# UNRAVELLING RECQ HELICASE FUNCTION IN GENOME STABILITY USING STRAND-SEQ

by

Zeid Hamadeh

B.Sc., The University of Western Ontario, 2018

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Genome Science and Technology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

October 2022

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Unravelling RecQ helicase function in genome stability using Strand-seq

submitted by Zeid Hamadeh in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Genome Science and Technology

**Examining Committee:**

Dr Peter Lansdorp, Professor, Medical Genetics, UBC
Supervisor

Dr Judy Wong, Professor, Pharmaceutical Sciences, UBC
Supervisory Committee Member

Dr Peter Stirling, Associate Professor, Medical Genetics, UBC
Supervisory Committee Member

Dr Gregg Morin, Associate Professor, Medical Genetics, UBC
University Examiner

Dr Wan Lam, Professor, Pathology and Laboratory Medicine, UBC
University Examiner

**Additional Supervisory Committee Members:**

Dr Martin Hirst, Professor, sMicrobiology and Immunology, UBC
Supervisory Committee Member

## Abstract

Helicases are a highly conserved family of motor proteins responsible for interacting with and unwinding canonical and non-canonical DNA and RNA structures. The RecQ class of helicases, known to suppress illegitimate recombination, are implicated in aging and cancer with four of the five human RecQ helicases directly linked to genome instability syndromes characterized in some cases by strong cancer predisposition or premature aging. While no human disease has been associated with the RECQL5 helicase, loss of this gene in cells is known to result in elevated double strand breaks (DSBs) and sister chromatid exchange events (SCEs), a phenotype of genome instability similar to what is observed in RecQ helicase-linked diseases of strong cancer predisposition. Until recently, studying SCEs has been limited to cytogenetic assays that map at megabase resolution. I used single cell template strand sequencing (Strand-seq) to map SCEs as changes in template strand orientation before and after loss of RECQL5 at kilobase resolution. I generated over 20 single and double knockout models for RECQL5 as well as BLM, WRN and RECQL1 helicases using CRISPR-Cas9 in the human haploid cell line, KBM7, and mapped SCEs to the genome using custom bioinformatic approaches to improve resolution and accuracy of SCE detection. I performed enrichment analysis to show SCEs are frequently occurring near actively transcribed genes with guanine quadruplexes (G4s) and common fragile sites further supporting the role of these helicase genes in suppressing inappropriate recombination at specific genomic elements. I also developed novel bioinformatic approaches to generate genotype-specific call sets for copy number alterations (CNAs), inversions, and translocations. Uncovering the role of DNA helicases in DNA repair and replication pathways is critical for understanding their significance in cancer and aging. Stand-seq offers a unique method to study helicases by mapping the location of SCEs arising in their absence.

## Lay Summary

DNA helicases are essential genes for repairing DNA damage and preventing mutations from occurring in the genome of cells. The gene class, RecQ helicases, have been implicated in aging and cancer due to their association with rapid aging syndromes that are susceptible to cancer. RECQL5 is one gene in this class that has remained understudied for its role in DNA repair. Until recently, studying DNA repair has been limited to molecular methods which suffer from limited resolution and throughput. I used a novel single cell sequencing method, known as Strand-seq, to identify genomic regions prone to DNA repair. I developed novel wet-lab and bioinformatic methods to improve the overall quality of DNA repair studies in single cells using Strand-seq. I found that that specific regions in the genome are troublesome for replication and RecQ helicases have a protective role in the faithful replication of DNA in these areas.

## Preface

This Ph.D. dissertation contains five chapters. Research was conducted under the supervision of Dr. Peter M. Lansdorp in the Genome Science and Technology graduate program at UBC. The literature review and experimental design of this thesis was performed by Zeid Hamadeh, with input from Dr. Peter M. Lansdorp. This thesis comprises both published and unpublished work performed by the author.

**Chapter 1**

A version of Section 1.3 in Chapter 1 has been published as a review article. Z. Hamadeh and P. Lansdorp, "RECQL5 at the Intersection of Replication and Transcription," *Frontiers in Cell and Developmental Biology*, vol. 8. Frontiers Media S.A., p. 324, 25-May-2020. I performed the conceptualization, background research and manuscript writing for this reviw article with assistance and guidance from Dr. Peter M. Lansdorp.

**Chapter 2**

Section 2.2.2 from Chapter 2 has been published. V. C. T. Hanlon *et al.*, "Construction of Strand-seq libraries in open nanoliter arrays," *Cell Reports Methods*, vol. 2, no. 1, p. 100150, Jan. 2022. Vincent C.T. Hanlon, Daniel D. Chan, Yanni Wang and I were responsible for protocol development and troubleshooting. Vincent C.T. Hanlon and Carl-Adam Mattsson were responsible for data curation and analysis. Diana C.J. Spierings contributed to writing the manuscript and protocol development. Robin J.N. Coope made the bespoke nanoliter arrays and other bespoke equipment. Peter M. Lansdorp. conceived the approach, contributed to protocol development, troubleshooting and writing the manuscript. Vincent C.T. Hanlon was also responsible for writing the manuscript and I assisted with revisions during manuscript writing.

Figures 2.8 and 2.10 were adapted from this publication and are based on data and methods that I was responsible for developing and generating along with the other co-authors listed above.

**Chapter 3**

A version of Chapter 3 has been published. Z. Hamadeh, V. C. T. Hanlon, and P. M. Lansdorp, "Mapping of sister chromatid exchange events and genome alterations in single cells," *Elsevier Methods*, 2022. I was responsible for conceptualization, methodology, data curation, visualization and writing the manuscript with assistance and guidance from Peter M. Lansdorp. Peter M. Lansdorp and Vincent Hanlon contributed to conceptualization and methodology.

**Chapter 4**

Chapter 4 consists of unpublished data generated and analyzed by myself.

# Table of Contents

## List of Tables

## List of Figures

## List of Abbreviations

ASHLEYS*: A*utomatic *S*election of *H*igh-quality *L*ibraries for the *E*xtensive anal*Y*sis of *S*trand-seq

AWC: Always-Watson-Crick

BCL: Binary Base Call

BIR: break induced replication

BrdU: Bromodeoxyuridine

C: Crick

CFS: Common fragile site

CI: Confidence interval

CNA: Copy-number alterations

CO: Crossover product

CIN: Chromosomal instability

crRNA: CRISPR RNAs

CS: CRISPR scores

dHJ: double Holliday junction

DSB: Double stranded break

DSBR: DSB repair

FISH: Fluorescence in situ hybridization

FOI: Features of interest

FPKM: Fragments per kilobase of processed transcript per million fragments mapped

G1: Gap 1

G2: Gap 2

G4: Guanine quadruplex

HiFi: High Fidelity

HR: Homologous recombination

INDELS: Insertions or deletions

KIX: Kinase-inducible domain interacting domain

LOH: Loss of heterozygosity

MAPQ: Mapping quality score

M: Mitosis

MNase: Micrococcal nuclease

mRNA: Messenger RNA

MSI: Microsatellite instability

nCO: Non-crossover product

NHEJ: Non-homologous end joining

NIS: Nucleotide instability

OEB: One-ended break

OP: One pot

PQS: putative G-quadruplex sequences

QC: Quality control

RNAPII: RNA polymerase II

RQC: RecQ C-terminal

S: DNA synthesis

SCEs: Sister chromatid exchange events

SDSA: Synthesis-dependent strand annealing

ssDNA: Single strands of DNA

SRI: Set2-Rpb1 interacting

SV: Structural variant

TRC: Transcription replication conflict

W: Watson

WGS: Whole-genome sequencing

XP: Xeroderma pigmentosum

## Acknowledgements

First, I would like to thank my research supervisor (Peter Lansdorp) for his continued support throughout my time in graduate school. Peter has consistently encouraged me to be inquisitive and relentless in my research endeavours to be the best researcher I can be. I would like to thank my supervisory committee (Drs. Judy Wong, Peter Stirling, Martin Hirst), for providing invaluable insight and guidance on my doctoral work. I would also like to thank several lab members and research collaborators. Thank you Dr Victor Guryev and Dr Diana Spierings (European Research Institute for the Biology of Ageing at the University of Groningen in the Netherlands) for providing additional insight and support upon my visit to ERIBA in 2019. Thank you Yanni Wang, Vincent Hanlon, Geraldine Aubert and Daniel Chan for all your thoughtfulness, encouragement and guidance in troubleshooting different experimental protocols.

I would also like to thank UBC, the Genome Science and Technology program and program coordinator, Sharon Ruschkowski, for providing me with all the resources needed throughput my graduate studies and allowing me to meet all the amazing graduate students that have contributed to my graduate school experience. Specifically, I would like to thank the following graduate students: James Wells, Kathryn Lande, Makoto Kishida, Vanessa Porter, Erin Marshall, Avery Noonan, Omer Riyadh, Jesse Ward-Bond, Emma Moreside and Charu Sankaran.

Lastly, I would like to thank my family who have always supported me. My parents and siblings have been there every step of the way throughout my pursuit of higher education and graduate studies.

# Chapter 1: Introduction

## 1.1 DNA, DNA damage and DNA repair in cancer

Cancer is a disease of the genome [1]. The human genome is constantly challenged to repair mistakes caused by endogenous and exogenous stressors to avoid genetic alterations that may disrupt gene function and perturb normal cell growth [2]–[5]. Cells that have acquired the ability for unregulated cell division proliferate indefinitely and acquire additional genetic alterations in the process [5]. The accumulation of genetic alterations, or genome instability, is a driving force in oncogenesis and, in turn, can contribute to the progressive deterioration of normal cell function [2]. Although genome instability is a characteristic of almost all human cancers and considered a defining hallmark, the amount and type of genomic instability in tumour genomes differ substantially across tumour types and cell types [4], [6]. Furthermore, the precise source of genome instability can stem from nearly all DNA transactions: replication, transcription, the cell-cycle, repair, and recombination [3]. In this chapter I provide an overview of DNA damage and repair in Section 1.1, the contribution of DNA repair to genome instability in Section 1.2 and the role of RecQ helicases in genome stability in Section 1.3.

### 1.1.1 Sources of DNA damage and associated DNA repair

DNA repair is the process by which damage to DNA is repaired to prevent mutations from propagating after DNA replication and cell division. DNA is constantly subjected to damage from endogenous and exogenous sources which by some estimates are as high as one million genetic lesions per day in a single cell [7]. Therefore, constant repair is needed to correct these changes and avoid the accumulation of mutations that can otherwise lead to diseases [7]. Many DNA repair pathways have evolved to address the breadth and diversity of possible DNA

damage. There are exogenous mutagens such as alkylating agents, X-rays and UV radiation capable of cross-linking or chemically modifying base pairs as well as endogenous sources of DNA damage such as DNA secondary structures that can obstruct or stall DNA replication (Figure 1.1) [3], [7]. The repair pathways required to address these lesions can be broadly grouped into two functional classes depending on the size of the lesion (Figure 1.1) [3], [7], [8].

Small lesions affecting one strand such as alkylated bases or deaminated bases often involve repair pathways that are minimally invasive and quick to repair the lesion (Figure 1.1) [3], [7], [8]. These lesions are typically repaired by excision repair which encompasses base excision repair, nucleotide excision repair and mismatch repair pathways, all of which involve the recognition, excision and replacement of mismatched or damaged nucleotides (Figure 1.1) [3], [7].

Alternatively, large lesions affecting both strands such as a double stranded break (DSB) or one-ended break (OEB) require repair by non-homologous end joining (NHEJ) or homologous recombination (HR) (Figure 1.1) [3], [7], [8]. In NHEJ, ends across a break are fused back together whereas in HR, an identical template molecule is used to recombine or join different strands or even molecules of DNA, generating complex joint molecules to reconstruct the native configuration across the break [3], [7], [8].

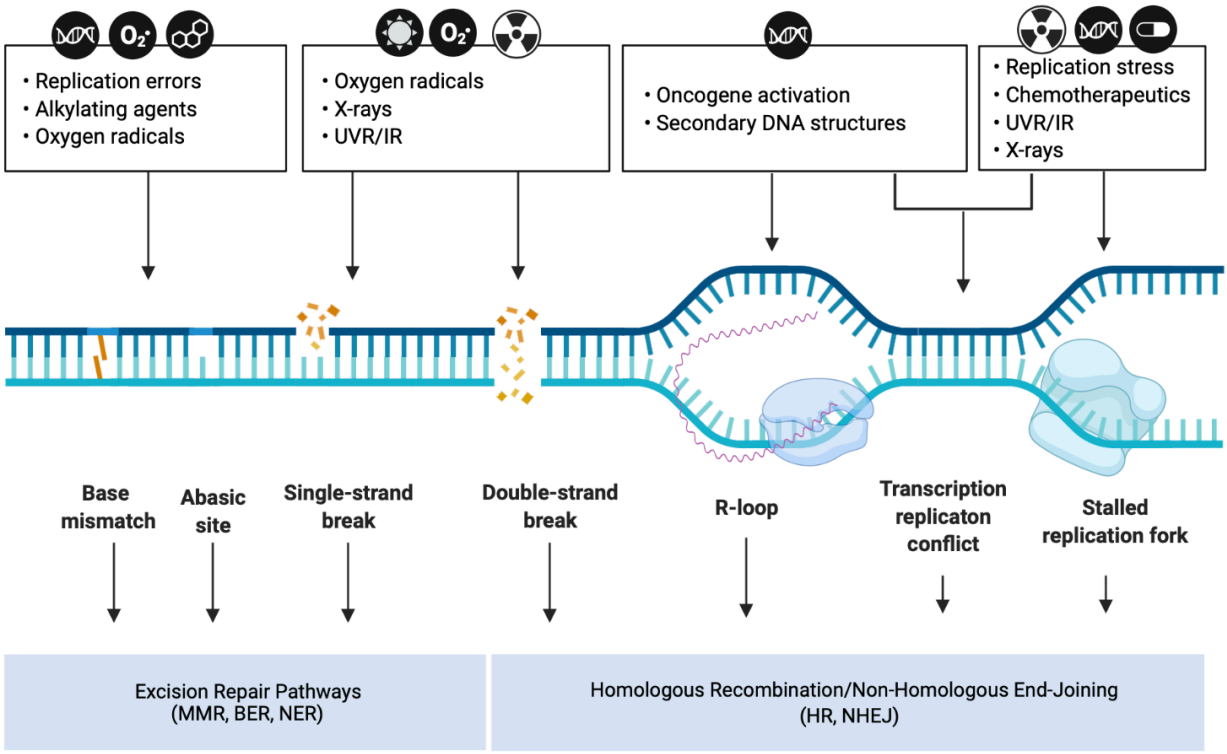



**Figure 1.1 Examples of DNA-damaging agents and the associated DNA lesions and repair pathways. Ionizing radiation (IR), ultraviolet radiation (UVR), base excision repair (BER), mismatch repair (MMR), nucleotide excision repair (NER). Created with BioRender.**

### 1.1.2 Repair of DNA double-strand breaks

When cells encounter DNA damage or replication stress that leads to a DSB, HR and NHEJ are essential for faithful DNA repair. NHEJ predominates during G1 because cells have yet to replicate their DNA and cannot access the redundancy of genetic material required as a template for faithful DNA repair by homologous recombination (HR) [9]. HR is the preferred pathway during DNA replication and there are three main steps (Figure 1.2). Firstly, 3' ssDNA overhangs are formed through end resection coordinated by the MRE11-RAD50-NBS1 complex at the DSB (Figure 1.2, step 2). Exposed ssDNA is bound by RPA, which is replaced by RAD51, to form RAD51-ssDNA nucleofilaments. These RAD51 nucleofilaments search for identical sequences present on nearby replicated sister chromatids or homologous chromosomes and

invade one or both complementary strands on the donor molecule to form a D-loop or double Holliday junction (dHJ), respectively (Figure 1.2, step 3 and 4a). Finally, strand extension of the invaded strand can occur either by synthesis-dependent strand annealing (SDSA) in the case of D-loop formation (step 3b) or through canonical DSB repair (DSBR) in the case of dHJ formation. Canonical DSBR occurs at the risk the forming hazardous crossover (CO) products where either sister chromatid or homologous chromosome donor molecules exchange strands of DNA between molecules (Figure 1.2). SDSA proceeds until the extended 5' ssDNA strand can reanneal with the template DNA of the other resected end of the DSB and continue gap filling and polymerization (Figure 1.2, step 4b). In canonical DSBR, the risk of forming CO products in turn, are a marker of genome instability [10]. In the case where a homologous chromosome is used as the template molecule opposed to a sister chromatid, the heterozygosity of deleterious alleles on one homolog may be lost if that allele is used to repair the DSB containing the healthy allele, leading to a null phenotype [10]. When dHJs form, the BLM-TOPOIIIa-RMI1/2 complex can promote convergent migration of the two HJs to produce a hemicatenane structure (Figure 1.2, step 5a) that can be processed by TOPOIIIa forming non CO (nCO) products [10]. Alternatively, structure-selective resolvases such as the SLX1/4 and MUS81-EME1 endonucleases can cleave both junctions either symmetrically or asymmetrically to form nCO and CO products, respectively [10]. Efforts to limit the risk of CO products aim to favor the DSB repair pathway that leads only to D-loop formation and SDSA. For example, disrupting D-loops before the other overhang of resected DNA anneals with the non-hybridized strand of donor DNA would bias DSBR pathways towards nCO products.

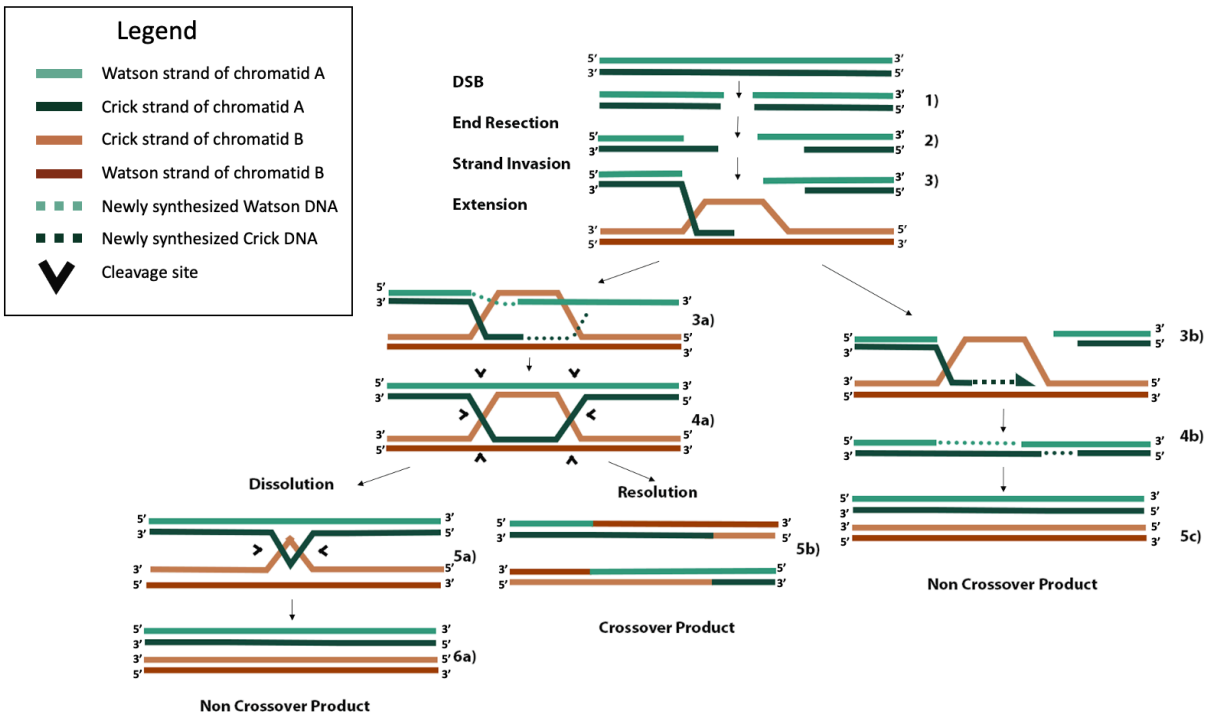



**Figure 1.2 Homologous recombination schematic of different repair pathways.**
**Three initial steps that are common to all pathways include end resection of 3′ overhangs, strand invasion of one or both overhangs with homologous donor DNA, and extension of annealed overhang (steps 1–3). When only one resected end of the DSB performs invasion, a D-loop is formed, and extension proceeds by synthesis-dependent strand annealing where one overhang is extended until there is sufficient homology to hybridize with the other resected end, gaps are filled in, and nCOs are produced (steps 3b, 4b, 5c). When the second resected end also hybridizes to the available strand in a D-loop, a dHJ is formed (steps 3a, 4a). Processing of dHJs can proceed by promoting convergent migration of the structure until a small hemicatenane structure is formed (step 5a), which can be cleaved by topoisomerases into an nCO product (step 6a). Alternatively, asymmetric cleavage of the dHJ by non-specific resolvases can result in a CO product (step 5b). The information from this figure was extracted from the works of Smith et al. (2007); West et al. (2016), and Rickman and Smogorzewska (2019).**

## 1.2 Mechanisms of genomic instability

After the discovery of the structure of DNA in 1953, the first clear connection between the processes of mutation and carcinogenesis was made shortly after [2]. The observation that polycyclic aromatic hydrocarbons in chimney soot could both alter the DNA molecule and initiate scrotal cancer led to the first association between mutagenesis and carcinogenesis [2]. Further evidence came from the discovery of two disorders of DNA repair and the associated extreme predisposition to cancer [2]. Xeroderma pigmentosum and Lynch syndrome reinforced

the role of genome instability in cancer initiation whereby deficiencies in DNA repair led to the accumulation of mutations that can disrupt the coding sequence of tumour suppressors or disrupt the regulatory apparatus of oncogenes, giving rise to unscheduled proliferation [2]. In 1999, it was shown that dysregulated oncogene expression could cause replication instability, suggesting that oncogene induced proliferation can perturb DNA replication and further contribute to the accumulation of DNA damage [11]. In 2005, DNA repair was introduced as an anticancer barrier in early-stage tumorigenesis and a target in late-stage tumours [2]. Genome stability has since been well defined as an essential cellular property needed for cells to faithfully preserve and transmit DNA between cell divisions [2], [5], [6], [12], [13]. Genome stability encompasses DNA repair using all the pathways discussed in Section 1.1.1 [2], [5], [6], [12], [13]. Conversely, genome instability describes the progressive deterioration of a cell's capacity for DNA repair characterized by an increased mutation rate and the accumulation of somatic mutations [2], [5], [6], [12], [13]. Genome instability is a functional property and enabling hallmark in cancer and aging [2], [5], [6], [12], [13]. Here, I discuss the historical discovery of genome instability and its documented causes and consequences as they relate to the fundamental cellular processes discussed in Section 1.1. Efforts to characterize the causes and consequence of genome instability have continued to highlight DNA replication, transcription, and repair as major sources of genome instability.

### 1.2.1 Replication-associated genomic instability

DNA replication is the process by which double-stranded DNA is copied to produce two identical DNA molecules [7]. DNA replication is an essential DNA transaction in every cell that is necessary for allowing two daughter cells to inherit the same genetic information. DNA replication is semi-conservative, such that each strand of DNA in the original molecule act as a

template for replication [7]. In short, DNA replication begins when the MCM2-7 helicase complex is assembled onto DNA at origins of replication, followed by the recruitment of other replication factors, helicases and polymerases to form the replisome complex [14], [15]. Next, the MCM2-7 helicase complex slides along the chromosome on both sides of the replication origin to unwind the double-stranded DNA helix into two single strands of DNA (ssDNA) [7]. ssDNA is then used as a template for synthesis of a complementary, nascent strand by the replisome complex, consisting of DNA polymerase, at each replication fork [7]. The DNA polymerases can only synthesize DNA in a 5' to 3' direction, thus one polymerase on each fork synthesizes DNA continuously while the other synthesizes small, separate Okazaki fragments that are eventually processed and ligated together after synthesis, known as the leading and lagging strand, respectively [7]. Replication is a highly dynamic process that is strictly regulated to ensure billions of nucleotides are copied accurately [7].

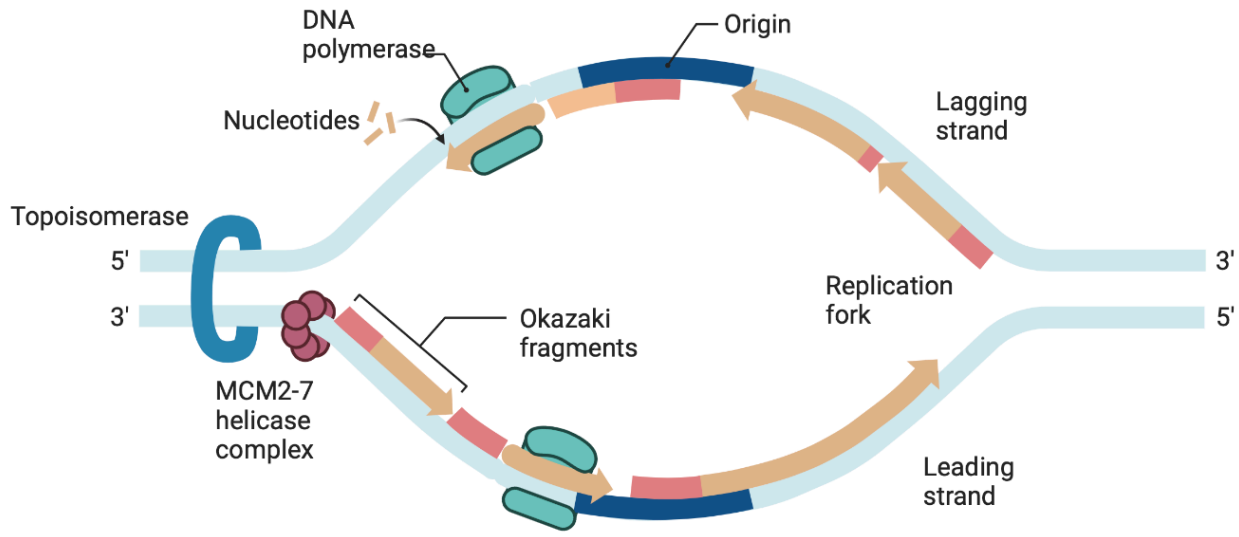



**Figure 1.3 Schematic of DNA replication.**
**The replisome complex forms at replication origins and consists of the MCM2-7 helicase complex and DNA polymerases. Each replisome slides along the chromosome in opposite directions to form both replication forks. Each replication fork uses ssDNA as a template for synthesis of a complementary, nascent strand by DNA polymerase on the leading strand. The lagging strand synthesizes Okazaki fragments. Adapted from BioRender templates.**

Faulty DNA replication can result in mutations either directly, in the form of mismatched base pairs, or indirectly, by stalled replication forks triggering breakage, rearrangement or the missegregation of chromosomes [3]. Frequent replication fork stalling can force cells into senescence or apoptosis by preventing entry into mitosis [3], [7], [8], [16]. Any condition that compromises the fidelity of DNA replication or replication fork speed is collectively referred to as replication stress and is the primary cause of genome instability [17]. DNA replication is thus inextricably linked with genome stability as DNA repair and recombination are needed to mediate any replication stress [13], [17]. There are both endogenous and exogenous sources of replication stress that are capable of stalling or even collapsing replication forks such as nicks, or breaks in one strand of the DNA molecule, can dismantle the replication fork as the fork proceeds over the break [18], [19]. Here, I discuss other sources of replication fork stalling and collapse.

#### 1.2.1.1 Replicating regions of DNA capable of forming guanine quadruplexes

Some areas of the genome are more troublesome for replication. For example, guanine-rich DNA in the genome can form alternatively folded secondary structures that may obstruct replication machinery and lead to replication fork stalling [3], [20]. For example, guanine quadruplexes (G4s) are DNA structures that form when certain guanine-rich sequences self-anneal via Hoogsteen hydrogen bond base pairing of guanine bases to form guanine quartets (G-quartets) which stack together to form G4s (Figure 1.4A-B) [21]. Each guanine in a G4 is thus stabilized by four hydrogen bonds opposed to the three bonds that typically form between guanine and cytosine nucleotides (Figure 1.4A) [21]. There are more than 700,000 guanine rich

sequences in the genome suggested to have G4-forming potential. Guanine rich sequences believed to have G4-forming potential are known as putative G-quadruplex sequences (PQSs) and are computationally defined using specific motifs [22]. Typically, a PQS contains at least four stretches of 2 or more consecutive guanine nucleotides separated by stretches of one to seven nucleotides in length known as loops [22]. This motif would be defined as, $G_{2+}N_{1-7}G_{2+}N_{1-7}G_{2+}N_{1-7}G_{2+}$ [22]. Motifs can also vary in the size of individual loops or the number of guanine stretches [22]. For example, Marsico et al. 2019 identifies potential G4s that are used in Chapter 4 by defining PQSs with the following motif: $G_{2+}N_{1-12}G_{2+}N_{1-12}G_{2+}\ N_{1-12}G_{2+}$ [22].

These methods reveal sites of the genome that may form potential G4 structures however, studies using antibodies generated against G4s to identify G4 structures have revealed only ~1–2% of these sequences form structures *in vitro*, suggesting an equilibrium exists between PGS and G4s [21]. Other approaches for detecting G4 structures rely on cations such as potassium or lithium, or G4 ligands to stabilize G4 structures in either an intact cell or cellular lysate [22], [23]. Therefore, methods for detecting G4s are dependent on either cation or G4 ligand concentration and thus the caveat of these approaches is that these factors may alter the equilibrium between PQSs and G4s from native *in vivo* conditions (Figure 1.4C). Considering this caveat in defining G4s, all mentions of G4s in Chapter 4 are considered potential G4s.

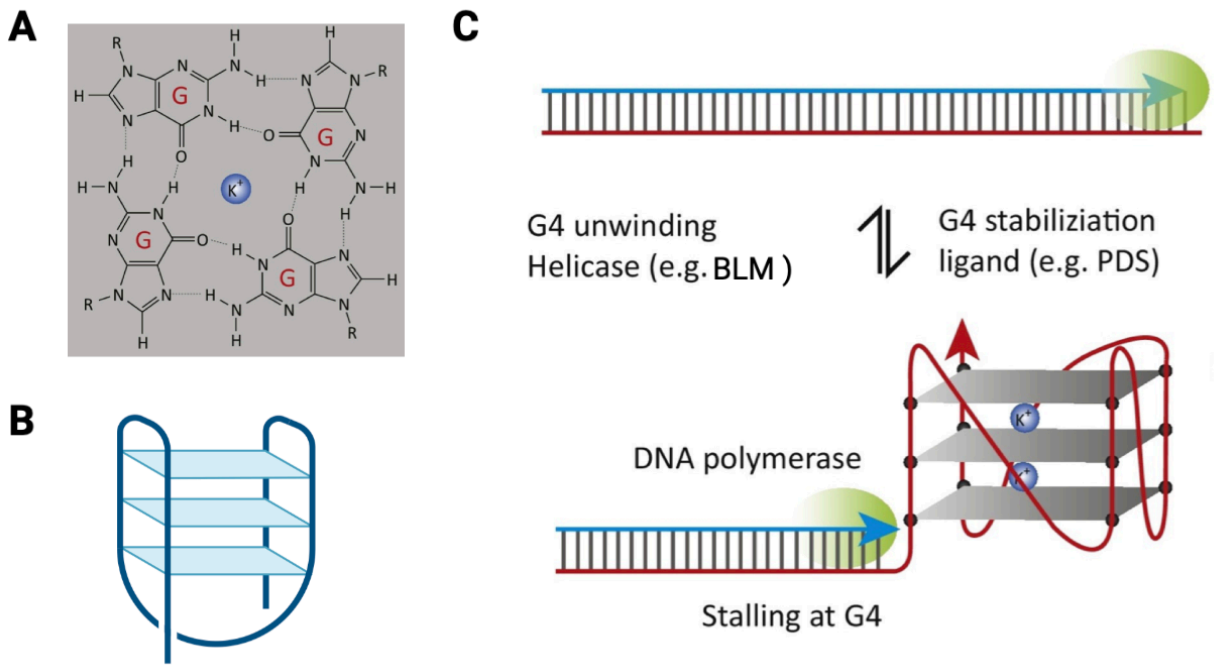



**Figure 1.4 Guanine quadruplex (G4) structure and schematic depicting the impact of G4s on DNA replication.**
**(A) A G-quartet is formed by Hoogsten hydrogen bonding between four guanines making up a square planar configuration. (B) A G4 is formed by 2 or more stacked G-quartets. (C) During DNA replication, G4s are unwound by DNA helicases to allow progression of the DNA polymerase. Insufficient helicase activity or stabilization of G4s by G4 ligands (e.g. pyridostatin) can result in replication fork stalling. Figure adapted from Kwok et al., 2017 and created using BioRender.**

These structures are implicated in DNA replication because they create barriers in the DNA template that require dismantling by DNA helicases for faithful replication progression (Figure 1.4C) [3], [24], [25]. During DNA replication, G4s are often unwound by DNA helicases to allow for progression of the DNA polymerase. Insufficient helicase activity or stabilization of G4s by G4 ligands, such as pyridostatin (PDS), can result in replication stress and replication fork stalling [3], [21], [25]. When the barrier cannot be resolved properly, replication fork stalling results in an increase in ssDNA and the subsequent recruitment of DNA repair factors such as RAD51 and ATR [3], [16]. These factors initiate the replication stress response that serves to reverse the replication fork by resecting and hybridizing nascent strands to form a four-way molecule known as a “chicken foot” structure in an attempt to bypass the barrier and restart

replication [3], [26], [27]. At this point, if replication fork restart is not possible, the fork will be cleaved with endonucleases to trigger repair by HR using the sister chromatid as a template [3], [16], [26].

G4s are also implicated in transcriptional regulation due to the abundance of potential G4s in human embryonic stem cells that are lost during lineage specification [21]. Zyner et al., 2022 also found potential G4s not lost during differentiation but preserved in both embryonic and downstream lineages are associated with genes involved in essential cellular functions [21]. They also found that G4 stabilization could effectively delay stem cell differentiation, further supporting the notion that G4s may be important genomic structural features linked to cellular differentiation and transcriptional regulation [21].

#### 1.2.1.2 Replicating regions of DNA capable of forming other secondary structures

Other areas of the genome more troublesome for replication include highly repetitive DNA sequences capable of forming higher-order non-B DNA structures. For example, TA dinucleotide repeats form cruciform secondary structures that may stall replication forks in a length dependent manner. Such cruciform structures require unwinding by DNA helicases to avoid initiating the replication stress response via RAD51 and ATR [28]. There are nearly a dozen types of higher-order non-B DNA structures that have been described, including Z-DNA, R-loops (discussed in Section 1.2.2.2), three-way and four-way joint molecules, all with variable biological and pathogenic significance [3].

#### 1.2.1.3 Replicating common fragile sites

Common fragile sites (CFSs) are regions of the genome that are prone to replication fork stalling and breakage [29], [30]. CFSs exist in all human genomes and are often found in large, transcriptionally active genes or oncogenes undergoing aberrant transcription [29], [30]. Replication fork stalling and subsequent breakage at these sites is thought to trigger genomic rearrangements as high frequencies of sister chromatid exchange events (SCEs), translocations, and insertions or deletions (INDELS) have been observed in these regions, especially in cancer genomes [29]–[31].

### 1.2.2 Transcription-coupled genome instability

Transcription is the process by which the sequence information of a gene that is stored in DNA is copied into a new molecule of messenger RNA (mRNA) that can then exit the nucleus and be translated into a protein [7]. Transcription is initiated as enzymes and proteins are recruited by transcription factors to unwind and open up the DNA double helix and allow for the RNA synthesizing complex, the RNA polymerase holoenzyme, to bind and form the transcription initiation complex inside a transcription bubble (Figure 1.5)[7]. This complex slides along the DNA template, pushing the transcription bubble along the length of the gene and synthesizing the entire mRNA molecule corresponding to the gene being transcribed (Figure 1.5) [7]. Splicing of mRNA to the mature messenger RNA occurs before the molecules exit the nucleus serve as the basis for the synthesis of protein by the process of translation in the cytoplasm of cells [7].

f

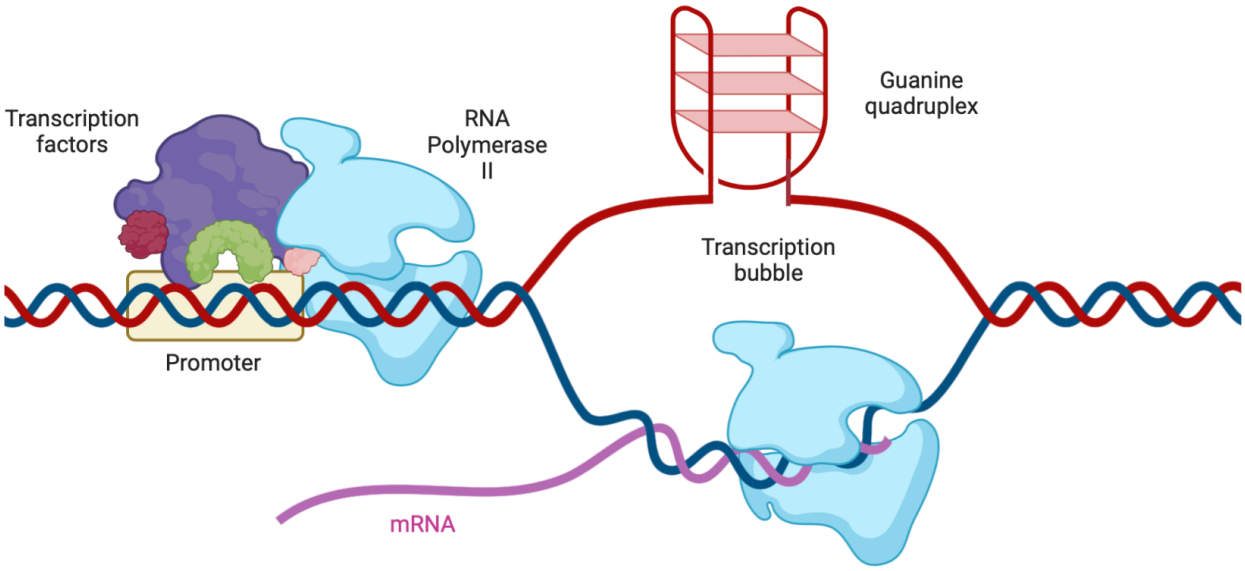



**Figure 1.5 Schematic of transcription by RNA polymerase II. Adapted from BioRender.**

Transcriptional activity has the potential to contribute to genome instability with the most notable evidence of this connection being the correlation between transcriptional activity and the accumulation of mutations in that gene [32]. Mutations that disrupt the coding or regulatory sequence of a gene can undoubtedly alter gene expression and have phenotypic consequences [32]. Transcription-coupled genome instability mainly stems from the collision of transcription and replication machineries or the formation of DNA:RNA hybrid molecules during transcription known as R-loops which are discussed in Section 1.2.2.2 [32]. Additionally, the presence of certain motifs in the ssDNA that is exposed during transcription can result in the formation of stable secondary structures, such as G-quadruplexes or hairpin loops, that can also pose a barrier to transcription and replication machinery (Figure 1.5).

#### 1.2.2.1 Collision between transcription and replication machinery

The same DNA substrate can simultaneously undergo replication and transcription. This opens up the possibility of colliding transcription and replication machineries on the same region of DNA, which can result in DSBs [33], [34]. Transcription-replication collisions can be head-on, where the direction of the replication fork and the transcription bubble are opposite, or they can be co-directional, where the direction of the replication fork and the transcription bubble are the same [33], [34]. Although co-directional collisions are still troublesome for the accurate completion of both processes, head-on collisions are considered more deleterious because of their ability to dissociate transcription machinery and stall or collapse replication forks [33], [34].

#### 1.2.2.2 Co-transcriptional R-loops

During DNA transcription, RNA-DNA hybrid molecules known as R-loops can form when the newly synthesized RNA molecule hybridizes to the coding DNA strand [3], [33], [35]. These RNA-DNA hybrids prevent displaced DNA strands from re-annealing and result in long stretches of ssDNA, both of which can cause issues for a cell [36]. ssDNA is prone to DNA damage and is also a major signal for DNA repair pathways as discussed in the Section 1.2.2.1 [36]. R-loops also can pose as a barrier to replication, resulting in replication fork stalling and/or collapse [36]. Interestingly, the presence of certain motifs when the non-complementary strand of ssDNA is exposed can result in the formation of stable secondary structures, such as G-quadruplexes, that can also pose a barrier to replication machinery [3].

### 1.2.3 Genome instability syndromes

The discovery of Lynch syndrome and Xeroderma pigmentosum helped establish the role of mutations and DNA repair in cancer [2]. To date, many disorders of DNA repair have been discovered. Such disorders can be classified on the basis of the DNA repair pathway that is deficient and the associated type of genomic instability.

Deficiency in nucleotide or base excision repair pathways causes nucleotide instability (NIS) defined by frequent single nucleotide polymorphisms (SNPs) and small insertions or deletions. Xeroderma pigmentosum (XP) is an example of this whereby the inability to repair UV radiation-induced thymidine dimers results in NIS and extreme skin cancer predisposition.

Deficiency in mismatch repair genes causes microsatellite instability (MSI) which is defined by the frequent expansion or contraction of short nucleotide repeats known as microsatellites. Lynch syndrome is an example of this whereby the inability to correct mismatched nucleotides introduced by DNA polymerase during replication results in MSI and the accumulation of SNPs. 3-4% of all colorectal cancers are characterized by MSI.

Deficiency in the repair of large lesions such as DSBs or OEBs causes chromosomal instability (CIN) which is defined by frequent structural rearrangements and alterations in copy number. Bloom syndrome, Werner syndrome and Rothmund-Thompson syndrome are examples of three disorders of DSB repair characterized by elevated levels of somatic structural rearrangements, premature aging, and cancer predisposition due to deficiency in one of three RecQ helicases, although the types of cancer associated with these diseases vary. BRCA1/2 are also involved in DSB repair and mutations in these genes are responsible for 3% and 10% of breast and ovarian cancers, respectively [19], [37]. p53 has been shown to play a regulatory role

in HR and is known to be mutated in over 50% of all cancers [19], [37]. It is well accepted that deficiency in DSB repair is a major contributor to genome instability and cancer predisposition.

### 1.2.4 Investigating genome instability

The type and extent of genome instability in tumour genomes has significant implications for patient prognosis and therapeutic treatment [38]. Among the types of genomic instability, CIN is considered a driving force in oncogenesis. Currently, detection of CIN is done using a variety of technologies that vary in sensitivity and throughput. Here, I discuss the use of cytogenetic techniques for the detection of DSBs, autoradiography and differential staining for the detection of SCEs, sequencing studies for the genome-wide profiling of SVs and Strand-seq studies for the high-resolution mapping of structural variants (SVs) and SCEs.

#### 1.2.4.1 Cytogenetic studies for the detection of DSBs

There are several methods that have been used to quantify the degree of genome instability in cells using DSBs as a surrogate measure. For example, in response to DSBs, cells rapidly phosphorylate H2AX, the minor histone H2A variant, to produce γH2AX [39]. γ-H2AX staining of DSBs has been used to infer levels of replication stress in cells [39]. However, γ-H2AX staining remains an indirect monitor of DSB formation and thus positive staining does not always represent DSB formation nor does the timing of positive staining fully correlate with DSB repair [39], [40]. Fluorescence in situ hybridization (FISH) has also been used to identify and count translocations and aneuploidy and metaphase spreads can be used to count chromosome aberrations [41]. However, these methods may not be informative when used on cell lines that exhibit low levels of genomic instability.

#### 1.2.4.2 Autoradiography studies for the detection of SCEs

Alternative to DSBs, the number of SCEs in cells are a useful indicator of genomic instability as increased levels are a diagnostic phenotype for cancer-prone disorders such as Bloom Syndrome [42]. Most spontaneous SCE events in wild-type yeast cells were proposed to reflect repair of gaps in ssDNA or lesions in the DNA template that cause a template strand switch or a controlled DSB [12]. Therefore, they reflect the degree of replication stress that a cell has experienced and can be used as a precise measure of genomic instability [43]. SCEs are error-free exchanges of genetic material between replicated sister chromatids and thus are not considered deleterious [43]. In rare cases, the template switching of SCEs can involve unequal crossing over between chromatids during DSB repair and result in the gain or loss of genetic material [25], [43]. The crossing over between homologs instead of sister chromatids during DSB repair can result in a loss of heterozygosity (LOH) however, this is also considered a rare outcome of DSB repair [25], [43]. Several methods for identifying and counting SCEs are discussed in this section.

The first observation of SCEs in cells came from autoradiography studies [44]. Plant cells grown in the presence of tritium ($^{3}$H) labelled thymidine allow for the incorporation into newly synthesized (nascent) DNA strands during replication [44]. Autoradiography of metaphase spreads distinguishes the levels of radioactivity between each sister chromatid such that some, but not all, chromosomes would display a differential staining pattern between chromatids [44]. Some chromatids would display a switch from light to dark staining, while the sister chromatid would show the inverse pattern, indicating an SCE has occurred between sister chromatids [44].

### 1.2.4.3 Differential cytogenetic staining of sister chromatids for SCE detection

The thymidine analogue bromodeoxyuridine (BrdU) was found to produce a similar chromosome labelling pattern to $^{3}$H-labelled thymidine that can then be detected using Hoechst and Giemsa dyes (Figure 1.6) [45]. Several studies found the presence of variable BrdU concentrations in the cell culture medium had no effect on SCE frequency suggesting BrdU incorporation does not induce SCEs [46].

