EPIGENETIC DETERMINANTS OF NUCLEOTIDE EXCISION REPAIR EFFICIENCY

by
ARDA ÇETİN

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Master of Science

Sabancı University
December 2020

EPIGENETIC DETERMINANTS OF NUCLEOTIDE EXCISION REPAIR EFFICIENCY

Approved by:

Date of Approval: Dec 22, 2020

ABSTRACT

EPIGENETIC DETERMINANTS OF NUCLEOTIDE EXCISION REPAIR EFFICIENCY

ARDA ÇETİN

MOLECULAR BIOLOGY, GENETICS, and BIOENGINEERING M.A. THESIS, DECEMBER 2020

Thesis Supervisor: Asst. Prof. Ogün Adebali
Thesis Supervisor: Asst. Prof. Öznur Taştan Okan

Keywords: nucleotide excision repair, histone modifications, chromatin, machine learning, UV lights, skin cancer

Passing through the atmosphere, UV components of sunlight reach the earth's surface. Long exposures of cells to UV-A and UV-B result in cellular dysfunctionalities by causing DNA damage. Nucleotide excision repair (NER) is a mechanism that identifies and removes bulky DNA adducts such as UV-induced dipyrimidines. NER consists of two sub-pathways with respect to its damage recognition step: global (G-NER) and transcription-coupled repair (TC-NER). TC-NER takes place on the transcribed strand of the genes, whereas G-NER is globally active throughout the genome. It has been reported some chromatin states affect the efficiency of NER which consists of G-NER and TC-NER, in combination. TC-NER is associated with transcription and related genomic features. However, epigenetic factors affecting the G-NER efficiency has been underexplored. Here, we processed the genome-wide datasets derived from DNA damage and repair maps as well as histone modification maps of the three cell lines. With the genomic DNA damage, repair, and histone modification datasets, we built machine learning models to reveal epigenetic factors that can be predictive of NER and particularly G-NER efficacy. Our models resulted in high accuracy prediction of DNA repair potential of the genomic regions. We suggest that cells' epigenetic architecture is likely the key determinant of global DNA repair bias, therefore, mutagenesis in cancer.

ÖZET

KESİP ÇIKARMALI DNA ONARIM ETKİNLİĞİNİN EPİGENETİK BELİRLEYİCİLERİ

ARDA ÇETİN

MOLEKÜLER BİYOLOJİ, GENETİK ve BİYOMÜHENDİSLİK YÜKSEK LİSANS TEZİ, ARALIK 2020

Tez Danışmanı: Dr. Ogün Adebali
Yardımcı Tez Danışmanı: Dr. Öznur Taştan Okan

Anahtar Kelimeler: kesip çıkarmalı DNA onarımı, histon modifikasyonları, kromatin, makine öğrenmesi, UV ışınları, cilt kanseri

Atmosferden geçerek güneş ışığının UV bileşenleri yeryüzüne ulaşır. Hücrelerin UV-A ve UV-B'ye uzun süre maruz kalması, DNA hasarına neden olarak hücresel işlev bozukluklarına neden olur. Nükleotid ekzisyon onarımı (NER), UV ile indüklenen dipirimidinler gibi büyük DNA eklentilerini tanımlayan ve ortadan kaldıran bir mekanizmadır. NER, hasar tanıma adımına bağlı olarak iki alt yolaktan oluşur: global (G-NER) ve transkripsiyona bağlı onarım (TC-NER). TCR, genlerin transkripsiyon zincirinde yer alırken, G-NER genom boyunca iki DNA zincirinde de aktiftir. Bazı kromatin yapıları, G-NER ve TC-NER'den oluşan NER'in işleyişini etkilediğini bildirilmiştir. TC-NER, yapıları transkripsiyon ve ilgili genomik aktivitelerle ilişkilidir. Bununla birlikte, G-NER verimliliğini etkileyen epigenetik faktörler yeterince araştırılmamıştır. Bu çalışmada, DNA hasarı ve onarım haritalarından ve üç hücre hattının histon modifikasyon haritalarından türetilen genom çapında veri kümelerini işledik. Genomik DNA hasarı, onarımı ve histon modifikasyon veri kümeleriyle, NER ve özellikle G-NER'in etkinliğini tahmin edebilecek epigenetik faktörleri ortaya çıkarmak için makine öğrenimi modelleri oluşturduk. Modellerimiz, genomik bölgelerin DNA onarım potansiyelini tahmin etmede yüksek doğruluk sağladı. Makine öğrenmesi yaklaşımımız, histon modifikasyonlarının kromatine, hasar türüne ve hücre tipine bağlı olarak, NER işleyişine etkisini göstermiştir. Hücrelerin epigenetik mimarisinin muhtemel global DNA onarım eğiliminin, dolayısıyla kanserde mutagenezin belirleyicisi olduğunu öne sürüyoruz.

ACKNOWLEDGEMENTS

I would like to start giving my gratitude to Prof. Ogün Adebali for providing me a chance to study and explore the computational biology and bioinformatics field. He accepted me to work on one of his projects and thanks to him I gained new perspectives and learned lot about not only doing a research but also about life. I am happy to be a part of his group and his professional support which, hopefully, I can continue to get in the future. I would like to express my appreciation to my co-advisor Prof. Öznur Taştan for being always kind, positive, and supportive. I appreciate for her help and feedback. Thanks to her we can be able to finish this project. I would like to specially thanks to Afshan Nabi, who is a master student in computer science department, for being kind, easy-going, and her endless help. Although we started our collaboration at the end of my first year, we have been through many obstacles and difficulties. Beside her friendship, she was so generous that helped in building a machine learning model for the project even if this is not her thesis project. I would like to thank all the members of ADEBALILAB; Cem Azgari, Aylin Bircan, Burak İşlek, Berkay Selçuk, and Sezgi Kaya for their continues support, scientific comments, positivity, and their friendship during my short but meaningful academic journey. I am also very appreciative to Erhan Ekmen and İrem Akülkü for helping me during this process. They were supportive and encouraging to me and I am glad that I have met them at the Sabanci University. Therefore, I am also thankful for Sabanci University to provide all its opportunities and enable me meet with generous people. Especially, I owe a great depth of gratitude to the best trio of my life, Burak Çavuşoğlu, Emre Alver, and Ufuk Kanat for being my best and most supportive friends and always being on my side not only in good times but in the bad times as well. Their infinite aids and their courage allow me to do many things that I cannot ever imagine. I knew them before I knew myself hence, they know me better than I know myself. We have grown up together and experienced many things together. Together, we were able to stand strong against what life throws at us and hopefully in the future it could continue like this as well. I would like to thank all of them wholeheartedly for being my brothers. Genuinely, I am so lucky to have a great family who support me in any circumstances. Without them, I could not achieve and have many things in this life. If I have some freedom, chances to act, and can pursue my career, I owe them. I am thankful for them for supporting and encouraging me to do master's degree and helping me in all ways. They always put me priority, act considering me and dedicated their life to me for making me a good and decent man. I could not have completed this project without their love and assist. Therefore, I would like to dedicate this research project first to my family and then, to anyone who truly accepts and loves me for the way I am.

To my beloved family and friends
Sevgili aileme ve arkadaşlarıma

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

UV	Ultraviolet
NER	Nucleotide excision repair
G-NER	Global nucleotide excision repair
TC-NER	Transcription coupled nucleotide excision repair
HeLa	Henrietta Lacks' cervical cancer cell line (1951)
NGS	Next generation sequencing
HTS	High throughput sequencing
Damage-seq	Damage sequencing method
XR-seq	Excision repair sequencing method
ChIP-seq	Chromatin immunoprecipitation sequencing method
DNase-seq	DNase I hypersensitivity sequencing method
[6-4]PP	6-4 pyrimidine primidone photo products
CPD	Cyclobutene pyrimidine dimers
DDR	DNA damage response
PTM	Post translational modifications
KMT	Histone lysine methyltransferase
TFIIH	Transcription factor II H
DDB2	DNA damage binding protein 2
DNA	Deoxyribonucleic acid
UV-DDB	UV-damaged DNA-binding protein
HAT	Histone acetyltransferase

1. INTRODUCTION

Cell survivability depends on the maintenance and proper operations of the cellular activities. Cellular activities enable cells to stay alive and provide cellular homeostasis. Despite the occurrence of thousands of cellular activities within the cell, most of the time they operate properly in healthy cells. Perhaps, as cellular activities are running, various events and cellular processes could go wrong within the cell due to the cell's complexity or environmental factors. Most interestingly, all these highly complex cellular events are arranged by the regulatory mechanisms to ensure the sufficiency and the integrity of these cellular processes and their products. Moreover, cells also have cellular defense mechanisms against the internal and external agents which can disrupt the cellular functioning through DNA damaging (1 - 9). These events and regulations are governed in an organelle called the nucleus. Depending on the expression of genes inside the nucleus, the functioning of many cellular activities can change. A genome is the cell's guidebook where all the information necessary for cellular events lies for building and maintaining the cell. The genome, the collection of all genetic materials of a cell, consists of chromosomes and the number of chromosomes found within a cell may vary from organism to organism. Chromosomes structurally consist of chromatins and they are the complex of DNA and proteins. The proteins, which constitute chromatins, are called nucleosomes (6). The nucleosomes are the complex which is formed by nine histone proteins (H1, H2A, H2B, H3, and H4) and a DNA helix (8).

Each layer building the genome is needed for providing cellular functionality and structural operations. It is also necessary for genome to fit inside the cell with an organized manner. The utmost evolutionary reason for genomic condensation is that the DNA at its naked form is larger than the size of a cell. Therefore, the condensed DNA was an evolutionary step for the beginning of a prokaryotic life. The compaction of the genome is provided by the coherent functioning of various molecules including nucleosomes, chromatin remodelers, chaperones, etc. These arrangements of chromatins for enabling or disabling cellular operations and providing genomic integrity for cells are mainly governed by the interaction of histones, hi-

stone modifications, DNA helix, and chromatin remodelers. Besides, compaction is also necessary for protecting DNA from physical damages and making it more manageable for controlling the cellular events. Despite this evolutionary mechanism of cells to defend itself from a wide variety of agents, DNA might still get malfunctioned. Thus, all cells have developed similar mechanisms which protect, and fix DNA defects created by the harmful agents (10).

Organisms are exposed to various internal and environmental agents throughout their life (9). These agents can damage DNA and then, affect the cell's viability depending on the rate of exposure. Cells having damaged DNA may experience deficiencies in some cellular activities. These deleterious impacts of DNA damage on the cells can also result in genomic instabilities (1). Consequently, these DNA originated abnormalities or genomic instabilities may give rise to diseases such as cancer (11). Another impact of these agents on the cells is the possibility of preserving UV damaged DNA in the daughter cells as hereditary information. Worse, the impaired genome is transmitted to offspring of an organism, if damage cannot be detected and repaired by the DNA regulatory mechanisms. Therefore, wide range of DNA defense mechanisms developed to repair DNA damages and evade the risks of transmitting disrupted genetic material to the progenies. If a damage on the genome cannot be detected because of the malfunctioning of corresponding repair mechanism, the damage will cause severe problems in an organism such as mutation formation and fast aging (1). Thus, regulation of DNA defense mechanisms is as important as the maintenance of the genome. That is the reason why every cell also has cell cycle checkpoints to ensure about the integrity of the cell.

The ancestral organisms developed varied DNA defense mechanisms and with the speciation (12, 13). Those are used against various genomic errors or damages. It is highly important for a cell to detect and respond to damage instantly and properly once it is encountered (14). However, some DNA repair pathways are often activated more due to the presence of particular lesions in the genome (13).

Although there are many environmental and internal agents, which are potential threats for a cell, some of them are more frequently encountered such as UV radiation, ionizing radiation, free radicals and oxidation, hydrolysis, and the alkylating agent such as cisplatin (2-9). UV light is one of the most seen DNA damage inducing agents. Sun emits UV radiations (UV-A, UV-B and UV-C) and only a small fraction of UV-A and B can reach to the earth surface (15). Despite the personal precautions taken for protecting the body against sun's deleterious impacts, UV light results in bulky dipyrimidines.

UV damage is one of the major causes of skin melanoma in humans. Even tiny

portions of UV radiation may be enough for cells to cause damage on the genome depending on the exposure rate. The genome is structurally vulnerable against the UV light due to genomic composition. These UV-induced DNA damages result in the formation of DNA bulky adducts such as 6-4 pyrimidine pyrimidone photoproducts ([6-4]PPs) and cyclobutane pyrimidine dimers (CPD) on the genome (16, 17). To cope with these deleterious impacts of UV light on the genome, various evolutionary DNA protecting mechanisms have been developed. Besides, another damage forming agent is a platinum-based drug called cisplatin, which is administered as an anti-cancer drug against the number of cancers (18). Cisplatin is a chemotherapy drug which generate cisplatin damages on the genome without distinguishing between normal or cancer cells. The nucleotide excision repair pathway is the sole mechanism which repairs the bulky DNA adducts on the genome caused by the UV-B light and cisplatin drug DNA damages (19).

Many endogenous and exogenous agents disrupt the genome in different ways. Depending on their targets, variations in their molecular pathways, and functioning, different DNA regulatory mechanisms continuously operates to keep the cell as healthy as possible. Proofreading, mismatch repair and DNA damage repair mechanisms maintain the genome in different ways by fixing dissimilar DNA errors and/or damage products (10, 12). Excision repair is a damage repair mechanism. Excision repair pathways are initiated once oxidized bases, alkylated bases, DNA adducts, or DNA photoproducts kind of single strand originated DNA damages were identified through the DDR process (2). Excision repair pathways eliminate DNA damage in three general steps starting by the recognition of damage, removal of damage through dual excision, and the resynthesis of correct nucleotides. Excision repair includes two sub-repair mechanisms called nucleotide excision and base excision repair. The main differences between BER and NER are their components, the origin of the DNA damage, DNA damage repair targets, and the discarded fragments lengths (9, 13).

It was suggested that contrast to BER, NER pathways are specific to some damages which substantially changes the helix structure of DNA (19). The nucleotide excision repair (NER), which has been detailly studied by many researchers, discards mainly UV- and chemical-induced DNA adducts. Specifically, in placentals, NER is the sole repair machinery correcting the bulky UV-induced DNA damages such as DNA photoproducts and pyrimidine dimers. On the other hand, some of the organisms, such as some bacteria, yeasts, and plants, directly repair the UV lesions by the enzyme called photolyase. The first phase of NER begins with the identification of damage products similar to the other repair mechanisms. The bulky DNA photoproducts and pyrimidine dimers are either recognized by XPC and DDB or RNA

polymerase. The recognized damage adducts are excised at both ends of the damage by XPG and XPF. Afterwards, these dual-cut single stranded DNA fragments are removed from the genome by TFIIH and XPC then DNA ligase and polymerase seals DNA breaks. The malfunctioning of any step of the NER pathways causes severe health problems and diseases such as Cockayne's syndrome and Xeroderma pigmentosum (20).

Although NER has evolved both in prokaryotes and eukaryotes, NER is not completely conserved between the organisms (21). NER may include similar mechanisms for damage repair across organisms but there are still differences in the components and the efficiency of NER in damage repair. For example, the number of repair factors found in NER machinery is around 15-16 in humans, but this number is lower in prokaryotes (15, 22, 23, 24). Conversely, unlike humans, additional helices proteins involve in the NER process in prokaryotes (24). Also, the length of the incision region varies between the prokaryotes and eukaryotes. In prokaryotes, the size of the replaced oligomer is between 12 and 13 nucleotides long whereas it can vary in length from 24 to 32 nucleotides in the eukaryotes (15, 20, 22, 23, 25 - 28).

NER pathways further divided into two sub pathways with respect to their damage recognition steps: global nucleotide excision repair (G-NER) and transcription coupled nucleotide excision repair (TC-NER). These two repair mechanisms differ in which strand of DNA they function and the way of damage detection. These two factors affect the rate of repair in two NER sub-mechanisms (29). TC-NER removes bulky DNA lesion more effectively and rapidly than the G-NER (30). Two sub-pathways use different constituents to sense the DNA damage but the later steps of both NER pathways employ common factors. In the TC-NER pathway, the identification of the bulky DNA adducts, and recruitment of repair complexes are mainly mediated by the error-free polymerase. RNA Polymerase stalls once it encounters damage and recruits repair complex proteins to the damaged region (31, 32). Conversely, polymerase is not involved in the damage detection process of G-NER pathway instead UV-DDB and XPC mediate DNA damage identification and recruit NER factors to the damage region. So, TC-NER pathway requires few molecules for the damage recognition step. Therefore, TC-NER pathway damage detection is less time demanding and more effective process than G-NER pathway when G-NER cannot efficiently repair the lesion on transcribed strands (33). Even though the general steps of both NER pathways are similar, they also possess small differences in the steps of repair. The reason lies under the differences between both NER pathways. TC-NER is activated if a damage is detected during transcription. However, unlike the TC-NER pathway, G-NER activity does not only rely on transcription. Conversely, it can also be activated in the replication process as well. As

the name implies, TC-NER removes the damages found on DNA transcribed strands as it is a transcription dependent repair mechanism. On the contrary, the G-NER pathway removes damages found on both strands (33).

