
Genome-wide analysis of DNA damage and repair



Vom Fachbereich Biologie der Technischen Universität Darmstadt
zur
Erlangung des akademischen Grades
eines Doctor rerum naturalium
genehmigte Dissertation von

M.Sc. Bioinf. Wei Yu
aus JiangXi, China

Berichterstatter (1. Referent): Prof. Dr. M. Cristina Cardoso
Mitberichterstatter (2. Referent): Prof. Dr. Barbara Drossel

Tag der Einreichung: 15.10.2014
Tag der mündlichen Prüfung: 17.12.2014
Darmstadt 2015
D17

1. Preface

The structure of the present thesis was designed in such a way that each of two projects was described as independent chapters in the results section. The general introduction part included the biological and methodological background information relevant for both parts. Emphasis was placed on the introduction of next-generation sequencing techniques. Material and Methods were also written for both parts, since they are largely overlapping. The two major projects are presented in a way that resembles the publications in preparation. Each project has its own specific introduction and results section. One project analyzed the distribution and repair of DNA double strand break by its proxy γ H2AX after ionizing radiation, the second projects studied the induction and repair of cyclobutane pyrimidine dimers in response to ultraviolet C light. Correspondingly, the general discussion and outlook summarized the result of both result chapters, connected both projects parts and described potential ideas to be validated in future.

2. Summary

In the first part of this thesis I studied the genome wide distribution of γ H2AX, H2AX and H3 under physiological conditions and after 10 Gy X-ray exposure in HepG2 cells. This was done using a chromatin immune-precipitation approach coupled to massively parallel DNA sequencing (ChIP-seq). This method enables the mapping of sequences that are coupled to the above-mentioned histones to the genome and thus allows studying the DNA damage response with high resolution in a genome-wide manner. Using these data I could show that under physiological conditions neither H3, H2AX nor γ H2AX are randomly distributed, but all three histone variants are overrepresented in euchromatic regions. But the relative γ H2AX abundance (compared to H2AX) is inverted with a weak enrichment in heterochromatic areas. After exposure to ionizing radiation (IR) euchromatic areas show an overrepresentation at early time points, positively correlated to high GC content, transcription and H3K36me3 histone marks. In contrast at 24h an inverted γ H2AX distribution becomes apparent, with correlation to H3K9me3 histone marks, low GC content and non-genic regions. Detailed analysis revealed that the expression level influences the phosphorylation levels at early time points in genic regions and that intermediately expressed genes show the strongest response. The analysis of repetitive elements revealed that different repetitive elements respond either according to their GC content, e.g. ALUs, or independent of their GC content but rather directed by their secondary structure, e.g. satellite repeats.

The second part of the thesis was aimed to study the genome-wide distribution of cyclobutane pyrimidine dimers (CPDs) following UVC exposure. Therefore a modified DNA immunoprecipitation technique was developed to combine with high-throughput sequencing that provided strand specific information (ssDIP-seq). The induction and persistence of a major DNA photo-lesion CPDs are thought to affect transcription, induce mutagenesis and finally contribute to skin cancer. Since CPDs can be repaired by two different sub-branches of the nucleotide excision repair pathway (NER), namely a XPC dependent global genomic (GG-NER) and CSB dependent transcription coupled NER (TC-NER), we studied the CPD repair in a NER proficient (HaCaT) and a XPC deficient cell line. The XPC^{-/-} cells show higher levels of endogenous copy number variations than HaCaT cells and thus support the idea that repair deficiency might contribute to genomic aberrations. Chromosomes 16, 17, 19, X with high densities of microsatellites show resistance of CPD repair in a chromosome-specific manner. The motifs of CPD hotspots are confirmed as continuous di-pyrimidine dimers and CPD distribution analysis revealed a non-random dispersal with preferential enrichment in repetitive regions especially in microsatellite and low complexity repeats. In genic regions, CPDs are distributed in a strand-specific manner and CPDs are overrepresented in the anti-sense strand rather than the sense strand, gradient increase from transcription start to stop site of the sense strand and anti-correlated to the expression levels. The chromatin feature analysis around the CPD hotspots shows

that condensed chromatin does not inhibit the formation of CPDs but hinders the repair process. Furthermore, histone marks for euchromatin are underrepresented around CPDs and heterochromatin is slightly enriched. This validates that a majority of un-repaired CPDs are located inside of heterochromatic regions and are deplete in regions with euchromatin histone modifications. And this tendency is enhanced in repair deficient cells at late repair time points.

3. Zusammenfassung

Im ersten Teil dieser Arbeit untersuchte ich in HepG2 Zellen die genom-weite Verteilung von γ H2AX, H2AX und H3 sowohl unter physiologischen Bedingungen als auch nach einer Bestrahlung mit 10Gy Röntgenstrahlung. Diese Untersuchungen wurden mittels Chromatin-Immunopräzipitation gefolgt von Hochdurchsatzsequenzierung durchgeführt (ChIP-seq). Dieser Ansatz ermöglicht es DNA Fragmente anzureichern die an die oben genannten Histon gebunden sind und diese im Genom zu kartieren, sodass sich die Untersuchungen mit hoher Auflösung genom-weit durchführen lassen. Auf diesen Daten basieren konnte ich zeigen, dass unter physiologischen Bedingungen weder H3, noch H2AX noch γ H2AX zufällig im Genom verteilt sind, sondern, dass alle drei Histon Varianten überrepräsentiert in euchromatischen Bereichen vorliegen. Betrachtet man allerdings den relativen Anteil an γ H2AX in Bezug auf H2AX so stellt man eine umgekehrte Verteilung mit einer leichten Überrepräsentation im Heterochromatin fest. Nach der Exposition gegenüber ionisierender Strahlung zeigen euchromatische Genomregionen bei frühen Zeitpunkten eine höhere Häufigkeit für γ H2AX an. Dies ist korreliert mit einem hohen GC Gehalt, aktiver Transkription und einer H3K36me3 Histonmethylierung. Im Gegensatz dazu zeigt sich nach 24h eine inverse γ H2AX Verteilung, mit einer positiven Korrelation zu H3K9me3 Histonmodifikationen, einem niederen GC Gehalt und nicht-codierenden Regionen. Eine detaillierte Analyse zeigte, dass die Stärke der Transkription die γ H2AX Phosphorylierung in Genregionen bei frühen Zeitpunkten beeinflusst. Die Analyse von Repetitiven Elementen zeigte, dass diese entweder entsprechende ihrem GC Gehalt (z.B. Alu Elemente) wie die anderen genomischen Elemente sich in Bezug auf ihre γ H2AX Antwort verhalten, oder unabhängig vom GC Gehalt verhalten, wobei dann die γ H2AX Antwort von der z.B. kompakten Sekundärstruktur der repetitive Elemente geprägt ist (z.B. Satellite Repeates).

Der zweite Teil der Arbeit zielte darauf ab die Genome weite Verteilung der DNA-Läsion Cyclobutane Pyrimidin Dimer (CPD) zu untersuchen. Dafür entwickelte ich eine modifizierte Version der Immunopräzipitations Technik, welche Strangspezifische DNA Sequenzinformationen liefert (ssDIP-seq). Die Induktion und Persistenz einer der wichtigsten DNA-Photo-Schäden CPDs sind dafür bekannt, dass sie die Transcription beeinflussen, Mutagenese induzieren und schlussendlich zu Hautkrebsentstehung beitragen. Da CPDs durch zwei unterschiedliche Unter-Reparaturwege der Nukleotid Exzisions Reparatur (NER) entfernt werden können, nämlich der XPC abhängigen global genomischen (GG-NER) und der CSB abhängigen transkriptionsgekoppelten (TC-NER), untersuchten wir die CPD Reparatur in einer NER Profizienten (HaCAT) und einer XPC defizienten Zell Linie. Die XPC^{-/-} Zellen zeigten eine erhöhte Menge an endogenen „Copy Number Variations“ als die Zelllinie HaCaT. Damit wird die Theorie unterstützt, dass DNA Reparaturdefizienz zu genomischen Aberrationen führt. Die Chromosomen 16, 17, 19 und X weisen eine hohe Dichte an Mikrosatellit DNA auf und erweisen sich im Verlauf der CPD Reparatur als sehr reparatur-resistant. Die Analyse der

CPD Hotspot Motive bestätigte, dass CPDs bevorzugt in kontinuierlichen Di-Pyrimidinen liegt. Diese liegen nicht zufällig verteilt, bevorzugt in repetitiven Elementen des Genoms, vor allem in Mikrosatelliten und „low Complexity repeats“. In Genen sind CPDs strangspezifisch verteilt, wobei CPDs im nicht-codierenden Strang überrepräsentiert sind. Die CPD Verteilung nimmt vom Transkriptionsstart zum Transkriptionsende zu und ist insgesamt korreliert zum Expressionsspiegel. Die Analyse der Chromatinumgebung um die CPD Peaks herum zeigt, dass kondensiertes Chromatin die Induktion der CPDs nicht verhindert, aber den Reparaturprozess behindert. Darüber hinaus sind genom-weit euchromatische Markierungen in der Nachbarschaft zu CPDs unterrepräsentiert, während heterochromatische Markeirungen leicht angereichert sind. Diese bestätigt, dass nicht-reparierte CPDs vorwiegend in heterochromatischem Genomregionen liegen.

Contents

1. General Introduction	4
DNA damage response.....	4
Chromatin structure and histone modifications.....	6
Genome-wide studies by next-generation sequencing.....	7
ChIP-seq	12
FAIRE-seq.....	14
DNase-seq	14
2. Aims of the work.....	16
3. Materials and Methods.....	17
Cell culture and irradiation	17
Immunofluorescence staining	17
High content screening	17
DNA extraction and quantification	18
Slot blot for CPD	18
CPD ssDIP-seq library preparation.....	19
Mapping, RPKM and peak calling.....	20
4. Result.....	21
4.1.Genome-wide analysis of DNA double strand damage and response	21
Introduction	21
Result.....	22
4.2.Genome-wide analysis of strand-specific cyclobutane pyrimidine dimer induction in NER repair proficient and deficient cell lines.....	40
Introduction	40
Result.....	45
5. Final conclusion and outlook.....	69
6. Reference	73
7. Appendix	80

List of Figures

Figure 1. Schematic illustration of nucleotide excision repair pathway.....	5
Figure 2. Schematic illustration of Illumina sequencing.....	8
Figure 3. Illustration of single-end, paired-end and mate-pair sequencing.....	10
Figure 4. Schematic illustration of ChIP-seq procedure for γ H2AX.....	12
Figure 5. Flow chart of DNase-seq and FAIRE-seq protocol.....	14
Figure 6. Characterization and validation of cellular system and experimental strategy.....	22
Figure 7. Establishment and validation of genome segmentation.....	25
Figure 8. Chromatin dynamics before and during DDR.....	27
Figure 9. The response of H3, H2AX, γ H2AX with GC content.....	29
Figure 10. Response of genic and non-genic elements to IR.....	32
Figure 11. Correlation of γ H2AX distribution during DDR to genomic features.....	35
Figure 12. Correlation of γ H2AX distribution with repetitive elements.....	39
Figure 13. Schematic illustration of ssDIP-seq procedure.....	44
Figure 14. Characterization and validation of cellular system and experimental strategy.....	47
Figure 15. Copy number variations in XPC ^{-/-} and HaCaT cell lines.....	50
Figure 16. Characterization of CPD distribution in genome-wide.....	54
Figure 17. Strand-specific CPD distribution in genic regions.....	60
Figure 18. The correlation of expression level with CPD abundance in metagene.....	62
Figure 19. RPKM of CPD in transcribed and non-transcribed strand.....	63
Figure 20. Chromatin structure around CPD peaks.....	66
Figure 21. Histone modification level around CPD peaks.....	68

List of Tables

Table 1: Overviews of genomic features	38
Table 2: Genes with amplified copy number in melanogenesis pathway	51
Table 3: Length percentage of CPD peaks overlapped by genomic elements and corresponding coverage of genomic elements in reference genome.....	56
Table 4: Length percentage of genomic elements overlapped by CPD peaks	56
Table 5: Retrieved DNase-seq, FAIRE-seq and histone modification data from Encode project.....	64
Table 6: Summary of γ H2AX and CPD genome-wide distribution and repair kinetics features.....	72

1. General Introduction

The DNA damage response

DNA (deoxyribonucleic acid) is the basic and vital genetic element to store and inherit information, which is essential for life. However, DNA is also vulnerable to be damaged by endogenous (e.g. reactive oxygen species) and exogenous (e.g. ultraviolet light, ion radiation, genotoxic chemicals) agents which form different DNA lesions such as 8-oxoguanine, 6-4 photoproduct (6-4 PP), cyclobutane pyrimidine dimers (CPD) and DNA double strand break (DSB) (Cooke et al., 2003; De Bont and van Larebeke, 2004; Hoeijmakers, 2001).

The consequence of DNA damage is diverse and comprises acute effects that arrest the cell cycle in G1, S, G2 and M phase (Bartek and Lukas, 2001), inhibition of transcription (Ljungman and Zhang, 1996; Mei Kwei et al., 2004), blocking of DNA replication and induction of apoptosis (Ljungman and Zhang, 1996; Lopes et al., 2001). The long term effects are the induction of irreversible mutations (Cooke et al., 2003) and chromosome aberrations (Asaithamby et al., 2011; Janssen et al., 2011) that contribute to carcinogenesis and aging (Hoeijmakers, 2009; Klaunig et al., 2010).

To repair different kinds of DNA damage, cells utilize a variety of repair pathways according to the type of DNA damage. The Base Excision Repair (BER) pathway repairs base lesions that are induced e.g. by oxidation, alkylation, deamination or depurination/depyrimidination (Robertson et al., 2009) or strand breaks. This pathway is further sub-divided into short-patch BER when only a single nucleotide is exchanged and long-patch BER when several nucleotides are replaced. Nucleotide Excision Repair (NER) pathway counteracts bulky adducts such as 6-4 Photoproducts (64-PP) and CPDs, which are the two major lesions induced by the UV component of the sunlight. Also the NER pathway is sub-divided in two distinct sub-pathways. The choice which pathway is activated depends on the position of the lesion and the resulting recognition. Global genome NER (GG-NER) repairs lesions in the entire genome, mediated by the helix distortion that is recognized by XPC protein. In contrast transcription-coupled NER (TC-NER) repairs lesion specifically in the transcribed strand of genes, mediated by stalled RNA polymerase II that recruits CSB protein to remove the damage(de Boer and Hoeijmakers, 2000; de Laat et al., 1999; Scharer, 2013) (for an overview see Figure 1). For the repair of double strand breaks, there are two independent repair pathways: Homologous Recombination (HR) and Non-Homologous End Joining (NHEJ). HR repairs DSBs in late S and G2 phase and allows repair with higher fidelity, while NHEJ is the major repair pathway to repair DSB in any cell cycle stage, fast but less accurate (Rothkamm et al., 2003). However, NHEJ repairs DSB with higher rates of recombination

and translocation (Chapman et al., 2012; Jackson, 2002; Jakob et al., 2011; Lobrich and Jeggo, 2007).

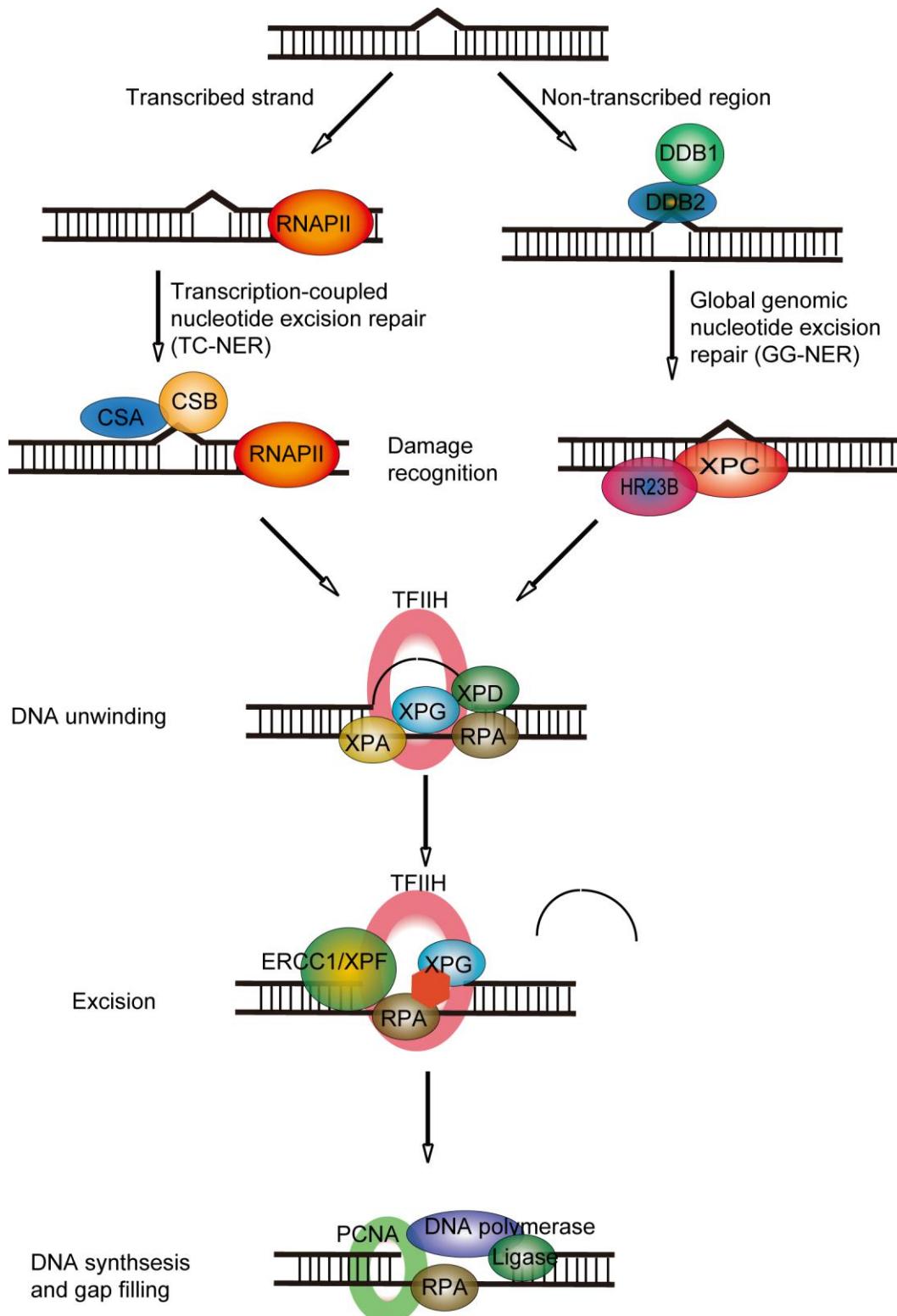


Figure 1. Schematic illustration of the Nucleotide Excision Repair pathway. This includes transcription-coupled repair and global genome repair, which differ on the recognition of the CPD and on the location of the damage. These distinct initiations are followed by common steps to unwind double strands and recruit proteins to cleave the strand with CPDs. Then synthesize new

DNA follow by ligation. The figure was modified according to (Cleaver et al., 2009) and (Fousteri and Mullenders, 2008).

Chromatin structure and histone modifications

In humans 147 base pairs (bp) of DNA are wrapped around an octamer of core histones containing 2x (H2A, H2B, H3, H4) to form nucleosomes, which are the fundamental unit of chromatin. The structure and dynamic of chromatin, which are regulated by histone tail modifications (such as H3K9me3) play a central role in nuclear and cellular functions. Chromatin of eukaryotes can be divided into two distinguishable functional compartments: heterochromatin and euchromatin. Heterochromatin is characterised by low GC content, low density of genes, repression of transcription, condensed chromatin structure and is usually located in the periphery of the nucleus. It can be further divided into two sub-categories: constitutive and facultative heterochromatin. Constitutive heterochromatin contains repetitive elements, forms centromeres and telomeres, as well as constantly repressed genes. Facultative heterochromatin also forms a compact chromatin structure, however, can turn into de-condensed or open structure under specific, temporal and spatial conditions, e.g. during development (Jost et al., 2012; Oberdoerffer and Sinclair, 2007; Trojer and Reinberg, 2007). In contrast to heterochromatin, euchromatin is defined as genomic regions, which are enriched in active genes and have an open chromatin structure.

Each of the core histones has a tail that sticks out of the nucleosome and is termed the N-terminal tail. These tails provide a platform for a large number of modifications; for instance methylation, acetylation, phosphorylation. These modifications play an important role in gene regulation and higher order chromatin structure. They dictate and orchestrate the recruitment of regulatory proteins that influence the fundamental biological functions such as transcription and replication. The histone modifications are not uniformly distributed and each histone modification is assumed to be functional. Histone modifications show dynamic and rapid exchange to regulate cellular functions during DNA replication, transcription and repair. Two potential functions of histone modification were proposed: the interference of clustering between nucleosomes to unravel condensed chromatin structure to establish an open chromatin environment. Another function is the modification of specific sites of the histone-tail to recruit or exclude the binding of non-histone proteins in order to localize and orchestrate enzymatic activities for DNA repair, replication, transcription etc to specific sites on the chromatin (Kouzarides, 2007; Zhou et al., 2011). For example, methylation of histone H3 lysine 9 is beneficial to the binding of HP1 protein and propagates heterochromatic spreading (Lachner et al., 2001).

As one of the fundamental functions of chromatin is to facilitate DNA repair, it shows a dynamic response to DNA damage. A model was proposed that chromatin concerted to access damage, prime the environment for repair and restore the chromatin structure after the DNA repair was finished (Misteli and Soutoglou, 2009; Soria et al., 2012). Chromatin remodelling upon activation of the cell cycle checkpoint and DNA repair is ATP dependent (Lans et al., 2012; Price and D'Andrea, 2013). However, the mechanism of chromatin remodelling response to DNA damage is still poorly understood. There remain many open questions such as which proteins or complexes are involved in the damage response of chromatin remodelling, consequences of failure of chromatin remodelling and its correlation to carcinogenesis.

Genome-wide studies by next-generation sequencing

Next-generation sequencing technologies (NGS) are constituted of three related techniques, which are represented by three companies: 454/Roche, SOLiD (Sequencing by Oligonucleotide Ligation and Detection) and Illumina (Metzker, 2010; Shendure and Ji, 2008). Among them the technique developed by Illumina is the most widely used and covers more than 90% of the current publications. The basic idea of NGS comes originally from shotgun sequencing techniques, which cut the long nuclear DNA into random small DNA fragments then sequence them by short Sanger sequencing and assemble the original genome according to the randomly generated overlaps. All three techniques can provide high coverage of genome-wide sequencing (named deep sequencing), but each technique has its own advantage and is based on different principles. The 454/Roche technique generates a single strand DNA (ssDNA) library and anneals it to agarose beads followed by emulsion PCR. Each single bead is arrayed into a plate and is subsequently sequenced by pyrosequencing, where the release of pyrophosphate after each incorporation of a nucleotide by DNA polymerase is converted to light. The technique developed by SOLiD also uses emulsion PCR coupled to magnetic beads to form clonal beads. Then sequencing is performed by ligation with a 2 base encoding system. These ligated primers are coupled to fluorescent dyes that are used for the read out (Luo et al., 2012; Mardis, 2008). These techniques produces sequence results of different lengths (each DNA sequence is termed read) that 454/Roche technique produces ~400 bps reads which is currently the longest, in contrast to 25~35 bps for the SOLiD technique. Whereas the Illumina technique can produce reads with a maximum length of 150bps (HiSeq 2500) and the total number of reads can be more than 200 million, which is the highest number among the three techniques (Bentley et al., 2008). Therefore, the 454/Roche technique is mostly used in combination with Illumina for *de novo* genome sequencing (Liu et al., 2012; Mardis, 2011). Since the Illumina

sequencing technique is the most widely used and is the sole method utilized in this work. The Illumina sequencing technique will be introduced in detail here.

The Illumina sequencing technique is based on sequencing by synthesis (SBS), the general procedure of library preparation is, firstly to fragment the DNA into 200~500 bps fragments. Then the DNA fragment is ligated to pre-designed Y-shape adaptors. Size selection (200~500 bps) is achieved by gel selection then the library is amplified by PCR. The sequencing procedure within the Illumina sequencing machine is starting by denaturing the DNA fragment library and hybridization of the single stranded library to complementary sequences immobilized on the surface of a small chip termed flowcell. After an isothermal amplification of the individually hybridized library molecules (cluster generation) the sequence is deciphered by a reversible dye-terminator-sequencing technique. Therefore four dNTPs with different fluorescent tags and cleavable terminator groups are added to the universal primer located in the immobilized adapter. After each nucleotide incubation step the fluorescence signal from each cluster is recorded and used to assign an incorporated base. Then the terminator is released and a new nucleotide incorporated. The details of Illumina sequencing are illustrated in Figure 2.

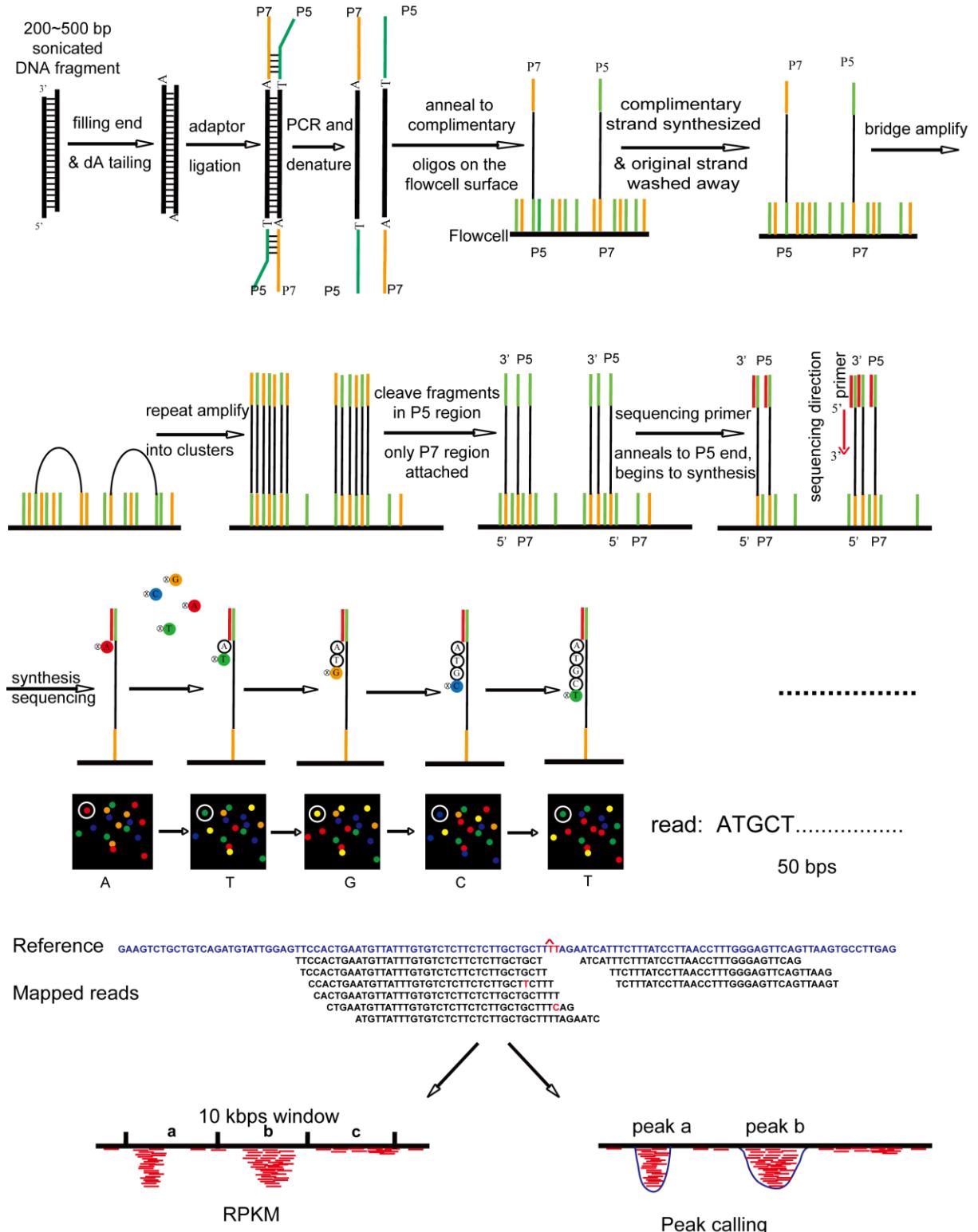


Figure 2. Schematic illustration of the Illumina sequencing method. Purified and sonicated DNA (200~500 bps) is end repaired to ligate a universal adapter to both ends. These adapters are used to amplify the library. Then the DNA is denatured into single strands, anneal to the complementary DNA on the surface of the flowcell. Then the complementary strand is synthesized and bridged amplification is used to generate clonal clusters that are then cleaved with the P5 end attached to the flowcell and only fragments with P7 attached to the flowcell are retained. Sequencing primers are annealed to the P5 end and reversible dye terminator sequencing is performed with fluorescently labelled dNTP. After each incorporation cycle the fluorescence is recorded (base calling), before the terminator and fluorescence tag is removed. After 50 cycles the

tag is read completely and the DNA sequences are mapped to the reference genome. This is divided into 10 kbps intervals and the RPKM is calculated for each interval. Alternatively peak calling is performed for each sample.

Illumina sequencing can be used for different applications, such as whole genome sequencing, *de novo* genome sequencing, exome sequencing, mRNA sequencing (mRNA-seq), micro-RNA sequencing (miRNA-seq) and protein-DNA binding sequencing (Chromatin immunoprecipitation coupled with high-throughput DNA sequencing, ChIP-seq). Whole genome sequencing is usually used to identify novel genetic sequences including mutations (single nucleotide polymorphism, SNPs) and copy number variations (CNVs) or structural variations (insertions, deletions, recombination, translocations) (Ng and Kirkness, 2010). For the purpose of detecting genetic variations, people usually apply either single-end sequencing, which reads one end of the DNA fragments, pair-end sequencing which reads both sides of the DNA fragments or mate-pair sequencing, which uses circularized DNA fragments first and then split them into pieces and sequencing two ends of the fragment, which includes genomic fusion fragments spanning larger distance than 2~5 kbps, to some percentage (illustrated in Figure 3) (Van Nieuwerburgh et al., 2012). If two ends of the same DNA fragment are mapped to genomic regions which have a distance of more than 2~5 kbps a potential genomic fusion site within this DNA fragment is detected.

