Computational studies of DNA damage and repair of alkylating agents and UV light

Hanna Kranas

TESI DOCTORAL UPF / Year 2023

THESIS SUPERVISORS
Dra. Nuria López-Bigas & Dr. Abel González-Pérez

Department of Medicine and Life Sciences

To my grandpa,

I miss you and our science chats.
There is this science thing I did that I want to tell you about...

Acknowledgements

I firmly believe that we rise by lifting others. Finally, I am here! But how much you all had to rise, if you lifted me up so high?! And you are still so high up above me. Thank you for showing me the way. This thesis wouldn't be possible without a bunch of amazing people, that lifted, and lifted, and lifted me up. Lifted my spirits and mood, as well as my mind. I hope to one day be able to give it back, or at least pay it forward.

Firstly, I thank you. Yes, YOU, the reader! Thank you for opening my thesis and granting my words the honor of your attention, even if just for this sentence.

There is no thesis without a supervisor. And this one had two great, supportive ones.

Nuria, thank you for bringing this compassionate component into your leadership, and always putting the person first! I am grateful for all the kind opportunities for growth you granted me (and supported me through), both academic and science-adjacent. Thank you for building this amazing lab, and letting me spend 5 splendid years here, working with you.

Abel, I cannot imagine this PhD without you, your insights, and your calm, understanding presence. To say I enjoyed our discussions would be an understatement! Thank you for sharing your knowledge so freely, and letting me learn in so many aspects from you.

Although officially I had two supervisors, unofficially I feel like I had many, many more! I was mentored in some way by each member of the BBGLab, both past and present. Thank you all for the pleasure of getting to know you and all your brilliant points of view!

There are two BBGLab mentors that I want to call out here (positively of course!). Both are vital to the projects of this thesis and even more so to my life during the PhD. Erika, Ferran, now is your turn!

Every padawan needs a great master to show them the path of the force. Ferran!!! As we regularly have our lifting-spirits-up sessions, I'm sure you are aware of at least a small part of my gratitude for you. And I'm sure you are equally aware that a paragraph, or even a page in the thesis is not nearly enough to begin to express it. How about we leave it at that, and set up pronto another one of our amazing chats, so I can start there?

Oh, Dear! Dearest Awesome Erika. The Vitamin Person™, one and only, there is absolutely no one like you. Thank you for all the loving support you offer, and for letting me sometimes turn the tables and be there for you as well. You are an absolute blessing to this world and an inspiration for how to show up, notice, care for, and lift other people. And a freakishly smart scientist at that!

PhDs of the lab - we mentor each other, we support each other, we cry and laugh together. It is no easy journey, but you make the time pass sweeter.

First, current PhDs: Katyayani, you are my favorite, very kind, understanding, and crazy fun neighbor. I love our chats and walks, and I always feel they are too short! Stefano, your beautiful intense energy is one of a kind! Can't wait for another one of our challenging discussions. Give a hug to the awesome Martina, and I am looking forward to your newest project! Ferriol, I might call you a "PhD baby", but I appreciate your mature views on science and the joyful proactiveness you bring to the lab, really a lot. Raquel, you are like a sister to me; I can't decide if younger or older (I guess it depends on the time!). What I mean is that connecting with you has been one of the best surprises at the end of the PhD. Expect more to come in one of my podcast messages!

To my dear past PhD mates: thank you for showing me the ropes, both in the lab and in Barcelona; you made the uneasy thing of starting in a new place much more enjoyable. First, the jokester trio: Joan, Oriol, Pepe! I loved coming into the lab in the morning and seeing you three fooling around, you could brighten the harshest of mornings! Joan, thank you for sharing your project with me and offering me guidance. Oriol, you might have called me your grandma, but you are the one whose understanding advice I needed the most! Pepe, my PhD twin, you know I don't have enough words to thank you for sharing this journey with me (which you are doing to this day!). Ines, thanks for all the lunches, coffees, and music exchanges over the years; you really knew when I needed one of them and showed up. Claudia, one thing I am sad about regarding my PhD journey is that we couldn't see each other more in real life instead of online. Thank you for experiencing the little joys of life so beautifully and sharing it with others.

To the rest of my dear labmates, past and present: I love the lab, and the atmosphere we all created together. I believe there is not even a single person I dislike. I love how each of you is so unique, and shows up authentically, to build together such a great place. Even if your name is not listed here, it does not mean I care less about you or your contribution to my life over these last years. It just means that there are so many of you, and I don't think I have enough space to write here. But! You can expect me to either come tell you soon or send you a message, to thank you directly. Just you wait!

Now let's switch gears shortly a bit to the more official, but no less important side. Thank you to CONTRA for offering funding, great training, and chances to meet great people. Thanks to all ESRs for sharing in this journey, with special thanks to Monica, Paula, Michael, Nico, and Mandi. Thanks to the PIs and organizers involved, for all their effort. I also want to thank Sohrab Shah's lab for the great time I had when they hosted me there.

I want to take this opportunity to also thank the members of the thesis tribunal, and international mention reviewers, which are likely the only people to read most of this thesis in depth. Thank you for your time, and your reviews!

Huge thanks to the IRB, as an institution, for taking me in and taking care of my PhD. Huge thanks also to the people that make up IRB. From the administrative side, special thanks go to Leyre and Alba, thank you for the kindest support. Thanks to my TAC committee, for guiding this thesis over the years: Robert Castelo, Travis Stracker, Marc A. Martí-Renom. Thanks to the student councils for your tireless work in supporting students; and thanks to the council team I was a part of for moving the needle a bit. Thanks to everyone who built and supported the actions of the Mental Wellbeing team. Thanks to you I learned a lot and was given amazing chances to advocate (a tiny bit) for the improvement of mental health in academic institutions. Equally, thanks to everyone that kindly listened to me rant about this topic for hours!

Thanks also to everyday people I met at IRB and who so happily shared their experiences with me. Starting with a few great PIs (Direna, Alejo, Salva, Manuel, and many others) - although I was not your student, you cheered me on. Thank you for that.

Leaving IRB for a little moment, I want to thank 2 important groups for my formation as a scientist. First, the ISCB-SC RSG-Spain, I loved learning from all of you. Secondly, the Awesome Leaders. Each one of you deserves your own note; I hope I can give it to you the next time we meet. You made me better.

Thanks to the many people that I interacted with, with some for longer than others. Thanks to the IRB Rock Band! I miss you, and I will come to hug all of you soon, and hopefully, we can jam together again. This was such an unexpected thing, but unforgettable, and even better so - I am so glad I met you all; friendships forged over music are definitely ones of the best.

IRB wouldn't be the same without what I consider people of my PhD core team. Thank you for taking me as your friend, all your patience, fun trips, good rants, and overall just sticking together these 5 years. Adrí, I consider you and Marina M. to simply be my people, I love you both a lot. Now let's go have some cheese. Paula, your energy to fight for what is good always inspires me; thank you for standing up for me too. Clara, thank you for always showing up in your abundant, kind, and creative energy; it refueled my mood many times. Nico, you are just the absolute dearest, always. Pep, thank you for always welcoming me with open arms. Marina S., I really wish I could have seen more of you; whenever you visit it's an absolute joy, we miss you. Diego, thanks for always offering your thought-through perspective both on the PhD and life, it has helped me a lot. Liza, I'm so happy we both made it! Elena, we might not be in touch anymore, but I am forever grateful for the time we lived together. And for introducing me to the two special people.

Irene and Tere! You showed me that if there is a will, there is a way; no distance can keep us apart. Thank you both for your beautiful friendship. Tere, thank you for always pointing me in the right direction. Irene, thank you for your deeply compassionate view of me.

My dear friends from Poland. Maciszeiro, Monika, Magda, and Igor. It's been so many beautiful years already, and I look forward to many more: of fun, of support, and of love. Kacpi, thank you for the lovely visits, and a shared perspective of growing up together; and thank you for introducing me to Jaz, and her radiant kindness.

Ramoneczku, I'm not sure you know this, but this thesis is also yours. Thank you, from the bottom of my heart, for this shared journey of the last few years. It's simply not possible to put into words everything I want to say to you. But let me at least start with this: Thank you. I love how we supported each other through all the hard parts. I loved getting to know you in-depth, learning from you, and growing together. It was, just like you, absolutely glorious. Our paths might be diverging now, but the time they coincided will always hold a very special space in my heart. I always, always wish you the best! Moltes gracies i ciudate, cuqi.

To my closest - thank you for being there always, through good and the bad. Thank you for being my main supporters through all my struggles, cheerleaders through all the successes, lending the ear and advice whenever needed. If there is something else after this life, I want you to be my family there too. The most, I want to acknowledge three people:

Ika i Wtk, mom and dad. Thank you for all the opportunities and care you have given me, that lead me to where I am today. Thank you for being my parents, and choosing to also become my friends. Kocham Was bardzo. Mom, thank you for always being my sounding board. Tato, dziękuję za wszystkie prezenty i ciasta, wiem co chcesz mi nimi powiedzieć.

kreska_, my sister. One and only. The force of nature. Not sure you know this, but you have the longest and most successful streak of inspiring me. Thank you for sharing everything with me so openly, tirelessly showing new perspectives, and letting me be who I am. And thank you for being the number one cheerleader of my scientific career too. Si, a ja ziemniaki. Kocham Cię!

And last, this might sound cheesy, but I also want to thank myself. Those who know, know that this was not easy; not just because of the PhD but also because of life. Everyone goes through hard times, and I went through them too. Hania, thank you for persevering, surviving, and thriving.

Abstract

Living cells are exposed to naturally occurring DNA-damaging agents that promote the loss of genome integrity, threatening the proper functioning of the cell. To counteract the toxic accumulation of DNA damage, organisms have acquired mechanisms of DNA repair, comprising multiple pathways charged with correcting the different types of damage a genome can accumulate. The aim of this thesis is the study of DNA damage caused by alkylating agents and UV light and their repair. First, we describe the first nucleotide-resolution alkylating damage maps in humans. Second, we present a novel approach to partition the genome by UV DNA damage repair activity aimed at studying the determinants of the UV mutagenic process that is unbiased by genomic features. Both projects contribute to expanding our knowledge of how different parts of DNA damage response interact with chromatin architecture and basic cell processes, like DNA replication or transcription.

Resum

Les cèl·lules estan exposades a agents naturals que danyen l'ADN, posant en risc el seu funcionament. Per tal de contrarestar la toxicitat del dany genòmic, els éssers vius han adquirit mecanismes de reparació de l'ADN dedicats a corregir els diferents tipus de dany. L'objectiu d'aquesta tesi és estudiar l'acumulació i reparació del dany genòmic causat pels agents alquilants i la radiació ultraviolada. Primer, descrivim els primers mapes de dany genòmic induït per agents alquilants en humans. A continuació, presentem una nova aproximació per a dividir el genoma en base a la reparació del dany genòmic causat per la radiació ultraviolada amb l'objectiu d'estudiar, de manera no esbiaixada, els determinants mutacionals associats. Tots dos estudis contribueixen a ampliar el coneixement de com l'acumulació de dany genòmic i la seva reparació s'associen amb les característiques de la cromatina i altres processos cel·lulars com la replicació de l'ADN o la transcripció.

Table of Contents

1. Introduction

1.1. DNA alteration processes

The importance of nucleic acids for currently existing organisms and the cells that make them is absolutely clear. Deoxyribonucleic acid (DNA) not only works as a 'memory molecule', storing a cookbook with recipes for the development, functioning, and even death of the cell; it also provides heredity directions, crucial for the continuation and evolution of life. While originally thought otherwise, DNA is a 'volatile' and reactive molecule [1, 2]. This duality of the DNA molecule - its importance, and a need to keep it intact versus the ease with which it can undergo changes - is right at the center of an essential biological balance. Balance, which our cells constantly need to try their best to get right. The balance between some changes – mutations – appearing, necessary for variability and evolution in the Darwinian sense; and preservation and protection of crucial information, necessary for the proper functioning of the cells and organisms [3, 4]. The reactivity of the DNA itself is only one side of this problem. On the other side are the unforgiving environments that it is exposed to.

Both the environment outside the cell, as well as the one inside it, house many potentially disruptive forces to DNA [2, 1, 5]. We call these kinds of disruptions 'DNA damage' or 'DNA lesions'. The ones caused by the forces coming from outside the cell are termed 'exogenous', and from within it - 'endogenous' - damages. Apart from there being many sources, there are also many different types of DNA damage, characterized by differences in disruption to the molecules' structure. Given the prevalence of DNA damage, DNA's constancy as an inheritable instruction manual only makes sense in the light of the presence of extensive maintenance keeping it in check. This means organisms and cells needed to evolve specialized mechanisms in place to: protect the DNA from those destructive forces; repair the damage if it occurs; if non-repairable (or happening during an important task) pass through it without introducing grave errors; if needed, stop everything else, allowing more time for repair; if all else fails - destroy the cell.

This tiered, hierarchical, organized, and very specialized answer of the cell to threats to the integrity of the DNA is often called the DNA Damage Response (DDR) [5, 4].

DDR encompasses: pathways of DNA repair, each charged with the repair of specific lesions; the DNA damage tolerance mechanisms, allowing the cell to continue - not without risk - when encountering unrepaired lesions during e.g. replication; and cell-cycle checkpoints, allowing for arresting the cell in the cell cycle, buying more time to repair or tolerate the DNA damage. If DDR is not successful, the cell should enter one of the cell death pathways, to avoid potentially malfunctioning. The success of DDR does not mean that the DNA remains unchanged, though. The cost of maintaining the genome perfectly is extremely high - too high to be worth it [6]. Many of the steps of the process are imperfect and can introduce permanent DNA sequence errors, called 'mutations', that the cell can still function with, but with varying degrees of consequences. The accumulation of mutations can lead to malfunctioning of the cell, early senescence, and over-increased proliferation, and has been implicated as an important factor in aging and cancer, among others [2, 6].

In this chapter, I introduce the different steps of both DNA damage and the cell's response to it, describing the sources of DNA lesions, DNA damage types, and the different elements of the DDR response. Unless otherwise specified, the focus of the chapter is on human biology. I highlight damage resulting from 2 sources pivotal for understanding this thesis: UV light, the most frequent exogenous damaging agent; and alkylators (which come from a variety of both endo- and exogenous sources), specifically ones used as chemotherapies.

1.1.1. DNA Damage

In this section the reader will find examples of DNA-damaging agents and their sources, as well as selected lesion types that they can produce.

1.1.1.1. Sources

DNA damage may come from a variety of sources that are broadly categorized into the factors coming from processes happening naturally within the cell (endogenous) and those from the environment outside of the cell (exogenous). Some examples of sources of endogenous lesions include natural cell metabolism products, reactions with molecules like water or reactive oxygen species (ROS), and accidental erroneous work of some substrates [5, 2]. Environmental factors come from a variety of lifestyle components, as well as exposure to viruses, radiation, and occupational mutagens [5]. This is a simplified and imperfect characterization though: some lesions categorized as coming from endo-, can also sometimes be from exogenous sources; some sources might be so ever-present along the human population and hard to study, that it becomes very difficult to tell where exactly they originate. Additionally, natural cell chemistry

can change the exogenously induced lesion into another one (e.g. a methylated base being cut out, producing a base-less site in the backbone of the DNA) making this categorization even more complex [5].

1.1.1.1.1. Endogenous

Every day, our cells are suffering unavoidable DNA damage which is a byproduct of their metabolism. Types of lesions introduced in the bases as a consequence of cell metabolism include methylation (e.g. spontaneous actions of nonenzymatic methylators), oxidation (oxidizing substances interacting with DNA) and hydrolysis (due to ever-present water molecules) [5, 2]. Although one cannot completely get rid of these reactions, some lifestyle changes have been shown to give a positive impact [6].

Interactions of the DNA with water molecules, termed hydrolysis, cause spontaneous modifications of the bases. These include depurination, depyrimidination, cytosine deamination, and others [7]. Some of these modifications result in atypical bases, usually easily recognized and repaired. An important example is the deamination of 5-methyl-cytosine (5mC, a DNA modification naturally present at CpG sites of inactive genes), which produces thymine [7]. The result is a T-G mispairing. Its repair is fairly inefficient and results in 5mC>T mutations [7]. In the case of depurination and depyrimidination, the bond linking the DNA backbone and the base gets spontaneously hydrolyzed. This leads to a site without any base, an abasic site [5, 2].

When Reactive Oxygen Species (ROS), like superoxide radicals or hydrogen peroxide, come in contact with DNA, they can oxidize it [5, 2]. This can produce many different damaging outcomes. Examples include modified bases (e.g. 8-Oxoguanine, resulting in G>T and C>A mutations), a saturation of pyrimidine rings, lipid peroxidation, and even breaks in the DNA backbone [5]).

An endogenous source, less known and frequent, but relevant for this thesis, is endogenous alkylators [5]. Alkylators are a very broad category, including practically any source of additional alkyl groups on DNA. One of the alkyl groups is methylation. The methylation of DNA serves many important functions. Due to that, there are agents in the cell tasked with methylating (alkylating) DNA [8]. Their actions, when happening in the right places, do not constitute damage, but a controlled and needed alkylation. An example of a nonenzymatic methylator present in human cells is S-adenosylmethionine (SAM) [9, 8]. However, SAM is also prone to spontaneous methylation in some guanines and adenines. More on alkylators, especially exogenous ones, is explained in the dedicated exogenous section below.

Finally, it is important to note that processes like replication or repair, while tightly regulated and nearly perfect, are not completely error-free [3]. It is estimated that during replication once every 10^4–10^7 bases, a mistake is introduced [3, 4]. Repair, when encountering a mismatch, often does not know what was the original base - and corrects based on the most likely scenario, but not necessarily the actual one [7]. Some repair mechanisms sacrifice the quality of repair - and a few bases at that - to repair a very cytotoxic type of damage, or prevent cell death [4]. Finally, some existing lesions might cause important protein complexes bound to DNA to get irreversibly trapped (e.g. Topoisomerase I) at the site, causing them to become DNA damage too [4]. Although not categorizable as mistakes, some steps of DNA damage repair entail changing a lesion to another one, e.g. an improper base might get cleaved during repair, producing an abasic site that should be further repaired [10].

1.1.1.1.2. Exogenous

Although exogenous (coming from outside environments) suggests a degree of control over our exposure to these DNA-damaging sources, some of them are ubiquitous, making them hard to evade. Lifestyle has a big influence: commonly known problematic substances are tobacco and alcohol, but diet (e.g. fungi infesting peanuts, grains, and corn stored in humid conditions, which produces dangerously mutagenic aflatoxins) can have a huge impact too [5, 6]. Radiation is inescapable, whether from UV exposure on a sunny day, during medical procedures (e.g. X-rays), or due to frequent flights (space radiation) [11]. Some infections and viruses are implicated as a risk factor for cancer (e.g. Hepatitis B virus). One's job environment might include dangerous, mutagenic substances (e.g. asbestos, vinyl chloride, benzene) [12]. Or one's health might require the use of medicine that might damage the DNA of healthy cells in their body (e.g. chemotherapies) [13, 14].

Of note, some of the mentioned sources do not generate DNA lesions themselves, but rather through their metabolites, produced by the interaction with oxidases of cytochrome p450 [5]. Although cytochrome p450 is charged with deactivating compounds into more easily excretable and less harmful forms, for some molecules the effect is opposite, producing forms that are more active, mutagenic or cytotoxic. As the description of all the potential exogenous DNA damage sources is out of the scope of this thesis, next I focus on the two of the highest importance for understanding this work: the most mutagenic part of the sunlight spectrum - UV (ultraviolet) light, and a few selected alkylators.

UV radiation

The ever-present sunlight, together with its many benefits, brings also risks. An important one is the ultraviolet component of the sunlight spectrum (covering around 5.4% of the sunlight composition) [5]. The UV wavelength spectrum can be split into three components: UV-A, UV-B, and UV-C. Each of these components is characterized by a slightly different wavelength range. Differences in wavelengths result, among other things, in slight differences in produced lesions and their proportions. Importantly, the most damaging UV-C is luckily heavily filtered by the ozone layer [5].

UV-induced damage comes in two main forms: direct DNA damage due to the excitation of the molecule, and indirect damage, coming from photosensitizers (other, photo-excited molecules) transferring the energy to the DNA [15, 16, 17]. In this thesis, we focus on the direct UV damage, in the form of pyrimidine dimer photoproducts. As UV-A – which constitutes the major UV contributor to the sunlight spectrum – does not induce as many pyrimidine dimers, we next focus on UV-B and UV-C.

Pyrimidine dimer photoproducts formed by UV-B light on DNA are bulky, helix-distorting lesions [18, 19, 5]. UV-C produces highly similar lesions as UV-B, in a shorter period of time, although in slightly different proportions [5]. Due to that, it is frequently used in experimental studies of mutagenesis by UV and sunlight exposure.

Due to the nature of the damage induced by UV light, most photoproducts are induced only while the light source is currently present. Without light, direct damage does not happen. As mentioned before, indirect damages, indicated as consequences of mostly UV-A exposure, have been reported [15, 16, 17]. For example, photo-excited melanin is indicated in indirectly generating ‘dark’ pyrimidine dimers. UV is also known to provoke oxidative stress.

Luckily, although exposure to UV radiation cannot be completely omitted (at least without serious risks to health), its damaging effects can be reduced by the proper use of sunscreen, and regulating the time spent in the heavy summer sun.

Alkylating agents

Alkylators, or alkylating agents, is a broad name for any substance that adds alkyl (methyl, ethyl, butyl, propyl...) groups to DNA. Alkylators are usually categorized as mono- or bi-functional, depending on whether they carry 1 or 2 reactive groups, allowing them to interact with 1 or 2 sites in the DNA, respectively [20].

Monofunctional alkylators mostly damage DNA by adding alkyl groups, with varying affinity, to nitrogens and some oxygens of the base rings. Bifunctional ones can additionally link two bases together between (crosslinks) or within the strands, or link a DNA base with a protein [5].

Natural, endogenous alkylators are, for example, the already mentioned DNA-methylating molecules like SAM. Exogenous alkylating agents come from many different sources, and notable examples include metabolites of benzopyrene present in tobacco smoke or biomass burning; historically known toxic agents like mustard gas; chemotherapeutics [5].

For this thesis, monofunctional so-called model alkylators and alkylators used in chemotherapy are of special interest and will be mostly described next. Model alkylators are alkylating agents that have been extensively used in experimental studies due to their potency. A prominent example is MMS (methyl methanesulfonate), which methylates bases. Among monofunctional alkylators commonly used in chemotherapy, the most known one is likely Temozolomide [21, 20].

One might wonder why DNA-damaging agents are used in the therapy of cancer if they inflict damage on healthy cells as well. The reasoning behind this comes from the high genetic instability that cancer cells exhibit. Many cancers have some repair pathways impaired, while others have messed up their cell cycle checkpoints. Even if they don't carry defects in DNA repair pathways, the fact that they go through the cell cycle faster than normal cells is detrimental to their ability to repair DNA. The rationale is that this may result in apoptosis. Normal, healthy stem cells are thought to be able to remain quiescent, with enhanced DNA repair capability. But unfortunately, high damage, even efficiently repaired, caused by prolonged exposure can leave scars. Chemotherapies are infamous for their strong side effects, caused by the depletion of stem cell populations [13]. There are also (very rare) reports of secondary malignancies, especially of blood cells, with causation pointing to mutations induced by chemotherapeutic treatment used many years before for a different tumor [14].

1.1.1.2. Types of damage

As indicated above, apart from many sources, DNA damage also comes in various structural shapes. In this section, I list a few categories of distortions to the structure of DNA, focusing on the examples most relevant to this thesis.

Types of distortions induced by the damage to the DNA range from small modifications of bases, through larger helix disturbances, to toxic breaks. Small modifications

are usually additions or losses of tiny groups on the bases. Sometimes, the loss encompasses the whole base, without affecting the backbone - this is termed an abasic site. Lesions with a stronger impact on the helix structure include large, bulky adducts and crosslinks between or within the strands. Finally, the DNA backbone can experience breakage, which can affect one, or worse, both of the strands.

1.1.1.2.1. Bulky adducts

Bulky lesions include any larger distortions of the helix, bulging out from the DNA. These are often adducts formed by an external molecule binding to the DNA. These are large and often formed by metabolites of e.g. aflatoxin or benzopyrene [22]. Some smaller bulky adducts are formed by a reaction caused by an outside agent on the DNA. In this group belong the UV-induced pyrimidine dimer photoproducts.

Pyrimidine dimers, as their name suggests, are formed by two adjacent pyrimidines (Cs and Ts) getting linked together. The two most frequent UV dimers of this type are cyclobutane pyrimidine dimers (CPDs) and pyrimidine(6-4)pyrimidone photoproducts (6-4PPs) [19]. CPDs are about more abundant than 6-4PPs, although this proportion slightly changes depending on the wavelength [5]. These two photoproduct types are distinct in the place where the dimer bond is formed. CPDs are characterized by the cyclobutane ring linking the two pyrimidines. In 6-4PPs, the two pyrimidines are bonded through the C6 position with the C4 of the other.

The 6-4PP structure is more disruptive to the helix. Luckily, the repair mechanism charged with its repair has a higher affinity for 6-4PPs than other pyrimidine dimers, and it is repaired faster [23]. All bulky lesions are predominantly repaired with Nucleotide Excision Repair.

1.1.1.2.2. Alkylations

As introduced before, alkylations include a large range of modifications including any addition of alkyl groups to the bases. The most frequently added groups are methyl and ethyl. Alkylating damage is thus in general of a small size – depending on the size of the alkyl group – compared to bulky lesions.

Alkyl groups can be added at various bases, and happen mostly at adenine and guanine [9]. The groups are also attached to various atoms of the base rings. Most frequently at nitrogens, producing e.g. N-methylpurines (NMPs) and oxygens, resulting in e.g. O6-methylguanine (m6G). The propensity to generate more specific alkyl group additions in some atoms or bases than others is dependent on the drug. Nevertheless, three frequently encountered NMPs are 3-methyladenine (m3A), 7-methylguanine

(m7G), and 1-methyladenine (1mA) [9]. In fact, m7G and m3A have been found to constitute up to 90% of the MMS-induced lesions, with m7G appearing 8-fold more than m3A [24]. These two methylations are also frequently induced by Temozolomide, although in slightly different proportions: m7G constitutes 70% and m3A only 9% of all lesions [21].

Alkylating lesions are repaired by a host of repair mechanisms. m1A and m6G are predominantly repaired through dedicated direct reversal mechanisms. Base Excision Repair is tasked with the removal of m3A and m7G. Nucleotide Excision Repair has been potentially implicated for some alkylating lesions as well [9, 8].

1.1.1.2.3. Crosslinks, abasic sites, and strand breaks

Other types of distortions to DNA that are not at the center of this thesis, but important to develop a bit more are DNA crosslinks, abasic sites, and strand breaks.

Some alkylators (and many other damaging agents, termed crosslinking agents) carry the ability to cross-link the two DNA strands [20, 25]. These types of lesions are broadly termed inter-strand crosslinks (ICLs). With the involvement of both strands, their repair is not trivial and involves many DNA repair pathways [5, 25].

Abasic sites (also called AP sites, where AP stands for apurinic/apyrimidinic) are single-nucleotide gaps in the backbone, with only the base gone. As mentioned earlier, they can be generated due to hydrolysis of the bond between the backbone and the lost base. Alternatively, they are a natural product of Base Excision Repair, where a modified base needs to be excised from the backbone. This means that they are also repaired through the next steps of the same pathway.

The backbone of the DNA can also break. Single Strand Breaks (SSBs) involve just one strand and are quite common. They are also necessarily produced during many cell processes. In fact, after the excision of the modified base from the backbone that induces an AP site, the next step of Base Excision Repair involves inducing a SSB. SSBs are usually quite easily and efficiently re-ligated. Double Strand Breaks (DSBs) on the other hand are much more problematic for the cells. Repair pathways do not have a way of knowing which pairs of ends fit together, although some of them attempt to utilize homology at the break sites [4]. Nevertheless, DSB repair often comes at either a loss or relocation of the DNA fragment.

1.1.2. DNA Damage Response (DDR)

Each day, mutagenic processes – both endogenous and exogenous – leave tens of thousands of different lesions in the DNA of each of our cells [2]. This amount of damage highlights how crucial – and incredibly efficient – the response of the cell to it is. Without repair and other damage tolerance mechanisms, DNA and the precious information encoded within it would decay at such a rate that it would be rendered useless [26]. Curiously, the DNA molecule seems almost built, primed with repair in mind. All spontaneous deamination events of the 4 main bases happen to yield very atypical products that can be easily recognized and repaired [7]. (An important exception is 5mC, which when deaminated yields thymine). The double helix structure carries twice the same information, automatically suggesting the basis for most of the DNA repair mechanisms: using the untouched strand as a template for the repair of the damaged one .

Cells can recruit different elements of the DNA damage response (DDR) upon need. Dedicated signaling pathways are activated upon detection of high DNA damage, upregulating the repair proteins. Some repair mechanisms operate constitutively, while others are triggered at certain events (e.g. upon failure of a cellular process or at least an interference) [26]. There are repair mechanisms linked to transcription, that arrive quickly after transcriptional machinery clashes with a lesion [27]. When repair is not able to handle the damage before replication, tolerance mechanisms like translesion strand synthesis (TLS) are employed to continue this process [28]. Finally, cells have dedicated checkpoints throughout the cell cycle, stopping it if DNA damage is found, to allow the other players of the DDR time to cope with the damage [4]. This section concisely outlines the components of the DDR with a special detail for the specific elements related to the handling of UV- and alkylation-induced lesions.

1.1.2.1. DNA Repair

With the variety of structural changes that DNA lesions produce, a one-size-fits-all solution does not seem appropriate. The cells have evolved multiple specialized repair mechanisms suited best for different DNA damage types.

1.1.2.1.1. Direct reversal of damage

Direct reversal of UV damage with photolyases is probably the first discovered and described repair mechanism [22]. As the name suggests, in this approach the lesion on the DNA is directly reversed by a dedicated enzyme. This repair approach is useful when highly mutagenic or cytotoxic lesions are present, and a need for a very rapid response arises. Some of the enzymes able to perform this type of repair can survive,

but some of those have been named ‘suicide enzymes’ as they get destroyed in the process [4]. In the second case, one can imagine that this is quite costly - one molecule gets sacrificed to repair one damaged site only.

The 2 very important mechanisms of direct damage reversal in humans are MGMT and ALKBH2-3 [8]. MGMT is tasked with the transfer of the methyl group from the m6G lesion to itself, destroying itself in the process of repairing the base [4, 26]. ALKBH components are involved in the repair of m1A lesions [9]).

1.1.2.1.2. Base Excision Repair

There are two incision-based repair mechanisms: Base Excision Repair (BER) and Nucleotide Excision Repair (NER, described in the next section). The main mechanical difference between the two comes from how the damage is excised. After the excision, the process looks similar: DNA polymerase fills the gap, using the other strand as a blueprint, and the DNA ligase brings together the free ends [10].

The main difference in the excision of the damage between the two Excision Repair mechanisms is driven by the structural differences of the lesions that those repair mechanisms dedicate themselves to (although there might be a degree of overlap). BER focuses on small lesions. It repairs bases with a small change in them (alkylation, deamination, oxidation) or AP sites [10].

The BER pathway starts with a group of glycosylases, each dedicated to the recognition of a slightly different group of (slightly structurally similar) modified bases. These glycosylases scan the DNA, flipping the bases out of the chain one by one, checking each for modifications. Once they encounter a modified base, they remove it from the backbone, leaving an AP site. AP sites are next recognized by the AP endonucleases, which nick the DNA backbone at the site. The nick is filled in by DNA polymerase and ligated to the other free end by DNA ligase, bringing the DNA at the site back to its original structure [10]. (Of note, this is a huge simplification for the needs of this thesis. Note that the details of processing the abasic site can differ depending on whether the dominant short-patch (generating a single nucleotide gap) BER pathway is involved, or the long-patch (2-10 nucleotides gap) BER [10].)

The glycosylase responsible for detecting m3A and m7G is the mammalian methyl purine DNA glycosylase (MPG, further called AAG) [10]. For purposes of this thesis, we assume most abasic sites are processed by the APE1 AP endonuclease.

1.1.2.1.3. Nucleotide Excision Repair

NER is a dedicated repair mechanism for any large distortions of the DNA double helix, so-called bulky lesions introduced above [29]. To achieve that, NER machinery does not just cut out the affected bases, as the bulky lesion is rather too large to handle, but a larger oligonucleotide fragment that surrounds it. This makes it very versatile, permitting the removal of bulky adducts of very different types. While mostly handling the ever-present UV-induced adducts, it can also handle those produced by the metabolism of aflatoxin, benzopyrene, and platins, among many others [22].

Rather than a single enzyme, like in the case of BER, NER consists of a large multi-protein complex that checks the DNA for the large, helix-disrupting lesions [29]. The current paradigm is that NER detects lesions in two ways: probing the DNA or being actively recruited when needed [29]. When the damage is found, the complex cuts the phosphodiester backbone at the same strand, and both sides of the lesion, initiating the excision of a roughly 30 nucleotides-long (in humans) fragment [27, 23]. DNA helicase helps uncouple the oligonucleotide from the DNA so that the DNA polymerase and ligase can fulfill their roles in filling and closing the gap [27].

Importantly, apart from the global NER mechanism based on probing the DNA for damage at any site, NER has another sub-pathway, which recognizes the damage differently. The Transcription-Coupled NER (TC-NER) allows additional recruitment of this repair pathway when damage stalls RNA polymerases (RNA Pol II) during transcription [29, 27, 23]. Through this coupling, RNA polymerase is backed up, the repair complex is recruited to repair the lesion on the template strand, and transcription can be restarted right after [29].

In global NER, the damage is recognized primarily by the XPC (Xeroderma pigmentosum, complementation group C) protein [23, 18]. XPC promotes the assembly of the complex responsible for the dual incision, although it does not form part of it. In the case of TC-NER, the stalled RNA Polymerase II together with the CSB (Cockayne syndrome type B) translocase actively recruits and assembles all of the incision core complex factors [23, 18]. The final complex is identical for both pathways.

The XPC and CSB damage recognition proteins are frequently implicated as crucial for two diseases (that their names are derived from: Xeroderma Pigmentosum and Cockayne Syndrome). Mutations in XPC inactivate the global NER pathway, leaving the cells proficient only in the transcription-coupled repair. In the opposite manner, CSB inactivating mutations affect the transcriptional coupling of NER, leaving the

cells to rely purely on the global pathway [23, 18].

Excision is performed systematically by a group of factors, including the Transcription factor II H (TFIIH). TFIIH is a complex crucial for both transcription initiation and repair. Importantly, TFIIH carries away the excised oligomers with the lesion [18]. These excised oligomers have a half-life of about 10 minutes before they get degraded [19].

When it comes to the repair of the UV-induced pyrimidine dimers, the contributions of the two pathways to the repair of the two damage types are different. CPDs do not disturb the helix structure as much as 6-4PPs. This makes them escape recognition by XPC, which relies on strong helix distortions. (Binding of other factors helps recruit the XPC factor there, but is not very efficient and out of the scope of this thesis [29, 23, 18].) Thus, 6-4PPs are more efficiently repaired and prioritized by the global NER. CPD repair is mainly left to TC-NER as the unnoticed lesions still stall RNA Pol II [23, 18].

1.1.2.1.4. Other DNA repair mechanisms

There are quite a few more repair mechanisms that, although not the focus of this thesis, deserve a mention. These include Mismatch Repair (MMR) tasked with fixing mononucleotide mismatches; Homologous Recombination (HR, also called recombinational repair) that repairs e.g. DSBs using the sister chromatids; Non-Homologous End Joining (NHEJ) which is an error-prone approach to ligating DSBs; Alternative End Joining (alt-EJ), and Fanconi Anemia (FA) Pathway [4, 5].

1.1.2.2. Damage tolerance mechanisms

Tolerance mechanisms come into play when damage disrupts central cell processes, and if not dealt with quickly it might lead to catastrophic consequences for the cell. This can happen e.g. during replication, where unrepaired lesions and delay caused by them might lead to cell death. Sometimes, repair mechanisms are unable to cope with a heavy damage load in a timely manner. This is where translesion polymerases, also known as DNA damage bypass, come to the aid [28].

Translesion polymerases (TLSPs) are special polymerases that are able to synthesize DNA over lesions. This comes at a cost of accuracy: substitutions and deletions. TLSPs specialize, i.e. tend to synthesize with a slightly better accuracy over some lesions than others [28]. Usually, TLSPs are employed very shortly, before the precise polymerase takes over [26].

1.1.2.3. Cell cycle checkpoints

In some phases of the cell cycle, damage poses more of an issue than in others. Hence, cells have checkpoints for DNA damage at 3 different times of the cell cycle. These checkpoints delay the progression of the cell cycle to buy more time for the repair machinery to handle the damage [4]. Specifically, these checkpoints can block transitions from G1 to S, from G2 to M, or slow down the S phase [4, 26]. Failures in repair, tolerance, and inability to exit the cell cycle arrest due to damage, lead the cell to die.

1.2. Consequences of the loss of integrity of the DNA

Although not a focus of this thesis, this chapter serves as a brief highlight of consequences of DNA damage at different levels (more can be found in [6]). As briefly indicated in the previous chapter, DNA damage can have many problematic outcomes not only for the molecule but also for the cells and whole organisms. DNA damage can be deleterious and can lead to mutations. Apart from causing enough of a crisis to the cell to induce cell death mechanisms, it can have other effects, both for cells and for whole organisms.

The impact of DNA damage induced loss of integrity of the DNA can be most appreciated through DNA repair diseases (e.g. Xeroderma Pigmentosum or Cockayne Syndrome). Some alterations to repair pathways are not survivable. Others produce diseases or syndromes that strongly increase the risk of cancer [30]. DNA damage is also an important factor in organ functioning, neurodegenerative diseases, infertility, frequent aircraft travel, and potential space exploration [24, 6]).

Unrepaired DNA damage can lead to deletion of bases or substitutions in the DNA of the daughter cells. There are many different types, but overall we call those permanent errors ‘mutations’ or ‘alterations’ of the DNA sequence. They range from small ones (e.g. single base substitutions, in-dels) through middle ones (multiple-base-substitutions and indels) to large ones (structural variants). Somatic mutations have been mapped both in cancer cells [31, 32] as well as healthy tissues [33, 34]. The vast majority of these mutations are considered harmless. Some mutations, however, can have an array of effects on the functioning of the cells, like changes in protein structure, silencing of the genes, and changes in expression. Moreover, mutations have been implicated as a factor in tumorigenesis [6, 35, 36], and some other diseases.

Finally, DNA damage and its many consequences – including mutations, as well as the persistence of the lesions themselves [37], and repercussions of cell's response to the damage – are considered strong contributors to aging [6].

1.3. Genomic studies of DNA Damage and Repair

Knowing the potential dire consequences of the DNA damage induced loss of integrity of the genome, one might expect a vested interest in studying the related processes. In fact, the topic has been approached from many angles throughout the years, leading to recent development of assays allowing to precisely locate the DNA lesions induced by an array of DNA damaging agents, and the activity of different DNA repair pathways. In this chapter, I first present a short historical overview of the studies of DNA damage and repair, to then focus on the next-generation sequencing-based, whole-genome, and high-precision strategies for mapping locations of the two stages of the mutagenic process: the DNA damage and the activity of its repair.

1.3.1. History of DNA damage and repair studies

This section serves to paint a very general picture of the history of the DNA damage and repair field and is based on a few important reviews. For a deep dive and specific references, I refer the reader to these prominent works [38, 22, 39].

The mutagenic effects of different radiation sources (Ionizing Radiation and UV) on the cells, as well as cells' potential to recover from them (photoreactivation), were already noticed between the 1920s and 1940s. Curiously, the scientists had a vague idea of the damage and repair mechanisms of the cells but did not yet know that the subject of damage was the DNA. The potential repercussions of mutagenesis for the structure of the DNA became clearer with the discovery of the double helix structure in 1953. That finding was followed by a mass of molecular experiments, many of them extensively using mutagenesis as a tool to study e.g. gene function. But at the same time, the interest in understanding mutagenesis itself was quite low.

The first successfully detected DNA damage was a CPD (cyclobutane pyrimidine dimer) induced by UV irradiation, in 1967. The different mechanisms of the repair of UV-induced DNA lesions were elucidated in the 1940-50s (photoreactivation, not present in human cells), and 1980-90s (nucleotide excision repair, present in human cells). In the meantime, many methods for quantification and detection of various kinds of DNA damage were being developed. Two notable examples are the still popular, so-called 'comet assay' (single cell gel electrophoresis assay, 1980s), which visualizes

the amount of DNA single-strand breaks in each observed cell; and the LM-PCR (ligation-mediated PCR, 1990s) that was used for mapping sites of enzymatic clipping of CPDs by endonuclease from DNA. The problem with approaches like the comet assay is that while allowing for an overall relative quantification of damage it does not provide information about the location of the damage in the DNA. On the other hand, methods similar to LM-PCR can provide the position of the damage with even nucleotide precision but are limited in breadth, providing this information for specific genomic regions only.

The first methods combining higher resolution and wide breadth appeared during the 2010s, using CHIP-Chip approaches. First, genome-wide maps of UV lesions in yeast were generated using chromatin immunoprecipitation with an antibody against CPDs, combined with microarrays. Soon after the first map of UV-induced damage in humans at a chromosome scale appeared. The resolution was further improved by adapting the idea to CHIP-seq in 2017 [40], which while resulting in many interesting insights, was still not at the level of nucleotide resolution. The use of sequencing was the right bet though, and many truly high-precision maps of DNA Damage and repair have been developed since the advent of next-generation sequencing (NGS).

1.3.2. Genome-wide high-resolution DNA damage and repair maps

Together with the first technologies of NGS the possibility appeared to map elements both genome-wide and at high-resolution. The sequencing methods on their own could not read the damaged DNA bases (either not recognizing other bases than basic A, C, G, and T, or even being hindered by the bulky damage). Thus, in the last decade or so, the efforts to combine the historical knowledge on recognizing DNA damage and repair with the promises of the NGS skyrocketed [41, 42, 39]. Many successful methods have been developed, using various strategies, at different resolutions, and for different species, resulting in damage and/or repair maps for many mutagens. Next, I will summarize a few common NGS-based strategies undertaken in several selected damage and repair mapping efforts, outlined in Table 1.1 and visualized in Figure 1.1. Finally, I will specifically detail three methods and associated datasets of high importance for the work presented in this thesis: HS-damage-seq and XR-seq employed to map UV-induced lesions and their repair, and an alkylation mapping method called NMP-seq.

The efforts in NGS-based mapping of DNA damage and its corresponding repair can be divided into two main strategy types: end marking and lesion reversal or bypass. End

Method, dataset	Methodology	Damage type	Map type	Kinetics	Resolution
XR-Seq [18]	lesion reversal, antibody-based	UV-C (CPD, 6-4PP)	Repair	-	excised fragment
XR-seq [23]	lesion reversal, antibody-based	UV-C (CPD, 6-4PP)	Repair	+	excised fragment
CPD-seq [43]	end marking (break), enzymatic cleavage	UV-C (CPD)	Damage	+	single nucleotide
Damage-seq [44]	end marking (pol block), antibody-based	cisplatin and oxaliplatin adducts	Damage	-	single nucleotide
Pt-XR-seq [44]	lesion reversal, antibody-based	cisplatin and oxaliplatin adducts	Repair	-	excised fragment
NMP-seq [24]	end marking (break), enzymatic cleavage	MMS (m7G, m3A)	Damage	+	single nucleotide
tXR-seq [45]	lesion bypass, antibody-based	UV-C (CPD); Benzopyrene (BPDE-dG adduct)	Repair	-	excised fragment
HS-Damage-seq [19]	end marking (pol block), antibody-based	UV-C (CPD, 6-4PP)	Damage	+	single nucleotide

Table 1.1: *Damage and repair mapping methods and datasets outlined in this thesis. The table includes the methodology used, type of damage mapped, map type, resolution, and whether kinetics are included in the dataset. Datasets of special importance have the method names bolded.*

marking takes its name from the preparation of reads so that one end is located close to where the lesion was, allowing to, later on, infer from the sequenced and mapped reads its precise location. Lesion reversal/bypass methods rely on either reversing the un-sequenceable damage or bypassing it with low-fidelity translesion polymerases so that a read can be generated containing the location of the damage inside its sequence.