It has been mentioned above that CPD and [6-4]PP damage products are formed on the genome after cells are exposed to UV after a certain amount of time. CPD and [6-4]PP are solely repaired by NER pathways, but these two damage types differ in their composition, repair rate, and genome-wide distributions. It has been shown on UV treated cells that unlike [6-4]PPs, CPDs are more uniformly distributed and abundantly found over the genome (34). Moreover, [6-4]PPs are localized at the DNA regions where the DNA wraps around the nucleosomes (35). Despite the relative less abundance of [6-4]PPs on the genome compared to CPDs (36), they are repaired at a faster rate by the G-NER and CPD damages are repaired relatively effectively by TC-NER, but the exact reason of that repair tendencies is not explained yet (37, 38, 39). Besides, unlike [6-4]PPs, CPDs are repaired more effectively at the transcribed strand because their repair strongly influenced by TC-NER (39). Therefore, [6-4]PP do not have strand-specific repair bias like CPD does.

Chromatin can act as a barrier for NER machinery in many instances because they possess strict regulations to control complex cellular events. However, some chromatin structures can facilitate the repair. Therefore, chromatin, in general, requires nucleosome rearrangements or modifications to allow transcription, replication, or DNA repair events. Chromatin structure may also restrict the operations of polymerase enzymes or other molecules on the genome. However, in case of the need of cellular activities, structural chromatin restrictions are lightened by the sets of regulatory molecules, such as histone modifiers, histone methylases, histone acetylases, histone deacetylases, chaperons, and other molecules. In a decade, some chromatin structure elements, which were thought to be related to repair, have been investigated (40). Of those, chromatin remodelers were suggested to indirectly interact with NER machinery. This indirect interaction between chromatin remodelers and NER rely on the modifications of histones. Until now, only a small part of the histone modifications (histone markers) has been associated with the NER and it is confirmed that different enzymes add diverse modifications to various amino acid residues of histones. The N-terminal tail acetylation of H3 and H4 histone proteins by HATs were reported to be closely related with the chromatin remodelers functioning such as H3K9ac. Yu et al. (2005) revealed that within a half an hour, histone acetylation process begins at a repressed region in the UV-treated yeast cells (41). The NER related factors and their mechanisms for damage detection vary depending on the organism. For instance, in mammals, transcription factor E2F1 and UV-DDB guides Gcn5 HATs to the DNA damage region (40, 42). Gcn5

then adds acetylation group to the 9th lysine of H3 and this enables NER proteins to access DNA damage with the help of chromatin remodelers. UV-DDB also indirectly recruits chromatin remodelers to the DNA damage region as well. These chromatin remodelers recognize H3K9 histone acetylation and act as an attachment site for NER protein complexes at the damage site (40, 42). These findings indicate that chromatin remodelers and HATs are some of the NER associated factors which facilitate NER process by loosening chromatin structures.

In addition to acetylation, another histone PTM related with NER mechanism is histone methylations. Histone methyltransferases (HMTs) add methyl groups to amino acid residues of histones. They participate in the regulation of many biological processes including DNA damage repair. Even though histones can be methylated at a wide range of amino acid residues of H3 and H4 proteins, none of these methylations have been directly related to repair. H3K79 and H4K20 methylations are some of the characterized markers with the NER (44, 45). Researchers reported increased formation of UV damage lesions in the yeast genome, where these modifications or HMT coding genes were mutated (43, 44). The effect of H3K79 on repair was only related to G-NER not with the TC-NER. It has been proposed that this modification may loosen the chromatin structure at the G-NER regions (44) or can act as a binding site for G-NER complex. Furthermore, H4K20 methylation serves as a docking site for damage detection protein in double stranded break repair but not much is known about its effect on NER (45). These suggests a possible relationship between these histone modifications and the NER machinery. Nevertheless, the explicit effect of histone methylations on NER is not fully comprehended. On the other hand, histone phosphorylation and ubiquitylation modifications are other PTMs which are suggested to be associated with the UV or cisplatin induced bulky adducts repair.

Many factors affecting NER have been shown over the years but some of them, such as histone markers, were not explicitly associated with NER yet. Previous studies revealed that NER's deficiencies in cells are a serious threat for cell's vitality and survivability. NER is one of the major DNA damage repair mechanisms which enables the genome to be stabilized and maintained from various internal and external agents. It has been unveiled that defective NER can foster the aging process and disorders like Xeroderma pigmentosum and cancer (20). Furthermore, research conducted on stem cells suggested that absence of an element from a NER pathway complex contributes to the development of Signature 8 mutations (46). Besides, cancer cells, which have distorted NER pathway, are more likely to respond to certain treatments. Therefore, the deficiency of NER in a cell can provide a significant tool to diagnose and treat the particular type of cancer cells.

The integrity of the genome highly depends on the removal of DNA lesions because the unstable structure of the genome is likely to cause structural and functional abnormalities which can lead to severe cellular problems or life-threatening issues as mentioned above. NER machinery can be comprehended better if factors affecting NER, such as chromatin accessibility and states, are unveiled more detailly (47). As an example, chromatin structure, histone markers, histone remodelers, and DNA damage repair mechanisms are intertwined and dependent quantities. So that the compression of DNA molecules using nucleosomes is not enough to preserve the genetic material of the cells alone hence genetic defense mechanisms have been evolved to undo the effects of deleterious agents. However, DNA damage repair mechanisms rely on both chromatin accessibility and protein availability which is regulated through DDR. Similarly, chromatin structures require histone modifications and chromatin remodelers for regulating their chromatin states. These interrelated quantities are the constitutive parts of a big machine. Yet, the interaction between the NER and histone modifications are not completely known (48, 49). Therefore, the NER mechanism and its effectors need investigation for its complete comprehension.

Genome-wide DNA damage and repair profiles might provide remarkable information for understanding factors affecting NER mechanisms. Fortunately, the advancement of NGS technologies accelerated many studies and resulted in the production of many high throughput data in that field. The NGS technologies are not only limited to genome-wide damage and its repair maps but also lots of other genome-wide high throughput methods were developed such as ChIP-seq and DNase-seq. Frankly, all NGS methods possess similar protocols because they developed based on common NGS technologies. For instance, ChIP-seq is an important and frequently applied method in many fields of biology. It can be utilized to find the regions bound by proteins or protein modification of interest along the genome. ChIP-seq starts with the cross-linking of protein of interest to the genome before digesting DNA with DNase enzyme. This step is followed by the precipitation of shredded DNA, removal of protein of interest, purification of DNA sequence, and sequencing these DNA fragments. On the other hand, DNase I hypersensitivity regions can be pinpointed throughout the genome using DNase-seq method. It is a crucial method to exhibit the handiness of chromatin structures or observe the gene regulatory regions along the genome. The eXcision Repair sequencing (XR-seq) method is only one of the products of the advanced NGS technologies. The eXcision Repair sequencing (XR-seq) has been developed for obtaining DNA damage repair maps throughout the genome (33). NER mechanism detects, binds, and cuts the damage region of the DNA from both ends, results in dual-incised short fragments, and the correct

sequences are added and ligated for sealing the gap. The novelty of XR-seq method is its capability of capturing these duel-cut short-lived fragments of DNA before they get degraded inside the cell (33). Moreover, XR-seq can identify the individual bases of dual incised short-lived DNA fragments at single nucleotide resolution hence, it provides reliable results. XR-seq method has been applied on different living organisms such as human, lemur, *D. Melanogaster*, *S. Cerevisiae*, *E.coli* and *A. thaliana* (21, 33, 50 - 54). Detecting the short-lived fragments in single nucleotide resolution starts with the immunoprecipitation of the excision product with alpha TFIIH, followed by the adaptor ligation, precipitation, damage reversal, amplification, and next generation sequencing (33). Next, with the similar approach, the damage sequencing (damage-seq) has been developed for detecting the particular position of the UV damage lesion on the genome at single nucleotide base resolutions (55). Damage-seq method yields different outputs and respectively has more complex procedures than XR-seq. In this method, unlike XR-seq, the length of fragments is bigger and end repair is required previously. Biotin is attached to the end of damage included fragments and it is used as a label for differentiating between the damaged and damage-free DNA fragments (33, 55). However, unlike XR-seq's specificity for identifying the damaged point, damage-seq is highly sensitive method in terms of detecting the exact damage position, yet, damaged bases need to be pinpointed through bioinformatics analysis by seeking adjacent sequence, which is not found at 5' end of the amplicon, using reference genome.

Recently, high throughput sequencing methods have been utilized frequently in various fields of biology, including bioinformatics, forensics, genetics, and biotechnology, for many reasons such as identification or validation of cellular pathways and complexes. XR-seq and damage-seq are specially tailored methods, which have been specifically developed for acquiring genome-wide repair and damage maps, with high precision (33, 55, 56). It has been reported that the genome-wide cisplatin repair and damage maps, which were generated by XR-seq and damage-seq methods respectively, were not overlapping with each other (57). Damage repair (NER) maps, which were generated by XR-seq, exhibited non-uniform repair hot and cold-spot distributions throughout the genome. Conversely, damage-seq data showed relatively homogeneous damage distribution patterns along the genome (33, 56). A damage-seq experiment clarified that the heterogeneous damage repair was solely the NER preference because damage had not been repaired at every genomic region equally (55, 56). Therefore, to explore the causes of this NER repair tendency, factors affecting the NER efficiency were started to be investigated. It was proposed that the active regions on the genome are repaired preferentially and more than the less functional regions and this is also valid for the NER repair as well (55). The

heterogeneous pattern of damage repair maps was associated with the chromatin states and it is also reported that NER repairs the damage rapidly and with ease at open chromatin regions (21, 25, 30, 33, 55, 56, 58). Similarly, other epigenetic factors, such as histone modifications, histone remodelers etc. were stated to affect NER tendency in damage repair (40, 45, 59). TC-NER overshadows G-NER efficiency because of its effective and rapid functioning which were mentioned in the above paragraphs. Thus, G-NER effectiveness and functioning at structurally different genomic regions also require clarification. To the best of our knowledge no research has been conducted before for linking the histone modifications and G-NER machinery efficiency using a machine learning application. Therefore, NER and the NER related factors must be identified and elucidated comprehensively to understand the NER better. Here, we aimed to explore the relationship between histone modifications and NER in order for explaining one of the causes of previously shown non-uniform genome wide NER pyrimidine dimers and photoproducts repair preference (59). We have applied machine-learning techniques to identify the epigenetic determinants of NER.

2. PURPOSE

UV radiations, which are emitted from the sun, are the major cause of skin melanoma and many other skin related syndromes. Only in USA, 196,060 people were diagnosed with skin melanoma in 2020 and approximately, 100,350 of those were invasive and 95,710 had non-invasive skin melanoma according to the statistics of 2020 Skin Cancer Foundation.

The sole UV lesion repair machinery in human cells is nucleotide excision repair and this repair machinery decreases the chances of developing skin-related diseases such as skin cancer. The repair maps of NER had revealed heterogeneous repair throughout the genome whereas the genome-wide distributions of UV damage lesions were homogeneous. We developed a machine learning model to predict the DNA repair positions along the genome using histone modifications. To achieve this, we compare the effect of histone modifications on the NER prediction with the chromatin states. To train and test the model, we provided four types of next generation sequencing data, which are eXcision Repair sequencing, damage sequencing, DNase I hypersensitivity sites sequencing, and chromatin immunoprecipitation sequencing, to the model. We used three cell lines (HeLa, GM12878, and NHF1), which are derived from different primary tissues, to show the relationship between the histone markers and NER. Thus, this study will emphasize the interaction between NER mechanism and histone markers as well as show the efficiency of G-NER, which is overshadowed by TC-NER, in damage repair.

3. METHOD

3.1 Cell culture, UV irradiation and cisplatin exposure

The XR-seq, damage-seq, ChIP-seq, and DNase-seq NGS data for HeLa, NHF1, and GM12878 cell lines were retrieved from an online publicly available database Gene Expression Omnibus (GEO). Therefore, any questions related with the experimental setup and the data protocol should be answered by visiting each data's corresponding GEO accession numbers. During the research, ENCODE portal was frequently used for data searching due to its user-friendly interface (ENCODE Project Consortium 2012, https://www.encodeproject.org/).

3.2 Data Collection

To monitor the model prediction performance in the different cell lines, HeLa, GM12878, and NHF1 cells were used. These HeLa cells were exposed to UV radiation for 12 minutes. Since TC-NER mechanism is not activated yet within 12 minutes by the cell, genome-wide UV damage (damage-seq) and its repair maps (XR-seq) were generated within this time. Therefore, our HeLa cell XR-seq data does not include TC-NER machinery repair activity. However, the remaining two cell lines include both TC-NER and G-NER mechanisms for the damage repair. These three cells either include different NER machineries or damage types (UV or cisplatin damage). In GM12878, damage-seq and XR-seq 1.5hr data were retrieved from GSE98025 and GSE82213, respectively. NHF1 cell line's XR-seq data generated after one-hour UV exposure were from GSE67941 and DNA repair after four

hours XR-seq data were from GSE76391. The both 1- and 4-hours UV treatment data on NHF1 cells line could be accessed at GSE98025.

In addition to XR-seq and damage-seq data in three cell lines, ChIP-seq and DNase-seq data were also utilized for the analysis. ChIP-seq data were needed for mapping the positions of histone markers genome-wide in three cell lines. The ENCODE IDs of each histone modification data (ChIP-seq) can be found in Table 3.1. DNase-seq data for the HeLa and GM12878 cell lines were accessed from GSE32970 and GSE32970, sequentially. The DNase-seq datum for the normal skin fibroblast cell line (NHDF) was obtained from GEO ID GSE2969.

Table 3.1 The list of histone modifications which were analyzed in this project and their ENCODE IDs

	HeLa ENCODE ID	GM12878 ENCODE ID	NHF1 ENCODE ID
H3K4me3	ENCSR340WQU; ENCSR000DUA; ENCSR000AOF	ENCSR057BWO; ENCSR000DRY	ENCSR000DPR
H3K36me3	ENCSR000DTZ; ENCSR000AOD	ENCSR000DRW; ENCSR000AKE	ENCSR000APP
H3K27me3	ENCSR000DTY; ENCSR000APB	ENCSR000DRX; ENCSR000AKD	ENCSR000APO
H3K4me2	ENCSR000AOE	ENCSR000AKG	ENCSR000APQ
H3K27ac	ENCSR000AOC	ENCSR000AKC	ENCSR000APN
H3K79me2	ENCSR000AOG	ENCSR000AOW	ENCSR000ARW
H2AFZ	ENCSR000AQN	ENCSR000AOV	-
H3K4me1	ENCSR000APW	ENCSR000AKF	ENCSR000ARV
H3K9ac	ENCSR000AOH	ENCSR000AKH	ENCSR000APS
H3K9me3	ENCSR000AQO	ENCSR000AOX	ENCSR000ARX
H4K20me1	ENCSR000AOI	ENCSR000AKI	ENCSR000ARJ

3.3 Bioinformatics and Statistical analysis

3.3.1 Data preparation

The XR-seq, damage-seq, ChIP-seq, and DNase-seq data used in this project were retrieved from GEO and ENCODE websites. Three different human cell lines were analyzed in this project, HeLa, GM12878 and NHF1. Data retrieval and fasta format conversions were performed using *fastq-dump* followed by the quality control through *FASTQC* for all sequencing methods. All four types of NGS data were preprocessed before given as an input to the machine learning model. Depending on the layout of NGS data (single- or paired-ended), the data retrieval and format conversions require alterations accordingly in data preprocessing step. The data preprocessing

steps will be elucidated in the following paragraphs.

3.3.2 ChIP-seq and DNase-seq bioinformatics preprocessing

Some DNase-seq data were paired-ended. Preprocessing the paired-ended data differ in some steps than the single-ended data such as downloading SRA files, SRA-fastq file format conversions and the read alignments to the reference genome. Similar protocol was applied for both ChIP-seq and DNase-seq preprocessing and they begin with quality control for checking read lengths, GC content, and adaptors using *FASTQC* tool. Individual reads' alignment to the reference genome (hg19) is performed using *bowtie2*. ".sam"-".bam" to ".bed" file format conversions are performed using *samtools* and *bamtools*. In case of read duplicates removal, Pickard's *MarkDuplicates* command was used if necessary. Chromosome name and coordinate sorting before mapping samples on the genome is performed either *sort* command or *bedtools sort*. *Bedtools* intersect is used for mapping DNase-I hypersensitivity or histone marker coordinates on 5kb windowed hg19 human reference genome. These steps were followed by data labeling for distinguishing between the different data types and RPKM calculations using mapped reads' count numbers.

3.3.3 Damage-seq bioinformatics preprocessing

Some damage-seq data were paired-ended therefore, data retrieval (using *fastq-dump*), adaptor removal, and read alignment steps were altered accordingly. Preprocessing steps differ depending on the NGS methods. In damage-seq, the preprocessing begins with quality control (*FASTQC*). Adapter filtering/trimming step is necessary for XR-seq and damage-seq because these NGS methods are amplicon based. *Cutadapt* tool was utilized for adaptor detection (adaptor sequence: GACTGGTTCCAATTGAAAGTGCTCTTCCGATCT). In damage-seq, reads having adaptors must be discarded from the dataset. Adaptor filtering is followed by read alignments to reference human genome (hg19) using *bowtie2*, ".sam"-".bam" to ".bed" file conversions using *samtools* and *bamtools*, sequentially, removing duplicates using Pickard's *MarkDuplicates*, damage location filtering and identification using *bedtools flank* and *slop*, finding pyrimidines, sorting damage positions and chromosome information (chr names and coordinates) using *sort* command or *bedtools sort*,

mapping damage positions on the 5kb windowed human reference genome (hg19) using *bedtools intersect*. These steps were followed by data labeling for distinguishing between the different data types and conversion from damage-seq data mapped read count numbers to RPKMs.

