Fall 2022

Genomic Analysis Workflows to Identify Mutational Signatures and Structural Variations in OVCAR8 Cells and Rad51d-deficient Mouse Embryonic Fibroblasts

Manli Yang

Genomic Analysis Workflows to Identify Mutational Signatures and Structural Variations in OVCAR8 cells and Rad51d-deficient Mouse Embryonic Fibroblasts

By

Manli Yang

Master of Public Health
Xinxiang Medical University, 2019

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Science in

Pharmaceutical Science

College of Pharmacy

University of South Carolina

2022

Accepted by:

Michael D. Wyatt, Director of Thesis

Phillip J. Buckhaults, Reader

Eugenia Broude, Reader

Cheryl L. Addy, Interim Vice Provost and Dean of the Graduate School

DEDICATION

This work is dedicated to my mother, Yuqin Liu; father, Weiguo Yang; boyfriend, Dingyang Chen for their love that supported, guided and protected me through this journey.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my advisor, Dr. Michael D. Wyatt, for his limitless support and enthusiasm that were instrumental to my success. The experience and knowledge that I gained by working alongside him is invaluable to me, and I will always be grateful. I also thank Dr. Douglas L. Pittman for giving me the opportunity to join this lab and to work on the Rad51d-deficient MEFs project. I am also thankful to my committee member, Dr. Phillip J. Buckhaults, he gave me a lot of practical and useful guidance in genomic data analysis, which is very important to my project. I am especially grateful to my colleagues: Mr. Jacob Massey, Mr. Lucius Matthew for their technical assistance and friendship. I am also very appreciative of the support of other faculty members in the Department of Drug Discovery and Biomedical Sciences: Dr. Phillip J. Buckhaults for his assistance on the structural variation analysis of my OVCAR8 cell line project and for his patience to help a graduate student with no bioinformatics background to start genomic data analysis, Dr. Diego Altomare, for his collaborative efforts on the DNA extraction of my nanopore sequencing project. I would like to thank my family – Yuqin, Weiguo, Dezhong, Jirong – for their love, support, and patience as I worked to complete this goal, and to my fellow graduate students – Sana, Angie – for their support and friendship. Finally, to Dingyang Chen, for the love, support, accompany and snuggles that helped me going through tough situation.

ABSTRACT

Specific genomic profiles associated with exposure to DNA damaging agents have been identified in cancer related genes, revealing that mutational patterns can be carcinogen- specific. However, there have been limited efforts to demonstrate similar clear relationships for endogenous mutational processes. One endogenous source of mutations is contamination of the deoxynucleotide pool with damaged bases that, if incorporated into DNA, cause mutations.

The Catalog of Somatic Mutations In Cancer (COSMIC) represents an international effort to characterize mutations in cancer. A mutational signature is the pattern of mutations generated by a mutational process. Each mutational process contains two parts: DNA damage and DNA repair. If DNA repair is successful, the DNA will be restored to its native sequence. However, if damaged bases are not faithfully removed and replaced, then damaged bases can be misinterpreted by DNA polymerases to cause mutations. A mutational signature refers to mutations that occur within a defined adjoining sequence 5’ and 3’ to the base of altered sequence.

Nudt15 and Nudt18 are part of a nucleotide hydrolase superfamily of enzymes that may play diverse and independent roles in modulation of nucleotide pools. The founding member of the family, Nudt1 (hMTH1) hydrolyzes the mutagenic deoxynucleotide, 8-oxo-deoxyguanosine-triphosphate, to a deoxy

nucleoside monophosphate, thus preventing mutagenesis caused by this damaged dNTP precursor. However, little is known about the endogenous substrates of Nudt15 and Nudt18. This study was undertaken with two purposes. First, there is a need to develop long and short-read sequencing and mutational profile analysis workflows for COSMIC mutational signature and structural variation analysis. Second, the application of the workflow was applied to studying mutational signatures and structural variations in two genetic models, namely in Nudt15 and Nudt18 knockout ovarian cancer cells, and in Rad51d deficient murine embryonic fibroblasts. This study reports Illumina sequencing and mutational patterns analysis workflows to analyze mutational signatures and structural variations in ovarian cancer cells with reads shorter than 5000 base pairs. The results showed that the overall mutational patterns in ovarian cancer single-cell colonies include COSMIC signature 5. Structural variations (SVs) are large scale genomic changes, including insertions, deletions, duplications, inversions, and translocations. This study also reports Nanopore sequencing and structural variation analysis workflows to detect structural variations in Rad51d-deficient mouse embryonic fibroblasts (MEFs). Nanopore sequencing generated 652,000 base-called reads containing 4,404,372,528 bases by using 1 MinION flow cell sequencing. The results establish workflow pipelines for analysis of mutational signatures and for determining structural variations using long-read Nanopore sequencing.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

OVCAR ..... Ovarian Carcinoma Cells

COSMIC ..... Catalogue of Somatic Mutations in Cancer

NUDIX ..... Nucleoside Diphosphates Linked to Moiety-X

GATK ..... Genome Analysis Toolkit

VCF ..... Variant Call Format

CRISPR ..... Clustered Regularly Interspaced Short Palindromic Repeats

LTF ..... Late Treatment Failure

MDR ..... Multidrug Resistance

CHAPTER 1

INTRODUCTION

1.1 Mutational signatures

Mutations are changes in the DNA sequence of an organism. Mutational signature as an analytical principle was first proposed in 2012 (Nik-Zainal et al. 2012) (Zou et al. 2018). Accurate DNA repair pathways are essential for maintaining the integrity of the genome. If DNA repair is not successful mutations will be generated. The accumulation of non-random mutations in a process that produces a discoverable signature is now known to cause cancer. Mutations are generated in cells resulting from exposure to exogenous agents and from defects in endogenous processes, such as metabolic pathways like defective DNA repair. Different mutational processes generate specific combinations of mutation types, termed mutational signatures (Nik-Zainal et al. 2012). Mutational signatures have gained considerable attention in recent years and there are diverse algorithms to extract mutational signatures and although each algorithm has its own mathematical method, the results are similar. (Alexandrov et al. 2013) (Dees et al. 2012). Single base substitutions (SBS), also known as single nucleotide variants, are defined as a replacement of a certain nucleotide base. Considering the pyrimidines of the Watson-Crick base pairs, there are only six different possible substitutions: C>A, C>G, C>T, T>A, T>C, and T>G. These SBS classes can be further expanded considering the surrounding nucleotide context. Current SBS signatures have been identified using 96 different sequence contexts,

considering not only the mutated base, but also the one base immediately 5' and 3' neighboring the altered base. Mutational signatures have been detected in many different cancer types, revealing 49 different single base substitution signatures (Alexandrov et al. 2020), further classifying the signatures that may represent associated—but distinct— endogenous DNA damage causes and/or DNA repair pathway insufficiencies.

1.2 Nucleotide pool and NUDT15 (nudix hydrolase 15)

The biological origin of mutational signature is genomic instability and in this regard the nucleotide pool for DNA and RNA synthesis plays significant role in genomic instability (Bester et al. 2011). There are exogenous and endogenous factors that may affect the nucleotide pool. Reactive oxygen species (ROS) are generated during metabolic processes, nucleotide pool is at high risk of being oxidized by ROS, guanine is more susceptible to be oxidized by ROS comparing with other nucleobases. There are two paths by which ROS damage of guanine can cause mutations: 1) the most common ROS damage to DNA is the addition of oxygen to the C-8 carbon, thus generating 8-oxo-7,8-dihydroguanine (GO); 2) exposure of 2′-deoxyguanosine 5′-triphosphate (dGTP) to ROS, 8-oxo-dGTP will be generated after oxidation (Kasai et al. 1984) (Mo et al. 1992) (Kamiya et al. 1995) (Nakabeppu et al. 2007). It has been demonstrated that dGTP in the nucleotide pool is more susceptible to be oxidized than guanine in DNA (Kamiya et al. 1995) (Nakabeppu et al. 2017). In mammalian cells, the defense against 8-oxo-dGTP being incorporated into DNA is an enzyme called MTH1 that hydrolyzes 8-oxo-dGTP to 8-oxo-dGMP or 8-oxo-dGDP. MTH1 is also known as nudix hydrolase 1 (Nudt1), nucleoside diphosphate-linked moiety X-type

(NUDIX) (Nishii et al. 2021). The nucleoside diphosphates linked to moiety-X (NUDIX) are a superfamily of hydrolytic enzymes, and their function is to hydrolyze phosphorylated nucleosides (Nishii et al. 2021). If NUDT1 activity is disrupted, single base substitutions resulting from 8-oxo-dGTP incorporation into DNA may be generated. After two rounds of replication, the incorporation of GO increases the occurrence of A:T to C:G or G:C to T:A transversions (Figure 1.1).

Besides NUDT1, another member of the NUDIX family is NUDT15, which has gained considerable attention in recent years. It has been demonstrated that NUDT15 plays a significant role in thiopurine metabolism. Thiopurines are anticancer drugs for childhood acute lymphoblastic leukemia and also used as immunosuppressants (Wyatt et al. 2018). Thiopurines including azathioprine (AZA), 6-mercaptopurine (6-MP), and thioguanine (TG), are prodrugs that must be metabolized to 6-thioguanosine triphosphate (d6-TGTP) (Figure 1.2). The incorporation of d6-TGTP into DNA may cause cell death or mutations in survivors (Cara et al. 2004), for example SBS 87 (Li et al. 2020). There is concern about thiopurine-induced secondary cancers. To better understand thiopurine-induced secondary cancers, researchers have focused on thiopurine metabolism, for example, it is known that low level of thiopurine methyltransferase (TPMT) activity has been proposed as a significant risk factor for thiopurine-induced toxicity such as myelosuppression. In addition to TMPT, NUDT15 hydrolyzes d6-TGTP to d6-TGMP and it has been demonstrated that NUDT15 deficiency is highly associated with thiopurine intolerance, especially in Asian populations (Tanaka et al. 2021). Previous study also demonstrated that NUDT18 prevents 8-oxoguanine to be incorporated into DNA and RNA by degrading 8-

oxoguanine-containing nucleoside diphosphates (Takagi et al. 2012). Therefore, studies about NUDT15 and NUDT18 are needed to investigate endogenous pathways and novel endogenous substrate of NUDT15 and NUDT18.

1.3 Homologous Recombination (HR) repair deficiency and mutational signatures.

DNA damage, for example double-strand breaks, can induce HR when a double-stranded copy of the sequence is available. HR plays an important role in replication, repair, and other processes. HR processes are described stepwise in the following manner (Figure 1.3): 1) a DSB is converted to a recombination substrate by degradation of a single-strand end, leaving a 3' single-stranded overhang typically several hundred base pairs in length; 2) the Rad51 nucleoprotein filament loads onto the recombination substrate, which promotes strand invasion and DNA repair synthesis; 3) Holliday junction resolution. HR is an error-free DNA repair process. However, if cells are HR-deficient, cells will undergo non-homologous end joining (NHEJ) to repair double-strand breaks. NHEJ is an error-prone repair process, which will lead to mutations, chromosome rearrangements in the genome, even causing cell death (Wyatt and Pittman, 2006) (Christine S. Walsh, 2015).

