METHODS DEVELOPMENT FOR TARGETED NANOPORE SEQUENCING

by

Timothy Gilpatrick

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

March, 2022

© 2022 Timothy Gilpatrick

All rights reserved

Abstract

This dissertation focuses on methods development for "third-generation" (long-read) sequencing technologies. With an emphasis on nanopore sequencing, this work discusses strategies and applications for targeted sequencing of select genomic loci. The methods described here make extensive use of the CRISPR/Cas9 system for target enrichment, adapting these tools to ligate sequencing adaptors at desired loci. We use this approach to evaluate at numerous features salient to human neoplasia: DNA methylation, structural variation, point mutations, and chromatin accessibility. The methods are then applied to cell lines and primary patient tissue; and these genomic features are evaluated and compared.

Thesis Committee

Primary Readers

Winston Timp (Primary Advisor)
Assistant Professor
Department of Biomedical Engineering
Johns Hopkins Whiting School of Engineering

Michael Schatz
Associate Professor
Department of Computer Science
Johns Hopkins Whiting School of Engineering

Alternate Readers

Ralph Hruban
Professor
Department of Pathology
Johns Hopkins University School of Medicine

Acknowledgments

Thanks to the Timp Lab, both the bossman at the helm (Winston) as well as all the other members who helped me through the process of my thesis. Thanks to the folks at Oxford Nanopore for all of their input and help in method development. Thanks to all of the Timp Lab collaborators for helping this take shape. And a huge thanks to my friends and family for supporting me through the thesis process.

Table of Contents

Abstract ii

Thesis Committee iii

Acknowledgements iv

Table of Contents v

List of Tables ix

List of Figures xi

1 Introduction 1

2 History of Targeted Sequencing 15

2.1 Introduction 15

2.2 NextGen Targeted Sequencing 16

2.3 Long-read Targeted Sequencing 16

2.4 Nanopore Cas9 Targeted Sequencing (nCATS) 21

3 Targeted Sequencing in Breast Cancer **27**

3.1 Introduction 27

3.2 Results 28

3.2.1 Single Nucleotide Variant Detection 31

3.2.2 CpG Methylation 34

3.2.3 Structural Variants 36

3.3 Discussion 41

3.4 Methods 42

3.4.1 Cell culture and DNA prep 42

3.4.2 Patient Tissue and Mouse Xenograft 43

3.4.3 GuideRNA design 43

3.4.4 Ribonucleoprotein Complex Assembly 43

3.4.5 Cas9 Cleavage and Library Prep 44

3.4.6 Sequencing 45

3.4.7 Data Analysis 45

3.5 Supplementary Material 47

4 TERT Sequencing in Thyroid Cancer **69**

4.1 Introduction 69

4.2 Results 72

4.2.1 CpG Methylation Studies 72

4.2.2 Chromatin Immunoprecipitation 75

4.2.3 TERT Transcriptional Analysis 76
4.3 Discussion 79
4.4 Methods 81
4.4.1 Cell lines, culture conditions, and TERT mutation status 81
4.4.2 DNA Isolation and Bisulfite Modification 82
4.4.3 Bisulfite Sequencing PCR of TERT Promoter 82
4.4.4 Bisulfite Library Preparation and Sequencing 83
4.4.5 Data Processing for Methylation 83
4.4.6 Chromatin Immunoprecipitation (ChIP) Analysis . . . 84
4.4.7 TERT Expression qRT-PCR 84
4.4.8 Nanopore Cas9 targeted sequencing 85
4.4.9 Data processing for methylation from nanopore data . 85
4.4.10 Allele-specific transcription characterization 86
4.5 Supplementary Materials 86

5 Additional Applications and Future Directions of Targeted Sequencing 92

5.1 Introduction 92
5.2 Gene Localization 93
5.2.1 Background 93
5.2.2 Results 94
5.2.3 Discussion 98
5.3 Targeted NanoNOME 99

5.3.1 Background 99
5.3.2 Results 101
5.3.3 Discussion 102
5.4 In vitro transcribed guideRNAs 105
5.4.1 Background 105
5.4.2 Results 106
5.4.3 Discussion 108
5.5 Methods 108
5.5.1 Sequencing 108
5.5.2 Locating Gene Insertions 109
5.5.3 GpC methyltransferase treatment for targeted nanoNOME 109
5.5.4 Generating in vitro transcribed guideRNAs 110
5.6 Supplementary Material 110

6 Discussion and Conclusion **114**

List of Tables

3.1 **Regions for Breast Cancer Studies** Type of aberration and size for the initial panel of 10 loci targeted 28

3.2 **Coverage Summary** Coverage at all loci for each MinION sequencing run 50

3.3 **Off-target Analysis** Off-target Analysis with SURVIVOR . . . 51

3.4 **MDA-MB-231 Single Nucleotide Variants** SNVs identified in MDA-MB-231 cell line enrichment data 53

3.5 **Sniffles calls in breast cell lines** Indels called at deletion locations in breast cell lines 59

3.6 **Sniffles calls in GM12878** Indels called at large heterozygous deletions in GM12878 59

3.7 **Comparing BRCA1 alleles in GM12878** Indels called between the assemblies of GM12878 alleles 60

3.8 **Breast Cancer GuideRNAs** GuideRNA sequences and target loci 63

4.1 **TERT mutations in cell lines** Mutation status of the TERT promoter in four thyroid cell lines 72

4.2 **Primers for TERT Analysis** Bisulfite, ChIP and cDNA primers for TERT 86

5.1 **Read Counts for CHO insertions** CHO clone identifier and total sequencing reads for the corresponding sample 94

5.2 **Targeted NanoNOME, regional coverage** The max coverage achieved within the targeted for the respective gene 101

List of Figures

1.1 **DNA in a cell** Illustration showing layers of DNA organization 3

1.2 **Long Read Sequencing Methods** Images showing how Oxford Nanopore and SMRT sequencing operate 7

2.1 **NextGen Target Enrichment** Schematic image of two leading methods for target enrichment with NextGen sequencing . . 17

2.2 **Long-read Target Enrichment** Summary of target enrichment methods using Cas9 for long-read sequencing 19

2.3 **Schematic of Cas9-targeted sequencing (nCATS)** Cartoon detailing steps of Cas9 enrichment 22

3.1 **On-target Performance Summary** On-target coverage for one region in cell line and tissue data, and summary of on-target metrics. 30

3.2 **Single Nucleotide Variants** Single Nucleotide Variant (SNV) analysis using nCATS data 33

3.3 **Methylation Analysis** Highlights of CpG methylation analysis using nCATS data 37

3.4 **Structural Variation** Structural variation analysis from nCATS data 39

3.5 **Example Off-Target Cleaving from BRCA1 guideRNA** Off-target coverage/reads and pairwise alignment between guideRNA and off-target site 48

3.6 **Stranded Information** Using strand information to validate nanopore-identified variants 49

3.7 **Persisting False Positive SNV** Details on a false positive single nucleotide variant (SNV) in a thymidine-dense homopolymeric region 52

3.8 **Tumor Loss of Heterozygosity** Two sites in tumor tissue demonstrating loss of heterozygosity on chr17 54

3.9 **Comparison of nCATS methylation Data with WGBS data** Methylation line plots, read-level plots and per-CpG plots for five loci in GM12878 enrichment data 55

3.10 **Breast cell line methylation** Read-level methylation plots for all methylation-associated loci in three breast cell lines 56

3.11 **Transcript level comparison in breast cell lines** RNA-seq expression data for five methylation-associated genes 56

3.12 **Primary tissue methylation** Read-level methylation plots for captured loci in primary breast tissue 57

3.13 **Chromosome 5 deletion in breast cell lines** Reads at a small (< 10kb) common structural variant on chromosome 5 58

3.14 **Per-allele coverage around large deletions** Coverage plots for large heterozygous deletions in GM12878 60

3.15 **Reads at the BRCA1 locus for GM12878** Reads and coverage in full-length BRCA1 enrichment data 61

3.16 **Validation of novel BRCA1 indels** Comparison of BRCA1 nanopore reads to PacBio reads at unannotated indels 62

4.1 **Tert Methylation with Bisulfite Amplicons** Methylation data in thyroid cell lines using tiled bisulfite amplicons 74

4.2 **TERT methylation with nCATS Data** Allele-specific methylation patterns in thyroid cell lines determined from targeted nanopore sequencing data 75

4.3 **Chromatin Immunoprecipitation in Thyroid Cell lines** ChIP-qPCR and ChIP-Sanger for CTCF, MYC and GABPA 77

4.4 **TERT Transcript Analysis** qPCR and Sanger sequencing of the TERT transcript 78

4.5 **ChIP-Sanger for Histone Modifications** H3K4me3 and H3K27me3 ChIP paired with Sanger sequencing for a TERT promoter-associated amplicon in thyroid cell lines 87

4.6 **Sanger sequencing of BCPAP cDNA** Exon 2 sequencing of TERT cDNA, containing a heterozygous mutation in BCPAP . 88

5.1 **Cas9 Targeted Sequencing for Insert Localization** Cartoon showing the strategy and resulting reads using targeted cleavage to locate insertion points 96

5.2 **Expression Vector Map** Map showing the genes and regulatory elements in the assembled expression vector sequence 97

5.3 **NanoNOME Schematic** Cartoon detailing the exogenous labeling strategy of the NOMEseq/nanoNOME 100

5.4 **NanoNOME Data at KRAS and GPX1** CpG methylation and accessibility of these two genes in GM12878 103

5.5 **Kernel-Smoothed NanoNOME** Read-level CpG methylation and kernel-smoothed accessibility data at the KRAS promoter 104

5.6 **In vitro transcribed guideRNA panel** Coverage and reads for a 500kb enriched region around genes commonly harboring deletions in pancreatic ductal adenocarcinoma (PDAC) 107

5.7 **Targeted versus whole-genome nanoNOME** Comparing CpG methylation and GpC methylation calls between targeted and whole-genome nanoNOME data 111

Chapter 1

Introduction

DNA is a data storage system, packaged within the nucleus of every living cell that carries the blueprints for how to make all the machinery and infrastructure required for the cell to operate. To perform different tasks or take on specialized functions (especially in the case of multicellular organisms), cells will utilize different parts of their genomic blueprints. The DNA (and associated histone proteins) is organized into a condensed form termed chromatin, and different regions of chromatin are "open" or "closed" to meet the functional needs of the cell. This organizational and regulatory system can broadly be described under the umbrella term "epigenetics". In the present era we recognize that epigenetics takes a number of forms with varying amounts of plasticity (**Figure 1.1**). One key mechanism controlling chromatin state is the nucleosome, which describes the octamer of histone proteins and the 147nt of DNA wrapped around it (Klemm, Shipony, and Greenleaf, 2019). The arrangement of histone proteins, as well as post-translational modifications to these proteins help to control DNA accessibility through numerous mechanisms including steric hindrance of transcription factor binding and

initiation of chromatin remodeling (Allis and Jenuwein, 2016). Another crucial epigenetic feature is that of DNA methylation. In mammalian systems, this occurs at cytosine nucleotides, specifically those directly preceding a guanine (CpG dinucleotide). The default status of CpGs in the genome appears to be methylated, with demethylation occurring at the start of genes as a regulatory mechanism. This is supported by the observation that many genes possess a region of increased CpG density around gene start sites (CpG islands) (Bird et al., 1985). It has long been understood that CpG methylation is associated with transcriptional silencing (Razin and Riggs, 1980), but where DNA methylation sits in the hierarchy of epigenetic control has become a point of debate. Recent studies have suggested that DNA methylation is not the primary mechanism of gene silencing; it was observed that the presence of nucleosomes (without activation marks) is required for DNA methylation to occur (Ooi et al., 2007). This suggests that DNA methylation may be a less plastic and more stable form of epigenetic modification. Our ability to explore these features in the context of an organism's full genome has been richly enhanced by technological developments in the area of nucleic acid sequencing.

The last four decades have seen rapid advancements in tools to sequence nucleotides (determine the order of A,C,T,G). These tools have become critical parts of the arsenal of both science and medicine. In the mid 1970s, the first wave of DNA sequencing strategies were presented. This included a method described by Sanger and Coulson which involved DNA synthesis with the use of chain-terminating nucleotides ("Sanger sequencing") (Sanger,

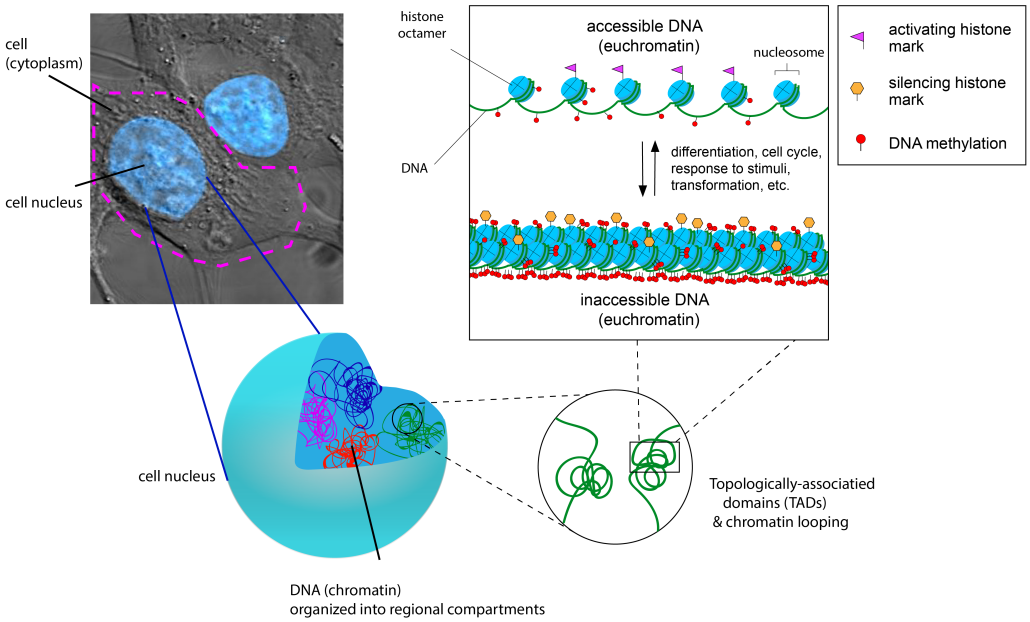


Figure 1.1: DNA in a cell Illustration showing layers of DNA organization. (Bottom Left): Cell nucleus showing regional domains of chromatin. (Bottom Right): Topologically-associated Domains (TADs) and chromatin looping. (Top Right): Cartoon showing histone post-translational modifications and CpG methylation in both active (euchromatin) and inactive (heterochromatin) states. The cell image was taken from the wikipedia commons. Public Domain, contributed by TenOfAllTrades https://commons.wikimedia.org/w/index.php?curid=1583880

Nicklen, and Coulson, 1977) and a method from Maxam and Gilbert using chemicals to induce nucleotide-specific partial cleavage (Maxam and Gilbert, 1977). In both cases, separate reactions were used for each of the four bases, and terminally radio-labelled DNA were size-separated using electrophoresis, allowing researchers to determine the order of nucleotides based on the resultant fragment sizes. By the late 1980s, automated instrumentation existed that were performing fluorescence-based Sanger sequencing, and capable of generating 1000nt /day (Smith et al., 1986). 1990 saw the official launch of the human genome project (HGP), which employed Sanger-based sequencing systems as the workhorse. Throughout the 1990s, sequencing the human genome was occurring at a number of academic genome centres, churning out tens of millions of nucleotides per day (Shendure et al., 2017). The spirit of competition was also a driving force in this process as both private companies and government-funded institutions worked to generate the human genome. This culminated with the first draft of the human genome released in 2001, and a completed sequence released in 2004 (International Human Genome Sequencing Consortium, 2004) .

During the execution of the HGP, newer strategies for nucleotide sequencing were developed in both the public and private sectors. This led to the development of massively parallel "Next Generation" (NextGen) sequencing (NGS), providing alternative approaches to the electrophoretic separation strategies that underlie Sanger sequencing. The key change that enabled massively parallel sequencing is the ability to multiplex large numbers of sequencing reads in the same reaction (> 1 billion for the Illumina® HiSeq).

This multiplexing of sequencing reactions is achieved by immobilizing DNA strands on a 2D surface, which allows the same volume of reagents to be used for all sequencing reactions. Further, NGS methods no longer rely on size separation of DNA by size, but instead detect the incorporation of nucleotides during *in vitro* DNA strand replication. Three strategies that have persisted to present day for identifying nucleotide incorporation with NGS. The first detects the pyrophosphate molecule released during nucleotide incorporation through a luciferase reaction (pyrosqeuencing) (Ronaghi et al., 1996), and a similar method came more than a decade later which detects the release of hydrogen ions during this reaction (ion semiconductor sequencing) (Toumazou and Purushothaman, 2010). Both of these methods require separate addition of the four nucleotides sequentially to detect nucleotide incorporation. A third method works through the detection of fluorescently labeled nucleotides (Braslavsky et al., 2003; Mitra et al., 2003), and a big advancement was the introduction of reversible "terminators" (Ruparel et al., 2005; Seo et al., 2005) to ensure only only single nucleotides incorporated at a time– enabling simultaneous detection of all four fluorescently-labeled nucleotides. This last approach laid the foundation for the company Solexa, founded in 1998 and purchased by Illumina in 2007, which has become the sequencing industry's dominating player. As a result of the rapid development and market competition, the cost of sequencing has fallen dramatically. Generating whole-genome sequencing data for an individual human- a project that in recent memory took more than a decade and cost nearly 3 billion dollars, is now done routinely in a few days for a few thousand dollars. This vastly increased the accessibility of sequencing, leading the methods to be adopted by academic and private

institutions worldwide.

NGS has become a critical component of modern science, but innovation in DNA sequencing has not ceased. NGS methods rely on PCR amplification in order to generate DNA "clusters", such that hundreds of identical copies of DNA are being measured simultaneously. This amplifies the sequencing signal, but requires the DNA to be broken into small pieces, and leads to a destruction of any modified nucleotides that were present. Further, the number of nucleotides in a row that can be evaluated (the "read length") by these clusters are limited to a few hundred, as the signal becomes unreliable beyond these lengths as the DNA clusters become out of sync. The newer technologies that have arisen to overcome these challenges are the so-called "long-read sequencing" strategies. At the crest of the current wave of long-read sequencing technologies (sometimes called "third-generation") are nanopore sequencing from Oxford Nanopore Technologies and single-molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio) (**Figure 1.2**) . Nanopore sequencing functions by collecting electrical data as DNA is cranked through DNA through a protein pore in an electric field (Jain et al., 2016). In contrast, PacBio operates by performing DNA synthesis within a nano-sizedwell (termed a "zero mode waveguide"), which has optical properties that permit detection of the sequential incorporation of fluorescent nucleotides at the single molecule scale (Travers et al., 2010). Long-read methods which can produce reads 30-50kb for hi-fidelity SMRT sequencing (consensus circular sequencing or CCS) (Amarasinghe et al., 2020), and have been reported at lengths greater than 2 megabases for nanopore sequencing reads (Payne et al., 2019).