The methods can also be divided based on the methodology used for recognizing the damage. One big group of methods uses enzymatic cleavage of the lesion for its recognition (like the LM-PCR approach mentioned above), while the other one is based on labeling and pulling down of the reads containing the damage (like CHIP-Chip and CHIP-seq approaches) - usually with specialized antibodies, or using chemical labeling. Enzymatic methods are considered more accurate, as the antibodies used in immunoprecipitation can be biased toward the recognition of damage within specific nucleotide contexts with varying affinities.

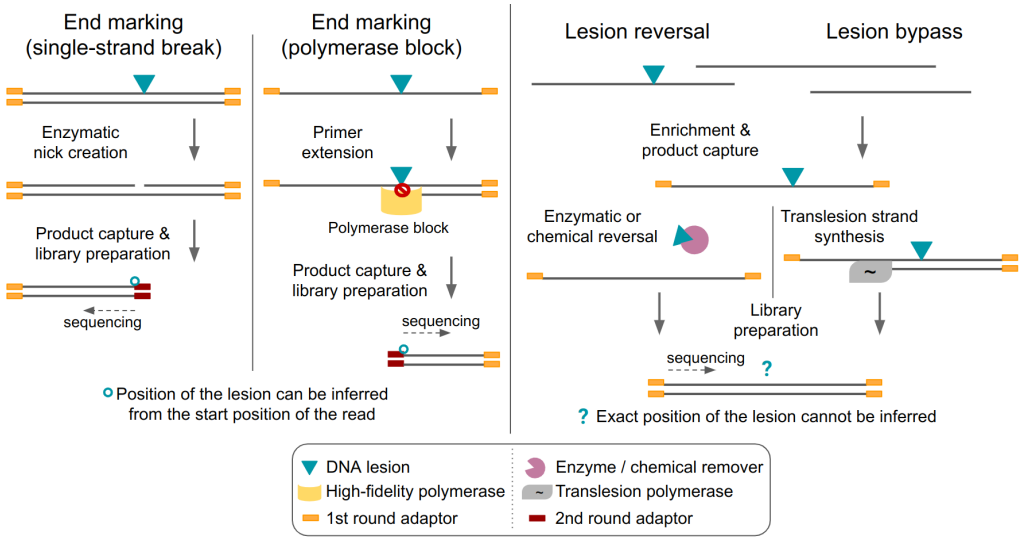

Figure 1.1: *Schematic, visual representation of described damage-mapping strategies.*

1.3.2.1. End marking

End marking can be achieved in two manners: a polymerase block or a generation of a single-strand break. Polymerase block induced by a big enough lesion (or a label attached to it) causes the synthesis of the strand and the corresponding read to end there. A single-strand break can mark the end of the read when generated close to the damaged site, followed by adapter ligation. End marking methods generate single base pair or single nucleotide precision damage maps.

Notable examples of approaches based on polymerase stalling are Damage-seq [44] and HS-Damage-seq [19]. They utilize specialized anti-damage antibodies to immunoprecipitate DNA carrying the damage. The next step is the end marking: the primer extension with a high-fidelity polymerase is performed, and when the polymerase blocks on the lesion, the extension ends there, marking the location of the lesion. HS-Damage-seq takes its name from 'high-sensitivity' due to the addition of a second enrichment step. Subtractive hybridization depletes the reads containing both adapters (which means they were read through completely by the polymerase) carrying non-damaged DNA. A strong side of these sister methods is that they can be adapted to study any polymerase-blocking lesions given the availability of specific antibodies.

The end-labeling single-strand breaks usually are generated using enzymatic cleavage of the damage from the DNA. The differences between enzymatic approaches come mostly from enzymes used for nick creation. For example, CPD-seq [43] and NMP-seq [24], aimed at mapping UV-induced CPD and MMS-produced lesions respectively,

follow roughly the same protocol and even use the same endonuclease (APE1) to cut the DNA after cleaving the damaged base, diverging in the lesion cleavage enzyme: CPDs by T4endoV and NMPs by AAG.

1.3.2.1.1. HS-Damage-seq mapping UV photoproducts in human fibroblasts

In high-sensitivity damage sequencing (HS-Damage-seq, [19]), extracted DNA is first sheared through sonication, followed by end repair (for background noise reduction) and ligation of first-round adapters to both ends of the produced double-stranded fragments (Figure 1.2). Strands are separated, and fragments containing damage are immunoprecipitated using specialized antibodies, depleting the undamaged DNA from the sample. Primer extension with a high-fidelity polymerase is performed next, from the 3' end of the fragment. When the polymerase encounters the lesion, it gets blocked, and the extension ends there. The primer has Biotin attached, allowing for streptavidin-based purification, and the non-extended fragments are discarded through subtractive hybridization. For the captured, proper fragments the second-round adaptor is added, and PCR is used for the second-strand synthesis and amplification, forming the library for the sequencing.

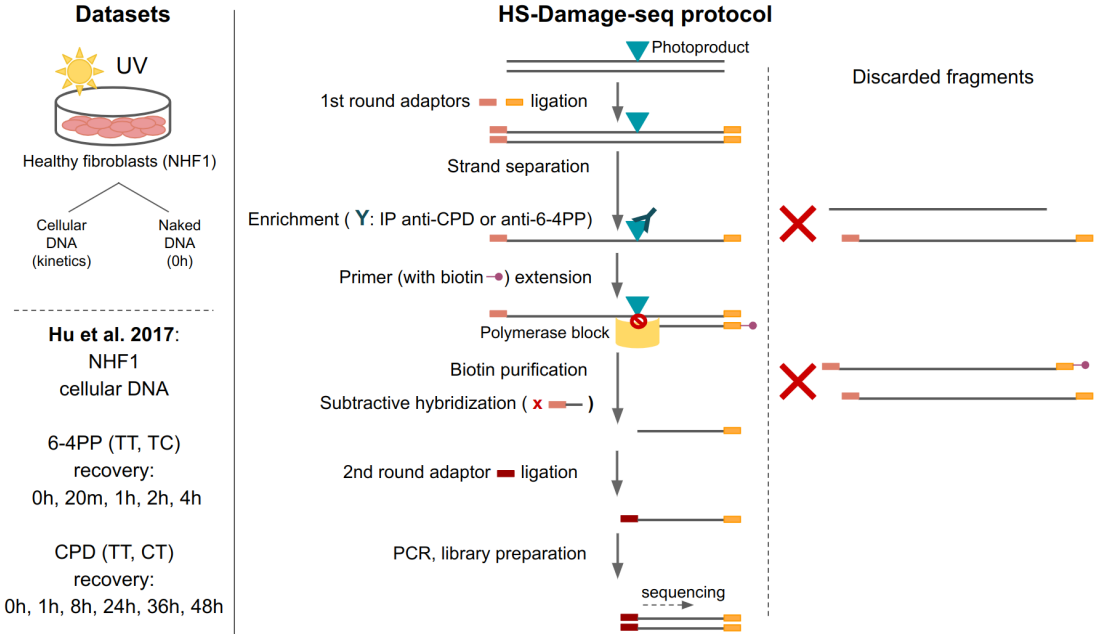

Figure 1.2: *HS-damage-seq protocol outlined in depth, with datasets of interest.*

HS-Damage-seq was designed and utilized to map the most frequent photoproducts – CPDs, and 6-4PPs – in healthy human fibroblasts as well as "naked" (stripped of nucleosomes, proteins, etc.) DNA extracted from the cells, shortly (10 or 20s) irradiated with UV-C. Together with the method, the authors published the datasets they generated, covering the 2 most abundant di-pyrimidines for each photoproduct

type. Importantly, for the cellular DNA (non-naked) they generated these damage maps at multiple time points after exposure, giving unique insights into the changing landscape of UV damage as time progresses and repair happens. 6-4PPs in TT and TC context were mapped at 0h, 20m, 1h, 2h, 4h post-exposure, and CPDs in TT and CT at 0h, 1h, 8h, 24h, 36h, and 48h. Authors propose an approach called 'subtractive HS-damage-seq' to infer repair in each time interval, in between every two consecutive time points, by simply subtracting one landscape from the other.

Of note, due to the nature of technologies used in these experiments, HS-Damage-seq (and similar methods reliant on PCR and sequencing) have to deal with amplification and saturation effects. Thus, they do not provide a total, accurate quantification of damage along the genome in the population of cells, but rather a relative measure. This has important implications for directly comparing one landscape to another and makes it impossible to infer the exact, total amount or rate of repair.

1.3.2.1.2. NMP-seq mapping of MMS-induced lesions in yeast

N-methylpurine sequencing (NMP-seq, [24]) starts by sonicating the DNA, after which the adapters are ligated to double-stranded DNA fragments (Figure 1.3). Next, a nick is created in the site of the lesion, through enzymatic means: AAG (3-alkyladenine DNA glycosylase) cleaves the base, leaving an AP site, to be next cut by APE1 (Apurinic/apyrimidinic endonuclease 1) in order to leave a single-strand break immediately upstream of where the damage was. Strands are separated, and a secondary, double-stranded adaptor with Biotin is ligated to the free 3'OH nicked ends created by the enzyme. This allows undamaged strands and the other side of the nicked fragments to be filtered out when pulling down the Biotin with streptavidin. PCR is performed for second-strand synthesis and amplification to prepare the library for NGS (using the Ion Torrent Proton sequencing platform).

NMP-seq is the first and so far only (at the time of writing this thesis) method designed for mapping alkylating damage. It was employed to map N-methylpurines m3A and m7G produced by a short (10m) treatment with the alkylating agent MMS in yeast. To understand the impact of BER, MMS yeast damage maps were generated at different time points after exposure in both wild-type (WT) and BER-impaired (mag1Δ mutant) yeast cells. In both cases, there was a clearly observable reduction in the number of mapped lesions. In the BER-impaired strain, the effects of NER repair could be appreciated. The authors additionally generated a damage map of chronically (3h) MMS-treated mag1Δ strain.

Although NMP-seq is the only method specifically created for that task, in principle,

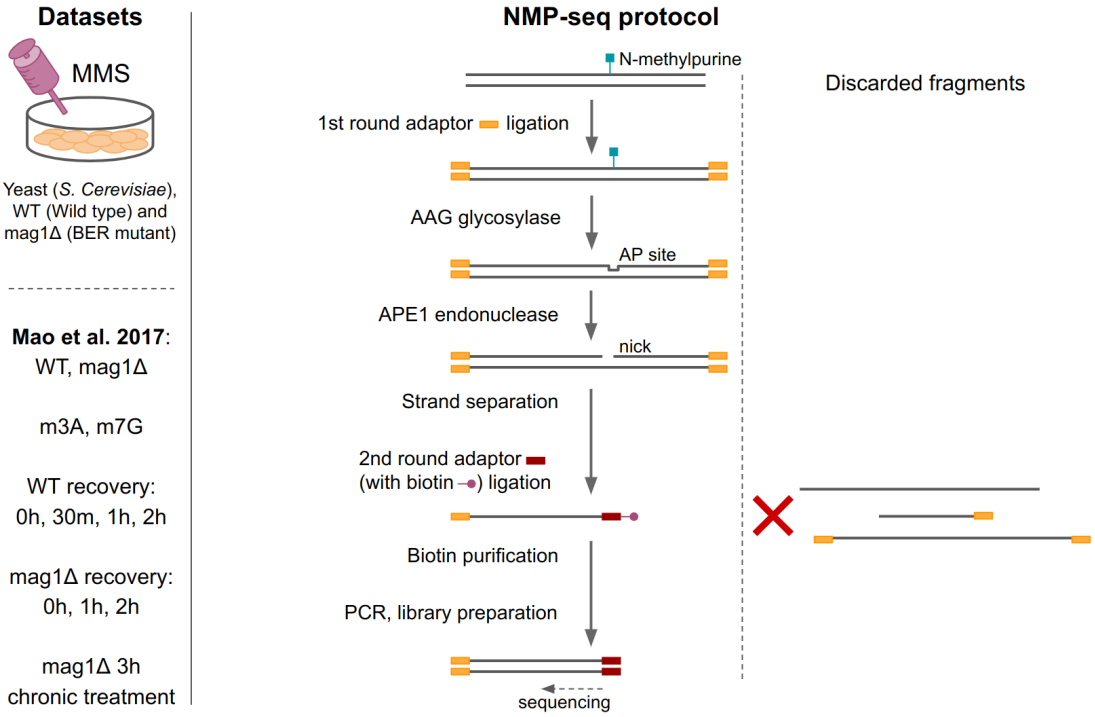

Figure 1.3: *NMP-seq protocol outlined in depth, with datasets of interest.*

lesions produced by alkylators could be also mapped by AP-site or single-strand break detection approaches (mentioned in one of the sections below), provided an enzymatic or chemical way to convert one into the other.

1.3.2.2. Lesion reversal and bypass

Strategies based on lesion reversal and bypass come in handy when the DNA damage is too small to cause polymerase block, when no reliable enzymatic approaches are available to process it, or to map excised oligonucleotide fragments, as in the case of NER. In lesion reversal, the damage is directly reversed to a sequenceable nucleotide, so that it can be processed by the high-fidelity polymerase. The rest of the methods bypass the lesion altogether with Translesion Strand Synthesis (TLS) utilizing low-fidelity polymerases. Lesion reversal and bypass strategies are characterized by a slightly worse - although still high - resolution of read insert size. While this might not be the best for locating the damage, it works perfectly for NER repair activity mapping, aided by damage sequence specificity. For this reason, next, I describe only the XR-seq (XR standing for eXcision Repair) family of methods [18, 44, 45].

XR-seq methods ingeniously utilize the facts that the NER machinery excises the damaged DNA fragments together with bulky lesions, and that in humans the length of these fragments, although short, is sequenceable. First, the DNA bound to the NER proteins is pulled down by dedicated antibodies. Next, similarly to Damage-seq

approaches, isolated fragments are purified by immunoprecipitation with specific anti-damaged-DNA antibodies, to further enrich the ones carrying the damage of interest. The main difference between XR-seq [18] and Pt-XR-seq [44] comes in the type of damage mapped, and hence antibodies used at this stage (against photoproducts or Pt-d(GpG) diadducts) as well as an enzyme used for lesion reversal to allow for mapping the read (photolyase and sodium cyanide, respectively). The tXR-seq (t standing for translesion) extends these methods to currently non-reversible lesions repaired by NER (like BPDE-DNA adducts) by swapping the lesion reversal for lesion bypass by translesion DNA synthesis (TLS) polymerases [45].

1.3.2.2.1. XR-seq mapping of NER of UV-induced lesions in human cells

As described above, Excision Repair sequencing (XR-seq, [18]) begins with immunoprecipitation of TFIIH complex and excised oligonucleotides carried by it (Figure 1.4). After adapter ligation to the single-stranded DNA fragments, another round of immunoprecipitation is performed, with anti-CPD and anti-6-4PP antibodies. Next, photoproducts are reversed with the dedicated photolyases, so that the high-fidelity polymerase can perform uninterrupted during PCR. After that, the samples are purified on the gel, and only fragments of sizes corresponding to small, excised oligonucleotides are used for sequencing.

XR-seq has been used for mapping UV-induced lesions in two important, consecutive papers.

The first one [18] published the method together with single time point (1h) damage maps after a 20s-long UV pulse. The treated cell populations included healthy human fibroblasts, as well as cells from patients with diseases of two subpathways of NER: Xeroderma Pigmentosum, and Cockayne Syndrome. This unique dataset elucidated differences in repair by global and transcription-coupled NER as compared to the cells in which both mechanisms are intact.

A year later [23], the authors followed with a dataset exploring the kinetics of NER in healthy human fibroblasts, this time performing XR-seq at different times after exposure to the mutagen. Repair of CPDs was registered at 1h, 4h, 8h, 16h, 24h, and 48h, and that of 6-4PPs at 5m, 20m, 1h, 2h, and 4h. As the approach comes from the same group that devised and inferred the repair of photoproducts by subtractive HS-damage-seq, it is only natural to compare their resolution: XR-seq is more precise and can measure repair "down to 0.1% of damage for a given region" (citing the authors of [19]), while subtractive HS-damage-seq capabilities end at 10%. But XR-seq has some caveats too - due to the short time in which excised fragments are

degraded, it represents only around 10m 'snapshot' of the repair process before the given time point of measurement; additionally, XR-seq provides only information for each photoproduct overall (e.g. CPDs) while subtractive HS-damage-seq separates the information by di-pyrimidines (e.g. CPDs in TTs and CPDs in CTs).

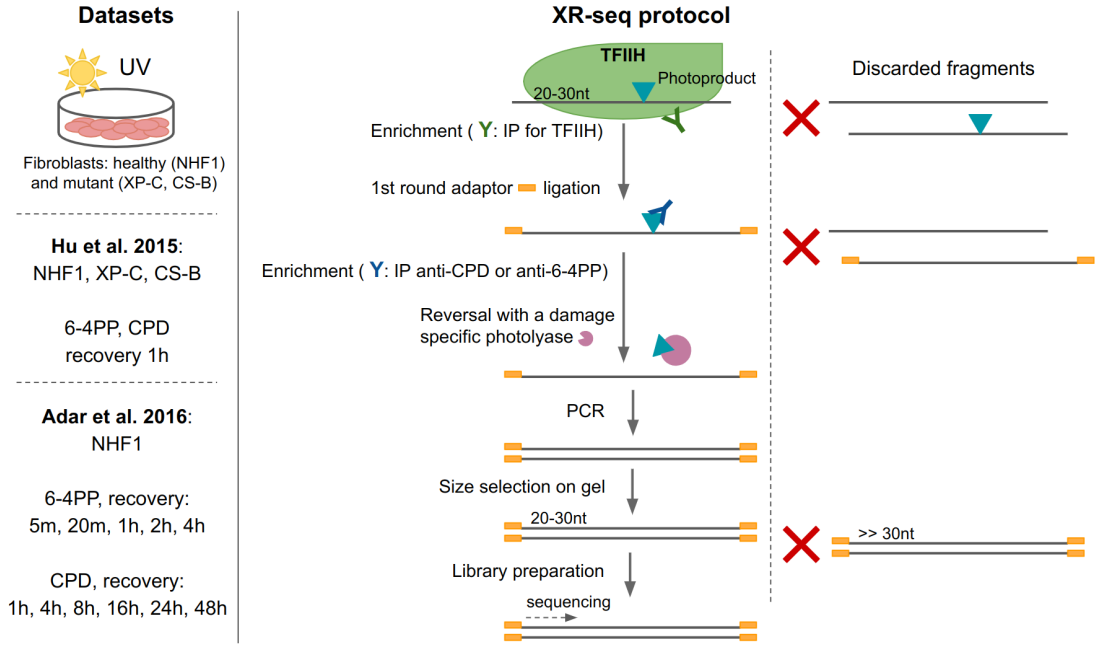

Figure 1.4: *XR-seq protocol outlined in depth, with datasets of interest.*

1.3.2.3. Other methods: detection of breaks, and third-generation sequencing

While out of the scope of this thesis, it is worth mentioning the existence of approaches aimed at detecting abasic sites (SSiNGLe-AP [46]), single-strand (GLOE-seq [47], SSiNGLe [48]), and double-strand breaks (END-seq [49]). Apart from providing crucial information on the location of these endogenous events in the cells, as well as their appearance after exposure to external factors, they could also be used for mapping other DNA lesions that are chemically or enzymatically convertible into one of these events (e.g. NMPs). GLOE-seq, aimed at the detection of free 3'OH single-strand breaks has shown proof of this concept.

Finally, while still in their infancy and also not covered in this thesis, one cannot ignore the promising potential of third-generation sequencing approaches for applications in DNA adductomics. Also called long-read sequencing, PacBio SMRT-seq and Nanopore methodologies have been applied to directly – without any prior handling – detect naturally occurring modifications on DNA in a genome-wide manner. This holds great promise for the direct detection of DNA damage. In fact, there are already two SMRT-based methods, one mentioned above, mapping single-strand breaks (SSiNGLe, SMS variant [48]) and RADAR-seq [50], employed for the task of

detecting ribonucleotides and UV-lesions in small, bacterial genomes. While PacBio SMRT holds an important limitation in mapping bulky lesions, due to the use of polymerases and the potential of a block occurring, Nanopore seems to be well suited for the task, with first successes in recognition of 8oxodG, 6mA, 5mC sites [51].

1.4. Studying mutational processes

Despite decades of accumulated studies on DNA repair machinery, a genomic perspective has only very recently become within reach, thanks not only to the development of massively parallel sequencing techniques but also the annotation of genomic elements.

Once obtained the whole-genome landscape of a mutational process - whether through a damage map, repair map, or mutations gathered from many datasets - one can start to probe it from different angles. I will focus mostly on its relation with genomic and chromatin features, and the effects of impairment of dedicated repair mechanisms.

1.4.1. Genomic features

The DNA - the genome - inside the nucleus is subject to many processes, interactions, and modifications. The differential activity of these processes along the genome creates regions with distinct characteristics, which we call broadly genomic 'features'. A few examples of features follow:

- the sequence composition of a given fragment,
- whether the sequence encodes a gene or a regulatory region,
- whether the given region is transcribed (or expressed) or not,
- when does this DNA fragment undergo replication,
- whether a protein is bound to DNA,
- how accessible or folded is this part of the genome,
- where within the 3D space of the nucleus is this sequence located,
- whether the DNA, or the histones it is wrapped around, carry a modification or a mark.

Note, that many of the features are specific to the tissue and will have a degree of variability depending on the cell type.

DNA damage and repair happen in the context of these, and other features of the genome. Hence, there is a vested interest in developing an understanding of the interplay of DNA damage deposition, and DNA repair activity with the genomic

features, resulting in particular mutational patterns. Moreover, elucidating the factors affecting the mutagenic process is an important step for improving modeling the expected mutation rate for driver discovery.

Since mutations are the consequence of unrepaired damage, the study of mutations provides an archaeological record of the activities of damage and repair. Mutations may thus be exploited for the purpose of understanding the interaction of genomic features with the mutagenic process. Indeed, many correlations between these processes have been found already. Mutation rates have been shown to be influenced by replication timing, expression, chromatin conformation, and accessibility (measured by the DNAse I Hypersensitive Sites, DHSs) [52, 53, 54, 55, 40, 56, 57, 58] - mostly features happening at a large scale. Associations with smaller-scale features have been shown as well, including intron-exon boundaries [59], transcription factor binding sites [60, 61] nucleosome-linkers alternation, nucleosome DNA wrapping orientation [62], among others [63]. In fact, features of various scales have been shown to affect damage, repair, and mutation rates [64].

One of the most prominent small-scale genomic features so far in the study of mutagenesis is the nucleotide context of the mutation: in the case of single base substitutions this is defined by the nucleotide change (reference and alternate allele) and the 5'/3' nucleotides flanking the substitution.

When considering all mutations from a tissue, one important factor needs to be taken into consideration: they are generated by a mixture of various mutagenic processes. Hence, came the idea of decomposing the mutational signals and correlating them with the metadata of various tissues and samples. Different mutagenic processes are known to have different preferences towards generating mutations at certain nucleotide contexts. Hence, the frequency profiles based on these nucleotide contexts are used to for the decomposition task. The decomposed parts of the mutational profile are termed 'mutational signatures' [65]. Mutational signatures highlight these sequence context differences of various mutagenic processes. Interpreting the signatures within the sample history context, and with knowledge of the biology underlying some mutagenic processes, the etiology of some signatures has been recognized (with varying degrees of certainty) [66, 67, 68]. Examples relevant to this thesis include the set of mutational signatures SBS7 (single base substitutions, 7a-d) strongly associated with UV light exposure, and SBS11 indicated as potentially related to Temozolomide treatment [65, 69, 70].

Mutation rate variability analyses have been carried out for different signatures,

helping to elucidate important differences in associations with chromatin features depending on the underlying mutational process [62, 63]. Although they can serve as a useful first step, these correlations are still difficult to interpret mechanistically. For that, one needs to go to the two steps of the mutagenic process that can be directly affected by the features: damage distribution and DNA repair activity. Exploiting available data (damage and repair maps), these two parts of the mutagenic processes have been explored within the context of features like exon-intron boundaries or nucleosomes [62, 71]. Interestingly, a few clear differences in the interplay with genomic features have been found for some damage types. For example, damage formation with respect to the nucleosomes and linkers shows opposite patterns for UV-induced lesions and benzopyrene adducts [62, 63].

Of note, most of these interactions have been explored using two main approaches. The first consists of correlating the mutations/damage/repair rates across different genomic regions (usually obtained by dividing the genome into chunks) with the representation of features across the same regions. This approach has been used both for large-scale (e.g., expression, replication timing) and small-scale features (e.g., DHS). Another approach, more frequently applied to study fine-grained features which cover relatively small parts of the genome (and contain therefore comparatively few mutations), is based on stacking specific size windows around the feature [63]. Then, it is possible to compare the observed mutation rate of the mutagenic process accumulated across the stacked features with that expected on the basis of the sequence context obtained from the same stacked features. Importantly, all studies so far start by mapping the genomic features of interest and then probing the components of mutagenesis with respect to them. Until recently [72], no approaches aimed at partitioning the genome based on underlying differences in the intensity of the mutational processes had been carried out.

Genomic feature associations with damage, repair, and mutation rates produced by UV light have been studied quite extensively. On the other hand, far less is known (mostly due to a lack of data) on features interacting with the mutagenic processes produced by chemotherapeutic or alkylating agents. I review both in the next sections, with a special focus on UV damage repair.

1.4.2. Genomic features influencing UV mutagenesis

1.4.2.1. Damage formation

Regarding sequence contexts, CPD damage happens more frequently in TT di-pyrimidines, and 6-4PPs form predominantly in TT and TC contexts [19]. UV CPD

lesions have been found to be distributed uniformly across chromatin states, regardless of the di-pyrimidine context [19]. 6-4PP formation seems to be mostly uniform, with a slight increase in active and poised promoter and repetitive chromatin states, and a slight decrease in the heterochromatin state [19]. Moreover, the 3D conformation of the genome inside the nucleus seems to entail a relative increase in UV damage formation on the sites closest to the nucleus periphery, and a protective effect resulting in decrease in UV lesions at the center of the chromatin in the nucleus [23, 40, 73]. The wrapping of the DNA around the nucleosomes seems to also experience this protective function against CPD formation, with fewer UV lesions in DNA where the minor groove is facing 'in' towards the nucleosome [43, 62].

Many studies [43, 19, 71, 74, 75] implicated that various Transcription Factors (TFs) bound on the DNA might decrease or increase UV lesion deposition along the genome. This variability in effect seems to be caused by TF-specific changes to the structure of the DNA induced upon binding and hence is dependent on the TF, type of the damage, strand, and position relative to the TF binding motif [19, 71, 74].

1.4.2.2. Repair activity

Not only damage, but also repair can be sometimes biased towards specific sequence patterns. The XR-seq study of 6-4PPs found TCs to be the most frequently damaged di-pyrimidines, in opposition to HS-damage-seq [23]. This suggests that while in both contexts the 6-4PP damage is highly deposited along the genome, 6-4PPs in TC contexts are repaired more efficiently.

One might assume that as Global NER shows higher activity than TC-NER in the repair of 6-4PPs ,this takes place rather uniformly along the genome than the repair of CPDs, where TC-NER plays a more important role. In reality, the repair of both lesions exhibits variability along the genome, at different scales, and associates to the presence of different chromatin features. This makes sense in light of the postulated 'access, repair, restore' model of repair, stating the remodeling of chromatin necessary for improving access of the repair machinery [23, 29].

The works of [18, 76], using both healthy and mutant human fibroblasts, confirmed the higher efficiency of CPD repair in all transcribed regions (including enhancers), especially on the transcribed strand, and positively correlating with the RNA levels of transcripts. The contribution of TC-NER is made evident through a simple comparison of repair in mutant cells. In XPC, lacking the global NER pathway, the observed effect is even higher. In CSB, where TC-NER is turned off, the CPD repair is happening quite uniformly along the genome.

These findings have been expanded by studies of the kinetics of XR-seq [23] and subtractive-HS-damage-seq [19] in healthy human fibroblasts. For CPDs, the complementarity of the two datasets highlights the high repair that correlates well with decreases in the amount of damage in the transcribed strand. For 6-4PPs and global NER, these effects seem to be driven not by transcription, but by higher accessibility to the regions: damage levels decrease with time at DNaseI hypersensitivity sites. Both types of repair exhibit higher activity in active (open and acetylated) chromatin states at early time points. On the other hand, lower, but persistent through time, repair activity is observed in inactive (heterochromatic, repressed, repetitive regions) states. (This persistence shows up as an increase in repair at later times in comparison to the active chromatin states). The reason for the increased repair of inactive chromatin regions at later time points is not clear - it might be just that other regions have been mostly repaired already, and repair can move on to the de-prioritized parts of the genome. Alternatively, as postulated above, there might be an active remodeling of chromatin and nucleosomes, improving accessibility of repair in these sites [23, 29].

3D structure of the chromatin inside the nucleus seems to be a relevant factor for NER efficiency too [23, 73]. Although no effect was found for very early UV repair, at 2h after irradiation the nucleus-centric regions exhibited increased repair compared to the outskirts [73]. Nucleosomes seem to affect NER too [23], both at the level of nucleosomes-linkers, as well as whether a minor or a major groove of DNA faces the nucleosome [43, 62, 75]. Specifically, nucleosomes bound to the DNA seem to impede NER's access to it [23]. The presence of bound transcription factors exhibits a similar negative effect on NER [61, 71].

Locally, NER repair seems to be enhanced whenever there is active replication of the region, although on a larger scale, repair of early replicating regions is faster, especially for CPDs [77]. This effect seems to be most pronounced in early replicating inactive and transcribed chromatin states, and might be related to the increased accessibility of the said regions [77]. Similarly, most likely due to improved accessibility, exons relative to introns exhibit patterns of more efficient repair (especially in the case of global NER) [76].

1.4.3. Features of the alkylation-based mutagenesis

1.4.3.1. Damage formation

Little is known about alkylating damage formation in relation to genomic features. Currently, the best source is the yeast MMS NMP-seq data [24]. NMP-seq mapped m3A and m7G formation was slightly higher on the non-transcribed strand. This effect

was stronger for m3A and was magnified in a chronically treated, BER-deficient strain.

NMP-seq was further used to characterize the effects of 2 bound transcription factors [78]. In both cases, TF binding was found to impact the generation of the lesions, with a clear reduction in the formation of m7Gs. (Of note, at specific positions of the TF motifs, the formation was sometimes instead increased.)

1.4.3.2. Repair activity

Similarly as in the case of NER, BER seems to be affected by the chromatin structure. Nucleosome-bound regions experience impeded BER of m7Gs, and the minor groove-inwards positioned DNA exhibited lower repair (likely due to lower accessibility, as indicated by DNase-seq) [24, 62, 63]. These effects have been found to be further modulated by the presence of various histone marks, especially acetylation, associated with the unwrapping of nucleosomes [24]. The presence of bound transcription factors inhibits BER as well, although the width of the region around it that is affected is smaller (likely due to the smaller size of BER compared to NER) [78].

Exploration of the BER-deficient mag1Δ strain produced interesting insights: for example some time after MMS exposure (allowing the action of repair), m3A lesions on the transcribed strand progressively disappeared [24]. This suggests an involvement of TC-NER as a fall-back mechanism for m3A repair (at least in yeast), further supported by results for other repair mutants generated in the same study.

2. Objectives

The thesis's main objective is to broaden the understanding of the interplay between DNA damage caused by different agents, its repair, and general cellular processes. This may be separated into two specific objectives tackled by each of the projects:

Alkylating damage maps

1. Obtain nucleotide-resolution, genome-wide maps of DNA damage for several alkylating agents in human cells. This includes:
    - Develop an end-to-end automated computational pipeline coupled to an experimental alkylation damage mapping library preparation that yields the location of alkylated bases along the human genome (i.e. alkylation damage maps).
    - Compare these alkylation damage maps in human cells with those obtained in yeast and with the body of knowledge on alkylation lesions.
    - Uncover the human sequence context preferences for damage formation of two alkylating agents, MMS and TMZ, and their genomic distribution, as a first step to apply the damage maps to the study of the dynamics of repair of alkylation lesions.

UV Repair states

2. Develop a novel approach to study DNA repair dynamics across the genome that is not constrained by the prior mapping of genomic features. This includes:
    - Establish a normalization and filtering protocol to appropriately compare existing UV damage and snapshot repair maps.
    - Develop a method to infer UV damage repair based on maps of UV-generated damage at consecutive time points and encode this inferred repair (and snapshot repair) in a continuous manner along the entire human genomic sequence.
    - Exploit the kinetics of inferred repair activity (and snapshot repair) to segment the human genome into regions with distinct UV-generated DNA damage repair activity, thus obtaining UV-induced damage DNA repair states.
    - Study the composition of genomic features underlying the differences in UV damage repair activity across these DNA repair states.

3. Methodology

Given that this thesis is a computational, bioinformatics thesis, the pipelines developed as part of it are considered results, not methodology.

3.1. Experimental AB-seq protocol for alkylation damage mapping

Here we describe the technical details of our team's approach to generate ready-to-sequence alkylation damage libraries at nucleotide resolution for different time-from-exposure conditions and damaging agents along with the necessary controls. All AB-seq (Alkylation BER sequencing) experiments described in this section were performed by other lab members: Erika López-Arribillaga, Katyayani Anshu, Nuria Samper or Morena Pinheiro.

3.1.1. Cell lines and reagents

Human cell line RPE-1 hTERT were grown in DMEM:F12 (Invitrogen, 31330-095) at 37°C in 5% CO2 in the corresponding media supplemented with 10% FBS (Gibco, 100822139) and Penicillin-Streptomycin (Gibco, 15140122). Cells were treated with the indicated concentrations for the indicated times with Temozolomide (Sigma, T2577) from a 103 mM stock diluted in DMSO, cell-culture grade DMSO (PanReac AppliChem A3672,0100), or MMS (Sigma, 129925-5G), in serum-free media. Cells were washed twice with PBS prior to analysis or recovery. For recovery, cells were incubated in DMEM:F12 10% FBS after wash out of the treatment for the indicated time before processing.

3.1.2. Damage map library preparation

Human genomic DNA (gDNA) was extracted using the DNeasy Blood and Tissue Kit (Qiagen, 69504), following the manufacturer's instructions. During sample preparation, all purifications between reactions were performed using AMPure XP beads (Beckman Coulter, A63880). For sample preparation, the gDNA was digested with Fragmentase (NE BioLabs, M0348) followed by a PreCR Repair reaction (NE BioLabs, M0309S) and A-tailing (NE BioLabs, E7546), the first P7 adaptor (with protected modified ends) was ligated with a T4 Ligase (NE BioLabs, M0202S). Two simultaneous enzymatic reactions were performed (1) with hAAG (NE BioLabs,

M0313S) to leave an abasic site where a m7G or m3A was and (2) with hAPE1 (NE BioLabs, M0282S) to nick the abasic sites, leaving ligatable 3'-OH groups. The remaining DNA was subjected to dephosphorylation of 5' ends (Quick CIP NE BioLabs M0525S), next the DNA was denatured by heating at 95ºC for 5 min and immediately cooled on ice. Finally, a double-stranded biotinylated second P5 adaptor ligation was performed by a Quick DNA Ligase (NE BioLabs, E6056S). The second P5 adaptor should only ligate at the cut site APE1 has left, and those fragments were captured by streptavidin beads (Streptavidin C1 Dynabeads; Invitrogen, 65001), to perform a PCR (KAPA High Fidelity Polymerase; Roche, 7958927001) before sequencing in Illumina platform (NovaSeq 6000 with 150 bp pair-end reads).

3.1.3. Generated datasets

4 samples comprising a full dataset were prepared as follows: 30 min treatment with 10 mM MMS, no recovery, paired with a vehicle: H2O and 30 min treatment 5 mM TMZ, no recovery, paired with a vehicle: 5% DMSO. Vehicle-treated samples were used as controls and are called untreated henceforth.

3.1.3.1. Multiplex Indexed Adaptors

In the sequences of adaptors below, the following abbreviations are used: '5Phos' is a phosphorylated 5' end, that can bind to a free 3'OH; '3ddC' stands for dideoxycytosine terminator at the 3' end that blocks polymerase extension; '*T' represents an extra hanging T (to be paired with an A added in the gDNA fragment during A-tailing) with a phosphorothioate bond to prevent degradation; '5Biosg' is 5' biotin; and '3InvdT' stands for inverted dT at 3' end that inhibits both the extension by polymerases as well as degradation.

Untreated

P7 UDI001-forward

/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCGCGGTTATCTCGTATGCCGTCTTCTGCTTG/3ddC/

P7 UDI001-reverse

CAAGCAGAAGACGGCATACGAGATAACCGCGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T

P5 UDI001-forward

/5Biosg/AATGATACGGCGACCACCGAGATCTACACAGCGCTAGACACTCTTTCCCTACACGACGCTCTTCCGATC/3InvdT/

P5 UDI001-reverse

/5Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCTAGCGCTGTGT
AGATCTCGGTGGTCGCCGTATCATT/3ddC/

MMS

P7 UDI003-forward

/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGACTTGGATCT
CGTATGCCGTCTTCTGCTTG/3ddC/

P7 UDI003-reverse

CAAGCAGAAGACGGCATACGAGATCCAAGTCCGTGACTGGAGTTCAGAC
GTGTGCTCTTCCGATC*T

P5 UDI003-forward

/5Biosg/AATGATACGGCGACCACCGAGATCTACACCGCAGACGACACTCTT
TCCCTACACGACGCTCTTCCGATC/3InvdT/

P5 UDI003-reverse

/5Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCGTCTGCGGTGT
AGATCTCGGTGGTCGCCGTATCATT/3ddC/

DMSO

P7 UDI005-forward

/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCCACTGATCT
CGTATGCCGTCTTCTGCTTG/3ddC/

P7 UDI005-reverse

CAAGCAGAAGACGGCATACGAGATCAGTGGATGTGACTGGAGTTCAGAC
GTGTGCTCTTCCGATC*T

P5 UDI005-forward

/5Biosg/AATGATACGGCGACCACCGAGATCTACACAGGTGCGTACACTCTT
TCCCTACACGACGCTCTTCCGATC/3InvdT/

P5 UDI005-reverse

/5Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTACGCACCTGTGT
AGATCTCGGTGGTCGCCGTATCATT/3ddC/

TMZ

P7 UDI007-forward

/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAAGCTAGATCT
CGTATGCCGTCTTCTGCTTG/3ddC/

P7 UDI007-reverse

CAAGCAGAAGACGGCATACGAGATCTAGCTTGGTGACTGGAGTTCAGAC
GTGTGCTCTTCCGATC*T

P5 UDI007-forward

/5Biosg/AATGATACGGCGACCACCGAGATCTACACACATAGCGACACTCTT
TCCCTACACGACGCTCTTCCGATC/3InvdT/

P5 UDI007-reverse

/5Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTCGCTATGTGTGT
AGATCTCGGTGGTCGCCGTATCATT/3ddC/

3.1.4. LC-MS/MS experimental set-up

Cells were treated with the drug, washed and had gDNA extracted using DNeasy Blood and Tissue Kit (QIAGEN, Ref.69504), treated with RNaseA and purified with AMPure XP magnetic beads at a 1:1.8 DNA:beads ratio and eluted in 100 μL EB buffer. Samples were incubated for 30 min at 100ºC in presence of 0.01 N HCl, and put on ice. To isolate the most labile modifications, we transferred samples to Amicon Ultra 0.5 ml 3K columns, centrifuged extensively, and rescued the methylated bases-containing flow-throughs, or "thermal fraction".

The fraction which did not pass the membrane was subjected to enzymatic hydrolysis at 37ºC with calf intestinal Alkaline Phosphatase, DNase I and Snake Venom Phosphodiesterase I. This digestion yielded the "enzymatic fraction". Samples were then desalted with acetonitrile prior to liquid chromatography-tandem mass spectrometry analysis (LC-MS/MS). This was performed on a LTQ-Orbitrap XLmass spectrometer combined with an EASY-nLC 1000. Desalted samples were injected into the columns and were distinctly separated through reversed-phase chromatography by a homemade analytical column of 50 cm with an inner diameter of 75 μm and packed with 4 μm Hydro-RP of 80 Å. The MS operation mode comprised a positive ionization with a 2 kV nanospray and a temperature of 200ºC.

Full MS scans were taken with the following parameters: one micro scan, 6×104 resolution, and mass range between 100 and 700 m/z. Fragmentation was achieved

by collision-induced dissociation (CID) at a normalized collision energy of 35 and 25 NCE for bases and DNA/RNA nucleosides, respectively. The isolation window was narrowed to 2.0 Th and activation time was set to 10 ms. The data were collected and acquired in Xcalibur software. Calibration curves were constructed using commercially available standards for modified and unmodified bases/deoxynucleosides and acquired data from the untreated and treated samples analyzed with Skyline software, using the fragment areas and retention times of each compound.

3.2. Computational part of AB-seq

Here we describe the technical details of our computational approach to derive genomic maps of alkylation damage at nucleotide resolution starting from the AB-seq damage libraries described above. All the steps below – unless explicitly indicated otherwise – are integrated into an in-house computational pipeline that can be efficiently run parallel on multiple cores on the cluster, with just a configuration file provided. The version of the pipeline code used in the thesis can be viewed in the `GitHub` repository at `https://github.com/bbglab/ABseq-PIPE`. (Note that at the time of writing the thesis, as this work remains unpublished, so does the repository. Once we generate all planned maps, the pipeline will be published alongside them, and the code will be made fully open and available. In the meantime, the code can be viewed upon request.)

The pipeline was implemented in `python3` for human reference genome version hg19 canonical. Pipeline includes an environment outlining necessary versions of packages and tools.

3.2.1. Workflow step by step

The pipeline was run for a library comprising 4 samples (MMS, TMZ, untreated, DMSO). It was executed using 49 cores and 250GB total, utilizing the threading options in external tools whenever possible, and the parallelization with a `Python multiprocessing` package `Pool` class whenever possible.