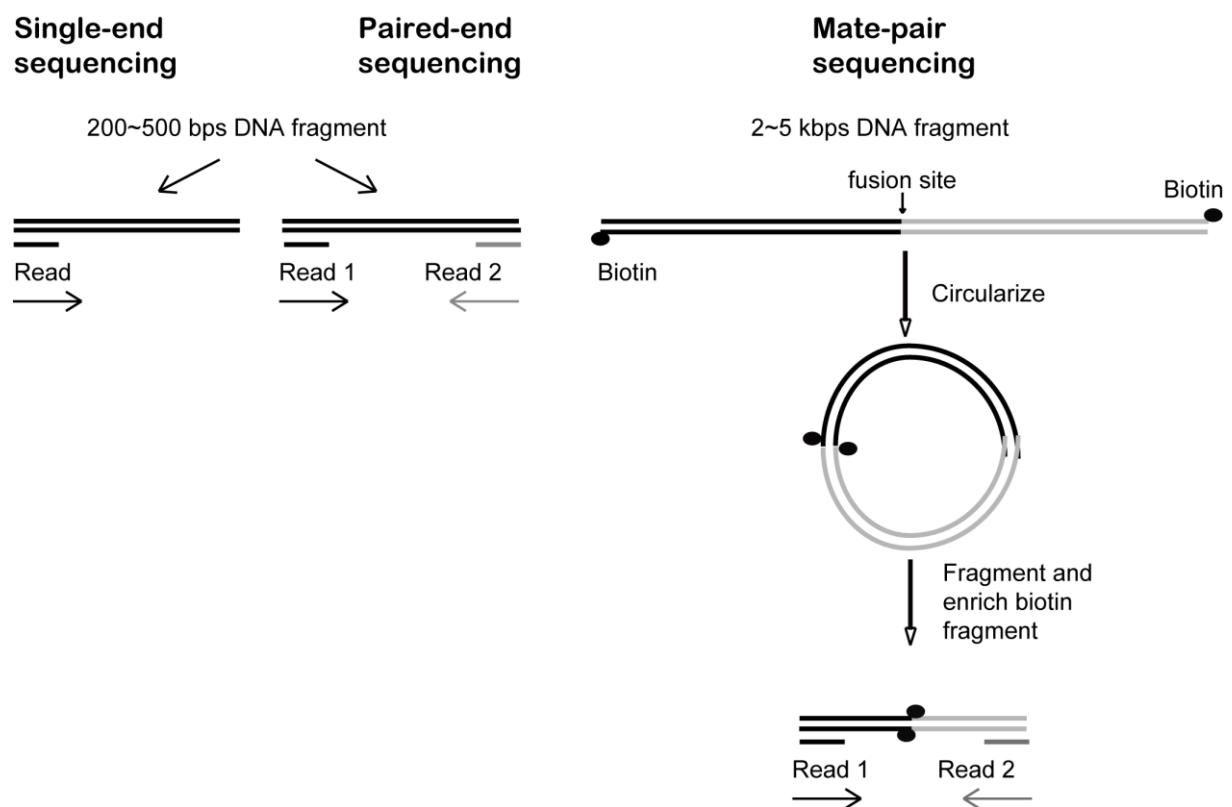


Figure 3. Illustration of single-end, paired-end and mate-pair sequencing. Single-end sequencing provides sequences of one end of DNA fragment. Whereas, both paired-end and mate-

pair sequencing detect the sequences of both ends which also cover the spanning distance information for uncovering genomic fusion site.

If sequencing samples are from species without public reference genomes, such as hg19 for human (*homo sapiens*) or mm9 for mouse (*mus musculus*), *de novo* genome sequencing is usually used to assemble a reference genome for instance the *de novo* genome sequencing for giant panda (Li et al., 2010). However, both whole genome sequencing and *de novo* genome sequencing highly rely on the sequencing depth and coverage (Sims et al., 2014), which are limited in terms of costs and time for most research groups. And *de novo* genome assembly software tools are still imperfect, due to the large portion of repetitive elements in the genome (Zhang et al., 2011).

Due to these disadvantages of *de novo* genome and whole genome sequencing, an alternative option is exome sequencing, which targets subsets of the genome, the exome as protein coding region. This is a powerful and cost-effective tool for discovering diseases related genetic variations and driver genes. Exome sequencing firstly captures exons through hybridization to biotinylated DNA or RNA baits which are designed for exons and is followed by biotin-streptavidin-based pull-down then amplification and sequencing are done in regular ways (Bamshad et al., 2011; Bras and Singleton, 2011). Since exons constitute only 1% of the human genome, ~30 Mbps, exome sequencing can provide ultra-deep and cost-effective sequencing result and is now widely used in clinics and research (Huang, 2011; Priya et al., 2012). However, exome sequencing only covers exons hence the variations in non-exon regions such as regulatory elements can't be assessed.

RNA-seq (mRNA-seq) is an approach for transcriptome profiling, which can provide precise transcript quantification and identify new isoforms. RNA-seq converts RNAs into a cDNA fragment library followed by massive parallel sequencing. Compared to traditional microarrays, this approach has several advantages. It provides a very high dynamic range in terms of the expression level (>8,000 fold) from transcripts to genes and can distinguish isoforms, allelic expression (Wang et al., 2009), even if these variations were previously unknown. RNA-seq can also be used to discover novel non-coding RNAs through sequencing of RNAs after filtering out ribosomal RNAs and mRNA (Cabili et al., 2011). miRNAs are highly conserved and regulatory molecules, which play an important role in many cellular process, such as let-7 and lin-4 miRNAs regulate target mRNA for degradation (Bagga et al., 2005). miRNA-seq as a sequencing method is used to survey the expression of miRNAs and discover novel miRNAs (Creighton et al., 2009; Morin et al., 2008).

There are many others deviations and approaches based on next-generation sequencing such as MeDIP-seq, Methyl-seq for the analysis of the methylation status of the DNA (Weber

et al., 2005), RIP-seq for identification of polycomb-associated RNAs (Zhao et al., 2010). To map spatially and temporally ordered DNA replication, Repli-seq (Hansen et al., 2010) can be applied. In the following paragraphs the approaches used in my projects are described in more detail: ChIP-seq, DNase-seq and FAIRE-seq.

ChIP-seq

Chromatin immunoprecipitation (ChIP) coupled with high-throughput DNA sequencing (ChIP-seq) has become a powerful and widely used approach for unbiased identification of genomic binding sites of given transcription factors, histone modifications or chromatin modifying complexes *in vivo* (Johnson et al., 2007). In the ChIP-seq procedure, the DNA for sequencing is isolated from cross-linked complexes with the protein of interest. After fixation, the procedures of ChIP-seq for transcription factor and histones modification are identical: chromatin or DNA is fragmented (e.g. by sonication, or enzymatic). Then antibodies coupled to magnetic beads are used to pull down the complex of chromatin, which is enriched in the given protein. DNA is reverse-cross-linked and further purified. Then the isolated DNA fragments are used for the preparation of sequencing libraries (Kharchenko et al., 2008). The illustration of ChIP-seq procedure for γ H2AX is shown in Figure 4. The ENCODE project provides guidelines to perform standard ChIP-seq experiments, data analysis and sequencing quality control (Landt et al., 2012), which are accepted as the gold standard in terms of experimental quality.

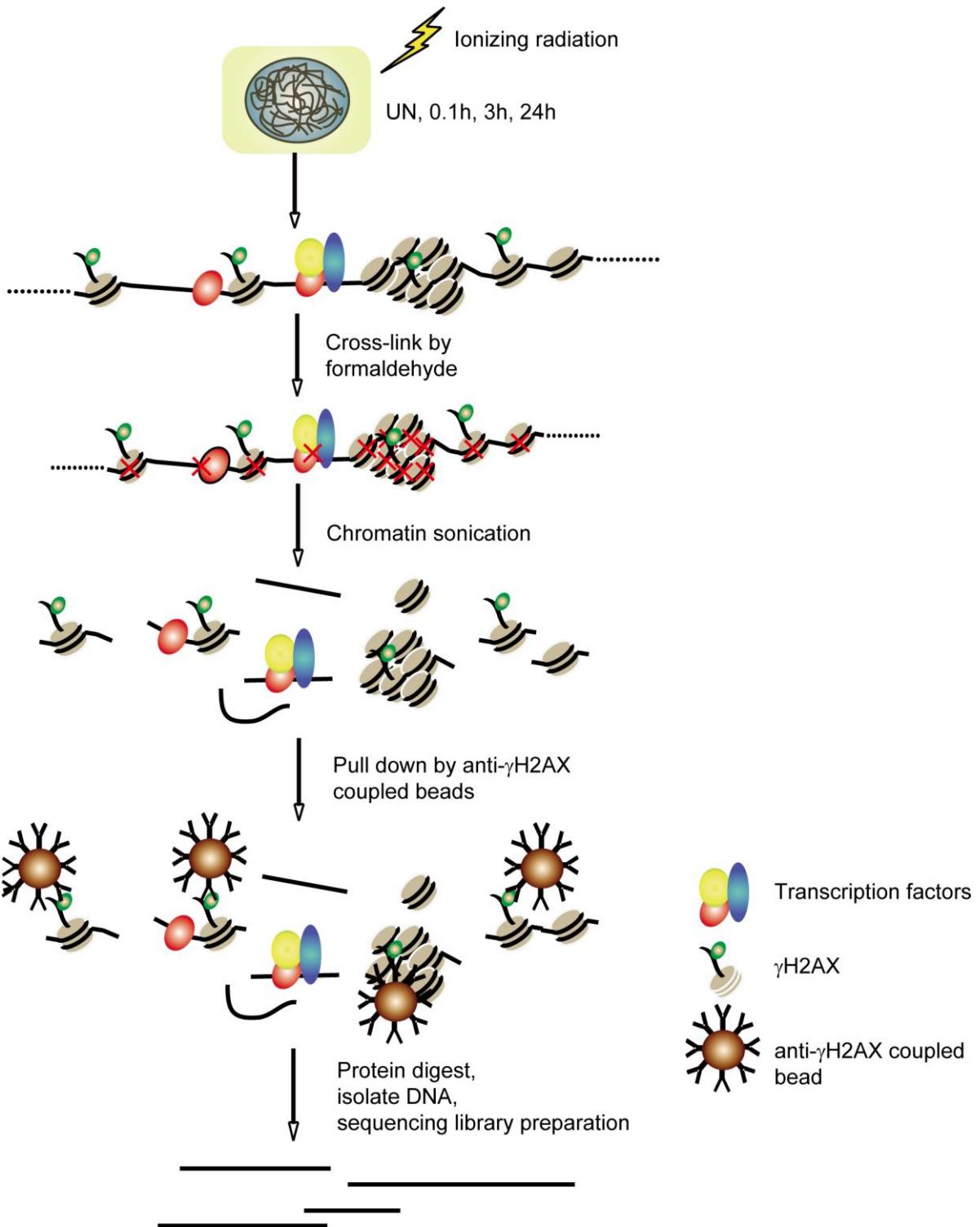


Figure 4. Schematic illustration of ChIP-seq procedure for γ H2AX. Upon IR, cross-link cells by formaldehyde, isolate nuclei, sonicate chromatin into fragments, pull down chromatin fragments with γ H2AX by anti- γ H2AX coupled magnetic beads, reverse cross-link and purify DNA fragment, prepare sequencing library.

Compared to microarray based ChIP-Chip, ChIP-seq has higher resolution due to the limited hybridisation targets in the arrays. Also ChIP-seq has a lower background noise. ChIP-Chip does not offer whole genome coverage and normally excludes repetitive regions and sites without complete sequence knowledge. However, ChIP-seq can not avoid the disadvantages

that the method is biased towards GC-rich content, higher requirement of depth of sequencing and cost (Park, 2009). Since double strand DNA is pulled down by antibodies for sequencing, ChIP-seq can't differentiate strand-specific binding, which is important for the identification of CPD sites. Therefore, to reveal strand specificity of CPD sites, we designed strand-specific damaged DNA immunoprecipitation followed by massively parallel DNA sequencing (ssDIP-seq).

FAIRE-seq

In 2007 Jason D. Liebs group firstly published the FAIRE protocol (Formaldehyde-Assisted Isolation of Regulatory Elements) for isolation of nucleosome-depleted DNA from human chromatin (Giresi et al., 2007). Massive parallel sequencing could then identify these regions. The procedure was termed FAIRE-seq and uses the fact that chromatin, which is highly covered by proteins after cross-linking is retained in the organic phase during phenol-chlorophorm extraction. Protein-free DNA stays in the aqueous phase and highly correlates to DNasel hypersensitive sites, transcriptional start sites, and active promoters, which are associated with regulatory activity (Nagy et al., 2003; Simon et al., 2012; Tu et al., 1996). The sequenced protein-free DNA indicates the genomic regions with open chromatin. The protocol for FAIRE-seq contains isolations of nuclei, shearing of DNA, phenol/chloroform extraction of DNA fragments, which separates free DNA and nucleosomes, recover and purify DNA from aqueous phase, followed by library preparation and sequencing (Giresi et al., 2007).

DNase-seq

Deoxyribonuclease I (DNase I; EC no.: 3.1.21.1) is a endonuclease, which digests nucleosome-depleted DNA, whereas DNA regions tightly wrapped around nucleosomes are more resistant. Sites for preferential digestion were termed DNase I hypersensitive (HS) sites (Boyle et al., 2008; Song and Crawford, 2010; Song et al., 2011; Tewari et al., 2012). These sites correspond to cis-regulatory elements such as promoters or enhancers of actively transcribed genes (Wu et al., 1979). Combination of DNase I digestion with next-generation sequencing was named DNase-seq. This method provides genome-wide detection of open chromatin regions and is depicted in the flow chart (Figure 5, left). In brief: cells are lysed to release nuclei. These are digested with a optimized concentration of DNase I. Then DNA ends are repaired and ligated to biotinylated linker 1, digest with the restriction endonucleases Mmel (cut 20 bp after TCCRAC), which is used to produce short fragments to which the linker 2 can be ligated. These double ligated fragments are then pulled down by streptavidin-coated dyna beads, amplify by PCR and sequenced (Song and Crawford, 2010).

Both approaches reveal the functional cis-regulatory elements and chromatin accessibility in the genome. However, DNase-seq provides higher resolution compared to FAIRE-seq due to the Mmel digestion. The 20 bp fragments, which are much smaller than DNA fragments used in FAIRE-seq, result in a higher resolution. The combination of DNase-seq and FAIRE-seq was used to map open chromatin, which covers 9% of the total human genome and tends to locate at or near transcription start sites and overlap with CTCF binding sites. These open chromatin showed conserved patterns among seven cell lines (Song et al., 2011). The procedures for DNase-seq and FAIRE-seq are illustrated in Figure 5.

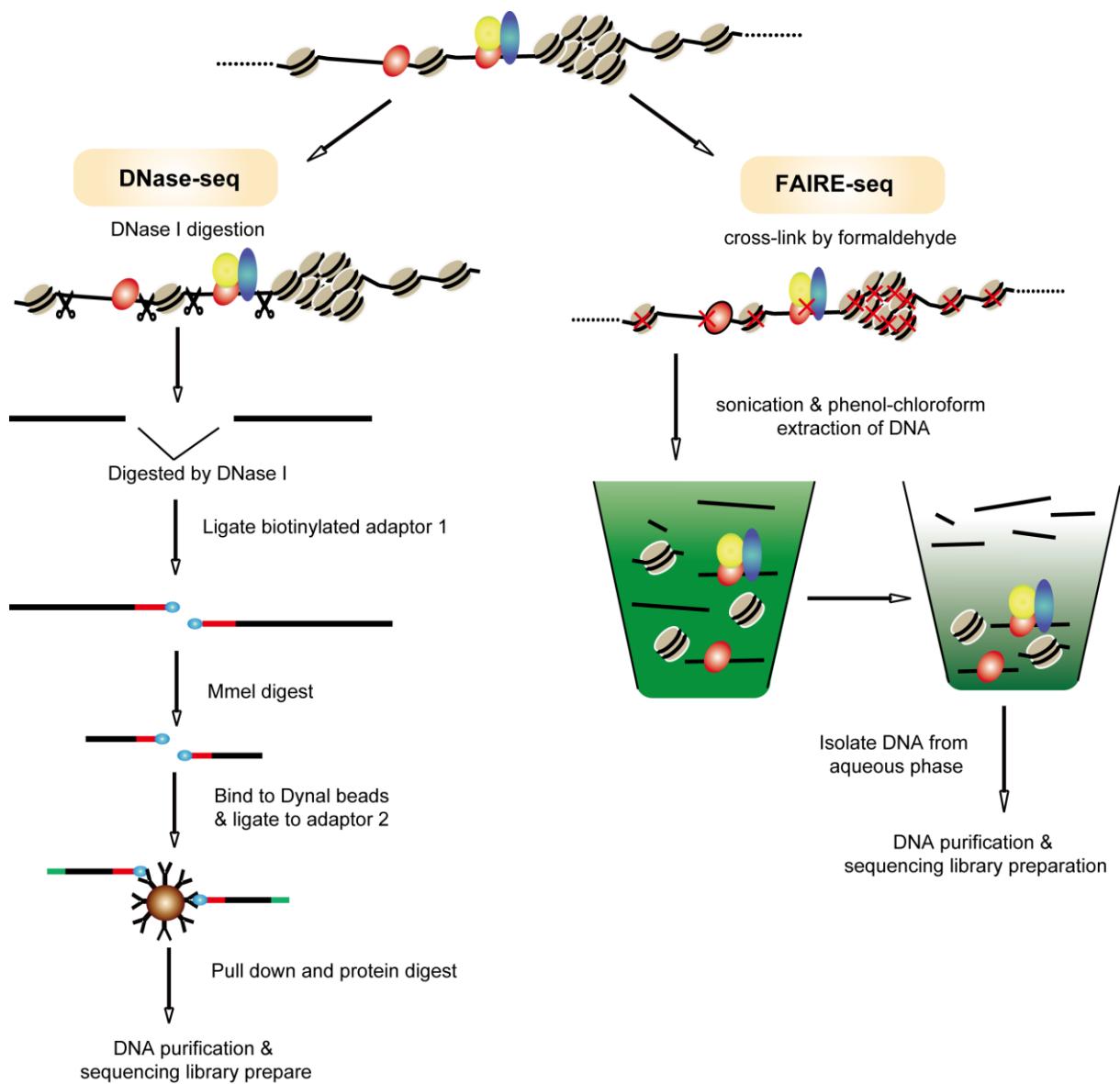


Figure 5. Flow chart of the DNase-seq and FAIRE-seq protocol. Left: DNase-seq; Right: FAIRE-seq, modified from Giresi et al., 2007 and Song et al., 2011. For details see text.

2. Aims of the work

In this study, I identified the effects of different kinds of radiation (ionizing radiation and UVC radiation) on histone modification patterns and chromatin states based on genome-wide analysis or within defined genomic loci. The aims of this study are to fulfil following purposes:

- I. Genome-wide investigation of double strand breaks distribution under physiological conditions and in response to IR in human carcinoma cells.
- II. To reveal the correlation of IR induced DNA lesions and genomic features in a genome-wide manner
- III. Comprehensive analysis of the DNA damage distribution as well as repair kinetics in response to UV exposure in a strand-specific way.
- IV. Analysis of CPD repair kinetics in different genetic backgrounds including a NER proficient human cell line HaCaT as well as a GG-NER deficient cell line XPC^{-/-} (cancer-prone) respectively.

3. Materials and Methods

Cell culture and irradiation

XPC^{-/-} cells XP4PA-SV-EB (Emmert et al., 2000) and CSB^{-/-} cells CS1AN (Mayne et al., 1986) were cultured in DMEM (4.5 g/L glucose, Biochrom AG) supplemented with 10% fetal calf serum (Biochrom AG). HaCaT cells were cultured in RPMI-1640 (Sigma) and kept in a humidified environment containing 5% CO₂ at 37°C. Cells were exposed to 12 J/m² UVC (single wavelength 254 nm) in UV crosslinker (CL-100 Ultraviolet Crosslinker, UVP), and collected before or at 0.1 or 24 hours post irradiation.

Immunofluorescence staining

HaCaT, XPC^{-/-}, CSB^{-/-} cells were exposed to a single dose of 12 J/m² UVC and post-incubated for 0.1, 1, 3 ,6, 24, 48, 72 and 96 hours after exposure. Then the medium was removed and cells were washed twice with PBS, and fixed in 3.7% FA PBS for 15 minutes at RT. Cells were washed with PBS three times and permeabilized with 0.5% Triton X100 in PBS for 20 minutes at RT. Coverslip with cells were placed on ice, washed 3 times with ice cold PBS, incubated in ice cold 88% MeOH for 5 minutes and washed 3 times with ice cold PBS. Cells were treated with RNase A in PBS (RNase 1 µL/mL PBS) for 30 minutes at 37°C, washed twice in PBS and were blocked with fish skin gelatine (FSG) in PBS (10 µL FSG in 1 mL PBS) for 30 minutes at 37°C. Cells were incubated with 1:200 diluted mouse anti-CPD antibody (KTM53, Cat. No. MC-062, Kamiya Biomedical Company) in 0.25 µL DNase, 10 µL 4% BSA PBS, 10 µL DNase 2X buffer for 90 minutes in 37°C. Then the slides were washed 3 times in PBS, 5 minutes each, incubated with 1:400 diluted secondary anti-mouse IgG antibody conjugated to Cy3 (Cat. No. 715-166-151, Jackson) in PBS for 60 minutes at 37°C. Finally the slides were washed 3 times in PBS 0.05% Tween 20, 5 minutes each, then in three times in PBS 5 minutes each. Cells were counterstained with 36 nM DAPI for 15 minutes at RT, washed once in PBS and mounted in Vectashield (Vectorlabs).

High content screening

A high content imager (Operetta, Perkin Elmer) was used to screen fluorescence intensity for different time points. The intensities were normalized to the unirradiated time point. The analysis pipeline used to filter out cells with bad quality (out of focus) was as followed: find nuclei by DAPI, select cells base on morphology, which filters out cells, which nucleus area is out of range of 180 to 800 µm² and roundness less than 0.7. After filtering, the average fluorescence intensity of DAPI and anti-CPD labelled Cy3 were calculated and outputted for each cell. The average and standard deviation of Cy3 intensity for all cells were calculated for each time points.

DNA extraction and quantification

Cells were harvested after 12 J/m² UVC exposure at defined time points and DNA was extracted according to the manual of the QiaAmp DNA Mini Kit (Cat. No. 51304, Qiagen). DNA was eluted in 100 µL EB buffer for each P100 plate and stored at -20°C. DNA concentrations were measured by Quant-iT™ dsDNA HS Assay Kit (Cat. No. Q32851, Invitrogen).

Slot blot for CPD

Nitrocellulose membranes (0.2 um pore, BIO-RAD) were saturated in ddH₂O for 5 minutes and soaked in 20X SSPE (3 M NaCl, 200 mM NaH₂PO₄, 20 mM EDTA and adjusted to pH7.4) buffer for 5minutes. Then the membranes were placed in the slotblot machine on top of a prewetted filter paper soaked in 20X SSPE. 200 µL of 20X SSPE for each well were loaded and pulled through the membrane slowly by vacuum. 150 ng of DNA were diluted in 100µl of TE buffer and a final concentration of 50 mM NaCl was adjusted. The samples were boiled for 5 minutes at 95°C, chilled immediately on ice for 5 minutes and 100 µl of 20X SSPE were added per sample. The denatured DNA was loaded in the blotter and incubated at room temperature for 15 minutes. Then the samples were slowly pulled through the membrane by vacuum. After the solution had passed through the membrane, 200 µl 20X SSPE were added to wash the membrane. The membrane was then taken out of the blotter and DNA was denatured by placing the membrane on top of a filter paper soaked in 0.4 M NaOH for 15 minutes. Then the membrane was rinsed briefly in 5X SSPE for 5 minutes, unspecific binding was blocked by incubating the membrane in PBS buffer with 0.2%Tween and 5% non-fat milk for 30 minutes at RT. Then the membrane was washed twice in 5X SSPE. CPDs were detected by incubation of the membrane in anti-CPD antibody (1:1000 diluted in PBS buffer with 0.2% Tween 4°C) over night at 4°C. The membrane was washed 3 times for 20 minutes each in PBS buffer with 0.2%Tween. Then the membrane was incubated in PBS buffer contain in anti-mouse IgG Alexa Fluor 647 (Cat. No. A-21236, Molecular Probes), diluted 1:5000 in PBS/0.2% Tween for 1 h at room temperature. Then the membrane was washed 3 times 20 minutes each in PBS buffer. The membrane was imaged using a STORM imager and the CPD density of each band was quantified by ImageJ.

Control DNA staining by methylene blue

DNA samples were treated as described above, but 1.5 µg of DNA were diluted in 300 µl of 20x SSPE buffer. This solution was loaded to the membrane as described above. Then the DNA was stained with 0.2% methylene blue solution for 30 minutes in RT. The membranes were washed in PBS and imaged with the Fuji Imager under white light.

CPD ssDIP-seq library preparation

DNA was sheared by ultrasound treatment (Bioruptor™ UCD200) with a treatment time of 60 min in total (10 min per cycle, 5 sec sonication and 15 sec stop). This resulted in DNA fragments of 150 ~ 550 bps fragments. Sequencing libraries were prepared by NEBNext ChIP-seq prep reagent set for Illumina (NEB #6200S) with a modified protocol. 10 µg DNA was used for the preparation of the sequencing libraries for each sample. End repair of DNA was done by a mixture of Klenow DNA polymerase (1 U/µl) in 10X phosphorylation reaction buffer, dNTP mix, T4 DNA polymerase, T4 polynucleotide kinase (50 µl final volume) for 30 minutes at 20°C. Then DNA was purified from the reaction by QIAquick PCR purification Kit (Cat. No. 28104, Qiagen) and eluted into 40 µl EB buffer. dA-tailing of the end repaired DNA was done by adding Klenow fragment exo⁻ (3' to 5'), dATP, NEB buffer 2 and incubation at 37°C for 60 minutes. DNA was purified by the MinElute PCR purification Kit (Cat. No. 28004, Qiagen) and eluted into 10 µl EB buffer. Illumina adaptors were ligated to the dA-tailed DNA by mixing eluted DNA with Quick T4 DNA ligase, 2x Quick ligation reaction buffer, adaptor oligos and incubation of the mix for 60 min at room temperature (RT). The DNA was isolated again following the MinElute protocol and eluted in 20 and size selected by agarose gel electrophoresis to fragments of 200~500 bps. DNA fragments were isolated from the agarose using the QIAquick gel Extraction Kit (Cat. No. 28704, Qiagen) and eluted into 50 µl EB buffer. Size selected DNA fragments were diluted to 400 µl TE buffer, followed by heat denaturation at 95°C for 10 minutes and were immediately chilled on ice for 10 minutes. IP was performed in 100 µl of 1x ChIP buffer (10X ChIP buffer: 100 mM pH 7.0 NaPOH (mix by 577 µl 1 M Na₂HPO₄ and 423 µl 1 M NaH₂PO₄), 1.4 M NaCl and 0.5% Triton X100), containing 10 µl anti-thymine dimer antibody (Clone KTM53, Cat. No. MC-062, Kamiya Biomedical Company) and denatured DNA. This mix was incubated overnight at 4°C with rolling. Magnetic beads were (anti-mouse IgG; Dynabeads® M-450, Prod. No. 110.01, DYNAL) prepared by taking 20 µl slurry per sample and wash it 3 times in 1 ml PBS+0.01% BSA. Then the beads were re-suspended in 1x IP buffer, and 50 µl beads slurry were added to the IP mix. Beads were incubated at 4°C for 4 hours, then precipitated using the magnetic particle concentrator (DYNAL MPC®-M, Prod. No. 120.09, DYNAL) to remove the unbound fraction. Beads were washed 3 times with 1 ml of 1x IP buffer followed by re-suspending the beads in 250 µl Proteinase K buffer (50 mM Tris-Cl pH8, 100 mM NaCl, 1 mM EDTA, 0.5% SDS). Then 10 µl of Protease K (20 mg/ml; Cat. No. P2308-500MG, Sigma-Aldrich) was added and incubated at 50°C for 3 hours with 800 rpm shaking. Immuno-enriched DNA was purified using the QIAquick PCR purification Kit and eluted in 25 µl of EB. DNA was amplified using the 2x Phusion® HF master mix (Cat. No. M0531S, NEB) with 18 cycles of PCR (30 sec initial denaturation at 98°C; 10 sec denaturation at 98°C; 30 sec annealing at 65°C; 30 sec extensions at 72°C; repeat 18 cycles; 5 minutes final extensions at 72°C and hold at

4°C). DNA was purified by MinElute PCR purification Kit and eluted into 10 µl EB buffer for sequencing. All the enzymes were provided in NEBNext ChIP-seq prep reagent set for Illumina (Cat. No. 6200S, NEB). All the sequencings were based on single-end sequencing and output 50bps reads result.

Mapping, RPKM and peak calling

To reveal the distribution of CPDs and γ H2AX, the human genome reference sequence was divided into 10,000 base pairs intervals, which results in total in 286,729 intervals. These sequencing result, all 50 bps reads, were filtered for redundancy so that multiple reads with the same sequence were treated as one unique read. Mapping non-redundant reads to the human reference genome (hg19, downloaded from UCSC browser) allowed up two mismatches by the SOAP2 software (Li et al., 2009). Due to the techniques, which include variations of DNA sonication, multiple steps of sequencing library preparation, the depth of sequencing and antibody efficiency and the number of total sequencing reads for each sample is varied. To be comparable among samples, for each 10 kbps interval, RPKM values (Mortazavi et al., 2008) were calculated: the number of reads uniquely mapped within the 10 kbps window divided by the total mapped reads in million and the length of window in kbps. Higher value of RPKM indicates more CPDs or γ H2AX per kbps and higher portion of cells with damage in this 10 kbps window region.

In order to reduce the sequencing bias to GC content and genome accessibility, DNA isolated from unirradiated sample (Input sample) were treated as input-control. To normalize by the input sample for each 10 kbps interval, normalized histone occupancy or CPD (also named relative CPD or γ H2AX abundance) are calculated as [(histone or CPD interval RPKM / input interval RPKM) - (histone or CPD average RPKM/ input average RPKM)] where “interval” is a 10 kb window (unless stated differently) and “average” is the overall genome RPKM average value of all intervals in each corresponding dataset. Positive values represent enrichment of the indicated sequence tag, whereas negative values represent under-representation. For the genome-wide plotting, data was smoothed by 25 intervals unless stated differently.

To provide higher resolution of CPD location, MACS version 1.4 (Zhang et al., 2008) was used to call CPD peaks for HaCaT and XPC^{-/-} cell lines, where the input sample is used as a control. The algorithm was used with the default setting and results were further filtered for peaks with fold enrichment above two for XPC^{-/-} to overcome the duplicated XPC^{-/-} input samples as control. For HaCaT cells, the same setting except the p-value was used, whereas the latter one was raised from default ($1e^{-05}$) to $1e^{-10}$, due to the single HaCaT input sample that was available as control.

4. Result

4.1. Genome-wide analysis of DNA double strand damage and response

Introduction

Double strand breaks (DSBs) are serious cytotoxic and mutagenic DNA lesions, which if inappropriately repaired can lead to severe consequence, for instance, permanent chromosome aberration, cell cycle arrest or apoptosis (Olive, 1998). This also induces mutations and genomic recombination (Bross et al., 2000) that can finally lead to cancer, cell aging or cell death (Hoeijmakers, 2001, 2007). Ionizing radiation (IR) is a common approach in tumour therapy. Therefore, it is important to understand how cancer cell respond to ionizing radiation and how they repair DSBs induced by ionizing radiation on a genome-wide manner. Upon induction of a DSB, the surrounding chromatin is phosphorylated at the serine 139 of histone H2AX (γ H2AX), a H2A variant that replaces H2A in 10-20% of all nucleosomes. The phosphorylation is accomplished by kinases of the PI3 family (ATM, ATR und DNA-PKcs) and appears very rapidly within several kilobase pairs (kbps) around the DSB (Chan et al., 2002). Phosphorylation reaches its maximum approximately 10 min after DSB induction and γ H2AX foci are believed to represent DSBs 1:1 (Rogakou et al., 1998). Therefore, it was used as an early marker to localize DNA double strand breaks (Kuo and Yang, 2008; Lobrich et al., 2010). Some previous experiments indicated the distinguishable distribution of H2AX between heterochromatin and euchromatin by immunofluorescence microscopy (Cowell et al., 2007). Some groups used artificial systems to enzymatically produce DSBs at known chromosomal positions for human cells (Iacovoni et al., 2010) or yeast (Kim et al., 2007). However, as the enzymatic activity cannot be turned off, such approaches can not monitor DNA repair over time and, as the sites are ectopically engineered, they also can hardly reflect the situation in the whole genome. Therefore, IR was used to induce DSBs, mimicking treatment received by patients.