BRCA1 and BRCA2 are tumor suppressor genes. Mutations in BRCA1/2 are responsible for approximately 40% of inherited breast cancers and more than 80% of inherited breast and ovarian cancers (Mehrgou, A., & Akouchekian, M. 2016). Normally, BRCA1/2 are involved in double-strand breaks repair via error-free homologous recombination. During HR repair, BRCA1 is involved in the resection of recombination overhang, and BRCA2 controls the process of loading RAD51 onto

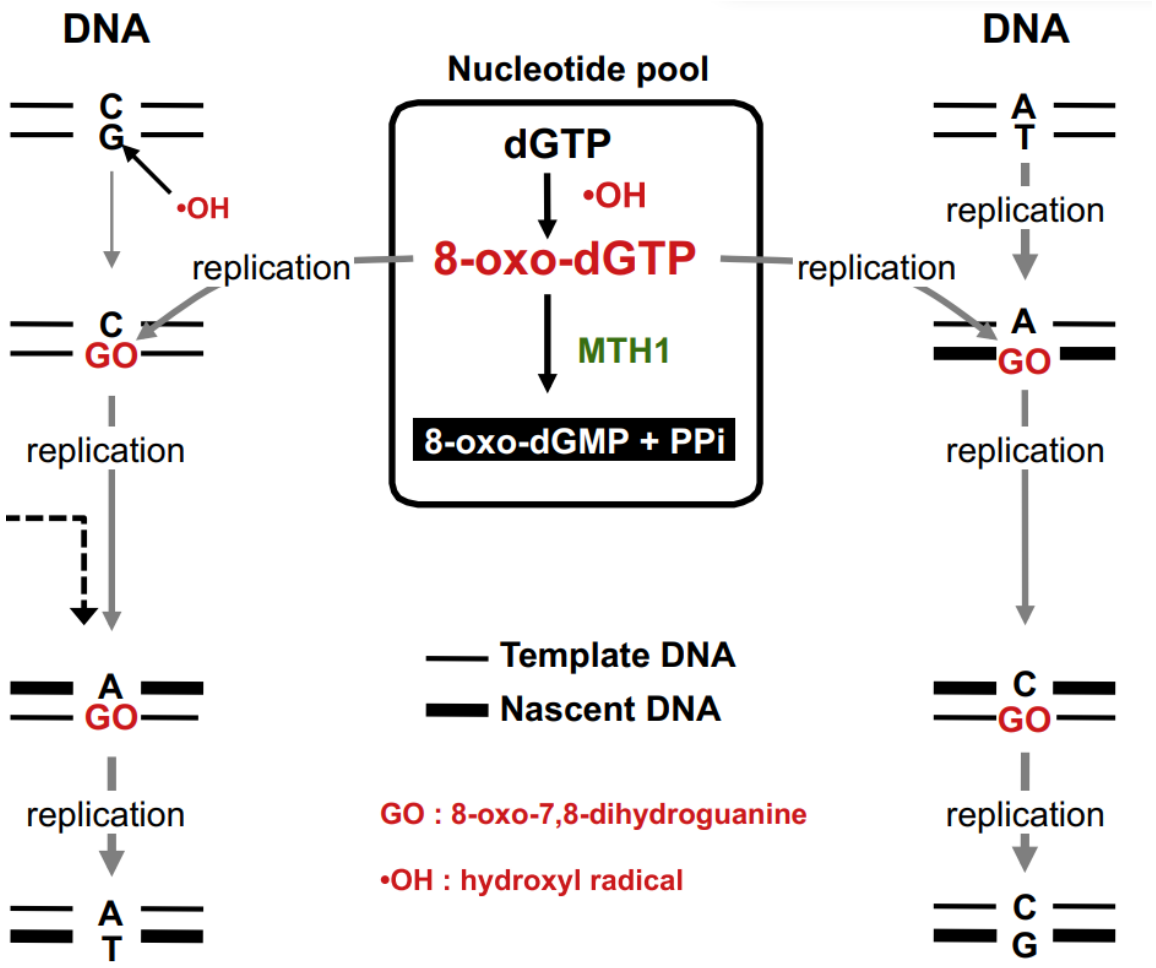
single-strand DNA. Moreover, during S phase, BRCA1, BRCA2, and Rad51 proteins colocalize to nuclear foci. In HR defective cells, the colocalization of Rad51 and BRCA2 into nuclear foci is abrogated (Prakash et al, 2015). In normal cells, DNA damaging agents cause DNA damage, then error-free DNA repair process for example HR repair correct the damage and suppress mutations. However, in HR-deficient cells, there is no functional HR, instead NHEJ acts as a backup repair solution for DNA DSBs. As mentioned earlier, NHEJ is an error-prone repair process, therefore a mutational signature will be generated. Nik-Zainal et al have demonstrated that single base substitution (SBS) signature 3 results from BRCA1 or BRCA2 deficiency. They also demonstrated that SBS3 is associated with BRCA1 promoter methylation in breast, pancreatic, and ovarian cancers. In pancreatic cancer, patients who are sensitive to platinum therapy usually exhibit SBS3 mutations (Golan et al. 2021). Together with associated small insertions and deletions, structural variations, SBS3 has been proposed as an indicator of homologous recombination-deficiency based repair, probably NHEJ. Besides this single base substitution signature, homologous recombination-deficiency based DNA damage repair manifests predominantly as small indels and structural variations due to abnormal DNA double-strand break repair (Nik-Zainal et al. 2012 Cell). As we know that both BRCA1/2 and Rad51d are HR-related genes, it is therefore of interest to determine if RAD51D deficiency produces a mutation signature that is similar to or distinct from BRCA1/2 deficiency.

1.4 Structural Variations

In recent years, rearrangement signatures have been proposed and studied, applying in the subtype's categorization of breast cancer (Morganella et al. 2016) (Nik-

Zainal et al. 2016) and the clinical applications of mutational signatures have been started (Davies et al. 2017). Structural variations (SVs) are large scale genomic changes, including insertions, deletions, duplications, inversions, and translocations. SVs are a major source of genome variation and contributes substantially to cancers and other diseases. Compared with single nucleotide polymorphisms (SNPs), large SVs are responsible for ten times more variant bases in the whole genome and are thirty times more likely to affect gene expression (Coster et al. 2019). Therefore, SV analysis will benefit the study of disease etiology on genome level. However, SVs are hard to detect, until recently, the large size and complicated genomic contexts of SVs have limited their study. Traditional short-read sequencing generates reads of 75–300 bp, whereas SVs range from 50 bp to several mega bases, meaning that SVs cannot typically be captured within a short-read. Therefore, long-read methods, which can read from several thousand to millions of bases, are important for structural variation analysis. Jain et al. report Nanopore sequencing and assembly of ultra-long reads from the human GM12878 Utah/Ceph cell line to the reference genome using the MinION (Oxford Nanopore Technologies) nanopore sequencer and Canu software. Nanopore sequencing generated 14,183,584 base-called reads containing 91,240,120,433 bases with a read N50 (the read length such that reads of this length or greater sum to at least half the total bases) of 10,589 bp by using 39 MinION flow cells. The results showed that Nanopore long-reads benefit structural variation detection but show relatively low-accuracy in single base substitution detection because nanopore sequencing has relatively low-coverage and -accuracy. However, ultra-long reads improve phasing and assembly contiguity and close gaps in the human reference genome. This is because

ultra-long reads can increase assembly continuity significantly; with their ability to resolve complicated SVs, ultra-long reads could be assembled and phased. The assembly of the 4-Mb major histocompatibility complex (MHC) locus in its entirety was achieved with this method, which also allowed prediction of telomere length (Jain et al. 2018). What's more, Nanopore sequencing can detect methylation in the genome with high accuracy. These data suggest that ultra-long read nanopore sequencing enables the analysis of regions of the human genome that were previously intractable.



(Yusaku et al, 2017)

Figure 1.1: MTH1 (Nudt 1): Modulation of nucleotide pools. Solid gray lines: mutagenic pathway. MTH1 hydrolyzes 8-oxo-dGTP to 8-oxo-dGMP act as defense metabolic pathway, thus preventing their incorporation into DNA.

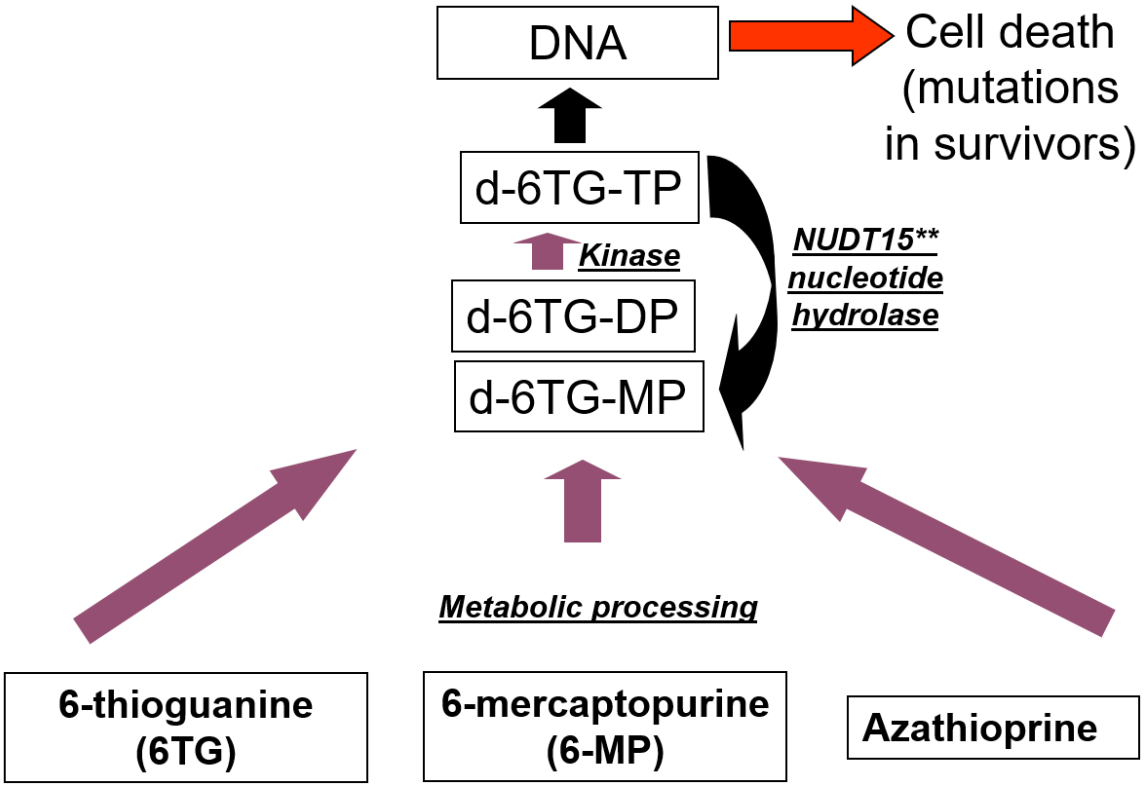
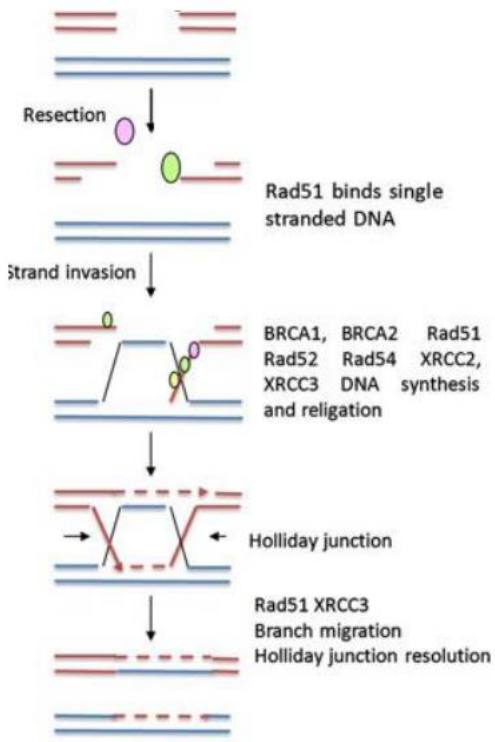


