**UNRAVELLNG RECQ HELICASE FUNCTION IN GENOME**

**STABILITY USING STRAND-SEQ**

by

Zeid Hamadeh

B.MSc., 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)

July 2022

© Zeid Hamadeh, 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:**

|  |
| --- |
| Peter Lansdorp, Medical Genetics, UBC |
| Supervisor |
| Judy Wong, Medical Genetics, UBC |
| Supervisory Committee Member |
| Peter Stirling, Medical Genetics, UBC |
| Supervisory Committee Member |
|  |
| University Examiner |
|  |
| University Examiner |

**Additional Supervisory Committee Members:**

|  |
| --- |
| Martin Hirst, Microbiology and Immunology, UBC |
| Supervisory Committee Member |
|  |
| Supervisory Committee Member |

# Abstract

Helicases are a highly conserved motor protein responsible for interacting with and unwinding canonical and non-canonical DNA structures. The RecQ class of helicases, known to suppress illegitimate recombination, are implicated in aging and cancer with three of the five human RecQ helicases directly linked to premature aging syndromes characterized by strong cancer predisposition. 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. We 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. We 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. We performed enrichment analysis to show SCEs are frequently occurring near actively transcribed genes with guanine quadruplexes and common fragile sites further supporting the role of these genes in suppressing inappropriate recombination at specific genomic elements. We also developed novel bioinformatic approaches to generate genotype-specific call sets for 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

It **must not** exceed 150 words.

The lay summary will explain the key goals and contributions of the scholarly work in terms that can be readily understood by the general public. It is common for research ethics boards and funding agencies to require abstracts that are accessible for the lay public. University scholars must be able to communicate effectively with the communities they serve and to demonstrate the relevance of their scholarship to the public good. The lay summary provides an educational and developmental opportunity for students to see their work as part of the greater knowledge of the broader community.

# Preface

This is the Preface. As of May 2013, **all** theses and dissertations are **required** to have a preface, which appears immediately after the Abstract and before the Table of Contents. The preface **must** include a statement indicating your contribution to the following:

* Identification and design of the research program,
* Performance of various parts of the research, and
* Analysis of the research data.

In addition, if any part of your research was partly or wholly published in articles, was part of a collaboration, or required the approval of one or more of UBC’s Research Ethics Boards, you must include that information in your preface as well. This information would be included as follows:

* A statement indicating the relative contributions of all collaborators and co-authors of publications (if any), emphasizing details of your contribution and stating the proportion of research and writing conducted by you
* A list of any publications arising from work presented in the thesis, and the chapter(s) in which the work is located. **Do not** include the actual publications at any point in your thesis.
* The name of the particular UBC Research Ethics Board, and the Certificate Number(s) of the Ethics Certificate(s) obtained, if ethics approval was required for the research. **Do not** include the actual Ethics Certificate(s) at any point in your thesis. If you are not sure if your research requires ethics approval, please contact the [Office of Research Services](http://www.ors.ubc.ca/home) as early in your research process as possible.

Examples of statements regarding collaborations or lists of publications arising from the research:

Chapter 2 is based on work conducted in UBC’s Maple Syrup Laboratory by Dr. A. Apple, Professor B. Boat, and [student’s name]. I was responsible for tapping the trees in forests X and Z, conducted and/or supervised all boiling operations, and performed frequent quality control tests on the product.

or

A version of chapter 3 has been published. [Student’s name], Apple, A. and Boat. B. (2010) Frequency of Quality Testing in Syrup Creation. Maple Science J. 255:139-144. I conducted all the testing and wrote most of the manuscript. The section on “Testing Implements” was originally drafted by Boat, B.

If your research did not require ethics approval, was conducted as part of a collaboration, was not previously published in whole or in part, and was designed, carried out, and analyzed by you alone, your preface may be very short. The contents of the preface must be verified by your supervisor. Please note that preface requirements may change over time; please check the [Faculty of Graduate Studies website](http://www.grad.ubc.ca/current-students/dissertation-thesis-preparation) to ensure that you are following the most current requirements.