With the development of next-generation sequencing techniques, which provide the possibility to sequencing whole genomes in one experiment. Combined chromatin immunoprecipitation with massively parallel DNA sequencing (ChIP-seq) can detect the binding sites of the proteins of interest in a genome-wide fashion. Therefore we utilized the ChIP-seq technique to survey the dynamics of the double strand break marker γ H2AX in a whole genome approach in human cells after 10 Gys of X-ray exposure at different time points. This study was performed in the hepatocellular carcinoma cell line HepG2 since annotation data for regulatory elements is included in the Encode project (2004; Qu and Fang, 2013) and the cell line was reported to have a stable karyotype. Here the genome-wide profiles of γ H2AX, H2AX and H3 through ChIP-seq were produced to comprehensively

understand the double strand break response and repair kinetics upon IR at early, middle and late time points.

Result

Experimental setup and validation

10 Gy X-ray was used to trigger the DNA damage response at early (0.5 hour), middle (3 hours) and late (24 hours) time points after radiation. DNA cross-linked with γ H2AX, H2AX, H3 was immunoprecipitated and sequenced from HepG2 cells. The pure DNA (without immuno-precipitation) of cells in physiological condition was treated as Input. To quantify the γ H2AX *in vivo*, confocal imaging of immunofluorescence stained cells was performed (Figure 6. A, B). The γ H2AX signal significantly increased at 0.5 hour and decreased overtime. It's interesting that, even without IR, γ H2AX signal still can be detected. This can partly be attributed to endogenous unrepaired DSBs (Vilenchik and Knudson, 2003) or replication associated damage. Therefore, the distribution of γ H2AX was analyzed in physiological condition (unirradiated) as well. The γ H2AX response was also tested by western blot. The intensity of γ H2AX normalized to H2AX showed the same repair tendency as by immune-fluorescence (Figure 6. C). This confirms that after 10 Gy X-ray HepG2 cells induce a 4-8 fold in γ H2AX levels at 0.5 hours that decreased to 1.5-2 fold at 24 hours. Due to the fact that γ H2AX is the phosphorylated form of H2AX, we validated the antibody specificity for γ H2AX that can only react with γ H2AX but not with H2AX. For this slot blot γ H2AX and H2AX peptides were utilized to test the monoclonal antibody γ H2AX and polyclonal antibody H2AX on phosphorylated and non-phosphorylated synthetic H2AX peptides. The γ H2AX antibody showed no cross-reactivity with the H2AX peptide. In contrast to the polyclonal H2AX antibody showed similar reactivity with phosphorylated and non-phosphorylated H2AX peptide (Figure 6. D), therefore, for the ChIP-seq data analysis, the relative γ H2AX abundance was normalized to H2AX and/or the Input. The experiments were performed by Alexander Rapp and Francesco Natale, results are presented here for completion of result.

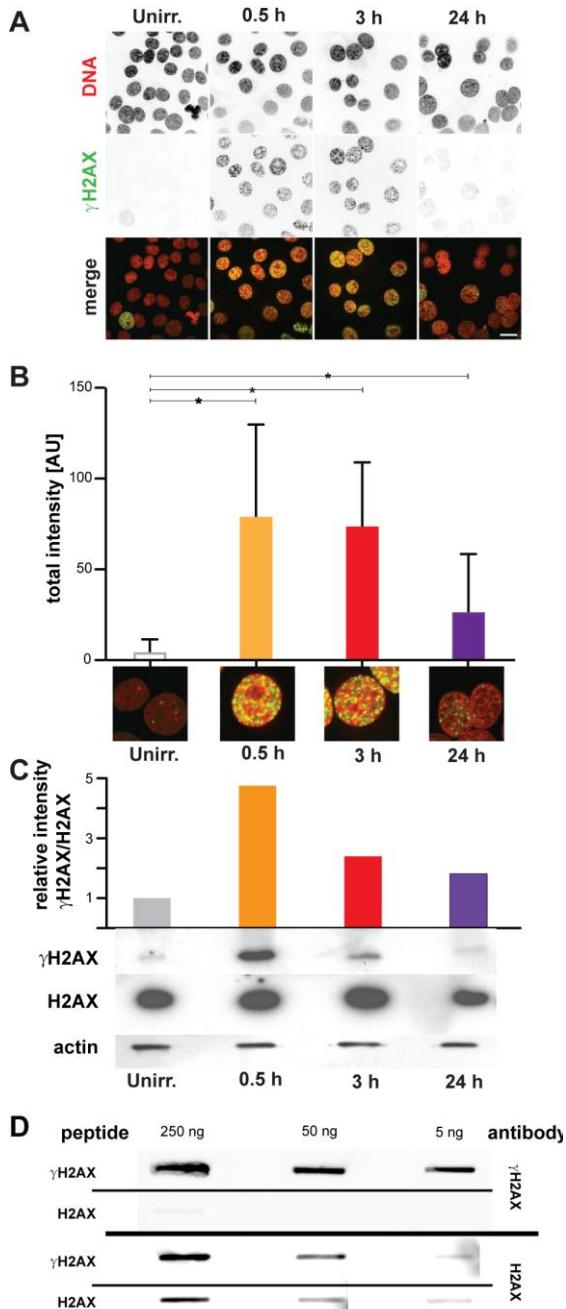


Figure 6. Characterization and validation of cellular system and experimental strategy. To characterize the DDR in terms of γH2AX, HepG2 cells were exposed to 10 Gy X-rays and were incubated as indicated. (A) Immunofluorescence confocal analysis of γH2AX foci (green) before and after exposure to IR. DNA counterstaining (red): propidium iodide. Bar: 20 μm. Total fluorescence intensity (arbitrary units) is presented in (B), with exemplary pictures shown below each bar. Error bars: SD. *: significantly different from the mean of unirradiated cells (one way ANOVA, p<0.0001). (C) Western blot analysis of γH2AX (top blot) and H2AX (middle blot) before and after exposure to IR. Loading control: β-actin (bottom blot). Intensity (relative ratio: γH2AX/H2AX) of the chemiluminescent signal is shown. (D) Slotblot analysis testing γH2AX and H2AX antibody specificity. No cross-reactivity of γH2AX antibody with H2AX peptide could be observed (top) with increasing amount of peptides used for immunization (data from Alexander Rapp and Francesco Natale).

Segmentation of the genome based on histone marks H3K9me3 and H3K36me3

As it is known that the human genome can be divided into the two components heterochromatin (HC) and euchromatin (EC) according to their different phenotype in which euchromatin is characterized with open chromatin structure, high GC level, active transcription and a more central location in the nucleus. In contrast to heterochromatin which has compact chromatin, low GC content, less or not active transcription and located in the periphery of the nucleus (Jost et al., 2012). Two histone modifications H3K9me3 and H3K36me3 were retrieved from published data (Ernst et al., 2011; Kolasinska-Zwierz et al.,

2009) for marking heterochromatin and euchromatin compartments (Kolasinska-Zwierz et al., 2009) (GEO accession number is GSM936090) respectively. Due to the broad distribution of γH2AX in the genome, peak calling could not be applied for the γH2AX distribution as it is used to find short distributions eg. transcription factor localisation. Therefore we divided the genome into 10 kbps intervals. Reads per kilobase pair per million (RPKM) value was utilized to quantify the two histone marks and further gate the distribution based on the principle that there is no overlapping genomic regions between heterochromatin and euchromatin. Genomic regions, which have high H3K9me3 and low H3K36me3 abundance were defined as heterochromatin and opposite with euchromatin. In total, 42,478 EC and 27,664 HC 10 kbps windows were assigned and thus ~25% genome was covered (Figure 7. A). For instance, the genomic region in chromosome 18 is mutually exclusive with H3K9me3 and H3K36me3 (Figure 7. B). The reason for less HC might be due to the pericentromeric regions which usually are HC, but can't be detected by uniquely mapping since the high percentage of repeat sequences prevents unique mapping.

To validate the segmentation result, several features of HC and EC were calculated for the segmented parts and all of them showed a significant difference. The GC content for these two components is significantly different, with higher GC percentage in EC (Figure 7. C) and lower in HC. The HepG2 expression data was retrieved from published data (accession number GSE30240, GEO). The gene expression value was assigned to each 10kbps interval which EC include 6,887 genes in contrast to 536 in HC. The expression level is significantly higher in EC bins than HC bins and the heatmap shows the higher gene expression level that correlates with higher H3K36me3 signals and the inverse correlation is true for the H3K9me3 bins (Figure 7. D, E). Since DNase-seq and FAIRE-seq data provide the information of chromatin accessibility, RPKM values for EC and HC were calculated from retrieved DNase-seq (GSM736639 from GEO) and FAIRE-seq data (GSM864354). EC and HC show distinguishable level of chromatin accessibility that chromatin is more open in EC than HC (Figure 7. F, G). And the track of H3K36me3 is close to DNase-seq (Figure 7. B second and third row), which corresponds to defined EC regions. All genomic features confirm that the genomic binary segmentation based on the H3K36me3 and H3K9me3 abundance is able to define HC and EC regions and represents their functional diversity. In the following analysis, the comparison of DSB signal distribution in HC and EC is based on this definition.

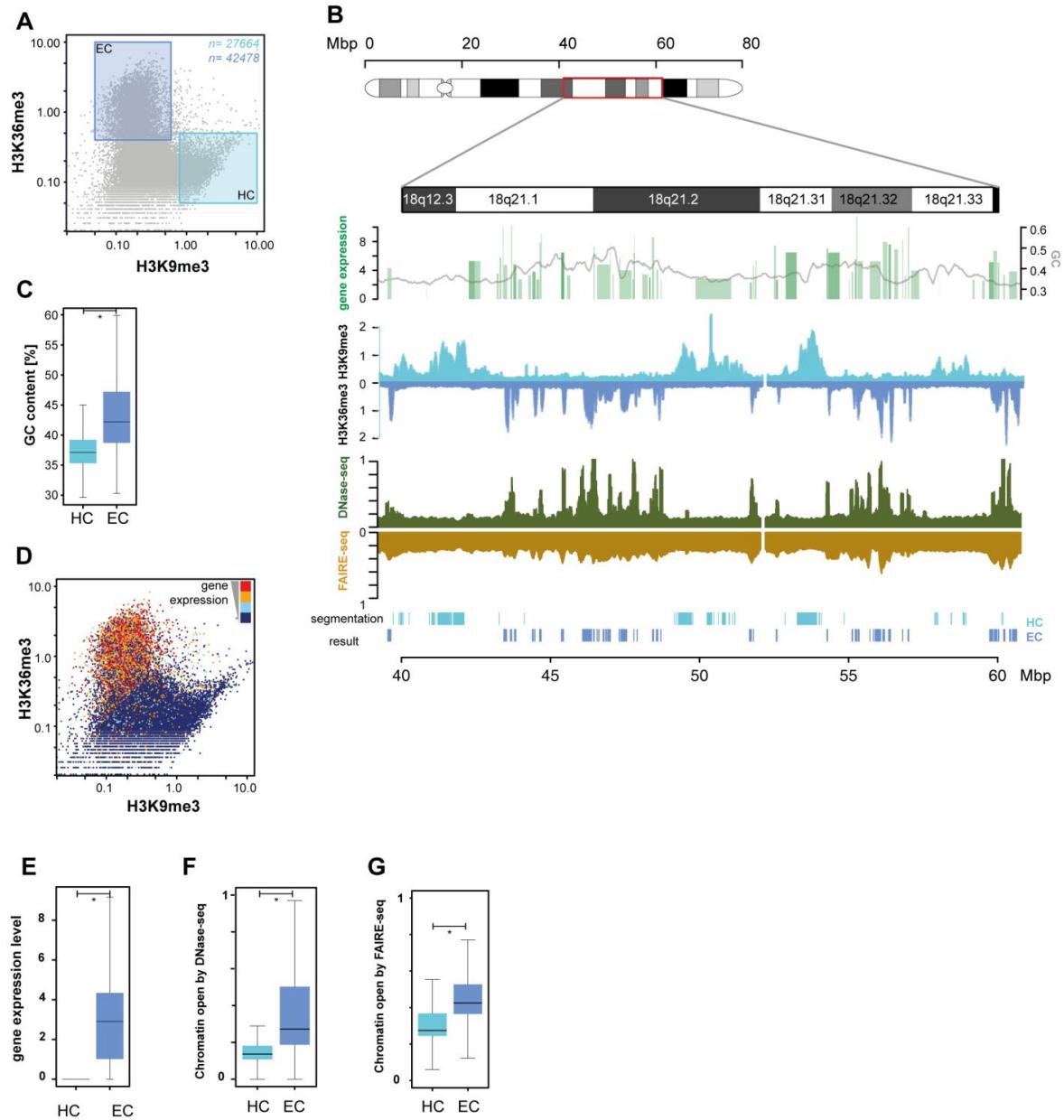


Figure 7. Establishment and validation of genome segmentation.(A) Gating of “euchromatic” (EC) and “heterochromatic” (HC) compartments. Normalized HepG2 ChIP-seq H3K9me3 and H3K36me3 levels (RPKM) in 10 kbps genomic intervals (grey dots) are compared. Intervals presenting high H3K36me3 (>0.6) and low H3K9me3 (<0.4) levels were assigned to EC compartment (dark blue) whereas intervals presenting low H3K36me3 (<0.5) and high H3K9me3 (>0.8) were assigned to HC compartment (light blue). The total number of intervals resulting from the gating is shown. (B) Exemplary 20 Mbp region on chromosome 18 showing the segmentation results. Top: GC content (grey line) and HepG2 expression data (green bars; increasing intensity depicts genes presenting multiple transcripts). Second row: H3K9me3 (dark grey) and H3K36me3 (light grey) ChIP-seq profiles. Third row: DNase-seq (dark green) and FAIRE-seq (dark yellow). Bottom: segmentation results of EC (dark blue) and HC (light blue). (C) GC content of EC and HC compartments. (D) HepG2 expression density scatter plot. Normalized H3K9me3 and H3K36me3 levels are presented as in (A). HepG2 gene expression levels were assigned to each genomic interval and are represented by a heat-map with increasing expression from blue to red. (E) Quantification of cumulative gene expression in EC and HC compartments is shown as boxplots. (F) RPKM of EC and HC from DNase-seq to measure chromatin openness. (G) RPKM of EC and HC from FAIRE-seq. All boxes and whiskers represent 25-75 percentile and 3 times the interquartile range (IQR), respectively. Significant differences were tested by the Wilcoxon-Mann-

Whitney test: * $p<2.2\times10^{-16}$

Under physiological condition, both H2AX and γ H2AX distribution follow the H3 track, however relative γ H2AX abundance to H2AX is inverted

Since γ H2AX can be detected even in non-irradiated cells, we were able to study γ H2AX distribution under physiological condition. The distribution of H2AX, H3 and γ H2AX were analyzed for each chromosome in physiological condition (Figure 8. A, B, D, G).

The profile of H2AX is similar to that of histone H3 in the whole chromosome. Both H3 and H2AX are found underrepresented in HC and overrepresented in EC. This also correlates to the GC content, meaning that H3 abundances as well as H2AX are enriched in regions with higher GC content and higher gene expression levels (Figure 8. I, J first column). This indicates that the higher density of histones in EC instead of higher compacted chromatin structure might be due to the shorter length of linker sequence between nucleosomes and thus be responsible for the increased histone abundance. Additionally the frequent eviction and reoccupation of histones in nucleosomes during transcription causes a higher density of nucleosome (Valouev et al., 2011). Overall γ H2AX follows the same distribution as H2AX and H3 when normalized to the input DNA. However the relative level of γ H2AX, which represents phosphorylated H2AX rely on the occupancy of H2AX as a basis. Therefore, the relative abundance of γ H2AX was additionally normalized to the H2AX density. The distribution of relative abundance of γ H2AX is opposite to the abundance of H2AX. Since less γ H2AX is found in EC regions compared to the overrepresented H2AX which results in an underrepresentation of the relative γ H2AX density in EC compared to HC. This implies that the γ H2AX signal is more or less resistant to be repaired in HC regions (Figure 8. F). This discrepancy between HC and EC is further elucidated in the following analysis.

Overall, the H2AX and γ H2AX track the distribution profile of core histones (H3) that is overrepresented in EC and underrepresented in HC. However, the γ H2AX abundance relative to H2AX is less in EC and higher in HC.

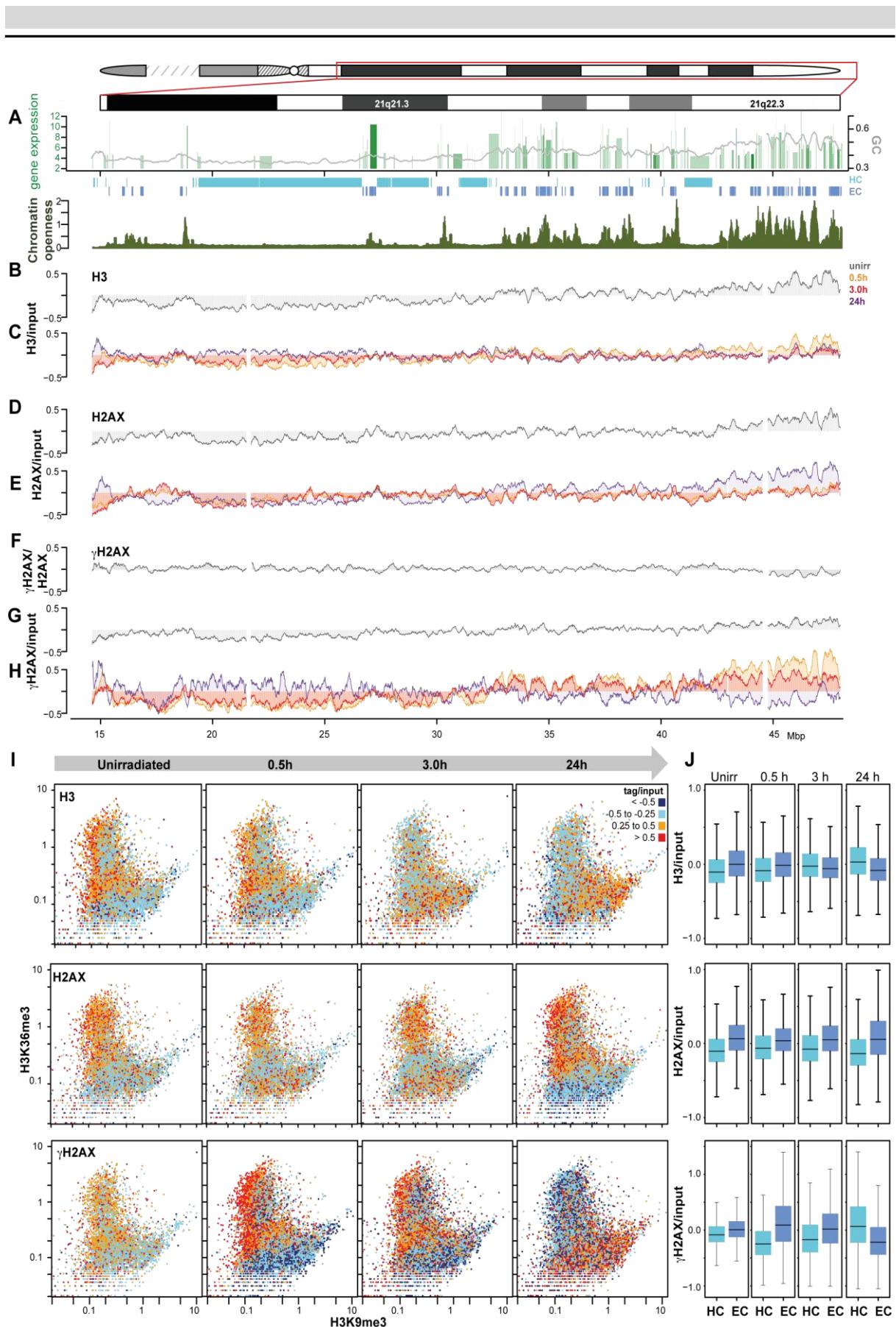


Figure 8. Chromatin dynamics before and during DDR. (A) Exemplary 35 Mbp region on chromosome 21 showing GC content (grey line), HepG2 expression data (green bars; as in Fig. 7. B) and gated EC and HC intervals. DNase-seq (dark green, A second row) (B-G) ChIP-seq histone

profiles under physiological conditions and post IR. The early (0.5h, orange) mid- (3.0h, red) and late (24h, purple) stages of DNA damage response (DDR) as well as the physiological state (Unirr, grey) are shown. (H) Genome-wide scatter plots (obtained as in Fig. 7. A) showing the dynamics of histone H3, H2AX, or γ H2AX occupancy after IR as well as under physiological conditions. Relative occupancy is presented as a heat-map, increasing from blue to red. (I) Quantification and comparison of histone H3, H2AX and γ H2AX occupancy in EC and HC compartments during all stages of DDR as well as under physiological conditions. All boxes and whiskers represent 25-75 percentiles and 3x IQR, respectively.

The dynamics of nucleosome components response to IR are varied

The chromatin structure of cells response dynamically to IR due to damage sensor and DNA repair process that requires remodelling of chromatin around the DNA damage sites. This also implies that the chromatin structure needs to be restored after repair (Chiolo et al., 2011; Soria et al., 2012). Therefore, the profiles of the two histones H2AX and H3 were produced in response to IR at 0.5 hour, 3 hours and 24 hours, which correspond to early, middle, late repair. The kinetic distribution of these two histones shows both temporal and spatial differences. At early repair time after IR, there is no significant difference in the distribution of H3 until at 3 hours the H3 signal decreases in EC regions and stays altered up to 24 hours (Figure 8. B, C). However, relative to the genome-wide average abundance, the H3 abundance is slightly overrepresented in HC regions 24 hours post IR (Figure 8. J, first row). For H2AX, the fast response to IR decreases its levels at 0.5 hour post IR in EC regions and further decrease at 3 hours post IR. And 24 hours after exposure, the abundance of H2AX is restored and shows the initial overrepresentation in EC regions. Whereas, there is no strong difference in HC regions between early and late repair time. On the other side, the slopes of a linear regression for the correlation of H3 and H2AX to the GC content are 2.5 and 1.3 respectively before IR. These slopes changes to 2.1 and -0.3 at 0.5 hour after IR and restores to 3.6 for H2AX, but is still -0.1 and thus anti-correlated to the GC content for H3 at 24 hours post IR (Figure 9. first and second row). This fast eviction of H2AX was also reported in a paper that described the fast release of H2AX from damaged chromatin (Ikura et al., 2007).

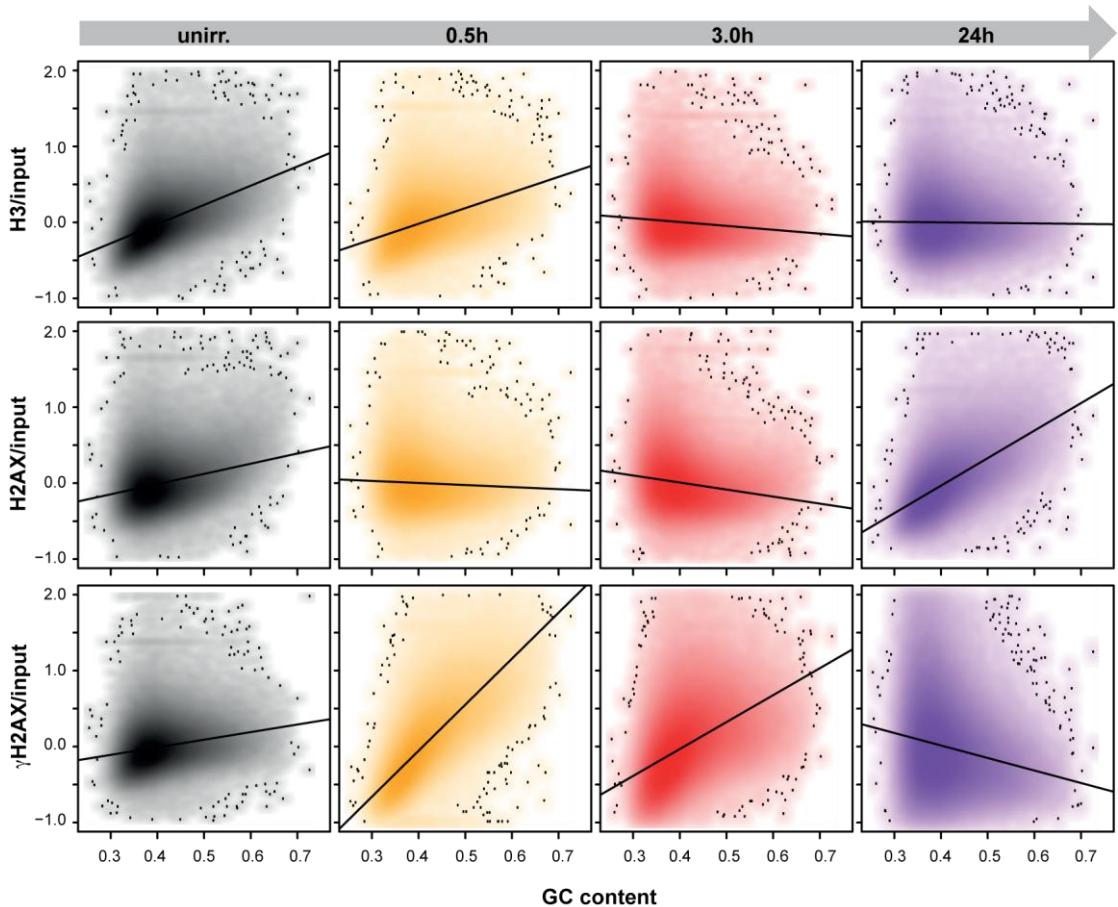


Figure 9. The response of H3, H2AX, γ H2AX with GC content. Distribution of GC content versus corresponding normalized H3, H2AX and γ H2AX abundance for each 10 kbps interval. The darker of colour the more signals are found in. The dots represent the 100 most dispersed signals from the average. Black lines: linear regressions..

The discrepancy of dynamic distribution between the two components of the nucleosome reveals that H2AX has different functional roles in response to IR. Histone H2AX is more related to the damage response and repair with fast eviction and restoration to keep an efficient sensor platform to trigger the DNA damage response. Histone H3 shows a delayed response to IR and might take longer time to restore. This also implies that the different components of a nucleosome respond to IR in independent ways and show variation in eviction and reloading.

After IR, γ H2AX is promptly formed in EC regions and decrease according to time, but the residual γ H2AX is overrepresented in HC regions at late repair stage

In order to analyze the phosphorylated H2AX signal response to IR, the γ H2AX profiles at three time points 0.5 hour, 3 hours, 24 hours were produced. The γ H2AX profiles show enhanced signal turnover in both EC and HC regions. After exposure to IR, γ H2AX is first overrepresented in EC regions at early stages of DNA repair (Figure 8. H) and the

discrepancy between EC and HC is enlarged compared to physiological conditions. During the repair process, damages in EC regions are preferentially repaired with the total amount of DSB decrease. Consequently 24 hours post IR, the residual DSBs are preferentially located in HC regions and less γ H2AX is left in EC regions. This tendency is further confirmed by generating a heatmap for EC and HC regions (third row of Figure 8. I). It becomes clear that higher levels of γ H2AX signal are found in H3K36me3 rich regions at early repair time point and regions with a relatively higher level of γ H2AX shift to H3K9me3 rich regions, corresponding to HC, at 24 hours. The box plots for HC and EC regions show the same change in distribution from early to late repair stages. Higher levels of γ H2AX in EC 0.5 hour post IR are observed. These levels then decrease until 24 hours post IR, when HC retains higher γ H2AX signal levels (third row of Figure 8 J). The corresponding slopes, calculated for the correlation of γ H2AX with the GC content shows same tendency: 6.1 at 0.5 hour, 3.5 at 3 hours and further decrease to -1.6 (Figure 9. third row).

The EC is characterized with higher GC content, higher gene expression level and more open chromatin accompanied by a higher density of H2AX and H3 under physiological condition. According to previous results, after IR in EC regions, H2AX promptly decrease to a level close to the overall genome-wide average. However at the same time, γ H2AX shows higher levels in HC indicating less H2AX but higher percentage of phosphorylated H2AX. This might be due to the chromatin structure of EC as it is more open to facilitate the kinase to access and phosphorylate histone H2AX around DSBs. And it is also possible that more DSBs are induced in EC regions. However in HC regions, to repair the DSBs, the chromatin needs to be firstly decondensed and DSB sites need to be relocated out of HC regions (Chiolo et al., 2011; Jakob et al., 2011). This process requires time and appropriate functional proteins that cascade to fulfil the steps preceding actual DNA repair that eventually result in delaying the DSB repair in HC. Therefore, γ H2AX can still be detected 24 hours later, which indicates the unrepaired DSBs. It's also confirmed that chromatin state influence the repair kinetic of DSB (Geuting et al., 2013).

In genic regions, the chromatin status determine the kinetics of nucleosome turnover and γ H2AX response to IR

To investigate the kinetics of histone turnover and γ H2AX response to IR in genic regions, metagene occupancy was calculated in a way that, for each transcript, the gene body region was divided into 100 bins according to the length of the gene (variable bin width) and extended to 10kbps upstream and downstream of gene body were also sub-divided into 100 bins each (fixed bin width). The RPKM for each transcript was calculated for each genic bin

and finally the average RPKM of all transcripts in each bin was calculated and normalized for H3, H2AX and γ H2AX at 0.5 hour, 3 hours and 24 hours. In addition we distinguished the transcripts again whether they are located in EC or HC and separately calculated the metagene profiles. The number of transcripts in these two components is greatly different since 11,590 transcripts were found in EC and 767 in HC respectively.

Compared EC with HC regions, γ H2AX showed similar levels of occupancy in the gene body and flanking regions under physiological conditions. However, for transcripts in EC compartments, higher levels of H2AX were found in the gene body and flanking regions compared to transcripts in HC (Figure 10. A, second column). This might indicate that H2AX is enriched in highly transcribed regions and hence is correlated to the transcription level.

Upon IR, histone H3 and H2AX show no significant difference between gene body and flanking regions at early time points. However γ H2AX is promptly formed in gene bodies and flanking regions of gene located in EC in contrast to transcripts in HC, where the phosphorylation is more restricted to the gene body, but not the flanking regions (Figure 10. A, third column). The difference can be attributed to the discrepancy of chromatin states between transcripts in HC and EC. The γ H2AX signal can't spread from the gene body to flanking regions due to the compact chromatin state in HC. Instead in EC, with open chromatin γ H2AX tends to spread to upstream and downstream flanking regions. This might be due to cohesin binding that antagonizes γ H2AX spreading and helps to isolate active genes (Caron et al., 2012). It suggests that the higher order of chromatin organisation influences the spreading behaviour of γ H2AX depending on the surrounding environment. In open chromatin it is easier to transfer γ H2AX signal to the neighbourhood than in condensed and closed chromatin, a fact that would also partially contribute to the explanation HC regions are delayed in repair. Three hours after IR, γ H2AX level decrease in both, the gene body and flanking regions, with the exception of an area around transcription start site (TSS) in HC regions. This reduction confirms the fast repair process along genic regions. However, at 24 hours, transcripts in HC show higher level of H3 in gene body and flanking regions than transcripts in EC. And this discrepancy indicates that H3 was not completely restored even after 24 hours post IR. The rearrangement of H3 within and around gene body is anti-correlated to the H2AX level, which is higher in transcripts of EC and lower in transcripts of HC. At the same time, γ H2AX is dephosphorylated or replaced by un-phosphorylated H2AX in transcripts of EC but is retained in transcripts of HC. This means the delayed or unrepaired DSB in HC compartments, which are silent genic regions.