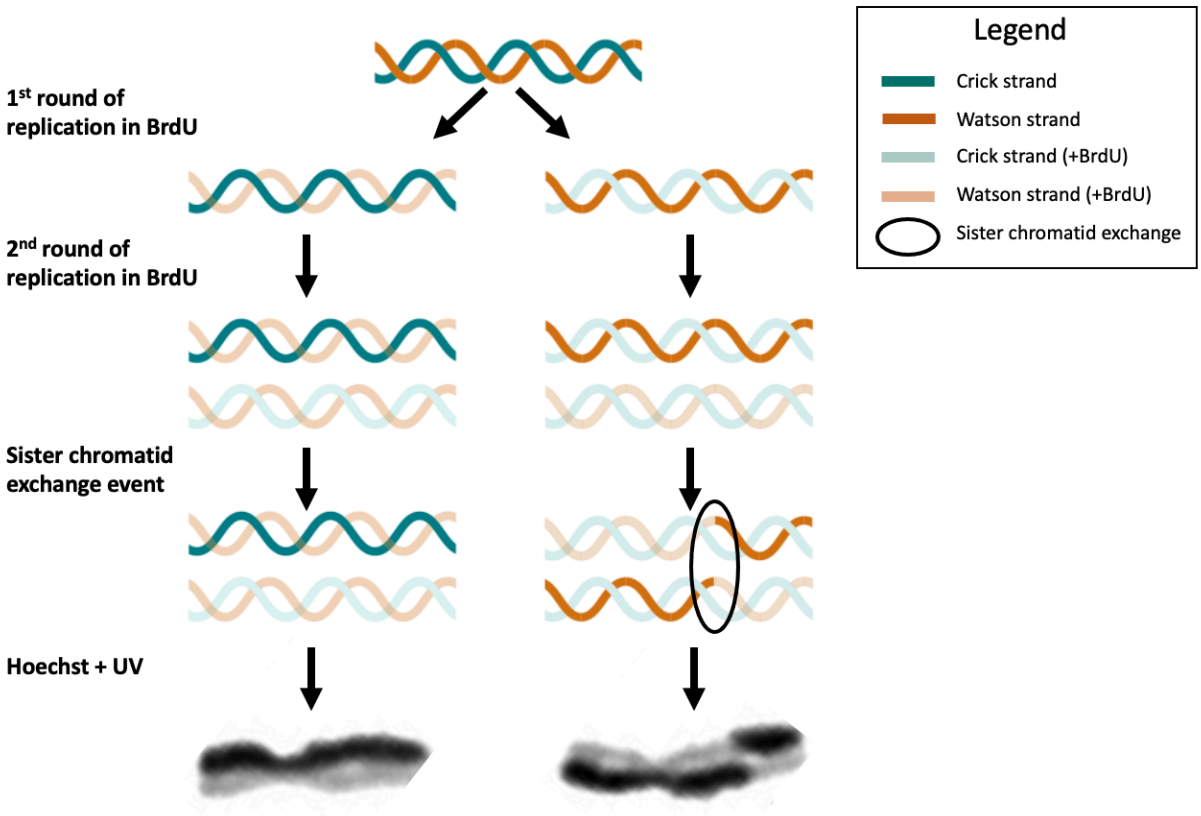



**Figure 1.6 Schematic of sister chromatid staining assay.**
**After 2 rounds of DNA replication in the presence of BrdU, staining with Hoechst 33258 and exposure to UV light reveal differential staining pattern between sister chromatids. Solid orange and teal lines represent normal DNA strand and faded orange and teal lines represent BrdU-substituted DNA strand. Circle highlights point of exchange between sister chromatids. (Left) Normally, one sister stains uniformly dark (top) and the other uniformly light (bottom). (Right) The effect of an SCE following the second round of DNA synthesis in the presence of BrdU. With exchange, light and dark staining regions switch at the point of exchange (circle). Created with Biorender.com**

Both autoradiography and differential staining of sister chromatids have several limitations. First, they suffer from poor resolution in their ability to map individual SCEs to the

genome at megabase scale [25]. SCEs can be localized to a region spanning several megabases making it difficult to investigate if SCEs are randomly occurring throughout the genome [25]. Second, there is no widely available software for the automation of SCE counting, thus these approaches are heavily reliant on manual curation for the counting of individual SCEs. This can be very time consuming for the analysis of hundreds of cells. Lastly, both DSBs and SCEs are merely a surrogate measure of genome instability and do not capture the full mutational landscape of cells. Sequencing-based studies to characterize somatic mutations in cells address these limitations.

#### 1.2.4.4 Sequencing studies for the profiling of somatic mutations

Bulk whole-genome sequencing (WGS) can be used to characterize the genome instability of cells by identifying somatic mutations [47]–[49]. Identifying all somatic mutations in cancer genomes can reveal alterations present in the coding or regulatory sequences of genes and the mutational processes implicated in their formation [50]. Bioinformatic tools known as "callers" can be used to call somatic mutations such as single nucleotide polymorphisms (SNPs), small insertions or deletions (INDELs) less than 50 bp, and structural variants (SVs) referring to any structural rearrangement greater than 50 bp in WGS data and generate comprehensive mutation callsets that are cross-referenced against a germline callset to distinguish somatic mutations from germline mutations [51]. The technologies for WGS can be broadly classified into two categories: short-read sequencing and long-read sequencing [52].

Short-read sequencing is dominated by Illumina and is considered the gold standard for any large-scale, clinical grade sequencing with highly accurate SNP and INDEL calling and base-calling accuracy exceeding 99.9% (Table 1.1) [52]. However, there are many challenges associated with detecting SVs when using short-read WGS data [52]. The main limitation with

short-read sequencing is reads less than 400 bases long are too short to detect more than 70% of the SVs in the human genome that lie within DNA that is inaccessible to assembly or variant discovery because of repeat-rich DNA or atypical GC content (Table 1.1) [52]. Few reads map in these regions and reads that do map to these regions may be incorrectly aligned if they fall within the span of a repetitive stretch of DNA [52]. In fact, repeat-rich DNA makes up 45% of the human genome and interestingly, these areas are among the most mutable consisting of many SVs such as inversions, duplications and translocations [3]. Therefore, SV callers can suffer from low accuracy due to errors in base-calls, alignment, or assembly resulting in either false negative or false positive SV calls [52].

On the other hand, long-read sequencing is dominated by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), both of which can generate continuous sequencing reads ranging from 10 kilobases to several megabases in length (Table 1.1) [52]. Long-reads are capable of traversing highly repetitive regions of the genome to reveal complex SVs that would typically go undetected by short-read sequencing methods and have contributed to the most complete assemblies of the human genome [52], [53]. These technologies typically have lower base-calling accuracy than Illumina short-read sequencing which limits the accuracy of SNP and INDEL calling (Table 1.1) [52]. PacBio can generate either continuous long reads that are typically between 1 and 100 Kb in length with between 85-92% base calling accuracy or High Fidelity (HiFi) reads which can exceed 99% base calling accuracy (Table 1.1) [52]. ONT can generate ultra-long reads that can exceed several Mb in length with between 87-98% base calling accuracy (Table 1.1) [52]. The biggest limitation of both long read technologies is the amount of starting DNA required. PacBio HiFi sequencing typically requires several micrograms of DNA, or millions of cells, for input whereas ONT require 1 microgram of input material (Table 1.1)

[54]. Some low-input PacBio protocols can use as little as 5 nanograms of input DNA however, this still equates to thousands of cells for input without any preamplification (Table 1.1) [54]. Additionally, high sequencing costs limit their general use [52].

| | Oxford Nanopore | Pacific Biosciences | Illumina WGS |
|---|---|---|---|
| Starting material required | ~1 µg | ~5 ng | 1-500 ng |
| Read length | ~1 Kb - 2 Mb | ~1-100 Kb | ~100-400 bp |
| Base calling accuracy | 87-98% | 85-99% | 99.9% |
| Applications for SV discovery | Intermediate to large SVs (>2 Kb) | Intermediate to large SVs (>2 Kb) | SNPs, INDELS, and small SVs (< 300 bp) |

**Table 1.1 Whole genome sequencing technologies and associated features.**

For short-read and to a lesser extent, long read sequencing methods, a major limitation of bulk WGS is the sensitivity for identifying rare mutations occurring in a small fraction of the bulk population of cells analyzed. Mutations present in a low fraction of cells, or with a low variant allelic frequency (VAF), will create computational challenges for distinguishing rare mutations from sequencing alignment errors or false positive calls [55]. This problem can partly be addressed by investigating single cells independent from the bulk population as I discuss in the next section.

#### 1.2.4.5 High resolution SCE and SV mapping using Strand-seq

Our research group developed a single cell sequencing-based approach to address some of the issues of bulk WGS discussed above in Section 1.2.5.4 [56]. Single-cell DNA template strand sequencing (or Strand-seq) can be used to detect several complex types of SVs at the single-cell level [49], [56]–[58]. While other single-cell WGS (scWGS) techniques are also capable of detecting some of these SVs, the unique preservation of native template strand orientation in Strand-seq reads permits the improved detection of copy-neutral rearrangements

such as SCEs, inversions and translocations that typically evade detection otherwise [49], [56]–[58]. In fact, Strand-seq is the only sequencing technology capable of detecting SCEs [49], [56], [57]. Additionally, a multi-platform comparison of inversion calling among Strand-seq and other short-read and long read WGS technologies revealed that Strand-seq was the only technology that provides data that can be used to make highly reliable inversion calls and inversion calls exceeding 50 Kb in size on its own [49].

One of the limitations of Strand-seq is that it requires dividing cells and thus nondividing or apoptotic cells cannot be studied. Additionally, like most scWGS technologies, the resolution of how finely SVs can be mapped to the genome is proportional to the fraction of genomic DNA that is captured in a single cell Strand-seq library and the depth of the subsequent sequencing [59]. This resolution far exceeds the resolution of cytogenetic approaches but is generally lower than that of bulk WGS approaches [56], [57].

In short, Strand-seq exploits the semi-conservative nature of DNA replication to incorporate the thymidine analog, BrdU, into the newly synthesized strand to allow for the distinction between template and nascent DNA strands [56], [57]. DNA fragments with BrdU can be selectively degraded by treatment with Hoechst and UV irradiation before PCR amplification [56], [57]. PCR amplification after degradation of nascent strand DNA fragments allows for the selective amplification of template strand reads [56], [57]. Illumina WGS is used to generate directional libraries with reads mapping to the reference genome in the orientation of the native parental DNA template strands [56], [57]. Unlike other single cell sequencing techniques, Strand-seq libraries harbor unique signatures of intra-chromosomal template strand changes that represent orientation-dependent SVs or SCEs [56], [57]. With the latest library preparation protocols up to 25% of the genome in a single cell can be captured in a Strand-seq

library resulting in SCEs being mapped to the genome several orders of magnitude more precise than what has been shown using cytogenetics [60].

Genomic signatures of SCEs in knockout models have been shown to elucidate the functions of helicases in DNA repair and genome stability. For example, deficiency in the BLM helicase has been shown to elevate SCE levels near G4s in actively transcribed genes, implicating BLM in unwinding G4s in specific genomic contexts [25]. Therefore, I wanted to investigate additional RecQ helicase functions using Strand-seq.

## 1.3 RecQ helicases in genome stability

### 1.3.1 Genome instability syndromes associated with RecQ helicases

Of the five RecQ helicases, *RECQL1, BLM, WRN* and *RECQL4* are associated with specific diseases of genome instability such as RECON syndrome, Bloom Syndrome (BS), Werner Syndrome (WS) and Rothmund-Thompson Syndrome (RTS), respectively (Table 1.2) [61]. Despite sharing functional roles, the genetic disorders associated with these helicases exhibit a unique set of clinical and cellular features, further supporting the non-redundant role of these genes (Table 1.2) [61]. The main clinical phenotype associated with RECON syndrome are short stature, progeroid facial features, skin photosensitivity and a moderately increased breast cancer risk (Table 1.2) [61]. Cells deficient in the RECQL1 helicase exhibit hypersensitivity to DSBs and genotoxic agents (Table 1.2) [61]. The main clinical phenotype associated with WS are features of premature aging such as osteoporosis, cataracts and loss of hair as well as early onset of sarcomas and mesenchymal tumours (Table 1.2) [61]. Cells deficient in the WRN helicase exhibit premature replicative senescence and are also hypersensitive to genotoxic agents that perturb DNA replication (Table 1.2) [61]. The main clinical features associated with BS are

dwarfism, mental retardation, microcephaly, immunodeficiency and predisposition for all cancer types (Table 1.2) [61]. Cells deficient in the BLM helicase are also hypersensitive to genotoxic agents that perturb DNA replication and exhibit elevated levels of DSBs and SCEs (Table 1.2) [61]. The main clinical features associated with RTS are growth retardation, skeletal dysplasia, sparse scalp hair, hypogonadism and early onset of osteosarcomas. Cells deficient in this helicase also exhibit a hypersensitivity to genotoxic agents (Table 1.2) [61].

Whereas *RECQL5* remains to be associated with a specific disorder [12]. In a group of 50 mice deficient in the murine homolog of RECQL5, Recql5, nearly 50% developed cancer within 22 months compared to 6% in wildtype mice [62]. Additionally, cells deficient in RECQL5 display a phenotype of chromosomal instability resulting in elevated SCEs and DSBs similar to cells deficient in most of the other RecQ helicases [62]. Unique to RECQL5 is a C-terminal domain consisting of multiple protein-protein interaction motifs that are believed to help RECQL5 regulate DNA repair intermediate structures resulting from the collision of DNA transcription and replication machinery [63].

| **Syndrome** | **Main clinical features** | **Main cellular features** | **Cancer predisposition** |
|---|---|---|---|
| RECON syndrome (RECQL1) | Short stature, progeroid facial features, skin photosensitivity | Hypersensitivity to DSBs genotoxic agents | Moderately increased breast cancer risk |
| Bloom syndrome (BLM) | Dwarfism, mental retardation, microcephaly, immunodeficiency | Elevated DSBs and SCEs | Early onset, all types |
| Werner syndrome (WRN) | Premature aging, loss of hair, short stature, osteoporosis | Premature replicative senescence, telomere erosion, hypersensitivity to DSBs genotoxic agents | Early onset of sarcomas and mesenchymal tumors |
| Rothmund-Thomson syndrome (RECQL4) | Growth retardation, skeletal dysplasia, sparse scalp hair, hypogonadism | Hypersensitive to DSBs genotoxic agents | Early onset of osteosarcomas |

**Table 1.2 Comparison of RecQ helicase-associated genome instability disorders.**
**Main clinical features, cellular features and cancer predisposition phenotype for each syndrome was retrieved from Abu-Libdeh et al. (2022).**

### 1.3.2 Biochemical characterization of RecQ helicases

Helicases are a highly diverse class of motor proteins that use ATP to unwind or translocate strands of nucleic acids [12], [64]. Helicases can be classified as DNA or RNA helicases, depending on their substrate [65]. The RecQ helicases are one highly conserved class of DNA helicases from bacteria to complex eukaryotes known best for preventing inappropriate recombination [12]. Bacteria and lower eukaryotes have only one RecQ orthologue, *RecQ*, whereas humans have 5 RecQ genes, each with a unique gene structure, suggestive of functional divergence (Figure 1.7A).

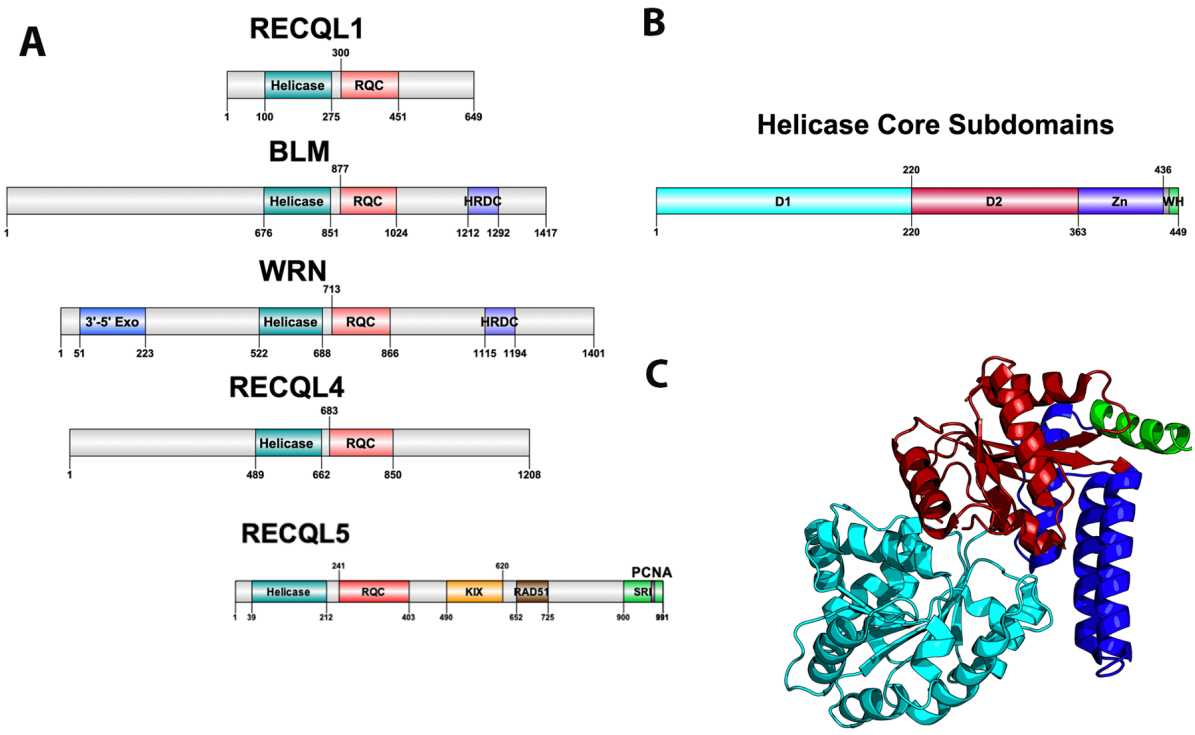



**Figure 1.7 Structure of RecQ helicases**
**(A) Domain architecture of all five RecQ helicases and isoforms of the RECQL5 helicase, aligned by core helicase and RQC domains. (B) Subdomains of the core helicase domain of RECQL5. Zn refers to the Zn-binding domain and WH refers to the winged helix-like structure of RECQL5. (C) Cartoon structure**

**diagram of the core helicase domain, colored by subdomain. Gene structure diagrams were designed using Domain Graph (DOG), and the protein structure was designed using PyMol with the crystal structure used in Newman et al. (2017). Data on gene structure was also retrieved from Croteau et al. (2014).**

All of the most abundant isoforms of the RecQ helicases share two common domains: the core helicase domain and the RecQ C-terminal (RQC) domain which together make up the catalytic core of the enzyme (Figure 1.7 B-C). Some members additionally contain a helicase and RNaseD C-terminal (HRDC) domain with a function that remains unclear but appears not to be essential for helicase activity [66]. Within the core helicase domain there are three subdomains, N and C terminal RecA like core domains (D1 and D2) and a $Zn^{2+}$ binding domain, followed by a winged helix (WH) responsible for interacting with DNA (Figure 1.7 B-C). It is the catalytic core helicase domain that is responsible for unwinding dsDNA, translocating ssDNA and in some cases, remodeling of non B-DNA structures that may arise during transcription, repair and replication [65].

*RECQL1* was the first RecQ helicase to be discovered in 1994 and was mapped to chromosome 12p12 [67]. *RECQL1* encodes a 649 amino acid protein that has been found to have low tissue specificity and is detected in nearly all tissues, making it the most abundant of the five helicases, yet little is known about its molecular functions in mammalian cells [61], [68], [69]. In vitro, RECQL1 has been shown to perform a variety of functions including unwinding a DNA structures such as G4s, catalyze branch migration of Holliday junctions and D-loops and promote single-strand DNA annealing (SSA) [61]. To accomplish these biological functions, RECQL1 has been shown to interact with PARP1, RPA, RAD51, TOP3α, EXO1, MSH2/6, MLH1-PMS2 and Ku70/80, implicating this protein mainly in DSB repair pathways yet its exact role remains enigmatic.

*BLM* was first cloned in 1995 and determined to be expressed in all tissues with notably strong expression in bone marrow and lymphoid tissues [69]. *BLM* was mapped to chromosome

15q26 and consists of multiple structural domains including a conserved catalytic helicase core domain, a N-terminal domain involved in regulation and oligomerization of BLM and a C-terminal region consisting of multiple protein interaction domains [70]. BLM has been implicated in DSB repair by a variety of functions. Namely, BLM has been shown to play a role in replication fork restart, DNA end resection, displacement of RAD51 from nucleoprotein filaments, disassembly of D-loops, and dissolution of Holliday junctions and other HR intermediates [71].

*WRN* was first cloned in 1996, guided by prior linkage analyses, and was mapped to the p-arm of chromosome 8 [72]. WRN is a 162 kDa protein that contains a central helicase core and is the only RecQ helicase to consist of an exonuclease domain which has been shown to cleave the 3′ ends of DNA (Figure 1.7A) [73]. WRN has been implicated in DSB repair, BER and the replication stress response by promoting replication fork reversal and restart. It has also been speculated that WRN can unwind certain secondary DNA structures [73]. For example, WRN has been shown to unwind cruciform structures formed by TA dinucleotide repeats that in the absence of WRN, result in replication fork stalling, cleavage by the MUS81 nuclease and massive chromosome shattering [28].

*RECQL5* was first cloned by Kitao *et al.* and was identified as a RecQ helicase based on homology with other characterized RecQ helicases [74]. In humans the gene is ubiquitously expressed in all tissues tested with notably strong expression in the testis and pancreas [74]. *RECQL5* was mapped to chromosome 17q25 and found to be alternatively spliced in 19 variant forms with three variant forms (α, β and γ) being the most predominant [62], [74]. The α and γ forms are less common variants that are truncated at the C-terminus and have only D1 and D2 helicase subdomains without the $Zn^{2+}$ binding domain that is essential for helicase activity [62].

Therefore, these truncated forms are deficient in helicase activity and only have strand annealing function [62]. The more common variant across all tissues, RECQL5β (referred to hereinafter as RECQL5), is a 120 kDa protein with 991 amino acids containing all three core helicase subdomains and an extended C-terminal that is different from other RecQ helicases and contains several regions essential for specific protein-protein interactions (Figure 1.7) [62], [75]. It remains unclear to what degree different isoforms of RECQL5 play a role in different cell types.

Crystal structures of RECQL5 have revealed D1 and D2 helicase subdomains that are highly similar to other RecQ helicases, whereas a helical hairpin motif in the $Zn^{2+}$ binding domain is significantly longer than that of any other RecQ helicase [66]. Additionally, the C-terminal of RECQL5 lacks a winged helix immediately following the $Zn^{2+}$ binding domain and instead has a positively charged alpha helix [66]. Both unique structures in the core catalytic unit are believed to confer selectivity in the DNA-binding capacity of RECQL5 compared to other RecQ helicases. Newman *et al.* showed that this region contributes to a higher specificity in RECQL5 for non-duplex DNA such as ssDNA, hairpin loops in dsDNA and forked DNA structures, all of which could occur as transcription intermediates [66].

Within the C-terminus of RECQL5 are two domains responsible for protein interactions [66]. The kinase-inducible domain interacting (KIX) domain and Set2-Rpb1 interacting (SRI) domain were isolated from full-length RECQL5 constructs and were shown to be required for the interaction between RECQL5 and RNA polymerase II (RNAPII) (Table 1.3). Using purified proteins, Hu *et al.* demonstrated that RECQL5 is capable of binding and inhibiting RAD51-mediated D-loop formation, an interaction discovered to require a motif between residues 652 and 725. Electron microscopy revealed RECQL5 can remove RAD51 from ssDNA in a reaction

dependent on ATP hydrolysis and the ssDNA-binding protein, RPA. Several other stimulatory interactions are summarized in Table 1.3 and are discussed in further detail below.

| Protein | Region | Function | Reference |
|---|---|---|---|
| FEN1 | ND | Stimulates FEN1 endonuclease activity | [76] |
| Mre11 | ND | Inhibits Mre11 activity | [77] |
| NBS1 | ND | ND | [77] |
| PCNA | 541-991 | Promotes conjugation of PCNA with SUMO2 | [78] |
| TOPO IIa | ND | Stimulates TOPOIIa decatenation activity | [79] |
| TOPO IIIa | ND | ND | [80] |
| RAD50 | ND | ND | [77] |
| RAD51 | 652-725 | Disrupts RAD51 nucleofilaments | [62] |
| RNAPI | ND | ND | [81] |
| RNAP II | KIX, SRI | Inhibits rate of RNAPII transcript elongation | [35], [82] |
| SWI/SNF complex | ND | ND | [83] |
| WRN | ND | Stimulates helicase activity of WRN | [75] |

**Table 1.3 Protein-protein interactions reported for RECQL5**

### 1.3.3 RECQL5 gene function

#### 1.3.3.1 Role of RECQL5 in double-stranded DNA break repair

Cells deficient in RECQL5 display a phenotype of genome instability and elevated CO products in the form of SCEs. In 2007, Hu *et al.* discovered that RECQL5 interacts with and disrupts RAD51 nucleofilaments similar to BLM and Sgs1 in yeast, a landmark finding that supported a model of HR where RAD51-dependant pathways are susceptible to CO products and RECQL5 and BLM are regulators of this pathway in humans [62]. However, the synergistic phenotype of genome instability in *RECQL5*$^{-/-}$ *BLM*$^{-/-}$ double knockouts was the first evidence that these genes may have non-overlapping roles as well. It was later shown *in vivo* that RECQL5 is essential for this disruptive interaction with RAD51 and its ability to form D-loops [84].

Bringing these observations together, Olsen *et al.* proposed a model of HR in which increased levels of RECQL5 reduce repair efficiency in the presence of a dsDNA donor molecule, whereas repair efficiency is significantly increased in the presence of a ssDNA donor [18]. This supports the notion that RAD51 is essential for strand invasion and by disrupting these nucleofilaments, RECQL5 is limiting the formation of D-loops and subsequent dHJ formation [85]. Given that RECQL5 gene amplification and deficiency have both been associated with cancer predisposition, it is possible RECQL5 is required at a suitable level to permit sufficient RAD51-mediated strand invasion for HR repair without an excess of D-loop formation biasing outcomes towards dHJ and CO products [9], [18], [86].

#### 1.3.3.2 Role of RECQL5 in replication stress

During replication, the replisome encounters many stressors that may hinder faithful chromosome duplication [87]. This replication stress may slow or even stall the replication fork and activate certain pleiotropic DNA repair genes to form intermediate molecules in an effort to prevent further damage from occurring [87]. These replication stress pathways serve to resolve these substructures of DNA that may arise during replication fork stalling [88]. As a typical by-product of replication fork stalling, the accumulation of exposed ssDNA occurs as RPA is depleted across multiple stalled forks [16]. This accumulation and subsequent depletion of free RPA serves to activate ATR kinase and the replication stress response which serves to recruit DNA repair machinery and stabilize the stalled fork before too much ssDNA is exposed (Figure 1.8)[88]. Most importantly, it serves to prevent new origins from firing and further RPA depletion and associated ssDNA exposure from leading to global replication fork stalling and

replication catastrophe [16]. Forks that fail to restart may lead to replication fork collapse and DSBs, activating canonical DSBR pathways [16].

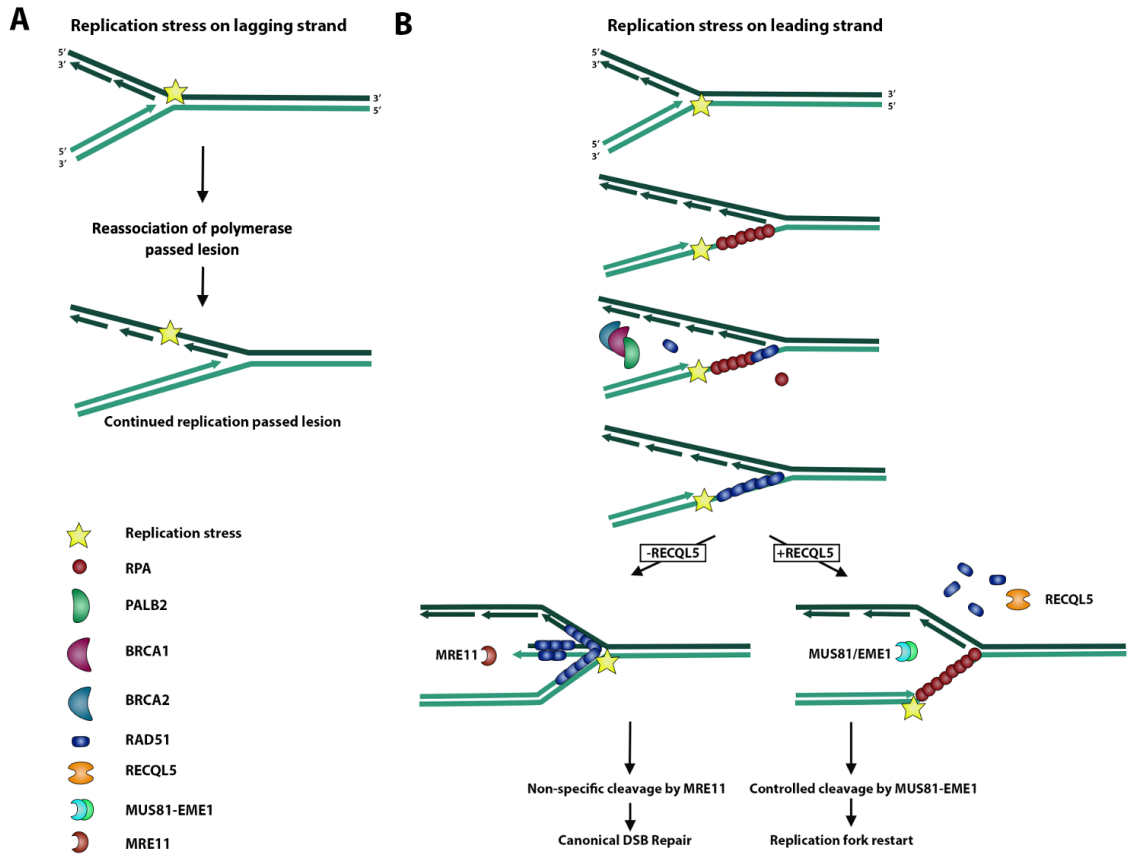



**Figure 1.8 Role of RECQL5 in replication stress response.**
**Replisome can encounter a replication stress-inducing lesion on either a lagging or leading strand. (A) When the replisome encounters a lesion on a lagging strand, DNA polymerase is able to bypass the lesion by dissociating from the Okazaki fragment and reassociating to form a new fragment ahead of the lesion. (B) Replisome encountering a lesion on a leading strand may lead to replication fork uncoupling, whereby the polymerase is stalled and dissociated from the replication helicase, which continues unwinding DNA and exposing ssDNA. Exposed ssDNA and simultaneous depletion of free RPA serve to activate ATR signaling of the replication stress response and recruit BRCA1/2 and PALB2 to begin exchange of RPA for RAD51. RAD51 actively promotes reversal of the replication fork and formation of a regressed arm whereby newly synthesized DNA strands anneal to each other, allowing for non-specific cleavage by the MRE11 endonuclease and subsequent canonical DSB repair. In the presence of RECQL5, RAD51 is removed from ssDNA, and the MUS81/EME1 endonuclease complex is recruited to allow controlled cleavage and replication fork restart.**

RECQL5 has been implicated in this stress response because of the finding that cells deficient in RECQL5 are hypersensitive to the Topoisomerase I inhibitor, camptothecin, which leads to impaired replication and experience an exaggerated phenotype of genome instability

[84]. Additionally, RECQL5 associates with the replisome factor, PCNA, and persists at sites of stalled replication forks [81]. This involvement of RECQL5 in resolving replication stress could in part be attributed to its ability to stimulate the endonuclease, FEN1, and coordinate the cleavage events needed for replication fork restart [76].

The interaction of RECQL5 with RAD51 also serves an important role in processing stalled replication forks as RAD51 has a pleiotropic function in both HR and replication stress [30]. Upon replication stress, stalled replication forks accumulate ssDNA and RAD51 stabilizes this DNA with the support of BRCA2, similar to how RAD51 binds ssDNA on the resected ends of a DSB in DSBR (Figure 1.8)[87]. Electron microscopy studies were performed to study replication fork reversal in the presence and absence of the stabilizing filament, RAD51, its loading partner, BRCA2, and the processing endonuclease, MRE11 (Figure 1.8). These studies revealed that RAD51 independently promotes replication fork reversal and that RAD51 and BRCA2 together protect against reversed fork degradation by MRE11 (Figure 1.8)[27]. Despite the protective role of RAD51 against MRE11-mediated reversed fork cleavage, overexpression of RAD51 created a phenotype of excessive fork stabilization and impaired replication fork restart, suggesting an appropriate balance of RAD51 stabilized replication forks is sufficient for replication restart [27]. Considering RECQL5 removes RAD51 filaments in DSBR, Di Marco *et al.* examined the role of RECQL5 in replication stress and showed that in addition to removing RAD51 filaments from reversed replication forks, RECQL5 recruits and stimulates the MUS81-EME1 endonuclease complex to promote cleavage and replication restart of difficult to replicate regions (Figure 1.8)[30]. Taken together, these findings support a model of RECQL5 in balancing the intermediate structures in DSBR and the replication stress response.

#### 1.3.3.3 Role of RECQL5 in transcription and regulating transcription-replication stress

A protein-protein interaction unique to RECQL5 and believed to be critical to its function is that between RECQL5 and the RNAPII complex [35], [82]. Cells deficient in RECQL5 display elevated levels of transcription, increased RNAPII-bound chromatin and increased DSBs associated with transcribed loci, suggesting that RECQL5 has more of an inhibitory role in this interaction [89]–[91]. Furthermore, RECQL5 loss increased the ratio of RNAPII associated with promoter-proximal regions relative to the gene body of a subset of over 5000 genes examined, whereas overexpression reversed this ratio [31]. However, there was no change in overall mRNA produced, suggesting that transcription elongation rate was affected opposed to transcription initiation [31]. For 80% of the transcribed genes in a genome wide assay, Saponaro *et al.* created an *in vivo* model to synchronize transcript cycles and measure elongation rate of individual genes and showed that depletion of RECQL5 significantly increased this value whereas overexpression reduced it. In the absence of RECQL5, sites of elevated transcript elongation were enriched for DSB breaks [31]. Together, these findings suggest that RECQL5 is an inhibitory RNAPII elongation factor and that deficiencies in RECQL5 lead to increased rates of RNAPII-mediated transcript elongation, higher levels of RNAPII pausing or arrest and overall transcription-induced genome instability. This form of transcription-associated genome instability appears to also be associated with replication since Li *et al.* showed many of the DSBs in this model accumulate during S-phase and associate with RNAPII transcribed loci. This phenotype was relieved in the presence of a transcription inhibitor further supporting the association of replication and transcription machinery driving DSBs and genome instability [90]. Together these findings support a model of transcription-associated genome instability where RECQL5 is limiting the

collision of transcription and replication machinery by slowing the elongation rate of transcription.

Another source of transcription-associated genome instability is the formation of R-loop structures at sites of active transcription during replication. The formation of ssDNA from negative supercoiling behind transcription allows RNA invasion forming a R-loop, making it difficult for replication machinery to continue [92]. RECQL5-bound RNAPII was shown to stimulate conjugation of SUMO2 to the replicative factor, PCNA, another one of its binding partners [92]. Conjugated SUMO2-PCNA is capable of interacting with the histone chaperone protein, CAF1, and depositing repressing histone marks in a CAF1-dependant manner therefore reducing chromatin accessibility and effectively dislodging RNAPII from DNA [92]. This was confirmed by showing that cells deficient in RECQL5 are TRC and DSB prone and that overexpressing SUMO2-PCNA or CAF1 rescued this phenotype [92]. Additionally, RECQL5 was shown to mediate replication fork restart at sites of stalled replication forks near R-loops by limiting RAD51-mediated replication fork reversal and recruiting the MUS81-EME1 endonuclease complex for appropriate processing of stalled replication [26]. These findings support a role for RECQL5 in limiting TRCs. There is evidence it does so both proactively by either inhibiting transcript elongation near sites of replication or remodeling chromatin to dislodge RNAPII from DNA and retroactively by limiting RAD51-mediated replication fork reversal and promoting MUS81-EME1 cleavage and replication fork restart [26], [30], [92].

It is clear that RECQL5 serves as an important regulator of DNA repair intermediate structures that may arise during DNA damage, replication stress and transcriptional stress. This essential regulatory role of RECQL5 is further highlighted by the observed elevated RECQL5 expression and gene amplification in urothelial carcinoma of the bladder and breast cancers [67],

[93], [94]. However, the nature of DNA lesions that are preferentially repaired using RECQL5, the choice of RECQL5 over alternative RecQ helicases for repair of various DNA lesions and the role of expression levels in such choices remain to be elucidated. The finding of significant cancer predisposition in mice models deficient in RECQL5, support that perturbation of RECQL5 levels in either direction can contribute to oncogenesis [62]. Yet it remains unclear to what degree RECQL5 is the only factor regulating these processes and how RECQL5 contributes to oncogenesis or provides a backup function to other essential DNA repair genes. There is evidence of some overlapping function, specifically with other RecQ helicases. For example, in comparison to BLM, RECQL5 shares a similar phenotype of genome instability, but there is sufficient evidence that RECQL5 suppresses SCEs and DSBs even in the presence of BLM [42]. Shared protein-protein interactions between RECQL5 and BLM, such as with RAD51, likely correspond to overlapping functions whereas interactions unique to RECQL5 such as that with RNAPII may provide useful insight into the unique functions of RECQL5 [85], [90].

It is well documented that helicases play essential roles in HR. However, the genetic pleiotropy that is prevalent among these helicases poses a challenge to the identification of their roles in genome stability. For example, the BLM helicase has been shown to a play a role in DNA end resection, RAD51 displacement, D-loop disassembly, and double Holliday junction dissolution. Yet these functions are both pro- and anti-recombinogenic functions, making it unclear to what extent BLM is acting redundantly with other helicases, or whether these functions are regulated for specific purposes. Answers to these questions will give us a better understanding of genome stability and possible ways to manipulate helicase activities and achieve useful therapeutic outcomes in cancer.

## 1.4 Research scope, hypothesis and objective

The larger body of research on other RecQ helicases supports further study of RECQL5 in parallel with other RecQ helicases given their anti-recombinogenic activity in SCE formation. Given that SCEs are a useful indicator of genome instability and genome instability is a driver of oncogenic transformation, it will be of interest to map the location of such events as it was shown that BLM preferentially prevents SCEs near transcribed genes and G-quadruplex motifs [25]. There may be specific motifs or substructures of DNA that RECQL5 preferentially localizes to and protects against genome instability. Therefore, our hypothesis is that different types of DNA lesions and structures may require different RecQ helicases to be resolved.

The primary objective of my thesis is specifically to uncover unique roles of RECQL5 and other RecQ helicases in SCE formation and the maintenance of genome stability using novel bioinformatic and wet-lab techniques. In Chapter 2, I discuss new ways to perform DNA repair studies using Strand-seq and improvements compared to previous methods. In Chapter 3, I discuss the design, implementation, and performance of SV bioinformatic callers using Strand-seq data. In Chapter 4, I perform exhaustive enrichment analysis of the SCEs in RecQ helicase deficient cell lines and draw conclusions about the roles of different RecQ helicases in DSB repair. In Chapter 5, I discuss conclusions, limitations, and future directions of my work.

A better understanding of the role of RecQ helicases in genome stability will yield novel information about molecules and pathways involved in recombination and SCE formation. Such information is essential to elucidate currently poorly understood medical conditions and inform therapeutic strategies in cancer.

# Chapter 2: New OP-Strand-seq pipeline for studying DNA repair

## 2.1 Introduction

The main goal of this chapter is to introduce novel approaches for performing DNA repair studies using Strand-seq. Standardized methods to study DNA repair is a prerequisite to elucidate the role of DNA repair in cancer and aging. Three implementations discussed in this chapter are the generation of CRISPR-Cas9 haploid KO lines of RecQ helicases, the development and use of the "One-pot" Strand-seq protocol and the development and use of a Strand-seq library quality classification system.

### 2.1.1 Original Strand-seq protocol

Single-cell template strand sequencing (Strand-seq) is a sequencing technique developed in 2012 for the selective sequencing of a dividing daughter cell's parental template strands used during DNA replication [56], [57]. As discussed in Section 1.2.5.5, this method relies on the directionality of DNA based on its 5'-3' orientation to preserve read directionality and permit the detection of orientation-dependent structural variants such as inversions and translocations that would otherwise be very challenging to detect using other single cell sequencing approaches [56], [57]. This approach was first developed using diploid cells [56], [57].

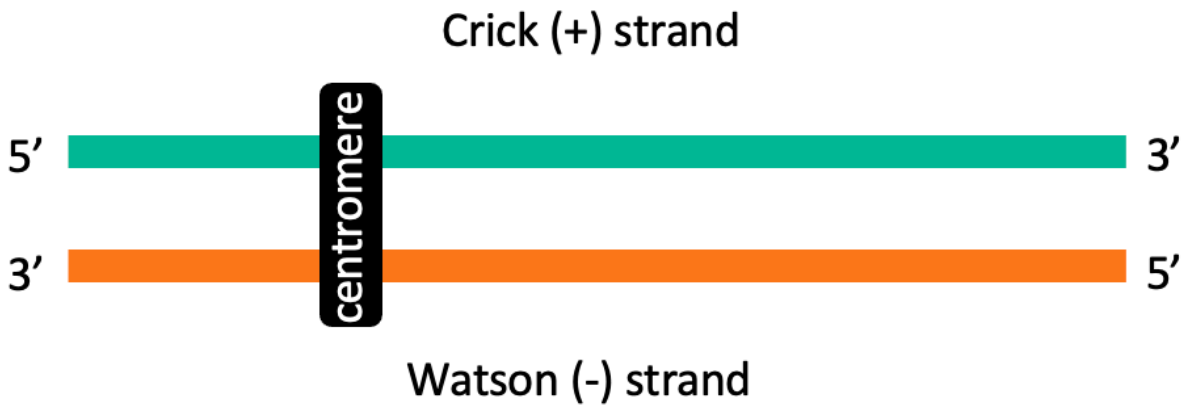



**Figure 2.1 Standardized definition of Watson and Crick strands.**

Strand-seq exploits the semi-conservative nature of DNA replication to identify parental DNA template strands in daughter cells following DNA replication and cell division [56], [57]. In genetics, the two strands making up the DNA double helix are commonly referred to as the “Watson strand” and “Crick strand” [95]. However, labeling of these strands relative to the reference genome can be done in two ways according to the position of specific genomic landmarks such centromeres. To date, no universal nomenclature to define each strand exists and this has led to multiple definitions being used [95]. For example, Cartwright and Graur, 2011 defined the Watson strand as the strand with its 5’-end at the short-arm telomere and the Crick strand as its complement [56], [95]. Alternatively, Falconer et al., 2010 defined the Crick strand as the strand with its 5’-end at the short-arm telomere and the Watson strand as its complement (Figure 2.1) [56], [95]. The Falconer et al., 2010 definition is the one used in this thesis (Figure 2.1). The Crick strand has also been synonymously referred to as the plus strand or the 5’-3’ strand of the refence assembly whereas the Watson is also considered the minus strand or 3’-5’ strand of the reference assembly (Figure 2.1) [56], [95].

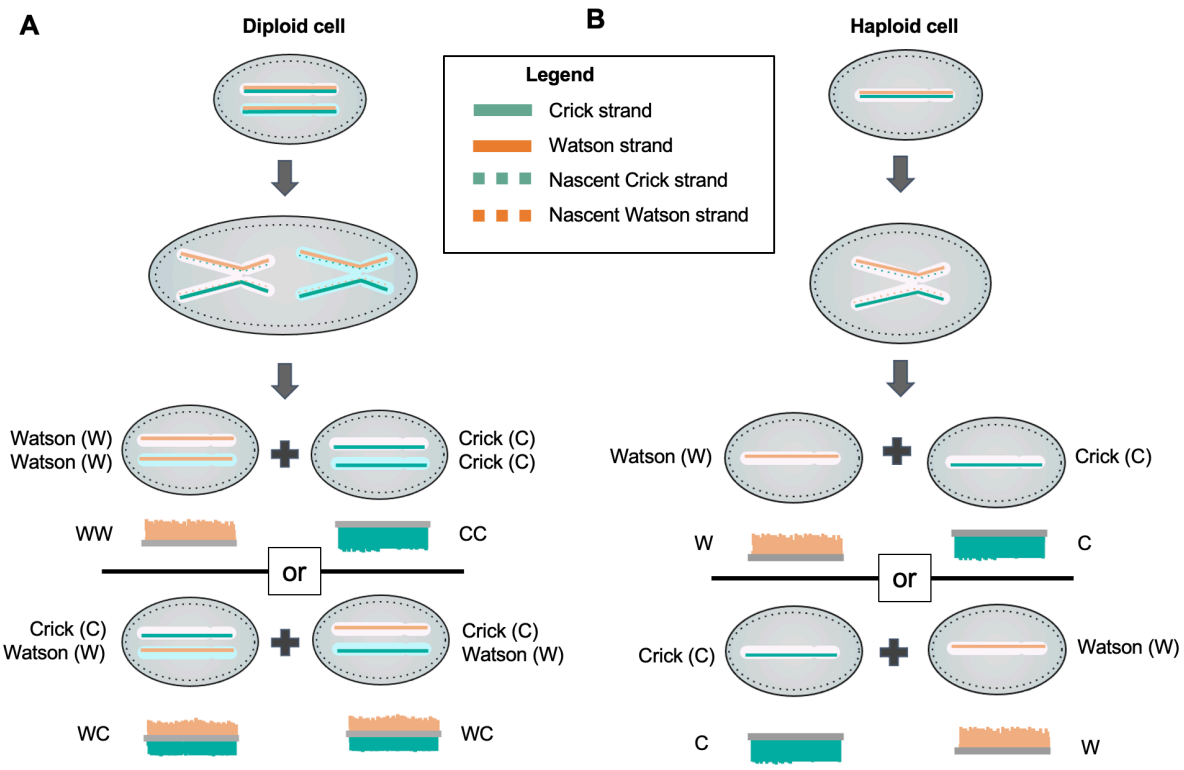



**Figure 2.2 Strand inheritance patterns and associated Strand-seq ideograms.**
**After one round of DNA replication in the presence of BrdU, each chromosome is hemi-substituted with BrdU. After cell division, each daughter cells can inherit either template strand for each sister chromatid upon random segregation of sister chromatids. Ideograms for one chromosome in a diploid cell (A) and a haploid cell are shown (B). Dotted lines represent strands of DNA with BrdU incorporation.**

In DNA replication, each parental template strand serves as a template for newly synthesized nascent DNA such that dividing diploid cell replicates both Watson (W; minus or 3'-5' strand of reference assembly) and Crick (C, plus or 5'-3' strand) strands of each homologous chromosome in a semi-conservative fashion (Figure 2.2A) [56], [57]. After a cell divides, the two daughter cells can inherit opposing template strands for each homolog (e.g. all W reads for one homolog and all C reads for the other homolog; W-C) or the same template strand for each homolog (e.g. all W reads for both homologs; W-W, or all C reads for both homologs; C-C), generating three possible patterns of template strand inheritance for the two homologous chromosomes in a given diploid daughter cell (W-W, C-C, W-C; Figure 2.2A). In a haploid cell, strand inheritance of one homolog can only produce one of two possible DNA

template strand inheritance patterns for each chromosome in a daughter cell (W or C, Figure 2.2B).