3.3.4 XR-seq bioinformatics preprocessing

The quality of reads coming from the XR-seq was checked using *FASTQC*. Quality control step is necessary to understand the data reliability and improve data quality. Quality control checks reads' GC content, presence of adaptors, and their lengths. During this operation, if data type related problems or bad sequencing methodology originated issues were detected in the data, the preprocessing is shaped accordingly to amend data quality. In case of paired-ended data, the following three steps with specially designed paired-ended reads parameters should be applied: downloading SRA files and fastq file conversion using fastq-dump, adaptor removal, and aligning reads to the hg19 genome. The 5' end adaptors were removed from repair data using *cutadapt* tool (adaptor sequence: TGGAATTCTCGGGTGCCAAGGAACTCCAGTNNNNNNACGATCTCGTATGCCGTCTTCTGCTTG). In XR-seq, damage and pyrimidine identification steps do not exist. Adaptor filtering was followed by aligning reads to human reference genome (hg19) using *bowtie2*, conversion from ".sam"-."bam" files to ".bed" file format using *samtools* and *bamtools*, sequentially, removing duplicates using Pickard's *MarkDuplicates* if necessary, sorting chromosome names and coordinates using *sort* command or *bedtools sort*, and mapping repair positions on 5kb windowed human reference genome (hg19) using *bedtools* intersect tool. These steps were followed by data labeling for distinguishing between the different data types and conversion from repair read count numbers to RPKMs.

3.3.5 Gene annotation

Each gene inside the hg19 genome was divided into 4 repetitive gene segments such as 5kb upstream of transcription start sites (U_TSS), 5kb downstream of transcription start sites (D_TSS), 5kb upstream of transcription end sites (U_TES) and intergenes by performing series of filtering and data edit-

ing steps explained in the subsequent paragraphs. In order to attain the hg19 (GrCh378) genome's gene annotations, the GTF file was downloaded from this source (ftp://ftp.ensembl.org/pub/grch37/release-98/gtf/homo_sapiens/).

Initially, all the genes of hg19 genome were filtered. In the below steps, the strand information of genes was preserved. During filtering step, each three gene segments were decided to be accepted as 5kb long, and any overlapping regions and 10kb or shorter genes were removed from the data using a python script. The transcription start (TSS) and transcription end (TES) sites of each remaining genes were determined, and a second filtering step was applied such that overlapping genes and genes found within the 5kb vicinity of a gene were discarded (using *bedtools merge* integrated with *awk* commands). At the end, there were three separate BED files which include the information of either 5kb U_TSS, 5kb D_TSS or 5kb U_TES. Each of these BED files were used as a reference genome for mapping NGS reads on three gene segments.

The intergenic regions were obtained with the help of a genic regions (location of genes). The genic regions were utilized as in their original state (without applying any of the above-mentioned filtering steps) for detecting the intergenes correctly. We have eliminated the unmappable regions from the dataset by matching mappable damage-seq data coordinates with gene annotation file coordinates. This is because if a read could be sequenced correctly, this indicates that this read is mappable. Also, we have added 5kb to the upstream or downstream of each gene's TSS coordinates by considering their strand information for evading the risk of including gene regulatory regions as a part of intergenic regions. To do this, we have assumed that gene regulatory regions span 5kb long area on the genome. At the end, this BED file was used for mapping NGS reads on the intergene segments.

3.3.6 Mapping reads on genes and intergenes

XR-seq, damage-seq, ChIP-seq, DNase-seq data samples were mapped on the hg19 reference genome, which had been divided into three gene and intergene segments, using *bedtools intersect -counts*. The *bedtools intersect* parameters used for mapping reads on four genomic segments were same as the parameters utilized at the whole genome preprocessing steps.

3.3.7 Data subsampling

Additionally, all NGS data reads were randomly subsampled according to the smallest read number across all four NGS data. Therefore, read coverages differences of four NGS data were evened up as much as possible. This step was performed for observing the effect of equal read numbers on the model prediction.

3.3.8 Data representation: RPKM

RPKM refers to read per kilobase per million mapped reads. RPKM calculation was required to be able to compare dissimilar sequencing methods and interpret knowledge from them. This is a data normalization method for resolving the different sequence coverage issues. So that RPKM enables us to deduce reliable information from the different sequencing methods by correcting different sequencing depth and length problems. The count information from four sequencing data were converted into numerical values before feeding them to the model. Using the below formula, read count values can be converted into RPKMs.

$$\text{RPKM} = \frac{Number\,of\,reads\,mapped\,to\,genome}{\frac{Size\,of\,bins}{1000} x \frac{Total\,number\,of\,reads}{10^6}}$$

After the labelling process, above calculation was performed on ChIP-seq, XR-seq, damage-seq and DNase-seq bed files for having better read representation.

3.3.9 Data representation: Peak Calling

To select optimum read representation in the analysis, the Peaks and read count numbers were assessed on model (data not shown). As a result, count numbers were chosen to represent the reads.

To apply PeakCalling, MACS2 tool (https://github.com/taoliu/MACS) was utilized. The sample and input data (BAM) for histone markers were found at GEO with their corresponding GEO accession number. The parameters for *macs2* were

callpeak, *–format* as BAM and *-g* as *hs* (hs indicates hg19 human genome).

If a sequencing datum includes biological replicates, they were merged as a single file. XR-seq and damage-seq data were further filtered and normalized for acquiring repair regions where there is damage. For filtering damage data, if there is no damage at a given 5kb region, it was discarded. Then, the repair divided by its corresponding filtered damage for each given 5kb genomic segment. After a series of trials and errors this filtering and normalization step were selected. First, if a coordinate does not include damage that coordinate was not only removed from damage-seq, but also removed from other three NGS data. Secondly, each damage repair type (XR-seq: CPD and [6-4]PP) was divided by its corresponding damage types (Damage-seq: CPD and [6-4]PP).

3.3.10 Data and the result demonstrations

The preprocessing data analysis were demonstrated via different types of plots. The pairs, scatter, histogram, box plots, and heat maps which were created using ggplot2 library of programming language R. Only the scatter plots were included in this research due to their sufficiency (see Appendix A). On the other hand, the model prediction results were illustrated in Python programming language.

3.3.11 Data and code availability

The scripts run in this research are available at https://github.com/CompGenomeLab/globalNERepair

4. RESULT

4.1 Results

4.1.1 Strategy for predicting damage normalized repair levels in open and closed chromatin using histone modifications

The Figure 4.1 details the steps of building a machine learning model that can predict the level of NER. Using four different types of sequencing data types, we constructed a dataset for predicting damage normalized repair regions where there are high or low repair levels. Nevertheless, the main dataset, which we focused on for predicting repair levels, is the ChIP-seq data (histone markers/modifications). The genome was divided into 5kb windows (bins). Every read was mapped on the human reference genome (hg19) and then, read count numbers of each sequencing data type were acquired at the given bins. In order to overcome the depth of coverage and breadth of coverage problems between the cross-NGS data, a normalization method called RPKM as applied on the mapped read's count values. To learn more about the RPKM calculation please refer to Methods section and see "RPKM calculation" subheading. RPKMs were calculated from the read's count numbers and RPKM values of DNase hypersensitivity sites, (9 - 12, 14, 60, 61) histone modifications, repair levels (from XR-seq) and damage levels (from damage seq) were used as inputs for the model after the additional data filtering and normalization steps (Fig. 4.1).

DNase hypersensitivity sites were utilized to classify 5kb windows as either open (top third percent, highest, RPKM values) or closed (bottom third percent, low-

est, RPKM values) chromatin. Similar to DNase-seq, the repair values, which were normalized by damage values, were used to designate windows as high (top thirty percent of RPKMs) or low (bottom thirty percent RPKMs) binary repair levels. RPKM values of histone markers were used as predictors for the repair levels. Instead of using RPKM values of histone modifications directly for the predictions, we decided to bin the RPKM values for each modification into four quartiles for ease of interpretability. A cross-chromosome training and testing approach was used: 80% of the genome was used for model training and the remaining 20% was used for testing the machine learning models. Test samples were further segregated as open or closed chromatin if needed and the same model was used to evaluate performance on both states (Fig. 4.1).

Preprocessed data were directly used in three types of scatter plot analysis: G-NER – histone markers, G-NER – DNase hypersensitivity sites, and histone markers – DNase hypersensitivity sites. Chromatin accessibility and NER scatter plots were showing positive correlations in both UV damage types. However, CPD damage NER repair had the highest positive correlation with the chromatin states (see Fig. A1. 17 - 18). The scatter plots analysis between UV damage G-NER repair and histone modifications also displayed positive correlations in HeLa (see Fig. A1. 34 - 93). Between the UV lesion repairs, CPD damage repair had the highest positively increasing association with the histone markers (33). In both UV-induced photoproduct repair, H3K36me3 and H3K36me3_1's RPKM distribution patterns demonstrated fewer positive correlations than the rest of the modifications excluding H3K9me3, H3K27me3, and H3K27me3_1 (see Fig. A1. 34 - 93). Similar positive trends were observed between the two chromatin states (euchromatin & heterochromatin) and same set of histone markers too. However, H3K36me3 were less positively correlated compared to other gene activity related modifications. H3K9me3 and H3K27me3 were the markers of inactive genes (59). Although they also presented some degree of positive correlation with the G-NER, they were not as significant as the linear positive sample distributions of other markers (see Fig. A1. 34 - 93). The sole unexpected correlations were from H3K36me3. H3K36me3 localizes at the active genes and functions for gene activations like the H3K4 methylations (62). However, H3K36me3s did not have the same pattern with H3K4me3 (see Fig. A1. 37 - 39, 45 - 46). Normally, H3K36me3 is highest at gene bodies and starts to lose enrichments as moving to upstream and downstream of gene bodies (59, 62). Thus, its insufficient positive correlation with the G-NER was not anticipated. Any kinds of sequencing data related problems for H3K36me3s might be the reason for this correlation results. On the other hand, UV radiation might have affected the distribution of epigenetic markers within the cells which may end up with unex-

pected results. In general, these raw data correlation plots may indicate that histone modifications may be the determinants of G-NER efficiency.

Not only testing the genome-wide effect of histone markers on the repair, but also the RPKM distributions of each histone marker at the four 5kb long genomic segments were tested (Fig. 4.2). Four genomic segments are 5kb upstream of TSS (U_TSS), 5kb downstream of TSS (D_TSS), 5kb upstream of TES (U_TES) and intergenic regions. For each of these regions, the same preprocessing, filtering, normalization, and training-testing strategy were applied.

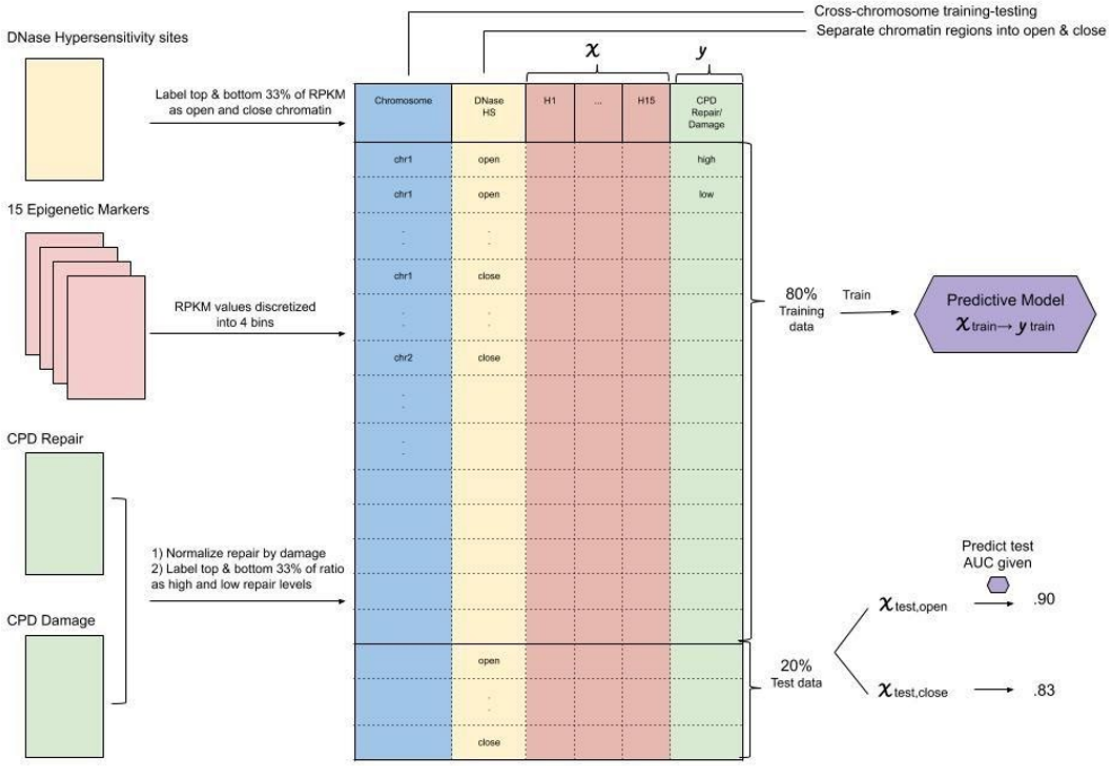


Figure 4.1 The demonstration of model development matrix which was used to predict the damage repair.

Chromosome name and chromosome start, and end site coordinates are found at the first three columns. DNase I hypersensitivity is found in the second column. Histone markers' discretized RPKM values can span from 9 up to 15 columns depending on the cell line. The last column of the matrix consists of filtered and damaged normalized binary repair values (High or low repair).

In Figure 4.2, the model training data were used for predicting CPD repair levels in the whole- and compartmentalized genome for HeLa cells. The RPKM distributions of each histone modification at the whole genome and four genomic segments (bins), where there is high or low damage repair, were exhibited by split violin plots (Fig. 4.2). At the whole genome (first panel), we found out that the histone markers

were localized throughout the genome nearly independent from the heterogeneous activity of G-NER. Even though most markers did not show an evident enrichment at high/low damage repair regions, there were also small but undeniable histone marker enrichments at these sites too such as H3K79me3, H3K27ac, H3K9ac, H3K4me3, H3K4me3_1 (Fig. 4.2).

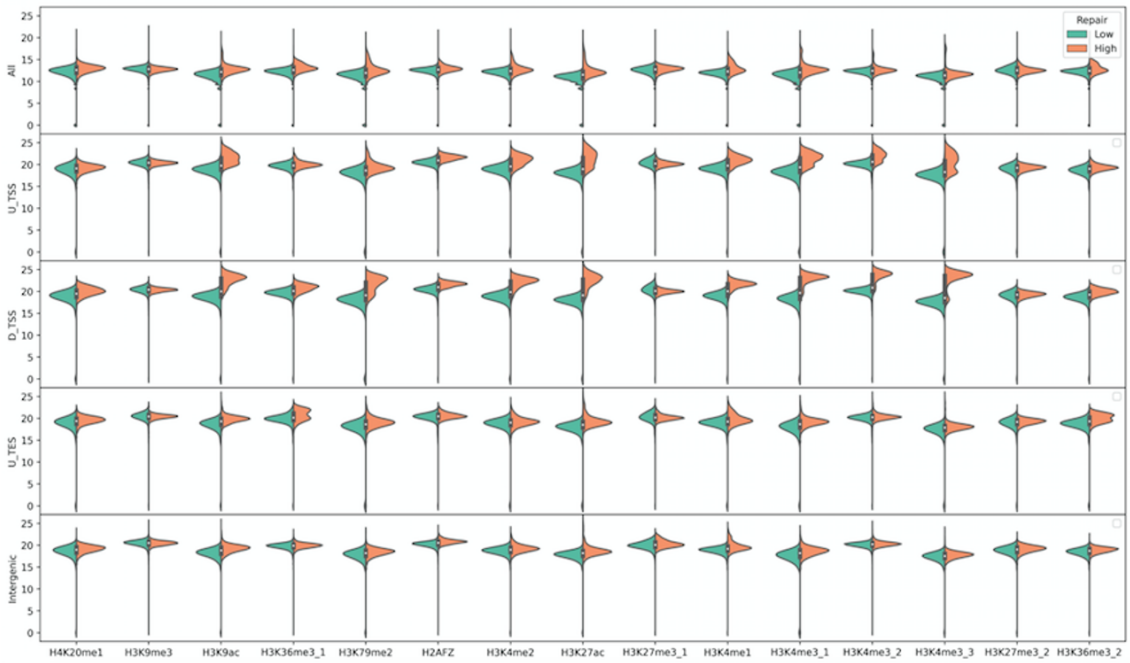


Figure 4.2 The illustration of model training set histone marker's RPKM distributions' in split violin plots.