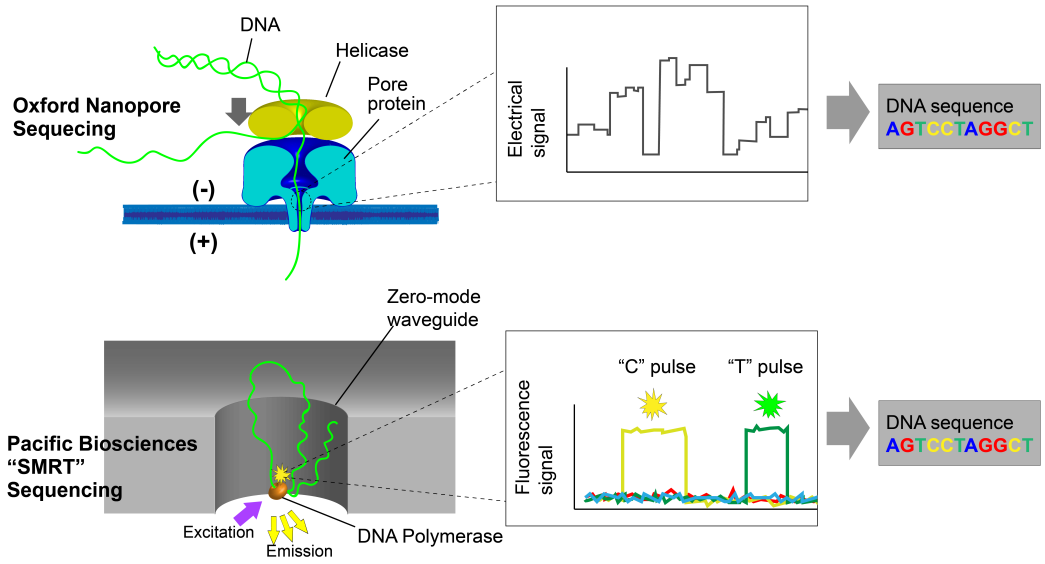


Figure 1.2: Long Read Sequencing Methods Top: Portrayal of how DNA passes through a protein pore in an electric field and the collection of electrical data which is converted into DNA sequence. Bottom: Portrayal of PacBio SMRT sequencing, where a tethered DNA polymerase copies DNA at bottom of nano-sized well and detection of sequential fluorescent nucleotides

Long-read methods enable the sequencing of single molecules of native DNA. By entirely avoiding PCR amplification, modified nucleotides are preserved– which are able to be detected in sequencing data from both nanopore (Simpson et al., 2017) and SMRT (Flusberg et al., 2010) sequencing. DNA methylation can be determined with NGS, but requires deamination of unmodified CpGs with chemicals or enzymatic treatments, which reduces the complexity of the nucleotide alphabet. Another advancement resulting of long-read sequencing is the increased ease of interrogating hard to map regions; enhancing our understanding of structural variation in the human genome (Chaisson et al., 2018; Audano et al., 2019) and improving our ability to interrogate hard-to-map regions such as repetitive elements and centromeres (Miga et al., 2020).

One field that has been dramatically affected through the advancement of sequencing strategies is the study of human cancers. Cancer is an inherently heterogenous disease with a variety underlying causative mechanisms. A number of unifying features of cancers have been identified, termed the "hallmarks of cancer". One such hallmark is instability of the genome, as cancer cells lose their ability recognize genomic insult and cease proliferation (Hanahan and Weinberg, 2011). This instability is often reflected directly in the coding sequence, (e.g. inactivating mutations in a tumor suppressor gene (Knudson et al., 1976)), and is also reflected in the epigenome (e.g. hypermethylation leading to silencing of tumor suppressor genes (Feinberg and Tycko, 2004)). The high-accessibility of sequencing in the current era has spurred efforts to characterize patient tumors and to cultivate databases of cancer-associated mutations (Ainscough et al., 2016). Further, sequencing offers exquisitely sensitive tools which can be used to detect cancer initiation or surveil for recurrence (Cohen et al., 2018). Long read strategies have been beneficial in the cancer realm as well, helping to decipher the structural variation landscape of human neoplasia (Dixon et al., 2018).

This sets the stage for the next chapter where we begin to focus not on whole-genome or whole-transcriptome sequencing, but rather methods to only sequence smaller regions e.g. single genes or sets of genes. In the subsequent chapters I describe method development and application of targeted strategies with for long-read sequencing (primarily nanopore sequencing). This work largely centers on applications to the study of human cancer, with sections decribing our work on breast cancer as well as thyroid cancer. Finally, there

is also a section describing ongoing development work to use these tools for interrogating large hotspots for cancer-causing deletions, applications in the bio-pharmaceutical to locate insertion sites of transgenes, and a method for evaluating chromatin state from nanopore sequencing data. Together this work builds on our existing repertoire of sequencing strategies to offer new tools and computational approaches for investigating features of the genome.

References

Klemm, Sandy L, Zohar Shipony, and William J Greenleaf (2019). "Chromatin accessibility and the regulatory epigenome". en. In: *Nat. Rev. Genet.* 20.4, pp. 207–220.

Allis, C David and Thomas Jenuwein (2016). "The molecular hallmarks of epigenetic control". en. In: *Nat. Rev. Genet.* 17.8, pp. 487–500.

Bird, A, M Taggart, M Frommer, O J Miller, and D Macleod (1985). "A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA". en. In: *Cell* 40.1, pp. 91–99.

Razin, A and A D Riggs (1980). "DNA methylation and gene function". en. In: *Science* 210.4470, pp. 604–610.

Ooi, Steen K T, Chen Qiu, Emily Bernstein, Keqin Li, Da Jia, Zhe Yang, Hediye Erdjument-Bromage, Paul Tempst, Shau-Ping Lin, C David Allis, Xiaodong Cheng, and Timothy H Bestor (2007). "DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA". en. In: *Nature* 448.7154, pp. 714–717.

Sanger, F, S Nicklen, and A R Coulson (1977). "DNA sequencing with chain-terminating inhibitors". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 74.12, pp. 5463–5467.

Maxam, A M and W Gilbert (1977). "A new method for sequencing DNA". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 74.2, pp. 560–564.

Smith, L M, J Z Sanders, R J Kaiser, P Hughes, C Dodd, C R Connell, C Heiner, S B Kent, and L E Hood (1986). "Fluorescence detection in automated DNA sequence analysis". en. In: *Nature* 321.6071, pp. 674–679.

Shendure, Jay, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston (2017). "DNA sequencing at 40: past, present and future". en. In: *Nature* 550.7676, pp. 345–353.

International Human Genome Sequencing Consortium (2004). “Finishing the euchromatic sequence of the human genome”. en. In: *Nature* 431.7011, pp. 931–945.

Toumazou, Christofer and Sunil Purushothaman (2010). “Sensing apparatus and method”. Pat. 7686929.

Braslavsky, Ido, Benedict Hebert, Emil Kartalov, and Stephen R Quake (2003). “Sequence information can be obtained from single DNA molecules”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 100.7, pp. 3960–3964.

Mitra, Robi D, Jay Shendure, Jerzy Olejnik, Edyta-Krzymanska-Olejnik, and George M Church (2003). “Fluorescent in situ sequencing on polymerase colonies”. en. In: *Anal. Biochem.* 320.1, pp. 55–65.

Ruparel, Hameer, Lanrong Bi, Zengmin Li, Xiaopeng Bai, Dae Hyun Kim, Nicholas J Turro, and Jingyue Ju (2005). “Design and synthesis of a 3’-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 102.17, pp. 5932–5937.

Seo, Tae Seok, Xiaopeng Bai, Dae Hyun Kim, Qinglin Meng, Shundi Shi, Hameer Ruparel, Zengmin Li, Nicholas J Turro, and Jingyue Ju (2005). “Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 102.17, pp. 5926–5931.

Jain, Miten, Hugh E Olsen, Benedict Paten, and Mark Akeson (2016). “The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community”. en. In: *Genome Biol.* 17.1, p. 239.

Travers, Kevin J, Chen-Shan Chin, David R Rank, John S Eid, and Stephen W Turner (2010). “A flexible and efficient template format for circular consensus sequencing and SNP detection”. en. In: *Nucleic Acids Res.* 38.15, e159.

Amarasinghe, Shanika L, Shian Su, Xueyi Dong, Luke Zappia, Matthew E Ritchie, and Quentin Gouil (2020). “Opportunities and challenges in long-read sequencing data analysis”. en. In: *Genome Biol.* 21.1, p. 30.

Payne, Alexander, Nadine Holmes, Vardhman Rakyan, and Matthew Loose (2019). “BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files”. en. In: *Bioinformatics* 35.13, pp. 2193–2198.

Simpson, Jared T, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi, and Winston Timp (2017). “Detecting DNA cytosine methylation using nanopore sequencing”. en. In: *Nat. Methods* 14.4, pp. 407–410.

Flusberg, Benjamin A, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson A Clark, Jonas Korlach, and Stephen W Turner (2010). "Direct detection of DNA methylation during single-molecule, real-time sequencing". en. In: *Nat. Methods* 7.6, pp. 461–465.

Chaisson, Mark J P, Ashley D Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J Gardner, Oscar Rodriguez, Li Guo, Ryan L Collins, Xian Fan, Jia Wen, Robert E Handsaker, Susan Fairley, Zev N Kronenberg, Xiangmeng Kong, Fereydoun Hormozdiari, Dillon Lee, Aaron M Wenger, Alex Hastie, Danny Antaki, Peter Audano, Harrison Brand, Stuart Cantsilieris, Han Cao, Eliza Cerveira, Chong Chen, Xintong Chen, Chen-Shan Chin, Zechen Chong, Nelson T Chuang, Christine C Lambert, Deanna M Church, Laura Clarke, Andrew Farrell, Joey Flores, Timur Galeev, David Gorkin, Madhusudan Gujral, Victor Guryev, William Haynes Heaton, Jonas Korlach, Sushant Kumar, Jee Young Kwon, Jong Eun Lee, Joyce Lee, Wan-Ping Lee, Sau Peng Lee, Shantao Li, Patrick Marks, Karine Viaud-Martinez, Sascha Meiers, Katherine M Munson, Fabio Navarro, Bradley J Nelson, Conor Nodzak, Amina Noor, Sofia Kyriazopoulou-Panagiotopoulou, Andy Pang, Yunjiang Qiu, Gabriel Rosanio, Mallory Ryan, Adrian Stütz, Diana C J Spierings, Alistair Ward, Annemarie E Welch, Ming Xiao, Wei Xu, Chengsheng Zhang, Qihui Zhu, Xiangqun Zheng-Bradley, Ernesto Lowy, Sergei Yakneen, Steven McCarroll, Goo Jun, Li Ding, Chong Lek Koh, Bing Ren, Paul Flicek, Ken Chen, Mark B Gerstein, Pui-Yan Kwok, Peter M Lansdorp, Gabor Marth, Jonathan Sebat, Xinghua Shi, Ali Bashir, Kai Ye, Scott E Devine, Michael Talkowski, Ryan E Mills, Tobias Marschall, Jan O Korbel, Evan E Eichler, and Charles Lee (2018). "Multi-platform discovery of haplotype-resolved structural variation in human genomes". en.

Audano, Peter A, Arvis Sulovari, Tina A Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, Annemarie E Welch, Max L Dougherty, Bradley J Nelson, Ankeeta Shah, Susan K Dutcher, Wesley C Warren, Vincent Magrini, Sean D McGrath, Yang I Li, Richard K Wilson, and Evan E Eichler (2019). "Characterizing the Major Structural Variant Alleles of the Human Genome". en. In: *Cell* 176.3, 663–675.e19.

Miga, Karen H, Sergey Koren, Arang Rhie, Mitchell R Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A Logsdon, Valerie A Schneider, Tamara Potapova, Jonathan Wood, William Chow, Joel Armstrong, Jeanne Fredrickson, Evgenia Pak, Kristof Tigyi, Milinn Kremitzki, Christopher Markovic, Valerie Maduro, Amalia

Dutra, Gerard G Bouffard, Alexander M Chang, Nancy F Hansen, Amy B Wilfert, Françoise Thibaud-Nissen, Anthony D Schmitt, Jon-Matthew Belton, Siddarth Selvaraj, Megan Y Dennis, Daniela C Soto, Ruta Sahasrabudhe, Gulhan Kaya, Josh Quick, Nicholas J Loman, Nadine Holmes, Matthew Loose, Urvashi Surti, Rosa Ana Risques, Tina A Graves Lindsay, Robert Fulton, Ira Hall, Benedict Paten, Kerstin Howe, Winston Timp, Alice Young, James C Mullikin, Pavel A Pevzner, Jennifer L Gerton, Beth A Sullivan, Evan E Eichler, and Adam M Phillippy (2020). "Telomere-to-telomere assembly of a complete human X chromosome". en. In: *Nature*.

Hanahan, Douglas and Robert A Weinberg (2011). "Hallmarks of cancer: the next generation". en. In: *Cell* 144.5, pp. 646–674.

Knudson Jr, A G, A T Meadows, W W Nichols, and R Hill (1976). "Chromosomal deletion and retinoblastoma". en. In: *N. Engl. J. Med.* 295.20, pp. 1120–1123.

Feinberg, Andrew P and Benjamin Tycko (2004). "The history of cancer epigenetics". en. In: *Nat. Rev. Cancer* 4.2, pp. 143–153.

Ainscough, Benjamin J, Malachi Griffith, Adam C Coffman, Alex H Wagner, Jason Kunisaki, Mayank Nk Choudhary, Joshua F McMichael, Robert S Fulton, Richard K Wilson, Obi L Griffith, and Elaine R Mardis (2016). "DoCM: a database of curated mutations in cancer". en. In: *Nat. Methods* 13.10, pp. 806–807.

Cohen, Joshua D, Lu Li, Yuxuan Wang, Christopher Thoburn, Bahman Afsari, Ludmila Danilova, Christopher Douville, Ammar A Javed, Fay Wong, Austin Mattox, Ralph H Hruban, Christopher L Wolfgang, Michael G Goggins, Marco Dal Molin, Tian-Li Wang, Richard Roden, Alison P Klein, Janine Ptak, Lisa Dobbyn, Joy Schaefer, Natalie Silliman, Maria Popoli, Joshua T Vogelstein, James D Browne, Robert E Schoen, Randall E Brand, Jeanne Tie, Peter Gibbs, Hui-Li Wong, Aaron S Mansfield, Jin Jen, Samir M Hanash, Massimo Falconi, Peter J Allen, Shibin Zhou, Chetan Bettegowda, Luis Diaz, Cristian Tomasetti, Kenneth W Kinzler, Bert Vogelstein, Anne Marie Lennon, and Nickolas Papadopoulos (2018). "Detection and localization of surgically resectable cancers with a multi-analyte blood test". en. In: *Science*, eaar3247.

Dixon, Jesse R, Jie Xu, Vishnu Dileep, Ye Zhan, Fan Song, Victoria T Le, Galip Gürkan Yardımcı, Abhijit Chakraborty, Darrin V Bann, Yanli Wang, Royden Clark, Lijun Zhang, Hongbo Yang, Tingting Liu, Sriranga Iyyanki, Lin An, Christopher Pool, Takayo Sasaki, Juan Carlos Rivera-Mulia, Hakan Ozadam, Bryan R Lajoie, Rajinder Kaul, Michael Buckley, Kristen Lee,

Morgan Diegel, Dubravka Pezic, Christina Ernst, Suzana Hadjur, Duncan T Odom, John A Stamatoyannopoulos, James R Broach, Ross C Hardison, Ferhat Ay, William Stafford Noble, Job Dekker, David M Gilbert, and Feng Yue (2018). "Integrative detection and analysis of structural variation in cancer genomes". en. In: *Nat. Genet.* 50.10, pp. 1388–1398.

Chapter 2

History of Targeted Sequencing

2.1 Introduction

Targeted sequencing is an important strategy for examining only select genomic loci; as it avoids some of the costs of generating, storing, and analyzing whole-genome sequencing data. The burgeoning era of long-read sequencing has prompted the development of new enrichment strategies compatible with long-read technologies (Pham et al., 2016; Gabrieli et al., 2018; Giesselmann et al., 2019; Gilpatrick et al., 2020; Tsai et al., 2017). This chapter provides a brief and highly abridged history of target-enrichment strategies, with a focus on amplification-free long-read sequencing. The chapter concludes with a deeper delve into nanopore sequencing with sequencing adaptors ligated directly to DNA ends made by Cas9 endonuclease cleavage (Gilpatrick et al., 2020), which is the centerpoint strategy underlying much of the work presented in this dissertation.

2.2 NextGen Targeted Sequencing

For NextGen sequencing, target enrichment is typically achieved by capture with hybridization probes or via selective amplification with targeted primers (**Figure 2.1**) (Kozarewa et al., 2015). Hybridization capture uses oligonucleotides (matrix-bound or solution phase) to pull-down desired target sequences) (Tewhey et al., 2009). Hybridization has been shown to work with long-read technologies (Eckert et al., 2016; *Hybridization-capture for nanopore sequencing v1 (protocols.io.zxyf7pw)*), but this approach typically involves shearing the DNA (Eckert et al., 2016; *Hybridization-capture for nanopore sequencing v1 (protocols.io.zxyf7pw)*) causing a limit on the read-length (typically about 10kb) and therefore not fully capitalizing on the long-reads. And although PCR amplification is commonly used for some long-read applications, it not only limits read lengths, but also obliterates modified nucleotides losing information that could be gleaned from long-read sequencing data (Simpson et al., 2017; Flusberg et al., 2010).