Strand-seq begins with the incorporation of the thymidine analog, BrdU, into the newly synthesized strand to allow for the distinction between template and nascent DNA strands (Figure 2.2) [56], [57]. DNA fragments with BrdU can be selectively degraded by treatment with Hoechst and UV irradiation before PCR amplification (Figure 2.3 ) [56], [57]. PCR amplification after degradation of nascent strand DNA fragments allows for the selective amplification of template strand reads (Figure 2.3 ) [56], [57]. Pools of Strand-seq libraries can then be loaded onto any Illumina sequencing instrument for paired end Illumina whole-genome sequencing (WGS) [56], [57]. Illumina WGS of Strand-seq libraries generates directional libraries with reads mapping to the reference genome in the orientation of the native parental DNA template strands (Figure 2.3 ) [56], [57]. The original protocol can construct several dozen Strand-seq cells at a time using a liquid-handling platform for automation [56], [57].

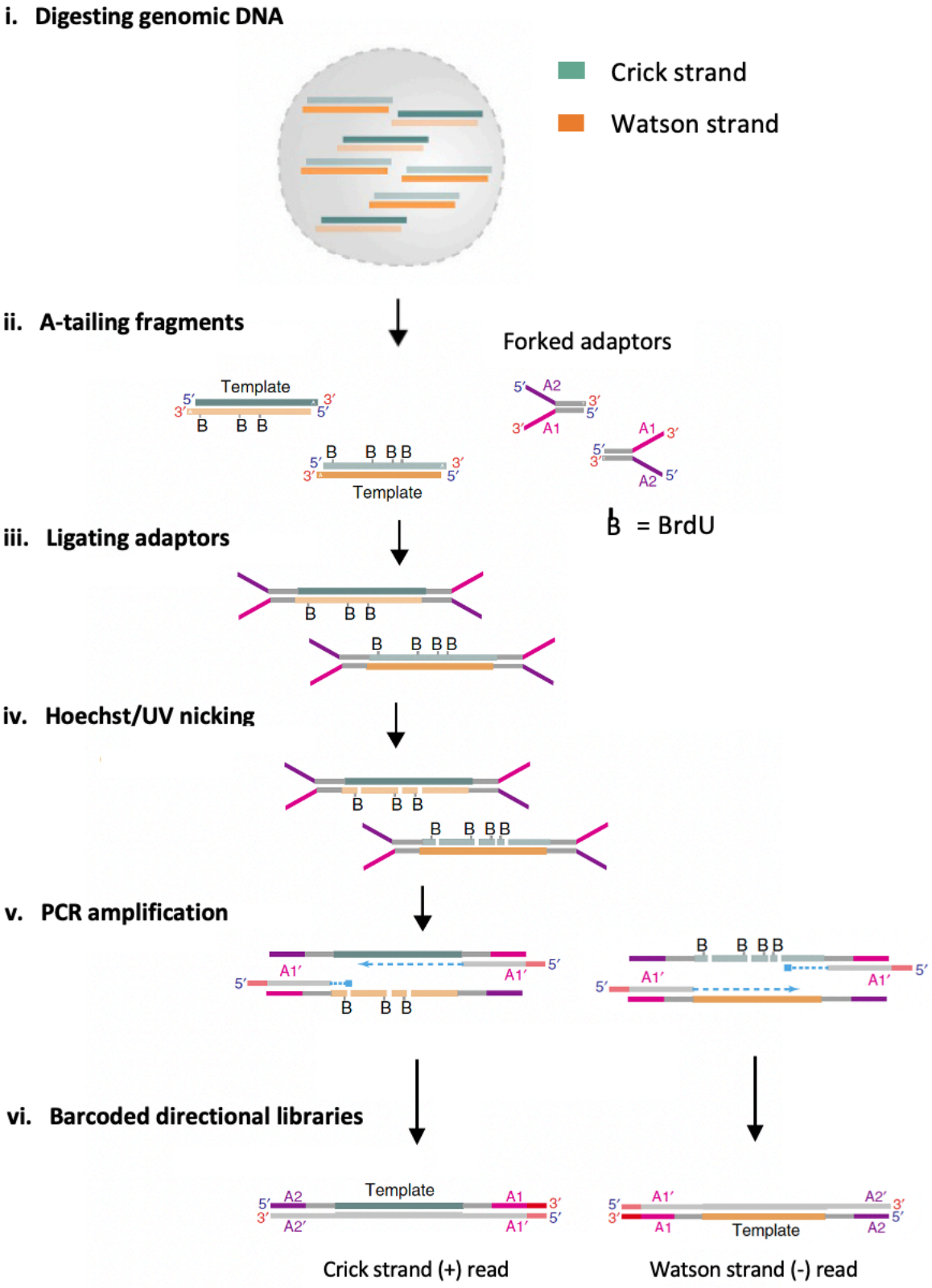

i. Digesting genomic DNA
Crick strand
Watson strand
ii. A-tailing fragments
Forked adaptors
Template
5′
3′
B B B
B B BB
Template
5′ A2 3′
3′ A1 5′
A1 3′
5′ 3′
A2 5′
B = BrdU
iii. Ligating adaptors
B B B
B B BB
iv. Hoechst/UV nicking
B B B
B B BB
v. PCR amplification
A1′
A1′
5′
5′
B B B
B B BB
A1′
A1′
5′
5′
vi. Barcoded directional libraries
Template
5′ A2 A1 3′
3′ A2′ A1′ 5′
5′ A1′ A2′ 3′
3′ A1 Template A2 5′
Crick strand (+) read
Watson strand (-) read

**Figure 2.3 Principle of single-cell DNA template strand sequencing.**
**(i) Chromosomes with BrdU substituted DNA are fragmented. (ii) DNA fragments are A-tailed. (iii) A-tailed DNA fragments are ligated to universal forked adaptors. (iv) Hoechst and UV photolysis create nicks at BrdU sites (v) Nicks prevent PCR amplification of nascent strands but allow selective amplification of the original intact template strand. (vi) The resulting libraries are directional, containing the template strand in its original genomic orientation in all amplified fragments. The hexamer barcode (red) is introduced by the PE 2.0 primer during PCR amplification. The directional library contains the A2 adaptor at the 5′-end and the A1 adaptor at the 3′-end of the template strand. Multiple single-cell libraries are pooled and sequenced on an Illumina platform. Figure adapted from Sanders et al. (2017).**

### 2.1.2 Applications of original Strand-seq method for studying DNA repair

One of the unique applications of Strand-seq is the ability to putatively identify SCEs and complex SVs by pinpointing changes in template strand inheritance (herein referred to as strand state switches) [25], [49], [56]–[58], [96]. After Illumina sequencing is performed on Strand-seq libraries to generate short-read sequencing data for each cell, reads are assigned a Watson or Crick designation based on their mapping orientation during alignment to the reference genome (Figure 2.4) [56], [57]. Duplicate reads are removed and strand state switches are pinpointed as changes in strand state genotype within a chromosome [25], [49], [56]–[58], [96]. Approximate coordinates and strand state genotype information across the breakpoint are identified for downstream analysis (Figure 2.4). The resolution of how finely these events can be mapped to the genome is proportional to the fraction of genomic DNA that is captured in a single cell Strand-seq library and the depth of the subsequent sequencing.

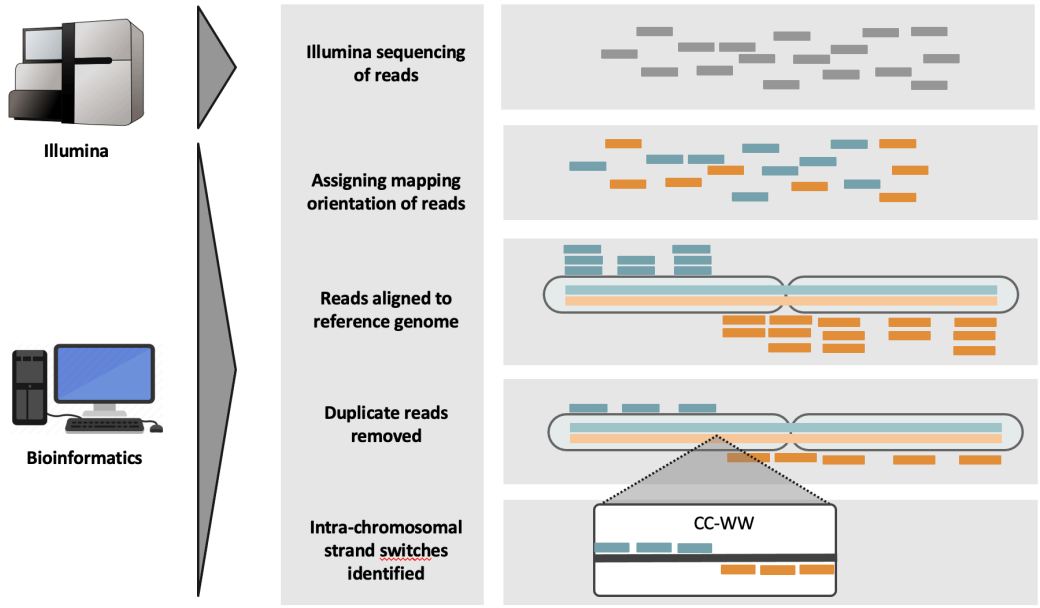



**Figure 2.4 Principle of identifying strand state switches in Strand-seq data.**
**Illumina WGS is performed on Strand-seq libraries to generate short-read sequencing data for each cell. Reads from each cell are assigned a Watson (orange) or Crick (teal) designation based on their mapping orientation during alignment to the reference genome. Duplicate reads are removed, and intra-chromosomal strand state switches are pinpointed and coordinates for breakpoints are identified for downstream analysis.**

Strand state switches suggest a SCE or a SV has occurred [25], [59], [97]. Different classes of SVs can be identified by integrating read depth, strand state switch genotypes and haplotype information present in Strand-seq data [25], [59], [97]. However, such SVs need to be distinguished from error-free SCE events and this can be a challenge considering the common features among both SCEs and SVs (Figure 2.5) [59]. For example, an SCE can be identified by collecting breakpoints that are not recurring in multiple libraries, are not associated with changes in read count and only affect one homolog (e.g. WW-WC or WC-CC; Figure 2.5) [25], [59], [97]. An inversion can be identified by collecting two neighboring breakpoints that are recurring in multiple libraries, are not associated with read count changes and can affect one or both homologs depending on if its heterozygous or homozygous and haploid or diploid (e.g. diploid homozygous inversion; CC-WW-CC or diploid heterozygous inversion; WC-WW-WC or

haploid inversion; C-W-C; Figure 2.5) [25], [59], [97]. Custom bioinformatic approaches for identifying different classes of SVs in Strand-seq library is thoroughly discussed in Chapter 3.

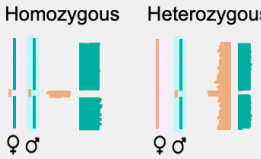
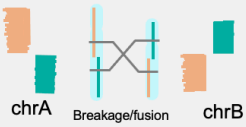
| | Sister chromatid exchange event | Inversion | Duplication | Translocation |
|---|---|---|---|---|
| **Diploid cell** |  |   |  |   |
| **Haploid cell** |  |  |  |   |
| **Number of breakpoints** | 1 breakpoint | 2 breakpoints on same chromosome | 1 or 2 breakpoint(s) | 2 breakpoints on different chromosomes |
| **Copy-number changes** | No | No | Yes | No |
| **Allelic fraction** | 1 cell | Multiple cells | Multiple cells | Multiple cells |
| **Genotype of breakpoint(s)** | WW-WC (diploid), W-C (haploid) | CC-WW-CC (diploid, homozygous), WC-WW-WC (diploid, heterozygous), C-W-C (haploid) | WWC-WC (diploid), C-WC (haploid) | WW-WC + CC-WC (diploid) WC + WC (haploid) |

**Figure 2.5 Features of different structural variants in haploid and diploid Strand-seq libraries.**
**Table providing an example of the Strand-seq chromosome ideograms, number of breakpoints, copy-number changes, allelic fraction and example genotype of breakpoints associated with an SCE and three classes of SVs in diploid and haploid cells.**

The genomic signatures of recombinatory events such as SCEs and SVs have been shown to elucidate specific helicase functions. For example, deficiency in the BLM helicase has been implicated in toxic unrestrained recombination resulting in up to 10 times more SCEs than a healthy cell [25]. Mapping SCEs from Bloom Syndrome patient cells to the genome revealed enrichment near G4 motifs in actively transcribed genes, suggesting the BLM helicase is responsible for resolving G4 motifs that likely arise during transcription replication conflicts [25]. Strand-seq data can be used to discover unique and essential roles of DNA helicases in DSB repair and reveal how faulty DSB repair gives rise to genomic instability.

### 2.1.3 Limitations of original Strand-seq method for studying DNA repair

One of the primary limitations of the original Strand-seq method is the cost and throughput. Each single cell library cost ~US$13 with a throughput of 96 cells per experiment, each of which takes four days to complete [57], [60]. Additionally, there is high variability in the quality of Strand-seq libraries resulting in poor Strand-seq library characteristics [57], [60]. The characteristics of poor Strand-seq libraries include unevenness and sparsity in sequencing coverage as well as a high proportion of non-directional reads (herein referred to as background reads) that do not retain native directionality of the template strand they originated from [57], [59], [60]. All three of these characteristics worsen the resolution in which breakpoints for putative SVs can be mapped to the genome of a Strand-seq library [57], [59], [60]. Another limitation lies with the bioinformatic approaches for calling different SV classes in Strand-seq libraries. Mainly, there are few Strand-seq-specific bioinformatic tools for calling SVs and like most SV callers, they suffer from high false positivity rate due to the challenges associated with calling SVs, which will be discussed in Section 3.1.3 [59]. Strand-seq libraries may experience a higher proportion of background reads and this confuses SV callers resulting in either false positive calls, false negative calls or poor breakpoint resolution [59]. Therefore, I focus on addressing these limitations in our novel methods for performing DNA repair studies using Strand-seq.

## 2.2 Methods

### 2.2.1 Knockout model generation in haploid cell line using CRISPR-Cas9

To improve the resolution of breakpoint coordinates, I chose the haploid cell line, KBM7, for the generation of RecQ knockout lines. Haploid cells offer three advantages over diploid cells for the detection of SCEs and SVs using Strand-seq in Chapter 3. First, having one haplotype improves the resolution at which SVs can be mapped to the genome (Figure 2.6) [59]. In a haploid cell, strand inheritance of one homolog can only produce one of two possible DNA template strand inheritance patterns for each chromosome in a daughter cell (W or C; Figure 2.6B). In contrast, diploid cells will show three strand inheritance patterns (WW, CC, and WC; Figure 2.6A). This makes intra-chromosomal changes in strand state less ambiguous in haploid cells because the breakpoint can theoretically be placed directly between a Crick and a Watson read to represent a W-C transition (Figure 2.6B). In a diploid cell there is uncertainty of whether a Watson read, for example, at the breakpoint is part of the WC segment or the WW segment for a WC-WW transition, resulting in a breakpoint confidence interval (CI) of at least one read plus gaps on either side of the W read (Figure 2.6A). Secondly, it is easier to generate knockout lines in a cell line with only one haplotype because only one allele must be altered. Finally, for the same genome coverage of a single cell Strand-seq library, sequencing costs are reduced two-fold.

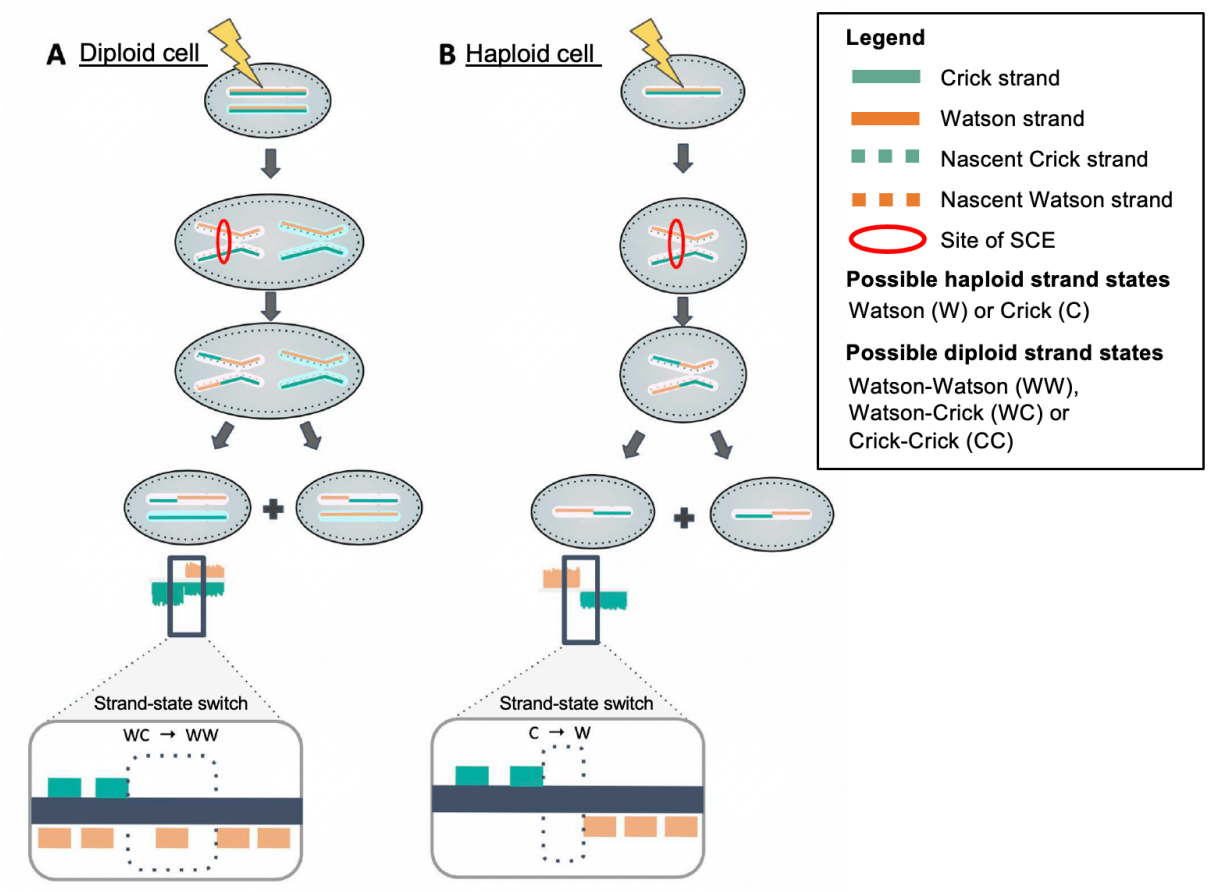



**Figure 2.6 Strand inheritance patterns and associated Strand-seq ideograms for a sister chromatid exchange (SCE).**
**(A) Strand inheritance patterns and associated Strand-seq ideograms in a diploid cell and (B) a haploid cell. Solid orange and teal lines represent normal DNA strand and dotted orange, and teal lines represent BrdU-substituted DNA strand. Circle highlights point of exchange between sister chromatids.**

Haploid cells such as the KBM7 haploid line also have one major limitation in that they are prone to endoreduplication in culture to become diploid, resulting in mixed populations of haploid and diploid cells [98]. Without intervention, haploid cell cultures often become fully diploid within 10-20 passages [98]. To counter this tendency, I used an approach described by Beigl et al. for assessing the ploidy of cell cultures [98]. For this purpose, the DNA content of nuclei from selected cell cultures was compared to the DNA content of parental haploid lines by

FACS (Figure 2.7). Once a cell culture was discovered to contain diploid cells, smaller cells were sorted by FACS to enrich for haploid cells (Figure 2.7B) [98].

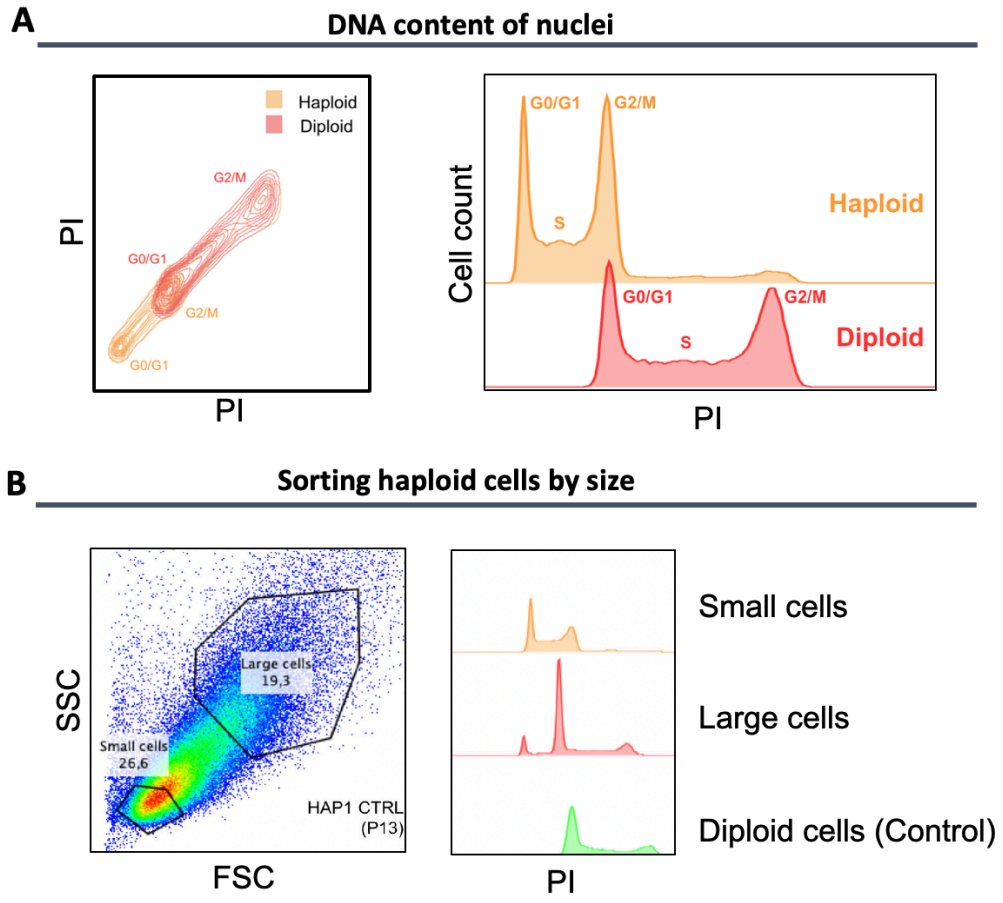



**Figure 2.7 Distinguishing haploid and diploid cells.**
**(A) Haploid and diploid KBM7 nuclei can be distinguished using DNA staining with using Propidium Iodide (PI) using flow cytometry. (B) Haploid and diploid KBM7 cells can be distinguished and enriched by FACS sorting based on light scatter properties related to size. Reproduced from Beigl et al., 2020.**

### 2.2.1.1 Cell culture

Cell lines were grown in Iscove's Modified Dulbecco's Medium (IMDM, StemCell, Vancouver, Canada) with 10 % Fetal Bovine Serum (FBS, Gibco, ThermoFisher, Canada) and 1% penicillin-streptomycin (Gibco). The doubling time for KBM7 cells is 22 hours [99]. Cultures were passed every other day. Once a month, cell lines were sorted using the protocol described in Figure 2.7 to enrich for haploid cells.

#### 2.2.1.2 CRISPR-Cas9 guide RNA design and electroporation

I used the Alt-R® CRISPR-Cas9 system (IDT, Coralville, Iowa 52241 USA) to generate knockout clones for *RECQL1, WRN, BLM* and *RECQL5* helicase as well as double knockout lines for *BLM/RECQL5* in the haploid KBM7 cell line. I designed multiple CRISPR RNAs (crRNA) to target either exon 1 or exon 2 for each gene using the built-in CRISPR RNA design tool from IDT to maximize predicted on-target efficiency and minimize predicted off-target efficiency and ordered these crRNAs from IDT (Table 2.1).

| Gene | Chromosome | Strand | Sequence | PAM | On-Target Score | Off-Target Score |
|---|---|---|---|---|---|---|
| RECQL1 | 12 | - | TTTAGAGGATTCTGATGCCG | GGG | 55 | 44 |
| BLM | 15 | - | GTTGGGTAGAGGTTCACTGA | AGG | 70 | 69 |
| BLM | 15 | + | AATCGGAATAGGCAAGCTTC | CGG | 67 | 82 |
| BLM | 15 | - | GTTGGGTAGAGGTTCACTGA | AGG | 70 | 69 |
| WRN | 8 | - | CAAGCAACATTTTTAATCCC | TGG | 63 | 31 |
| RECQL4 | 8 | - | AAGAGTCCACAGTCTACGCC | AGG | 62 | 79 |
| RECQL5 | 17 | + | ATGGTCGCACTCTCCTGTAA | AGG | 70 | 72 |
| RECQL5 | 17 | - | CTCTTTTAAGACGCCTTTAC | AGG | 61 | 68 |

**Table 2.1 CRISPR-Cas9 gRNA sequences designed for RecQ helicases**
**Guide sequences were designed using the built-in CRISPR-Cas9 gRNA design tool from IDT.**

Next, I followed the Alt-R® CRISPR-Cas9 system RNP Electroporation and Neon Transfection protocol from IDT. Two crRNAs were hybridized with a fluorescently labelled tracrRNA-ATTO (IDT) and assembled into ribonucleoprotein complexes with the Cas9 protein (IDT; Figure 2.8). CRISPR-Cas9 ribonucleoprotein complexes were electroporated into cells using the recommended settings: 1600 V, 10ms pulse width, 3 pulses (Figure 2.8). After 24 hours single cells, positive for tracrRNA fluorescence, were sorted by FACS into individual

wells of 96 well plates (Figure 2.8). After 5 days plates were inspected, and media was changed. Colonies were grown up to allow for KO screening by Sanger sequencing (Figure 2.8).

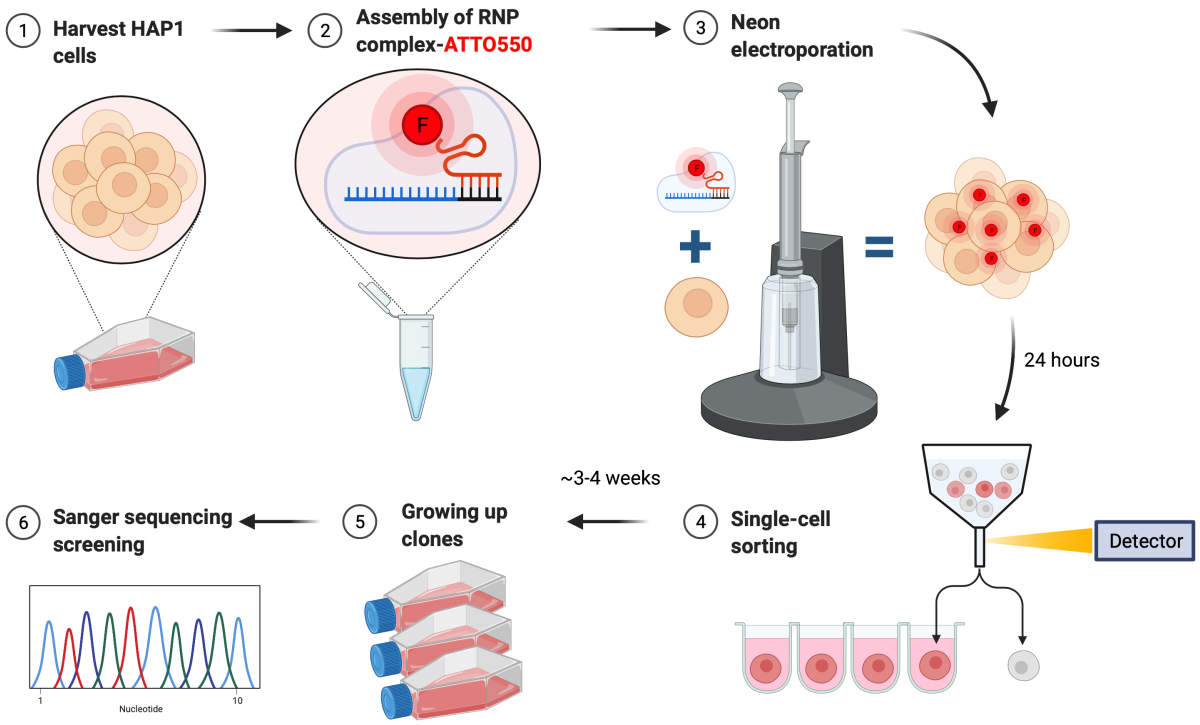



**Figure 2.8 Protocol for generating KBM7 CRISPR-Cas9 knockout cell lines. Generated using BioRender.**

### 2.2.1.3 Validation of CRISPR-Cas9 KOs

DNA from growing colonies was isolated and segments flanking the gRNA sequences were amplified by PCR for Sanger sequencing to identify and characterize frameshifting mutations (Figure 2.8). Sequence information for PCR primers is shown in Appendix Table 1. I collected Sanger sequencing data from control cells and used the tool, ICE (Synthego, Redwood City, CA), to characterize frameshifting mutations. ICE uses Sanger sequencing chromatograms from an edited sample and a control sample to identify frameshifting insertions or deletions < 21

bp or large insertions or deletions greater than 21 bp. This procedure was used to isolate one clone for *RECQL1* and three clones for *BLM, WRN, RECQL5* and *BLM/RECQL5* double KO cells (Figure 2.9). All *WRN* KO clones were found to consist of entirely diploid cell cultures due to the original ancestral KO cell after cloning being diploid. Therefore, *WRN* KO clones could not be enriched for haploid cells because there were none to begin with. *BLM, RECQL5* and *BLM/RECQL5* clones were functionally validated using the differential cytogenetic staining assay for the detection of SCEs in metaphase spreads discussed in Section 1.2.4.3. Ten to twenty metaphase spreads for each KO clone were analyzed for SCE staining (Figure A2.1). This assay revealed significant increases in SCE frequency upon knockout of *BLM, RECQL5* and *BLM/RECQL5* in comparison to our control WT cell line as would be expected for these KO phenotypes (Figure A2.1).

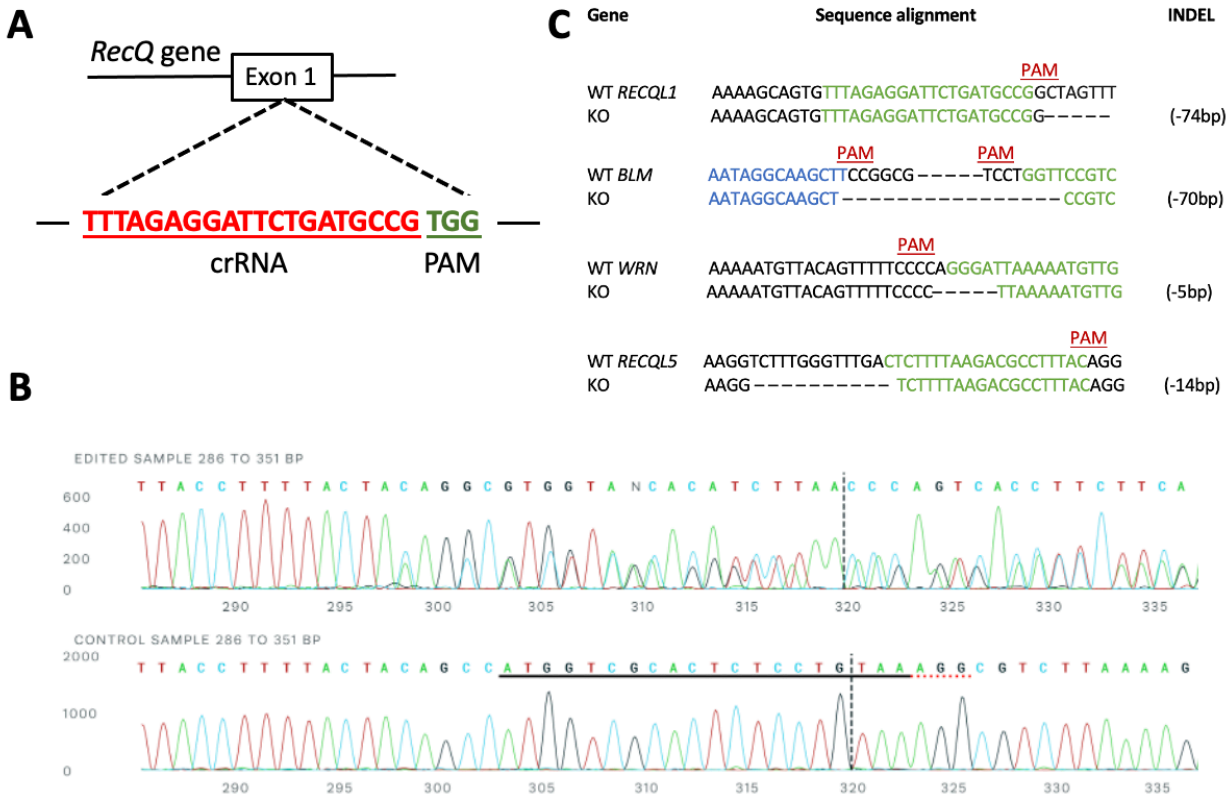



**Figure 2.9 Screening KBM7 CRISPR-Cas9 knockout cell lines.**

**(A) Each crRNA was designed to target exon 1 of each RecQ helicase gene. (B) Sanger sequence data from edited samples was compared against non-edited samples using the mutation characterising tool, ICE (Synthego, Redwood City, CA) (C) Examples of frameshifting mutations for each RecQ helicase.**

### 2.2.2 Construction of "one-pot" Strand-seq libraries

As previously discussed in Section 2.1.2, Strand-seq libraries provide directional genomic information that is essential for assembling haplotype-resolved genomes and comprehensive SV calling. However, the library preparation protocol described in Sanders et al., 2017 is costly, slow and suffers from low genomic coverage. The protocol typically requires up to 4 days to prepare 96 libraries at a cost of ~$1,300 USD (~$13 USD/cell) capturing at most, 5% of the genome in a single cell [57]. I developed a modified version of the original version of the Strand-seq protocol, known as "one-pot" (OP)-Strand-seq, to improve the cost efficiency, quality and throughput of Strand-seq [60].

#### 2.2.2.1 Library preparation of OP-Strand-seq

The new OP-Strand-seq protocol is summarized side-by-side with the original Strand-seq protocol in Figure 2.10. Several aspects of library preparation between the OP-Strand-seq protocol and the original Strand-seq protocol remain the same. In short, the thymidine analog, BrdU, is incorporated into nascent strand synthesis during DNA replication and single cells hemi-substituted with BrdU are sorted into individual wells by FACS (Figure 2.10) [57]. BrdU incorporated strands are selectively degraded by UV irradiation and Hoechst to allow for the selective amplification of template strands by PCR (Figure 2.10) [57].

The library preparation of OP-Strand-seq differ from that of the original Strand-seq method described in Sanders et al., 2017 in three main ways. Firstly, the reagent volume was reduced between 500- to 1,000-fold to improve the efficiency of the enzymatic steps involved in

library preparation such as digestion and ligation. The reduction in individual reaction volumes also permitted the increase in the relative concentration of DNA fragments which further supports enzymatic efficiency. For example, the ligation of adapter to each other to form adaptor dimers rather than to DNA fragments is a common limitation of enzymatic efficiency and by increasing the relative concentration of DNA fragments in smaller reaction volumes, adaptor dimers are less likely to form [57], [100]. Secondly, genomic DNA was fragmented in bulk using micrococcal nuclease (MNase) to reduce the overall variability and GC-bias in standard library digestion and unevenness of sequencing coverage and a higher proportion of background reads (Figure 2.13). Lastly bead clean-up steps are replaced with thermolabile protease treatments due to the loss of DNA reads associated with bead clean-up purification steps (Figure 2.10).

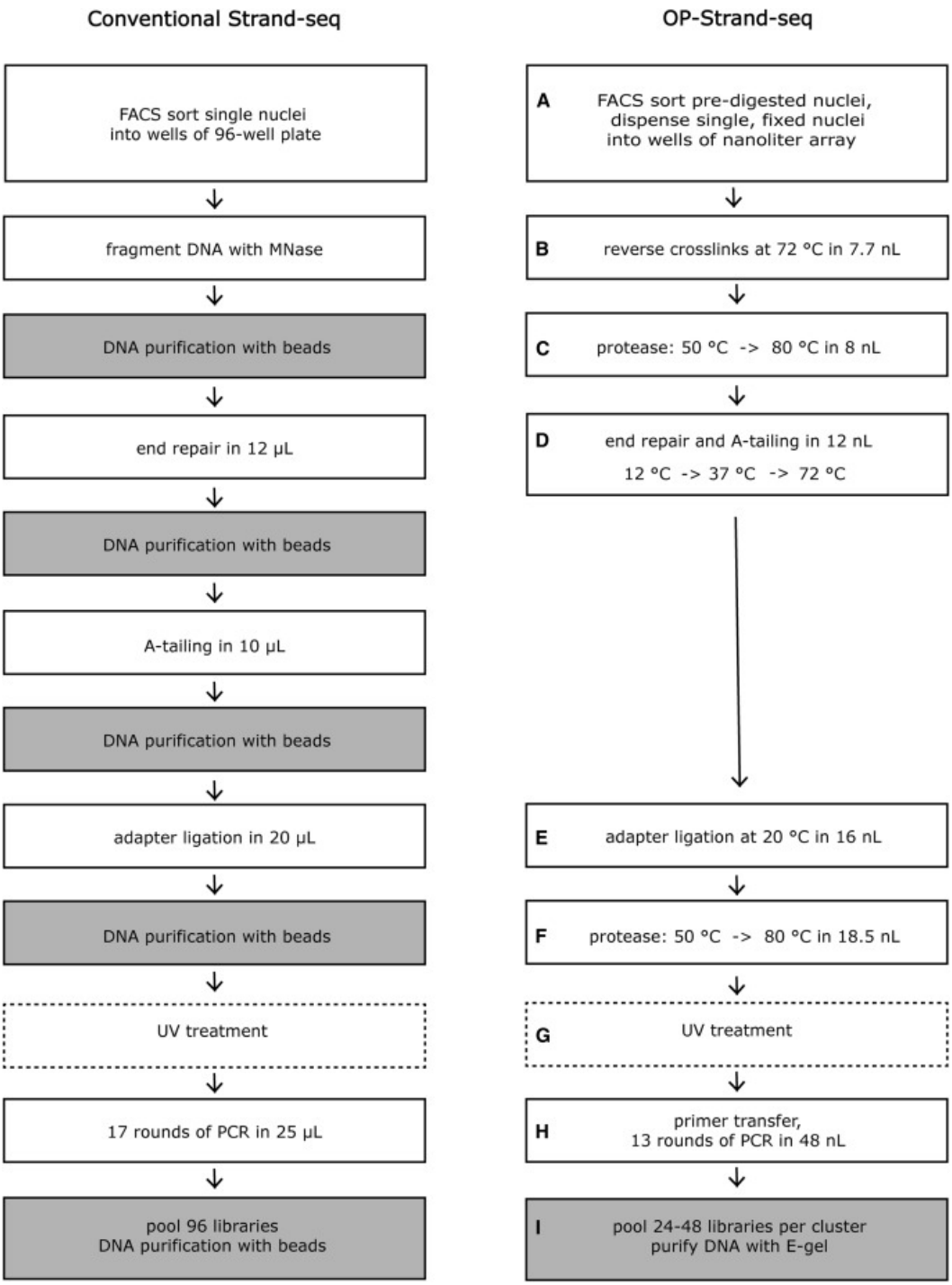



**Figure 2.10 Comparison of the original Strand-seq protocol (left) with the OP-Strand-seq method (right). Adapted from Hanlon et al., 2021.**

#### 2.2.2.2 Illumina whole genome sequencing

Sequencing libraries were pooled together to allow for size selection to remove primer and adapter dimer contamination. Specifically, 1 μL of pooled library sample was run on a 2% E-Gel EX Agarose Gel and DNA fragments >200 bp were selected and purified with the Zymoclean Gel DNA Recovery Kit. Purified samples were loaded onto a NextSeq 550 instrument for paired end 75 bp sequencing following standard Illumina guidelines for denaturation and dilution [57], [60]. This approach allows the selective sequencing of the parental DNA template strands and the generation of directional libraries with reads mapping to the reference genome in the orientation of the native parental DNA template strands.

#### 2.2.2.3 Bioinformatic pre-processing

The output of Illumina sequencing are Binary Base Call (BCL) files that require demultiplexing to generate separate FASTQ files for each library. Adaptor sequences are removed from FASTQ files using Cutadapt (cutadapt-v4.1) using default parameters and reads shorter than 30 bp are removed. Libraries were aligned to the GRCh38 human reference with Bowtie2 (bowtie2-v2.4.5) using default parameters and duplicate reads were removed with Picard (picard-v2.27.3) using default parameters to generate BAM files that are sorted using Samtools (samtools-v1.15.1) [60]. In total, 3873 Strand-seq libraries were sequenced across 21 independent sequencing experiments. Indeed, the coverage and quality of sequencing libraries is variable therefore, a quality control (QC) step is needed to discard poor quality Strand-seq libraries, and this was originally done by manually characterising the quality of each library in an experiment. I would typically exclude sequencing libraries with low coverage (<50k reads, <25 Reads Per Mb), high proportion of reads mapping to the wrong template strand (herein referred to as background), and uneven sequencing coverage (herein referred to as spikiness).

Background reads or increased bin-to-bin variation (spikiness) in read depth or sporadic gaps (spikiness) in read density can result from too many cycles of BrdU incorporation. For manual QC, there is no threshold for the level of background or spikiness that is considered acceptable, these metrics are assessed in unison by domain experts.

### 2.2.3 Classifier for automated quality control of OP-Strand-seq libraries

The reduced costs and increased throughput of the OP-Strand-seq method poses unique challenges to scalability. Currently, a quality control (QC) step is needed to discard poor quality Strand-seq libraries, and this is done only by domain experts capable of manually characterising the quality of each library in an experiment. There is one automated QC method that was recently developed for the *A*utomatic *S*election of *H*igh-quality *L*ibraries for the *E*xtensive anal*Y*sis of *S*trand-seq data (ASHLEYS). ASHLEYS uses pretrained models to categorize Strand-seq libraries using the original Strand-seq method with 92% accuracy. However, this method can only classify haploid OP-Strand-seq libraries with an accuracy of 83.4% according to our estimates (Figure 2.14). Therefore, I developed a novel classifier to automate the selection of good quality haploid OP-Strand-seq libraries.

#### 2.2.3.1 Training random forest model to classify Strand-seq library quality

First, I manually annotated the quality of 3873 OP-Strand-seq libraries and performed an 80:20 split to generate a training and test set for developing a classifier. Then I collected the following metrics from aligned BAM files for each library to be used as features in our classifier: coverage, background, evenness, and spikiness. Coverage is simply a metric of depth of sequencing or the amount of the genome that has been captured during sequencing (Figure 2.11).

Coverage is calculated by multiplying the number of reads in a library, $n_{reads}$, by the average read length, $x_{read\ length}$, and dividing by 100 (Equation 2.1).

$$\boldsymbol{C = \frac{(n_{reads} * \ x_{read\ length})}{100}} \tag{2.1}$$

Background, as previously mentioned, refers to a metric devised to calculate the proportion of non-directional reads (Figure 2.11). The relative proportion of Crick reads in WW segments is added to the relative proportion of Watson reads in CC segments and divided by 2 (Equation 2.2).

$$\boldsymbol{b = \frac{(\frac{\sum_{t=1}^{T-1} WWc}{\sum_{t=1}^{T-1} WWw}) + (\frac{\sum_{t=1}^{T-1} CCw}{\sum_{t=1}^{T-1} CCc})}{2}} \tag{2.2}$$

Spikiness is a metric that assesses the bin-to-bin variability in sequencing coverage (Figure 2.11). Each bin read count, $x_t$, is subtracted from the next adjacent bin read count and averaged across the genome (Equation 2.3).

$$\boldsymbol{s = \frac{\sum_{t=1}^{T-1} |x_{t+1} - x_t|}{\sum_{t=1}^{T} x_t}} \tag{2.3}$$

Evenness is a metric that assesses the genome wide variability in sequencing coverage (Figure 2.11). The genome is split into Mb-sized bins and the median value for binned read counts, $\mu_{reads\ per\ Mb}$, is calculated. The absolute difference between each binned read count, $x$, and the median value is used to calculate a Z score for each bin that is averaged across the genome (Equation 2.4).

$$E = \frac{\sum_{t=1}^{T-1} \frac{\left| x - \mu_{reads\ per\ Mb} \right|}{\sigma}}{\sum_{t=1}^{T} n_{bins}} \quad (2.4)$$

Together, I used these four features to predict the quality classifications of good Strand-seq libraries using a random forest algorithm from the R package, Caret [101].

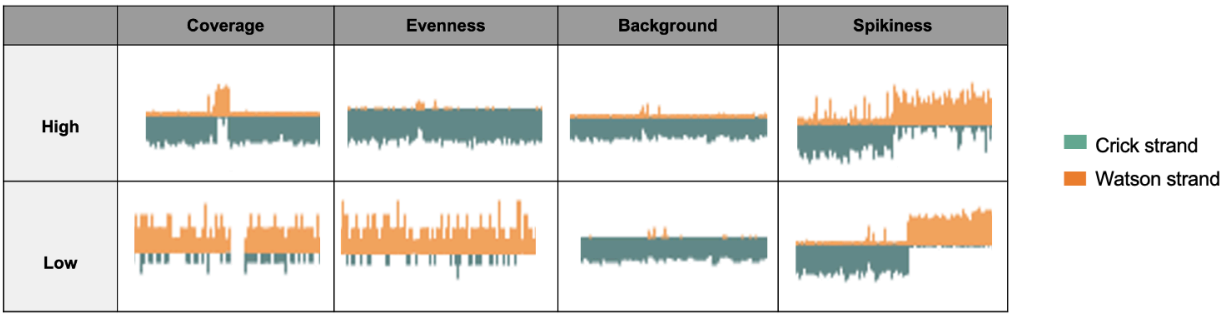



**Figure 2.11 Examples of Strand-seq chromosome ideograms showcasing differences in library quality features.**

## 2.3 Results

### 2.3.1 Comparison of breakpoint resolution between haploid and diploid cells

As previously mentioned, Strand-seq was first developed with the use of diploid cells. I explained at the beginning of Section 2.2.1 why haploid cells would pose several advantages over diploid cells for the downstream analysis of SCEs and SVs. One advantage I discussed is the theoretical improvement in strand state switch breakpoint resolution in haploid cells versus diploid cells given the same depth of sequencing (herein referred to as sequencing effort). I confirmed that for the same sequencing effort, the resolution of strand state switch breakpoints is higher in haploid cells than in diploid cells as indicated by a lower breakpoint confidence interval (Figure 2.12). Therefore, SCEs and SVs are mapped to the genome at higher resolution than in diploid cells for the same cost.