# Table of Contents

Abstract iii

Lay Summary iv

Preface v

Table of Contents vii

List of Tables xiii

List of Figures xiv

List of Symbols xv

List of Abbreviations xvi

Glossary xvii

Acknowledgements xviii

Dedication xix

Chapter 1: Introduction 20

1.1 Overview 20

1.1.1 The human genome 20

1.1.2 Genetic disease 20

1.1.2.1 Cancer 20

1.1.2.2 Aging 20

1.1.2.3 Other genetic diseases 20

1.1.3 DNA transactions 20

1.1.3.1 DNA damage and repair 20

1.1.3.2 DNA replication 21

1.1.3.3 DNA transcription 21

1.1.3.4 The cell-cycle 21

1.1.4 Mechanisms of genome instability 21

1.1.4.1 Replication-associated genomic instability 21

1.1.4.1.1 Replication fork stalling and collapse 21

1.1.4.1.2 Replication stress and fragile sites 21

1.1.4.1.3 Incomplete DNA replication and ultrafine bridges 21

1.1.4.2 Transcription-coupled genome instability 21

1.1.4.2.1 Collision between transcription and replication machinery 21

1.1.4.2.2 Co-transcriptional R-loops 21

1.1.4.3 Genome instability syndromes 21

1.1.5 Mechanisms of DNA repair 21

1.1.5.1 Repair of small DNA lesions 21

1.1.5.2 Types of small DNA lesions and repair 21

1.1.5.3 Point mutations 21

1.1.5.4 Repair of large DNA lesions 21

1.1.5.5 Types of DNA lesions and repair 21

1.1.5.6 Structural variants 21

1.2 Challenges in studying DNA repair 22

1.2.1 Methods of detection 22

1.2.1.1 Cytogenetics/FISH 22

1.2.1.2 Challenges and limitations 22

1.2.2 High resolution SCE mapping using Strand-seq 22

1.2.2.1 Technology 22

1.2.2.2 Bioinformatic tools 22

1.3 Thesis rationale, objective and research scope 22

1.3.1 Chapter 2: New OP-Strand-seq pipeline for studying DNA repair 22

1.3.2 Chapter 3: Bioinformatic tools for uncovering multiple SV classes 22

1.3.3 Chapter 4: Role of RecQ helicases and SCEs and genome stability 22

1.3.4 Chapter 5: Conclusions and future directions 22

Chapter 2: New OP-Strand-seq pipeline for studying DNA repair 23

2.1 Introduction 23

2.1.1 Original Strand-seq protocol 23

2.1.2 Applications of original Strand-seq method for studying DNA repair 24

2.1.3 Limitations of original Strand-seq method for studying DNA repair 25

2.1.3.1 Background limits breakpoint resolution 25

2.1.3.2 Constraints on throughput of manual library quality selection 25

2.1.3.3 Need for automated library classification system 25

2.2 Methods 25

2.2.1 Knockout model generation in haploid cell line using CRISPR-Cas9 25

2.2.2 High throughput “one-pot” protocol 26

2.2.2.1 Library preparation 27

2.2.2.2 Illumina whole genome sequencing 27

2.2.2.3 Bioinformatic pre-processing 27

2.2.3 Bioinformatic classifier for sorting Strand-seq libraries 27

2.3 Results 27

2.3.1 Comparison of breakpoint resolution between haploid and diploid cells 27

2.3.2 Comparison library classification accuracy to existing methods 28

2.4 Discussion 28

2.4.1 Applications of novel methods for DNA repair studies 28

2.4.2 Conclusions 28

Chapter 3: Structural variant callers 29

3.1 Introduction 29

3.1.1 Rationale for improving SV callers in the field to study intra-tumour heterogeneity 29

3.2 Methods 29

3.2.1 SCEs calling 29

3.2.1.1 Bioinformatic approaches to SCE detection 29

3.2.1.2 Genotype-specific differences in SCE frequency 29

3.2.2 CNAs calling 29

3.2.2.1 Bioinformatic approaches to CNAs detection 29

3.2.2.2 Genotype-specific differences in CNAs frequency 29

3.2.3 Translocation calling 29

3.2.3.1 Bioinformatic approaches to translocations detection 29

3.2.3.2 Genotype-specific differences in translocations frequency 29

3.3 Results 29

3.3.1 Compare accuracy (precision/recall) of different callers to existing tools 29

3.4 Discussion 29

3.4.1 Applications for future studies in uncovering complete mutational landscape 29

Chapter 4: Role of BLM and RECQL5 in DNA repair using mainly SCE data 30

4.1 Introduction 30

4.1.1 Role of DNA helicases in DNA repair 30

4.1.2 Non-redundant roles of RecQ helicases 30

4.2 Methods 30

4.2.1 Generation of comprehensive SV call sets 30

4.2.2 Bioinformatic tools for running enrichment tests between genetic elements 30

4.2.3 Genetic element datasets 30

4.2.3.1 Collection of essential/non-essential genes for KBM7 cell line 30

4.2.3.2 Collection of actively/non-actively transcribed genes for KBM7 cell line 30