Figure 1.2: The role of NUDT15 in thiopurine metabolism. Thiopurines including azathioprine (AZA), 6-mercaptopurine (6-MP), and thioguanine (TG), are prodrugs that must be metabolized to 6-thio-deoxyguanosine triphosphate (d6-TGTP). The incorporation of d6-TGTP into DNA may cause cell death or mutations in survivors, for example SBS 87. NUDT15 hydrolyzes d6-TGTP to d6-TGMP, which prevents the toxic overaccumulation of 6-TG into the genome. Therefore, Nudt15 plays important role in thiopurine metabolism.



(Cerbinskaite et al. 2012)

Figure 1.3: Homologous recombination.

CHAPTER 2

EXTRACTED SINGLE BASE SUBSTITUTION (SBS) SIGNATURE FROM NUDT15 KNOCKOUT AND NUDT18 KNOCKOUT OVARIAN CANCER CELLS

2.1 Introduction

NUDT15 has nucleotide hydrolase activity that can prevent the incorporation of potentially mutagenic bases into DNA. Yet, the only known substrate established thus far for NUDT15, d6-TGTP, results from an exogenous treatment, thiopurines. The naturally occurring substrate for NUDT15 remains unknown. Another NUDIX family member, NUDT18, is even less characterized regarding its activities and functions. To determine whether NUDT15 deficiency results in increased mutagenesis, we explored determining mutational signatures in cells with defined knockouts of NUDT15. Another graduate student, Jacob Massey, generated and characterized the NUDT15 and NUDT18 knockout ovarian cancer cells (OVCAR8), which is described in his thesis. Here, I utilize the NUDT15 and NUDT18 knockout cells to determine whether deficiency in these nucleotide hydrolases causes mutational signatures. This is also related to thiopurine treatments discussed in Chapter 1, because COSMIC mutational signature SBS 87 in humans is thought to be caused by thiopurine exposure (Figure 2.1). The goal of this chapter is to determine mutations and mutational signatures in the OVCAR8 cancer cells in the absence of NUDT15 and NUDT18.

2.2 Methods

In our lab, we have Nudt15 and Nudt18 knockout ovarian cancer cells – OVCAR8 cells. OVCAR8 single-cell colonies corresponding to three genotypes were propagated and DNA isolated. S51 and S52 are parental OVCAR8 cells incubated with Cas9 but without a gRNA; S53 and S54 are OVCAR8 Cas9 cells with Nudt15 KO gRNA; S55 and S56 are OVCAR8 cells with Nudt18 KO gRNA. After cells were grown to confluence, a Qiagen kit (DNeasy Blood & Tissue Kits) to extract DNA products for Illumina sequencing (Figure 2.2).

The bioinformatics pipeline is as follows. Illumina sequencing generated fastq files; next, BWA (Burrows-Wheeler Aligner) was used to perform alignment and generate a sam file; the picard program takes a sam (Sequence Alignment Map) file as input and converts it to a bam file; picard was also used to build a bam index (bai file). The GATK program was used to call variants and generate vcf files; MutationalPatterns is a mutational signature analysis software that takes a vcf file as input to analyze single base substitution signatures, double base substitution signatures and indel signatures (Figure 2.3). This study chiefly focused on single base substitution signatures.

2.3 Results and Discussion

As mentioned earlier, there are six different possible substitutions: C>A, C>G, C>T, T>A, T>C, and T>G, considering not only the mutated base, but also the base immediately 5’ and 3’ to the central base, which produces 96 different

mutational contexts. To extract mutational signatures, a 96-trinucleotide mutation count matrix was made. Table 2.1 shows as an example part of the mutation count matrix for S51. This table shows the substitutions type and the trinucleotide contexts. Next, this matrix was used to plot the 96 mutational profiles of each sample. Figure 2.4 shows the 96 mutational profiles also known as mutational signatures extracted from S51 and S52, S53 and S54, S55 and S56, separately. Comparing mutational patterns of wild-type ovarian cancer cells with mutational patterns of Nudt15 knockout and Nudt18 knockout ovarian cancer cells, the results show that there are no specific mutational patterns related to Nudt15 and Nudt18 knockout in OVCAR8 cells.

To estimate the similarity between the mutational signatures extracted from our samples with the established COSMIC mutational signatures, we calculate the cosine similarity. High cosine similarity (>0.85) means two mutational signatures are similar. As we can see in Figure 2.5, mutational signatures of OVCAR8 samples have high cosine similarities with the COSMIC signature 5, which are around 0.9.

Considering that the total number of mutations in single-cell colonies can be decomposed into background mutations (including mutation patterns generated from cell culture processes, mutational patterns of cancer cell line) and mutations caused by Nudt15/18 knockout.

$$N_{subclone\text{-}Nudt15/18\ knockout} = N_{background} + N_{Nudt15/18\ knockout}$$

To determine whether we could detect mutational signatures that result from Nudt15 or Nudt18 knockout, for the next step, we will use wild-type ovarian cancer

single-cell colonies to act as control. The difference between the total number of mutations in NUDT15/18 knockout groups will be measured and compared to the number of mutations in wild-type groups (Jill E. Kucab, 2019).

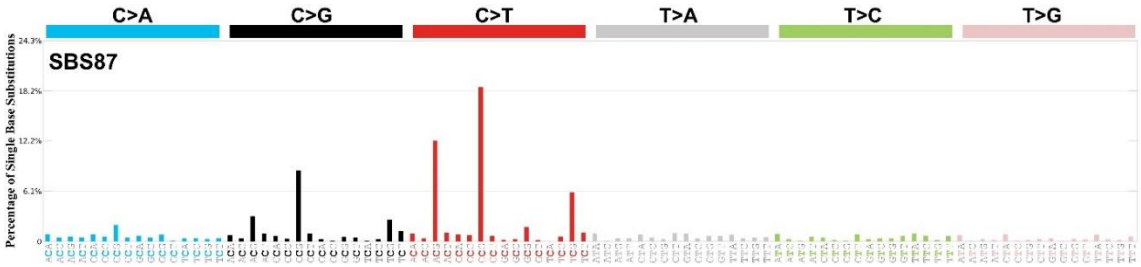


Figure 2.1: Single Base Substitution (SBS) signature 87. There are six different possible substitutions: C>A, C>G, C>T, T>A, T>C, T>G. These SBS classes can be further expanded considering the nucleotide context. SBS signatures were identified using 96 different contexts, considering not only the mutated base, but also the one base immediately neighboring 5' and 3' to the central nucleotide. SBS 87 is shown as an example. The etiology of SBS87 is associated with thiopurine chemotherapy treatment.

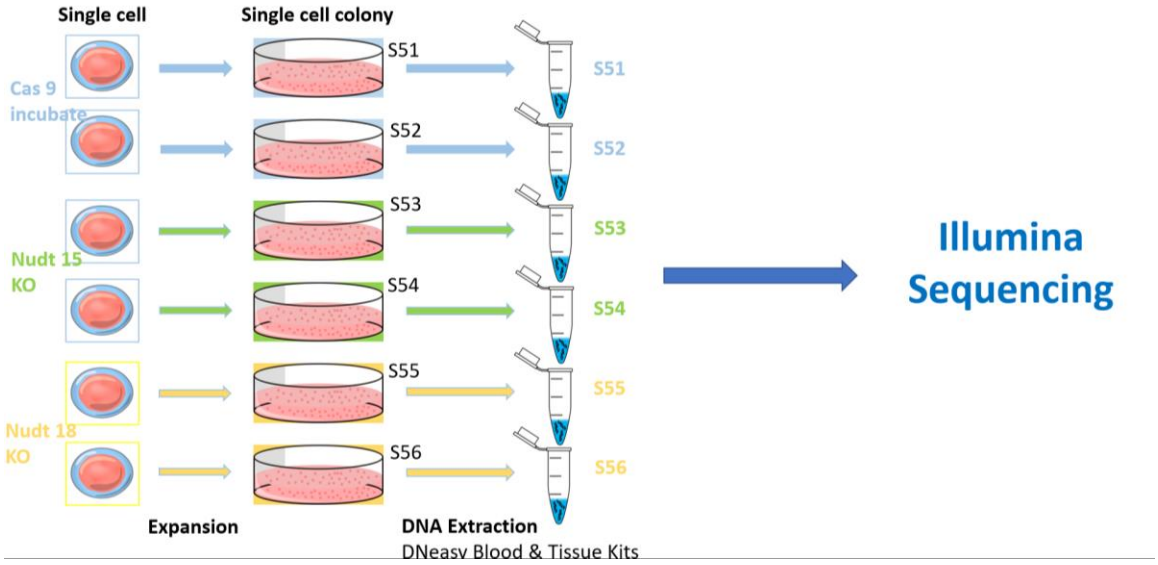


Figure 2.2: Workflow showing colony selection of 6 single-cell colonies of ovarian cancer cells. S51 and 52 are wild-type OVCAR8 cells incubated with Cas9; S53 and S54 are OVCAR8 cells with Nudt15 deleted by Nudt15 gRNA; S55 and S56 are OVCAR8 cell lines with Nudt18 deleted by Nudt18 gRNA. After cells were grown to confluence, a Qiagen kit was used to extract DNA products, then Illumina sequencing was performed.