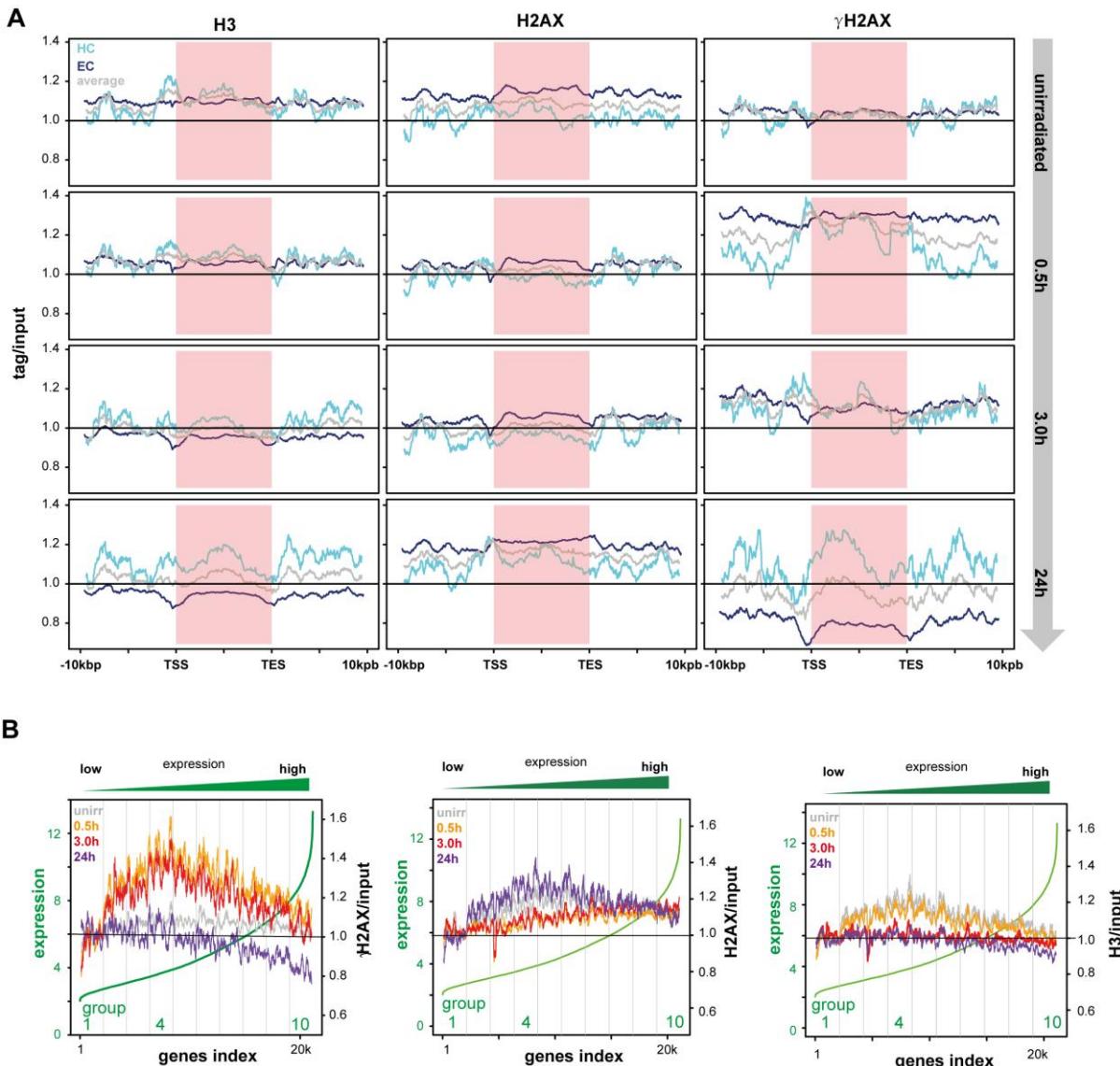


Figure 10. Response of genic and non-genic elements to IR. (A) Metagene profiles showing normalized H3, H2AX or γ H2AX occupancy over gene bodies and flanking regions in EC (dark blue) or HC (light blue) compartments. Grey line shows the average occupancy value. Smoothing: 25 (EC) and 35 (HC) intervals. TSS: transcription start site; TES: transcription end site. (B) Correlation between γ H2AX (left), H2AX (middle), H3 (right) levels and gene expression before and after IR. HepG2 genes are sorted on the x-axis with increasing expression (green curve). γ H2AX, H2AX, H3 levels, normalized to input, calculated over each gene body (from TSS to TES) are shown for early (0.5h, orange) middle (3h, red) and late (24h, purple) stages of DDR and physiological levels (unirradiate, grey). 1, 4-5, 10: groups of genes presenting expression values falling into 10th, 31st to 50th and 91st to 100th percentile of HepG2 expression distribution.

The expression level of transcript is not linearly correlated with γ H2AX, H2AX and H3 abundance

To reveal the relationship between gene expression and γ H2AX levels at different time points after IR exposure, all transcripts (21,325) were divided into ten groups according to their expression from low (class 1) to high (class 10) and the normalized γ H2AX in gene bodies

were calculated for each transcript. Before IR, there is no significant difference among the 10 groups of transcripts except slightly decreased γ H2AX in low and high groups (Figure 10. B. Left, grey line). After IR, at early repair time of 0.5 hours and middle time 3 hours, γ H2AX signal forms in gene body regions but it is interesting to note that the γ H2AX level is not linear correlated to the expression level even the low expression transcripts with the lowest level of γ H2AX. High expression transcripts also have low levels of γ H2AX. However, the highest level of γ H2AX was found in genes with middle expression level transcripts (class 4), while higher expression genes showed decreasing levels of γ H2AX. At 24 hours after IR, lower levels of γ H2AX are found in all groups of transcripts, with the highly expressed transcripts retaining the lowest levels of γ H2AX (Figure 10. B. Left).

The fact that highly expressed transcripts show counteraction of γ H2AX formation in gene bodies might be due to two reasons: frequently transcribed genes are usually located in EC regions that have overrepresented levels of H2AX and loading and expeditiously decreased in response to IR. It also might be attributed to competition of the transcription machinery, which is frequently found in highly expressing genes. For instance it was reported that the binding of polymerase II inhibits endogenous γ H2AX formation in chromatin domains (Iacovoni et al., 2010). The lowest level of γ H2AX is found in low expression genes that can be explained by the condensed chromatin state inhibiting the spreading of γ H2AX and thus lowering the amount of H2AX. However, the medium expression level genes exhibit the highest level of γ H2AX since there is no condensed chromatin and frequently bound transcriptional complex to repress the formation of γ H2AX. Therefore, chromatin state and transcription play complex roles in the formation of γ H2AX along the gene body whereas how they influence the spreading of γ H2AX still needs to be further elucidated.

Under physiological condition, highly expressed genes show slightly elevated levels of H2AX and lower level of H3 but medially expressed genes retain the highest level of H3 and slightly elevated levels of H2AX. After exposure to IR, H2AX promptly decreases in medially expressed genes but not in highly expressed genes. At late repair time, all genes are restored the H2AX levels as before irradiation. Whereas, H3 shows a delayed response upon IR and exhibits no changes after 0.5 hours. From 3h to 24h after IR, H3 levels decrease in a majority of genes whereas the genes with higher expression levels show less H3 at 24h which is different compared to physiological condition suggesting that H3 can not be restored for a long time upon IR or might be replaced by alternatives such as H3.3 that can not be detected by the H3 antibody (Figure 10. B. middle and right).

Genome-wide correlation of γ H2AX and genomic features after exposure to IR

To uncover the relationship of γ H2AX signal distribution and different genomic features, the genome-wide abundance of 90 different genomic features were retrieved and calculated for each 10kbps intervals. Most data was generated in HepG2 cells, but not all. The sources and details for all genomic features are listed in Table 1. The Spearman's rho correlation for each genomic feature with γ H2AX at each time point was calculated and plotted into a heat map (Figure 11. A). Positive correlation is represented as green box and anti-correlation is shown as red boxes. Agglomerative hierarchical clustering was used to cluster genomic features based on pairwise spearman's rank correlation between all features and the distance of $1 - |\rho|$ for dissimilarity is displayed in the dendrogram (Figure 11. B).

Both the correlation matrix (heat map) and the clustering of genomic features show the positive correlation of the γ H2AX response to euchromatic features at early time points and anti-correlation at late time points. For instance, H3K36me3, Pol II, GC content, early replication timing and H3K4me2 are positively correlated to the γ H2AX abundance at 0.5 and 3 hours but are anti-correlated at 24 hours. Conversely for the heterochromatin related features, such as AT content, H3K9me3, Topo CAT and G banding, which are anti-correlated with the γ H2AX signal distribution at early and middle time points, but they become positively correlated at later time points. These findings further confirm previous result that γ H2AX is overrepresented in EC regions at early time point and residual γ H2AX preferentially in HC regions.

It is interesting that the γ H2AX distribution is closely correlated to the distribution of the tumor suppressor BRCA1 binding sites at early time points (Figure 11), but this correlation is lost at later time points. BRCA1 is a genome guardian maintaining genomic stability that co-localize with γ H2AX at the arrested replication fork in S phase (Paull et al., 2000), which indicates that this co-localization occurs at early time points in response to IR. Rad21 shows a similar relationship with the γ H2AX distribution at the different time points (Bauerschmidt et al., 2010), which we also find in the correlation analysis. MBD4 is also positively correlated at early time and anti-correlated at late time point which suggests MBD4 plays a potential role in DNA repair (Bellacosa, 2001).

However for known common fragile sites, recombination sites, evolutionary break points regions (EBR) and break points hotspot, there is neither strong positive nor negative correlation with the γ H2AX distribution at the different time points. This might be due to the discrepancy of data resolution or different cellular backgrounds. The relationship of γ H2AX distribution and recombination sites needs to be further validated.

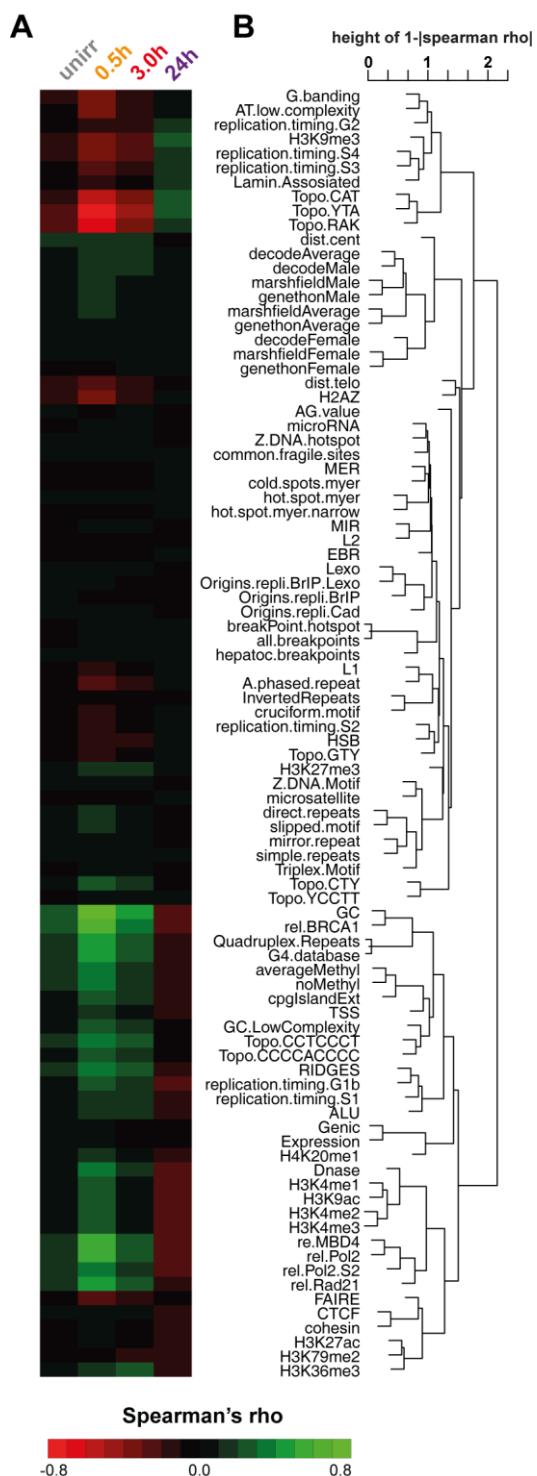


Figure 11. Correlation of γ H2AX distribution during DDR to genomic features.
(A) Heatmap of the correlation (spearman rho) of the distribution of 90 genomic features to the γ H2AX distribution of all four time-points. The genomic features are ordered according to an agglomerative hierarchical clustering (B), based on the dissimilarities of the features among each others.

Feature	Cell Type	Type of Data	Data Source / Reference
General Features			
G-banding	human	% shading	UCSC Genome Brower
Distance to the telomere	Hg19	Distance in bp	UCSC Genome Brower
Distance to the centromere	Hg19	Distance in bp	UCSC Genome Brower
Purine percent	Hg19	Percentage	Calculation by ourself
GC content	Hg19	Percentage	Calculation by ourself
AG content	Hg19	Percentage	Calculation by ourself
DNase	HepG2	DNase-seq	GSM816662

FAIRE	HepG2	FAIRE-seq	GSM864354
CpG island	Hg19	count	UCSC Genome Brower
Transcription			
No of Transcription start sites	Hg19	count	UCSC Genome Brower
miRNA	human	count	miRBase (Griffiths-Jones et al., 2008)
TSS microRNA	Hg19	Distance in bp count	UCSC Genome Brower miRBase: tools for microRNA genomics
Expression rel.Pol2	HepG2	ChIP-seq	GSM822284
rel.Pol2_S2	HepG2	ChIP-seq	GSM935543
RIDGES Genic region	human	coordinates	http://r2.amc.nl
	Hg19	count	UCSC Genome Brower
DNA Methylation			
Cumulative Methylation	HepG2	Microarray	GSM999338
no of Methylation sites	HepG2	Microarray count	GSM999338
relative MBD4 abundance	HepG2	ChIP-seq	GSM1010740
Histone Modifications			
H2A.Z	HepG2	ChIP-seq	GSM733774 (Consortium et al., 2012)
H3K4me1	HepG2	ChIP-seq	GSM798321 (Consortium et al., 2012)
H3K36me3	HepG2	ChIP-seq	GSM733685 (Consortium et al., 2012)
H3K9me3	HepG2	ChIP-seq	GSM1003519 (Consortium et al., 2012)
H3K14Ac	IMR90	ChIP-seq	GSM521881 (Consortium et al., 2012)
H3K79me2	HepG2	ChIP-seq	GSM733641 (Consortium et al., 2012)
H3K27ac	HepG2	ChIP-seq	(Consortium et al., 2012)
H3K27me3	HepG2	ChIP-seq	(Consortium et al., 2012)
H3K4me2	HepG2	ChIP-seq	(Consortium et al., 2012)
H3K4me3	HepG2	ChIP-seq	(Consortium et al., 2012)
H3K9ac	HepG2	ChIP-seq	(Consortium et al., 2012)
H4K20me1	HepG2	ChIP-seq	(Consortium et al., 2012)
DNA Sequence Elements			
Alu repeats	human	count	RepeatMasker, (Smit, 1996-2010)
MIR repeats	human	count	RepeatMasker, (Smit, 1996-2010)
LINE1 repeats	human	count	RepeatMasker, (Smit, 1996-2010)
LINE2 repeats	human	count	RepeatMasker, (Smit, 1996-2010)
MER repeats	human	count	RepeatMasker, (Smit, 1996-2010)
AT_Low_complexity	human	count	RepeatMasker, (Smit, 1996-2010)
GC_Low_complexity	human	count	RepeatMasker, (Smit, 1996-2010)
simpleRepeat	human	count	RepeatMasker, (Smit, 1996-2010)
G-Quadruplex	human	count	RepeatMasker, (Smit, 1996-2010)
Forming_Repeat			RepeatMasker, (Smit, 1996-2010)
Z-DNA_Motif	human	count	(Cer et al., 2011)
Z.DNA.hotspot	human	count	(Cer et al., 2011)
Inverted_Repeat	human	count	(Cer et al., 2011)
Cruciform_Motif	human	count	(Cer et al., 2011)
Direct_Repeat	human	count	(Cer et al., 2011)

Slipped_Motif	human	count	(Cer et al., 2011)	
Mirror_Repeat	human	count	(Cer et al., 2011)	
Triplex_Motif	human	count	(Cer et al., 2011)	
A-Phased_Repeat	human	count	(Cer et al., 2011)	
microsatellite	human	count	RepeatMasker, (Smit, 1996-2010)	
DNA Replication				
Replication timing	GM12801	RepliSeq	GSM923440 (Hansen et al., 2010)	
Origins of replication by lambda exonuclease digestion (Lexo)	HeLa	genomic array	(Karnani et al., 2010)	
Origins of replication by anti-bromodeoxyuridine IP (BrIP)	HeLa	genomic array	(Karnani et al., 2010)	
Origins of replication by common anti-bromodeoxyuridine IP and lambda exonuclease digestion (Lexo+BrIP)	HeLa	genomic array	(Karnani et al., 2010)	
Origins of replication (Ori Cadoret)	HeLa	genomic array	(Cadoret et al., 2008)	
Topoisomerase (CAT)	motif	Hg19	<i>Density</i>	(Arlt and Glover, 2010)
Topoisomerase (CTY)	motif	Hg19	<i>Density</i>	(Arlt and Glover, 2010)
Topoisomerase (GTY)	motif	Hg19	<i>Density</i>	(Arlt and Glover, 2010)
Topoisomerase (RAK)	motif	Hg19	<i>Density</i>	(Arlt and Glover, 2010)
Topoisomerase (YCCTT)	motif	Hg19	<i>Density</i>	(Arlt and Glover, 2010)
Topoisomerase (YTA)	motif	Hg19	<i>Density</i>	(Arlt and Glover, 2010)
Breakpoint and Recombination Hotspots				
Hotspots of recombination motif (CCCCACCCC)	Hg19	count	(Myers et al., 2008)	
Hotspots of recombination motif (CCTCCCT)	Hg19	count	(Myers et al., 2008)	
Evolutionary breakpoint regions (EBR)	Hg19	count	(Larkin et al., 2009)	
Homologous synteny blocks (HSB)	synteny	Hg19	count	(Larkin et al., 2009)
hotspot myer		Hg19	count	(Myers et al., 2008)
hotspot myer narrow		Hg19	count	(Myers et al., 2008)
cold spots myer		Hg19	count	(Myers et al., 2008)
common fragile sites		human	coverage	(Fungtammasan et al., 2012)
hepatocellular breakpoints		human	count	(Beroukhim et al., 2010)
Breakpoints hotspots		human	count	(Beroukhim et al., 2010)
all breakpoints		human	count	(Beroukhim et al., 2010)
Decode Average		human	count	UCSC Genome Browser (Kong et al., 2002)
Decode Female		human	count	UCSC Genome Browser (Kong et al., 2002)
Decode Male		human	count	UCSC Genome Browser (Kong et al., 2002)
Marshfield Average		human	count	UCSC Genome Browser (Broman et al., 1998)
Marshfield Female		human	count	UCSC Genome Browser (Broman et al., 1998)

Marshfield Male	human	count	UCSC Genome Browser (Broman et al., 1998)
Genethon Average	human	count	UCSC Genome Browser (Dib et al., 1996)
Genethon Female	human	count	UCSC Genome Browser (Dib et al., 1996)
Genethon Male	human	count	UCSC Genome Browser (Dib et al., 1996)
DNA Binding Factors			
SMC3 (cohesin)	HepG2	ChIP-seq	GSM935542
Insulator CTCF	HepG2	ChIP-seq	GSM733645
Lamina Associated Domain	Tig3ET	Coverage	(Guelen et al., 2008)
rel.BRCA1	HepG2	ChIP-seq	(Consortium et al., 2012)
rel.Rad21	HepG2	ChIP-seq	(Consortium et al., 2012)

Table 1: Overviews of genomic features

γ H2AX response to IR in repetitive elements is not linear to GC content

In order to evaluate the γ H2AX levels in repetitive element regions, four well characterized repetitive elements (Alu, LINE, Satellite and LTR) were selected for analysis. Alu elements are usually associated with high GC content and gene-rich regions, in contrast to LINE elements and Satellites that are associated with low GC content and gene-poor regions. The preferential location of LINE elements are in intergenic regions or introns, however, Satellites are located in peri-telomeric and pericentromeric regions (de Koning et al., 2011). Since only uniquely mapped reads were counted, a majority of repetitive elements can't be mapped. To address this problem, multiple mapping results were retrieved again and if a read is mapped to different genomic regions (e.g. Alu, LINE, LTR) then this read was discarded. But if all these regions belong into same class of repetitive element (e.g. Alu, Alu, Alu), then this read was counted into the respective class of repetitive element (e.g. Alu). The fraction relative to the number of total reads mapped in all four classes and the genomic coverage was calculated.

Compared to the physiological condition, Alu elements are overrepresented at early and middle repair time points and underrepresented at later time points. Compared to Alu, LINE and Satellite repeats show exactly opposite tendencies. The LTR elements show no significant change according to time (Figure 12. A). In order to elucidate whether these differences could be attributed to the average GC content of each repetitive element, we calculated the genome wide mean RPKM values in 10 kbps windows (sorted by the GC content; solid lines in Figure 12.B) for each time point and compared them to the RPKMs found for each repetitive element based on the corresponding GC content (symbols in Figure 12.B). It can be seen that LINEs show a behaviour that is in agreement with the genome-wide GC adjusted behaviour. While in contrast Satellites do not show behaviour according to

their GC content. Satellite repeats show lower γ H2AX occupancy than its corresponding average GC content in early, middle repair time points and physiological condition but higher γ H2AX in late time. LTRs do not show a significant change of γ H2AX occupancy during the DDR response, while in contrast Alu elements again behave according to their average GC content. This might be due to the location of different repetitive elements that Satellites usually distribute in highly repetitive regions (peri-telomeric and pericentromeric regions), which are constituted of highly condensed chromatin structure. Even if DSBs are induced within Satellites, γ H2AX can't promptly spread into these regions and requires longer time to repair. Whereas Alu is usually located in GC and gene rich regions with open chromatin structure that allows γ H2AX to spread and dismiss easily.

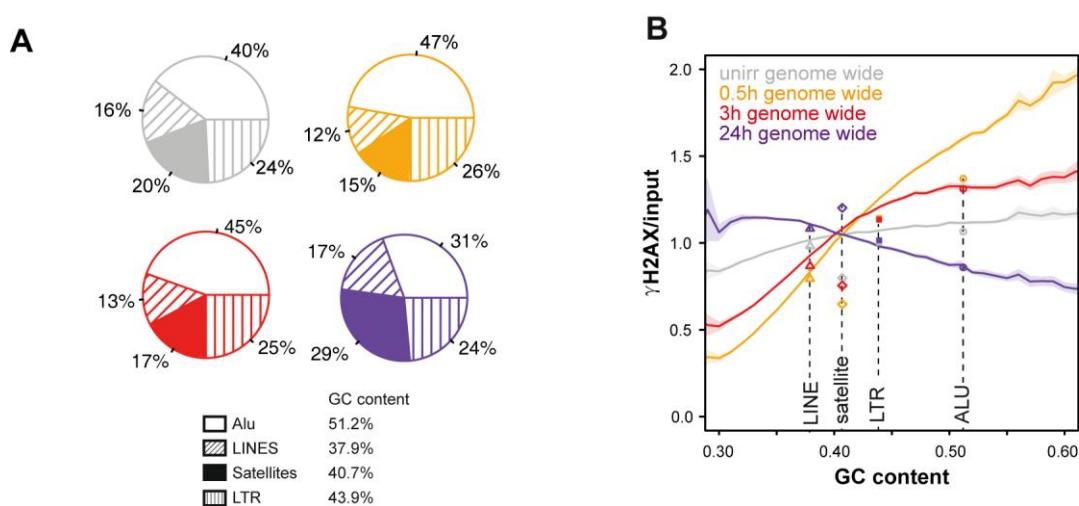


Figure 12. Correlation of γ H2AX distribution with repetitive elements. (A) Distribution of sequence tags in repetitive elements for 4 time points. (B) DDR for repetitive elements in the context of their GC value. The lines represent the genome wide distribution of reads sorted with increasing GC value. The repetitive elements are shown at their GC position. LINE and Alu elements show a DDR according to their GC content, whereas LTRs and satellites do not.

4.2. Genome-wide analysis of strand-specific cyclobutane pyrimidine dimer induction in NER repair proficient and deficient cell lines

Introduction

An important type of DNA damage is cyclobutane pyrimidine dimer (CPD), which is formed between the C5 and C6 atoms of two adjacent pyrimidine bases. The major source of CPD is short wavelength ultraviolet light (UV) (Tornaletti and Pfeifer, 1996). Deficiencies in CPD repair are associated to serious diseases including Xeroderma pigmentosum (XP), Cockayne syndrome (CS) and skin cancer (Garinis et al., 2005; You et al., 2001). Another type of UV-associated DNA lesion is the 6-4 photo product (6-4 PP), which is formed between the C6 and C4 of two adjacent pyrimidine bases and preferentially occurs at TC and CC sequence (Hauser et al., 1986).

Previous studies revealed that unrepaired CPD lesions modulate vital cellular functions, including induction of cellular apoptosis, blocking of DNA replication *in vivo* and cell cycle arrest (Donahue et al., 1994; Lo et al., 2005; Tornaletti et al., 1999). The induction of CPD lesions in the genome is non-homogenous and the repair of CPD is highly selective for transcribed active genes (Hanawalt, 1991). For the silent genes or non-transcribed domains, CPD lesions persist and remain unrepaired for a long time (van Hoffen et al., 1993). In mammalian cells, UV induced CPDs are believed to be one major source of mutagenic lesions due to the high abundance and slow repair. However, the mutations contributed by 6-4 PP can-not be completely excluded as well, although these lesions are induced at a significantly reduced level and are repaired with a faster kinetic (Pfeifer et al., 2005; Protic-Sabljic et al., 1986). The hallmark of mutation induced by UVB and UVC is C to T transition at pyrimidine sequences (Pfeifer et al., 2005; You et al., 2001) and the fingerprint mutations induced by UVA also include G to T transitions (Kappes et al., 2006). Due to the atmospheric ozone, UVC is filtered out completely and thus is not present in the natural environment. Most of the UVA irradiation and a small portion of UVB can reach the earth's surface. However, the damage and depleting of the ozone layer by emission of chemicals from industries in recent years allows more UV radiation, especially UVC, to reach the earth and the skin of humans. Hence, UV induced DNA damage is gaining more attentions.

Another effect besides the induction of mutations in coding genes is the alteration of expression levels. One mechanism how CPDs can alter the expression level of genes is that binding of transcription factors to the corresponding promoter is impaired by persisting CPDs induced in the promoter sequence. This can lead to a reduction of binding ranging from 11 to 60-fold for several transcription factors E2F, p53, NfkB, Ap-1 and NF-Y *in vitro* (Tommasi et

al., 1996). The inhibition of binding reveals a potential way to influence the gene expression and transcription. Furthermore, unrepaired CPDs arrest the RNA polymerase II in the transcribed strand of genes during the transcription process (Tornaletti et al., 1999). A large amount of CPDs arrests the cell cycle in order to repair the damage or induce apoptosis. And unrepaired CPDs can also partially prevent the DNA replication by blocking DNA polymerase I (Setlow et al., 1963; Taylor and O'Day, 1990). The inhibition of replication is different in leading strand and lagging strand with TT pyrimidine dimer (Svoboda and Vos, 1995) *in vitro*. Another group reported the equal replication efficiency on the lagging strand and 20% decreased efficiency on the leading strand (Carty et al., 1996). Compared to CPDs, 6-4 PPs have a higher ability to block replication and transcription due to a stronger change of the structural conformation of DNA (Batista et al., 2009).

Since CPD occurs at adjacent pyrimidine dimers in a single strand, the repair preference exists between transcribed and non-transcribed strand. Mellon et al. reported that in the DHFR gene of CHO cells, the efficiency of repair showed a significant difference in both strands (Mellon et al., 1987). They also reported the preferentially repair of active genes in human cells (Bohr et al., 1985; Mellon et al., 1986). However, no difference was reported between the repair of transcribed and non-transcribed strand for the N-methylpurines gene (Scicchitano and Hanawalt, 1989). These results indicated the strand-specific and gene-specific repair kinetics of CPD, which will be further investigated in a genome-wide manner.

A large number of enzymes play vital roles in the repair of CPD or 6-4 PP. The simplest and most effective repair enzyme is photolyase. Organisms utilize photolyase specific binding to CPD (CPD photolyase) or 6-4 photo products (6-4 photolyase) that can trigger a photoreactivation process, which photolyase binds first to the CPD site independent of light, and flips the dimer out of the helix into the center of active domain and forms the photolyase-DNA complex. A chromophore-cofactor MTHF or 8HDF absorbs a photon from light and transfers the energy to another chromophore-cofactor flavin which further transfers the electron to the double bond of pyrimidine dimers and splits the dimer into normal pyrimidines (Sancar, 2003).

However, placental mammals repair CPD and 6-4 PP through more complex but less efficient nucleotide excision repair pathway (NER) instead of photolyase due to the loss of the photolyase gene during evolution. The Nucleotide Excision Repair (NER) pathway, which is divided into two sub-pathways: global genome nucleotide excision repair (GG-NER) and transcription-coupled repair (TC-NER) (Costa et al., 2003). The TC-NER pathway specifically repairs CPD in transcribed strands of actively transcribed genes, whereas GG-NER repairs CPDs in the whole genome including the regions repaired by TC-NER, however the repair rate of GG-NER is much slower than that of TC-NER (Hanawalt, 2002).

In human cells, the deficiency in NER leads to one of three autosomal recessive inherited diseases: Xeroderma Pigmentosum (XP), Cockayne Syndrome (CS), or Trichothiodystrophy (de Boer and Hoeijmakers, 2000; Lehmann, 2003). The patients with these diseases show common features such as hypersensitivity towards UV light. Xeroderma Pigmentosum disease is characterized by deficiency of GG-NER, hyper-mutability, pigmentation and more than 10,000-fold increased risk of cancer in sunlight exposed areas of the skin (Niedernhofer et al., 2011). Especially in the sunlight exposed tissues: such as eyes, lips and tip of tongue (Cleaver et al., 2009). The XPC protein is a DNA damage sensor, which plays a key role in the recognition of damaged DNA in the global-genome repair pathway. And the XPC protein is the initiator of the nucleotide excision repair and involves a two-step mechanism of damage recognition, which primarily detects damage by the XPC-HR23B complex, followed by damage verification by XPA (Sugasawa et al., 1998).

CSA and CSB proteins are involved in the transcription-coupled repair pathway that recognizes CPDs by stalled RNA polymerase II in the sense strand of actively transcribed genes. People with deficiency of CSA and/or CSB suffers from Cockayne Syndrome (CS). These patients are characterized by sunlight sensitivity, pigmentary changes, premature aging and neuro-developmental abnormalities but exhibit normal skin cancer risk (de Boer and Hoeijmakers, 2000; Kraemer et al., 2007; Lehmann, 2003).

Until now, most studies reported CPD location in one or two specific genes for human cells. However, CPDs are widely distributed in genome and the repair kinetic of CPDs in single genes can not be representative for the whole genome. Hence, it is necessary to understand, which genes and genomic regions are sensitive to be damaged and form CPDs, and which genes or which genomic regions are resistant to repair in a genome-wide manner. Therefore the consequence of unrepaired CPDs in genomic regions needs be further investigated. Until today only two groups used chromatin immunoprecipitation combined with microarrays to detect CPDs in yeast cells (Teng et al., 2011) and another group utilized the same technique to map CPD hotspots in human chromosomes 1 and 6 (Zavala et al., 2014). However, both reports were based on repair proficient cells and highly relied on the pre-designed probes for certain genomic regions, which only covered a small portion of the genome and provided no strand specific information. The large portion of repetitive elements can not be analyzed due to limitations of the microarrays used. CPD distribution and repair kinetics in repair deficient cells, which are responsible for XP and CS disease are poorly defined in genome-wide studies. Furthermore, the CPDs distribution in a strand-specific manner has never been proved on a genome-wide level.