3.2.1.1. Main damage processing pipeline

Files corresponding to the same sample re-sequenced at different times, machines, and/or lanes are merged in the same order for both read1 and read2. Reads are deduplicated in a paired-end manner using a `bbmap clumpify.sh` tool as follows:

```
clumpify.sh threads=49 in1=DMSO_1.fastq.gz in2=DMSO_2.fastq.gz
 ↪  out1=DMSO_1.dedup.fastq.gz out2=DMSO_2.dedup.fastq.gz
 ↪  dedupe
```

Trimming low sequencing quality bases (less than 15) and adaptor sequences (P7 forward adaptor, preceded by A, on 3' side for read1 and P5 reverse adaptor on 3' side for read2), discarding sequences ending up shorter than 15 nucleotides is performed:

```
cutadapt --times 3 --overlap 9 -j 49 -m 15 --nextseq-trim=15
 ↪  --pair-filter=any -a AGATCGGAAGAGCACACGTCTGAA... -A
 ↪  AGATCGGAAGAGCGTCGTGTAGGG... -o
 ↪  DMSO_1.TRIMMED_PAIR.minSEQQ15.fastq.gz -p
 ↪  DMSO_2.TRIMMED_PAIR.minSEQQ15.fastq.gz
 ↪  DMSO_1.dedup.fastq.gz DMSO_2.dedup.fastq.gz
```

Alignment of reads to the human hg19 reference genome is done in a paired-end mode, forcing the reads to be separated by a maximum insert size of 2kb as expected from the fragmentation.

```
bowtie2 -x Bowtie2Index/genome -1
 ↪  DMSO_1.TRIMMED_PAIR.minSEQQ15.fastq.gz -2
 ↪  DMSO_2.TRIMMED_PAIR.minSEQQ15.fastq.gz -S
 ↪  DMSO.TRIMMED_PAIR.minSEQQ15_both.sam -p 49 --maxins 2000
 ↪  --no-mixed --no-discordant
```

From produced alignment files, bam files are generated considering only properly mapped pairs, filtered for mapping quality (discarding less than 15), sorted and indexed. Read1 file generation (for read2 `-f 131` used instead):

```
samtools view -F 12 -f 67 -b
 ↪  DMSO.TRIMMED_PAIR.minSEQQ15_both.sam >
 ↪  DMSO_1.TRIMMED_PAIR.minSEQQ15_1.bam -@ 49
```

Read1 mapping quality filtering (for read2 `-f 131` used instead):

```
samtools view -@ 49 -q 15 -F 12 -f 67 -bS
 ↪  DMSO_1.TRIMMED_PAIR.minSEQQ15_1.bam >
 ↪  DMSO_1.TRIMMED_PAIR.minSEQQ15_1.minMAPQ15.bam
```

Read1 sorting (same for read2):

```
samtools sort DMSO_1.TRIMMED_PAIR.minSEQQ15_1.minMAPQ15.bam -@
 ↪  49 -m 5G -o
 ↪  DMSO_1.TRIMMED_PAIR.minSEQQ15_1.minMAPQ15.sort.bam
```

Read1 indexing (same for read2):

```
samtools index
 ↪  DMSO_1.TRIMMED_PAIR.minSEQQ15_1.minMAPQ15.sort.bam -@ 49
```

The total overlap of each read1 with problematic regions is calculated using bedtools and awk. The definition of problematic regions (`all_disallowed.bed`) is the same for both projects and can be found below in the methods chapter 3.3.1.2 of the Repair States framework.

```
intersectBed -abam
 ↪  DMSO_1.TRIMMED_PAIR.minSEQQ15_1.minMAPQ15.sort.bam -b
 ↪  all_disallowed.bed -bed -wao | cut -f1,2,3,4,6,16 | awk
 ↪  'BEGIN{{ FS=OFS=\"\t\" }} {{ a[$1 FS $2 FS $3 FS $4 FS $5]
 ↪  += $6 }} END{{ for (i in a) print i, a[i] }}'
```

Duplicate reads (with respect to chromosome, start, end, and strand) are dropped. Next, the reads are filtered to keep only those overlapping problematic regions less than the set threshold (10%). Reads aligned to the mitochondrial genome are discarded, and only ones on chromosomes are considered (chr1-22, chrX, and chrY) next.

Positions of lesions are inferred from the reads (if reads do not point to the outside of the reference genome assembly), assuming that they are immediately upstream of the start of read1, and on the complementary strand. Additionally, the pentamer corresponding to the reference sequence of the inferred lesion position and the 2bp contexts on each side are retrieved. The information is saved into bed-like files and sorted with bedtools.

```
sortBed -i DMSO1.bed_unsorted -faidx human.hg19.genome >
 ↪  DMSO1.bed
```

Finally, the inferred damage positions are deduplicated. The whole above process is accompanied by saving outputs and statistics about the most crucial steps, and is then automatically summarized into a table of read counts through all of it.

3.2.1.2. Integrated downstream validation analyses

3.2.1.2.1. Context sequence plots

Total counts of bases and two types of contexts (tri- and pentanucleotide) are calculated. Next, all positions with contexts including an unknown base (N) are discarded. Positions are collapsed over both strands. The total genomic amount of bases, trinucleotides, and pentanucleotides is calculated with overlaps from the forward strand of the reference genome. The total counts of bases and contexts within each sample are normalized by the total genomic counts (summed over both strands). The normalized counts are put into frequencies (the sum of all becomes 1). All three options: raw counts, normalized counts, and frequencies can be visualized on either single-sample (for bases and both contexts) or treated-control pair plots (bases and triplets only).

3.2.1.2.2. Genomic damage distribution plots

The generation of genomic chunks is the same for both projects and can be found below in the methods chapter about the Repair States framework. For each sample, from the final, position-deduplicated files, from the collapsed strands, the number of mapped positions within each consecutive 1Mb chunk of the human genome is counted. In each chunk, the base-position count is normalized by the genomic count of the bases on both strands within the chunk. Both the raw and normalized counts can be plotted for any given genomic interval (here for chunks 0-249, corresponding to chromosome 1 of the human genome, and 150-200 for the zoom-in). Disallowed chunks (high coverage (at least 40%) by problematic regions, defined in 3.3.1.2.5) are marked in light gray on the plot, if present. Later on, the chromosome ideogram generated on the NCBI Genome Decoration Page (`https://www.ncbi.nlm.nih.gov/genome/tools/gdp`) was added manually.

3.2.2. Other analyses

Analyses included in this section do not form part of the integrated AB-seq analysis pipeline. All code and version requirements can be accessed in the `GitHub` repository at `https://github.com/bbglab/ABseq-analyses`. (Note that at the time of writing the thesis, as this work remains unpublished, so does the repository. Once we generate all planned maps, this code will be published alongside them, and will be made fully open and available. In the meantime, the repository can be viewed upon request.)

3.2.2.1. Similarity of context frequencies and landscapes

The genome-normalized trinucleotide context frequencies are analyzed with a custom `R` script (adapted from Maria Andrianova, a postdoc in the lab), which includes cosine similarity calculation, PCA analysis, and visualization. All analyzes are performed on the matrix with rows representing samples, columns the triplets, and values the genome-normalized frequencies.

Cosine similarity of the context frequency profiles is calculated using the `cosine` function from the `lsa` package and plotted using `pheatmap` library. PCA is performed using the `prcomp` function. Plots are generated using `ggplot2`, for the first two principal components: scores for the samples with noted the variance explained, and loadings of the triplet contexts in the principal component.

We also performed variants of these analyses for two different conditions. First, for comparison with published MMS damage data [24], only G and A-centric contexts were available, so we filtered, recalculated, plotted the frequencies, and performed cosine similarity and PCA for both yeast and the human AB-seq data using only those contexts. In the second case, the comparison with the COSMIC mutational signature SBS11 (GRCh37 version, [65, 69, 70]), the mutational frequency values of 96 pyrimidine-centered channels were summed into corresponding triplets (e.g. A[C>A]A, A[C>G]A and A[C>T]A become 'ACA+TGT' triplet). The AB-seq TMZ data was then represented in a similar format, summing the two triplet frequencies together (so ACA and TGT become 'ACA+TGT'), and this format was used for the cosine similarity and PCA analyses.

Finally, in an analysis of genomic damage distribution landscapes, instead of cosine similarity we applied correlation. We gathered total corrected counts of mapped sites in either 1Mb or 100kb chunks along the whole genome in each sample into a vector. Vector correlations were calculated using the `cor` function, and plotted using `pheatmap` library.

3.2.2.2. Representation of LC-MS/MS results

For each experiment replicate, we obtained picomoles of the modifications (m7G, m3A) in the thermal fraction and of the non-modified DNA nucleosides (dG, dA) in the enzymatic fraction. Each of the picomol values was divided by the injected percentage of its corresponding fraction. These corrected values were then used to calculate the ratios of modification to the non-modified nucleoside (m7G/dG and m3A/dA) and finally transformed into percentages.

3.3. Repair states framework for UV DNA damage

Here we detail all technical aspects of our bioinformatics framework for segmentation of the genome into UV repair states, characterized by differences in dynamics of NER repair. The work described in this subchapter was performed in close collaboration with Ferran Muiños and is based on the work of Joan Frigola. All the steps outlined below are contained in code repositories with automated snakemake and nextflow pipelines for processing the data, as well as python scripts, jupyter notebooks, and a singularity distribution. The current versions of the code and package requirements can be accessed in the following GitHub repositories:

Preprocessing:
`https://github.com/bbglab/repair-states-preprocess`
Repair states learning:
`https://github.com/bbglab/repair-states-hdphmm`
Downstream analyses:
`https://github.com/bbglab/repair-states-analyses`

(Note that at the time of writing the thesis, as this work remains unpublished, so does the repository. Once the project is published, the updated code will be published alongside it, fully open and available. In the meantime, the repository can be viewed upon request.)

First, I describe the needed processing of input damage and repair data that will be used for the segmentation. Next, I outline the modeling steps needed to obtain that segmentation. After that, I explain the steps of postprocessing of the obtained repair states segmentation, and, finally, the analyses annotating various datasets on top of that segmentation.

3.3.1. Input data preprocessing

All of the preprocessing steps are included in the `https://github.com/bbglab/repair-states-preprocess` data repository. Steps are grouped into 4 automated `snakemake` pipelines. In this section, I describe the steps alongside examples of commands included in the pipelines.

3.3.1.1. Chunking the genome

Genomic chunks along the whole genome were generated using `bedtools`, according to the hg19 chromosome lengths reference file with only canonical chromosomes, for each given chunk size:

```
bedtools makewindows -g human.hg19.genome -w 100000 >
↪  chromosomes_100kb_division.txt
```

3.3.1.2. Problematic regions

We gathered three sources of potential confounders for sequencing including regions we consider low-trust. We call this set 'problematic regions'. The three categories include low-complexity regions, blacklisted regions, and regions with mappability issues.

3.3.1.2.1. Low complexity regions

Regions of low complexity (many repeats, or consisting mostly of tracts of one base) are problematic for sequencing and aligning to the genome. To download low complexity regions we accessed the UCSC table browser at `https://genome.ucsc.edu/cgi-bin/hgTables` and changed the following terms: 1) assembly: *Feb. 2009 (GRCh37/hg19)*, 2) group: *Repeats*, 3) track: *RepeatMasker*, 4) table: *rmsk*, 5) region: *genome*, 6) filter: click on *create*, on the new page paste *Low_complexity* after *"repClass does match"*, and click *submit*, 7) output format: BED - browser extensible data, 8) output filename: paste *hg19_low_complexity_regions.gz*, 9) file type returned: *gzip compressed*.

Regions in non-canonical chromosomes were filtered out, and the remaining ones were sorted with `bedtools` according to the genome chromosome lengths reference file.

3.3.1.2.2. UCSC Excludable blacklisted regions

UCSC provides two blacklists of regions that can cause trouble for genome alignment tools. The first list – DAC – gathers regions identified as artifactual in multiple different tissues and cell types. The Duke list contains regions that have proven to be burdensome for short-sequence tag signal detection.

The excludable regions were downloaded from `http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeMapability/`, and noncanonical chromosomes were filtered out. Next, the regions from both lists were sorted together and finally merged into non-overlapping regions using `bedtools merge`.

The two blacklists are recommended to be used with a complementary mappability filter (see below).

3.3.1.2.3. UCSC 36mer alignability

Issues with mappability might mean that a region from the reference genome does not align to itself; or aligns to multiple different sites in the genome. The

information on the alignability of 36mers along the genome was downloaded from `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeCrgMapabilityAlign36mer.bigWig`, and exported to BedGraph format using `bigWigToBedGraph`. Any region without a score or a score different than 1 (signifying only one match) was discarded, and so were noncanonical chromosomes. The remaining regions (signifying great alignability) were complemented using `bedtools` to obtain alignability-problematic regions:

```
bedtools complement -i mappable.bed -g human.hg19.genome >
↪  unmappable.bed
```

Finally, the regions with mappability issues were sorted using `bedtools`.

3.3.1.2.4. Unified problematic regions and disallowed chunks

All three problematic lists were merged using `bedtools merge` into a single list of non-overlapping problematic regions, and next sorted.

To obtain the base-pair overlap of each genomic chunk of a given size with the problematic regions, we intersected the two datasets first:

```
bedtools intersect -wao -a chromosomes_100kb_division.txt -b
↪  problematic.bed -sorted -g human.hg19.genome
```

Next, we summed the problematic overlaps for each specific chunk. Each chunk with 40% or more problematic overlap is flagged as disallowed.

3.3.1.2.5. Disallowed chunks

Chunks that are considered disallowed from the analyses are chunks either highly covered by problematic regions (see above), or the ones that have a count of 0 pyrimidine pairs (defined right below) in the reference genome or a count of 0 of damage at any 0h time point of the damage data (processing in 3.3.1.4). All the disallowed chunks are zero-ed in analyses.

For the AB-seq damage distribution plots (described in 3.2.1.2.2), the definition of the disallowed chunks includes only the chunks highly covered by the problematic regions.

3.3.1.3. Pyrimidine pairs

For the set of the 3 pyrimidine pair contexts of interest (TT, CT, and TC) we performed a search for overlaps with the reference genome. To obtain the occurrences on the

other strand, we added the reverse complements of the three patterns to the search. The search for all 6 patterns was performed on the forward strand of the hg19 reference genome. All found sites with the strand and marked specific di-pyrimidine were sorted by the starting position and saved as a bed file.

Next, we wanted to filter the pyrimidine pairs by the problematic regions. The premise of the filtering lies in the idea of how the damaged sites are mapped - through the mapping of R1 reads from HS-damage-seq.

Hence, first, we simulated 50bp-long 'pseudo-reads' next to each pyrimidine pair, in the same orientation that we would obtain if the damage was mapped there using the HS-damage-seq protocol. We generated the pseudo-reads iteratively for each chromosome using `bedtools`. Afterward, we sorted the reads using the chromosome lengths reference file.

```
grep -P "^chr1\t" pyrimidine_pairs_positions.bed | bedtools
↪  flank -i stdin -g human.hg19.genome -l 0 -r 50 -s |
↪  bedtools sort -faidx human.hg19.genome
```

Next, iteratively for each chromosome, we calculated the total problematic regions coverage of each pseudo-read:

```
grep -P "^chr1\t" pyrimidine_pairs_pseudoreads.bed | bedtools
↪  coverage -a stdin -b problematic.bed -sorted -g
↪  human.hg19.genome | cut -f1,2,3,4,5,6,8,10
```

Finally, we extracted the positions of pyrimidine pairs back from the pseudo-reads and filtered them using a maximum threshold of 60% coverage (30bp) for the pseudo-read overlap of problematic regions. The resulting files were sorted again.

```
grep -P "^chr1\t" pyrimidine_pairs_pseudoread_overlaps.bed |
↪  bedtools flank -i stdin -g human.hg19.genome -l 2 -r 0 -s
↪  | awk '$NF < 0.6' | bedtools sort -faidx human.hg19.genome
↪  | cut -f1,2,3,4,6
```

After the filtering, the remaining pyrimidine pairs were chunked, one chromosome at a time, to obtain the sorted per-chunk count for each di-pyrimidine:

```
grep -P "^chr1\t" filtered_pyrimidine_pairs.bed | bedtools
 ↪  intersect -wao -a <(grep -P "^chr1\t"
 ↪  chromosomes_100kb_division.txt) -b stdin -sorted -g
 ↪  human.hg19.genome | awk -F '\t' '{{ print
 ↪  $1"\t"$2"\t"$3"\t"$7"\t"$NF }}' | awk -F '\t'
 ↪  '{{a[$1"\t"$2"\t"$3"\t"$4] += $NF}} END{{for (i in a)
 ↪  print i"\t"a[i]}}' | bedtools sort -faidx
 ↪  human.hg19.genome
```

The sorted, per-di-pyrimidine chunked count data was gathered into a coherent data frame, with one row per chunk, and filtered pyrimidine pair counts represented in the columns.

3.3.1.4. HS-damage-seq damage

The HS-damage-seq used in this thesis [19] was downloaded from GEO using the accession number GSE98025. Both replicates for specific conditions used in the thesis were extracted (CPDs at 0h, 1h, 8h, 24h, 48h; 6-4PPs at 0h, 20m, 1h, 2h, 4h) from the compressed directory for cell line NHF1 (Normal Healthy Fibroblasts). Note that we did not use the available 36h time point data for CPDs, as the authors [19] indicated a high similarity of the 36h and 48h profiles.

For each file, we obtained the positions of the damaged di-pyrimidines (located in positions 4-5 in the original files) using `bedtools`. Finally, the resulting file was sorted.

```
zcat damage_file.bed.gz | awk -v FS='\t' -v OFS='\t' '{{$4 =
 ↪  "0\t0\t" $4}} 1' | bedtools slop -i stdin -g
 ↪  human.hg19.genome -b -4 -s | bedtools sort -faidx
 ↪  human.hg19.genome | cut -f1,2,3,6
```

For each damage position, we extracted the corresponding di-pyrimidine sequence from the reference genome. We performed a 'sanity check' - checked the correspondence of the mapped positions to all available pyrimidine pairs in the reference genome. With this fitting properly we moved on to filtering of the damages. For this, we used the pseudo-read filtered pyrimidine pairs. The damage positions for each file passed this step only if they were found in a position of an allowed pyrimidine pair. This was performed in a per-chromosome manner.

```
bedtools intersect -f 1.0 -r -u -wa -a <(grep -P "^$chr1\t"
 ↪  damage_positoins.bed) -b <(grep -P "^$chr1\t"
 ↪  filtered_pyrimidine_pairs.bed) -sorted -g
 ↪  human.hg19.genome
```

Each filtered damage file was chunked, the count of damaged di-pyrimidine types was obtained for each chunk, and the chunks were sorted using `bedtools`.

```
bedtools intersect -wao -a chromosomes_100kb_division.txt -b
 ↪  filtered_damage.bed -sorted -g human.hg19.genome | awk -F
 ↪  '\t' '{{ print $1"\t"$2"\t"$3"\t"$7"\t"$NF }}' | awk -F
 ↪  '\t' '{{a[$1"\t"$2"\t"$3"\t"$4] += $NF}} END{{for (i in a)
 ↪  print i"\t"a[i]}}' | bedtools sort -faidx
 ↪  human.hg19.genome
```

The chunked counts from files that constituted 2 replicates of the same condition (i.e. CPDs at 0h, replicates A and B) were summed. All conditions corresponding to a given damage type were gathered in a data frame, where each row represented a chunk, and columns the counts of di-pyrimidines in all conditions. As the used cell line was of male origin, we doubled the counts corresponding to X and Y chromosomes to make the damage counts across sex chromosomes comparable to that across autosomes.

At this stage, we generated the disallowed flags for each chunk (see 3.3.1.2.5) using the 0h damage, pyrimidine pair counts, and problematic regions. In the next steps, all the chunks with the disallowed flag were zero-ed.

3.3.1.4.1. Damage counts normalization

We set out to reduce the impact of the genome composition, and the experimental saturation that HS-damage-seq tends to suffer from (see 1.3.2.1.1) on the damage counts. First, we added 1 pseudo-count to the filtered damage count and filtered pyrimidine pair count of the chunk. Next, we divided the filtered damage counts of a specific di-pyrimidine in each condition by the corresponding filtered pyrimidine pair count in the chunk. Finally, we divided the genome-normalized damage counts by the total sum of all counts in this condition. This constitutes what we further refer to as 'normalized damage score', and can be also thought of as genome-normalized relative damage counts.

3.3.1.4.2. Inferring total repair from normalized damage

To obtain a measure of total repair happening in a specific time interval, we focused on the normalized damage scores between any two consecutive time points. For each damage type and di-pyrimidine, in each chunk, we calculated the divergence of the scores between the start and end of each interval. Divergence is computed as a log fold change of normalized damage scores in the start time point divided by those at the end time point. It is given by the following equation: $Div(t_0, t_1) = \log d(t_0) - \log d(t_1)$, where t_0 represents the start time of the interval, t_1 represents the endpoint of the

interval, and $d(t)$ represents the normalized damage score at time t. Next, we refer to the normalized damage scores' divergences as Inferred Repair.

3.3.1.5. XR-seq repair

For inferred repair, see above. This subsection is focused on snapshot repair.

The XR-seq data from [23] was recovered from GEO with the accession number GSE76391. The files for conditions of interest for NHF1 cells were collected (CPDs at 1h, 4h, 8h, 16h, 24h, 48h; 6-4PPs at 5m, 20m, 1h, 2h, 4h) from the compressed directory. Each condition was separated into two replicates (1 and 2) and by the counts on a given strand. We sorted each of these subfiles using `bedtools` and the chromosome lengths definition file. Next, each subfile was intersected with the genomic chunks using `bedtools`.

```
bedtools intersect -wao -a chromosomes_100kb_division.txt -b
 ↪  xr_file.bed -sorted -g human.hg19.genome | cut -f1,2,3,7,8
```

The total repair score of a chunk was calculated by multiplying the 25bp density score (provided in the original file) of a given XR-seq annotated fragment by its total overlap with the chunk and dividing by 25, and then summing over all fragments overlapping the chunk. These total repair scores were summed over the files that constituted 2 replicates and separate strands of the same condition (i.e. CPDs at 0h, replicates 1 and 2, strands MINUS and PLUS). Finally, all conditions were gathered in a data frame so that a row represents a chunk, and each column has the repair score of each condition. As the repair density was already normalized by chromosome coverage, we did not need to double the sexual chromosome scores. As the XR-seq density scores capture the NER excised fragments, whose half-life is estimated to be around 10m (see 1.1.2.1.3 and 1.3.2.2.1), they represent a detailed picture of repair at a specific moment rather than total repair along a longer time interval. Hence, we further refer to XR-seq repair scores as Snapshot Repair.

3.3.2. Correlation of damage and mutations

3.3.2.1. Processing of mutations

We retrieved the 'simple_somatic_mutation' files from ICGC [79, 31] for two skin cancer datasets: SKCA-BR (Skin Adenocarcinoma Brazil) and MELA-AU (Skin Cancer Australia). The mutation files were filtered to select only single base substitutions on canonical chromosomes. The records were deduplicated, and sorted with `bedtools`.

```
cat simple_somatic_mutation.SKCA-BR.tsv | awk -F '\t'
 ↪  '$14=="single base substitution"' | awk -F '\t' '{{print
 ↪  "chr"$9"\t"$10"\t"$16"\t"$17"\t"$8 }}' | sort | uniq | awk
 ↪  -v FS='\t' -v OFS='\t' '{{$2 = $2-1 "\t" $2}} 1'| sed
 ↪  's/^chrMT/chrM/g' | input.bedtools sort -faidx
 ↪  human.hg19.genome
```

We were interested in samples that exhibited signs of high UV mutational activity. We calculated the percentage of the fraction of mutations in each sample that happens to be in the most frequent UV mutational contexts (C>T and G>A). We plotted the sample fractions for both datasets (Figures 3.1 and 3.2) and decided on the threshold of a minimum of 70%. The samples not meeting this threshold were discarded.

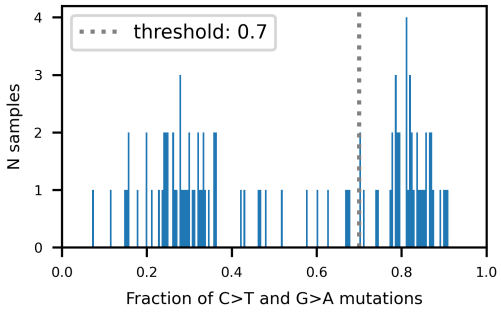

Figure 3.1: *Histogram of SKCA-BR samples representing the fractions of potential UV mutations (C>T and G>A) that they carry.*

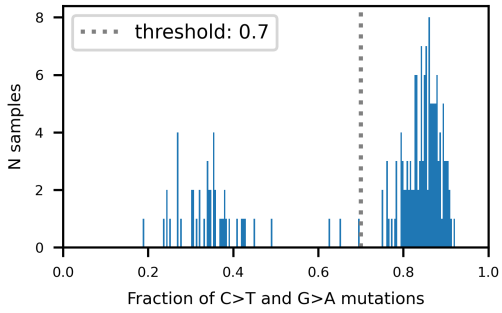

Figure 3.2: *Histogram of MELA-AU samples representing the fractions of potential UV mutations (C>T and G>A) that they carry.*

Next, from the high-UV exposed samples we obtained just the mutations within the di-pyrimidine contexts same as our damage data. This included XTT, TTX, XCT, CTX, XTC, and TCX (where the middle nucleotide is the reference allele of the base substitution), as well as their reverse complements. The mutations with the triplets were sorted according to the chromosome lengths reference file.

We chunked and filtered each of the filtered mutation files using `bedtools`, counting the number of mutations in each context per chunk.

```
bedtools intersect -wao -a chromosomes_100kb_division.txt -b
 ↪  UV_context_mutations.SKCA-BR.bed -sorted -g
 ↪  human.hg19.genome | awk -F '\t' '{{ print
 ↪  $1"\t"$2"\t"$3"\t"$7"\t"$NF }}' | awk -F '\t'
 ↪  '{{a[$1"\t"$2"\t"$3"\t"$4] += $NF}} END{{for (i in a)
 ↪  print i"\t"a[i]}}' | bedtools sort -faidx
 ↪  chromosomes_100kb_division.txt
```

Finally, we gathered the mutations from both datasets and summed their counts in

chunks. Apart from the context-specific counts, we calculated the count of all UV mutations found per chunk and made sure to sort the file.

3.3.2.2. Plots aligning the damage, repair, and mutations

To visualize a small part of the data along a fraction of the genome (Figure 4.21), we gathered the following 100kb-chunked data tracks: filtered TT pyrimidine pairs, filtered non-normalized CPD TT damage counts at 0h, normalized CPD TT damage scores at 0h and 48h, TT-corrected snapshot repair score at 1h, and UV-context mutations. All the data tracks were zero-ed in disallowed chunks.

We plotted the tracks along a 90-100Mb part of the human chromosome 1, which did not have at this resolution any disallowed chunks. The normalized damage score tracks were represented within the same y-axis minimum and maximum bounds for easier comparison.

The data explainers in the A part of the figure, as well as the chromosome ideogram in B, were both added manually. Chromosome 1 ideogram was generated on the NCBI Genome Decoration Page (`https://www.ncbi.nlm.nih.gov/genome/tools/gdp`)

3.3.2.3. Correlations calculations

For each damage type, we calculated two different correlations: 1) between damage at 0h and damages at the following time points, 2) between UV-context mutations and the damage at the earliest and latest time point. These correlations were calculated for both non-normalized damage, and the normalized damage scores, at 100kb resolution.

First, we calculated the $\log_{10}$ of both variables (i.e. mutations and damage at 0h). The non-finite values were masked. Using the `scipy.stats.linregress` function, we calculated the Pearson correlation (R) coefficient between each two variables.

The damage-damage correlations were next saved as a table (presented in Appendix, Tables B.1 and B.2). The mutation-damage correlations were represented on KDE density plots, together with the space covered by the two $\log_{10}$ variables in question.

3.3.2.4. Subsampling correlations

To explore the correlations of the UV-context mutations with all the damage time points and how they vary along the 100kb-chunked genome, we performed sub-sampling correlations.

First, we $\log_{10}$-transformed both variables of interest and masked non-finite values. We

bound the values of the two variables in the same chunk together. Next, we shuffled the chunks. The shuffled chunked genome was then split into sub-samples of size 40. The smallest window, below the size of 40, was discarded.

In each window, we used the `scipy.stats.linregress` function to calculate the R correlation coefficient between the two sub-sampled variables. We plotted the correlation coefficients for all windows as boxplots, one boxplot per one damage time point. This was performed for all damage types, and both non-normalized as well as normalized damage scores.

3.3.3. Sticky HDP HMM-based repair states partitioning

Steps of the HDPHMM mentioned in the section are included in the following `GitHub` data repository: `https://github.com/bbglab/repair-states-hdphmm`. The code can be executed using `Nextflow (v21.04)` pipeline with the `Python` interpreter in a `Singularity (v3.7.3)` container where all the software dependencies have been resolved. A few steps including the clustering of the hidden states into repair states, reproducibility analyses, and graphical representation of the framework are instead available in the `https://github.com/bbglab/repair-states-analyses` repository.

3.3.3.1. Modeling assumptions

In our framework we conceptualize the repair states as intrinsic states of genomic chunks which determine characteristic repair dynamics. We assume that the generative process that links a repair state with the observable repair dynamics is stochastic, meaning that each repair state can result in a range of observable repair dynamics, although with some characteristic propensities that allow us to tell apart one state from another.

In order to carry out the modeling, we need first a way to encode the observable repair dynamics using the inferred repair and snapshot repair data per chunk. We assume that the observable repair dynamics have been produced in accordance with the following generative process: each genomic chunk is associated with a hidden state (unobservable latent variable of the model) that determines the distribution that the observable repair dynamics have been randomly sampled from. We also assume that hidden states have some level of persistence, meaning that a chunk is more likely to be in a hidden state if the adjacent chunks are in the same hidden state. This sort of persistence is convenient when fitting a model on sequential data, as it primes the model to come up with parsimonious solutions that are often in accordance

with biology: in our case, since the properties of genomic chunks are not expected to be independent when the chunks are in contact or close proximity, we deem this persistence an adequate modeling assumption.

3.3.3.2. Rationale behind sticky HDP HMM as a method of choice

A common choice in the field for this type of modeling is to employ Markov processes with a specified number of hidden states (Hidden Markov Model or HMM). There are well-known, efficient methods to fit these models. However, with the development of efficient learning methods for more complex probabilistic graphical modeling (e.g. variational inference) alternative methods have become a good practical choice.

Here we adopt the so-called sticky Hierarchical Dirichlet Process Hidden Markov Model (sticky HDP HMM). This method has several conceptual and practical advantages compared with the classical HMM approach:

- It can automatically learn an optimal number of hidden states.
- It is compatible with several distributional assumptions for the observable emissions.
- It is conceptually very flexible, allowing fine-tuning e.g. the degree of persistence of the hidden states.
- There are efficient implementations of learning methods already packaged as ready-to-use `Python` APIs, like in the package "`bnpy`: Bayesian nonparametric machine learning for python", the API of choice in our analysis (`https://github.com/bnpy/bnpy`).

3.3.3.3. Encoding of observable repair dynamics

For modeling we encoded the observable repair dynamics from both sources (inferred and snapshot) corresponding to each chunk the following way. We computed the 5-quantile (herein "quantiles") of the chunk relative to all the chunks reported for the given repair track. Quantile 0th henceforth gathers chunks with the lowest values (representing the lowest repair scores) in the given data track and 4th with the highest.

For inferred repair, we obtained quantile encodings for each time interval, each damage type, and each di-pyrimidine context, totaling 16 tracks. For snapshot repair, we computed the quantiles for each time point and type of damage, amounting to 11 tracks. This way, the data emitted by each hidden state comes at a total of 27 tracks with quantile values. We further binarized the quantile data so that each preceding track spans 5 binary channels, giving a total of 135 binary channels. For the disallowed chunks, the observable repair dynamics are encoded as the zero vector.

3.3.3.4. Dirichlet Process

In order to specify the generative process whereby hidden states are related to one another that allows us to carry out an effective learning of the optimal number of clusters, we need to introduce the Dirichlet Process. The Dirichlet Process (DP) can be formally defined as a distribution on probability measures on a measurable space Θ, uniquely defined by a base probability measure H and a concentration parameter $\gamma > 0$ and characterized by the following property: if $G \sim DP(\gamma, H)$ then for any finite partition $A_1, ..., A_K$ of Θ, $(G(A_1), ..., G(A_K))$ are distributed as a Dirichlet distribution with parameters $(\gamma H(A_1), ..., \gamma H(A_K))$.

However technically sound, the DP can be more explicitly understood by using a more constructive definition based on the so-called stick-breaking construction. This is accomplished in a three-step process:

1. First an infinite countable partition of the unit interval $\{\beta_k\}_{k\geq 1}$ is randomly generated by sampling $v_k \sim Beta(1, \gamma)$ and combining them into the weights $\beta_k = v_k(1-v_{k-1})(1-v_{k-2})...(1-v_1)$ for each $k \geq 1$. This is also known as the Griffiths-Engen-McCloskey (GEM) or stick-breaking process and it is usually denoted as $\beta \sim GEM(\gamma)$.

2. Then for each component k we draw $\Theta_k \sim H$ from the base distribution H.

3. With these ingredients we can then assemble a discrete distribution on the countable set $\{\Theta_k\}_{k\geq 1}$ with probabilities β_k for each k.

Intuitively, randomly sampling from a DP always produces countable discrete distribution such that: 1) each possible value that the distribution can take (atom) is sampled from the base distribution H and 2) the variance of the probability mass of the atoms depends on the concentration parameter γ.

3.3.3.5. Generative model

We will denote y_t and Z_t the observable repair dynamics and the hidden state, respectively, corresponding to chunk t (Figure 3.3). The observable repair dynamics y_t takes a binary 135-vector as value, whilst the hidden state Z_t takes a positive integer $k \geq 1$ as value. We denote by Θ_k the vector of parameters governing the emission distribution corresponding to hidden state k, i.e. if $Z_t = k$ then $y_t \sim F(\Theta_k)$, where F is the joint probability mass function corresponding to the product of 135 independent Bernoulli random variables, each component representing one of the encoding tracks of the observable repair dynamics.

To model the transitions between states in adjacent chunks we need to introduce the distributions π_k defined for each possible hidden state: these distributions encode the probabilities to transition from state k to any other possible state. In more concrete terms, the hidden state in chunk $t+1$ is assumed to be randomly sampled from π_{Z_t}. Note that since π_k are countable, discrete probability mass functions, they can be realized as random draws from a DP. In the generative modeling approach adopted here, each π_k is randomly drawn from a DP with base distribution H_k and concentration parameter γ_k, where: $H_k = \frac{1}{\alpha+\kappa}(\alpha\beta + \kappa\delta(k))$, $\gamma_k = \alpha + \kappa$, β being the a priori distribution on the hidden states randomly drawn from $GEM(\gamma)$, $\delta(k)$ is the delta-distribution giving probability mass of 1 to k, and κ (Greek letter kappa) is the self-transition bias (governing how sticky transitions between hidden states are) and α is an additional concentration parameter [80].

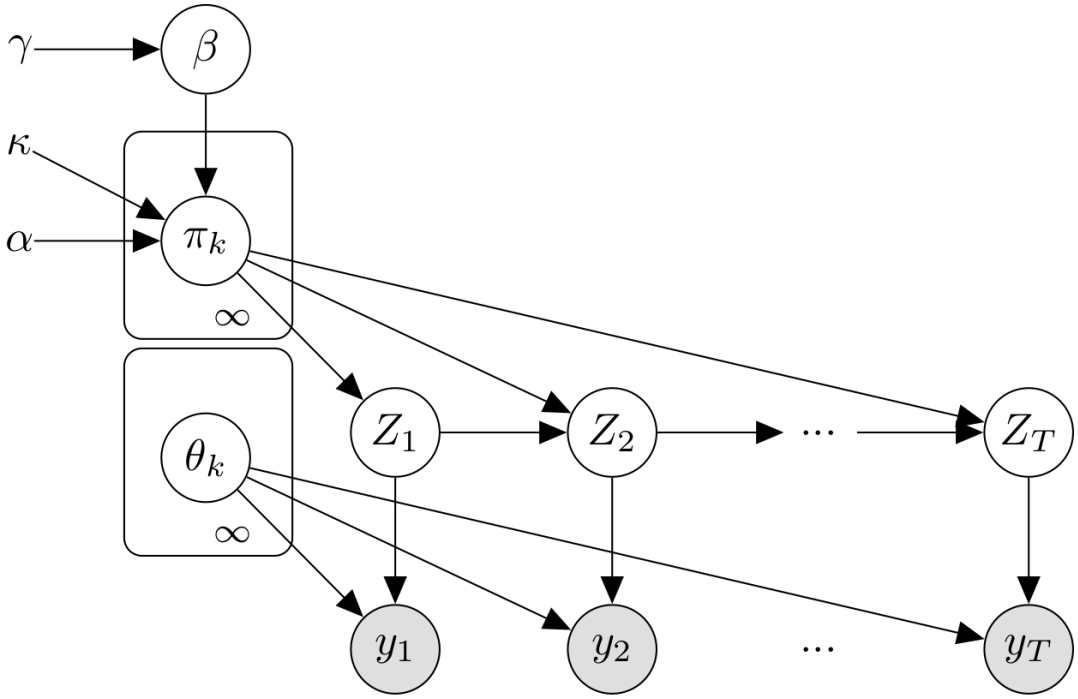

Figure 3.3: *Plate diagram representing the sticky HDP HMM model. Top-left: stick-breaking process generating the β prior used for modeling the transition probabilities between hidden states. Middle-left: generative process for the transition probability vectors based on the stickiness parameter κ (kappa), the concentration parameter α, and the β prior. Right: the observable repair activity per chunk y is randomly emitted from a distribution that depends on the hidden state and the parameters governing each hidden state; the hidden states Z are randomly sampled from the transition probability vector π corresponding to the preceding state in the sequence of chunks.*

3.3.3.6. Learning

The learning is carried out in two steps: mean-field variational inference and hidden state allocation to genomic chunks.

3.3.3.6.1. Mean-field variational inference

Variational (Bayesian) inference stands for the learning strategy whereby the full intractable posterior distribution on the latent factors Z is approximated with a tractable distribution $q(Z)$ known as the variational distribution. In this setting, the set of parameters governing $q(Z)$ is fitted via an optimization strategy that is equivalent to the minimization of the Kullback-Leibler divergence between the true, intractable posterior and the variational distribution. In mean-field variational inference, the variational distribution $q(Z)$ is assumed to be fully factorized over some partition of the latent factors.

To learn the set of hidden states, their transition probabilities, and their respective emission parameter vectors, we resorted to a mean-field variational inference algorithm, known as "memoized online variational inference" that is already implemented as a `Python` API [81].

3.3.3.6.2. Default configuration of the mean-field variational algorithm

The main hyperparameters for the variational inference step are:

- γ governing the stick-breaking leading to the β prior; set at 5 across all runs;
- α governing the variance of the transition probability vectors of each state; set at 0.5 across all runs;
- κ governing the self-transition bias between hidden states (stickiness) was set at 100 for the main run (but we carried out reproducibility analyses for the values 200 and 300 as well).

Other technical configuration values worth mentioning are:

- The initial number of hidden states that are subsequently pruned by the variational inference step was set to $N = 50$.
- Number of batches. Before execution, each sequence of observable binary vectors is randomly assigned to one batch, then the variational inference algorithm visits batches one at a time in random order. The number of batches was set to $B = 10$.
- Number of laps. Each full pass through the complete set of batches is a lap. The number of laps was set to $L = 300$.
- Moves. The optimization strategy to escape local optima in the gradient ascent process was set to "merge-shuffle".

3.3.3.6.3. Viterbi's algorithm

Using the outputs of the variational inference step, we can then run Viterbi's algorithm to carry out a final allocation step of hidden states to each genomic chunk. Viterbi's algorithm is a dynamic programming algorithm that finds the most likely sequence of hidden states Z resulting in the sequence of observable emissions y by using the learned transition probabilities between hidden states and the emission multivariate Bernoulli distributions corresponding to each hidden state. The output of this step gives us the map of hidden states across genomic chunks.

3.3.3.6.4. Python API

The mean-field variational inference and Viterbi's algorithm steps were run using the "Bayesian nonparametric machine learning for python" or "`bnpy`" package: `https://github.com/bnpy/bnpy`.

3.3.3.7. Hidden states' hierarchical structure

The hidden states learned may in part reflect heterogeneity that is not biological, rather technical, explained by e.g. data sparsity, different repair activity mixtures being included in the same genomic chunks or by means of the learning algorithm itself. With that regard, we deemed the number of hidden states that we learned a good upper bound of the true number of repair states. Motivated by this, we aimed to discover the hierarchical structure of the hidden states by clustering them into a lower number of states based on the similarity of their repair activity.

First, we investigated the disallowed chunks within the hidden states. All disallowed chunks in all runs were grouped together in one hidden state that did not cover any other chunks. We renamed this state as 'disallowed' and excluded it from most analyses, including clustering of the hidden states.

Upon allocation of hidden states across genomic chunks, we stacked the divergence scores derived from the inferred repair and snapshot repair data across genomic chunks belonging to the same hidden state. As a result, for each hidden state, we derived a vector of mean divergence scores across the inferred repair data tracks, and mean of ranks across the snapshot repair data tracks. Using these vectors we computed the hierarchical agglomerative clustering based on the mean divergence score vectors.

Agglomerative hierarchical clustering method: Unweighted Pair Group Method with Arithmetic Mean (UPGMA) with Euclidean distance (`https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html#scipy.cluster.hierarchy.linkage`).

From the topological information of the resulting hierarchy, we could compute different flat clusterings of hidden states, one for each number of flat clusters we wanted to consider (granularity). Flat clusters were assigned using the `fcluster` function from `scipy.cluster.hierarchy`.

3.3.3.7.1. Clustering the hidden states into repair states

We found the high number of hidden states allocated by the model to be difficult to interpret. Hence we used agglomerative clustering to cluster the hidden states into an interpretable number of final 'repair states'. We performed the clustering for granularities varying between 2 and the number of hidden states learned in each case. Next, we needed to decide on a number of flat clusters that represent most coherently the substructure of the hidden states.

For each of the granularities, we inspected the matrix of the means of scores representing the vectors used for the clustering, together with the resulting clustering dendrogram, and the hidden states colored by the assigned flat clusters. We looked for the highest possible granularity level where different flat clusters would continue to reflect distinct repair activity profiles as per the vectors of mean scores. We also verified if the learned transition probabilities between hidden states were compatible with this clustering, checking the coherence of frequent within-cluster transitions. Based on these two criteria, we decided on a final set of 12 repair states.

We saved the dendrogram, the ordering of the clustered hidden states, and the clusters corresponding to repair states for use in further analyses. Note that the 12 repair states everywhere in this thesis are already represented as re-numbered and ordered by average mutation count (more in 3.3.4.1). Both the matrix with the dendrogram and the transitions were plotted with `seaborn clustermap`.

3.3.3.8. Graphical representation of the framework

To plot the figure representing the repair states partitioning framework step by step along a small genome part (Figure 4.23), we gathered a set of 100kb-chunked data tracks from 93rd to 98th Mb of chromosome 1 (marked in orange, did not have any disallowed chunks). The tracks included: normalized CPD TT damage scores at 8h and 24h, inferred CPD TT repair between 8-24h, snapshot repair at 1h, and quantile encodings of the two aforementioned repair tracks.

For inferred and snapshot repair, on the right of the plotted track fragments, we represented the distribution of all the scores for this track. Horizontal lines represent the cut-offs of the five quantiles within this distribution.

Below the plotted fragment, we added the whole chromosome 1 colored by the hidden states and repair states assignment. Above the plotted fragment, we added the chromosome ideogram, generated on the NCBI Genome Decoration Page (`https://www.ncbi.nlm.nih.gov/genome/tools/gdp`).

3.3.3.9. Reproducibility analysis

In order to validate our approach and assess the amount of technical variability that our clustering method might introduce in the results, we set out to run the clustering method across several configurations and using different input datasets.

We carried out two types of comparisons:

- Compare runs fit across different stickiness levels (κ hyperparameter value)
- Compare runs fit using full data (all replicates combined) with partial data (single replicates)

To assess the concordance between the different resulting hidden state partitions of the genome, we devised a method that checks the consistency not just at the level of inferred hidden states, but also across the entire hierarchical structure underlying the repair dynamics.

