Investigating the interplay between DNA damage frequency and distribution in the genome

by

Y. Lucia Wang

Department of Pharmacology and Therapeutics

McGill University, Montréal, Québec, Canada

August 2022

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Science

Table of Contents

Abstract

DNA damage often has a negative connotation but has an important function in the clinic. There are many classes of chemotherapeutics which function by non-specifically damaging DNA. However, while these drugs are still commonly used as first-line therapies, not all patients respond positively. *Currently, we are unable to reliably predict patient response to specific drug regimens*. **We hypothesize that both the frequency and the distribution of drug-induced damage in the genome are necessary to predict patient drug response.** However, measuring DNA damage distribution in the genome is challenging and only recently have a few methods been developed. The goal of this thesis is to validate, develop, and assess DNA damage sequencing methods for their potential to predict cellular fate upon treatment with damaging agents. I have approached the problem from three perspectives: 1) Developing a novel method that will allow interstrand crosslinks (ICLs), the most lethal type of crosslink damage, to be sequenced at a nucleotide resolution; 2) Validating a recently described method for sequencing abasic (AP) sites and applying this method to understand the effect of an important DNA damage repair (DDR) accessory factor, *CUX1*, on DNA damage distribution in mammalian cells; and 3) Comparing the distribution of double strand breaks (DSBs) across different cell genomes using publicly accessible sequencing data obtained by recently described DSB sequencing methods.

First, we optimised a new protocol using synthetic oligonucleotides that would enable the specific recovery and future sequencing of ICLs. Furthermore, we validated the differential drug response of our three cancer cell lines in response to cyclophosphamide, an ICL-inducing chemotherapeutic. These cell lines will serve as excellent models to address whether DNA damage patterns in the genome correlate to cellular response. Moving forwards, our method to

recover and sequence ICLs could be applied to perform the first ever whole genome sequencing of ICLs.

My second aim was to validate a novel method called snAP-seq which should allow genome-wide sequencing of AP-sites. We synthesized a chemical probe that can specifically label AP-sites in the genome and confirmed site-specific capture of AP-sites. We then applied this method to investigate the DDR pathway, investigating the effects of knocking down a DDR accessory factor on damage distribution. Because the DDR accessory factor of interest, CUX1, is a transcription factor protein, we hypothesize that the distribution of AP-sites upon CUX1 knockdown will not be uniformly distributed.

Aim three makes use of data obtained via other published DNA damage sequencing methods. Specifically, there are multiple recently described methods available to sequence DSBs which are a highly destructive form of DNA damage. However, these methods are currently not commonly used and there has been no comparison between them. As such, using the sequencing data from these papers, we have investigated the distribution of damage across the entire genome and demonstrated that damage distribution in each chromosome is not random. Furthermore, we have also extrapolated patterns in DNA damage distribution relating to cancer genes and specific genes involved in the DDR pathway.

The results of this thesis confirm that snAP-seq and currently published DSB sequencing methods can be successfully used to study both AP-site and DSB formation across the genome. Furthermore, we also developed a promising new method for sequencing ICLs. In the future, these methods can be applied to provide a better understanding of the interplay between damage frequency and distribution. By improving our ability to predict patient response to

chemotherapeutic regimens, we aim to reduce the number of treatments a patient must undergo to achieve a positive response, thereby improving their overall quality of life.

Résumé

Les dommages à l'ADN ont souvent une connotation défavorable malgré leur fonction importante en milieu clinique. Il existe de nombreuses classes de produits chimiothérapeutiques (CT) qui endommagent l'ADN. Cependant, bien que ces médicaments soient couramment utilisés, ce ne sont pas tous les patients qui réagissent de façon positive. *Actuellement, nous ne sommes pas capables de prédire de manière fiable la réponse des patients à des schémas thérapeutiques.* **Nous postulons que la fréquence et la distribution des dommages provoqués par les médicaments dans le génome sont essentiels pour prédire la réponse des patients.** Cependant, mesurer la distribution des dommages à l'ADN dans le génome est un défi et ce n'est que récemment que quelques méthodes ont été développées. L'objectif de cette thèse est de valider, développer et d'évaluer les méthodes de séquençage des dommages à l'ADN afin de déterminer leur potentiel à prédire les modifications cellulaires lors d'un traitement avec des agents nocifs. J'ai examiné le problème de trois façons : 1) Développer une nouvelle méthode pour séquencer les réticulations interbrins (RIB) ; 2) Valider une méthode récemment décrite pour le séquençage des sites abasiques (AP) et l'application de cette méthode pour comprendre l'effet d'un facteur accessoire, CUX1, impliqué dans la réparation des dommages à l'ADN (RDA) sur la distribution des dommages ; et 3) Comparer la distribution des cassures double brin (CDB) sur différents génomes cellulaires en se fondant sur des données de séquençage qui ont été obtenues par des méthodes de séquençage récemment décrites.

Tout d'abord, j'ai optimisé un nouveau protocole utilisant des oligonucléotides qui la récupération et le séquençage spécifique des RIBs. De plus, nous avons validé la réponse différentielle de trois lignées cellulaires en réponse au cyclophosphamide, un agent CT induisant les RIBs. Ces lignées cellulaires serviront comme modèles pour déterminer si les modèles de

dommages à l'ADN dans le génome sont reliés à la réponse cellulaire. À l'avenir, notre méthode pourrait être appliquée pour effectuer le tout premier séquençage du RIBs dans le génome entier.

Mon deuxième objectif était de valider une nouvelle méthode, snAP-seq, qui permet le séquençage pangénomique des sites-AP. Nous avons synthétisé une sonde chimique qui étiquète spécifiquement les sites-AP dans le génome et nous avons confirmé l'indentification spécifique des sites-AP. Nous avons ensuite appliqué cette méthode pour étudier la RDA en étudiant les effets de la suppression d'un facteur accessoire sur la distribution des dommages. Étant donné que le facteur accessoire, CUX1, est un facteur de transcription, nous pensons que la distribution des sites-AP lors de l'inactivation de CUX1 ne sera pas uniformément distribuée.

Le troisième objectif utilise des données obtenues par d'autres méthodes publiées de séquençage des CDB. Cependant, ces méthodes ne sont pas couramment utilisées et il n'y a pas de comparaison entre elles. Ainsi, en utilisant ces données de séquençage, nous avons étudié la distribution des dommages sur l'ensemble du génome et démontré que la distribution des dommages n'est pas aléatoire. En outre, nous avons également extrapolé des schémas de distribution des CDB liés aux oncogènes et aux gènes spécifiquement impliqués dans le RDA.

Les résultats de cette thèse confirment que snAP-seq et les méthodes de séquençage CDB peuvent être utilisées avec succès pour étudier à la fois les sites-AP et les CDB à travers le génome. Nous avons également développé une nouvelle méthode pour le séquençage des RIBs. À l'avenir, ces méthodes pourront être appliquées pour fournir une meilleure compréhension de l'interaction entre la fréquence et la distribution des dommages d'ADN. En perfectionnant notre capacité à prédire la réponse des patients aux régimes CT, nous visons à réduire leur thérapie tout en maintenant les effets bénéfiques et ainsi d'améliorer leur qualité de vie.

Acknowledgements

The completion of my Master's degree has not been easy, so I would like to acknowledge and thank all of the people who have helped me and supported me along the way. First, I would like to thank my supervisor, Prof. Maureen McKeague, for giving me the opportunity to work in her lab over the last four years. I will forever be grateful for her kindness, patience, and encouragement, especially as I struggled with my projects. Her flexibility and encouragement for me to explore multiple avenues taught me how to approach a question from many perspectives and to never give up. I have never had a mentor who is so open and supportive of her students in all of their endeavours. Maureen, your support of not just my science, but my passion for writing has shaped me into the type of scientist I have always wanted to be. I will move forward into my PhD with all of the skills that you have helped me develop. Thank you so much!

I would like to thank everyone who has helped me with my project. Thank you to my advisor, Prof. Jean-François Trempe, as well as my committee members, Prof. Sarah Kimmins and Prof. Jason Tanny. Though we did not meet often, their feedback and guidance were invaluable for shaping the direction of my project. I would also like to thank Johanna Krebs who really helped me to develop the first chapter of my thesis. Without her I would have needed a lot more time to complete it. I would also like to thank the Department of Pharmacology and Therapeutics and all of the faculty and employees who have helped me along the way. I would especially like to thank Tina Tremblay, Cathy Shang Kuan, and Nadee Buddhiwickrama for all of their support and encouragement and for helping to make my Master's journey a lot less scary. Thank you also to NSERC and the FRQ-NT for helping to fund my studies.

Thank you to all of my lab mates! Bruk, Olivia, Negin, and Ayo, your constant encouragement was a ray of sunshine when it felt like everything was failing. Eiman, I am so

grateful for all of your guidance – your level headedness really helped me push through the hard times. Micaela, thank you for being in the lab late at night with me and for always cheering for my successes and sharing yours. Michael and Serge, thank you for always being around to chat. Son, thank you for always challenging me in ways I don't expect. Omma and Maira, thank you for being such sources of positivity in the lab. Janeva, thank you for being such a great student to mentor. Jath, thank you for always being around to listen, whether it be about life or science. Thank you all – without you, this experience would not have been so vibrant and fun.

Finally, I would like to thank my friends and family for all of their support. All of the Js (Jason, Jayson, Jon, June, and Jess), David, Liam, and Hanshi, thank you for making Montreal feel like home. You are all so wonderful in your own special ways and never fail to make me laugh. Carrie, thank you for being you. I don't think I would have made it out of undergrad, much less through my Master's without you. You are a constant source of entertainment and relatability, and I will always cherish the memories we've made. Jerry, thank you for staying up with me through the late nights and cheering for me from afar, though I think you sabotaged me a little bit by introducing me to League. To the friends I don't have space to thank by name (Wendell), just know I cherish you all and appreciate all of the support you have given me! To my parents and Carolyn, thank you for challenging me and pushing me to be the best version of myself. Thank you for telling me the hard truths and believing in me even when I didn't. I appreciate everything that you have done for me. Nathalie, thank you for helping me with my French abstract and for always asking interesting and provocative questions about my work. Finally, Seb, thank you for being my number one fan and for listening to me ramble about my project even when you didn't understand a single word that was coming out of my mouth. 我爱你们!

Contribution of Authors

All of the chapters presented in this thesis are authored by Y. Lucia Wang and edited by Prof. Maureen McKeague.

Chapter 2: Crosslinks Approach

All experimental work was performed by Y. Lucia Wang. The primers used were designed by Johanna Krebs.

Chapter 3: AP-Site Approach

All experimental work was performed by Y. Lucia Wang, except for the validation of the CUX1 knockdown which was performed by Elise Vickridge (Nepveu Lab), the synthesis of the HIPS probe which was done by Serge Hirka (McKeague Lab) and Kaleena Basran (Luedtke Lab), and the PAGE gels which were run together with Jathavan Asohan (Sleiman Lab).

Chapter 4: Bioinformatics Comparison of Double Strand Break Sequencing Methods

Interpretation of the results obtained from the raw bioinformatics data was performed by Y. Lucia Wang. All of the bioinformatics data processing was performed by Malinda Huang.

List of Abbreviations

4-HPCP	4-hydroperoxycyclophosphamide
AB/AM	Antibiotic-antimycotic
ALL	Acute lymphoblastic leukemia
AP	Abasic
ARP	Aldehyde reactive probe
BER	Base excision repair
BLESS	Direct *in situ* breaks labelling enrichment on streptavidin
BLISS	Breaks labeling *in situ* and sequencing
bp	Base pair
ChIP-seq	Chromatin immunoprecipitation sequencing
CHOP	Cyclophosphamide, vincristine, daunorubicin, and prednisone combination treatment
DDR	DNA damage repair
DNMT1	DNA methyltransferase I
dNTP	Deoxynucleotide triphosphate
DSB	Double strand break
EDTA	Ethylenediaminetetraacetic acid
EGFR	Epidermal growth factor receptor
ELISA	Enzyme-linked immunosorbent assay
ExoIII	Exonuclease III
FA	Fanconi anemia
FBS	Fetal bovine serum

FORMA	Forma Water Jacketed CO_2 Incubator
G-Nor-G	N7-guanine interstrand crosslink
HIPS	Hydrazino-*iso*-Pictet-Spengler
IC_{50}	Half maximal inhibitory concentration
ICL	Interstrand crosslink
MMR	Mismatch repair
NER	Nucleotide excision repair
NGS	Next generation sequencing
PBMC	Peripheral blood mononuclear cell
PBS	Phosphate buffered saline
Ph+	Philadelphia chromosome
Poly(dI-dC)	Poly(deoxyinosinic-deoxycytidylic) acid
RPMI 1640	RPMI 1640 medium (mod.) 1X with L-glutamine
sBLISS	In-suspension breaks labeling *in situ* and sequencing
TE	Trisaminomethane ethylenediaminetetraacetic acid
TFIIH	Transcription factor IIH
THPTA	Tris-hydroxypropyltriazolylmethylamine
Tris	Trisaminomethane
Tris-HCl	Trisaminomethane hydrochloride
UNG	Uracil-DNA-glycosylase

List of Figures

List of Tables

Chapter 1: Introduction

Cancer as a disease

Cancer is defined by the uncontrolled proliferation of cells [1]. This can happen through gain-of-function mutations in protooncogenes and through loss-of-function mutations in tumor-suppressor genes [2]. While oncogenes and tumor-suppressor genes function in opposite manners, requiring activating and silencing mutations, respectively, the end result is the same: increased cell proliferation and reduced cell death. Unsurprisingly, the majority of genes falling into either of these classes are implicated in the control of cell growth [2,3].

Oncogenes and tumor-suppressor genes can be viewed as two sides to the same coin, both aiming to maintain the regular growth and turnover of healthy cells. Oncogenes are typically referred to as protooncogenes when they are in their native state (i.e., non-mutationally activated and contributing to normal cell growth), as they only result in uncontrollable cell division upon mutation [4,5]. These genes tend to encode for proteins that stimulate cell division and inhibit differentiation [5], such as growth factor receptors, signal transducers, and transcription factors [2]. For example, epidermal growth factor receptors (EGFR) are a family of transmembrane glycoproteins which function as receptor tyrosine kinases. Upon binding of the epidermal growth factor ligand, EGFR signalling will be activated, leading to the stimulation of several signaling cascades [6]. These subsequently activated pathways can then modulate the growth, differentiation, migration, and survival of cancer cells [6]. In the case of mutations in *EGFR*, the development of malignant cells is often seen. In fact, expression of these mutated EGFR proteins has been observed in a plethora of cancers ranging from breast to esophageal cancer [6].

Looking at the development of cancer from a different perspective, mutations in tumor-suppressor genes are also often implicated. Tumor-suppressor genes exist to inhibit cell proliferation and tumor development. Thus, when they are lost or inactivated via mutation, abnormal proliferation leading to tumor growth may be observed as the negative regulators of the cell cycle are lost [7]. One of the most well-known tumor-suppressor genes that is often mutated in cancer is *p53*. In fact, it is estimated that *p53* mutations may be involved in up to 50% of all cancers and it is seen in cancers ranging from leukemias to glioblastomas [7]. In its native form, *p53* plays an essential role in mitigating DNA damage by controlling cellular arrest, apoptosis, and cellular proliferation [2]. As such, when *p53* is inactivated, cells will no longer stop dividing in response to DNA damage, leading to an accumulation of damage and mutations which may purport further malignancy [2]. In addition, *p53* inactivation may also result in a reduction of apoptosis as it plays a role in downregulating the expression of the oncogene *bcl2* which inhibits apoptosis. Thus, damaged cells are able to continue proliferating despite genomic instability furthering the potential for malignancy [2]. Altogether, a select few mutations are enough to result in the development of cancer as mutations in both protooncogenes and tumor-suppressor genes can tip the scales in favour of the uncontrolled proliferation of malignant cells.

The development of chemotherapeutics and personalized medicine

While cancer may be a devastating disease, many scientists have tackled the problem of treating it over the years through the development of chemotherapeutics. Given that these malignancies are the result of uncontrolled cell proliferation, the goal of chemotherapy is twofold, aiming to both slow or stop the growth of cancerous cells as well as inducing cell death. Historically, chemotherapy has taken advantage of the characteristic rapid cell division of malignant cells as they often do not stop the cell cycle to repair incurred damage, thus leading

more rapidly to unsustainable levels of genomic instability and subsequent cell death. As such, first generation chemotherapies were not designed to be targeted treatments, but rather damaged all cells ubiquitously and relied on healthy cells to be able to repair the damage (Table 1) [8,9]. However, this type of drug therapy is not ideal as the body also contains normal rapidly dividing cells, such as those in the hair follicles and bone marrow [9]. For this reason, many classical chemotherapeutics result in the well known side-effects of alopecia and decreased hematopoiesis, among others [9]. Furthermore, traditional chemotherapies also carry the risk of latent secondary cancer development arising from the side effects of primary treatment [10]. Thus, novel chemotherapies targeted specifically towards cancerous cells have been, and continue to be, developed.

Table 1. Classical DNA-damaging chemotherapeutics, their indications, and their mechanisms of action.

Drug family	Drug classification	Mechanism of action	Cancer indications
Alkylating agents	Alkyl sulphonates	Alkylation resulting in adenine-guanine crosslinks [11,15]	Chronic myelogenous leukemia [12]
	Nitrogen mustards	Alkylation and crosslinking of DNA, particularly at the N-7 position of guanine [13,15]	Leukemias, lymphomas, sarcomas, neuroblastoma, and breast, nasopharyngeal, lung, and ovarian cancer [12,13]
	Nitrosureas	Alkylates and crosslinks DNA [12,15]	Brain, pancreatic, and hematopoietic cancer [12]
	Triazines	Alkylates guanine in DNA [12]	Melanoma, sarcoma, Hodgkin’s lymphoma,

			neuroblastoma, and colon and breast cancer [12,14]
	Aziridines	Crosslinks DNA and may form intermediates which damage DNA [12,15]	Leukemias, multiple myeloma, and ovarian, breast, gastrointestinal, brain, and bladder cancer [12]
Antimetabolites	Pyrimidine analogues	Incorporated into DNA, disrupting elongation or binds to and inhibits thymidylate synthase [16,17]	Leukemias and pancreatic, breast, colorectal, lung, and bladder cancer [17]
	Purine analogues	Inhibits nucleotide synthesis and metabolism or interrupts the processing and elongation of DNA [18,19]	Leukemias and lymphomas [18]
	Folate antagonists	Inhibits dihydrofolate reductase or inhibits purine and pyrimidine synthesis via thymidylate synthase inhibition [20]	Leukemias, lymphomas, mesotheliomas, non-small cell lung cancer, and breast, head and neck, pancreatic, stomach, colorectal, and bladder cancer [21]

DNA Crosslinkers	Platinum-based drugs	Crosslinks DNA, preferentially binding to guanine [15]	Lymphomas, sarcomas, and lung, breast, head and neck, brain, testicular, and lung cancer [12,15]

Targeted chemotherapies share the same goal as traditional cancer drugs in that they aim to stop cancer growth and eradicate malignant cells. However, they have the significant advantage of incurring fewer side effects, since non-cancerous cells should not be affected by drug treatment [9]. As such, targeted chemotherapies have become increasingly popular over the past two decades. One of the largest classes of targeted chemotherapies are tyrosine kinase inhibitors. These drugs bind to aberrant kinases arising from specific cancerous mutations to inhibit their phosphorylation activities. For example, Imatinib (Gleevec®) is a popular tyrosine kinase inhibitor used to treat patients with the Philadelphia chromosome (Ph+). Ph+ patients experience a reciprocal translocation on chromosomes 9 and 22 wherein the *BCR* and *ABL* genes are fused [22]. This leads to an overactive tyrosine kinase which permits uncontrolled cell proliferation, as is commonly seen with the pathogenesis of hematopoietic cancers [23]. Thus, in Ph+ patients, Imatinib can be used to reduce the activity of the BCR-ABL kinase, thereby mitigating the uncontrolled cell division.

<u>The current cancer treatment landscape for acute lymphoblastic leukemia</u>

While it is true that targeted chemotherapies provide many benefits over traditional chemotherapies, not all patients are eligible for them. This is because targeted therapies require specific mutations to be present and cancer is an extremely heterogenous disease. As such, traditional chemotherapies still make up a large majority of frontline cancer therapies. For

example, acute lymphoblastic leukemia (ALL) is a blood cancer with a first-line chemotherapy regimen consisting of only classical chemotherapies.

ALL is an aggressive cancer that accounts for approximately 80% of all childhood cancers and 20% of all adult leukemias [24,25]. Given its prevalence, ALL has been very well-studied over the past few decades. Arising from the malignant transformation and proliferation of lymphoid progenitor cells in the bone marrow, ALL results in the accumulation of immature lymphoid cells in the circulation and can be fatal within months if left untreated [25,26]. While there are certain chromosomal aberrations such as the BCR-ABL fusion gene which may predispose an individual to developing ALL, the majority of ALL cases manifest as *de novo* malignancies [25].

Current frontline treatment for ALL has three stages: induction, consolidation, and long-term maintenance; largely relying on the use of DNA damaging drugs [25-28]. Most commonly, cyclophosphamide, an alkylating agent, is used in combination with vincristine, a vinca alkaloid, daunorubicin, a topoisomerase inhibitor, and prednisone, an immune system suppressant [25]. Together, this treatment regimen is known as CHOP. In CHOP, each drug has a role to play in order to halt cancer proliferation: cyclophosphamide is metabolized in the liver, forming an active phosphoramide mustard which creates inter- and intra- strand N7-guanine crosslinks [29-31]; vincristine inhibits mitosis at metaphase by binding and inactivating tubulin [32]; daunorubicin stabilizes the DNA-topoisomerase II complex, preventing DNA unwinding during DNA replication [33]; and prednisone reduces inflammation and suppresses the body's immune response [34].

Generally, CHOP is very effective, and the cure rate of pediatric ALL is now approaching 90% [35]. However, with the current deeper understanding that we now have

regarding the heterogeneity of cancers and the different genetic alterations that can lead to the same phenotype, this one size fits all method of treatment is largely outdated and can lead to a decrease in patients' quality of life without achieving complete remission [36-40]. Thus, cancer therapy is now shifting towards a more patient-oriented personalized approach. In line with this, there is a need for the identification novel biomarkers which can be used to stratify patients and predict whether they will respond positively to frontline treatment, or whether other therapies should be used instead.

DNA damage as a disease biomarker for personalized cancer therapy

One useful biomarker of treatment efficacy may be the presence of interstrand crosslinks (ICLs), which are produced by many traditional chemotherapeutics to exert their cancer cell killing effects. In particular, cyclophosphamide, a commonly used bifunctional alkylating agent, produces N7-guanine ICLs (G-Nor-G) which are hypothesized to be lethal as these ICLs constitute a complete block of DNA replication [41]. G-Nor-G may be a useful biomarker as activated cyclophosphamide preferentially attacks the N7 position of guanine, creating an adduct which can be specifically measured (Figure 1).