2.3 Long-read Targeted Sequencing

In order to achieve enrichment while avoiding PCR amplification and maintaining large DNA fragments, new strategies have been described for use with long-read technologies. Several of these strategies employ the CRISPR/Cas9 system, a bacterial endonuclease (Cas9) that uses a guideRNA to introduce double strand breaks (DSBs) at specific loci (Sternberg et al., 2014). The first application of Cas9 cleavage for long-read target enrichment was described

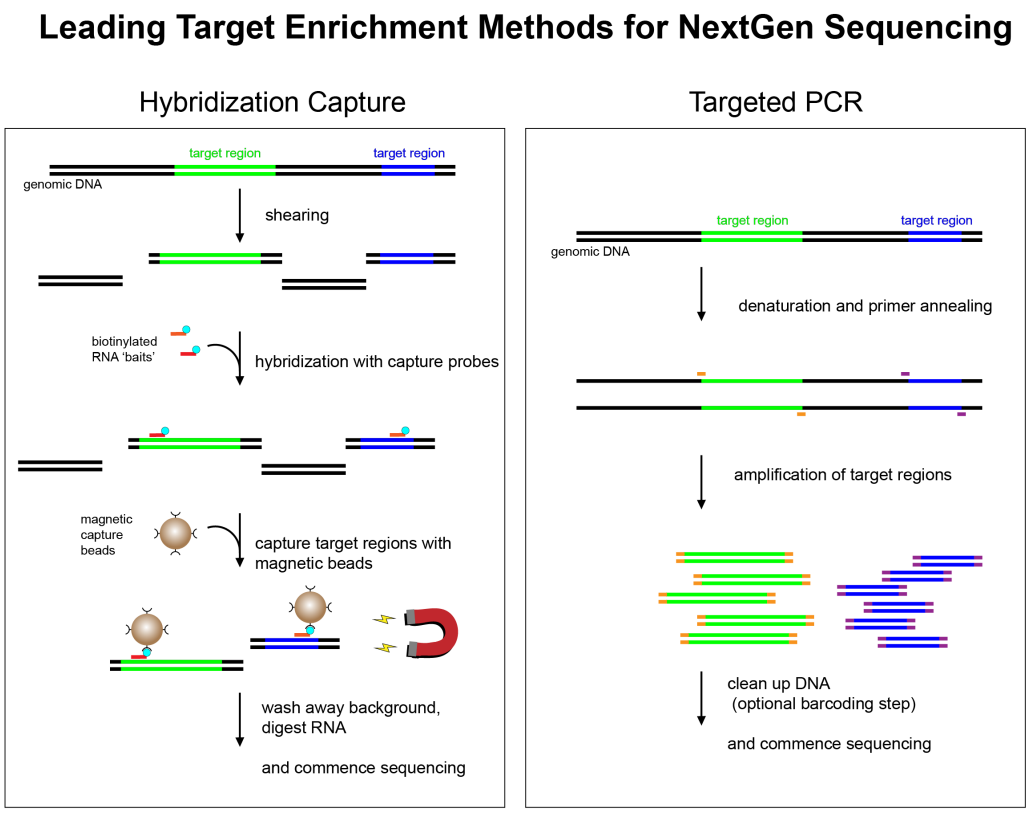


Figure 2.1: NextGen Target Enrichment Images showing the general principles behind hybridization capture and targeted amplification

for SMRT sequencing (Tsai et al., 2017), and used ends made by Cas9 cleavage sites to ligate adaptors for magnetic bead capture (**Figure 2.2**). For nanopore sequencing, early iterations of Cas9 enrichment used two cuts to excise a region, and removed the background DNA by size selection ("CATCH-seq", **Figure 2.2**) (Gabrieli et al., 2018). The CATCH-seq strategy worked well for small genomes, but for large genomes such as homo sapiens, amplification was necessary to achieve significant on-target enrichment. Further iterations of targeted enrichment took advantage of the fact that the Cas9/guideRNA complex remains tightly bound to target DNA after cleavage, and employed affinity handles (e.g. biotin or 6xHis) on the guideRNA/Cas9 complex to pull out the associated DNA via bead capture (**Figure 2.2**). The on-target coverage from these Cas9 pull-down strategies was still lower than desired (50-100X local coverage in the human genome from 3ug of input DNA – Gilpatrick, unpublished results), motivating the continuing development of enrichment methods. This led to "Cas9-targeted sequencing" (CATS), wherein sequencing adaptors are ligated directly to ends created by CRISPR/Cas9 cleavage (Gilpatrick et al., 2020) (**Figure 2.2, 2.3**). This strategy exploits the 5' phosphorylated ends created by Cas9 cleavage for unique ligation of sequencing adaptors. This approach can be applied to either mode of long-read sequencing, but this thesis centers on its use with nanopore sequencing (nCATS).

In addition to molecular biology strategies for regional enrichment, it is worth noting recent publications describing computational-based methods for target enrichment with nanopore sequencing, sometimes referred to as "adaptive sequencing" (Kovaka et al., 2020; Payne et al., 2020). Computational

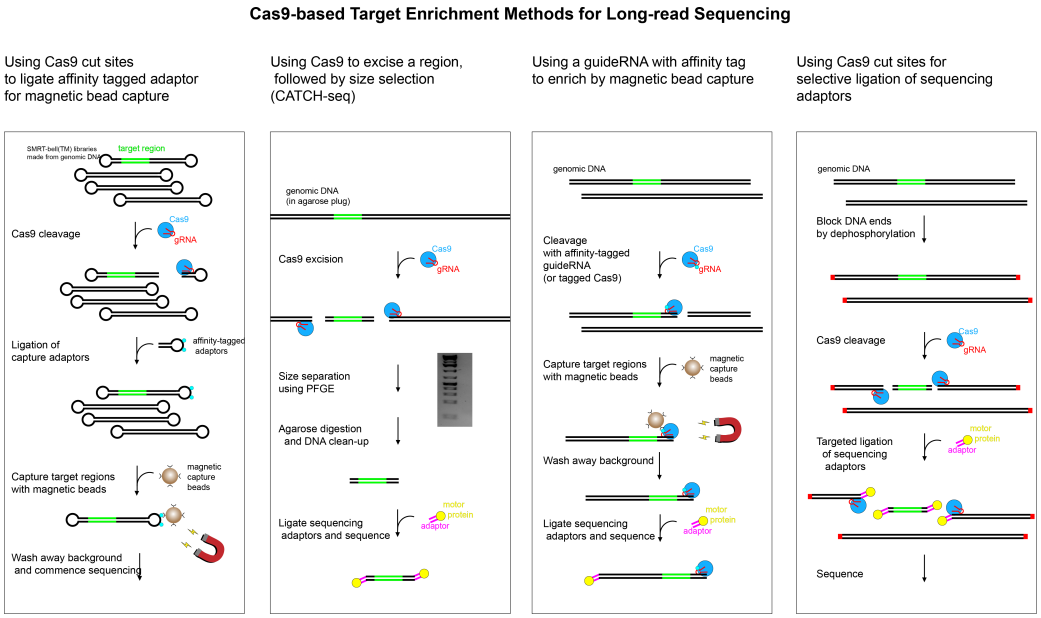


Figure 2.2: Long-read Target Enrichment Summary of target enrichment methods using Cas9 for long-read sequencing (1) Method described in (Tsai et al., 2017) (2) Method described in (Gabrieli et al., 2018) (3) Enrichment using affinity tag on guideRNA (Gilpatrick, unpublished) (4) Method described in (Gilpatrick et al., 2020)

enrichment is done by aligning the nanopore signal while the nucleic acid strand is still translocating through the pore, and deciding either (a) to continue sequencing the strand if it represents desired sequence or (b) reverse the current in the pore and eject the DNA, permitting another molecule to be sequenced. There are some potential advantages to computational enrichment These include (1) no DSBs are introduced, creating less of a restriction in the size of the molecules that can be sequenced and (2) computational enrichment avoids the designing, ordering and testing of guideRNAs, which can be iterative, labor intensive, and expensive. Despite these drawbacks, the nCATS molecular method is still outperforming adaptive sequencing at present, generating about an order of magnitude more coverage (>400X versus 20-30X).

The Cas9 method has been demonstrated to generate coverage greater than 400X at each of 10 sites with sizes ranging from 12 to 24kb (Gilpatrick et al., 2020), and can be used to target up to a 100 sites simultaneously without affecting performance. Enrichment for even larger fragments can be achieved through Cas9 enrichment (we tested up to 84kb), but these larger fragment sizes show a significant drop in coverage, especially at the center of the region due to read drop-off. There have already been numerous applications for this method described including for the study of repeat expansions (Giesselmann et al., 2019), DNA methylation (Giesselmann et al., 2019; Gilpatrick et al., 2020), single nucleotide changes (Gilpatrick et al., 2020), structural variation (Gilpatrick et al., 2020; Watson et al., 2020), and to study off-target cutting activity of Cas9 (Höijer et al., 2020).

2.4 Nanopore Cas9 Targeted Sequencing (nCATS)

A more detailed image of how nCATS this strategy works is shown in (**Figure 2.3**). This method employs the "Sequencing by Ligation" kits from ONT (e.g. the current kit LSK-109), wherein sequencing adaptors are attached to DNA ends via 'TA-ligation'. In this process a 3' (dA) overhang on DNA hybridizes to a 3' (dT) overhang on sequencing adaptors– allowing for a DNA ligase to form phosphodiester bonds between 3' hydroxyl groups and 5' phosphates. With the standard (genomic) sequencing workflow, all DNA ends are bestowed with phosphorylated 5' ends and 3' (dA) overhangs during DNA "end-prep" permitting widespread ligation with the 3' (dT) on adaptors. Critically, in the nCATS approach, all 5' phosphate ends are first removed with a phosphatase, preventing widespread adapter ligation. Next, Cas9 cleavage is used to introduce double strand breaks adjacent to regions of interest. After the universal addition of 3'(dA) tails, this reaction yields sites properly end-prepped for adapter ligation (3'(dA) overhangs and with 5' phosphorylation). Sites of Cas9 cleavage are enriched among these properly end-prepped sites, yielding approximately 5% of sequencing reads originating at sites of Cas9 cleavage.

The CRISPR/Cas9 system natively uses two separate RNA species for targeting DNA in order to introduce a double-strand break (DSB). The DNA-targeting component is called the CRISPR-RNA (crRNA, 40nt), which contains a 20 nucleotide sequence complementary to the target DNA. Another RNA, termed the trans-activating CRISPR RNA (tracrRNA, 75nt) binds to the crRNA, forming a two-RNA "hybrid". The resulting crRNA:tracrRNA

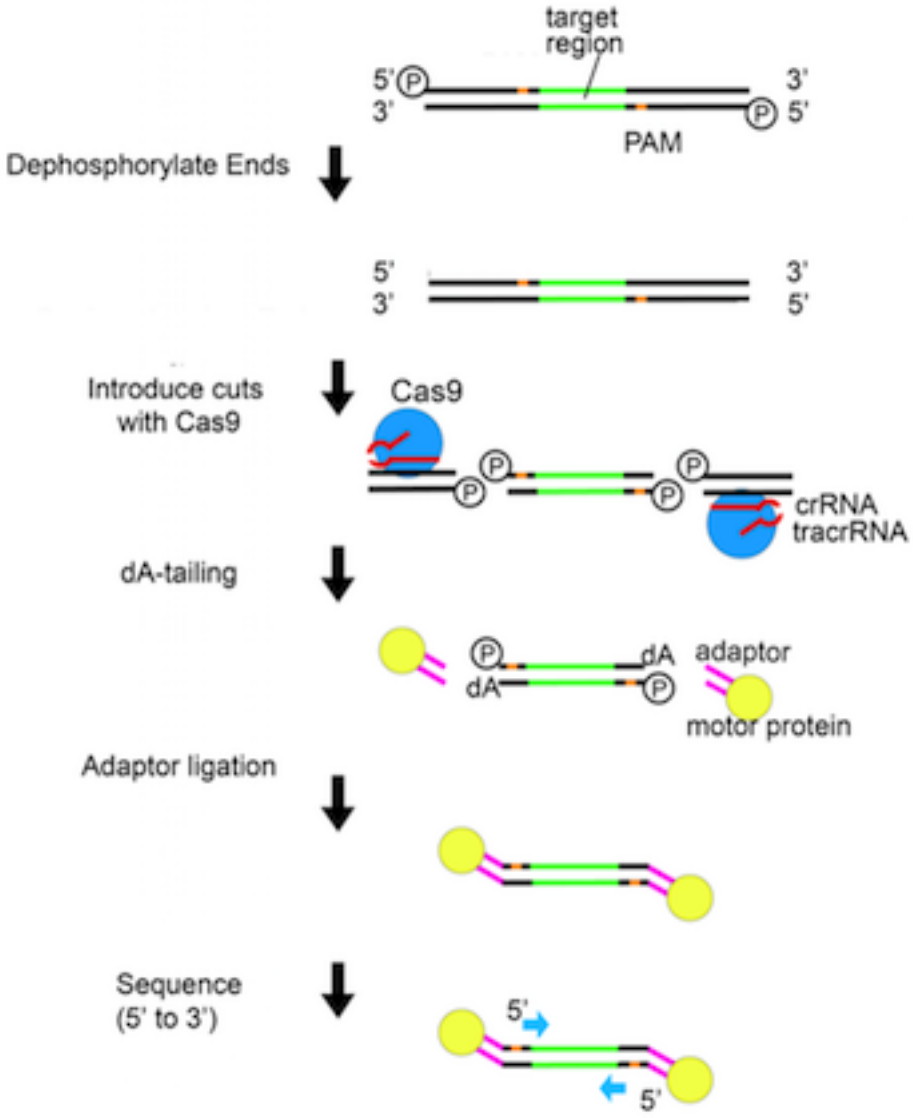


Figure 2.3: Schematic of Cas9-targeted sequencing (nCATS)
Detailed cartoon demonstrating the steps of nCATS enrichment. (1) Dephosphorylation of all 5′ ends (2) Introduction of DSB at select loci using CRISPR/Cas9 (3) 3′(dA) tailing (4) Ligation of sequencing adaptors and motor protein to DNA ends (5) DNA sequencing (5′ to 3′)

hybrid has come to be known as the 'guideRNA'. The guideRNA forms a secondary structure that is recognized by the Cas9 enzyme, forming the ribonucleoprotein complex (RNP). In order for Cas9 to introduce DSBs at the desired targets, there must also be an adjacent 3' nucleotide motif, the so-called "protospacer-adjacent motif" (PAM). For Cas9, this motif is two sequential guanine nucleotides preceded by any nucleotide (NGG). Without this obligate PAM recognition the Cas9-RNA RNP is unable to interrogate the DNA for crRNA complementarity (Sternberg et al., 2014).

During cleavage with the CRISPR/Cas9 system, the Cas9 protein remains bound to the 5' side of the gRNA after DNA cleavage (Sternberg et al., 2014), resulting in preferential ligation of adaptors onto the 3' side of the cut. This means that the orientation of the guideRNA relvant to the target site is important, as nanopore DNA sequencing commences in the 5' > 3' direction (**Figure** 2.3). To sequence both strands of DNA, We therefore flank the target region on both sides with guideRNAs oriented towards the region of interest, ensuring coverage on both DNA strands. We found the coverage could be greatly enhanced by using multiple guideRNAs at each cut site. Having multiple cuts not only increased the chances of having a high-performing guideRNA, but also was observed to have a synergistic cooperation of using multiple RNPs simultaneously cutting within a few hundred nucleotides of one another (Gilpatrick et al., 2020). It is not however, strictly necessary, as we found we were able to enrich to 50X coverage even with a single cut from a single guideRNA. Once targeted studies using the nCATS method were performing consistently, we next executed a number of validation studies to

explore how the data compared with existing data generated both by whole-genome nanopore and NextGen sequencing. In the chapters that follow we delve into the biological insight that was uncovered by the application of this approach, including applications to both clinical diagnostics as well as fundamental research into DNA biology.

References

Pham, Thang T, Jun Yin, John S Eid, Evan Adams, Regina Lam, Stephen W Turner, Erick W Loomis, Jun Yi Wang, Paul J Hagerman, and Jeremiah W Hanes (2016). "Single-locus enrichment without amplification for sequencing and direct detection of epigenetic modifications". en. In: *Mol. Genet. Genomics* 291.3, pp. 1491–1504.

Gabrieli, Tslil, Hila Sharim, Dena Fridman, Nissim Arbib, Yael Michaeli, and Yuval Ebenstein (2018). "Selective nanopore sequencing of human BRCA1 by Cas9-assisted targeting of chromosome segments (CATCH)". en. In: *Nucleic Acids Res.* 46.14, e87.

Giesselmann, Pay, Björn Brändl, Etienne Raimondeau, Rebecca Bowen, Christian Rohrandt, Rashmi Tandon, Helene Kretzmer, Günter Assum, Christina Galonska, Reiner Siebert, Ole Ammerpohl, Andrew Heron, Susanne A Schneider, Julia Ladewig, Philipp Koch, Bernhard M Schuldt, James E Graham, Alexander Meissner, and Franz-Josef Müller (2019). "Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing". en. In: *Nat. Biotechnol.* 37.12, pp. 1478–1481.

Gilpatrick, Timothy, Isac Lee, James E Graham, Etienne Raimondeau, Rebecca Bowen, Andrew Heron, Bradley Downs, Saraswati Sukumar, Fritz J Sedlazeck, and Winston Timp (2020). "Targeted nanopore sequencing with Cas9-guided adapter ligation". en. In: *Nat. Biotechnol.*

Tsai, Yu-Chih, David Greenberg, James Powell, Ida Höijer, Adam Ameur, Maya Strahl, Ethan Ellis, Inger Jonasson, Ricardo Mouro Pinto, Vanessa C Wheeler, Melissa L Smith, Ulf Gyllensten, Robert Sebra, Jonas Korlach, and Tyson A Clark (2017). "Amplification-free, CRISPR-Cas9 Targeted Enrichment and SMRT Sequencing of Repeat-Expansion Disease Causative Genomic Regions". en.

Kozarewa, Iwanka, Javier Armisen, Andrew F Gardner, Barton E Slatko, and C L Hendrickson (2015). "Overview of Target Enrichment Strategies". en. In: *Curr. Protoc. Mol. Biol.* 112, pp. 7.21.1–7.21.23.

Tewhey, Ryan, Masakazu Nakano, Xiaoyun Wang, Carlos Pabón-Peña, Barbara Novak, Angelica Giuffre, Eric Lin, Scott Happe, Doug N Roberts, Emily M LeProust, Eric J Topol, Olivier Harismendy, and Kelly A Frazer (2009). "Enrichment of sequencing targets from the human genome by solution hybridization". en. In: *Genome Biol.* 10.10, R116.