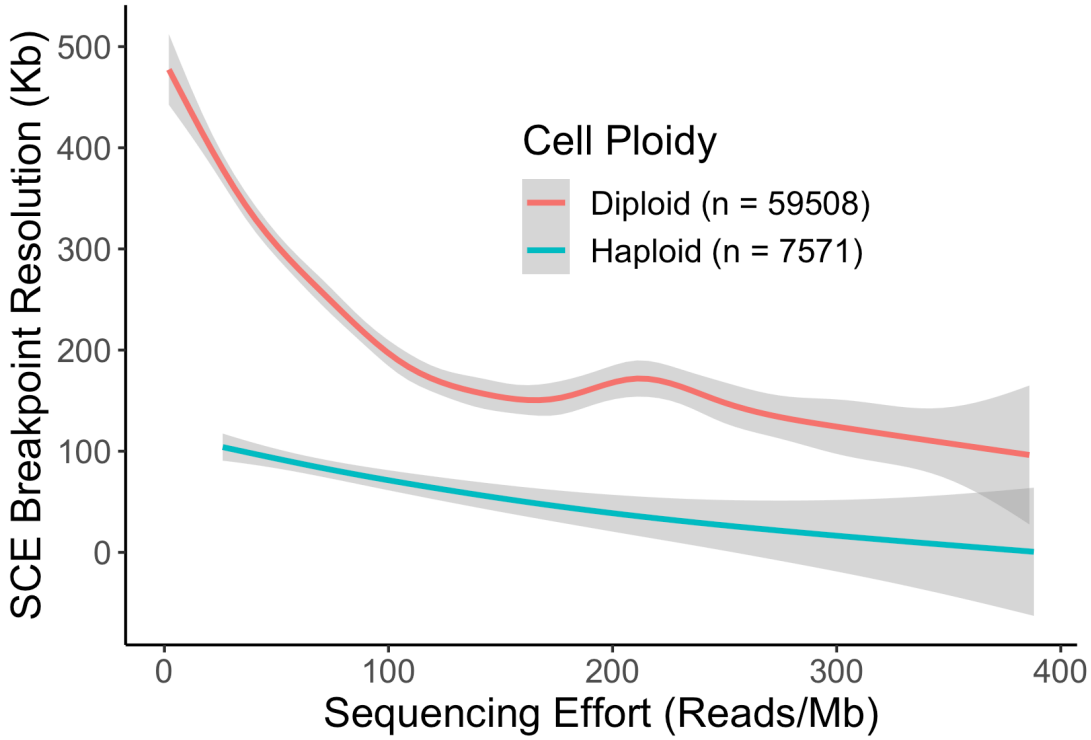



**Figure 2.12 SCE breakpoint resolution relative to sequencing effort for haploid and diploid cells.**

### 2.3.2 Improved cost, throughput, and quality of the OP-Strand-seq protocol

Next, I wanted to demonstrate the improved cost, quality, and throughput of the OP-Strand-seq method. OP-Strand-seq can produce between 6 and 16-fold more libraries than the original Strand-seq protocol [57], [60]. The cost per library has also been reduced to 15% of the original cost per library using the original Strand-seq protocol, at $2 versus $13 per library [57], [60].

To assess the quality of Strand-seq libraries, I primarily focused on assessing the average complexity of the sequencing library. Complexity refers to the number of unique sequencing reads captured in one library and, using single cells, directly reflects the percentage of the genome that was captured in the library [102]. Low complexity sequencing libraries have many of the same reads and deeper sequencing would only yield more duplicate reads that would be removed during pre-processing, resulting in wasted sequencing costs [102]. High complexity

sequencing libraries offer more unique reads and thus higher genomic coverage when sequenced deeper [102]. One bioinformatic tool, known as *PreSeq,* was developed for assessing complexity in individual sequencing libraries by devising a function to model the expected genomic coverage at increasing sequencing efforts [102]. I used *PreSeq* to show that the OP-Strand-seq libraries have ~4-fold greater complexity on average than libraries made with the original protocol capturing up to 25% of the haploid genome per cell [57], [60].

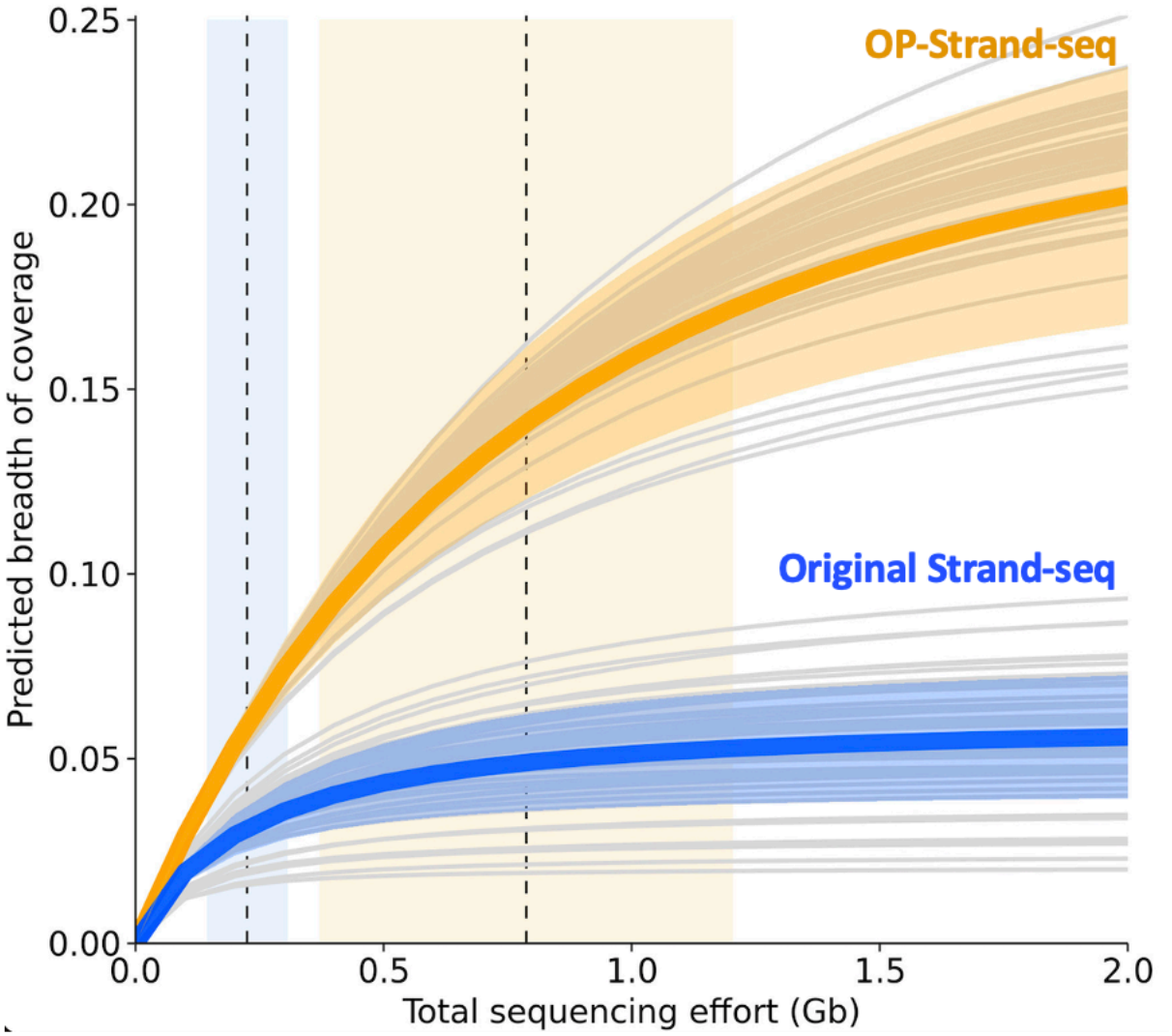



**Figure 2.13 Complexity curves for libraries made with OP-Strand-seq and original Strand-seq.**

**Individual libraries are show as gray lines. Complexity mean and standard deviation of libraries are shown in yellow for those produced with OP-Strand-seq blue for those produced with original protocol. Sequencing effort mean and standard deviation shown with dashed vertical lines. Breadth of coverage is the fraction of the haploid reference genome covered by at least one read fragment. Complexity estimates were made using Preseq. Adapted from Hanlon et al, 2021.**

### 2.3.3 Testing random forest model in classifying Strand-seq library quality

Our classifier was trained on 3,873 OP-Strand-seq haploid libraries to accurately predict good quality libraries using the methods described in Section 2.2.3. Model training included feature selection and training error estimation using 10 iterations of class-balanced 10-fold cross-validation. I showed that the four selected features can distinguish the quality of Strand-seq libraries (Figure 2.14A). Then I assessed the performance of our classifier in comparison to the only existing Strand-seq library quality classifier [103]. Model performance was assessed on an independent test dataset ($n = 893$). The random forest model I trained showed an accuracy of 96.6% versus 83.4% for ASHLEYs on the same set of OP-Strand-seq libraries (Figure 2.14B).

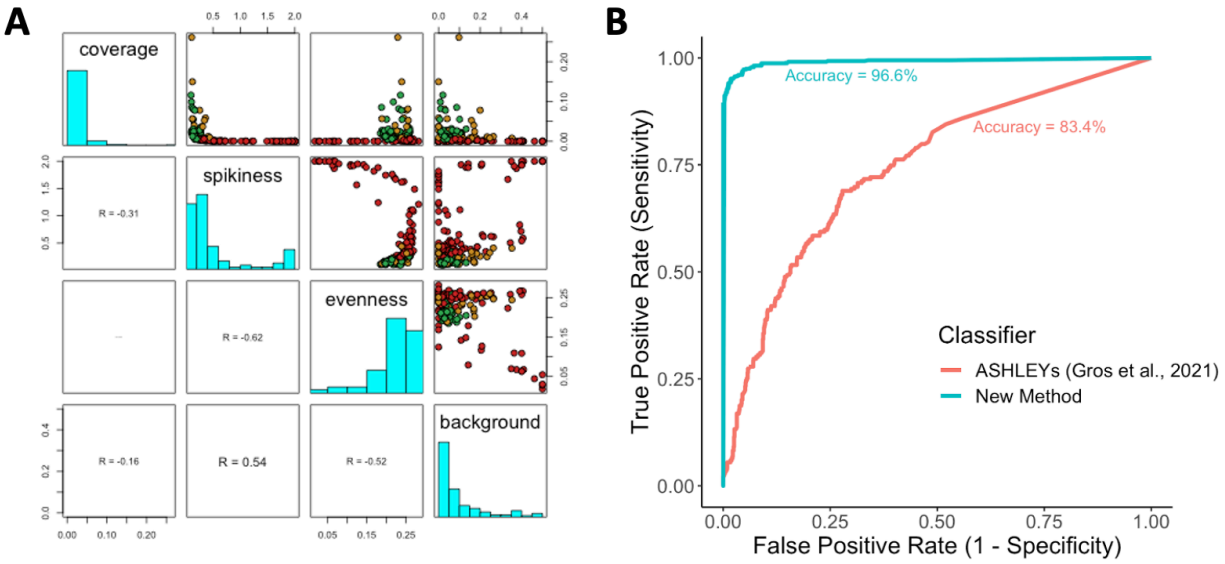



**Figure 2.14 Feature selection and model performance of Strand-seq library classifier.**
**(A) Properties of OP-Strand-seq libraries. I calculated the coverage, spikiness, evenness, and background. High-quality libraries are shown in green, medium-quality libraries are shown in yellow and low-quality libraries are shown in red. (B) ROC curves assessing performance of random forest classifier and ASHLEYS on the same test dataset of OP-Strand-seq libraries. Accuracy is labelled beneath each curve.**

Considering the accuracy of ASHLEYs on our test set was lower than what has been reported by Gros et al., I wanted to investigate the possibility of overfitting on a novel datatype. Our test dataset consisted of OP-Strand-seq libraries whereas ASHLEYs was trained and tested on libraries made using the original Strand-seq protocol [103]. I re-tested our model and ASHELYS on both OP-Strand-seq libraries and original Strand-seq libraries (Figure 2.15). As expected, I found ASHLEYs had an accuracy of 91% on original Strand-seq data compared to an accuracy of 74% using our random forest classifier (Figure 2.15) [103].

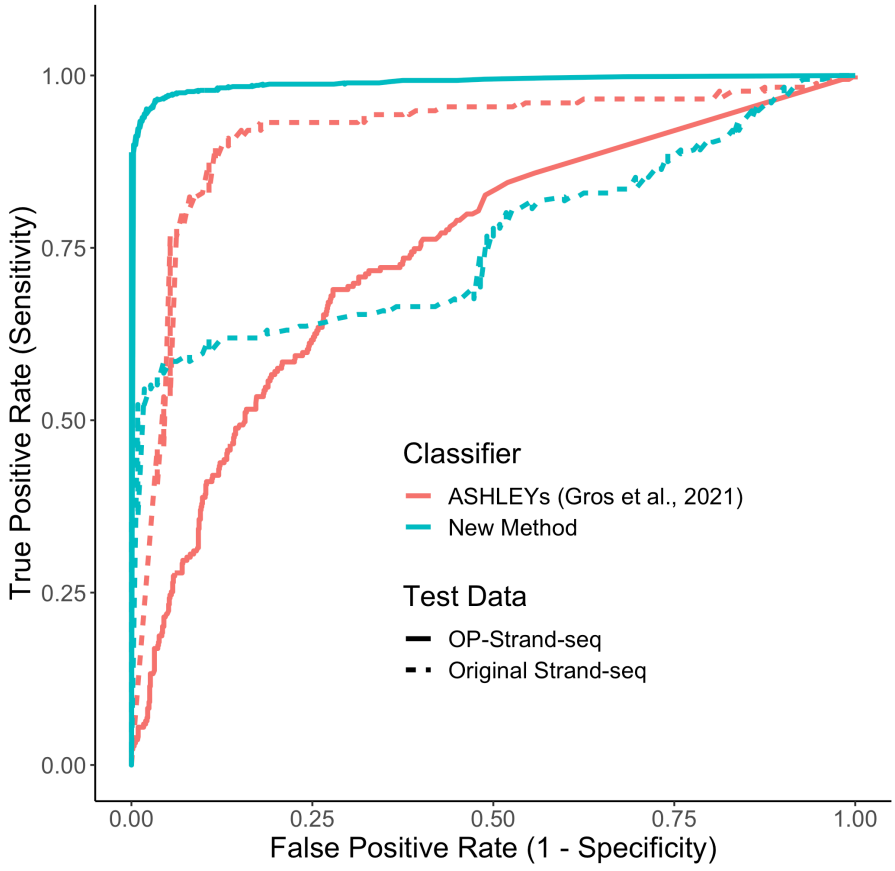



**Figure 2.15 ROC curves assessing performance of random forest classifier and ASHLEYS. Each classifier was tested on two datasets: OP-Strand-seq libraries and original Strand-seq libraries. Accuracy is labelled beneath each curve.**

The classifiers shown here appear to be overfitted to the data used for training and limit their universal application. Increasing the size of the training sets will undoubtedly help mitigate overfitting. However, it should be noted that the quality of Strand-seq libraries may depend on the desired application. For example, accurate copy-number analysis in Strand-seq libraries requires uniform sequencing coverage, and thus library evenness could be weighted more heavily than other sequencing metrics when assessing library quality for this application. Alternatively, SCE analysis may be able to accommodate higher spikiness or lower sequencing coverage in comparison to SV analysis, especially when it comes to analyzing smaller SVs that require high sequencing coverage. Thus, a multi-variable definition of library quality that is dependent on the desired application may be necessary for the accurate classification of Strand-seq library quality.

## 2.4 Discussion

Strand-seq can be used to facilitate DNA repair studies by revealing essential roles of DNA helicases in DSB repair. However, there are several limitations of this approach as discussed in Section 2.1. The variable quality of sequencing libraries can worsen the resolution of breakpoints for SCEs and SVs which poses a significant challenge for both detection and downstream enrichment analysis of breakpoint coordinates. Therefore, I introduced three novel implementations in Section 2.2 to address these and improve the overall quality of DNA repair studies that can be performed using Strand-seq.

The first implementation intends to harness the unique qualities that haploid cells possess in order to cut sequencing costs and improve breakpoint resolution. Because haploid cells only possess one set of chromosomes, they have smaller breakpoint confidence intervals than diploid cells, given the same sequencing effort. The second implementation describes a revised library

preparation protocol for Strand-seq known as the OP-Strand-seq protocol. OP-Strand-seq can produce 6 to 16-fold more libraries than the original Strand-seq protocol at 15% of the original cost with ~4-fold greater complexity, capturing up to 25% of the haploid genome per cell on average. These improvements in library quality and throughput create challenges for scalability and thus necessitate the automation of QC to discard poor quality Strand-seq libraries. Therefore, the third implementation is a random forest classifier that can accurately classify good quality OP-Strand-seq libraries with 96.6% accuracy and original Strand-seq libraries with 74%. This accuracy is higher than what has been reported by the ASHLEYs classifier for OP-Strand-seq libraries although lower for classifying original Strand-seq libraries.

Together, these three implementations allowed us to generate thousands of good quality OP-Strand-seq libraries. These libraries possess improved breakpoint resolution for SCE and SV calling as discussed in Chapter 3. Improved resolution of the SCE callsets generated in Chapter 3 will increase the power of our enrichment analysis in Chapter 4. Improved breakpoint resolution of SCEs is essential for performing enrichment analysis because SCE-triggering structures, such as G4s, can occur frequently throughout the genome (~ 8.6 Kb on average) and performing enrichment analysis with SCEs with poor resolution, or large confidence intervals, would result in increased noise because of the high likelihood of permutated SCE regions overlapping with G4s due to their large size. These implementations will undoubtedly help clarify the role of specific helicases in resolving different kinds of replication barriers by investigating the genomic context of finely mapped SCE coordinates. These studies will yield novel information about molecules and pathways involved in recombination and sister chromatid exchange mechanisms in mammalian cells.

# Chapter 3: Structural variant callers

## 3.1 Introduction

The purpose of this chapter is to introduce the bioinformatic tools I developed for the comprehensive discovery of somatic structural variants (SVs) in individual cells. Specifically, I introduce novel approaches to screen Strand-seq libraries for SCEs, CNAs, and translocations.

### 3.1.1 Structural variants

As discussed in Chapter 1, SVs contribute greater diversity at the nucleotide level between two human genomes than any other form of genetic variation [49]. SVs are defined as genetic variants that rearrange, delete, or amplify a sequence of DNA greater than 50 bp in length and are grouped into classes based off the nature of their rearrangement [104]. SVs are considered separate from insertions and deletions less than 50 bp which are commonly referred to as INDELs. SVs include duplications, deletions, translocations and inversions and they can be broadly classified as either copy-neutral (inversions and translocations) and copy-number changes (deletions and duplications) [105]. They encompass key mutational processes in cancer that can drive oncogenesis and tumour development by, for example, altering oncogene copy number, disrupting tumour-suppressor genes or creating fusion genes that allow one gene to hijack the regulatory sequences of another gene [106]. In fact, a given SV is 53x more likely to have a phenotypic consequence on gene expression than a given SNP or small INDEL (< 50 bp) and at least 30% of cancers contain at least one pathogenic SV [107], [108]. Somatic SVs are abundant in cancer genomes and are considered a major source of genomic instability [108].

### 3.1.2 Structural variant discovery

Bulk WGS technologies makes it possible to identify SVs above a minimum VAF [58]. However, scWGS technologies can characterize rare SVs below the minimum VAF and are capable of resolving subclonal SV heterogeneity in tumours [58]. There are three main approaches for detecting SVs using scWGS data: mapping-based, mapping-free, and assembly-based approaches [109]. In mapping-based approaches, each class of SV experiences a unique pattern of read mapping that can be used to infer the underlying mutation [109]. Mapping based approaches encompass signatures of read depth, read pairs, and split read mapping [104]. Changes in read depth suggest a CNA has occurred [104]. Discordant read pairs arise when the distance or orientation of reads differs from what is expected and suggest an inter-chromosomal translocation or inversion has occurred [104]. Split reads span the breakpoints of SVs can be used to detect small SVs [104]. Long reads, as discussed in Section 1.2.5.4, are also capable of resolving large SVs using mapping-based approaches [54]. Mapping-free methods detect SVs by comparing WGS data between different genomes [109]. Lastly, assembly-based approaches involve the reference-free *de novo* genome assembly of sequenced reads into larger contig assemblies to allow for accurate SV detection [109]. Assembly-based methods have integrated multiple sequencing approaches including short read, long read, bulk and scWGS technologies and are considered the most accurate and comprehensive approach for SV discovery, albeit, at high costs due to computing resources and integrated WGS technologies [49], [109].

### 3.1.3 Structural variant discovery challenges

Despite their relevance, SVs have remained an understudied source of genetic variation due to the technological challenges associated with detection. As discussed above, SVs

experience unique patterns of read mapping however, different classes of SVs can still evade detection by confusing sequencing assembly and SV detection algorithms in different ways [109]. These patterns are difficult, and in some cases virtually impossible, to detect if the regions spanning the SV is associated with repetitive DNA, uneven and sparse sequencing coverage, or multiple SVs are overlapping or nested within one another [109].

SV detection algorithms break down within repetitive DNA, which is heavily enriched for SVs and highly abundant in the genome [49], [104], [109]. In fact, tandem repeats, microsatellites, and inverted repeats, among other repeat-rich DNA elements, make up approximately 45% of the genome [3]. Repetitive DNA can be prone to false SV discovery due to read mapping errors [104], [109]. For example, tandem duplications create multiple possible alignments resulting in some repeats may going unmapped, or other types of mapping errors (Figure 3.1) [104], [109]. Additionally, copy-neutral SVs such as inversions are often flanked by highly repetitive sequences that evade read alignment all together resulting in unmapped regions flanking the inversion [104], [109]. These SVs evade detection all together because they preserve the same underlying DNA content and individual reads cannot span the full length of the rearrangement or the breakpoints of the rearrangement to indicate its presence [104], [109].

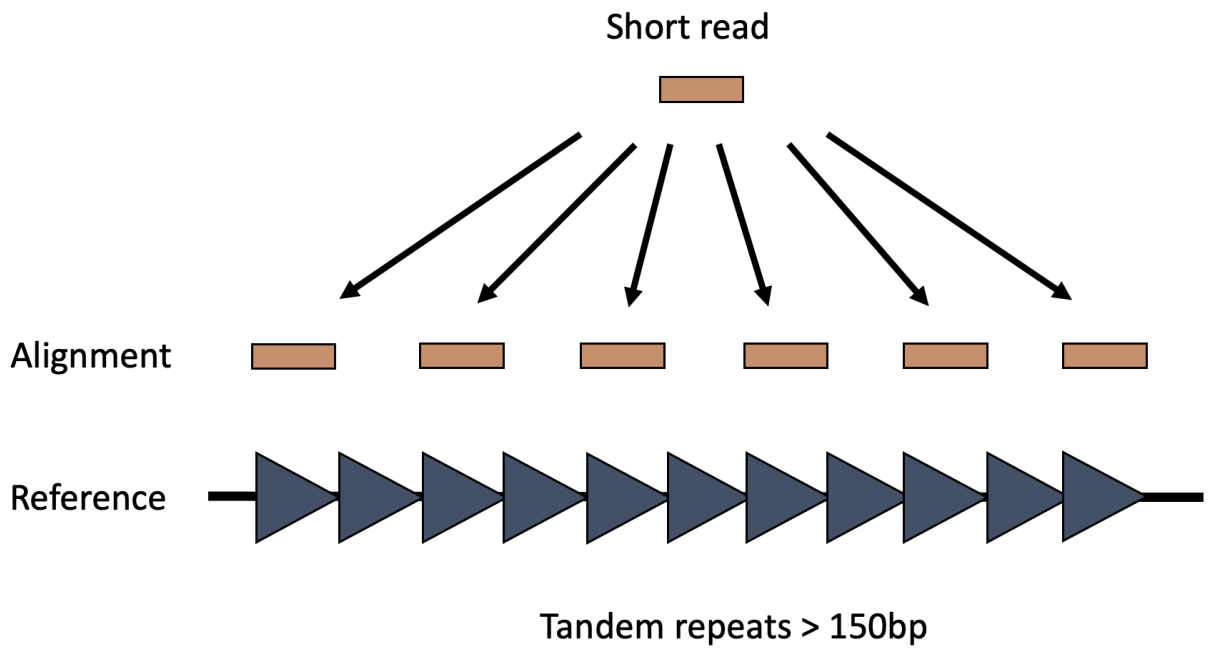



**Figure 3.1 Multiple possible mapping patterns of reads within tandem duplications.**
**A sequencing read (orange) can align to multiple positions within tandem repeats (blue) that exceed the length of short reads.**

Copy-number SV detection that relies on read depth changes also breaks down with uneven or sparse sequencing coverage [104]. Reads that are GC-rich may be preferentially amplified during the PCR step of library preparation and can thus create false positive calls that inappropriately appear as read depth changes [110], [111]. Copy-number SVs may also evade read depth changes when sequencing coverage is sparse resulting in false negative calls [110], [111].

Lastly, multiple SVs that are overlapping or nested within each other, such as inversion-duplications, account for many SVs and give rise to complex mapping patterns that are difficult to resolve [109].

In Chapter 2, I described the unique advantages of Strand-seq over other WGS methods and how read directionality in a sequencing library can be used to pinpoint different classes of SVs and SCEs. This is strictly a Strand-seq-specific quality that has been exploited to facilitate

comprehensive SV discovery [49], [58]. As a new technology, the accessibility of Strand-seq is directly related to the availability of bioinformatic tools capable of exploiting the directionality of template reads for comprehensive SV discovery. To date, there are no standardized methods for comprehensive full-spectrum SV discovery in Strand-seq data. Here, I introduce the general framework for mapping SCEs, translocations, and CNAs to facilitate the development of standardized bioinformatic tools that can be used by the wider SV discovery community.

## 3.2 Methods

As discussed in Chapter 2, Strand-seq is a method in which parental DNA template strands are sequenced in single daughter cells to generate directional libraries. One of the unique applications of Strand-seq is the ability to identify SCEs as well as complex SVs in the genome of cells by pinpointing changes in template strand inheritance, haplotypes as well as copy numbers [25], [56], [58]. Here, I introduce a conceptual framework for the comprehensive discovery of SCEs, translocations, and CNAs in Strand-seq data. I used the R package, *BreakpointR,* to identify intra-chromosomal strand state switches, or breakpoints, that can be used for downstream SCE and translocation analysis. I also used the R package, *AneuFinder*, to identify intra-chromosomal changes in read count that can be used for downstream CNA analysis. The resolution of how finely these events can be mapped to the genome and the accuracy in which they can be called are among the most important indicators for assessing the performance of our bioinformatic tools and are discussed in Section 3.3.

### 3.2.1 Identifying strand-state change breakpoints

I used the R package, *BreakpointR*, to analyze BAM files from good quality individual Strand-seq libraries that were selected during the manual QC step discussed in Section 2.2.2.3. From the 3873 KBM7 OP-Strand-seq libraries sequenced in Chapter 2 collected across 21 independent sequencing experiments, we retained 1684 good quality Strand-seq libraries from all RecQ KO lines for SV analysis that passed our QC step. *BreakpointR* is a tool for identifying intra-chromosomal changes in strand-state genotype in individual cells (Figure 3.2A) [97]. It functions by forming a sliding window of a user-defined bin size (e.g. 20 mapped reads in a given library) at the beginning of each chromosome and calculating the percentage of Watson reads in the first half of the bin compared to the second half to determine the change in Watson reads, or ΔW value, for that location (Figure 3.2B) [97]. The bin on each chromosome then slides one read over at a time to calculate the ΔW at each point of that chromosome until it has shuffled across the whole chromosome and the highest ΔW values can be identified and the coordinates refined to call changes in intra-chromosomal strand state genotype with high confidence (Figure 3.2C)[97]. Strand state change breakpoints are genotyped to identify the exact nature of the strand state transition and coordinates are refined using Fisher's exact test (Figure 3.2D) [97]. The coordinates for these breakpoints can be used for downstream analysis of different SVs [57], [59].

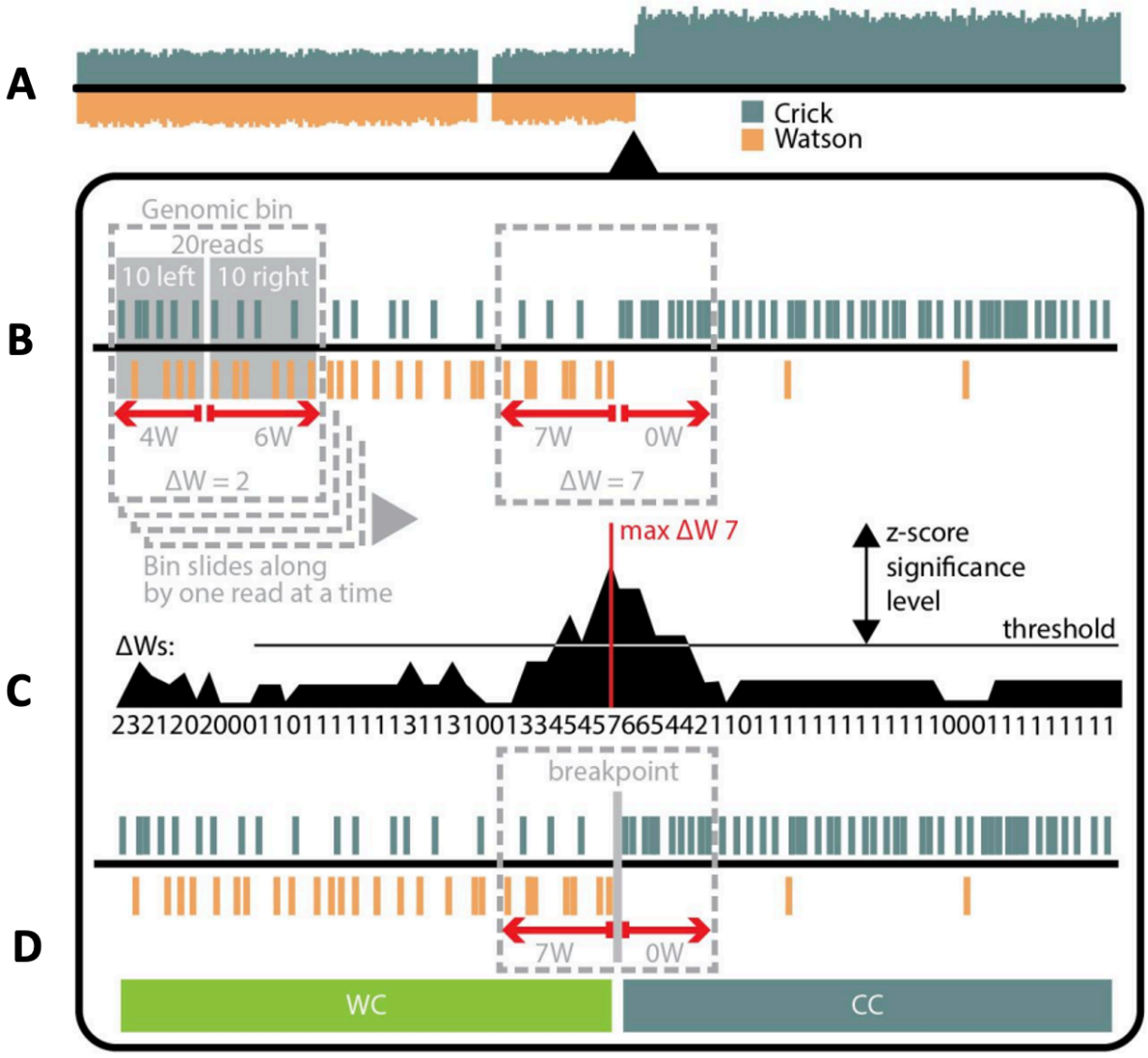



**Figure 3.2 BreakpointR algorithm**
**(A) Binned read counts for a chromosome where vertical bars denote number of 'Crick' (C; teal) reads and 'Watson' (W; orange) reads in each bin. (B) User-defined bin of 20 reads is split in half and the number of W reads in the left portion of the bin are subtracted from the number of W reads in the right portion (ΔW). The bin advances one read at a time and is dynamically resized to accommodate changes in read density and sequencing coverage seen. (C) Peak calling is then applied to search for high-confidence peaks in the ΔW scores. Peak confidence is determined using z-score statistics to test for significance above a user-defined threshold. (D) Significant peaks are considered putative breakpoints that mark the location of template-strand-state changes. Using these breakpoints to define segments for strand state assignment. The strand state is tested between all putative breakpoints by measuring the total number of W and C reads in the segment and assigning the most-probable template-strand-state using the Fisher's exact test. A breakpoint is retained only if two neighboring segments show different template-strand-state; otherwise the breakpoint is removed and the two segments are merged. Figure adapted from Porubsky et al., 2020.**

### 3.2.2 Bioinformatic approaches for SCE detection

As discussed in Chapter 2, haploid and diploid cells have unique signatures for SCEs. Therefore, I developed a ploidy-based SCE caller to refine breakpoint calls generated by *BreakpointR* to those that most likely represent SCEs and omit breakpoints that represent other SVs or false positive calls. This caller has three main steps.

First, I use the ploidy of a cell to identify impossible genotypes across a breakpoint for an SCE. For example, SCE breakpoints in a diploid cell can only affect one homolog because two SCEs occurring in the same position on two homologs in the same cell is very unlikely and more likely to represent a homozygous SV. Therefore, SCE breakpoints can only exhibit the transition of one homolog in a diploid cell, meaning the other homolog retains the same strand state genotype (i.e. WW→WC, CC→WC, WC→CC, WC→WW; Figure 3.3). By extension, breakpoints that affect both homologs are deemed homozygous breakpoints (i.e. WW→CC, CC→WW) and are omitted for a diploid cell (Figure 3.3). Conversely, SCE breakpoints in a haploid cell can only resemble a homozygous breakpoint (i.e. WW→CC, CC→WW) because they only have one homolog, so all other breakpoint genotypes are omitted (Figure 3.3). Often, mis-aligned reads can result in "diploid" appearing genotypes (i.e. WC) where both Watson and Crick reads are aligning to the genome in the same area of a haploid cell. The sources of mis-aligned reads are addressed in the next step of our SCE caller.

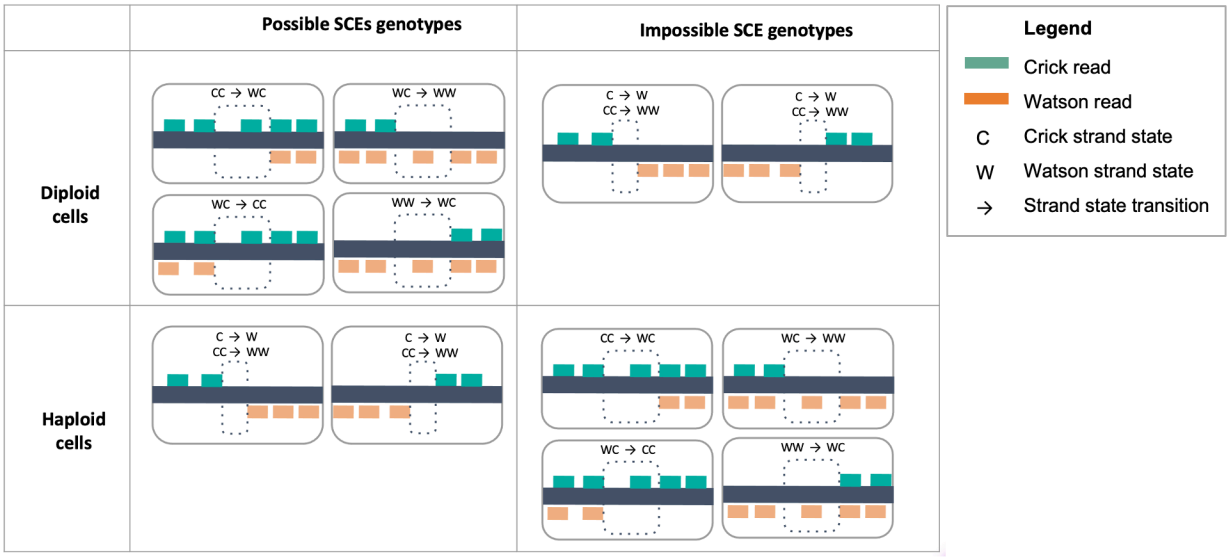



**Figure 3.3 Examples of possible and impossible SCE breakpoint genotypes for haploid and diploid cells. SCE breakpoints in a diploid cell can only affect one homolog and thus can only exhibit the transition of one homolog in a diploid cell, meaning the other homolog retains the same strand state genotype. Breakpoints that affect both homologs are deemed homozygous breakpoints and are considered unlikely to represent an SCE in a diploid cell. In a haploid cell, SCE breakpoints can only resemble a homozygous breakpoint because they only have one homolog, so all other breakpoint genotypes are considered unlikely to represent an SCE.**

Second, the co-occurrence of two neighboring events is also considered to stem from misaligned background reads and breakpoints occurring within 2 Mb of each other on the same chromosome in the same cell are omitted. These likely correspond to false positive calls stemming from background reads due to mapping errors. Mapping errors may occur due to highly repetitive DNA content such as centromeres and result in small segments of reads mapping to both strands in every cell, otherwise known as "Always-Watson-Crick" (AWC) regions (Figure 3.4). These regions were identified individually from observing small spikes (< 2Mb) of AWC regions recuring in all libraries (Figure 3.4). These regions are blacklisted from the final call so that SCEs called in these regions are omitted. I devised a custom blacklist created for the KBM7 cell line however, it is strongly recommended that a new blacklist is generated if other cell lines are used.

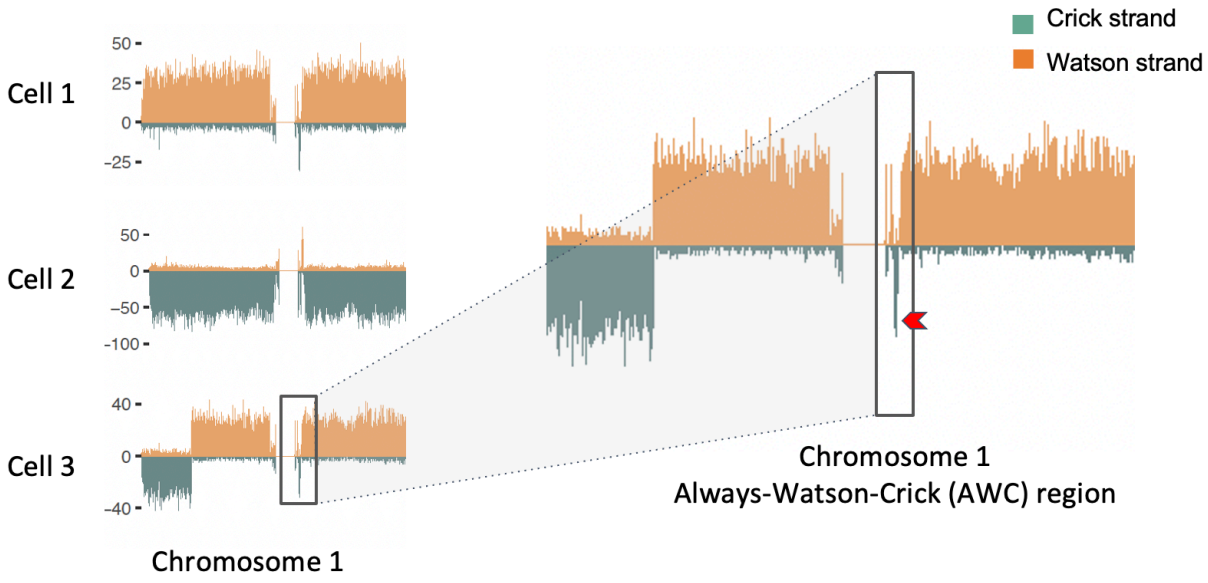



**Figure 3.4 Example of Always-Watson-Crick region on ideograms of binned read counts for chromosome 1.**

Lastly, recurring breakpoints in multiple single cell libraries that are not in close proximity to other breakpoints or centromeres likely correspond to non-SCE SVs. For example, translocations can result in one breakpoint recurring in the same position in multiple cells (Figure 3.5).

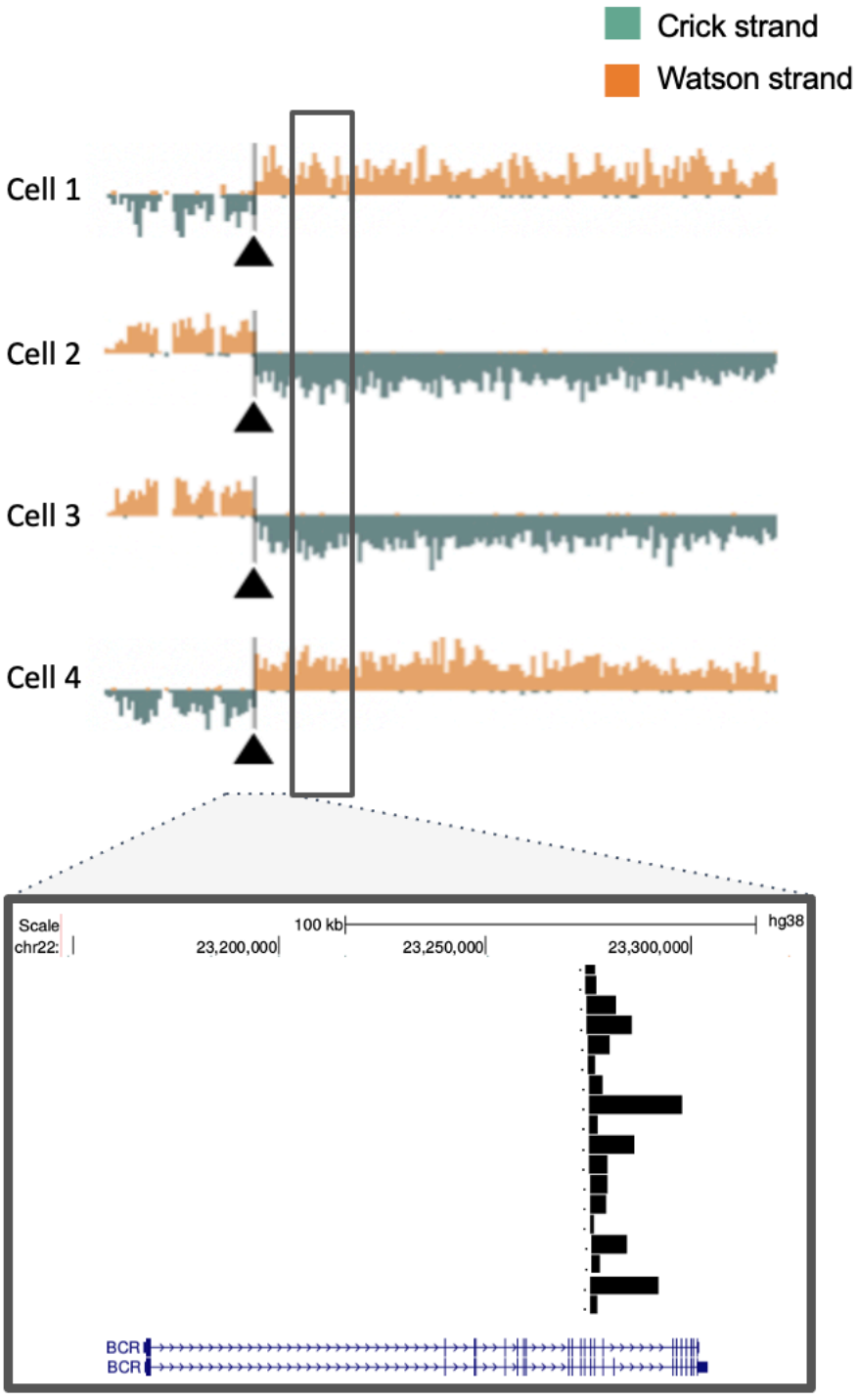



**Figure 3.5 Example of breakpoints recurring in multiple cells that correspond to translocation breakpoint**

A a custom script derived from an auxiliary function of BreakpointR known as "hotspotting" was used to find regions of the genome where the density of breakpoints significantly exceeds a null gaussian distribution of events to identify recurring breakpoints in many libraries (Figure 3.6). Most often, these events correspond to SVs however, in some circumstances, these "hotspots" can correspond to many SCEs occurring in the same area of the genome in multiple cells. However, hotspots made up of SCE or SVs can only be distinguished by uploading bed-formatted read count files generated by *BreakpointR* to the UCSC Genome Browser [112] to identify whether the breakpoints for SCEs are clustering near each other in different cells or if SVs have produced identical breakpoints in multiple cells.

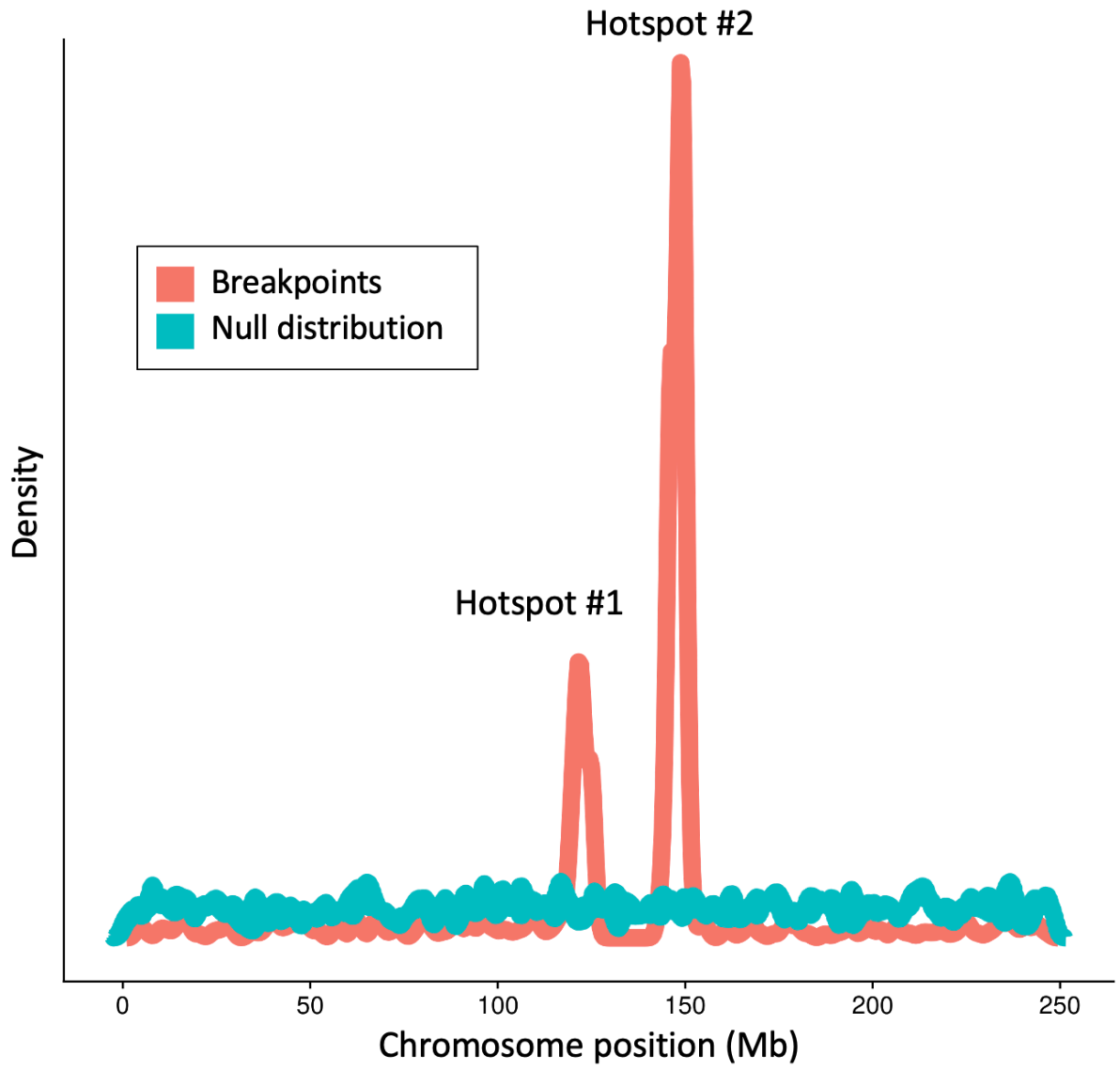



**Figure 3.6 Density distribution of SCE hotspots.**
**Red line shows density distribution of breakpoints across chromosome 1. Blue line shows simulated random distribution of breakpoints. Two peaks correspond to two breakpoint hotspots.**

### 3.2.3 Bioinformatic approaches for translocations detection

Strand-seq libraries can reveal the presence of translocations. Translocations are relevant to genomic instability because they can disrupt gene function and regulation by rearranging genomic architecture with one of the more notable examples of this being the Philadelphia chromosome that is formed when the q-arms of chromosomes 9 and 22 fuse and form the *BCR*-

*ABL1* fusion gene that is known for initiating leukemogenesis [106], [113], [114]. Translocations arise when simultaneous DSBs occur on separate chromosomes and aberrant repair of both breaks results in the fusion of separate chromosome segments [114].