3.3.3.9.1. Contingency tables

Given two flat clusterings of hidden states with the same granularity (N) produced by separate runs of our clustering method (run1 and run2), we can create an $N \times N$ contingency table that provides the counts of genomic chunks mapping to each pair i, j of flat clusters, belonging to run1 and run2 flat clustering indices, respectively.

Note that the indices in the flat clustering need not be compatible with each other. We can then relabel indices in both flat clusterings to render the trace (sum of the diagonal entries) of the contingency table as high as possible. Such a reindexing provides the most likely compatible indexings relative to the chosen granularity level N. The higher the counts in the diagonal, the stronger the evidence that both flat clusterings are consistent.

Given two hidden state allocations with their respective clusterings based on mean divergence observable repair, and a granularity level N, we can build the reindexed contingency table and score the proportion of counts in the diagonal via a generalized MCC score.

3.3.3.9.2. Generalized MCC score

Given an $N \times N$ contingency table with compatible indexes and a query flat cluster index F we can render a 2×2 table by collapsing all the flat clusters different from F into one class. With such a contingency table we can then compute an MCC score, which measures how consistently F-labeled chunks are allocated to the same F class of hidden states. We can proceed with all possible flat cluster indices, computing the MCC of the corresponding collapsed 2×2 contingency table, then averaging out all the MCC values into what we call a "generalized MCC score".

This method provides a valuable means to test at which granularities the highest reproducibility levels are reached. In other words, we want to assess with which granularities clustered states correspond well to one another when allocated using slightly perturbed data and learning configurations. With this information, we can then go about defining informed criteria as to what amount of the inferred and snapshot repair heterogeneity is biological rather than technical and apply yet another post-processing step to define prototypical repair state families. The rationale would be the following: we would like the repair states to be as granular as possible, but also as reproducible as possible across perturbations of the standard clustering algorithm.

3.3.3.9.3. Reproducibility across replicates

To obtain the input inferred and snapshot repair for replicates, we simply processed the HS-damage-seq and XR-seq data in the same way, only without the step of summing over the replicates. Next, we matched the inferred repair from A replicate of HS-damage-seq with the snapshot repair of replicate 1 of XR-seq and did the same for B and 2.

We ran the repair state clustering method across 3 chunk-size configurations (1Mb, 100kb, and 10kb) with the input repair divergence scores derived from the following settings (rest of configuration parameters kept as in the default configuration):

- Standard run:
  - full HS-damage-seq (replicate A + replicate B)
  - full XR-seq (replicate 1 + replicate 2)
- Damage-seq replicate A and XR-seq replicate 1
- Damage-seq replicate B and XR-seq replicate 2

3.3.3.9.4. Reproducibility across stickiness levels

We ran the repair state clustering method across 3 chunk-size configurations (1Mb, 100kb, and 10kb) with the input repair divergence scores derived from the full data

(rest of configuration values kept as in the configuration by default) across 3 kappa stickiness levels: 100, 200 and 300.

3.3.4. Post-processing of repair states

All the code necessary for the analyses indicated below is available in `https://github.com/bbglab/repair-states-analyses` repository, and includes a mixture of `python` scripts and snakemake workflows.

3.3.4.1. Ordering and naming states by mutations

After obtaining the 12 final repair states, we needed a simple way to name and order them. To do that, we calculated the average count of UV-related mutations of chunks belonging to each state. We then sorted and numbered the repair states from the highest average count of mutations (RS1) to the lowest (RS12). We chose a color palette for the ordered states. Both the ordering and the palette are kept consistent throughout all the presented analyses.

3.3.4.2. Repair state logos

With the final repair states assigned, we set out to visualize their repair dynamics using the original input data tracks (inferred and snapshot repair). We aimed to represent the repair activity for each damage and di-pyrimidine pair (whenever available) along all intervals/time points.

To this end, for each repair state, we generated the following 'logo' of repair activity. The subplots (henceforth 'boxes') order corresponds to: first inferred CPD TT repair, second inferred CPD CT repair, then snapshot CPD repair, followed by two boxes of inferred repair for 6-4PPs (TT and TC), and finally snapshot 6-4PP repair. The last box in the logo represents mutations.

In each box, the values in the chunks assigned to a specific repair state are represented, along all the time intervals (or points). The values of specific chunks are plotted as scatters. The small, connected boxplots, each represent the median and 25th and 75th percentiles of the values for this repair state at a specific time.

In inferred repair boxes, we represented the divergence scores. For snapshot repair, we presented the ranks of TT-corrected repair scores (ranked along all genomic chunks and scaled from 0 to 1). In the mutations box, we plotted the ranks of the UV-related mutations.

The full logos of the repair states were plotted as rows, one underneath the other,

ordered from RS1 to RS12. The plots were arranged with `matplotlib`.

3.3.4.3. Comparison of orderings

We aimed to verify how the ordering of the repair states by average mutation counts corresponds to their repair activity. First, we generated the orderings based on each of the logo boxes (see above). For inferred repair, in each repair state, we summed the means of values in all intervals for the given damage type and di-pyrimidine. In the case of snapshot repair, in each repair state, we summed the means of ranks of the values for all time points for the given damage type. Each ordering was generated by sorting the states from the lowest sum to the highest.

The repair state orderings were compared with the mutations-based one using the Jaro-Winkler metric implemented as `jaro_winkler_metric` function in the `python jaro` package (`https://pypi.org/project/jaro-winkler/`). Importantly, this metric takes into account the transpositions, which is important for the order comparisons here. Jaro-Winkler metric in this implementation ranges from 0 to 1, with 0 being zero similarity, and 1 the exact same ordering.

3.3.4.4. Repair state transitions

To visualize how frequently one repair state 'changes' to another, we counted the times each 'origin' state transitioned to any other 'destination' one. Then, we normalized the transition count by the size of the destination state (in chunks). These normalized transition frequencies were plotted, indicating the origin repair state as 'from' and the destination state as 'to'. The plots were generated using `seaborn clustermap`.

3.3.5. Processing of genomic features

For all annotations, non-canonical chromosomes were discarded. All regions in the feature file were always sorted using `bedtools` and the chromosome lengths definition file, and merged, before overlapping with the 100kb genomic chunks.

3.3.5.1. RefSeq and LADs

The hg19 RefSeq annotations (including genes, exons, TSS, TES, and 2kb next to the TSS), as well as CpG Islands and LADs, were taken from the `COORDS` directory of the `ChromHMM tool` [82] distribution. The coordinates of RefSeq and CpG Islands provided were originally taken from the UCSC genome browser. LAD locations in human fibroblasts provided in this distribution are originally from [83]. Each of these annotations was summarized as a simple base-pair overlap of the feature with each 100kb chunk.

3.3.5.2. Chromatin states

We downloaded the chromatin states generated by the core 15-way chromHMM model (Table 3.1) for the E055 cell line [84, 85].

Label	Name	Description
TssA	Active TSS	Enriched in TSS of actively transcribed genes
TssAFlnk	Flanking Active TSS	Enriched in immediate neighborhood of TSS of actively transcribed genes
TxFlnk	Transcribed state at gene 5' and 3'	Enriched at 5' (downstream of TSS) and 3' (upstream of TES) of actively transcribed genes
Tx	Strong transcription	Enriched in gene bodies of transcribed genes
TxWk	Weak transcription	Enriched in gene bodies of transcribed genes
EnhG	Genic enhancers	Enriched in gene bodies of transcribed genes
Enh	Enhancers	Not enriched at TSSs
ZNF/Rpts	ZNF genes & repeats	Enriched for ZNF genes and satellite repeats
Het	Heterochromatin	Enriched at heterochromatin regions and centromeric and telomeric repeats
TssBiv	Bivalent/Poised TSS	Enriched in TSS of repressed genes
BivFlnk	Flanking Bivalent TSS/Enh	Enriched around TSS of repressed genes
EnhBiv	Bivalent Enhancer	Not enriched at TSSs
ReprPC	Repressed PolyComb	Enriched at gene bodies of repressed genes
ReprPCWk	Weak Repressed PolyComb	Enriched at gene bodies of repressed genes
Quies	Quiescent/Low	No marks

Table 3.1: *Chromatin State descriptions for the 15-way chromHMM core model. Source: Table S3 reference from [85]*

In the Roadmap Epigenomics Consortium [86, 85], E055 denotes a cell line corresponding to foreskin fibroblast primary cells (closest cell type to the one used to produce the XR-seq and HS-damage-seq maps). For each of the 15 chromatin states, we calculated their base-pair overlap with each 100kb chunk.

3.3.5.3. Genomic features mapped through CHIP-seq

The CHIP-seqs for histone marks were downloaded from two different datasets, and two slightly different cell types (specified below). Some of the histone mark types repeated in the two sets. We decided to include both and check for any differences. Apart from histone marks, we also downloaded the CTCF transcription factor CHIP-seq data.

We processed all CHIP-seqs in the same manner. Whenever possible, we chose the

file corresponding to peaks from 2 biological replicates. The peaks were filtered for the $-log_{10}$ q-values (qValue) threshold higher than $-log_{10}(0.05)$ and the enrichment (signalValue) threshold higher than 1.5. For all histone marks, we used the base-pair overlap of the merged, filtered peaks with each 100kb chunk.

3.3.5.3.1. Roadmap Epigenomics

We downloaded the hg19 histone marks available for the E055 (foreskin fibroblast primary cells, same one as chromatin states) cell line from Roadmap Epigenomics [86, 85] in the narrowPeak format. These marks included: DNase, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, and H3K9me3.

3.3.5.3.2. ENCODE

The other set of histone marks for the hg19 reference was downloaded from ENCODE [87, 88, 89, 90] for the GM23248 cells (arm skin primary fibroblasts). The retrieved marks were H2AFZ, H3F3A, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K79me2, H3K9ac, H3K9me2, and H4K20me1.

Files downloaded correspond to these identifiers in the ENCODE portal [88] (`https://www.encodeproject.org/`): ENCFF126IUT, ENCFF885LXS, ENCFF371OCE, ENCFF289CWH, ENCFF599FDC, ENCFF251CWC, ENCFF113GWR, ENCFF300JYE, ENCFF765TBF, ENCFF642UWE, ENCFF194AYL.

3.3.5.3.3. CTCF sites

Sites of bound CTCF in hg19 were downloaded from ENCODE for lower leg skin tissue. The corresponding ENCODE portal [88] identifier is ENCFF637FOR.

3.3.5.4. Expressed genes

3.3.5.4.1. Expression in fibroblasts

The expression (given by RPKM – Reads Per Kilobase of transcript, per Million mapped reads – values per gene) table in two fibroblast samples was downloaded from (Table S1 from [91]). We averaged the expression in each gene over the two samples by taking the median of their RPKM values. Next, for every chunk, we multiplied the base-pair overlap of the gene with the chunk by its median expression. The final score of the chunk was the sum of these scores over all genes overlapping the chunk.

3.3.5.4.2. GTEx Constitutively expressed genes

The median gene-level TPM (transcripts per million) expression across 54 non-tumor tissue types was downloaded from the GTEx Portal [92]. The dataset

includes two types of skin tissue: suprapubic non-sun-exposed skin (GTEx code "skin_not_sun_exposed_suprapubic") and lower leg sun-exposed-skin (GTEx code "skin_sun_exposed_lower_leg").

Downloadable file: `https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz`

For the two skin tissue types, we selected the genes with median TPM > 50 across the GTEx dataset as "constitutively active" in the tissue. From the two gene sets we then retrieved the hg19 genomic coordinates of all the exonic sequences (CDS, 5-UTR, 3-UTR from Gencode v31) mapping to the canonical transcripts of genes in the gene sets (Ensembl canonical transcripts v97) as a BED file.

Finally, the constitutively expressed genes, regardless of the sun exposure, were encoded in each chunk as a simple base-pair overlap for both lists.

3.3.5.5. Replication timing

BigWig Repli-seq files for BJ cell line (Foreskin Fibroblasts) each containing replication timing tags signal for a given phase were retrieved from UCSC (`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/`).

We exported each file to bed format using `bigWigToBedGraph`. In total, we obtained 6 data tracks corresponding to the 6 cell cycle phases: G1b, G2, S1-4. For each of the data tracks, we calculated the 75th percentile of the tag density values, and cut off all values below this threshold. This thresholded tracks were then encoded in each chunk as a simple base-pair overlap of the high tag density.

3.3.6. Analyses

3.3.6.1. Percent of genome covered by each DNA repair state

To explore the coverage of the whole genome by the repair states, we used the assignments including the disallowed state. With matploltib we plotted a stacked bar chart for the whole genome, with each stacked bar representing one state, and the percentage of the genome it covers. The same approach was repeated for all chromosomes, one chromosome at a time.

3.3.6.2. Contribution of NER pathways to DNA repair states

To elucidate the contributions of the two NER pathways (transcription-coupled and global) we used the XR-seq data for 3 cell lines [18]. Apart from healthy human fibroblasts (NHF1) the dataset also has data for two repair-mutant cell lines, one each in components of either of the two pathways. XPC refers to Xeroderma Pigmentosum cells, which are proficient for the transcription-coupled and deficient for global NER. CSB, referring to Cockayne Syndrome, is the opposite. This XR-seq dataset with 1h time point for the two mutant cell lines and the NHF1 cells was retrieved from GEO using GSE67941 accession number. The dataset was processed in the same manner as the rest of XR-seq to obtain chunked snapshot repair.

As one of the cell lines was of female origin (GM16095 - CSB), and others of male, we excluded the X and Y chromosomes from the further analysis. Disallowed chunks were excluded as well.

To compare the distribution of repair at 1h exhibited by the three cell lines in each of the repair states, for each of them we plotted `violinplots` of their snapshot repair (measured as TT-corrected repair scores) along all of the states. These plots were generated separately for each damage type, with `matplotlib`.

3.3.6.3. Feature enrichments

For each feature represented in the chunks (for details refer to 3.3.5) the process of calculating the enrichments was the same. We went state by state and calculated the following values:
A: the number of base pairs covered by the repair state,
B: the sum of the scores of the feature in all chunks,
C: the sum of the scores of the feature in chunks covered by the repair state,
D: number of bases in the genome (defined by summing the chromosome sizes from the reference file).

The fold enrichment of the repair state for a specific feature is then defined as $(C/A)/(B/D)$. Of note, the features were zero-ed in all the disallowed chunks.

Next, we set out to visualize the fold enrichments. For that, we grouped the features by similar categories (as represented in the Results). In each group, the features were reordered based on the correlation of their values. (Apart from replication timing features, that were ordered by the order of phases of the cell cycle).

We plotted the reordered features in each group against all repair states, representing

the fold enrichments on a `seaborn clustermap` as a heatmap. Importantly, we represented only enrichments (values above 1), and no depletions. The same representation was used in the junction analysis, but for hidden states instead of repair states, and with added the hidden state clustering dendrogram.

3.3.6.4. Junction analysis

Starting from the logo-based hierarchical clustering of hidden states (UPGMA with Euclidean distance) and the chromatin state enrichments (features) for each hidden state, we devised a way to regress the hidden states against the chromatin features. Our method yields, for each hidden state, an associated sequence of characteristic features with enrichment annotations.

First, from the logo-based hierarchical clustering we extracted the binary (rooted) tree representing its topology. This tree has hidden states as leaves. Each internal node of the tree (herein junction) splits the set of hidden state descendants into two sets of hidden states. For each chromatin state and each junction, we carried out a group comparison using the enrichment values by computing the difference of the means of each group. For each junction in the hierarchy topology, we selected the 3 features with the highest difference of means (characteristic features) if the difference was higher than 1.5.

Note that in this step many possible fitting approaches could be potentially applied to separate hidden state partitions by using the chromatin state features as explanatory variables.

Finally, since each hidden state is reached from the root of this hierarchy topology through a unique sequence of junction turns, we can pinpoint whether at each turn the path takes the enriched or depleted side relative to the characteristic features of the junction. Thus each hidden state can be described as a sequence – sorted from root to leaf – of characteristic feature sets with enrichment/depletion effect size annotations.

The junction analysis was also visually represented on a dendrogram plot from `scipy.cluster.hierarchy`, with the top features annotated at each junction, together with the differences. The direction of the difference is indicated by an arrow to avoid confusion.

4. Results

The main objective of this thesis is to gain more insight into the causes of variability of mutation rate along the human genome by searching for the mutual influences of DNA damage, repair, and the features of the genome that result in particular mutational patterns. We investigate different types of damage known to be a source of mutations: the maps generated in our group of alkylating agents, like MMS and TMZ (the second one commonly used in chemotherapy), and publicly available damage maps of UV light. The combined knowledge from both of those projects helps us disentangle the interactions of the (epi)genome, damage, and repair mechanisms that often lead to the creation of mutations in cancer patients' genomes.

4.1. AB-seq: maps of alkylating DNA Damage in human cells

To obtain a damage map, a unified analysis pipeline comprising both a lab protocol and a bioinformatics processing framework tailored to process the specific data produced are needed. To facilitate the comprehension of this chapter focused on my work developing the bioinformatics part, I also introduce the corresponding experimental part developed in the lab. This chapter focuses on describing the steps of the AB-seq (Alkylation BER sequencing) protocol in a high-level manner and the motivation for including them, as well as downstream analyses of the data. Detailed technical descriptions of all tools and parameters included in the pipeline can be found in the Methods chapter of the thesis.

4.1.1. Experimental in-house alkylation mapping protocol

While alkylation damage maps have been successfully produced in yeast cells [24], at the time of writing this thesis there is no publicly available nucleotide-precision map of alkylator-induced damage in human cells. To our knowledge, the preliminary results of the AB-seq method described here constitute the first map of this kind, for two alkylating agents, MMS and TMZ, in comparison with their corresponding controls: untreated DNA and DMSO-treated DNA.

The experimental part of the AB-seq protocol (Figure 4.1) is inspired by a previous one implemented to map MMS-induced methylation in yeast [24]. First, the DNA of cells

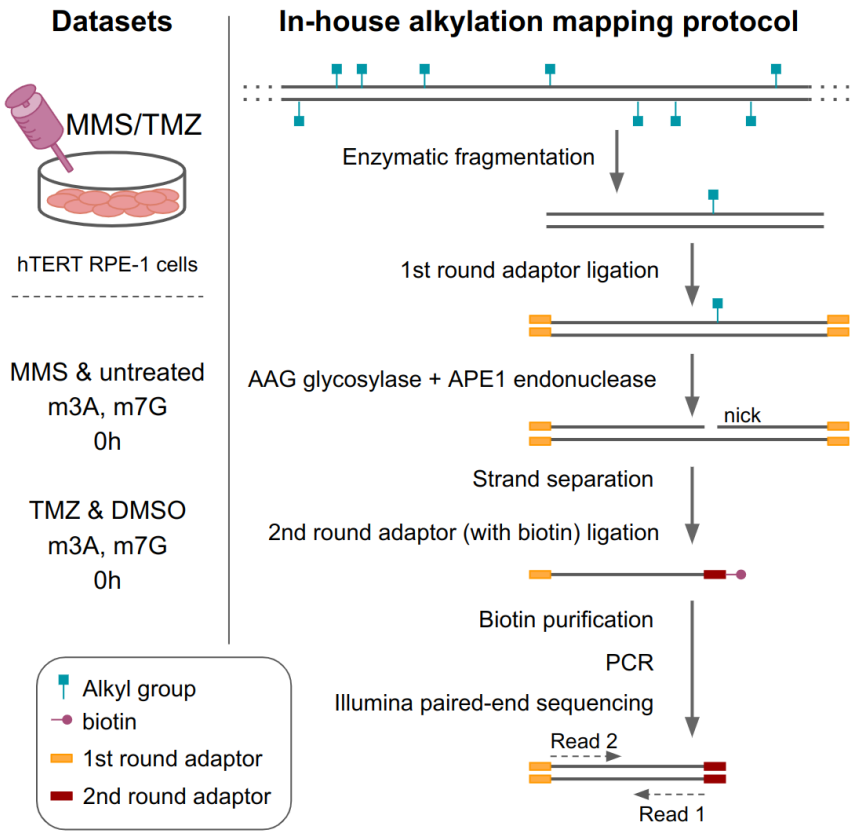

Figure 4.1: *Conditions tested and schematic representation of the experimental AB-seq protocol developed in the lab for the mapping of alkylation damage in human cells.*

subjected to damaging agent or control conditions is extracted and fragmented with a mix of restriction enzymes to shear the DNA into fragments of similar sizes and with no relevant sequence bias at the cut site. Double-stranded DNA fragments are then ligated to first-round adaptors, and a nick is introduced in the place of the lesion, by the successive action of a glycosylase and an endonuclease (AAG and APE1). Next, the strands are separated, and second-round, biotinylated adaptors are ligated to the free 3'OH ends that have been released by the endonuclease at the site of the excised base. After biotin purification, the obtained fragments are subject to PCR amplification and standard Illumina sequencing.

4.1.2. Computational in-house damage processing pipeline

The reads from the sequenced libraries need to be computationally processed to obtain the precise genomic positions of the alkylating lesions e.g. produce the actual damage maps. To this end, I have developed a highly-parallelized in-house processing pipeline (Figure 4.2).

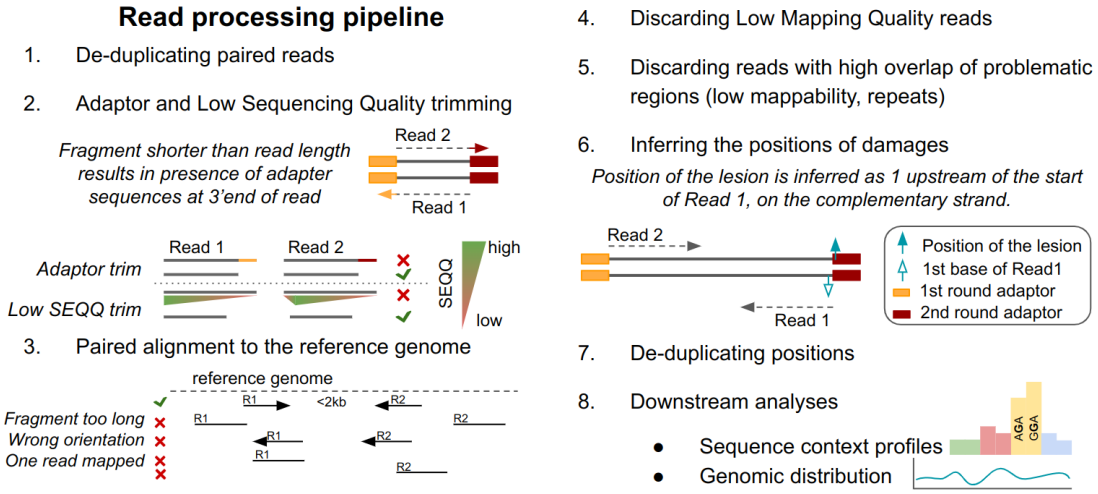

Figure 4.2: *Schematic of the computational steps of the AB-seq method developed in the lab for the mapping of alkylation damage in human cells.*

4.1.2.1. Read processing

Lists of reads sequenced from each sample are usually distributed across several files depending on the total sequencing output. These need to be merged in the proper order for both paired-end reads files. After that, the reads are deduplicated in a paired-wise manner (to discard only duplicates in both reads, highly likely to be PCR duplicates). Next, with one read1 and read2 file for each sample, each file is trimmed for low sequencing mapping quality bases (potential sequencing errors), and for specific adaptor sequences (e.g. for fragments shorter than the expected sequencing insert size, the adaptor sequence from the other side might show up at the end of the read). Trimmed reads that might be too short for the next step are discarded.

After trimming artificial and potentially erroneous sequences that might interfere with the process, the read1 and read2 sequences are aligned to the reference genome. In the pipeline, based on paired-end sequencing, only pairs of reads that map in the proper orientation and with the expected insert size given the previously described fragmentation approach are kept. After the alignment, the pairs of reads with low mapping quality (e.g. low probability of a proper match to the reference) are filtered out.

As the experimental part of AB-seq is designed so that the original location of the damage is immediately adjacent to the start of read1, the next steps utilize just read1 sequences. Those are first deduplicated. To ensure the high quality of mapped positions, the reads with a high overlap of regions problematic for their alignment (e.g. low mappability, highly repetitive, more in Methods) are discarded. The position

of the damage is inferred by taking the base immediately upstream of the end of read1 in the opposite strand. Positions are deduplicated. The counts of reads at all the processing steps are summarized in an automatically-generated table (Table 4.1) with the following columns for each sample:

- Paired Reads: the number of pairs of reads received,
- Pair-deduplicated Reads: the count of unique pairs after pair-informed deduplication,
- Trimmed: counts the pairs left after sequencing quality and adaptor trimming,
- Overall Alignment Rate: the percentage of reads mapped to the reference genome (according to set bowtie2 filters),
- Mapped in proper pair: number of read pairs mapped in proper orientation and distance between each other,
- MAPQ>=15: pairs of reads with mapping quality score at least 15 (probability of an incorrect match equals $\frac{1}{10\sqrt{10}}$),
- Read-deduplicated: the number of read1 reads that are unique,
- Overlapping <10% Problematic: read1 reads that overlap less than a given percentage threshold of problematic regions,
- Valid Positions: positions of damage inferred from read1 reads considered valid (on chromosomes of interest and mapping within the assembly),
- Position-deduplicated: the number of unique inferred positions.

Sample	Paired Reads	Pair dedup. Reads	Trimmed	Overall Alignment Rate	Mapped in proper pair
DMSO	870906	650177	646674	91.95%	594648
TMZ	4771523	3893584	3851689	94.03%	3621916
Untreated	2837921	2006593	2000028	92.64%	1852757
MMS	3714684	3257220	3227443	93.68%	3023388

Table 4.1: *Per-sample read counts at all preprocessing steps of the computational processing of AB-seq results, for 4 samples: DMSO, TMZ, untreated, MMS.*

Sample	MAPQ>=15	Read dedup.	Overlap <10% Problematic	Valid Positions	Position dedup.
DMSO	551487	525275	346467	345883	343582
TMZ	3332198	3102463	2180663	2178119	2158870
Untreated	1724564	1634986	1112599	1110945	1102042
MMS	2766826	2649981	1865808	1864369	1856451

Table 4.1: *(*Continued*)*

The pipeline includes basic visualization summary plots of the produced damage maps. For example, the frequencies of damage across the four nucleotides and their triplet and pentameric contexts (e.g., including one or two bases on each side of the damaged one; see below) are compared across treated and control samples. Moreover, the overall distribution of damage sites along the genome, for different bin sizes is computed and visualized.

4.1.2.2. Sequence context analyses

It is known that alkylating agents mostly produce damage in G's and A's [9]. The results of NMP-seq (that serves as an inspiration for AB-seq) confirmed that mapping damage using a BER-like approach maps lesions mostly in these two bases [24]. Hence, in the first analysis of the damage maps, we explored both bases and larger sequence contexts in which the AB-seq mapped damage. We inspected raw counts of the bases, trinucleotides, or pentamers corresponding to mapped positions; and we also looked at normalized counts (corrected by the expected genomic counts of the same contexts) and at their frequencies.

4.1.2.2.1. Modified bases

We started by inspecting plots of raw base counts (Figure 4.3). First, looking at the raw counts we noticed that controls have consistently fewer reads overall than the treated samples. This is not the case when looking at specific bases, where sometimes controls have more reads. When interpreting this result, it is important to keep in mind that treated samples are enriched for the signal, damage in Gs and As, and thus in relative terms the proportion of Cs and Ts with damage is lower than in the untreated. In other words, a lower absolute number of reads in those bases does not mean that there are fewer of them in the treated cells than in untreated cells. It just means we capture preferentially, and enrich successfully, for the actual damages.

Importantly, for both alkylator-treated samples, we saw an important increase of positions mapping in G's and a potential one in A's (relative to C's and T's), compared to controls. Across TMZ-treated samples, both G's and A's counts were higher than their counterparts for DMSO. While for MMS the increase in reads mapping to A's is not as obvious while looking at the raw counts, one can see it if interpreting the frequencies (Figure 4.4). Frequencies are normalized both by the genomic context (allowing the proper study of actual enrichments in bases or contexts over the genomic composition) and the total number of reads sequenced in the sample (helping with proper interpretation in light of experimental variability and saturation). Of note, the frequencies of the 4 bases in the control samples were considerably flatter than in the

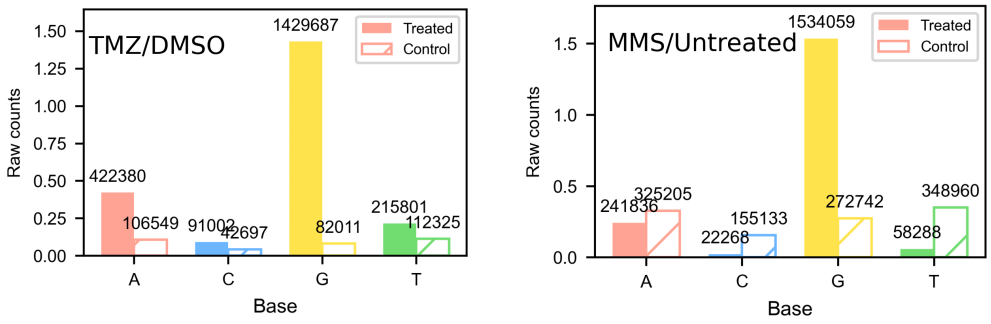

Figure 4.3: *Raw counts of bases mapped in treatment (filled color bars) and corresponding control (hatched bars) AB-seq samples. Left: Temozolomide and DMSO; right: MMS and untreated.*

treated ones. If one visually compares the frequency levels of A and T in the untreated sample – and sees they are practically the same – and then notices that the frequency of A's is higher than that of T's in MMS, the presence of modifications in A's becomes obvious.

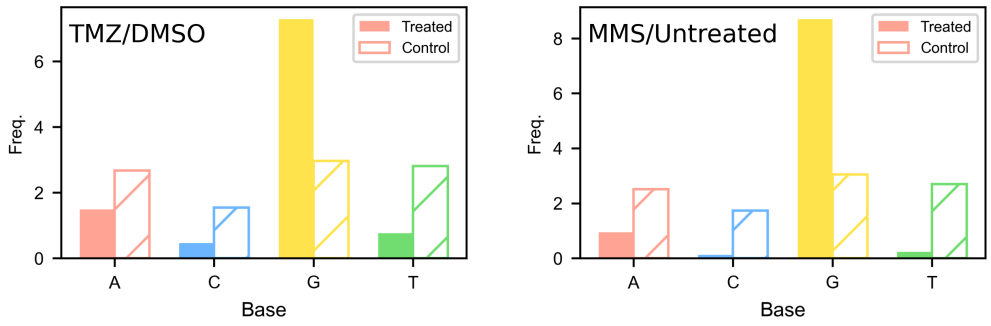

Figure 4.4: *Frequencies of genome-normalized bases mapped in treatment (filled color bars) and corresponding control (hatched bars) AB-seq samples. left: Temozolomide and DMSO; right: MMS and untreated.*

Relation to other datasets

To validate the result, we looked at related datasets from the literature and the lab. We revisited the MMS damage mapping in yeast paper [24] and saw similar results: high enrichment in G's and minor in A's upon treatment. Additionally, we checked the liquid chromatography tandem mass spectrometry (LC-MS/MS) results of DNA extracted from drug-treated cells in vivo that provided us with percentages of modified to unmodified nucleotides in the samples (Figure 4.5). No modified bases were detected for the controls. For the treated, consistently for both drugs, m7G was more abundant than m3A. With this, we confirmed that the AB-seq damage mapping

outcomes, with respect to mapped bases, are in agreement with previous works, and that the G and A reads in treated samples probably reflect the expected proportions of m7G and m3A adducts.

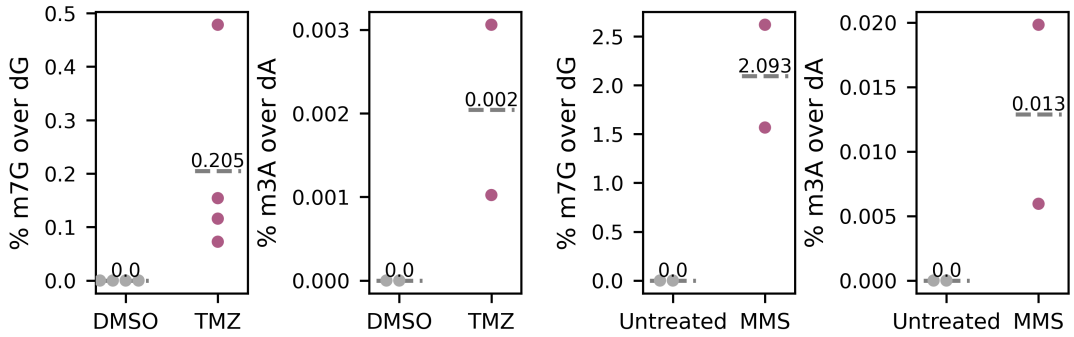

Figure 4.5: *Results from LC-MS/MS of treated and control samples in the experiments in the lab, used to quantify the induced m7G and m3A damage. Each dot represents a replicate. Means are shown with dashed lines.*

The high concentration found in G's was consistent with the literature: most BER-recognizable, MMS- and TMZ-induced damages are m7G. Next, we wanted to understand how the ratios of m7G and m7A look for both drugs and how the results from the two methods – damage maps and LC-MS/MS – correspond to each other. For the damage mapping, first, we calculated the fold increase of mapped bases in treated over what is expected from the control (Figure 4.6), to then calculate G/A ratios of those (results were similar for raw and normalized frequency counts). For LC-MS/MS, we calculated the ratios of means represented in (Figure 4.5). Interestingly, there was a higher ratio of m7G to m3A in MMS (7.564 damage map, 162.43 LC-MS/MS) versus TMZ (4.398 and 100.54). The differences between the methods might be due to the differences in the metrics behind the ratios (that represent different things: for maps, we compared fold changes of bases over control, and for LC-MS/MS we compared means of percentages mapped over the whole genome) as well as different sensitivities and experimental settings of the methods. For the differences between agents, a potential source could be that – although the treatment times were the same – the doses differed. It would be surprising though that this would cause such a pronounced difference in the ratio of mapped damages in two bases. Taking into account that MMS is a direct methylator, and TMZ needs to be metabolized into one, we suggest a slightly different preference or mode of action for the modification deposition of the two agents. In addition, the difference could also be explained in part by the mechanism of nucleophilic substitution both agents have: while both are monofunctional methylating agents, MMS is as S_N2-type methylator and TMZ is an S_N1-type methylator [9].

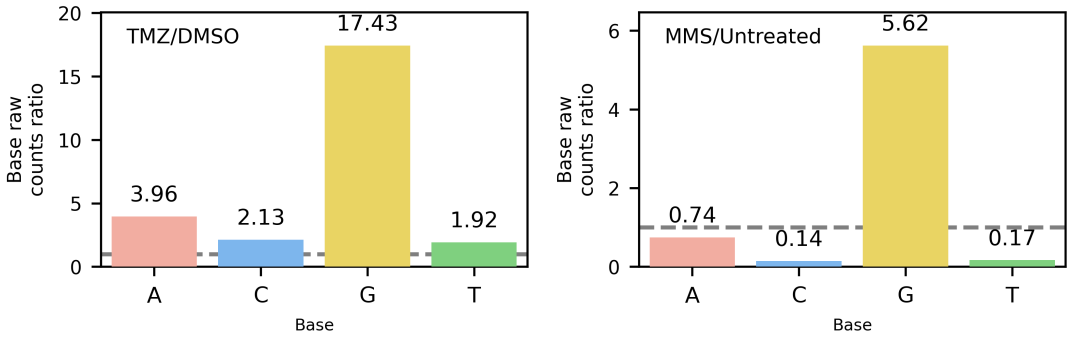

Figure 4.6: *Fold changes of bases in the treated sample over control, for raw count types of two treatments from AB-seq experiments: left TMZ/DMSO, right MMS/untreated. The dashed line points out 1 (equality).*

4.1.2.2.2. Trinucleotide damage patterns

Trinucleotide context analysis serves to understand the local sequence context preferences of the damage formation and to compare them to mutational signature patterns produced by the damaging agent if those are available. It is important to note, that the trinucleotide context in the generated control samples was not perfectly flat (which might be due to the glycosylase or endonuclease sequence bias, or due to some endogenous damage, or other damage introduced during the handling of the samples). But, when comparing raw counts, it became obvious that the control pattern is still flatter and smaller than that observed in the treated samples (TMZ and DMSO on Figire 4.7, MMS and Untreated on Figure 4.8).

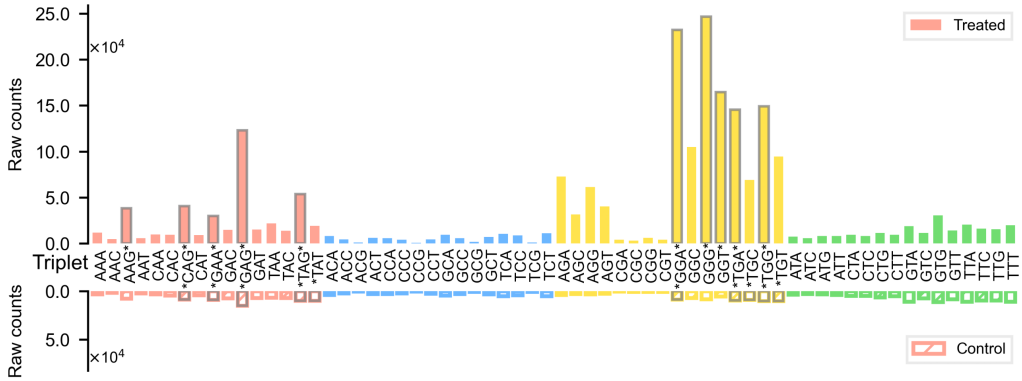

Figure 4.7: *Raw counts of triplets mapped in Temozolomide treatment (filled color bars, ticking to the top) and corresponding DMSO control (hatched bars, ticking to the bottom) AB-seq samples. Grey markings on the bars and stars next to the triplet indicate it to be within the top 5 highest ones either amongst the A-centric or G-centric contexts.*

To properly infer the most enriched contexts, we inspected genome-normalized frequencies (TMZ amd DMSO on Figure 4.9, MMS and Untreated on Figure 4.10).

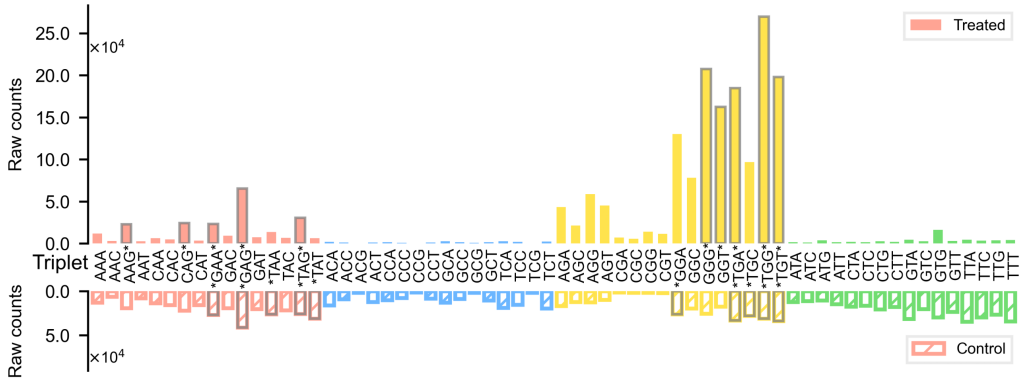

Figure 4.8: *Raw counts of triplets mapped in MMS treatment (filled color bars, ticking to the top) and corresponding untreated control (hatched bars, ticking to bottom) AB-seq samples. Grey markings on the bars and stars next to the triplet indicate it to be within the top 5 highest ones either amongst the A-centric or G-centric contexts.*

There was a very clear enrichment in G-centric triplets in the treated samples: the top 5 TMZ triplet contexts were GGG, GGA, GGT, GGC, and TGG; for MMS those were GGG, TGG, GGT, TGT, and TGA. The most frequent triplets were slightly different for the two alkylating agents. While all TMZ top 5 were enriched in the MMS-treated sample too, only 3 of them were found in the top 5 of MMS, and in a different order. This added to the suggestion of a small, but present, difference in the sequence context preferences between the MMS and TMZ.

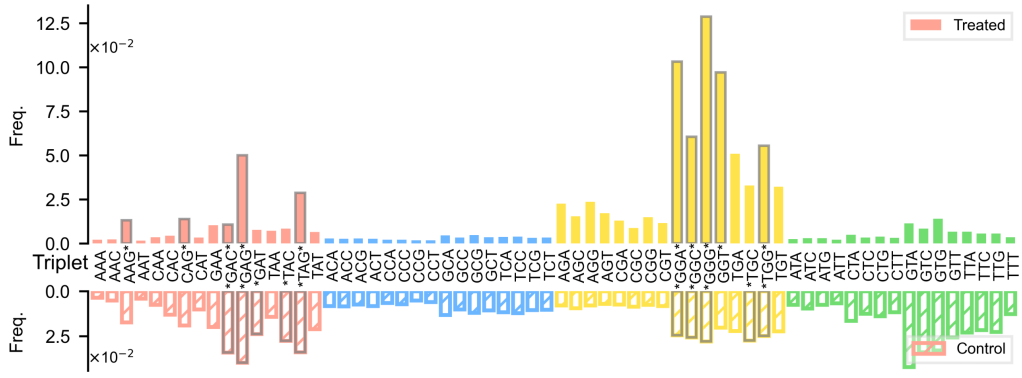

Figure 4.9: *Frequencies of genome-normalized triplets mapped in Temozolomide treatment (filled color bars, ticking to the top) and corresponding DMSO control (hatched bars, ticking to bottom) AB-seq samples. Grey markings on the bars and stars next to the triplet indicate it to be within the top 5 highest ones either amongst the A-centric or G-centric contexts.*

To quantify the differences between control, MMS, and TMZ trinucleotide patterns, we calculated the cosine similarity of the genome-normalized triplet frequency profiles (Figure 4.11). A clear conclusion emerged: the two control profiles were nearly

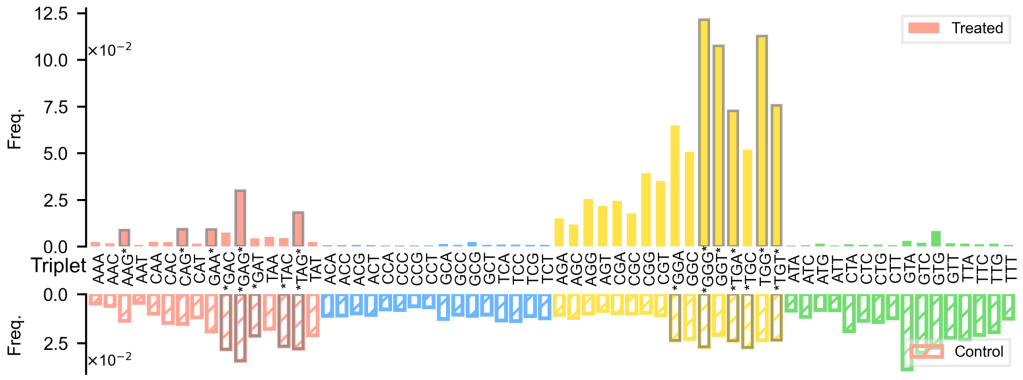

Figure 4.10: *Frequencies of genome-normalized triplets mapped in MMS treatment (filled color bars, ticking to the top) and corresponding untreated control (hatched bars, ticking to bottom) AB-seq samples. Grey markings on the bars and stars next to the triplet indicate it to be within the top 5 highest ones either amongst the A-centric or G-centric contexts.*

identical, and the treatment ones were highly similar to each other; while the control and treatment profiles showed only some similarity between themselves. This is not surprising - the background measured in controls is somewhat present in the treatment. The similarities between the drugs are expected as well - both are alkylators that deposit the same modifications.