Eckert, Sabine E, Jackie Z-M Chan, Darren Houniet, The Pathseek Consortium, Judy Breuer, and Graham Speight (2016). "Enrichment by hybridisation of long DNA fragments for Nanopore sequencing". en. In: *Microb Genom* 2.9, e000087.

Lee, Isac, Rachael Workman, Josh Zhiyong, and Winston Timp. *Hybridization-capture for nanopore sequencing v1 (protocols.io.zxyf7pw)*.

Simpson, Jared T, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi, and Winston Timp (2017). "Detecting DNA cytosine methylation using nanopore sequencing". en. In: *Nat. Methods* 14.4, pp. 407–410.

Flusberg, Benjamin A, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson A Clark, Jonas Korlach, and Stephen W Turner (2010). "Direct detection of DNA methylation during single-molecule, real-time sequencing". en. In: *Nat. Methods* 7.6, pp. 461–465.

Sternberg, Samuel H, Sy Redding, Martin Jinek, Eric C Greene, and Jennifer A Doudna (2014). "DNA interrogation by the CRISPR RNA-guided endonuclease Cas9". en. In: *Nature* 507.7490, pp. 62–67.

Kovaka, Sam, Yunfan Fan, Bohan Ni, Winston Timp, and Michael C Schatz (2020). "Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED". en.

Payne, Alexander, Nadine Holmes, Thomas Clarke, Rory Munro, Bisrat Debebe, and Matthew Loose (2020). "Nanopore adaptive sequencing for mixed samples, whole exome capture and targeted panels". en.

Watson, Christopher M, Laura A Crinnion, Sarah Hewitt, Jennifer Bates, Rachel Robinson, Ian M Carr, Eamonn Sheridan, Julian Adlard, and David T Bonthron (2020). "Cas9-based enrichment and single-molecule sequencing for precise characterization of genomic duplications". en. In: *Lab. Invest.* 100.1, pp. 135–146.

Höijer, Ida, Josefin Johansson, Sanna Gudmundsson, Chen-Shan Chin, Ignas Bunikis, Susana Häggqvist, Anastasia Emmanouilidou, Maria Wilbe, Marcel den Hoed, Marie-Louise Bondeson, Lars Feuk, Ulf Gyllensten, and Adam Ameur (2020). "Amplification-free long read sequencing reveals unforeseen CRISPR-Cas9 off-target activity". en.

Chapter 3

Targeted Sequencing in Breast Cancer

3.1 Introduction

The Cas9-based targeted nanopore sequencing strategy described in the previous chapter was first validated and tested by comparing enrichment data to existing data on the GM12878 cell line. The GM12878 lymphoblast cell line has been extensively characterized for numerous features, saliently including annotated variants (Eberle et al., 2016; Zook et al., 2016) and whole-genome bisulfite methylation data (ENCODE Project Consortium, 2012).

We then applied these strategies to assess genetic and epigenetic changes in breast cell lines, a breast cancer cell line xenograft, and primary patient tissue. For cell lines, we used three breast epithelial cell lines: the non-tumorigenic MCF-10A, the ER(+) line MCF-7, and the triple-negative breast cell line MDA-MB-231. We selected three aberrations known to occur in neoplasia for further investigation: changes in DNA methylation, point mutations, and structural variations. We targeted ten genomic loci in our initial panel, with sizes ranging

from 12-24kb (**Table 3.1**). For evaluating single nucleotide mutations we selected three cancer-associated genes (TP53, KRAS, and BRAF) with annotated mutations in the MDA-MB-231 cell line (Forbes et al., 2017). Five regions for methylation studies (KRT19, SLC12A4, GSTP1, TPM2, and GPX1) were selected by identifying sites with differential methylation (Lee et al., 2018) in these breast cell lines. The whole-genome nanopore data (Lee et al., 2018) was also used to identify two candidate large deletions (6-8kb) for study. An 11th region, the BRCA1 locus, was included in later sequencing runs (GM12878, and primary patient samples) to test our ability to capture larger regions (>80kb), and to evaluate this method for sequencing highly repetitive regions (Welcsh and King, 2001).

Location/Gene - [Aberration]	size(kb)
GPX1 - [DNA methylation]	13.6
GSTP1 - [DNA methylation]	17.8
KRT19 - [DNA methylation]	18.1
SLC12A4 - [DNA methylation]	24.4
TPM2 - [DNA methylation]	19.6
chr5 deletion - [Structural variant]	18.7
chr7 deletion - [Structural variant]	20.0
BRAF - [Single Nucleotide Variant]	12.3
KRAS - [Single Nucleotide Variant]	16.7
TP53 - [Single Nucleotide Variant]	16.1

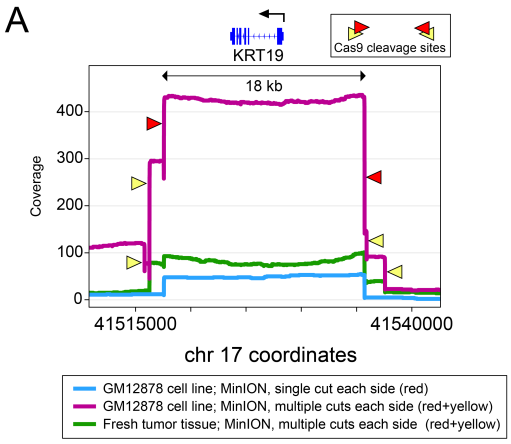
Table 3.1: Regions for Breast Cancer Studies Type of aberration and size for the initial panel of 10 loci targeted

3.2 Results

In our initial experiments, we used a single guideRNA flanking each site in each of the four cell lines (GM12878, MCF-10A, MCF-7 and MDA-MB-231).

Targeted libraries were prepared from 3ug of starting DNA and each sample run on a separate minION flow cell. This resulted in coverage ranging from 18X to 846X (**Supplementary Table** 3.2). We attributed this highly variable coverage between regions to differing on-target cutting efficiency and off-target binding of the guideRNAs. Subsequently, we experimented with a combination of multiple guideRNAs flanking each locus and found this significantly improved median coverage. For example, at the KRT19 locus, with multiple guides the coverage increased to 407X versus 47X with single guides (**Figure** 3.1). Using multiple guides at all loci yielded a median regional coverage of 680X (**Figure** 3.1), and greater than 400X at all sites when using cell line DNA (**Supplementary Table** 3.2).

From GM12878 MinION data, the percentage of 'on-target' reads, was 1.8% with a single guideRNA cut flanking each site per site and 4.6% with the multi-guideRNA panel (**Figure** 3.1). Genome-wide coverage analysis found the off-target reads to be distributed randomly across the genome, indicating they result primarily from ligation of nanopore adaptors to random breakage points. For example, in the GM12878 cell line with single guideRNAs flanking each site, after quality filtering alignments (MAPQ > 30) there were only 2 genomic sites outside target regions where coverage reached 25X. Both of these are at repetitive peri-centromeric sites and contain reads with lower mapping quality (MAPQ 30-50), suggesting the increased coverage to result from alignment errors in these poorly mappable regions. We did note the occurrence of some off-target cleaving with the inclusion of guideRNAs designed to flank the BRCA1 locus (**Supplementary Table** 3.3), which we



B

10 sites (12-24kb)

	GM12878		Breast Tumor
	one cut each side	multiple cuts (2 or 3) each side	
	minION	minION	minION
aligned reads	75822	235328	611992
on-target reads	1370	10854	2702
on-target percent	1.81%	4.61%	0.44%
median coverage	80	680	93

Figure 3.1: On-target Performance Summary A. Local coverage at the KRT19 gene using both single and multiple cuts in GM12878 cell line DNA, and using multiple cuts in primary patient tissue derived DNA. B. Total reads and on-target rate for the same samples described in (A).

attribute to the abundance of repetitive regions (Welcsh and King, 2001) at this locus resulting in increased homology with other genomic loci. Investigation of one off-target site found it highly resembled one of the BRCA1 guideRNAs (**Supplementary Figure** 3.5).

With this panel of guides in hand, we tested the assay's performance in tissue samples: normal human breast tissue, a breast cancer cell-line-derived xenograft, and a human breast tumor/normal pair. In tissue from a reduction mammoplasty (normal) and cell-line-derived mouse xenograft we measured a median coverage of 162X/312X; and from the paired primary tumor/normal sample with limited input we achieved median coverage of 93X/70X (**Figure** 3.1, and **Supplementary Table** 3.2).

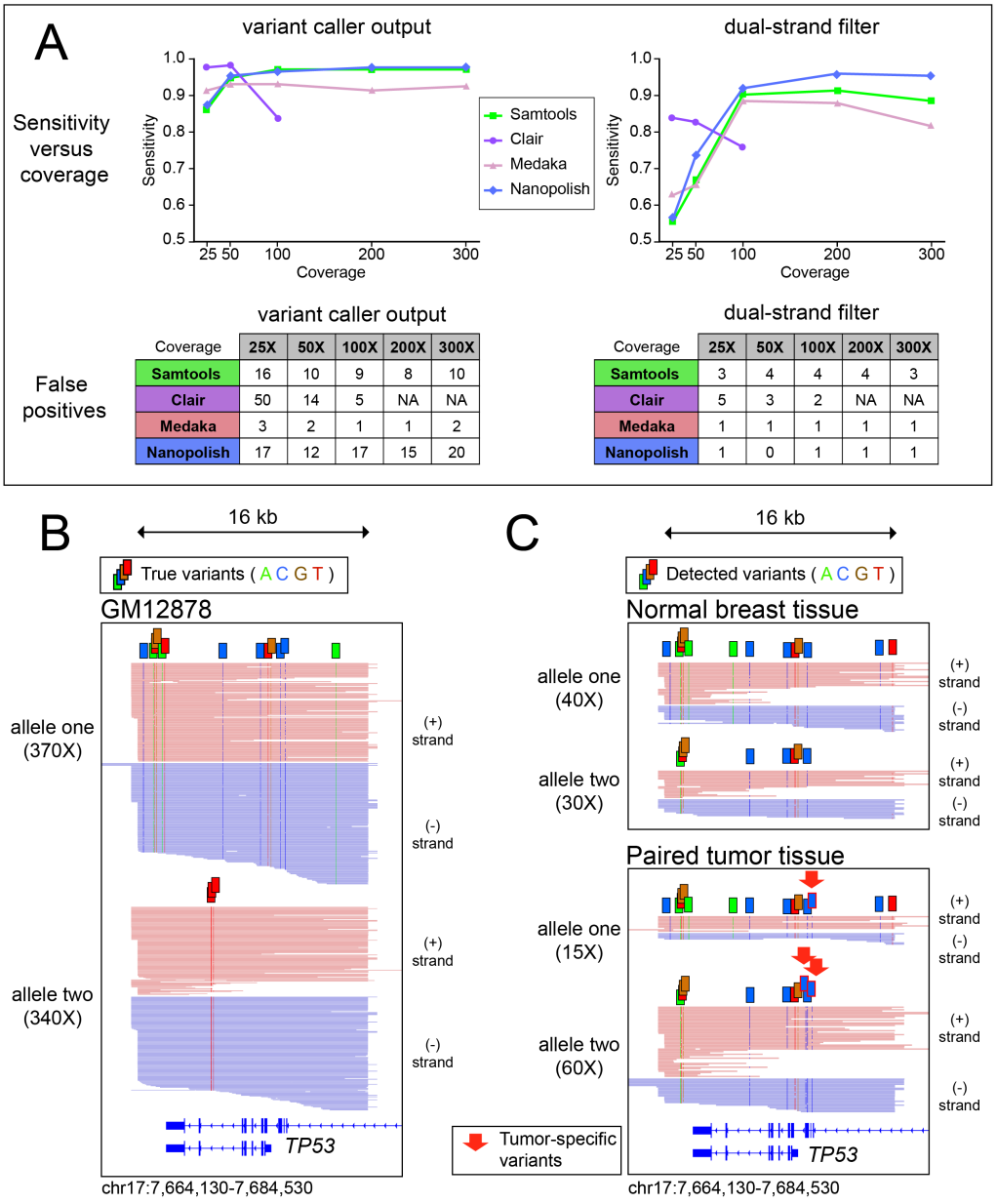
3.2.1 Single Nucleotide Variant Detection

Nanopore sequencing still has intrinsically high error rates (5-10%) due to the inability of the basecaller to distinguish between some k-mers and the difficulty in discriminating signal events in repetitive regions (e.g. homopolymers). We applied the high-coverage data achieved through the nCATS protocol to explore how this affected the ability to call variants from nanopore data. To simplify analysis, we limited this to the evaluation of single nucleotide substitutions. There are numerous tools that currently exist for calling variants, and we selected four for comparison: (1) the Samtools/Bcftools package (Li, 2011), which generates genotype likelihoods from alignment data (2) Clair (Luo et al., 2019), which uses a deep neural network for variant calling from alignment data, (3) Medaka(ONT), a tool from Oxford Nanopore which also uses a neural network algorithm, and (4) Nanopolish (Simpson et al., 2017), which uses a hidden Markov model to interrogate the raw electrical data as well as alignment data.

For initial validation, we used the GM12878 cell line and the platinum genome dataset (Eberle et al., 2016) as ground truth for single nucleotide variants (SNVs). We benchmarked SNVs over the 8 enriched loci without large deletions (total size of 140kb) wherein a total of 174 annotated SNVs are annotated in GM12878. To explore the relationship between coverage and variant calling efficiency, we subsampled the aligned data to coverage of 300X, 200X, 100X, 50X and 25X (see methods). During filtering we selected for reads spanning the region, and maintained balanced coverage between both DNA strands.

We found that at lower coverage data (25X and 50X) Clair had the greatest sensitivity (0.98). However, the current model for Clair was trained and assessed on whole genome data only up to 100X coverage; and above this coverage it no longer functioned. Medaka showed peak sensitivity of 0.93 at both 50X and 100X coverage, with sensitivity remaining robust at higher coverage. Samtools variant calling and Nanopolish variant calling both increased in sensitivity up to 200X coverage, at which point they plateaued with sensitivities of 0.97 and 0.98, respectively (**Figure** 3.2).

One important caveat of the raw output of these variant caller pipelines is the persistence of false positives, limiting the use of this method for *de novo* SNV discovery. On inspection, we noted many false positives to occur on only one strand (**Supplementary Figure** 3.6), suggesting the basecaller having systematic issues with the sequence of k-mers on one strand but not on the other. Thus, we implemented a filter requiring variants to be supported by reads from both strands("dual-strand filter"). This filter caused a decrease in sensitivity, especially at lower coverage. But strikingly this filter eliminated nearly all false positive variant calls (**Figure** 3.2), yielding a set of high-confidence variants. The dual-strand filter performed best with 200X coverage using nanopolish variant calling (Sensitivity: 0.96, F1score: 0.97), with the sole false positive variant existing in a thymidine-dense homopolymer region (**Supplementary Figure** 3.7). We then applied WhatsHap (Patterson et al., 2015), a weighted haplotype assembler that uses statistical information as well as coverage depth to assign reads into parental haplotypes using SNVs detected in long-read data. A graphical depiction of detected variants is



Coverage	25X	50X	100X	200X	300X
Samtools	16	10	9	8	10
Clair	50	14	5	NA	NA
Medaka	3	2	1	1	2
Nanopolish	17	12	17	15	20

Coverage	25X	50X	100X	200X	300X
Samtools	3	4	4	4	3
Clair	5	3	2	NA	NA
Medaka	1	1	1	1	1
Nanopolish	1	0	1	1	1

Figure 3.2: Single Nucleotide Variants (A) Plot of sensitivity versus coverage using four tools to call single nucleotide variants from enrichment data in GM12878 for a 140kb region containing 174 annotated SNVs (B) Visual representation of high-confidence variants detected by nanopolish in the MinION data from GM12878 for the captured region around TP53, reads phased into homologous alleles using WhatsHap. (C) High-confidence variants identified in primary tissue from a tumor/normal pair, red arrows used to demarcate tumor-specific variants.

shown in **Figure** 3.2B, highlighting the identification and phasing of variants in the captured region of TP53 in GM12878. All 17 of the annotated SNVs were detected in this region with no false positives. We then applied this variant caller pipeline to our data from the MDA-MB-231 cell line to detect cancer-associated mutations. Across the captured regions of three cancer-associated genes (BRAF, KRAS, and TP53) nanopolish called 42 high-confidence SNVs (**Supplementary Table** 3.4), including 2 of the 3 annotated in the COSMIC database for MDA-MB-231 (Tate et al., 2019) (the third variant was detected, but at a lower frequency in this aneuploid line and thereby did not pass dual-strand filtering). We applied this variant calling pipeline to a paired tumor/normal breast tissue sample, and phased the reads into haplotypes with WhatsHap (Patterson et al., 2015). We noted the presence of tumor-specific variants that were identified through this approach, as well as strong variation in the number of reads per haplotype in the TP53 region implying an imbalanced copy number in tumor cells (**Figure** 3.2C). We captured two other regions on the same chromosome and observed similar chromosomal imbalance with additional mutations in the tumor sample (**Supplementary Figure** 3.8).

3.2.2 CpG Methylation

We next evaluated CpG methylation, which can be measured from the electrical data produced by nanopore sequencing (Simpson et al., 2017). As previously mentioned, for all of the sites studied for methylation, there was pre-existing data suggesting either differential methylation in the breast cell

lines being studied (Lee et al., 2018). Differentially methylated regions were prioritized if there was evidence of changes in transcript levels in existing RNA-seq data (Messier et al., 2016), and further filtered to select genes with prognostic implications in human cancer (Kabir, Rönnstrand, and Kazi, 2014; Martignano et al., 2016; Wang et al., 2018).