Hence, it's critical to identify CPD sites and differentiate the strand that contains the CPD in a genome-wide scale and uncover the difference between normal and NER deficient cell lines.

With the development of next-generation sequencing techniques, these techniques can provide high resolution and high throughput sequence information. Since CPDs occurs in one strand of DNA, distinguishable pathways like GG-NER for non-transcribed strands and TC-NER for transcribed strands are used to repair CPDs. It's necessary to specify the strand with CPDs by sequencing. However, the traditional ChIP-seq technique fails to differentiate the strand after sequencing since it precipitates the double strand DNA fragments. Hence, the ChIP-seq procedure was modified so that sonicated DNA fragments are ligated to adaptors first, then denatured into single strands and pulled down by the thymidine dimer antibody coupled to magnetic beads. This modified protocol was termed "strand-specific damaged DNA immunoprecipitation followed by massively parallel DNA sequencing" (ssDIP-seq) and is illustrated as ssDIP-seq in Figure 13.

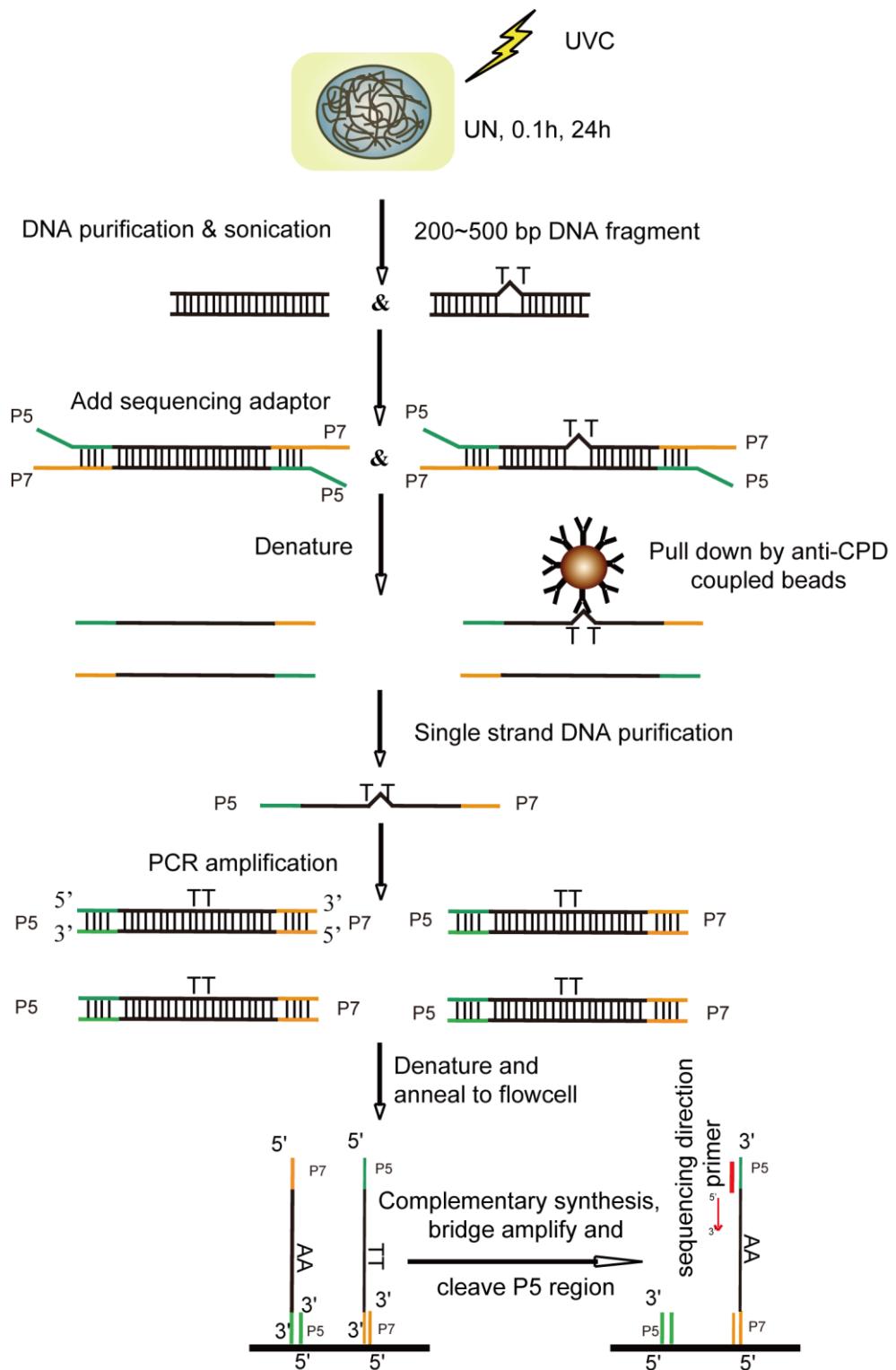


Figure 13. Schematic illustration of ssDIP-seq procedure. Purified and sonicated dsDNA containing CPDs (200~500 bp) are ligated to adaptors. Then the adapter ligated DNA fragments are denatured into ssDNA and subjected to immune-precipitation using the anti-CPD antibody and magnetic beads. ssDNA is purified and amplified to dsDNA with the Illumina compatible ends P7 and P5 and annealed to the flowcell. The sequence is retrieved after standard cluster generation and sequencing by synthesis. Since the ssDNA with the same direction as CPD contained strand will be cleaved, the opposite strand will be taken as template to start the synthesis and sequencing. The output sequence is read in the same direction as the CPD containing strand.

Three cell lines were selected for exposure to 12 J/m² UVC (254nm) irradiation, HaCaT cells (keratinocyte) that are proficient in NER. In contrast, the XPC deficient skin fibroblast cell line XP4PA-SV-EB (XPC^{-/-}) which is deficient in GG-NER, and the CSB deficient skin fibroblast cell line CS1AN (CSB^{-/-}) were used. The XPC^{-/-} cell line is SV40 transformed cell line with the TG di-nucleotide deletion at position of 1483 and 1484 of the XPC gene. The coding frame shifts and results into a 430 amino acid long truncated XPC protein which deletes about half the size of wild-type protein (Emmert et al., 2000; Legerski and Peterson, 1992; Li et al., 1993). The CSB^{-/-} cell line is also a SV40 transformed cell line (Mayne et al., 1986).

In this work, NER proficient and deficient cells were exposed to 12 J/m² UVC and DNA was isolated from cells either immediately (0.1 hour) or late (24 hours) to prepare the sequencing library of ssDIP-seq. As controls input DNA of all three cell lines under physiological condition were used. Until now, sequencing result for cell lines HaCaT and XPC^{-/-} were finished while the results for the CSB^{-/-} will be processed in the near future, therefore, here only the analysis and the results of HaCaT and XPC^{-/-} cells will be presented in this thesis.

High resolution mapping data of CPD sites are revealed in a strand-specific way and the relationship of CPD sites with genomic features are explored. Furthermore, the chromatin state around CPD hotspot sites is investigated. The difference of CPD distribution and repair kinetics between NER proficient and deficient cell lines is discussed.

Results

Both CSB and XPC deficient cell lines show partial CPD repair, while the NER proficient cell line HaCaT shows complete repair

To compare the CPD repair kinetics of repair deficient and proficient cell lines *in vivo*, immunofluorescence staining (IF) for CPD was performed and quantified by high throughput screen with a high content screening microscope (Operetta, Perkin Elmer) for CSB^{-/-}, XPC^{-/-} and HaCaT cells at time points UN, 0.1, 1, 3, 6, 24, 48 hours after a single dose of 12 J/m² UVC irradiation (Figure 14. A). Since most of the cells induced apoptosis after 48 hours in the NER deficient cell lines, the two late time points 72 and 96 hours were stained for HaCaT cells only.

It can be seen that HaCaT cells repair CPDs normally. The amount of CPDs decreases according to time and is close to the background level after 48 hours (Figure 14. A). Whereas both NER deficient cell lines revealed a partial reparability but reaches a plateau at 24 hours after irradiation. To further validate the repair ability of the three cell lines at three time points unirradiated (UN), 0.1 hour, 24 hours corresponding to physiological condition, early, later

repair time point were chosen and isolated DNA was analyzed by slot blot combined with immuno-CPD staining. Due to the sensitivity variation of the approaches it can be explained that a lower level of background intensity was detected by slot blot, the folds of CPD induced immediate after UV are also different. With a 6~8 fold increase of the CPD intensity by immune-fluorescence staining and a 11~15 fold increase detected by slot blotting at 0.1 hour after exposure (Figure 14. B, C, D).

Overall both approaches confirm the CPD repair deficiency of CSB^{-/-}, XPC^{-/-} and the NER proficiency of the HaCaT cells. For the IF result, immediately after exposure to UVC, more than 6 times higher levels of CPDs are induced in all three cell lines (CSB^{-/-} 6.25; XPC^{-/-} 7.38; HaCaT 6.19). After two days, CPDs are close to completely repaired in HaCaT cells (16% unrepairs) whereas 41%, 57% remain unrepairs in CSB^{-/-} and XPC^{-/-}.

Immediate and 24 hours after UV, isolated DNA from HaCaT and XPC^{-/-} cell was used to prepare ssDIP-seq libraries and DNA for input samples (unirradiated cells) was treated as control for DNA sequencing without pulldown.

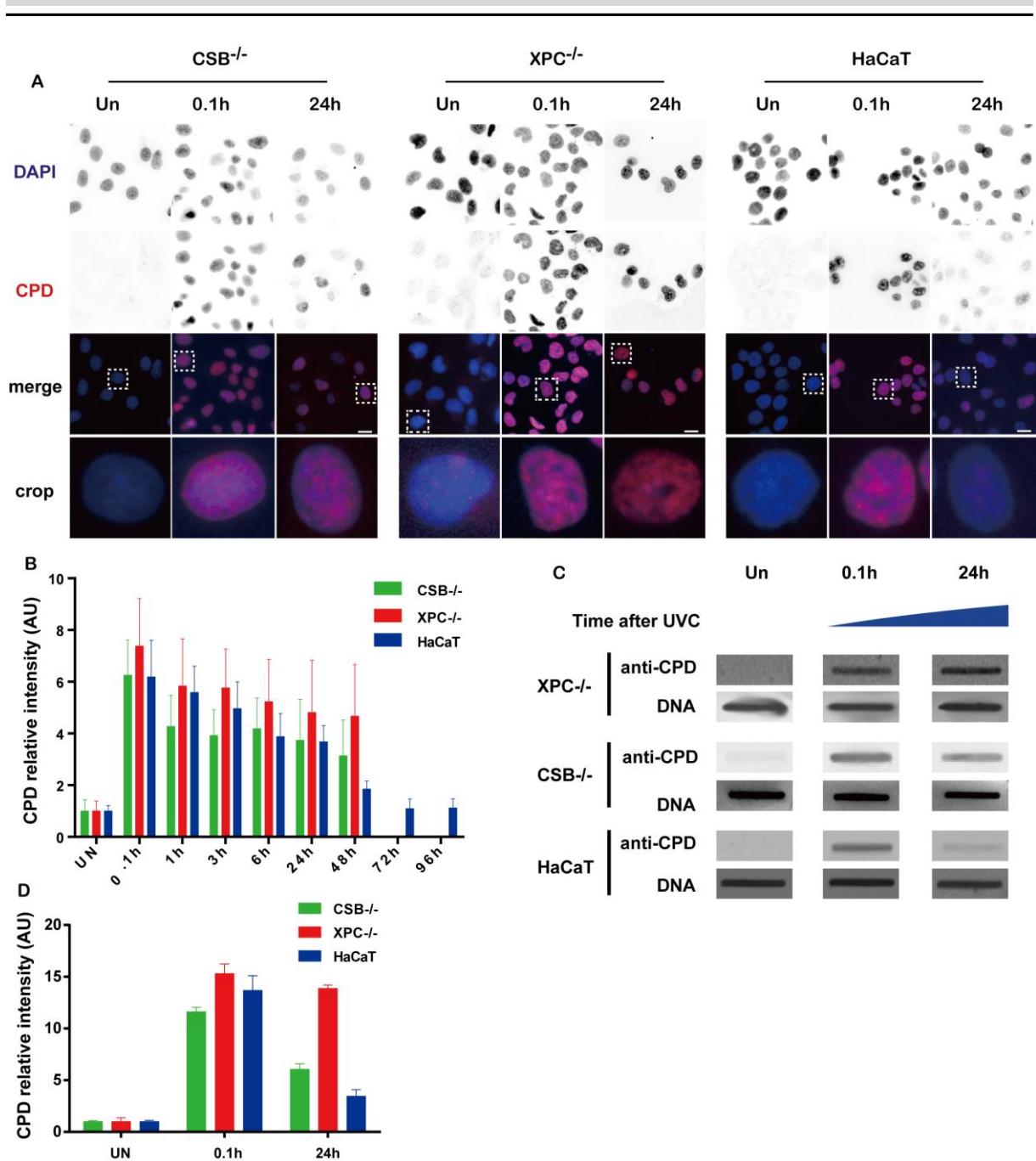


Figure 14. Characterization and validation of cellular system and experimental strategies.

To characterize the repair of CPDs in XPC^{-/-}, CSB^{-/-} and HaCaT cell, these were exposed to 12 J/m² UVC (254 nm) and incubated as indicated. (A) Examples of immunofluorescence staining for CSB^{-/-} (left); XPC^{-/-} (middle); HaCaT (right) cells. CPD staining (Cy 3, red), DNA: DAPI. Bar: 20 μ m. (B) Average CPD immunofluorescence intensity (arbitrary units) is calculated by immunofluorescence high throughput analysis at different time points after exposure to UVC. Intensities are normalized to unirradiated cells. (C) Slot blot results for XPC^{-/-}, CSB^{-/-} and HaCaT at unirradiated, early, late repair time points. 150 ng DNA was used for anti-CPD immuno-staining, 1.5 μ g DNA for methylene blue staining as DNA loading control. (D) Quantification of slot blot after normalizing to corresponding DNA amount. Error bars represent standard deviation.

XPC^{-/-} cells show higher levels of endogenous copy number variations than HaCaT cells

It is known that chromosome aberrations and genome instability contribute to oncogenesis and are characteristics of most cancers (Negrini et al., 2010). Some deletions and duplications of chromosomal segments, such as copy-number variations (CNVs), which are defined as DNA segments with sizes that vary from 1 kb or longer and have variable copy number relative to the reference genome, have been associated with evolution, genetic diversity between individual humans and susceptibility or resistance to human diseases (Iafrate et al., 2004; Redon et al., 2006; Stankiewicz and Lupski, 2010). For instance, the higher copy number of gene CCL3L1 is related with the locus susceptibility to HIV integration (Gonzalez et al., 2005). Even within normal individuals, CNVs occur frequently. Among the large number of individuals from four populations, the regions with variable copy numbers cover 12% of the human genome (Iafrate et al., 2004). Current models indicate that replication, homologous recombination (HR) and non-homologous end-joining (NHEJ) contribute to the formation of CNVs and tend to be increased in the vicinity of low copy repeat (LCR) regions (Hastings et al., 2009).

Recently researchers revealed that the deficiency or reduction of the XPC gene is not only involved in skin cancer but also closely relates with bladder and lung cancer (Cheo et al., 1999; Dai et al., 2014; Hollander et al., 2005). Although bladder and lung tissues are not exposed to UV light, the function of XPC is not restricted to UV induced NER pathway only. Melanomas are often associated with the presences of frequent chromosomal aberrations. Whereas, melanocytic nevi shows no chromosomal aberration (Bauer and Bastian, 2006). To investigate whether the deficiency of XPC is related to genome instability, I analyzed the genomic DNA (input sample) of the HaCaT and XPC^{-/-} cell lines to detect copy number variation by whole genome DNA sequencing. This is the first time that CNVs data are shown for the XPC^{-/-} cell line.

CNVs were detected in the input sample (genomic DNA) by the software control-FREEC (Boeva et al., 2012). In the XPC^{-/-} cell line, the total length of regions with higher copy numbers is 704,738,571 bps (22.4% of the genome). And chromosomal losses are 529,925,485 bps (16.8% of the genome) (Figure 15. A). Overall 39.2% of the genome shows copy number variations, which is significantly higher than the 12% in normal populations. Among these regions 5,383 genes are within gained regions and 3,272 genes are located in lost regions. The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis by WebGestalt (Wang et al., 2013) for gain and loss genes base on a p-value of 0.001 was used to identified 26 genes in the copy number gain population that belonged to the melanogenesis pathway (hsa04916), whereas no gene in the lost regions belonged to

this pathway. The regions with copy number variation and genes included in pathways melanogenesis (KEGG ID: 4916) can be seen in (Table 2). 68 genes were assigned to pathways in cancer in gain regions in contrast to 51 genes in loss regions. However, both chromosomal gains and losses contain oncogenes and tumor repressors.

Especially, chromosome 8 shows losses in the short arm and gains in the long arm. This 8q region (115 M-130 M) shows a copy number increase of 4 to 11 fold and contains the important oncogene c-Myc and downstream target PVT1 which is low expressed in normal tissue, but over-expressed in transformed cells (Carramus et al., 2007). The high copy number of the PVT1 gene may explain the over-expression level. In the downstream location of the oncogene c-Myc, there is a cluster of microRNAs has-miR-1204, 1205, 1206, 1207, 1208 which are up-regulated through p53-dependent induction of PVT1 (Barsotti et al., 2012). The 8q region with the increased copy number was also reported in patients and found in gained regions of chromosome 8 in metastasizing primary melanomas (Aalto et al., 2001; Bastian et al., 1998; Curtin et al., 2005). The gains of this genomic region indicate a potential mechanism that contributes to the development of skin cancer in XPC deficient cells.

The average sizes of copy number variations in HaCaT, a spontaneous immortalized cell line, are larger than those in the XPC^{-/-} cell line. HaCaT cells show whole chromosome or whole arm depletions or amplifications, such as depletion of whole chromosome 4 and 21, short arm of chromosome 3 and 9, the amplification of chromosome 15 and 20, long arm of chromosome 1, 9 (Figure 15. B). In total, regions with altered copy numbers include 347,262,859 bps (11% of human genome) gains (enclosing 3,347 genes) and 631,641,303 bps (20% of human genome) of losses that contain 3,715 genes. 51 genes in gained regions and 54 genes in lost regions belong to pathways associated with cancer.

However, due to the fact that the XPC^{-/-} cell line is SV40 transformed, it can not be exclusively concluded that the deficiency of XPC results in this massive copy number variation since it might be partially induced by the SV40 transformation. Hence, the genomic regions with copy number variation need to be further validated in primary deficient cells or tissues.

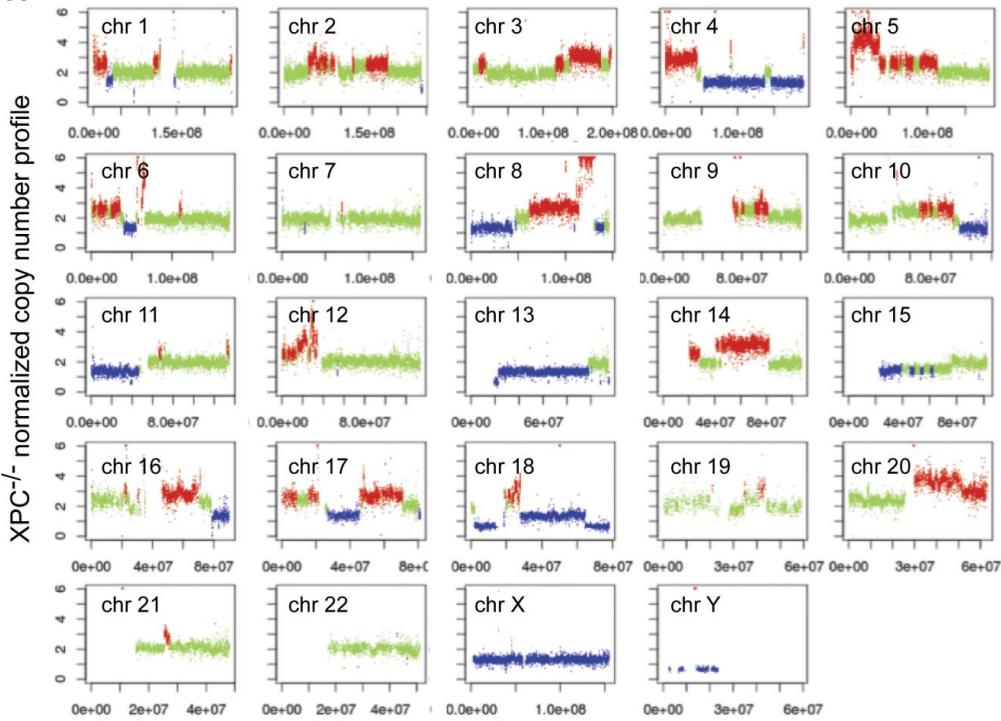
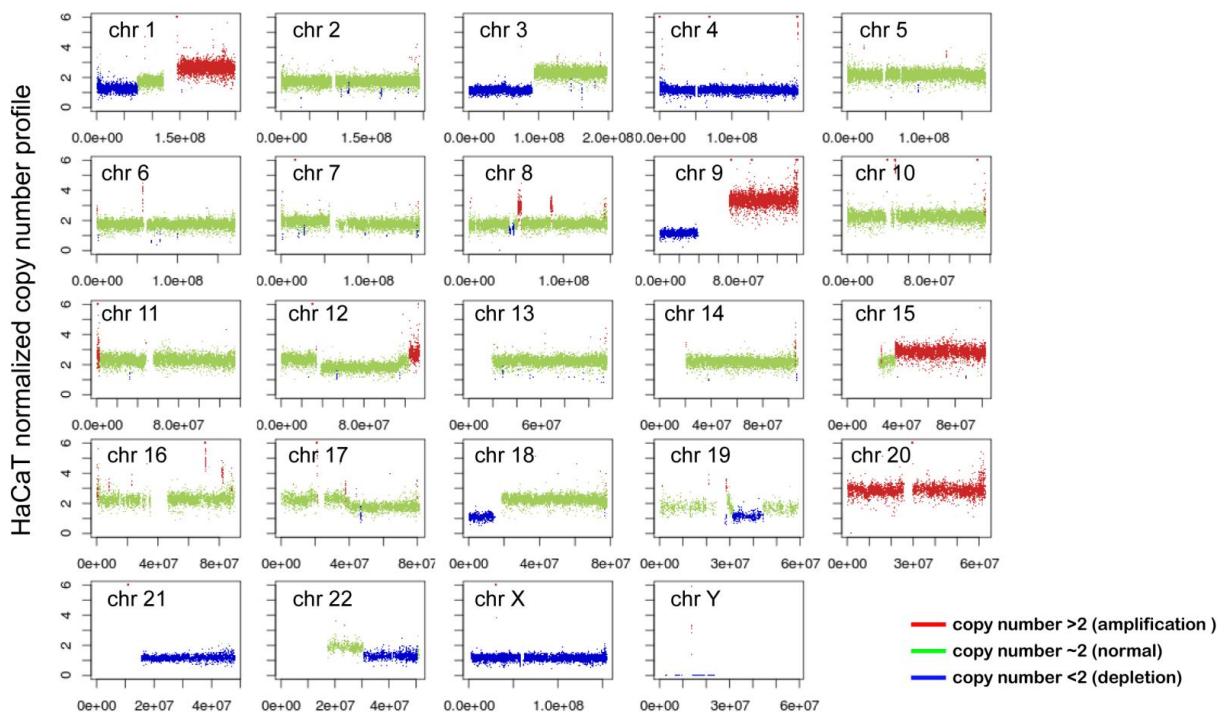
A**B**

Figure 15. Copy number variations in XPC^{-/-} and HaCaT cell lines. (A) XPC^{-/-} (B) HaCaT. Amplification (red); depletion (depletion); normal (green).

Symbol	Gene name	Refseq ID	Ensembl ID
NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog	4893	ENSG00000213281
ADCY4	adenylate cyclase 4	196883	ENSG00000129467

RAF1	v-raf-1 murine leukemia viral oncogene homolog 1	5894	ENSG00000132155
WNT7A	wingless-type MMTV integration site family, member 7A	7476	ENSG00000154764
EDN1	endothelin 1	1906	ENSG00000078401
ASIP	agouti signaling protein	434	ENSG00000101440
PRKCA	protein kinase C, alpha	5578	ENSG00000154229
GNAI3	guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 3	2773	ENSG00000065135
FZD6	frizzled family receptor 6	8323	ENSG00000164930
WNT2B	wingless-type MMTV integration site family, member 2B	7482	ENSG00000134245
CALM2	calmodulin 2 (phosphorylase kinase, delta)	805	ENSG00000143933
TCF7L1	transcription factor 7-like 1 (T-cell specific, HMG-box)	83439	ENSG00000152284
WNT4	wingless-type MMTV integration site family, member 4	54361	ENSG00000162552
PRKACG	protein kinase, cAMP-dependent, catalytic, gamma	5568	ENSG00000165059
WNT5B	wingless-type MMTV integration site family, member 5B	81029	ENSG00000111186
ADCY2	adenylate cyclase 2 (brain)	108	ENSG00000078295
ADCY5	adenylate cyclase 5	111	ENSG00000173175
GNAO1	guanine nucleotide binding protein (G protein), alpha activating activity polypeptide O	2775	ENSG00000087258
CALML6	calmodulin-like 6	163688	ENSG00000169885
GSK3B	glycogen synthase kinase 3 beta	2932	ENSG00000082701
GNAQ	guanine nucleotide binding protein (G protein), q polypeptide	2776	ENSG00000156052
WNT8B	wingless-type MMTV integration site family, member 8B	7479	ENSG00000075290
DVL1	dishevelled, dsh homolog 1 (Drosophila)	1855	ENSG00000107404
CAMK2G	calcium/calmodulin-dependent protein kinase II gamma	818	ENSG00000148660
DVL2	dishevelled, dsh homolog 2 (Drosophila)	1856	ENSG00000004975
ADCY7	adenylate cyclase 7	113	ENSG00000121281

Table 2: Genes with amplified copy number in melanogenesis pathway

Chromosomes show specific repair kinetics both in NER proficient and deficient cells

To investigate the chromosome-specific CPD repair in XPC^{-/-} cells, the total RPKM values for each chromosome are compared directly after UV exposure and 24 hours later. These values are normalized to the length of the corresponding chromosome in mega base pairs. Due to the sum of RPKM value for the whole sample, which is equal to the sum of RPKM values of all chromosomes is constant between immediate and after 24 hours upon UV exposure. For each chromosome, the normalized sum of RPKM values at 24 hours was subtracted from the normalized sum of RPKM values obtained at 0.1h after exposure. Values greater than zero indicate that the CPDs are relative resistant to repair in this chromosome. In this way, several chromosomes (chromosome 16, 17, 19, X) show resistance of repair in a chromosome-specific way (Figure 16. A red bars). All these chromosomes show a high density of simple sequence repeat (microsatellite), especially for chromosomes 16, 17 and

19. Chromosome 19 shows the highest amount of mononucleotide repeats in intron regions and a high density of AAT repeats. It also contains a large number of tetranucleotide repeats (AAAT, AAAG, AAAC and AAGG) in the non-coding regions(Subramanian et al., 2003). Although chromosome 19 has the highest abundance of genes per mega base pair, this chromosome is still resistant to repair. This suggests that the high density of simple sequence repeats overrules the TC-NER benefit on this chromosome. This indicates that the microsatellites are obstacle for efficient DNA repair and microsatellite frequency is closely related to CPD density at later time points. However, HaCaT cells show the opposite trend (Figure 16. A blue bars), which chromosome 19 exhibiting the lowest repair resistance, which might reflect the high repair efficiency of TC-NER in this chromosome for a large number of genes.

Overall these results indicate that the sequence in microsatellite regions of chromosome 16, 17, 19, X is prone to the induction of CPDs and resistant to repair as well. How the unrepaired CPDs in these several chromosomes influence the expression level of genes need to be further investigated through combining analysis of CPDs site and expression data before and after irradiation.

In both cell lines, the amount of CPDs is correlated to A/T, AT/AT, TA/TA, TT/AA content but anti-correlated to G/C content

To reveal the relationship of CPD distribution with the DNA content of the surrounding regions, for each 10 kbps intervals, the ratio of RPKM to input was calculated and the genome wide average of this ratio was subtracted. This reflects the relative abundance of CPDs. These normalized RPKM ratio values for each 10 kbps window is comparable across the whole genome within the same time point, but not comparable among different time points since the average ratio of the whole genome is different between the different time points. To access the correlation between A/T and G/C content of each 10 kbps window with the CPD abundance at the different time points we calculated the spearman's correlation between normalized RPKMs and A/T, G/C content as well as to the di-nucleotides for both cell lines (Figure 16. B). Since the human reference sequence is ordered in 5' to 3' direction, the di-nucleotide and its complementary di-nucleotide were summed up, for example TG+CA, then the density was calculated as the number of di-nucleotide divided by the length of the interval for each genomic bin. Thus the correlation value is based on the whole genome-wide correlation that includes 286,729 of 10 kbps intervals, the correlation was defined as significant if a spearman value is higher than 0.2 or lower than -0.2. This is in accordance to other genome-wide correlation studies (Fungtammasan et al., 2012). From early to late time points after UV exposure, for both cell lines, the correlation of A/T, G/C content tend to the

opposite direction that the correlation of CPD with A/T content increased whereas the correlation to the GC content decreases. The correlation of CPD with TT/AA, AT/AT and TA/TA increases while others di-nucleotide correlations decrease (Figure 16. B). The correlations of the three di-nucleotide types to the CPD persistence are consistent with the A/T content. The increasing correlation with the A/T content might be due to CPDs in repetitive elements which are A/T rich regions and are resistant to repair. This means that the repair kinetic of CPD sites is not homogeneous in the whole genome and is dependent on DNA sequence.

To uncover the differences of CPDs in genic and non-genic regions, the average of normalized CPD's RPKM is calculated (Figure 16. C). For XPC^{-/-}, the relative abundance of CPDs in genic and non-genic regions shows a significant difference from immediate to late time points after irradiation. It is consistent with the XPC^{-/-} cells genotype and the deficiency in GG-NER to result in unrepaired CPDs in non-genic regions. HaCaT cells however are proficient in both GG-NER and TC-NER and therefore the CPD reparability in both genic and non-genic regions is equal. Hence there is no significant difference at late time point for the distribution of CPDs between genic and non-genic regions (Figure 16. C).