Historically, the alkaline elution method has been used to quantify the production of ICLs on the basis that damaged DNA is more prone to fragmentation and will have a higher mean elution rate compared to untreated DNA. Using this method, researchers have shown a positive correlation between cyclophosphamide-based DNA damage and treatment outcome in patient samples, thus setting the precedence that G-Nor-G formation must, in some way, be related to a positive drug response [42-44]. However, the alkaline elution method cannot differentiate between ICLs and other forms of DNA damage, such as double strand breaks (DSBs), which may also arise from chemotherapeutic treatment. Building on this work, a novel mass

spectrometry technique was recently developed to detect G-Nor-G specifically *in vivo* and has been used to compare G-Nor-G formation between different patient populations [42,43]. Using this method, several other DNA adducts have also been explored as candidates for predictive markers of chemotherapy response [40]. Indeed, two clinical trials are underway which aim to determine whether DNA adduct formation following subtherapeutic dosing of chemotherapeutic drugs can act as a predictor for treatment response [40,45,46]. The preliminary results from clinical trial number NCT01261299 have shown that above a threshold of approximately 0.74 DNA adducts per 10^8 nucleotides, there is a strong correlation between the number of DNA adducts formed and positive patient response towards chemotherapy [47]. That being said, the patients experiencing a number of DNA adducts below this threshold demonstrate a variable response to chemotherapy (Figure 2). This indicates that while the quantity of DNA adducts

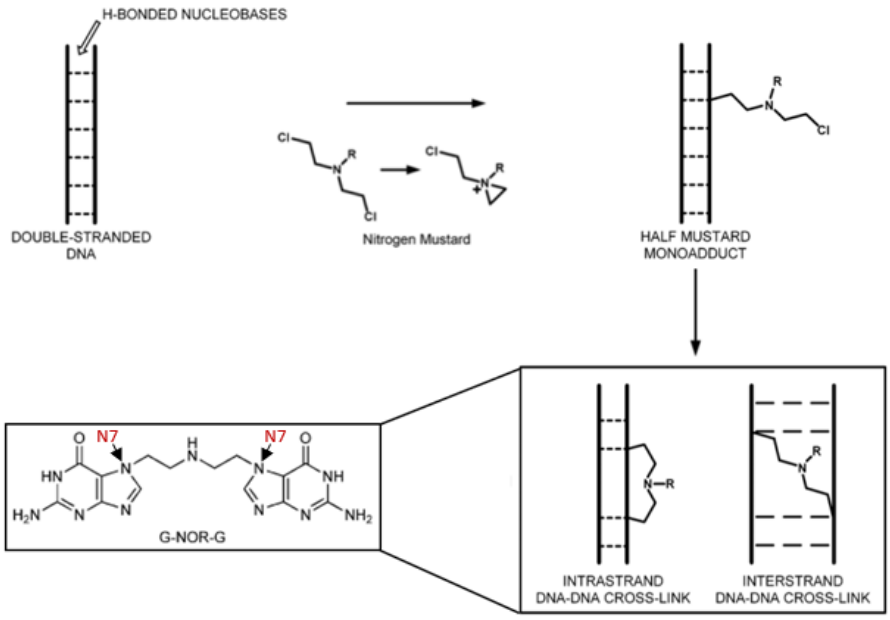


Figure 1. Formation of guanine crosslinks at the N7 position by a nitrogen mustard. The N7 position is favoured for this substitution reaction as it has the most nucleophilic character in biological systems [42,62]. Both inter- and intra-strand crosslinks can be formed, however interstrand crosslinks are the most physiologically relevant as they constitute an absolute blockage of DNA strand separation. Thus, they disrupt transcription and can lead to double strand breaks and eventual cell death [63]. Figure adapted from [64].

formed may be a promising indicator of treatment success, it alone does not allow for the consistent prediction of treatment outcome. Furthermore, while this mass spectrometry method is extremely sensitive and can quantify the number of G-Nor-G in biological samples, information about the sequence specific context of ICL distribution is lost. Thus, novel methods that allow for the investigation of G-Nor-G distribution at a nucleotide resolution are needed to further improve our understanding of how ICL quantity and distribution are interrelated.

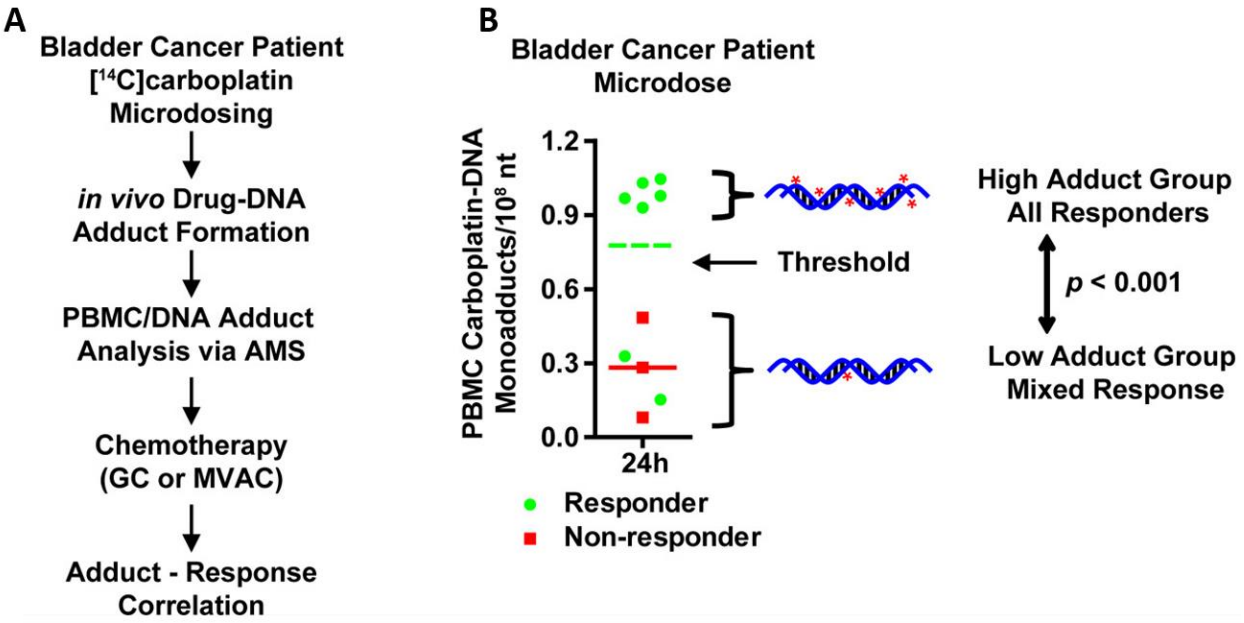


Figure 2. *In vivo* analysis of the correlation between the amount of DNA adduct formation and patient response. (A) Within the clinical trial, patients were given a subtherapeutic dose of carboplatin, an adduct forming chemotherapeutic. Blood samples were collected 24 hours post-treatment and peripheral blood mononuclear cells (PBMCs) were collected for DNA adduct quantification. The patients were then treated with the appropriate dose of chemotherapy for the clinician recommended regimen and patient response to the therapy was recorded. From there, a retrospective correlation was drawn between the number of DNA adducts formed and patient response. (B) A clear correlation is seen above a threshold of 0.741 PBMC carboplatin-DNA monoadducts per 10^8 nucleotides, however below this threshold we see mixed responses from the patients. Figure adapted from [40].

The DNA damage response

Given the ubiquitous nature of DNA damage, it is unsurprising that the human body has evolved many different repair mechanisms which constitute the DNA damage response (DDR). The specific DDR pathway involved in repair differs depending on the type of damage that is experienced. Table 2 summarizes the different types of damage and the pathways that repair them.

Table 2. DNA damage repair mechanisms and associated lesions [48].

DNA damage repair mechanism	**Primary lesions involved**
Mismatch repair	DNA base pair mismatches and indel loops arising from DNA replication
Base excision repair	Abnormal DNA bases (ex. Uracil in DNA) and simple monoadducts
Nucleotide excision repair	Bulky base adducts and UV photo-products which disrupt the DNA double helix
Non-homologous end joining	Radiation- or chemically- induced DSBs
Homologous recombination	DSBs in S phase, stalled replication forks, and ICLs
Fanconi anaemia pathway	ICLs
Direct reversal of DNA lesions	UV photo-products and O^6 alkylguanine adducts
ATM-mediated DDR signalling	DSBs
ATR-mediated DDR signalling	Single stranded DNA at DSB sites

CUX1: A DNA damage response accessory factor

While there are many proteins dedicated to the DDR, researchers have found that transcription factors may also play a key role [49-54]. In particular, studies suggest that the distribution of DNA damage is heavily influenced by the binding of transcription factors, particularly because cellular transcription stimulates the repair of damage because transcription is associated with an open chromatin conformation and is therefore more accessible to proteins involved in DNA repair [54]. As one key example, the Nepveu lab has shown CUX1, a member of the homeodomain transcription factor family, to act as either a transcriptional repressor or activator in a promoter dependent manner [55]. Similarly, CUX1 has also been paradoxically implicated in both tumor suppression and progression [56]. Thus, it is perhaps unsurprising that CUX1 plays a role in maintaining genome stability. Upon *CUX1* knockdown, it has been observed that there is a decrease in DDR gene expression. In particular, the expression of ATM and ATR which are two kinases involved in the recruitment of downstream effectors in response to DSBs is downregulated [56]. This abrogation of ATM and ATR signalling is even more striking when *CUX1* knockout cells are exposed to DNA damaging agents [56]. In addition to this, CUX1 also plays an important role as an accessory factor in the base excision repair (BER) pathway. Our collaborators have shown that in a *CUX1* knockdown glioblastoma cell model, there is an increase in abasic (AP) sites following treatment with the mono-alkylating agent temozolomide (Figure 3A) [57]. Furthermore, they have also shown that introducing ectopic expression of CUX1 after temozolomide treatment reduces AP-site formation (Figure 3B) [57]. Thus, it is clear that there is a link between CUX1 expression and DNA damage quantity.

As a transcription factor, CUX1 does not operate ubiquitously throughout the cell. Thus, it is possible that DNA damage arising from CUX1 mutations will not be uniformly distributed

throughout the genome. In fact, previous sequencing studies have shown that DNA damage distribution is never random and often there are genetic contexts that can inform damage patterns [58-61]. Much like ICLs, it is possible that AP-sites arising from CUX1 mutations may be concentrated in the genes and regulatory regions bound by CUX1. However, this has yet to be proven as AP-site quantification assays do not offer information about the sequence specific context of AP-site formation.

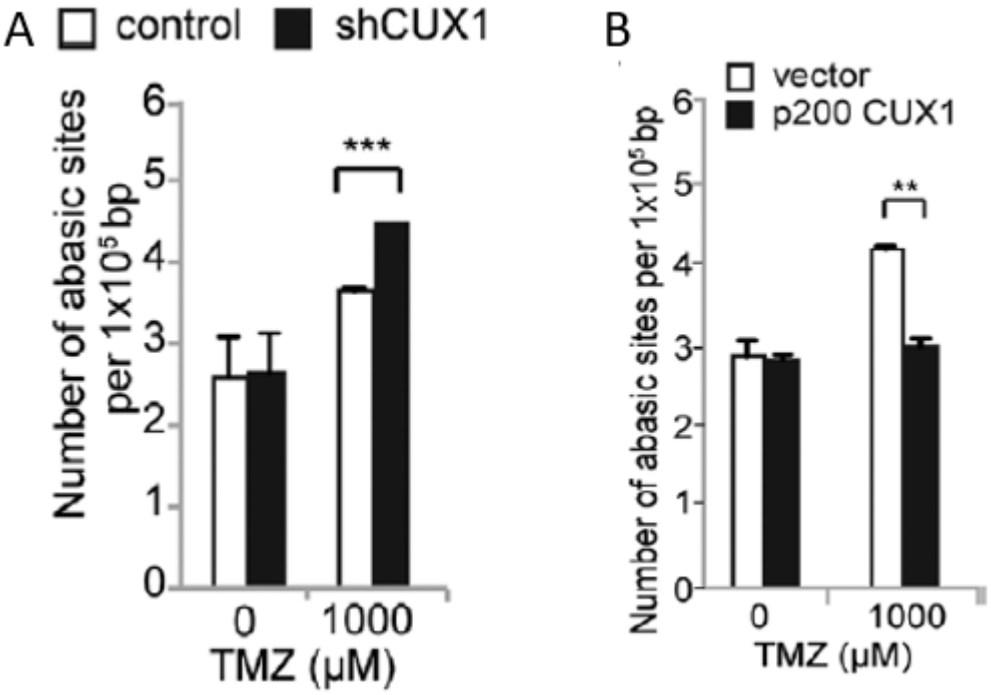


Figure 3. CUX1 modulation affects abasic site formation. In *CUX1* knockdown cells, (A) significantly more AP-sites form with temozolomide treatment, and (B) this number is significantly reduced with the reintroduction of CUX1. Figure adapted from [57].

Next generation sequencing as a tool for studying DNA damage and repair

Over the last decade, next generation sequencing (NGS) has revolutionized our ability to perform genomic research. Thanks to its ability to analyze many short sequences concurrently, NGS methods are much quicker and cheaper than the previously used Sanger sequencing

[65,66]. However, despite the whole genome sequencing capabilities of NGS, sequence information is still lost in the face of DNA damage as NGS still relies on polymerases which stall in the presence of DNA lesions [54,67,68]. Thus, novel methods had to be developed to address this shortcoming.

Methods for sequencing DNA damage

A loss of sequence information remains a general problem when it comes to investigating DNA damage. Our current understanding of mutations and their effect on cell viability is advanced, yet we still do not have a good understanding of how initial DNA damage due to mutagens is distributed across the genome [69]. This is largely due to the fact that conventional NGS does not provide any information about chemical damage and mis-inserts nucleobases when faced with a modified one [70]. Only in the past six years have novel technologies emerged, allowing us to identify the sequence and location of single base DNA damage events [61,71-76]. For example, Wu *et al*. [61] published a click-code-seq method in 2018 that has allowed for the sequencing of oxidative damage across the entire genome with single nucleotide-resolution by incorporating a "barcode sequence" (Figure 4). This novel method with its extremely high resolution has allowed 8-oxoguanine damage to be situated in its local sequence context for the first time ever. These results revealed that there is a preference at the nucleotide level for where damage will form [61].

Methods to sequence DNA damage arising from chemotherapeutics have also been developed. For example, Hu et al. [72] developed Damage-seq to map cisplatin adducts across the entire human genome. This method centers around the fact that bulky DNA adducts, such as those formed in response to cisplatin, block high fidelity DNA polymerases. DNA fragments containing cisplatin adducts were immunoprecipitated with anti-platinum antibodies, primers

were annealed to the selected DNA fragments, and then primer extension was done by PCR. All DNA fragments containing a cisplatin adduct caused the polymerase to fall off, leading to the generation of a shorter oligonucleotide where the 3′-end was the exact nucleotide containing the adduct. From there the second adapters could be ligated on and a sequencing library could be prepared.

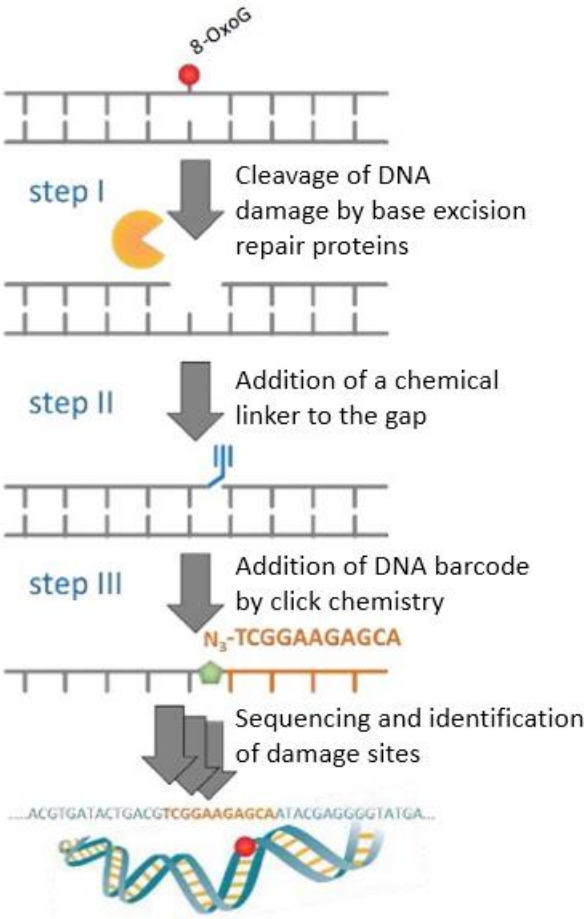


Figure 4. Schematic of the click-code-seq procedure. At the site of the DNA damage, an enzyme can be used to remove the adduct, leaving a gap in the DNA backbone which can then be filled by a base modified with a chemical linker. This linker will allow for the addition of a DNA sequence which can be identified after sequencing to determine the original damage site at a single nucleotide level resolution. Figure adapted from [61].

Working with the same type of damage, Hu et al. [73] also pioneered XR-seq which is a sequencing method designed to map the removal of cisplatin damage, providing information on damage repair patterns. XR-seq takes advantage of the fact that single-stranded fragments of DNA excised by the NER pathway are still associated with certain repair proteins such as transcription factor IIH (TFIIH). As such, these pieces of DNA can be enriched for using anti-TFIIH antibodies, allowing us to map the DNA sequences that have been targeted for repair. With these two methods, the authors showed that while the distribution of cisplatin adducts are mainly dictated by the underlying genomic sequences, the factors influencing repair are more heterogenous, correlating strongly with transcription and chromatin states. Together, these results support the idea that similar correlations between adduct formation and genomic sequence context may be found with other chemotherapeutics.

Sequencing double strand breaks

DSBs are a type of damage that is of particular interest because they involve the most complex repair pathways, namely non-homologous end joining and homologous recombination. Furthermore, DSBs constitute some of the damage most detrimental to genome stability. As such, much work has been done to develop methods that allow this type of damage to be sequenced at a nucleotide resolution. Certain established techniques, such as chromatin immunoprecipitation sequencing (ChIP-seq), have historically allowed for the indirect detection and sequencing of DSBs as ChIP-seq allows for the capture and sequencing of chromatin markers associated with DSB markers [77]. However, because this method uses DSB proxies to indicate the genomic positions of DSBs, ChIP-seq compromises both accuracy and resolution. Thus, since 2011, 18 new DSB sequencing methods have been published [77]. Notably, these methods all sequence DSBs directly, no longer relying on the detection of proteins associated

with DSB repair pathways [77]. As such, we now have tools that allow for a much higher resolution when determining the genomic contexts in which DSBs form.

Of these newly developed techniques, BLESS (Direct *In Situ* Breaks Labeling, Enrichment on Streptavidin and Next Generation Sequencing) [59], BLISS (Breaks Labeling *In Situ* and Sequencing) [60], sBLISS (In-suspension Breaks Labeling *In Situ* and Sequencing) [77], and DSBCapture [58] are the four techniques that label DSBs *in situ*. While the exact details of these methods differ, they share a similarity in that biotinylated adaptors are ligated directly to blunted DSB ends, providing a method to selectively enrich for DSBs. Furthermore, since the DSBs are labelled *in situ*, the risk of capturing DSBs formed artificially during the DNA extraction process is largely minimized [77]. Thus, the high-resolution maps of DSB formation across the genome produced by any of these four methods should yield relatively high-fidelity products that accurately reflect the endogenous locations of DSB formation.

Sequencing abasic sites

Many types of DNA lesions arise from exogenous exposures; however, some may also develop as an intermediate of a DDR pathway. For example, AP-sites are an intermediate product of the BER pathway. As such, mapping their distribution may provide interesting insights into both DNA damage and repair landscapes in the genome.

Upon their formation, AP-sites exist in an equilibrium of the closed-ring furanose and the highly reactive open-ring aldehyde (Figure 5) [78,79]. In organic synthesis, amine nucleophiles are very commonly used to form imines with aldehyde groups [80]. As such, in 1992, researchers produced a biotinylated probe – the Aldehyde Reactive Probe (ARP) – which contains a hydroxylamine group that reacts easily with the open-ring form of AP-sites. With this probe, the first enzyme-linked immunosorbent assay (ELISA) for quick and sensitive AP-site detection was

produced. This assay is relatively sensitive, able to detect between 1 and 40 AP-sites per 10^5 base pairs [82]. Furthermore, the biotin moiety allows tagged DNA to be selectively enriched. Thus, AP-seq was developed to sequence AP-sites genome wide [135]. However, the major drawback of using ARP is the mechanism of action by which ARP tags AP-sites. ARP tagging occurs via hydroxylamine condensation which is a reaction that takes place between the ARP hydroxylamine and any reactive aldehyde. For example, formylated bases, such as 5-formylcytosine and 5-formyluracil, have been shown to react with ARP and quantitative mass spectrometry measurements have shown that these formylated bases exist in a naturally higher abundance than AP-sites [80]. As such, the inability of ARP to discern between AP-sites and formylated bases poses a major issue for the correct analysis of AP-site distribution in the genome.

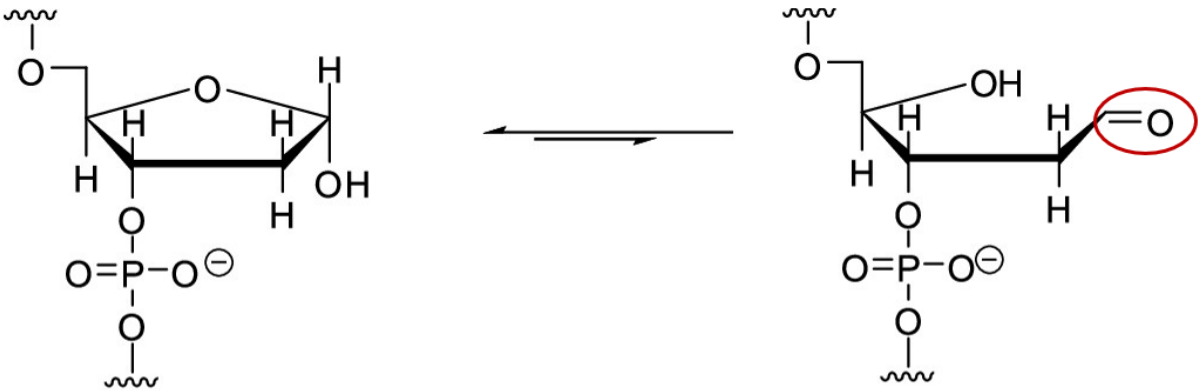


Figure 5. Abasic sites exist in equilibrium between two states. AP-sites exist more commonly in the closed-ring furanose form, but a small percentage of AP-sites will be converted to the open-ring aldehyde. The aldehyde, circled in red, is highly reactive and is commonly leveraged for reactions which allow AP-sites to be tagged with chemical probes. Figure adapted from [79].

In light of this, the Balasubramanian group has produced a novel probe which reacts with AP-sites via a Hydrazino-*iso*-Pictet-Spengler (HIPS) reaction instead. Using their HIPS probe,

the researchers developed a high-selectivity, single nucleotide resolution sequencing method that they used to map endogenous abasic sites across the human genome with high confidence [80].

Thesis objective and rationale

Many new sequencing methods have provided compelling evidence that DNA damage distribution is not random. However, these methods are new and have yet to be applied generally and they do not allow for the sequencing of all DNA damage types of interest. Furthermore, several groups have shown that certain types of drug-derived DNA damage quantities can be partially predictive of patient response to chemotherapy. That being said, the link between the frequency and distribution of DNA damage in the genome in relation to treatment response remains unclear. Thus, the main objective of this project will be to pursue several different methodologies to further validate and develop methods for DNA damage sequencing, working towards elucidating the relationships between DNA damage frequency, distribution, and response to drug treatment.

My first aim was to develop a method that would allow ICLs, the most lethal type of crosslink damage, to be sequenced at a nucleotide resolution. Despite advances in DNA damage sequencing, ICLs and their distribution remain a mystery. The end goal of this aim is to use this novel method to sequence the genomes of many cell lines to determine whether there is a pattern of ICL damage which can predict responses to cyclophosphamide treatment. I began by validating that we could produce the stable ring structures needed for our method using oligonucleotides. Following this, I began to optimize sonication parameters such that the same stable ring structures could be produced from genomic DNA. Finally, I established three model cell lines that were differentially sensitive to cyclophosphamide.

My second aim was to validate snAP-seq and apply it to investigate the effect of knocking down DDR accessory factor CUX1 on damage distribution. Modulation of CUX1 expression has been shown to have an effect on AP-site formation. However, while it is clear that the number of AP-sites changes with CUX1 expression, we do not know whether the distribution of damage changes as well. Towards this, I applied the snAP-seq method to a breast cancer cell line (MDA 231) to determine whether there are hotspots of damage formation which differ between wild-type and *CUX1* knockdown cells.

My third aim was to compare the distribution of DSBs across different cell genomes using publicly accessible sequencing data acquired by various DSB sequencing methods. This was done in collaboration with Malinda Huang who performed the bioinformatics analyses. Using the outputs generated from DSB sequencing data, I sought to determine whether there were common genes or genetic features that were disproportionately sensitive to DNA damage.

Through this project, I aim to provide more information regarding the interplay between DNA damage distribution and treatment response. As cancer is a heterogenous disease, it is erroneous to treat all patients in the same manner. By looking for biomarkers and predictive patterns of DNA damage, we move one step closer towards effective personalized medicine.

Chapter 2: Crosslinks Approach

Preface

Novel sequencing methods targeting DNA damage have become increasingly common, yet there is still no current technique that allows ICLs to be sequenced. Furthermore, while there are mass spectrometry methods that allow for ICL quantification [42,43], these methods are not sequence specific and do not allow ICLs to be distinguished from intrastrand crosslinks [42]. However, ICLs are a highly relevant form of DNA damage since they constitute an absolute replication block [41]. Thus, there is a need for a sequencing method that will allow ICL formation across the genome to be mapped.