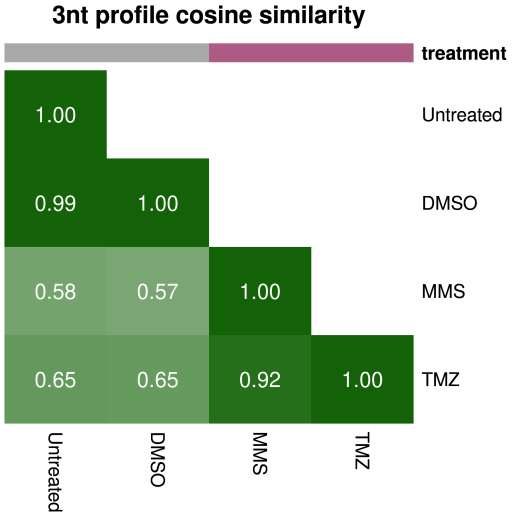

Figure 4.11: *Cosine similarity of the trinucleotide genome-normalized frequency profiles between the 4 AB-seq samples.*

Next, we set out to deepen the understanding of how these profiles separate from each other. Maria Andrianova from our lab performed Principal component analysis (PCA) using the genome-normalized frequency data (Figure 4.12). The two first principal components explained nearly all (99.6%) of the variance in the data. The first principal

component (PC1) clearly separated patterns of both controls from both treatments and negatively correlated with the latter (hence we represent -PC1 below). The triplets with the largest weights for PC1 in the direction of controls were mostly C, T and A-centric; while the high-weight contexts for the other – treatment – direction, were G-centric: -PC1: GGG, GGT, TGG, GGA, TGA, TGT. When inspected together with the plots of normalized frequencies of triplets shown above (TMZ and DMSO on Figure 4.9, MMS and Untreated on Figure 4.10), one can notice that these specific contexts concentrated in the high frequencies for all samples, but in different orders. The second principal component (PC2) separated the two treatments and highlighted the triplets with the highest loadings for both directions: -PC2, associated with MMS: TGG, TGT, CGG, CGT; and TMZ-related PC2 with GGA, GGG, GAG, and GGC. Again, viewing these high-weight PC2 triplets in the circumstance of the trinucleotide profiles above, many clearly varied in position when ordered by frequency between the TMZ and MMS. Additionally, some were on rather the low-frequency end but still differed in rank between the two samples. These distinct patterns not only further established the success of mapping alkylating damage, but also show the differences in induced adducts and potential context preferences of the two damaging agents.

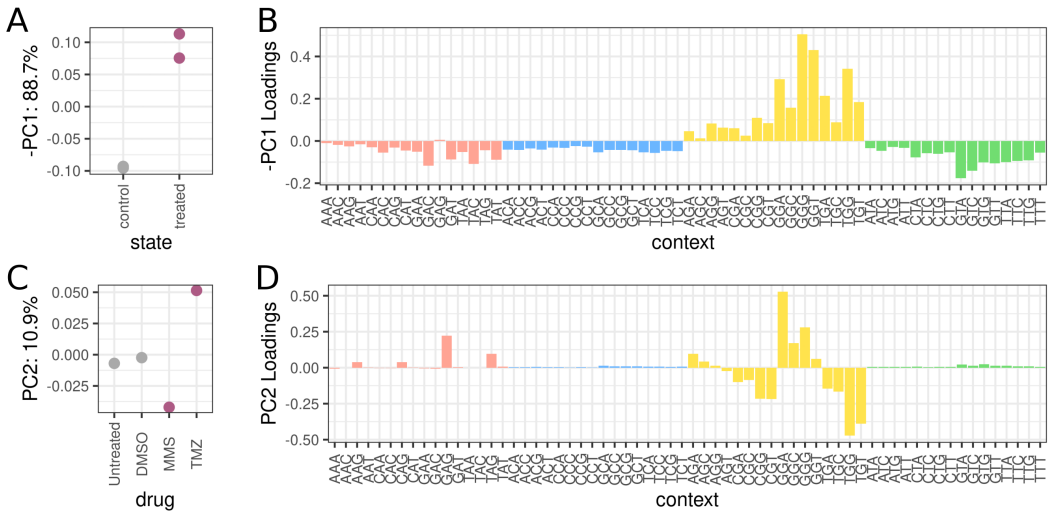

Figure 4.12: *PCA of the trinucleotide genome-normalized frequency profiles between the 4 AB-seq samples. A: scores of the samples on the -PC1; B: -PC1 loadings of triplets; C: scores of the samples on the PC2; D: PC2 loadings of triplets.*

Relation to other datasets

Similarly as for bases, we aimed to compare the AB-seq results to previously published data. We compared the AB-seq TMZ data with the COSMIC mutational signature SBS11 [65, 69, 70] associated with this treatment. The first 4 top trinucleotide TMZ contexts were reflected within the top 5 contexts of the SBS11 mutational signature

(Figure 4.13): T[C>T]C (GGA), A[C>T]C (GGT), C[C>T]C (GGG), T[C>T]T (AGA, not in top 5 of AB-seq data), G[C>T]C (GGC).

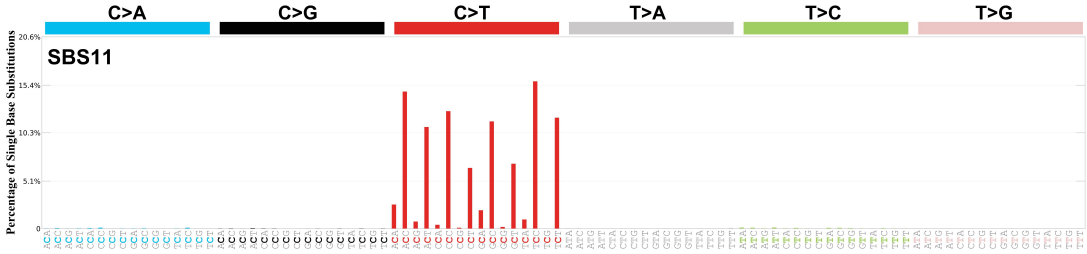

Figure 4.13: *COSMIC Mutational Single Base Substitution Signature 11, represented along the 96 mutational channels (6 pyrimidine-centered mutation types and 2 neighboring bases). SBS11 is suggested to be associated with Temozolomide treatment. Adapted from COSMIC [65, 69, 70].*

To quantify the closeness of the damage trinucleotide patterns to the SBS11 signature, we calculated the cosine similarity of the genome-normalized triplet frequency profiles (Figure 4.14). The data required a slight transformation: as for mutations we do not have a specific triplet available, but rather 3 mutational channels representing two triplets (e.g. A[C>A]A, A[C>G]A and A[C>T]A represent the sum of ACA and TGT triplets) we used the sum of the represented triplets for both cases and compared profiles this way. The mutational signature seemed to be highly similar to the TMZ AB-seq patterns, and not to the ones of DMSO. We concluded that the Temozolomide damage profile is coherent with the mutational signature potentially induced by this treatment.

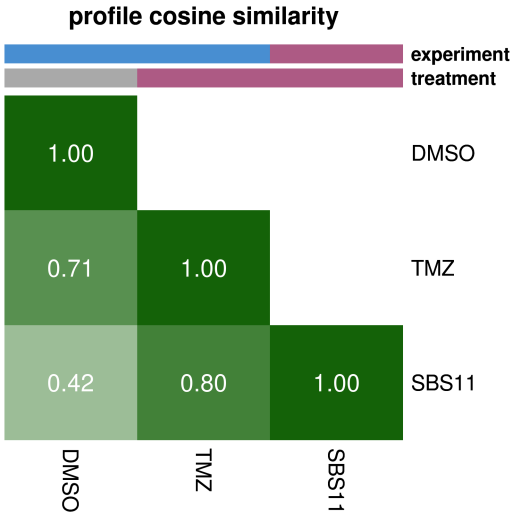

Figure 4.14: *Cosine similarity of the summed (based on similar mutational channels) trinucleotide genome-normalized frequency profiles of TMZ and DMSO AB-seq samples and the COSMIC Signature SBS11.*

In the case of MMS, no mutational signature in human cells is currently available. Instead, we looked at NMP-seq yeast damage data [24]. For that, we downloaded two 0h MMS-treated samples (0.2% and 0.4%, 10m MMS treatment or 23.6 and 47.2 mM respectively) and a noMMS control, extracted and counted the triplets for all of them, and performed the downstream analysis in the same manner as for the AB-seq data (adding the step of discarding of damages overlapping NotI restriction enzyme sites). After normalizing for the genome content (Figures 4.15 and 4.16), triplet contexts were quite flat compared to observed AB-seq results in human cells.

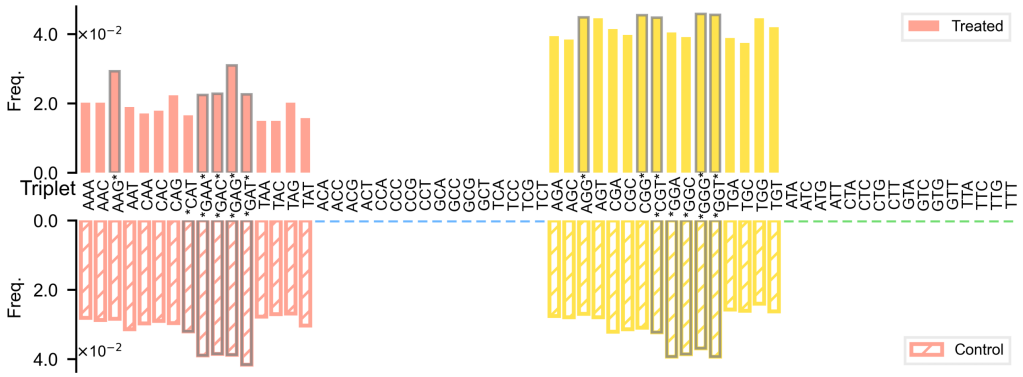

Figure 4.15: *Frequencies of genome-normalized triplets mapped in treatment MMS treated replicate 1 (filled color bars, ticking to the top) and corresponding no MMS control (hatched bars, ticking to bottom) [24] yeast samples. Only A and G-centric triplets were available. Grey markings on the bars and stars next to the triplet indicate it to be within the top 5 highest ones either amongst the A-centric or G-centric contexts.*

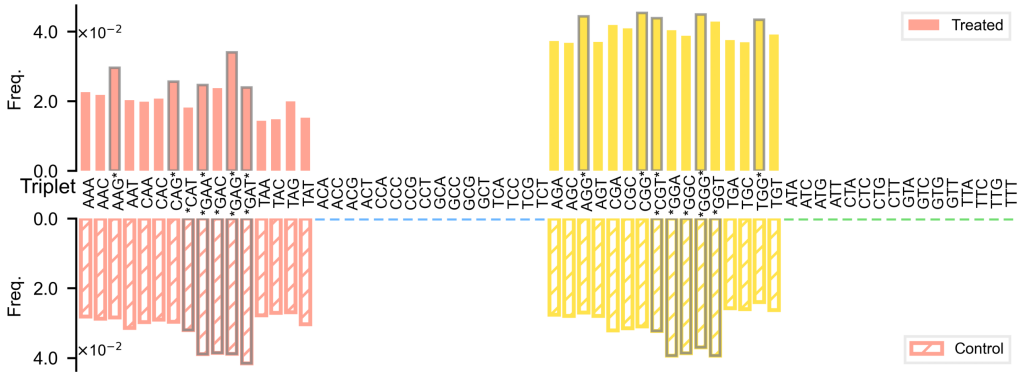

Figure 4.16: *Frequencies of genome-normalized triplets mapped in treatment MMS treated replicate 2 (filled color bars, ticking to the top) and corresponding no MMS control (hatched bars, ticking to bottom) [24] yeast samples. Only A and G-centric triplets were available. Grey markings on the bars and stars next to the triplet indicate it to be within the top 5 highest ones either amongst the A-centric or G-centric contexts.*

Moreover, when intending to find the top 5 A-centric and G-centric triplets from the AB-seq data (Figure 4.10) in each yeast sample, we found the overlap to be quite poor. Comparing AB-seq untreated top contexts to the yeast noMMS, one finds 3 in the top of noMMS for A's, and rest at the very tail; and 2 for G's (GGG, GGA) and similarly rest at the bottom of the frequency ranking. The same comparison for the MMS-treated samples yielded varying results depending on the MMS dose for the yeast: for the higher dose, the 4 top AB-seq A triplets could be found at the top of the list, while for the smaller dose just 3. For G triplets the overlap was the same in both cases: 2 at the top (GGG, TGG), but in a different order.

To understand the low correspondence between the trinucleotide MMS patterns in human obtained by AB-seq to the ones obtained by NMP-seq in yeast, we again utilized the cosine similarity and PCA, selecting only the A and G-centric parts of the genome-normalized triplet frequency profiles (Fig 4.17). To perform the comparison as fair as possible, similarly to the NMP-seq data, we filtered the AB-seq damage (including the genomic counts) to exclude any triplets overlapping NotI sites.

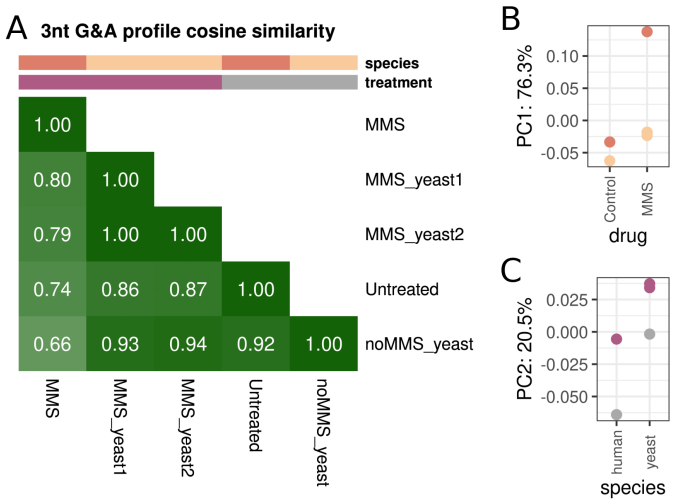

Figure 4.17: *Cosine similarity, B and C: PCA of the A and G-centric trinucleotide genome-normalized frequency profiles, excluding damages overlapping NotI sites, of MMS and untreated AB-seq samples (human) and the two MMS and no MMS yeast samples from [24]. For PCA, scores of the samples on the B: PC1 or C: PC2 are represented.*

Although the MMS signatures clustered close together in both cases and were highly similar, we were surprised by the very high similarity of the treated samples of one species and the opposing species controls. This disparity – in line with the low triplet overlap mentioned above – might be explained through a few paths: differences in the protocol (e.g. length or dosage of treatment, different sequencing platform used) or

differences between yeast and human (e.g. in chromatin accessibility to the damage, mechanistic DNA repair differences, cell response to the drug).

4.1.2.2.3. Pentanucleotide damage patterns

Two important findings from trinucleotide patterns - non-flat controls, and context differences for the two agents - were reflected when extending the context to two bases more and exploring the pentamer sequences (Appendix, Figures A.1, A.3, A.2, A.4). Similarly, both cosine similarity and PCA analysis using these extended contexts yielded a separation of control pentamer patterns from treatment on PC1, and a separation of treatments from two agents on PC2 (Appendix, Figure A.5). This further solidified the differences in the sequence preferences of the two alkylating agents.

4.1.2.3. Damage distribution along the genome

We aimed to visually represent the damage map and inspect the general distribution of damage along the genome in the 4 bases. To this end, we generated damage-distribution plots of base counts in 1Mb chunks along chromosome 1, and its zoomed-in fragment of Mb 150 to 200 (TMZ and DMSO on Figure 4.18, MMS and Untreated on Figure 4.19). All the counts represented were normalized by the genomic counts of the base in each chunk.

Knowing that C and T bases are expected to be almost exclusively background and not actual damage, it is important to notice that their distribution in treated samples was at around the same level or lower than in the controls. Of note, A's distribution in treated samples was substantially higher than what was expected from controls (where it would be overlapping the T's distribution). Visually, the landscapes of the bases, especially G's and A's, seemed quite similar both within the same sample as well as in the matched one.

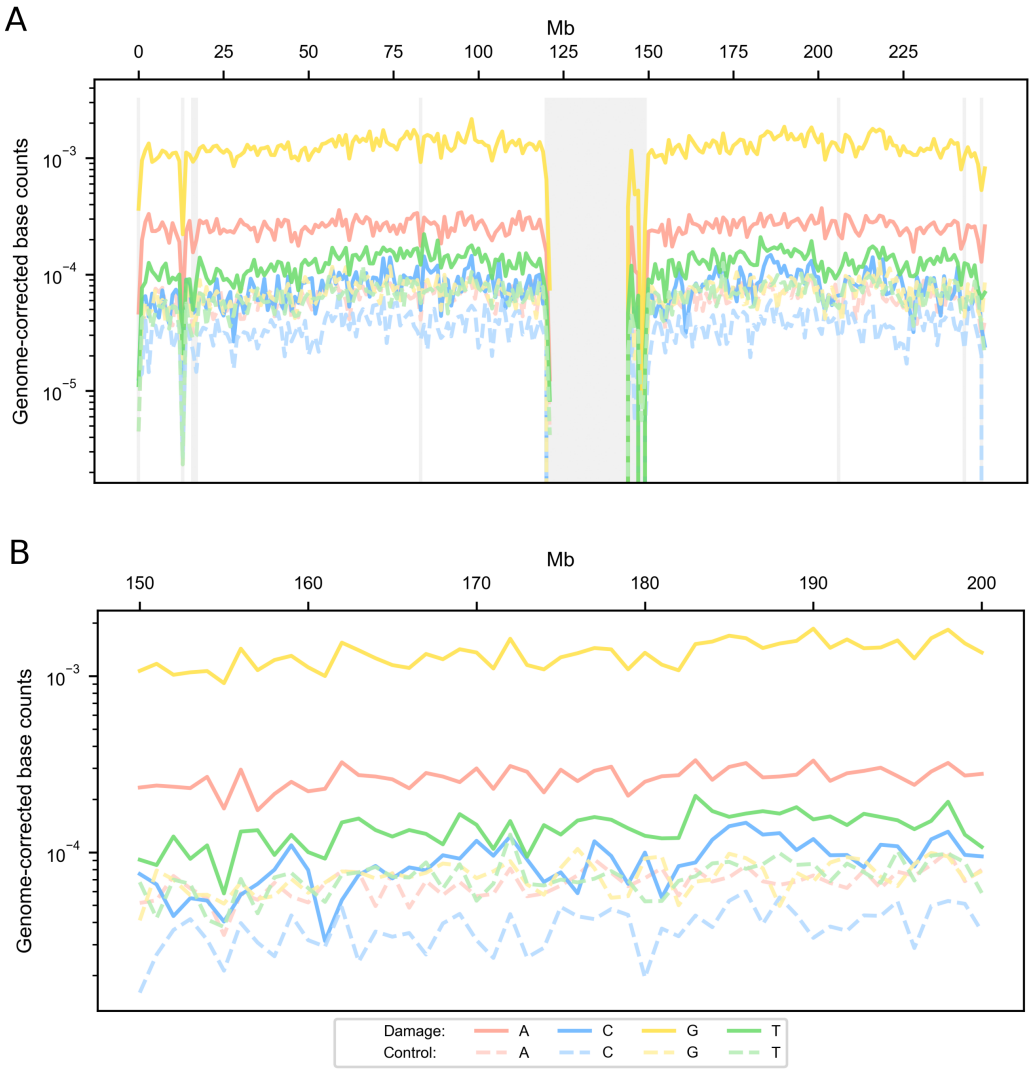

Figure 4.18: *1Mb chunk landscapes of the distributions of genome-corrected counts of bases mapped in a TMZ treatment sample and matched DMSO control. A: chromosome 1, B: zoomed-in view of 150Mb to 200Mb of chromosome 1.*

To characterize in depth this perceived affinity of the shapes of the landscapes along the whole genome, we evaluated their correlations. We took the vectors of total counts of damaged positions in chunks along the whole genome and compared them using the correlation metric (Figure 4.20). The analysis was repeated for 2 resolutions of different chunk sizes: 1Mb, and 100kb, using the corrected total counts (summed over all bases).

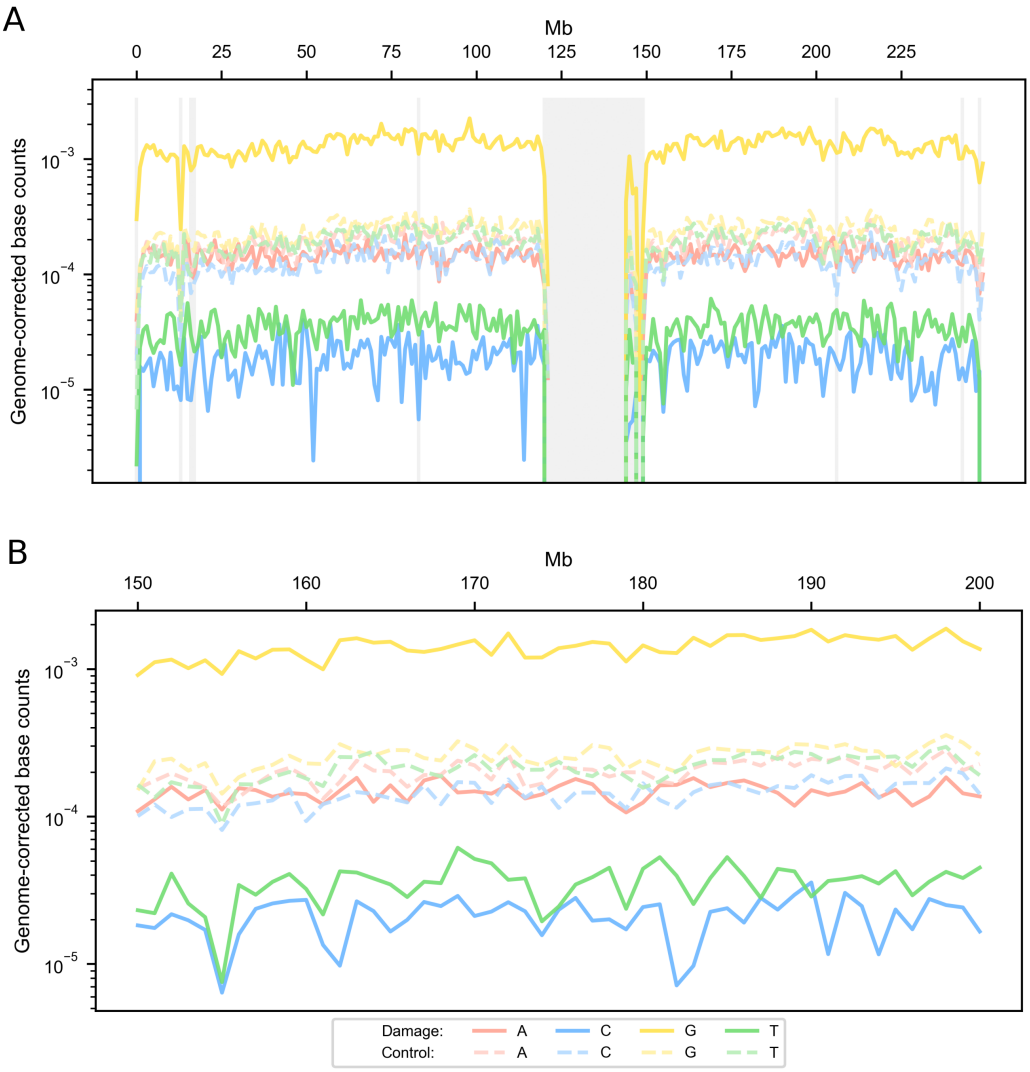

Figure 4.19: *1Mb chunk landscapes of the distributions of genome-corrected counts of bases mapped in MMS treatment sample and matched untreated control. A: chromosome 1, B: zoomed-in view of 150Mb to 200Mb of chromosome 1.*

The DMSO landscape seemed to differ the most from others. The untreated sample was closer to the treated ones. The treated sample profiles, as expected, were the most similar. The high similarities might point to either too high of a chunk size used for the differences to be seen, or too high sparsity of the data or noise-to-signal ratio for this type of analysis. Taken together, these results suggest that we were able to successfully build damage maps of alkylating agents and create sufficient opportunities for insightful analyses into damage formation in the context of features.

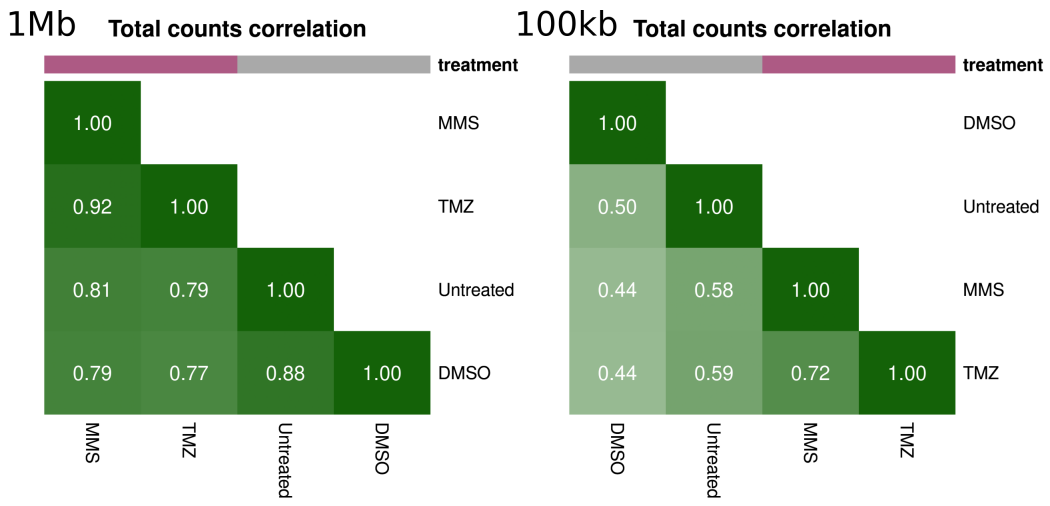

Figure 4.20: *Correlation of the corrected genomic distribution landscapes between the 4 AB-seq samples. Left: 1Mb, right: 100kb resolution..*

4.1.2.4. Next steps

There are many further, interesting analyses planned to be done utilizing this data. Next, the computational team will analyze the damage formation with respect to different (epi)genomic features. From the experimental side, we are currently working on adapting the protocol to the generation of maps at different times after the exposure. This way, cells have time to recover from the damage and repair some of it, and we can use this to infer and study the activity of repair. This study could be done both within specific features, as well as in a more comprehensive manner - with the repair states framework we developed and tested on UV damage maps.

4.2. Framework for genome partitioning into UV DNA Damage Repair States

Analyzing damage deposition and repair activity is crucial to understand mutagenesis, since mutagenesis is the end result of their interplay. These two components of the mutagenic process have long been known to be directly affected by genomic features. As outlined in the Introduction (1.4), this interplay has been explored already, but only in the context of specific features. The usual approach focuses on one specific feature and stratifies the genome according to it, to then regress mutation, damage, or repair rates against it. This is not very efficient if one wants to study multiple epigenetic features, or if a given feature does not cover the entire genome.

This chapter focuses on an idea with two goals: partitioning the genome solely by the activity of DNA damage repair, in a way that is biologically meaningful. It starts with the introduction to the UV damage and repair data. However rich and unprecedented, DNA damage and repair mapping data pose many challenges for a correct downstream analysis and interpretation. I explain the motivations – confounders – behind the necessary data processing.

Inspired by the idea of partitioning the genome into chromatin states ([82, 84, 85]), we reasoned that we could devise a strategy to partition the genome solely by the activity of DNA damage repair. With this repair-based, data-driven partitioning, we could undertake systematic surveys to uncover links to various features. I outline the computational HMM-based framework implementing this idea and finish by presenting the resulting states of repair activity along the genome (further termed: repair states) and a comprehensive exploration of their relations to various features. All the technical details and parameters are described in depth in the corresponding Methods chapter of this thesis.

Most of the results presented in this thesis are focused on the 100kb chunk size. Although we found the 10kb model to be quite reliable as well - and more useful for fine-grained analyses of smaller-sized features - the sparsity of the data at this level, and the consistency of results with the 100kb resolution, made us focus on the larger chunks.

4.2.1. DNA repair is a major influence on the distribution of UV mutations along the genome

We aimed to find a way to partition the human genome into segments of similar UV-induced damage repair dynamics. We started by exploring UV mutagenesis at three different levels: DNA damage (photoproducts), its repair (NER), and UV-induced mutations. Whole-genome sequencing nucleotide-precision datasets are available for all three (Figure 4.21 A).

4.2.1.1. Damage and repair maps as great tools to study UV mutagenesis

UV damage sites for cellular DNA kinetics mapped with HS-damage-seq were obtained from [19]. Overall, we obtained data for 20 conditions, defined by the combination of damage type (6-4PP or CPD), the di-pyrimidine contexts, and the time elapsed after UV light irradiation (apart from the 36h time point, see Methods 3.3.1.4). We binned the amount of damage mapped to equal-sized chunks of genomic sequence (100kb; heretofore genomic chunks). We noticed three confounders of the number of damage sites per genomic chunk. First, damage right after exposure (0h time point) closely follows the amount of "damageable sites" - corresponding pyrimidine pairs along the genome (as represented for CPDs in Figure 4.21 B). Secondly, more CPDs were mapped at 1h after irradiation than at 0h – unlikely with an agent like UV-light which produces DNA damage when present (we are not considering indirect damages here) – thus suggesting experimental saturation of the HS-damage-seq protocol at early time points following exposure (see 1.3.2.1.1). Third, the counts of damage sites were not reliable in genomic chunks with no damageable sites, or no damage mapped at 0h, or within those with more than 40% overlap of problematic regions (low mappability, repetitive, UCSC-blacklisted). We applied suitable corrections for these three confounders by filtering out damage sites close to large problematic regions and normalizing each di-pyrimidine count by its genomic context and total sites mapped at each time point.

NER activity, consisting of sequenced excised DNA fragments, measured by XR-seq snapshots for both damage types at different times after exposure to UV (amounting to 11 conditions) was downloaded from [23]. We corrected the chromosome-sequencing-depth normalized repair counts for each chunk by the genomic counts of TTs (as they constitute the vast majority of the reads), due to the dependency of XR-seq scores on total available damageable sites (as suggested by [76]).

To obtain mutations that are highly likely produced by UV light exposure, we

exploited skin cancer samples sequenced within ICGC (International Cancer Genome Consortium [79, 31]), and chose only samples where at least 70% of mutation calls were UV-attributable. Then, to make UV damage, repair, and mutations comparable, we filtered the mutations for UV-specific trinucleotide contexts, defined by the predominant di-pyrimidines for photoproducts – TT, CT, TC – forming a part of the mutation triplet.

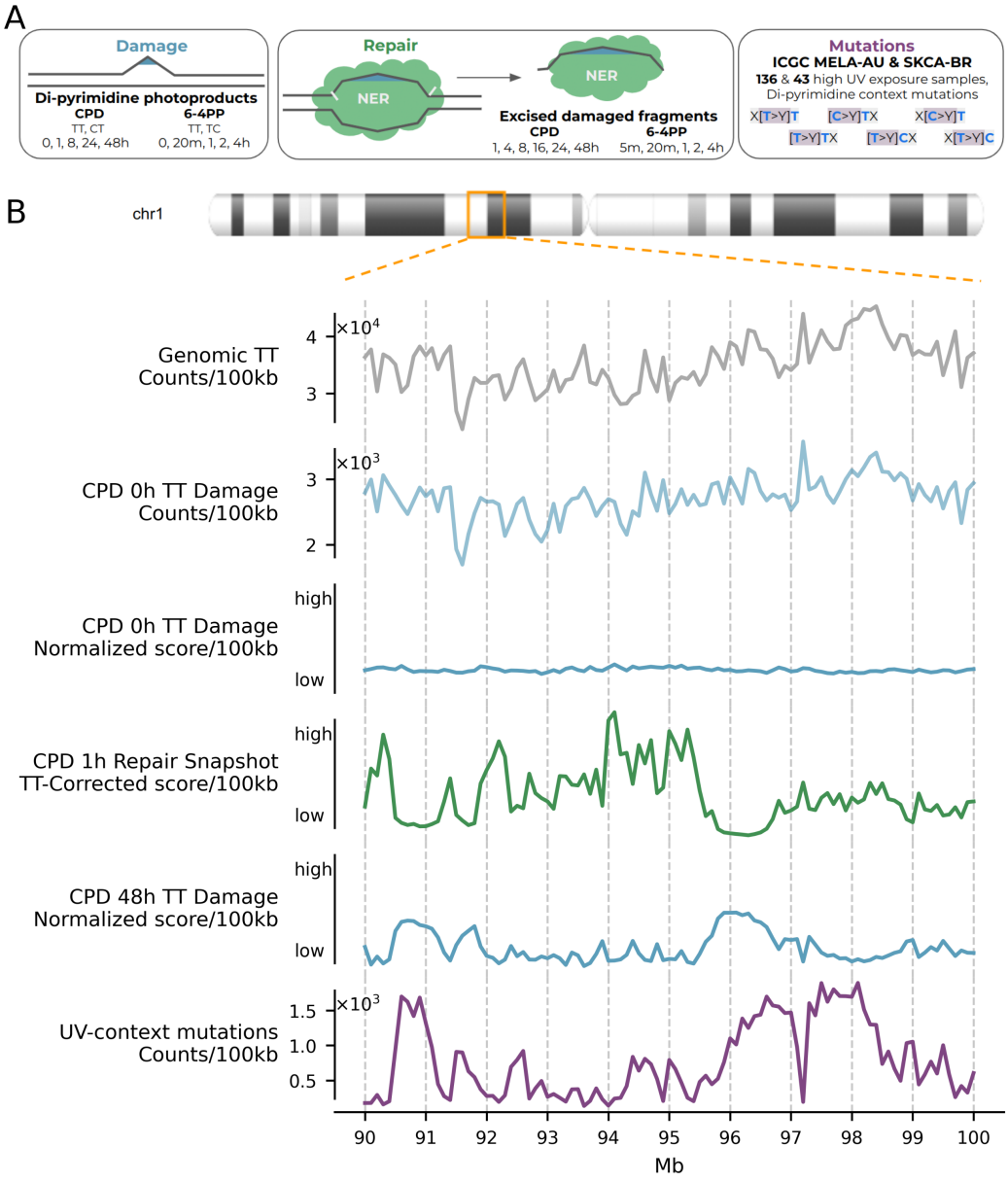

Figure 4.21: *Characteristics of the data used in this work to study UV mutagenesis. A) Datasets used at the three studied levels of mutagenesis. B) Data with various corrections plotted along the 90-100Mb of chromosome 1.*

To explore the relationship between the three levels of data associated with UV

mutagenesis and their experimental representations, we first plotted them along a small part of chromosome 1 (example for CPD TTs in Figure 4.21 B). We could appreciate visually the clear correspondence of uncorrected counts of damage sites at 0h with the number of available TT sites. On the other hand, after the corrective steps, the 0h damage landscape appeared quite flat, especially when compared to the distribution of the damage after 48h. This implies that, over time, the variability of the unrepaired damage along the genome increases. Moreover, the landscape of damage at 48h after UV exposure resembled the genomic distribution of mutations. The landscape of repair intensity measured 1h after exposure to UV tracked the inverted landscape of damage sites at 48h. This rapid exploration of these three types of data thus highlighted the well-known fact that DNA repair plays a predominant role in shaping the landscape of damage after exposure, rendering it similar to the observed distribution of mutations across tumors originating from the same tissue.

4.2.1.2. Unrepaired damage at late time points correlates better with mutations

To show the point established in the previous section more quantitatively, we performed correlation analyses. We correlated normalized damage scores at different time points with the mutations across all genomic chunks (Example for CPD TTs on Figure 4.22). We observed a poor correlation of mutations with the normalized damage at early time points, which improved at later time points. Additionally, the normalized damage scores at 0h had a very small value range. In line with increased variability, the scores of damage left at 48h covered an increased range.

Of note, when we performed the same correlation analysis for the non-normalized, raw counts, the correlations both between damage and mutations (as well as between the 0h damage and other damage time points (Appendix, Tables B.1, B.2)) were higher (Appendix, Figure B.1). It seems, as one might expect, that the amount of damageable sites has a degree of influence over the amount of produced mutations. We decided to correct for that effect, however, and explore the repair activity irrespective of the availability of sites to be damaged.

Other damage types and di-pyrimidine contexts exhibited the same trend of increased correlations with mutations at later time points (Appendix, Figures B.2, B.3, B.4). The correlations were slightly weaker than for CPD TTs.

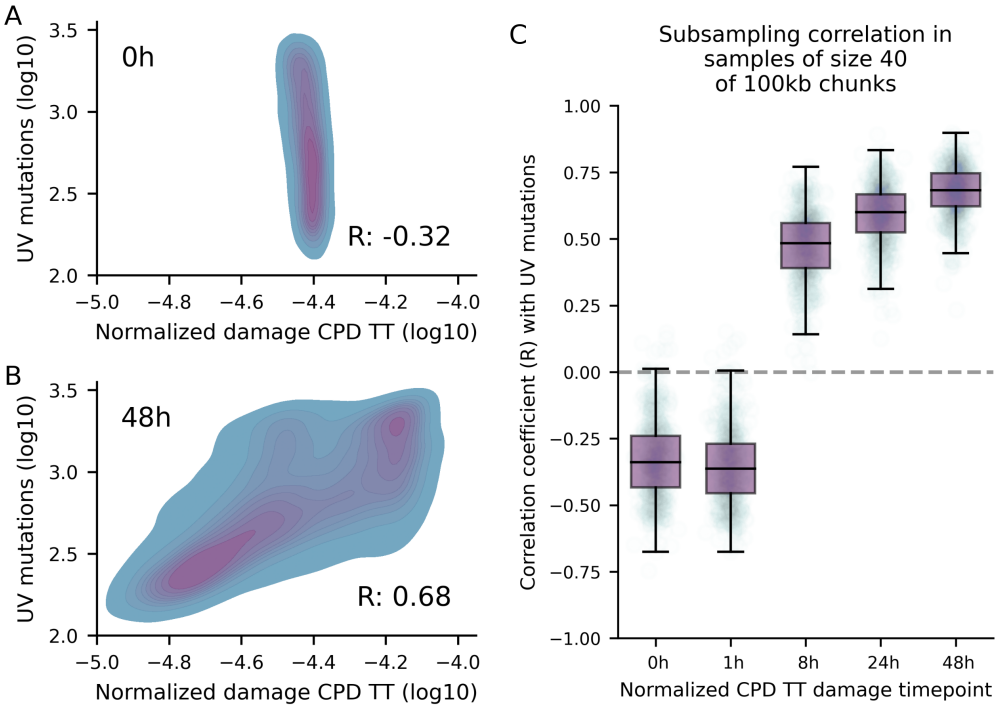

Figure 4.22: *Correlations of normalized CPD TT damage score with UV mutations within 100kb chunks. A) KDE plot of 0h damage with mutations, B) KDE plot of 48h damage with mutations, C) Subsampling correlation of damage at various time points with mutations, with a sample size of 40 shuffled chunks. Each point in the boxplot represents a single sample. R is the correlation coefficient.*

4.2.2. Segmentation of the genome into repair states

The activity of DNA repair machinery varies along the genome, and lesions in parts of the genome with a lower repair activity have a higher chance to remain unrepaired when the DNA undergoes replication. These damages then end up more likely set in the genome as a mutation. Due to that, we decided to focus our study on the activity of repair. The consecutive steps of encoding the repair data, constituting the repair state segmentation framework, are visualized in (Figure 4.23).

Although XR-seq data provides accurate information about the intensity of DNA repair, it can only be thought of as snapshots at a given time point, and not a measure of total repair between two points in time. The amount of DNA repair effected between the two time points in a genomic chunk can be inferred by subtracting the count of damage sites mapped in the chunk at a later time point from the count of damage sites mapped at the immediately prior time point (subtractive HS-damage-seq, see 1.3.2.1.1 and 1.3.2.2.1). This inference of the repair based on the damage is a complex and

challenging task, though.

For example, the issue with experimental saturation (see 1.3.2.1.1) makes subtractive HS-damage-seq problematic and opens up its results to misinterpreting e.g. the percentage of damage that is repaired in a specific time interval. To avoid this pitfall, we took a slightly different approach. Firstly, we used the normalized damage scores, to avoid the confounders explained in the previous chapter. Secondly, instead of inferring repair by subtracting the damage from two consecutive time points (further called the time interval), we calculated their divergence. Divergence is defined by the following expression: $Div(t_0, t_1) = \log d(t_0) - \log d(t_1)$, where t_0 and t_1 represent the start and end time points of the time interval, respectively, while $d(t)$ represents the normalized damage score at time t. The divergence score can be also understood as a log-fold change of normalized damage scores. As can be appreciated in (Figure 4.23), when the normalized, relative damage scores of the 8h and 24h CPDs overlap – meaning that the amount of damage stayed in a similar condition with respect to the rest of the genome – the divergence score equals 0. When the 24h damage scores are below those at 8h – meaning that the damage status of these chunks relatively decreased – divergence is positive and high. On the other hand, when the later time point damage is above that from an earlier time point (signifying a relative increase of the damage status), divergence scores are below 0 and low. From now on, we refer to the divergence scores – the repair inferred from the HS-damage-seq data – as 'inferred repair', and to the repair intensity directly measured through XR-seq as 'snapshot repair'.

To partition the genome by the repair activity, we sought a method that took two important factors into account: 1) the position of the chunks within the distribution of dynamics of inferred and snapshot repair along the whole genome, and 2) the relation of the neighborhood between chunks, so that the state assigned to a chunk influences the state of the neighboring one. We decided to go with Hidden Markov Models (HMMs). HMMs are a group of methods proven and tested time and time again for analyzing biological sequential data. One of the best-known examples of the use of this type of models in the field of biology is probably chromHMM [82, 84, 85]. ChromHMM was designed for 'chromatin state' discovery - the unsupervised learning task of identifying functionally different genomic regions based on the input of various types of chromatin features. In our case, we were interested in using the method for the task of discovering differentially repaired regions – 'repair states' – by inputting various tracks of DNA damage repair information (Figure 4.23).