4.2.3.3 Collection of experimentally reported guanine quadruplexes (G4) 30

4.3 Results 30

4.3.1 RecQ helicase KO genotype-specific differences in SCE frequency 30

4.3.2 Enrichment analysis of SCEs 30

4.4 Discussion 30

4.4.1.1 Role of RecQ helicases in SCE and G4 biology 30

4.4.1.2 Conclusions 30

Chapter 5: Conclusion 31

5.1 Discussion of results 31

5.1.1 Novel methods 31

5.1.1.1 Strand-seq library generation for DNA repair studies 31

5.1.1.2 Generation of complete SV call sets 31

5.1.1.3 Enrichment analysis of SV coordinates 31

5.1.2 Novel findings of RecQ helicases and implications on field of DNA repair 31

5.2 Strengths and limitations of the research 31

5.2.1 Strengths 31

5.2.1.1 Novel method allows for improved resolution of studying DNA repair events 31

5.2.1.2 New callers exceed accuracy performance of existing/published methods 31

5.2.2 Weaknesses 31

5.2.2.1 No molecular work in protein or RNA to deduce functional consequences of different SVs 31

5.2.2.2 Limitations on overall accuracy of callers 31

5.3 Discussion of any potential applications of the findings 31

5.3.1 Investigating RecQ helicases in the context of DNA repair 31

5.3.2 Using bioinformatic tools to uncover SVs in single cells 31

5.4 A description of possible future research directions 32

5.4.1 Resolving intra-tumour heterogeneity 32

Bibliography 33

Appendices 36

Appendix A 36

A.1 Sub-Appendix 36

A.2 Sub-Appendix 37

Appendix B 38

B.1 Sub-Appendix 38

# List of Tables

This is the List of Tables. This must be included if you have any tables in your thesis. If you do not have any tables in your thesis, you should delete this page. If you have taken any tables from other sources, please contact the [Copyright Office](http://copyright.ubc.ca/) as early in your writing process as possible.

Tables must be numbered consecutively. They can be numbered with the chapter number or with no reference to the chapter number. You **must** use the same numbering system for tables, figures, illustrations, etc. Avoid mixing numbering systems.

Your table captions should either be consistently in sentence case (i.e., only the first letter of the caption and any proper nouns are capitalized) or consistently in title case (i.e., most words in the caption are capitalized, with the general exception of prepositions, conjunctions and articles).

Table 2.1 A sample of a table in Chapter 2 5

Table 3.1 A sample of a table in Chapter 3 7

Table 3.2 Second sample of a table in Chapter 3 8

# List of Figures

This is the List of Figures. This must be included if you have any figures in your thesis. If you do not have any figures in your thesis, you should delete this page. If you have taken any figures from other sources, please contact the [Copyright Office](http://copyright.ubc.ca/) as early in your writing process as possible.

Figures must be numbered consecutively. They can be numbered with the chapter number or with no reference to the chapter number. You **must** use the same numbering system for tables, figures, illustrations, etc. Avoid mixing numbering systems.

Your figure captions should either be consistently in sentence case (i.e., only the first letter of the caption and any proper nouns are capitalized) or consistently in title case (i.e., most words in the caption are capitalized, with the general exception of prepositions, conjunctions and articles).

Figure 2.1 SCE breakpoint resolution relative to sequencing effort for haploid and diploid cells 29

# List of Symbols

If you are using specialized symbols in your thesis, list them here. Lists of symbols, lists of abbreviations, glossaries, etc., may go in any order you wish, so long as they appear after the Lists of Tables and/or Figures and before the Acknowledgements. If you do not have any specialized symbols in your thesis and do not need a list of them, you should delete this page.

# List of Abbreviations

If you are using abbreviations in your thesis, list them here. Lists of symbols, lists of abbreviations, glossaries, etc., may go in any order you wish, so long as they appear after the Lists of Tables and/or Figures and before the Acknowledgements. If you do not have any abbreviations in your thesis and do not need a list of them, you should delete this page.