Materials

Reagents

10 mM deoxynucleotide triphosphates (dNTPs), 5 mM magnesium chloride ($MgCl_2$), 10X Taq polymerase buffer, 1:10 Taq polymerase, 100-1,500 base pair (bp) DNA ladder, 0.5 M ethylenediaminetetraacetic acid (EDTA), and isopropanol were purchased from Bio Basic Inc. (Markham, ON, CA). RPMI 1640 medium (mod.) 1X with L-glutamine (RPMI 1640) was purchased from Corning Inc. (Corning, NY, USA). Trypan blue, phosphate buffered saline (PBS) (pH 7.4), and 100X antibiotic-antimycotic (AB/AM) was purchased from Gibco (Waltham, MA, USA). Forward and reverse primers as well as 1X trisaminomethane (Tris) EDTA buffer (TE) (10 mM Tris, 0.1 mM EDTA, pH 8.0) was purchased from Integrated DNA Technologies (Coralville, IA, USA). MilliQ water was obtained from the Department of Chemistry at McGill University (Montréal, QC, CA). 10X T4 DNA ligase buffer, T4 DNA ligase, and exonuclease III (ExoIII) was purchased from New England Biolabs (Ipswich, MA, USA). Resazurin sodium salt was purchased from Sigma-Aldrich (St. Louis, MO, USA). EcoRI

restriction enzyme, 10X EcoRI buffer and fetal bovine serum (FBS) was purchased from Thermo Fisher Scientific (Waltham, MA, USA). 4-hydroperoxycyclophosphamide (4-HPCP) was purchased from Toronto research chemicals (North York, ON, CA). Anhydrous ethanol was purchased from VWR International (Radnor, PA, USA).

Commercial kits

The Monarch® PCR and DNA Cleanup Kit (5 µg) was purchased from New England Biolabs (Ipswich, MA, USA). The Wizard® Genomic DNA Purification Kit was purchased from Promega (Madison, WI, USA).

Plasmids and cell lines

ALL-SIL cells were purchased from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (Braunschweig, Germany). CCRF-CEM cells were purchased from Cedarlane Laboratories (Burlington, ON, CA). Jurkat cells were a generous gift from the Sleiman Lab at McGill University (Montréal, QC, CA). pCS1748 plasmid was obtained from the McKeague Lab at McGill University (Montréal, QC, CA).

Methods

Formation of circular DNA

250, 330, and 400 base pair (bp) oligonucleotides with an added EcoRI restriction site were synthesized using 100 µL PCR reactions containing 1 µL of 1 µM pCS1748 plasmid stock, 1 µL of 10 µM corresponding forward and reverse primers (Table 3), and 97 µL of master mix composed of 2 µL of 10 mM dNTP mix, 3 µL of 5 mM $MgCl_2$, 10 µL of 10X Taq buffer, 1µL of 1:10 Taq polymerase, and 80 µL of milliQ water. A negative control was also run using the same master mix and primers, but without the plasmid template. The cycle parameters were 2 minutes

at 95˚C; 30 cycles of 95˚C for 30 seconds, 55˚C for 30 seconds, 72˚C for 30 seconds; and 5 minutes at 72˚C, in a Mastercycler Nexus X2 Thermocycler.

Table 3. Primer sequences used for oligonucleotide synthesis. The underlined sequence indicates the EcoRI restriction site which was added.

Forward primer	AGTCTGAATTCTGATATTTAAGTTAATAAACGGTCTTCA
Reverse primer (250 bp)	GTAGTGAATTCCCTTCTATTTCAAATTCATGTCCAT
Reverse primer (330 bp)	GTAGTGAATTCATGCAAATGGTAATGGGCC
Reverse primer (400 bp)	GTAGTGAATTCTCTGGAATGTCGGCGG

After PCR, the samples and negative control were run on a 1.5% agarose gel with a 100-1,500 bp DNA ladder for 60 minutes at 95 V. Then, the gel was imaged with a Bio-Rad Gel Doc XR+ Imaging System to verify the successful formation of the oligonucleotides. If successful, the linear DNA samples were purified using the Monarch® PCR & DNA Cleanup kit, following the manufacturer's instructions. Briefly, the PCR samples were diluted in a 5:1 ratio of DNA Binding Buffer:sample. Then, the sample was loaded onto a column and spun through. The column was washed twice with 700 µL DNA Wash Buffer and then transferred to a clean 1.5 mL microcentrifuge tube. The PCR product was then eluted in 15 µL of TE buffer. The purity of each PCR product was determined using a NanoDrop Lite Spectrophotometer.

To form the rings, the purified PCR samples were split into 1 µg samples and digested with 1 µL EcoRI enzyme in a buffer consisting of 2 µL 10X EcoRI buffer and 10 µL milliQ water for 4 hours at 37˚C to obtain linear DNA fragments with sticky ends. The samples were heated to 65˚C for 20 minutes following the digestion to inactivate the enzyme. Each digested sample was then added to new PCR tubes containing 3 µL 10X T4 DNA ligase buffer to reach a DNA concentration greater than 0.03 µg/µL and the remaining DNA was left untreated to serve as a linear control. Then, 0.2 µL of T4 DNA ligase was added along with 8 µL milliQ water.

The samples were left at room temperature for 20 minutes followed by an incubation at 65˚C for 10 minutes to deactivate the ligase. To verify the success of the ligation, the products were loaded onto a 1.5% agarose gel along with the corresponding linear DNA as a control. The gel was run for 60 minutes at 95 V and then imaged using a Bio-Rad Gel Doc XR+ Imaging System.

To verify the ring formation, the product was split in half to obtain two samples of approximately 0.01 μg/μL of circularized DNA. To one half, 0.5 μL of ExoIII was added, and the other half was maintained as a control. The samples were all incubated for one hour at 37˚C and then 0.5 M EDTA was added to a final concentration of >11 mM to stop the reaction and the enzyme was inactivated at 70˚C for 30 minutes. The final product was run on a 1.5% agarose gel alongside the undigested controls and a 100-1,500 bp ladder for 60 minutes at 95 V and then imaged on a Bio-Rad Gel Doc XR+ Imaging System.

Mammalian tissue culture

All cell culture procedures were performed in a biosafety cabinet using proper sterile technique. FBS and AB/AM were added to RPMI 1640 to a final concentration of 10% FBS and 1% AB/AM to make complete media. The complete media was warmed in a 37˚C water bath and once warm, 9 mL was placed in a T-25 cell culture flask. A freezer stock of Jurkat cells containing 1 mL of 1 x 10^6 cells/mL was thawed and added to the media in the cell culture flask. Then, the flask was placed in a Forma Water Jacketed CO_2 Incubator (FORMA) at 37˚C with 5% CO_2 overnight. After 16 hours, the cells were transferred into a 15 mL Falcon tube and spun in a centrifuge at 3,000 x g for 3 minutes. The supernatant was removed, and the cells were resuspended in 10 mL of fresh warm media and transferred to a new T-25 cell culture flask. The cells were left in the 37˚C FORMA cell incubator at 5% CO_2 for four days. After four days of growth, the cells were counted using the cell counting function of the Tecan Spark microplate

reader. 20 µL of cell suspension was mixed thoroughly with 20 µL of Trypan Blue and then 10 µL of this mixture was loaded onto a Cell Chip™. This produced an output of both the number of cells per millilitre as well as the percent cell viability.

The cell suspension was transferred from the cell culture flask into a sterile Falcon tube and centrifuged at 3,000 x g for 3 minutes. The supernatant was removed, and the cells were resuspended in 5 mL of fresh warm media. Given the cell concentration calculated from the hemocytometer, the cells were diluted to approximately 200,000 cells/mL in a new T-25 cell culture flask with fresh warm media for a final volume of 10 mL. Then, they were placed in the 37˚C FORMA cell incubator at 5% CO_2. This process was repeated every four days.

The mammalian cell culture protocol described above was followed for both the ALL-SIL and CCRF-CEM cell lines as well. All three cell lines (Jurkat, ALL-SIL, and CCRF-CEM) were maintained in culture simultaneously such that the passage numbers aligned for all subsequent experiments.

Extraction of genomic DNA

Approximately 3.5 x 10^6 total cells were harvested per cell line (Jurakt, ALL-SIL, and CCRF-CEM), placed in a 1.5 mL microcentrifuge tube, and centrifuged at maximum speed (13,000 rpm) for 10 seconds. The supernatant was removed, and the cells were washed using 200 µL PBS. Then, the genomic DNA was extracted using the Wizard® Genomic DNA Purification Kit as per the manufacturer's instructions. Briefly, 600 µL nuclei lysis solution was added to the cell suspensions and mixed by pipetting up and down until no visible cells remained. 3 µL RNase solution was added to each sample and the samples were mixed by inverting before being incubated at 37˚C for 30 minutes. After the incubation, the samples were cooled at room temperature for 5 minutes. Then, 200 µL protein precipitation solution was added to each

sample, they were vortexed vigorously for 20 seconds and then chilled on ice for 5 minutes. To pellet the proteins, the samples were centrifuged for 4 minutes at maximum speed. The supernatants containing the DNA were carefully transferred to clean 1.5 mL microcentrifuge tubes and 600 µL room temperature isopropanol was added to each one. These mixtures were mixed gently by inversion until a visible mass of thread-like DNA could be observed in each tube. Once visible, the mixtures were centrifuged for one minute at 13,000 rpm to pellet the DNA and then the supernatant was slowly decanted from each sample. 600 µL 70% ethanol was added to each tube to wash the DNA followed by another centrifugation step at 13,000 rpm for one minute. The ethanol was aspirated from each sample using a Pasteur pipette, taking care not to disturb the DNA pellets. Then, the tubes were inverted on a clean KimWipe and allowed to air dry for 15 minutes. Finally, 100 µL of 1X TE buffer was added to each dried pellet and the samples were incubated at 65˚C for 1 hour. During this hour, the mixtures were agitated every 10 minutes by gently flicking the tube. Once rehydrated, the DNA concentrations and purities were determined using a NanoDrop Lite Spectrophotometer and the samples were stored at -20˚C.

Sonication of genomic DNA

Extracted genomic DNA was sonicated to produce fragments between 250 and 400 bp in size to reflect the sizes of the oligonucleotides that were circularized. The genomic DNA was thawed and then diluted to a concentration of 40 ng/µL in a total volume of 100 µL. These samples were then loaded into a QSonica Q800R Sonicator chilled to 4°C. The samples were sonicated for increasing amounts of time with varying amplitudes. For each condition, the total on time was adjusted as well as the intervals in which the sonicator turned on and off and the amplitude.

Making 1000X and 10X resazurin stocks

Resazurin stocks were made to ensure that fresh resazurin dye at the working concentration would be available for each viability assay. 1000X resazurin stocks were made by dissolving 0.1 g of resazurin sodium salt in 9 mL PBS. The solution was then sterilized using 0.22 μM syringe filters and dispensed into 1 mL aliquots. Each aliquot was wrapped in tin foil and stored at -20°C indefinitely.

When ready to perform a viability assay,10X resazurin stocks were made by diluting 1 mL of 1000X resazurin in 99 mL of PBS. This solution was then aliquoted in 15 mL falcon tubes, wrapped in tin foil and stored at 4°C. These stocks were utilizable for 6 months.

Establishment of cell-line susceptibility to 4-hydroperoxycyclophosphamide

The cytotoxicity of 4-HPCP was determined using a resazurin cell viability assay. Jurkat and CCRF-CEM cells were seeded in a 96-well plate at a concentration of 25,000 cells per well in 50 μL complete RPMI 1640 media. ALL-SIL cells were seeded in a 96-well plate at a concentration of 50,000 cells per well in 50 μL complete RPMI 1640 media. All cells were plated identically such that there were six replicates for each drug concentration to be tested as well as the positive and negative controls. Then, the cells were left to rest for a minimum of 2 hours in the cell incubator. After the two hours, a 200 μg/mL stock solution of 4-HPCP was prepared by dissolving 1 mg of 4-HPCP in 5 mL complete RPMI 1640 media. Dilutions were created by performing a 1:8 serial dilution for 8 total 4-HCPC concentrations (0, 0.00001, 0.0008, 0.006, 0.05, 0.4, 3.1, 25, and 200 μg/mL). 50 μL of each 4-HCPC drug concentration was added to the appropriate wells for all three cell lines (Figure 6). The cells were incubated with the drug treatment for 72 hours.

	1	2	3	4	5	6	7	8	9	10	11	12
A	water											
B	Dye only	0 µg/mL	0.000048 µg/mL	0.000381 µg/mL	0.00305 µg/mL	0.0244 µg/mL	0.195 µg/mL	1.56 µg/mL	12.5 µg/mL	100 µg/mL	Triton-X	
C												
D												
E												
F												
G												
H												

Figure 6. Loading scheme for resazurin cell viability assay. The Triton-X column was the positive control and received 100 µL media without any drug, same as the 0 µg/mL negative control cells. In the dye only column, 100 µL PBS was added. Water was added to the surrounding cells (indicated in grey) to prevent uneven evaporation in the assay wells.

After 72 hours, 2 µL Triton-X 100 was added to each of the Triton-X positive control wells to obtain a final concentration of approximately 2%. A working stock of resazurin was made by diluting 8 mL of 10X resazurin in 4 mL of PBS. 40 µL of the working resazurin stock was added to each well except for the wells in row G. The cells were incubated for 3 hours before fluorescence measurements were taken with a microplate reader (Excitation/Emission: 550 nm/590 nm).

Resazurin assay data and statistical analysis

Fluorescence data obtained from the resazurin assay was exported and analyzed in Excel and plotted in GraphPad Prism v9.0.2. In Excel, the values obtained from the "dye only" column were averaged to give an average background. This average background value was subtracted from each experimental row. Then, the values in row G were subtracted from the background-free values in each corresponding column (ex. $B3 - G3$, $B4 - G4$, etc.). These adjusted values were then imported into GraphPad Prism v9.0.2 in an XY Table with 5 replicates for each Y-value. The 4-HPCP concentrations were transformed to logarithmic values using the GraphPad "Transform" function and normalized to the fluorescence value corresponding to the lowest drug dose (i.e., the value representing no cell death). Finally, the data was fitted to a non-linear

regression curve, specifically the "[Inhibitor] vs. response – Variable slope (four parameters)" equation provided in GraphPad, to produce a normalized dose-response curve. The statistical significance of the difference in drug sensitivity between the three cell lines was calculated by comparing the fold shift of the IC50s and their associated confidence intervals using GraphPad.

Results

Towards a method for sequencing crosslinks: Step 1 - Formation of DNA Rings

Evidence to date indicates that ICLs are relatively rare in the genome, yet they constitute an extremely important form of damage which can result in cell death if left unrepaired. As such, there is a need for a sequencing method that will allow ICLs to be distinguished from other forms of damage (i.e., intrastrand crosslinks, DSBs, etc.), selectively enriched, and sequenced. We sought to develop a method to sequence ICLs at a single nucleotide-level resolution using the scheme illustrated below (Figure 7). We designed our method in a way that would enable the specific capture of only the ICL sequences, and not DSBs or undamaged DNA.

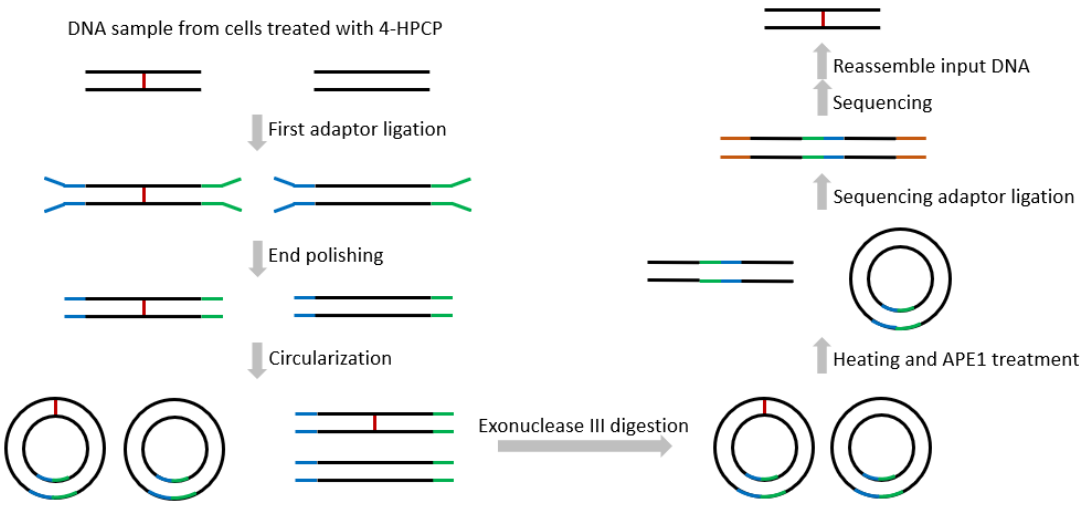


Figure 7. Proposed method for the single-nucleotide resolution sequencing of interstrand crosslinks. The basis of this proposed method is that a singular ICL, as indicated by the red line, can be isolated per ring formed such that each adduct can be mapped directly back to its site of origin.

The proposed scheme starts with genomic DNA sample obtained from cells treated with cyclophosphamide. The DNA must first be fragmented into short pieces and adaptors would be ligated to the ends for later identification and mapping to the genome. These same adaptors are designed to allow circularization of the linear DNA fragments through the generation of specific "sticky ends". Then, the formed DNA rings would be heated as alkylated G-Nor-G crosslinks are much less stable and are readily depurinated through heating, leaving an AP-site (85). Previous work in the McKeague lab has shown that the treatment of these AP-sites with APE1, an apurinic/apyrimidinic endonuclease involved in the base excision repair pathway, efficiently yields a strand break. Thus, only the DNA rings containing ICLs would become linearized enabling sequencing adaptors to be ligated to the ends. These sequencing adaptors would serve as a barcode, allowing for the identification of the damage site at nucleotide resolution after sequencing. All DNA rings that were not re-linearized would remain as rings that would not be amplified or read through sequencing. With this "ring-formation" enrichment method, we would thus be able to obtain a genome wide map of damage caused by cyclophosphamide treatment, allowing for the identification of any damage distribution patterns.

Towards the development of this sequencing method, we first needed to determine the conditions for and efficiency of producing double stranded DNA rings from linear fragments. As such, we first designed primers to amplify three random sequences that were 250, 330, and 400 bp in size from a plasmid template. The three linear DNA fragments of different sizes were successfully created by PCR (Figure 8), confirming that 30 rounds of PCR were sufficient. We next used these linear fragments to generate circular DNA. To increase the efficiency of the ligation, we induced sticky ends by digesting the EcoRI sites which were embedded in the primers used for PCR (Table 3).

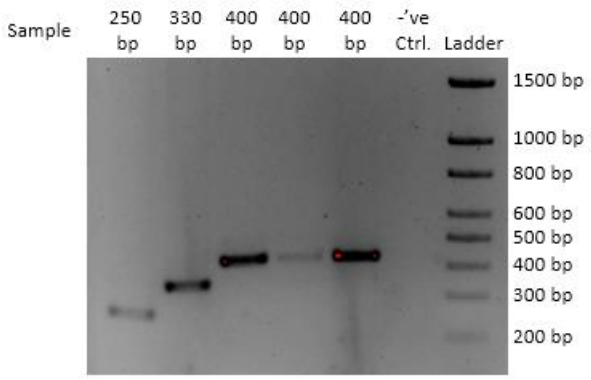


Figure 8. Generation of three different sized linear DNA fragments by PCR. 250, 330, and 400 bp linear DNA fragments were successfully generated from the pCS1748 plasmid. The 400bp sample was run in triplicate to test a new Taq polymerase and verify its efficiency.

The presence of the sticky ends resulted in the successful generation of multiple constructs of varying sizes in the presence of T4 DNA ligase (Figure 9). However, while the gel indicated that the small linear DNA fragments had ligated to form larger constructs, it did not differentiate between linear and circular DNA structures.

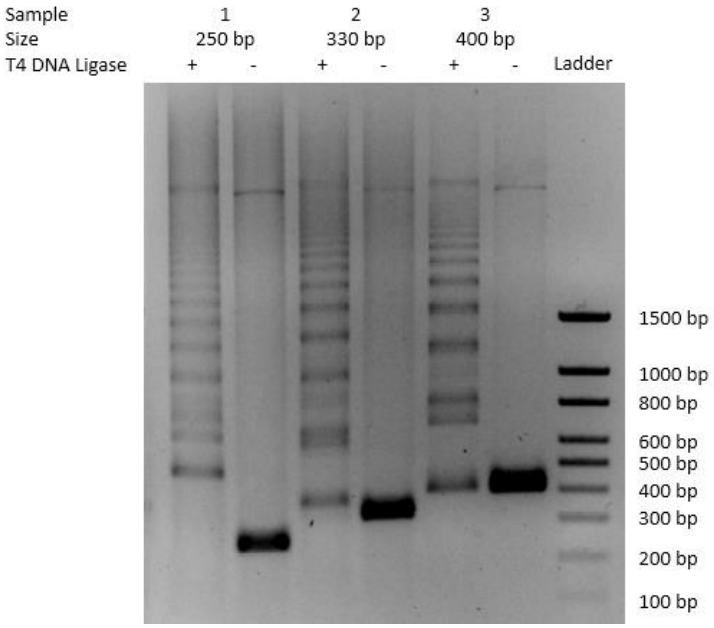


Figure 9. Circular DNA formed by ligation of EcoRI sticky ends. The addition of ligase resulted in the formation of many different sized constructs, as shown by the appearance of multiple bands in the lanes with added T4 DNA ligase versus the control lanes which contained only the EcoRI digested linear DNA. The large band seen at the top of each lane is assumed to be a large linear product resulting from the self-ligation of many linear pieces of DNA together since all of the PCR amplified constructs contained the same sticky ends resulting from EcoRI digestion.

To confirm the identity of the constructs observed in Figure 9 as rings, each sample was digested with ExoIII, an exonuclease enzyme which digests any DNA with a free 3′-hydroxyl terminus. All of the bands remaining after digestion were therefore determined to be rings stable to degradation (Figure 10). The bands of the smallest size matched with the corresponding linear DNA sizes in the case of the 330 bp and 400 bp samples. For the 250 bp sample, it appeared that two linear pieces of DNA ligated together and then formed a 500 bp ring which could be due to a steric preference. This idea is further supported by the presence of multiple bands remaining for each sample despite the exonuclease digestion, indicating that there may be the formation of non-monomeric rings for each DNA fragment size.

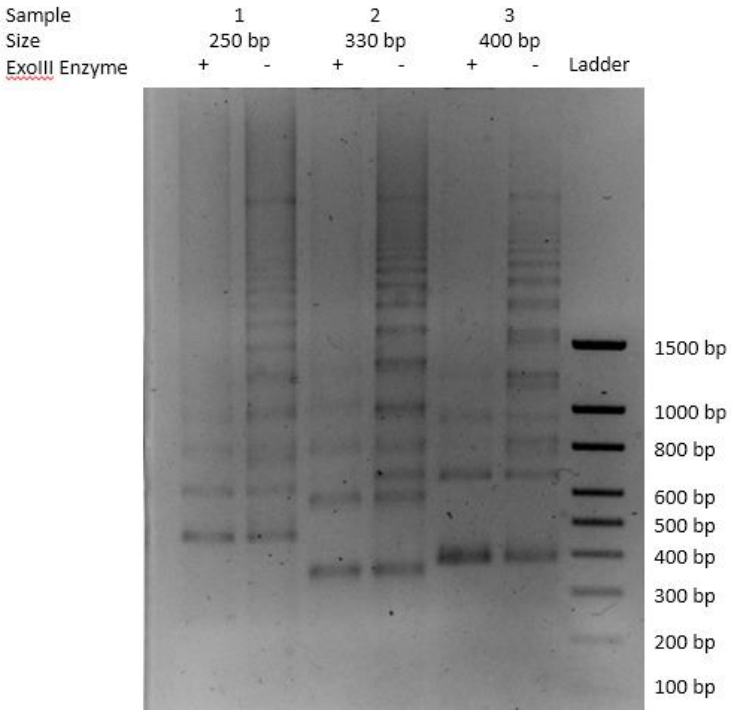


Figure 10. Confirmation of circular DNA presence. The bands shown in the lanes after degradation with ExoIII indicate the presence of fully circularized DNA. This is further confirmed by the size of the bands which either match the size of the corresponding linear DNA or are an exact multiple of the corresponding linear DNA size. In each sample lane, there is also the presence of multiple bands of different sizes which suggests the presence of many different sized rings, particularly as the additional bands seem to be multiples of the original linear band size. It is possible that the sizing does not match perfectly as larger circular DNA may travel slightly differently through the agarose gel.

Extraction of genomic DNA from acute lymphoblastic leukemia cell lines for biomarker testing

For future testing of the sequencing method on the genomic DNA of our model system, the conditions for DNA extraction using the commercially available Promega kit had to be optimized. The manufacturer's information indicated that a total yield of between 15 and 30 μg could be obtained from 3 x 10^6 human K562 cells. However, there was no directive for ALL-SIL, CCRF-CEM, or Jurkat cells specifically. Thus, to confirm the applicability of this number of cells to our cell lines, the extraction was performed with 3.5 x 10^6 cells on three separate occasions, and consistent yields totalling between 10 and 20 μg were obtained. Furthermore, the A260/A280 value was consistently above 1.80, indicating good purity. Thus, it was confirmed that for our cell lines, approximately 3.5 x 10^6 cells is a sufficient number for DNA extraction with good yield and purity.