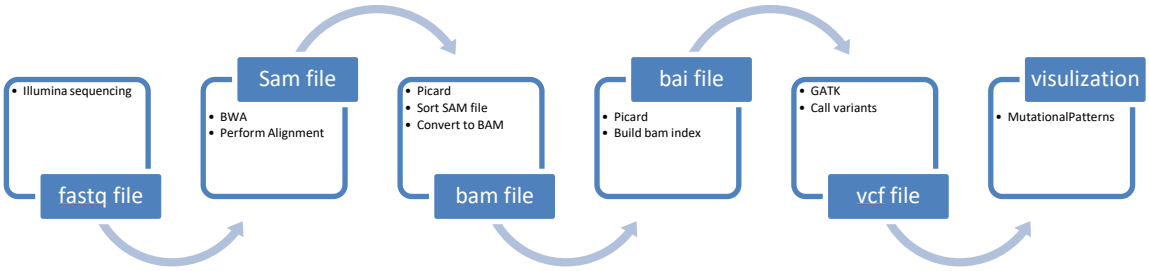
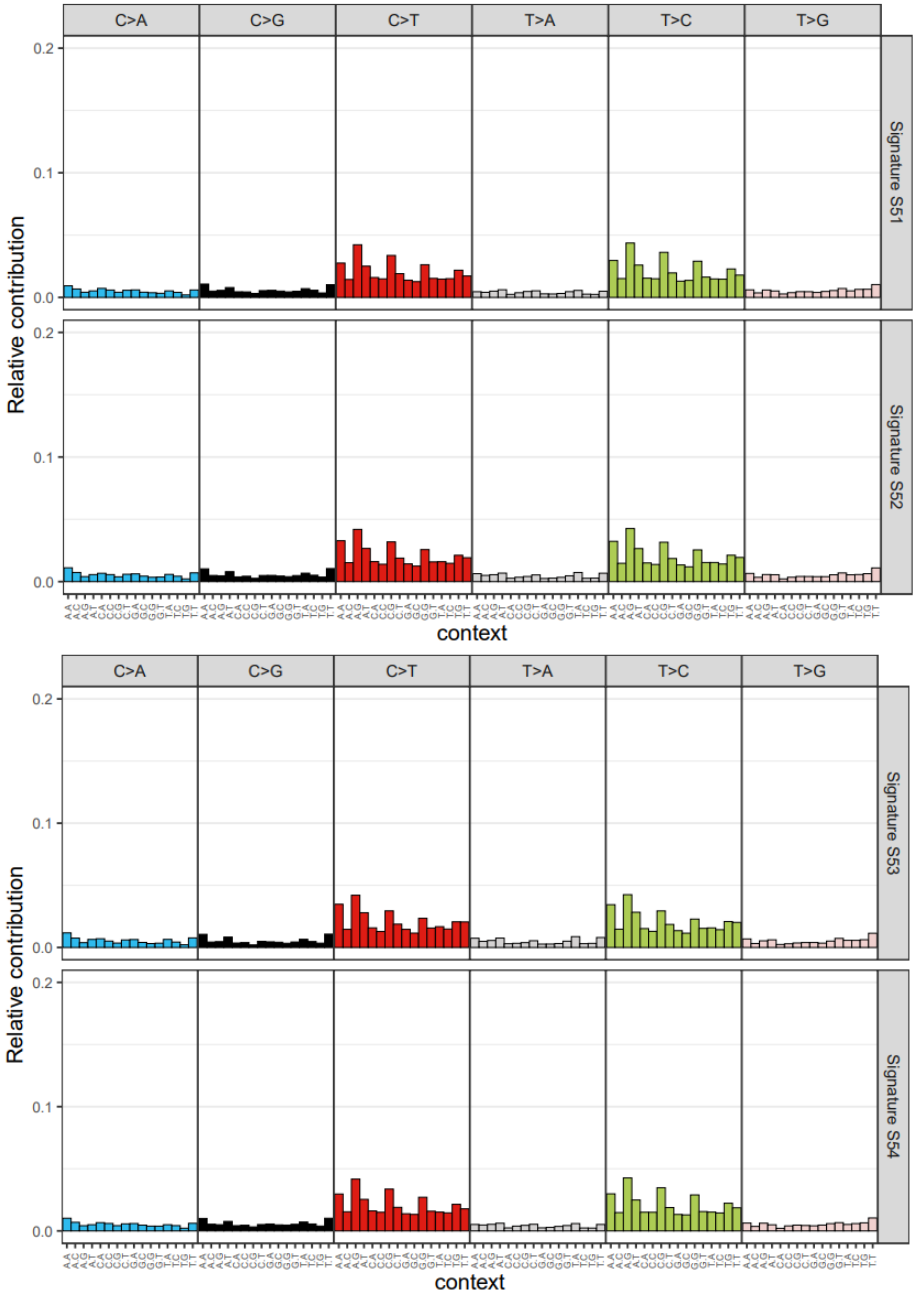


Figure 2.3: Mutational signature analysis pipeline. Illumina sequencing generates a fastq file; next, alignment using BWA generates a sam file; picard takes a sam file as input and converts it to a bam file; picard also builds a bam index (bai file); gatk is used to to call variants and generate a vcf file; MutationalPatterns is a mutational signature analysis software that takes a vcf file as input to analyze single base substitution signatures, double base substitution signatures and indel signatures.

Table 2.1 Substitution type and the trinucleotide contexts table of S51

```
> print(cancer_signatures)
        Substitution.Type Trinucleotide
A[C>A]A               C>A           ACA
A[C>A]C               C>A           ACC
A[C>A]G               C>A           ACG
A[C>A]T               C>A           ACT
C[C>A]A               C>A           CCA
C[C>A]C               C>A           CCC
C[C>A]G               C>A           CCG
C[C>A]T               C>A           CCT
G[C>A]A               C>A           GCA
G[C>A]C               C>A           GCC
G[C>A]G               C>A           GCG
G[C>A]T               C>A           GCT
T[C>A]A               C>A           TCA
T[C>A]C               C>A           TCC
T[C>A]G               C>A           TCG
T[C>A]T               C>A           TCT
A[C>G]A               C>G           ACA
A[C>G]C               C>G           ACC
A[C>G]G               C>G           ACG
A[C>G]T               C>G           ACT
C[C>G]A               C>G           CCA
C[C>G]C               C>G           CCC
C[C>G]G               C>G           CCG
C[C>G]T               C>G           CCT
G[C>G]A               C>G           GCA
G[C>G]C               C>G           GCC
G[C>G]G               C>G           GCG
G[C>G]T               C>G           GCT
T[C>G]A               C>G           TCA
T[C>G]C               C>G           TCC
T[C>G]G               C>G           TCG
T[C>G]T               C>G           TCT
```

Table 2.1: This table shows the substitution type and the trinucleotide contexts.


C>A
C>G
C>T
T>A
T>C
T>G
Signature S51
Signature S52
Signature S53
Signature S54
Relative contribution
0.0
0.1
0.2
context

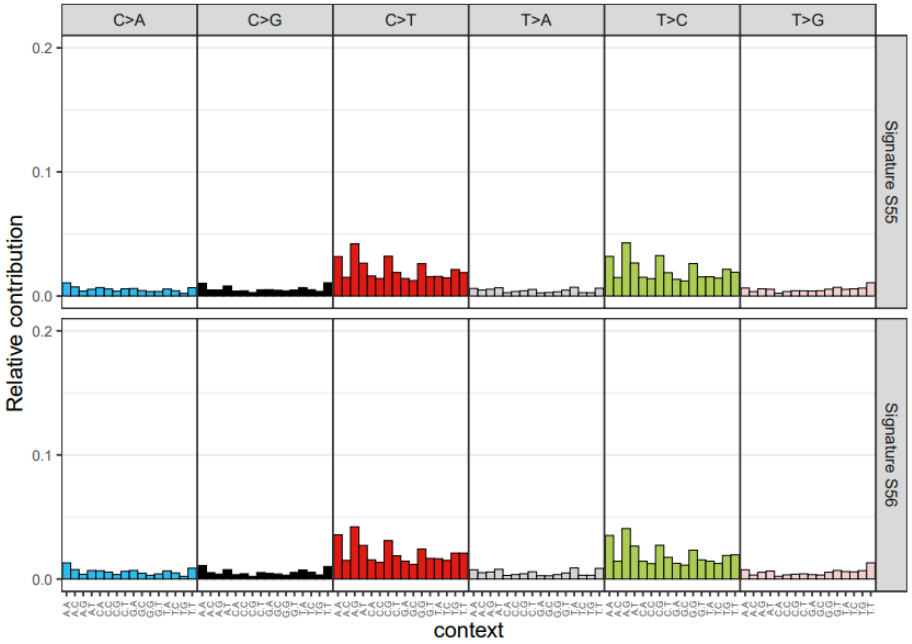


Figure 2.4: Mutational signatures in Nudt15 knockout and Nudt18 knockout OVCAR8 single-cell colonies. The X axis represents 96 trinucleotide contexts, Y axis represents the relative contribution of each trinucleotide context to the mutational burden in that sample.

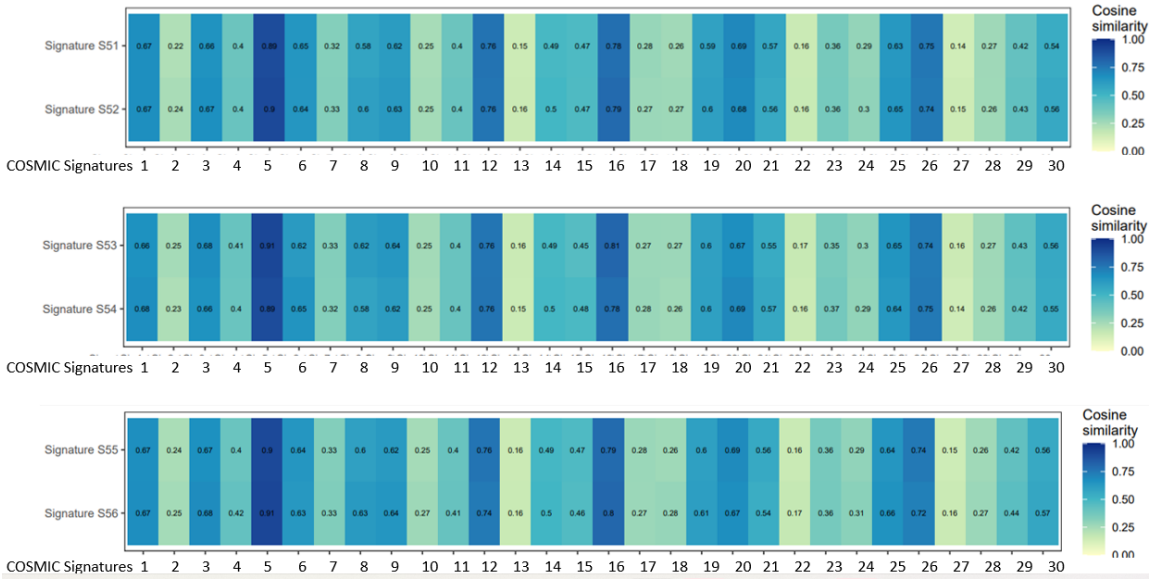


Figure 2.5: Cosine Similarity. Mutational signatures extracted from OVCAR8 samples have high cosine similarities of ~0.9 with the COSMIC signature 5.