In a Strand-seq library, a translocation can be visualized as two strand state switches in two chromosomes recurring in multiple libraries such that the fused segments from one chromosome would follow the same inheritance pattern of its translocation partner (Figure 3.7) [58]. This difference in inheritance pattern would only be present in 50% of independent assortment combinations [58].

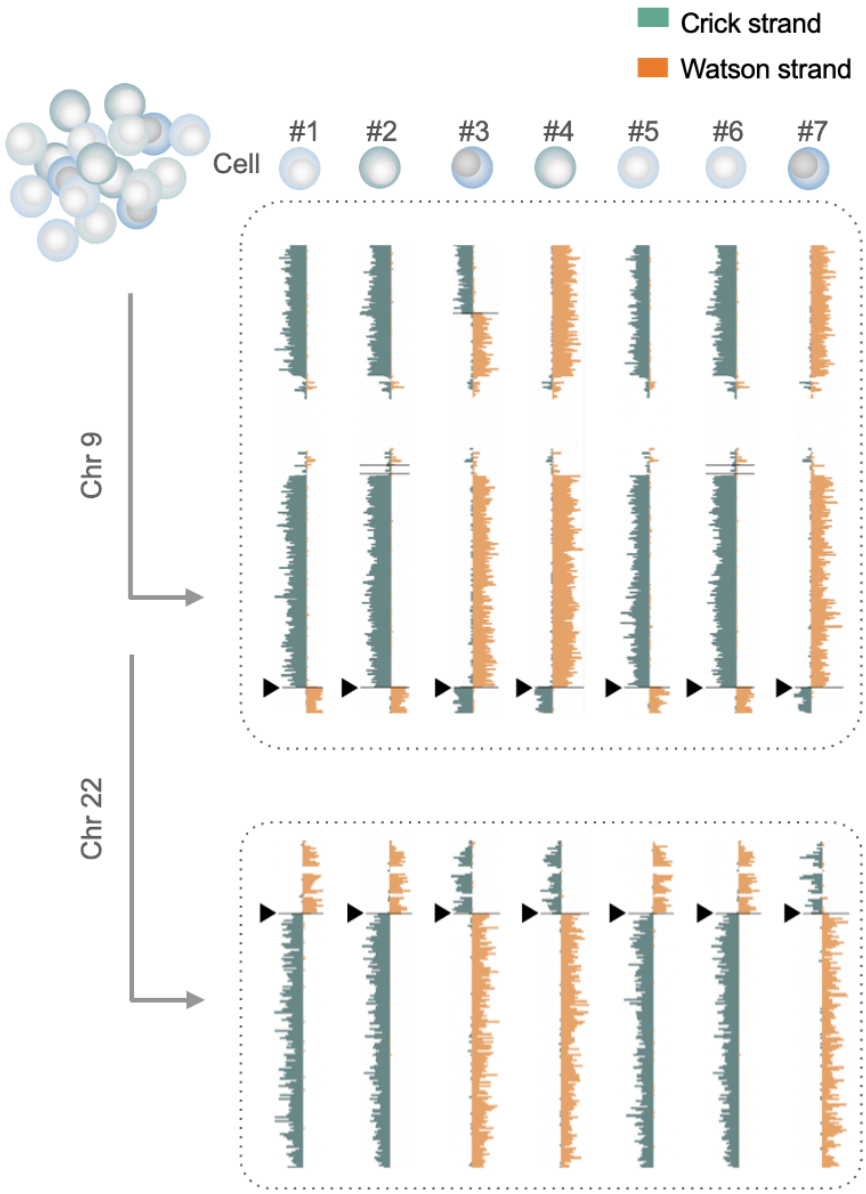
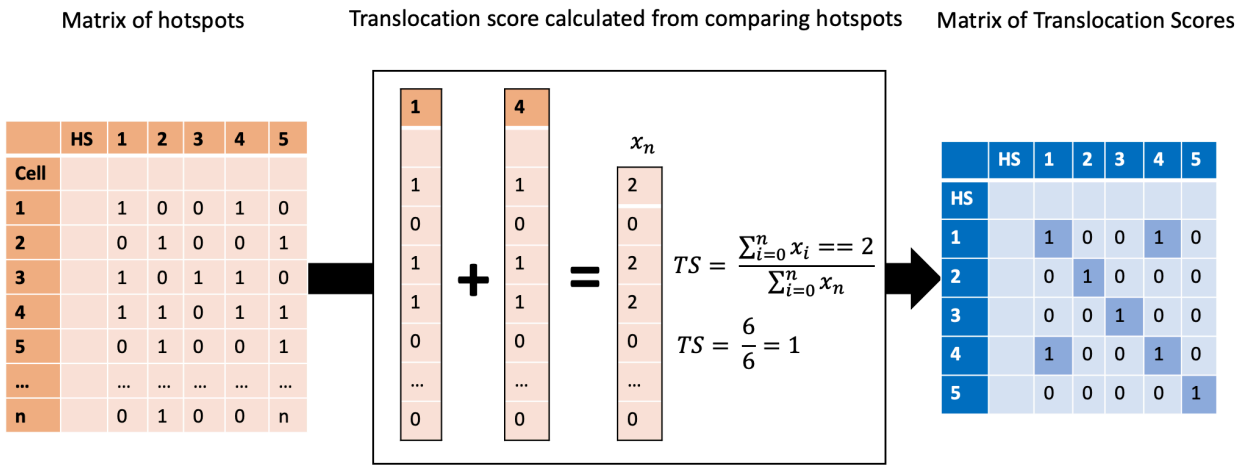



**Figure 3.7 Philadelphia chromosome translocation signature in Strand-seq libraries.**
**Schematic of template strand inheritance patterns present in 7 KBM7 cells. Black arrows point to translocation breakpoint in chromosome 9 and 22.**

Here, I used a custom algorithm to quantify the frequency of recurring strand state switch breakpoints to identify probable translocations (Figure 3.8). I devised a matrix of size $n * m$, where $n$ refers to the number of cells and $m$ refers to the number of hotspots found in all Strand-

seq libraries (Figure 3.8). For each hotspot, I quantify how similar it is to other hotspots in terms of the libraries involved in making up that hotspot. I calculate a Translocation Score, $TS$, for each pair of hotspots by adding each column together and taking the sum of shared libraries divided by the sum of libraries involved (Figure 3.8). The TS from each comparison is placed into new matrix of $m * m$. When I plot this matrix as a heatmap, regions of the genome that resemble probable translocations are highlighted (Figure 3.9). Only one region is highlighted and that reveals the Philadelphia chromosome translocation which is a known translocation in the KBM7 cell line (Figure 3.9).




| | HS | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Cell | | | | | | |
| 1 | | 1 | 0 | 0 | 1 | 0 |
| 2 | | 0 | 1 | 0 | 0 | 1 |
| 3 | | 1 | 0 | 1 | 1 | 0 |
| 4 | | 1 | 1 | 0 | 1 | 1 |
| 5 | | 0 | 1 | 0 | 0 | 1 |
| ... | | ... | ... | ... | ... | ... |
| n | | 0 | 1 | 0 | 0 | n |

| | HS | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| HS | | | | | | |
| 1 | | 1 | 0 | 0 | 1 | 0 |
| 2 | | 0 | 1 | 0 | 0 | 0 |
| 3 | | 0 | 0 | 1 | 0 | 0 |
| 4 | | 1 | 0 | 0 | 1 | 0 |
| 5 | | 0 | 0 | 0 | 0 | 1 |

**Figure 3.8 Algorithm for calling translocations in Strand-seq libraries.**
**A matrix of size $n * m$ is generated, where $n$ refers to the number of cells and $m$ refers to the number of hotspots found in all Strand-seq libraries. A Translocation Score, $TS$, is calculated for each pair of hotspots by adding each column together and taking the sum of shared libraries divided by the sum of all libraries involved. The TS from each comparison is placed into new matrix of $m * m$. High scores in this matrix resemble probable translocations.**

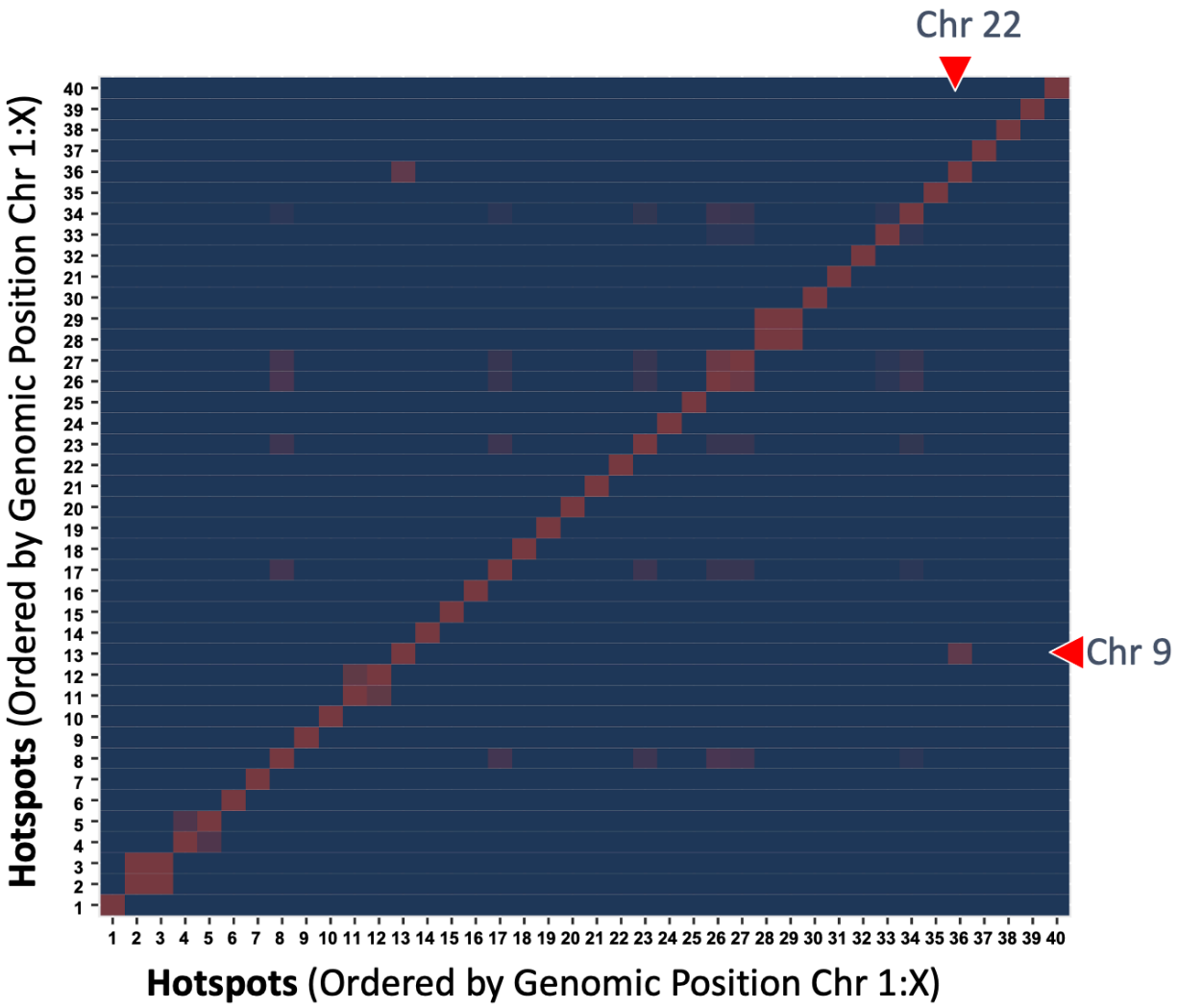



**Figure 3.9 Heatmap of Translocation Score matrix.**
**Columns and rows are organized from chromosome 1 to chromosome X. Low scores are shown in blue, high scores are highlighted in red. The position of hotspots on chromosome 9 and 22 is shown with red arrows.**

### 3.2.4 Bioinformatic approaches to CNAs detection

I used the R package, *AneuFinder,* for the discovery of CNAs in haploid Strand-seq libraries [96]. Using *AneuFinder* on haploid Strand-seq libraries offers one main advantage over other CNA analysis tools that rely solely on changes in read count density to flag changes in copy-number in diploid cells [109]. As previously mentioned, haploid cells can only have reads mapping to one template strand (W or C) for a given chromosome so segments that have reads

mapping to both template strands must be duplicated regions (Figure 3.10). A duplication can present with or without associated changes in strand state genotype in a Strand-seq library (Figure 3.10). Non-tandem duplicated chromosome segments present this way in 50% of cases due to random segregation of sister chromatids and can be reliably detected as a WC region in a haploid cell (Figure 3.10). Two additional steps after running *AneuFinder* were taken to refine the output of this program to putative somatic CNAs.

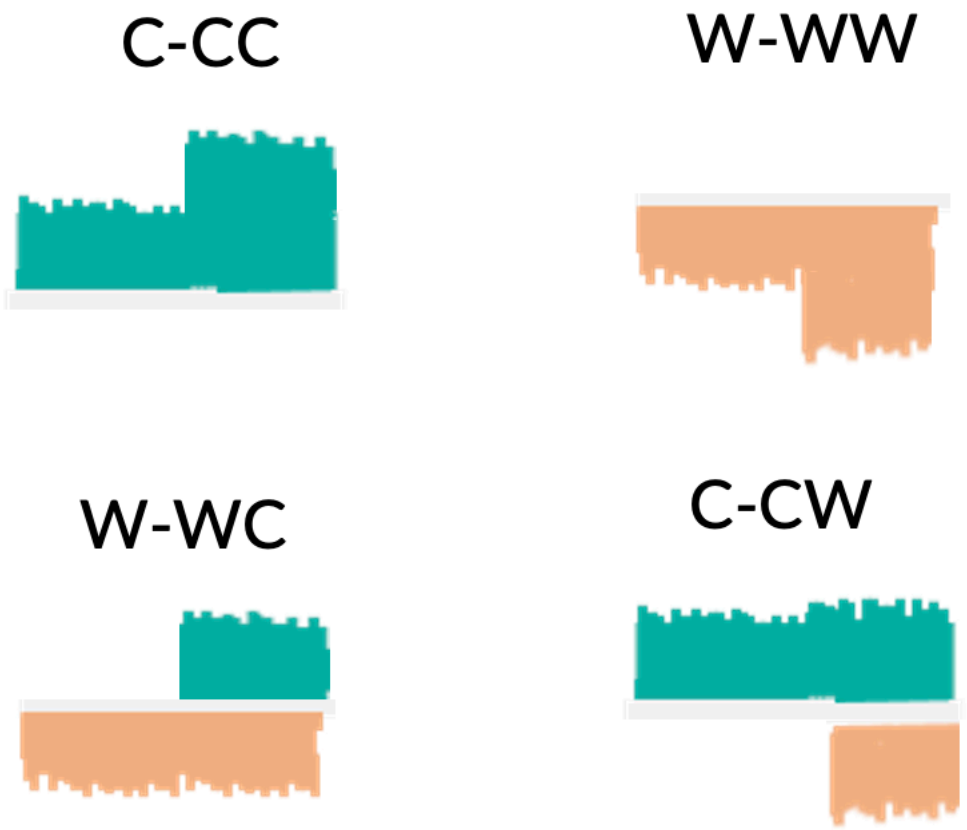



**Figure 3.10 Strand-seq ideograms of CNAs with associated strand-state switches in haploid cells.**

First, I intended to eliminate "germline" events that were present in WT cells prior to the generation of KO lines. If events are present in WT and KO lines, they would not be considered somatic CNAs. I generated a composite BAM file of all WT KBM7 libraries to be used as a reference for *AneuFinder* to normalize binned read counts to a "germline" control using the

parameter ‘variable.width.reference = composite_wt_file.bam’. This step removes CNA calls present in WT cells.

Second, I intended to eliminate false positive calls by focusing on CNAs exceeding 20 Mb in size. This is because it is computationally challenging to distinguish whether changes in read-count are due to amplification biases or are in fact true CNAs [115]. Therefore, this step allows us to validate each event by visually inspecting changes in read count density on chromosome ideograms generated by *BreakpointR* that are large enough to verify as real CNAs. Breakpoints that were marked by a notable change in read count density in *BreakpointR* ideograms were added to the comprehensive CNA callset. Next, I re-ran *AneuFinder* on individual BAM files for WT cells without the use of a variable width reference to identify somatic CNAs that have occurred in WT cells and are not present in all WT cells.

The following settings in *AneuFinder* were used: low-quality alignments (mapping quality score (MAPQ) < 10) and duplicate reads were excluded and read counts in 2 Mb variable-width bins were determined with a 10-state Hidden Markov Model with copy-number states: null-, mono-, di-, tri-, tetra-, penta-, hexa-, septa-, and octasomy. I also set ‘strandseq = TRUE’ and ‘gc.correction = TRUE’.

## 3.3 Results

### 3.3.1 Genome-wide screening for SCEs

I mapped 14,879 SCEs from 1684 KBM7 cells across all RecQ KO lines. I show a genome-wide distribution of SCEs and found strong correlations ($R = 0.93$, $p < 0.001$) between the number of SCEs on each chromosome and chromosome length, suggesting that on a global

scale, these events are randomly distributed across the genome in accordance with previous findings of SCEs from Bloom Syndrome patient cells (Figure 3.11) [25]

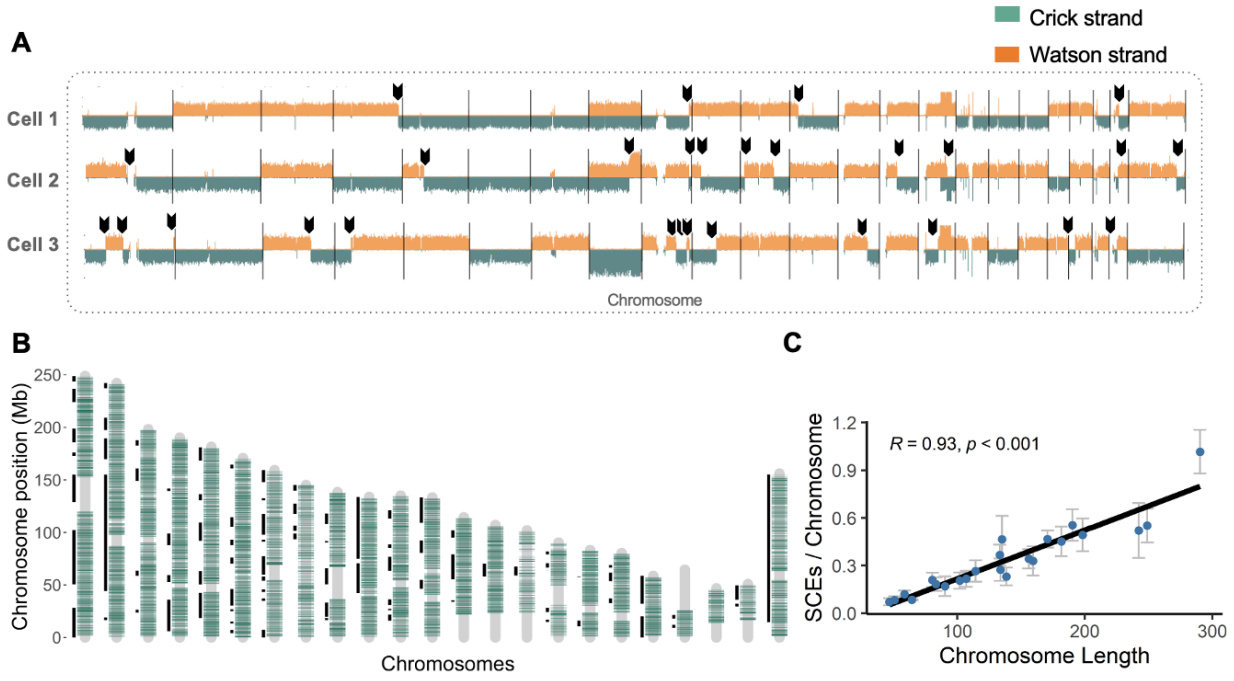



**Figure 3.11 Mapping of SCEs in single cells.**
**(A) Mapping of SCEs in three haploid KBM7 cells using Strand-seq. Directional chromosome ideograms show reads mapping to the Crick (positive) strand of the reference genome in green and reads mapping to the Watson (negative) strand in orange. SCEs are identified as a change in template strand state within a chromosome (arrowheads). (B) Genome-wide summary of SCE density. (C) Correlations between average numbers of SCEs/chromosome/library and chromosome size.**

I found a median resolution for all SCEs of 19.2 Kb (Figure 3.12A). Next, I assessed the accuracy of our bioinformatic approach to detect SCEs using a manually curated benchmark set of SCEs that were generated as a golden standard to compare the performance of different methods. I ran *BreakpointR* with and without any processing steps used the benchmark set to calculate precision and recall at multiple measures of overlap between the SCE calls made by each method and our benchmark set (Figure 3.12B).

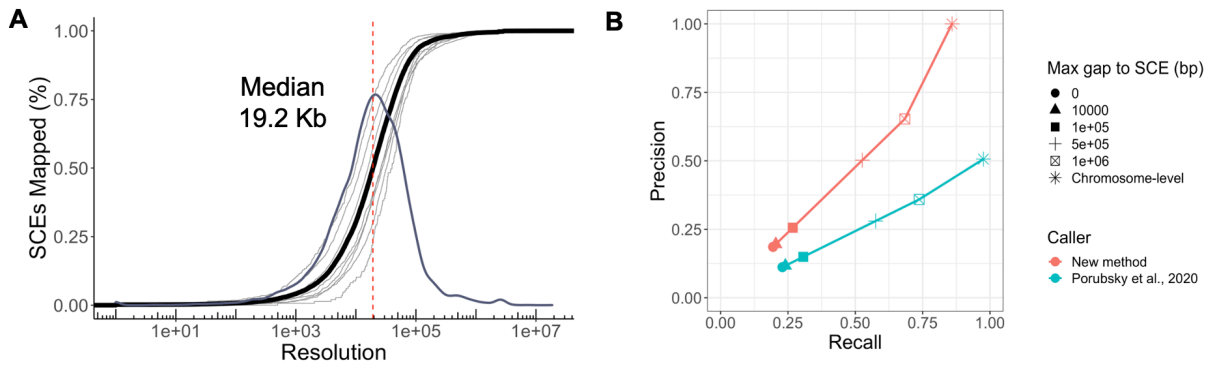



**Figure 3.12 SCE mapping resolution and accuracy.**
**(A) Resolution of SCE mapping in KBM7 cells. Lines represent cumulative density of the total number of SCEs mapped at resolution values indicated below. (B) Precision and recall comparing performance of BreakpointR alone and with additional processing steps mentioned above to remove false positive calls. This plot assesses how many SCEs are appropriately called with direct overlap or within a maximum gap of 10 Kb, 100 Kb, 500 Kb, 1 Mb and on a chromosome-ideogram level.**

Next, we investigated differences in SCE frequency between RecQ KO KBM7 cell lines using good quality Strand-seq libraries. As stated in Section 2.2.2.3, 1684 good quality Strand-seq libraries collected over 21 independent sequencing experiments were pooled for this analysis. There were several genotype-specific differences in SCE levels. Upon knockout of *BLM* or *WRN* helicase, the number of SCEs per haploid genome rose by 2.34 and 1.5-fold, respectively (Table 3.1, Figure 3.13). Upon knockout of *BLM/RECQL5*, there was also an increase in SCE levels by 2.3-fold that was not significantly different than for *BLM* knockout cells alone (Table 3.1, Figure 3.13). There was no significant change in SCE levels upon knockout of *RECQL1* and *RECQL5* (Table 3.1, Figure 3.13).

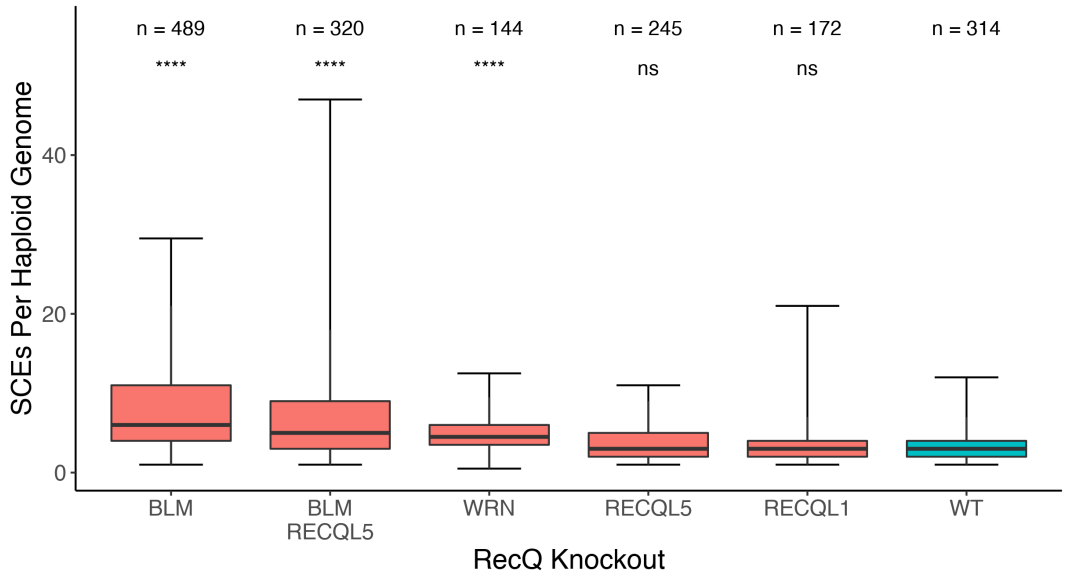



**Figure 3.13 Number of SCEs detected per haploid genome in a single cell division for RecQ helicase single and double knockouts in the KBM7 cell line.**
**Number of cells analyzed (n) is shown above. Statistical significance was evaluated using a two-sample t-test where WT cells are the control group. **** $p < 0.001$, ** $p < 0.01$, not significant (ns) $p > 0.05$.**

| Knockout line | Number of cells | % Haploid cells | Total SCEs | Mean SCEs/haploid genome | Standard deviation SCEs/haploid genome | Standard error of the mean | Fold change in SCEs/haploid genome |
|---|---|---|---|---|---|---|---|
| BLM | 489 | 70% | 5841 | 7.76 | 5.38 | 0.24 | 2.34 |
| BLM/ RECQL5 | 320 | 98% | 2456 | 7.62 | 7.01 | 0.39 | 2.30 |
| RECQL1 | 172 | 43% | 909 | 3.39 | 2.25 | 0.17 | 1.02 |
| RECQL5 | 245 | 93% | 967 | 3.66 | 1.99 | 0.13 | 1.10 |
| WRN | 144 | 0% | 1434 | 4.98 | 2.21 | 0.18 | 1.50 |
| WT | 314 | 99% | 1049 | 3.32 | 1.78 | 0.10 | 1.00 |

**Table 3.1 Comparison of SCE frequency by genotype**

To investigate the large variation in SCE levels in individual BLM and BLM/RECQL5 KO cells (Table 3.1), SCE levels in relation to ploidy were analyzed (Figure 3.14). *WRN* KO cells were omitted from this analysis since all *WRN* KO cells were found to be diploid (Table 1).

After correction for genome content, a significant increase in SCEs in diploid compared to haploid cells with the same targeted disruption of the *BLM* gene was found (Figure 3.14, Table 3.2). Haploid *BLM/RECQL5* KO cells had significantly higher SCE levels than diploid *BLM/RECQL5* KO cells with the caveat that there were only 6 diploid cells (Figure 3.14, Table 3.2). There were no changes in SCE levels between haploid and diploid cells for *RECQL5, RECQL1* and WT cells (Figure 3.14, Table 3.2).

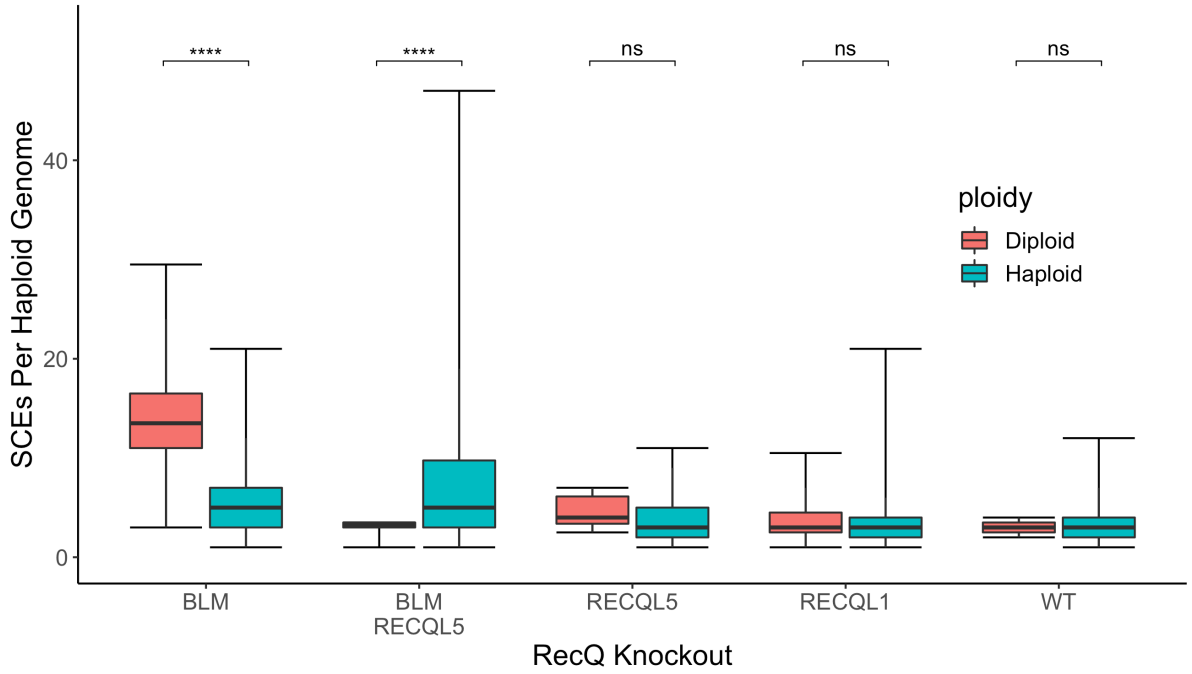



**Figure 3.14 Number of SCEs detected per haploid genome in a single cell division for RecQ helicase single and double knockouts in the KBM7 cell line grouped by ploidy of cells.**
**Statistical significance was evaluated using a two-sample t-test between haploid and diploid cells of the same gene knockout. **** $p < 0.001$, ** $p < 0.01$, not significant (ns) $p > 0.05$**

| Knockout line | Ploidy | # Of cells | # Of SCEs | Mean SCEs/haploid genome | Standard deviation SCEs/haploid genome | Standard error of the mean | Fold change in SCEs/haploid genome | Adjusted p-value for dip/hap comparison |
|---|---|---|---|---|---|---|---|---|
| BLM | 1 | 341 | 1748 | 5.13 | 2.87 | 0.16 | 1.55 | |
| BLM | 2 | 148 | 4093 | 13.83 | 4.92 | 0.40 | 4.61 | 2.28E-48 |
| BLM/ RECQL5 | 1 | 314 | 2421 | 7.71 | 7.04 | 0.40 | 2.33 | 2.25E-07 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| BLM/ RECQL5 | 2 | 6 | 35 | 2.92 | 0.97 | 0.40 | 0.97 | |
| RECQL1 | 1 | 74 | 257 | 3.47 | 2.92 | 0.34 | 1.05 | |
| RECQL1 | 2 | 98 | 652 | 3.33 | 1.58 | 0.16 | 1.11 | 1.00E+00 |
| RECQL5 | 1 | 229 | 825 | 3.60 | 2.00 | 0.13 | 1.09 | |
| RECQL5 | 2 | 16 | 142 | 4.44 | 1.60 | 0.40 | 1.48 | 3.14E-01 |
| WRN | 1 | 0 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | |
| WRN | 2 | 144 | 1434 | 9.96 | 2.21 | 0.18 | 3.32 | N/A |
| WT | 1 | 314 | 1037 | 3.30 | 1.78 | 0.10 | 1.00 | |
| WT | 2 | 2 | 12 | 3.00 | 1.41 | 1.00 | 1.00 | 1.00E+00 |

**Table 3.2 Comparison of SCE frequency by genotype grouped by ploidy of cells**

After using BreakpointR's auxiliary hotspotting function to identify recurring breakpoints and remove non-SCE SVs by examining read distributions on the UCSC Genome Browser, I confirmed the presence of regions with significantly clustered SCEs. One was near the FRA20A CFS (Figure 3.15). This site had 22.3 SCEs/Mb relative to only 0.89 SCEs/Mb on average across the whole genome (Figure 3.15).

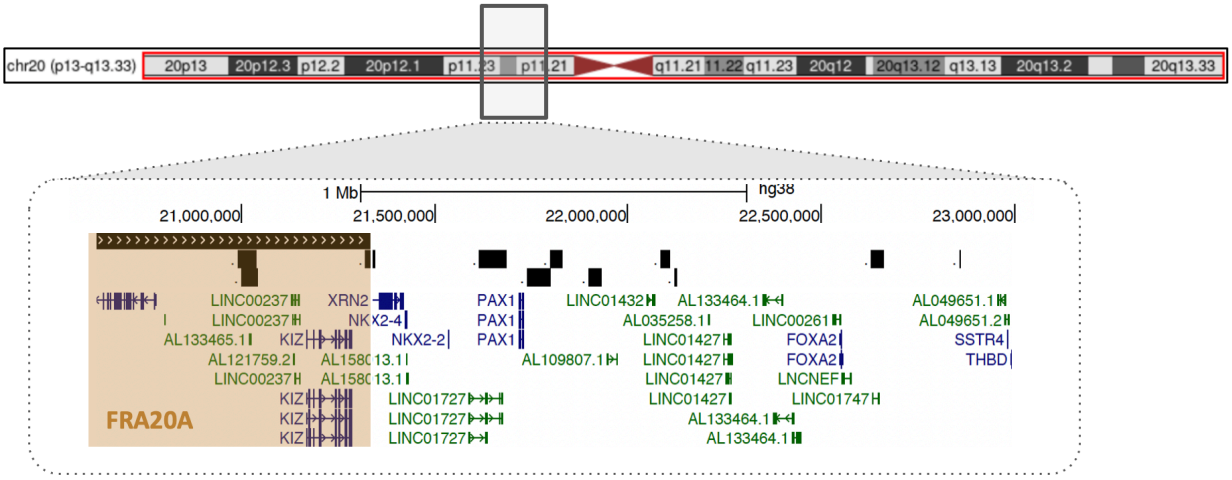



**Figure 3.15 UCSC Genome Browser example of SCE hotspot within FRA20A common fragile site**
**CFS shown in orange box. Confidence intervals for detected SCEs from individual cells are depicted as black bars. Coordinates for CFS were collected from Kumar et al., 2019** [116].

### 3.3.2 Genome-wide screening for translocation

As mentioned in Section 3.2.3, the Philadelphia chromosome translocation was found in our dataset using our custom algorithm. The breakpoint confidence interval for this translocation is 18.7 Kb and 13.7 Kb for chromosome 9 and 22, respectively (Figure 3.16). Conventional methods for genome wide screening for translocations such as karyotyping and Interphase FISH have a resolution of 5 Mb and 50 to 100 kb, respectively [117]. Although some methods such as ChromPET do have base-pair resolution, they require prior knowledge for targeted sequencing whereas our method is a naïve approach for searching genome-wide for unknown translocations [117].

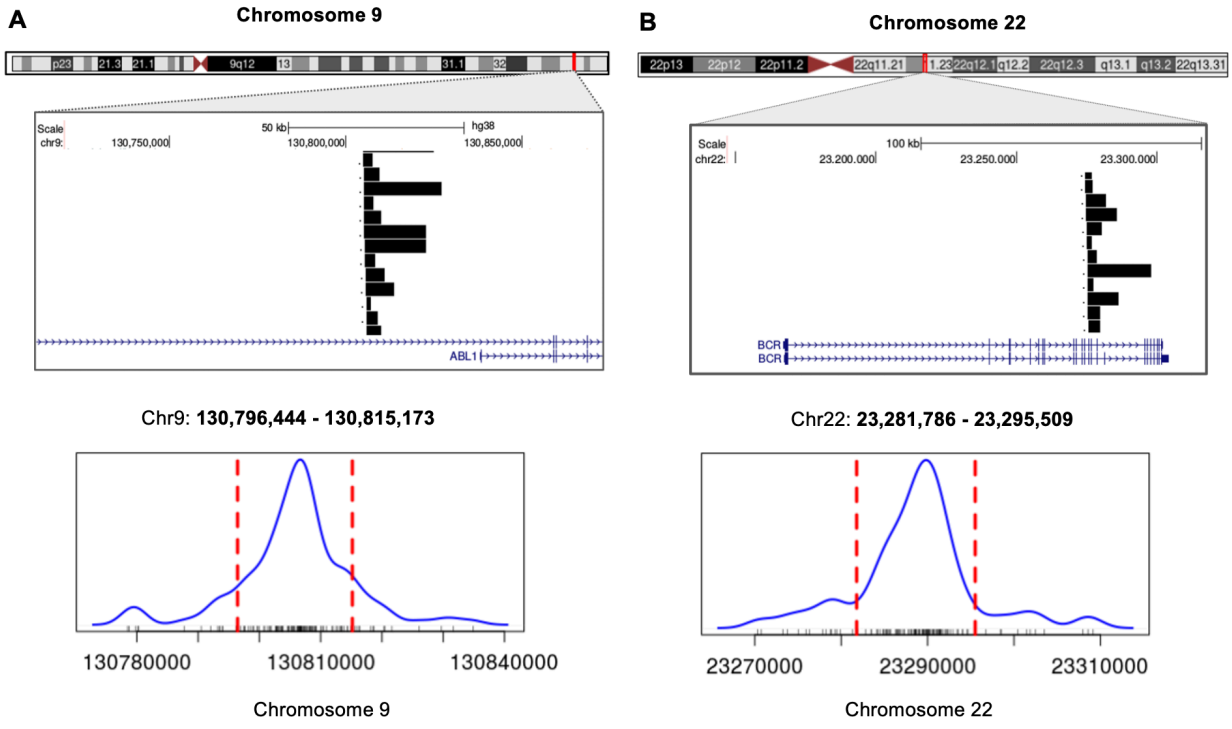



**Figure 3.16 Translocation resolution for Philadelphia chromosome breakpoints.**
**(A) UCSC Genome Browser image of translocation breakpoints for chromosome 9 shown with confidence intervals for 13 breakpoints from individual cells sorted by start position (black bars) (top). Density of all breakpoint intervals for chromosome 9 with 99% confidence interval shown in dashed red line (bottom). B) UCSC Genome Browser image of translocation breakpoints for chromosome 22 shown with 12 example confidence intervals for breakpoints from individual cells sorted by start position (black bars) (top). Density**

**of all breakpoint intervals for chromosome 22 with 99% confidence interval shown in dashed red line (bottom).**

### 3.3.3 Genome-wide screening for CNAs

I first searched for germline CNAs present in all our WT cells that were not removed by *AneuFinder*. Two "germline" duplications in WT cells were detected using a composite BAM file of WT cells. Next, I looked for somatic CNAs that were present in single WT cells only and verified each one independently by inspecting read depth and strand state switches on binned read count ideograms (Figure 3.17A). 30 somatic CNAs in WT cells were found. All together I found 178 somatic duplications (blue) and 451 somatic deletions (red) in single and double RecQ helicase knockout KBM7 lines (Figure 3.17B).

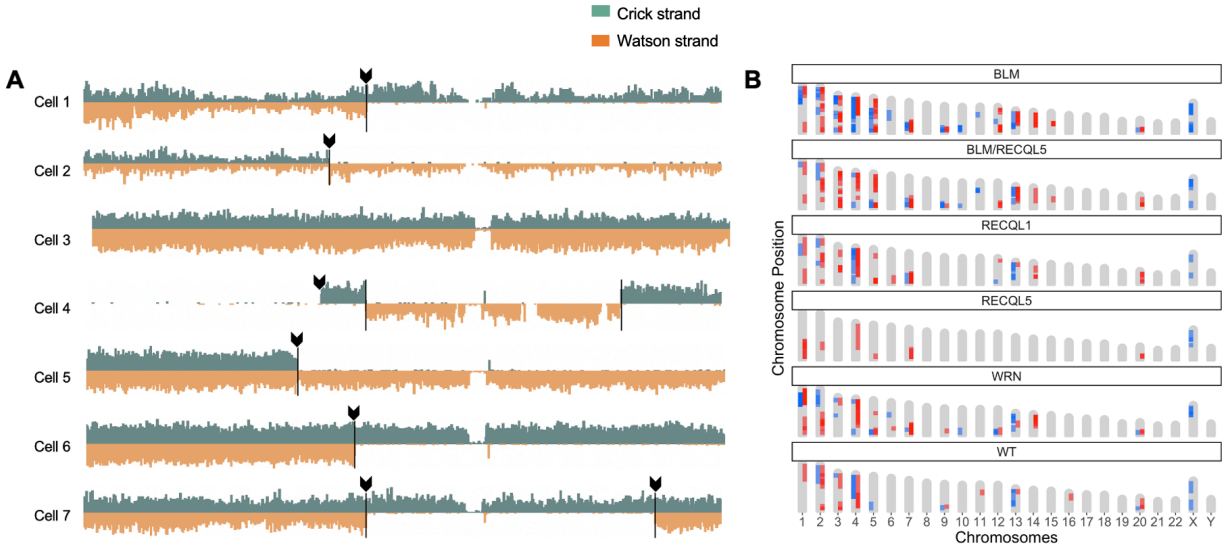



**Figure 3.17 Mapping of CNAs in single cells.**
**(A) Genome-wide summary of CNAs in a composite BAM file of all WT cells. (B) Examples of somatic duplications and deletions from seven cells on chromosome 3. (C) Genome-wide summary of CNAs from separate RecQ knockout lines. Duplications shown in blue. Deletions shown in red.**

CNA frequency was calculated by the number of CNAs per cell and showed a significant increase in deletions upon knockout of *BLM* or *WRN* and only a slightly significant increase upon knockout of *BLM/RECQL5* compared to WT cells (Figure 3.18A). There was also a

significant increase in the frequency of duplications upon knockout of *WRN* and a slightly significant increase upon knockout of *BLM* and *WRN/RECQL5* (Figure 3.18A). No changes in CNA frequency were observed upon knockout of RECQL1 and WRN (Figure 3.18A). Interestingly, the resolution at which we can observe CNA breakpoints allows us to discern clonally derived events from events that have occurred independently in the same region in two different cells. A representative example of this is shown where two CNA breakpoints on chromosome X ideograms appear to have derived from the same parent cell but the UCSC Genome Browser plotting of each cell's reads highlights two distinct breakpoints in and around the PCDH11X gene suggesting these events are not clonally derived but arose independently (Figure 3.18B).