One way to visualize the methylation data while maintaining the single-read-level information is to use read-level plots. Methylation data for one locus (KRT19) is shown in **Figure 3.3A**, with four additional genes (GSTP1, GPX1, SLC12A4, and TPM2) plotted in **Supplementary Figure 3.9**. We compared nanopore methylation patterns with existing whole genome bisulfite sequencing (WGBS) data in GM12878 (ENCODE Project Consortium, 2012) using smoothed (loess) line plots (**Figure 3.3B**, and **Supplementary Figure 3.9**). Directly comparing per-CpG methylation **Supplementary Figure 3.9C** at each locus, we observed per-CpG methylation largely clustered at points reflecting completely methylated or unmethylated sites, with an aggregate per-CpG correlation of 0.81 (Pearson). We applied this strategy to our data from breast cell lines, looking for regions with differential methylation. One gene that showed clear differences is the keratin family member gene: KRT19. KRT19 is known to be upregulated in breast cancer (Kabir, Rönnstrand, and Kazi, 2014), and detection of KRT19 mRNA has been used to identify micrometastasis of breast cancer to lymph nodes (Noguchi et al., 1996) and to detect circulating tumor cells (Wang et al., 2018). We observed that KRT19 remains largely methylated in the non-tumorigenic MCF-10A cell line, but becomes hypomethylated in both of the transformed cell lines, MCF-7 and MDA-MB-231 (**Supplementary**

Figure 3.10). This is correlated with the observed increased transcript level for KRT19 in the transformed cell lines (**Supplementary Figure 3.11**, GEO: GSE75168). Further, we note the pattern of methylation in MDA-MB-231 is largely maintained in mouse xenografts that are derived from this cell line (**Supplementary Figure 3.10** and **Supplementary Figure 3.12**). One unanticipated result was in our evaluation of the paired tumor/normal patient sample, where we found that the primary patient tumor had a dramatic allele-specific hypomethylation of KRT19 on the haplotype with increased copy number (**Figure 3.3C**, and **Supplementary Figure 3.12**). This suggests a possible mechanism for the increased expression in tumor cells (Kabir, Rönnstrand, and Kazi, 2014; Noguchi et al., 1996; Wang et al., 2018), and highlights a benefit of this approach as allele-specific methylation would have been difficult to evaluate without the enhanced ability to phase haplotypes provided by long-read sequencing.

3.2.3 Structural Variants

We next applied this method to evaluate structural variations by confirming the presence of candidate deletions identified in whole genome nanopore sequencing data (Lee et al., 2018). We selected two deletions present in the MDA-MB-231 and MCF-7 breast cancer lines and absent in the MCF-10A cell line, and designed guideRNAs to flank breakpoints by 5kb. Plotting reads in IGV showed both deletions as heterozygous in MDA-MB-231 and homozygous in MCF-7 (**Figure 3.4A** and **Supplementary Figure 3.13**). The alignment data was passed to the "Sniffles" variant caller (Sedlazeck et al., 2018), which

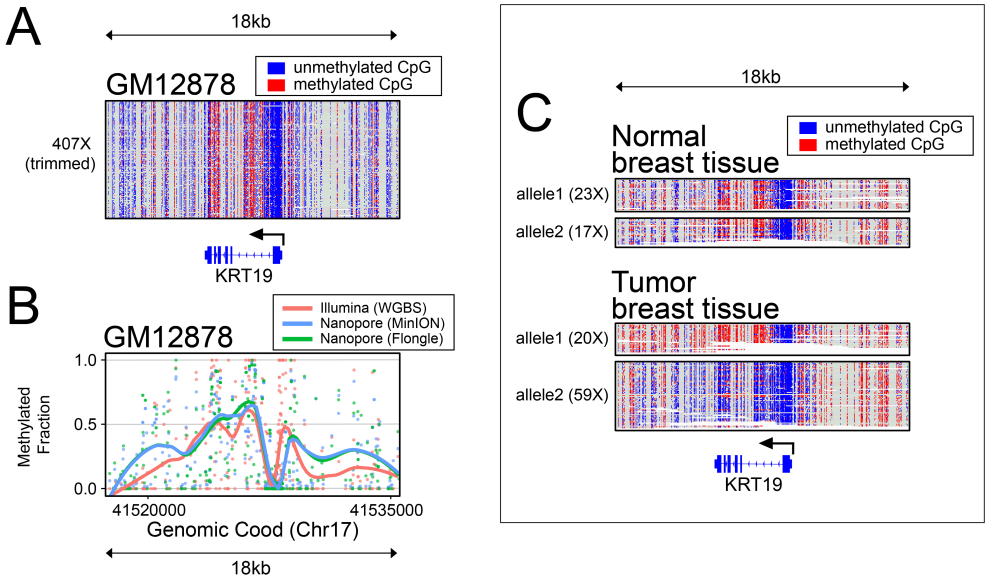


Figure 3.3: Methylation Analysis (A) Read-level plots showing methylation patterns in GM12878 from minION data at the KRT19 locus. (B) Methylation calls (points) and line plots at the same locus as in (A) showing smoothed (loess) methylation calls from whole genome bisulfite sequencing on the Illumina platform (GEO: GSE86765), compared with methylation calls from minION and flongle targeted nanopore sequencing. (C) Haplotype phased methylation calls in primary patient tissue and paired tumor at the KRT19 locus.

identified the breakpoints and zygosity of both deletions (**Supplementary Table 3.5**). We also performed methylation studies on these regions but did not note any difference in methylation patterns between the deleted and not-deleted allele. To explore the use of this method for targeting larger fragments of DNA, we enriched for regions harboring large (>70kb) heterozygous chromosomal deletions. We identified three large heterozygous deletions in GM12878 from available 10X Genomics data through the Genome In a Bottle (GIAB) Consortium project (Zook et al., 2016). Two heterozygous deletions with sizes of 70kb and one 155kb. GuideRNAs were designed to flank the deletion breakpoints by 5kb, resulting in reads of 10 kb on the deleted allele, and spanning the region between cut sites (80kb/165kb) on the non-deleted allele. We phased the reads into parental alleles using WhatsHap (*Whatshap: fast and accurate read-based phasing. bioRxiv. 2016*) and compared read lengths and read counts achieved from each allele. Interestingly, we found that the allele containing the deletion, with the correspondingly shorter distance between the cut sites, demonstrated an order of magnitude higher number of reads (**Figure 3.4B** and **Supplementary Figure 3.14**). This reflects a bias against achieving reads >50kb, likely introduced during DNA purification, library preparation, or delivery to the pore. To confirm this size-bias, we performed similar parental-allele segregation on sites without SVs and did not observe bias towards either parental allele. The alignment data for GM12878 was passed to the Sniffles variant caller (Sedlazeck et al., 2018), which identified all 3 of the deletions within 10nt from the annotated breakpoints in existing GIAB data (**Supplementary Table 3.6**). We adjusted Sniffles parameters to call SVs as heterozygous if an allele was supported by even a very low amount (0.1%)

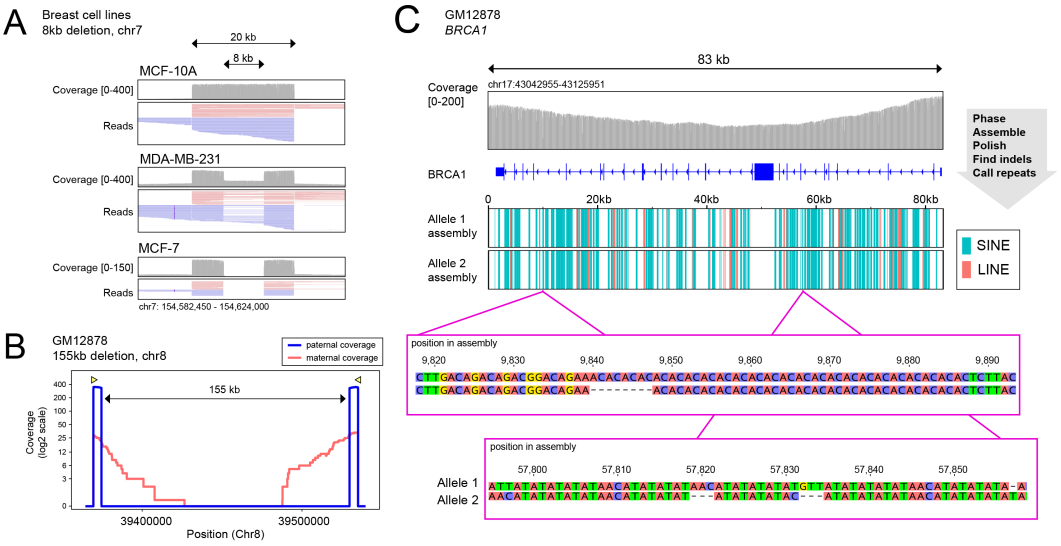


Figure 3.4: Structural Variation (A) Reads around an 8kb deletion in chromosome 7 present in MCF-7 and MDA-MB-231, and absent in MCF-10A. (B) Coverage on each parental allele in the region of a large (155kb) heterozygous deletion in GM12878. (C) Top: Coverage at the BRCA1 locus from DNA extracted using Circulomics CBB kit. Middle: LINE and SINE components identified by RepeatMasker on each of the BRCA1 allele assemblies. Bottom: Three indels discovered between BRCA1 assemblies not annotated in platinum genome data set for GM12878.

of reads, as the imbalance of reads from the two alleles caused the software to initially identify these deletions as homozygous (see Methods).

Finally, we targeted the BRCA1 gene, because of the well-documented association of this gene with familial breast cancer (Welcsh and King, 2001). BRCA1 is also an attractive target for long-read sequencing because of the abundance of hard to map repetitive Alu elements (Deininger, 2011). To capture the entire BRCA1 gene (distance between flanking guideRNAs: 84kb) further adjustments to DNA extraction were needed. Initial MinION sequencing runs from 3ug extracted GM12878 gDNA resulted in only 10 sequencing reads spanning the entire region with many smaller fragments (**Supplementary Figure 3.15**). We found that using the Circulomics NanoBind kit for DNA

extraction resulted in an increase to nearly 30 reads completely spanning the BRCA1 gene, with coverage between guideRNAs ranging between 100X and 200X (**Figure 3.4C** and **Supplementary Figure 3.15**). We phased the BRCA1 reads into two alleles using WhatsHap (*Whatshap: fast and accurate read-based phasing. bioRxiv. 2016*) with *de novo* called high-confidence variants found with nanopolish. We then built an assembly for each of the two alleles using the Flye assembler (Kolmogorov et al., 2019) and polished the assemblies using Racon (Vaser et al., 2017) and Medaka(ONT) (see Methods). This resulted in full length assemblies for each of the two alleles, within which we identified the presence of SINEs (e.g. Alu-elements) and LINEs using RepeatMasker (*RepeatMasker Open-4.0*). We compared the two assemblies for variants differing between alleles and found numerous indels and single nucleotide changes (**Supplementary Table 3.7**) using the minimap2 suite (Li, 2018). After filtering for homopolymer regions (which nanopore sequencing is known to still have difficulty resolving (Simpson et al., 2017)), we found 10 indels of at least 3 nucleotides between the two assemblies (**Supplementary Table 3.7**). Seven of these ten were annotated in the platinum genome data set for GM12878 (Eberle et al., 2016). The remaining three not-annotated indels are within large repetitive regions (**Figure 3.4C**), making it difficult to map reads from conventional short read sequencing. To validate these indels, we compared against recently released whole genome PacBio data for GM12878 (SRA: SRR9001768-SRR9001773) (Zook et al., 2016) which confirmed each of these three small unannotated indels between alleles (**Supplementary Figure 3.16**).

3.3 Discussion

Because of the low cost to entry and small footprint of the instrument, and the ability to sequence targeted regions of the genome with long-reads, this assay has the potential to be widely utilized as a tool for identifying single nucleotide changes, evaluating DNA methylation, and studying structural variation. We were even able to apply this to clinical tissue despite the relatively high DNA input requirements (3 micrograms). We show that single nucleotide variants in regions of interest can be queried with the nCATS protocol, although there are persisting limitations, as evinced by the few SNVs not detected by this approach. We found that by using only high-confidence variants, we were able to phase nanopore sequencing reads into parental alleles using WhatsHap (*Whatshap: fast and accurate read-based phasing. bioRxiv. 2016*), permitting haplotype resolution of high-coverage nanopore data. As basecalling and variant-calling algorithms continue to improve we anticipate higher future performance for surveillance and identification of mutations. We also highlight the use of nCATS to detect and validate structural variants. It is only with the advent of long-read sequencing that the great diversity of structural variation in human genomes has been appreciated (Audano et al., 2019; Chaisson et al., 2018), and this method provides a dynamic approach to evaluate genomic rearrangements, including large structural variants and hard-to-map repetitive regions (Dixon et al., 2018). Importantly, because nanopore sequencing interrogates the DNA strand rather than sequencing "by-synthesis", we can simultaneously profile methylation in these loci, providing biological as well as diagnostic insight into the epigenome, which is

commonly disrupted in human neoplasia (Timp and Feinberg, 2013). The high sequencing depth granted by this method is especially useful to characterize genetically and epigenetically heterogeneous samples typically obtained from clinical samples; giving us insight into the frequency of different mutations and epigenetic changes present.

3.4 Methods

3.4.1 Cell culture and DNA prep

Cell lines were obtained from ATCC: MCF10A (CRL-10317), MCF7 (HTB-22), MDA-MB-231 (HTB-26); or Coriell institute: CEPH/UTAH Pedigree 1463 (GM12878). Cells were cultured according to recommended protocols. Briefly, all cell lines were maintained at 37°C in 5% CO2. The GM12878 cell line was grown in high-glucose RPMI media supplemented with 10% fetal calf serum (FCS), penicillin-streptomycin antibiotics (pen-strep), and L-glutamine. MCF-7 and MDA-MB-231 were grown hi-glucose DMEM media supplemented with 10% FCS, pen-strep, and L-glutamine. MCF-10A cells were grown in hi-glucose DMEM media supplemented with 5% horse serum, pen-strep, L-glutamine, epidermal growth factor, insulin, hydrocortisone, and cholera toxin. DNA was extracted from cells, using either the MasterPure kit (Lucigen, MC85200), or the Nanobind kit (Circulomics, NB-900-001-01) and stored at 4°C until use. DNA was quantified using the Qubit fluorometer (Thermo) immediately before performing the assay.

3.4.2 Patient Tissue and Mouse Xenograft

All human samples were collected with appropriate approval from the Johns Hopkins institutional review board. The primary breast tumor was identified as ER/PR+ by immunohistochemistry and snap frozen. Mouse experiments were conducted with prior approval from JH-IACUC. Mouse xenografts were generated by injecting 106 ER/PR/HER2-negative MDA-MB-231 breast cancer cells into the mammary fat pad of athymic mice. Tumors were collected 6-8 weeks later and frozen immediately as small chunks. The snap frozen tissue was ground under liquid nitrogen using a CryoMill (Retch) and DNA extracted using MasterPure kit (Lucigen, MC85200).

3.4.3 GuideRNA design

Guide RNAs were assembled as a duplex from synthetic crRNAs (IDT, custom designed) and tracrRNAs (IDT, 1072532). Sequences are provided in **Supplementary Table 3.8**. The crRNAs were designed using IDT's design tool and selected for the highest predicted on-target performance with minimal off-target activity. The gRNA duplex was designed to introduce cuts on complementary strands flanking the region of interest. For methylation studies and SNV studies, the target size between gRNAs was 12-24 kb; for deletions, the gRNAs were designed to flank the suspected breakpoints by 5kb.

3.4.4 Ribonucleoprotein Complex Assembly

Prior to guide RNA assembly, all crRNAs were pooled into an equimolar mix, with a total concentration of 100uM. The crRNA mix and tracrRNA were

then combined such that the tracrRNA concentration and total crRNA concentration were both 10uM. The gRNA duplexes were formed by denaturation for 5 minutes at 95°C, then allowed to cool to room temp for 5 minutes on a benchtop. Ribonucleoprotein complexes (RNPs) were constructed by combining 10pmol of gRNA duplexes with 10pmol of HiFi Cas9 Nuclease V3 (IDT, 1081060) in 1X CutSmart Buffer (NEB, B7204) at a final volume of 30uL (conc: 333nM), incubated 20 minutes at room temperature, then stored at 4°C until use, up to 2 days.

3.4.5 Cas9 Cleavage and Library Prep

3ug of input DNA was resuspended in 30uL of 1X CutSmart buffer (NEB, B7204), and dephosphorylated with 3uL of Quick CIP enzyme (NEB, M0508) for 10 min at 37C, followed by heating for 2 minutes at 80C for CIP enzyme inactivation. After allowing the sample to return to room temp, 10uL of the pre-assembled 333nM Cas9/gRNA complex was added to the sample. In the same tube, 1uL of 10mM dATP (Zymo, D1005) and 1uL of Taq DNA polymerase (NEB, M0267) were added for A-tailing of DNA ends. The sample was then incubated at 37C for 20min for Cas9 cleavage followed by 5 minutes at 72C for A-tailing. Sequencing adaptors and ligation buffer from the Oxford Nanopore Ligation Sequencing Kit (ONT, LSK109) were ligated to DNA ends using Quick Ligase (NEB, M2200) for 10 min at room temp. The sample was cleaned up using 0.3X Ampure XP beads (Beckman Coulter, A63881), washing twice on a magnetic rack with the long-fragment buffer (ONT, LSK109) before eluting in 15uL of elution buffer (ONT, LSK109). Sequencing libraries were

prepared by adding the following to the eluate: 25uL sequencing buffer (ONT, LSK109), 9.5uL loading beads (ONT, LSK109), and 0.5uL sequencing tether (ONT, LSK109). A detailed step-wise description of the enrichment method is available on protocols.io (https://www.protocols.io/view/cas9-enrichment-for-nanopore-sequencing-68ihhue)

3.4.6 Sequencing

Samples were run on a MinION (ver 9.4.1) flow cell or Flongle flow cell (ver 9.4.1 pore), using the MK1B or GridION sequencer.