CHAPTER 3

STRUCTURAL VARIATION ANALYSIS OF NUDT15 KNOCKOUT AND NUDT18 KNOCKOUT OVARIAN CANCER CELLS

3.1 Introduction

The goal of the work in this chapter is to analyze structural variations in ovarian cancer cells to investigate whether CRISPR Cas 9 cause specific structural variations in NUDT15 and NUDT18 knockout ovarian cancer cells. There are several questions to be answered. Firstly, what are the structural variations? Secondly, how are structural variations called? Structural variation analysis workflows using Illumina and Nanopore sequencing will be presented separately. The structural variation pipeline was used to analyze six OVCAR8 colonies using Illumina sequencing data, because we want to analyze whether CRISPR Cas 9 caused specific structural variations in the parental OVCAR8 cancer cells. Structural variations are large scale genomic changes, including insertions, deletions, duplications, inversions, and translocations (Figure 3.1) (Geòrgia Escaramís et al. 2015). SVs are a major source of genomic variation in the human genome and are responsible for ten times more variant bases than single nucleotide polymorphisms (SNPs) and are thirty times more likely to affect gene expression (Korbel et al. 2007).

3.2 Methods

Illumina sequencing produced fastq files; next, alignment was achieved using BWA to generate sam files; picard can take sam file as input and convert it to bam file; then we can use gatk to call variants and generate vcf file; QIAGEN CLC platform is a powerful genomic data analysis platform, structural variations caller tool is included in CLC platform, we used this tool to analysis structural variations with reads shorter than 5000 bp. This tool can take a vcf file as input to analyze many kinds of structural variations, including insertions, deletions, tandem duplications, and inversions by chromosomes (Figure 3.2).

3.3 Results

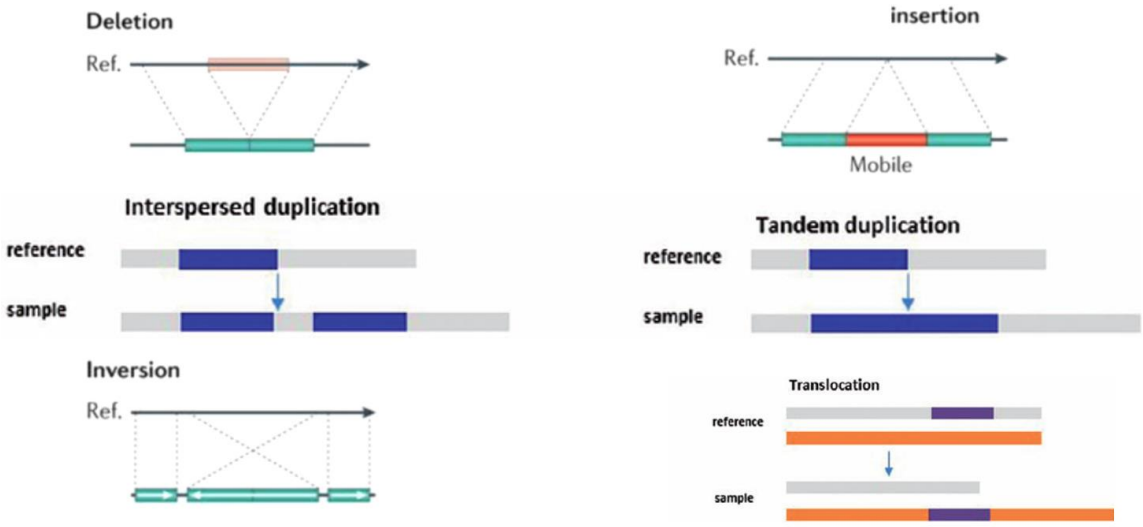
There are examples of insertion (A), deletion (B) and inversion (C) extracted from ovarian cancer cells (Figure 3.3). Figure 3.4 shows the structural variation analysis results counted by chromosome in the S51 single cell colony. The X axis represents chromosome number; Y axis represents the number of structural variations; different colors represent different kinds of structural variations, blue represents total number of structural variations, orange represents insertions, gray represents deletions, yellow represents tandem duplications, and light blue represents inversions. Comparing different types of structural variations in six single-cell colonies, we found that inversions and deletions are the two main types of structural variations in Cas 9-parental, Nudt15 knockout and Nudt18 knockout OVCAR8 cells.

Next, the number of structural variations by each type in every OVCAR8 single-cell colonies were calculated. It was found that the number of inversions in Nudt15 knockout and Nudt18 knockout groups are much higher than that in the Cas

9 parental control group (Figure 3.5).

1.4 Discussion

The results showed that the number of inversions in Nudt15 knockout and Nudt18 knockout groups are much higher than that in the Cas 9 incubated wild-type groups. For the next step, we will compare the sequences of the primers used to knockout NUDT15 and NUDT18 gene in OVCAR8 cells with the sequences at the ends of the inversions.



(Geòrgia Escaramís et al. 2015)

Figure 3.1: Different types of structural variations, such as deletions (top left), insertions (top right), interspersed and tandem duplications (middle), and inversions and translocations (bottom).

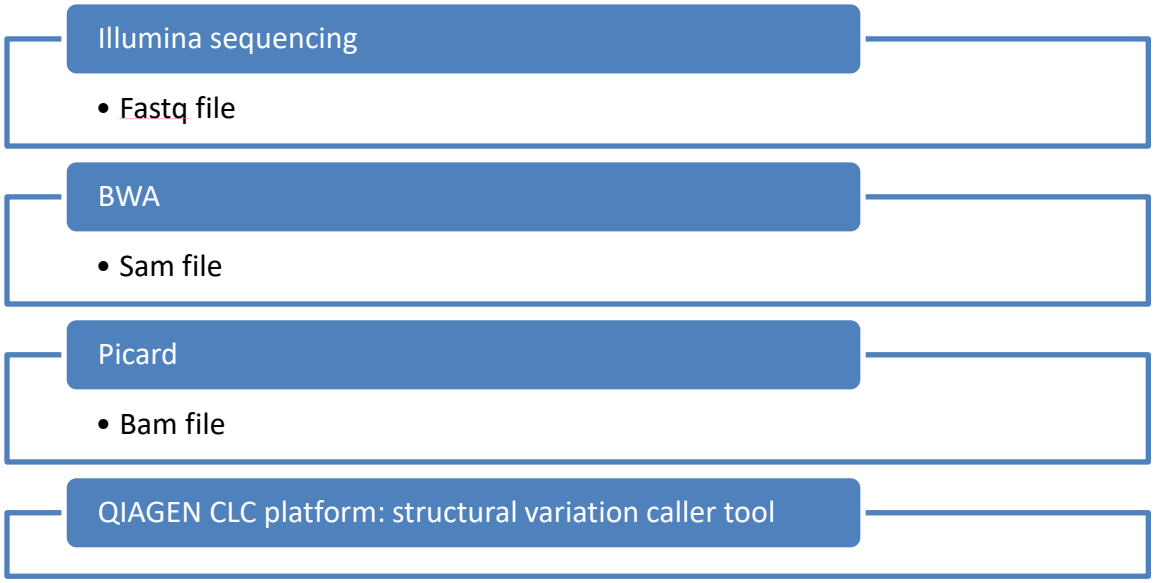
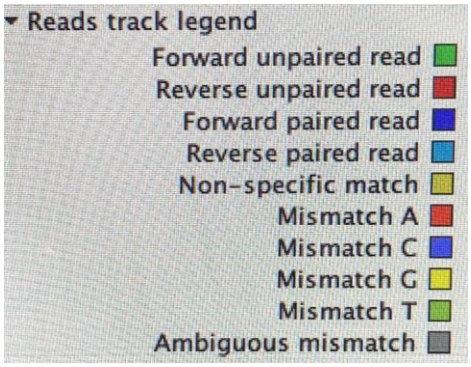
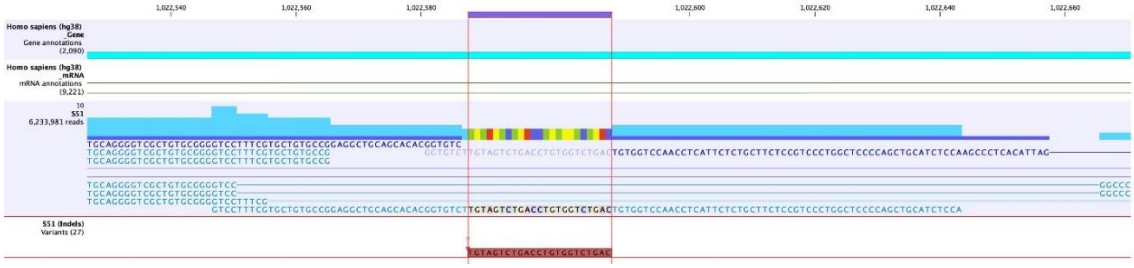


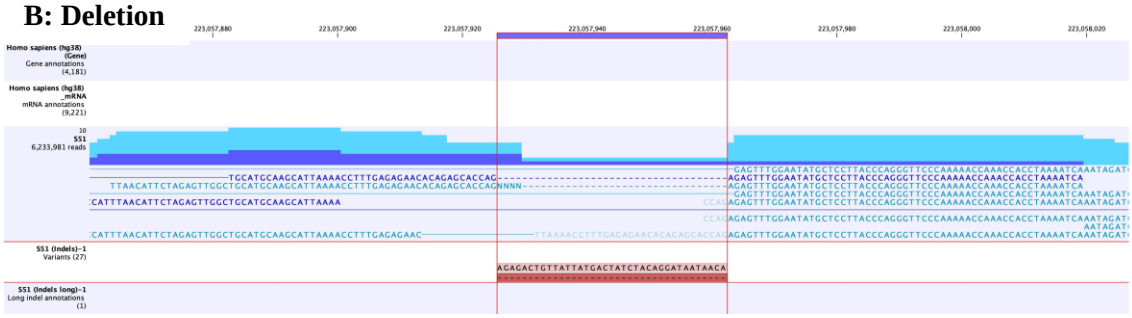
Figure 3.2: Structural variation analysis pipeline of NUDT15 and NUDT18 knockout OVCAR8 cells.