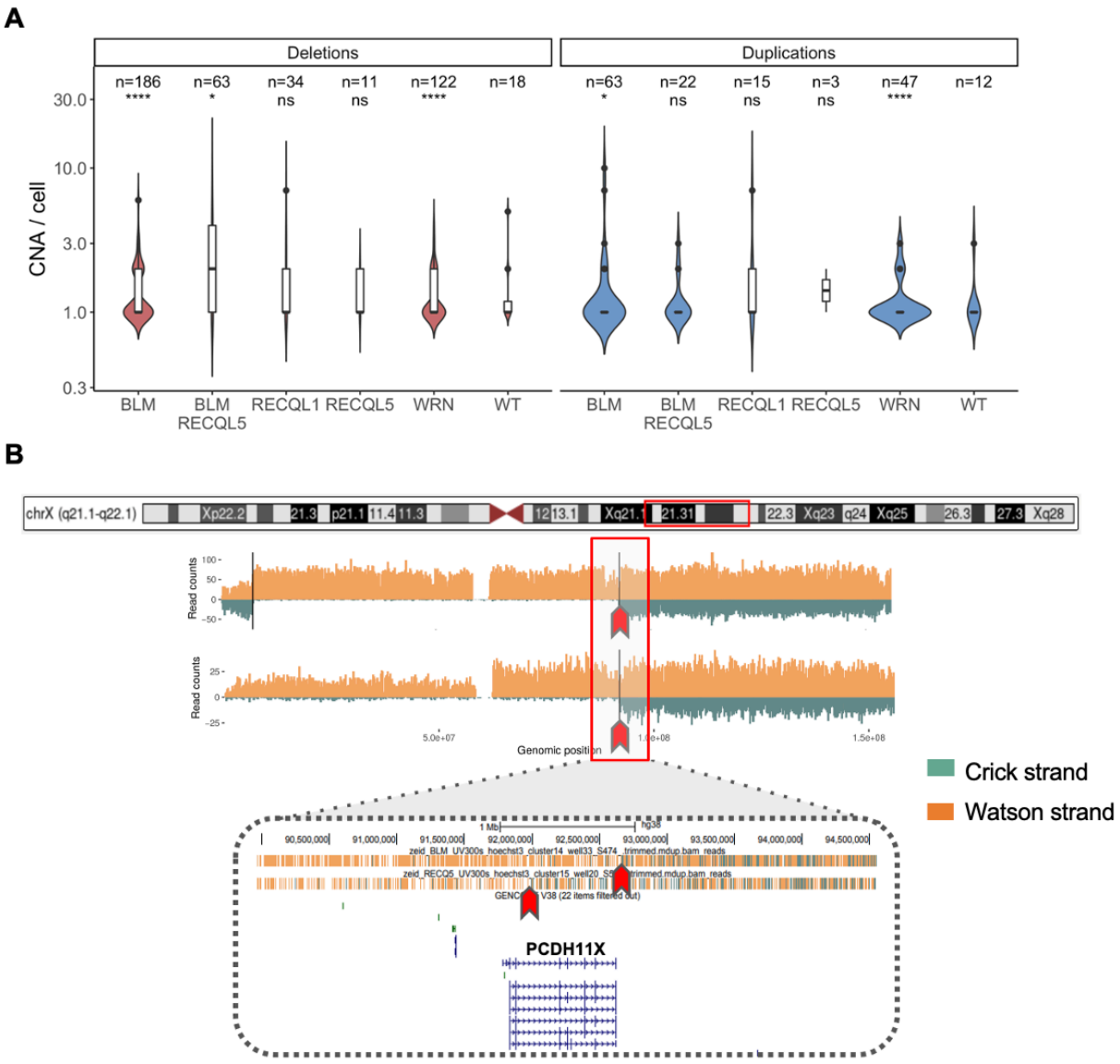



**Figure 3.18 Analysis of somatic CNAs in single cells.**
**(A) Frequency of CNAs in single RecQ-deficient cells with the number of cells analyzed listed above. (B) Example duplication on chromosome X reveals distinct duplication breakpoints within PCDH11X gene when examined on UCSC Genome Browser. Number of CNAs analyzed (n) is shown above. Statistical significance was evaluated using a two-sample t-test where WT CNAs are the control group. **** $p < 0.001$, ** $p < 0.01$, not significant (ns) $p > 0.05$**

## 3.4 Discussion

Strand-seq offers a novel approach for interrogating the genome of single cells for chromosomal instability by mapping SCEs, translocations, and CNAs to the genome at kilobase resolution.

In our datasets of high confidence SCE calls, I show there are notable genotype-specific differences in SCE frequency that I was able to quantify. There was a significant increase in SCE frequency upon knockout of *BLM* and *BLM/RECQL5* but there was no significant increase in *RECQL5* KO cells or between *BLM* and *BLM/RECQL5* KO cells suggesting BLM and RECQL5 may have redundant functional roles in preventing SCE formation. This contradicts previous findings that these enzymes have non-redundant roles in suppressing sister chromatid recombination [62]. Changes in SCE levels in *BLM* KO cells were lower than what has previously reported with cells from patients with Bloom Syndrome or murine cells lacking BLM [25]. There was also a large standard deviation of SCE frequency in *BLM/RECQL5* KO cells [25]. The number of SCEs in diploid cells was increased more than the expected two-fold relative to haploid *BLM* KO cells. These differences were significant suggesting that the haploid nature may alter the sensitivity to BLM helicase deficiency. Possible explanations for this remain unclear but it seems possible that doubling the DNA content in a cell may lead to more than double the replication conflicts due to saturation of the DNA repair machinery needed to prevent SCE formation. This trend was likely not observed in in *BLM/RECQL5* KO cells because there were only 6 diploid cells in our analysis leading to underpowered statistical tests.

I initially show that on a global scale, these events appear to be randomly distributed across the genome. However, I also found regions of the genome harboring a significant number of SCEs relative to neighboring regions known as SCE hotspots, suggesting that on a local scale

these events are perhaps not randomly distributed. This is consistent with previous findings that SCEs can occur more frequently in CFSs due to replication stress and replication fork stalling [25]. However, it remains unclear what factors in our cells may predispose cells to having these SCE hotspots and thus further studies are needed to identify if specific genetic elements that may be contributing to the occurrence of these events. It should be noted to improve the resolution of SCE breakpoints, we can increase the depth of sequencing albeit at increased sequencing costs. This approach will allow SCEs to be mapped more finely to genetic regions that are problematic for replication and transcription machinery.

Furthermore, I did find some genotype-specific differences in CNA frequency. From the significant increase in somatic deletions and duplications from *BLM* and *WRN* KO lines, these genes likely play a role in preventing somatic CNAs. Interestingly, two well defined CNA breakpoints from two different cells were found to occur within the same gene, PCDH11X. The resolution of these breakpoints revealed these CNAs were far enough apart within this gene to be considered two separate events arose in two cells independently rather than in one original ancestral cell. This may suggest this gene is problematic for replication and prone to aberrant repair outcomes in the form of CNAs however, I did not observe any other instances of this. Together, these findings support the role of WRN and BLM helicases in preventing SCEs and CNAs.

# Chapter 4: Role of BLM and RECQL5 in DNA repair

## 4.1 Introduction

Since their discovery in the mid 1970s, many DNA helicases have emerged as important DNA repair proteins [118]. Their critical role in DNA repair pathways is highlighted by their association with aging and cancer-prone disorders [118]. Many helicases assist in normal DNA repair by unwinding and resolving alternatively folded DNA structures that can arise during transcription, recombination or repair of DSBs [3]. Repair of DSBs is a highly dynamic process that involves a balance between multiple, distinct pathways. During HR, a wide variety of DNA structures can be formed between broken DNA ends and homologous donor template molecules that need to be resolved by specific nucleases and helicases. Mutations in such enzymes can result in failure to properly resolve intermediate joint molecules and aberrant repair outcomes such as gene conversion and loss of heterozygosity may contribute to genomic instability, cancer predisposition and the progressive deterioration of normal cell function [8], [19], [119]. The RecQ class is of particular interest because mutations in four of five genes in this class are linked to disorders of genome instability, characterized in some cases by aging or cancer predisposition. Cells deficient in these helicases exhibit genome instability evidenced by high spontaneous somatic mutation rates, loss of heterozygosity and an elevated frequency of SCEs. Although RECQL5 has not yet been associated with a specific disorder, recent studies support the hypothesis that RECQL5 can resolve intermediate DNA repair structures such as stalled replication resulting from the collision of replication machinery with lesions or secondary structures that would otherwise result in template switching and SCEs. Researchers have also shown an additive phenotype when RECQL5 is knocked out with another helicase in the class,

BLM, further supporting the non-redundant role of these genes in DNA repair. Therefore, I wanted to investigate the unique role of RECQL5 relative to other genes in the RecQ class.

As discussed in Chapter 1, there are many challenges studying the precise role of DNA repair and associated enzymes. This in part stems from genetic pleiotropy of the many enzymes involved in DNA repair [3]. Helicases with overlapping functions may have redundant roles or they may be regulated for specific use. For example, Srs2 is a yeast helicase responsible for regulating RAD51 displacement from ssDNA and promoting crossover-avoidance [118]. In humans, all 5 members of the RecQ class are homologs of Srs2, each of which have been shown to have both unique and overlapping roles in DNA repair, G4 unwinding, end resection and RAD51 displacement activity [118]. Secondly, the outcome of faulty DNA repair in the form of complex somatic SVs remain elusive to conventional short-read sequencing technologies

As discussed in Chapter 2, new sequencing approaches such as Strand-seq can help uncover helicase function by mapping the location of complex genome alterations in cells in which helicase function is lost [120]. Strand-seq was developed in 2012 and has been used to map sister chromatid exchange events (SCEs) to the genome at kilobase resolution in cells deficient in the BLM helicase from a Bloom Syndrome patient [25], [56]. This study confirmed that murine cells lacking the BLM helicase have up to 10-fold more SCEs compared to healthy controls [25]. SCEs in BLM cells were enriched in known fragile sites and near G4 forming motifs in the genome, providing further support for the notion that the BLM helicase is required to prevent exchanges of DNA strands during recombination and repair at specific genomic locations [25]. SCEs are a useful indicator of genomic instability and indicate replication fork stalling and template switching have occurred due to gaps in single stranded DNA or DNA lesions [43]. It is well documented that the genomic context plays a role in biasing repair

outcomes because of replication impeding structures, such as G-quadruplexes, and non-allelic recombination susceptible areas, such as LCRs, that arise in repetitive areas of the genome [3], [8]. Therefore, we hypothesize that SCEs are not randomly distributed across the genome upon KO of each RecQ helicase. In this chapter, I test this hypothesis and clarify the role of specific helicases in resolving different kinds of replication barriers by investigating the genomic context of SCEs.

I used Strand-seq to map SCEs in KBM7 cell lines deficient in RECQL1, BLM, WRN, RECQL5 and BLM/RECQL5 together. I collected 14,879 SCEs from 1684 cells across all cell lines in order to identify regions in the genome where SCEs are more likely to occur. These studies yielded novel information about molecules and pathways involved in recombination and sister chromatid exchange mechanisms in mammalian cells. Such information is essential to elucidate currently poorly understood medical conditions and inform therapeutic strategies in cancer.

## 4.2 Methods

### 4.2.1 Generation of comprehensive SCE and SV call sets

It was previously shown that SCEs can be detected using Strand-seq and that elevated rates of spontaneous SCEs can be induced by knocking out RecQ helicases (Chapter 3). Several hundred single-cell Strand-seq libraries were generated for each of the RecQ KO cell lines across 21 independent sequencing experiments and pooled together for this analysis. I introduced accurate and scalable bioinformatic approaches to identify and map SCEs to their exact locations in the genome (Chapter 3). I showed that SCE rate per haploid genome for cells deficient in BLM, WRN and BLM/RECQL5 were ~2x higher SCE rates than in WT cells. Here, my

comprehensive SCE call sets was used to investigate if SCEs are not randomly distributed across the genome. As an additional control, I collected SCEs using the methods discussed in Chapter 3 from Strand-seq libraries generated from the EBV-transformed lymphocyte cell line, NA12878, using the same library preparation methods discussed in Chapter 2.

### 4.2.2 Bioinformatic tools for assessing the enrichment of SCEs with genetic elements

I wanted to investigate the possibility that SCEs may occur preferentially at certain genomic features of interest (FOI). However, statistically assessing the relation between a set of genomic regions and other genetic FOIs poses a few challenges. For one, the position of each SCE is unique to the cell it occurred in, and it is extremely rare for two SCEs to occur in the exact same position in two cells, thus the location of SCEs cannot be directly compared across cells. Secondly, it is difficult to assign a significance level to the degree of association between two sets of genetic regions [121]. For example, if nearly all SCEs overlap with one FOI it might be reasonable to assume that SCEs are associated with this FOI however, this could in fact be coincidental or owing to the abundance or the size of the FOI in the genome [121]. For this, an expected (random) distribution is needed for comparison to say for certain if SCEs are in fact associated with a given FOI. Lastly, when using a randomization-based approach for devising a random distribution, we need to account for the complexity of the human genome [121]. For example, the human genome is made up of 46 separate molecules in diploid cells and 23 molecules in haploid cell and each of which consists of unmapped gaps in the human reference genome that need to be accounted for when assessing the overlap of any genetic elements [121].

Here, I introduce a robust approach to statistically assess the association between SCEs and a FOI using a permutation model to simulate an expected background frequency of

association. I have used several permutation models to generate random distributions of SCEs; however, I found, *RegioneR,* to be a highly reliable bioinformatic tool for implementing a permutation tests involving genomic regions [121]. Briefly, with each permutation test the number of overlaps between SCEs from a particular cell line and a FOI were counted as the experimentally reported value (Figure 4.1). Next, all SCE coordinates were randomly shuffled along the same chromosome and overlaps are recounted (Figure 4.1). Annotated assembly gaps are excluded from possible shuffled SCE coordinates to prevent shuffling into regions where few to zero reads map. This permutation is repeated 1000 times to generate a distribution of randomly simulated overlaps, or permuted values (Figure 4.1). The significance levels for each enrichment analysis were calculated as follows. Both the experimentally reported or observed value and permuted values were normalized to the median permutated value to determine the relative enrichment of experimentally reported SCEs over an expected, randomized distribution (Figure 4.1). Any experimental overlap that lies outside of the 95% confidence interval of the randomized distribution was given a *p*-value below 0.05 and was deemed significant (Figure 4.1). Experimental overlaps lying outside of the entire permuted range were given a *p*-value below 0.001 (Figure 4.1).

For each RecQ KO line, a permutation analysis was performed to calculate the frequency of SCE regions overlapping with an FOI and compared it against the expected background frequency. Next, the significance of enrichment or depletion from one KO line to that of WT cells was compared in order to draw conclusions based on trends that were present in a KO line but not in our WT line.

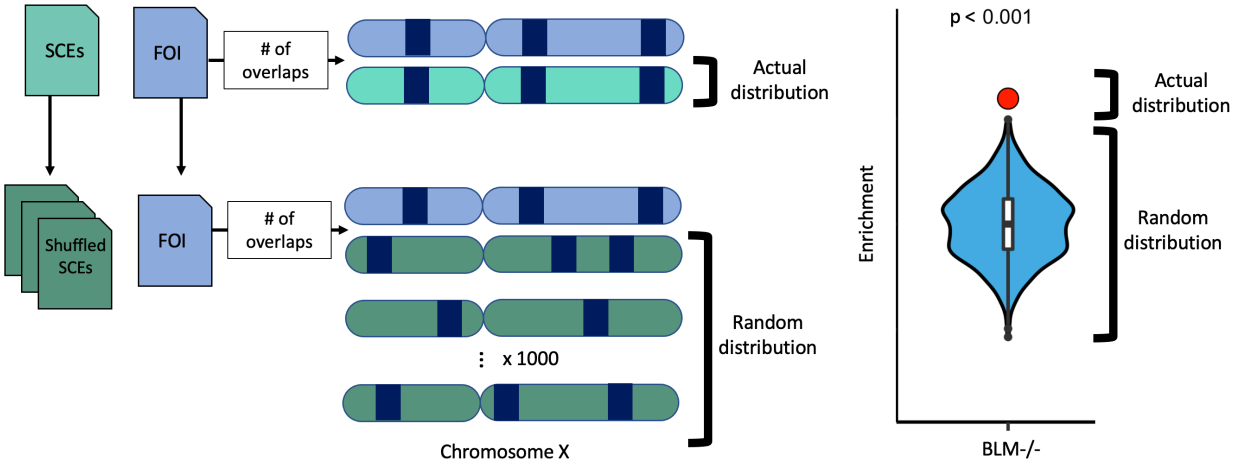



**Figure 4.1 Enrichment analysis workflow using permutation tests.**
**Experimental overlaps of SCEs with a FOI are counted and compared to a randomized distribution of overlap between permuted SCEs and the same FOI. Both the experimentally reported value and permuted values were normalized to the median permutated value to determine the relative enrichment of experimentally reported SCEs (red dot) over an expected, randomized distribution (blue violin plot). Experimental overlaps lying outside of the entire permuted range were given a *p*-value below 0.001.**

### 4.2.3 Genetic element datasets

Here, the association between SCEs and genes with potential G4s was investigated. Transcriptional activity is a well-known cause of genome instability due to transcription-replication conflicts and R-loop formation. Multiple RecQ helicases are also known to bind and unwind G4 structures in vitro. Additionally, G4s have been shown to frequently associate with gene promoters and gene bodies [24]. Below I discuss how I obtained datasets for each of these FOI.

#### 4.2.3.1 Collection and clustering of genes from KBM7 cell line

To investigate the association of genes with SCEs, gene body annotations were obtained from Ensembl (GRCh38.p13) and genes were divided by both transcriptional activity and gene essentiality (Figure 4.3). First, transcriptional activity was assessed using KBM7 RNA-seq data obtained from Rodríguez-Castañeda et al [122]. Genes with the number of fragments per

kilobase of processed transcript per million fragments mapped (FPKM) values of 0 were designated as non-transcribed genes (n = 21,244) (Figure 4.3A). The remaining 19,320 genes with an FPKM > 0 were sorted by FPKM and split into three even groups: lowly expressed genes (n = 6,440), moderately expressed genes (n = 6,440) and highly expressed genes (n = 6,440) (Figure 4.2, Figure 4.3A).

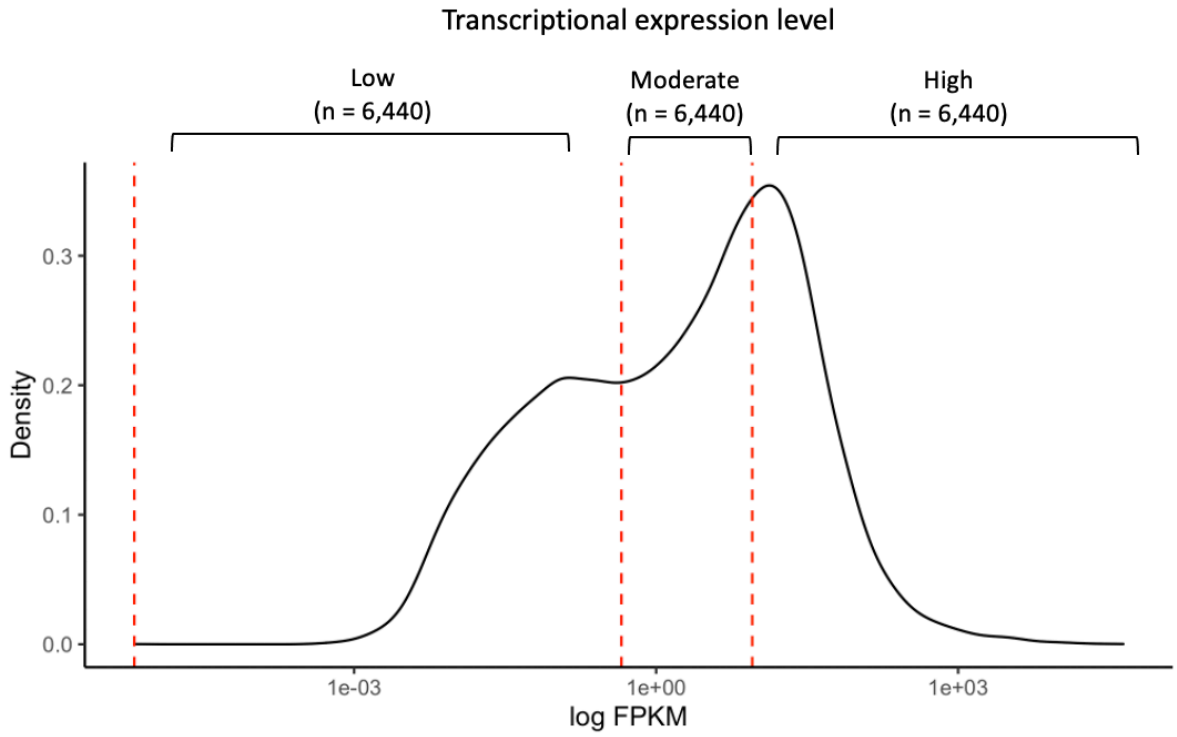



**Figure 4.2 Density distribution of FPKM values across three transcriptional activity levels. Dotted red lines mark boundaries of FPKM values for each level of transcriptional activity.**

Next, genes were also clustered by essentiality for proliferation and survival in the human cancer cell line, KBM7 [123]. Gene essentiality was determined using data from a genome-wide sgRNA library screen that assesses the fitness cost associated with inactivation of each gene [123]. A CRISPR scores (CS) is assigned to each gene as the $\log_2$ fold change in the abundance of sgRNAs targeting that gene and significance values were assigned. Using a cut-off of $p < 0.05$

identified 2,165 genes as essential with the remaining 20,840 genes deemed non-essential (Figure 4.3B).

The overlap of genes clustered by transcriptional activity with genes clustered by essentiality was investigated. I found 81% of essential genes are considered highly expressed versus 45% of non-essential genes (Figure 4.3 C-D). I also found 7% of essential genes were lowly expressed versus 20% of non-essential genes (Figure 4.3 C-D).

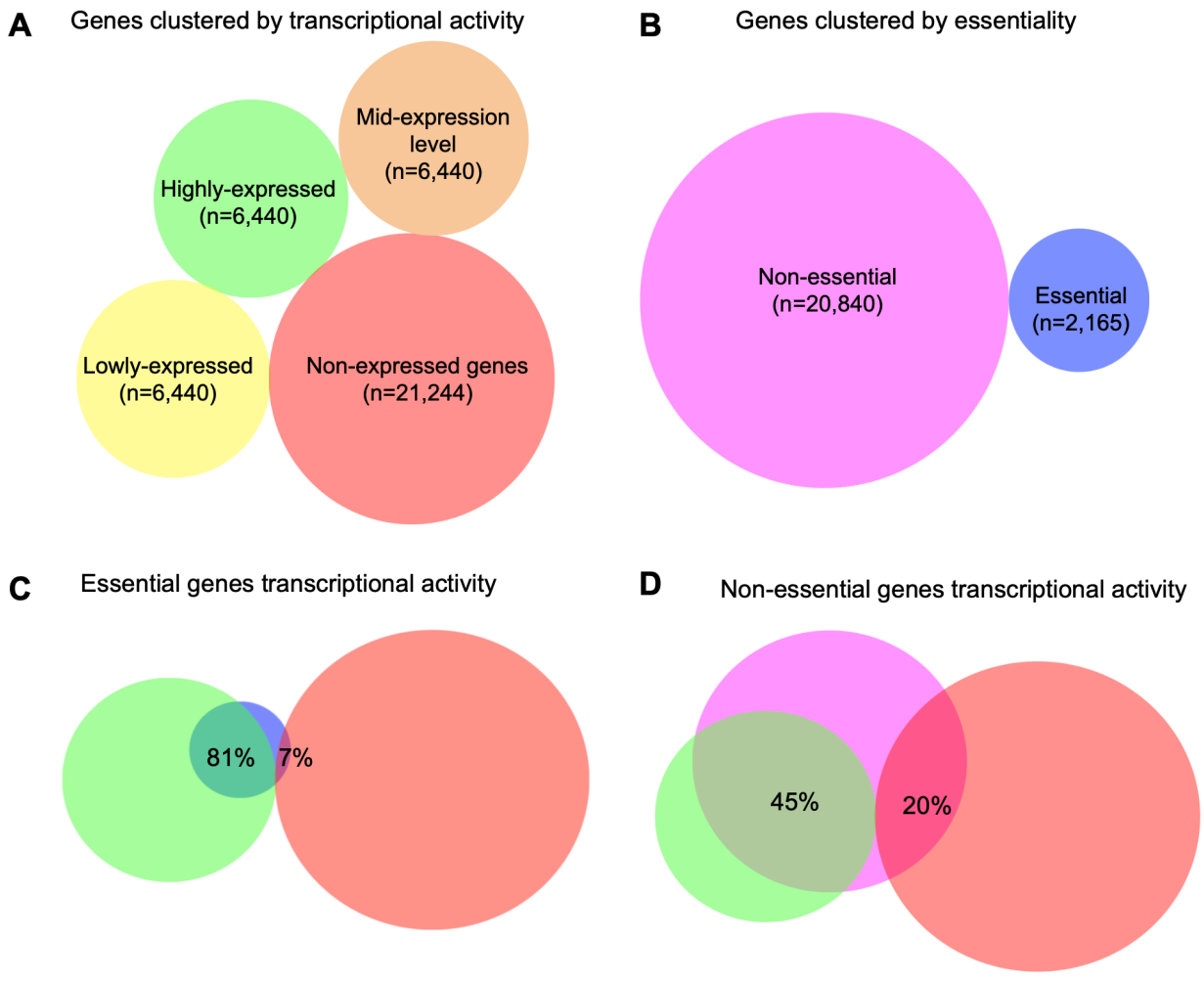



**Figure 4.3 Venn diagrams of genes clustered by transcriptional activity and gene essentiality.**
**A Transcriptional activity of genes in KBM7 cell line. B Essentiality of genes for proliferation in KBM7 cell line. C Overlap of essential genes with highly and lowly expressed genes. Percentage of essential genes that are shared with each group of transcriptional activity is shown. D Overlap of non-essential genes with highly and lowly expressed genes. Percentage of non-essential genes that are shared with each group of**

**transcriptional activity is shown. RNA-seq data was obtained from Rodríguez-Castañeda et al. (2018) and gene essentiality was determined from CS values obtained from Wang et al. (2015).**

#### 4.2.3.2 Collection of experimentally reported potential G4s

To investigate the association of G4s with SCEs, datasets of putative G4 structures in the genome were obtained from a G4 sequencing method that identified structures forming under physiological $K^+$ conditions at PQSs in the genome [22]. Marsico et al., 2019 defined PQSs using the following sequence motif: $G_{2+}N_{1-12}G_{2+}N_{1-12}G_{2+}N_{1-12}G_{2+}$ [22]. As discussed in Chapter 1, an equilibrium exists between PGS and G4s and methods for detecting G4s may alter the equilibrium between PQSs and G4s from native *in vivo* conditions. Therefore, all mentions of G4s herein Chapter 4 are considered potential G4s.

## 4.3 Results

Collisions between replication machinery and transcription machinery or secondary structures such as G4s are potential sources of SCEs. Therefore, I investigated how frequently SCEs occur at G4 quadruplexes and protein coding genes in RecQ KO cells and WT KBM7 cells by performing permutation analyses. I also incorporated a large dataset of SCEs from the diploid EBV-transformed B-lymphocyte cell line, NA12878, to serve as an additional control for ploidy. Enrichment analysis performed using gene datasets was performed using a 100 Kb SCE size cut-off. However, G4s can occur more frequently than genes throughout the genome (~ 8.6 Kb on average) and performing enrichment analysis with SCEs with large confidence intervals would result in increased noise because of the high likelihood of permutated SCE regions overlapping with G4s due to their large size. Therefore, enrichment analysis performed using G4 datasets was performed using a 10 Kb SCE size cut-off.

### 4.3.1 Association between SCEs and genes containing potential G4 structures

There were non-significant SCE enrichment with protein coding genes in all cell lines (Figure 4.4A). Interestingly, SCE enrichments for G4s in RECQL1 and BLM cell lines were detected but no such trend were found in WT cell lines, suggesting some sort of relationship exists between RECQL1, BLM and G4s (Figure 4.4).

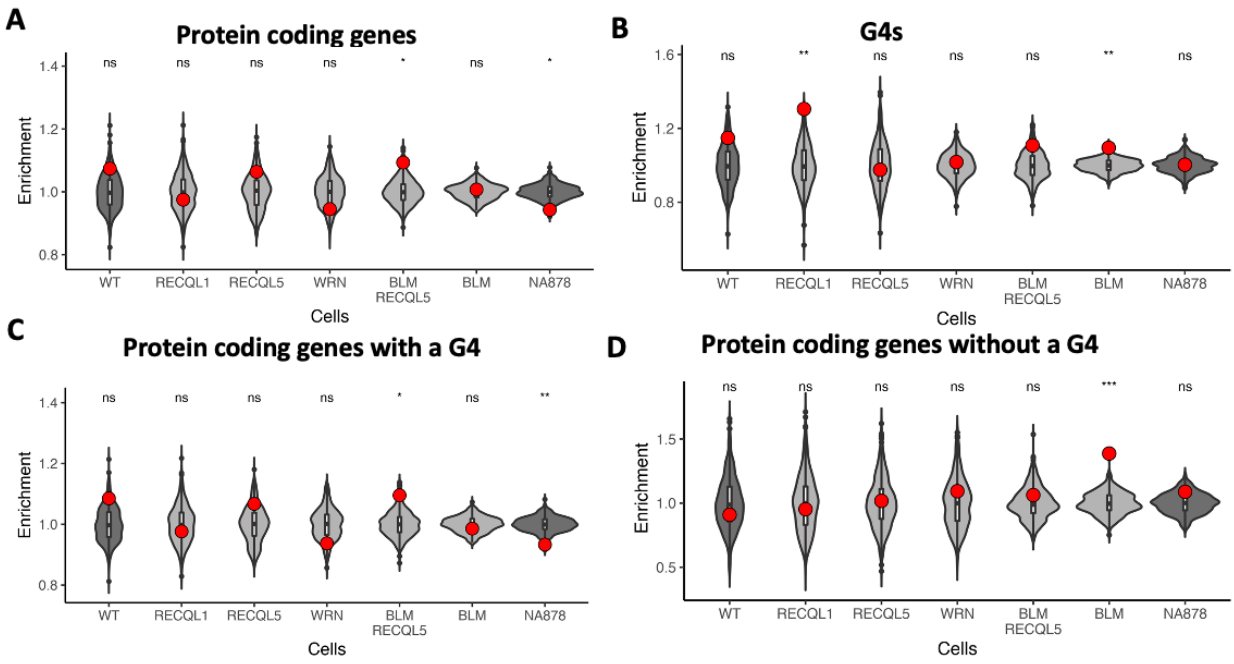



**Figure 4.4 SCE enrichment at protein coding genes and G4 quadruplexes.**
**SCE enrichment patterns for (A) protein coding genes (B) G4s (C) protein coding genes with at least one G4 (D) protein coding genes without a G4. Normal cell lines are indicated in blue, RecQ KO cell lines in red. Violin plots represent the expected range for random overlap. Red dots represent overlap of SCE with FOI described above each plot. P-values calculated from permutation test described in Section 4.2.2. **** $p < 0.001$, ** $p < 0.01$, not significant (ns) $p > 0.05$.**

It should be noted the violin plots for permutated data show more narrow distributions in the BLM, BLM/RECQL5 and NA12878 cell lines than in other cell lines. This is because of the higher numbers of SCEs that were collected from these cell lines. To exclude the possibility that SCE enrichment values in the cell lines are caused by the higher numbers of SCEs, I repeated the analysis for the same number of SCEs across all cell lines. A constant number of randomly

sampled SCEs were selected from each cell line and the enrichment analysis was repeated (Figure A3.2). The permutated ranges appeared more similar across cell lines when constant numbers of SCEs across cell lines are used however, the enrichment patterns across the different cell lines were conserved (Figure A3.2). Therefore, I conclude that the higher number of SCEs analysed does not confound enrichment analysis.

### 4.3.2 Association between SCEs and gene essentiality

Next, I wanted to focus on unique subsets of genes and the possible synergistic role of potential G4s in SCE formation. Therefore, my dataset of genes ranked by essentiality was used to divide genes into six categories: essential genes, non-essential genes, essential genes containing at least one potential G4, essential genes not containing any potential G4s, non-essential genes containing at least one potential G4 and non-essential genes not containing any potential G4s. I detected SCE enrichments for essential genes only in the RECQL1 cell line, but no such trend in other cell lines, including the WT and NA12878 lines (Figure A3.3A). I also found no significant SCE enrichments in non-essential genes across all cell lines, except for a minor depletion in the NA12878 cell line (Figure A3.3B). Next, I found that significant SCE enrichment for essential genes in the RECQL1 cell line was conserved regardless of potential G4 presence, suggesting potential G4 structures may not impact SCE formation in the absence of RECQL1 (Figure A3.3 C-D). Strikingly, for both essential and non-essential genes without at least one potential G4 structure I found significant SCE enrichments in BLM cell lines (Figure A3.3 D-F). These results indicate that SCEs in RECQL1 deficient cells may be due to the active transcription of essential genes and that SCEs in BLM deficient cells may preferentially occur in genes without potential G4 structures.

Although, SCE enrichment in the RECQL1 cell line seemed to be independent of potential G4 presence, I wanted to investigate if potential G4 "strandedness" has any effect on SCE formation by separating genes into those with potential G4s on the template strand and those with potential G4s on the coding strand and repeating SCE enrichment analyses for both. I found significant SCE enrichment for essential genes with potential G4s on the template strand in the RECQL1 cell line but not for essential genes with coding strand potential G4s (Figure A3.4 C-D). This trend was not observed in WT cells or in the RECQL1 cells line with respect to non-essential genes with potential G4s on the transcribed or coding strands (Figure A3.4 C-F). Together, these results suggest that the SCE enrichments for essential genes with potential G4s in RECQL1 deficient cells is mainly caused by potential G4 structures on actively transcribed template strands.

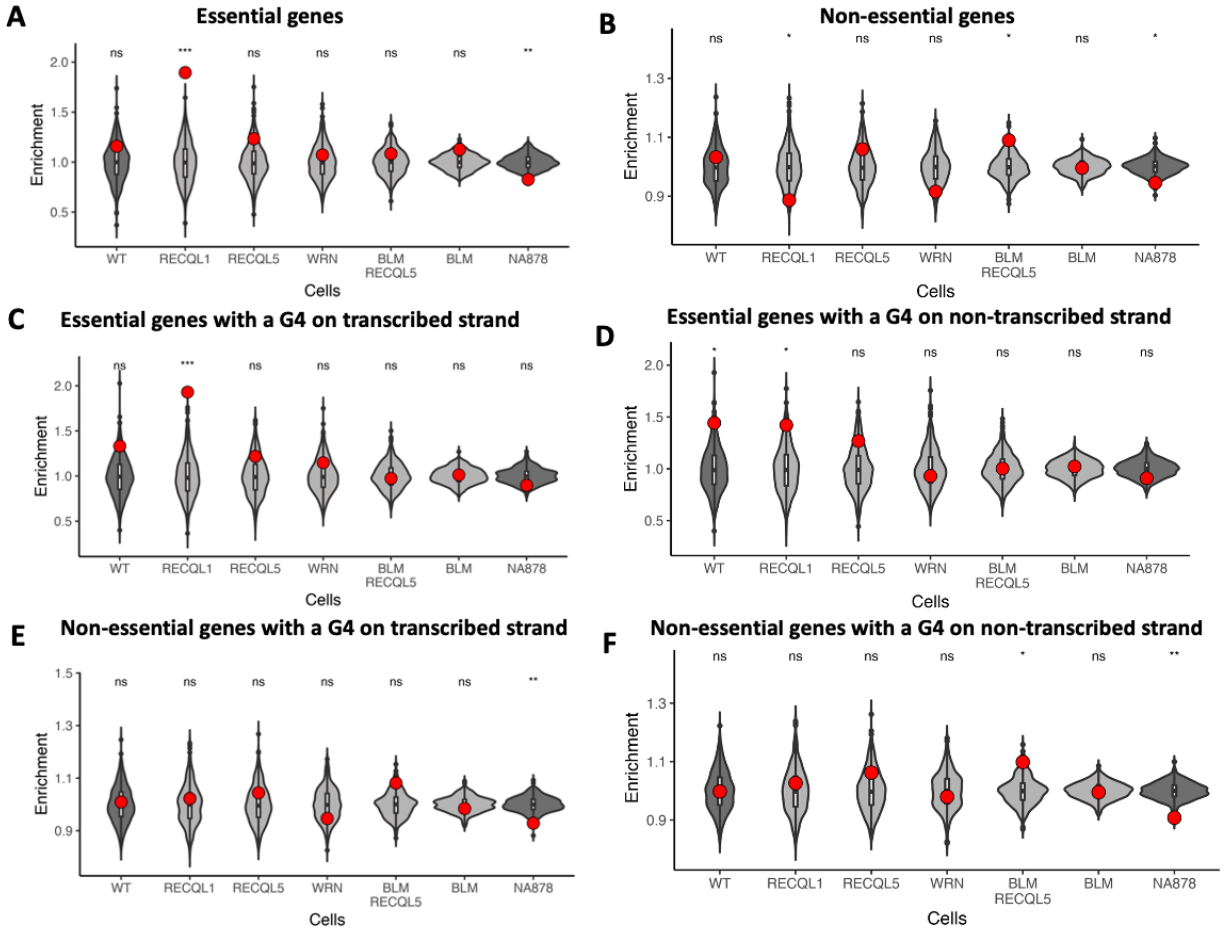



**Figure 4.5 SCE enrichment at essential and non-essential genes with and without G4 quadruplexes.**
**SCE enrichment patterns for (A) essential genes (B) non-essential genes (C) essential genes with a G4 (D) essential genes without at least one G4 (E) non-essential genes with a G4 (F) non-essential genes without at least one G4. Normal cell lines are indicated in blue, RecQ KO cell lines in red. Violin plots represent the expected range for random overlap. Red dots represent overlap of SCE with FOI described above each plot. P-values calculated from permutation test described in Section 4.2.2. **** $p < 0.001$, ** $p < 0.01$, not significant (ns) $p > 0.05$.**

### 4.3.3 Association between SCEs and gene transcriptional activity

After establishing an association between RECQL1 and SCE formation in essential genes with potential G4s on the transcribed strand, I wanted to investigate the effect of transcriptional activity on SCE formation. As discussed in Section 4.2.3.1, I used RNA-seq data from KBM7 cells to group genes into four categories by transcriptional activity: highly expressed genes,

moderately expressed genes, lowly expressed genes, and non-expressed genes (Figure 4.3). I found a modest relationship between SCE enrichment and transcriptional activity for certain KO lines. Mainly, I detected significant SCE enrichments for highly transcribed genes in WT, WRN and BLM/RECQL5 lines that were not present at any other transcriptional level. Strikingly, this trend was the most pronounced for BLM/RECQL5 cell lines and the least pronounced with the WT cell line (Figure A3.4). I also found a significant SCE enrichment for moderately and lowly expressed genes in BLM cell lines that was not present in WT cells (Figure A3.4). Taken together, these results indicate there is a modest association between transcriptional activity and SCE formation in BLM and BLM/RECQL5 cell lines.

To further elucidate the association between transcriptional activity and SCE abundance for BLM and BLM/RECQL5 cell lines, I wanted to investigate the possible synergistic role of potential G4s and transcriptional activity on SCE abundance. Interestingly, I found extremely significant SCE enrichment with highly transcribed genes containing at least one potential G4 in the BLM/RECQL5 cell lines but not for highly transcribed genes not containing any potential G4s (Figure A3.5 A-B). This trend was less pronounced in WT cells (Figure A3.5 A-B) suggesting that SCE formation in cells deficient in BLM and RECQL5 may preferentially occur at highly transcribed genes containing at least one potential G4. Interestingly, this trend was reversed in the BLM cell lines, such that SCE enrichment was only present with highly transcribed genes not containing any potential G4s but not with highly transcribed genes containing at least one potential G4, suggesting that SCE formation in cells deficient in BLM may preferentially occur at highly transcribed genes without any potential G4s (Figure A3.5 A-B). There was also significant SCE enrichment in lowly transcribed genes containing at least one potential G4 in the BLM cell lines that was not present in lowly transcribed genes not containing

any potential G4s (Figure A3.5 E-F). This trend was not observed in WT cells (Figure A3.5 E-F). Taken together, these results indicate that BLM may play multiple roles in preventing SCE formation.

The results described above suggest multiple roles for BLM and BLM/RECQL5 in suppressing SCE formation at highly and lowly transcribed genes, with a possible synergistic effect of potential G4 presence in some cases. Therefore, I wanted to further investigate the role of potential G4 "strandedness" on SCE formation by separating transcriptionally grouped genes into those containing potential G4s on the template strand and those containing potential G4s on the coding strand and repeating SCE enrichment analyses. Enrichment patterns looked remarkably similar regardless of the strand of potential G4s, suggesting potential G4 "strandedness" does not impact the enrichment of SCE formation for BLM and BLM/RECQL5 cell lines

I wanted to continue identifying specific features of genes that may predispose areas of the genome to replication stress. Large, transcriptionally active genes are often prone to replication stress and are frequently described as common fragile sites. Therefore, I grouped genes into those in top and bottom $50^{th}$ percentiles for gene size and repeated SCE enrichment analyses. Our enrichment analysis considers gene size by using an expected (random) distribution for the comparison of SCE overlap with genes. For example, if larger genes were more likely to overlap with SCEs due to their larger size, this trend would be present in our random distribution and thus would be reflected in the enrichment score. I detected SCE enrichments for large, highly transcribed genes among BLM/RECQL5 cell lines, a trend that was also present in WT cells (Figure A3.7). I also detected significant SCE enrichments for small, highly transcribed genes in BLM and BLM/RECQL5 cell lines, a trend that was not present in

WT cells (Figure A3.7). Interestingly, SCE enrichments were also observed in large, lowly transcribed genes in BLM and to a lesser extent, RECQL1, cell lines and SCE depletions were observed in large, non-expressed genes in RECQL1 cell lines (Figure A3.7). Both trends were not present in WT cells (Figure A3.7). Together, these results suggest large, highly transcribed genes are normally prone to SCE formation regardless of RecQ helicase deficiency, however, cells may require BLM and to a lesser extent, RECQL1, to prevent SCE formation in small, highly transcribed genes.

### 4.3.4 Association between SCEs and transcriptional activity and gene size

The results described above suggest BLM and RECQL1 may have unique roles in suppressing SCE formation at small, highly transcribed genes and large, lowly transcribed genes. Next, I wanted to investigate if there was a synergistic effect of potential G4 presence with gene size and transcriptional activity. Therefore, I first separated large genes grouped by four transcription levels into two categories: large genes containing at least one potential G4 and large genes not containing any potential G4s. Enrichment patterns looked remarkably similar regardless of the presence of potential G4s, suggesting potential G4 presence in large genes does not impact the enrichment of SCE formation for BLM and RECQL1 cell lines (Figure A3.8). Then, I separated small genes grouped by four transcription levels into two categories: small genes containing at least one potential G4 and small genes not containing any potential G4s. Enrichment patterns also looked similar regardless of the presence of potential G4s, suggesting potential G4 presence in small genes also does not impact the enrichment of SCE formation for BLM and RECQL1 cell lines (Figure A3.9).

### 4.3.5 Association between SCEs and gene function and size

Based on the results in Figure 4.5, I found an association between RECQL1 and essential genes containing at least one potential G4 on the transcribed strand. Therefore, I wanted to investigate any potential synergistic effects of gene size and gene function, so I grouped essential genes into six categories: essential genes containing at least one potential G4, essential genes not containing any potential G4s, large essential genes containing at least one potential G4, large essential genes not containing any potential G4s, small essential genes containing at least one potential G4 and small essential genes not containing any potential G4s. Strikingly, I detected significant SCE enrichments in large essential genes containing at least one potential G4 and small essential genes not containing any potential G4s in the RECQL1 cell lines, suggesting RECQL1 has multiple functions in suppressing SCE formation (Figure 4.6 C-F). SCE enrichments were also detected in small essential genes regardless of potential G4 presence in BLM cell lines (Figure 4.6 E-F). Both trends were less pronounced in WT cells (Figure 4.6 C-F).

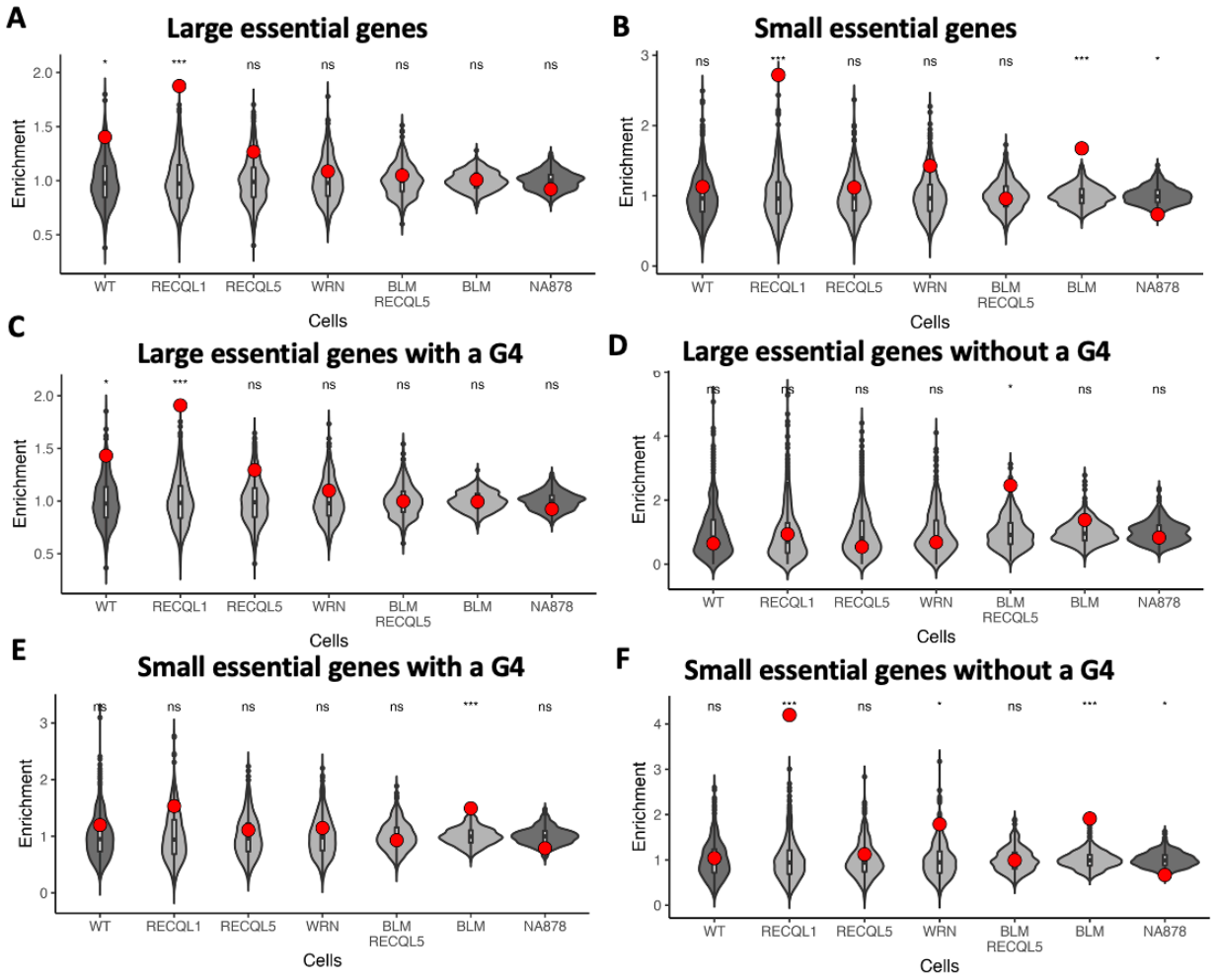



**Figure 4.6 SCE enrichment at large and small essential genes with potential G4 quadruplexes.**
**SCE enrichment patterns in (A) large essential genes (B) small essential genes (C) small essential genes containing at least one potential G4 and (D) small essential genes not containing any potential G4s. Normal cell lines are indicated in blue, RecQ KO cell lines are indicated in red. Violin plots represent the expected range for random overlap. Red dots represent overlap of SCEs with FOI described above each plot. P-values calculated from permutation test described in Section 4.2.2. **** $p < 0.001$, ** $p < 0.01$, not significant (ns) $p > 0.05$.**

## 4.4 Discussion

SCEs are a useful indicator of genome instability, but their exact mechanism remains incompletely understood [25]. Several RecQ helicases have been shown to play a role in suppressing SCE formation however, mapping SCEs to fine coordinates in the genome has been a major limitation of using cytogenetic detection methods to uncover the role of RecQ helicases

in SCE formation [25]. I used Strand-seq and custom bioinformatic methods discussed in Chapters 2 and 3 to map SCEs to the genome at kilobase resolution and here, I investigated the non-random distribution of SCEs in the genome. This approach allowed us to identify areas of the genome that are more troublesome for replication and prone to SCE formation and reveal protective functions of several RecQ helicase in preserving genome integrity in these areas. Using our customized SCE enrichment analysis pipeline, I show that SCEs in several RecQ helicase deficient cell lines are frequently occurring in subsets of genes with a possible synergistic role of potential G4 presence. I found strong SCE enrichments for large, essential genes containing a potential G4 on the transcribed strand in RECQL1 deficient cells. This is consistent with previous findings that RECQL1 can bind to and unwind potential G4s in the promoter regions of genes [124]. Additionally, RECQL1 has been shown to unwind other DNA structures such as replication fork joint molecules and promote ssDNA annealing and branch migration of dHJs and D-loops [61]. Considering I found strong SCE enrichments for small, essential genes not containing any potential G4s, its possible RECQL1 may suppress SCEs in these regions through one of the above-mentioned functions. These results are of interest considering the recently discovered genome instability syndrome, RECON syndrome, associated with RECQL1 [61]. RECON is characterized by progeroid facial features, small facial features, skin photosensitivity, xeroderma, and slender, elongated thumbs [61]. The fact that so many SCEs occur in specific subsets of actively transcribed genes in RECQL1 deficient cells, suggests that elevated SCE rates in RECON syndrome patient cells may be increasing somatic mutation rates within these gene to produce the associated clinical phenotype of RECON syndrome.