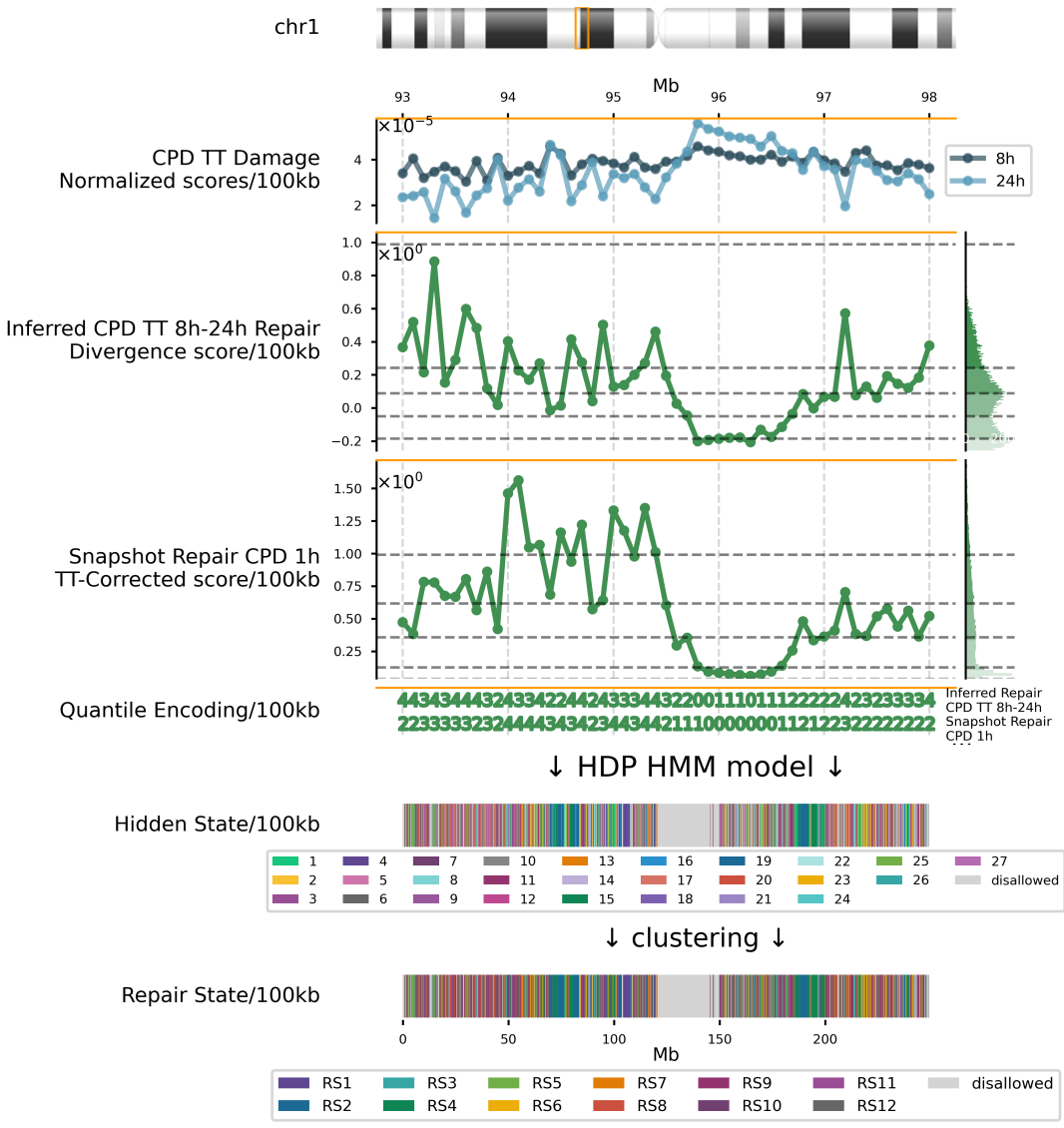

Figure 4.23: *The framework for the segmentation by repair states: from the inference of repair from the damage mapping data, through encoding the repair tracks as quantiles, and the final assignment of the hidden and repair states.*

The specific HMM model we were intent on using required an input that was encoded in a binarized manner. To binarize a particular repair data track, we split each distribution of inferred or snapshot repair points (each point a chunk) into 5 quantiles (0-4). Then, the specific repair computed for a chunk was encoded as a binary sequence of five elements (one per quantile) with 1 in the element of the sequence corresponding to the quantile to which the chunk belonged and 0 in all others (Figure 4.23). As both damage types have different time intervals, and the three di-pyrimidines seem to exhibit different behavior, we encoded all conditions separately, when possible. This encoding was performed along all time intervals and all chunks (Figure 4.24). We gathered all

the data tracks and we zeroed out the disallowed chunks (as defined in 3.3.1.2.5).

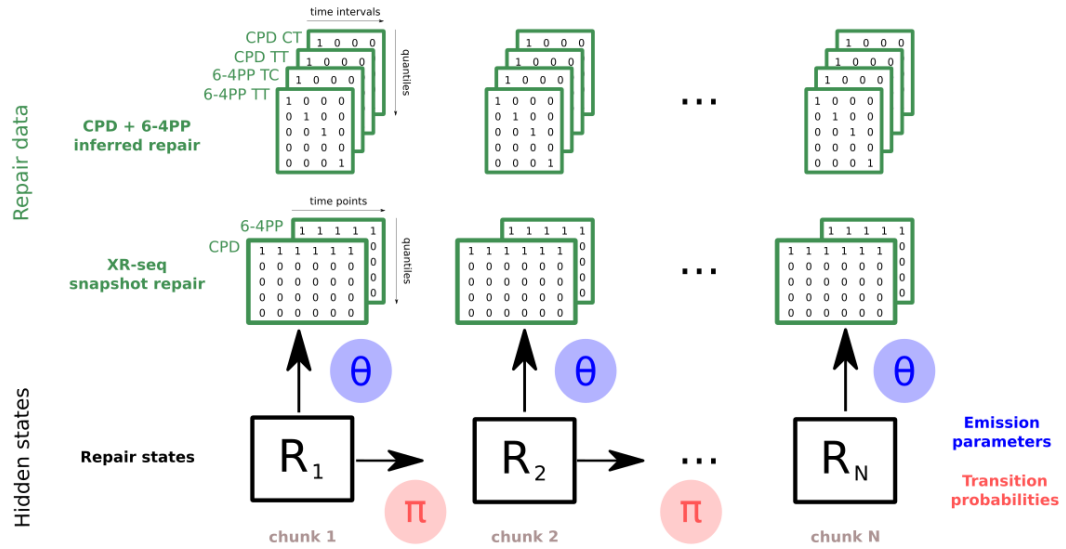

Figure 4.24: *Graphical representation of the HDP HMM model and its use for repair state partitioning, along the sequence of chunks along the genome. In green: encoded repair data tracks, serving as observations. In black: hidden repair states.*

The encoded repair input matrix for the model can be thought of as observations made as we go through the genomic chunks (in green, Figure 4.24), which are emitted by some hidden states (R_1 to R_N on Figure 4.24) of the genome that we do not know. The goal of the HMM is to uncover those hidden states. Here, we employed a sticky Hierarchical Dirichlet Process HMM (sticky HDP-HMM, [93, 80, 81]). The HDP part is used to infer the number of states (N), probabilities for an observation to be emitted from each state (emissions, Θ), and transition probabilities (π) between states from the data. 'Sticky' comes from the propensity of states to 'stick' together. Once the model found a parsimonious set of states from the observed data, we ran Viterbi's algorithm to allocate the states across chunks. More on the model and used parameters can be found in the Methods 3.3.3.

Starting with genomic chunks of size 100kb, sticky HDP-HMM allocated some 30 hidden states (Figure 4.23). This number includes one empty, 'disallowed' state, formed by all the disallowed chunks encoded as zero vectors, pulled together by the model. This disallowed state is discarded from analyses that follow.

For the sake of interpretability, we decided to further reduce the number of final repair states. We did so by clustering the hidden states based on the similarity of their repair kinetics (Figure 4.25), as follows. We calculated the mean of each data track across all chunks belonging to a specific hidden state. For inferred repair, we

used a mean of divergence scores, while for snapshot repair it was the mean of ranks, standardized from 0 to 1. The vectors of mean repair tracks representing each hidden state were clustered with the aim of obtaining a number of clusters ranging from 5 to 16. After exploring the clusterings (Figure 4.25) and transition matrices (Appendix, Figure B.5), we decided on 12 clusters as the most coherent. Those 12 clusters are further considered as and termed 'repair states'.

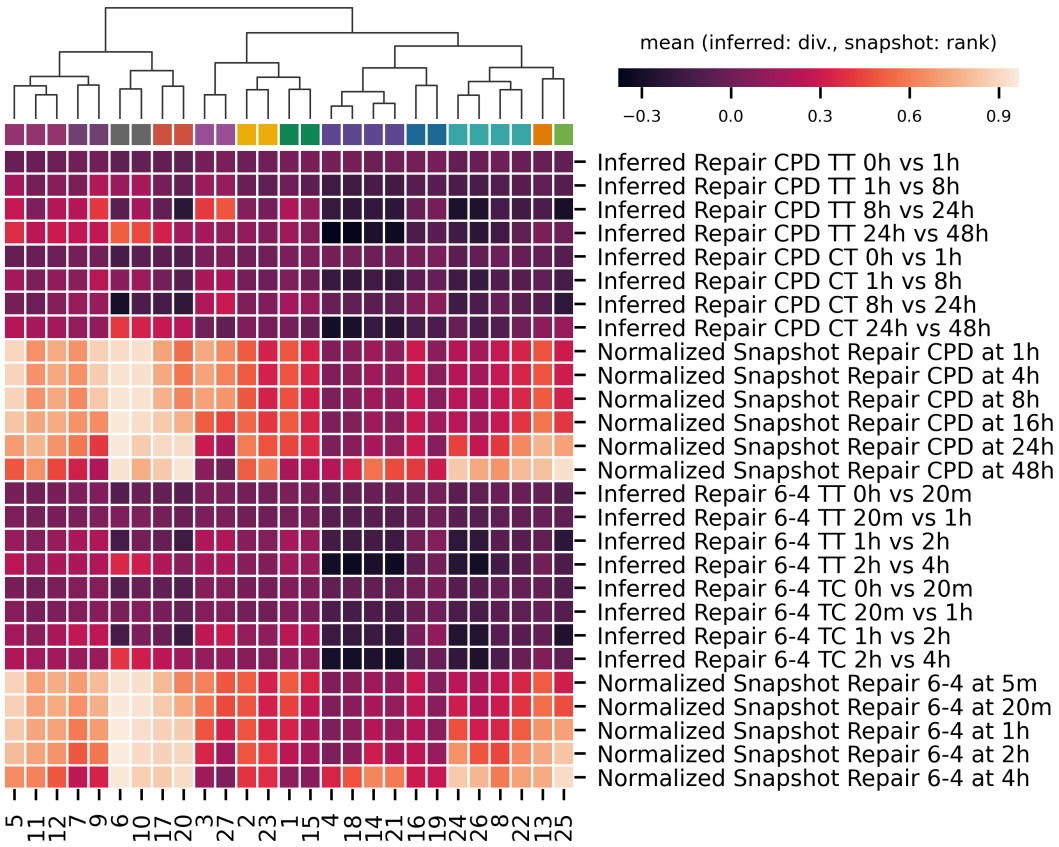

Figure 4.25: *Clustering of the hidden states based on vectors of means of each data track into 12 final repair states. For inferred repair data tracks, the mean of divergence scores was used. For snapshot repair, the mean of ranks of normalized scores was used.*

4.2.2.1. Reproducibility of the hidden state assignments

To gauge the quality of the models and states we performed a few reproducibility analyses. In the first one, we compared the assignment of the hidden states separating the data by replicates, to assess the coherence for a given chunk size (10kb, 100kb, 1Mb, Appendix Table B.4 and Figures B.7, B.8, B.9). We found the replicate reproducibility to be lower the smaller the chunk size - as expected when the data is relatively sparse. Still, a large proportion of state assignments was reproduced even for 10kb (45% for 12 clusters, compared to 70% for 100kb). In the second analysis, we explored different levels of the stickiness parameter to understand its influence on the

assignment of hidden states. Importantly, regardless of how sticky – with a stronger trend to merge similar states – the model was, the number of resulting states was similar (Appendix, Table B.3 and Figure B.6).

In summary, in this section, we obtained the repair-based partitioning. Next, we needed to check whether it was successful in terms of uncovering differences in repair dynamics along the genome.

4.2.3. Repair states reflect qualitatively different repair dynamics along the genome

What is a repair state? In order to study the differences between the repair states and best interpret their biological meaning, we wondered how they could be visually represented so that we can gather all the relevant pieces of evidence together. To do so, we decided to go back to the original input data tracks - repair inferred from damage, and the snapshot repair per condition.

Inferred repair is represented as is, meaning, we used the actual divergence scores obtained from the normalized remaining damage scores within each time interval. Negative and low scores can be interpreted as repair speed being lower, and conversely, positive scores suggest a more intense repair. The distribution of divergence scores and normalized snapshot repair (see below) of all chunks of a specific repair state for consecutive time intervals across di-pyrimidine types constitute that repair state's 'logo'. In the inferred repair parts of logos, the scores of each chunk were represented as light-colored scatters, above which the line and the box both represent the median and 25th and 75th percentiles of the scores.

To visualize the snapshot repair, we first ranked the values in the chunks at a specific time and condition along the whole genome. We used ranks, and not original scores, to more easily visualize the differences in repair. The lower the rank, the lower the repair. Aligning the ranks along all time points we obtained snapshot repair boxes of the logos. There, the ranks of each chunk were represented as light-colored scatters. The dashed line is the median and the box marks the 25th and 75th percentiles of the ranks.

This way, for each repair state, we dedicate one row – a 'logo' – with multiple columns (boxes), representing a specific repair data type throughout all the intervals (inferred repair) or sampling times (snapshot repair). We sorted the boxes first by the damage type (CPD, then 6-4PP). Within those, we first presented the inferred repair for both

di-pyrimidines, followed by snapshot repair.

Additionally, we were interested in visualizing the distribution of the number of UV-related mutations across repair states alongside their logos. Thus, we added a small boxplot with the distribution of mutations across the genomic chunks in each state at the end of each logo. We ranked the mutations similarly as with the snapshot repair, and represented the ranks of each chunk as scatters, with the median and 25th and 75th percentiles marked with the state-indicated color point and box. The repair states were then sorted and numbered from the highest average count of mutations to the lowest.

To exemplify this representation, we present two CPD TT subsamples of repair state logos from two sides of the spectrum: the one with the most (RS1) and least (RS12) mutations (Figure 4.26).

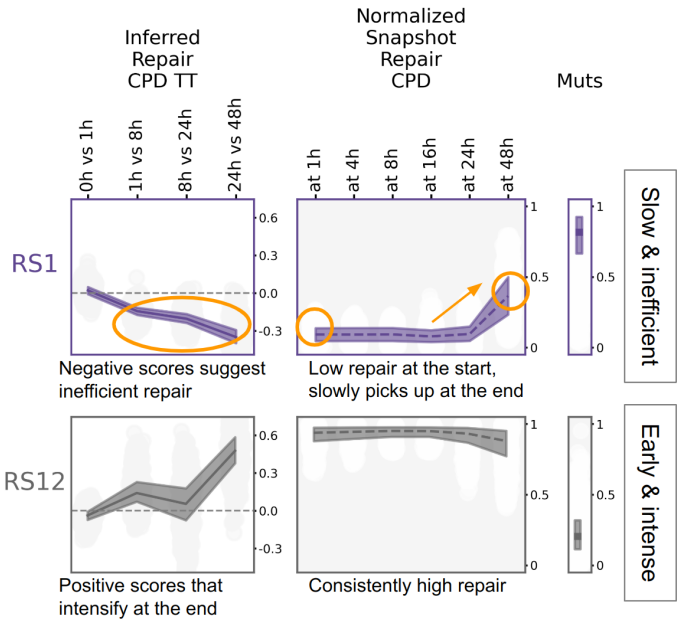

Figure 4.26: *Subsample example of state logos, for CPD TTs and states RS1 and RS12, with explainers guiding the interpretation. For inferred repair, the divergence scores of each chunk are represented on the Y-axis. For snapshot repair and mutations, ranks scaled to the 0-1 range are represented on the Y-axis. Scatterplots represent values of chunks, darker lines indicate the median, and shaded bound mark the 25th and 75th percentiles of values.*

Figure 4.27 presents the logos of all repair states in the human genome numbered RS1 through RS12. The three states represented at the top of the figure possess clear differences in their logos, reflective of their underlying DNA repair dynamics: RS1 shows the lowest DNA repair intensity among repair states, which becomes more apparent over time and begins to increase only at the end of the measured period;

RS2 also presents low repair intensity, with the chunks ranking hovering around 0 throughout all the intervals; and RS3 while similar to RS2 has repair picking up the tempo earlier and stronger.

Apart from the differences between the states, we want to note the variability of repair activity within some of them. By 'within' we mean the distinctive ways in which each damage type or different di-pyrimidines are repaired. For most of the states, the logos for the snapshot repairs deviate slightly between the two damage types. 6-4PPs have shorter time points (max 4h versus 48h for CPDs), which might partially explain this, like in the case of RS2 where the shorter 6-4PP snapshot course could form part of the longer one of the other damage type. Nevertheless, this is not as clear for all of the states. For RS4, the snapshot repair of 6-4PPs clearly trails down, while for CPDs it is fairly maintained until 16h, to undergo a short bump of activity before going down. This discrepancy makes sense in the light of two NER pathways and their preferences: although there are overwhelmingly more CPDs, the global NER's high affinity for more toxic 6-4PPs may result in less intense overall repair of CPDs at certain time points, unless detected by TC-NER during transcription.

When focusing on repair differences of di-pyrimidines, a few conclusions emerged. In the case of 6-4PPs, the logos of inferred repair for both di-pyrimidines within the state are nearly indistinguishable. One can notice a similar trend of very high similarity for TT di-pyrimidine logos between both damage types. This would suggest that CPDs in TTs and 6-4PPs in TTs and TCs get repaired in a highly similar fashion. However, we know (see the paragraph above) that there are small differences in the way the two damage types are processed by the snapshot repair. It is possible that repair inferred as divergence is less sensitive than the direct measurement of snapshot DNA repair, as subtractive HS-damage-seq is considered a less sensitive method of measuring repair than XR-seq [19]. Additionally, XR-seq takes all di-pyrimidines into account at once, and we see some slight differences in the repair of CPDs of the two di-pyrimidines (TT and CT). Hence, we trust the first conclusion, going in line with the literature: 6-4PPs and CPDs are repaired in slightly different manners.

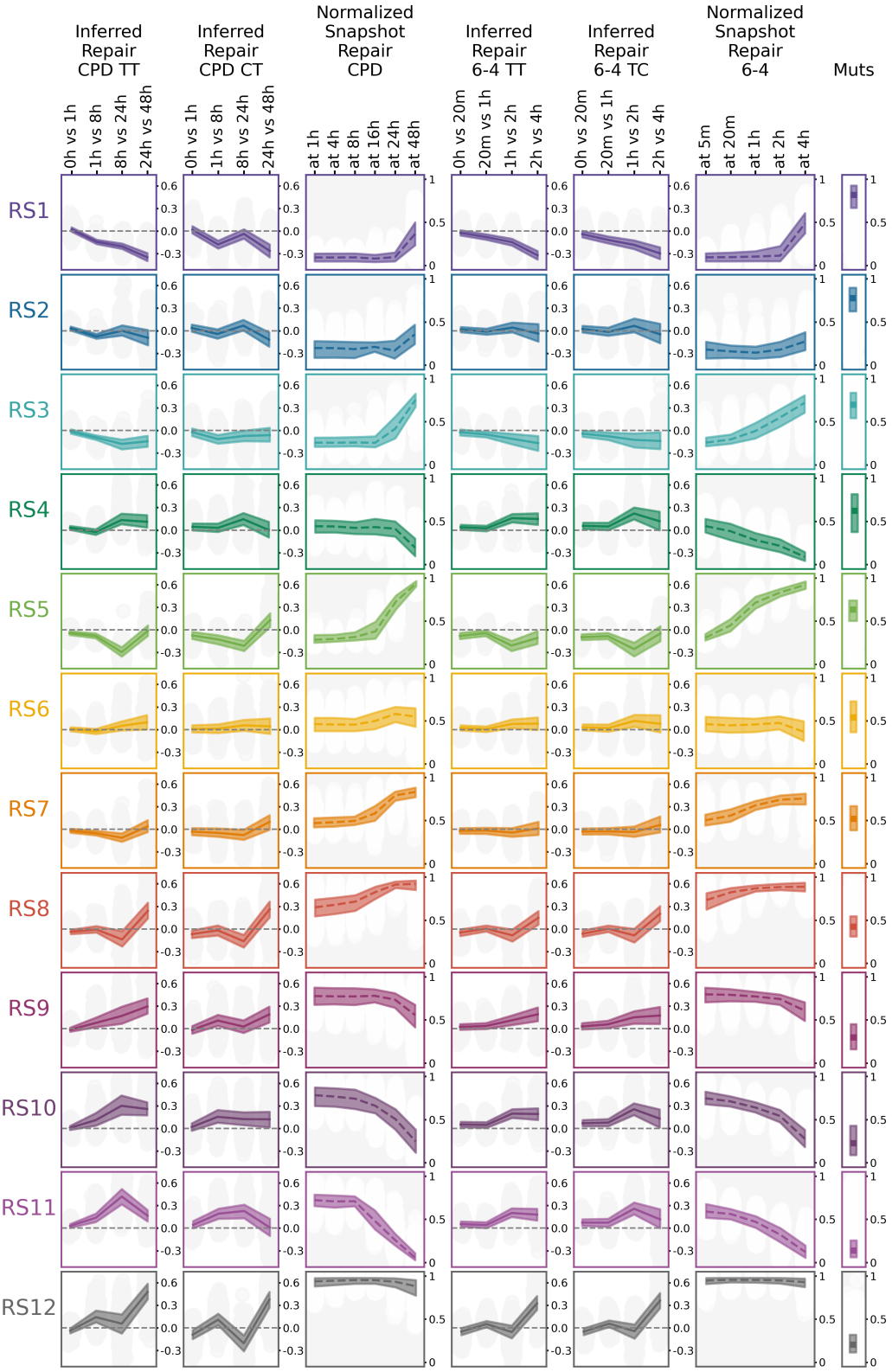

Figure 4.27: *Logos of all 12 repair states, across all repair data modalities, with the mutation rank on the right. For inferred repair, the divergence scores of each chunk are represented on the Y-axis. For snapshot repair and mutations, ranks scaled to the 0-1 range are represented on the Y-axis. Scatterplots represent values of chunks, darker lines indicate the median, and shaded bound mark the 25th and 75th percentiles of values.*

As mentioned, in some cases we could notice the differences in the inferred repair of TT and CT CPD di-pyrimidines. The most obvious example would be RS11, where TT repair intensity rises sharply early after exposure, and then decreases moderately, while CTs experience a lower intensity of repair - the bump is softer and flatter. This could reflect underlying differences in the preference of NER for the repair of CPDs of the two types of di-pyrimidines.

4.2.3.1. The mutation rate of repair states relates closely to their repair activity

We noticed that the spectrum of the repair states visually roughly follows the order established by the distribution of mutations. As expected, the states at the top – with more mutations – seem to have a later repair onset, as well as less intense repair overall, in stark contrast to the ones at the bottom. We set out to measure how well the order by mutations corresponds to the repair efficiency (Table 4.2).

We needed to encode the repair efficiency in a simple manner, to be able to order the repair states following this variable. To do so, for each box of a logo, we calculated the mean of each data track (ranks in case of snapshot repair) and summed them for each state. The states were sorted from the lowest sum – representing the repair activity in a given box – to the highest. Then, we compared the produced orderings to the one by mutations. We used the Jaro-Winkler metric, which takes into account transpositions (Table 4.2). While most of the orderings resulting from individual boxes were quite similar to the one established by mutations, interestingly, the snapshot repair of 6-4PP differed the most. Overall, the orderings by the efficiency of repair of CPDs corresponded better to the one by mutations, suggesting a higher contribution of this damage type to UV mutation formation.

We also wanted to understand the frequency of transitions between each pair of the final repair states. We calculated the number of transitions to each state and weighted it by the total size of that state, to be plotted as a heatmap (Figure 4.28). First, we noted that as expected, each repair state ‘sticks’ to itself - tends to transition more to itself than other states. A second observation was that, quite interestingly, repair states at the very edges of the ordering (either with very high or very low respective amounts of mutations) transition very little to other states. This seems especially true for the mutationally high states (RS1-RS2). On the other hand, more transitions between different types of states are observed for repair states at the middle of the mutation rate spectrum, mainly between each other. It seems that ‘extreme’ low activity states form larger, connected stretches, and are harder for the repair to get into, and get out of.

All the results presented so far support the claim that the repair states reflect genuine

Logo box	Ordering	Jaro-Winkler metric
Inferred repair CPD TT	RS1, RS3, RS5, RS7, RS2, RS8, RS6, RS4, RS9, RS12, RS10, RS11	0.89
Inferred repair CPD CT	RS1, RS5, RS3, RS7, RS2, RS8, RS6, RS12, RS4, RS9, RS10, RS11	0.89
Inferred repair 6-4PP TT	RS1, RS5, RS3, RS7, RS8, RS2, RS6, RS12, RS9, RS4, RS11, RS10	0.86
Inferred repair 6-4PP TC	RS1, RS5, RS3, RS7, RS2, RS8, RS6, RS12, RS9, RS4, RS11, RS10	0.86
Snapshot repair CPD	RS1, RS2, RS3, RS4, RS5, RS6, RS11, RS7, RS10, RS9, RS8, RS12	0.96
Snapshot repair 6-4PP	RS1, RS10, RS11, RS12, RS2, RS3, RS4, RS5, RS6, RS7, RS8, RS9	0.67

Table 4.2: *Orderings of states by various states logo boxes, and their corresponding Jaro-Winkler metric comparing them to the ordering by mutations. In the used implementation, the Jaro-Winkler score of 0 represents no match, and 1 is a perfect match.*

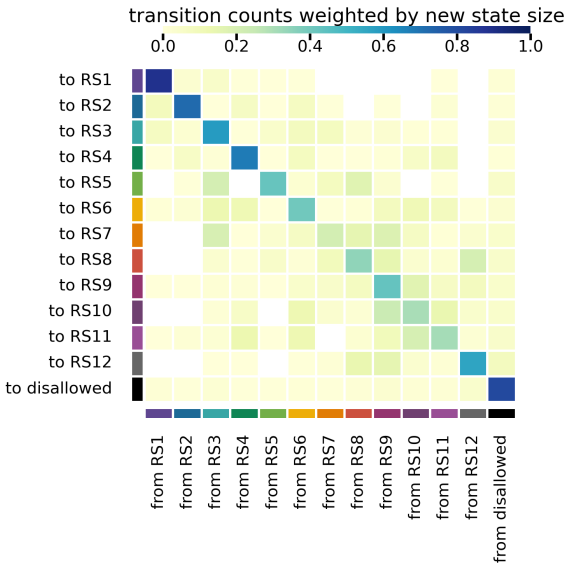

Figure 4.28: *Heatmap of transition frequencies between the 12 repair states and the disallowed one.*

differences in terms of the DNA damage repair activity operative across the genome. Consequently, we deem our approach successful at the first of the goals: partitioning the genome into distinct repair regimes.

4.2.4. Repair states distribution across genomic regions

We were curious about how the repair states distribute along the genome. Is there a dominant repair state? Do some states cover specific genomic elements?

To further characterize the repair states, we checked the percentage of the genome each of them covered (Figure 4.29). While some states covered a larger portion of the genome (RS1, RS9), the overall genome coverage by repair states was surprisingly homogenous (as compared to e.g. chromatin states). When repeating the same analysis over each chromosome separately (Appendix, Figure B.10) we noted similar distributions of repair states across chromosomes. The noteworthy exception was the higher RS12 coverage of chromosomes 19 and 22. These chromosomes are characterized by small size, high gene density, and, importantly, a central location within the nucleus [94].

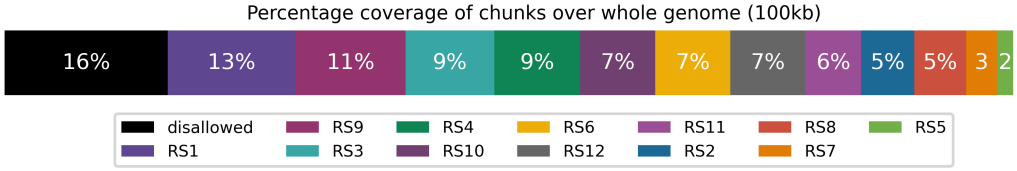

Figure 4.29: *Percentage coverage of the human genome by repair states, in 100kb chunks.*

Next, we sought to explore the states at the gene scale. We calculated fold enrichments of different elements from RefSeq data, as well as CpGIslands, within the state with respect to the genome as a whole (Figure 4.30). Repair states with more intense repair (RS8-RS12) were enriched for exons, and transcription start and end sites, with RS12 showing the highest enrichment. RS12 stood out compared to other intensely repaired states due to the high enrichment in CpG islands, associated with promoter sequences. Interestingly, RS5 seemed to exhibit some enrichment for these regions as well.

We reasoned that the observed relationship between these repair states and transcriptionally active regions is underpinned by transcription-coupled NER. Thus, we next sought to explore the contributions of both NER pathways to the repair states. To this end, we exploited publicly available XR-seq data [18] of two mutant cell lines: XPC (Xeroderma Pigmentosum, TC-NER proficient) and CSB (Cockayne Syndrome, TC-NER deficient, global NER active), together with a normal cell line (NHF1). The snapshot repair of these three cell lines was measured at 1h after UV exposure. We pre-processed the dataset in the same manner as the rest of the XR-seq data. Finally, we compared the TT-corrected counts of snapshot UV-damage repair for the three cell lines across repair states (Figure 4.31).

The first observation derived from this comparison was that the differences in repair intensity observed across repair states appear to be driven, primarily, by TC-NER. Global NER also showed a trend (albeit much smaller) of increase that correlates with

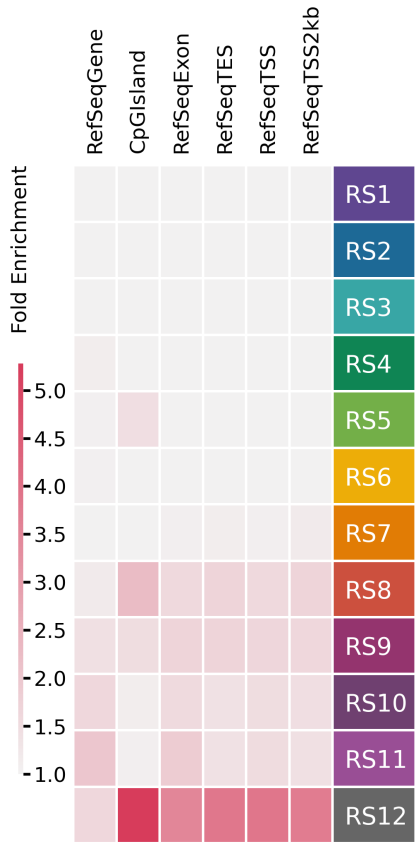

Figure 4.30: *Fold enrichments of various genomic regions (RefSeq and CpGIslands) in repair states as compared to the whole genome.*

the increase in overall repair intensity across repair states. This was quite surprising, as the original study using this same data found the repair in CSB cells to be uniform along the genome. This upward trend could be potentially explained by recent studies suggesting that global NER can be also stimulated by transcription [95, 96]. The trend, while present, was slightly less prominent for 6-4PPs in CSB and NHF1 cells. States between RS4 and RS9 'broke out' of the trend. This might be a subtle reflection of the fact that 6-4PPs are predominantly repaired by the global pathway.

In RS9-12 states, the snapshot repair exhibited by TC-NER was markedly higher than the one by the global pathway. The situation was reversed for the other states, although the effect was subtle. This goes in line with the enrichments of RS9-12 in exons, TSS, and TES regions (Figure 4.30), and might suggest a higher contribution of TC-NER to the repair activity of these states. (This did not appear to be the case for RS8, also enriched for these elements.) Interestingly, RS12 seemed to have the highest repair activity of all states, regardless of the NER status of the cell. This suggests that the regions of the genome covered by this highly efficient repair state are intensely repaired

by both pathways of NER.

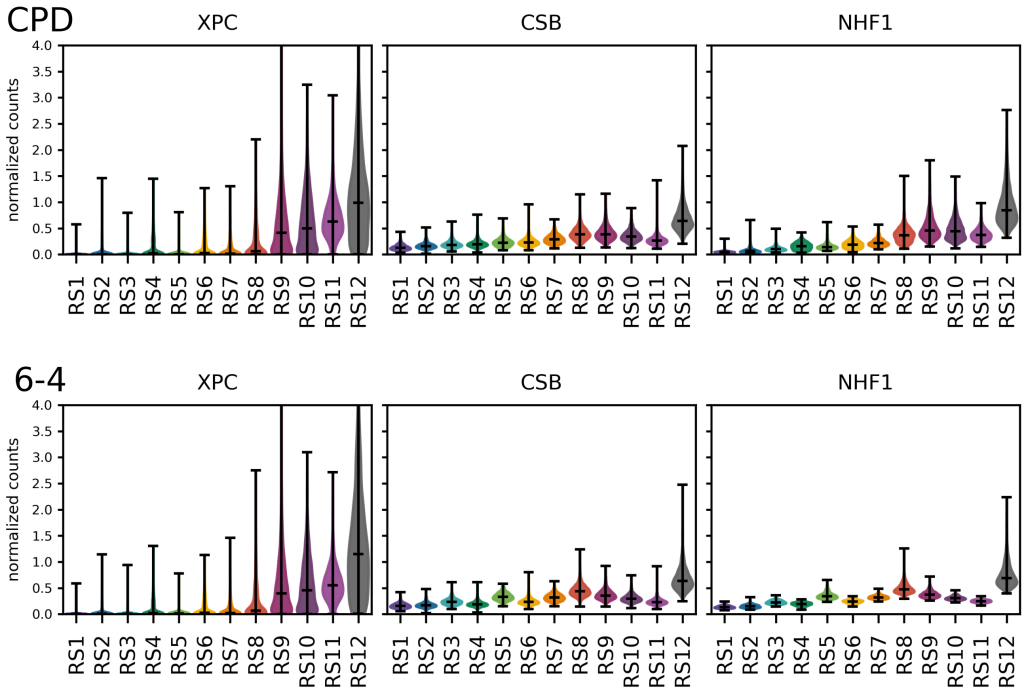

Figure 4.31: *TT-normalized score of snapshot repair at 1h across states for three cell lines: XPC (TC-NER proficient) and CSB (global NER proficient) mutants, and NHF1 (proficient in both NER pathways). Plotted values were cut off at 4 for ease of comparison, and as most of the data concentrated within this window. The plots with the full data range can be seen in Appendix, Figure B.11*

In summary, the partition of the human genome by repair states recapitulates our knowledge of the two NER pathways, in particular with respect to likely transcribed regions.

4.2.5. Repair states reflect the underlying features of the genome

To further explore the biological relevance of repair states, we set out to study their degree of overlap with several features of the genome. The three different sets of features we investigated were: chromatin states and marks, replication timing data, and other epigenetic features (related to expression and chromatin structure). All of the features were taken from the closest cell type available (outlined in the methods). First, we calculated a score for each feature tracking its representation in each genomic chunk. (Usually, that score was a simple overlap. Specific ways of scoring each feature can be found in 3.3.5.) Then, we calculated the fold enrichment of the feature within the chunks assigned to the specific state with respect to the genome as a whole.

4.2.5.1. Influence of chromatin states and histone marks

It is known that chromatin states influence repair in different ways [19]. It would be useful to know whether chromatin states might underlie the repair states.

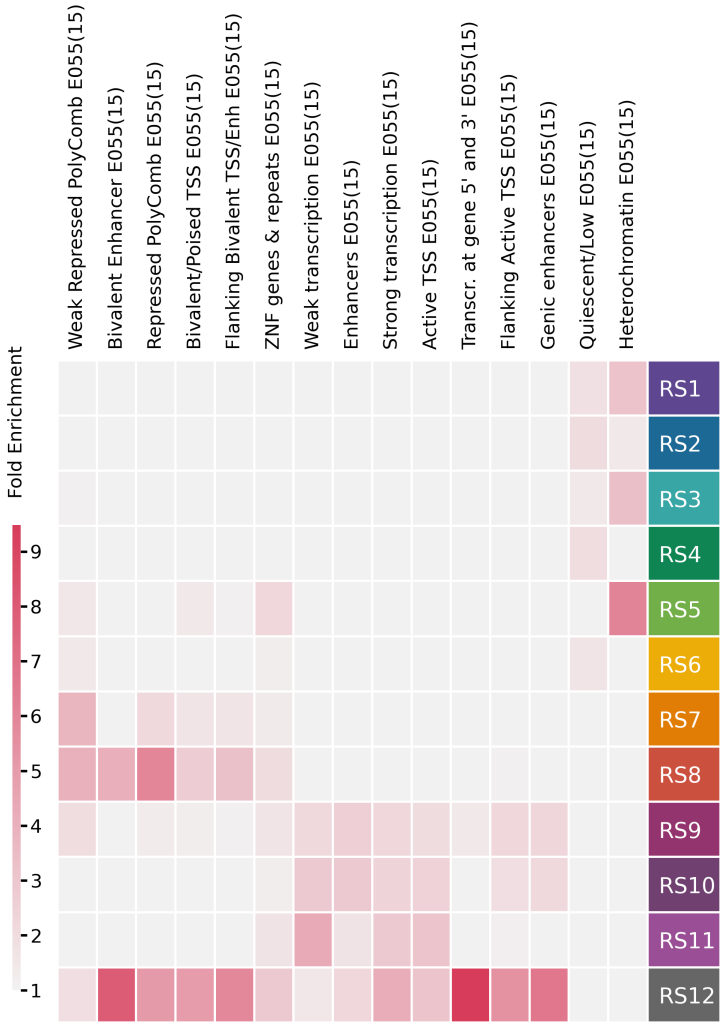

Figure 4.32: *Repair state fold enrichments of 15 canonical chromatin states from the E055 (Foreskin Fibroblast Primary Cells) Roadmap Epigenomics cell line.*

To this end, we explored the correspondence between repair states and chromatin states (Figure 4.32). We noticed a clear pattern of enrichment for Quiescent and Heterochromatin segments across repair states with less intense repair (RS1-6) Repressed and Bivalent chromatin states across repair states of medium repair intensity (RS5-9), and Transcribed and Enhancer regions across repair states with highest repair intensity (RS9-12). Interestingly, RS12 seemed to be strongly enriched for both the Transcribed and Enhancer as well as Repressed and Bivalent states. We did not observe a one-to-one correspondence of any repair state to the chromatin state, with some repair states showing enrichment for the same chromatin states. This suggests that while

chromatin states may constitute part of the underpinning of repair states (as established here and in [19]), more genomic features must be at play in their definition.

As chromatin states alone could not explain the repair states, we next examined the enrichments in histone marks (some used as inputs for the chromHMM chromatin states) across repair states. We gathered 18 histone mark-cell type combinations from ENCODE [87] and Roadmap Epigenomics [85] (Figure 4.33).

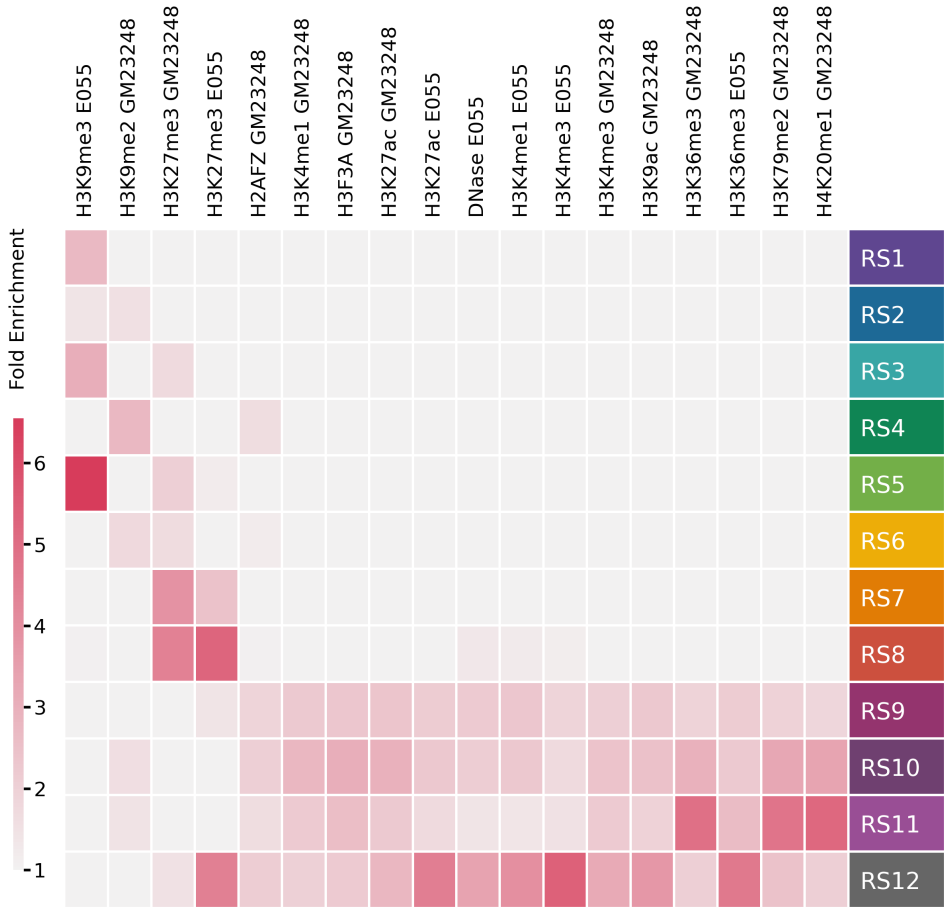

Figure 4.33: *Repair state fold enrichments of histone marks and DNase from the E055 (Foreskin Fibroblast Primary Cells) Roadmap Epigenomics and GM23248 (Arm Fibroblasts Primary Cells) ENCODE cell lines.*

Low repair activity was enriched in histone marks associated with heterochromatin, gene deserts, and gene expression silencing (H3K9me3, H3K9me2 [97]). Repair states with intermediate repair intensity were enriched in the H3K27me3 mark, related to transcriptional shutdown. The remaining repair states were enriched in a variety of histone marks related to chromatin accessibility, transcription, gene bodies, and regulatory elements, in agreement with previous findings in works analyzing HS-damage-seq and XR-seq data [19, 23]. Interestingly, while RS8 appeared enriched

for expressed regions (Figure 4.30) – same as RS9-12 – it was clearly distinguished from the other gene-rich repair states by its enrichment for chromatin states (Figure 4.32) and chromatin marks (Figure 4.33) related to bivalent and repressed regions. This may explain the clearly lower normalized repair snapshot score of this state in the XPC mutant cell line (Figure 4.31). RS11 exhibited the strongest enrichments for HK20me1 (highly transcribed genes), H3K79me2 (transcriptionally active genes), and one of the H3K36me3 (exons) tracks. RS12 was also enriched for H3K36me3 (although measured from a different cell type), with additional contributions of one of the H3K4me1 tracks and H3K4me3 (TSS of actively transcribed genes, and priming for rapid gene activation), as well as H3K27me3 (transcriptional shutdown) and its antagonistic H3K27ac. Interestingly, the 'bivalent' occupation of the same locus by the H3K4me3 and H3K27me3 frequently happens at important developmental genes. In summary, the observations made on enrichment for histone marks across repair states extend the previous ones on chromatin states. Still, we could not define well all of the repair states. We wondered which other known covariates could help explain them. Next, we explored replication timing, expression, and features related to the chromatin structure inside the nucleus.

4.2.5.2. Influence of the replication timing

The timing of replication is one of the most critical determinants of the mutation rate [58, 55, 54, 72]. Active replication in the region is thought to promote DNA repair regardless of the phase of the cell cycle [77]. But early replicating genomic regions tend to be repaired more than late replicating areas [77]. We reasoned that replication timing should be one factor underpinning the genome segmentation on repair states. To this end, we checked the fold enrichments (Figure 4.34) of replication timing tags mapped across genomic chunks for interphase phases of the cell cycle (G1b, S1-4, and G2).