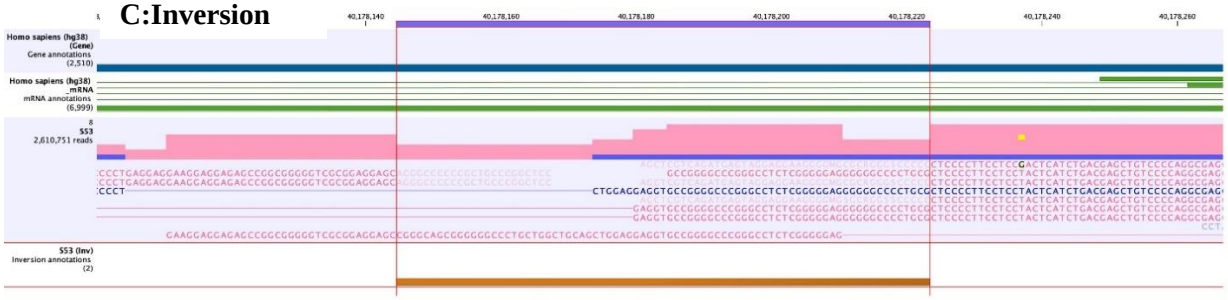
A: Insertion



B: Deletion



C:Inversion



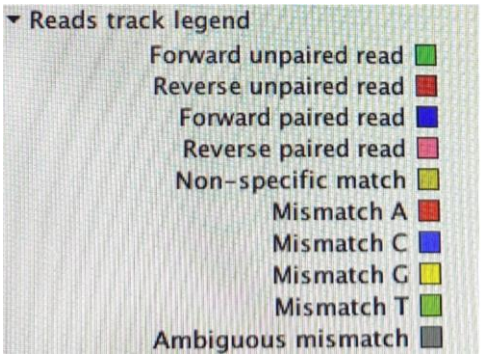


Figure 3.3: Insertion, deletion and inversion examples in ovarian cancer cells detected by structural variation analysis tool.

S51	chromosome	1	2	3	4	5	6	7	8	9	10	11
Total # variants		23	20	16	15	39	32	20	18	13	30	13
Insertion		13	12	7	8	13	14	14	6	8	18	8
Deletion		8	7	4	4	18	9	6	12	2	11	4
Tandem Duplication		2	1	5	3	8	9	0	0	2	1	1
Inversion		0	0	0	0	0	0	0	0	1	0	0

S51	chromosome	12	13	14	15	16	17	18	19	20	21	22	X
Total # variants		11	10	9	9	9	28	2	16	24	19	7	2
Insertion		5	5	3	6	5	16	1	8	16	6	3	1
Deletion		5	5	3	3	4	7	0	6	6	7	3	1
Tandem Duplication		1	0	3	0	0	5	1	2	2	6	1	0
Inversion		0	0	0	0	0	0	0	0	0	0	0	0

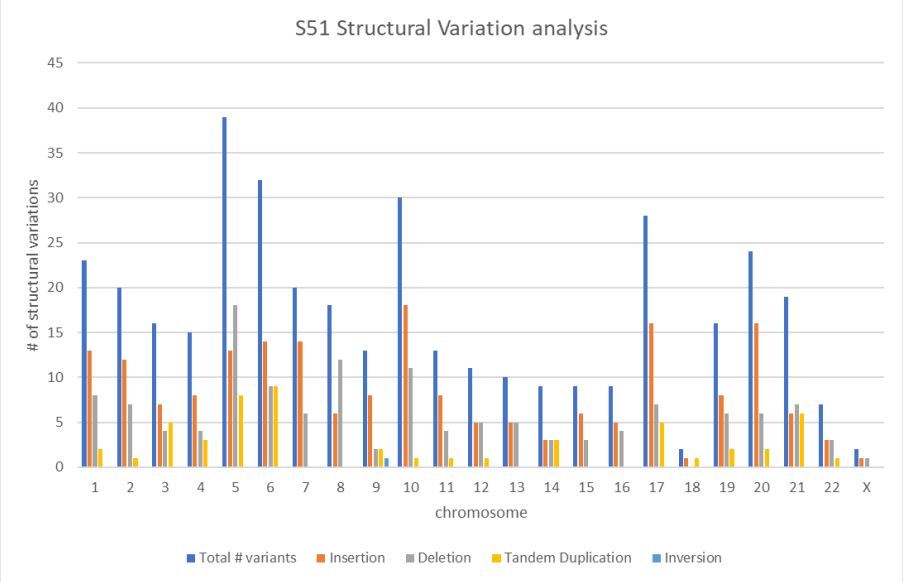


Figure 3.4: Structural variation results of the S51 single-cell colony. X axis represents chromosome number; Y axis represents the number of structural variations.

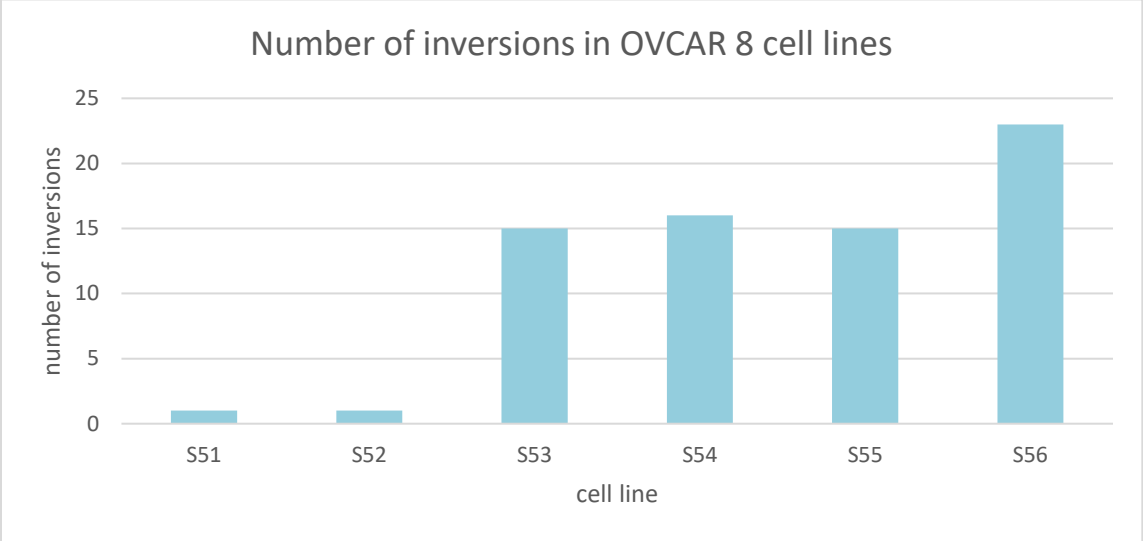


Figure 3.5: the number of inversions in 6 single-cell colonies of OVCAR8 cells. S51, S52: wild-type OVCAR8 cells; S53, S54: NUDT15 knockout OVCAR8 cells; S55, S56: NUDT18 knockout OVCAR8 cells.

CHAPTER 4

STRUCTURAL VARIATION ANALYSIS OF RAD51D-DEFICIENT MOUSE EMBRYONIC FIBROBLASTS

4.1 Introduction

In Chapter 3, we introduced structural variation analysis pipeline for Illumina short-read sequencing. In this chapter, structural variation analysis pipeline for nanopore ultra-long read sequencing is discussed. Why do we need nanopore ultralong read sequencing? Because traditional short-read sequencing generates reads of 75–300 bp (Figure 4.1), large structural variations range from 50 bp to several megabases, meaning that Illumina sequencing is good at detecting small structural variations but nanopore long read sequencing is good at resolving big and complex SVs (Kosugi et al. 2019). For example, Figure 4.1 shows a 1054 bases insertion. A recent study published by the Telomere-to-Telomere (T2T) Consortium has reported completing a challenging 8% of the unresolved human genome using Pacbio HiFi and Oxford Nanopore sequencing with ultra-long reads (Zahn et al. 2022). Although the initial Human Genome Project was declared “complete” for more than 20 years, technology limitations prevented some regions of the human genome from being resolved. This most recent publication has reported the success of long-read sequencing methods to complete these challenging regions.

Previous work has demonstrated that Rad51d-deficient MEFs have chromosome instability (Figure 4.2) and are extremely sensitive to DNA damaging

agents (Figure 4.3). The source of this chromosomal instability resulting from HR deficiency remains unknown and under-investigated. Therefore, I established a pipeline for long-read structural variations analysis, and then used this pipeline to detect structural variations in Rad51d-deficient MEFs, and MEFs treated with DNA damaging agents.

4.2 Methods

Firstly, colony selection prepared 2-3 colonies for each genotype of MEFs. two colonies for Rad51d-proficient MEFs (C53), two colonies for Rad51d-deficient MEFs (258), two colonies for Rad51d-deficient MEFs (310) (Figure 4.4). Structural variation analysis pipeline for nanopore sequencing with reads longer than 5000 bp. Here is the basic workflow: first, nanopore sequencing generates a fast5 file; then guppy performs base calling and generates a fastq file. Next, the SV aligner – minimap2 aligns nanopore long reads to a reference genome; Then an SV caller program called sniffles was used to detect structural variations across the whole genome. The key steps here are align reads and call variants (Figure 4.5).

4.3 Results

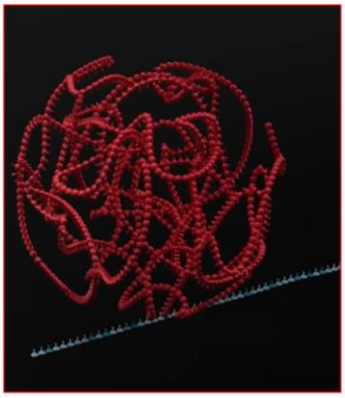
After nanopore sequencing, we checked sequencing and mapping statistics, which gave us a quality estimation of the experiment and suggestions for preparing future experiments. Our preliminary experiment design used one flowcell to sequence one colony of Rad51d-deficient MEFs for 96 hours. In Table 4.1, we can see that the average coverage is 1.

Minimap2 was used to perform mapping and the mapping results showed that 86.97% of reads were mapped, and 93.92% of bases were mapped (Table 4.2). In this figure, we can see that the most of the read's lengths were between 0 – 46000 bp. Reads lengths are more concentrated within a range of 0 – 4000 bp (Figure 4.6).