Sonication of genomic DNA can achieve desired fragment sizes

Towards the goal of using our method for whole genome sequencing, we needed to be able to produce DNA fragments between 250 and 400 bp in size. This was achieved through sonication which produced a nice spread of fragment sizes. Several conditions were tested, but a 15 second on/off interval for a total on time of 3 minutes at 40% amplitude was deemed to produce the best fragment sizes (Figure 11). However, the spread is quite large, so additional optimization will be required. Since NGS library preparation is quite strict in terms of input size, we want to limit the amount of DNA that will not be sequenced from our sample. Thus, we will aim to reduce the spread such that the DNA fragment sizes will not exceed the range of 250 to 400 bp.

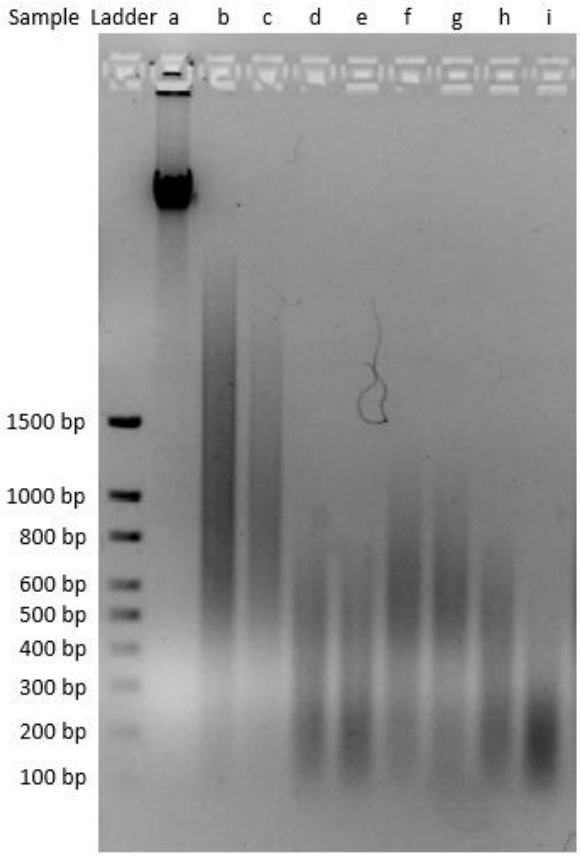


Figure 11. Optimization of sonication conditions. The bands show the spread of DNA fragment sizes following sonication. Lane a contains the native genomic DNA; lane b shows 30 seconds of sonication in 5 second intervals at 100% amplitude; lane c shows 1 minute of sonication in 5 second intervals at 100% amplitude; lane d shows 3 minutes of sonication in 15 second intervals at 40% amplitude; lane e shows 5 minutes of sonication in 15 second intervals at 40% amplitude; lane f shows 2 minutes of sonication in 1 minute intervals at 30% amplitude; lane g shows 3 minutes of sonication in 1 minute intervals at 30% amplitude; lane h shows 5 minutes of sonication in 1 minute intervals at 30% amplitude; and lane i shows 10 minutes of sonication in 1 minute intervals at 30% amplitude. The spread of fragment size was determined by looking at the length of the band. Lanes d, e and h show fragment sizes that fall within the desired range.

The active cyclophosphamide metabolite is differentially cytotoxic to acute lymphoblastic leukemia cell lines

To test the sequencing method, the appropriate drug dose to use for the treatment of our cells to produce DNA damage was required. However, to our knowledge, there is no published half maximal inhibitory concentrations (IC_{50}) for 4-HPCP in ALL-SIL, CCRF-CEM, or Jurkat cell lines. To address this issue, we performed dose-response toxicity assays using a resazurin

cell viability assay. This assay was chosen because of its rapid, accurate, and simple workflow. It was previously confirmed in the lab that the presence of 4-HPCP did not affect the resazurin dye. Thus, the drug-containing media did not have to be removed prior to resazurin addition. We found this to be an issue with other assays such as the Promega CellTiter-Glo (data not shown) as the cells are suspension cells. Thus, when removing media, some cells were consistently removed despite centrifuging the plate, introducing a large standard deviation to the data.

The resazurin assay quantifies the number of live cells in a sample by detecting the activity of the mitochondrial respiratory chain. The resazurin dye begins as an oxidized non-fluorescent blue dye which is reduced to the red fluorescent dye resorufin. The amount of resorufin produced is directly proportional to the number of metabolically active, and therefore viable cells [82]. For the assay, an initial dose range of 0.000048 μg/mL to 100 μg/mL was chosen for 4-HPCP based on previous work done in the lab. This dose range was large enough to encompass both the IC_0 and IC_{100} (i.e., no cell death and complete cell death) of all three cell lines despite their differential susceptibility to 4-HPCP treatment (Figure 12). From Figure 12, the IC_{50} was determined to be 2.4, 0.3, and 0.7 ng/mL for All-SIL, CCRF-CEM, and Jurkat cells, respectively. This shows that the three cell lines are differentially sensitive to 4-HPCP treatment with Jurkat cells requiring an 8-fold higher dose of 4-HPCP to achieve the same IC_{50} response as CCRF-CEM cells.

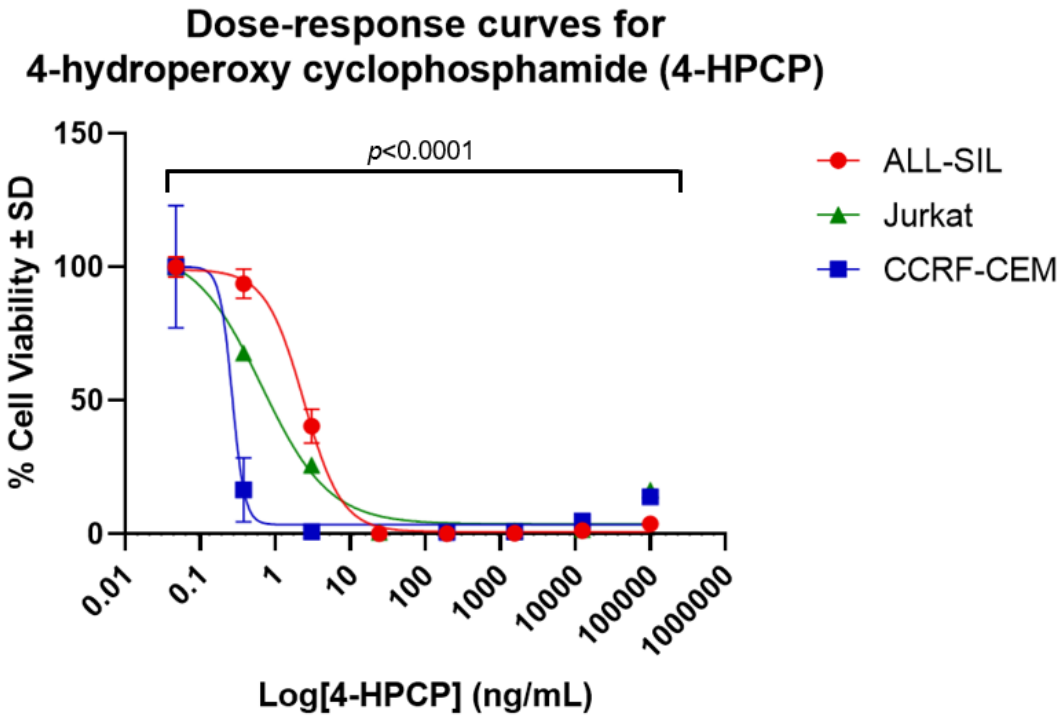


Figure 12. Dose-response curves measuring ALL-SIL, CCRF-CEM, and Jurkat cell susceptibility to the active cyclophosphamide metabolite. The average percent cell viability was calculated individually within each sample (n = 5 per sample). The dose range of 4-HPCP induced complete cell death above 100 ng/mL for all cell lines and, below 0.1 ng/mL, no cell death was observed. The R^2 values of each curve were 0.99, 0.92, and 0.97 for the ALL-SIL, CCRF-CEM, and Jurkat cells, respectively. GraphPad analysis of the fold shift of the IC50s and their confidence intervals revealed a statistically significant difference in drug sensitivity between the three cell lines. The error bars represent the standard deviation.

Summary of chapter

This chapter aimed to establish the framework for a novel ICL sequencing method using synthetic oligonucleotides. ICLs remain an elusive form of DNA damage as current methods of detection are either unable to distinguish between ICLs and intrastrand crosslinks or do not provide a high enough resolution to properly investigate the genomic context in which ICLs form. However, ICLs constitute a major type of damage that arises in response to chemotherapeutic damage. Thus, a method to pinpoint the genomic locations where ICLs form is needed.

The results presented in Chapter 2 show that we have successfully established the DNA circularization step, which is the most important as this is our proposed method to selectively enrich for DNA containing ICLs. Furthermore, we have shown that the circularizable DNA is of a size, specifically 250 bp to 400 bp, which aligns well with the requirements for NGS library preparation. Towards the application of this ICL sequencing method *in vitro*, we established the IC_{50}s of three ALL cell lines that are differentially sensitive to cyclophosphamide treatment. The larger implications of these findings will be discussed in Chapter 5.

Chapter 3: AP-Site Approach

Preface

AP-sites are a common form of damage that can arise either in response to DNA damaging agents or as an intermediate of the DDR, particularly the BER pathway [83,84]. Recent work has shown that the transcription factor, CUX1, plays an important role as an accessory factor in the BER pathway [49-54]. Interestingly, studies suggest that DNA damage distribution may be influenced by the binding of transcription factors [54]. Thus, it is possible that CUX1 expression may influence the pattern of DNA damage, specifically AP-sites, in the genome. snAP-seq is a novel sequencing technique that allows AP-sites across the genome to be sequenced at a nucleotide-resolution [80]. The application of snAP-seq in samples with either wild-type expression of CUX1 or a *CUX1* knockdown may help elucidate the role of this protein in influencing the distribution of AP-sites.

Materials

Reagents

Any reagents also used in Chapter 1 of this thesis may be found there. The HIPS probe was synthesized by Serge Hirka and Kaleena Basran from the McKeague and Luedtke labs, respectively, at McGill University (Montréal, QC, CA). The P7 adaptor was synthesized by Dr. Eiman Osman from the McKeague Lab at McGill University (Montréal, QC, CA). AMPure XP beads were provided by Daniel Shapoznikov from the Szyf Lab at McGill University (Montréal, QC, CA). Sodium hydroxide (NaOH) and trisaminomethane hydrochloride (Tris-HCl) were purchased from Bio Basic Inc. (Markham, ON, CA). All synthetic oligonucleotides and P5 adaptors were purchased from Integrated DNA Technologies (Coralville, IA, USA). Shrimp alkaline phosphatase, CutSmart buffer, NEBuffer 2, rCutSmart buffer, DNA polymerase I large

(Klenow) fragment, blunt/TA ligase master mix, uracil-DNA glycosylase (UNG) and Q5 hot start high fidelity master mix were purchased from New England Biolabs (Ipswich, MA, USA). Copper bromide (CuBr), tris-hydroxypropyltriazolylmethylamine (THPTA), biotin-PEG3-azide, doxycycline, Tween-20, poly(deoxyinosinic-deoxycytidylic) acid (Poly(dI-dC)), and puromycin were purchased from Sigma-Aldrich (St. Louis, MO, USA). Streptavidin MagneSphere® Paramagnetic Particles were purchased from Thermo Fisher Scientific (Waltham, MA, USA).

Commercial kits and materials

DNA Damage Quantification Kit – AP-site Counting was purchased from Dojindo Molecular Technologies, Inc. (Rockville, MD, USA). Cytiva Amersham MicroSpin G-25 Columns were purchased from Global Life Sciences (Twickenham, UK). Amicon Ultra-0.5 mL 10K centrifugal filters were purchased from Thermo Fisher Scientific (Waltham, MA, USA). NEBNext Ultra II DNA Library Preparation Kit and NEBNext Library Quant Kit for Illumina® was purchased from New England Biolabs (Ipswich, MA, USA). ssDNA/RNA Clean & Concentrator was purchased from Zymo Research (Irvine, CA, USA). LightCycler® 480 Clear Multiwell qPCR Plate was kindly provided by Jathavan Asohan from the Sleiman Lab at McGill University (Montréal, QC, CA).

Cell lines

The human breast cancer cell line, MDA 231, was kindly provided by the Nepveu Lab at McGill University (Montréal, QC, CA).

Methods

Mammalian tissue culture

All cell culture procedures were performed in a biosafety cabinet using proper sterile technique. FBS, AB/AM and puromycin were added to RPMI 1640 to a final concentration of 10% FBS, 1% AB/AM and 0.5 µg/mL puromycin to make complete media. The complete media was warmed in a 37˚C water bath and once warm, 9 mL was placed in a T-25 cell culture flask. A freezer stock of MDA 231 cells containing 1 mL of 1.5 x 10^6 cells/mL was thawed and added to the media in the cell culture flask. Then, the flask was placed in a Forma Water Jacketed CO_2 Incubator (FORMA) at 37˚C with 5% CO_2 overnight. After 16 hours, the media was replaced and the cells were left in the 37˚C FORMA cell incubator at 5% CO_2 for four days. After four days of growth, the cells were observed under the microscope and confluency was estimated.

When the cells reached between 80% to 100% confluency, they were split in a 1:10 ratio as follows. The media was removed by vacuum and the cells were gently washed with 2 mL PBS. The PBS was removed and 2 mL of pre-warmed trypsin was added. The cells were placed back in the 37˚C incubator for 5 minutes until they detached and then the flask was brought back into the biosafety cabinet. 8 mL of media was added to the flask and the cells were gently broken apart by pipetting up and down. 1 mL of the cell suspension was then added to a new T-25 cell culture flask and 9 mL of fresh media was added. Then, they were placed back in the 37˚C FORMA cell incubator at 5% CO_2. This process was repeated every four days.

Induction of CUX1 knockdown

The MDA 231 cells contain a lentiviral Tet-On system for doxycycline-inducible knockdown of *CUX1*. To activate the *CUX1* knockdown, the cells were passaged as described

above and then doxycycline was added to a final concentration of 100 ng/mL. The cells were placed back in the 37˚C FORMA cell incubator at 5% CO_2 and were left for four days.

Extraction of genomic DNA

Approximately 3.5 x 10^6 cells were harvested, placed in a 1.5 mL microcentrifuge tube, and centrifuged at maximum speed (13,000 rpm) for 10 seconds. The supernatant was removed, and the cells were washed using 200 µL PBS. Then, the genomic DNA was extracted using the Wizard® Genomic DNA Purification Kit as per the manufacturer's instructions. Briefly, 600 µL nuclei lysis solution was added to the cell suspension and mixed by pipetting up and down until no visible cells remained. 3 µL RNase solution was added and the sample was mixed by inverting before being incubated at 37˚C for 30 minutes. After the incubation, the sample was cooled at room temperature for 5 minutes and then 200 µL protein precipitation solution was added. The sample was vortexed vigorously for 20 seconds and then chilled on ice for 5 minutes. To pellet the proteins, the sample was centrifuged for 4 minutes at maximum speed. The supernatant containing the DNA was carefully transferred to a clean 1.5 mL microcentrifuge tube and 600 µL room temperature isopropanol was added. This mixture was mixed gently by inversion until a visible mass of thread-like DNA could be observed. Once visible, the mixture was centrifuged for 1 minute at maximum speed to pellet the DNA and then the supernatant was slowly decanted. 600 µL 70% ethanol was added to wash the DNA followed by another centrifugation step at maximum speed for 1 minute. The ethanol was aspirated using a Pasteur pipette, taking care not to disturb the DNA pellet. Then, the tube was inverted on a clean KimWipe and allowed to air dry for 15 minutes. Finally, 100 µL of 1X TE buffer was added and the sample was incubated at 65˚C for 1 hour. During this hour, the mixture was agitated every 10 minutes by gently flicking the tube. Once rehydrated, the DNA concentration and its purity were

determined using a NanoDrop Lite Spectrophotometer and the sample was immediately treated with the HIPs probe.

Quantification of AP-sites

The number of AP-sites in wild-type and *CUX1* knockdown MDA 231 cells was quantified using the Dojindo DNA Damage Quantification Kit, following manufacturers instructions. Briefly, DNA was extracted from the cell samples and diluted to 100 ng/µL. Then, 10 µL of each sample was immediately incubated with 10 µL of ARP solution for 1 hour at 37°C. The labelled samples were flowed through a Filtration Tube cup, prewashed twice with TE. The DNA samples were then applied to the Filtration Tube cups and the purified samples were collected.

To determine the number of AP-sites in the DNA, the ARP-labelled samples were diluted with TE and added to the well plate. The Standard ARP-DNA Solutions were added as well and then the plate was placed in the fridge overnight. After the overnight incubation, the wells were washed and then the data was collected by plate reader.

Preparation of AP-site containing oligonucleotides

Synthetic oligonucleotides containing AP-sites cannot be ordered. Thus, the AP-site must be generated using UNG, a glycosylase that removes deoxyuracil bases from DNA. Two different AP-site containing oligonucleotides were produced: 1) AP-ODN, a short single-stranded oligonucleotide containing one AP-site; and 2) AP-DNA, a longer double stranded DNA oligonucleotide containing one AP-site (Table 5). Before UNG treatment, 10 µL of 10 µM forward and reverse AP-DNA were added to a microcentrifuge tube and topped up with TE buffer to a final concentration of 1 µM each. This mixture was then placed in a heat block set to

95°C for 2 minutes and then the heat block was turned off and the mixture was left to cool slowly to room temperature. Once annealed, the AP-DNA oligonucleotide was ready for use.

To produce an AP-site, 10 µL of 1 µM annealed AP-ODN and 10 µL of 1 µM annealed AP-DNA were mixed with 1 µL of 10X UNG reaction buffer and 1 µL of UNG, respectively in a thin-wall PCR tube. The PCR tubes were then incubated for 30 minutes at 37°C followed by a 10 minute incubation at 95°C to inactivate UNG.

Table 4. Oligonucleotide sequences used for snAP-seq. The "**U**" indicates a deoxyuracil which is enzymatically removed to yield an AP-site at that position.

AP-ODN	AGCGACA**U**ATCTTGT
AP-DNA	AGCGACA**U**ATCTTGTAGATC/FAM/
GCAT DNA	AGCGACATATCTTGTAGATC/FAM/
GCAT dsDNA (Top)	GGCCACCACCCGCACATACTCTGGTACGATTACGAACACAGCC CGACACCACCTCTAATGAACGTCGCTTATAGTGATTAACGCCC CGTAGACACCATGG
GCAT dsDNA (Bottom)	CCATGGTGTCTACGGGGCGTTAATCACTATAAGCGACGTTCATT AGAGGTGGTGTCGGGCTGTGTTCGTAATCGTACCAGAGTATGTG CGGGTGGTGGCC
Custom P7 Adaptor (Top)	/2′-5Me/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Custom P7 Adaptor B1 (Bottom)	/5′-Phosphate/GATCGGAAGAGCACACGTCTGAACTCCAGTCACAT CACGATCTCGTATGCCGTCTTCTGCTTG/Spacer C3/
Custom P7 Adaptor B2 (Bottom)	/5′-Phosphate/GATCGGAAGAGCACACGTCTGAACTCCAGTCACCG ATGTATCTCGTATGCCGTCTTCTGCTTG/Spacer C3/
Custom P5 Adapter (Top)	GAATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACA CGACGCTCTTCCGATCT

Custom P5 Adapter (Bottom)	/5′-Phosphate/GATCGGAAGAGCG
Library Amplification Primer (Forward)	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA
Library Amplification Primer (Reverse)	CAAGCAGAAGACGGCATACGAGAT

Validation of the HIPS probe

To validate that the HIPS probe binds to AP-sites, 10 µL of AP-ODN was mixed with 10 µL 10 mM HIPS probe 1 in sodium phosphate buffer (40 mM, pH 7.4) and incubated for 2 hours at room temperature. Then, the sample was sent for mass spectrometry analysis.

Annealing snAP-seq sequencing adaptors

The P7 and P5 adaptors were resuspended and then 15 µM stocks were made for each. The complementary top and bottom P7 and P5 oligonucleotides were combined in TE buffer in equal proportions and heated to 95°C for 2 minutes in a heat block. The heat block was then turned off and the samples were left in the heat block to cool to room temperature.

snAP-seq sequencing of MDA 231 wild-type and CUX1 knockdown genomic DNA

Genomic DNA extracted from wild-type and *CUX1* knockdown MDA 231 cells were treated identically. 10 µL of 100 ng/µL genomic DNA was mixed with 10 µL 10 mM HIPS probe in 40 mM sodium phosphate buffer (pH 7.4) for two hours at room temperature. After two hours, the DNA was diluted to 40 ng/µL in 100 µL total volume and sonicated with a QSonica sonicator, prechilled to 4°C. The sonicator parameters were set as 15 seconds ON followed by 15 seconds OFF for a total ON time of 5 minutes at 40% amplitude. After sonication, the samples

were purified with the Cytiva Amersham Microspin G-25 columns as per the manufacturer's instructions. Briefly, the resin in the columns was resuspended by vortexing and then the bottom closure of the columns was removed. Then, the column was placed in the supplied collection tube and spun for one minute at 735 x g to compact the resin and remove the storage buffer. The prepared columns were placed in a fresh DNase-free 1.5 mL microcentrifuge tube and the samples were slowly loaded to the top-center of the resin bed, taking care to load the samples onto the high side of the resin. To elute the samples, the columns were spun for two minutes at 735 x g. The purified DNA was then incubated with 25 µL 250 µM CuBr (prepared fresh), 50 µL 1.25 mM THPTA, and 10 µL 500 µM biotin-PEG3-azide (all in molar excess) for 2 hours at 37°C. During the last 30 minutes of the click reaction, two Amicon Ultra-0.5 mL 10K filters were prewashed with 500 µL water by adding the water to the filter and spinning the columns for 30 minutes at 14,000 x g. After the completion of the click reaction, the samples were loaded into the prepared Amicon Ultra-0.5 mL 10K centrifugal filters and spun for 15 minutes at 14,000 x g. The samples were then washed with 450 µL water followed by 450 µL 10 mM Tris-HCl buffer (pH 7.4); during each wash step, the columns were spun for 15 minutes at 14,000 x g. To elute the samples, the columns were flipped upside down and inserted into clean Amicon collection tubes, containing 50 µL 10 mM Tris-HCl buffer (pH 7.4), and were spun for 2 minutes at 14,000 x g. The purified samples were stored overnight at -20°C.

The next day, the samples as well as a GCAT dsDNA oligonucleotide (Table 4) to be used as a positive control were thawed on ice and then the annealed custom P7 sequencing adapter (Table 4) was ligated on using the NEBNext Ultra II DNA library preparation kit according to manufacturer's instructions. Briefly, the samples were combined with 3 µL NEBNext Ultra II End Prep Enzyme Mix and 7 µL NEBNext Ultra II End Prep Reaction Buffer

for a total volume of 60 µL in 0.1X TE buffer. The samples were then pipetted up and down 10 times to mix thoroughly and then they were placed in a thermocycler with the lid set to 80°C for 30 minutes at 20°C, followed by 30 minutes at 65°C. 2.5 µL of the annealed custom P7 adaptor (15 µM), 30 µL NEBNext Ultra II Ligation Master Mix, and 1 µL NEBNext Ligation Enhancer were then added to the samples and mixed thoroughly by pipetting up and down 10 times. Note that at this step, the *CUX1* knockdown and wild-type samples were treated with the custom P7 adaptor annealed with the custom P7 bottom adaptor B1 and B2, respectively. The samples were incubated for 15 minutes at 20°C in a thermocycler with the heated lid turned off. 1.5 µL shrimp alkaline phosphatase and 10 µL 10X rCutSmart buffer were added to the samples and incubated for 30 minutes at 37°C, followed by heat inactivation at 65°C for five minutes. The samples were then added to 140 µL of AMPure XP beads and left at room temperature for 5 minutes. The tubes were then placed in a magnetic bead stand and the supernatants were discarded. 200 µL 80% ethanol was added to the tubes and incubated for 30 seconds before being removed. This was then repeated once more and then the beads were left to air dry for 5 minutes. 48 µL 0.1X TE buffer was then added to the beads and the samples were vortexed and then left at room temperature for 5 minutes. Then, the tubes were placed again on a magnetic stand and the supernatant was transferred to a new, clean tube. The GCAT oligonucleotide was then stored in the fridge until the P7 adaptor PCR and P5 adaptor annealing step.