3.4.7 Data Analysis

Code and pipelines used in data analysis is available online at

```
https://github.com/timplab/Cas9Enrichment
```

Basecalling was performed using the GUPPY algorithm (Version 3.0.3) to generate FASTQ sequencing reads from electrical data. Reads were aligned to the human reference genome (Hg38) using Minimap2 (Li, 2018). Per-nucleotide coverage was determined using samtools, and clustered using the 'bincov' script of the SURVIVOR software package (Jeffares et al., 2017). On-target reads were defined as those which aligned within 20kb of a guideRNA site. Average coverage per region is the average of coverage of all bases between the innermost guideRNA sites, using coverage found by samtools. De novo variant calling was performed using samtools (Li, 2011), Clair (Luo et al., 2019), Medaka (*medaka: Sequence correction provided by ONT Research*) or nanopolish (Simpson et al., 2017). For validation, we compared SNV calls to

those annotated for GM12878 as part of the platinum genome dataset (Eberle et al., 2016). To achieve different coverage values for validation of GM12878 data, each region was subsampled at random using samtools to achieve 300X coverage with reads balanced on each strand. The reads were then further subsampled to achieve the lower coverage values of 200X, 100X, 50X and 25X. Sensitivity was calculated as correctly called SNVs (true positives) out of all true SNVs (true positives plus false negatives). The F1 score is included as a measure of overall test accuracy, calculated as the harmonic mean of precision and recall. High-confidence variants were generated by an additional filter requiring variants to be supported by reads from both strands. Bam alignment files were split into reads aligning to forward strand and reverse strand, and variant calls performed were performed on each set of reads separately. Variants were only included in the high-confidence set if they were called in forward strand reads alone, reverse strand reads alone, and the complete data set. Segregation of reads into parental alleles was performed with WhatsHap (Patterson et al., 2015), using only de novo called high-confidence variants. For patient tumor tissue, reads phased into haplotypes using only the variants identified from paired normal tissue. CpG methylation calling on nanopore data was performed using nanopolish (Simpson et al., 2017). Methylation calling on existing WGBS GM12878 data (GEO: GSE86765) (ENCODE Project Consortium, 2012) was performed using the bismark software tool (Krueger and Andrews, 2011). The bismark output files were processed using the bsseq R package (Hansen, Langmead, and Irizarry, 2012), and Pearson correlation coefficient was calculated using base R. RNA-seq data of MCF-10A, MCF-7, and MDA-MB-231 were downloaded from GEO (Accession: GSE75168) in

the form of RNA counts. Deletions were called using the structural variant caller Sniffles (Sedlazeck et al., 2018), set to find deletions with a minimum size of 100bp. In the instance of the very large (>70kb) heterozygous deletions in GM12878, the allelic size bias caused the ploidy to be incorrectly called as homozygous. To correct this, we used the option

`"--min_homo_af"`

set to 99.9, which ensured a deletion was called as heterozygous if supporting reads for an allele were present at a rate as low as one in one thousand. For assembly of the BRCA1 region, reads were first split into haplotypes with WhatsHap (Patterson et al., 2015). A draft assembly for each allele was built using the Flye assembly tool (Kolmogorov et al., 2019), with default parameters for nanopore reads. Draft assemblies were then corrected by using four iterative rounds of polishing with the Racon error-correction software (Vaser et al., 2017), with the score for matching bases ("-m") increased to 8 and the score for mismatching bases ("-x") decreased to -6. A final round of polishing was performed using the Medaka consensus tool (ONT) with default parameters. The assemblies were surveilled for indels using the paftools helper script of the Minimap2 suite (Li, 2018).

3.5 Supplementary Material

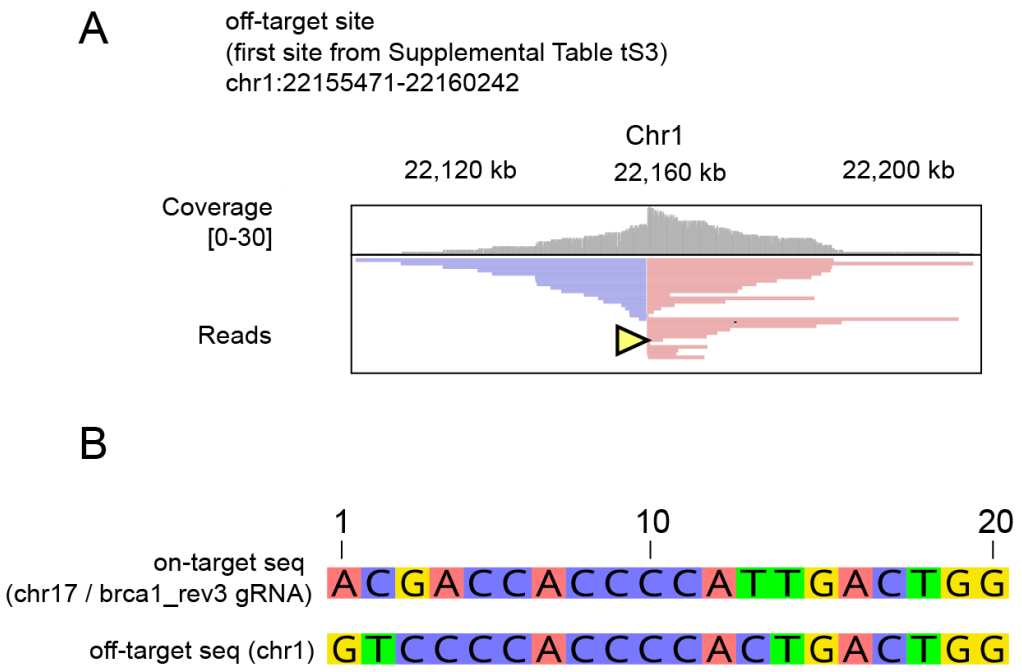


Figure 3.5: Example Off-Target Cleaving from BRCA1 guideRNA (Top) Off-target coverage/reads. (Bottom) pairwise alignment between guideRNA and off-target site.

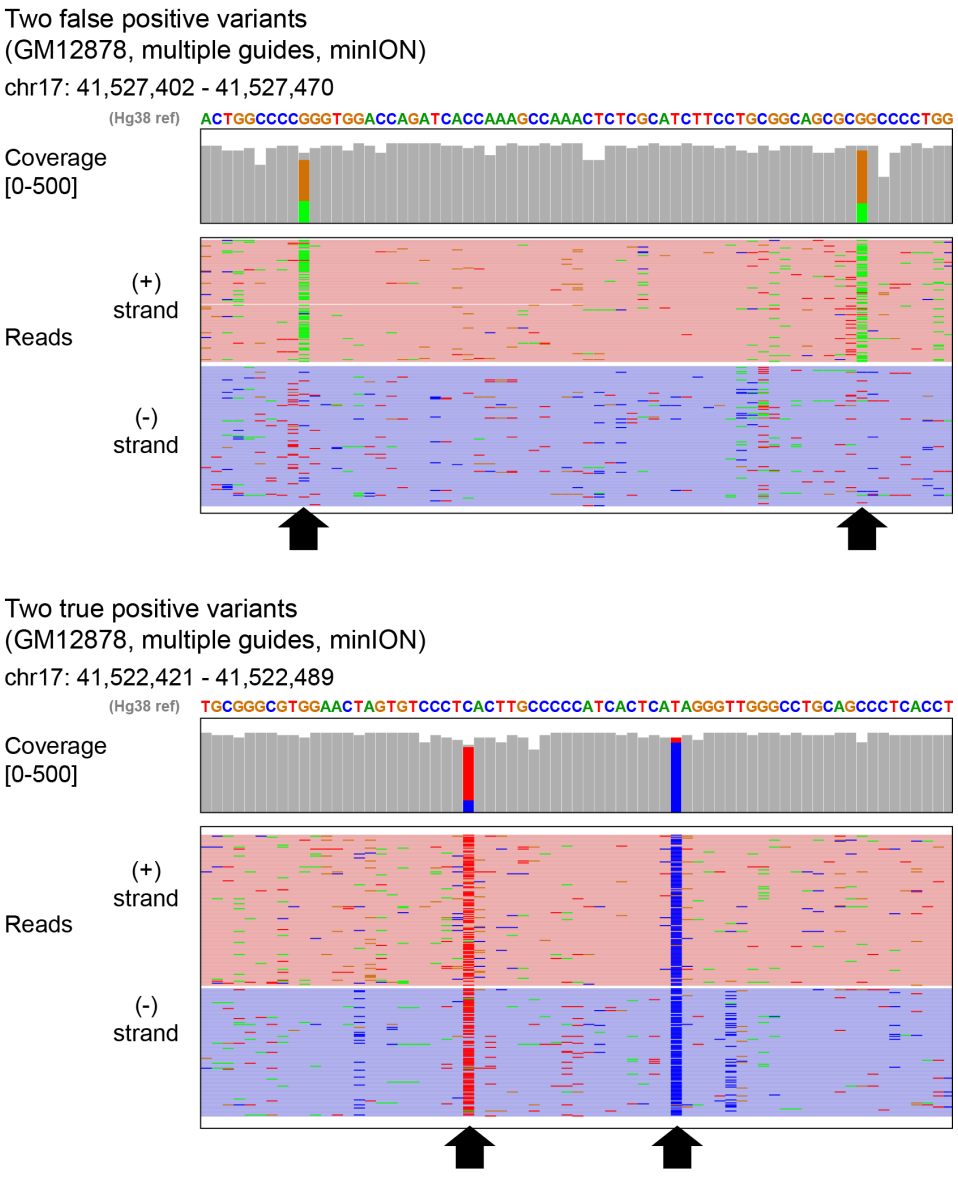


Figure 3.6: Stranded Information (Top) Example of two false positive variants resulting from a sequencing error on only one strand. (Bottom) Two real variants which are supported by data on both strands.

SINGLE CUT EACH SIDE

		GM12878			MCF-10A			MDA-MB-231			MCF-7		
Cell line													
Total aligned reads		75822			55883			191492			66487		
Total on-target reads		1370			4484			4145			1550		
On-target Percent		1.81%			8.02%			2.16%			2.33%		
Median Avg Cov (10 sites)		80			271			249			94		
LOCUS	**region size (kb)**	**Average Coverage**	**Region Read count**	**(% of on-target reads)**	**Average Coverage**	**Region Read count**	**(% of on-target reads)**	**Average Coverage**	**Region Read count**	**(% of on-target reads)**	**Average Coverage**	**Region Read count**	**(% of on-target reads)**
GPX1	13.6	136	176	12.85%	656	792	17.66%	417	548	13.22%	106	130	8.39%
GSTP1	17.8	69	95	6.93%	140	187	4.17%	229	335	8.08%	68	99	6.39%
KRT19	18.1	47	67	4.89%	111	139	3.10%	249	347	8.37%	88	112	7.23%
SLC12A4	24.4	176	253	18.47%	586	761	16.97%	416	561	13.53%	332	425	27.42%
TPM2	19.6	108	155	11.31%	338	469	10.46%	212	342	8.25%	65	88	5.68%
chr5 deletion	18.7	18	29	2.12%	63	88	1.96%	181	256	6.18%	66	90	5.81%
chr7 deletion	20.0	73	102	7.45%	339	456	10.17%	342	517	12.47%	126	168	10.84%
BRAF	12.3	45	65	4.74%	203	262	5.84%	189	278	6.71%	50	76	4.90%
KRAS	16.7	87	130	9.49%	198	291	6.49%	249	361	8.71%	100	132	8.52%
TP53	16.1	236	298	21.75%	846	1039	23.17%	467	600	14.48%	181	230	14.84%

MULTIPLE CUTS EACH SIDE

		GM12878			Normal Breast Tissue			Xenograft (MDA-MB-			Normal Breast Tissue			Tumor Breast Tissue		
Sample																
Total aligned reads		235328			166047			433710			247110			611992		
Total on-target reads		10854			3116			5880			1667			2702		
On-target Percent		4.61%			1.88%			1.36%			0.67%			0.44%		
Median Avg Cov (10 sites)		680			162			312			70			93		
LOCUS	**region size (kb)**	**Average Coverage**	**Region Read count**	**(% of on-target reads)**	**Average Coverage**	**Region Read count**	**(% of on-target reads)**	**Average Coverage**	**Region Read count**	**(% of on-target reads)**	**Average Coverage**	**Region Read count**	**(% of on-target reads)**	**Average Coverage**	**Region Read count**	**(% of on-target reads)**
GPX1	13.6	767	1046	9.64%	241	358	11.49%	484	724	12.31%	123	205	12.30%	101	193	7.14%
GSTP1	17.8	599	1138	10.48%	144	307	9.85%	324	673	11.45%	80	198	11.88%	117	354	13.10%
KRT19	18.1	407	722	6.65%	158	290	9.31%	288	537	9.13%	41	89	5.34%	79	203	7.51%
SLC12A4	24.4	1056	1464	13.49%	208	393	12.61%	278	516	8.78%	95	214	12.84%	165	374	13.84%
TPM2	19.6	752	1325	12.21%	167	298	9.56%	295	540	9.18%	68	141	8.46%	66	193	7.14%
chr5 deletion	18.7	435	816	7.52%	178	295	9.47%	304	474	8.06%	59	128	7.68%	63	141	5.22%
chr7 deletion	20.0	444	628	5.79%	117	203	6.51%	320	486	8.27%	63	112	6.72%	85	179	6.62%
BRAF	12.3	1058	1824	16.80%	268	464	14.89%	569	942	16.02%	178	317	19.02%	256	530	19.62%
KRAS	16.7	649	963	8.87%	149	281	9.02%	254	453	7.70%	72	140	8.40%	149	379	14.03%
TP53	16.1	710	928	8.55%	149	227	7.28%	343	535	9.10%	69	123	7.38%	76	156	5.77%

Table 3.2: Coverage Summary Top: Coverage at 10 loci in GM12878, MCF-10A, MCF-7 and MDA-MB-231 using single cut on each side of region of interest. Bottom: Coverage at 10 loci in GM12878 and primary tissue samples (Normal breast, Xenograft, and Tumor/Normal paired) with multiples cuts on each side of region of interest.

SURVIVOR OFF-TARGET bincov analysis
(finding all off-target pileups with more than 25X coverage, min distance between 1kb)
GM12878, MinION using all breast panel guides + BRCA1 guides

Using only UNIQUE reads (no multi-mappers) ALL SITES WITH COVERAGE of AT LEAST 25X				
chr	start	stop	max coverage	(on-target sites colored orange)
chr1	22155471	22160242	29	
chr1	23510087	23516325	30	
chr1	143210893	143227300	97	
chr1	230323467	230331685	32	
chr2	64300610	64305215	27	
chr2	185547715	185588879	86	
chr3	49345992	49386662	810	(GPX1)
chr4	28968110	28997498	59	
chr4	76674985	76685747	38	
chr4	165120084	165125858	29	
chr5	1869090	1915959	119	
chr5	58356373	58416308	470	(chr5 deletion)
chr6	42228233	42256265	84	
chr7	109276989	109307088	101	
chr7	140766524	140802004	1104	(BRAF)
chr7	142756553	142776937	64	
chr7	154592510	154617821	481	(chr7 deletion)
chr8	69219197	69229030	42	
chr9	35672425	35704226	811	(TPM2)
chr11	67550587	67618650	651	(GSTP1)
chr12	25208649	25277043	718	(KRAS)
chr12	95015201	95025437	30	
chr14	97202222	97209242	37	
chr14	105593745	105599458	29	
chr15	73858753	73863754	30	
chr16	46382684	46406850	179	
chr16	67944237	67978278	1125	(SLC12A4)
chr16	84808469	84844355	73	
chr17	7664714	7690276	744	(TP53)
chr17	28504500	28509976	30	
chr17	41493463	41538564	430	(KRT19)
chr17	43038193	43127260	189	(BRCA1)
chr21	8436628	8449619	30	
chr22	44317694	44326791	38	
chrM	1	17219	100	

Table 3.3: Off-target Analysis Off-target analysis for the GM12878 sequencing run using multiple guideRNAs flanking each site, using the bincov tool from SURVIVOR (Jeffares, D. C. et al. Nat. Commun. 8, 14061 (2017). On-target loci are colored orange. Max coverage shows the highest coverage reached in the specific locus.

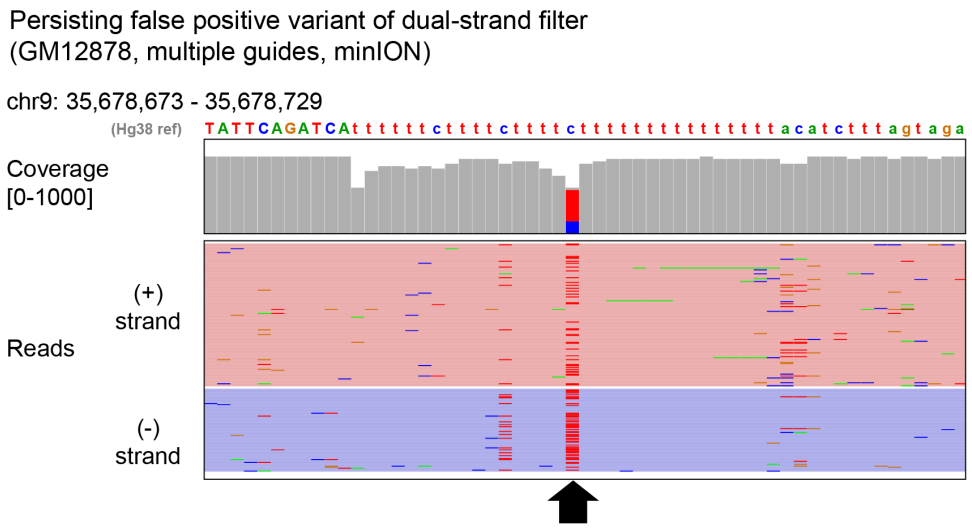


Figure 3.7: Persisting False Positive SNV The single false positive variant from high-coverage sequencing data that passes dual-strand filtering. This variant is present in a highly thymidine-dense region. Note this variant falls within a repetitive region of the genome masked by RepeatMasker, thus the lowercase reference.