4.4 Discussion

After sequencing and alignment, we can use CLC platform tools to call structural variations in reads shorter than 5000 bp, and at the same time, using sniffles to call structural variations in reads longer than 5000 bp (Figure 4.7). By using 258 single colony MEFs, we finish this pipeline, considering the coverage is low, we will post these data later after we add more coverages to it.

Traditional short-read sequencing generates reads of 75–300 bp



This picture shows a 1054 -base insertion

https://www.youtube.com/watch?v=N8KPwWdQTAw&ab_channel=OxfordNanoporeTechnologies

Figure 4.1: The figure shows the visualization of a large (1054 bp) insertion (red, right figure).

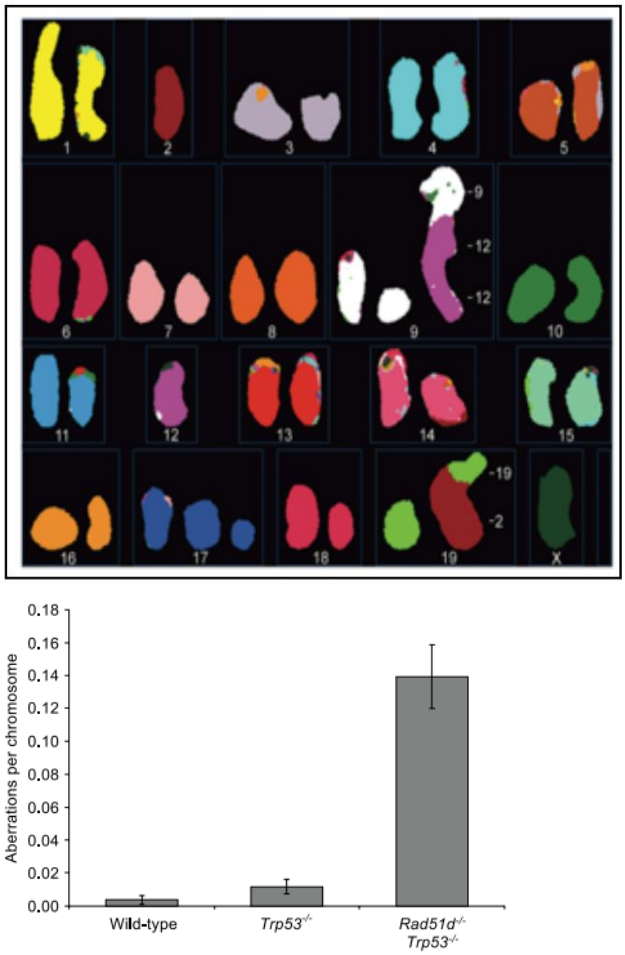


Figure 4.2: genome instability of Rad51d-deficient MEFs. Figure is from (Smiraldo et al. 2005).

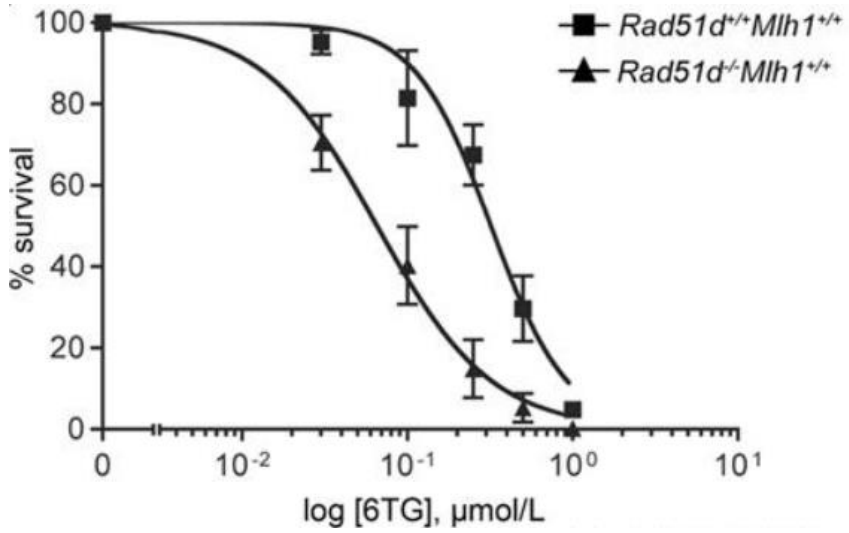


Figure 4.3: Rad51d-deficient MEFs are sensitive to thiopurine treatment. Figure is from (Rajesh et al. 2011).

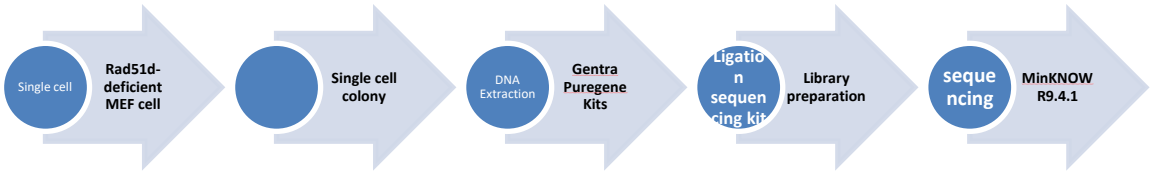


Figure 4.4: Workflow for long-read sequencing of DNA from RAD51-proficient and deficient cells.

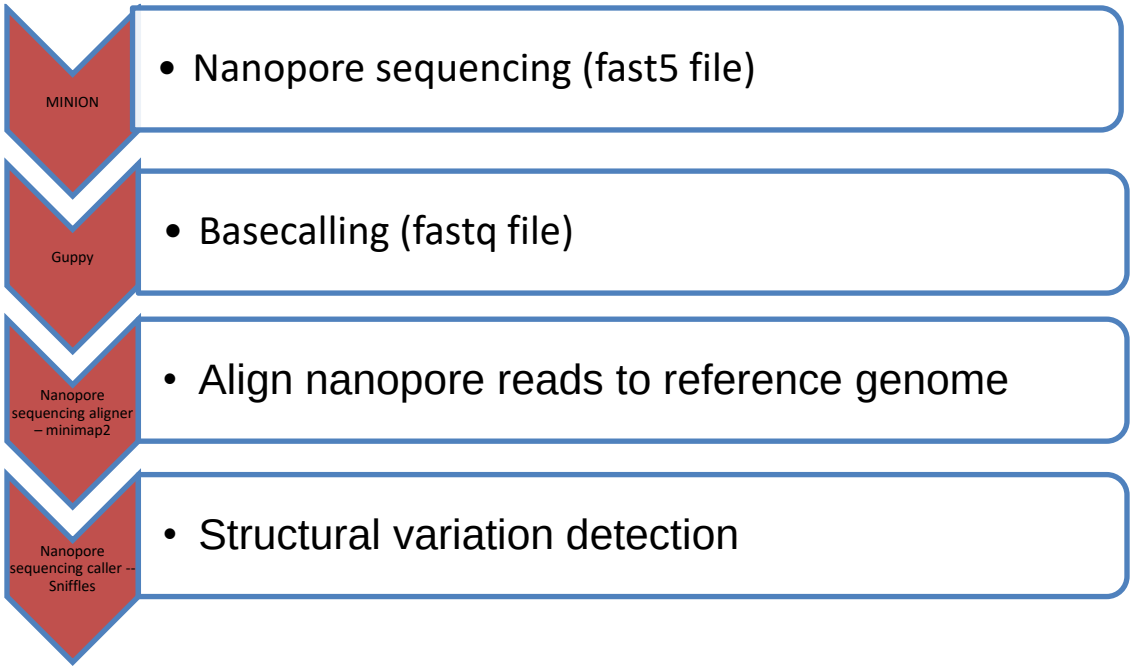


Figure 4.5: Structural variation analysis pipeline for nanopore sequencing data.

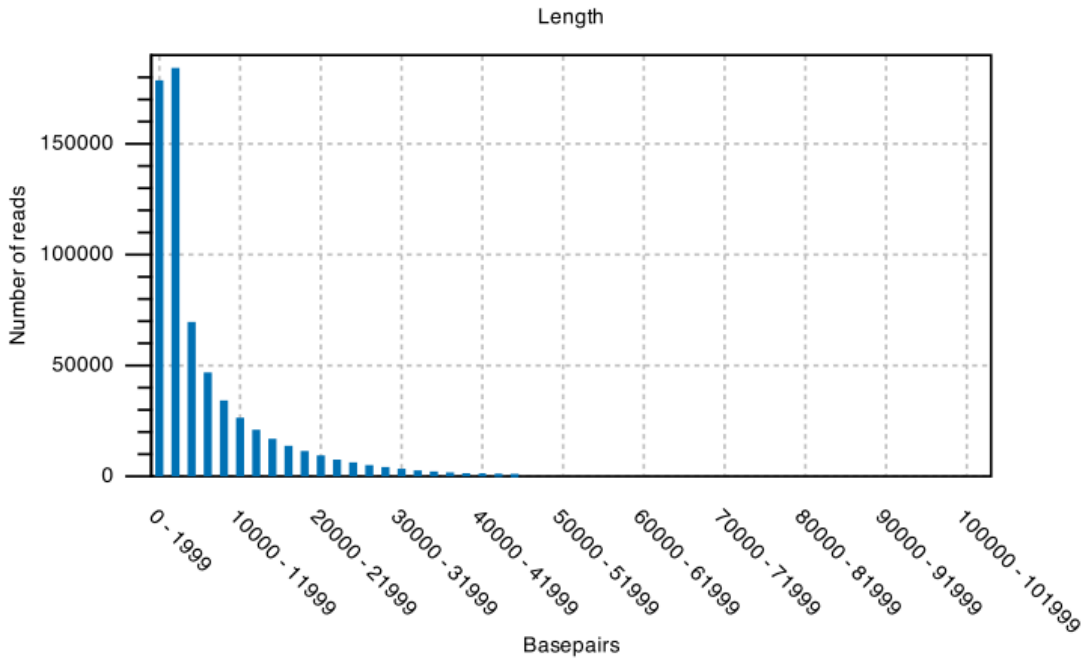


Figure 4.6 Distribution of read lengths.

