at a unique site in a viral genome. *Biochemistry* **29**: 7024-7032.

Woodbine L, Brunton H, Goodarzi AA, Shibata A, Jeggo PA. 2011. Endogenously induced DNA double strand breaks arise in heterochromatic DNA regions and require ataxia telangiectasia mutated and Artemis for their repair. *Nucleic acids research* **39**: 6986-6997.

World Health O. 2005. IARC, International agency for research on cancer WHO.

Wu Y, Berends MJW, Mensink RGJ, Kempinga C, Sijmons RH, van der Zee AGJ, Hollema H, Kleibeuker JH, Buys CHCM, Hofstra RMW. 1999. Association of hereditary nonpolyposis colorectal cancer–related tumors displaying low microsatellite instability with MSH6 germline mutations. *The American Journal of Human Genetics* **65**: 1291-1298.

Wyatt MD, Pittman DL. 2006. Methylating agents and DNA repair responses: Methylated bases and sources of strand breaks. *Chemical research in toxicology* **19**: 1580-1594.

Wyrick JJ, Roberts SA. 2015. Genomic approaches to DNA repair and mutagenesis. *DNA Repair* **36**: 146-155.

Xu C. 2018. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and structural biotechnology journal* **16**: 15-24.

Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**: 1033.

Yadon AN, Van de Mark D, Basom R, Delrow J, Whitehouse I, Tsukiyama T. 2010. Chromatin remodeling around nucleosome-free regions leads to repression of noncoding RNA transcription. *Mol Cell Biol* **30**: 5110-5122.

Yarragudi A, Parfrey LW, Morse RH. 2007. Genome-wide analysis of transcriptional dependence and probable target sites for Abf1 and Rap1 in Saccharomyces cerevisiae. *Nucleic Acids Research* **35**: 193-202.

Yu BP. 1994. Cellular defenses against damage from reactive oxygen species. *Physiological reviews* **74**: 139-162.

Yu S, Evans KE, van Eijk P, Bennett M, Webster RM, Leadbitter M, Teng Y, Waters R, Jackson SPP, Reed SH. 2016. Global Genome Nucleotide Excision Repair is Organised into Domains Promoting Efficient DNA Repair in Chromatin. *bioRxiv*: 050807.

Yu SR, Owen-Hughes T, Friedberg EC, Waters R, Reed SH. 2004. The yeast Rad7/Rad16/Abf1 complex generates superhelical torsion in DNA that is required for nucleotide excision repair. *DNA Repair* **3**: 277-287.

Yu SR, Smirnova JB, Friedberg EC, Stillman B, Akiyama M, Owen-Hughes T, Waters R, Reed SH. 2009. ABF1-binding Sites Promote Efficient Global Genome Nucleotide Excision Repair. *Journal of Biological Chemistry* **284**: 966-973.

Yu SR, Teng YM, Waters R, Reed SH. 2011. How Chromatin Is Remodelled during DNA Repair of UV-Induced DNA Damage in Saccharomyces cerevisiae. *Plos Genetics* **7**.

Yu Y, Deng Y, Reed SH, Millar CB, Waters R. 2013. Histone variant Htz1 promotes histone H3 acetylation to enhance nucleotide excision repair in Htz1 nucleosomes. *Nucleic Acids Res* **41**: 9006-9019.

Yu YC, Teng YM, Liu HR, Reed SH, Waters R. 2005. UV irradiation stimulates histone acetylation and chromatin remodeling at a repressed yeast locus. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 8650-8655.

Zentner GE, Kasinathan S, Xin B, Rohs R, Henikoff S. 2015. ChEC-seq kinetics discriminates transcription factor binding sites by DNA sequence and shape in vivo. *Nature communications* **6**: 8733.

Zhang X, Yu Q, Olsen L, Bi X. 2012. Functions of protosilencers in the formation and maintenance of heterochromatin in Saccharomyces cerevisiae. *PloS one* **7**: e37092.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**: R137.

Zhang Z, Pugh BF. 2011. High-resolution genome-wide mapping of the primary structure of chromatin. *Cell* **144**: 175-186.

Zhao C, Tyndyk M, Eide I, Hemminki K. 1999. Endogenous and background DNA adducts by methylating and 2-hydroxyethylating agents. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **424**: 117-125.

Zheng CL, Wang NJ, Chung J, Moslehi H, Sanborn JZ, Hur JS, Collisson EA, Vemula SS, Naujokas A, Chiotti KE et al. 2014. Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes. *Cell Reports* **9**: 1228-1234.

Zhou BB, Elledge SJ. 2000. The DNA damage response: putting checkpoints in perspective. *Nature* **408**: 433-439.

Zhu LJ, Gazin C, Lawson ND, Pages H, Lin SM, Lapointe DS, Green MR. 2010. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11**: 237.

Zou X, Owusu M, Harris R, Jackson SP, Loizou JI, Nik-Zainal S. 2018. Validating the concept of mutational signatures with isogenic cell models. *Nature communications* **9**: 1744.

Zámborszky J, Szikriszt B, Gervai JZ, Pipek O, Póti Á, Krzystanek M, Ribli D, Szalai-Gindl JM, Csabai I, Szallasi Z. 2017. Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene* **36**: 746.