50 µL Magnesphere streptavidin beads were prewashed three times with 1X binding buffer. Then, they were resuspended in 48 µL 2X binding buffer. 2 µg poly(deoxyinosinic-deoxycytidylic) acid was added to the DNA samples and then the samples were added to the prewashed streptavidin beads. The tubes were vortexed and then incubated for 15 minutes at room temperature. The beads were then washed six times with 500 µL 1X binding buffer and

then 100 µL NaOH (100 mM) was added to the samples, and they were incubated for 10 minutes at room temperature to denature the DNA. The beads were then washed three times with 100 µL NaOH (100 mM) followed by three washes with 500 µL 1X binding buffer. The DNA was then eluted off the beads in 50 µL NaOH (100 mM) by incubating for 15 minutes at 70°C. Immediately after the 15 minutes, the reaction was quenched with 25 µL Tris-HCl (500 mM, pH 7.0). A fresh sample of 75 µL prewashed Magnesphere streptavidin beads was then prepared by washing three times with 1X binding buffer and then resuspending in 75 µL 2X binding buffer. 2 µg of poly(deoxyinosinic-deoxycytidylic) acid was then added to the beads with the neutralized DNA eluents and the samples were incubated for 15 minutes at room temperature. The supernatant was then collected from the beads and the samples were purified using the Zymo Research ssDNA/RNA Clean & Concentrator kit according to manufacturer's guidelines. Briefly, 150 µL DNA/RNA Binding Buffer was added to each sample and each sample was mixed by vortexing. Note that the Zymo-Spin IICR Column step was omitted and 150 µL 100% ethanol was added to the samples directly. They were mixed and then transferred to the Zymo-Spin IC Column and centrifuged for 30 seconds at 14,000 x g. The flow-through was discarded and then 400 µL DNA/RNA Prep Buffer was added to the columns. They were spun for 30 seconds at 14,000 x g and the flow through was discarded. 700 µL DNA/RNA wash buffer was added to the columns, and they were spun for 30 seconds at 14,000 x g. This wash step was repeated again with 400 µL DNA/RNA Wash Buffer. The columns were then transferred to clean tubes and 15 µL DNase/RNase-Free Water was added to the column matrix and the samples were eluted by centrifuging for 30 seconds at 14,000 x g.

6 µL dNTPs (1 mM), 3 µL of the custom P7 top adaptor (10 µM) (Table 4), 3 µL 10X NEBuffer 2, and 3 µL nuclease-free water was added to the 15 µL of purified samples, or 15 µL

of the prepared GCAT dsDNA oligonucleotide control, for a total volume of 30 µL. The samples were heated for one minute at 95°C, annealed for 30 seconds at 65°C and then 0.5 µL of the Klenow fragment of DNA polymerase was added to the samples and they were incubated for 30 minutes at 37°C. The samples were then purified with the Monarch PCR and DNA Cleanup kit as per manufacturer's instructions. Briefly, the samples were diluted in a 5:1 ratio of DNA Binding Buffer:sample. Then, the samples were loaded onto a column and spun through. The columns were washed twice with 700 µL DNA Wash Buffer and then transferred to clean 1.5 mL microcentrifuge tubes, and the samples were eluted in 22.5 µL Tris-HCl (10 mM, pH 7.4). To these samples, 25 µL Blunt/TA ligase master mix and 2.5 µL of the annealed custom P5 adapter (Table 4) were added and incubated for 30 minutes at 20°C. The samples were then added to 75 µL AMPure XP beads, vortexed, and incubated for 5 minutes at room temperature. Then, the beads were washed twice with 200 µL 80% ethanol, left to dry for 5 minutes, and then the samples were eluted in 20 µL Tris-HCl (10 mM, pH 7.4). The samples were then added to 12.5 µL Q5 High-Fidelity 2X Master Mix along with 1.25 µL of 10 µM forward and reverse library amplification primers (Table 4) in thin-walled PCR tubes. Then, the samples were placed in a thermocycler and the following cycle parameters were followed: 30 seconds at 98˚C; 30 cycles of 98˚C for 10 seconds, 52˚C for 30 seconds, 72˚C for 30 seconds; and 2 minutes at 72˚C followed by a hold at 4°C.

After preparing the sequencing libraries, they were quantified with he NEBNext Library Quant Kit for Illumina following the manufacturer's instructions. Briefly, the kit reagents were thawed on ice and all components were vortexed and then spun down. The NEBNext Library Quant Master Mix (with primers) was prepared by adding 100 µL to 1.5 mL NEBNext Library Quant Master Mix. The NEBNext Library Quant Dilution Buffer (10X) was prepared by diluting

it 1:10 with nuclease free water and then vortexing to mix. The amount of buffer prepared was calculated such that there was 1.2 mL of buffer per library to be amplified. The libraries and GCAT dsDNA oligonucleotide control were then diluted 1:1000 in 1X NEBNext Library Quant Dilution Buffer for a final volume of 1 mL and mixed by vortexing. Two additional dilutions of 1:10,000 and 1:100,000 were also created by performing a 1:10 serial dilution down from the 1:1000 dilution for a final volume of 100 µL per dilution. The DNA standards and diluted libraries were then prepared by combining 48 µL NEBNext Library Quant Master Mix with 12 µL DNA standard or library dilution such that each standard and library could be tested in triplicate. A no template control was also included in addition to the DNA standards by adding 12 µL NEBNext Library Quant Dilution Buffer to the 48 µL NEBNext Library Quant Master Mix in place of the DNA standards. The reactions were all mixed by pipetting up and down 5 times and then 20 µL of each standard, sample and control were loaded onto the qPCR plate in triplicate and the plate was subsequently sealed with a clear film. The plate was spun down briefly for 30 seconds at 3,000 x g to remove any bubbles formed when loading and to collect all of the samples to the bottom of the wells. The qPCR plate was then loaded into the Roche LightCycler 96 qPCR machine and the following cycling conditions were followed using the SYBR Green channel: 1 minute at 95°C followed by 35 cycles of 15 seconds at 95°C and 45 seconds at 63°C.

Statistical analysis of qPCR library validation data

The resulting qPCR data was exported, and the Roche LightCycler Software was used to generate graphs of the average Cq values across the samples. The individual Cq values for each replicate were also imported into GraphPad Prism v9.0.2 and unpaired t tests were run to

determine whether the differences in Cq values between the samples were statistically significant.

Results

Doxycycline treatment knocks down CUX1 expression

Following 4 days of doxycycline treatment, *CUX1* was confirmed to be knocked down via a western blot performed by our collaborators in the Nepveu Lab (Figure 13A). In line with this knockdown confirmation, we also observe that there is approximately a 6-fold increase in the number of AP-sites per 10,000 bp in the *CUX1* knockdown line (Figure 13B).

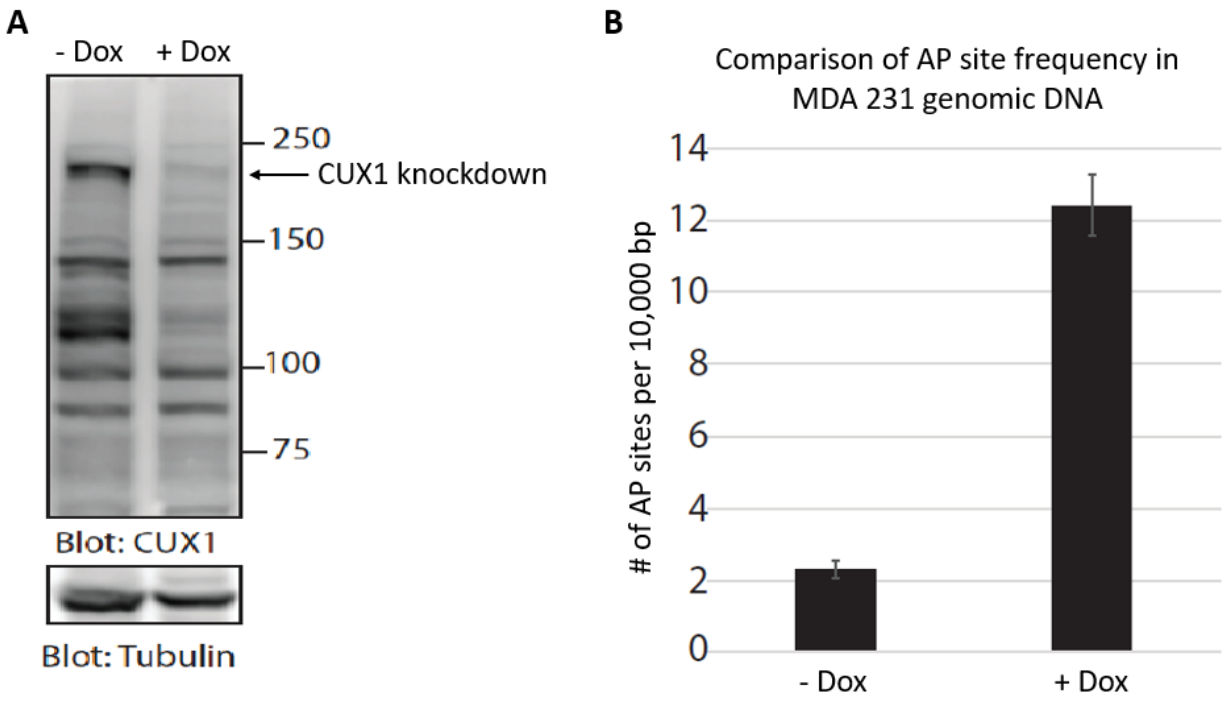


Figure 13. Confirmation of CUX1 knockdown and quantification of abasic sites. (A) With the addition of doxycycline, we see an almost 100% reduction in CUX1 protein expression, indicating a successful knockdown. (B) Extracted genomic DNA from both wild-type and *CUX1* knockdown cells was immediately treated with ARP and the quantity of AP-sites was determined by ELISA assay.

Synthetic generation of an AP-site for method validation

To validate the HIPS probe labelling method, we needed a robust synthetic model system that could be used to test the probe. However, AP-sites cannot be produced using solid phase synthesis due to their instability under synthetic conditions. As such, we used deoxyuracil DNA to artificially generate an AP-site. UNG is a glycosylase that removes deoxyuracil from DNA, leaving behind an AP-site. Since the AP-site cannot be visualized on a gel, the sample was subsequently treated with APE1 which is an enzyme that cleaves the phosphate backbone of DNA at AP-sites. We expected to see bands at 73 and 26 bp following UNG and APE1 treatment. Some cleavage was seen with the UNG alone; this is because UNG has some endonuclease activity [85]. Comparing the intensity of the bands, the addition of APE1 yields more cleavage at the AP-site which is expected (Figure 14). The cleavage is not 100% efficient as APE1 has reduced activity on ssDNA [86], however the increase in cleavage shows that the AP-site is, indeed, being produced by UNG treatment.

Validation of the HIPS probe binding to AP-sites

Mass spectrometry analysis indicated that within our sample, there was both bound and unbound AP-site containing oligonucleotide. Specifically, the exact masses corresponded to m/z 4456.8125 and m/z 4651.9766, respectively, indicating that the HIPS probe labelling reaction is not occurring at 100% efficiency. These peaks matching both the unbound oligonucleotide as well as the oligonucleotide with the HIPS probe attached at the AP-site, respectively, confirm that the condensation reaction between the HIPS probe and the open aldehyde ring of the AP-site is occurring as expected (Figure 15).

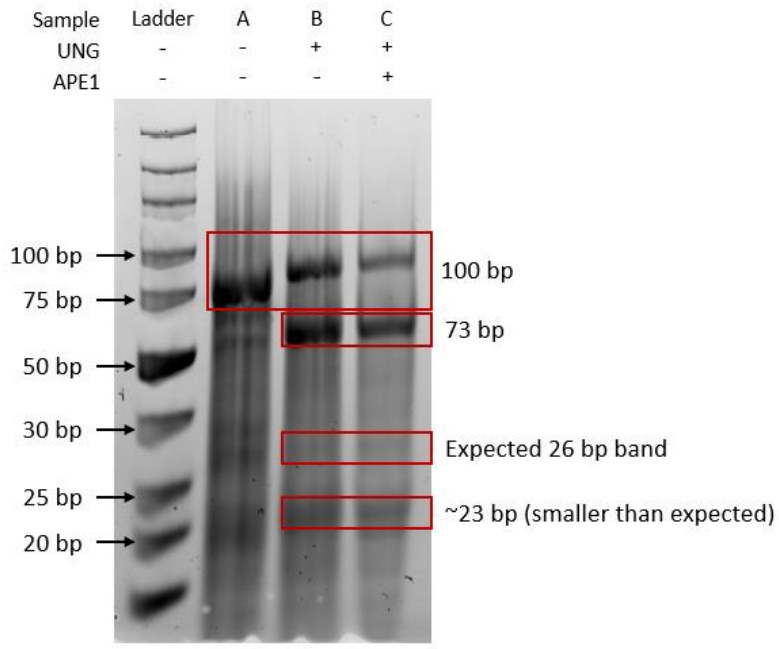


Figure 14. UNG treatment yields an abasic site at the expected location. To confirm UNG-based AP-site formation, APE1, an enzyme that cleaves AP-sites, was added to the samples and incubated for 15 minutes. UNG treatment produces some strand cleavage, but also produces the desired AP-site in high enough quantity that this protocol can be used when an AP-site must be artificially produced.

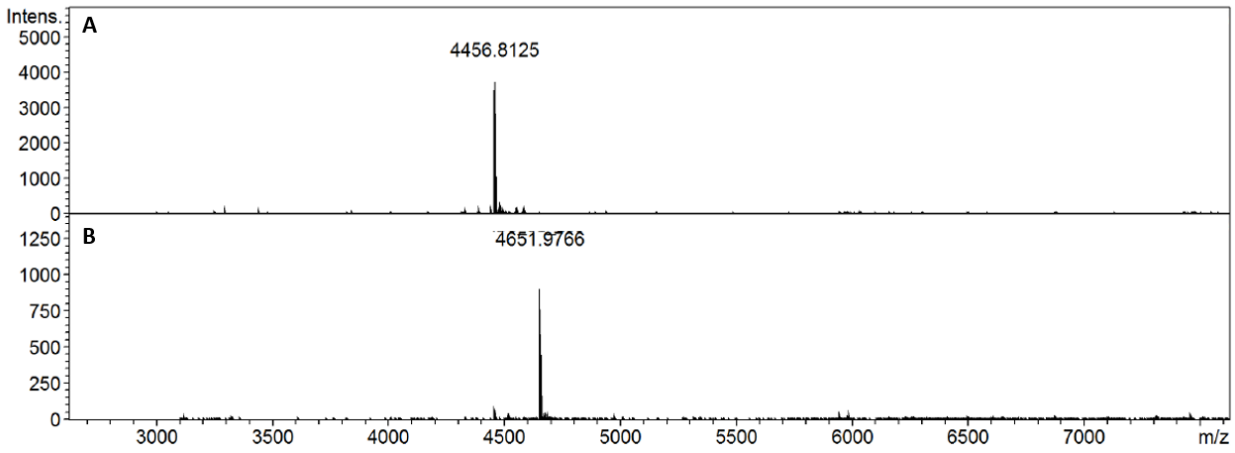


Figure 15. Mass spectrometry confirmation that the HIPS probe binds properly to abasic sites. (A) The expected mass of the ssDNA oligonucleotide containing the AP-site (m/z 4456.8125) matches the observed weight. (B) The expected mass of the AP-site containing oligonucleotide bound to the HIPS probe (m/z 4651.9766) matches the observed weight. We expected a mass change of 231.2807 m/z since this is the molecular weight of the probe, however the difference calculated from the mass spectra is 195.1641. The 18.11 mass difference observed here is attributed to the loss of water as the HIPS probe binds to the AP-site via a condensation reaction.

snAP-seq validation by qPCR amplification of sequencing libraries

Before sending samples for sequencing, we validated the snAP-seq enrichment strategy on the genomic DNA samples from the wild-type and *CUX1* knockdown MDA 231 cells. Specifically, to ensure that the HIPS probe enrichment resulted in sufficient DNA for library preparation, we performed qPCR using the library prepared primers. If the probe is working properly, only genomic DNA that was enriched with the HIPS probe should be tagged with the adaptors that contain the primers necessary for amplification.

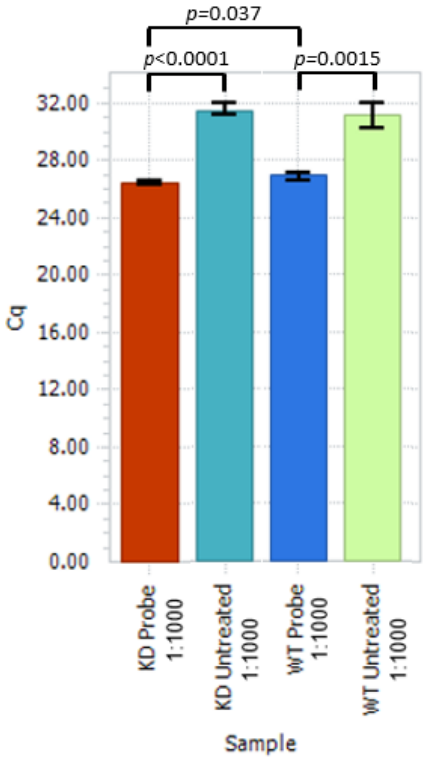


Figure 16. Cq values for qPCR amplification of CUX1 knockdown and wild-type sequencing libraries prepared using the snAP-seq method. Each sample was diluted 1:1000 before being loaded into the qPCR plate. Each bar represents the data averaged from three technical replicates and the error bars represent the standard deviation. Across the x-axis, the samples (i.e., the CUX1 knockdown (KD) and wild-type (WT) genomic DNA that were either treated with the probe or left untreated) are listed. The probe-treated and untreated samples show similar trends across the *CUX1* knockdown and wild-type samples. There were statistically significant differences in Cq values between the probe-treated and untreated samples as well as between the KD and WT samples, indicating a significantly different amount of DNA in these samples, respectively. Additional data including the other library dilutions, the standards, and the controls are available in Supplementary Figure 1.

The Cq values show a similar trend between both the probe treated and untreated CUX1 knockdown and wild-type samples (Figure 16). The GCAT oligo positive control shows that the P5 and P7 adaptor ligation steps worked as expected and the Cq values are comparable to those of the first three standards. Furthermore, the no template control shows that there was no non-specific amplification. The non-treated samples have Cq values similar to the no template control, indicating that little to no DNA was pulled down without the probe which is as expected (Supplementary Figure 1).

Summary of chapter

Previous studies suggest that transcription factors may heavily influence the distribution of DNA damage and our collaborators have shown that knocking down the transcription factor and BER accessory factor CUX1 results in an increase in AP-sites in a cell. Thus, in this chapter, we aimed to use snAP-seq, a newly published AP-site sequencing method, to determine the effects CUX1 on the distribution of AP-sites across the genome.

The results presented in Chapter 3 show that we successfully tagged AP-sites in both synthetic oligonucleotides and genomic DNA using the HIPS probe. The HIPS probe provides a more specific way to tag AP-sites as it only reacts with the aldehyde presented by the open-ring form of AP-sites and not other aldehydes present in the cell. Furthermore, we have shown that we are able to use a click reaction to biotinylate the probe such that AP-site containing DNA can be pulled out of a genomic DNA sample and enriched. As such, we were able to selectively ligate on sequencing adaptors to the AP-site containing genomic DNA to prepare a sequencing library. The larger implications of these findings will be discussed in Chapter 5.

Chapter 4: Bioinformatics Comparison of Double Strand Break Sequencing Methods

Preface

DSBs are an extremely harmful form of damage as they are not easily repaired and can easily cause genomic instability. Thus, DSBs have been studied for many years. While we understand some of the intricacies of DSB repair, very little is known about their distribution across the genome, both endogenously and in response to DNA damaging agents. Thus, many different DSB sequencing methods have been published over the last 10 years. However, despite the plethora of methods available for DSB sequencing, no comparison has been made between these methods. By comparing the trends that can be extracted from the data published by these papers, it may be possible to determine whether these methods are able to robustly reveal trends in DSB formation across the genome.

Materials

Data sets were obtained from the papers detailing DSBCapture, BLISS, sBLISS and BLESS DSB sequencing methods [58-60,77]. All four methods sequenced the DSBs using paired-end Illumina sequencing. A bioinformatics pipeline was developed by Malinda Huang to evaluate the relative frequency of DSBs across each chromosome as well as the abundance of DSBs in cancer-related genes and genes involved in the DNA damage response. The code is available at https://github.com/MalindaH/DNA-Break-Analysis.

Methods

DSBCapture sequencing data and statistical analysis

DSBCapture sequenced the DSBs in NHEK cells with no drug treatment. The sequencing data was downloaded from Gene Expression Omnibus using the accession code GSE78712 and Illumina adaptors were removed using cutadapt. Bowtie was then used to align the sequences to

the human reference genome hg19 and alignments in the blacklist regions of the human genome were removed. Poisson p-values were calculated for every 10,000 bp window on the genome to identify the important windows of high DSB abundance.

BLISS sequencing data and statistical analysis

BLISS sequenced the DSBs in U2OS cells that were treated with either DMSO (control) or etoposide, a DSB inducing chemotherapeutic. The sequencing data was downloaded from the NCBI Sequence Read Archive using the accession code SRP099132 and the reads were filtered using umi_tools to identify an 8 nucleotide unique molecular identifier embedded in the sequencing adapters. The filtered reads were then cleaned up by removing the adapter sequences with cutadapt and alignments in the blacklist regions of the human genome were removed. These reads were then aligned to the human reference genome hg38 using Bowtie and DSBs were mapped across the genome in 10,000 bp windows. The important windows of high DSB abundance were established by calculating hypergeometric p-values for both the drug-treated and untreated samples.

sBLISS sequencing data and statistical analysis

sBLISS sequenced the DSBs in TK6 cells that were treated with either DMSO (control) or etoposide, a DSB inducing chemotherapeutic. The sequencing data was downloaded from Gene Expression Omnibus using the accession code GSE145598 and aligned to the human reference genome hg38 using Bowtie. Alignments in the blacklist regions of the human genome were then removed. Hypergeometric p-values were calculated for each 10,000 bp window on the genome to identify important windows of high DSB abundance in both the drug-treated and untreated samples.

BLESS sequencing data and statistical analysis

BLESS sequenced the DSBs in HeLa cells treated with either DMSO (control) or aphidicolin, a drug that induces DSBs by inhibiting DNA polymerase. The sequencing data was downloaded from the NCBI Sequence Read Archive using the accession code SRP018506 and prepared by removing the primer sequences using cutadapt. Bowtie was then used to align the reads with the human reference genome hg19 and alignments in the blacklist regions of the human genome were removed. DSBs were mapped across each chromosome in 10,000 bp windows and significant regions of high DSB abundance were determined by calculating hypergeometric p-values for both the drug-treated and untreated samples.

Determination of cancer genes and DNA repair genes that are more sensitive to double strand breaks

A list of cancer genes identified by the Cancer Gene Census was used for these analyses and a p-value for the window from the transcription start site to the transcription end site was calculated. The $-\log_{10}$(p_value) was then plotted to rank the relative sensitivity of the cancer genes. The same analysis was repeated on a list of identified DNA repair genes which was manually compiled [87].

Results

Mapping of double strand break damage across the genome

The probability of finding a DSB at each position of each chromosome was calculated and plotted to determine whether the distribution of the DSBs is random or if there are certain hotspots for damage. If the probability of finding DSBs was truly random, we would expect the plot of probable DSB locations to resemble a normal distribution. However, we found instead

that there are specific locations in the chromosomes that show a higher probability of DSB formation (Figure 17), and this trend was universal across the genome (data not shown).