RPKMs plotted on both the whole genome and four genomic segments where there is high or low CPD damage repair. The first panel (All) indicates all RPKM distributions of the corresponding histone modifications. The second panel (U_TSS) indicates the 5kb upstream of the transcription start site. The third is the 5kb downstream of the transcription start site. The fourth and fifth are 5kb upstream of the transcription end site and intergenes, sequentially. The positioning of the genes at sense and antisense strands were considered during data extraction and calculation. The green histograms represent low repair at given genomic regions and the red histograms declare the high repair at given genomic regions.

However, at U_TSS and D_TSS, some markers showed evidently higher RPKM distributions in regions where there are high repair rates. For example, H3K9ac, H3K79me2, H3K4me2, H3K27ac, H3K4me1 and H3K4me3 were enriched in 5kb U_TSS regions where global NER efficiency is high (Fig. 4.2). A similar trend was also observed in D_TSS which represents the part of a gene body. This observation is consistent with previous published data which show the listed histone markers are known to be enriched either at enhancer and/or promoter regions and involved in transcription initiation or elongation events (45, 59, 62, 63).

Conversely, at U_TES and intergenic regions, the distributions of histone modification RPKM values did not exhibit large differences between the regions where repair is high or low. The pattern of histone marker enrichments at U_TES and intergenic regions was similar with their enrichment at the whole genome (first panel). On the other hand, the 5kb vicinity of TSS, where repair sites are hotspots, depicted higher RPKMs. (Fig. 4.2). Most histone markers were genomic segment dependently increased or decreased at the repair sites but, still, some histone markers, such as H3K36me3, H3K27me3, and H3K79me2, were found at the genomic regions with relatively similar degrees (see Fig. A1. 1-4). However, findings show that histone marker enrichments may also depend on damage type and the repair pathway. For instance, H3K9me3 and H3K27me3s had higher RPKM values at regions where there are low repair levels (Fig. 4.2). Besides, the same experiment conducted with other damage types exhibited slight differences for some histone markers found at the same genomic segments (see Fig. A1. 1).

4.1.2 Histone modifications can be used to predict genome wide repair levels across cell-line, damage types and repair pathways

We trained our model using all histone modifications to predict genome-wide repair levels across 3 different cell types: Hela, GM12878 and NHF1. See Figure. A1. 5 for the repair prediction scores of histone markers on GM12878 and NHF1. GM12878 cell had been subjected to cisplatin damage and had both TC-NER and G-NER for repairing inflicted damage. Similar to the GM12878, NHF1 had the same NER machineries but it was exposed to the same damage source with the HeLa. For brief information about cell lines and their features please visit Table 4.1. For the Hela cell-line, repair was measured within 12 minutes after the UV exposure, during this time frame only the global NER pathway is known to be active. Therefore, the HeLa cell line provides an environment, which lacks the repair effects of TC-NER, to monitor the efficiency of G-NER alone. In the data, cisplatin damage had two biological replicates but in UV damage bulky adducts CPD and [6-4]PP alone had two biological replicates hence, there were four replicates in HeLa and four replicates in NHF1. Therefore, it is possible to see the same damage repair more than one time in the analysis. For instance, two biological replicates of the UV damage repair appear as [6-4]PP replicate A and [6-4]PP replicate B in Figure 4.3.

Here, we would like to show that histone modifications can be used to accurately

Table 4.1 The cell lines and their corresponding damage sources and lesions, and repair pathways involved

Cell Type	Damage Type	Repair Pathway Involved
HeLa	UV (CPD, [6-4]PP)	Global NER
NHF1	UV (CPD, [6-4]PP)	Global + TC-NER
GM12878	Cisplatin	Global + TC-NER

predict repair levels in HeLa cells (Fig. 4.3) as we as in other cell lines (see Fig. A1. 5). In all three cell lines, there were good agreement between the AUC values of damage repair biological replicates similar to the positive correlation observed in the scatter plots between G-NER and histone modifications (see Fig. A1. 4, 34-93). However, there were small differences in the repair prediction accuracies between these cell lines. In HeLa cells, the NER prediction accuracies were slightly higher than other two cells. This high UV damage repair prediction scores in HeLa cells may be due to the high number of model inputs provided. In HeLa cell line, some histone modifications were represented by more than one (different runs of same ChIP-seq data) hence, providing more information to model may increase the repair estimation.

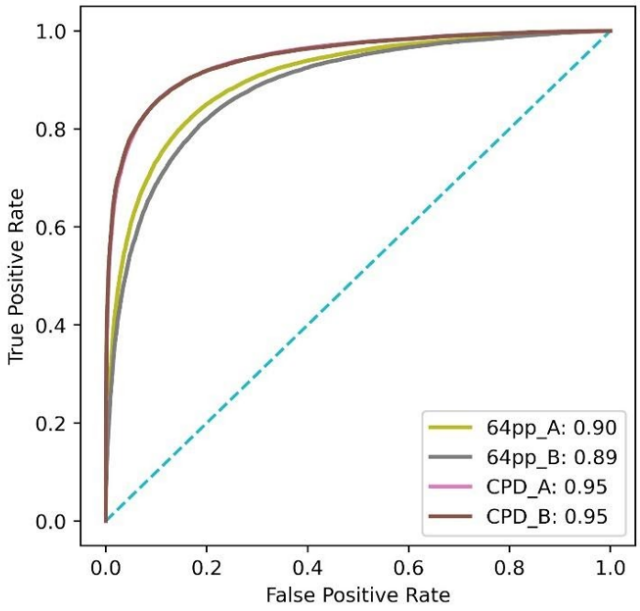


Figure 4.3 The demonstration of model UV damage bulky adducts repair predictions in HeLa cell.

The x-axis represents the correct repair estimations performed all along the genome by the model and y-axis shows wrong repair predictions exhibited as correct by the model. The vertical blue dotted lines indicate the AUC threshold (0.5) meaning that it is not informative. The grey and yellowish colored curves are the plotted recall values. The small box within the box shows the corresponding AUC score (prediction accuracy) with respect to the damage type.

Remarkably, the prediction accuracy was highest for the CPD damage repair in Figure 4.3 (test AUC 0.95 in both Hela cells and NHF1 cells). For [6-4]PP, the repair prediction accuracy was lower than that seen for CPD (Fig. 4.3). However, repair pathways (global NER for HeLa and global + TCR NER for NHF1) did not seem to affect the model prediction performance alone; cell line specific features were likely to affect these model prediction scores at this time. The same is also true for NER repair predictions in NHF1. Moreover, in the scatter plot analysis of G-NER and histone markers CPD damage repair demonstrated more linear positive trend with histone markers compared to other damage types (see Fig. A1. 34 - 63). The difference between the accuracy of CPD and [6-4]PP could be explained by the dissimilarities in CPD and [6-4]PP damage composition, their repair efficiency, and the repair pathways that repair them.

Since there is only one damage type in GM12878, we only assessed the cisplatin damage and its biological replicates' repair performance. Finally, the cisplatin damage repair level predictions (AUC = 0.92) in GM12878 cells were also high for both replicates. Importantly, a distinct higher or lower repair accuracy was detected in none of the cell lines. (Fig. 4.3).

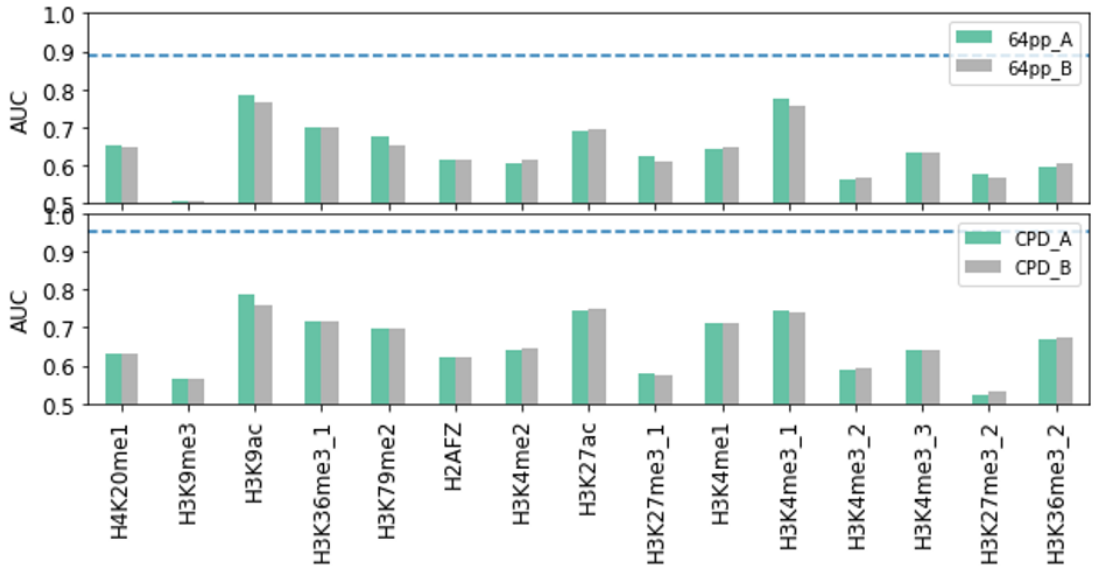


Figure 4.4 The effect of all and individual histone modifications in HeLa cell line repair.

The blue horizontal dots imply the model prediction accuracy for the repair, when all histone modifications together. x-axis is AUC score (accuracy of model) and y-axis is the histone modifications used as model input for prediction in the HeLa cells. The bars represent the individual markers' contribution to damage repair prediction when they are used to predict alone. Green and grey boxes represent the contribution of individual histone markers to the prediction of two biological replicates of damage repair, replicate A and replicate B, respectively.

Next, we trained decision stumps using individual histone modifications to establish a baseline of classification performance. The impact of all histone modifications to the overall model performance was also displayed in the graph using blue horizontal dots. AUC scores of these decision stumps are shown in Figure 4.4 and Figure A1. 5. Not surprisingly, we observed that a model that integrates all features (dotted horizontal line) outperforms better than models using single features. However, it is interesting to note that the models using individual histone modifications are also able to predict G-NER repair levels (Fig. 4.4). In Hela cells, decision stumps individually using H3K9ac, H3K27ac, and H3K4me3_1 had AUC scores close to 0.8. There was a discordance between the AUCs of some histone modifications such as H3K4me3_1, H3K4me3_2, H3K4me3_3, H3K27me3, and H3K27me3_2. As stated above, some histone markers were included more than one, such as H3K4me3_2 and H3K4me3_3, in the same cell line (HeLa and GM12878). The reason was providing more input to the model to enable it to interpret more reliable information. These were supposedly the same histone markers but published by different lab groups or different runs of the same experiment. The inconsistency of H3K4me3_2 and H3K4me3_3 in HeLa or H3K4me3_1, H3K4me3_2 and H3K4me3_3 in GM12878 were likely due to the differences in the cell physiological states, such as cell cycles, or lab protocols, which the experiments were performed accordingly, or the possible experimental errors. Unlike model prediction results, HeLa cell preprocessed raw data scatter plot analysis between the G-NER, and histone modifications exhibited nearly similar positive correlations independent from the repetition of same histone markers in the same cell line (see Fig. A1. 34 - 93). This is plausible because the data content and quality between the preprocessed and training-testing may differ.

It is also interesting to note that despite the different cell types and histone marker prediction values, the pattern of AUC prediction was overall similar for all cell types. Almost the same set of histone markers were contributed to both damage type repair in NHF1 and HeLa. The histone marker correlations with G-NER were also coherent with the repair predictions in other two cell lines as well (Fig. A1. 5). For instance, for both damage types, in HeLa cells, H3K9ac and H3K4me3_1 had higher prediction accuracies while H3K9me3 and H3K27me3 had lower prediction accuracies (Fig. 4.4). Similarly, in NHF1 cells, H4K20me1, H3K4me1, and H3K4me3 showed high prediction accuracy while H3K9me3 and H3K27me3 showed low prediction accuracies. As mentioned, there are different histone marker contribution rates and genomic segment preference of them in the model prediction although the overall contribution pattern of histone modifications to model repair predictions was same (Fig. 4.2 and 4.4). It is well confirmed that different cell types may have different patterns of histone modifications and this is part of what makes one cell type unique

from another despite all cells from the same individual having the same genome. To extend this, our results suggest that the pattern of cell-type specific histone modification contributions also manifests itself and, apparently, this cell-type specific patterns of the histone modifications are predictive for repair efficiency levels (Fig. 4.4). These are also in line with our scatter plots findings between the G-NER and each histone marker (see Fig. A1. 34 - 93). H4K20me1, H3K4me1, H3K4me3 and H3K4me2 exhibited relatively higher prediction accuracies at both bulky DNA adducts repair in NHF1 cells (see Fig. A1. 5).

Finally, for cisplatin damage, the decision stumps (histone modifications), which show high prediction accuracies, include H3K36me3_1, H3K79me3, H3K27ac, H3K9ac and H3K36me3 in GM12878 cells (see Fig. A1. 5).

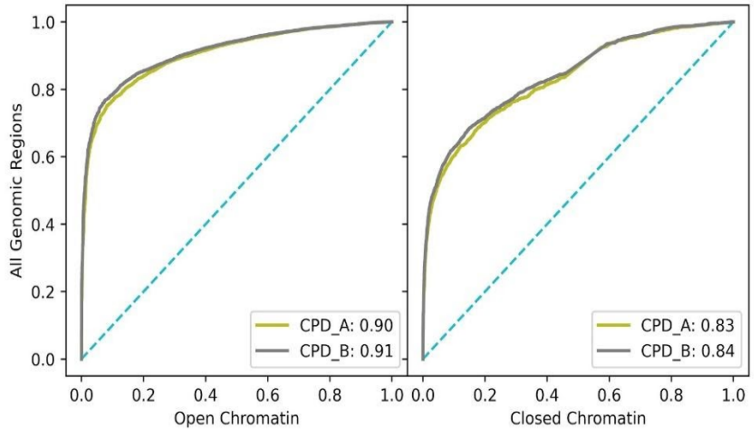
As we mentioned before, the RPKMs of each histone modification were binned into quartiles (0,1,2,3). By doing this, we can quantify the levels of histone modifications and train the model easily. For instance, if a given 5kb region in the genome has 0 value for H3K9ac then, we know that as compared to the genomic distribution of H3K9ac, this region has a low level of H3K9ac. Similarly, a genomic region with a 3 for H3K9ac has high levels of this modification.

Decision stumps are based on a single cutoff of the level of a single histone modification. For instance, for H3K9ac, a possible decision stump might decide on the repair levels by saying that a genomic region with H3K9ac value 0 or 1 is a low repair region and H3K9ac value 2 or 3 is a high repair region.

The fact that some decision stumps, can make predictions for two bulky DNA adducts (UV damage) repair levels up to AUC score 0.8, in case of H3K9ac, H3K4me3, and H3K27ac, in HeLa cells. It is surprising and likely suggests a close relationship between these histone modifications and G-NER in HeLa cells (Fig. 4.4). Slightly less (AUC score = 0.76) contribution of H3K9ac was also monitored in GM12878 (see Fig. A1. 5). In all cases, technical replicates were in good agreement. This suggests that the patterns that we found are likely to reflect the reality (see Appendix A).

4.1.3 Repair levels are predicted better for open chromatin regions compared to closed chromatin regions

A)



B)

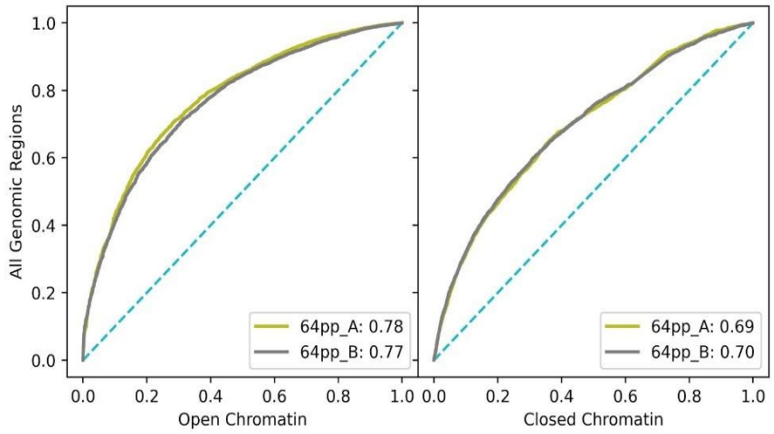


Figure 4.5 The demonstration of model repair predictions at different chromatin states in HeLa cells.

The x-axis represents the state of chromatins and y-axis shows AUC score. The vertical blue dotted lines indicate the AUC threshold (0.5) meaning that it is not informative. The grey and yellowish colored curves are the plotted recall values. The small boxes within each box show the AUC score (prediction accuracy) of the model with respect to the damage type which was repaired. A) is for model CPD damage repair prediction at both chromatin states and B) is for the same analysis but for [6-4]PP damage repair.