Replication timing corresponded closely to the repair states, with a 'fade' going through the consecutive phases of the cell cycle. All 4 states with intense repair (RS9-12) correlated with early replication (G1b and S1). RS12 exhibited high enrichment in G1b, which is considered very early replication as this phase of the cycle is focused on growth and preparation for the synthesis of DNA. RS1-3 on the other hand corresponded to late replication (S4 and G2). The least intense repair state RS1 was strongly enriched in delayed replication in the G2 phase (cell cycle phase focused on growth and preparation for cell division). Interestingly, one state stood out of this clear fade pattern - RS5, exhibiting later replication than the states around it. RS5 was also the most Heterochromatin-enriched repair state, and heterochromatic

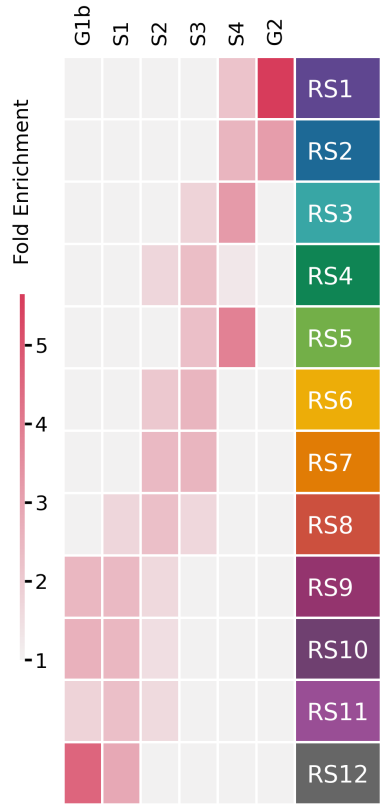

Figure 4.34: *Repair state fold enrichments of replication timing tags from the BJ (Foreskin Fibroblast Cells) ENCODE cell line.*

regions close to lamina tend to have later replication times [97]. Taking all these results into account, they suggest that the priority of DNA repair closely follows the priority of replication.

4.2.5.3. Influence of the chromatin 3D structure and expression

Next, we explored other features of the (epi)genome that did not fit the previous categories (Figure 4.35).

The first group (Figure 4.35) consisted of features related to the 3D structure of the chromatin inside the nucleus. LADs (lamina-associated domains) are regions of heterochromatin in close contact with the nuclear lamina. We noticed the enrichment for LADs across low and intermediate repair intensity repair states (RS1-6), which were also enriched for the Heterochromatin state (with the exception of RS4 and the addition of RS7). The rest of the states, exhibiting high repair intensity, are not enriched in LADs. This suggests a more centric placement of the parts of the genome covered by these repair states inside the nucleus. This is consistent with observations of increased activity of NER in nucleus-centric regions 2h after exposure [73].

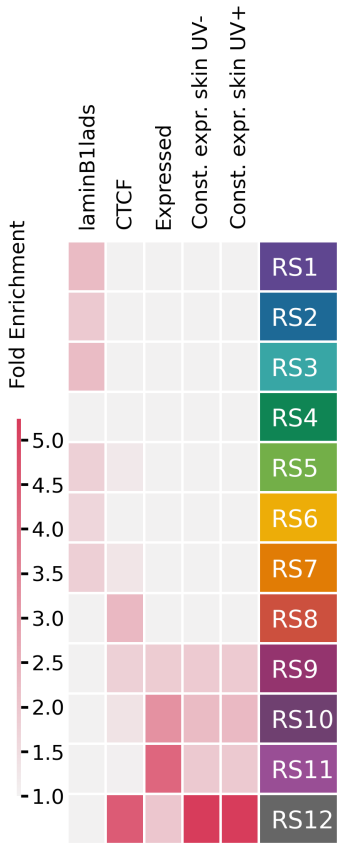

Figure 4.35: *Repair state fold enrichments of chromatin structure features (LADs and CTCF) and expression features (protein-coding expression, and constitutively expressed genes locations in the skin under UV exposure or not). LADs were taken from Tig-3 (Lung fibroblasts cells) cell line, CTCF from the lower leg skin tissue in ENCODE. Expression in protein-coding genes was averaged over two patient juvenile skin fibroblast samples. Genes constitutively expressed in the skin were taken from GTEx data on the median gene-level expression by tissue.*

CTCF is a chromatin-binding factor that functions in the maintenance of the chromatin structure, with effects that can be both activation or repression of transcription, overall regulating and promoting interactions between regulatory elements such as promoters and enhancers. High enrichment in CTCF found in RS8, RS9, and especially RS12 may thus imply an overrepresentation of regulatory chromatin interactions across these three repair states.

The second group of features (Figure 4.35) related to transcriptional activity across repair states. We explored the overall expression of protein-coding genes in skin fibroblasts, as well as coverage of states by constitutively expressed genes in the same tissue. We found that intensely repaired states RS10 and RS11 were the most highly enriched for the overall expression. Curiously, RS12, the most intensely repaired state, appeared less enriched than RS11 and RS10 for expressed regions. Nevertheless, RS12

was the most highly enriched repair state for a set of genes that are constitutively expressed in the skin, irrespective of exposure to UV light.

In summary, the analysis of the overrepresentation of a number of genomic features across repair states reveals known associations linking less intensely repaired states, –bearing more mutations– to repressed, packed, inaccessible, close to the lamina and late replicating genomic regions. Conversely, as also known, intensely repaired states appear enriched for parts of the genome that are predominantly genic, expressed, transcribed, accessible, internal, and early replicating. Moreover, the gradient of repair states – representing an inferred order of UV damage repair intensity agnostic of any genomic feature – reproduces the spectrum of certain genomic features, such as replication timing, chromatin accessibility, and transcriptional activity. Interestingly, while these and other genomic features explain quite well the repair states partition, some features are less clear in this respect. The repair states might thus be regarded as a framework to systematically test the strength of the influence of genomic features with the intensity of DNA repair.

4.2.5.4. Feature composition of states

What makes a state? What tells one state apart from another? We wondered to what extent the repair states can be characterized by the patterns of enrichment in different sets of genomic features.

To address these questions we devised an approach to compare, in a hierarchical manner, the repair states against the enrichments in sets of genomic features. We focused on chromatin states, and for this analysis, we went back to the more granular 27 hidden states that make up the 12 repair states.

First, we sought to assess the fidelity of the correspondence between the hidden states and the chromatin states. We were seeking to answer two questions: do the smaller hidden states belonging to a given repair state reproduce the same enrichment pattern? Do different enrichment patterns converge to the same repair state? We represented the chromatin state enrichments for the 27 hidden states together with the repair-activity-based clustering dendrogram (Figure 4.36). The hidden state numbers were colored by the repair state color. We found that in general the hidden states that cluster together based on repair activity (and especially make up the same final repair state) tend to be enriched in the same features. However, we also noticed that the chromatin state enrichments do sometimes fail to separate distinct hidden states in this visual analysis. This ambiguity underscores a limitation for a complete explainability of the repair activity in terms of chromatin state enrichments alone.

Figure 4.36: *Hidden state fold enrichments of 15 canonical chromatin states from the E055 (Foreskin Fibroblast Primary Cells) Roadmap Epigenomics cell line, with the repair-dynamics-based hidden state clustering dendrogram. Hidden states are colored by the repair state they are clustered into.*

Next, we sought to determine the combination of chromatin state enrichment features that best characterize each hidden state. We conceptualized our solution as follows. We seek paths that traverse the dendrogram (representing the hierarchical clustering based on repair activity) starting at the root of the dendrogram and ending at each hidden state leaf. Such a path can be described as a sequence of binary decisions as to which of the two possible downward edges to choose, through all traversed nodes (hereto "junctions"). These two edges at the junction always define two sets of leaves (hidden states). Thus, at each junction, we want to know how to make a decision to reach a set of hidden states with a specific state of interest. To do so, we compared the enrichments of the two sets of hidden states separated by each junction. For each chromatin state one by one, we calculated the means of enrichment in the two sets. Next, we calculated the difference of the means in the group. Once the differences

of the means were calculated for each chromatin state, we selected the top ones. As a result, for each hidden state, we provide a collection of features that represent the characteristic chromatin state biases consistently with the hierarchical structure that underlies the repair dynamics.

We represented the results atop the dendrogram (Figure 4.37). At each junction, the top 3 features with the absolute difference of means above 1.5 were highlighted. The arrow indicates the directionality of the sign of the difference at the junction - e.g. for the topmost junction, the group of states on its left side is more enriched (positive difference of means) in Bivalent Enhancer than the right-side group.

The annotated dendrogram clearly visualized the relationships between the chromatin states and the hierarchy of repair states, at various levels. In most cases, for the hidden states clustering into the same repair state, there was no feature significantly separating them under the used threshold. This suggests a high coherence of the repair states not only in repair dynamics but also in chromatin state composition. However, there are 5 cases where separation could be observed, namely RS9, RS10, RS12, RS8, RS3. When inspecting the highlighted features (Figure 4.37) together with the hidden state enrichments (Figure 4.36), we could see the coherence between the two pictures.

We believe that this simple method helps capture the between- and in-state differences better than inspecting the enrichments by eye. The analysis presented here provides a systematic way to find a representative set of features (chromatin states) alongside their importance that best explains the repair states consistently within their underlying hierarchical structure. We plan to extend this analysis to include more features in the future and delve deep into characterizing each repair state.

Finally, we close this chapter, concluding that we find the spectrum of repair activity to be closely related to the spectrum of features of the human genome in matching tissue types. This way we fulfilled our second goal. Not only was the presented segmentation of the human genome by repair successful, but we also found it to be biologically meaningful.

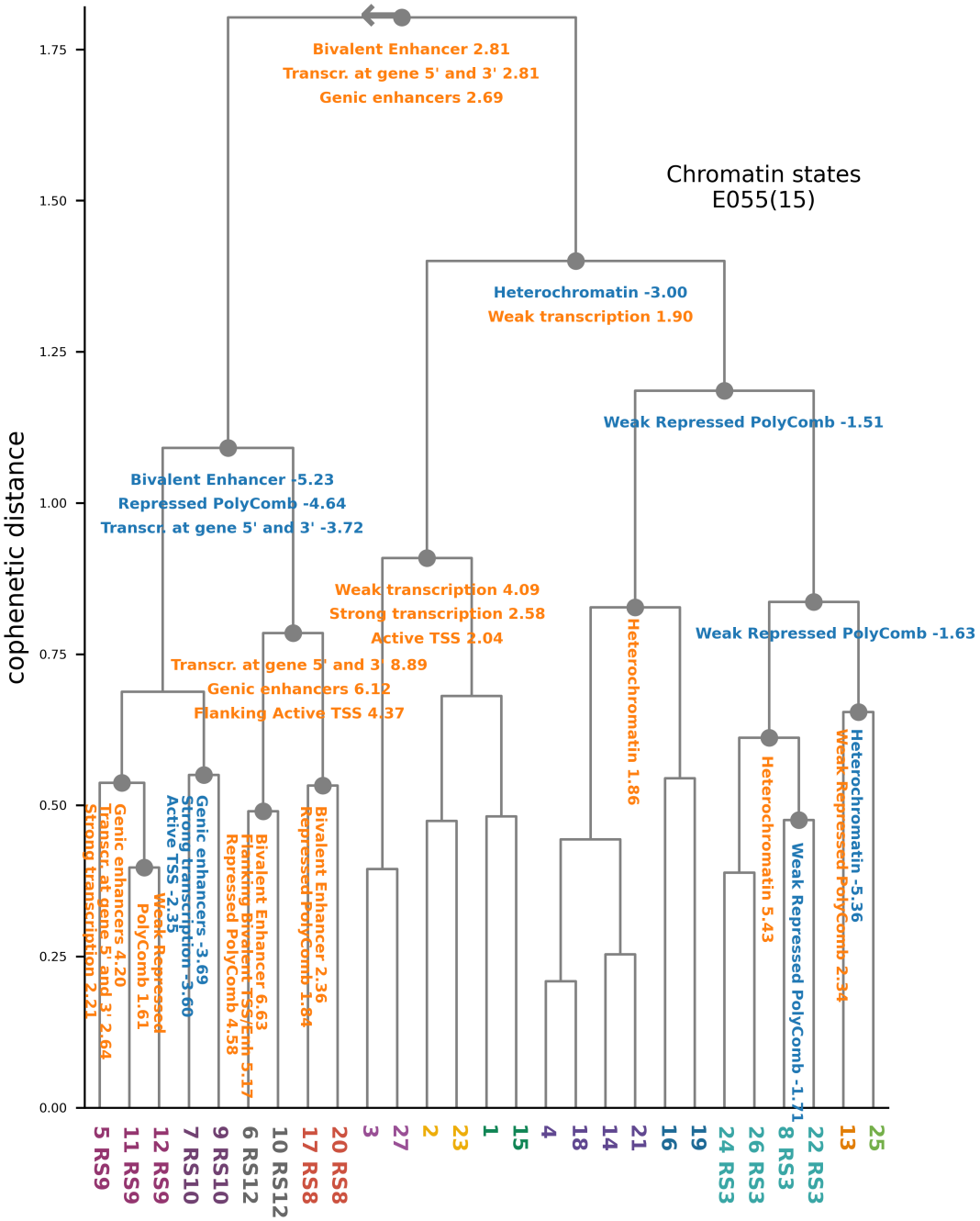

Figure 4.37: *Junction analysis annotated atop repair-dynamics-based hidden state clustering dendrogram. At each junction, a difference of the means of the two groups of hidden states below is calculated, using the fold enrichments of 15 canonical chromatin states from the E055 (Foreskin Fibroblast Primary Cells) Roadmap Epigenomics cell line. Hidden states are colored by the repair state they are clustered into. Arrow indicates that the direction related to the sign of the difference (meaning higher enrichments) printed is always the left branch.*

5. Discussion

In this chapter, I first discuss the two projects separately and conclude with a common discussion of the implications of this thesis as a whole.

5.1. Alkylation damage mapping in human cells

Many chemotherapeutic agents still in use are alkylators - they act by overwhelming the cancer cells with alkylating damage. This is due to the fact that cancer cells frequently are deficient in specific DNA repair mechanisms. Many healthy, DNA repair-proficient cells, likely get destroyed by chemotherapy too, but they can generally withstand this damage better. Even so, if they survive, they can be left with a mutational footprint [65, 13]. Some of these mutations have been associated with a subsequent clonal expansion and malignant transformation [14]. Moreover, alkylators can be also found in nature, both endo- and exogenous [5, 20, 98]. Hence, studying the impacts of exposure to alkylating agents on both cancerous and healthy cells in the body has implications for our understanding of basic cell biology, as well as therapy development and use.

The AB-seq damage mapping approach presented in this thesis, while closely inspired by yeast NMP-seq [24], is the first applied to mapping alkylating damage in human cells. When comparing the results from the two protocols, we found some discrepancies. The frequencies of modified bases were the same, with expected modifications happening predominantly in G's and A's. The most salient differences were in the sequence contexts of the damaged positions. While the frequencies of damaged bases detected by AB-seq varied across trinucleotide contexts –relative to the frequency of trinucleotides in the genome– the damaged bases found by NMP-seq were distributed fairly uniformly across trinucleotide contexts. We speculated on the few possible causes of this deviation.

The causes might be biological - due to differences between the model systems used in the two protocols: yeast and human cells. Since our analysis of context-specific damage takes into account the normalization by the genomic background, the effect of the nucleotide genome composition should not be at play here. One possible explanation for the reported disparity could be differences related to how the damage interacts with DNA and how the repair mechanisms operate in the two species.

There might be a divergence in repair mechanics charged with cleaning up alkylating lesions, strong enough to leave an observable mark during the treatment time. In both species, BER is known as one of the main alkylation-repair mechanisms [5, 20, 98], contributions of NER have been reported in yeast, and so far remain not known in humans [24]. The two species also differ in the epigenome (e.g. usage of the chromatin marks; some can exist in one species and not in another [99, 100]). Hypothetically, this could have an effect on damage deposition and repair mechanics as well.

The disparity could have also originated due to a technical difference between the two protocols. In NMP-seq, cells were stored frozen prior to DNA extraction, while AB-seq samples were processed immediately after exposure to the alkylating agent. Importantly, the concentrations and lengths of treatments differed. In NMP-seq, two MMS concentrations were used: 0.2% (23.6 mM) and 0.4% (47.2 mM). In AB-seq we used less: 10mM. Additionally, NMP-seq treatment was only 10 minutes long, while we exposed the cells to MMS for 30 minutes. The treatment regimen in AB-seq was then roughly 5 times weaker, but 3 times longer. (However, one should also note that yeast cells have a cell wall as a protective mechanism, while eukaryotic cells do not.) This might have caused differences in how many lesions can get deposited on the DNA, and where. Furthermore, the longer treatment time allows for more repair to happen while the damage is still being deposited.

These causes for the protocol differences can be further explored by introducing some variability in the AB-seq. We could modify treatment times and doses, or generate maps in cells deficient in various repair mechanisms.

NMP-seq focused only on the damage induced by a model alkylator MMS. While convenient for protocol development, MMS is not in use in treatment anymore due to its potency and following toxicity. We applied AB-seq to map damage from both MMS and Temozolomide (frequently used in the treatment of e.g. glioblastoma). Our results constitute the very first map of TMZ damage ever, opening the opportunity to explore its unknown damage formation. Although we do not have previous damage data to compare with, there is a mutational signature potentially associated with TMZ treatment - SBS11. We found the coherence between the trinucleotide profile of this signature and that of the damaged bases to be surprisingly high. Given the fact that before a lesion has a chance to become a mutation there are several steps at which it may be repaired, one could expect the final mutational pattern to differ more. (Even more so when considering that m6G is the most mutagenic modification – hence likely highly contributing to SBS11 – introduced by TMZ, which we do not map in AB-seq,

as discussed further below.) Notwithstanding, this high coherence between SBS11 and AB-seq damage patterns might suggest that the mechanisms of repair of this type of damage do not have very different preferences across trinucleotide contexts.

Having mapped damage from two sources of alkylation, we had a unique opportunity to compare them. The lesions that both sources induced were, as expected, highly enriched for G's and A's, corresponding to expected m7G and m3A lesions. Interestingly, the G/A ratios of the lesions were different between the two agents. In MMS, we mapped significantly more G's than A's than for TMZ. There might be a technical reason explaining this - TMZ treatment was only 5mM, while MMS was two times more concentrated. There might be a biological reason contributing as well, because of the metabolism differences of the two drugs. MMS is a direct methylator, while TMZ first needs to be metabolized by chemical hydrolysis to an actively methylating molecule (MTIC, and further AIC and a methyldiazonium cation [21]). Their chemistry is also slightly different, TMZ being an S_N1-type methylator, and MMS an S_N2-type one [9]. Their modes of action might thus differ. We believe this to be highly likely, as we also observe different trinucleotide context preferences between the two agents.

There are a few important limitations regarding the AB-seq damage maps. First, the backgrounds mapped in DMSO and untreated samples are not flat, even though we do not expect any damage. We have explored different strategies for blocking unspecific damage to improve this, but we could not successfully resolve this issue or figure out the source for those unspecific reads. Secondly, these backgrounds in controls clearly contribute to the signal observed in the treated samples. While the contributions are negligible for damage in G's, they might impair fine-grained analyses of the infrequent damages in A's. One possible strategy to circumvent this limitation and provide an estimate of the amount of true damage per trinucleotide consists in decomposing the trinucleotide profile of damage counts as a weighted sum of two signals, the inferred damage and the observed background, from which the probability that a given damage site at a specific trinucleotide is true damage can be computed. Therefore, we could assign a damage-or-not probability for each trinucleotide context. This relatively simple rationale can be extended to consider not one but several background processes that might operate simultaneously – although with different intensities – in each treated/untreated case, using techniques that have been previously applied to model mutagenesis as a mixture of elementary mutational processes [65, 70].

Other considerations pertain to types of lesions induced by alkylating agents that are

not currently mapped by AB-seq. The most toxic lesion induced by many alkylating agents is likely m6G [20, 98, 9]. m6G is repaired by direct reversal by MGMT [20, 98, 5], making it difficult to map with enzymatic approaches such as AB-seq. Moreover, so far AB-seq was used for mapping damage from mono-functional alkylators. Bi-functional alkylators (such as platinum-based drugs), apart from adding alkyl groups to a single base, can also form bonds between two different bases [20, 5]. These very disruptive and helix-distorting lesions are inter- and intra-strand crosslinks (ICLs). Due to their complicated structure, BER cannot deal with this type of damage, rendering AB-seq – which relies on BER-related lesion processing for its recognition – useless for their detection. AB-seq in its current form is applicable only to mono-alkylating lesions. Mapping inter or intra-strand lesions left by bi-functional alkylators will probably require a very different approach.

This project has many future plans and applications and constitutes an important direction for the lab. The version presented here focused on 2 agents and 1 time point, in 1 cell line. These dimensions are being actively expanded. We are planning to test multiple alkylating agents, both mono- and bi-functional. We are already extending the MMS and TMZ maps to multiple recovery time points after exposure. We are preparing to map the damage in TMZ-sensitive and resistant colorectal cancer cell lines and in repair-pathway-mutated cells. We would like to also follow the alkylator-treated cells until the mutation formation. This multidimensionality will allow us to study the alkylating damage formation, repair activity, mutation variability, and alkylator differences from many angles.

Finally, a compelling and crucial set of computational analyses relating the mapped TMZ and MMS damage formation with features of the genome is out of the scope of this thesis. Nevertheless, we believe the AB-seq protocol work presented here serves as a crucial first step and unfolds the path to many future feature analyses. One of the naturally following directions would be using the alkylation data with the repair state discovery framework presented next.

5.2. UV repair state discovery

Mutagenesis, including damage formation and repair activity, has been extensively studied in the context of interactions with features of the genome (see 1.4). However, the idea presented in these studies of partitioning the genome by the features has been an important limitation of the approaches hitherto. Some features do not cover the whole genome, leading to uncharacterized gaps in the analysis. Moreover, certain

features may be overlooked, due to lack of prior knowledge. Developing approaches that are not constrained by known features is a crucial next direction in the study of the determinants of DNA repair and the mutagenic process.

The 'repair states' framework presented in this thesis proposes, for the first time, an ab initio partitioning of the genome not biased by any prior knowledge of features of the genome. Thus, this repair-data-driven partitioning does not require any feature information. We start only with the genome itself - and the repair activity along it.

Encoding repair activity is no simple feat. We do not have a measure of total repair, but either snapshots of repair activity in a given moment or damage landscapes at different time points. While some measure of total repair can be inferred from the consecutive damage landscapes, as explained in the Results, this has its important caveats. We needed to perform several normalizations, calculate divergences between each two time points, and still take care when interpreting the inferred repair scores. Additionally, the approach we used required binarized data, which we provide by calculating the quantiles of each data track. However, although this ranking is computationally convenient, there are a few aspects worth keeping in mind. The first, obvious one is the information loss - instead of seeing all the scores' values, we get 5 categories representing each data track only. Moreover, in the case of inferred repair, when the original damage landscapes are close to each other, we will obtain a small range of divergence scores, reflecting little repair activity happening. However, the quantile labels will continue to underscore differences, even if the true effect between the top and bottom chunks is very small. Thus, caution should be taken when interpreting inferred repair divergence scores and quantile encodings between close time points, or for long recovery times when there is little damage left. Moreover, this means that not all the inferred repair tracks carry the same amount of information; it might be worth considering for the choice of time points of damage maps in the future.

We successfully segmented the genome into 12 repair states. Repair states are mathematical abstractions obtained from the repair dynamics of the genome, and are reflective of them. (By 'repair dynamics' we mean how differential repair activity changes through time after exposure). We use repair states to understand these genomic differences in the dynamics of the repair process.

The repair states did not only exhibit differences between each other but also a degree of variability in the repair of the different damage types and contexts within the same state. It is important to keep in mind that we segmented a continuous process. Thus, one can see the repair states as a spectrum of the continuum of repair intensities. A

strong argument for this is that states with similar dynamics tend to transition between each other more frequently than more different states. It is important to acknowledge that the repair states presented here were built at a certain resolution (100kb). This means that this segmentation likely captures larger aspects of repair dynamics, rather than smaller, higher-resolution ones, and needs to be interpreted as such.

With the feature-agnostic repair state segmentation done, we could comprehensively probe many genomic features for their distribution across repair states. We confirmed an inverse correlation between repair intensity and mutation rate across repair states. We found the repair state spectrum to be closely related to several genomic features. Low-intensity repair states correlated with heterochromatic, low-accessibility regions, undergoing late replication. This goes in line with the idea that NER, as a rather large protein complex, might be hindered by its size for accessing tightly-packed DNA [78, 23]. On the other hand, high-intensity repair states corresponded predominantly to areas with a high density of genes and regulatory regions, actively transcribed, and early-replicating. These repair states showed a higher intensity of TC-NER repair than all others, linking the results with the mechanistic explanation. Finally, repair states with intermediate repair intensity were enriched for bivalent, repressed, and intermediate replicating regions. We observed the clear-cut correspondence between the order of replication and the intensity of NER repair.

Importantly, we established that repair states, and the differences in repair activity they represent, are not a simple reflection of chromatin states. Chromatin states vary in size, with the Quiescent state covering most of the genome [82, 84], while repair states are distributed quite uniformly. This suggests that repair states are rather governed by a combination of many features, including chromatin states and others (e.g. replication timing), that together converge into this fairly homogeneous higher-level landscape. In summary: there is no single genomic feature that underlies the repair dynamics; rather, what we uncovered with the repair states are the composites of various features.

Even after including all features explored in this thesis, some repair states seemed indistinguishable from each other. This could be explained in two ways. Similar states could simply be an artifact of the model, matching some slight differences in the dynamics, when the underlying biological mechanics of the two states are the same. We believe this explanation is not very likely, given our repair states reproducibility analyses from independent replicates. Alternatively, we just might not know yet the genomic features that distinguish the repair states from each other.

Many of the aforementioned repair and feature relations are already established in the

literature. Even though we start with repair, instead of features, we arrive at similar observations. We successfully recapitulated them, taking an entirely different approach - this means we added consistency and coherence to existing knowledge.

Ideally, we would like to produce repair states at different levels of resolution, especially in smaller chunk sizes. However, currently, there is one important limitation to this: the achievable resolution hinders studying the very fine-grained features (e.g. nucleosome positioning) known for their considerable impact on repair. This limitation is only temporary though, as its source is not in the framework itself, but in the sparsity of available data. We believe this limitation can be instead presented as an opportunity for developing more damage maps with higher resolution as deeper sequencing damage data is produced. This data is bound to appear shortly, as panel-like approaches utilizing deep sequencing have been recently presented [101, 102]. With time and further developments, we will be able to probe smaller chunks and likely reveal smaller-scale repair states and their interactions with fine-grained features. Additionally, the repair states analysis at a smaller scale could benefit from using strand-resolved information. Currently, we aggregate all the input data over both strands. With smaller chunks, we could extend the models by separating the strand information by the activity of transcription or sense of replication. Knowing the impact of these two features, this could bring interesting new insights into the repair state analysis.

5.3. General considerations

Studying repair and its interplay with various features of the genome, and correlations with mutations fuels the general development of genomics and advances knowledge of cell biology. Thanks to AB-seq, for the first time we have insight into alkylation damage (and soon repair) in human cells. Next, AB-seq damage maps can be employed for different types of analyses. This would include already developed approaches [63, 64] (more outlined in 1.4) as well as the repair states.

There are quite a few things to contemplate when re-purposing the repair states framework for the analysis of AB-seq data. First, the damage types are different - instead of helix-distorting bulky lesions, there are small methyl or ethyl groups added to a single base. Additionally, the nature of the source of the damage is widely different. UV pulse is very time-constrained and short, and once taken away, there should be no new damage appearing. (Of note, this is a simplification, as there are reports of so-called dark CPDs that seem to occur long after the UV exposure

[15, 16, 17]). Alkylators are chemical agents that need to enter the nucleus, some of them need to get metabolized, and finally need to be washed off. This makes the exact moment when the damage stops being generated by chemical agents hard to pinpoint. Due to these differences in kinetics, the chemical treatments are longer. What we consider the '0h' time point, supposed to represent the damage formation, is likely already affected by repair. Moreover, these different damage types are repaired by different pathways. NER is a large complex with an elaborate damage recognition system. BER recognition is simpler and relies on two single enzymes. Differences between the damage types and repair machinery need to be carefully considered and suitable adjustments need to be made to the framework. If successfully applied to AB-seq data, repair states analysis would open the possibility of comparisons of repair states from these widely different conditions.

Of note, we do not plan to directly map the intensity of alkylation repair, as has been done for UV damage with XR-seq. Due to the nature of the BER repair, this is a challenging problem. The implication of this is that we would only have the repair inferred from alkylating lesions mapped through time. While not presented in this thesis, we did previously generate the UV repair states without the snapshot repair. The results were coherent, although we did observe a loss of quality of the repair states partitioning. This will be something we will have to take into account.

There are other damage and repair maps publicly available (some outlined in 1.3.2), although lacking the time-resolved component. Once there are more maps generated at different time points, the repair states framework could be employed to analyze them. It would be intriguing to compare repair states produced for two different damaging agents, whether or not they share the same repair mechanism. On the other hand, to compare the contributions of different repair pathways to the repair of one specific damage type, we would need to map the repair of the damage under various repair-deficient backgrounds. This happens to be planned for AB-seq damage maps, intertwining the contributions of the two projects closely.

An important factor to consider when using damage and repair mapping to study mutagenesis is the differences in the setting. In UV damage/repair mapping, cells are treated with only one, short (10-20s) UV-C light pulse [18, 23, 19]. The main components of sunlight that we are most exposed to are UV-A and UV-B [5]. While UV-C is the most mutagenic of the three, its contributions in everyday sunlight exposure are negligible due to the ozone layer filtering [5]. While to a lesser extent, UV-B, and less so UV-A, also lead to di-pyrimidine photoproduct formation, of similar

types [5]. One cannot ignore the potential differences in the damage formation that come from the difference in the energy of the radiation. Moreover, mutations are usually produced by repeated, long-term exposures. Our cells are exposed to damaging UV light every day, with varying intensities throughout. In the case of chemotherapy, the single treatment usually lasts for hours and is repeated multiple times over the course of a few weeks. Likely, the mechanics of damage formation and repair activity are intertwined in time and space and affect each other. Additionally, DNA Damage response is reported to be regulated by the circadian clock, meaning that DNA repair might act slightly differently depending on the time of day [103]. The experimental settings for both damage mapping datasets include only a single, and considerably shorter treatment, albeit with a stronger dose of the mutagen. This is useful for separating the process to understand the components of mutagenesis in an easier setting. In the future, there should be a focus on understanding these crucial exposure components of the mutagenic process. This could be done by generating damage and repair maps under varying exposures - both in terms of repetitions, as well as lengths. Nevertheless, we believe that the current damage maps are highly useful already. They can be thought of as probability distributions for mutational generation, supported by increasing correlations with mutations after exposure to the mutagen, that we presented for UV.

6. Conclusions

We believe that both projects constituting this thesis advance our study of how different mechanisms of DNA repair interact with basic processes of the cell, such as DNA replication, transcription, and chromatin structure maintenance. Both projects have opened interesting future paths for the study of the determinants of the DNA damage and dynamics of DNA repair of different types of lesions. We present the following conclusions in the two projects:

Alkylating damage maps

- We have presented AB-seq (Alkylation BER sequencing), an end-to-end genome-wide nucleotide-precision method for mapping alkylation, comprising an experimental DNA-damage capture library preparation with a computational pipeline to precisely map the captured damage sites to the human genomic sequence.
- The analysis of AB-seq libraries for two alkylating agents, MMS and TMZ, rendered enrichments in bases corresponding to the most prevalent known alkylating lesions (m7G, m3A).
- Methylation maps for MMS and TMZ uncover slight differences in the sequence context preferences of the two agents, specifically in tri-nucleotides, suggesting potentially different paths to damage formation.
- These methylation maps open the door for analyses of the dynamics of repair of the damage generated by these two alkylating agents, and the determinants underlying both the deposition of methyl groups and their repair. In particular, combined with the use of mutant human cells, these maps will shed light on the contribution of different DNA repair systems to the correction of alkylation damage.

UV Repair states

- We presented a novel approach to partition the genome according to the dynamics of DNA repair, named DNA repair states.
- Applying this approach to time-course UV-light damage and repair maps, we obtained UV damage DNA repair states, that represent genomic regions with differences in their kinetics of repair of two types of UV-induced photoproducts.
- The two NER pathways contribute to the differentiation between repair states,

with an important contribution of both to DNA repair states with the highest repair intensity.

- We showed that the spectrum of repair intensity of DNA repair states clearly anti-correlates with the rate of UV-generated mutations observed in melanomas, especially for CPDs.
- DNA repair states with low repair intensity are enriched for repressed, packed, inaccessible, close to the lamina and late replicating parts of the genome. Conversely, high-intensity repair states are enriched for predominantly genic, expressed, transcribed, accessible, internal, and early replicating parts of the genome.
- Nevertheless, we show that the DNA repair states obtained at this resolution (100Kb) are not explained by a single, or simple combination of a few genomic features; rather, the interaction between several genomic features underlies the differences in repair intensity across DNA repair states.
- New sets of genomic features can be systematically probed for their association with the intensity of DNA repair using these DNA repair states.
- The repair states framework here could be used for the analysis of the repair of other types of DNA damage, as well as analysis of smaller, local-scale determinants, once higher-coverage damage maps become available.

7. Bibliography

[1] Friedberg EC. DNA damage and repair. Nature. 2003 Jan;421(6921):436–440. Number: 6921 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature01408.

[2] Lindahl T. Instability and decay of the primary structure of DNA. Nature. 1993 Apr;362(6422):709–715. Available from: https://www.nature.com/articles/362709a0.

[3] Ganai RA, Johansson E. DNA Replication—A Matter of Fidelity. Molecular Cell. 2016 Jun;62(5):745–755. Publisher: Elsevier. Available from: https://www.cell.com/molecular-cell/abstract/S1097-2765(16)30140-X.

[4] Giglia-Mari G, Zotter A, Vermeulen W. DNA Damage Response. Cold Spring Harbor Perspectives in Biology. 2011 Jan;3(1):a000745. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3003462/.

[5] Chatterjee N, Walker GC. Mechanisms of DNA damage, repair, and mutagenesis. Environmental and Molecular Mutagenesis. 2017;58(5):235–263. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/em.22087. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/em.22087.

[6] Vijg J. From DNA damage to mutations: All roads lead to aging. Ageing Research Reviews. 2021 Jul;68:101316. Available from: https://www.sciencedirect.com/science/article/pii/S1568163721000635.

[7] Kow YW. Repair of deaminated bases in DNA. Free Radical Biology and Medicine. 2002 Oct;33(7):886–893. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0891584902009024.

[8] Peng Y, Pei H. DNA alkylation lesion repair: outcomes and implications in cancer chemotherapy. Journal of Zhejiang University Science B. 2021 Jan;22(1):47–62. Available from: https://doi.org/10.1631/jzus.B2000344.

[9] Fu D, Calvo JA, Samson LD. Balancing repair and tolerance of DNA damage caused by alkylating agents. Nature Reviews Cancer. 2012 Jan;12(2):104–120. Available from: https://www.nature.com/articles/nrc3185.

[10] Krokan HE, Bjørås M. Base Excision Repair. Cold Spring Harbor Perspectives in Biology. 2013 Apr;5(4):a012583. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3683898/.

[11] Delhomme TM, Munteanu M, Buonanno M, Grilj V, Biayna J, Supek F. Proton and alpha radiation-induced mutational profiles in human cells. bioRxiv; 2023. Pages: 2022.07.29.501997 Section: New Results. Available from: https://www.biorxiv.org/content/10.1101/2022.07.29.501997v2.

[12] Keshava N, Ong Tm. Occupational exposure to genotoxic agents. Mutation Research/Reviews in Mutation Research. 1999 Sep;437(2):175–194. Available from: https://www.sciencedirect.com/science/article/pii/S1383574299000836.

[13] Pich O, Muiños F, Lolkema MP, Steeghs N, Gonzalez-Perez A, Lopez-Bigas N. The mutational footprints of cancer therapies. Nature Genetics. 2019 Dec;51(12):1732–1740. Number: 12 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41588-019-0525-5.

[14] Pich O, Cortes-Bullich A, Muiños F, Pratcorona M, Gonzalez-Perez A, Lopez-Bigas N. The evolution of hematopoietic cells under cancer therapy. Nature Communications. 2021 Aug;12(1):4803. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-021-24858-3.

[15] Noonan FP, Zaidi MR, Wolnicka-Glubisz A, Anver MR, Bahn J, Wielgus A, et al. Melanoma induction by ultraviolet A but not ultraviolet B radiation requires melanin pigment. Nature Communications. 2012 Jun;3:884. Available from: https://www.nature.com/articles/ncomms1893.

[16] Premi S, Wallisch S, Mano CM, Weiner AB, Bacchiocchi A, Wakamatsu K, et al. Chemiexcitation of melanin derivatives induces DNA photoproducts long after UV exposure. Science. 2015 Feb;347(6224):842–847. Publisher: American Association for the Advancement of Science. Available from: https://www.science.org/doi/10.1126/science.1256022.

[17] Lawrence KP, Delinasios GJ, Premi S, Young AR, Cooke MS. Perspectives on Cyclobutane Pyrimidine Dimers—Rise of the Dark Dimers†. Photochemistry and Photobiology. 2022;98(3):609–616. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/php.13551. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/php.13551.

[18] Hu J, Adar S, Selby CP, Lieb JD, Sancar A. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. Genes & Development. 2015 Jan;29(9):948–960. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. Available from: http://genesdev.cshlp.org/content/29/9/948.

[19] Hu J, Adebali O, Adar S, Sancar A. Dynamic maps of UV damage formation and repair for the human genome. Proceedings of the National Academy of Sciences of the United States of America. 2017 Jun;114(26):6758–6763. Available from: https://www.pnas.org/doi/full/10.1073/pnas.1706522114.

[20] Drabløs F, Feyzi E, Aas PA, Vaagbø CB, Kavli B, Bratlie MS, et al. Alkylation damage in DNA and RNA—repair mechanisms and medical significance. DNA Repair. 2004 Nov;3(11):1389–1407. Available from: https://linkinghub.elsevier.com/retrieve/pii/S156878640400148X.

[21] Zhang J, F G Stevens M, D Bradshaw T. Temozolomide: Mechanisms of Action, Repair and Resistance. Current Molecular Pharmacology. 2012 Jan;5(1):102–114. Available from: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1874-4672&volume=5&issue=1&spage=102.

[22] Li W, Sancar A. Methodologies for detecting environmentally induced DNA damage and repair. Environmental and Molecular Mutagenesis. 2020;61(7):664–679. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/em.22365. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/em.22365.

[23] Adar S, Hu J, Lieb JD, Sancar A. Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. Proceedings of the National Academy of Sciences. 2016 Apr;113(15):E2124–E2133. Publisher: Proceedings of the National Academy of Sciences. Available from: https://www.pnas.org/doi/10.1073/pnas.1603388113.

[24] Mao P, Brown AJ, Malc EP, Mieczkowski PA, Smerdon MJ, Roberts SA, et al. Genome-wide maps of alkylation damage, repair, and mutagenesis in yeast reveal mechanisms of mutational heterogeneity. Genome Research. 2017 Oct;27(10):1674–1684. Available from: https://genome.cshlp.org/content/27/10/1674.long.

[25] Hashimoto S, Anai H, Hanada K. Mechanisms of interstrand DNA crosslink repair and human disorders. Genes and Environment. 2016 May;38(1):9. Available from: https://doi.org/10.1186/s41021-016-0037-9.

[26] Alberts B. Molecular biology of the cell. Sixth edition ed. New York, NY: Garland Science, Taylor and Francis Group; 2015.

[27] Spivak G. Nucleotide excision repair in humans. DNA repair. 2015 Dec;36:13–18. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4688078/.

[28] Vaisman A, Woodgate R. Translesion DNA polymerases in eukaryotes: what makes them tick? Critical Reviews in Biochemistry and Molecular Biology. 2017 Jun;52(3):274–303. Available from: https://www.tandfonline.com/doi/full/10.1080/10409238.2017.1291576.

[29] Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JHJ. Understanding nucleotide excision repair and its roles in cancer and ageing. Nature Reviews Molecular Cell Biology. 2014 Jul;15(7):465–481. Number: 7 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nrm3822.

[30] Lehmann AR, McGibbon D, Stefanini M. Xeroderma pigmentosum. Orphanet Journal of Rare Diseases. 2011 Nov;6(1):70. Available from: https://doi.org/10.1186/1750-1172-6-70.

[31] Hudson (Chairperson) TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. Nature. 2010 Apr;464(7291):993–998. Number: 7291 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature08987.

[32] Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics. 2013 Oct;45(10):1113–1120. Number: 10 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/ng.2764.

[33] Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science. 2015 Sep;349(6255):1483–1489. Available from: https://www.science.org/doi/10.1126/science.aab4082.

[34] Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. High burden and pervasive positive selection of somatic mutations in normal human skin. Science. 2015 May;348(6237):880–886. Publisher: American Association for the Advancement of Science. Available from: https://www.science.org/doi/10.1126/science.aaa6806.

[35] Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 2000 Jan;100(1):57–70. Publisher: Elsevier. Available from: https://www.cell.com/cell/abstract/S0092-8674(00)81683-9.

[36] Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011 Mar;144(5):646–674. Publisher: Elsevier. Available from: https://www.cell.com/cell/abstract/S0092-8674(11)00127-9.

[37] Aitken SJ, Anderson CJ, Connor F, Pich O, Sundaram V, Feig C, et al. Pervasive lesion segregation shapes cancer genome evolution. Nature. 2020 Jul;583(7815):265–270. Available from: https://www.nature.com/articles/s41586-020-2435-1.

[38] Sloan DB, Broz AK, Sharbrough J, Wu Z. Detecting Rare Mutations and DNA Damage with Sequencing-Based Methods. Trends in Biotechnology. 2018 Jul;36(7):729–740. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0167779918300763.

[39] Bohm KA, Wyrick JJ. Damage mapping techniques and the light they have shed on canonical and atypical UV photoproducts. Frontiers in Genetics. 2023;13. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2022.1102593.

[40] García-Nieto PE, Schwartz EK, King DA, Paulsen J, Collas P, Herrera RE, et al. Carcinogen susceptibility is regulated by genome architecture and predicts cancer mutagenesis. The EMBO Journal. 2017 Oct;36(19):2829–2843. Publisher: John Wiley & Sons, Ltd. Available from: https://www.embopress.org/doi/full/10.15252/embj.201796717.

[41] Amente S, Scala G, Majello B, Azmoun S, Tempest HG, Premi S, et al. Genome-wide mapping of genomic DNA damage: methods and implications. Cellular and Molecular Life Sciences. 2021 Nov;78(21):6745–6762. Available from: https://doi.org/10.1007/s00018-021-03923-6.

[42] Salk JJ, Kennedy SR. Next-Generation Genotoxicology: Using Modern Sequencing Technologies to Assess Somatic Mutagenesis and Cancer Risk. Environmental and Molecular Mutagenesis. 2020;61(1):135–151. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/em.22342. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/em.22342.

[43] Mao P, Smerdon MJ, Roberts SA, Wyrick JJ. Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. Proceedings of the National Academy of Sciences. 2016 Aug;113(32):9057–9062. Publisher: Proceedings of the National Academy of Sciences. Available from: https://www.pnas.org/doi/10.1073/pnas.1606667113.