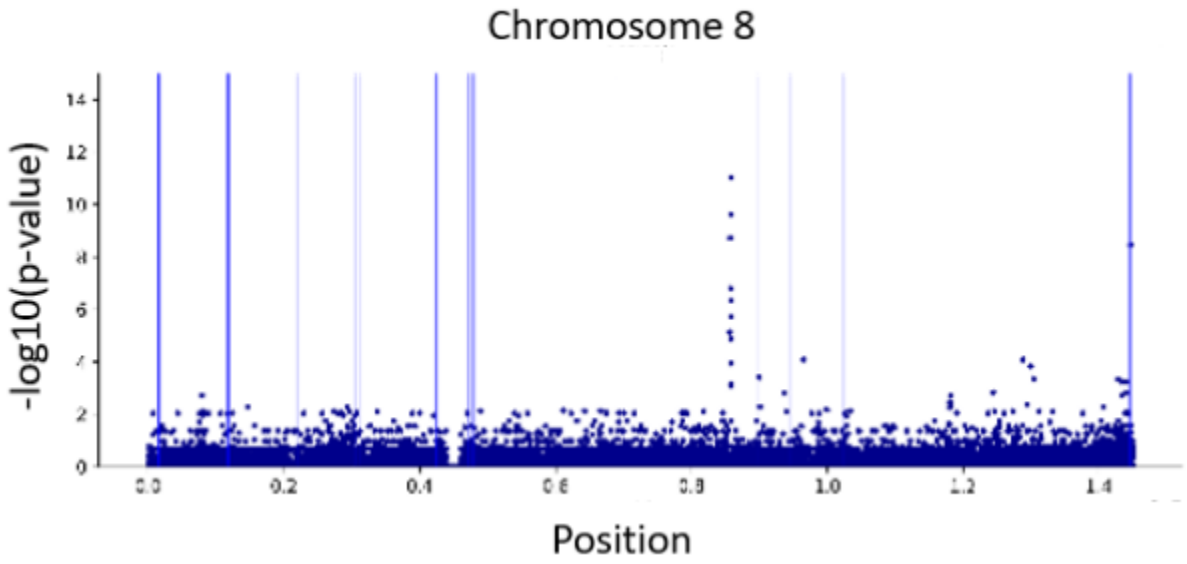


Figure 17. Map of the probability of double strand break formation across chromosome 8. The positions along the chromosome are split into 20,000 nt bins to make the data more easily legible. The blue vertical lines indicate transcription start sits as we were interested in determining whether the transcription start sites are more drug sensitive. While this exact trend was not observed for every chromosome, we did find that there were very distinct locations in every chromosome where the probability of finding a DSB was significantly higher.

Comparison of oncogene sensitivity to double strand breaks in treated and untreated cell lines

BLISS, sBLISS, BLESS, and DSBCapture applied *in vitro*. Specifically, BLISS used U2OS cells which are human osteosarcoma cells; sBLISS used TK6 cells which originate from a human B-cell lymphoma; BLESS used HeLa cells which are human cervical cancer cells; and DSBCapture used NHEK cells which are normal human epidermal keratinocytes. It should be noted that the NHEK cells are non-cancerous.

Within the treated cell populations, there are several cancer genes that show a higher sensitivity towards DSB formation in both the TK6 and HeLa cell lines (Supplementary Figure 2). In particular, above a relative break abundance of 0.2, there appears to be several similarly

sensitive cancer genes in the TK6 and HeLa populations. A similar trend is observed in the untreated controls where the same cancer genes are sensitive to DSB formation above a threshold of 0.1 (Supplementary Figure 3). Between the treated and untreated population, 14 of the 15 most sensitive genes are the same. In contrast, the U2OS cells show higher sensitivity in only one gene (Supplementary Figure 2). This observation is consistent in the untreated U2OS cells as well, where the same gene, *OLIG2*, represents the only oncogene that is highly sensitive to DSB formation in the U2OS cell population (Supplementary Figure 3).

Comparison of DNA repair gene sensitivity to double strand breaks in treated and untreated cell lines

Similar to the cancer genes, we compared the relative number of DSBs mapped within a list of DNA repair genes. We then ranked these genes based on their overall sensitivity to DSB formation using the calculated relative break abundance. Within the treated cell populations, there are several DNA damage repair genes that show the same trend of a higher sensitivity towards DSB formation in all three cell lines (Supplementary Figure 4). Overall, the TK6 cells showed a higher sensitivity towards DSB formation in more genes. A similar trend is observed in the untreated cell populations, with 18 of the top 20 most sensitive genes matching those in the treated cell populations (Supplementary Figure 5).

Summary of chapter

This chapter aimed to compare the distribution of DSBs across different cell genomes using the data produced by newly published DSB sequencing methods. While these publications have shown a robust ability to capture and sequence ICLs *in situ*, there are, to date, no comparisons between the results produced by each of these methods. As such, we compared the

data from these methods to determine if the DSB patterns were consistent across different method types.

The results presented in Chapter 4 show that the published DSB sequencing methods produce similar results in terms of the trends in DSB formation across the genome. In particular, our data comparison shows that there is a large similarity between the most sensitive cancer and DNA damage repair genes in both the drug treated and untreated cell populations. Furthermore, the results indicate that DSB formation across the genome is not random and that there are hotspots for DSB formation. Taken together, this data suggests that DSB formation patterns may indeed be indicative of cellular response to drug treatment. A further discussion of the larger implications of these findings can be found in Chapter 5.

Chapter 5: Discussion

A novel approach to understanding the distribution of interstrand crosslinks

Towards the development of an interstrand crosslink sequencing method

Sequencing ICLs is not a trivial problem as they are not easily labelled in the genome. Theoretically, antibodies could be used to pulldown DNA bound to proteins involved in the ICL-repair pathway. However, these pathways are also implicated in other cellular processes. Thus, this method would not provide the accuracy and resolution that is required to truly probe the genomic context of ICL formation. Alternatively, DSB sequencing methods such as BLESS and BLISS could be used wherein DSBs would be used as a proxy for ICL formation since DSBs are an intermediate of ICL repair. However, this again would not allow us to confidently sequence only ICLs in the genome. Our proposed method is unique in that it allows for ICLs to be specifically identified and enriched. That said, the use of rings in sequencing is not altogether new. Circle-seq, a sequencing method used for CRISPR/Cas9 off-target screening [88] employs a similar method where sheared genomic DNA is circularized and then the DNA of interest is enriched as only circular DNA containing Cas9 cleavage sites can be relinearized. Given that Circle-seq is commonly used for CRISPR/Cas9 screens, this provides a positive indication that our proposed circularization method should function to isolate ICLs specifically.

Towards the goal of developing a new method to sequence any ICLs, we used a clinically relevant model system, mapping G-Nor-G crosslinks specifically. We established a method using a restriction digest and ligation to form circular DNA. The results showed that stable ring structures could be formed (Figure 9). While the selective formation of only one ring size was not achieved, this can be mitigated in the future through the incorporation of the histone-like protein HU in the ligation process. This protein is a prokaryotic DNA-binding protein involved

in DNA replication [89] and has previously been shown to dramatically change the pitch of the DNA helix once bound [90]. As such, we hypothesize that HU may be able to bind the 250 bp fragment of DNA, pulling the ends closer together such that the increased proximity of the two EcoRI sticky ends promotes the formation of the monomeric 250 bp ring. Furthermore, HU may also help to favour the formation of monomeric rings by reducing the strain that can arise from the required bending angle. In fact, from the literature, rings of less than 250 bp have been made before, with 126 bp being approximately the smallest size that can be made without torsional or bending constraints [91,92].

The importance of the small ring size cannot be overstated as it is the basis of our planned sequencing method which would allow for the identification of the exact nucleotide that was damaged (see Figure 5). According to previous work done using mass spectrometry methods, it has been determined that anywhere from 2 to 18 G-Nor-G crosslinks per 10^6 nucleotides may be found at any time after treatment with cyclophosphamide [42,93]. Thus, smaller ring sizes are preferable as they reduce the likelihood of multiple ICLs being found in the same ring. Furthermore, the majority of NGS platforms require inputs between 200 and 400 bp for high quality reads [94]. Thus, a small ring size will not only reduce the risk of capturing multiple ICLs in one ring but will also better align with the requirements of NGS library preparation kits, making the whole genome sequencing process much more streamlined. This will also ensure that our sequencing data is high quality for downstream processing.

Establishing model systems for the in vitro sequencing of interstrand crosslinks

With regards to the application of this sequencing method *in vitro*, the end goal is to use this method to determine whether drug-sensitive and drug-resistant cells show different patterns of DNA damage. As such, we first needed to establish differentially sensitive cell lines.

Cyclophosphamide is used in many different cancer therapy regimens, however in the treatment of ALL, it remains a frontline therapy. Thus, the three cell lines chosen were all cells originating from T-lymphoblastic leukemia patients. The Genomics of Drug Sensitivity in Cancer Project provides an experimental IC_{50} for these three cell lines, namely 10 µg/mL, 15.7 µg/mL, and 24.4 µg/mL, for Jurkat, CCRF-CEM, and ALL-SIL cells respectively. However, our data reflects a much lower IC_{50}, indicating an increased sensitivity towards drug treatment. This may be because we used 4-HPCP instead of the native cyclophosphamide drug which is a prodrug. In the body, cyclophosphamide is metabolized in the liver by the enzyme CYP2B6 [95]. This produces the intermediate 4-hydroxycyclophosphamide which can then diffuse into cells where it spontaneously decomposes into the active phosphoramide mustard [95,96]. The compound 4-HPCP is an analogue of 4-hydroxycyclophosphamide, differing by only one ether group, and has been shown to decompose into the same active phosphoramide mustard [96]. Since cells of a lymphoblastic origin lack the cytochrome P450 enzymes needed to activate cyclophosphamide, it is possible that the literature values are much higher because the prodrug form of cyclophosphamide is much less potent. Our treatment of the cell lines with 4-HPCP on the other hand may reflect a higher sensitivity towards the drug because the compound is already in the form that can easily diffuse into cells and produce the active crosslinking mustard which induces apoptosis.

While our results reflect the literature in that the three cell lines show a differential sensitivity to 4-HPCP treatment, it should be noted that our results indicate CCRF-CEM is more sensitive than Jurkat cells while the literature showed the opposite trend. Again, this may be due to the fact that we used the active 4-HPCP compound as opposed to the cyclophosphamide prodrug. However, the observed drug sensitivity may also be tied to certain characteristic

mutations of these cell lines. In particular, both cell lines contain loss of function mutations in genes involved in the mismatch repair (MMR) pathway, namely *MSH2 and MSH6* in Jurkat cells and *MLH1* and *PMS2* in CCRF-CEM cells. While we are primarily interested in the interstrand crosslinks formed by cyclophosphamide, it is also true that cyclophosphamide can form monoadducts which may recruit MMR proteins during the repair process [97,98]. Within our two cell lines, the variant allele frequency indicates that there is a higher rate of mutation among Jurkat cells (Table 5).

Table 5. Variant allele frequency of loss of function mutations in acute lymphoblastic leukemia cell lines.

Cell Line	**Gene**	**Variant Allele Frequency (27,28)**
Jurkat	*MSH2*	98%
	MSH6	96%
CCRF-CEM	*MLH1*	54%
	PMS2	44%

Loss of function mutations in MMR-associated genes have been associated with an increased resistance to monofunctional alkylating agents such as cisplatin [99-101]. Specifically, MMR-deficient cells are able to accumulate DNA damage without triggering cell death [102,103]. While cyclophosphamide is classified as a bifunctional alkylating agent, it is capable of forming similar adducts. Thus, it is not unreasonable to hypothesize that increased deficiency in MMR-pathway proteins may lead to increased 4-HPCP resistance. As such, it would make sense for us to observe a higher resistance to 4-HPCP in Jurkat cells since they have a higher frequency of mutated MMR genes (Table 5). However, it is important to remember that the cell

is a complex interconnected system. Therefore, it is likely that at least some of the other mutations in these cell lines are also of importance. As such, it should be noted that the Jurkat cell line contains approximately 3 times as many mutations and copy number alterations in oncogenes and tumor suppressor genes as compared to the CCRF-CEM cell line [104,105]. These additional mutations may also confer some additional resistance to 4-HPCP treatment.

Now that a working range of 4-HPCP concentrations for the three ALL cell lines has been established, the presence of G-Nor-G crosslinks will be established and quantified to confirm the ability of 4-HPCP to create the crosslinks that we are aiming to detect with our sequencing method. To do so, we can use AP-sites as a proxy for G-Nor-G ICLs. The guanines involved in the G-Nor-G are not very stable and are readily depurinated through heating [91,106], leaving AP-sites which can be readily detected and quantified by commercial kits. Thus, we would be able to determine the extent to which 4-HPCP damages the cellular DNA. However, since G-Nor-G crosslinks are fairly rare, occurring in the range of 2 to 18 crosslinks per 10^6 bp, it is possible that the number of ICLs will be outside the range of detection of these kits [106]. In this case, the mass spectrometry method established by Malayappan *et al.* [42] could be used as it has already been used to successfully quantify G-Nor-G crosslinks *in vitro* with a sensitivity in the desired range.

Having established a protocol for circular DNA formation and a working concentration of drug for each of our three cell lines, we will continue to optimize the steps of our method such that whole genome sequencing may be performed. As the ultimate goal of this project is to provide a biomarker-based tool to predict the best course of treatment for individual ALL patients, future work will also involve the replication of these experiments with clinical samples to determine whether trends observed *in vitro* are generalizable to a patient population.

The impact of CUX1 on AP-site formation

CUX1 knockdown confirmed to increase AP-site frequency

CUX1 is a protein containing a DNA binding CUT domain that stimulates the enzymatic activities of the 8-oxoguanine DNA glycosylase OGG1 [107]. As such, it functions as an accessory factor in the BER pathway. Previous work by our collaborators has confirmed that knocking down *CUX1* results in an increase in AP-sites in cell lines undergoing stress (i.e., experiencing oxidative damage or following treatment with mono-alkylating agents). This was confirmed using an AP-site ELISA assay which shows a significant increase in the number of AP-sites in our model cell line following doxycycline-induced knockdown of *CUX1*.

HIPS probe binds to AP-sites

The HIPS probe was designed to react specifically with the open-ring reactive aldehyde form of AP-sites and not other aldehydes in the cells. Based on our mass spectrometry data, we observe that the probe reacts with the synthetic AP-site. The reaction was not 100% efficient as, despite treating the DNA with a molar excess of the probe, we still observe some unbound oligonucleotide in our sample. This may be because, at equilibrium, only about 1% of AP-sites reside in the reactive open-ring aldehyde form, and the other 99% remains in the closed-ring furanose form [79]. In this case, the probe would not be able to react with the closed-ring furanose form. If this is the case, many libraries will have to be prepared from each sample to ensure that a robust coverage of all locations in the genome containing AP-sites are covered.

Liu et al. demonstrated that the HIPS probe binds specifically to AP-sites and not other aldehyde-containing species, but we will perform similar studies to ensure the robust synthesis and purification of the probe.

qPCR validation of the snAP-seq method shows probe pulldown of DNA containing AP-sites

Before sending the libraries for sequencing, we must validate that the libraries were properly enriched for AP-sites and prepared with the appropriate sequencing adaptors. Using qPCR, we confirmed that the HIPS probe was successful in tagging AP-sites in the genomic DNA and that this tagged DNA was able to be pulled down out of the samples and enriched using the snAP-seq method. In addition, there was significantly more DNA present in the libraries originating from probe-treated samples, which further confirms the high sensitivity of the HIPS probe for AP-sites. Surprisingly, the amount of DNA in both the *CUX1* knockdown and wild-type samples seemed comparable, though the p-value generated by an unpaired t test suggests that the small difference in average Cq between these samples is still statistically significant. Based on previous work by our collaborators, we expected to see a more dramatic difference between the number of AP-sites in the *CUX1* knockdown cells and the wild-type cells, because they have shown, using an ELISA-based test, that there are quantitatively more AP-sites in this knockdown model. However, it should be noted that this ELISA-based method uses ARP which, as previously mentioned, does not react solely with AP-sites. Thus, it may be that *CUX1* knockdown does not actually impact AP-site formation, but rather prompts the formation of more reactive aldehyde species which can react with ARP such that the total apparent number of AP-sites detected is artificially inflated.

snAP-seq presents a novel method that will allow us to specifically distinguish between AP-sites and other aldehyde species in the genome for the first time. As such, future work will involve sequencing these libraries to determine whether the increase in AP-sites that was previously noted is actually attributable to AP-site formation, or if *CUX1* is involved in the

suppression of aldehyde species production, such that a knockdown of *CUX1* artificially suggests an increase in AP-sites.

Double strand break sensitivity across the genome

Distribution of genomic locations sensitive to double strand break formation

Having calculated and plotted the probability of DSB formation across the entire genome, it is clear that the distribution of DSB-sensitive locations is not random. While the location of the DSB-sensitive positions is not the same for each chromosome, there was a universal trend where the average probability of DSB formation was $-\log_{10}(2)$ or lower, except for in very specific locations where the probability increased dramatically. We also plotted the transcription start sites in each genome as we were curious whether this might be a genomic feature that may be more sensitive to DSB formation. However, based on our data, it does not appear that there is any strong relationship between transcription start site and a higher probability for DSB formation. This is in contrast to other forms of DNA damage, such as 8-oxoguanine, where there is a distinct drop in 8-oxoguanine abundance at transcription start sites [61].

Sensitivity of cancer genes to double strand break formation following drug treatment

When we compared the relative abundance of DSBs across the human genome using the BLISS, sBLISS, BLESS, and DSBCapture methods, 10 cancer genes consistently emerged. These 10 genes which are *OLIG2, CXCR4, RPL5, TAL2, PCBP1, HOXD13, MYOD1, TLX3, DUX4L1,* and *ID3I* largely fall into the two categories of developmentally regulated transcription factors and genes involved in increasing gene activity [108,115,129-134]. Broadly, this aligns with the main characteristic of cancer which is the uncontrolled proliferation of cells.

Interestingly, only *OLIG2*, a basic helix-loop-helix transcription factor which is expressed in the CNS during embryonic development [108], showed major sensitivity in the

U2OS cells which were treated with etoposide. It is expected to see a greater similarity between the U2OS and TK6 cells as they were treated identically, and the sequencing methods and data analysis protocols were largely the same, yet across the entire panel of cancer genes, only *OLIG2* showed the same sensitivity to DSBs in both U2OS and TK6 cells. One possible explanation is that the U2OS cells are osteosarcoma cells whereas the TK6 cells are lymphoblasts. As such, only the TK6 cells are derived from a cancer that is typically treated frontline with etoposide [109,110]. Thus, it is possible that the U2OS cells have a naturally lower response to etoposide as is reflected by the lower number of DSB sensitive cancer genes.

HIST1H3B and *PCBP1* are two other genes with an increased sensitivity towards DSB formation. These two genes are involved in controlling gene activity, with *HIST1H3B* encoding the canonical histone H3 protein [111,112] and *PCBP1* encoding for a multifunctional RNA-binding protein involved in post-transcriptional gene regulation. With regard to *HIST1H3B*, the histone H3 protein for which it encodes is important in the epigenetic regulation of gene expression [113]. In particular, K27 and G34 are two amino acids located in the N-terminal tail f the histone H3 protein and they are critical sites for methylation which indicates active gene regions [111]. Given its role in controlling gene expression, it is perhaps unsurprising that mutations in histone H3 have been found in pediatric high-grade gliomas and subsequent tumors in adolescents [114]. Therefore, it may be possible that generally, in drug sensitive cells, *HIST1H3B* has a higher relative sensitivity to DSB formation in response to chemotherapeutic treatment since this gene has a general involvement in increasing gene expression across the genome. Similarly, *PCBP1* is involved in regulating the alternative splicing, translation, and RNA stability of many cancer-related genes [115], making it a logical target for chemotherapy-induced DNA damage and, indeed, we do observe an increased relative abundance of DSBs.

Overall, we see similar genes exhibiting increased sensitivity to DSB formation in all cell types and drug treatments. This indicates that DSB formation in the genome is not random and is dictated by the existence of higher sensitivity regions in specific gene loci.

Sensitivity of cancer genes to double strand break formation in untreated cell lines

In the untreated cell populations, we see a similar trend with nine of the top 10 most sensitive genes overlapping with the drug-treated cell populations. However, the change in the order of the most sensitive genes may be due to the addition of the NHEK cell data from the DSB-capture sequencing. One possible reason behind this similarity between the treated and untreated cell populations may be because we are looking at cells treated with chemotherapeutics without an established IC_{50}. If we were to repeat these experiments with cell lines and the chemotherapeutics that would be clinically used to treat those cancer types, it is possible that we would see a larger difference between the treated and untreated cells. Furthermore, I think additional patterns in DSB distribution could be revealed by comparing the sequencing results for drug-sensitive and drug-resistant cell lines. In addition, these results may be more dramatic by using a drug-resistant cell line developed from a drug-sensitive cell line. As such, we would be able to observe the differences in drug sensitivity in cells with the exact same genetic makeup.

Sensitivity of genes involved in the DNA damage response to double strand break formation following drug treatment

When we compared the relative abundance of DSBs across the human genome, eight DNA damage repair genes consistently emerged in both the treated and untreated cell populations, namely *RPA4, RECQL4, H2AX, SPO11, CETN2, UBE2T, MPLKIP,* and *NEIL1*. Compared to the cancer genes, the DNA repair genes are not as easily divided into two categories. *CETN2, RPA4, NEIL1, and UBE2T* are involved in the nucleotide excision repair

(NER), BER, and Fanconi anemia (FA) pathways, respectively [87]. *MPLKIP* and *RECQL4* are genes that are commonly defective in diseases associated with sensitivity to DNA damaging agents [87]. These two genes are thought to be involved in cell cycle regulation, specifically mitosis and cytokinesis, and chromosome segregation, respectively [116,117]. Finally, *H2AX* is a gene involved with chromatin structure and modification [87], and *SPO11* is an endonuclease involved in the meiotic recombination pathway [87,118].

The data we analyzed used drugs that mainly produce DSBs; these include etoposide, a topoisomerase II inhibitor [119], and aphidicolin, a DNA polymerase alpha inhibitor [120]. As such, it is not surprising that genes involved in the FA pathways are more sensitive to DSB formation since these are the repair pathways most heavily implicated in DSB repair. As such, future work will examine cellular resistance from drug treatment correlates with these genes being differentially damaged or expressed.

In the literature, it has also been found that etoposide-related DNA damage also recruits BER proteins for repair [121]. However, in our data, the U2OS and TK6 cells which were both treated with etoposide show that *RPA4* and *NEIL1* are much more sensitive to DSB formation in the U2OS cells, than in the TK6 cells. As mentioned previously, the U2OS cells are osteosarcoma cells whereas the TK6 cells are lymphoblasts. As such, only the TK6 cells are derived from a cancer that is typically treated frontline with etoposide [113,114]. Thus, in this case, it is possible that *RPA4* and *NEIL1* in the TK6 cells have already adopted resistance to DSB formation as compared to the same genes in the U2OS cell line.

When comparing *MPLKIP* and *RECQL4,* both genes are associated with Rothmund-Thompson syndrome and a non-photosensitive form of trichothiodystrophy, respectively [87]. These diseases are characterized by an increased risk of DNA damage and, therefore, and

increased sensitivity to DNA damaging agents. Because these genes are significantly impacted by damage, it is possible that they alone would serve as good biomarkers for chemotherapeutic outcome. Future work will examine the damage patterns and expression of these genes compared to the sensitivity of cancers to treatment. We will test whether there is a change in the relative abundance of DSBs in these two genes in drug-sensitive and drug-resistant matched control cell lines.

It is also interesting that *H2AX*, a gene involved in chromatin structure and modification, is highly sensitive to DSB formation, and is the third most sensitive gene in our analysis (Figure 16). The repair of DSBs by non-homologous end joining and homologous recombination is initiated by the phosphorylation of Ser-139 on the minor histone H2A variant H2AX, leading to the production of γH2AX [122]. Thus, perhaps the increased sensitivity of *H2AX* to DSBs has evolved as a cancer treatment resistance mechanism to prevent the repair of the DSBs which are meant to induce apoptosis in cancerous cells. In line with this, studies have also shown that DNA methyltransferase I (DNMT1) often colocalizes with γH2AX to maintain epigenetic markers, such as methylation, in DNA that is newly synthesized through the repair process. However, changes in gene expression have been shown to arise from DNMT1 recruitment, suggesting that DNA accessibility can be altered following a DSB event [123-125]. More specifically, hypermethylation catalyzed by DNMT1 can restrict access to important damage sensitive hotspots, leading to chemotherapeutic resistance. In fact, several studies have shown that resistance to DNA adduct forming chemotherapeutics can be linked to DNA hypermethylation [126-128]. Thus, perhaps *H2AX* has such a high sensitivity to DSB formation because its role in possible drug resistance is two-fold. In other words, it is possible that the cell's safety mechanism to ensure apoptosis in response to drug treatment is to damage the gene implicated in

multiple facets of the repair process such that rapidly-dividing cancer cells lose two methods by which they can avoid cell death and develop resistance.