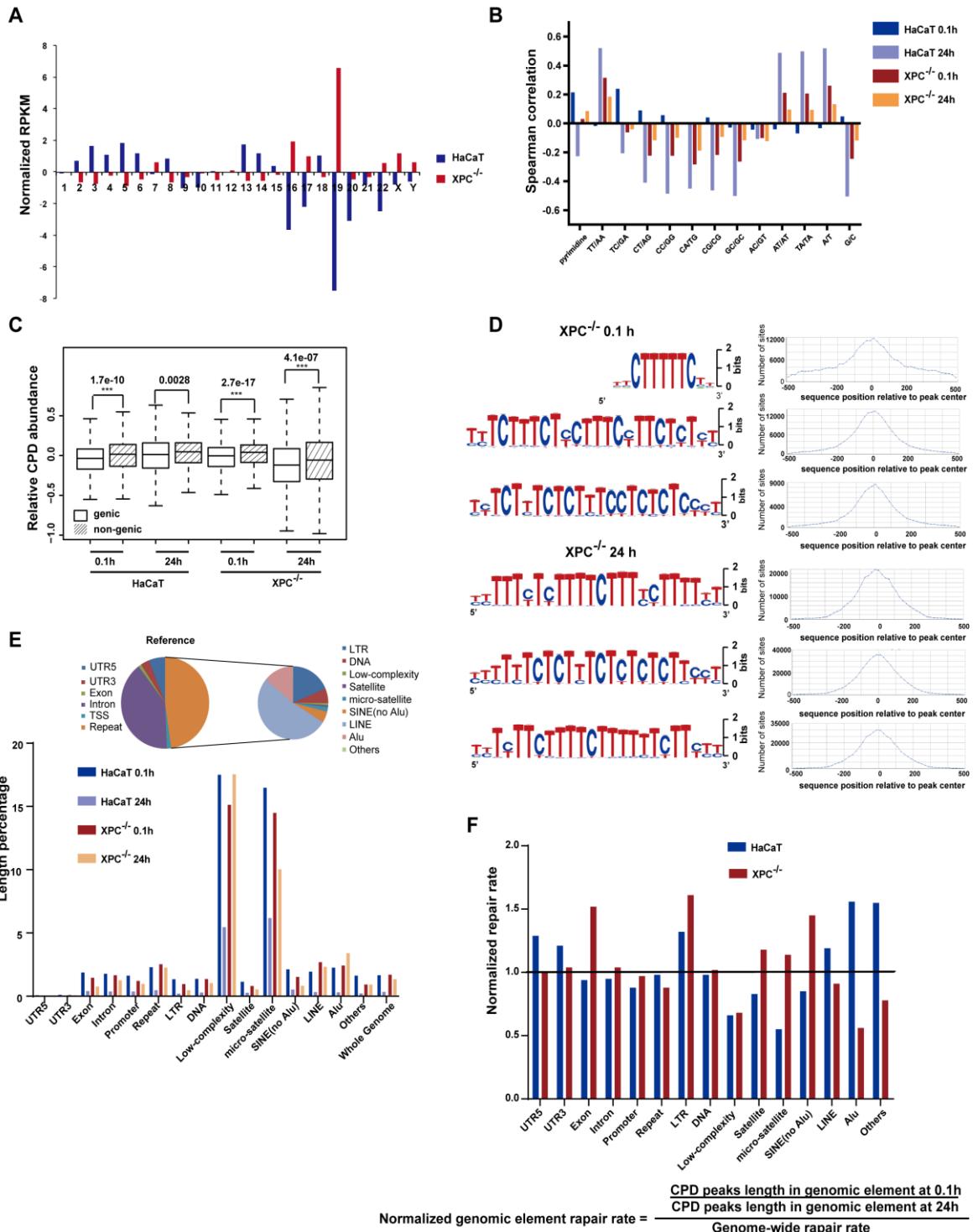


Figure 16. Genome-wide characterization of CPD distribution. (A) chromosome-specific CPD repair resistance. For each chromosome, the sum of RPKM values at 0.1h are subtract with the values obtained for 24h, normalized to the chromosomal length in mbps. Bar: HaCaT (blue), XPC^{-/-} (red). (B) The Spearman correlation of CPD abundance with dimer density in 10 kbps intervals. HaCaT 0.1h (dark blue), 24h (light blue); XPC^{-/-} 0.1h (dark red), 24h (orange) (C) CPD abundance in genic and non-genic regions. Genic (blank box); Non-genic (stripe box). All boxes and whiskers represent 25-75 percentiles and 3x IQR, respectively. (D) Left: CPD motifs for XPC^{-/-} 0.1h and 24h. Height is normalized to the motif frequency. Right: number of sites within peaks of the presented motif. (E) Percentage of length in genomic elements occupied by CPD peaks. HaCaT 0.1h (dark blue), 24h (light blue); XPC^{-/-} 0.1h (dark red), 24h (orange). Top panel: percentage of genomic elements in reference genome (hg19). (F) Repair rate of genomic elements relative to overall genome repair rate. HaCaT (blue); XPC^{-/-} (red). Bottom: formula to calculate repair rate for

genomic elements.

The motifs of CPD peaks are continuous di-pyrimidine dimers and CPD peaks are preferentially located in introns and repetitive regions, especially in micro-satellite and low complexity sequence

All above described results are based on RPKM calculation for 10 kbps genomic intervals. For CPDs an alternative approach was applied that allows a more precise localisation of CPD sites in chromatin. This approach is called peak-calling and is realised with an algorithm called MACS (Zhang et al., 2008) and the settings were described in the methods section. After peak-calling, the total length of CPD peaks is ~50 Mbps directly after UV exposure and is similar for both HaCaT (49,947,430 bps) and XPC^{-/-} (48,507,810 bps). This is consistent with the immunofluorescence staining and slot blotting result that revealed that the same dose of UVC induces similar amount of CPDs among the three cell lines. However, 24 hours after UV, the total length of peaks for XPC^{-/-} is about 38 Mbps, which is significantly higher than the area covered in HaCaT cells with is 10 Mbps due to the deficiency of GG-NER in XPC^{-/-} cells. This again could be verified by the immunofluorescence results (Figure 14). The motifs were derived from the CPD peaks and discovered by the computational pipeline RSAT (Thomas-Chollier et al., 2012) confirmed that continuous di-pyrimidine dimers (TT, CT, TC, CC) motif occurs within the highly significant CPD peaks, especially at the centre of the peak regions (Figure 16. D). The threshold for motif detection was set so that at least 75% of the peaks contained the predicted motif. The CPD motifs for HaCaT showed the same results as for the XPC^{-/-} cells.

In order to further understand the genomic location of CPD peaks, genomic features were assigned to each base pair of the peaks based on their position. The length percentage of the CPD peaks overlapped by genomic elements was calculated and compared to the coverage of the corresponding genomic elements in the reference genome (Table 3). The genic elements include promoters (5 kbps upstream of transcription start site), 5' prime untranslated regions (UTR5'), exons, introns and 3' prime untranslated regions (UTR3'). Compared to genic elements, the length percentage of CPD peaks overlapped by repetitive elements is significant higher than the coverage of repeats in the reference genome. Especially for low complexity and microsatellites, which are close to 10 fold enriched, only the LTR is lower than the corresponding part in the reference genome. Whereas the length percentage of CPD peaks within genic regions is less or close to the coverage of genic elements in reference genome.

	XPC ^{-/-} 0.1h	XPC ^{-/-} 24h	HaCaT 0.1h	HaCaT 24h	reference
UTR5	0.19%	0.19%	0.26%	0.20%	5.51%
UTR3	0.17%	0.16%	0.21%	0.18%	2.93%
Exon	0.92%	0.60%	1.15%	1.22%	0.97%
Intron	39.69%	38.06%	41.28%	43.24%	36.84%
Promoter	1.09%	1.12%	1.44%	1.63%	1.40%
Repeat	71.73%	81.18%	62.97%	64.33%	43.64%
LTR	5.09%	3.14%	6.89%	5.21%	8.13%
DNA	2.49%	2.42%	2.45%	2.49%	2.83%
Lowcomplexity	4.12%	6.06%	4.64%	7.04%	0.42%
Satellite	0.18%	0.15%	0.25%	0.30%	0.35%
microsatellite	8.37%	7.33%	9.25%	16.88%	0.89%
SINE_noAlu	2.14%	1.47%	2.92%	3.44%	2.17%
LINE	39.70%	43.63%	27.82%	23.33%	22.74%
Alu	9.53%	16.86%	8.61%	5.53%	6.03%
Other	0.05%	0.06%	0.08%	0.05%	0.08%

Table 3: Length percentage of CPD peaks overlapped by genomic elements and corresponding coverage of genomic elements in the reference genome.

Furthermore, the length percentage of genomic element covered by CPD peaks relative to the total length of each genomic element was calculated as well. The total length of CPD peaks occupies about 1.6% of the whole genome in both cell lines at 0.1 hour after UV exposure. Due to the deficiency of the GG-NER pathway in XPC^{-/-}, the percentage only decreased slightly to 1.28% compared to the 0.34% in HaCaT cells after 24 hours. Among different repetitive elements, low complexity and microsatellite repeats are strongly overrepresented in both cell lines. Even after 24 hours, in HaCaT cells, there are still about 6% of these two elements, which are covered by CPD peaks. Compared to genome-wide CPD peaks coverage, LINE and Alu are slightly overrepresented whereas LTR, DNA, satellite repetitive elements are underrepresented in XPC^{-/-} cell line (Figure 16. E and Table 4).

	XPC ^{-/-} 0.1h	XPC ^{-/-} 24h	HaCaT 0.1h	HaCaT 24h
UTR5	0.05%	0.04%	0.08%	0.01%
UTR3	0.09%	0.07%	0.11%	0.02%
Exon	1.46%	0.76%	1.88%	0.41%
Intron	1.66%	1.26%	1.78%	0.38%
TSS	1.20%	0.98%	1.63%	0.38%
Repeat	2.53%	2.26%	2.29%	0.48%
LTR	0.96%	0.47%	1.34%	0.21%
DNA	1.35%	1.04%	1.37%	0.29%
Low complexity	15.12%	17.54%	17.51%	5.46%
Satellite	0.81%	0.54%	1.15%	0.28%
microsatellite	14.49%	10.02%	16.48%	6.18%
SINE_noAlu	1.52%	0.82%	2.13%	0.52%
LINE	2.69%	2.33%	1.94%	0.33%
Alu	2.43%	3.40%	2.26%	0.30%
Other	0.92%	0.93%	1.63%	0.22%
total	1.62%	1.28%	1.66%	0.34%

Table 4: Length percentage of genomic elements overlapped by CPD peaks

Both ways of calculation confirm the preferentially location of CPD peaks in repetitive elements, especially in low complexity and microsatellite elements, which suggests the distribution of CPDs is not stochastic.

The repair deficient and proficient cell lines show a discrepancy in the repair rate of CPDs since HaCaT cells can decrease the CPD peaks coverage by 4.87 fold, while in contrast XPC^{-/-} cells only decrease the coverage by 1.27 fold within the first 24 hours post UV exposure. To further reveal the variations of repair kinetics in all genomic features, the fold change of the decrease was calculated for each genomic feature and normalized to the whole genome average repair rate (4.87 for HaCaT and 1.27 for XPC^{-/-}). If the value is higher than the one, this indicates that the repair rate in the specific genomic feature is higher than the genome wide average and vice versa. The repair rate of repetitive elements, especially low complexity elements, which is consistently lower than average and indicates the repetitive elements are resistant to repair in both cell lines (Figure 16. F). However, not all the repetitive elements have the same lower repair rate such as LTR elements show a higher repair rate in both cell lines, compared to the genome-wide average, which means CPDs in LTR elements were fast repaired. Other elements present different repair rates in HaCaT compared to XPC^{-/-} cells. CPDs in satellite, microsatellite and SINE sequences are repaired slower in XPC^{-/-} cells, whereas faster in HaCaT, opposite to LINE and ALu elements. In XPC^{-/-} cells, compared to previous result that microsatellites are overrepresented within CPD peaks, microsatellite shows a slightly higher repair rate than the genome-wide average. But microsatellites are still more than 10% of regions that are covered by CPD peaks after 24 hours of repair. While in contrast in LINE and Alu repeats about 2% of the regions are covered by CPD peaks immediately after exposure, however the percentages stay the same or slightly increased at late time points due to the slower repair rate in XPC^{-/-} (Table 4).

CPDs in exons were repaired faster in XPC^{-/-} due to the proficiency in TC-NER. Relative to HaCaT cells, which are proficient in the total NER pathway, the repair rate is still not equal in each feature, which the repair rate of repetitive region is close to the overall genome-wide average. However, CPDs in low complexity regions, satellite, microsatellite and SINE elements are also repaired slowly. LINE, LTR and ALu on the other hand are fast repaired (Figure 16. F). This result indicates the potential mechanism to influence the repair efficiency, which might be due to the higher order chromatin environment around these elements. And there is discrepancy of repair rate between repair proficient and deficient cell lines.

Overall, analysis of the location of all CPD peaks reveals that CPD peaks are overrepresented within repetitive DNA regions, such as microsatellites and low complexity. Due to the fact that all data analysis is based on the uniquely mapped reads, even 48% of human genome is non-repetitive, 80% of human genome can be mapped by 30 bp per sequencing

tag and 86% by 70 bp (Rozowsky et al., 2009). Although we sequenced 50 bps per tag, there are still some repetitive regions with are highly repeated tandem repeat, which cannot be mapped uniquely. Therefore, the amount of CPDs in repetitive regions is still under-estimated.

Since the XPC^{-/-} cells have a high risk of cancer after expose to sunlight, the high percentage of resistant CPDs in microsatellite provides hints of the potential mechanism that contributes to cancer. Previous reports have shown that microsatellite instability (MSI) result in colon cancer (Boland and Goel, 2010) and cutaneous tumors (Hussein and Wood, 2002) since the deletion of microsatellites result in frame shifts or mis-sense mutations in genes (Duval et al., 1999; Oda et al., 2005). However, the mechanism of induction of MSI is still unknown. CPDs can induce microsatellite associated mutations or deletion during the replication process (Ikehata and Ono, 2011; Pfeifer et al., 2005). Here a hypothesis is proposed that the microsatellites located in the exon regions where enriched by unrepaired CPDs in the non-transcribed strand to induce frame shifts or mis-sense mutations of genes during replication and result in disruption of tumor suppressor genes or up-regulation of oncogenes to promote skin cancer development. However, this hypothesis needs to be further validated by experiments.

CPDs are preferentially induced in the anti-sense strand and repaired faster in sense strand

It is known that CPDs are lesions formed at adjacent thymine or cytosine bases in DNA via photochemical reactions. CPDs in transcribed strand are recognized by RNA polymerase (Pol) II and induce stalling of transcription (Brueckner et al., 2007). Then transcription-coupled repair (TC-NER) is triggered to efficiently eliminate CPD lesions by recruiting the CSB protein to Pol II. Cells utilize two NER repair sub-pathways to eliminate CPDs located in transcribed (or sense) and non-transcribed (or anti-sense) strands. CPDs in non-transcribed strands are repaired by GG-NER instead of TC-NER, whereas the repair efficiency of TC-NER is higher than GG-NER and this higher repair efficiency of TC-NER causes CPDs are preferentially repaired in the transcribed strand (Hanawalt, 1991; Sweder and Hanawalt, 1992). However, all previous results are based on experiments of single or only a few genes. No evidence was shown that the same mechanism is valid for all genes. Therefore, it is necessary to study the induction and repair of CPDs in the sense and anti-sense strand in a genome-wide approach. Our method ssDIP-seq possesses the advantages of strand-specific sequencing, which is depicted in the flowchart in Figure 13.

To explore whether the distinguishable strand-specific CPDs occur at all genes, metagene analysis was performed for all transcripts (annotation information was retrieved from UCSC browser). This was done for early and late time points after UV exposure in both XPC^{-/-} and HaCaT cells. The schematic calculation of metagene occupancy is shown in Figure 17. A. The metagene analysis shows that CPD RPKM values are higher in the anti-sense strand than the sense strand, even slightly higher than in the flanking regions. This is consistent for both early and late repair time points (Figure 17. B, C).

Previous reports showed the selective repair of CPDs near the transcription start site in both strands for the JUN gene (Tu et al., 1996). Our metagene result for all transcripts confirms these findings and show that around the transcription start site (TSS) and transcription end site (TES), the RPKM values are lower compared to the surrounding regions. It can be hypothesized that, if a large number of transcription factor complexes permanently occupy the transcription start sites this can lead to a shielding effect that prevents the DNA from getting hit in the first instance. Also it is reasonable to believe that the high local density of RNA polymerase II, which is a CPD sensor by itself, leads to an enhanced repair rate of the TSS compared to less occupied gene regions. These factors either rapidly identify CPD and trigger the repair process or the binding of factors counteract the formation of CPDs, which alter the DNA structure. Especially for some general transcription factors (e.g. TFIIH), which are components of both the transcription and damage repair (NER) machinery. The same tendency of lower levels of damage in TES indicates the potential function of transcription ending sites where cells have to rapidly repair the damage. It might be due to the fact that cells need to ensure precise and complete transcription termination (Kuehner et al., 2011). Again the TES site is known for the higher density of Pol II before it disengaged from the template.

Previous studies on CPD distribution have already shown DNA repair heterogeneity along the gene body and its variations among genes. Active genes are repaired faster than inactive genes and transcribed strands are repaired faster than non-transcribed strand. Even within one gene, the repair rates diminished from 5'end towards 3'end (Gao et al., 1994; Tu et al., 1996; Wei et al., 1995). In our metagene analysis we show that less CPDs occur in genes than in flanking regions in the sense strand, whereas, the anti-sense strand has a higher amount of CPDs. In both strands the amount of CPD is gradient increasing along the gene body from start to end, which can be also explained by the diminished local concentration of transcription factors and Pol II, which act as sensors and trigger the CPD repair. In contrast to genic regions, there is no significant difference between sense strand and anti-sense strand at upstream and downstream regions of genes (Figure 17. B, C), which demonstrates that the strand specific repair is exclusively active in transcribed regions.

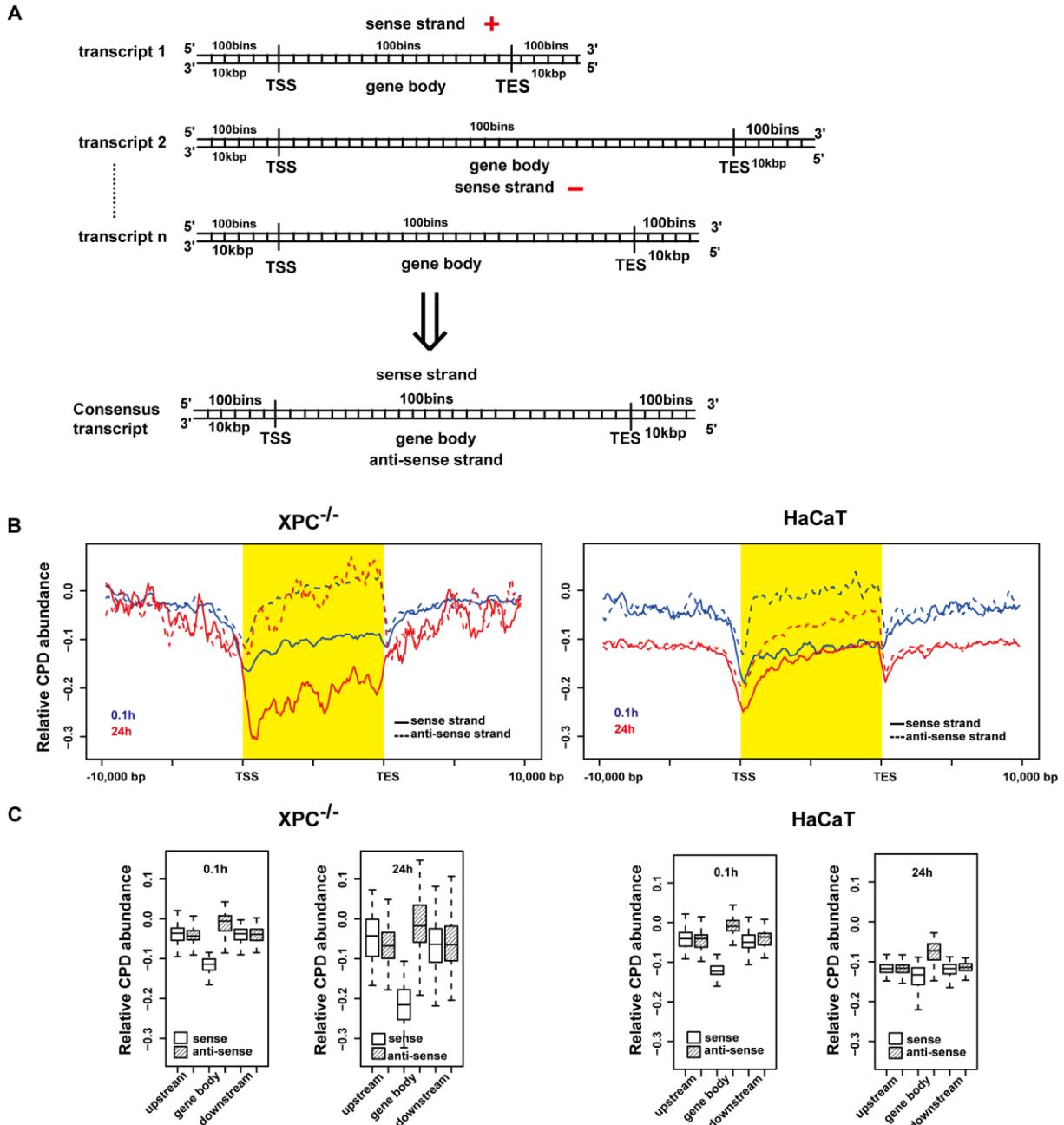


Figure 17. Strand-specific CPD distribution in genic regions. (A) Schematic calculation of CPD metagene. (B) Relative CPD abundance in metagenes of sense strand (filled line) and anti-sense strand (dash line) at 0.1h (blue) and 24h (red). Left: XPC^{-/-}; Right: HaCaT. (C) Relative CPD abundance in upstream, gene body, downstream in sense strand (blank box) and anti-sense strand (stripe box) at 0.1h and 24h. Left: XPC^{-/-}; Right: HaCaT. All boxes and whiskers represent 25-75 percentiles and 3x IQR, respectively.

For the HaCaT cell line, both genic regions and flanking regions, the CPD level gradually decreases from 0.1h to 24h after UVC, which confirms the proficient repair of the HaCaT cells. In contrast, the CPDs in the anti-sense strand and flanking regions of XPC^{-/-} are resistant to repair. HaCaT cells show a higher level of CPDs in the non-transcribed strand

compared to the transcribed strand and higher levels of CPDs than in the flanking regions as well (Figure 17. B, right).

To investigate whether the expression level of genes influences the CPD induction and repair, all transcripts (25,680 for HaCaT and 31,947 for XPC^{-/-}) were sorted and divided into 10 classes according to their expression level (HaCaT expression data was retrieved from GEO with accession No. GSM960278, GSM960286, GSM960294; XPC^{-/-} expression data was produced by Stephan Grulich in our lab). The metagenes of sense strand (transcribed strand) and antisense strand (non-transcribed strand) for each class of genes were calculated separately. In HaCaT cells, genes with the highest expression level showed the lowest level of CPDs in their transcribed strand. In contrast the higher level of CPDs is induced in the sense strand of inactive genes as well. But there is no significant difference in the anti-sense strands and the flanking regions among all classes of transcripts (Figure 18. A, B). This indicates that the expression level of genes only affects the level of CPD signals in the transcribed strand but not in the non-transcribed strand. It might be explained that the highly expressed genes have a higher frequency and concentration of transcription factors including pol II that function as damage sensors, which enable the faster repair rate to reduce the CPD levels. This tendency is attenuated in XPC^{-/-} cells that slightly differ between lowly and highly transcribed transcripts (Figure 18. C, D). It is interesting to note that, in HaCaT cells, after 24 hours, the non-transcribed strand of highly expressed genes retains the highest level of CPDs whereas the transcribed strand of medially expressed genes also have the highest level of CPDs. Here it is explained that, for the highly expressed genes, during transcription the DNA is unwound, which renders the non-transcribed strand into a single strand status that delays the repair due to the fact that repair factors require the double strand structure for efficient initialisation (Schaeffer et al., 1993; Scrima et al., 2008). Also single stranded nucleic acids form a preferred target for the CPD induction compared to the double strand (Becker and Wang, 1989). However, for the inactive genes, which are usually located in facultative heterochromatin, to repair the damage in the anti-sense strand, the condensed chromatin has to be opened up to trigger the repair process, which requires time to remodel the chromatin structure. Nevertheless, the medially expressed genes, which neither exhibit constant unwinding of the DNA double strand nor condensed chromatin, therefore, CPDs in the transcribed strand of medially expressed genes are repaired more efficient.

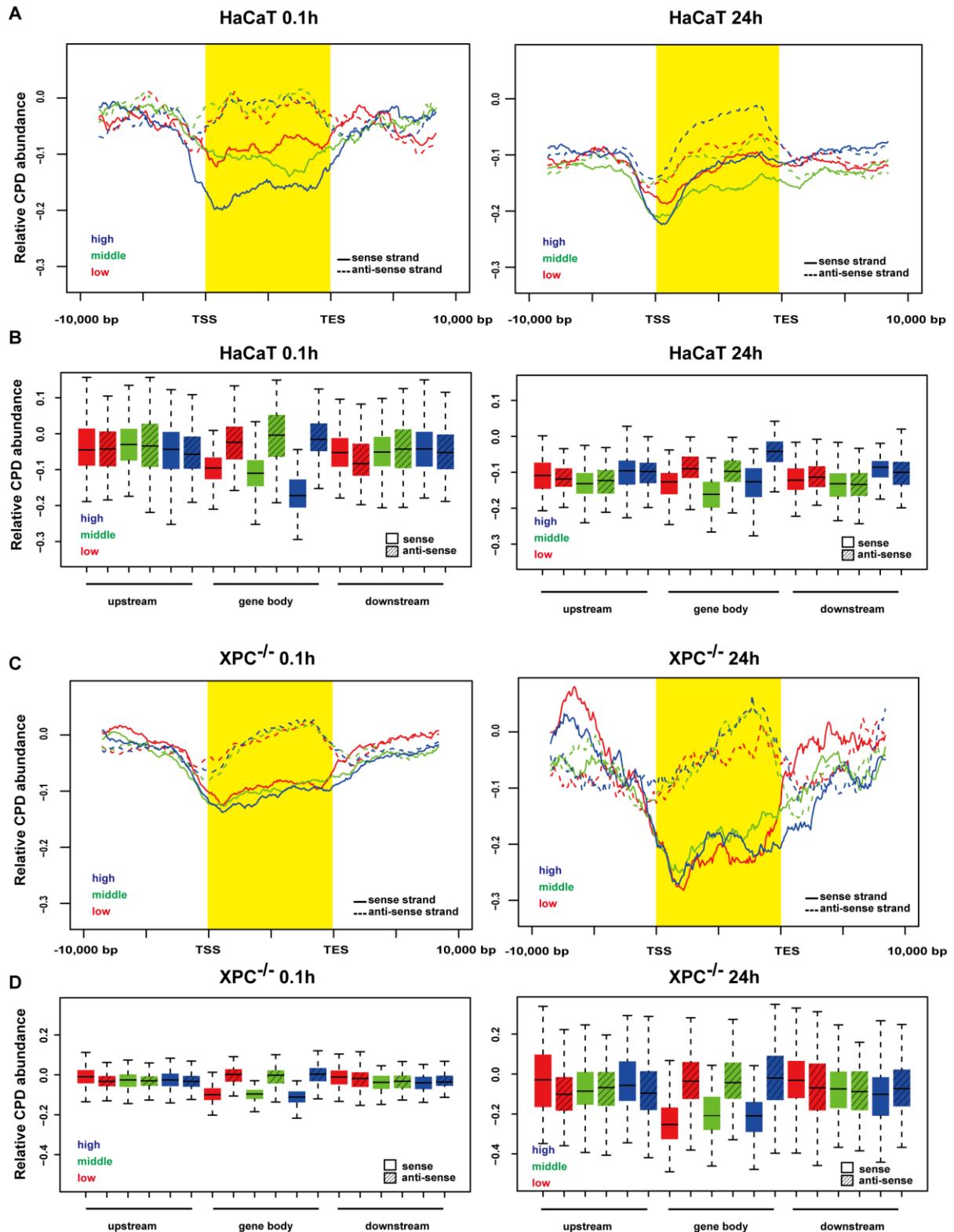


Figure 18. The correlation of expression levels with CPD abundance in metagenes. (A) Relative CPD abundance of high(blue), middle (green) and low (red) expressed transcripts in the sense strand (filled line) and anti-sense strand (dash line) of metagenes. Left: HaCaT 0.1h; Right: HaCaT 24h. (B) Relative CPD abundance of high (blue), middle (green) and low (red) expressed transcripts in the sense strand (filled line) and anti-sense strand (dash line) of upstream, gene body and downstream. Left: HaCaT 0.1h; Right: HaCaT 24h. (C) Relative CPD abundance of high(blue), middle (green) and low (red) expressed transcripts in the sense(filled line) and anti-sense strands (dash line) of metagenes. Left: XPC^{-/-} 0.1h; Right: XPC^{-/-} 24h. (D) Relative CPD

abundance of high (blue), middle (green) and low (red)expressed transcripts in the sense (filled line) and anti-sense strand (dash line) of upstream, gene body and downstream. Left: XPC^{-/-} 0.1h; Right: XPC^{-/-} 24h.All boxes and whiskers represent 25-75 percentiles and 3x IQR, respectively.

Furthermore, RPKM values for sense and anti-sense strand were calculated for all transcripts. The statistical analysis revealed that the difference between the two strands is significant. The non-transcribed strand has a higher RPKM value than the transcribed strand at both 0.1 hour and 24 hours after UVC in both cell lines. The p-value of the t-test is less than $3.1e^{-17}$ for 0.1 hour and $1.4e^{-11}$ for the 24 hours time point in XPC^{-/-} cells, in contrast to $8.2e^{-6}$ and $8.3e^{-5}$ in HaCaT. The discrepancy increases slightly from 0.1 hour to 24 hours since XPC^{-/-} have the normal ability of repair CPDs in transcribed strands but not in non-transcribed strands. However, the p-value for the input sample is 0.763 for XPC^{-/-} and 0.08 for HaCaT, which indicates there is no significant difference between the two strands in the control sample for both cell lines (Figure 19).

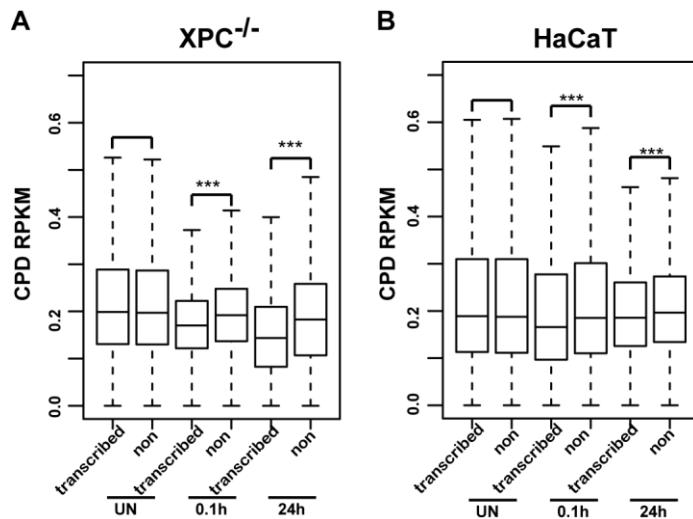


Figure 19. RPKM of CPDs in transcribed and non-transcribed strand. (A) XPC^{-/-}; (B) HaCaT. All boxes and whiskers represent 25-75 percentiles and 3x IQR, respectively.

Condensed chromatin structure influences the repair rather than inhibiting the formation of CPDs

DNase-seq and FAIRE-seq data are used to portrait the accessibility of chromatin. DNase-seq preferentially reveals the localisation of binding sites of active regulatory elements, such as transcription factors. FAIRE-seq on the other hands shows the position of linker regions. To find possible correlations between CPD location and chromatin conformation in terms of openness or condensation around the CPDs, published DNase-seq and FAIRE-seq data was retrieved from different cell lines and tissues, including the BJ (skin fibroblast cell), NHEK

(skin keratinocyte cell), GM12878 (B-lymphocyte cell), H1hESC (embryonic stem cell), K562 (leukemia cell) cells were retrieved from Encode project (Table 5).