ALL single nucleotide variants identified in MDA-MB-231 data by nanopolish in the captured regions for TP53, KRAS and BRAF

CHROM	POS	REF	ALT	QUAL	INFO	FORMAT	sample
chr7	140775567	G	A	746.0	BaseCalledReadsWithVariant=78;BaseCalledFraction=0.443182;TotalReads=170;AlleleCount=1;SupportFraction=0.472754	GT	0/1
chr7	140776301	G	C	944.2	BaseCalledReadsWithVariant=71;BaseCalledFraction=0.398876;TotalReads=172;AlleleCount=1;SupportFraction=0.48225	GT	0/1
chr7	140777136	G	A	714.2	BaseCalledReadsWithVariant=81;BaseCalledFraction=0.452514;TotalReads=173;AlleleCount=1;SupportFraction=0.48875	GT	0/1
chr7	140777206	A	T	69.3	BaseCalledReadsWithVariant=43;BaseCalledFraction=0.240223;TotalReads=173;AlleleCount=1;SupportFraction=0.561115	GT	0/1
chr7	140777956	C	G	1384.9	BaseCalledReadsWithVariant=79;BaseCalledFraction=0.438889;TotalReads=174;AlleleCount=1;SupportFraction=0.490326	GT	0/1
chr7	140778454	G	T	993.2	BaseCalledReadsWithVariant=90;BaseCalledFraction=0.48913;TotalReads=178;AlleleCount=1;SupportFraction=0.514656	GT	0/1
chr7	140780263	A	T	1231.7	BaseCalledReadsWithVariant=113;BaseCalledFraction=0.588542;TotalReads=186;AlleleCount=2;SupportFraction=0.925469	GT	1/1
chr7	140781426	T	C	1218.1	BaseCalledReadsWithVariant=91;BaseCalledFraction=0.457286;TotalReads=191;AlleleCount=1;SupportFraction=0.518763	GT	0/1
chr7	140781617	C	A	986.5	BaseCalledReadsWithVariant=94;BaseCalledFraction=0.47;TotalReads=192;AlleleCount=1;SupportFraction=0.508238	GT	0/1
chr7	140782627	T	C	657.4	BaseCalledReadsWithVariant=76;BaseCalledFraction=0.374384;TotalReads=195;AlleleCount=1;SupportFraction=0.574334	GT	0/1
chr7	140783600	T	C	1843.2	BaseCalledReadsWithVariant=81;BaseCalledFraction=0.397059;TotalReads=196;AlleleCount=2;SupportFraction=0.707989	GT	1/1
chr7	140783805	C	T	504.9	BaseCalledReadsWithVariant=82;BaseCalledFraction=0.398058;TotalReads=198;AlleleCount=1;SupportFraction=0.510465	GT	0/1
chr7	140785535	T	C	1065.3	BaseCalledReadsWithVariant=156;BaseCalledFraction=0.725581;TotalReads=206;AlleleCount=2;SupportFraction=0.900077	GT	1/1
chr7	140785890	C	T	1099.9	BaseCalledReadsWithVariant=193;BaseCalledFraction=0.897674;TotalReads=205;AlleleCount=2;SupportFraction=0.937688	GT	1/1
chr7	140786488	C	T	1253.3	BaseCalledReadsWithVariant=204;BaseCalledFraction=0.953271;TotalReads=205;AlleleCount=2;SupportFraction=0.925928	GT	1/1
chr7	140786713	A	G	719.9	BaseCalledReadsWithVariant=96;BaseCalledFraction=0.446512;TotalReads=206;AlleleCount=1;SupportFraction=0.497005	GT	0/1
chr7	140786943	T	C	922.3	BaseCalledReadsWithVariant=76;BaseCalledFraction=0.353488;TotalReads=206;AlleleCount=1;SupportFraction=0.500815	GT	0/1
chr7	140787279	C	T	917.1	BaseCalledReadsWithVariant=90;BaseCalledFraction=0.420561;TotalReads=205;AlleleCount=1;SupportFraction=0.538519	GT	0/1
chr7	140787402	C	T	1243.4	BaseCalledReadsWithVariant=195;BaseCalledFraction=0.915493;TotalReads=204;AlleleCount=2;SupportFraction=0.931547	GT	1/1
chr12	25239348	C	T	1220.8	BaseCalledReadsWithVariant=152;BaseCalledFraction=0.529617;TotalReads=276;AlleleCount=1;SupportFraction=0.600488	GT	0/1
chr12	25241785	C	G	1755.3	BaseCalledReadsWithVariant=124;BaseCalledFraction=0.457565;TotalReads=259;AlleleCount=1;SupportFraction=0.577746	GT	0/1
chr12	25241845	C	T	1278.6	BaseCalledReadsWithVariant=147;BaseCalledFraction=0.542435;TotalReads=259;AlleleCount=1;SupportFraction=0.601782	GT	0/1
chr12	25245347	C	T	1047.5	BaseCalledReadsWithVariant=117;BaseCalledFraction=0.45;TotalReads=250;AlleleCount=1;SupportFraction=0.490003	GT	0/1
chr12	25246187	A	C	861.2	BaseCalledReadsWithVariant=113;BaseCalledFraction=0.441406;TotalReads=245;AlleleCount=1;SupportFraction=0.569208	GT	0/1
chr12	25248730	C	T	516.8	BaseCalledReadsWithVariant=93;BaseCalledFraction=0.375;TotalReads=238;AlleleCount=1;SupportFraction=0.366001	GT	0/1
chr17	7666380	A	C	5993.8	BaseCalledReadsWithVariant=435;BaseCalledFraction=0.873494;TotalReads=477;AlleleCount=2;SupportFraction=0.956688	GT	1/1
chr17	7666871	T	C	2586.8	BaseCalledReadsWithVariant=423;BaseCalledFraction=0.844311;TotalReads=478;AlleleCount=2;SupportFraction=0.907531	GT	1/1
chr17	7667560	G	A	2481.9	BaseCalledReadsWithVariant=370;BaseCalledFraction=0.745968;TotalReads=474;AlleleCount=2;SupportFraction=0.905794	GT	1/1
chr17	7667611	C	T	4225.8	BaseCalledReadsWithVariant=422;BaseCalledFraction=0.850806;TotalReads=474;AlleleCount=2;SupportFraction=0.788236	GT	1/1
chr17	7667612	A	G	4225.8	BaseCalledReadsWithVariant=428;BaseCalledFraction=0.862903;TotalReads=474;AlleleCount=2;SupportFraction=0.797237	GT	1/1
chr17	7667762	A	G	4137.1	BaseCalledReadsWithVariant=417;BaseCalledFraction=0.840726;TotalReads=474;AlleleCount=2;SupportFraction=0.961707	GT	1/1
chr17	7668134	G	A	4024.8	BaseCalledReadsWithVariant=428;BaseCalledFraction=0.862903;TotalReads=473;AlleleCount=2;SupportFraction=0.955112	GT	1/1
chr17	7672246	T	C	4537.6	BaseCalledReadsWithVariant=403;BaseCalledFraction=0.832645;TotalReads=463;AlleleCount=2;SupportFraction=0.93749	GT	1/1
chr17	7673183	G	A	3209.1	BaseCalledReadsWithVariant=446;BaseCalledFraction=0.919588;TotalReads=463;AlleleCount=2;SupportFraction=0.9372	GT	1/1
chr17	7673781	C	T	5030.0	BaseCalledReadsWithVariant=390;BaseCalledFraction=0.802469;TotalReads=464;AlleleCount=2;SupportFraction=0.975668	GT	1/1
chr17	7674797	T	C	4154.4	BaseCalledReadsWithVariant=317;BaseCalledFraction=0.653608;TotalReads=464;AlleleCount=2;SupportFraction=0.964486	GT	1/1
chr17	7675327	C	T	2858.9	BaseCalledReadsWithVariant=364;BaseCalledFraction=0.753623;TotalReads=461;AlleleCount=2;SupportFraction=0.936794	GT	1/1
chr17	7675519	A	G	4121.6	BaseCalledReadsWithVariant=418;BaseCalledFraction=0.869023;TotalReads=459;AlleleCount=2;SupportFraction=0.931947	GT	1/1
chr17	7676154	G	C	2301.5	BaseCalledReadsWithVariant=271;BaseCalledFraction=0.566946;TotalReads=455;AlleleCount=2;SupportFraction=0.88552	GT	1/1
chr17	7676483	G	C	4211.8	BaseCalledReadsWithVariant=211;BaseCalledFraction=0.440501;TotalReads=456;AlleleCount=2;SupportFraction=0.871194	GT	1/1
chr17	7677867	C	A	4622.2	BaseCalledReadsWithVariant=378;BaseCalledFraction=0.784232;TotalReads=459;AlleleCount=2;SupportFraction=0.96723	GT	1/1
chr17	7679660	C	T	422.7	BaseCalledReadsWithVariant=104;BaseCalledFraction=0.213552;TotalReads=463;AlleleCount=1;SupportFraction=0.423232	GT	0/1

Table 3.4: MDA-MB-231 Single Nucleotide Variants Variants identified *de novo* using nanopolish from nanopore data at three loci (TP53, BRAF, KRAS.) The red boxes show the mutations annotated in COSMIC database, blue shows the five variants removed by dual-strand filtering

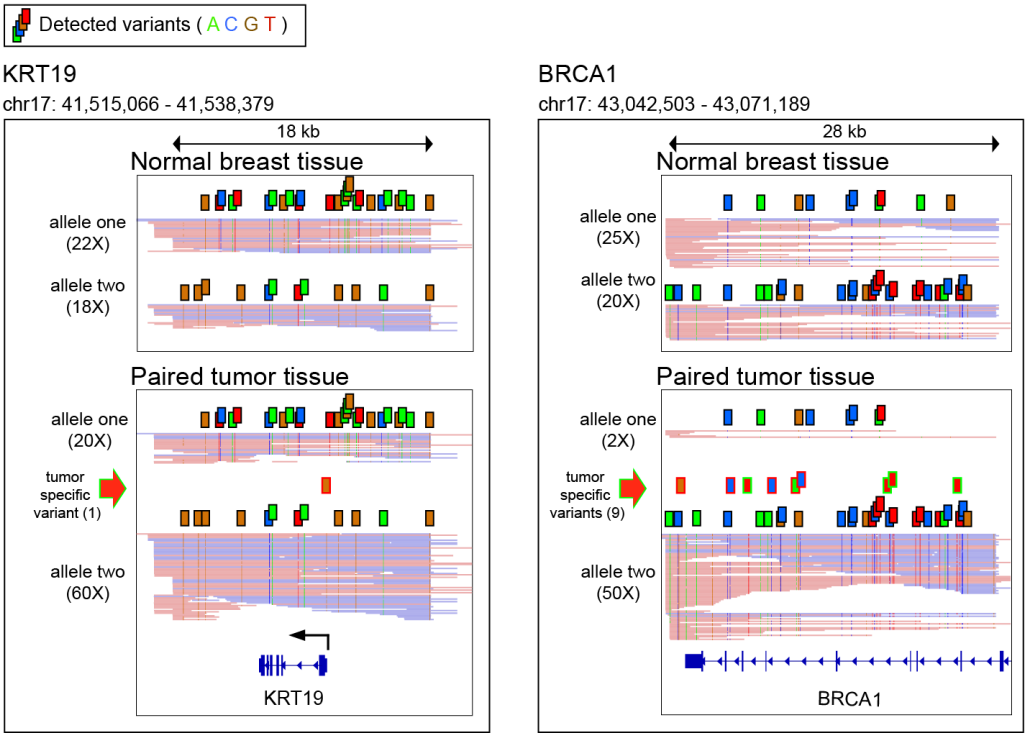


Figure 3.8: Tumor Loss of Heterozygosity Single nucleotide high-confidence variant calls (nanopolish passing dual strand filter) at two other enriched sites on chr17 (KRT19 and 30kb piece of BRCA1). Reads were phased to show only variants passing dual-strand filter using the 'phase-reads' module of nanopolish. Tumor reads were phased into haplotypes using only variants from the corresponding normal sample.

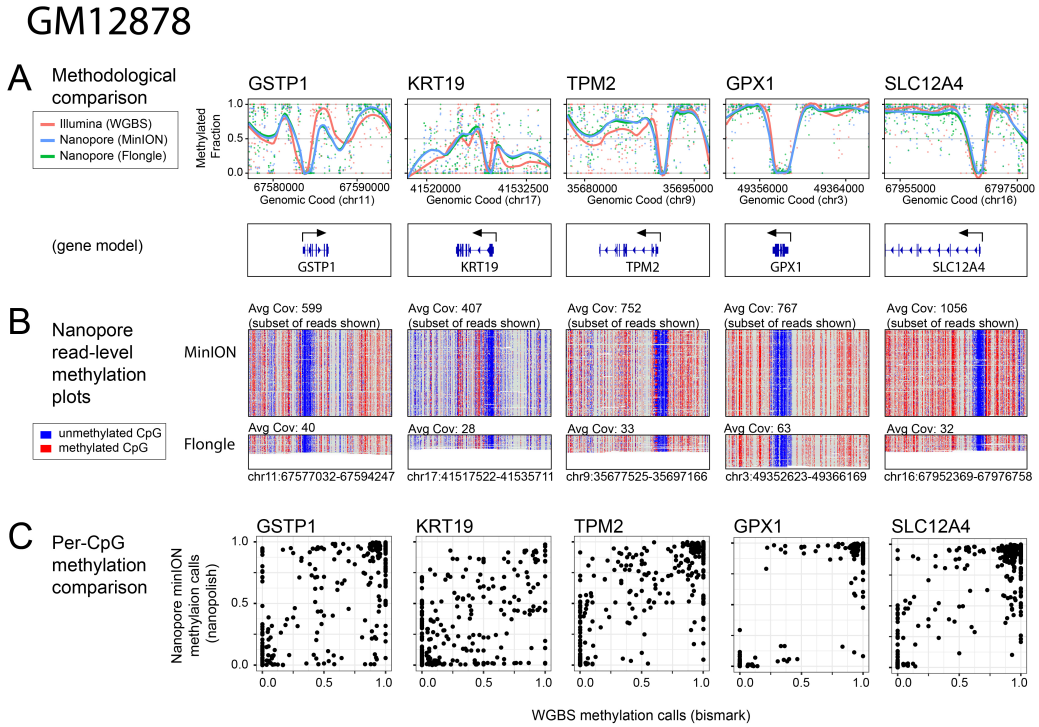


Figure 3.9: Comparison of nCATS methylation Data with WGBS data (A) Line and dot plot of methylation calls made by bismark (WGBS Illumina data: GEO: GSE86765) and nanopolish (Cas9-targeted nanopore data) at all CpGs in the targeted regions. Gene models plotted below for orientation. (B) Read-level methylation plots for five loci in GM12878. (C) Per-CpG scatter plot comparing methylation calls made by bismark (WGBS Illumina data: GEO: GSE86765) and nanopolish (Cas9-targeted nanopore data) at all CpGs in the targeted regions. r=0.81 across all 5 sites.

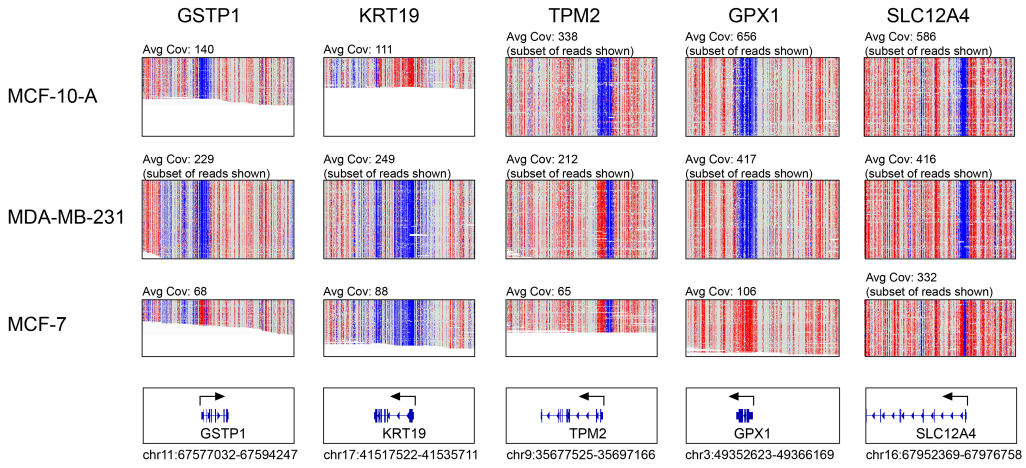


Figure 3.10: Breast cell line methylation Read-level methylation plots for all methylation-associated loci in three breast cell lines (MCF-10A, MDA-MB-231, MCF-7)

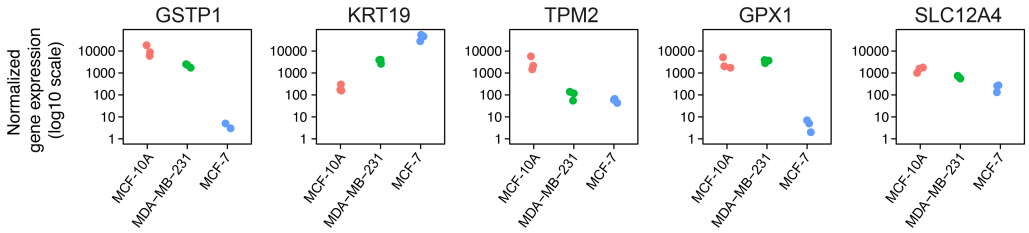


Figure 3.11: Transcript level comparison in breast cell lines Normalized expression data (read counts) for three breast cell lines (MCF-10A, MDA-MB-231, MCF-7) from existing RNA-seq data (GEO: GSE75168).

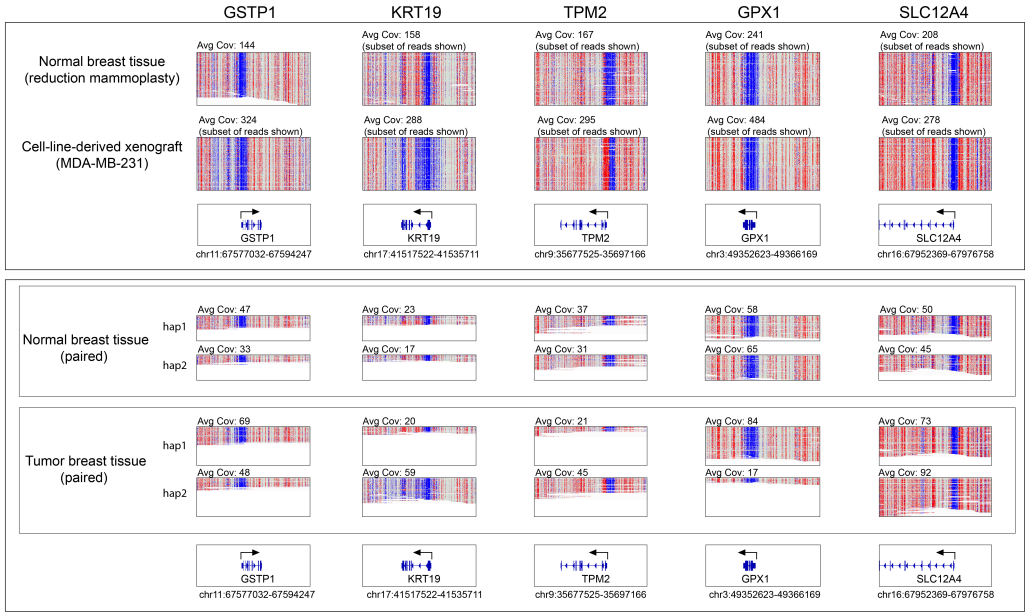


Figure 3.12: Primary tissue methylation Read-level methylation plots for five captured loci in fresh breast tissue (reduction mammoplasty, cell-line-derived xenograft, paired tumor/normal). Tumor/normal samples are segregated into haplotypes using only variants from the normal sample.