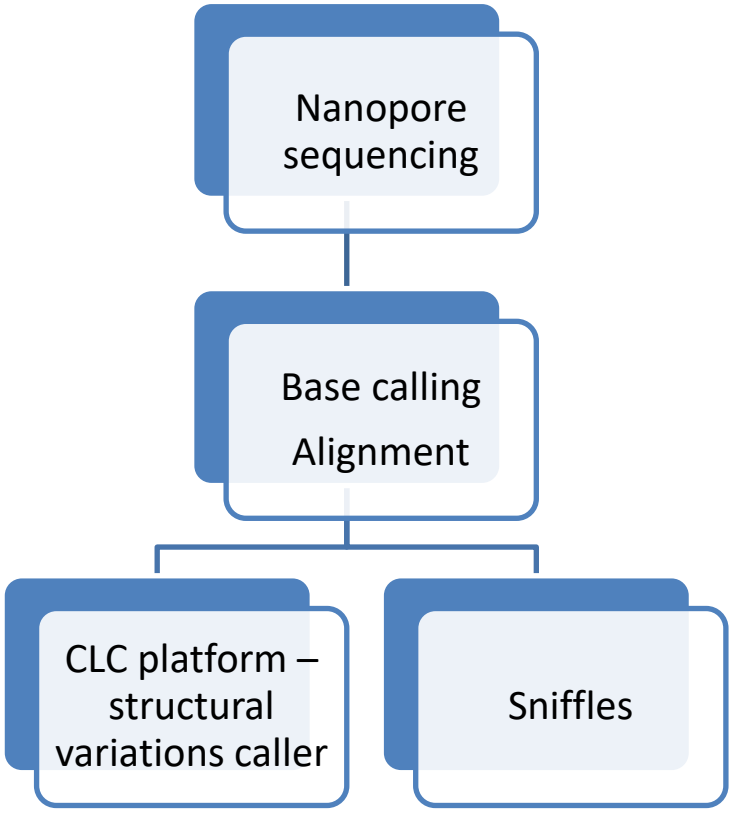


Figure 4.7 Structural variation analysis pipeline.

Table 4.1 Coverage statistics of Nanopore sequencing

2.2 Coverage statistics

Total reference length	2,723,431,143
Minimum coverage	0
Maximum coverage	13,974
Median coverage	1.00
Average coverage	1.48

Table 4.1 Coverage statistics of Nanopore sequencing.

- 816,940,284 positions have coverage below 1 (not shown in graph).
- 1,906,390,270 positions have coverage between 1 and 99.
- 100,589 positions have coverage above 99 (not shown in graph).
- Note that positions with an ambiguous nucleotide in the reference (i.e., not A, C,T or G), count as zero coverage regions, regardless of the number of reads mapping across them.

Table 4.2 Mapping statistics of nanopore sequencing.

	Count	Percentage of reads (%)	Average length	Number of bases
References	22	-	123,792,324.68	2,723,431,143
Mapped reads	567,050	86.97	7,294.68	4,136,446,203
Not mapped reads	84,950	13.03	3,153.93	267,926,325
Total reads	652,000	100.00	6,755.17	4,404,372,528

REFERENCES

1. Zou, X., Owusu, M., Harris, R. et al. Validating the concept of mutational signatures with isogenic cell models. Nat Commun 9, 1744 (2018). https://doi.org/10.1038/s41467-018-04052-8

2. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. Cell 149, 979–993 (2012).

3. Nik-Zainal, S. et al. The life history of 21 breast cancers. Cell 149, 994–1007 (2012).

4. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 3, 246–259 (2013).

5. Dees, N. D. et al. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 22, 1589–1598 (2012).

6. Alexandrov, L.B., Kim, J., Haradhvala, N.J. et al. The repertoire of mutational signatures in human cancer. Nature 578, 94–101 (2020). https://doi.org/10.1038/s41586-020-1943-3

7. Y. Nakabeppu, E. Ohta, N. Abolhassani MTH1 as a nucleotide pool sanitizing enzyme, Friend or foe? Free Radic. Biol. Med., 107 (2017), pp. 151-158

8. Nishii, R., Mizuno, T., Rehling, D., Smith, C., Clark, B. L., Zhao, X., Brown, S. A., Smart, B., Moriyama, T., Yamada, Y., Ichinohe, T., Onizuka, M., Atsuta, Y., Yang, L., Yang, W., Thomas, P. G., Stenmark, P., Kato, M., and Yang, J. J. (2021) NUDT15 polymorphism influences the metabolism and therapeutic effects of acyclovir and ganciclovir, *Nat Commun 12*, 4181.

9. Wyatt, M. D., Reilly, N. M., Patel, S., Rajesh, P., Schools, G. P., Smiraldo, P. G., and Pittman, D. L. (2018) Thiopurine-induced mitotic catastrophe in Rad51d-deficient mammalian cells, Environ Mol Mutagen 59, 38-48.

10. Cara, Carlos J., et al. "Reviewing the mechanism of action of thiopurine drugs: towards a new paradigm in clinical practice." Medical science monitor 10.11 (2004): RA247-RA254.

11. Li B, Brady SW, Ma X, et al. Therapyinduced mutations drive the genomic landscape

of relapsed acute lymphoblastic leukemia. Blood. 2020;135(1):41-55.

12. Tanaka Y, Saito Y. Importance of NUDT15 Polymorphisms in Thiopurine Treatments. J Pers Med. 2021 Aug 10;11(8):778. doi: 10.3390/jpm11080778. PMID: 34442422; PMCID: PMC8399029.

13 Takagi Y, Setoyama D, Ito R, Kamiya H, Yamagata Y, Sekiguchi M. Human MTH3 (NUDT18) protein hydrolyzes oxidized forms of guanosine and deoxyguanosine diphosphates: comparison with MTH1 and MTH2. J Biol Chem. 2012 Jun 15;287(25):21541-9. doi: 10.1074/jbc.M112.363010. Epub 2012 May 3. PMID: 22556419; PMCID: PMC3375575.

14. Wyatt, M. D. and Pittman, D. L. (2006) Methylating agents and DNA repair responses: Methylated bases and sources of strand breaks. Chem. Res. Toxicol. 19, 1580–1594, DOI: 10.1021/tx060164e

15. Christine S. Walsh, Two decades beyond BRCA1/2: Homologous recombination, hereditary cancer risk and a target for ovarian cancer therapy, Gynecologic Oncology, Volume 137, Issue 2, 2015, Pages 343-350, ISSN 0090-8258, https://doi.org/10.1016/j.ygyno.2015.02.017.

16. Mehrgou, A., & Akouchekian, M. (2016). The importance of BRCA1 and BRCA2 genes mutations in breast cancer development. Medical journal of the Islamic Republic of Iran, 30, 369.).

17. Prakash, R., Zhang, Y., Feng, W. & Jasin, M. Homologous recombination and human health: The roles of BRCA1, BRCA2, and associated proteins. Cold Spring Harbor Perspectives in Biology 7, (2015).

18.T Golan, GM O'Kane, RE Denroche, et al. Genomic Features and Classification of Homologous Recombination Deficient Pancreatic Ductal Adenocarcinoma Gastroenterology, 160 (6) (2021), pp. 2119-2132e9

19. Morganella, S. et al. The topography of mutational processes in breast cancer genomes. Nat. Commun. 7, 11383 EP (2016).

20. De Coster W, De Rijk P, De Roeck A, De Pooter T, D'Hert S, Strazisar M, Sleegers K, Van Broeckhoven C. Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. Genome Res. 2019 Jul;29(7):1178-1187. doi: 10.1101/gr.244939.118. Epub 2019 Jun 11. PMID: 31186302; PMCID: PMC6633254.

21. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47–54 (2016)

22. Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat. Med. 23, 517–525 (2017)

23. Jain, M., Koren, S., Miga, K. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36, 338–345 (2018).

24. A. Cerbinskaite, A. Mukhopadhyay, E.R. Plummer, N.J. Curtin, R.J. Edmondson Defective homologous recombination in human cancers Cancer Treat. Rev., 38 (2012), pp. 89-100

25. A.C. Bester, M. Roniger, Y.S. Oren, M.M. Im, D. Sarni, M. Chaoat, et al. Nucleotide deficiency promotes genomic instability in early stages of cancer development Cell, 145 (2011), pp. 435-446

26. H. Kasai, S. Nishimura, Hydroxylation of deoxyguanosine at the C-8 position by ascorbic-acid and other reducing agents, Nucleic Acids Res. 12 (1984) 2137–2145.

27. J.Y. Mo, H. Maki, M. Sekiguchi, Hydrolytic elimination of a mutagenic nucleotide, 8-oxodGTP, by human 18-kilodalton protein: sanitization of nucleotide pool, Proc. Natl. Acad. Sci. USA 89 (1992) 11021–11025.

28. H. Kamiya, H. Kasai, Formation of 2-hydroxydeoxyadenosine triphosphate, an oxidatively damaged nucleotide, and its incorporation by DNA polymerases. Steady-state kinetics of the incorporation, J. Biol. Chem. 270 (1995) 19446–19450.

29. Y. Nakabeppu, M. Behmanesh, H. Yamaguchi, D. Tsuchimoto, K. Sakumi, Prevention of the mutagenecity and cytotoxicity of oxidized purine nucleotides, in: M. Evans, M. Cooke (Eds.), Oxidative Damage to Nucleic Acids, Landes Bioscience, Springer Science+Business Media, Austin, TE, USA and New York, NY, USA, 2007, pp. 40–53.

30. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, Gomez C, Degasperi A, Harris R, Jackson SP, Arlt VM, Phillips DH, Nik-Zainal S. A Compendium of Mutational Signatures of Environmental Agents. Cell. 2019 May 2;177(4):821-836.e16. doi: 10.1016/j.cell.2019.03.001. Epub 2019 Apr 11. PMID: 30982602; PMCID: PMC6506336.

31. Geòrgia Escaramís, Elisa Docampo, Raquel Rabionet, A decade of structural variants: description, history and methods to detect structural variation, *Briefings in Functional Genomics*, Volume 14, Issue 5, September 2015, Pages 305–314, https://doi.org/10.1093/bfgp/elv014

32. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder

M. Paired-end mapping reveals extensive structural variation in the human genome. Science. 2007 Oct 19;318(5849):420-6. doi: 10.1126/science.1149504. Epub 2007 Sep 27. PMID: 17901297; PMCID: PMC2674581.

33. Kosugi, S., Momozawa, Y., Liu, X. et al. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol 20, 117 (2019). https://doi.org/10.1186/s13059-019-1720-5

34. Zahn LM. Filling the gaps. Science. 2022 Apr;376(6588):42-43. doi: 10.1126/science.abp8653. Epub 2022 Mar 31. PMID: 35357909.

35. P.G. Smiraldo, A.M. Gruver, J.C. Osborn, D.L. Pittman
Extensive chromosomal instability in Rad51d-deficient mouse cells
Cancer Res., 65 (2005), pp. 2089-2096

36. Rajesh P, Litvinchuk AV, Pittman DL, Wyatt MD. The homologous recombination protein RAD51D mediates the processing of 6-thioguanine lesions downstream of mismatch repair. Mol Cancer Res. 2011 Feb;9(2):206-14. doi: 10.1158/1541-7786.MCR-10-0451. Epub 2011 Jan 4. PMID: 21205838; PMCID: PMC3041871