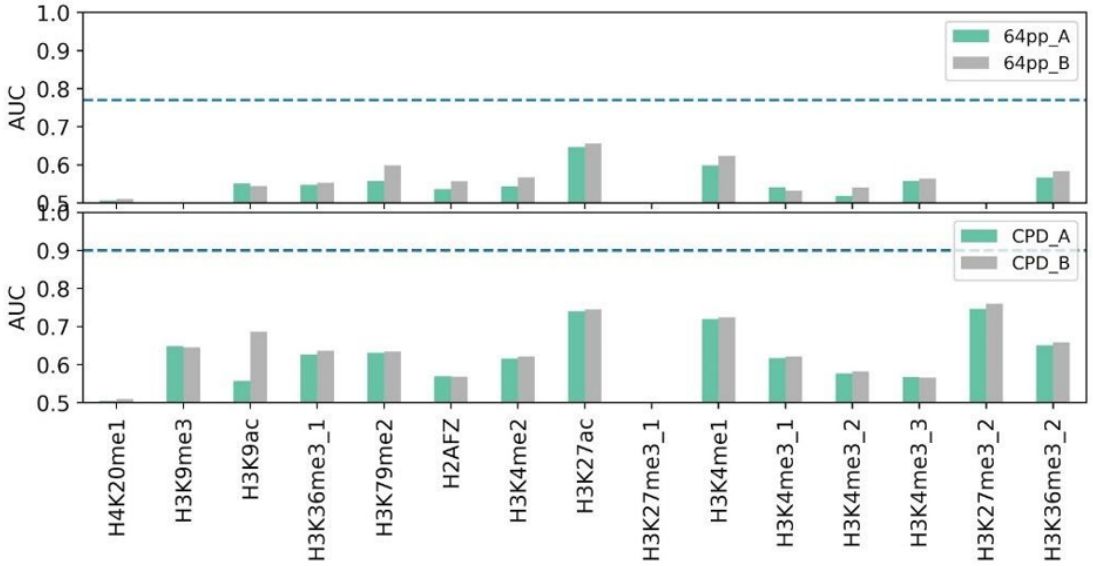
Next, we aimed to determine how chromatin accessibility affects the NER prediction levels (Fig. 4.5 and Fig. A1. 6). To do so, we separated only the test data into

open and closed chromatin regions. This is because separating training data as well did not enable us to observe chromatin state differences on the genome as clearly as applied at the test data (data not shown). The same model (trained irrespective of chromatin state, see Methods, Fig. 4.1) was then used to predict the repair levels on the regions designated as open and closed chromatin.

When comparing CPD and [6-4]PP damage repair in HeLa cells, the former had higher prediction accuracy irrespective of the chromatin type (Fig. 4.5). This is because of the dissimilarities in their chemical compositions and their strand specific repair rate differences. This same pattern was also seen when comparing repair level predictions which ignored the chromatin states (Fig. 4.2).

The scatter plot analysis also demonstrated a more linear positive trend for CPD damage repair at both histone markers and chromatins (see Fig. A1. 17, 18, 64 - 93). In HeLa cells, for both CPD and [6-4]PP bulky adducts, the accuracy of G-NER prediction efficiency was higher at open chromatin than at closed chromatin (Fig. 4.5). The same pattern of repair was also observed in GM12878 cell lines for the cisplatin damage repair (see Fig. A1. 6). However, in case of NHF1 cells, in both CPD and [6-4]PP damage types, the differences between the prediction accuracies at open and closed chromatin were smaller (see Fig. A1. 6). This might be explained by the cell physiological dissimilarities (genomic and cell cycle differences), possible cellular abnormalities after damage exposure or data related dissimilarities between the cell lines.

A)



B)

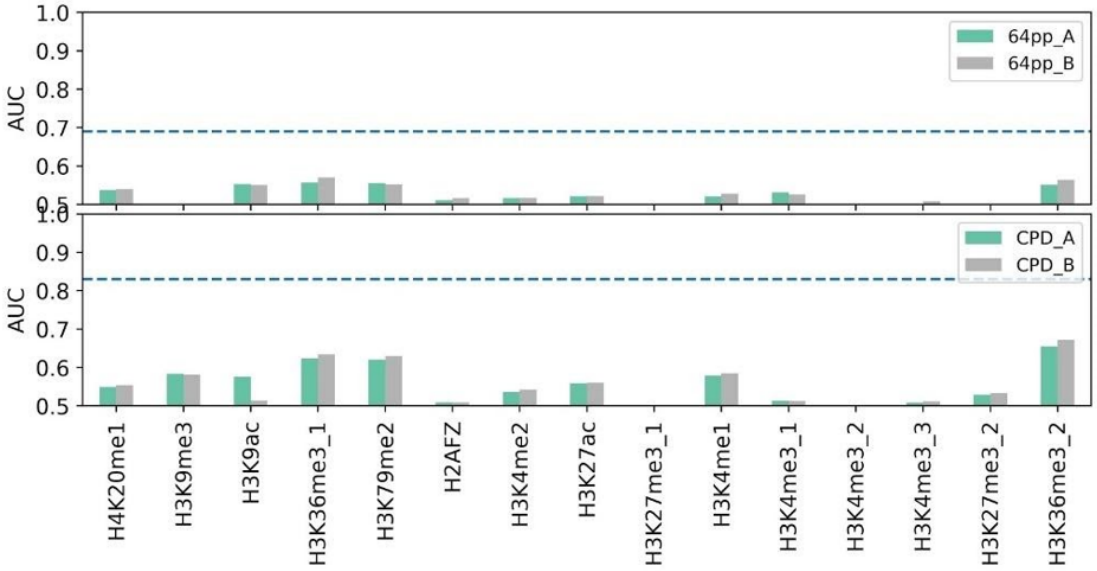


Figure 4.6 The effect of all and individual histone modifications on the repair efficiency in HeLa cell line with different damage and chromatin conditions.

The blue horizontal dots imply the prediction accuracy for repair when all histone modifications together incorporated. x-axis is AUC score (accuracy of model) and y-axis is the histone modifications used in HeLa cells as model input for prediction. The bars represent individual markers' contribution to damage repair prediction when they are used to predict alone. Green and grey boxes represent the contribution of individual histone markers to the prediction of two biological replicates of damage, replicate A and replicate B, respectively. A) exhibits the model prediction at open chromatin states, B) is for same analysis at closed chromatin states.

To understand which histone modifications are best to contribute to repair collectively or alone, we used decision stumps to predict repair levels at open and closed chromatin states separately. The results for HeLa cells are shown in Figure 4.6 and those for NHF1 and GM12878 are in Figure A1. 7. Apparently in all cell lines, damages were repaired preferentially better at open chromatin states which is coherent with the DNA damage repair and DNase hypersensitivity scatter plots (see Fig. A. 15 - 18). In all cases, irrespective of the chromatin type, cell type, damage type or repair pathway, a model incorporating all histone modifications performed clearly better at predicting repair levels than individual decision stumps as in Figure 4.4.

Comparison between open and closed chromatin regions for HeLa showed that the prediction accuracy was higher for open chromatins compared to closed chromatins for all decision stumps (Fig. 4.6). Moreover, comparisons between CPD and [6-4]PP damage types showed that accuracies of individual decision stumps were again higher for the former. This bulky DNA adduct repair pattern also holds for NHF1

cells and the HeLa cell scatter plot analysis as well (see Appendix A). The same result and explanation are valid for the CPD and [6-4]PP bulky DNA adducts repair rate differences which were mentioned at Figure 4.3 and 4.5 as well.

The prediction accuracy across two damage types were the same for open chromatin regions. For instance, in HeLa, both bulky DNA adducts (UV damage) repair predictions at open chromatins, H3K27ac and H3K4me2 had higher accuracies compared to their neighbors (Fig. 4.6). This suggests that, in each cell line, same histone modifications are related to repair irrespective of the damage type but, the chromatin states seem to affect that repair prediction. Thus, chromatin states is another significant factor and it is both coherent with the previous results and previously published studies. Coherent with the deduction, histone markers differed between open and closed chromatin regions (Fig. 4.6 and Fig. A1. 7). For instance, in the case of open chromatin, repair levels of CPD and [6-4]PP were best predicted by decision stumps using H3K27ac and H3K4me1. However, for the case of predicting repair levels for closed chromatin, the decision stump using H3K36me3 was the best predictor irrespective of the damage type (Fig. 4.6). In NHF1 cell line, H3K36me3, H4K20me1, and H3K79me2 were predictive for the repair in both chromatin states although they lost effect on repair at closed chromatin states (see Figure A1. 7). This variety in the histone marker contributions to repair may also imply the role of histone modifications functioning, which could depend on damage type and chromatin states, within the cell.

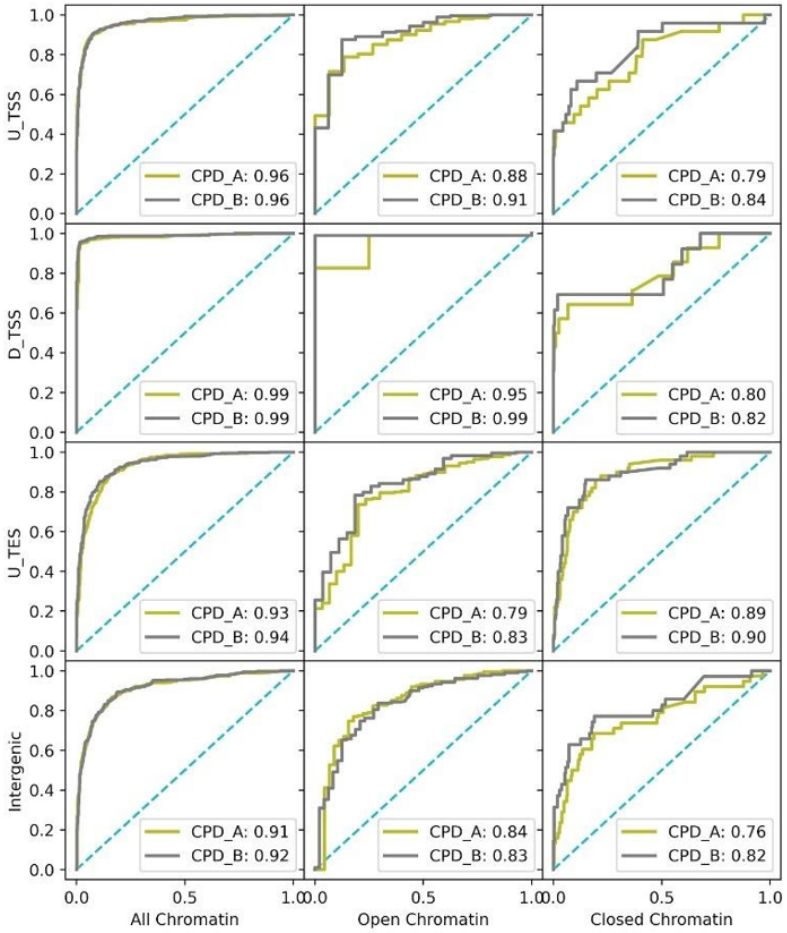
According to the scatter plots between the individual histone markers and DNase hypersensitivity sites in HeLa cell line, almost all histone markers, except H3K9me3, were at some degree positively correlated with the accessible chromatin regions. However, in these plots, especially H3K27me3 and H3K27me3_1 exhibited remarkably fewer positive correlations than the rest of the histone modifications. On the other hand, H3K36me3 and H3K36me3_1 were more positively correlated than H3K27me3s but not as strongly correlated as the other histone markers (see Figure A1. 19 - 33). The model repair prediction with H3K36me3 at the closed chromatin states did not match with its scatter plots analysis (Fig. 4.6). Similarly, H3K27me3 was enriched at the open states but not at close states compared to the findings from scatter plots. However, the reasons for the discordant findings between histone markers and chromatin accessibility can be due to the cell line specific distribution patterns of H3K27me3 and H3K36me3 (63) and impacts of damage exposures on the histone markers distributions. These possible reasons might also be valid for explaining the unexpected findings between the H3K36me3 and NER correlation analysis too (see Figure A1. 46 - 46, 60 - 61, 75 - 76, 90 - 91). Nevertheless, in general, the scatter plots of histone modifications with the chromatin states or repair

manifested compatibility with our raw data, data preprocessing and the knowledge from the literature.

The highest open to closed chromatin differences were observed in GM12878 cell lines (See Figure A1. 7). Cisplatin damage repair were best predicted by H3K27ac, H3K36me3, H3K36me3_2, and H3K79me2 but they also lost their predictive strength on repair at closed chromatin states (See Figure A1. 7). Therefore, there is an obvious change at histone modifications for predicting the repair levels for euchromatin and heterochromatin regions (Fig. 4.6). Repair levels can be predicted better for the euchromatin regions (64). This is likely due to the dynamic differences at the heterochromatin sites and the reorganization differences at these sites after the UV damage. The possible effect of histone markers on DNA repair is estimated from their functions that might be associated with repair. However, histone modifications are likely not to involve in DNA damage repair processes evidently.

4.1.4 There are differences in the model's repair predictions between the four genomic regions

A)



B)

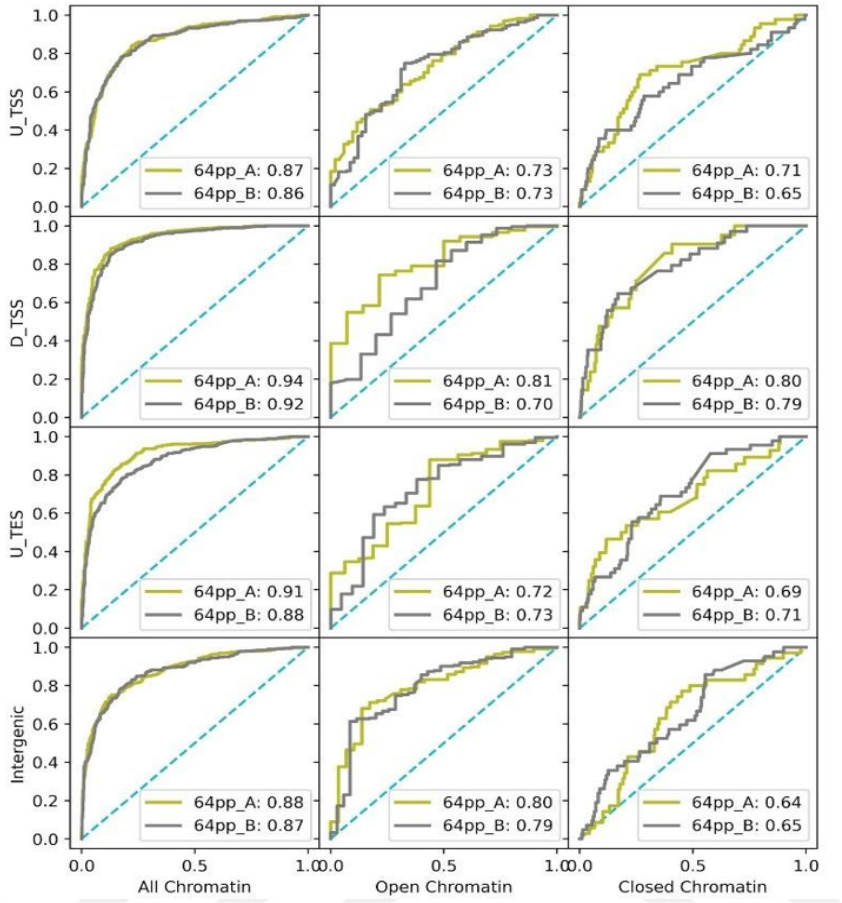


Figure 4.7 The demonstration of model repair prediction at different genomic segments and chromatin states.

The y-axis represents the four genomic segments and x-axis is for indicating at which circumstances the model function. The above 16 boxes are for model prediction for CPD damage and the below 16 are for [6-4]PP repair. The vertical blue doted lines indicate the AUC threshold (0.5) meaning that it is not informative. The grey and yellowish colored curves are the plotted recall values. The small boxes within each box show the AUC score (prediction accuracy) of the model for corresponding damage type.

To understand whether the prediction repair levels differs or not depending on the genomic regions, we generated datasets which contain samples from histone markers and repair corresponding to genomic segments 5kb upstream of TSS, downstream of TSS, upstream of TES and in intergenic regions, respectively, in Figure 4.7 (See Methods and Fig. 4.1).

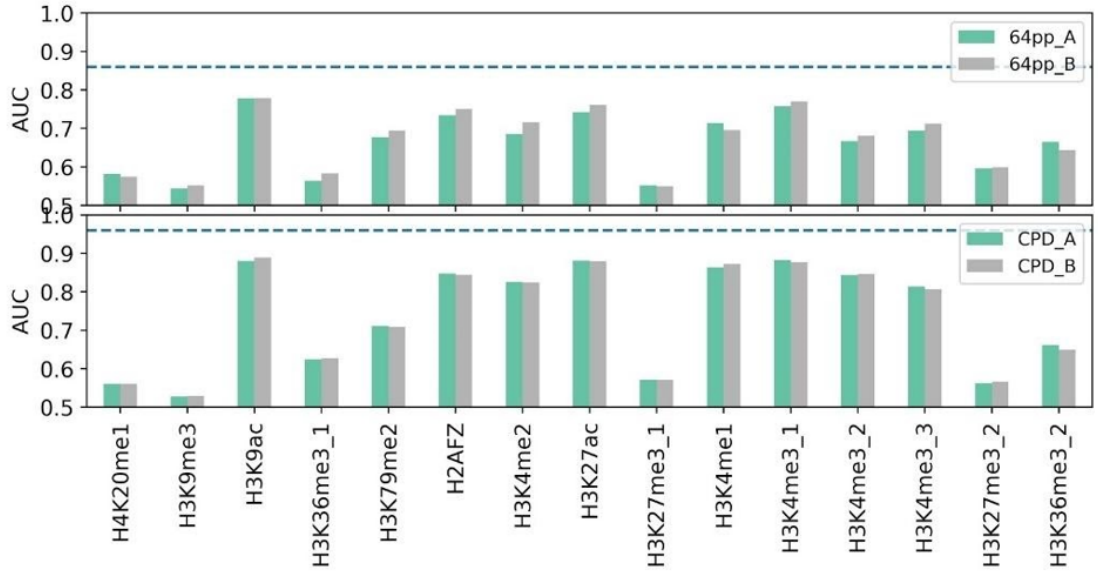
We found that in case of HeLa cells, for both CPD and [6-4]PP damage types, genomic segments (U_TSS, D_TSS and U_TES) had higher prediction accuracies for NER levels compared to the intergenic regions (Fig. 4.7). This is because

intergenic regions were less informative for the model. Among the genomic regions, D_TSS seems to have the highest prediction accuracy in both damage types. These patterns hold across damage types (UV and Cisplatin) and cell lines (HeLa, NHF1, GM12878) (See Figure A1. 8).