I also found unique SCE enrichments for highly transcribed genes in BLM/RECQL5 deficient cell lines that was dependent on potential G4 presence. Further investigation revealed that these SCEs occurred preferentially in highly transcribed genes with potential G4s on the transcribed strand. This trend was reversed for the BLM deficient cell lines, such that SCE were uniquely and preferentially occurring in highly transcribed genes not containing potential G4s. Neither trend was observed for RECQL5. Additionally, lowly transcribed genes containing at least one potential G4 require BLM to prevent SCE formation. Together, these findings support the notion that BLM is responsible for unwinding potential G4 structures in actively and non-transcribed genes and preventing SCEs in actively transcribed genes without potential G4s, whereas RECQL5 is exclusively responsible for preventing SCEs in actively transcribed genes with potential G4s. This suggests in the absence of RECQL5, BLM may compensate by unwinding potential G4s at actively transcribed genes and preventing SCEs at other actively transcribed genes, but in the absence of BLM, RECQL5 is unable to compensate for both roles of BLM in suppressing SCEs. This is consistent with the fact that there is a high frequency of G4 motifs in genes, their ability to form potential G4 structures during the formation of transcription bubbles and that both helicases have been shown to unwind potential G4 structures *in vitro* [3], [25], [125]. However, this is inconsistent with previous findings from van Wietmarschen *et al.* 2018 that found the BLM helicase unwinds potential G4s near actively transcribed genes containing potential G4s. One possible explanation for this is these previous findings relied on the use of G4 motifs that were identified as having G4-forming potential whereas my SCE enrichment analysis uses whole genome experimental maps of potential G4 structures collected from Marsico *et al.,* 2019. These differences between G4 motifs and G4 structures could explain the difference in SCE enrichment observed in activeley transcribed genes containing potential

G4s from Bloom Syndrome patient cells from van Wietmarschen *et al.* 2018 and my BLM deficient cells because not all G4 motifs form G4 structures.

Based on these results, I propose that the BLM helicase has multiple roles in preventing SCEs near actively transcribed genes, one of which is unwinding potential G4 structures that could possibly be shared by RECQL5. Indeed, potential G4 structures at sites of transcription can act as barriers to replication forks, resulting in persistent fork stalling and collision with replication machinery [3]. These collisions are a known source of genome instability, yet the exact mechanism has remained incompletely understood. Cells that experience increased SCE formation at these sites are at higher risk for the formation of deleterious by-products such as LOH or aberrant structural rearrangements, both of which can disrupt the coding sequence of genes and perturb gene function, giving rise to the clinical phenotypes associated with the absence of the BLM helicase. I believe the shared and unique roles of RECQL1, BLM and RECQL5 in preventing SCEs at these sites is essential to preserve genome stability and act as an anti-cancer barrier.

# Chapter 5: Conclusion

## 5.1 Summary of results

The goal of my thesis was to develop new methods to investigate the role of RecQ helicases in genome stability. To this end, I have developed and implemented several methods for performing DNA repair studies in single cells using Strand-seq and identified several functions for RECQL1, RECQL5, BLM and WRN in genome stability.

In Chapter 1, I summarized the role of genome instability in cancer. Because there are many ways the human genome can repair itself against endogenous and exogenous stressors, there are also many ways in which DNA repair can break down [3], [4], [126]. Faulty DNA repair machinery can contribute to the accumulation of mutations in the genome and the progressive deterioration of normal cell function [3], [4], [126]. Genome instability is a characteristic of almost all human cancers yet the amount, type, and source of genomic instability in tumour genomes differ substantially across tumour types and cell types [3], [4], [6], [126]. The association between deficiencies in several RecQ helicases and premature aging syndromes characterized by extreme cancer predisposition provide some insight into the relationship between DNA repair and cancer [12]. RecQ helicases are essential DNA repair genes that are believed to be responsible for suppressing inappropriate recombination during DSB repair however, their exact functions remain incompletely understood [12].

In Chapter 2, I introduced novel methods for performing DNA repair studies in single cells using Strand-seq. I discuss the advantages and limitations of Strand-seq in comparison to other bulk WGS and scWGS techniques. Specifically, how the unique component of read directionality in Strand-seq libraries provides accurate detection of copy-neutral rearrangements

such as SCEs, inversions and translocations that typically evade detection otherwise [49], [56]–[58]. I highlighted the limitations associated with using diploid cells, the original Strand-seq protocol and a manual QC step for filtering Strand-seq libraries. I introduced three implementations to address these limitations and improve the overall quality of DNA repair studies performed using Strand-seq. First, I used haploid cells to cut sequencing costs and improve breakpoint resolution for SCE and SV calling. Next, I revised the original Strand-seq protocol to improve cost, scalability, and quality of sequencing libraries. To that end, OP-Strand-seq can produce 6 to 16-fold more libraries than the original Strand-seq protocol at 15% of the original cost with ~4-fold greater complexity per cell on average, capturing up to 25% of the haploid genome in a single cell. Lastly, to address variable sequencing library quality and facilitate high-throughput QC, I developed an automated method for sorting the quality of OP-Strand-seq libraries that exceeds the performance of existing methods. All together, these implementations allowed us to generate low-cost, good quality Strand-seq libraries for downstream SCE and SV calling and analysis in Chapter 3 and 4.

In Chapter 3, I discussed how to harness the unique capability of Strand-seq to call different classes of SVs. SVs are a major source of genomic instability and contribute to significant intra-tumour heterogeneity. I summarized the approaches that have historically been used for SV detection and the challenges associated with SV discovery. As a new technology, Strand-seq has very few bioinformatic tools for SV discovery and of the tools that do exist, they often require domain experts to process the data. The accessibility of Strand-seq to field is therefore directly related to the collection of bioinformatic tools capable of exploiting the directionality of template reads for comprehensive SV discovery. To this end, I introduced the general framework for tools to map SCEs, translocations, and CNAs in Strand-seq libraries. The

SCE caller I developed has improved precision over alternative methods, owing to the removal of nearly all false positive SCE calls, and is capable of generating highly accurate SCE callsets. To date, no translocation caller using Strand-seq data exists. The method I developed describes a novel approach for the genome-wide screening of germline translocations in cells. This method was capable of finely mapping the coordinates of the Philadelphia chromosome translocation to the genome. Lastly, I demonstrated the unique advantage of using haploid cells to call CNAs in Strand-seq data. I demonstrate how approximately half of all CNAs in haploid cells have read depth changes associated with unambiguous strand-state switches than with diploid cells resulting in more accurate CNA calling in haploid cells than diploid cells. I believe these methods will undoubtedly facilitate the development of bioinformatic tools that can be used by the wider SV discovery community.

In chapter 4, I investigated whether SCEs from different RecQ helicase deficient cell lines were non-randomly distributed across the genome. This approach allowed us to identify areas of the genome that are more troublesome for replication and prone to SCE formation as well as reveal specific functions of several RecQ helicase in preserving genome integrity in these areas. Using our customized SCE enrichment analysis pipeline, I identified a strong association between SCE formation and genes, with a possible synergistic role of potential G4 presence within genes. RECQL1 deficient cells exhibited strong SCE enrichments in large, essential genes containing a potential G4 on the transcribed strand and small, essential genes not containing any potential G4s, suggesting RECQL1 is capable of binding and unwinding potential G4s in large actively transcribed genes and suppressing replication fork stress in small, actively transcribed genes through an independent mechanism [124]. BLM/RECQL5 deficient cell lines exhibited SCE enrichments in actively transcribed genes containing at least one potential G4 on the

transcribed strand whereas BLM deficient cell lines exhibited SCE enrichments in actively transcribed genes not containing potential G4s. With neither trend observed in RECQL5 deficient cell lines, these findings suggest that BLM may be responsible for suppressing SCE formation in actively transcribed genes by unwinding potential G4s and another independent mechanism, whereas RECQL5 is, to a lesser extent, capable of suppressing SCEs by unwinding potential G4s in actively transcribed genes. In other words, this evidence supports a redundant role between BLM and RECQL5. I propose that the BLM helicase has multiple roles in preventing SCEs near actively transcribed genes, one of which is unwinding potential G4 structures that is shared by RECQL5. These findings are consistent with the fact that there is a high frequency of potential G4 motifs in genes that can form potential G4 structures during transcription and act as barriers to replication forks, resulting in persistent fork stalling and collision with replication machinery and SCE formation [3], [25], [125]. These collisions are a known source of genome instability and both helicases have been shown to unwind potential G4 structures *in vitro*, however, the exact mechanism has remained incompletely understood [3], [25]. It should be noted that SCEs are considered error-free and do not result in any genetic alterations however, SCEs are considered a marker of genome instability due to increased recombinogenic activity and the possibility of LOH if homologous chromosomes are used for DSB repair or structural rearrangements if crossing over is unbalanced [25], [43]. I believe the shared and unique roles of RECQL1, BLM and RECQL5 in preventing SCEs at these sites are essential to preserve genome stability and acting as an anti-cancer barrier.

## 5.2 Limitations and weaknesses

The importance of reproducibility, accuracy and scalability cannot be understated when evaluating different SV callers. The SCE and SV callers introduced in this thesis have unique advantages and specific improvements over alternative methods however, they still have limitations.

A major challenge with SV calling, as discussed in Section 3.1.3, is calling SVs when multiple SVs are overlapping or nested within one another [104], [109]. The same challenge is also true for SCE calling, as discussed in Section 3.2.2, and comes with distinguishing "hotspots" that correspond to many SCEs occurring in the same area of the genome in multiple cells from SV breakpoints. To date, this can only be done on an individual basis by uploading bed-formatted read count files generated by *BreakpointR* to the UCSC Genome Browser to identify whether the breakpoints for SCEs are clustering near each other in different cells or if SVs have produced identical breakpoints in multiple cells [112]. These regions are of particular interest because it is unclear whether regions susceptible to frequent SCE formation may also experience SV formation or if SVs such as translocations or inversions may be troublesome for replication and susceptible to frequent SCE formation.

The major limitation of our translocation calling approach is the accuracy of detection for non-germline translocations present in a fraction of cells analyzed remains unclear. I have neither observed any somatic translocations nor tested our algorithm on data containing a somatic translocation. Therefore, the applicability of our translocation for somatic translocations in Strand-seq data is limited.

A major limitation of our CNA caller is the minimum size threshold placed on CNA calls. As discussed in Section 3.1.3, CNAs calling in regions of the genome associated with uneven or sparse sequencing coverage disrupts expected read depth changes and can lead to false negative and false positive calls [104], [109]. Therefore, I used a minimum size threshold for CNAs of 20 Mb to avoid false discovery associated with small CNAs at the expense of missing all CNAs less than 20 Mb all together. These challenges can of course be overcome with improved sequencing that has higher and more even sequencing coverage.

## 5.3 Future applications

### 5.3.1 Uncovering precise mechanistic role of RecQ helicases in DSB repair

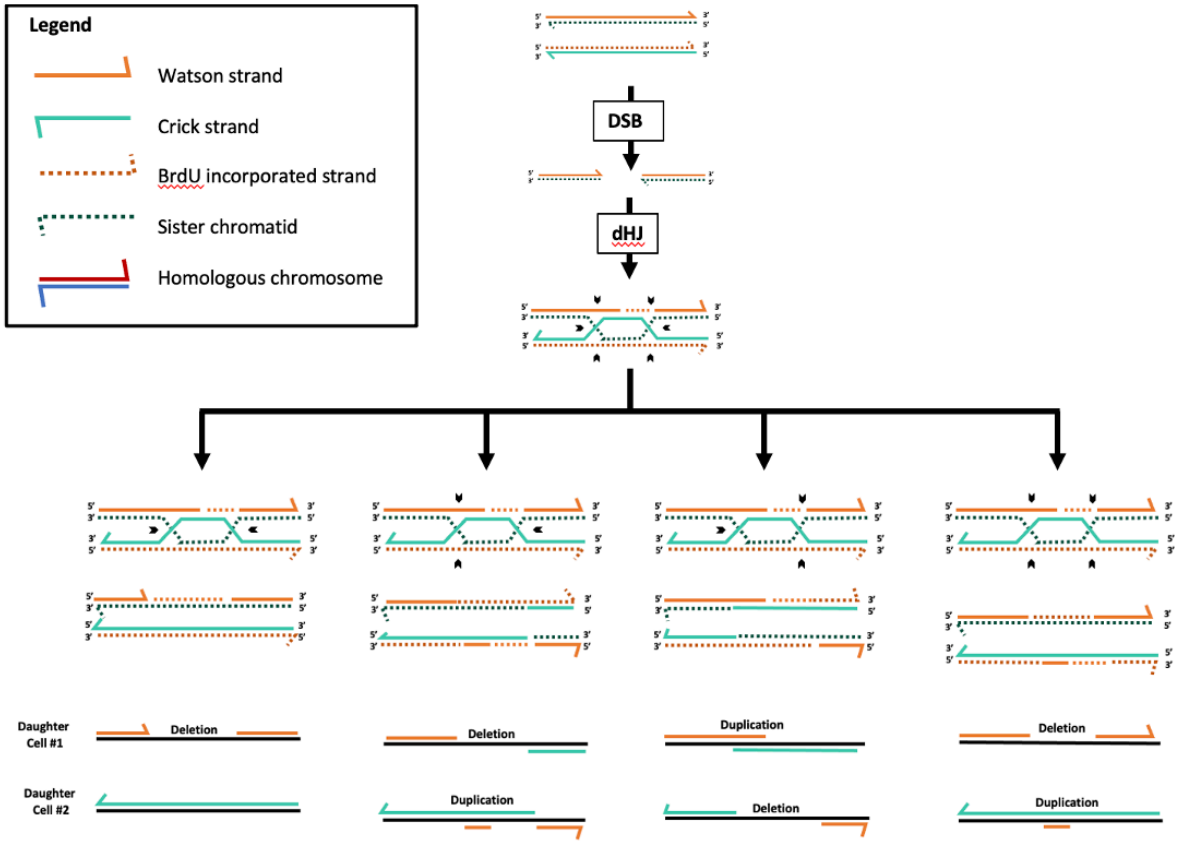
While my results show that certain subsets of genes containing secondary structure forming motifs are a cause of replication stress in absence of certain RecQ helicases, I was not able to identify the precise mechanistic function of these helicases in preventing replication fork stalling. I believe exposing our knockout lines to genotoxic compounds could inform our mechanistic understanding of RecQ helicase function. Using genotoxic compounds with known modes of action could inform the functional role of BLM and RECQL5 as compounds with different mechanistic action would likely yield different degrees of genome instability depending on gene function. However, this experiment poses several logistical challenges. Compounds that disrupt cell division, progression through the cell cycle or DNA replication may interfere with BrdU incorporation during Strand-seq library preparation. This type of experiment would need to ensure that cells have successfully divided once and incorporated BrdU into nascent strands after genotoxic compound exposure. This may involve testing multiple doses and incubation steps of

varying length during and after compound exposure followed by cell cycle analysis to assess progression through the cell cycle.

As more accurate, novel datasets of sequences capable of forming secondary structures are discovered, SCE enrichment analysis will reveal more genetic elements that are troublesome for replication and prone to replication fork stalling and collapse. Increasing the power of SCE datasets will also strengthen the evidence for the molecular insights of how these helicases function in distinct genomic contexts.

### 5.3.2 Improved resolution of strand state switches could reveal novel strand switches

In Chapter 2, I highlighted the importance of improving strand state switch breakpoint resolution for the comprehensive discovery of SCEs and SVs. Considering the breadth of possible DSB repair outcomes, it seems possible that improving breakpoint resolution may reveal unique Strand-seq signatures depending on when a DSB occurs and how it is cleaved and repaired by HR (Figure 5.1). Figure 5.1 describes several theoretical DSB outcomes depending on when the DSB occurs and whether it is repaired by SDSA, dHJ formation or break induced replication (BIR). Each repair pathway could likely produce district signatures of small deletions, duplications, and strand state switches in one or both daughter cells (Figure 5.1). These Strand-seq signatures could provide unique insight into small, novel repair events that have never been observed *in vivo* before.




| | | | | |
|---|---|---|---|---|
| **Crossover product** | No | Yes | Yes | Yes? |
| **Paired event** | No | Yes | Yes | Yes |
| **Signature cell #1** | Small het deletion | Small het deletion + strand state change | Small het duplication + Strand state change | Small deletion |
| **Signature cell #2** | NA | Small het duplication + Strand state change | Small het deletion + Strand state change | Small duplication |

**Figure 5.1 Possible DSB repair outcomes using dHJ formation and associated Strand-seq signatures. Schematic shows DSB occurring after DNA replication and RAD51-mediated donor search forms double Holliday junction (dHJ) and depending on cleavage and ligation sites may result in paired deletion and duplication events in respective daughter cells. Below table summarizes Strand-seq-specific signature associated with each dHJ outcome.**

### 5.3.3 Combinatorial approaches of DSB and SCE detection

In this thesis, I used SCE frequency and location as a surrogate marker for the frequency and location of stalled or collapsed replication forks and DSBs occurring within a cell. However, for the majority of DSBs, the location of SCE does not always indicate the exact location of the DSB [39], [40]. As discussed in Chapter 1, there are several techniques that been used to detect the frequency and location of DSBs in cells, including γH2AX staining or the mapping of γH2AX markers using ChIP-seq [39], [40]. However, the resolution of γH2AX staining and mapping is still limited by the fact that γH2AX spreads over a region of several megabases around a DSB [39], [40]. Alternatively, there are several sequencing-based techniques for the direct detection of DSBs. For example, END-seq and DSBCapture are two methods capable of capturing DSBs before they are repaired [127], [128]. However, there is only a small window for capturing unrepaired DSBs, resulting in many repaired DSBs going undetected [39], [40]. Thus, a combinational approach involving the detecting of SCEs and DSBs using multiple techniques may be very useful for the accurate detection and analysis of DSB formation and repair.

### 5.3.4 Resolving intra-tumour heterogeneity using combinatorial sequencing approaches

Chapter 3 highlights the importance of having a comprehensive survey of the mutational landscape in single cells of a tumour to reconstruct the clonal evolution and devise therapeutic strategies in cancer. I believe the methods for SV discovery introduced in this thesis have improved the baseline for detecting SVs and SCEs in single cells. However, as discussed in Section 5.2, these methods suffer from unique limitations that prevent them from being able to reliably identify all SV types. Therefore, it seems reasonable that a combinatorial approach involving multiple sequencing technologies and SV callers may fill this gap by combining

unique strengths while overcoming individual weaknesses of multiple methods. A single bioinformatic tool for the automated and accurate full-spectrum detection of all SV types will undoubtedly advance our understanding of cancer, aging and poorly understood medical conditions.

## 5.4 Conclusions

Genome instability is a characteristic of almost all human cancers and considered a defining hallmark that drives oncogenesis and uncontrollable proliferation. In this thesis I hypothesized that different RecQ helicases have unique functions in preserving genome stability. Therefore, I developed novel methods to investigate the role of RecQ helicases in genome stability and identified several functions for RECQL1 and BLM in preventing replication stress, replication fork stalling and ultimately the formation of SCEs in subsets of actively transcribed genes. I believe the shared and unique roles of these helicases in preventing SCEs at these sites are essential to preserve genome stability and act as an anti-cancer barrier. Such information will help elucidate currently poorly understood DNA repair processes and will inform novel therapeutic strategies in cancer.

## Bibliography


[1] L. E. MacConaill and L. A. Garraway, "Clinical Implications of the Cancer Genome," *J. Clin. Oncol.*, vol. 28, no. 35, p. 5219, Dec. 2010.

[2] P. A. Jeggo, L. H. Pearl, and A. M. Carr, "DNA repair, genome stability and cancer: A historical perspective," *Nature Reviews Cancer*, vol. 16, no. 1. Nature Publishing Group, pp. 35–42, 01-Jan-2016.

[3] J. Zell, F. R. Sperti, S. Britton, and D. Monchaud, "DNA folds threaten genetic stability and can be leveraged for chemotherapy," *RSC Chem. Biol.*, vol. 2, no. 1, pp. 47–76, Feb. 2021.

[4] S. Negrini, V. G. Gorgoulis, and T. D. Halazonetis, "Genomic instability — an evolving hallmark of cancer," *Nat. Rev. Mol. Cell Biol.*, vol. 11, no. 3, pp. 220–228, Mar. 2010.

[5] A. Aguilera and B. Gómez-González, "Genome instability: a mechanistic view of its causes and consequences," *Nat. Rev. Genet.*, vol. 9, no. 3, pp. 204–217, Mar. 2008.

[6] S. F. Bakhoum and D. A. Landau, "Chromosomal Instability as a Driver of Tumor Heterogeneity and Evolution," *Cold Spring Harb. Perspect. Med.*, vol. 7, no. 6, Jun. 2017.

[7] W. S. Klug, M. R. Cummings, C. A. Spencer, and M. A. Palladino, *Essentials of Genetics*, Ninth Edit. Pearson Education, 2017.

[8] J. Guirouilh-Barbat, S. Lambert, P. Bertrand, and B. S. Lopez, "Is homologous recombination really an error-free process?," *Frontiers in Genetics*, vol. 5, no. JUN. Frontiers Research Foundation, p. 175, 11-Jun-2014.

[9] W. D. Wright, S. S. Shah, and W. D. Heyer, "Homologous recombination and the repair of DNA double-strand breaks," *Journal of Biological Chemistry*, vol. 293, no. 27. American Society for Biochemistry and Molecular Biology Inc., pp. 10524–10535, 2018.

[10] S. C. West, M. G. Blanco, Y. W. Chan, J. Matos, S. Sarbajna, and H. D. M. Wyatt, "Resolution of recombination intermediates: Mechanisms and regulation," in *Cold Spring Harbor Symposia on Quantitative Biology*, 2016, vol. 80, pp. 103–109.

[11] D. W. Felsher and J. M. Bishop, "Transient excess of MYC activity can elicit genomic instability and tumorigenesis," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 7, pp. 3940–3944, Mar. 1999.

[12] K. A. Bernstein, S. Gangloff, and R. Rothstein, "The RecQ DNA Helicases in DNA Repair," *Annu. Rev. Genet.*, vol. 44, no. 1, pp. 393–417, Dec. 2010.

[13] W. D. Heyer, K. T. Ehmsen, and J. Liu, "Regulation of homologous recombination in eukaryotes," *Annu. Rev. Genet.*, vol. 44, pp. 113–139, Dec. 2010.

[14] M. W. Parker, M. R. Botchan, and J. M. Berger, "Mechanisms and regulation of DNA replication initiation in eukaryotes," *Crit. Rev. Biochem. Mol. Biol.*, vol. 52, no. 2, p. 107, Mar. 2017.

[15] M. Fragkos, O. Ganier, P. Coulombe, and M. Méchali, "DNA replication origin activation in space and time," *Nat. Rev. Mol. Cell Biol. 2015 166*, vol. 16, no. 6, pp. 360–374, May 2015.

[16] L. Toledo, K. J. Neelsen, and J. Lukas, "Replication Catastrophe: When a Checkpoint Fails because of Exhaustion.," *Mol. Cell*, vol. 66, no. 6, pp. 735–749, Jun. 2017.

[17] H. Gaillard, T. García-Muse, and A. Aguilera, "Replication stress and cancer," *Nat. Rev. Cancer 2015 155*, vol. 15, no. 5, pp. 276–289, Apr. 2015.

[18] H. C. Olson *et al.*, "Increased levels of RECQ5 shift DNA repair from canonical to alternative pathways," *Nucleic Acids Res.*, vol. 46, no. 18, pp. 9496–9509, 2018.

[19] R. Scully, A. Panday, R. Elango, and N. A. Willis, "DNA double-strand break repair-

pathway choice in somatic mammalian cells," *Nature Reviews Molecular Cell Biology*, vol. 20, no. 11. Nature Publishing Group, pp. 698–714, 01-Nov-2019.

[20] Y. Wang *et al.*, "G-quadruplex DNA drives genomic instability and represents a targetable molecular abnormality in ATRX-deficient malignant glioma," *Nat. Commun.*, vol. 10, no. 1, p. 943, Dec. 2019.

[21] K. G. Zyner *et al.*, "G-quadruplex DNA structures in human stem cells and differentiation," *Nat. Commun. 2022 131*, vol. 13, no. 1, pp. 1–17, Jan. 2022.

[22] G. Marsico *et al.*, "Whole genome experimental maps of DNA G-quadruplexes in multiple species," *Nucleic Acids Res.*, vol. 47, no. 8, pp. 3862–3874, May 2019.

[23] K. G. Zyner *et al.*, "G-quadruplex DNA structures in human stem cells and differentiation," *Nat. Commun. 2022 131*, vol. 13, no. 1, pp. 1–17, Jan. 2022.

[24] J. L. Huppert and S. Balasubramanian, "G-quadruplexes in promoters throughout the human genome," *Nucleic Acids Res.*, vol. 35, no. 2, pp. 406–413, 2007.

[25] N. van Wietmarschen, S. Merzouk, N. Halsema, D. C. J. Spierings, V. Guryev, and P. M. Lansdorp, "BLM helicase suppresses recombination at G-quadruplex motifs in transcribed genes," *Nat. Commun.*, vol. 9, no. 1, p. 271, Dec. 2018.

[26] N. Chappidi *et al.*, "Fork Cleavage-Religation Cycle and Active Transcription Mediate Replication Restart after Fork Stalling at Co-transcriptional R-Loops," *Mol. Cell*, Nov. 2019.

[27] S. Mijic *et al.*, "Replication fork reversal triggers fork degradation in BRCA2-defective cells," *Nat. Commun.*, vol. 8, no. 1, Dec. 2017.

[28] N. van Wietmarschen *et al.*, "Repeat expansions confer WRN dependence in microsatellite-unstable cancers," *Nature*, vol. 586, no. 7828, p. 292, Oct. 2020.

[29] R. Kumar, G. Nagpal, V. Kumar, S. S. Usmani, P. Agrawal, and G. P. S. Raghava, "HumCFS: A database of fragile sites in human chromosomes," *BMC Genomics*, vol. 19, no. 9, pp. 1–8, Apr. 2019.

[30] S. Di Marco *et al.*, "RECQ5 Helicase Cooperates with MUS81 Endonuclease in Processing Stalled Replication Forks at Common Fragile Sites during Mitosis," *Mol. Cell*, vol. 66, no. 5, pp. 658-671.e8, Jun. 2017.

[31] M. Saponaro *et al.*, "RECQL5 controls transcript elongation and suppresses genome instability associated with transcription stress," *Cell*, vol. 157, no. 5, pp. 1037–1049, May 2014.

[32] N. Kim and S. Jinks-Robertson, "Transcription as a source of genome instability," *Nat. Rev. Genet.*, vol. 13, no. 3, p. 204, Mar. 2012.

[33] Y. L. Lin and P. Pasero, "Transcription-Replication Conflicts: Orientation Matters," *Cell*, vol. 170, no. 4, pp. 603–604, Aug. 2017.

[34] L. Olavarrieta, P. Hernández, D. B. Krimer, and J. B. Schvartzman, "DNA knotting caused by head-on collision of transcription and replication," *J. Mol. Biol.*, vol. 322, no. 1, pp. 1–6, 2002.

[35] R. Kanagaraj *et al.*, "RECQ5 helicase associates with the C-terminal repeat domain of RNA polymerase II during productive elongation phase of transcription," *Nucleic Acids Res.*, vol. 38, no. 22, pp. 8131–8140, Dec. 2010.

[36] C. Rinaldi, P. Pizzul, M. P. Longhese, and D. Bonetti, "Sensing R-Loop-Associated DNA Damage to Safeguard Genome Stability," *Front. Cell Dev. Biol.*, vol. 8, Jan. 2020.

[37] S. Hedau *et al.*, "Novel germline mutations in breast cancer susceptibility genes BRCA1, BRCA2 and p53 gene in breast cancer patients from India," *Breast Cancer Res. Treat.*,

vol. 88, no. 2, pp. 177–186, 2004.

[38] T. Gemoll, G. Auer, T. Ried, and J. K. Habermann, “Genetic Instability and Disease Prognostication,” *Recent Results Cancer Res.*, vol. 200, p. 81, Sep. 2015.

[39] A. C. Vítor, P. Huertas, G. Legube, and S. F. de Almeida, “Studying DNA Double-Strand Break Repair: An Ever-Growing Toolbox,” *Front. Mol. Biosci.*, vol. 7, p. 24, Feb. 2020.

[40] M. Löbrich *et al.*, “Cell Cycle γH2AX foci analysis for monitoring DNA double-strand break repair: Strengths, limitations and optimization,” *Cell Cycle*, vol. 9, pp. 662–669, 2010.

[41] R. Torres-Ruiz, T. P. Grazioso, M. Brandt, M. Martinez-Lage, S. Rodriguez-Perales, and N. Djouder, “Detection of chromosome instability by interphase FISH in mouse and human tissues,” *STAR Protoc.*, vol. 2, no. 3, Sep. 2021.

[42] Y. Hu, X. Lu, E. Barnes, M. Yan, H. Lou, and G. Luo, “Recql5 and Blm RecQ DNA Helicases Have Nonredundant Roles in Suppressing Crossovers,” *Mol. Cell. Biol.*, vol. 25, no. 9, pp. 3431–3442, May 2005.

[43] C. Claussin *et al.*, “Genome-wide mapping of sister chromatid exchange events in single yeast cells using strand-seq,” *Elife*, vol. 6, Dec. 2017.

[44] J. H. Taylor, P. S. Woods, and W. L. Hughes, “THE ORGANIZATION AND DUPLICATION OF CHROMOSOMES AS REVEALED BY AUTORADIOGRAPHIC STUDIES USING TRITIUM-LABELED THYMIDINEE,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 43, no. 1, p. 122, Jan. 1957.

[45] S. A. Latt, “Microfluorometric detection of deoxyribonucleic acid replication in human metaphase chromosomes,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 70, no. 12, pp. 3395–3399, 1973.

[46] N. Van Wietmarschen and P. M. Lansdorp, “Bromodeoxyuridine does not contribute to sister chromatid exchange events in normal or Bloom syndrome cells,” *Nucleic Acids Res.*, vol. 44, no. 14, p. 6787, Aug. 2016.

[47] P. A. Audano *et al.*, “Characterizing the Major Structural Variant Alleles of the Human Genome,” *Cell*, vol. 176, no. 3, pp. 663-675.e19, Jan. 2019.

[48] A. Auton *et al.*, “A global reference for human genetic variation,” *Nature*, vol. 526, no. 7571. Nature Publishing Group, pp. 68–74, 30-Sep-2015.

[49] M. J. P. Chaisson *et al.*, “Multi-platform discovery of haplotype-resolved structural variation in human genomes,” *Nat. Commun.*, vol. 10, no. 1, p. 1784, Dec. 2019.

[50] G. Macintyre, B. Ylstra, and J. D. Brenton, “Sequencing Structural Variants in Cancer for Precision Therapeutics,” *Trends in Genetics*, vol. 32, no. 9. Elsevier Ltd, pp. 530–542, 01-Sep-2016.

[51] A. Kuzniar *et al.*, “sv-callers: A highly portable parallel workflow for structural variant detection in whole-genome sequence data,” *PeerJ*, vol. 1, Jan. 2020.

[52] G. A. Logsdon, M. R. Vollger, and E. E. Eichler, “Long-read human genome sequencing and its applications,” *Nat. Rev. Genet. 2020 2110*, vol. 21, no. 10, pp. 597–614, Jun. 2020.

[53] S. Nurk *et al.*, “The complete sequence of a human genome,” *Science (80-. ).*, vol. 376, no. 6588, pp. 44–53, Apr. 2022.

[54] J. Hård *et al.*, “Long-read whole genome analysis of human single cells,” *bioRxiv*, p. 2021.04.13.439527, Apr. 2021.

[55] S. L. Goldman, M. MacKay, E. Afshinnekoo, A. M. Melnick, S. Wu, and C. E. Mason, “The impact of heterogeneity on single-cell sequencing,” *Front. Genet.*, vol. 10, no. MAR, p. 8, 2019.

[56] E. Falconer *et al.*, “DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution,” *Nat. Methods*, vol. 9, no. 11, pp. 1107–1112, Nov. 2012.

[57] A. D. Sanders, E. Falconer, M. Hills, D. C. J. Spierings, and P. M. Lansdorp, “Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs,” *Nat. Protoc.*, vol. 12, no. 6, pp. 1151–1176, May 2017.

[58] A. D. Sanders *et al.*, “Single-cell analysis of structural variations and complex rearrangements with tri-channel processing,” *Nat. Biotechnol. 2019 383*, vol. 38, no. 3, pp. 343–354, Dec. 2019.

[59] Z. Hamadeh, V. C. T. Hanlon, and P. M. Lansdorp, “Mapping of sister chromatid exchange events and genome alterations in single cells,” *Elsevier Methods*, 2022.

[60] V. C. T. Hanlon *et al.*, “Construction of Strand-seq libraries in open nanoliter arrays,” *Cell Reports Methods*, vol. 2, no. 1, p. 100150, Jan. 2022.

[61] B. Abu-Libdeh *et al.*, “RECON syndrome is a genome instability disorder caused by mutations in the DNA helicase RECQL1,” *J. Clin. Invest.*, vol. 132, no. 5, Mar. 2022.

[62] Y. Hu *et al.*, “RECQL5/Recql5 helicase regulates homologous recombination and suppresses tumor formation via disruption of Rad51 presynaptic filaments,” *Genes Dev.*, vol. 21, no. 23, pp. 3073–3084, Dec. 2007.

[63] P. Khadka, D. L. Croteau, and V. A. Bohr, “RECQL5 has unique strand annealing properties relative to the other human RecQ helicase proteins,” *DNA Repair (Amst).*, vol. 37, pp. 53–66, Jan. 2016.

[64] D. L. Croteau, V. Popuri, P. L. Opresko, and V. A. Bohr, “Human RecQ Helicases in DNA Repair, Recombination, and Replication,” *Annu. Rev. Biochem.*, vol. 83, no. 1, pp.

519–552, Jun. 2014.

[65] Y. Wu, “Unwinding and Rewinding: Double Faces of Helicase?,” *J. Nucleic Acids*, vol. 2012, pp. 1–14, Jul. 2012.

[66] J. A. Newman, H. Aitkenhead, P. Savitsky, and O. Gileadi, “Insights into the RecQ helicase mechanism revealed by the structure of the helicase domain of human RECQL5,” *Nucleic Acids Res.*, vol. 45, no. 7, p. gkw1362, Jan. 2017.

[67] A. Arora *et al.*, “Clinicopathological and prognostic significance of RECQL5 helicase expression in breast cancers,” *Carcinogenesis*, vol. 37, no. 1, pp. 63–71, Jan. 2016.

[68] R. M. Brosh and V. A. Bohr, “Human premature aging, DNA repair and RecQ helicases,” *Nucleic Acids Res.*, vol. 35, no. 22, pp. 7527–7544, Dec. 2007.

[69] E. Sjöstedt *et al.*, “An atlas of the protein-coding genes in the human, pig, and mouse brain,” *Science (80-. ).*, vol. 367, no. 6482, Mar. 2020.

[70] J. A. Newman *et al.*, “Crystal structure of the Bloom’s syndrome helicase indicates a role for the HRDC domain in conformational changes,” *Nucleic Acids Res.*, vol. 43, no. 10, p. 5221, May 2015.

[71] E. Huselid and S. F. Bunting, “The Regulation of Homologous Recombination by Helicases,” *Genes (Basel).*, vol. 11, no. 5, May 2020.

[72] R. J. Monnat, “Human RECQ helicases: Roles in DNA metabolism, mutagenesis and cancer biology,” *Semin. Cancer Biol.*, vol. 20, no. 5, p. 329, Oct. 2010.

[73] J. A. Newman *et al.*, “Structure of the helicase core of Werner helicase, a key target in microsatellite instability cancers,” *Life Sci. Alliance*, vol. 4, no. 1, 2021.

[74] S. Kitao, I. Ohsugi, K. Ichikawa, M. Goto, Y. Furuichi, and A. Shimamoto, “Cloning of two new human helicase genes of the RecQ family: Biological significance of multiple

species in higher eukaryotes," *Genomics*, vol. 54, no. 3, pp. 443–452, Dec. 1998.

[75] V. Popuri, J. Huang, M. Ramamoorthy, T. Tadokoro, D. L. Croteau, and V. A. Bohr, "RECQL5 plays co-operative and complementary roles with WRN syndrome helicase," *Nucleic Acids Res.*, vol. 41, no. 2, pp. 881–899, 2013.

[76] E. Speina *et al.*, "Human RECQL5β stimulates flap endonuclease 1," *Nucleic Acids Res.*, vol. 38, no. 9, pp. 2904–2916, Jan. 2010.

[77] L. Zheng *et al.*, "MRE11 complex links RECQ5 helicase to sites of DNA damage," 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2677886/. [Accessed: 31-Mar-2020].

[78] R. Kanagaraj, N. Saydam, P. L. Garcia, L. Zheng, and P. Janscak, "Human RECQ5β helicase promotes strand exchange on synthetic DNA structures resembling a stalled replication fork," *Nucleic Acids Res.*, vol. 34, no. 18, pp. 5217–5231, Oct. 2006.

[79] M. Ramamoorthy *et al.*, "RECQL5 cooperates with Topoisomerase II alpha in DNA decatenation and cell cycle progression," *Nucleic Acids Res.*, vol. 40, no. 4, pp. 1621–1635, Feb. 2012.

[80] A. Shimamoto, "Human RecQ5beta, a large isomer of RecQ5 DNA helicase, localizes in the nucleoplasm and interacts with topoisomerases 3alpha and 3beta," *Nucleic Acids Res.*, vol. 28, no. 7, pp. 1647–1655, Apr. 2000.

[81] V. Urban, J. Dobrovolna, D. Hühn, J. Fryzelkova, J. Bartek, and P. Janscak, "RECQ5 helicase promotes resolution of conflicts between replication and transcription in human cells.," *J. Cell Biol.*, vol. 214, no. 4, pp. 401–15, Aug. 2016.

[82] O. Aygün, J. Svejstrup, and Y. Liu, "A RECQ5-RNA polymerase II association identified by targeted proteomic analysis of human chromatin," *Proc. Natl. Acad. Sci. U. S. A.*, vol.

105, no. 25, pp. 8580–8584, Jun. 2008.

[83] G. Zhou *et al.*, “Purification of a novel RECQL5-SWI/SNF-RNAPII super complex.,” *Int. J. Biochem. Mol. Biol.*, vol. 1, no. 1, pp. 101–111, 2010.

[84] Y. Hu, X. Lu, G. Zhou, E. L. Barnes, and G. Luo, “Recql5 plays an important role in DNA replication and cell survival after camptothecin treatment.,” *Mol. Biol. Cell*, vol. 20, no. 1, pp. 114–23, Jan. 2009.

[85] S. Paliwal, R. Kanagaraj, A. Sturzenegger, K. Burdova, and P. Janscak, “Human RECQ5 helicase promotes repair of DNA double-strand breaks by synthesis-dependent strand annealing,” *Nucleic Acids Res.*, vol. 42, no. 4, pp. 2380–2390, Feb. 2014.

[86] X. Zhu *et al.*, “Distinct prognosis of mRNA expression of the five RecQ DNA-helicase family members – RECQL, BLM, WRN, RECQL4, and RECQL5 – in patients with breast cancer,” *Cancer Manag. Res.*, vol. Volume 10, pp. 6649–6668, Dec. 2018.

[87] K. Rickman and A. Smogorzewska, “Advances in understanding DNA processing and protection at stalled replication forks.,” *J. Cell Biol.*, vol. 218, no. 4, pp. 1096–1107, Apr. 2019.

[88] J. C. Saldivar, D. Cortez, and K. A. Cimprich, “The essential kinase ATR: Ensuring faithful duplication of a challenging genome,” *Nature Reviews Molecular Cell Biology*, vol. 18, no. 10. Nature Publishing Group, pp. 622–636, 01-Oct-2017.

[89] K. Izumikawa *et al.*, “Association of human DNA helicase RecQ5β with RNA polymerase II and its possible role in transcription,” *Biochem. J.*, vol. 413, no. 3, pp. 505–516, Aug. 2008.

[90] M. Li, X. Xu, and Y. Liu, “The SET2-RPB1 interaction domain of human RECQ5 is

important for transcription-associated genome stability.," *Mol. Cell. Biol.*, vol. 31, no. 10, pp. 2090–9, May 2011.

[91] O. Aygün *et al.*, "Direct inhibition of RNA polymerase II transcription by RECQL5," *J. Biol. Chem.*, vol. 284, no. 35, pp. 23197–23203, Aug. 2009.

[92] M. Li, X. Xu, C.-W. Chang, L. Zheng, B. Shen, and Y. Liu, "SUMO2 conjugation of PCNA facilitates chromatin remodeling to resolve transcription-replication conflicts," *Nat. Commun.*, vol. 9, 2018.

[93] K. Patterson *et al.*, "Altered RECQL5 expression in urothelial bladder carcinoma increases cellular proliferation and makes RECQL5 helicase activity a novel target for chemotherapy," *Oncotarget*, vol. 7, no. 46, pp. 76140–76150, 2016.

[94] E. Chen *et al.*, "RECQL5 Suppresses Oncogenic JAK2-Induced Replication Stress and Genomic Instability," *Cell Rep.*, vol. 13, no. 11, pp. 2345–2352, Dec. 2015.

[95] R. A. Cartwright and D. Graur, "The multiple personalities of Watson and Crick strands," *Biol. Direct*, vol. 6, p. 7, Feb. 2011.

[96] B. Bakker *et al.*, "Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies," *Genome Biol. 2016 171*, vol. 17, no. 1, pp. 1–15, May 2016.

[97] D. Porubsky, A. D. Sanders, A. Taudt, M. Colomé-Tatche, P. M. Lansdorp, and V. Guryev, "BreakpointR: An R/Bioconductor package to localize strand state changes in Strand-seq data," *Bioinformatics*, vol. 36, no. 4, pp. 1260–1261, Feb. 2020.

[98] T. B. Beigl, I. Kjosås, E. Seljeseth, N. Glomnes, and H. Aksnes, "Efficient and crucial quality control of HAP1 cell ploidy status," *Biol. Open*, vol. 9, no. 11, Nov. 2020.

[99] M. Wetzler, M. Talpaz, R. A. Van Etten, C. Hirsh-Ginsberg, M. Beran, and R. Kurzrock, "Subcellular Localization of Bcr, Abl, and Bcr-Abl Proteins in Normal and Leukemic

Cells and Correlation of Expression with Myeloid Differentiation Bcr-Abl fusion proteins * c-ABL protooncogene proteins * im-munohistochemistry," 1993.

[100] W. Gao, B. Lai, B. Ni, and K. Zhao, "Genome-wide profiling of nucleosome position and chromatin accessibility in single cells using scMNase-seq," *Nat. Protoc. 2019 151*, vol. 15, no. 1, pp. 68–85, Dec. 2019.

[101] M. Kuhn, "Building Predictive Models in R Using the caret Package," *J. Stat. Softw.*, vol. 28, no. 5, pp. 1–26, Nov. 2008.

[102] T. Daley and A. D. Smith, "Predicting the molecular complexity of sequencing libraries," *Nat. Methods 2013 104*, vol. 10, no. 4, pp. 325–327, Feb. 2013.

[103] C. Gros, A. D. Sanders, J. O. Korbel, T. Marschall, and P. Ebert, "ASHLEYS: automated quality control for single-cell Strand-seq data," *Bioinformatics*, vol. 37, no. 19, pp. 3356–3357, Oct. 2021.

[104] M. Mahmoud, N. Gobet, D. I. Cruz-Dávalos, N. Mounier, C. Dessimoz, and F. J. Sedlazeck, "Structural variant calling: The long and the short of it," *Genome Biol.*, vol. 20, no. 1, pp. 1–14, Nov. 2019.

[105] A. E. Minoche *et al.*, "ClinSV: clinical grade structural and copy number variant detection from whole genome sequencing data," *Genome Med.*, vol. 13, no. 1, pp. 1–19, Dec. 2021.

[106] G. Macintyre, B. Ylstra, and J. D. Brenton, "Sequencing Structural Variants in Cancer for Precision Therapeutics," *Trends Genet.*, vol. 32, no. 9, pp. 530–542, Sep. 2016.

[107] C. Chiang *et al.*, "The impact of structural variation on human gene expression," *Nat. Genet.*, vol. 49, no. 5, pp. 692–699, May 2017.

[108] I. A. E. M. van Belzen, A. Schönhuth, P. Kemmeren, and J. Y. Hehir-Kwa, "Structural variant detection in cancer genomes: computational challenges and perspectives for

precision oncology," *npj Precis. Oncol. 2021 51*, vol. 5, no. 1, pp. 1–11, Mar. 2021.