Limitations of double strand break sequencing methods and subsequent data analysis

Our results indicate that DSB sequencing methods produce data that are largely comparable despite differences in methodology. Nevertheless, there are still several limitations in our analysis and when comparing the datasets. Firstly, while the data from these publications enabled the first comparison of damage sensitive genes, it is difficult to draw definitive conclusions regarding the distribution of damage across the genome due to the lack of high coverage data and replicates. In particular, DSBs are relatively infrequent in the genome since they are an extremely damaging form of damage and cells have highly evolved pathways that repair them in a timely manner. As such, a much larger number of cells and sequencing runs may be needed to collect enough data to have a high confidence in the obtained results. For example, the analysis of the distribution of DSBs across the chromosomes could only be done confidently with DSBCapture which had an average of 80 million reads per sample as compared to the 6 million reads per sample obtained using BLESS. Thus, while these methods represent a big step forward in mapping damage in the genome, there is still much work to be done to produce data sets that result in analyses with significance. Furthermore, "relative abundance" may not be the best way to compare data as we lose information as to whether drug treatment actually produces a different number of DSBs. Thus, future work will make use of "spike in" control sequences to obtain an "absolute break abundance" such that our data can be compared in a more meaningful way. Additionally, all of the cell lines used, except for the NHEK cells, are cancer cells. Thus, while there are clearly some trends that encompass damage in general, we are unable to determine whether these trends extend to healthy cells as well. The similarity in DSB sensitivity

of the NHEK cell genes in the untreated cell populations in DSBCapture seem to indicate that a similar trend may be observed, but more healthy cells, including patient samples, should be tested to determine whether this is a robust effect.

While our analysis indicates that all of the different DSB sequencing methods are valid strategies to obtain reliable maps of DSBs in the genome, we need more careful comparisons of the subtle differences between each cell line to be able to determine whether DNA damage patterns may be useful as a biomarker for treatment response. This is a difficult question to answer, especially as each method uses a different dose of drug, treatment time, treatment conditions, etc. Thus, to properly assess the importance of DNA damage frequency and distribution as a biomarker of drug sensitivity or efficacy, experiments with identical conditions should be conducted.

Limitations of the methodologies

The goal of this thesis was to pursue several different methodologies to validate and develop methods for DNA damage sequencing, working towards elucidating the relationships between DNA damage frequency, distribution, and response to drug treatment. While we were able to make strides towards this goal in three different ways, there are some overarching limitations. Firstly, while the sensitivity of all of these methods is relatively high, there are still issues in terms of capturing enough data to make significant conclusions and this is largely driven by the fact that these types of DNA damage are simply not abundant in the genome. Thus, current methods are limited by scale and future work will need to include ways to increase the scalability of these methods.

Another major issue is that all of the sequencing methods discussed require the treatment of each cell line with artificially high concentrations of drug. While researchers often use

clinically relevant doses of drug, these doses do not always produce high enough quantities of damage for sequencing. As of now, mass spectrometry is the only method that can reliably detect "real life" quantities of damage. As such, strategies such as microdosing patients, taking cell samples from them, and then sequencing the genomic DNA extracted from these cells will not be amenable to sequencing due to the low quantity of both the damaged DNA and number of cells. As an alternative approach, patient cells could be harvested and grafted into animal models. Once significant engraftment is achieved, the animal could be treated at a relevant dose. Following this, a much larger number of cells could be extracted. Another alternative would be to extract patient cells and treat them *in vitro* with high doses of drug. Each of these approaches does not perfectly mimic the treatment but may help to answer important questions regarding differential DNA damage patterns in a specific patient. For examples, if the pattern of DNA damage is reproducible at high doses of drug, then we will still be able to use these damage patterns as biomarkers for treatment response.

While much work still remains to be done to improve on the limitations listed above, the development and use of these novel damage sequencing methods represents an exciting new frontier of possibilities when it comes to understanding the response of the genome to assault by exogenous stressors. As these methods continue to be refined, they will allow us to answer many previously unanswerable questions regarding the exact effect of DNA damaging agents on the genome.

Chapter 6: Conclusion and Future Directions

Towards the goal of improving patient treatment by being able to predict a positive response to chemotherapy, the distribution of DNA damage across the genome needs to be better understood. The goal of my thesis was to provide more insight into the interplay between DNA damage distribution and treatment response from three different perspectives. First, I sought to develop a method to sequence a form of DNA damage, ICLs, that are heretofore unable to be sequenced. Second, I validated an AP-site sequencing method and worked towards applying it to determine the effect of a DDR accessory protein on damage distribution in the genome. Finally, I compared the datasets produced by currently published DSB sequencing methods to determine whether there are some patterns of damage that can be revealed by these methods.

Towards my first aim, I successfully laid the groundwork for a novel ICL sequencing method. Future work can build on our ring development strategy to specifically pull down and sequence ICLs. Towards aim two, I validated the snAP-seq labelling procedure using breast cancer cells and synthetic oligonucleotides. In the future, our lab will apply this method directly to our genomic DNA samples from our cell line models to ascertain whether *CUX1* knockdown will influence the distribution of AP-sites in the genome. Finally, in aim three, I compared published DSB sequencing methods, confirming that, despite procedural differences, each unique method is reliable. Furthermore, by comparing data generated by these methods against gene sets representing cancer genes and DNA damage repair genes, I have revealed a general pattern that the distribution of damage is not random and certain genes are consistently more sensitive to DSB formation than others.

This thesis generated a better understanding of the interplay between DNA damage distribution, DNA damage quantity, and patient response to chemotherapy, though there is still

much work to be done. For example, to determine whether there is overlap in genomic regions sensitive to specific types of damage, the DNA damage distribution profiles of different types of damage in the same cell line will be compared. Furthermore, it would be most clinically relevant to understand how different treatment regimens may result in different patterns of DNA damage. The difference in DNA damage patterns is particularly important given that cancer is often treated frontline with a cocktail of chemotherapies, each with a different mechanism of action. Thus, we must continue to push forwards with the development and usage of high-resolution sequencing methods that allow us to map all types of damage that may be generated by chemotherapeutic treatment.

Despite the negative connotations associated with the word “damage”, it is a tool that is clinically relevant today and will continue to be used in the treatment of cancer. However, many of our current frontline chemotherapies non-specifically damage all genomic DNA. Thus, it is imperative that we develop novel damage sequencing methods and use current damage sequencing methods in tandem to understand how the distribution of DNA damage in the genome may impact patient response to treatment. With studies already showing that there is some correlation between the quantity of DNA damage and patient response, we think that the missing piece for being able to truly predict patient response is the distribution of DNA damage in the genome. By developing DNA damage maps and better understanding genetic hotspots for damage, we move one step closer to being able to provide a personalized treatment plan for every patient, thereby improving their quality of life for the duration of treatment for a deadly disease.

References

1. World Health Organization. (n.d.). *Cancer*. https://www.who.int/health-topics/cancer.
2. Sahora, A., & Khanna, C. (2013). Cellular Growth/Neoplasia. *Canine and Feline Gastroenterology*, 61-69.
3. Weinberg, R. A. (1994). Oncogenes and tumor suppressor genes. *CA: a cancer journal for clinicians*, *44*(3), 160-170.
4. National Cancer Institute. (n.d.). Oncogene. Retrieved May 24, 2022 from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/oncogene.
5. Chial, H. (2008). Proto-oncogenes to oncogenes to cancer. *Nature education*, *1*(1), 33.
6. Seshacharyulu, P., Ponnusamy, M. P., Haridas, D., Jain, M., Ganti, A. K., & Batra, S. K. (2012). Targeting the EGFR signaling pathway in cancer therapy. *Expert opinion on therapeutic targets*, *16*(1), 15-31.
7. Cooper, G. M., Hausman, R. E., & Hausman, R. E. (2007). Tumor Suppressor Genes. In *The cell: a molecular approach* (Vol. 4, pp. 649-656). ASM press.
8. Noonan, K. L., Ho, C., Laskin, J., & Murray, N. (2015). The influence of the evolution of first-line chemotherapy on steadily improving survival in advanced non-small-cell lung cancer clinical trials. *Journal of Thoracic Oncology*, *10*(11), 1523-1531..
9. Baldo, B. A., & Pham, N. H. (2013). Adverse reactions to targeted and non-targeted chemotherapeutic drugs with emphasis on hypersensitivity responses and the invasive metastatic switch. *Cancer and Metastasis Reviews*, *32*(3), 723-761.
10. Rheingold, S. R., Neugut, A. I., Meadows, A. T. (2003). Therapy-Related Secondary Cancers. In *Holland-Frei Cancer Medicine*. 6th Edition. BC Decker.
11. Iwamoto, T., Hiraku, Y., Oikawa, S., Mizutani, H., Kojima, M., & Kawanishi, S. (2004). DNA intrastrand cross-link at the 5′-GA-3′ sequence formed by busulfan and its role in the cytotoxic effect. *Cancer science*, *95*(5), 454-458.
12. Colvin, M. (2003). Alkylating Agents. In *Holland-Frei Cancer Medicine*. 6th Edition. BC Decker.
13. Chen, Y., Jia, Y., Song, W., & Zhang, L. (2018). Therapeutic potential of nitrogen mustard based hybrid molecules. *Frontiers in pharmacology*, *9*, 1453.
14. Tayyab Imtiaz, M., Anwar, F., Saleem, U., Ahmad, B., Hira, S., Mehmood, Y., ... & Ismail, T. (2021). Triazine Derivative as Putative Candidate for the Reduction of Hormone-Positive Breast Tumor: In Silico, Pharmacological, and Toxicological Approach. *Frontiers in Pharmacology*, 1267.
15. Siddik, Z. H. (2002). Mechanisms of action of cancer chemotherapeutic agents: DNA-interactive alkylating agents and antitumour platinum-based drugs. *The cancer handbook*, *1*.

16. Pyrimidine Analogues. (2017). In *LiverTox: Clinical and Research Information on Drug-Induced Liver Injury*. National Institute of Diabetes and Digestive and Kidney Diseases.

17. Galmarini, C. M., Jordheim, L., & Dumontet, C. (2003). Pyrimidine nucleoside analogs in cancer treatment. *Expert review of anticancer therapy*, *3*(5), 717-728.

18. Purine Analogues. (2014). In *LiverTox: Clinical and Research Information on Drug-Induced Liver Injury*. National Institute of Diabetes and Digestive and Kidney Diseases.

19. Pettitt, A. R. (2003). Mechanism of action of purine analogues in chronic lymphocytic leukaemia. *British journal of haematology*, *121*(5), 692-702.

20. Bertino, J. R. (2009). Cancer research: from folate antagonism to molecular targets. *Best Practice & Research Clinical Haematology*, *22*(4), 577-582.

21. Robien, K., Boynton, A., & Ulrich, C. M. (2005). Pharmacogenetics of folate-related drug targets in cancer treatment.

22. Anguita, E., Valverde, F., Gonzalez, F. A., Gil, C., Mateo, M., Ferro, M. T., & Villegas, A. (1998). The first report of a Philadelphia chromosome and BCR/ABL rearrangement positive myeloproliferative disorder in a child with thrombocythemia. *Leukemia*, *12*(3), 442-444.

23. Heisterkamp, N., & Groffen, J. (1991). Molecular insights into the Philadelphia translocation. *Hematologic pathology*, *5*(1), 1-10.

24. Snodgrass, R., Nguyen, L. T., Guo, M., Vaska, M., Naugler, C., & Rashid-Kolvear, F. (2018). Incidence of acute lymphocytic leukemia in Calgary, Alberta, Canada: a retrospective cohort study. BMC research notes, 11(1), 104.

25. Terwilliger, T., & Abdul-Hay, M. J. B. C. J. (2017). Acute lymphoblastic leukemia: a comprehensive review and 2017 update. Blood cancer journal, 7(6), e577-e577.

26. Cortes, J. E., & Kantarjian, H. M. (1995). Acute lymphoblastic leukemia a comprehensive review with emphasis on biology and therapy. Cancer, 76(12), 2393-2417.

27. Gaynon, P. S., Desai, A. A., Bostrom, B. C., Hutchinson, R. J., Lange, B. J., Nachman, J. B., ... & Tubergen, D. G. (1997). Early response to therapy and outcome in childhood acute lymphoblastic leukemia: a review. Cancer: Interdisciplinary International Journal of the American Cancer Society, 80(9), 1717-1726.

28. Davis, A. S., Viera, A. J., & Mead, M. D. (2014). Leukemia: An overview for primary care. American family physician, 89(9), 731-738.

29. Panahi, Y., Fattahi, A., Nejabati, H. R., Abroon, S., Latifi, Z., Akbarzadeh, A., & Ghasemnejad, T. (2018). DNA repair mechanisms in response to genotoxicity of warfare agent sulfur mustard. Environmental toxicology and pharmacology, 58, 230-236.

30. Fleming, R. A. (1997). An overview of cyclophosphamide and ifosfamide pharmacology. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 17(5P2), 146S-154S.

31. Roy, U., & Schärer, O. D. (2016). Involvement of translesion synthesis DNA polymerases in DNA interstrand crosslink repair. DNA repair, 44, 33-41.

32. Moudi, M., Go, R., Yien, C. Y. S., & Nazre, M. (2013). Vinca alkaloids. International journal of preventive medicine, 4(11), 1231.

33. Gewirtz, D. (1999). A critical evaluation of the mechanisms of action proposed for the antitumor effects of the anthracycline antibiotics adriamycin and daunorubicin. Biochemical pharmacology, 57(7), 727-741.

34. Lossignol, D. (2016). A little help from steroids in oncology. *Journal of translational internal medicine*, *4*(1), 52-54.

35. Pui, C. H., Mullighan, C. G., Evans, W. E., & Relling, M. V. (2012). Pediatric acute lymphoblastic leukemia: where are we going and how do we get there?. *Blood, The Journal of the American Society of Hematology*, *120*(6), 1165-1174.

36. Kamel, H. F. M., & Al-Amodi, H. S. A. B. (2017). Exploitation of gene expression and cancer biomarkers in paving the path to era of personalized medicine. *Genomics, proteomics & bioinformatics*, *15*(4), 220-235.

37. Park, J. W., Kerbel, R. S., Kelloff, G. J., Barrett, J. C., Chabner, B. A., Parkinson, D. R., ... & Slamon, D. J. (2004). Rationale for biomarkers and surrogate end points in mechanism-driven oncology drug development. Clinical Cancer Research, 10(11), 3885-3896.

38. Kelloff, G. J., & Sigman, C. C. (2012). Cancer biomarkers: selecting the right drug for the right patient. Nature reviews Drug discovery, 11(3), 201-214.

39. Van't Veer, L. J., & Bernards, R. (2008). Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature, 452(7187), 564-570.

40. Stornetta, A., Zimmermann, M., Cimino, G. D., Henderson, P. T., & Sturla, S. J. (2017). DNA adducts from anticancer drugs as candidate predictive markers for precision medicine. *Chemical research in toxicology*, *30*(1), 388-409.

41. Deans, A. J., & West, S. C. (2011). DNA interstrand crosslink repair and cancer. *Nature reviews cancer*, *11*(7), 467-480.

42. Malayappan, B., Johnson, L. A., Nie, B., Panchal, D., Matter, B., Jacobson, P., & Tretyakova, N. (2010). Quantitative High-Performance Liquid Chromatography− Electrospray Ionization Tandem Mass Spectrometry Analysis of Bis-N 7-Guanine DNA− DNA Cross-Links in White Blood Cells of Cancer Patients Receiving Cyclophosphamide Therapy. Analytical chemistry, 82(9), 3650-3658.

43. Johnson, L. A. A., Malayappan, B., Tretyakova, N., Campbell, C., MacMillan, M. L., Wagner, J. E., & Jacobson, P. A. (2012). Formation of cyclophosphamide specific DNA adducts in hematological diseases. *Pediatric blood & cancer*, *58*(5), 708-714.

44. Hengstler, J. G., Fuchs, J., and Oesch, F. (1992) DNA strand breaks and DNA cross-links in peripheral mononuclear blood cells of ovarian cancer patients during chemotherapy with cyclophosphamide/carboplatin Cancer Res. 52, 5622– 5626

45. Pan, C. X. (2010, December – 2016, November). *Clinical Study of Microdosing Carboplatin in Lung and Bladder Cancer*. Identifier NCT01261299. https://clinicaltrials.gov/ct2/show/NCT01261299.

46. Kim, E. (2015, October 7 – 2021, March 30). *Oxaliplatin Microdosing Assay in Predicting Exposure and Sensitivity to Oxaliplatin-Based Chemotherapy*. Identifier NCT02569723. https://clinicaltrials.gov/ct2/show/NCT02569723.

47. Zimmermann, M., Wang, S. S., Zhang, H., Lin, T. Y., Malfatti, M., Haack, K., ... & Henderson, P. T. (2017). Microdose-induced drug–DNA adducts as biomarkers of chemotherapy resistance in humans and mice. *Molecular cancer therapeutics*, *16*(2), 376-387.

48. Jackson, S. P., & Bartek, J. (2009). The DNA-damage response in human biology and disease. *Nature*, *461*(7267), 1071-1078.

49. Vadnais, C., S. Davoudi, M. Afshin, R. Harada, R. Dudley, P.-L. Clermont, E. Drobetsky and Alain Nepveu. CUX1 Transcription Factor is Required for Optimal ATM/ATR-Mediated Responses to DNA Damage. Nucleic Acids Research, 40(10): 4483-4495. 2012.

50. Samarakkody, A. S., Shin, N. Y., & Cantor, A. B. (2020). Role of RUNX family transcription factors in DNA damage response. *Molecules and cells*, *43*(2), 99.

51. Frontini, M., Vijayakumar, M., Garvin, A., & Clarke, N. (2009). A ChIP–chip approach reveals a novel role for transcription factor IRF1 in the DNA damage response. *Nucleic acids research*, *37*(4), 1073-1085.

52. Bhoumik, A., Lopez-Bergami, P., & Ronai, Z. E. (2007). ATF2 on the double–activating transcription factor and DNA damage response protein. *Pigment Cell Research*, *20*(6), 498-506.

53. Giuliano, S., Cheli, Y., Ohanna, M., Bonet, C., Beuret, L., Bille, K., ... & Bertolotto, C. (2010). Microphthalmia-associated transcription factor controls the DNA damage response and a lineage-specific senescence program in melanomas. *Cancer research*, *70*(9), 3813-3822.

54. Mingard, C., Wu, J., McKeague, M., & Sturla, S. J. (2020). Next-generation DNA damage sequencing. *Chemical Society Reviews*, *49*(20), 7354-7377.

55. Liu, N., Sun, Q., Wan, L., Wang, X., Feng, Y., Luo, J., & Wu, H. (2020). CUX1, a controversial player in tumor development. *Frontiers in oncology*, *10*, 738.

56. Ramdzan, Z. M., & Nepveu, A. (2014). CUX1, a haploinsufficient tumour suppressor gene overexpressed in advanced cancers. *Nature Reviews Cancer*, *14*(10), 673-682.

57. Kaur et al. CUX1 Stimulates APE1 Enzymatic Activity and Increases the Resistance of Glioblastoma Cells to the Mono-Alkylating Agent, Temozolomide. Neuro-Oncology 20 (4):484-493. 2018.

58. Lensing, S. V., Marsico, G., Hänsel-Hertsch, R., Lam, E. Y., Tannahill, D., & Balasubramanian, S. (2016). DSBCapture: in situ capture and sequencing of DNA breaks. *Nature methods*, *13*(10), 855-857.

59. Crosetto, N., Mitra, A., Silva, M. J., Bienko, M., Dojer, N., Wang, Q., ... & Dikic, I. (2013). Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nature methods*, *10*(4), 361-365.

60. Yan, W. X., Mirzazadeh, R., Garnerone, S., Scott, D., Schneider, M. W., Kallas, T., ... & Crosetto, N. (2017). BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nature communications, 8*(1), 1-9.

61. Wu, J., McKeague, M., & Sturla, S. J. (2018). Nucleotide-resolution genome-wide mapping of oxidative DNA damage by click-code-seq. Journal of the American Chemical Society, 140(31), 9783-9787.

62. Stachowicz-Kuśnierz, A., & Korchowiec, J. (2016). Nucleophilic properties of purine bases: inherent reactivity versus reaction conditions. Structural Chemistry, 27(2), 543-555.

63. Huang, Y., & Li, L. (2013). DNA crosslinking damage and cancer-a tale of friend and foe. Translational cancer research, 2(3), 144.

64. Balcome, S., Park, S., Quirk Dorr, D. R., Hafner, L., Phillips, L., & Tretyakova, N. (2004). Adenine-containing DNA− DNA cross-links of antitumor nitrogen mustards. Chemical research in toxicology, 17(7), 950-962.

65. Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing?. *Archives of Disease in Childhood-Education and Practice, 98*(6), 236-238.

66. Ari, Ş., & Arikan, M. (2016). Next-generation sequencing: advantages, disadvantages, and future. In *Plant omics: Trends and applications* (pp. 109-135). Springer, Cham.

67. Panahi, Y., Fattahi, A., Zarei, F., Ghasemzadeh, N., Mohammadpoor, A., Abroon, S., ... & Ghasemnejad, T. (2018). Next-generation sequencing approaches for the study of genome and epigenome toxicity induced by sulfur mustard. *Archives of Toxicology*, *92*(12), 3443-3457.

68. Broyde, S., Wang, L., Rechkoblit, O., Geacintov, N. E., & Patel, D. J. (2008). Lesion processing: high-fidelity versus lesion-bypass DNA polymerases. *Trends in biochemical sciences, 33*(5), 209-219.

69. Zavadil, J., & Rozen, S. G. (2019). Experimental Delineation of Mutational Signatures Is an Essential Tool in Cancer Epidemiology and Prevention.

70. Yu, Y., Wang, P., Cui, Y., & Wang, Y. (2017). Chemical analysis of DNA damage. Analytical chemistry, 90(1), 556-576.

71. Li, W., Hu, J., Adebali, O., Adar, S., Yang, Y., Chiou, Y. Y., & Sancar, A. (2017). Human genome-wide repair map of DNA damage caused by the cigarette smoke carcinogen benzo [a] pyrene. Proceedings of the National Academy of Sciences, 114(26), 6752-6757.

72. Hu, J., Lieb, J. D., Sancar, A., & Adar, S. (2016). Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. Proceedings of the National Academy of Sciences, 113(41), 11507-11512.

73. Hu, J., Li, W., Adebali, O., Yang, Y., Oztas, O., Selby, C. P., & Sancar, A. (2019). Genome-wide mapping of nucleotide excision repair with XR-seq. Nature protocols, 14(1), 248.

74. Vitelli, V., Galbiati, A., Iannelli, F., Pessina, F., Sharma, S., & d'Adda di Fagagna, F. (2017). Recent Advancements in DNA Damage–Transcription Crosstalk and High-Resolution Mapping of DNA Breaks. Annual review of genomics and human genetics, 18, 87-113.

75. Baranello, L., Kouzine, F., Wojtowicz, D., Cui, K., Zhao, K., Przytycka, T. M., ... & Levens, D. (2018). Mapping DNA Breaks by Next-Generation Sequencing. In Genome Instability (pp. 155-166). Humana Press, New York, NY.

76. Liu, Z. J., Cuesta, S. M., van Delft, P., & Balasubramanian, S. (2019). Sequencing abasic sites in DNA at single-nucleotide resolution. Nature chemistry, 1.

77. Bouwman, B. A., Agostini, F., Garnerone, S., Petrosino, G., Gothe, H. J., Sayols, S., ... & Crosetto, N. (2020). Genome-wide detection of DNA double-strand breaks by in-suspension BLISS. *Nature protocols*, *15*(12), 3894-3941.

78. Thompson, P. S., & Cortez, D. (2020). New insights into abasic site repair and tolerance. *DNA repair*, *90*, 102866.

79. Mandi, C. S., Mahata, T., Patra, D., Chakraborty, J., Bora, A., Pal, R., & Dutta, S. (2022). Cleavage of Abasic Sites in DNA by an Aminoquinoxaline Compound: Augmented Cytotoxicity and DNA Damage in Combination with an Anticancer Drug Chlorambucil in Human Colorectal Carcinoma Cells. *ACS omega*, *7*(8), 6488-6501.

80. Liu, Z. J., Martínez Cuesta, S., van Delft, P., & Balasubramanian, S. (2019). Sequencing abasic sites in DNA at single-nucleotide resolution. *Nature chemistry*, *11*(7), 629-637.

81. Dojindo Molecular Technologies. (2018). *DNA Damage Quantification Kit -AP Site Counting- Technical Manual.* Dojindo Molecular Technologies, Inc. https://drive.google.com/file/d/1m6-oVraDlnW8Tkje67UK4kx2SdWZd-xt/view

82. Abcam. (2019). *Ab129732 – Resazurin Cell Viability Assay*. Abcam plc. https://www.abcam.com/ps/products/129/ab129732/documents/ab129732_Resazurin%20Cell%20Viability%20Assay_20190715_ACW%20(website).pdf

83. Sung, J. S., & Demple, B. (2006). Roles of base excision repair subpathways in correcting oxidized abasic sites in DNA. *The FEBS journal*, *273*(8), 1620-1629.