Data type	Cell line	GEO Accession No.	Tissue
Dnase-seq	BJ	GSM736518	skin fibroblasts
Dnase-seq	GM12878	GSM736620	B-lymphocyte
Dnase-seq	H1-hESC	GSM816632	embryonic stem cells
Dnase-seq	NHEK	GSM736545	epidermal keratinocytes
Dnase-seq	K562	GSM816655	leukemia
FAIRE-seq	GM12878	GSM864360	B-lymphocyte
FAIRE-seq	K562	GSM864361	leukemia
FAIRE-seq	H1-hESC	GSM864341	embryonic stem cells
FAIRE-seq	NHEK	GSM864338	keratinocyte
H3K9me3	K562	GSM733776	leukemia
H3K9me3	H1-hESC	GSM1003585	embryonic stem cells
H3K9me3	GM12878	GSM733664	B-lymphocyte
H3K9me3	NHEK	GSM1003528	keratinocyte
H3K27ac	K562	GSM733656	leukemia
H3K27ac	H1-hESC	GSM733718	embryonic stem cells
H3K27ac	GM12878	GSM733771	B-lymphocyte
H3K27ac	NHEK	GSM733674	keratinocyte
H3K4me3	K562	GSM733680	leukemia
H3K4me3	H1-hESC	GSM733657	embryonic stem cells
H3K4me3	GM12878	GSM733708	B-lymphocyte
H3K4me3	NHEK	GSM733720	keratinocyte

Table 5: Retrieved DNase-seq, FAIRE-seq and histone modification data from Encode project

The top ten percent of the CPD peaks in terms of fold enrichment were defined as hotspots of CPD peaks and the surrounding regions were analyzed for chromatin conformation. These CPD hotspots regions were used to calculate the heatmap of DNase-seq and FAIRE-seq signals in the surrounding of the CPD peak regions, which include an additional 5 kbps of up- and down-stream sequences around the centre of the peak (Figure 20. A). The whole region (10 kbps) was divided into 100 bins, each bin containing 100 bps. The tag density was calculated for each bin and tag density cannot be normalized to input samples since the tag density of the input samples is also affected by the chromatin accessibility. The profiles of the tag densities are presented in Figure 20. B.

For CPD hotspots in XPC^{-/-}, all compared cell lines show a consistent tendency that the chromatin in CPD hotspots is less accessible than flanking regions both in early and late repair time points (low colour intensity in the centre of the heat maps). This tendency is enhanced from 0.1 hour to 24 hours after UVC light, which is consistent with the above described analysis that CPDs are enriched in repetitive elements, which in turn form condensed chromatin regions. For both cell lines, the pyrimidine dimer density along the CPD hotspots was plotted as well and a significant increase of pyrimidine dimer density in the central of the CPD hotspots (Figure 20. C), which is correlated to decrease in chromatin accessibility. Directly after UVC exposure, the CPDs are immediately formed in less open

regions for both cell lines, which show lower levels in the DNase-seq and FAIRE-seq tracks (Figure 20. A). These are accompanied by higher pyrimidine dimer densities relative to the surrounding regions. It indicates that the DNA sequence primarily determines the CPD formation, even in condensed chromatin structures. However, from 0.1 hour to 24 hours, the tendency that CPD hotspots are in condensed chromatin gets enhanced in XPC^{-/-} cells due to the faster removal of CPDs in less condensed chromatin structures. CPDs in the condensed chromatin are not repaired as efficiently and therefore become more prominent in the later repair time-points. This tendency is consistent across different cell lines, which indicates that the positions of CPD hotspots are highly conserved. Especially, the chromatin compartment in pluripotent (H1hESC) and immortalized (NHEK) cells shares higher similarity than others such as differentiated cells (BJ, K563, GM12878), which was also reported in a previous studies that a potential link exists between chromatin accessibility and patterns of genomic variation (Thurman et al., 2012).

HaCaT cells show the same low accessibility of CPD hotspots in early time point, meaning that the majority of CPD hotspots are located in compact chromatin. But in contrast to XPC^{-/-} cells, HaCaT cells exhibit no enhancement of this tendency at late times due to the proficient GG-NER and TC-NER that is able to repair CPDs in both condensed and open chromatin. It is also shown that the pyrimidine dimer density in the central of CPD hotspots decreases from early to late repair time points (Figure 20. C. Left). Furthermore, RPKM values of DNase-seq and FAIRE-seq in CPD hotspots and flanking regions which are 5 kbps up and down-stream of hotspots were calculated as well (Figure 20. D). The boxplot of the RPKM values confirms that, in HaCaT cells, CPD peaks are located in slightly more condensed chromatin structures compared to their flanking regions. Whereas, in XPC^{-/-} cells, CPD hotspots show much less chromatin accessibility at 24 hours than 0.1 hour after UVC exposure.

Altogether, chromatin structure around CPD hotspots suggests that condensed chromatin cannot hinder the formation of CPDs, whereas CPDs in condensed chromatin are resistant to be repaired efficiently, especially in the XPC^{-/-} cell line, whereas in the HaCaT cells these CPDs are repaired.

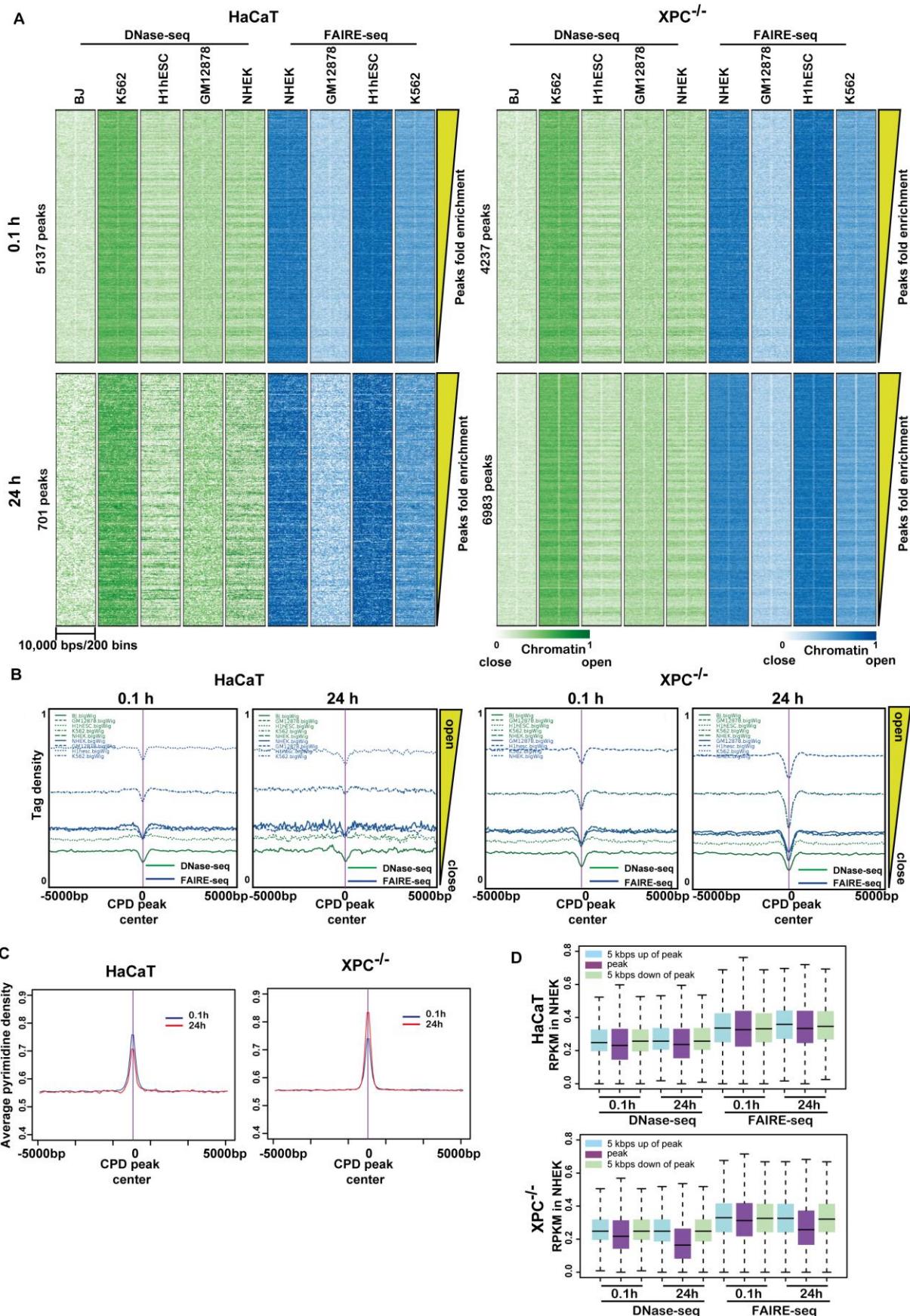


Figure 20. Chromatin structure around CPD peaks. (A) Tags density from DNase-seq (green) and FAIRE-seq (blue) in BJ (skin fibroblast), K562 (leukemia), H1hESC (embryonic stem cell), GM12878 (B-lymphocyte) and NHEK (keratinocyte) cell lines. Each row indicates one CPD peak

region extend to 10 kbps size and divided into 200 bins, each bin 50 bps length. The tags density was assigned to each bin and higher tag density presence with darker colour. (B) DNase-seq and FAIRE-seq tags density profile. CPD peak centre (purple line). HaCaT (left panels); XPC^{-/-} (right panels). Chromatin openness (yellow trigon) (C) Pyrimidine dimer density along CPD peaks and flanking regions in 100 bps resolution. 0.1h (blue); 24h (red). Left: HaCaT; Right: XPC^{-/-}. (D) Average RPKM of CPD peak, 5 kbps up peak and 5 kbps down peaks regions in DNase-seq and FAIRE-seq data. HaCaT (top); XPC^{-/-} (bottom). All boxes and whiskers represent 25-75 percentiles and 3x IQR, respectively.

Euchromatin histone modification marks are depleted and heterochromatin marker is slightly enriched in CPD peaks

As described previously, CPDs can be induced within condensed chromatin, which is usually marked with specific histone modification marks such as H3K9me3. To reveal the histone modification around CPD sites, the histone mark data for euchromatin (H3K4me3 and H3K27ac) and heterochromatin (H3K9me3) were retrieved from the GEO database (Table 5). The normalized RPKM values of these histone modification marks were analysed from 5 kbps upstream to 5 kbps downstream of each CPD peak location. This analysis was performed for the following cell lines NHEK, K562, GM12878 and H1hESC (Figure 21).

Since the length of the CPD peaks is around 1 kbps, which covers more than 5 nucleosomes, both peak and flanking regions shows similar histone modification levels due to the broad distribution of histone modifications. However, the euchromatin marks are consistently under-represented relative to the overall genome wide average in the CPD peaks and flanking regions. Conversely, the heterochromatin marker shows levels close to the genome-wide average. From early to late repair time, there is no significant difference in the HaCaT cell line. However, XPC^{-/-} cells exhibit lower euchromatin marker levels and higher heterochromatin marker levels in peaks than flanking regions from 0.1 hour to 24 hours after UVC exposure. This again validates that unrepaired CPDs are located inside of heterochromatic regions and are depleted in regions with euchromatin histone modifications. Yet data of histone modification markers were retrieved under physiological condition, hence the chromatin state might be dynamically responding to UV exposure. Therefore, the unrepaired CPDs within heterochromatin regions might be relocated to euchromatin for repair or might induce a change in the histone modification pattern.

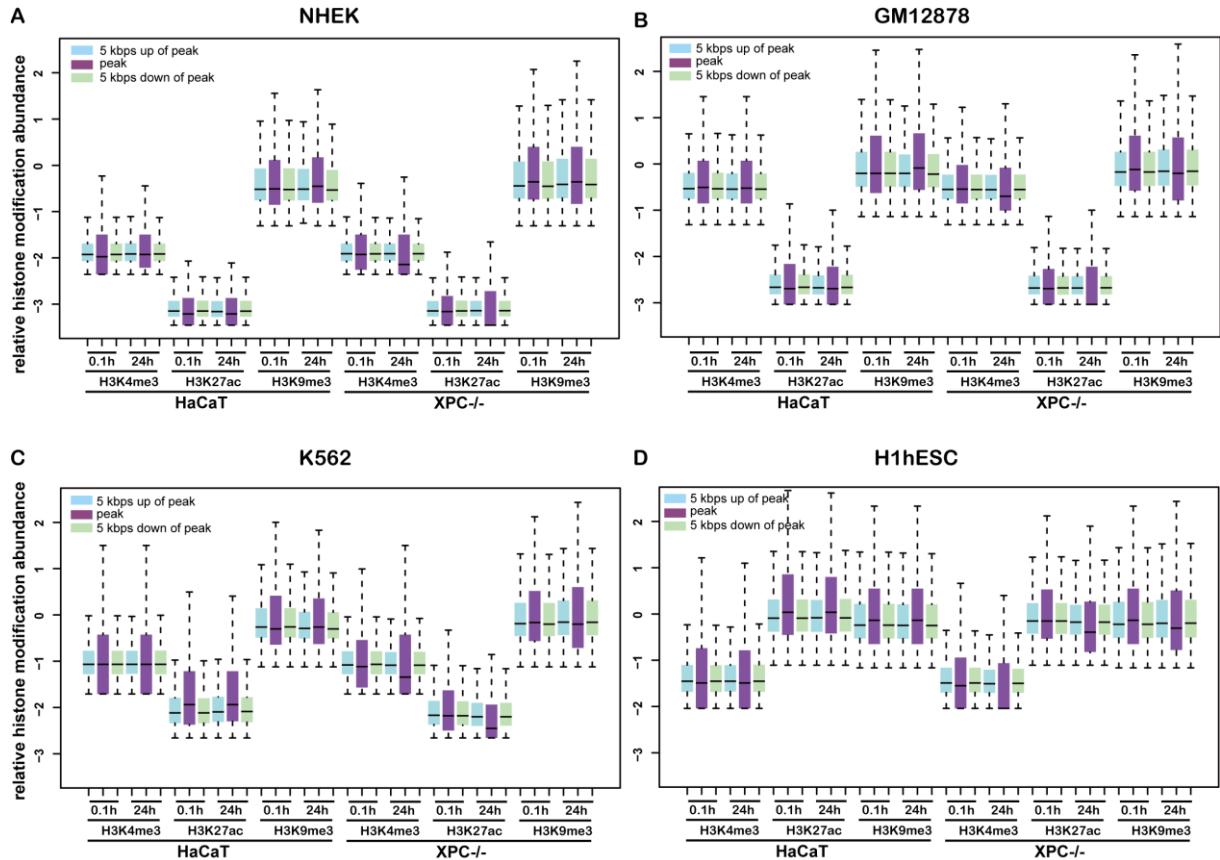


Figure 21. Histone modification level around CPD peaks. Cell lines: (A) NHEK (B) K562 (C) GM12878 (D) H1hESC. Left: HaCaT; Right: XPC^{-/-}. All boxes and whiskers represent 25-75 percentiles and 3x IQR, respectively.

5. Final conclusion and outlook

In this work I studied the distributions of H3, H2AX and γ H2AX under physiological conditions as well as repair kinetic of X-ray-induced DSBs in the human hepato-carcinoma cell line HepG2. H2AX and γ H2AX patterns show a non-random distribution in the human genome. Correlation analysis with multiple genomic features revealed that H2AX was enriched in active gene bodies. γ H2AX, in contrast, was underrepresented in these regions under physiological conditions, despite the high H2AX abundance. During early DNA repair, H2AX was promptly phosphorylated in transcriptional active chromatin. Conversely, at later repair time residual γ H2AX marked the heterochromatic compartments.

Since DSBs are critical damage to cells, NHEJ and HR pathways are involved in repairing these lesions in different cell cycle stages. However, cell cycle effects were not considered in our experiments. The presented results were obtained from unsynchronized cell populations that can utilize both HR to repair DSBs in late S and G2 phase and NHEJ for all phases. DSBs are repaired with high fidelity through HR whereas NHEJ repairs DSB with low fidelity that includes the risk of chromosomal translocations and genomic aberration. It is also known that 3D (three dimension) chromatin structure plays an important role for nuclear functions including DNA replication, translocation and gene expression. The chromatin structure changes after X-ray irradiation and IR induced DSBs do at least partially translocate to the outside of condensed chromatin regions for repair (Chiolo et al., 2011). Soria et al proposed a model that they termed prime repair and restore (Soria et al., 2012). If DSBs are induced in condensed chromatin, the level of condensation decreases quickly before the onset of DNA repair. After completion of repair, chromatin structure must be restored. Furthermore, the one dimension of DNA sequence might not provide enough evidence to infer genomic recombination sites. The genomic recombination may be due to the DSB in three dimensional proximity of chromatin structure (Jakob et al., 2009). Therefore, it is important to elucidate the different levels of 3D chromatin structure before irradiation and during DNA repair. It is necessary to figure out whether nucleosomes are depleted, how the 3D chromatin conformation changes, when and how the cells restore the original 3D structure.

Until now, most of researchers focused on revealing the proteins and pathways involved in chromatin remodelling, but there is no published data available for genome-wide chromatin remodelling in response to irradiation. Therefore, the genome-wide chromatin remodelling data might help to understand how cells utilized chromatin remodelling to repair damage. Our high-resolution image analysis of HepG2 cells upon IR at different times revealed that residual γ H2AX form clusters at late repair time points (data shown in manuscript). So far there exists no evidence to answer how these residual DSBs contribute to IR induced

translocation and thus further contribute to genomic aberration through the NHEJ pathway during repair process. Another question is whether the depletion of any DSB repair protein (eg. BRCA1) can increase the amount of recombination sites, especially in leukemia and lymphomas cells.

Not only the profiles of IR induced DSB were investigated, but also the comprehensive profiles of CPDs in repair deficient and proficient cell lines were produced and analyzed. The results reveal potential mechanism how CPDs are preferentially distributed in NER proficient and GG-NER deficient cell lines on a genome-wide scale. The modified ssDIP-seq technique was developed and approved as a valuable approach to identify the CPD location in a strand-specific manner. This method shows that CPDs are preferentially induced and delayed repaired in the anti-sense strand of both cell lines. The induction of CPDs in the sense strand is anti-correlated with gene expression levels. Low CPD densities are induced in highly expressed genes and vice versa. Genome-wide distribution of CPDs is not homogeneous and random, conversely CPDs are over-represented in repetitive elements especially in microsatellite and low complexity elements. CPD hotspots are located in highly condensed chromatin structure and euchromatin marks around CPD are under-represented, whereas heterochromatin marks are close to the genome average level. All these result provides hints that, upon UVC exposure, unrepaired CPDs in non-genic regions and anti-sense strand contribute to oncogenesis, which might be part of driving cells into a transformed state through the induction of mutagenesis, microsatellite instability and genomic aberration.

However, there are numerous challenges remaining for a complete understanding of the mechanism of repair and consequence of unrepaired CPDs. Based on our results, it is proposed that the induction of CPDs is mostly sequence dependent in non-genic regions but this effect is overridden by the molecular environment in genic regions. The target sequences are preferentially located in heterochromatin, repetitive, non-genic and AT rich regions. The condensed chromatin structure does not block the formation of damage, with CPDs are found in the centre of condensed chromatin. This tendency is decreased in repair proficient HaCaT cells, whereas enhanced in GG-NER deficient cells. The speculation is that HaCaT can trigger chromatin remodelling to access CPD inside of condensed chromatin and initiate the repair pathway to repair. Whereas, due to the defect of GG-NER and proficiency in TC-NER, XPC^{-/-} cells can only sense and repair CPD in the transcribed strand of genic regions but not in the non-transcribed strand and non-genic regions, which make up a large portion of genome. The CPDs are repaired in open chromatin but not in the condensed chromatin. Upon UV exposure, chromatin needs to be remodelled. This turns condensed chromatin into open chromatin in yeast and this process is related to histone H3 hyperacetylation (Yu et al.,

2005). The damage sensor DDB2 can promotes chromatin decondensation at UV-induced DNA damage (Luijsterburg et al., 2012), but no reports exists that XPC is essential for chromatin remodelling. However, DDB2 binds to damage sites, then recruits XPC. Therefore it is speculated that, in non-genic and heterochromatin regions, chromatin structure is altered around the CPD site by DDB2. Due to the defect of the XPC protein, CPDs, which are located in non-transcribed regions can not be repaired. This raises the questions whether cells can restore the chromatin structure if CPDs are not repaired. If not, what is the consequence of permanently decondensed heterochromatin, which might contribute to carcinogenesis. Similarly it is reported, that deficiency of the tumor suppressor protein BRCA1 causes the disruption of global heterochromatin integrity and leads to transcriptional over-expression of the tandem repeated satellite DNA and finally result in cancer (Zhu et al., 2011). Therefore, it can be speculated that XPC plays a similar role that maintains genomic stability while otherwise the chromatin structural aberration contributes to a high risk of cancer in XP patients. In contrast to $XPC^{-/-}$ cells, $CSB^{-/-}$ cells cannot repair CPDs in the transcribed strand of active genes, which is occupied by RNA polymerase II for transcription. The encounter of CPDs with RNA polymerase II will cause a stalling of the latter and normally recruits CSA and CSB to the damage sites. If CSB is defect, CPDs cannot be repaired by TC-NER but only GG-NER. The consequence of delayed repair of CPD in transcribed strands is poorly understood. Since the higher level of unrepaired CPDs is found in the anti-sense strand, it's proposed that most mutations in genic regions originated from unrepaired CPDs in the anti-sense strand, which are propagated to the sense-strand during DNA replication. This suggests that CPD repair in anti-sense strand is equally important. A high number of CPDs in the sense strand will lead to a cytotoxic effect, e.g. apoptosis, while a high level of CPD in the anti-sense strand will cause mutations.

There are also a lot of open questions that need to be addressed such as whether RNA polymerase II will be degraded or released from the transcribed strand and which replication and/or transcription blocked genes are critical for the activation of apoptosis. Upon UV exposure, $CSB^{-/-}$ triggers repression or transcription arrest in critical genes that activate apoptosis, which might contribute to exhausting of stem cells and lead to aging (Lopez-Otin et al., 2013). In contrast $XPC^{-/-}$ induces mutation instead of apoptosis, which cells can still survive through translesion synthesis pathways and propagate mutations to next the generation (Marteijn et al., 2014). And it's also unclear how the chromatin remodelling response to UV induced damage in repair deficient cells. To answer this question, micro-pore experiments will be performed on HaCaT, $XPC^{-/-}$ and $CSB^{-/-}$ cells with local UVC exposure (Suzuki et al., 2011). After staining CPDs at different time points after UVC, the intensity of heterochromatin marks and DAPI within these pores and the size of the CPD area will be calculated from early to late repair time points. Due to the spontaneous chromatin mobility, in

HaCaT cells, which are proficient in NER repair, the size and edge of the induced CPDs in micro-pore will disappear much faster than in repair deficient cells. If the deficient cell lines have no ability to relax condensed chromatin to sensor damage and are disabled to restore the chromatin structure since the damage is not repaired, the spot of induced CPDs within the micro-pore will disappear at a later time point compared to the proficient cells. The distribution of heterochromatin and euchromatin marks within the micro-pore regions will be investigated as well.

	γ H2AX	CPD
Physiological condition	Enriched in EC and under-represented in HC	None
Early repair time	Promptly formed in EC	Immediately formed and enriched in repetitive elements
Late repair time	Retained in HC	Repaired in HaCaT cells and retained in repetitive elements of XPC ^{-/-} cells
Strand specific	Undetermined	High in anti-sense strand and low in sense strand
Expression influence	Intermediately expressed genes with the highest level at early repair time	No difference in anti-sense strand and anti-correlated in sense strand
Repair kinetics	Fast repair in EC, delayed repair in HC	Repaired in genic and non-genic in HaCaT cells; Only repaired in sense strand of XPC ^{-/-}
Sequence motif	Undetermined	Continuous pyrimidine dimer
GC content	Correlated at early and anti-correlated at late	Anti-correlated

Table 6: Summary of the features of the genome-wide distribution of γ H2AX and CPD and their repair kinetics.

The comparisons of γ H2AX and CPD distribution and repair kinetics are summarized in Table 6. Overall, this work provides a genome-wide landscape to reveal DSB repair kinetics of cancer cells upon ionizing radiation and CPD repair kinetics in NER repair proficient and deficient cell lines after exposure to UVC. Two potential mechanisms are revealed that the unrepaired CPD in microsatellite and anti-sense strand might contribute to carcinogenesis.

6. Reference

- (2004). The ENCODE (ENCylopedia Of DNA Elements) Project. *Science* 306, 636-640.
- Aalto, Y., Eriksson, L., Seregard, S., Larsson, O., and Knuutila, S. (2001). Concomitant loss of chromosome 3 and whole arm losses and gains of chromosome 1, 6, or 8 in metastasizing primary uveal melanoma. *Investigative ophthalmology & visual science* 42, 313-317.
- Arlt, M.F., and Glover, T.W. (2010). Inhibition of topoisomerase I prevents chromosome breakage at common fragile sites. *DNA repair* 9, 678-689.
- Asaithamby, A., Hu, B., and Chen, D.J. (2011). Unrepaired clustered DNA lesions induce chromosome breakage in human cells. *Proceedings of the National Academy of Sciences of the United States of America* 108, 8293-8298.
- Bagga, S., Bracht, J., Hunter, S., Massirer, K., Holtz, J., Eachus, R., and Pasquinelli, A.E. (2005). Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell* 122, 553-563.
- Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature reviews Genetics* 12, 745-755.
- Barsotti, A.M., Beckerman, R., Laptenko, O., Huppi, K., Caplen, N.J., and Prives, C. (2012). p53-Dependent induction of PVT1 and miR-1204. *The Journal of biological chemistry* 287, 2509-2519.
- Bartek, J., and Lukas, J. (2001). Mammalian G1- and S-phase checkpoints in response to DNA damage. *Curr Opin Cell Biol* 13, 738-747.
- Bastian, B.C., LeBoit, P.E., Hamm, H., Brocker, E.B., and Pinkel, D. (1998). Chromosomal gains and losses in primary cutaneous melanomas detected by comparative genomic hybridization. *Cancer research* 58, 2170-2175.
- Batista, L.F., Kaina, B., Meneghini, R., and Menck, C.F. (2009). How DNA lesions are turned into powerful killing structures: insights from UV-induced apoptosis. *Mutation research* 681, 197-208.
- Bauer, J., and Bastian, B.C. (2006). Distinguishing melanocytic nevi from melanoma by DNA copy number changes: comparative genomic hybridization as a research and diagnostic tool. *Dermatol Ther* 19, 40-49.
- Bauerschmidt, C., Arrichiello, C., Burdak-Rothkamm, S., Woodcock, M., Hill, M.A., Stevens, D.L., and Rothkamm, K. (2010). Cohesin promotes the repair of ionizing radiation-induced DNA double-strand breaks in replicated chromatin. *Nucleic acids research* 38, 477-487.
- Becker, M.M., and Wang, Z. (1989). Origin of ultraviolet damage in DNA. *J Mol Biol* 210, 429-438.
- Bellacosa, A. (2001). Role of MED1 (MBD4) Gene in DNA repair and human cancer. *Journal of cellular physiology* 187, 137-144.
- Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53-59.
- Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature* 463, 899-905.
- Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2012). Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* 28, 423-425.
- Bohr, V.A., Smith, C.A., Okumoto, D.S., and Hanawalt, P.C. (1985). DNA repair in an active gene: removal of pyrimidine dimers from the DHFR gene of CHO cells is much more efficient than in the genome overall. *Cell* 40, 359-369.
- Boland, C.R., and Goel, A. (2010). Microsatellite instability in colorectal cancer. *Gastroenterology* 138, 2073-2087 e2073.
- Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132, 311-322.
- Bras, J.M., and Singleton, A.B. (2011). Exome sequencing in Parkinson's disease. *Clinical genetics* 80, 104-109.
- Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L., and Weber, J.L. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. *American journal of human genetics* 63, 861-869.
- Bross, L., Fukita, Y., McBlane, F., Demolliere, C., Rajewsky, K., and Jacobs, H. (2000). DNA double-strand breaks in immunoglobulin genes undergoing somatic hypermutation. *Immunity* 13, 589-597.
- Brueckner, F., Hennecke, U., Carell, T., and Cramer, P. (2007). CPD damage recognition by transcribing RNA polymerase II. *Science* 315, 859-862.
- Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development* 25, 1915-1927.
- Cadoret, J.C., Meisch, F., Hassan-Zadeh, V., Luyten, I., Guillet, C., Duret, L., Quesneville, H., and Prioleau, M.N. (2008). Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proceedings of the National Academy of Sciences of the United States of America* 105, 15837-15842.
- Caron, P., Aymard, F., Iacovoni, J.S., Briois, S., Canitrot, Y., Bugler, B., Massip, L., Losada, A., and Legube, G. (2012). Cohesin protects genes against gammaH2AX Induced by DNA double-strand breaks. *PLoS genetics* 8, e1002460.

- Carramus, L., Contino, F., Ferro, A., Minafra, L., Perconti, G., Giallongo, A., and Feo, S. (2007). The PVT-1 oncogene is a Myc protein target that is overexpressed in transformed cells. *Journal of cellular physiology* 213, 511-518.
- Carty, M.P., Lawrence, C.W., and Dixon, K. (1996). Complete replication of plasmid DNA containing a single UV-induced lesion in human cell extracts. *The Journal of biological chemistry* 271, 9637-9647.
- Cer, R.Z., Bruce, K.H., Mudunuri, U.S., Yi, M., Volfovsky, N., Luke, B.T., Bacolla, A., Collins, J.R., and Stephens, R.M. (2011). Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic acids research* 39, D383-391.
- Chan, D.W., Chen, B.P., Prithivirajsingh, S., Kurimasa, A., Story, M.D., Qin, J., and Chen, D.J. (2002). Autophosphorylation of the DNA-dependent protein kinase catalytic subunit is required for rejoining of DNA double-strand breaks. *Genes & development* 16, 2333-2338.
- Chapman, J.R., Taylor, M.R., and Boulton, S.J. (2012). Playing the end game: DNA double-strand break repair pathway choice. *Molecular cell* 47, 497-510.
- Cheo, D.L., Burns, D.K., Meira, L.B., Houle, J.F., and Friedberg, E.C. (1999). Mutational inactivation of the xeroderma pigmentosum group C gene confers predisposition to 2-acetylaminofluorene-induced liver and lung cancer and to spontaneous testicular cancer in Trp53^{-/-} mice. *Cancer research* 59, 771-775.
- Chiolo, I., Minoda, A., Colmenares, S.U., Polyzos, A., Costes, S.V., and Karpen, G.H. (2011). Double-strand breaks in heterochromatin move outside of a dynamic HP1a domain to complete recombinational repair. *Cell* 144, 732-744.
- Cleaver, J.E., Lam, E.T., and Revet, I. (2009). Disorders of nucleotide excision repair: the genetic and molecular basis of heterogeneity. *Nature reviews Genetics* 10, 756-768.
- Consortium, E.P., Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74.
- Cooke, M.S., Evans, M.D., Dizdaroglu, M., and Lunec, J. (2003). Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB J* 17, 1195-1214.
- Costa, R.M., Chigancas, V., Galhardo Rda, S., Carvalho, H., and Menck, C.F. (2003). The eukaryotic nucleotide excision repair pathway. *Biochimie* 85, 1083-1099.
- Cowell, I.G., Sunter, N.J., Singh, P.B., Austin, C.A., Durkacz, B.W., and Tilby, M.J. (2007). gammaH2AX foci form preferentially in euchromatin after ionising-radiation. *PLoS one* 2, e1057.
- Creighton, C.J., Reid, J.G., and Gunaratne, P.H. (2009). Expression profiling of microRNAs by deep sequencing. *Briefings in bioinformatics* 10, 490-497.
- Curtin, J.A., Fridlyand, J., Kageshita, T., Patel, H.N., Busam, K.J., Kutzner, H., Cho, K.H., Aiba, S., Brocker, E.B., LeBoit, P.E., et al. (2005). Distinct sets of genetic alterations in melanoma. *The New England journal of medicine* 353, 2135-2147.
- Dai, Q.S., Hua, R.X., Zeng, R.F., Long, J.T., and Peng, Z.W. (2014). XPC gene polymorphisms contribute to bladder cancer susceptibility: a meta-analysis. *Tumour Biol* 35, 447-453.
- de Boer, J., and Hoeijmakers, J.H. (2000). Nucleotide excision repair and human syndromes. *Carcinogenesis* 21, 453-460.
- De Bont, R., and van Larebeke, N. (2004). Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* 19, 169-185.
- de Koning, A.P., Gu, W., Castoe, T.A., Batzer, M.A., and Pollock, D.D. (2011). Repetitive elements may comprise over two-thirds of the human genome. *PLoS genetics* 7, e1002384.
- de Laat, W.L., Jaspers, N.G., and Hoeijmakers, J.H. (1999). Molecular mechanism of nucleotide excision repair. *Genes & development* 13, 768-785.
- Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., et al. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380, 152-154.
- Donahue, B.A., Yin, S., Taylor, J.S., Reines, D., and Hanawalt, P.C. (1994). Transcript cleavage by RNA polymerase II arrested by a cyclobutane pyrimidine dimer in the DNA template. *Proceedings of the National Academy of Sciences of the United States of America* 91, 8502-8506.
- Duval, A., Gayet, J., Zhou, X.P., Iacopetta, B., Thomas, G., and Hamelin, R. (1999). Frequent frameshift mutations of the TCF-4 gene in colorectal cancers with microsatellite instability. *Cancer research* 59, 4213-4215.
- Emmert, S., Kobayashi, N., Khan, S.G., and Kraemer, K.H. (2000). The xeroderma pigmentosum group C gene leads to selective repair of cyclobutane pyrimidine dimers rather than 6-4 photoproducts. *Proceedings of the National Academy of Sciences of the United States of America* 97, 2151-2156.
- Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shores, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43-49.
- Fousteri, M., and Mullenders, L.H. (2008). Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell research* 18, 73-84.
- Fungtammasan, A., Walsh, E., Chiaramonte, F., Eckert, K.A., and Makova, K.D. (2012). A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome? *Genome research* 22, 993-1005.
- Gao, S., Drouin, R., and Holmquist, G.P. (1994). DNA repair rates mapped along the human PGK1 gene at nucleotide resolution. *Science* 263, 1438-1440.