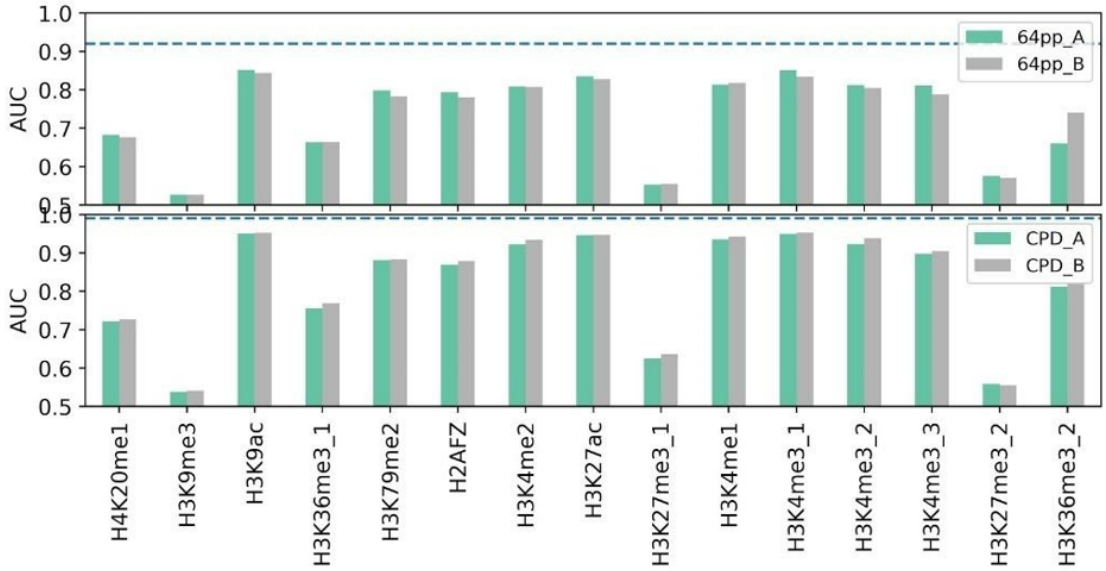
Apparently in all cell lines, damages were repaired better at open chromatin states and more likely to repair CPD damage, which are coherent with the scatter plots between DNA damage repair and DNase hypersensitivity sites as well as between DNA damage repair and histone modifications, although the score differences were not immense in Figure 4.7 (see Appendix A). In HeLa cells CPD repair, in case of U_TES, the model predicted G-NER repair with higher accuracy at closed chromatin than open regions (Fig. 4.7). In general, similar repair predictions were observed between the open and closed chromatin but still considerably dominant accuracies were observed at the open chromatin states such as CPD damage repair at D_TSS. However, observing a higher repair accuracy at closed chromatin regions is exceptional (Fig. 4.7). This is likely to be a special instance for HeLa cells at the U_TES regions or although our method had been checked, still this might be originated from a methodological error.

In general, for these different genomic regions, it appears that predictions for open chromatins were better than those for closed chromatins. In addition, differences were also monitored between biological replicates of the damage repairs due to their small sample sizes (Fig. 4.7 and Figure A1. 8).

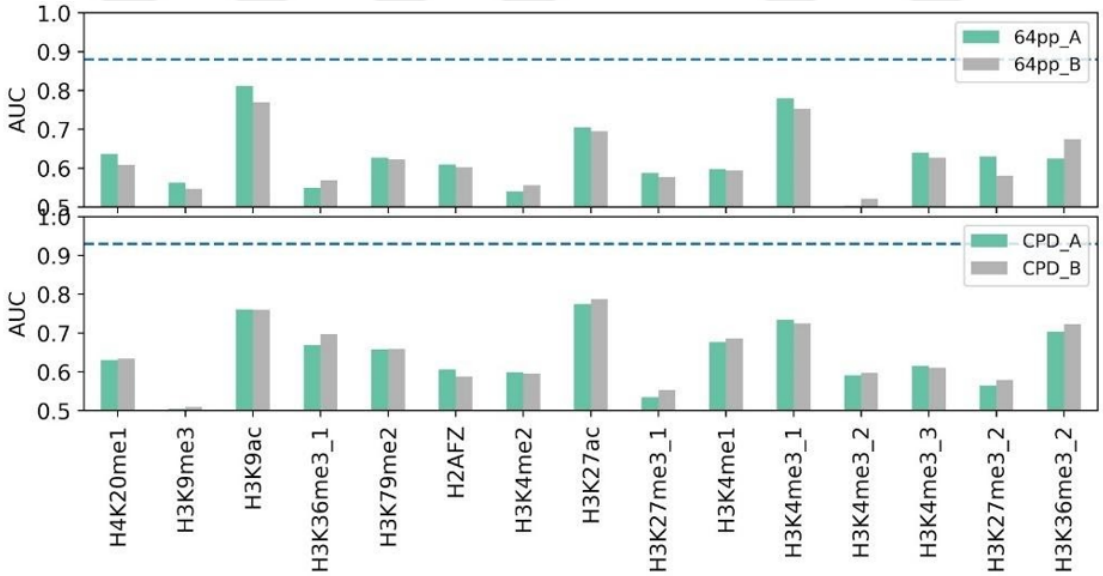
A)



B)


AUC
64pp_A
64pp_B
CPD_A
CPD_B
H4K20me1
H3K9me3
H3K9ac
H3K36me3_1
H3K79me2
H2AFZ
H3K4me2
H3K27ac
H3K27me3_1
H3K4me1
H3K4me3_1
H3K4me3_2
H3K4me3_3
H3K27me3_2
H3K36me3_2

C)


AUC
64pp_A
64pp_B
CPD_A
CPD_B
H4K20me1
H3K9me3
H3K9ac
H3K36me3_1
H3K79me2
H2AFZ
H3K4me2
H3K27ac
H3K27me3_1
H3K4me1
H3K4me3_1
H3K4me3_2
H3K4me3_3
H3K27me3_2
H3K36me3_2

D)

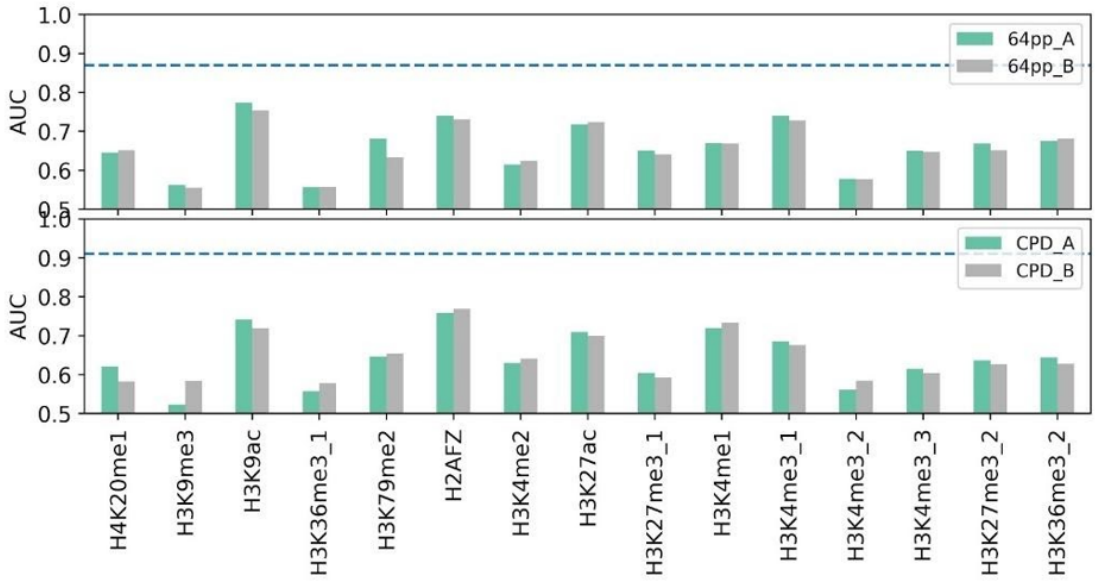


Figure 4.8 The effect of all and individual histone modifications on model repair in HeLa are illustrated

(see Figure A1. 9 for other cell lines). The blue horizontal dots imply the prediction accuracy for repair when all histone modifications incorporated. x-axis is AUC score (accuracy of model) and y-axis is the histone modifications used in HeLa as model input for prediction. The bars represent individual markers contribution to damage repair prediction. Green and grey boxes represent the contribution of individual histone markers to the prediction of two biological replicates of the damage repair, replicate A and replicate B, respectively. A) is for showing the effect of features on repair at U_TSS. B) is for showing the effect of features on repair at D_TSS C) is for showing the effect of features on repair at U_TES. D) is for showing the effect of features on repair at intergenes.

The influence of histone modifications on the repair prediction was analyzed individually and collectively with respect to their enrichments on the four genomic segments for HeLa cells in Figure 4.8. First, when all histone modifications were incorporated as features for the model, the prediction accuracies were changing between the 9.1 - 9.9 for CPD and 8.8 – 9.2 for [6-4]PP repair depending on their predictions at the genomic segments (Fig. 4.8). Therefore, histone modifications evaluated together for the repair exhibit considerably preferable results than utilizing them alone and, again, CPD predicted slightly better. These findings were also encountered in the previous model predictions and scatter plot analysis results too (see Appendix A). At U_TSS and D_TSS, the individual decision stumps had higher AUCs compared to U_TES and intergenic regions. This is consistent with the differences observed between the histones' marker RPKM distributions at the U_TSS and D_TSS compared to other genomic segments in Figure 4.2.

Moreover, it is interesting to note that the pattern of decision stumps was different across the genomic regions but nearly conserved for the damage types. For instance, irrespective of the damage type, for U_TSS regions, H3K9ac, H3K27ac had high AUC scores, while H3K9me3 and H3K36me3 had lower AUC scores (Fig. 4.8). The patterns of high and low AUCs seem to be similar for U_TSS and D_TSS, but different for U_TES and intergenic regions. This might be related to and reveal the histone modifications primary enrichment region preferences on the genome and their cellular functions. The scatter plot analysis and the previous literature (62) reported that H3K36me3 enriches at the gene body and loses enrichments as moving towards to TSS and TES (see Figure A1. 46 - 46, 60 - 61, 75 - 76, 90 - 91). Remarkably, this enrichment trend is exactly the stated enrichment pattern of H3K36me3, which can be seen in Fig. 4.8, but the chromatin accessibility incorporated results of H3K36me3 were a bit of controversial in our research (Fig. 4.6 and 4.7).

Comparing individual histone markers' repair prediction patterns to those seen for the predictions at all genomic regions (Fig. 4.8), it seems that the patterns observed for the whole genome are more than just being an average of every individual marker's effect on G-NER seen at each genomic segment. This is because gene annotation was performed by applying series of filtering steps, in other words significant amount of read information lost during these steps (see Methods). Moreover, some modifications together appear to enhance the repair prediction levels at specific genomic segments (Fig. 4.4). Of course, this is related with predicting the repair on more specified regions (genomic segments rather than whole genome) which enabled more detailed monitoring of histone modifications effect on the repair.

In case of HeLa cells, decision stumps that exhibited high AUC scores across genomic regions included H3K9ac, H3K27ac, and H3K4me3_1 (Fig. 4.8). Of these, H3K9ac and H3K27ac were also shown to be one of the predictors most important for a model that incorporated all histone modifications effect on the repair (Fig. 4.4 and Fig. 4.9). NHF1 and GM12878 exhibited intriguing similarities with the G-NER efficiencies of HeLa cell. So that, same histone modifications contributed the prediction of NER at same genomic segments as in HeLa except; H3K9me3 and H3K27me3 were better predictors at D_TSS, and H3K27me3 were better predictor at U_TES in GM12878. On the other hand, in NHF1, H3K9ac on intergenes worst and H3K9me3 on U_TES were best effectors for the NER repair efficiency (see Figure A1. 9).

4.1.5 Feature importance using SHAP differs for repair types and cell lines

So far, we have looked at the repair predictions levels across chromatin states and genomic regions. In all cases, we have used decision trees for individual histone modifications to estimate baseline prediction accuracies. In Figure 4.9, we took a deeper look at models incorporating all histone modifications to find the ones that are the most important for making decisions on repair levels. To do so, we used SHAP values which will allow us to judge the model feature performance and in other words, enhancer or suppressor impact of each histone markers on the repair.

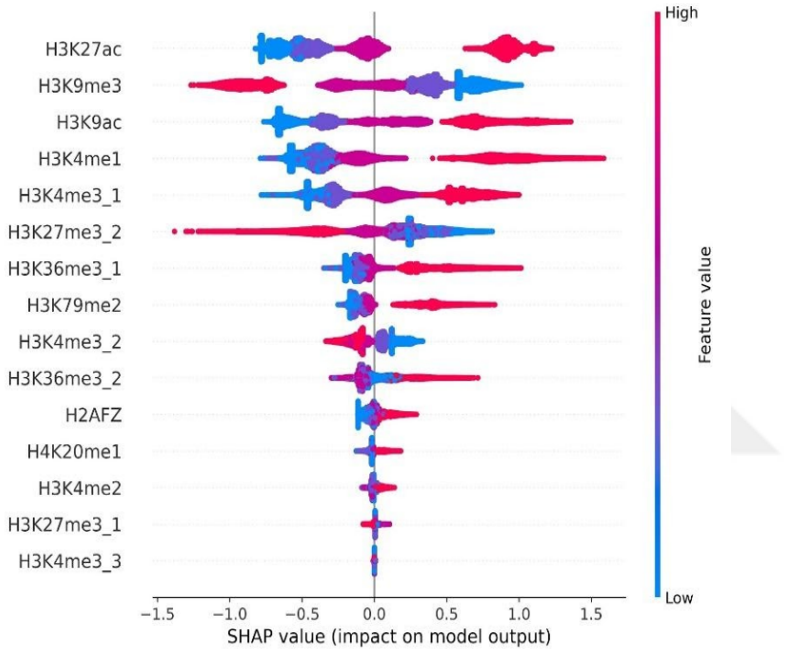


Figure 4.9 Demonstration of individual histone modifications (feature) importance for the model repair estimation performance.

All histone modifications were combined and introduced to develop the model.

An example SHAP plot depicting the prediction of CPD replicate A damage repair level in HeLa cells for whole genome and all chromatin states is shown in Figure 4.9. The features (histone modifications) were arranged in decreasing order of importance in the model incorporating all features. In this case, H3K27ac ranks the highest in importance of predicting repair levels. This evidently positive effect of H3K9ac on CPD damage repair acquired from SHAP value analysis (Fig. 4.9) exhibits us the direction of H3K27ac contribution at decision tree models in Figure 4.4 and 4.8.

Presence of H3K9me3 as the second most important feature was surprising, when we are looking at decision stumps for individual histone modifications, because it did

not exhibit strong prediction potential in different experimental setups (Fig. 4.2, 4.4, 4.6 and 4.8). It is also interesting to note that high values of H3K9me3 tend to lead the model to predict that a given genomic bin has low repair. This might help to explain why H3K9me3 is an important feature in combination with other histone modifications despite having no prediction potential by itself. Except some histone markers, the top four histone modifications did not differ in the repair prediction of other UV-induced bulky adducts and their biological replicates (see Appendix A).