[109] L. Denti, P. Khorsand, P. Bonizzoni, F. Hormozdiari, and R. Chikhi, "Improved structural variant discovery in hard-to-call regions using sample-specific string detection from accurate long reads," *bioRxiv*, p. 2022.02.12.480198, Feb. 2022.

[110] Y. Benjamini and T. P. Speed, "Summarizing and correcting the GC content bias in high-throughput sequencing," *Nucleic Acids Res.*, vol. 40, no. 10, pp. e72–e72, May 2012.

[111] D. Aird *et al.*, "Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries," *Genome Biol.*, vol. 12, no. 2, pp. 1–14, Feb. 2011.

[112] W. J. Kent *et al.*, "The human genome browser at UCSC," *Genome Res.*, vol. 12, no. 6, pp. 996–1006, Jun. 2002.

[113] S. Soverini *et al.*, "Next-generation sequencing improves BCR-ABL1 mutation detection in Philadelphia chromosome-positive acute lymphoblastic leukaemia," *Br. J. Haematol.*, vol. 193, no. 2, pp. 271–279, Apr. 2021.

[114] M. Nambiar and S. C. Raghavan, "How does DNA break during chromosomal translocations?," *Nucleic Acids Res.*, vol. 39, no. 14, p. 5813, Aug. 2011.

[115] X. F. Mallory, M. Edrisi, N. Navin, and L. Nakhleh, "Methods for copy number aberration detection from single-cell DNA-sequencing data," *Genome Biol.*, vol. 21, no. 1, pp. 1–22, Aug. 2020.

[116] R. Kumar, G. Nagpal, V. Kumar, S. S. Usmani, P. Agrawal, and G. P. S. Raghava, "HumCFS: a database of fragile sites in human chromosomes," *BMC Genomics 2019 199*, vol. 19, no. 9, pp. 1–8, Apr. 2019.

[117] Y. Shibata, A. Malhotra, and A. Dutta, "Detection of DNA fusion junctions for BCR-ABL translocations by Anchored ChromPET," *Genome Med.*, vol. 2, no. 9, p. 70, Sep. 2010.

[118] R. M. Brosh and S. W. Matson, “History of DNA Helicases,” *Genes 2020, Vol. 11, Page 255*, vol. 11, no. 3, p. 255, Feb. 2020.

[119] Y. Li *et al.*, “Patterns of somatic structural variation in human cancer genomes,” *Nature*, vol. 578, no. 7793, pp. 112–121, Feb. 2020.

[120] S. S. Ho, A. E. Urban, and R. E. Mills, “Structural variation in the sequencing era,” *Nat. Rev. Genet. 2019 213*, vol. 21, no. 3, pp. 171–189, Nov. 2019.

[121] B. Gel, A. Díez-Villanueva, E. Serra, M. Buschbeck, M. A. Peinado, and R. Malinverni, “regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests,” *Bioinformatics*, vol. 32, no. 2, pp. 289–291, Jan. 2016.

[122] F. Rodríguez-Castañeda *et al.*, “The SUMO protease SENP1 and the chromatin remodeler CHD3 interact and jointly affect chromatin accessibility and gene expression,” *J. Biol. Chem.*, vol. 293, no. 40, pp. 15439–15454, Oct. 2018.

[123] T. Wang *et al.*, “Identification and characterization of essential genes in the human genome,” *Science*, vol. 350, no. 6264, p. 1096, Nov. 2015.

[124] Y. Liu, X. Zhu, K. Wang, B. Zhang, and S. Qiu, “The Cellular Functions and Molecular Mechanisms of G-Quadruplex Unwinding Helicases in Humans,” *Front. Mol. Biosci.*, vol. 8, p. 1134, Nov. 2021.

[125] N. Maizels, “G4 motifs in human genes,” *Ann. N. Y. Acad. Sci.*, vol. 1267, no. 1, p. 53, 2012.

[126] A. Aguilera and B. Gómez-González, “Genome instability: a mechanistic view of its causes and consequences,” *Nat. Rev. Genet. 2008 93*, vol. 9, no. 3, pp. 204–217, Mar. 2008.

[127] N. Wong, S. John, A. Nussenzweig, and A. Canela, “END-seq: An Unbiased, High-

Resolution, and Genome-Wide Approach to Map DNA Double-Strand Breaks and Resection in Human Cells," *Methods Mol. Biol.*, vol. 2153, pp. 9–31, 2021.

[128] S. V. Lensing, G. Marsico, R. Hänsel-Hertsch, E. Y. Lam, D. Tannahill, and S. Balasubramanian, "DSBCapture: in situ capture and direct sequencing of dsDNA breaks," *Nat. Methods*, vol. 13, no. 10, p. 855, Oct. 2016.

# Appendix A

## A.1 Primers for CRISPR-Cas9 KO screening

| Gene | Chromosome | Forward PCR primer | Reverse PCR primer |
|---|---|---|---|
| RECQL1 | 12 | CCTTTGGCAAGGAGTTTGAA | GCAGGTAAAAGGAGGACCTG |
| BLM | 15 | TGGATTCTTTGCTCAGTTGGGA | TCTCTGTGTTTCCTGTCCTGC |
| BLM | 15 | CCGGACTCTGATTGGGCTTT | GCTAGATCAATGCGGACCGA |
| BLM | 15 | TGGGAATGACCTCTCAAAGC | AAGTGACTTTGGGGTGGTGT |
| WRN | 8 | CCTGTGAGGCATTGACATTTT | ATGCACATGTACCCCGATCT |
| RECQL4 | 8 | GGGTGGATGCCTTAGATGAG | CTCCTCCCACTTCCCTGTTT |
| RECQL5 | 17 | GGGTGGTTCTGCCACTAAAA | TACCGTGGCCTGGAATAGTT |
| RECQL5 | 17 | TGAGGCAGTTCACCTTCAGC | TGCCGCAGCGTTATGGTAT |

**Table A1.1 PCR primer sequences flanking CRISPR-Cas9 gRNA sequences designed for RecQ helicases**

## A.2 Functional validation of RecQ helicase KO cell lines

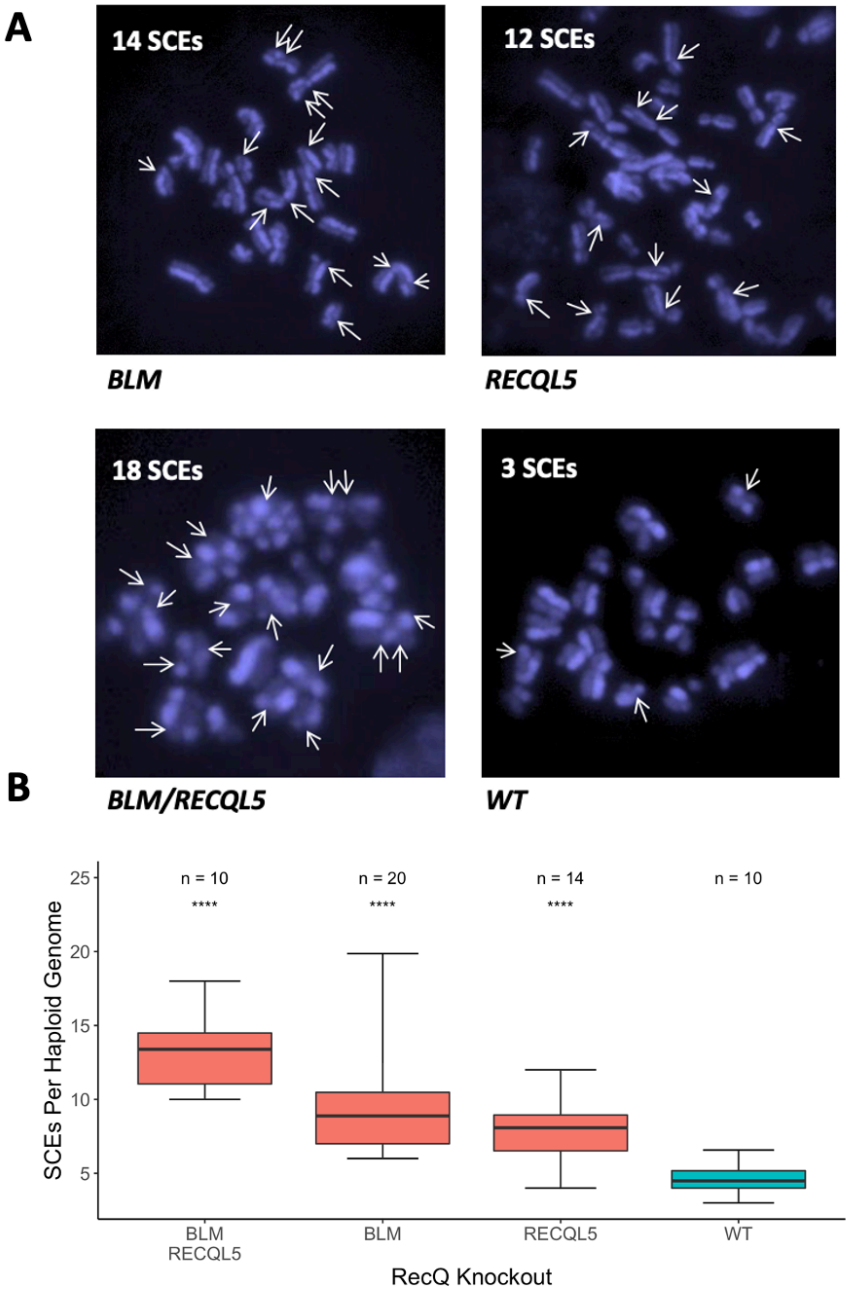



**Figure A2.1 Functional characterization of RecQ helicase KO clones using sister chromatid staining assay. (A) After 2 rounds of cell division in the presence of BrdU, staining with Hoechst 33258 and exposure to UV light reveal differential staining pattern between sister chromatids. White arrows highlights point of exchange between sister chromatids. (B) Number of SCEs detected per haploid genome in a single cell division for RecQ helicase single and double knockouts in the KBM7 cell line. Number of cells analyzed is listed above. Statistical significance was evaluated using a two-sample t-test where WT cells are the control group. **** $p < 0.001$, ** $p < 0.01$, not significant (ns) $p > 0.05$.**

## A.3 SCE enrichment analysis using the same number of SCEs across cell lines

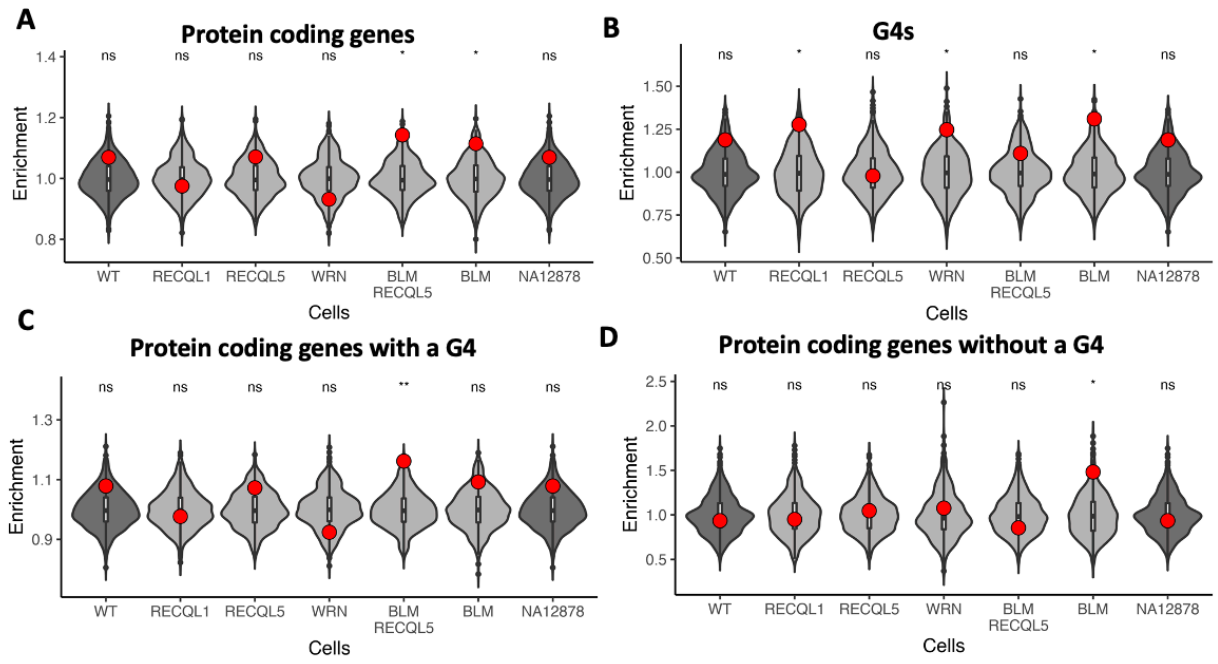



**Figure A3.2 SCE enrichment at protein coding genes and potential G4 quadruplexes using the same number of SCEs across cell lines.**
**SCE enrichment patterns for (A) protein coding genes (B) potential G4s (C) protein coding genes with at least one potential G4 (D) protein coding genes without a potential G4. Normal cell lines are indicated in blue, RecQ KO cell lines in red. Violin plots represent the expected range for random overlap. Red dots represent overlap of SCE with feature of interest. SCEs were randomly down sampled to a constant number for each cell line. P-values calculated from permutation test described in Section 4.2.2. **** $p < 0.001$, ** $p < 0.01$, not significant (ns) $p > 0.05$.**

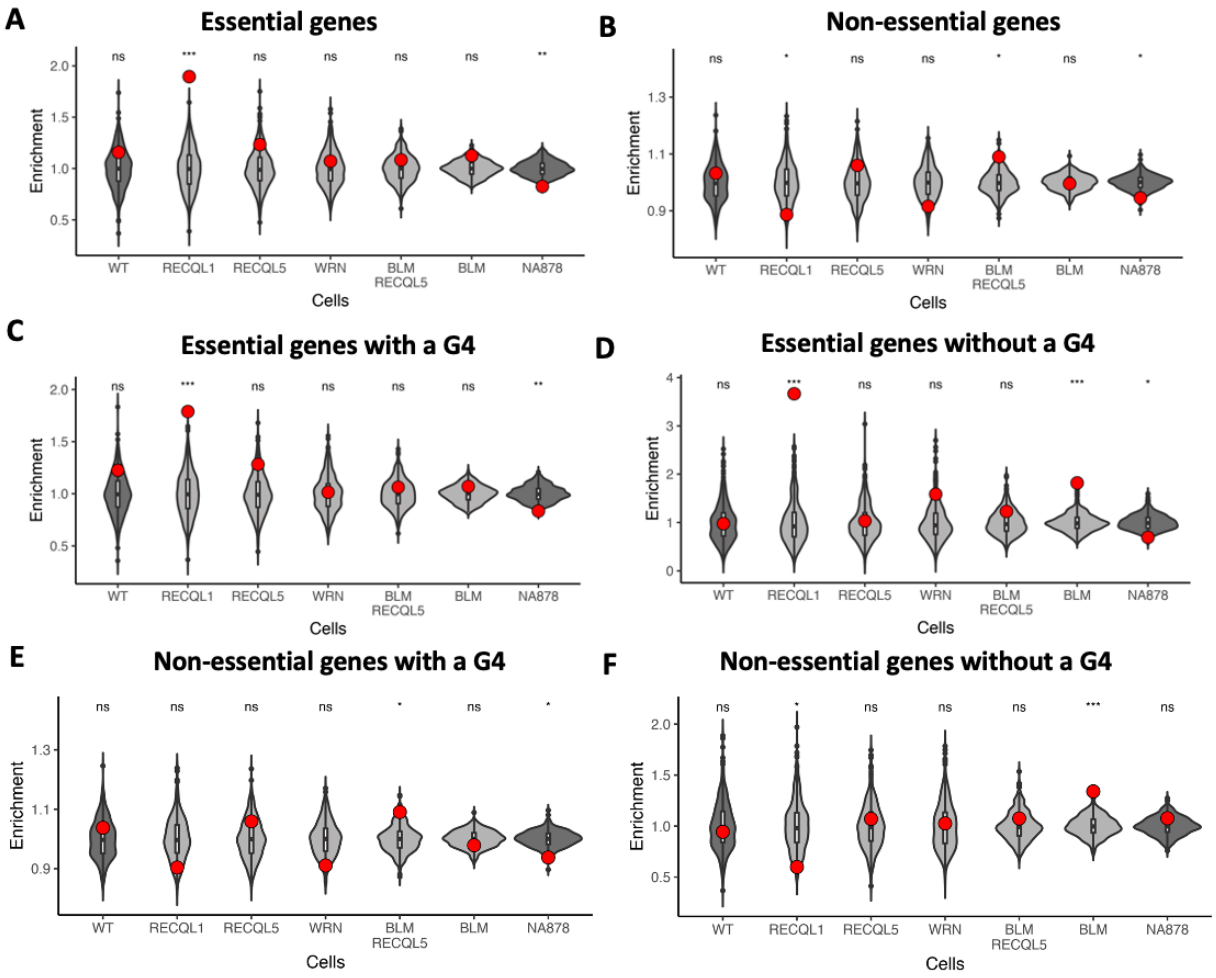



**Figure A3.3 SCE enrichment at essential and non-essential genes with and without potential G4 quadruplexes.**

**SCE enrichment patterns for (A) essential genes (B) non-essential genes (C) essential genes with a potential G4 (D) essential genes without at least one potential G4 (E) non-essential genes with a potential G4 (F) non-essential genes without at least one potential G4. Normal cell lines are indicated in blue, RecQ KO cell lines in red. Violin plots represent the expected range for random overlap. Red dots represent overlap of SCE with FOI described above each plot. P-values calculated from permutation test described in Section 4.2.2. **** $p < 0.001$, ** $p < 0.01$, not significant (ns) $p > 0.05$.**

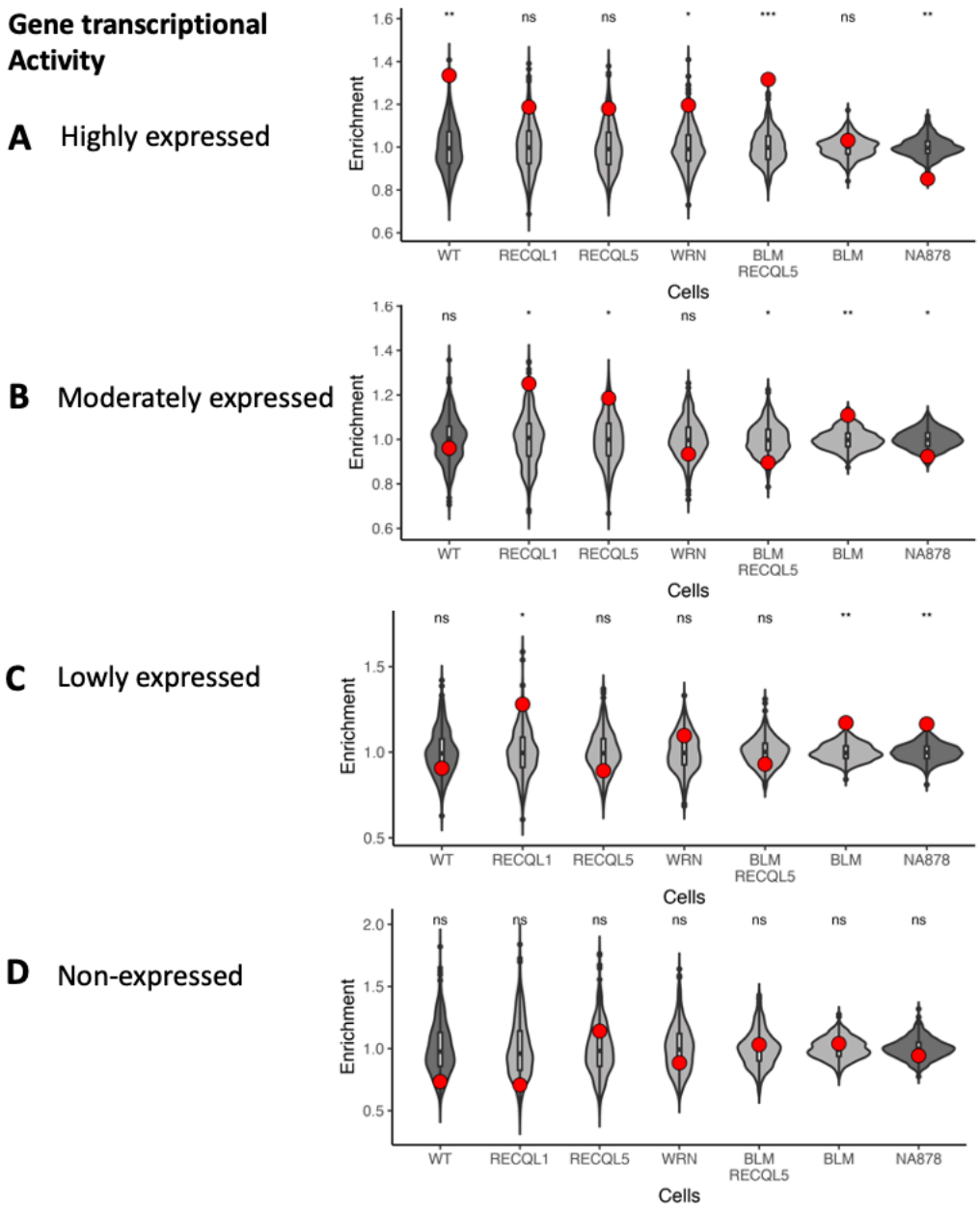



**Figure A3.4 SCE enrichment at genes grouped by transcriptional activity.**
**SCE enrichment patterns for (A) highly expressed genes (B) moderately expressed genes (C) lowly expressed genes (D) non-expressed genes. Normal cell lines are indicated in blue, RecQ KO cell lines in red. Violin plots represent the expected range for random overlap. Red dots represent overlap of SCE with FOI described next to each plot. Genes are grouped by transcriptional activity indicated on the left. Highly transcribed genes have FPKM values in the range of $8.61x10^{0}$ and $4.37x10^{4}$. Moderately expressed genes have FPKM values between the range of $4.02x10^{-1}$ and $8.61x10^{0}$. Lowly expressed genes have FPKM values between the range of $6.56x10^{-6}$ and $4.02x10^{-1}$. P-values calculated from permutation test described in Section 4.2.2. **** $p < 0.001$, ** $p < 0.01$, not significant (ns) $p > 0.05$.**

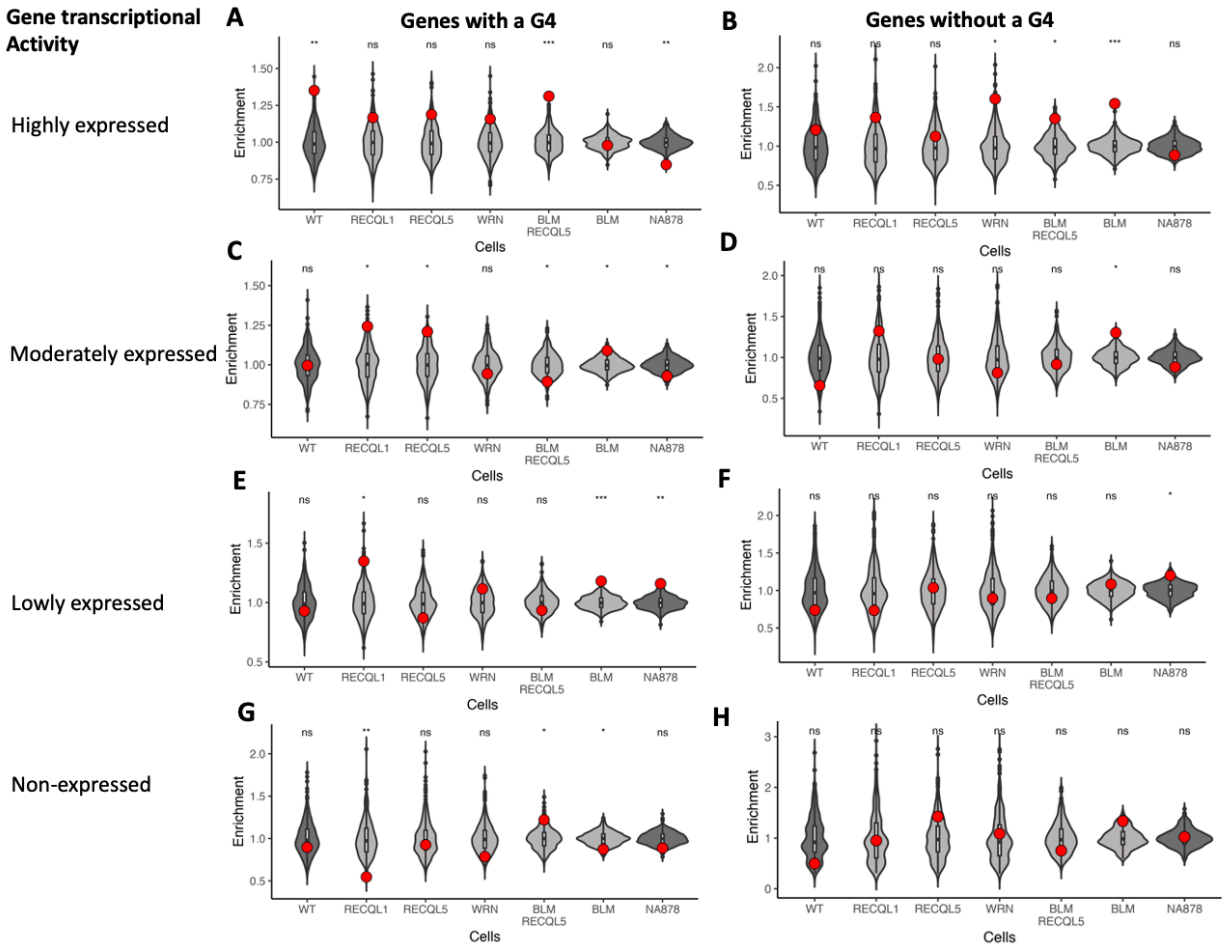



**Figure A3.5 SCE enrichment at transcriptionally grouped genes with potential G4 quadruplexes.**
**SCE enrichment patterns in (A) highly transcribed genes containing at least one potential G4 (B) highly transcribed genes not containing any potential G4s (C) moderately transcribed genes containing at least one potential G4 (D) moderately transcribed genes not containing any potential G4s (E) lowly transcribed genes containing at least one potential G4 (F) lowly transcribed genes not containing any potential G4s (G) non-transcribed genes containing at least one potential G4 and (H) non-transcribed genes not containing any potential G4s. Normal cell lines are indicated in blue, RecQ KO cell lines are indicated in red. Violin plots represent the expected range for random overlap. Red dots represent overlap of SCEs with FOI described next to each plot. Genes are grouped by transcriptional activity indicated on the left. Highly transcribed genes have FPKM values in the range of $8.61x10^{0}$ and $4.37x10^{4}$. Moderately expressed genes have FPKM values between the range of $4.02x10^{-1}$ and $8.61x10^{0}$. Lowly expressed genes have FPKM values between the range of $6.56x10^{-6}$ and $4.02x10^{-1}$. P-values calculated from permutation test described in Section 4.2.2. **** $p < 0.001$, ** $p < 0.01$, not significant (ns) $p > 0.05$.**

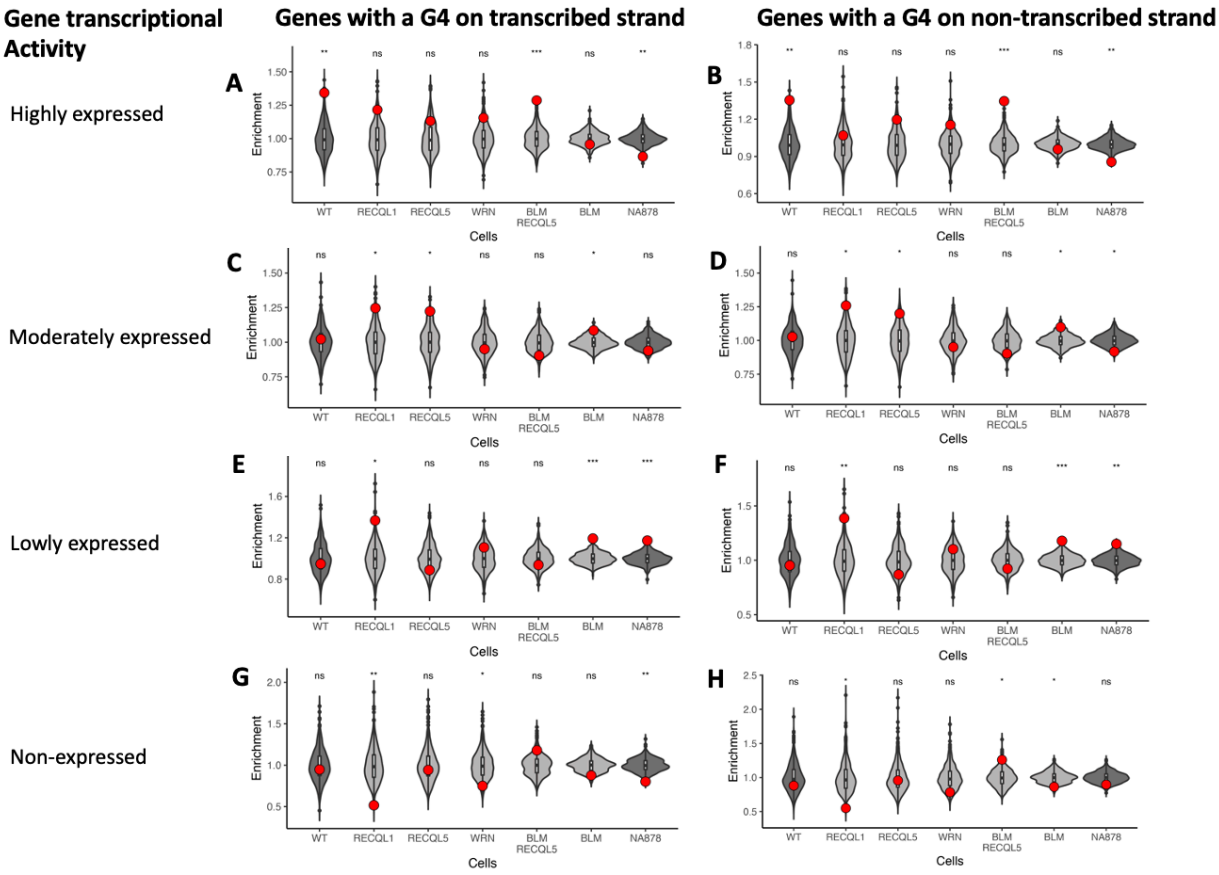



**Figure A3.6 SCE enrichment at transcriptionally grouped genes with potential G4s on coding and template strands .**

**SCE enrichment patterns in (A) highly transcribed genes containing at least one potential G4 on transcribed strand (B) highly transcribed genes containing at least one potential G4 on non-transcribed strand (C) moderately transcribed genes containing at least one potential G4 on transcribed strand (D) moderately transcribed genes containing at least one potential G4 on non-transcribed strand (E) lowly transcribed genes containing at least one potential G4 on transcribed strand (F) lowly transcribed genes containing at least one potential G4 on non-transcribed strand (G) non-transcribed genes containing at least one potential G4 on transcribed strand and (H) non-transcribed genes containing at least one potential G4 on non-transcribed strand. Normal cell lines are indicated in blue, RecQ KO cell lines are indicated in red. Violin plots represent the expected range for random overlap. Red dots represent overlap of SCEs with FOI described next to each plot. Genes are grouped by transcriptional activity indicated on the left. Highly transcribed genes have FPKM values in the range of $8.61 \times 10^{0}$ and $4.37 \times 10^{4}$. Moderately expressed genes have FPKM values between the range of $4.02 \times 10^{-1}$ and $8.61 \times 10^{0}$. Lowly expressed genes have FPKM values between the range of $6.56 \times 10^{-6}$ and $4.02 \times 10^{-1}$. P-values calculated from permutation test described in Section 4.2.2. **** $p < 0.001$, ** $p < 0.01$, not significant (ns) $p > 0.05$.**

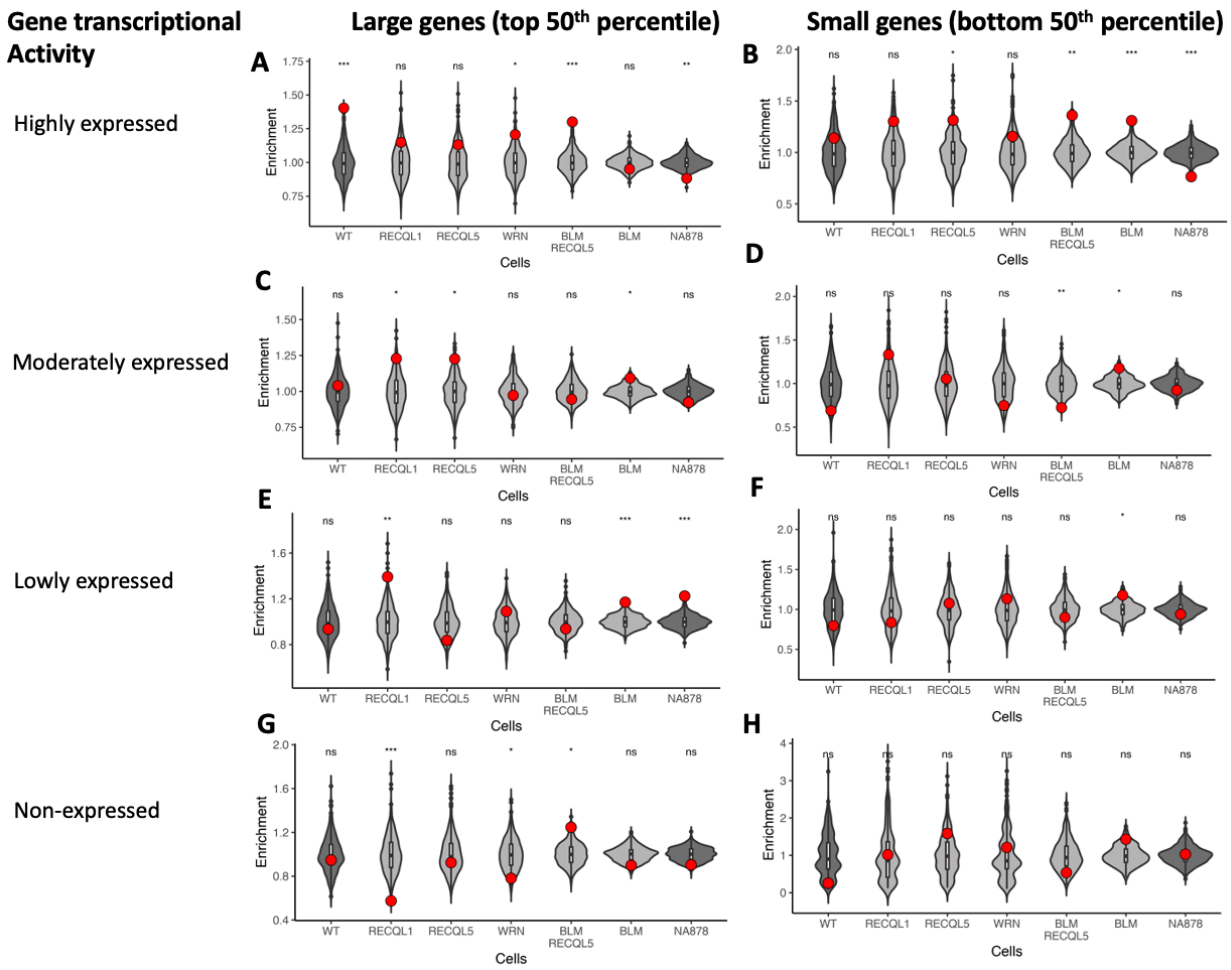



**Figure A3.7 SCE enrichment at transcriptionally grouped large and small genes.**
**SCE enrichment patterns in (A) highly transcribed large genes (B) highly transcribed small genes (C) moderately transcribed large genes (D) moderately transcribed small genes (E) lowly transcribed large genes (F) lowly transcribed small genes (G) non-transcribed large genes (H) non-transcribed small genes. Normal cell lines are indicated in blue; RecQ KO cell lines are indicated in red. Violin plots represent the expected range for random overlap. Red dots represent overlap of SCEs with FOI described next to each plot. Genes are grouped by transcriptional activity indicated on the left. Highly transcribed genes have FPKM values in the range of $8.61x10^{0}$ and $4.37x10^{4}$. Moderately expressed genes have FPKM values between the range of $4.02x10^{-1}$ and $8.61x10^{0}$. Lowly expressed genes have FPKM values between the range of $6.56x10^{-6}$ and $4.02x10^{-1}$. P-values calculated from permutation test described in Section 4.2.2. **** $p < 0.001$, ** $p < 0.01$, not significant (ns) $p > 0.05$.**

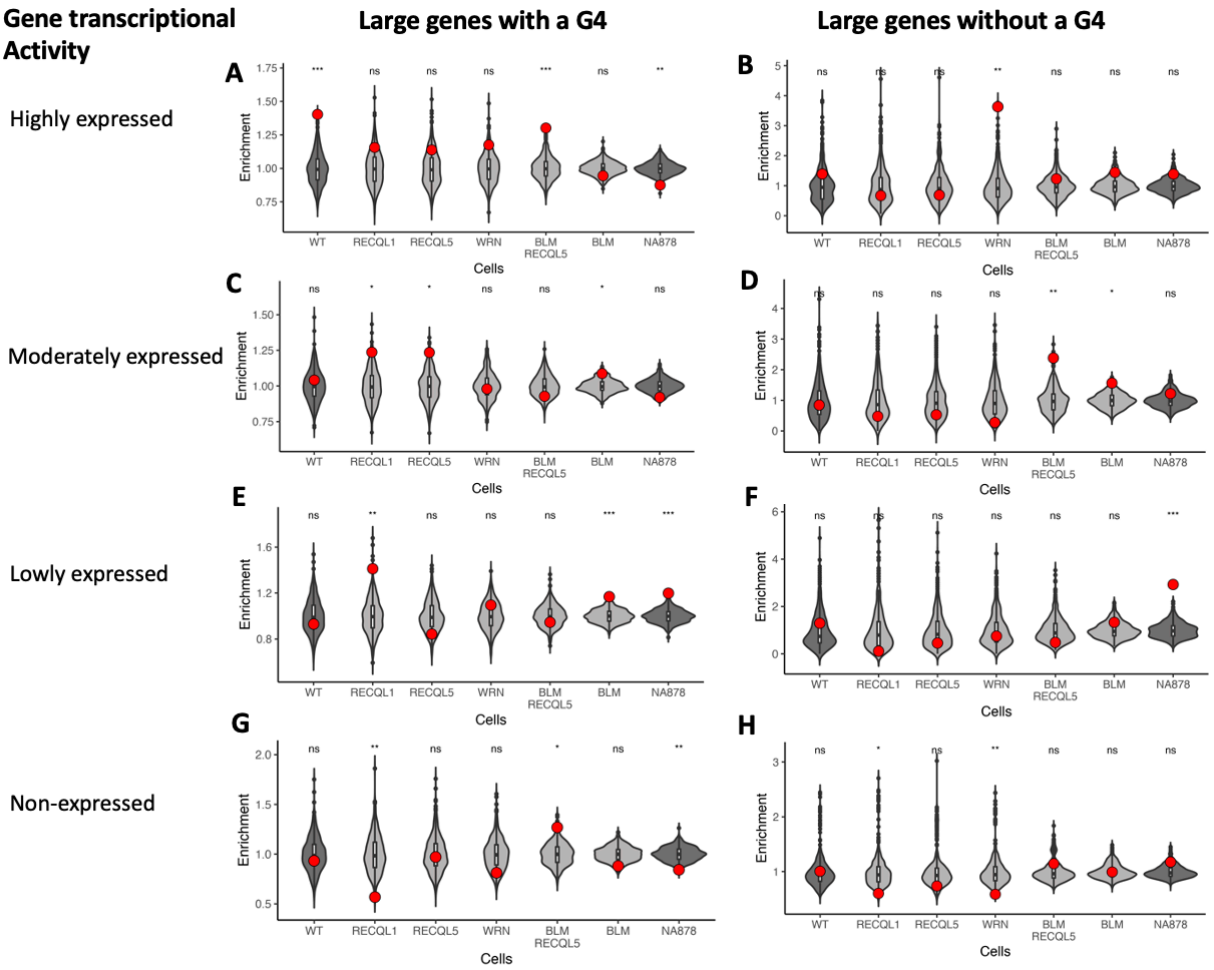



**Figure A3.8 SCE enrichment at transcriptionally grouped large genes with potential G4 quadruplexes. SCE enrichment patterns in (A) highly transcribed large genes containing at least one potential G4 (B) highly transcribed large genes not containing any potential G4s (C) moderately transcribed large genes containing at least one potential G4 (D) moderately transcribed large genes not containing any potential G4s (E) lowly transcribed large genes containing at least one potential G4 (F) lowly transcribed large genes not containing any potential G4s (G) non-transcribed large genes containing at least one potential G4 and (H) non-transcribed large genes not containing any potential G4s. Normal cell lines are indicated in blue, RecQ KO cell lines are indicated in red. Violin plots represent the expected range for random overlap. Red dots represent overlap of SCEs with FOI described next to each plot. Genes are grouped by transcriptional activity indicated on the left. Highly transcribed genes have FPKM values in the range of $8.61\text{x}10^{0}$ and $4.37\text{x}10^{4}$. Moderately expressed genes have FPKM values between the range of $4.02\text{x}10^{-1}$ and $8.61\text{x}10^{0}$. Lowly expressed genes have FPKM values between the range of $6.56\text{x}10^{-6}$ and $4.02\text{x}10^{-1}$. P-values calculated from permutation test described in Section 4.2.2. **** $p < 0.001$, ** $p < 0.01$, not significant (ns) $p > 0.05$.**

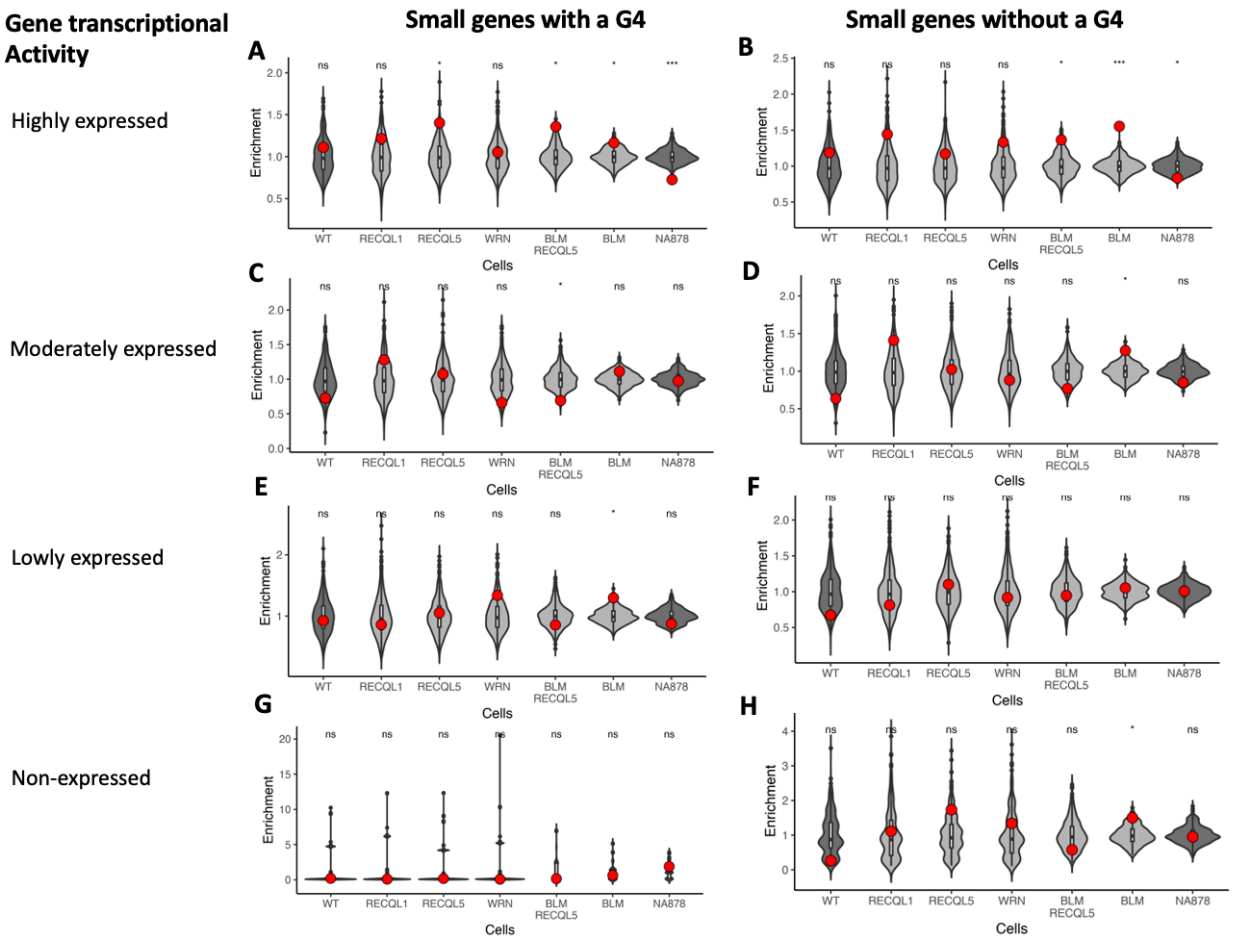



**Figure A3.9 SCE enrichment at transcriptionally grouped small genes with potential G4 quadruplexes. SCE enrichment patterns in (A) highly transcribed small genes containing at least one potential G4 (B) highly transcribed small genes not containing any potential G4s (C) moderately transcribed small genes containing at least one potential G4 (D) moderately transcribed small genes not containing any potential G4s (E) lowly transcribed small genes containing at least one potential G4 (F) lowly transcribed small genes not containing any potential G4s (G) non-transcribed small genes containing at least one potential G4 and (H) non-transcribed small genes not containing any potential G4s. Normal cell lines are indicated in blue, RecQ KO cell lines are indicated in red. Violin plots represent the expected range for random overlap. Red dots represent overlap of SCEs with FOI described next to each plot. Genes are grouped by transcriptional activity indicated on the left. Highly transcribed genes have FPKM values in the range of $8.61x10^{0}$ and $4.37x10^{4}$. Moderately expressed genes have FPKM values between the range of $4.02x10^{-1}$ and $8.61x10^{0}$. Lowly expressed genes have FPKM values between the range of $6.56x10^{-6}$ and $4.02x10^{-1}$. P-values calculated from permutation test described in Section 4.2.2. **** $p < 0.001$, ** $p < 0.01$, not significant (ns) $p > 0.05$.**