# Glossary

If you require a glossary of specialized terms for your thesis, list them here. Lists of symbols, lists of abbreviations, glossaries, etc., may go in any order you wish, so long as they appear after the Lists of Tables and/or Figures and before the Acknowledgements. If you do not require a glossary for your thesis, you should delete this page.

# Acknowledgements

This is the Acknowledgements section.

The following are usually mentioned in the Acknowledgements:

* Supervisor and committee
* Grant support
* Helpful fellow students, lab mates, etc.
* Family support

Acknowledgements may extend for more than one page, but should be no longer than two pages.

Example:

I offer my enduring gratitude to the faculty, staff and my fellow students at UBC, who have inspired me to continue my work in this field. I owe particular thanks to Dr. J. M. Schneider, whose penetrating questions taught me to question more deeply.

I thank Dr. O. Meyer for enlarging my vision of science and providing coherent answers to my endless questions.

Special thanks are owed to my parents, whose have supported me throughout my years of education, both morally and financially.

# Dedication

This is the Dedication.

The dedication must be listed in the Table of Contents. You may decide you do not want to have the word “Dedication” appear on this page. In Microsoft Word, to get “Dedication” to show up in the Table of Contents without appearing here, change the colour of the word “Dedication” to white. You will then need to change the font colour of “Dedication” back to black in the Table of Contents.

## Introduction

### Overview

#### The human genome

What is the human genome?

How big is it? How it fits in the central dogma of biology

#### Genetic disease

What are genetic diseases? Mutations in the human genome

Hereditary vs de novo mutations

Cancer and aging both stem from the accumulation of mutation rates

##### Cancer

What is cancer? Hallmarks

Genome stability and accumulation of mutation rate

Hereditary vs sporadic

##### Aging

What is aging? Hallmarks

Genome stability and accumulation of mutation rate

##### Other genetic diseases

Monogenic vs polygenic

#### DNA transactions

How do mutations occur?

##### DNA damage and repair

Exogenous vs endogenous

Different repair pathways and mutation types

##### DNA replication

##### DNA transcription

##### The cell-cycle

#### Mechanisms of genome instability

##### Replication-associated genomic instability

###### Replication fork stalling and collapse

###### Replication stress and fragile sites

###### Incomplete DNA replication and ultrafine bridges

##### Transcription-coupled genome instability

###### Collision between transcription and replication machinery

###### Co-transcriptional R-loops

##### Genome instability syndromes

#### Mechanisms of DNA repair

##### Repair of small DNA lesions

##### Types of small DNA lesions and repair

##### Point mutations

##### Repair of large DNA lesions

##### Types of DNA lesions and repair

##### Structural variants

### Challenges in studying DNA repair

#### Methods of detection

##### Cytogenetics/FISH

##### Challenges and limitations

#### High resolution SCE mapping using Strand-seq

##### Technology

##### Bioinformatic tools

### Thesis rationale, objective and research scope

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

#### Chapter 3: Bioinformatic tools for uncovering multiple SV classes

#### Chapter 4: Role of RecQ helicases and SCEs and genome stability

#### Chapter 5: Conclusions and future directions

## New OP-Strand-seq pipeline for studying DNA repair

### Introduction

The main goal of this chapter is to introduce novel approaches for performing DNA repair studies using Strand-seq. Establishment of standardized methods to study DNA repair is a prerequisite to elucidate the role of DNA repair in cancer and aging. Three novel implementations in this chapter include the development and use of the “One-pot” Strand-seq protocol, generating CRISPR-Cas9 haploid KO lines for Strand-seq and the classification of Strand-seq library quality.

#### Original Strand-seq protocol

Single-cell template strand sequencing (Strand-seq) is a technique discovered in 2012 for the selective sequencing of a dividing daughter cell’s parental template strands used during DNA replication. 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 conventional sequencing approaches.

A dividing diploid cell replicates both Watson (W; minus or 3’-5’ strand of reference assembly) and Crick (C, plus or 3’-5’ strand) strands of each homologous chromosome in a semi-conservative fashion. Strand-seq exploits the semi-conservative nature of DNA replication to incorporate the thymidine analog bromodeoxyuridine (BrdU) into the newly synthesized strand to allow for the distinction between template and nascent DNA strands. DNA fragments with BrdU can be selectively degraded by treatment with Hoechst and UV irradiation before PCR amplification [3]. PCR amplification after degradation of nascent strand DNA fragments allows for the selective amplification of template strand reads. Pools of Strand-seq libraries can then be loaded onto a Illumina MiSeq or NextSeq instrument for paired end Illumina whole-genome sequencing (WGS) [3], [4]. 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. Strand-seq libraries thus harbor unique signatures of intra-chromosomal template strand changes that represent orientation-dependent SVs. The original protocol can construct several dozen Strand-seq cells at a time using a liquid-handling platform for automation.