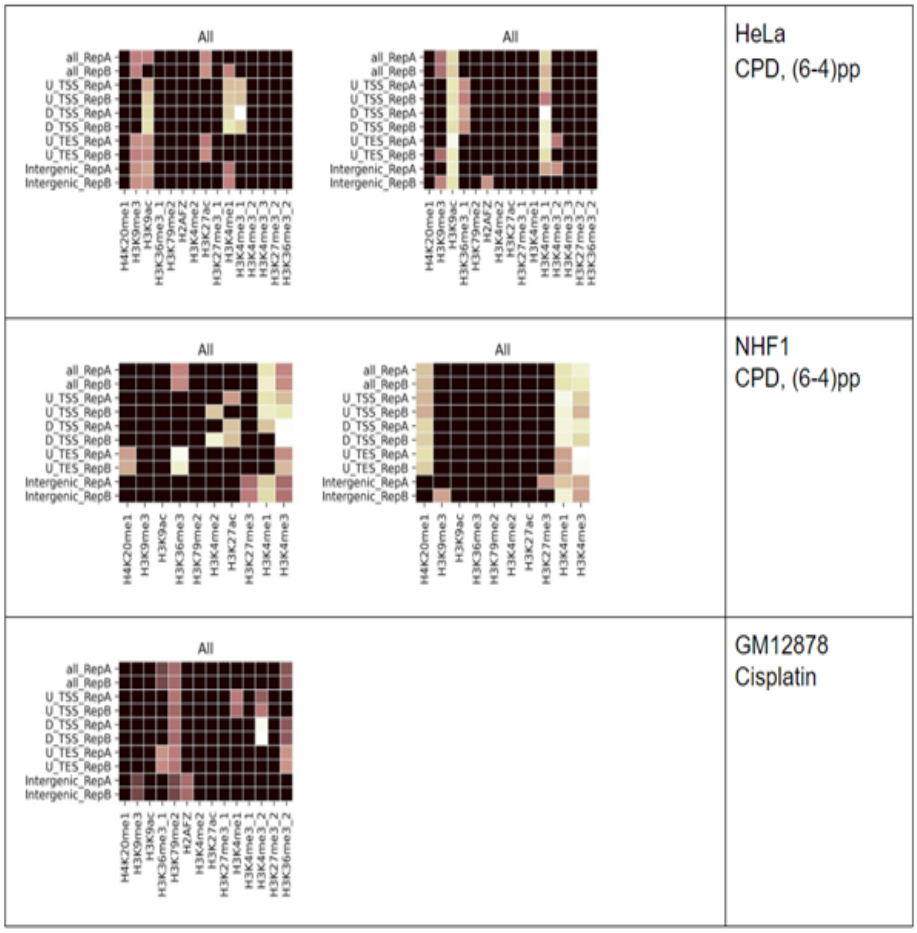


Figure 4.10 Three histone markers having the highest absolute SHAP values were selected for each genomic segment in every cell line.

Columns represent histone markers specific to the given cell line and rows represent genomic segments observed for each damage repair (UV and cisplatin) biological replicates. First two rows of every cell line are for observing contributions of individual markers to model estimations on whole genome. The color-coded boxes represent high absolute SHAP value. The light colors indicate their high significance for the prediction. Conversely, darker colors are vice a versa.

To understand whether the patterns of feature importance in predicting repair levels change across the cell lines, damage types, chromatin type, genomic regions, and the replicates, we calculated the absolute SHAP values for all models and highlighted top

3 most important histone modifications with a color code for each genomic segment (Fig. 4.10). Since representation of genomic segments and the method of showing SHAP scores of histone markers were not specific to reveal their relationship with the NER, only the model features' performances were assessed across conditions. A square corresponding to a given histone modification and a genomic region is light colored if it is among the top 3 most important features (highest absolute SHAP values). If a histone modification is not among the top 3 most important features for a given model, it is assigned a black color. The color spectrum between the white (highest) and brownish (lowest) colors inside the colored boxes reveals the magnitude of absolute SHAP values. Importantly, the absolute SHAP values do not provide information whether these three modifications are positively and negatively collaborated with the repair (Fig. 4.10).

It seems that patterns of histone modifications importance on model are defined by damage types and cell lines. For instance, in case of HeLa cells, CPD damage repair in H3K9ac, H3K9me3 and H3K4me1 seemed to be the most conserved and important histone markers across genomic regions. However, in the same cell type, for [6-4]PP damage, H3K9ac and H3K4me3_1 were mostly found across all genomic regions. For Cisplatin damage, H3K79me2 appeared to be conserved across genomic regions. Furthermore, no histone modification was common in three cell lines and damage types excluding H3K4me3 (almost conserved).

In case of NHF1 cells, CPD and [6-4]PP damage repairs showed different feature importance patterns (Fig. 4.10). However, it is important to bear in mind that the repair pathways in NHF1 and GM12878 were different from the HeLa cells. Moreover, it seems that despite the better prediction performance on open chromatin regions as compared to closed chromatin regions, the underlying features, which are important for making the prediction, remained almost unchanged within the damage types (Fig. 4.10). This is true across cell lines, damage types, and genomic regions.

Finally, there were differences between the marker patterns across genomic regions. For instance, at CPD damage repair in HeLa cells, H3K9me3 was not among the top 3 most important features for making repair predictions at U_TSS and D_TSS. Instead, H3K4me3_1 was the feature importance for making predictions at these regions. Similar differences were also presented in GM12878 and NHFl. Although there is no common histone marker found across all cell lines' same genomic segments or damage types; cell type and damage type specific conservations can be mentioned (Fig. 4.10).

5. DISCUSSION and CONCLUSION

The role of histone markers in DNA excision repair was researched before in 2016 on yeast cells (40). In this research, we have developed a ML model and used it as a tool to show another NER pathways associated factor which is histone markers. Moreover, we have studied G-NER, which is overshadowed by TC-NER inside the cell, alone and together with the TC-NER for exploring the effect of histone modifications on G-NER's efficiency. To perform these, various types of histone modifications were individually or collectively given to the ML model and that model used to estimate the NER damage repair regions all along the genome. Thus, it became possible to assess the predictive strength of each histone modification as well. If histone markers can be used to predict the presence or absence of repair positions correctly on the genome, that would unveil their interactions with the NER and prove that histone markers can be another effector for the NER machinery. Our ML model functions by assessing the relationship between genome-wide repair (XR-seq), damage (damage-seq), posttranslational histone modifications (chip-seq), and the (DNase-seq) chromatin accessibility (Fig. 4.1). During this process, many algorithms had been evaluated for algorithm selection and then, the chosen algorithm was tested with different inputs and parameters for the model development and optimization. Different filtering and normalization techniques were also tested on the preprocessed data for providing better model inputs and having more biologically meaningful findings (data not shown). After that ML model was utilized with four different types of omics data which had been previously preprocessed. To assess the model performance, we have conducted the time course analysis of [6-4]PP repair in NHF1 cells by training and testing the model with initial time (1hr) and late time points (4hr) (see Figure A1. 14). Moreover, to assess further, we ran our model with three different cell lines which have either different NER machinery or damage source (Table 4.1). We have explored histone modifications possible functions on the NER efficiency and found that some histone markers are potentially crucial for the damage repair. In the following paragraphs, novelty and limitations of the research, project findings, pros-cons of machine learning (ML), and conclusion will be spoken, sequentially.

Although computational approaches have been used in our field since the beginning of the advancement of NGS technologies, to the best of our knowledge, no ML model was cooperated with the NGS data for exploring the interactions between histone markers and the NER. Another novelty of this research is showing G-NER's efficiency alone on damage repair using a ML algorithm. We have succeeded to acquire high NER prediction accuracies using histone modifications but ML model development, understanding its findings, data quality, and experimental errors were our project limitations. The data were retrieved from publicly available databases and hence, the variation in the data quality, experimental procedure, experimental errors, and data processing requirements were the factors affecting our model reliability. These obstacles were eliminated as much as possible during this project by data preprocessing, filtering, and normalization steps but data related impacts on our results could not be completely avoided. Moreover, some histone markers can be seen more than one time in the analyses for the same cell line (GM12878 and HeLa). Their divergent effect on the analysis can be seen especially at Figure 4.6, 4.8 and 4.10. These modifications (e.g. H3K27me3, H3K36me3, or H3K4me3) were from the different ChIP-seq data which had been uploaded to the databases by various research labs. These ChIP-seq data were acquired either from different research labs or same research labs' different experimental runs. Therefore, the main reason of inconsistency in the findings between the same histone markers is the data dissimilarities and the errors during data generation. Similarly, in some cases, the two biological replicates of the same damage repair exhibited different prediction accuracies. They did not exhibit different accuracy results in Figure 4.3 and 4.5 but in Figure 4.7, the differences between the replicates increased in the repair prediction at U_TES and D_TSS genomic segments. This unstable repair accuracies can be both related with the increased specificity in the analysis and data differences between the replicates of same damage and repair data. As data were utilized at more specific analysis conditions, the chances of having analysis errors and seeing the impact of data quality on the model increase.

Three different scatter plot analysis were made by using the preprocessed raw data to demonstrate the correlations between the histone modifications-G-NER machinery, DNase I hypersensitivity sites-G-NER machinery, and DNase I hypersensitivity sites-histone modifications in HeLa cells (see Appendix A). The ML model repair prediction analysis and DNA damage repair-histone marker scatter plots revealed a positive tendency in the CPD damage repair. This high CPD repair accuracies and more positive correlations of histone markers with the CPD damage repair is a known phenomenon. Although the exact reasons were not unveiled yet, the effective repair rate of [6-4]PP and its particular localization at nucleosomal regions were

most probably the causes (34). The swift repair of [6-4]PP by both NER pathways cause less reads to be acquired during NGS because some portion of the [6-4]PP damage have already been repaired until the DNA sequencing.

In addition, all histone modifications, which are anticipated to be positively correlated, had linear positive sample distributions with the UV damage repair correlations except H3K36me3 (see Figure A1. 45 - 46, 60 - 61, 75 - 76, 90 - 91). H3K36me3 showed less positive correlation with the G-NER. H3K27me3 were also exhibited low positive correlation but it was expected to do so (59, 63). Considering the cell type specific differences and condition of the genome after DNA damage treatment could lead H3K27me3 and H3K36me3 to undergo genome-wide re-distribution (62, 63). Furthermore, the repair predictions and scatter plot analysis of H3K36me3s and H3K27me3_2 with chromatins were inconsistent. H3K36me3 contributed repair with low accuracy at both chromatin states whereas H3K27me3 contributed repair more at open regions with low accuracy. The AUC plots do not reveal the positive or negative contribution strength of model features in model prediction. Similarly, the effect of the presence or absence of a feature in the model prediction cannot be known in these graphs too. Therefore, it is not possible to interpret information about H3K36me3 and H3K27me3's role in the model prediction. Maybe the absence of H3K27me3 might be the reason for H3K27me3 to have more prediction accuracy on open chromatins. Similar explanation can be made about H3K36me3 as well. The presence of H3K36me3 at open chromatins may be less important for model but the absence of it (as a lone feature) on closed chromatins may contribute to model much more. On the other hand, scatter plots only depicted the RPKM values distribution trends on the plots but since ChIP-seq and DNase-seq data were carried out in healthy cells as opposed to XR-seq and damage-seq, the sequencing conditions of damage repair and epigenetics factors were dissimilar. This is because these results do not show the direct impact and real environmental instant of histone markers and chromatins on damage repair process. That is the reasons why inconsistent results were observed between the scatter plots and model repair predictions. Beside scatter plot analysis, the genome-wide model NER prediction analysis also revealed evidently important histone markers which we considered as potential epigenetic modifications for the NER machinery (see Results). H3K9me3, H3K4me3, H3K27ac, H3K36me3, H3K27me3, and the H3K79me2 were featured at almost all analysis. Except H3K9me3 and H3K27me3, other four are known with their enrichments at the active promoters, enhancers, or gene bodies. The H3K9me3 and H3K27me3 are signs for inactive genes and heterochromatin regions (59). Their positive or negative effects on the model repair predictions attracted our attention and their effects on repair were in line with their role in the other genomic events

as well. Thus, we believe that they are considerable epigenetic modifications.

The previously mentioned machine learning related limitations and the data quality divergences led the detection of some unexpected findings. Considering the previously published literature information of histone modifications, some of their behaviors in the model predictions were not anticipated. They exhibited damage, cell type, and chromatin state dependent effects on the repair predictions. Thus, it is revealed that the role of chromatins is not only on NER but also on histone markers as well. Also, some modifications enriched almost at all four genomic segments (U_TSS, D_TSS, U_TES, and intergenes) as opposed to their known genomic enrichment patterns. H3K4me3_1 and H3K9ac were enriched at all three gene segments and intergenes in HeLa cell (Fig. 4.8). However, they are known to be localized at the transcriptionally active promoters (59). We assumed that transcription, gene expression and DNA repair should have similar necessities to occur such as chromatin accessibility and protein availability (64). Therefore, we expected H3K9ac and H3K4me3 to enrich more at U_TSS and decreasing its enrichment gradually as moving through other genomic segments. Similarly, the unexpected chromatin state dependent distributions of H3K36me3 and H4K20me1, were observed in scatter plot correlation analysis (see Appendix). H3K36me3 was mostly enriched at D_TSS open chromatins and lost its enrichments at other genomic segments in HeLa (Fig. 4.8). Like H3K36me3, H4K20me1 contributed to model mostly at D_TSS and lost its enrichment at other genomic segments as well (Fig. 4.6 and 4.8). The correlation plots between H3K36me3, H4K20me1 and DNase hypersensitivity exhibited less positively increasing linear trends compared to H3K79me2 and H3K4me3 which had similar functions with H3K36me3 and H34K20me1 (see Figure A1. 22 - 24, 30 - 32). These noncoherent prediction and correlation analysis differences between the two histone markers and DNase hypersensitivity sites may be because of the ChIP-seq data coverage differences, less data processing on the raw data before the model training or UV damage related histone marker redistributions. Most of the histone markers were utilized on the model prediction collectively and individually. Interestingly, H3K9me3 and H4K20me1 contributed to model prediction in different ways. The participation of H4K20me1 to model as a lone feature increased its feature importance on the repair prediction but its incorporation to prediction with the other histone markers decreased its effect. The complete contrary role of H3K9me3 on the model were observed in results (Fig. 4.4, 4.6 and 4.8). Therefore, this may indicate that their possible role in the damage repair process depends on the absence or presence of other histone markers in the same genomic environment.

Each AUC plot shows recall and accuracy values of NER damage predictions in three cell lines (Fig. 4.3, 4.5, and 4.7) and their corresponding bar graphs and horizontal

blue dots are the qualitative representation of the effects of each decision stumps and all histone markers for the model repair predictions (Fig. 4.4, 4.6, and 4.8). The success of having high prediction accuracies from our model increased reliability of the NER repair predictions and our claims. Also, our specially selected cell lines, analysis, and developed XGBoost model enabled us to reveal relatively high G-NER efficiency independent from the effect of TC-NER. The analysis exhibited not only the effect of chromatins on G-NER damage repair but also the effect of chromatins on the histone modifications (Fig. 4.5, 4.6, and 4.7). All histone modifications contributed to model prediction dependent from the chromatin accessibility but some of them were also revealed damage type and cell type dependent effects on NER efficiencies (Fig. 4.4, 4.6, and 4.10). H3K9me3, H3K9ac, and H3K4me3 were important factors for UV damage repair predictions in HeLa. On the other hand, H4K20me1, H3K27me3, H3K4me1, and H3K4me3 were important factors in NHF1. Thus, some histone markers contributed to model cell type dependently. Similarly, H3K9me3, H3K36me3, H3K79me2, H2AFZ, and H3K4me3 were crucial markers for predicting cisplatin damage repair (Fig. 4.10). Irrelevant from their localization at the four genomics segments, H3K4me1 and H3K27ac seemed to be common histone markers for CPD repair whereas H3K4me3 contributed [6-4]PP damage repair predictions. Moreover, some histone markers were nearly equally involved in both UV damage adducts repair predictions, and they are: H3K4me3, H3K4me1 and H3K27ac (Fig. 4.10). Therefore, there is also damage-type specific model contributions.

Damage types may have different repair rates due to their dissimilar chemical compositions and genome-wide distributions hence, NER machineries cannot repair these damages with equal effort and time such as repair of CPD and [6-4]PP. Therefore, having damage type specific model contributions of histone markers may mediate NER mechanism in reality to recognize lesions and help the repair process. H3K9me3, H3K4me3 and H3K27ac were frequently seen in many model predictions analyses so they are probably remarkable factors for the NER process. H3K4me3, which affected the NER repair irrespective from cell type, was the only marker contributed to model in every condition (Fig. 4.10). Even though, in this research, most histone marker enrichment patterns on the gene regions were mostly in line with literature, there were also dissimilar enrichment patterns of some histone markers too (Fig. 4.8). The all NER associated damage and cell line specific histone markers were positioned at active gene regions except H3K9me3 and H3K27me3. H3K9me3 and H3K27me3 are repressors for the euchromatin structure (59, 62). Conversely, H3K36me3, and H3K79me2 enrich at gene body (59, 62). Unlike the expected enrichments of H3K36me3 and H3K79me2, they were found mainly at active enhancer and promoter regions of the genes (Fig. 4.8). As stated above, the experimental

result and literature differentiation can be associated with the consequence of unable to including UV treated ChIP-seq data to the research due to insufficient data availability. Considering published gene enrichment patterns of histone markers, their known functions, and the enrichment of two histone markers at the gene regulatory and start sites, we are still considering as having a potential for the repair. Furthermore, H3K9me3's contributions to model G-NER's prediction at four gene segments was not predictive (Fig. 4.8) and its enrichment in given 5kb windows were, in fact, decreasing the chances of having repair at that genomic regions (Fig. 4.9). This is because H3K9me3 is a gene inactivation associated marker within the cell (59) hence, we expect it to function as a suppressor for the damage repair as well. The above listed histone markers' contributions, which are cell, damage type, and chromatin accessibility dependent, on the G-NER are also significant to demonstrate its efficiency irrespective from the TC-NER. Although these findings were acquired with relatively good prediction accuracies, more analysis needed for unveiling exact molecular interaction between every histone modification and NER pathways. However, it is also important to bear in mind that, there is no literature information that rejects the findings of our model (45, 59, 62, 63, 65).