- Garinis, G.A., Mitchell, J.R., Moorhouse, M.J., Hanada, K., de Waard, H., Vandeputte, D., Jans, J., Brand, K., Smid, M., van der Spek, P.J., et al. (2005). Transcriptome analysis reveals cyclobutane pyrimidine dimers as a major source of UV-induced DNA breaks. *The EMBO journal* 24, 3952-3962.
- Geuting, V., Reul, C., and Lobrich, M. (2013). ATM release at resected double-strand breaks provides heterochromatin reconstitution to facilitate homologous recombination. *PLoS genetics* 9, e1003667.
- Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., and Lieb, J.D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome research* 17, 877-885.
- Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R.J., Freedman, B.I., Quinones, M.P., Bamshad, M.J., et al. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* 307, 1434-1440.
- Griffiths-Jones, S., Saini, H.K., van Dongen, S., and Enright, A.J. (2008). miRBase: tools for microRNA genomics. *Nucleic acids research* 36, D154-158.
- Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 453, 948-951.
- Hanawalt, P.C. (1991). Heterogeneity of DNA repair at the gene level. *Mutation research* 247, 203-211.
- Hanawalt, P.C. (2002). Subpathways of nucleotide excision repair and their regulation. *Oncogene* 21, 8949-8956.
- Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., and Stamatoyannopoulos, J.A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences of the United States of America* 107, 139-144.
- Hastings, P.J., Lupski, J.R., Rosenberg, S.M., and Ira, G. (2009). Mechanisms of change in gene copy number. *Nature reviews Genetics* 10, 551-564.
- Hauser, J., Seidman, M.M., Sidur, K., and Dixon, K. (1986). Sequence specificity of point mutations induced during passage of a UV-irradiated shuttle vector plasmid in monkey cells. *Molecular and cellular biology* 6, 277-285.
- Hoeijmakers, J.H. (2001). Genome maintenance mechanisms for preventing cancer. *Nature* 411, 366-374.
- Hoeijmakers, J.H. (2007). Genome maintenance mechanisms are critical for preventing cancer as well as other aging-associated diseases. *Mechanisms of ageing and development* 128, 460-462.
- Hoeijmakers, J.H. (2009). DNA damage, aging, and cancer. *The New England journal of medicine* 361, 1475-1485.
- Hollander, M.C., Philburn, R.T., Patterson, A.D., Velasco-Miguel, S., Friedberg, E.C., Linnoila, R.I., and Fornace, A.J., Jr. (2005). Deletion of XPC leads to lung tumors in mice and is associated with early events in human lung carcinogenesis. *Proceedings of the National Academy of Sciences of the United States of America* 102, 13200-13205.
- Huang, K. (2011). Exome sequencing expedites disease gene discovery. *Clinical genetics* 80, 133-134.
- Hussein, M.R., and Wood, G.S. (2002). Microsatellite instability and its relevance to cutaneous tumorigenesis. *J Cutan Pathol* 29, 257-267.
- Iacoboni, J.S., Caron, P., Lassadi, I., Nicolas, E., Massip, L., Trouche, D., and Legube, G. (2010). High-resolution profiling of gammaH2AX around DNA double strand breaks in the mammalian genome. *The EMBO journal* 29, 1446-1457.
- Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature genetics* 36, 949-951.
- Ikehata, H., and Ono, T. (2011). The mechanisms of UV mutagenesis. *J Radiat Res* 52, 115-125.
- Ikura, T., Tashiro, S., Kakino, A., Shima, H., Jacob, N., Amunugama, R., Yoder, K., Izumi, S., Kuraoka, I., Tanaka, K., et al. (2007). DNA damage-dependent acetylation and ubiquitination of H2AX enhances chromatin dynamics. *Molecular and cellular biology* 27, 7028-7040.
- Jackson, S.P. (2002). Sensing and repairing DNA double-strand breaks. *Carcinogenesis* 23, 687-696.
- Jakob, B., Splinter, J., Conrad, S., Voss, K.O., Zink, D., Durante, M., Lobrich, M., and Taucher-Scholz, G. (2011). DNA double-strand breaks in heterochromatin elicit fast repair protein recruitment, histone H2AX phosphorylation and relocation to euchromatin. *Nucleic acids research* 39, 6489-6499.
- Jakob, B., Splinter, J., Durante, M., and Taucher-Scholz, G. (2009). Live cell microscopy analysis of radiation-induced DNA double-strand break motion. *Proceedings of the National Academy of Sciences of the United States of America* 106, 3172-3177.
- Janssen, A., van der Burg, M., Szuhai, K., Kops, G.J., and Medema, R.H. (2011). Chromosome segregation errors as a cause of DNA damage and structural chromosome aberrations. *Science* 333, 1895-1898.
- Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316, 1497-1502.
- Jost, K.L., Bertulat, B., and Cardoso, M.C. (2012). Heterochromatin and gene positioning: inside, outside, any side? *Chromosoma* 121, 555-563.
- Kappes, U.P., Luo, D., Potter, M., Schulmeister, K., and Runger, T.M. (2006). Short- and long-wave UV light (UVB and UVA) induce similar mutations in human skin cells. *J Invest Dermatol* 126, 667-675.
- Karnani, N., Taylor, C.M., Malhotra, A., and Dutta, A. (2010). Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection. *Molecular biology of the cell* 21, 393-404.
- Kharchenko, P.V., Tolstorukov, M.Y., and Park, P.J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature biotechnology* 26, 1351-1359.

- Kim, J.A., Kruhlak, M., Dotiwala, F., Nussenzweig, A., and Haber, J.E. (2007). Heterochromatin is refractory to gamma-H2AX modification in yeast and mammals. *J Cell Biol* 178, 209-218.
- Klaunig, J.E., Kamendulis, L.M., and Hocevar, B.A. (2010). Oxidative stress and oxidative damage in carcinogenesis. *Toxicol Pathol* 38, 96-109.
- Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X.S., and Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature genetics* 41, 376-381.
- Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. (2002). A high-resolution recombination map of the human genome. *Nature genetics* 31, 241-247.
- Kouzarides, T. (2007). Chromatin modifications and their function. *Cell* 128, 693-705.
- Kraemer, K.H., Patronas, N.J., Schiffmann, R., Brooks, B.P., Tamura, D., and DiGiovanna, J.J. (2007). Xeroderma pigmentosum, trichothiodystrophy and Cockayne syndrome: a complex genotype-phenotype relationship. *Neuroscience* 145, 1388-1396.
- Kuehner, J.N., Pearson, E.L., and Moore, C. (2011). Unravelling the means to an end: RNA polymerase II transcription termination. *Nat Rev Mol Cell Biol* 12, 283-294.
- Kuo, L.J., and Yang, L.X. (2008). Gamma-H2AX - a novel biomarker for DNA double-strand breaks. *In Vivo* 22, 305-309.
- Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., and Jenuwein, T. (2001). Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* 410, 116-120.
- Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research* 22, 1813-1831.
- Lans, H., Marteijn, J.A., and Vermeulen, W. (2012). ATP-dependent chromatin remodeling in the DNA-damage response. *Epigenetics Chromatin* 5, 4.
- Larkin, D.M., Pape, G., Donthu, R., Auvil, L., Welge, M., and Lewin, H.A. (2009). Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome research* 19, 770-777.
- Legerski, R., and Peterson, C. (1992). Expression cloning of a human DNA repair gene involved in xeroderma pigmentosum group C. *Nature* 360, 610.
- Lehmann, A.R. (2003). DNA repair-deficient diseases, xeroderma pigmentosum, Cockayne syndrome and trichothiodystrophy. *Biochimie* 85, 1101-1111.
- Li, L., Bales, E.S., Peterson, C.A., and Legerski, R.J. (1993). Characterization of molecular defects in xeroderma pigmentosum group C. *Nature genetics* 5, 413-417.
- Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., et al. (2010). The sequence and de novo assembly of the giant panda genome. *Nature* 463, 311-317.
- Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25, 1966-1967.
- Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology* 2012, 251364.
- Ljungman, M., and Zhang, F. (1996). Blockage of RNA polymerase as a possible trigger for u.v. light-induced apoptosis. *Oncogene* 13, 823-831.
- Lo, H.L., Nakajima, S., Ma, L., Walter, B., Yasui, A., Ethell, D.W., and Owen, L.B. (2005). Differential biologic effects of CPD and 6-4PP UV-induced DNA damage on the induction of apoptosis and cell-cycle arrest. *BMC Cancer* 5, 135.
- Lobrich, M., and Jeggo, P.A. (2007). The impact of a negligent G2/M checkpoint on genomic instability and cancer induction. *Nat Rev Cancer* 7, 861-869.
- Lobrich, M., Shibata, A., Beucher, A., Fisher, A., Ensminger, M., Goodarzi, A.A., Barton, O., and Jeggo, P.A. (2010). gammaH2AX foci analysis for monitoring DNA double-strand break repair: strengths, limitations and optimization. *Cell Cycle* 9, 662-669.
- Lopes, M., Cotta-Ramusino, C., Pellicioli, A., Liberi, G., Plevani, P., Muzi-Falconi, M., Newlon, C.S., and Foiani, M. (2001). The DNA replication checkpoint response stabilizes stalled replication forks. *Nature* 412, 557-561.
- Lopez-Otin, C., Blasco, M.A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. *Cell* 153, 1194-1217.
- Luijsterburg, M.S., Lindh, M., Acs, K., Vrouwe, M.G., Pines, A., van Attikum, H., Mullenders, L.H., and Dantuma, N.P. (2012). DDB2 promotes chromatin decondensation at UV-induced DNA damage. *J Cell Biol* 197, 267-281.
- Luo, C., Tsementzi, D., Kyripides, N., Read, T., and Konstantinidis, K.T. (2012). Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS one* 7, e30087.
- Mardis, E.R. (2008). Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 9, 387-402.
- Mardis, E.R. (2011). A decade's perspective on DNA sequencing technology. *Nature* 470, 198-203.
- Marteijn, J.A., Lans, H., Vermeulen, W., and Hoeijmakers, J.H. (2014). Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat Rev Mol Cell Biol* 15, 465-481.
- Mayne, L.V., Priestley, A., James, M.R., and Burke, J.F. (1986). Efficient immortalization and morphological transformation of human fibroblasts by transfection with SV40 DNA linked to a dominant marker. *Exp Cell Res* 162, 530-538.
- Mei Kwei, J.S., Kuraoka, I., Horibata, K., Ubukata, M., Kobatake, E., Iwai, S., Handa, H., and Tanaka, K. (2004). Blockage of RNA polymerase II at a cyclobutane pyrimidine dimer and 6-4 photoproduct. *Biochemical and biophysical research communications* 320, 1133-1138.

- Mellon, I., Bohr, V.A., Smith, C.A., and Hanawalt, P.C. (1986). Preferential DNA repair of an active gene in human cells. *Proceedings of the National Academy of Sciences of the United States of America* 83, 8878-8882.
- Mellon, I., Spivak, G., and Hanawalt, P.C. (1987). Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian DHFR gene. *Cell* 51, 241-249.
- Metzker, M.L. (2010). Sequencing technologies - the next generation. *Nature reviews Genetics* 11, 31-46.
- Misteli, T., and Soutoglou, E. (2009). The emerging role of nuclear architecture in DNA repair and genome maintenance. *Nat Rev Mol Cell Biol* 10, 243-254.
- Morin, R.D., O'Connor, M.D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., et al. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome research* 18, 610-621.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* 5, 621-628.
- Myers, S., Freeman, C., Auton, A., Donnelly, P., and McVean, G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature genetics* 40, 1124-1129.
- Nagy, P.L., Cleary, M.L., Brown, P.O., and Lieb, J.D. (2003). Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. *Proceedings of the National Academy of Sciences of the United States of America* 100, 6364-6369.
- Negrini, S., Gorgoulis, V.G., and Halazonetis, T.D. (2010). Genomic instability--an evolving hallmark of cancer. *Nat Rev Mol Cell Biol* 11, 220-228.
- Ng, P.C., and Kirkness, E.F. (2010). Whole genome sequencing. *Methods Mol Biol* 628, 215-226.
- Niedernhofer, L.J., Bohr, V.A., Sander, M., and Kraemer, K.H. (2011). Xeroderma pigmentosum and other diseases of human premature aging and DNA repair: molecules to patients. *Mechanisms of ageing and development* 132, 340-347.
- Oberdoerffer, P., and Sinclair, D.A. (2007). The role of nuclear architecture in genomic instability and ageing. *Nat Rev Mol Cell Biol* 8, 692-702.
- Oda, S., Maehara, Y., Ikeda, Y., Oki, E., Egashira, A., Okamura, Y., Takahashi, I., Kakeji, Y., Sumiyoshi, Y., Miyashita, K., et al. (2005). Two modes of microsatellite instability in human cancer: differential connection of defective DNA mismatch repair to dinucleotide repeat instability. *Nucleic acids research* 33, 1628-1636.
- Olive, P.L. (1998). The role of DNA single- and double-strand breaks in cell killing by ionizing radiation. *Radiat Res* 150, S42-51.
- Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews Genetics* 10, 669-680.
- Paull, T.T., Rogakou, E.P., Yamazaki, V., Kirchgessner, C.U., Gellert, M., and Bonner, W.M. (2000). A critical role for histone H2AX in recruitment of repair factors to nuclear foci after DNA damage. *Current biology : CB* 10, 886-895.
- Pfeifer, G.P., You, Y.H., and Besaratinia, A. (2005). Mutations induced by ultraviolet light. *Mutation research* 571, 19-31.
- Price, B.D., and D'Andrea, A.D. (2013). Chromatin remodeling at DNA double-strand breaks. *Cell* 152, 1344-1354.
- Priya, R.R., Rajasimha, H.K., Brooks, M.J., and Swaroop, A. (2012). Exome sequencing: capture and sequencing of all human coding regions for disease gene discovery. *Methods Mol Biol* 884, 335-351.
- Protic-Sabljic, M., Tuteja, N., Munson, P.J., Hauser, J., Kraemer, K.H., and Dixon, K. (1986). UV light-induced cyclobutane pyrimidine dimers are mutagenic in mammalian cells. *Molecular and cellular biology* 6, 3349-3356.
- Qu, H., and Fang, X. (2013). A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project. *Genomics Proteomics Bioinformatics* 11, 135-141.
- Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. *Nature* 444, 444-454.
- Robertson, A.B., Klungland, A., Rognes, T., and Leiros, I. (2009). DNA repair in mammalian cells: Base excision repair: the long and short of it. *Cell Mol Life Sci* 66, 981-993.
- Rogakou, E.P., Pilch, D.R., Orr, A.H., Ivanova, V.S., and Bonner, W.M. (1998). DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *The Journal of biological chemistry* 273, 5858-5868.
- Rothkamm, K., Kruger, I., Thompson, L.H., and Lobrich, M. (2003). Pathways of DNA double-strand break repair during the mammalian cell cycle. *Molecular and cellular biology* 23, 5706-5715.
- Rozowsky, J., Euskirchen, G., Auerbach, R.K., Zhang, Z.D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M.B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology* 27, 66-75.
- Sancar, A. (2003). Structure and function of DNA photolyase and cryptochrome blue-light photoreceptors. *Chem Rev* 103, 2203-2237.
- Schaeffer, L., Roy, R., Humbert, S., Moncollin, V., Vermeulen, W., Hoeijmakers, J.H., Chambon, P., and Egly, J.M. (1993). DNA repair helicase: a component of BTF2 (TFIILH) basic transcription factor. *Science* 260, 58-63.
- Scharer, O.D. (2013). Nucleotide excision repair in eukaryotes. *Cold Spring Harb Perspect Biol* 5, a012609.
- Scicchitano, D.A., and Hanawalt, P.C. (1989). Repair of N-methylpurines in specific DNA sequences in Chinese hamster ovary cells: absence of strand specificity in the dihydrofolate reductase gene. *Proceedings of the National Academy of Sciences of the United States of America* 86, 3050-3054.
- Scrima, A., Konickova, R., Czyzewski, B.K., Kawasaki, Y., Jeffrey, P.D., Groisman, R., Nakatani, Y., Iwai, S., Pavletich, N.P., and Thoma, N.H. (2008). Structural basis of UV DNA-damage recognition by the DDB1-DDB2 complex. *Cell* 135, 1213-1223.
- Setlow, R.B., Swenson, P.A., and Carrier, W.L. (1963). Thymine Dimers and Inhibition of DNA Synthesis by Ultraviolet Irradiation of Cells. *Science* 142, 1464-1466.

- Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology* 26, 1135-1145.
- Simon, J.M., Giresi, P.G., Davis, I.J., and Lieb, J.D. (2012). Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nature protocols* 7, 256-267.
- Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews Genetics* 15, 121-132.
- Smit, A.F.A.H., R.; Green, P. (1996-2010). RepeatMasker Open-3.0.
- Song, L., and Crawford, G.E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* 2010, pdb prot5384.
- Song, L., Zhang, Z., Grasfeder, L.L., Boyle, A.P., Giresi, P.G., Lee, B.K., Sheffield, N.C., Graf, S., Huss, M., Keefe, D., et al. (2011). Open chromatin defined by DNasel and FAIRE identifies regulatory elements that shape cell-type identity. *Genome research* 21, 1757-1767.
- Soria, G., Polo, S.E., and Almouzni, G. (2012). Prime, repair, restore: the active role of chromatin in the DNA damage response. *Molecular cell* 46, 722-734.
- Stankiewicz, P., and Lupski, J.R. (2010). Structural variation in the human genome and its role in disease. *Annual review of medicine* 61, 437-455.
- Subramanian, S., Mishra, R.K., and Singh, L. (2003). Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome biology* 4, R13.
- Sugasawa, K., Ng, J.M., Masutani, C., Iwai, S., van der Spek, P.J., Eker, A.P., Hanaoka, F., Bootsma, D., and Hoeijmakers, J.H. (1998). Xeroderma pigmentosum group C protein complex is the initiator of global genome nucleotide excision repair. *Molecular cell* 2, 223-232.
- Suzuki, K., Yamauchi, M., Oka, Y., Suzuki, M., and Yamashita, S. (2011). Creating localized DNA double-strand breaks with microirradiation. *Nature protocols* 6, 134-139.
- Svoboda, D.L., and Vos, J.M. (1995). Differential replication of a single, UV-induced lesion in the leading or lagging strand by a human cell extract: fork uncoupling or gap formation. *Proceedings of the National Academy of Sciences of the United States of America* 92, 11975-11979.
- Sweder, K.S., and Hanawalt, P.C. (1992). Preferential repair of cyclobutane pyrimidine dimers in the transcribed strand of a gene in yeast chromosomes and plasmids is dependent on transcription. *Proceedings of the National Academy of Sciences of the United States of America* 89, 10696-10700.
- Taylor, J.S., and O'Day, C.L. (1990). cis-syn thymine dimers are not absolute blocks to replication by DNA polymerase I of Escherichia coli in vitro. *Biochemistry* 29, 1624-1632.
- Teng, Y., Bennett, M., Evans, K.E., Zhuang-Jackson, H., Higgs, A., Reed, S.H., and Waters, R. (2011). A novel method for the genome-wide high resolution analysis of DNA damage. *Nucleic acids research* 39, e10.
- Tewari, A.K., Yardimci, G.G., Shibata, Y., Sheffield, N.C., Song, L., Taylor, B.S., Georgiev, S.G., Coetzee, G.A., Ohler, U., Furey, T.S., et al. (2012). Chromatin accessibility reveals insights into androgen receptor activation and transcriptional specificity. *Genome biology* 13, R88.
- Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D., and van Helden, J. (2012). RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic acids research* 40, e31.
- Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. *Nature* 489, 75-82.
- Tommasi, S., Swiderski, P.M., Tu, Y., Kaplan, B.E., and Pfeifer, G.P. (1996). Inhibition of transcription factor binding by ultraviolet-induced pyrimidine dimers. *Biochemistry* 35, 15693-15703.
- Tornaletti, S., and Pfeifer, G.P. (1996). UV damage and repair mechanisms in mammalian cells. *Bioessays* 18, 221-228.
- Tornaletti, S., Reines, D., and Hanawalt, P.C. (1999). Structural characterization of RNA polymerase II complexes arrested by a cyclobutane pyrimidine dimer in the transcribed strand of template DNA. *The Journal of biological chemistry* 274, 24124-24130.
- Trojer, P., and Reinberg, D. (2007). Facultative heterochromatin: is there a distinctive molecular signature? *Molecular cell* 28, 1-13.
- Tu, Y., Tornaletti, S., and Pfeifer, G.P. (1996). DNA repair domains within a human gene: selective repair of sequences near the transcription initiation site. *The EMBO journal* 15, 675-683.
- Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z., and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. *Nature* 474, 516-520.
- van Hoffen, A., Natarajan, A.T., Mayne, L.V., van Zeeland, A.A., Mullenders, L.H., and Venema, J. (1993). Deficient repair of the transcribed strand of active genes in Cockayne's syndrome cells. *Nucleic acids research* 21, 5890-5895.
- Van Nieuwerburgh, F., Thompson, R.C., Ledesma, J., Deforce, D., Gaasterland, T., Ordoukhianian, P., and Head, S.R. (2012). Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. *Nucleic acids research* 40, e24.
- Vilenchik, M.M., and Knudson, A.G. (2003). Endogenous DNA double-strand breaks: production, fidelity of repair, and induction of cancer. *Proceedings of the National Academy of Sciences of the United States of America* 100, 12871-12876.
- Wang, J., Duncan, D., Shi, Z., and Zhang, B. (2013). WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic acids research* 41, W77-83.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews Genetics* 10, 57-63.

- Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L., and Schubeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature genetics* 37, 853-862.
- Wei, D., Maher, V.M., and McCormick, J.J. (1995). Site-specific rates of excision repair of benzo[a]pyrene diol epoxide adducts in the hypoxanthine phosphoribosyltransferase gene of human fibroblasts: correlation with mutation spectra. *Proceedings of the National Academy of Sciences of the United States of America* 92, 2204-2208.
- Wu, C., Bingham, P.M., Livak, K.J., Holmgren, R., and Elgin, S.C. (1979). The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. *Cell* 16, 797-806.
- You, Y.H., Lee, D.H., Yoon, J.H., Nakajima, S., Yasui, A., and Pfeifer, G.P. (2001). Cyclobutane pyrimidine dimers are responsible for the vast majority of mutations induced by UVB irradiation in mammalian cells. *The Journal of biological chemistry* 276, 44688-44694.
- Yu, Y., Teng, Y., Liu, H., Reed, S.H., and 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.
- Zavala, A.G., Morris, R.T., Wyrick, J.J., and Smerdon, M.J. (2014). High-resolution characterization of CPD hotspot formation in human fibroblasts. *Nucleic acids research* 42, 893-905.
- Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., and Shen, B. (2011). A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PloS one* 6, e17915.
- Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology* 9, R137.
- Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Molecular cell* 40, 939-953.
- Zhou, V.W., Goren, A., and Bernstein, B.E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nature reviews Genetics* 12, 7-18.
- Zhu, Q., Pao, G.M., Huynh, A.M., Suh, H., Tonnu, N., Nederlof, P.M., Gage, F.H., and Verma, I.M. (2011). BRCA1 tumour suppression occurs via heterochromatin-mediated silencing. *Nature* 477, 179-184.

7. Appendix

Abbreviations

(6-4)PP	6,4-photoproduct
BER	Base excision repair
BSA	Bovine serum albumin
ChIP -seq	Chromatin immunoprecipitation coupled with high-throughput DNA sequencing
CNV	Copy number variation
CPD	Cyclobutane pyrimidine dimers
CS	Cockayne syndrome
DAPI	4',6-diamidino-2-phenylindole
DNA	Deoxyribonucleic acid
DNase-seq	Combination of DNase I digestion with high-throughput DNA sequencing
DSB	Double strand break
EC	Euchromatin
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements coupled with high-throughput DNA sequencing
GG-NER	Global genome nucleotide-excision repair
γ H2AX	Phosphorylated at the serine 139 of histone H2AX
H3K27ac	Histone 3 lysine 27acetylation
H3K36me3	Histone 3 lysine 36 trimethylation
H3K4me3	Histone 3 lysine 4 trimethylation
H3K9me3	Histone 3 lysine 9 trimethylation
HC	Heterochromatin
HR	Homologous recombination
IF	Immunofluorescence
IP	Immunoprecipitation
IR	Ionizing radiation
NER	Nucleotide-excision repair
NGS	Next-generation sequencing
NHEJ	Non-homologous end joining
PBS	Phosphate buffered saline
RPKM	Reads per kilobase pair per million
SNP	Single nucleotide polymorphism
ssDIP-seq	Strand-specific damaged DNA immunoprecipitation followed by massively parallel DNA sequencing
TC-NER	Transcription-coupled nucleotide-excision repair
TES	Transcription end site
TSS	Transcription start site
UTR3'	3' prime untranslated regions
UTR5'	5' prime untranslated regions
UV	Ultraviolet
XP	Xeroderma pigmentosum

List of contributions

γ H2AX project:

Wei Yu	Data analysis
Alexander Rapp	Designed the project; Carried out experiments; Figures preparation
Francesco Natale	Designed the project; Carried out experiments; Figures preparation
Gisela Taucher-Scholz (GSI)	Designed the project
AG Wei Chen (MDC)	Performed sequencing
M. Cristina Cardoso	Designed the project

CPD project:

Wei Yu	Designed the project; Carried out experiments; Data analysis; Figures preparation
Stephan Grulich	Performed microarray experiment for expression of XPC ^{-/-} cells
Alexander Rapp	Designed the project; Developed ssDIP-seq protocol
AG Wei Chen (MDC)	Performed sequencing
M. Cristina Cardoso	Designed the project

Acknowledgement

Here I would like to thank Dr. M Cristina Cardoso and Dr. Alexander Rapp. For not only giving me the opportunity to carry out my thesis work in the lab but also for the selfless help, a lot of suggestions, daily discussion and the great patience to teach me how to do experiments.

I would like to also thank Dr. Barbara Drossel agreed to be my second supervisor without any hesitation.

I am also thankful to all the fellows in Cardoso lab for a fantastic working atmosphere, and helped me a lot especially helping to read the German letter. It's lucky and happy to work with you for three years. I like to thank Anne Lehmkuhl for preparation of cells, Francesco Natale for project discussion, and also neighbourhood Malini for proof reading and technical support, also Peng and Anne's help.

感谢父母家人的支持，以及朋友们的关心，为了梦想

Declaration – Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig angefertigt habe. Sämtliche aus fremden Quellen direkt oder indirect übernommene Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und noch nicht veröffentlicht.

Darmstadt, den 10. Oct 2014

Wei Yu

Curriculum vitæ

Name: Wei Yu
Address: Technische Universität Darmstadt
Schnittspahnstrasse 10
64287 Darmstadt, Germany
Date of birth: 01.09.1984
Place of birth: JiangXi China

Education:

2011.10-2014.09	PhD thesis at TU Darmstadt, Germany in the group of Prof. Cardoso "cell biology and epigenetics", under supervision of Dr. Rapp
2006.09-2009.06	Master of Bioinformatics under supervision of Associate Prof. Dr .Miao He
2001.09-2005.06	School of Life Science, Sun Yat-sen University, China Bachelor of Automation Major: Artificial intelligence College of Automation, Nanchang Institute of Aeronautical Technology (changed name to Nanchang Hangkong University in 2007), China

Working:

2009.03-2011.09	Bioinformatics assistant under supervision of Prof. Dr. Biaoyang Lin & Dr. Bingding Huang System System Biology Division, Zhejiang-California International Nanosystem Institute (ZCNI), Zhe Jiang University, Hang Zhou, China
-----------------	---

List of Publications

- 1) In preparation: **Combined genome-wide and 3D super-resolution nanoscopy analysis of the DNA damage response after ionizing radiation**
- 2) In preparation: **ssDIP-seq reveals cyclobutane pyrimidine dimers distribute and repair in strand-specific manner in NER proficient and deficient cells**
- 3) **Yu Wei, Li-Ran He, Yan Chao Zhao, Ma Him Chan,Meng Zhang and Miao He. Predicting and Function Analysis of Protein-protein Interaction Networks of Lung Cancer.** *Chin J Cancer.* 2013.
- 4) **Jin C, Yu W, Lou X, Zhou F, Han X, Zhao N, Lin B.** **UCHL1 Is a Putative Tumor Suppressor in Ovarian Cancer Cells and Contributes to Cisplatin Resistance.** *Journal of Cancer.* 2013.
- 5) **Ding D, Lou X, Hua D, Yu W, Li L, Wang J, Gao F, Zhao N, Ren G, Li L, Lin B.** **Recurrent Targeted Genes of Hepatitis B Virus in the Liver Cancer Genomes Identified by a Next-Generation Sequencing-Based Approach.** *PlosGenetic.* 2012
- 6) **Yu W, Jin C, Lou X, Han X, Li L, He Y, Zhang H, Ma K, Zhu J, Cheng L, Lin B.** **Global Analysis of DNA Methylation by Methyl-Capture Sequencing Reveals Epigenetic Control of Cisplatin Resistance in Ovarian Cancer Cell.** *PloS One.* 2011.
- 7) **Fang X, Yu W, Li L, Shao J, Zhao N, Chen Q, Ye Z, Lin SC, Zheng S, Lin B.** **ChIP-seq and Functional Analysis of the SOX2 Gene in Colorectal Cancers.** *OMICS.* 2010.

Conference contributions

Poster:

16. Jahrestagung der Gesellschaft für Biologische Strahlenforschung e.V. (GBS); Darmstadt, Germany; 2013

Title: Genome-wide analysis of DNA damage response

Poster:

Barcelona Conference on Epigenetics and Cancer: Challenges, Opportunities and Perspectives; Barcelona, Spain; 2013

Title: Combined genome-wide and 3D super-resolution nanoscopy analysis of the DNA damage Response

Poster:

30th Ernst Klenk Symposium in Molecular Medicine. DNA Damage Response and Repair Mechanisms in Aging and Disease; Cologne, Germany; 2014

Title: Genome-wide analysis reveals DNA damage preferential distribution and repair kinetics response to UV