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

One of the unique applications of Strand-seq is the ability to putatively identify complex SVs by pinpointing changes in template strand inheritance. 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 inheritance, or strand state genotypes, for each chromosome in a given daughter cell (W-W, C-C, W-C). Intra-chromosomal changes in strand state in a cell suggest some sort of SV has occurred and can be used to putatively identify different SV breakpoints [16]. Intra-chromosomal strand state segments can be genotyped and coordinates for the exact breakpoint of strand state changes can refined using Fisher's exact test [16]. Putative breakpoints can be used to call different SVs based on the strand state genotype, changes in read counts and the occurrence this breakpoint in multiple cells. For example, a sister chromatid exchange event (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). 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 (e.g. homozygous inversion; WW-CC-WW or heterozygous inversion; WW-WC-WW).

#### 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 only 96 cells. Additionally, there is high variability in the quality of Strand-seq libraries resulting in poor Strand-seq library characteristics. 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. 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. 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 poor Strand-seq characteristics. 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.

### Methods

Three novel approaches introduced here improve the ability to investigate DNA repair using Strand-seq.

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

The first implementation is to generate and use haploid knockout lines because haploid cells offer two main advantages over a diploid cell line. A cell line with only one haplotype is easier to generate knockout lines because only one allele must be altered to produce a bonified knockout. Secondly, having one haplotype improves the resolution in which SVs can be mapped to the genome. 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, C), versus three in a diploid cell (WW, CC, WC). This makes intra-chromosomal strand state changes less ambiguous because the breakpoint CI can be placed directly between a Crick and a Watson read whereas in a diploid cell there is uncertainty of whether a Watson read at the breakpoint is part of the WC segment or the WW segment, resulting in a breakpoint CI of one read plus gaps to the next adjacent read on both sides (Figure 2A).

We used CRISPR-Cas9 to generate knockout clones for RECQL1, BLM, WRN, RECQL5, and both BLM and RECQL5 and WRN and RECQL5 together in the haploid KBM7 cell line. Two guides were designed per gene with maximized predicted on-target efficiency and minimized predicted off-target efficiency [1]. Assembled ribonucleoprotein CRISPR-Cas9 complexes were electroporated into cells and after 24 hours single cells, positive for gRNA fluorescence, were sorted by FACS into individual wells of 96 well plates. Colonies were grown up and DNA isolated from expanded cultures was used for PCR amplification of segments flanking the gRNA sequences and subsequent Sanger sequencing to characterize frameshifting mutations. This procedure was used to isolate 1-4 distinct clones for RECQL1, BLM, WRN, RECQL5 single KO cells and WRN/RECQL5 and BLM/RECQL5 double KO cells.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Table 2.1 A sample of a table in Chapter 2

#### High throughput “one-pot” protocol

We developed a modified version of the original version of the Strand-seq protocol known as one-pot (OP)-Strand-seq to improve the quality and throughput of single cell sequencing libraries. This approach differs from the original method described in Sanders et al., 2017 in three 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 [2], [3]. 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. 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.

##### Library preparation

##### Illumina whole genome sequencing

##### Bioinformatic pre-processing

#### Bioinformatic classifier for sorting 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 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 is 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 74%. Therefore, we developed a novel classifier to automate the selection of good quality haploid OP-Strand-seq libraries.

First, we manually benchmarked 4000+ libraries to be split 80:20 for a training and test set. Then we 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. Background, as previously mentioned, refers to a metric devised to calculate the proportion of non-directional reads. Evenness is a metric that assesses the genome wide variability in sequencing coverage where the genome is split into Mb-sized bins whereas spikiness refers to the bin-to-bin variability in sequencing coverage. Together, we used these four features to predict the classifications of good Strand-seq libraries using a random forest algorithm from the R package, Caret.

### Results

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

For the same sequencing effort, the resolution of SCEs is higher in haploid cells than in diploid cells (p=2e-10).