The time course analysis on the model was performed to assess the performance and see the versatility of our model (see Figure A1. 14). Not surprisingly model gave best outputs (AUC = 0.91) for [6-4]PP repair predictions when it was trained with early time point and tested for early time (1hr) points again. This was a consistent output considering the machine learning algorithm's operation principle. Most interestingly, model which was trained with early time point but tested at late time point (4hr), exhibited relatively good performance (AUC = 0.78). These AUC scores exhibit that even in least favorable conditions model can evaluate true and false positives with relatively high accuracy.

We hope that our model can be one of the pioneers in this field to show researchers the benefits and importance of artificial intelligence applications for detecting and analyzing the cellular events . Although we are proud of our model and findings, utilizing ML algorithms have their own difficulties and benefits. One of the major toughness of using ML is it requires expertise of true algorithm selection for the analysis if an algorithm is not written from scratch for that specific research purpose. Understanding and interpreting the model outputs require a bit computer and algorithmic knowledge. To develop a model, sufficient input (data) should be provided for avoiding model from over- or under-fitting problems. Also, ML is susceptible to error and trustworthy results require assessment of model with different combinations of inputs and statistical analysis of model inputs. On the other hand, the major strengths of using ML are less time and effort needed to find interesting

patterns and results. Secondly, making mistake does not cost as much as errors done in the wet-lab and they are easily reversible. In addition, using machine learning does not require users to be in one place physically and it is accessible for running and editing at anytime and anywhere.

In conclusion, the good accuracies of our model during the tests allowed us to exhibit another effector of NER pathways, histone markers, and the efficiency of G-NER in bulky adducts repair. Thus, it is likely that the epigenetic determinants of G-NER bias are also key for the mutagenesis in cancer. We also achieved to demonstrate similar relationship, which had been observed between chromatins and NER previously, between histone markers and NER (47, 55, 64, 66, 67, 68). Beside their relationship with repair, H3K9me3, H3K4me3, H3K27ac, H3K36me3, H3K27me3, and H3K79me2 still need further detailed investigation for exploring their exact role in the NER repair (40). Moreover, a research designed to discover the relationship between chromatins, NER, and histone markers would provide better scientific proof to confirm their crosstalk as well even though chromatins and histone modifications were previously shown to be DNA repair related factors (69).

BIBLIOGRAPHY

1. De Bont, R., and van Larebeke, N. (2004). Endogenous DNA damage in humans: a review of quantitative data. Mutagenesis 19, 169–185. doi: 10.1093/mutage/geh025.

2. Ciccia A, Elledge SJ. (2010). The DNA Damage Response: Making it safe to play with knives. Molecular cell. 40:179–204.

3. Chen Y., Zhu WG. (2016). Biological function and regulation of histone and non-histone lysine methylation in response to DNA damage. Acta Biochimica et Biophysica Sinica. 48:603–16.

4. Ng SS., Yue WW., Oppermann U. et al. (2008). Dynamic protein methylation in chromatin biology. Cellular & Molecular Life Sciences Cmls. 66:407–22.

5. Freitag M. (2017). Histone Methylation by SET Domain Proteins in Fungi. Annual review of microbiology. 71:413–439.

6. Kinnaird A., Zhao S., Wellen KE. et al. (2016). Metabolic control of epigenetics in cancer. Nature Reviews Cancer. 16:694–707.

7. Morera L., Lübbert M., Jung M. (2016). Targeting histone methyltransferases and demethylases in clinical trials for cancer therapy. Clinical Epigenetics. 8:1–16.

8. Hyun K., Jeon J., Park K., et al. (2017). Writing, erasing and reading histone lysine methylations. Experimental & Molecular Medicine. 49:e324.

9. Lindahl, T., and Barnes, D. E. (2000). Repair of endogenous DNA damage. Cold Spring Harb. Symp. Quant. Biol. 65, 127–133. doi: 10.1101/sqb.2000.65.127.

10. Dexheimer, T. S. (2013). DNA repair pathways and mechanisms. In L. A. Matthews, S. M. Cabarcas, and E. Hurt (Eds.), DNA repair of cancer stem cells.

11. Curtin, N. J. (2012). DNA repair dysregulation from cancer driver to therapeutic target. Nat. Rev. Cancer 12, 801–817. doi: 10.1038/nrc3399.

12. Cooper, G. M. (2000). DNA repair. In The cell: A molecular approach (2nd ed.). Sunderland, MA: Sinauer Associates.

13. Kimball, J. W. (2015, October 31). DNA repair. In Kimball's biology pages.

14. Wei, S., Li, C., Yin, Z., Wen, J., Meng, H., Xue, L., & Wang, J. (2018). Histone methylation in DNA repair and clinical practice: new findings during the past 5-years. Journal of Cancer, 9(12), 2072–2081. doi: 10.7150/jca.23427.

15. Goosen N., Moolenaar GF. (2008). Repair of UV damage in bacteria. DNA Repair (Amst) 7:353–379.

16. Hoeijmakers JH. (2001) Genome maintenance mechanisms for preventing cancer. 411(6835):366-374. doi:10.1038/35077232.

17. Hoeijmakers JH. (2009). DNA damage, aging, and cancer. New England J Med. 361(15):1475-85.

18. Yimit A, Adebali O, Sancar A, Jiang Y. 2019. Differential damage and repair of DNA-adducts induced by anti-cancer drug cisplatin across mouse organs. Nat Commun 10: 309.

19. Reardon JT., Sancar A. (2005). Nucleotide excision repair. Prog Nucleic Acid Res Mol Biol 79:183–235.

20. Naegeli H., Sugasawa K. (2011). The xeroderma pigmentosum pathway: decision tree analysis of DNA quality. DNA Repair (Amst) 10:673–683.

21. Li W., Adebali O., Yang Y., Selby CP., Sancar A. (2018). Single-nucleotide resolution dynamic repair maps of UV damage in Saccharomyces cerevisiae genome. Proc Natl Acad Sci U S A 115: 570 E3408-E3415.

22. Truglio JJ., Croteau DL., Van Houten B., Kisker C. (2006). Prokaryotic nucleotide excision repair: the UvrABC system. Chem Rev 106:233–252.

23. Guzder SN., Habraken Y., Sung P., Prakash L., Prakash S. (1995). Reconstitution of yeast nucleotide excision repair with purified Rad proteins, replication protein A, and transcription factor TFIIH. J Biol Chem 270:12973–12976.

24. Mu, D., Wakasugi, M., Hsu, D. S., and Sancar, A. (1997). Characterization of reaction intermediates of human excision repair nuclease. J. Biol. Chem. 272, 28971–28979.

25. Hu J., Selby CP., Adar S, Adebali O., Sancar A. (2017). Molecular mechanisms and genomic maps of DNA excision repair in Escherichia coli and humans. J Biol Chem 292:15588–15597.

26. Sancar A., Rupp WD. (1983). A novel repair enzyme: UVRABC excision nuclease of Escherichia coli cuts a DNA strand on both sides of the damaged region. Cell 33:249–260.

27. Huang JC., Svoboda DL., Reardon JT., Sancar A. (1992). Human nucleotide excision nuclease removes thymine dimers from DNA by incising the 22nd phosphodiester bond 5' and the 6th phosphodiester bond 3' to the photodimer. Proc Natl Acad. Sci USA 89:3664–3668.

28. Canturk F., et al. (2016). Nucleotide excision repair by dual incisions in plants. Proc Natl Acad. Sci USA 113:4706–4710.

29. Kusakabe M., Onishi Y., Tada H., Kurihara F., Kusao K., Furukawa M., Iwai S., Yokoi M., Sakai W., Sugasawa K. (2019). Mechanism and regulation of DNA damage recognition in nucleotide excision repair. Genes Environ 41: 2.

30. Mao, P., Smerdon, M.J., Roberts, S.A. and Wyrick, J.J. (2016). Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution.

Proc. Natl. Acad. Sci. U.S.A., 113, 9057–9062.

31. Jung-Hoon Yoon, Louise Prakash, Satya Prakash. (2009). Highly error-free role of DNA polymerase n in the replicative bypass of UV-induced pyrimidine dimers in mouse and human cells, Proceedings of the National Academy of Sciences (2009), 106 (43) 18219-18224; 10.1073/pnas.0910121106.

32. Giglia-Mari, G., Zotter, A., Vermeulen, W. (2011). DNA damage response. Cold Spring Harbor perspectives in biology, 3(1), a000745.

33. Hu J., Adar S., Selby CP., Lieb JD., Sancar A. (2015). Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. Genes Dev 29:948–960.

34. R.M. Costa, V. Chigancas, R.S. Galhardo, H. Carvalho, C.F. Menck, The eukaryotic nucleotide excision repair pathway, Biochimie 85 (2003) 1083–1099.

35. You, Y. H., Lee, D. H., Yoon, J. H., Nakajima, S., Yasui, A., & 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(48), 44688–44694. https://doi.org/10.1074/jbc.M107696200.

36. Lo, H. L., Nakajima, S., Ma, L., Walter, B., Yasui, A., Ethell, D. W., & 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. https://doi.org/10.1186/1471-2407-5-135.

37. Niggli, H. J., & Cerutti, P. A. (1982). Nucleosomal distribution of thymine photodimers following far- and near-ultraviolet irradiation. Biochemical and biophysical research communications, 105(3), 1215–1223. https://doi.org/10.1016/0006-291x(82)91098-1.

38. Mitchell, D. L., Nguyen, T. D., & Cleaver, J. E. (1990). Nonrandom induction of pyrimidine-pyrimidone (6-4) photoproducts in ultraviolet-irradiated human chromatin. The Journal of biological chemistry, 265(10), 5353–5356.

39. Gale, J. M., & Smerdon, M. J. (1990). UV induced (6-4) photoproducts are distributed differently than cyclobutane dimers in nucleosomes. Photochemistry and photobiology, 51(4), 411–417. https://doi.org/10.1111/j.1751-1097.1990.tb01732.x.

40. Mao, P., & Wyrick, J. J. (2016). Emerging roles for histone modifications in DNA excision repair. FEMS yeast research, 16(7), fow090. https://doi.org/10.1093/femsyr/fow090.

41. Yu, Y., Teng, Y., Liu, H., Reed, S. H., & 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(24), 8650–8655. https://doi.org/10.1073/pnas.0501458102.

42. Guo, R., Chen, J., Mitchell, D. L., & Johnson, D. G. (2011). Gcn5 and E2F1 stimulate nucleotide excision repair by promoting H3K9 acety-

lation at sites of damage. Nucleic acids research, 39(4), 1390–1397. https://doi.org/10.1093/nar/gkq983.

43. Bostelman, L. J., Keller, A. M., Albrecht, A. M., Arat, A., & Thompson, J. S. (2007). Methylation of histone H3 lysine-79 by Dot1p plays multiple roles in the response to UV damage in Saccharomyces cerevisiae. DNA repair, 6(3), 383–395. https://doi.org/10.1016/j.dnarep.2006.12.010.

44. Chaudhuri, S., Wyrick, J. J., & Smerdon, M. J. (2009). Histone H3 Lys79 methylation is required for efficient nucleotide excision repair in a silenced locus of Saccharomyces cerevisiae. Nucleic acids research, 37(5), 1690–1700. https://doi.org/10.1093/nar/gkp003.

45. Evertts, A. G., Manning, A. L., Wang, X., Dyson, N. J., Garcia, B. A., & Coller, H. A. (2013). H4K20 methylation regulates quiescence and chromatin compaction. Molecular biology of the cell, 24(19), 3025–3037. https://doi.org/10.1091/mbc.E12-07-0529.

46. Jager, M., Blokzijl, F., Kuijk, E., Bertl, J., Vougioukalaki, M., Janssen, R., Besselink, N., Boymans, S., de Ligt, J., Pedersen, J. S., Hoeijmakers, J., Pothof, J., van Boxtel, R., & Cuppen, E. (2019). Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer. Genome research, 29(7), 1067–1077. https://doi.org/10.1101/gr.246223.118.

47. Corina Gsell, Holger Richly, Frédéric Coin, Hanspeter Naegeli. (2020). A chromatin scaffold for DNA damage recognition: how histone methyltransferases prime nucleosomes for repair of ultraviolet light-induced lesions, Nucleic Acids Research, Volume 48, Issue 4, (2020). Pages 1652–1668.

48. Mao, P., Brown, A.J., Esaki, S., Lockwood, S., Poon, G.M.K., Smerdon, M.J., Roberts, S.A. and Wyrick, J.J. (2018). ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. Nat. Commun., 9, 2626.

49. Brown, A.J., Mao, P., Smerdon, M.J., Wyrick, J.J. and Roberts, S.A. (2018). Nucleosome positions establish an extended mutation signature in melanoma. PLoS Genet., 14, e1007823.

50. Adebali O., Chiou YY., Hu J., Sancar A., Selby CP. (2017). Genome-wide transcription-coupled repair in Escherichia coli is mediated by the Mfd translocase. Proc Natl Acad Sci U S A 114: E2116-E2125.

51. Deger N., Yang Y., Lindsey-Boltz LA., Sancar A., Selby CP. (2019). Drosophila, which lacks canonical 507 transcription-coupled repair proteins, performs transcription-coupled repair. J Biol Chem 294: 18092-18098.

52. Yang Y., Hu J., Selby CP., Li W., Yimit A., Jiang Y., Sancar A. (2019). Single-nucleotide resolution analysis of nucleotide excision repair of ribosomal DNA in humans and mice. J Biol Chem 294: 210- 618 217.

53. Akkose U., Kaya O., Lindsey-Boltz L., Karagoz Z., Brown AD., Larsen PA., Yoder AD., Sancar A., Adebali O. (2020). Comparative analyses of two pri-

mate species diverged by more than 60 million years show different rates but similar distribution of genome-wide UV repair events. bioRxiv.

54. Oztas O., Selby CP., Sancar A., Adebali O. (2018). Genome-wide excision repair in Arabidopsis is coupled to transcription and reflects circadian gene expression patterns. Nat Commun 9: 1503.

55. Adar S., Hu J., Lieb JD., Sancar A. (2016). Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. Proc Natl Acad Sci USA 113:E2124–E2133.

56. Hu J., Adebali O., Adar S., Sancar A. (2017). Dynamic maps of UV damage formation and repair for the human genome. Proc Natl Acad Sci USA 114:6758–6763.

57. Hu, J., Lieb, J. D., Sancar, A., & Adar, S. (2016). Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. Proceedings of the National Academy of Sciences of the United States of America, 113(41), 11507–11512. https://doi.org/10.1073/pnas.1614430113.

58. Li W, et al. (2017). Human genome-wide repair map of DNA damage caused by the cigarette smoke carcinogen benzo[a]pyrene. Proc Natl Acad Sci USA 114:6752–6757.

59. Burçak Otlu, Can Firtina, Sündüz Keleş, Oznur Tastan. (2017). GLANET: genomic loci annotation and enrichment tool, Bioinformatics, Volume 33, Issue 18, (2017), Pages 2818–2828.

60. Berg, J. M., Tymoczko, J. L., and Stryer, L. (2002). DNA polymerases require a template and primer. In Biochemistry (5th ed., section 27.2.4). New York, NY: W. H. Freeman, 2002.

61. Goodsell, D. (2007). Thymine dimers. In RCSB PDB molecule of the month.

62. Huang, C., & Zhu, B. (2018). Roles of H3K36-specific histone methyltransferases in transcription: antagonizing silencing and safeguarding transcription fidelity. Biophysics reports, 4(4), 170–177. https://doi.org/10.1007/s41048-018-0063-1.

63. Wiles, E. T., & Selker, E. U. (2017). H3K27 methylation: a promiscuous repressive chromatin mark. Current opinion in genetics & development, 43, 31–37. https://doi.org/10.1016/j.gde.2016.11.001.

64. Stadler J., Richly H. (2017). Regulation of DNA Repair Mechanisms: How the Chromatin Environment Regulates the DNA Damage Response. Int J Mol Sci. 2017;18(8):1715. Published 2017 Aug 5. doi:10.3390/ijms18081715.

65. Jørgensen S., Schotta G., Sørensen CS. (2013). Histone H4 lysine 20 methylation: key player in epigenetic regulation of genomic integrity. Nucleic Acids Res.;41(5):2797-2806. doi:10.1093/nar/gkt012.

66. Mao P., Wyrick JJ., Roberts SA., Smerdon MJ. (2017). UV-induced DNA damage and mutagenesis in chromatin. Photochem Photobiol 93:216–228.

67. Gong F., Kwon Y., Smerdon MJ. (2005). Nucleotide excision repair in chromatin and the right of entry. DNA Repair (Amst) 4:884–896.

68. Smerdon, M.J. and Conconi, A. (1998). Modulation of DNA damage and DNA repair in chromatin. Prog. Nucleic Acid Res. Mol. Biol., 62, 227–255.

69. Spivak G. (2015). Nucleotide excision repair in humans. DNA repair, 36, 13–18. https://doi.org/10.1016/j.dnarep.2015.09.003.