84. Simonelli, V., Narciso, L., Dogliotti, E., & Fortini, P. (2005). Base excision repair intermediates are mutagenic in mammalian cells. *Nucleic acids research*, *33*(14), 4404-4411.

85. Hölz, K., Pavlic, A., Lietard, J., & Somoza, M. M. (2019). Specificity and efficiency of the uracil DNA glycosylase-mediated strand cleavage surveyed on large sequence libraries. Scientific reports, 9(1), 1-12.

86. Marenstein, D. R., Wilson III, D. M., & Teebor, G. W. (2004). Human AP endonuclease (APE1) demonstrates endonucleolytic activity against AP sites in single-stranded DNA. *DNA repair*, *3*(5), 527-533.

87. Wood, R., Lowery, M. (2020, June 10) *Human DNA Repair Genes*. Human DNA Repair Genes. https://www.mdanderson.org/documents/Labs/Wood-Laboratory/human-dna-repair-genes.html.

88. Tsai, S. Q., Nguyen, N. T., Malagon-Lopez, J., Topkar, V. V., Aryee, M. J., & Joung, J. K. (2017). CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nature methods*, *14*(6), 607-614.

89. Pinson, V., Takahashi, M., & Rouviere-Yaniv, J. (1999). Differential binding of the Escherichia coli HU, homodimeric forms and heterodimeric form to linear, gapped and cruciform DNA. Journal of molecular biology, 287(3), 485-497.

90. Broyles, S. S., & Pettijohn, D. E. (1986). Interaction of the Escherichia coli HU protein with DNA: evidence for formation of nucleosome-like structures with altered DNA helical pitch. Journal of molecular biology, 187(1), 47-60.

91. Ulanovsky, L., Bodner, M., Trifonov, E. N., & Choder, M. (1986). Curved DNA: design, synthesis, and circularization. Proceedings of the national academy of sciences, 83(4), 862-866.

92. Nilsson, M., Malmgren, H., Samiotaki, M., Kwiatkowski, M., Chowdhary, B. P., & Landegren, U. (1994). Padlock probes: circularizing oligonucleotides for localized DNA detection. Science, 265(5181), 2085-2088.

93. Benson, A. J., Martin, C. N., & Garner, R. C. (1988). N-(2-hydroxyethyl)-N-[2-(7-guaninyl) ethyl] amine, the putative major DNA adduct of cyclophosphamide in vitro and in vivo in the rat. Biochemical pharmacology, 37(15), 2979-2985.

94. Tan, G., Opitz, L., Schlapbach, R., & Rehrauer, H. (2019). Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific reports*, *9*(1), 1-7.

95. Emadi, A., Jones, R. J., & Brodsky, R. A. (2009). Cyclophosphamide and cancer: golden anniversary. *Nature reviews Clinical oncology*, *6*(11), 638-647.

96. Low, J. E., Borch, R. F., & Sladek, N. E. (1982). Conversion of 4-hydroperoxycyclophosphamide and 4-hydroxycyclophosphamide to phosphoramide mustard and acrolein mediated by bifunctional catalysts. *Cancer Research*, *42*(3), 830-837.

97. Shiraishi, A., Sakumi, K., & Sekiguchi, M. (2000). Increased susceptibility to chemotherapeutic alkylating agents of mice deficient in DNA repair methyltransferase. *Carcinogenesis*, *21*(10), 1879-1883.

98. Cai, Y., Wu, M. H., Ludeman, S. M., Grdina, D. J., & Dolan, M. E. (1999). Role of O 6-alkylguanine-DNA alkyltransferase in protecting against cyclophosphamide-induced toxicity and mutagenicity. *Cancer research*, *59*(13), 3059-3063.

99. Kondo, N., Takahashi, A., Ono, K., & Ohnishi, T. (2010). DNA damage induced by alkylating agents and repair pathways. *Journal of nucleic acids*, *2010*.

100. Colella, G., Marchini, S., d'Íncalci, M., Brown, R., & Broggini, M. (1999). Mismatch repair deficiency is associated with resistance to DNA minor groove alkylating agents. *British journal of cancer*, *80*(3), 338-343.

101. Fink, D., Aebi, S., & Howell, S. B. (1998). The role of DNA mismatch repair in drug resistance. *Clinical cancer research: an official journal of the American Association for Cancer Research*, *4*(1), 1-6.

102. Stojic, L., Brun, R., & Jiricny, J. (2004). Mismatch repair and DNA damage signalling. *DNA repair*, *3*(8-9), 1091-1101.

103. Karran, P. (2001). Mechanisms of tolerance to DNA damaging therapeutic drugs. *Carcinogenesis*, *22*(12), 1931-1937.

104. Wellcome Sanger Institute. (n.d.). *CCRF-CEM*. Cell Model Passports. https://cellmodelpassports.sanger.ac.uk/passports/SIDM00121.

105. Wellcome Sanger Institute. (n.d.). *Jurkat*. Cell Model Passports. https://cellmodelpassports.sanger.ac.uk/passports/SIDM01016.

106. Boysen, G., Pachkowski, B. F., Nakamura, J., & Swenberg, J. A. (2009). The formation and biological significance of N7-guanine adducts. Mutation Research/Genetic Toxicology and Environmental Mutagenesis, 678(2), 76-94.

107. Ramdzan, Z. M., Vickridge, E., Faraco, C. C., & Nepveu, A. (2021). CUT domain proteins in DNA repair and cancer. *Cancers*, *13*(12), 2953.

108. Lee, J. E., Ahn, S., Jeong, H., An, S., Myung, C. H., Lee, J. A., ... & Hwang, J. S. (2021). Olig2 regulates p53-mediated apoptosis, migration and invasion of melanoma cells. *Scientific reports*, *11*(1), 1-15.

109. Hake, S. B., & Allis, C. D. (2006). Histone H3 variants and their potential role in indexing mammalian genomes: the "H3 barcode hypothesis". *Proceedings of the National Academy of Sciences*, *103*(17), 6428-6435.

110. Lowe, B. R., Maxham, L. A., Hamey, J. J., Wilkins, M. R., & Partridge, J. F. (2019). Histone H3 mutations: an updated view of their role in chromatin deregulation and cancer. *Cancers*, *11*(5), 660.

111. Huang, T. Y., Piunti, A., Lulla, R. R., Qi, J., Horbinski, C. M., Tomita, T., ... & Saratsis, A. M. (2017). Detection of Histone H3 mutations in cerebrospinal fluid-derived tumor DNA from children with diffuse midline glioma. *Acta neuropathologica communications*, *5*(1), 1-12.

112. Yuen, B. T., & Knoepfler, P. S. (2013). Histone H3. 3 mutations: a variant path to cancer. *Cancer cell*, *24*(5), 567-574.

113. Hake, S. B., & Allis, C. D. (2006). Histone H3 variants and their potential role in indexing mammalian genomes: the "H3 barcode hypothesis". *Proceedings of the National Academy of Sciences*, *103*(17), 6428-6435.

114. Lowe, B. R., Maxham, L. A., Hamey, J. J., Wilkins, M. R., & Partridge, J. F. (2019). Histone H3 mutations: an updated view of their role in chromatin deregulation and cancer. *Cancers*, *11*(5), 660.

115. Huang, S., Luo, K., Jiang, L., Zhang, X. D., Lv, Y. H., & Li, R. F. (2021). PCBP1 regulates the transcription and alternative splicing of metastasis-related genes and pathways in hepatocellular carcinoma. *Scientific reports*, *11*(1), 1-14.

116. National Library of Medicine. (2022, May 13). *MPLKIP*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/gene/136647.

117. National Library of Medicine. (2022, July 3). *RECQL4*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/gene/9401.

118. Keeney, S. (2007). Spo11 and the formation of DNA double-strand breaks in meiosis. In *Recombination and meiosis* (pp. 81-123). Springer, Berlin, Heidelberg.

119. Hande, K. R. (1998). Etoposide: four decades of development of a topoisomerase II inhibitor. *European journal of cancer*, *34*(10), 1514-1521.

120. Spadari, S., Focher, F., Kuenzle, C., Corey, E. J., Myers, A. G., Hardt, N., ... & Pedrali-Noy, G. (1985). In vivo distribution and activity of aphidicolin on dividing and quiescent cells. *Antiviral research*, *5*(2), 93-101.

121. Singh, V., Johansson, P., Ekedahl, E., Lin, Y. L., Hammarsten, O., & Westerlund, F. (2022). Quantification of single-strand DNA lesions caused by the topoisomerase II poison etoposide using single DNA molecule imaging. *Biochemical and Biophysical Research Communications*, *594*, 57-62.

122. Mah, L. J., El-Osta, A., & Karagiannis, T. C. (2010). γH2AX: a sensitive molecular marker of DNA damage and repair. *Leukemia*, *24*(4), 679-686.

123. Mortusewicz, O., Schermelleh, L., Walter, J., Cardoso, M. C., & Leonhardt, H. (2005). Recruitment of DNA methyltransferase I to DNA repair sites. *Proceedings of the National Academy of Sciences*, *102*(25), 8905-8909.

124. Hayashi, K., Hishikawa, A., & Itoh, H. (2019). DNA damage repair and DNA methylation in the kidney. *American journal of nephrology*, *50*(2), 81-91.

125. Ha, K., Lee, G. E., Palii, S. S., Brown, K. D., Takeda, Y., Liu, K., ... & Robertson, K. D. (2011). Rapid and transient recruitment of DNMT1 to DNA double-strand breaks is mediated by its interaction with multiple components of the DNA damage response machinery. *Human molecular genetics*, *20*(1), 126-140.

126. Nyce, J. (1989). Drug-induced DNA hypermethylation and drug resistance in human tumors. *Cancer research*, *49*(21), 5829-5836.

127. Jing, D., Huang, Y., Liu, X., Sia, K. C., Zhang, J. C., Tai, X., ... & Mayoh, C. (2018). Lymphocyte-specific chromatin accessibility pre-determines glucocorticoid resistance in acute lymphoblastic leukemia. *Cancer Cell*, *34*(6), 906-921.

128. Chen, C. C., Lee, K. D., Pai, M. Y., Chu, P. Y., Hsu, C. C., Chiu, C. C., ... & Leu, Y. W. (2015). Changes in DNA methylation are associated with the development of drug resistance in cervical cancer cells. *Cancer cell international*, *15*(1), 1-9.

129. Knut and Alice Wallenberg Foundation. (n.d.). *OLIG2*. The Human Protein Atlas. https://www.proteinatlas.org/ENSG00000205927-OLIG2/pathology.

130. Xia, Y., Brown, L., Yang, C. Y., Tsan, J. T., Siciliano, M. J., Espinosa III, R., ... & Baer, R. J. (1991). TAL2, a helix-loop-helix gene activated by the (7; 9)(q34; q32) translocation in

human T-cell leukemia. *Proceedings of the National Academy of Sciences*, *88*(24), 11416-11420.

131. Ferrando, A. A., & Look, A. T. (2003, October). Gene expression profiling in T-cell acute lymphoblastic leukemia. In *Seminars in hematology* (Vol. 40, No. 4, pp. 274-280). WB Saunders.

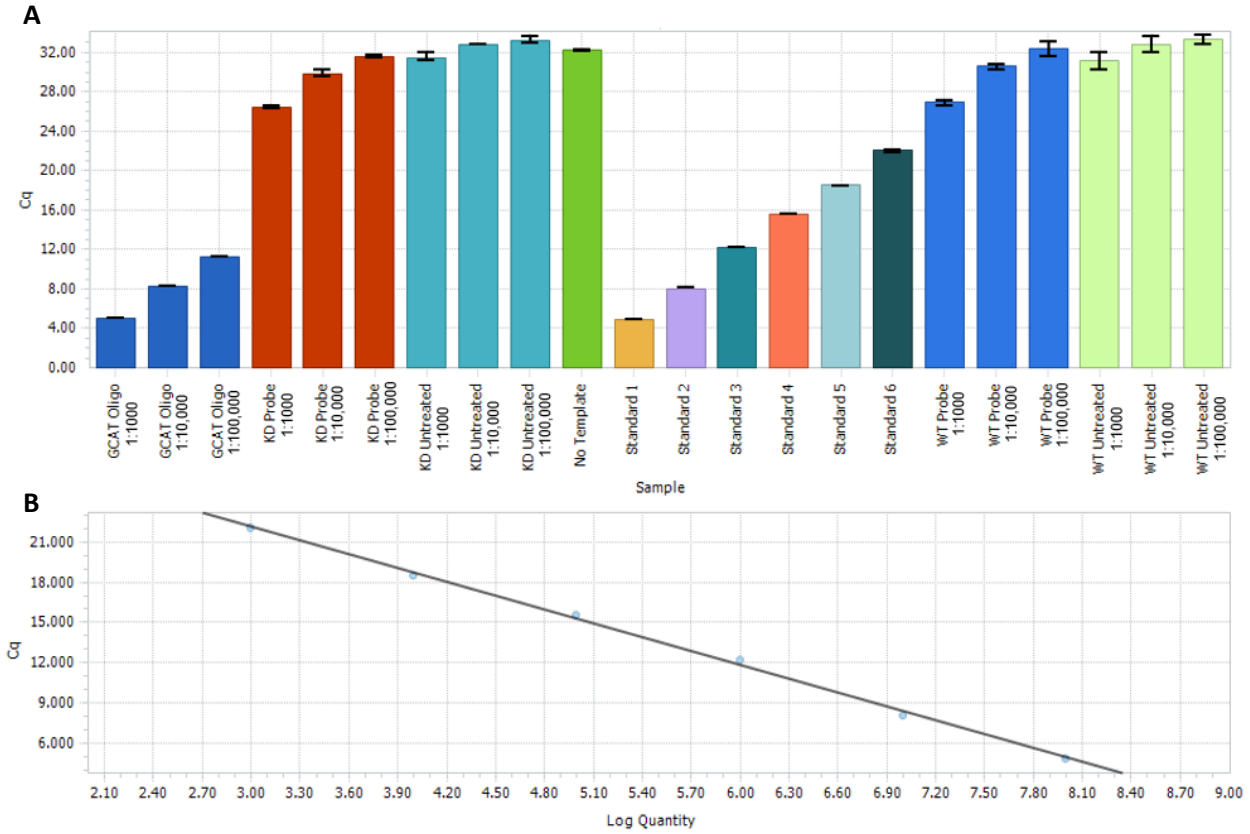
132. Folpe, A. L. (2002). MyoD1 and myogenin expression in human neoplasia: a review and update. *Advances in Anatomic Pathology*, *9*(3), 198-203.

133. Dekel, I., Magal, Y., Pearson-White, S., Emerson, C. P., & Shani, M. (1992). Conditional conversion of ES cells to skeletal muscle by an exogenous MyoD1 gene. *New Biol*, *4*(3), 217-224.

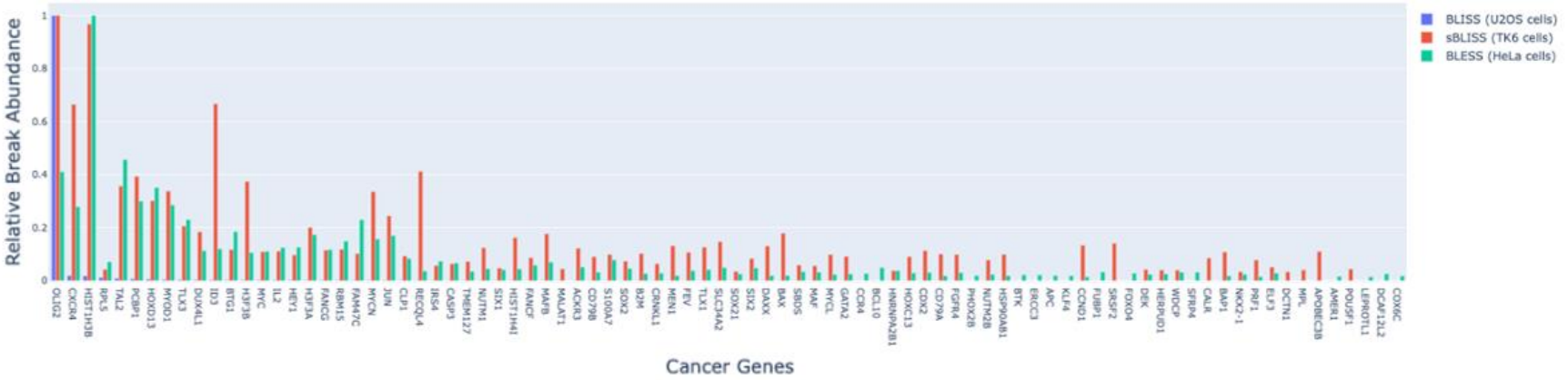
134. Zhao, X., Sun, M., Zhao, J., Leyva, J. A., Zhu, H., Yang, W., ... & Zhang, X. (2007). Mutations in HOXD13 underlie syndactyly type V and a novel brachydactyly-syndactyly syndrome. *The American Journal of Human Genetics*, *80*(2), 361-371.

135. Poetsch, A. R. (2020). AP-Seq: A Method to Measure Apurinic Sites and Small Base Adducts Genome-Wide. In *The Nucleus* (pp. 95-108). Humana, New York, NY.

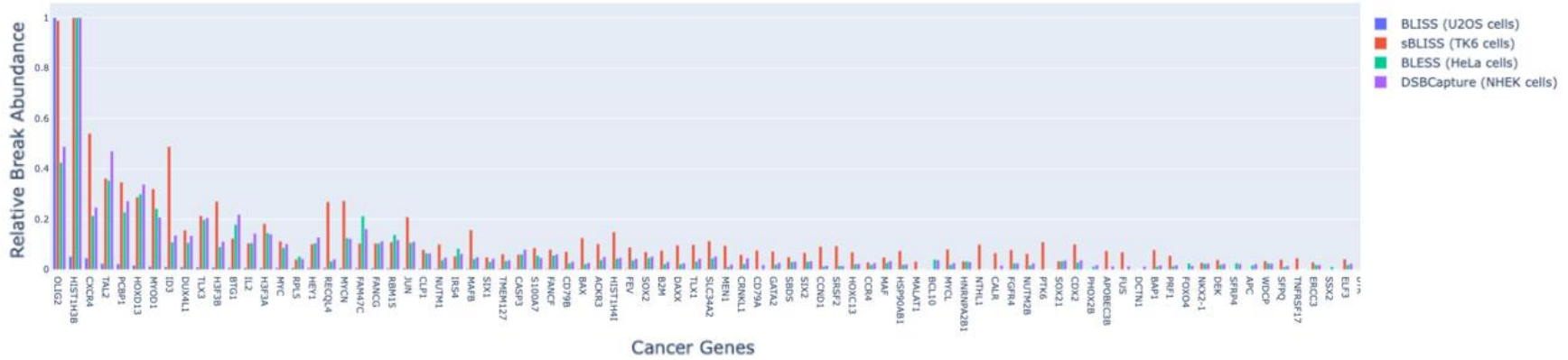
Appendix



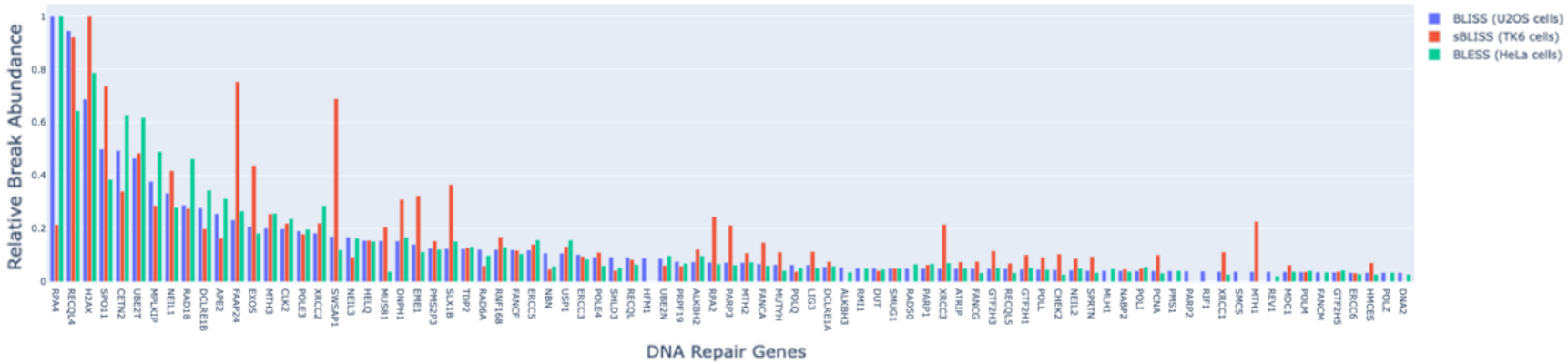
Supplementary Figure 1. Cq values for qPCR amplification of sequencing libraries prepared using the snAP-seq method compared to internal standards and controls – an expansion of Figure 16 in Chapter 3. (A) Each sample, standard, and control represent the data averaged from three technical replicates and the error bars represent the standard deviation. Across the x-axis, the different samples are listed along with the sample dilution ratio where applicable. The probe-treated and untreated samples show similar trends across the *CUX1* knockdown and wild-type samples. (B) A standard curve was generated using the Cq values produced by the standards. Using this curve, the log quantity of DNA in each sample could be calculated. However, the Cq values for all of the sample libraries fall outside of the range of the standards such that the percent recovery of the input DNA cannot be accurately calculated.



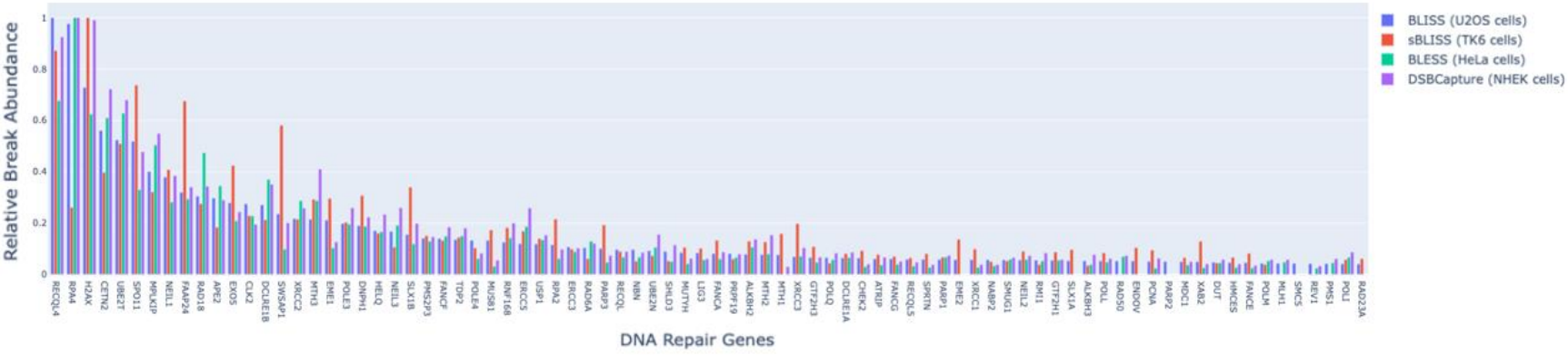
Supplementary Figure 2. Relative sensitivity of oncogenes to double strand breaks in drug-treated cell lines as determined by different double strand break sequencing methods. In both BLISS (n = 1) and sBLISS (n = 2), the cells were treated with etoposide, a chemotherapeutic that induces DSBs by complexing with topoisomerase II and DNA. In BLESS (n = 9) however, the cells were treated with aphidicolin, a DNA replication inhibitor that also produces DSBs. Across the three cell lines, there are several oncogenes that share a similar increase in sensitivity to DSB formation, suggesting that the distribution of DSBs in the genome is not random.



Supplementary Figure 3. Relative sensitivity of oncogenes to double strand breaks in untreated cell lines as determined by different double strand break sequencing methods. Each of the sequencing methods used a different cell line which may contribute to the differences in relative DSB sensitivity observed.



Supplementary Figure 4. Relative sensitivity of DNA damage repair genes to double strand breaks in drug-treated cell lines as determined by different double strand break sequencing methods. In both BLISS (n = 1) and sBLISS (n = 2), the cells were treated with etoposide, a chemotherapeutic that induces DSBs by complexing with topoisomerase II and DNA. In BLESS (n = 9) however, the cells were treated with aphidicolin, a DNA replication inhibitor that also produces DSBs. Across the three cell lines, there are several oncogenes that share a similar increase in sensitivity to DSB formation, suggesting that the distribution of DSBs in the genome is not random.



Supplementary Figure 5. Relative sensitivity of DNA damage repair genes to double strand breaks in untreated cell lines as determined by different double strand break sequencing methods. Across all four cell lines, there is largely a consensus in terms of the DNA damage repair genes that are the most sensitive to endogenous DSB formation.