[44] Hu J, Lieb JD, Sancar A, Adar S. Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. Proceedings of the National Academy of Sciences. 2016 Oct;113(41):11507–11512. Publisher: Proceedings of the National Academy of Sciences. Available from: https://www.pnas.org/doi/full/10.1073/pnas.1614430113.

[45] Li W, Hu J, Adebali O, Adar S, Yang Y, Chiou YY, et al. Human genome-wide repair map of DNA damage caused by the cigarette smoke carcinogen benzo[a]pyrene. Proceedings of the National Academy of Sciences. 2017 Jun;114(26):6752–6757. Publisher: Proceedings of the National Academy of Sciences. Available from: https://www.pnas.org/doi/full/10.1073/pnas.1706021114.

[46] Cai Y, Cao H, Wang F, Zhang Y, Kapranov P. Complex genomic patterns of abasic sites in mammalian DNA revealed by a high-resolution SSiNGLe-AP method. Nature Communications. 2022 Oct;13(1):5868. Available from: https://www.nature.com/articles/s41467-022-33594-1.

[47] Sriramachandran AM, Petrosino G, Méndez-Lago M, Schäfer AJ, Batista-Nascimento LS, Zilio N, et al. Genome-wide Nucleotide-Resolution Mapping of DNA Replication Patterns, Single-Strand Breaks, and Lesions by GLOE-Seq. Molecular Cell. 2020 Jun;78(5):975–985.e7. Publisher: Elsevier. Available from: https://www.cell.com/molecular-cell/abstract/S1097-2765(20)30195-7.

[48] Cao H, Salazar-García L, Gao F, Wahlestedt T, Wu CL, Han X, et al. Novel approach reveals genomic landscapes of single-strand DNA breaks with nucleotide resolution in human cells. Nature Communications. 2019;10. Publisher: Nature Publishing Group. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6925131/.

[49] Canela A, Sridharan S, Sciascia N, Tubbs A, Meltzer P, Sleckman BP, et al. DNA Breaks and End Resection Measured Genome-wide by End Sequencing. Molecular Cell. 2016 Sep;63(5):898–911. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1097276516302921.

[50] Zatopek KM, Potapov V, Maduzia LL, Alpaslan E, Chen L, Evans TC, et al. RADAR-seq: A RAre DAmage and Repair sequencing method for detecting DNA damage on a genome-wide scale. DNA Repair. 2019 Aug;80:36–44. Available from: https://www.sciencedirect.com/science/article/pii/S1568786419301351.

[51] Lucas MC, Novoa EM. Long-read sequencing in the era of epigenomics and epitranscriptomics. Nature Methods. 2023 Jan;20(1):25–29. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41592-022-01724-8.

[52] Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012 Aug;488(7412):504–507. Number: 7412 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature11273.

[53] Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015 Feb;518(7539):360–364. Number: 7539 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature14221.

[54] Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. Human mutation rate associated with DNA replication timing. Nature Genetics. 2009 Apr;41(4):393–395. Number: 4 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/ng.363.

[55] Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 Jul;499(7457):214–218. Number: 7457 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature12213.

[56] Supek F, Lehner B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. Cell. 2017 Jul;170(3):534–547.e23. Publisher: Elsevier. Available from: https://www.cell.com/cell/abstract/S0092-8674(17)30774-2.

[57] Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. 2015 May;521(7550):81–84.

[58] Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, et al. Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation. American Journal of Human Genetics. 2012

Dec;91(6):1033–1040. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3516607/.

[59] Frigola J, Sabarinathan R, Mularoni L, Muiños F, Gonzalez-Perez A, López-Bigas N. Reduced mutation rate in exons due to differential mismatch repair. Nature Genetics. 2017 Dec;49(12):1684–1692. Number: 12 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/ng.3991.

[60] Fredriksson NJ, Elliott K, Filges S, Eynden JVd, Ståhlberg A, Larsson E. Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature. PLOS Genetics. 2017 May;13(5):e1006773. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006773.

[61] Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. Nature. 2016 Apr;532(7598):264–267. Number: 7598 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature17661.

[62] Pich O, Muiños F, Sabarinathan R, Reyes-Salazar I, Gonzalez-Perez A, Lopez-Bigas N. Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. Cell. 2018 Nov;175(4):1074–1087.e18. Publisher: Elsevier. Available from: https://www.cell.com/cell/abstract/S0092-8674(18)31312-6.

[63] Gonzalez-Perez A, Sabarinathan R, Lopez-Bigas N. Local Determinants of the Mutational Landscape of the Human Genome. Cell. 2019 Mar;177(1):101–114. Publisher: Elsevier. Available from: https://www.cell.com/cell/abstract/S0092-8674(19)30234-X.

[64] Supek F, Lehner B. Scales and mechanisms of somatic mutation rate variation across the human genome. DNA repair. 2019 Sep;81:102647. Available from: https://www.sciencedirect.com/science/article/pii/S1568786419302009.

[65] Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013 Aug;500(7463):415–421. Number: 7463 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature12477.

[66] Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Reports. 2013 Jan;3(1):246–259. Publisher: Elsevier. Available from: https://www.cell.com/cell-reports/abstract/S2211-1247(12)00433-0.

[67] Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. Nature Genetics.

2015 Dec;47(12):1402–1407. Available from: https://www.nature.com/articles/ng.3441.

[68] Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, et al. Mutational signatures associated with tobacco smoking in human cancer. Science. 2016 Nov;354(6312):618–622. Publisher: American Association for the Advancement of Science. Available from: https://www.science.org/doi/10.1126/science.aag0299.

[69] Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Research. 2019 Jan;47(Database issue):D941–D947. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6323903/.

[70] Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020 Feb;578(7793):94–101. Number: 7793 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41586-020-1943-3.

[71] Frigola J, Sabarinathan R, Gonzalez-Perez A, Lopez-Bigas N. Variable interplay of UV-induced DNA damage and repair at transcription factor binding sites. Nucleic Acids Research. 2021 Jan;49(2):891–901. Available from: https://doi.org/10.1093/nar/gkaa1219.

[72] Salvadores M, Supek F. Cell cycle alterations associate with a redistribution of mutation rates across chromosomal domains in human cancers. bioRxiv; 2022. Pages: 2022.10.24.513586 Section: New Results. Available from: https://www.biorxiv.org/content/10.1101/2022.10.24.513586v2.

[73] Akköse U, Adebali O. The interplay of 3D genome organization with UV-induced DNA damage and repair. Journal of Biological Chemistry. 2023 May;299(5). Publisher: Elsevier. Available from: https://www.jbc.org/article/S0021-9258(23)00321-6/abstract.

[74] Elliott K, Boström M, Filges S, Lindberg M, Eynden JVd, Ståhlberg A, et al. Elevated pyrimidine dimer formation at distinct genomic bases underlies promoter mutation hotspots in UV-exposed cancers. PLOS Genetics. 2018 Dec;14(12):e1007849. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007849.

[75] Mao P, Wyrick JJ. Organization of DNA damage, excision repair, and mutagenesis in chromatin: A genomic perspective. DNA Repair. 2019 Sep;81:102645. Available from: https://www.sciencedirect.com/science/article/pii/S1568786419301983.

[76] Heilbrun EE, Merav M, Adar S. Exons and introns exhibit transcriptional strand asymmetry of dinucleotide distribution, damage formation and DNA repair. NAR Genomics and Bioinformatics. 2021 Mar;3(1):lqab020. Available from: https://doi.org/10.1093/nargab/lqab020.

[77] Huang Y, Azgari C, Yin M, Chiou YY, Lindsey-Boltz LA, Sancar A, et al. Effects of replication domains on genome-wide UV-induced DNA damage and repair. PLOS Genetics. 2022 Sep;18(9):e1010426. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1010426.

[78] Duan M, Sivapragasam S, Antony JS, Ulibarri J, Hinz JM, Poon GM, et al. High-resolution mapping demonstrates inhibition of DNA excision repair by transcription factors. eLife. 2022 Mar;11:e73943. Publisher: eLife Sciences Publications, Ltd. Available from: https://doi.org/10.7554/eLife.73943.

[79] Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, et al. The International Cancer Genome Consortium Data Portal. Nature Biotechnology. 2019 Apr;37(4):367–369. Number: 4 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41587-019-0055-9.

[80] Fox EB, Sudderth EB, Jordan MI, Willsky AS. A sticky HDP-HMM with application to speaker diarization. The Annals of Applied Statistics. 2011 Jun;5(2A):1020–1056. Publisher: Institute of Mathematical Statistics. Available from: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-5/issue-2A/A-sticky-HDP-HMM-with-application-to-speaker-diarization/10.1214/10-AOAS395.full.

[81] Hughes MC, Stephenson WT, Sudderth E. Scalable Adaptation of State Complexity for Nonparametric Hidden Markov Models. In: Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc.; 2015. Available from: https://proceedings.neurips.cc/paper/2015/hash/2e65f2f2fdaf6c699b223c61b1b5ab89-Abstract.html.

[82] Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature Methods. 2012 Mar;9(3):215–216. Number: 3 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nmeth.1906.

[83] Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008 Jun;453(7197):948–951. Number: 7197 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature06947.

[84] Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nature Protocols. 2017 Dec;12(12):2478–2492. Number: 12 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nprot.2017.124.

[85] Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb;518(7539):317–330. Number: 7539 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature14248.

[86] Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nature biotechnology. 2010 Oct;28(10):1045–1048. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3607281/.

[87] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep;489(7414):57–74. Available from: https://www.nature.com/articles/nature11247.

[88] Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Research. 2020 Jan;48(D1):D882–D889. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7061942/.

[89] Kagda MS, Lam B, Litton C, Small C, Sloan CA, Spragins E, et al.. Data navigation on the ENCODE portal. arXiv; 2023. ArXiv:2305.00006 [cs, q-bio]. Available from: http://arxiv.org/abs/2305.00006.

[90] Hitz BC, Lee JW, Jolanki O, Kagda MS, Graham K, Sud P, et al.. The ENCODE Uniform Analysis Pipelines. bioRxiv; 2023. Pages: 2023.04.04.535623 Section: New Results. Available from: https://www.biorxiv.org/content/10.1101/2023.04.04.535623v1.

[91] Reemann P, Reimann E, Ilmjärv S, Porosaar O, Silm H, Jaks V, et al. Melanocytes in the Skin – Comparative Whole Transcriptome Analysis of Main Skin Cell Types. PLOS ONE. 2014 Dec;9(12):e115717. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115717.

[92] Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics. 2013 Jun;45(6):580–585. Number: 6 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/ng.2653.

[93] Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet Processes. Journal of the American Statistical Association. 2006 Dec;101(476):1566–1581. Available from: https://www.tandfonline.com/doi/full/10.1198/016214506000000302.

[94] Boyle S, Gilchrist S, Bridger JM, Mahy NL, Ellis JA, Bickmore WA. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. Human Molecular Genetics. 2001 Feb;10(3):211–220. Available from: https://doi.org/10.1093/hmg/10.3.211.

[95] Kusakabe M, Onishi Y, Tada H, Kurihara F, Kusao K, Furukawa M, et al. Mechanism and regulation of DNA damage recognition in nucleotide excision repair. Genes and Environment. 2019 Jan;41(1):2. Available from: https://doi.org/10.1186/s41021-019-0119-6.

[96] Zheng CL, Wang NJ, Chung J, Moslehi H, Sanborn JZ, Hur JS, et al. Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes. Cell Reports. 2014 Nov;9(4):1228–1234. Publisher: Elsevier. Available from: https://www.cell.com/cell-reports/abstract/S2211-1247(14)00900-0.

[97] Black JC, Whetstine JR. Chromatin landscape. Epigenetics. 2011 Jan;6(1):9–15. Publisher: Taylor & Francis _eprint: https://doi.org/10.4161/epi.6.1.13331. Available from: https://doi.org/10.4161/epi.6.1.13331.

[98] Klapacz J, Pottenger LH, Engelward BP, Heinen CD, Johnson GE, Clewell RA, et al. Contributions of DNA repair and damage response pathways to the non-linear genotoxic responses of alkylating agents. Mutation research Reviews in mutation research. 2016;767:77–91. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4818947/.

[99] White CL, Suto RK, Luger K. Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions. The EMBO Journal. 2001 Sep;20(18):5207–5218. Publisher: John Wiley & Sons, Ltd. Available from: https://www.embopress.org/doi/full/10.1093/emboj/20.18.5207.

[100] O'Kane C, Hyland E. Yeast epigenetics: the inheritance of histone modification states. Bioscience Reports. 2019 May;39(5):BSR20182006. Available from: https://doi.org/10.1042/BSR20182006.

[101] Elliott K, Singh VK, Boström M, Larsson E. Base-resolution UV footprinting by sequencing reveals distinctive damage signatures for DNA-binding proteins. Nature Communications. 2023 May;14(1):2701. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-023-38266-2.

[102] Selvam K, Sivapragasam S, Poon GMK, Wyrick JJ. Detecting recurrent passenger mutations in melanoma by targeted UV damage sequencing. Nature Communications. 2023 May;14(1):2702. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-023-38265-3.

[103] Ashok Kumar PV, Dakup PP, Sarkar S, Modasia JB, Motzner MS, Gaddameedhi S. It's About Time: Advances in Understanding the Circadian Regulation of DNA Damage and Repair in Carcinogenesis and Cancer Treatment Outcomes. The Yale Journal of Biology and Medicine. 2019 Jun;92(2):305–316. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6585512/.

A. Appendix to Alkylating Damage Maps

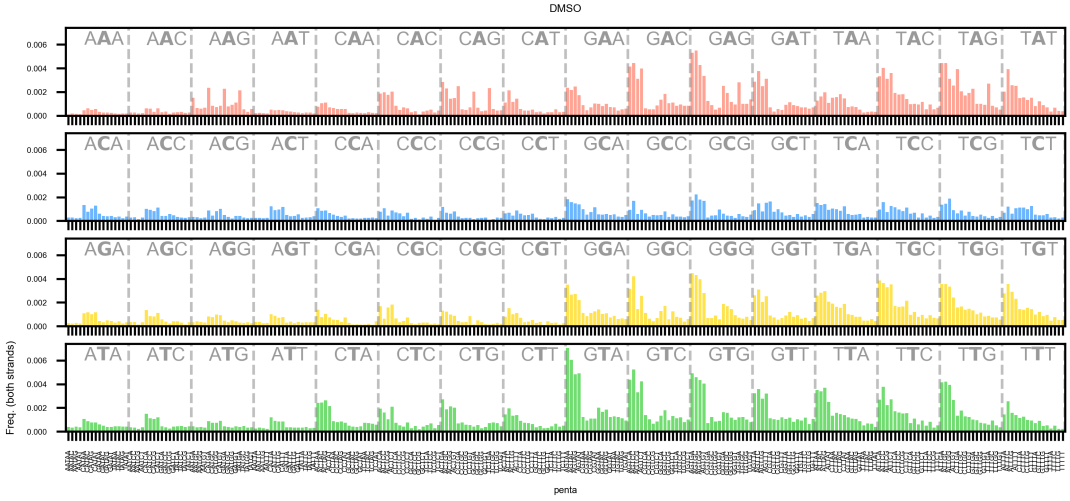

Figure A.1: *Pentanucleotide probability frequencies for the DMSO sample.*

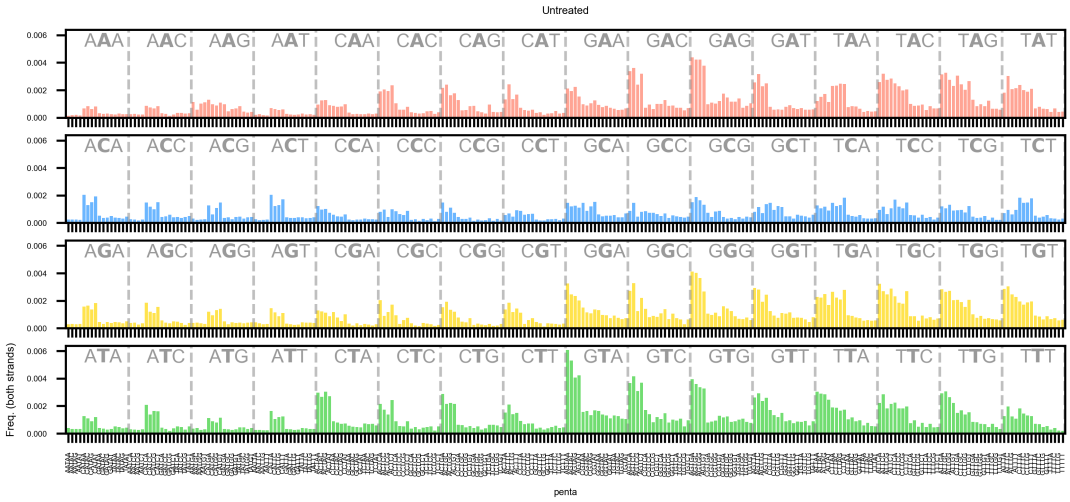

Figure A.2: *Pentanucleotide probability frequencies for the Untreated sample.*

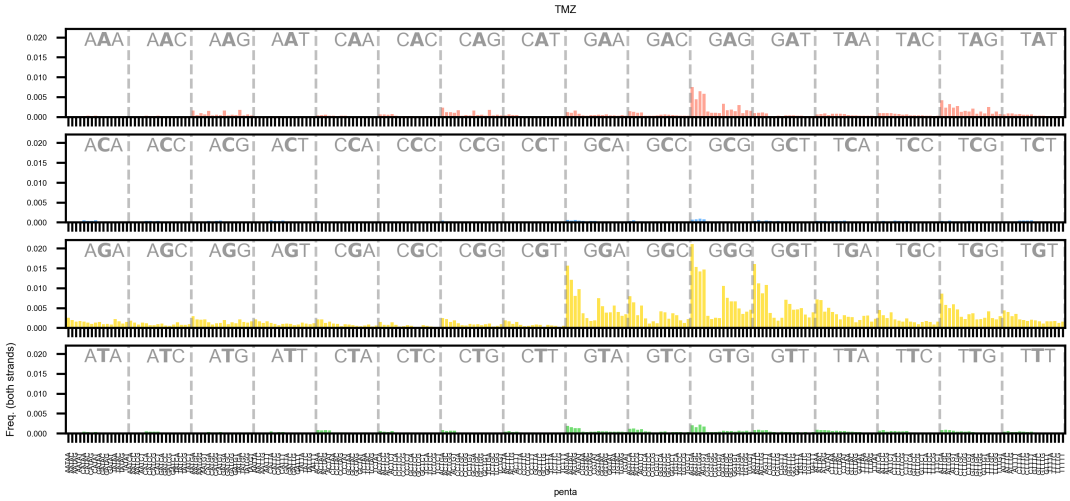

Figure A.3: *Pentanucleotide probability frequencies for the TMZ sample.*

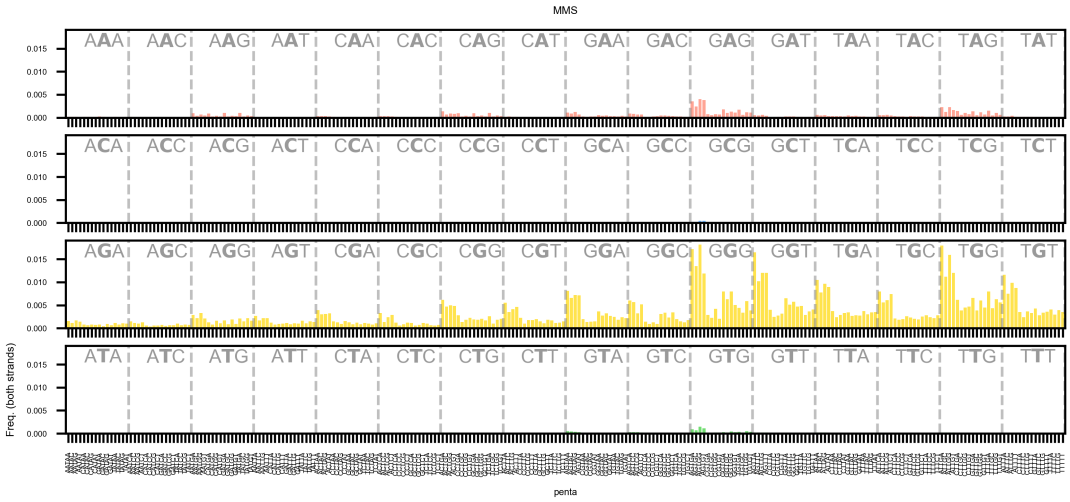

Figure A.4: *Pentanucleotide probability frequencies for the MMS sample.*

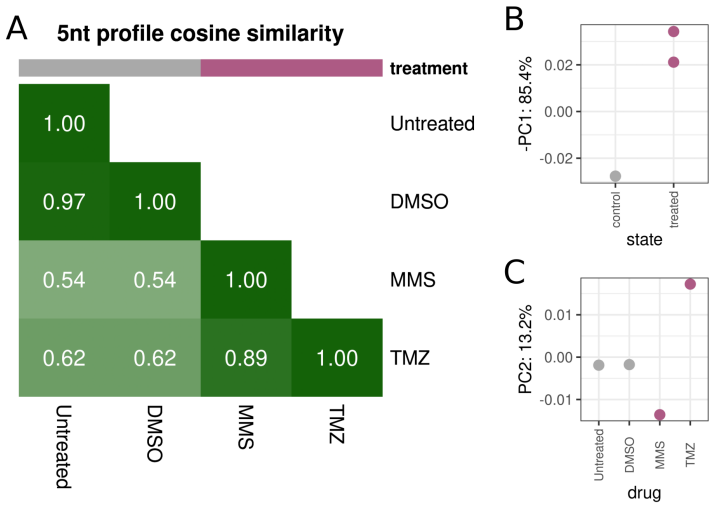

Figure A.5: *A: Cosine similarity, B and C: PCA of the pentanucleotide genome-normalized frequency profiles between the 4 AB-seq samples. For PCA, scores of the samples on the B: -PC1 or C: PC2 are represented.*

B. Appendix to UV Repair States

B.1. Damage and mutations correlations

	1h	8h	24h	48h
Normalized score CPD TT	0.86	0.24	-0.16	-0.34
Normalized score CPD CT	0.85	0.57	0.02	0.30
Raw counts CPD TT	0.98	0.88	0.47	0.45
Raw counts CPD CT	0.85	0.61	0.21	0.28

Table B.1: *Correlation coefficients of correlations between CPD 0h damage with other time points within 100kb chunks, for both normalized score and raw data counts.*

	20m	1h	2h	4h
Normalized score 6-4 TT	0.86	0.68	0.30	-0.15
Normalized score 6-4 TC	0.84	0.58	0.29	-0.07
Raw counts 6-4 TT	0.93	0.85	0.36	0.27
Raw counts 6-4 TC	0.92	0.77	0.50	0.19

Table B.2: *Correlation coefficients of correlations between 6-4PP 0h damage with other time points within 100kb chunks, for both normalized score and raw data counts.*

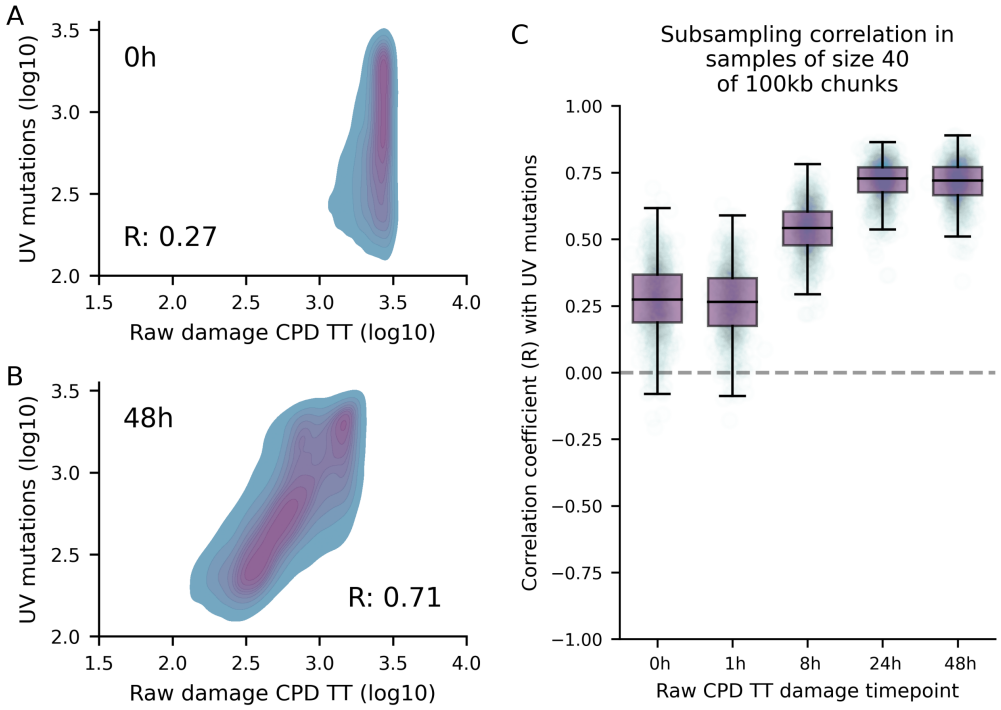

Figure B.1: *Correlations of raw CPD TT damage counts with UV mutations within 100kb chunks. A) KDE plot of 0h damage with mutations, B) KDE plot of 48h damage with mutations, C) Subsampling correlation of damage at various time points with mutations, with a sample size of 40 shuffled chunks. Each point in the boxplot represents a single sample. R is the correlation coefficient.*

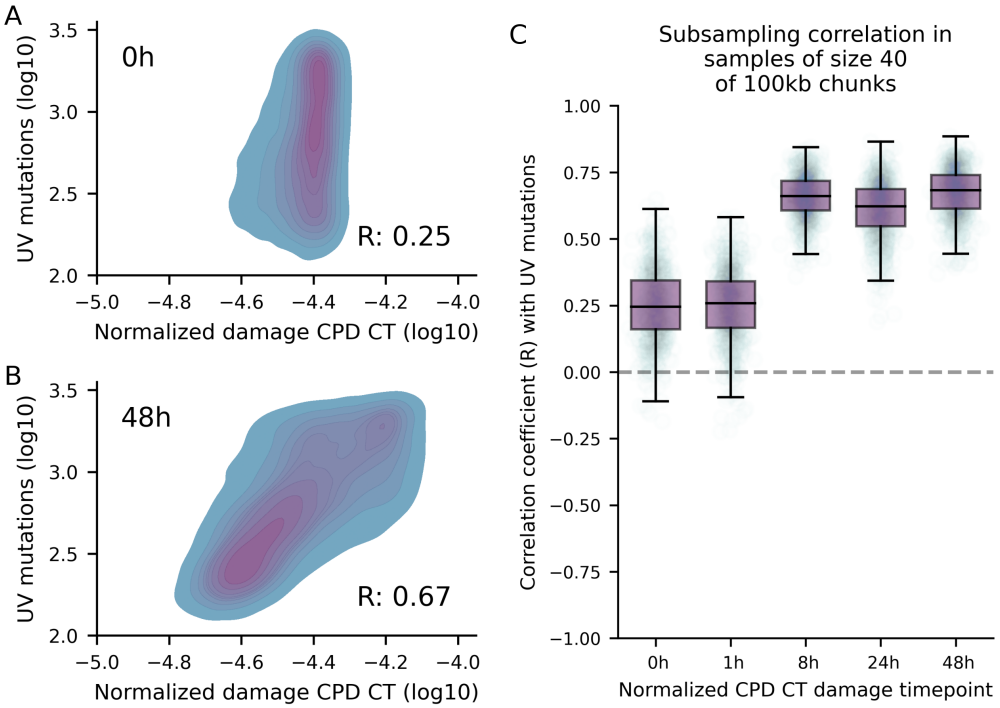

Figure B.2: *Correlations of normalized CPD CT damage score with UV mutations within 100kb chunks. A) KDE plot of 0h damage with mutations, B) KDE plot of 48h damage with mutations, C) Subsampling correlation of damage at various time points with mutations, with a sample size of 40 shuffled chunks. Each point in the boxplot represents a single sample. R is the correlation coefficient.*

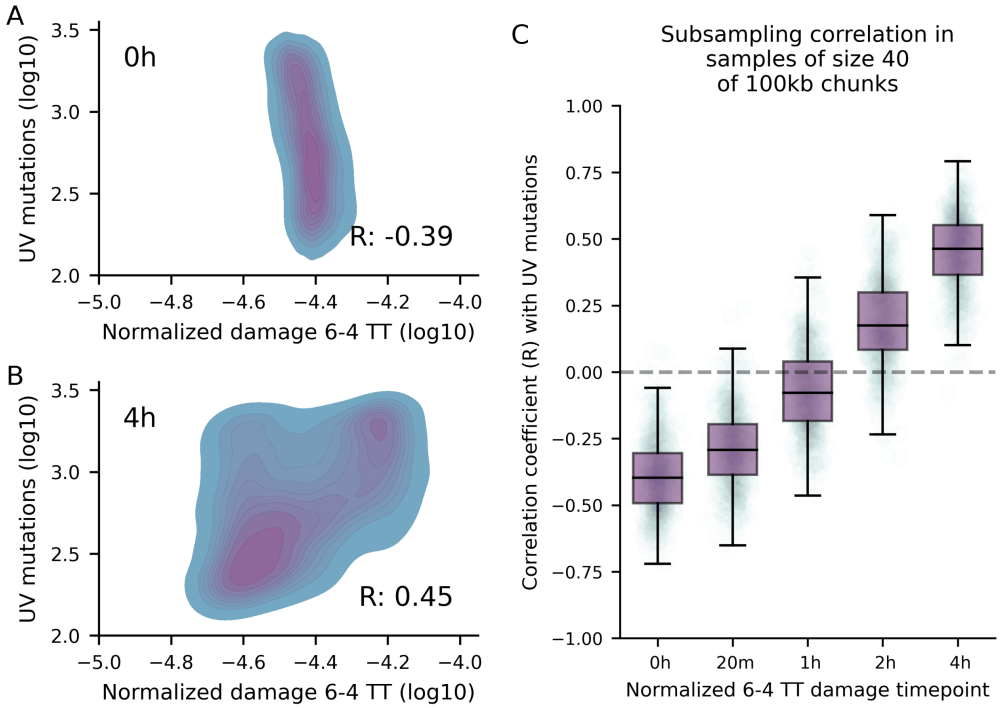

Figure B.3: *Correlations of normalized 6-4PP TT damage score with UV mutations within 100kb chunks. A) KDE plot of 0h damage with mutations, B) KDE plot of 48h damage with mutations, C) Subsampling correlation of damage at various time points with mutations, with a sample size of 40 shuffled chunks. Each point in the boxplot represents a single sample. R is the correlation coefficient.*

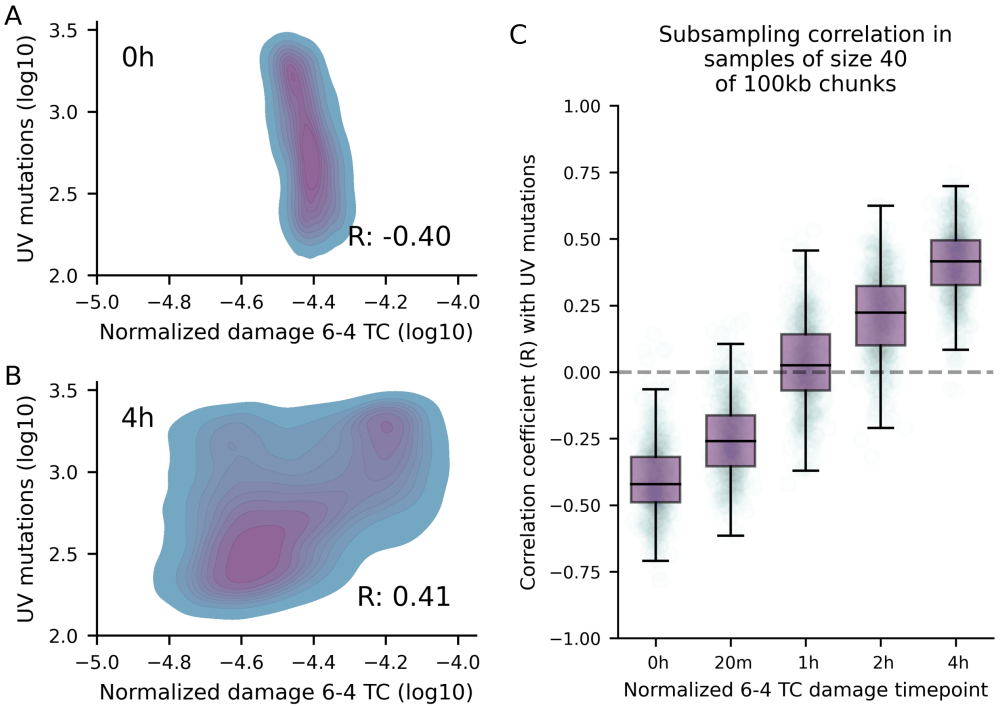

Figure B.4: *Correlations of normalized 6-4PP TC damage score with UV mutations within 100kb chunks. A) KDE plot of 0h damage with mutations, B) KDE plot of 48h damage with mutations, C) Subsampling correlation of damage at various time points with mutations, with a sample size of 40 shuffled chunks. Each point in the boxplot represents a single sample. R is the correlation coefficient.*

B.2. Hidden states transition probabilities

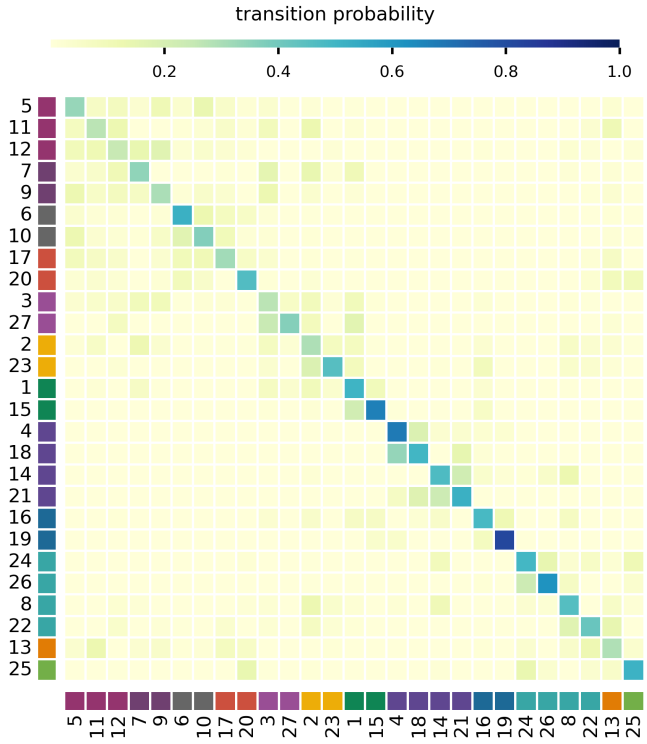

Figure B.5: *Transition probabilities between all hidden states from sticky HDP-HMM.*

B.3. Reproducibility analyses

Stickiness	Hidden states number (including disallowed)
1	28
2	27
3	27

Table B.3: *Number (including the disallowed state) of hidden states generated by sticky HDP-HMM at 100kb with differing stickiness parameters. Stickiness 1 (lowest) = kappa 100, 2 = kappa 200, 3 (highest) = kappa=300.*

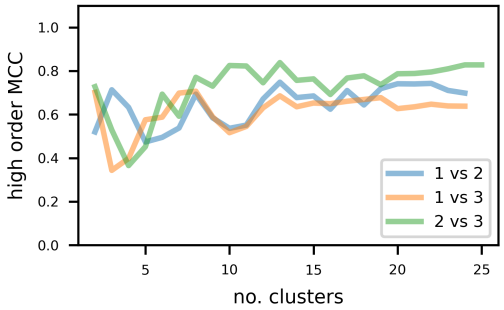

Figure B.6: *Reproducibility analysis of hidden state partitionings and their re-clusterings at 100kb with differing stickiness parameters. Stickiness 1 (lowest): kappa=100, 2: kappa=200, 3 (highest): kappa=300.*

Replicate	Chunksize	Hidden states number (including disallowed)
A_norm_1	100kb	26
B_norm_2	100kb	27
A_norm_1	1Mb	8
B_norm_2	1Mb	8
A_norm_1	10kb	28
B_norm_2	10kb	31

Table B.4: *Number (including the disallowed state) of hidden states generated by sticky HDP-HMM for the two data replicates. A and B are damage map replicates, 1 and 2 are repair map replicates.*

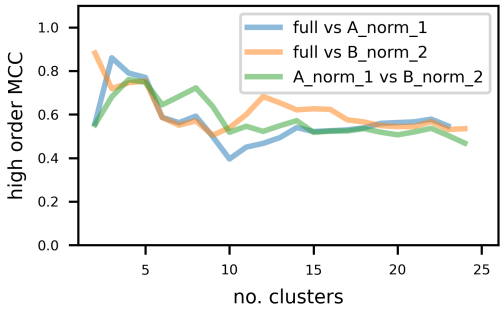

Figure B.7: *Reproducibility analysis of hidden state partitionings and their re-clusterings at 100kb with the two replicates and the main run. A and B are damage map replicates, 1 and 2 are repair map replicates.*

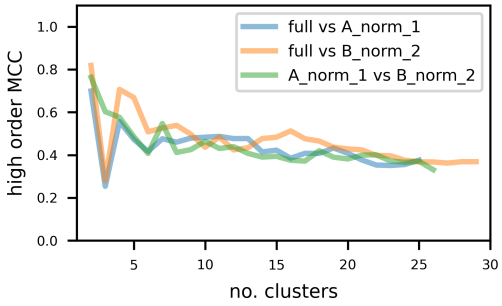

Figure B.8: *Reproducibility analysis of hidden state partitionings and their re-clusterings at 10kb with the two replicates and the main run. A and B are damage map replicates, 1 and 2 are repair map replicates.*

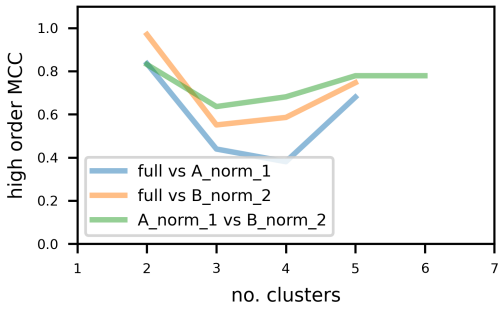

Figure B.9: *Reproducibility analysis of hidden state partitionings and their re-clusterings at 1Mb with the two replicates and the main run. A and B are damage map replicates, 1 and 2 are repair map replicates.*

B.4. Genomic features

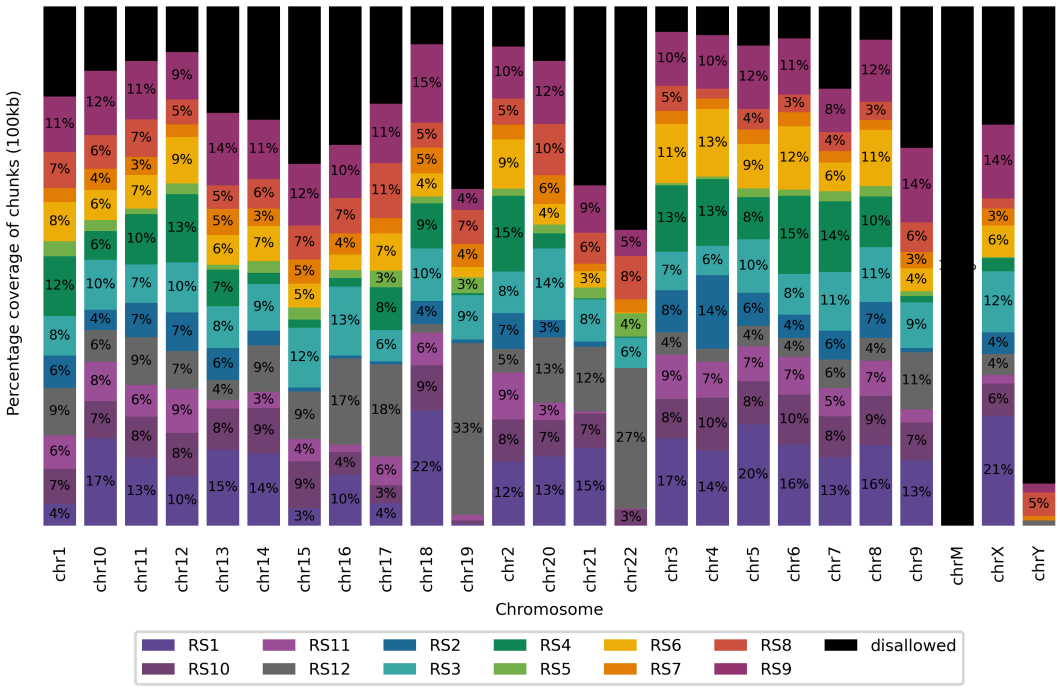

Figure B.10: *Percentage coverage of the human genome by repair states, on each chromosome, in 100kb chunks.*

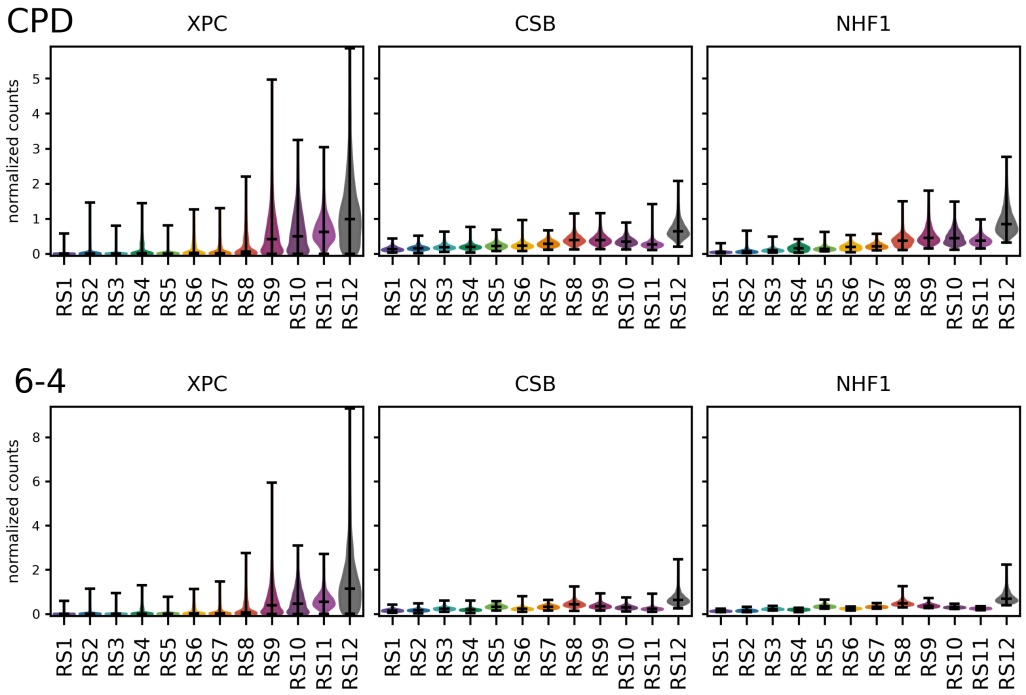

Figure B.11: *TT-normalized score of snapshot repair at 1h across states for three cell lines: XPC (TC-NER proficient) and CSB (global NER proficient) mutants, and NHF1 (proficient in both NER pathways).*