Chart

Description automatically generated

Figure 2.1 SCE breakpoint resolution relative to sequencing effort for haploid and diploid cells

#### Comparison library classification accuracy to existing methods

Diagram

Description automatically generated

Figure 2.2 Properties of Strand-seq libraries and ROC curves

### Discussion

#### Applications of novel methods for DNA repair studies

#### Conclusions

## Structural variant callers

### Introduction

#### Rationale for improving SV callers in the field to study intra-tumour heterogeneity

### Methods

#### SCEs calling

##### Bioinformatic approaches to SCE detection

##### Genotype-specific differences in SCE frequency

#### CNAs calling

##### Bioinformatic approaches to CNAs detection

##### Genotype-specific differences in CNAs frequency

#### Translocation calling

##### Bioinformatic approaches to translocations detection

##### Genotype-specific differences in translocations frequency

### Results

#### Compare accuracy (precision/recall) of different callers to existing tools

### Discussion

#### Applications for future studies in uncovering complete mutational landscape

## Role of BLM and RECQL5 in DNA repair using mainly SCE data

### Introduction

#### Role of DNA helicases in DNA repair

#### Non-redundant roles of RecQ helicases

### Methods

#### Generation of comprehensive SV call sets

#### Bioinformatic tools for running enrichment tests between genetic elements

#### Genetic element datasets

##### Collection of essential/non-essential genes for KBM7 cell line

##### Collection of actively/non-actively transcribed genes for KBM7 cell line

##### Collection of experimentally reported guanine quadruplexes (G4)

### Results

#### RecQ helicase KO genotype-specific differences in SCE frequency

#### Enrichment analysis of SCEs

### Discussion

##### Role of RecQ helicases in SCE and G4 biology

##### Conclusions

## Conclusion

### Discussion of results

#### Novel methods

##### Strand-seq library generation for DNA repair studies

##### Generation of complete SV call sets

##### Enrichment analysis of SV coordinates

#### Novel findings of RecQ helicases and implications on field of DNA repair

### Strengths and limitations of the research

#### Strengths

##### Novel method allows for improved resolution of studying DNA repair events

##### New callers exceed accuracy performance of existing/published methods

#### Weaknesses

##### No molecular work in protein or RNA to deduce functional consequences of different SVs

##### Limitations on overall accuracy of callers

### Discussion of any potential applications of the findings

#### Investigating RecQ helicases in the context of DNA repair

#### Using bioinformatic tools to uncover SVs in single cells

### A description of possible future research directions

#### Resolving intra-tumour heterogeneity

# Bibliography

[1] J. P. Concordet and M. Haeussler, “CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens,” *Nucleic Acids Res.*, vol. 46, no. W1, pp. W242–W245, Jul. 2018.

[2] 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.

[3] 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.

[4] 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.

This is the bibliographic section of your thesis. What you will use as the title of this section will depend on what style of referencing you are using. “Bibliography” refers to a general list of information about sources, including both sources you have cited and sources you have consulted but not cited. “References” refers to a list of sources cited; sources consulted but not cited are not included in a reference list. “Works Cited” is the name given to a reference list in the MLA style. There must be a single bibliography for the entire thesis. For information on a variety of citation styles, please see UBC Library’s [How to Cite Sources](http://learningcommons.ubc.ca/get-study-help/research-and-writing/citing-sources/) page. The bibliographic and citation style you use will be determined by the discipline you are in; consult your supervisor for advice on appropriate styles.

Author, A. 2010. Title, publication details, in style permitted by your discipline. Please consult a style manual appropriate for your discipline for further details.

Author, B. 2010. Title, publication details, in style permitted by your discipline. Please consult a style manual appropriate for your discipline for further details.

# Appendices

This title use Heading 1

This is Appendix A

[Type a space and apply style “Heading 7” to create the title.]

All appendices in the thesis must be presented together after the concluding chapter. Page numbering for appendices continues on from the body of the thesis, in sequence.

Appendices must be limited to supporting material genuinely subsidiary to the main argument of the thesis. Appendices must **only** include material that is referred to in the Body of the thesis.

The following are appropriate for inclusion in the appendices:

1. Additional details of methodology and/or data
2. Diagrams of specialized equipment developed
3. Copies of questionnaires or surveys used in the research. Please ensure that personal information (e.g., names or contact information of subjects and/or researchers) is removed and/or blacked out from your questionnaire/survey copies.

Sub-Appendix

This is Appendix A, Section 1.

[Sub-Appendix: use Heading 8]

Sub-Appendix

This is Appendix A, Section 2.

This is Appendix B.

Sub-Appendix

This is Appendix B, Section 1.