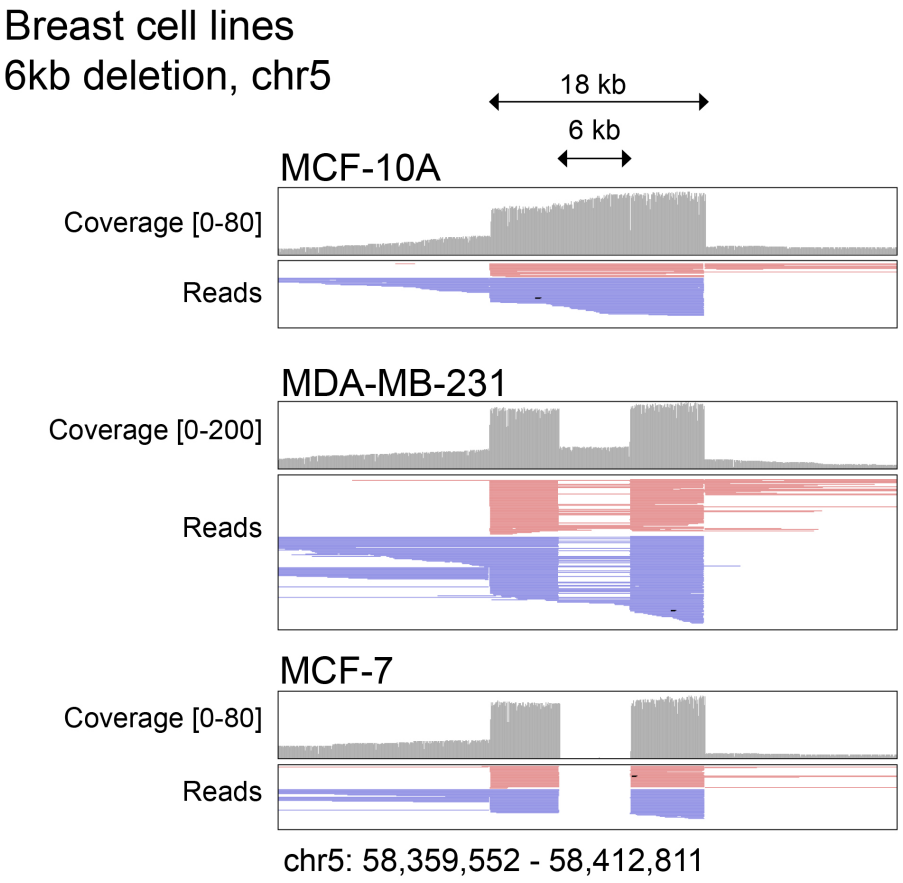


Figure 3.13: Chromosome 5 deletion in breast cell lines Reads at a small (< 10kb) common structural variant on chromosome 5 from breast cell line nanopore enrichment data (deletion at chromosome 7 is included as Figure 3.4A).

Breast Cancer Cell Lines - Sniffles Calls
(min size: 100bp; min support: 50 reads)

chr5 deletion

region between guideRNAs: 58378092-58396781

cell line	start site	size	GT
MCF-10A	[no SV called]	[]	[]
MDA-MB-231	58384160	6107	het
MCF-7	58384160	6110	homo

chr7 deletion

region between guideRNAs: 154593967-154613978

cell line	start site	size	GT
MCF-10A	[no SV called]	[]	[]
MDA-MB-231	154600417	7569	het
MCF-7	154600417	7569	homo

Table 3.5: Sniffles calls in breast cell lines Indels called at deletion locations in breast cell lines

GM12878 - Large Deletions

REFERENCE
Reference Calls (GIAB, 10X genomics LongRanger2.1 data)

chromosome	start site	size	GT
chr5	105096420	71545	het
chr6	78257486	69265	het
chr8	39374563	155140	het

CALLED
Sniffles Calls (min size: 100bp; min support: 50 reads)

chromosome	start site	size	GT*
chr5	105096412	71560	het
chr6	78257496	69263	het
chr8	39374556	155153	het

GT*: option "--min_homo_af" set to 99.9 to account for allelic imbalance

Table 3.6: Sniffles calls in GM12878 Indels called at large heterozygous deletions in GM12878

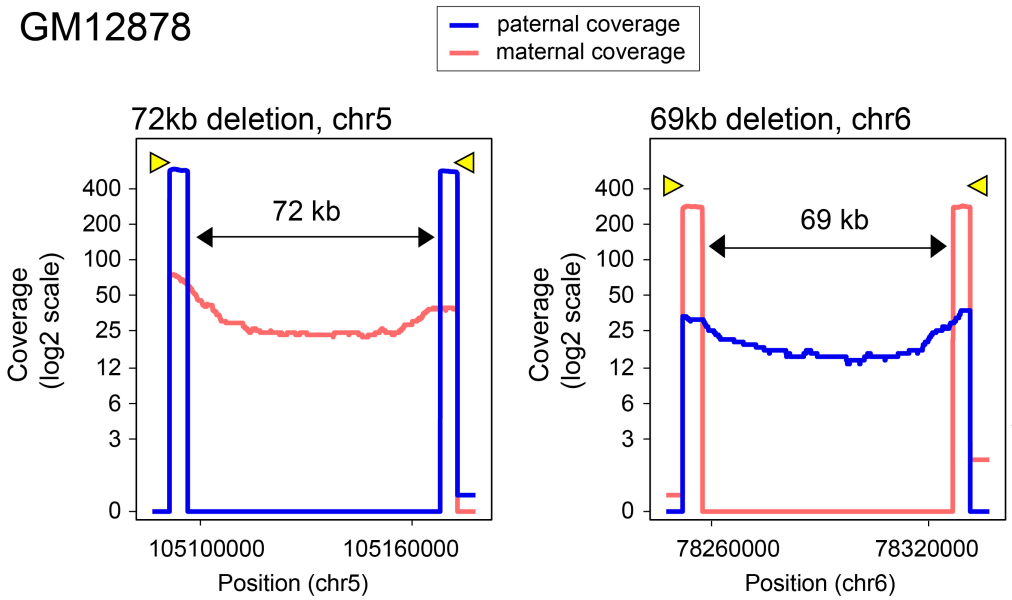


Figure 3.14: Per-allele coverage around large deletions Coverage plots around two large heterozygous deletions in GM12878. Yellow triangles show points of Cas9 cleavage. Blue lines show coverage of reads assigned to paternal haplotype and red lines show coverage of reads assigned to maternal haplotype. (In both cases, the distance between cuts on the deleted allele is 10kb and distance between cuts on non-deleted allele is 80kb).

Indels greater than or equal to three nucleotides between assemblies, removing indels from homopolymer length differences

position located in (Hg38)	platinum_genomes_2017_annotation (chr | pos | ref | alt)	Assembly_1_start_position	**Assembly_1_variant**	**Assembly_2_variant**	Assembly_2_start_position
43052820*	NA	9825	acacacac	-	9817
43059470	chr17 | 43059469 | CACA | C,CACAACA	16530	-	acaaca	16510
43071520	chr17 | 43071521 | G | GAATGTTCACTGTAACAATGCTTGT	28592	-	aatgttcactgtaacaatgcttgt	28580
43074719	chr17 | 43074719 | GGGGTT | G	31791	gggtt	-	31807
43075066	chr17 | 43075066 | C | CGGAA	32138	-	ggaa	32149
43086110	chr17 | 43086109 | T | TACACAC	43188	-	acacac	43201
43095105	chr17 | 43095105 | A | ACCT	52180	-	cct	52198
43100618*	NA	57703	aac	-	57735
43100630*	NA	57715	tgt	-	57744
43109042	chr17 | 43109041 | T | TCTATCTATCTACCTAC	66142	-	ctatctatctacctac	66167

*Blue coloration indicates that the indel is not presently annotated in the Genome in a Bottle or the Platinum Genomes data sets

Table 3.7: Comparing BRCA1 alleles in GM12878 Indels greater than or equal to three nucleotides between assemblies, removing indels from homopolymer length differences

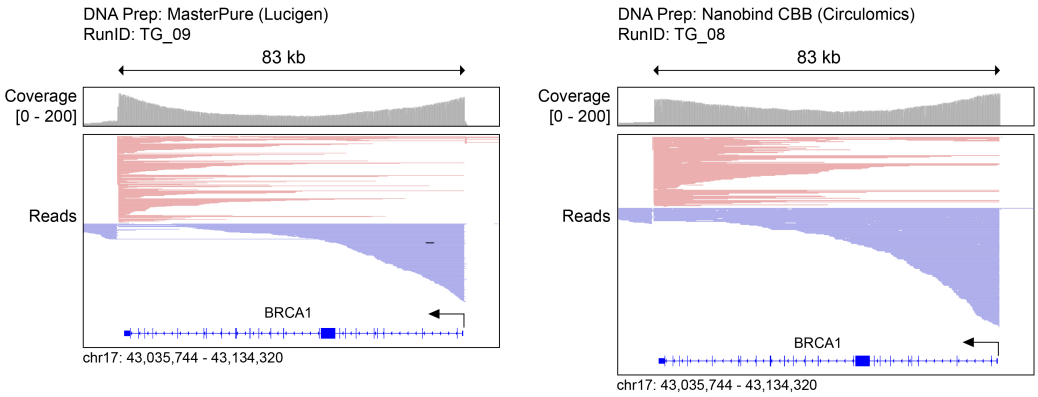


Figure 3.15: Reads at the BRCA1 locus for GM12878 Left: Reads from BRCA1 enrichment with DNA extracted using the Masterpure kit (Lucigen, Cat MC85200) Right: Reads from BRCA1 enrichment run with DNA extracted using the Nanobind kit (Circulomics, Cat NB-900-001-0)

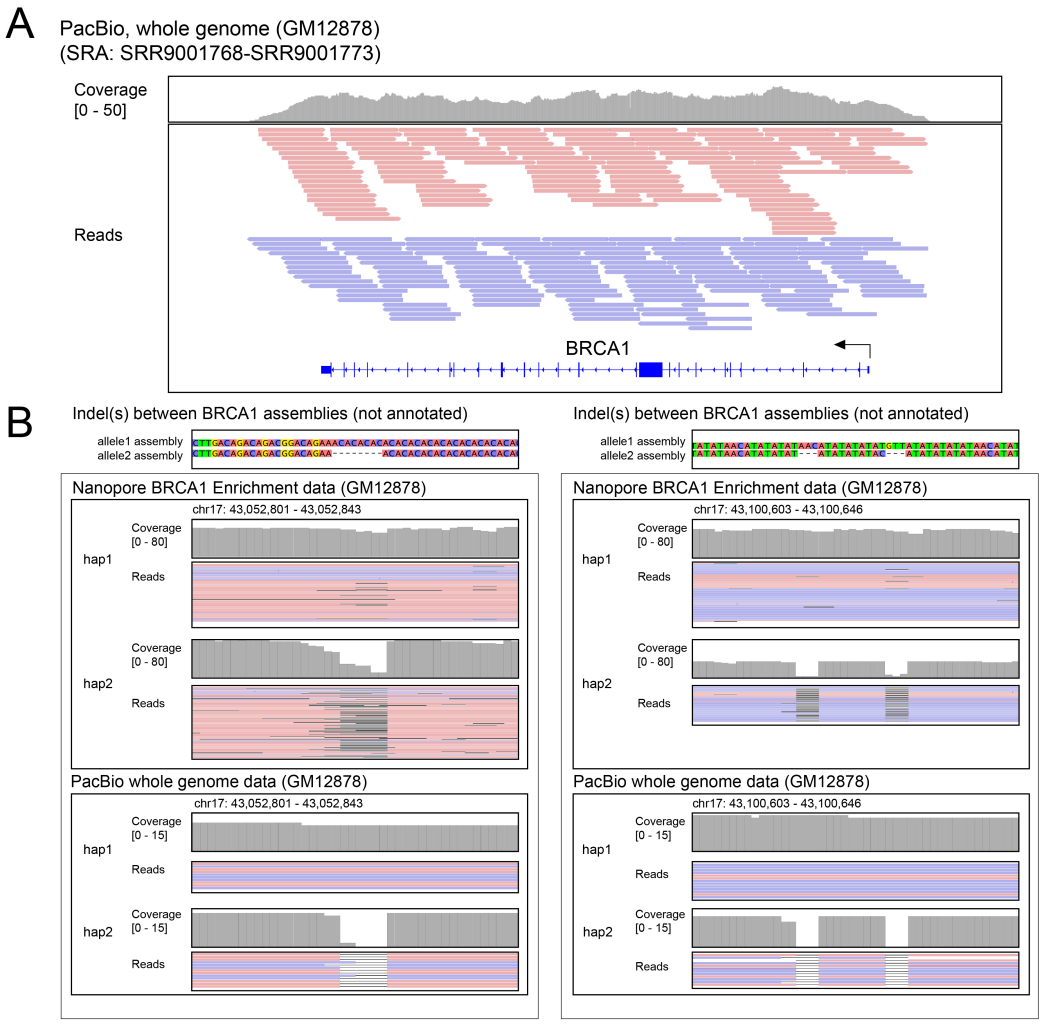


Figure 3.16: Validation of novel BRCA1 indels (A) Showing whole genome PacBio data around BRCA1 in GM12878 from publicly available data (SRA: SRR9001768 - SRR9001773) (B) Comparison of the three not annotated heterozygous indels found in GM12878 between Cas9-nanopore enrichment data (top) and whole-genome PacBio data (bottom).

Target Locus	guideRNA name	guideRNA sequence	target site
GPX1	gpx1_fwd1	GCAAGGAGGGGTCAATCACC	chr3:(+) 49352525
	gpx1_fwd2	AGGCCACATTGACCCTGTAC	chr3:(+) 49352623
	gpx1_rev1	TAACTTCACCTAGATCCTAT	chr3:(-) 49366169
	gpx1_rev2	CCTACCTTAATGATGATAAC	chr3:(-) 49366773
GSTP1	gstp1_fwd1	CCCGATGACGCACCTCGGAG	chr11:(+) 67576428
	gstp1_fwd2	GGGTATCCTAAGACACGTGT	chr11:(+) 67577032
	gstp1_fwd3	ACTGCATACAAGCCTCGTCT	chr11:(+) 67576525
	gstp1_rev1	GAATTATAGTGATACGGAAG	chr11:(-) 67594247
	gstp1_rev2	GTCAATAGCACCCCACAAGT	chr11:(-) 67597052
	gstp1_rev3	CCTCAAGTGCTCCCTACATC	chr11:(-) 67596359
KRT19	krt19_fwd1	GCCCCACTGTGGACAACTCA	chr17:(+) 41517522
	krt19_fwd2	CGTCCATCTCCACATTGACC	chr17:(+) 41516210
	krt19_fwd3	GTTCAGGGAACCACCTCTTG	chr17:(+) 41515773
	krt19_rev1	TACTCTCTAGCCCACCACTA	chr17:(-) 41535711
	krt19_rev2	GGCTCCCCAAGAGACGGCAT	chr17:(-) 41535896
	krt19_rev3	GCTCACCCCTCTTAACCACT	chr17:(-) 41537561
SLC12A4	slc12a4_fwd1	GACGTGTCATGAGCACCCGA	chr16:(+) 67952369
	slc12a4_fwd2	TGGGTTACTTACGGAAACAC	chr16:(+) 67952171
	slc12a4_rev1	TAGATCACTTTGACCACATC	chr16:(-) 67976758
	slc12a4_rev2	GCTTGCTAACGTGTCCACTG	chr16:(-) 67977275
TPM2	tpm2_fwd1	AACCAGTCCACCAAGCTTGT	chr9:(+) 35677525
	tpm2_fwd2	ACATACACCATGATTAGAGG	chr9:(+) 35676780
	tpm2_fwd3	GTGACACTTAAAGCCTTCAC	chr9:(+) 35676677
	tpm2_rev1	CGGACCCTAAAGTTATCCAA	chr9:(-) 35697166
	tpm2_rev2	CCAAGTGCCTTCATGCCCTA	chr9:(-) 35697594
	tpm2_rev3	CAGCCAATGCACTGGACGAT	chr9:(-) 35698646
chr5 brca 6kb deletion	chr5_brca_del_fwd1	ACCATAAATTTCCCCTCTAC	chr5:(+) 58378092
	chr5_brca_del_fwd2	ACTTCCAACTGTAGCTCAGT	chr5:(+) 58379108
	chr5_brca_del_fwd3	TAGTCAAGCTCTAACAGCCT	chr5:(+) 58379389
	chr5_brca_del_rev1	GAACGATTCTGGTTAGTCTA	chr5:(-) 58396781
	chr5_brca_del_rev2	TTGAAAGTACCATCTCGTGT	chr5:(-) 58395638
	chr5_brca_del_rev3	ATCATCAACTTTTGTCGTAA	chr5:(-) 58395287
chr7 brca 6kb deletion	chr7_brca_del_fwd1	TCTGAGAACGGCCTACATAT	chr7:(+) 154593967
	chr7_brca_del_fwd2	GGACTATTGATTATCAACTC	chr7:(+) 154593492
	chr7_brca_del_rev1	AAGACACTGTATGCGGAATG	chr7:(-) 154613978
	chr7_brca_del_rev2	AAACTCACAGCGGTCCCATG	chr7:(-) 154614920
BRAF	braf_fwd1	TCCTAAAGAAGGAACACGCT	chr7:(+) 140775214
	braf_fwd2	AGTTATGTACTGAGTATGCC	chr7:(+) 140774916
	braf_fwd3	AGGATTCTACGGTATAAACC	chr7:(+) 140773753
	braf_rev1	GACCAAGGATTTCGTGGTGA	chr7:(-) 140787552
	braf_rev2	AATGCAAGTTCTACCCATCA	chr7:(-) 140789303
	braf_rev3	TATCACCTGTATCACTTAGT	chr7:(-) 140789720
KRAS	kras_fwd1	GCTACAACTGAAGGCTATGA	chr12:(+) 25237978
	kras_fwd2	CCAGGACAACCCATGAGTAC	chr12:(+) 25237475
	kras_rev1	GCCACAACCGTCTGGACCCA	chr12:(-) 25254719
	kras_rev2	GCAAGATTAAGTCTCCGGGA	chr12:(-) 25255658
TP53	tp53_fwd1	CTCGTGTTCCTAAAATGAGG	chr17:(+) 7666028
	tp53_fwd2	GGGAACTTGAACACGACCAG	chr17:(+) 7666465
	tp53_rev1	CAACGTTGATACCATACTGG	chr17:(-) 7682126
	tp53_rev2	ATAGTCAGACTGGTCTTTAC	chr17:(-) 7682752
BRCA1	brca1_fwd1	ACCGAGACTCATCAACTCAC	chr17:(+) 43042955
	brca1_fwd2	CGTGTATTAACTCCATCATC	chr17:(+) 43042624
	brca1_rev1	TCGGTCCCTCAGAACACGAA	chr17:(-) 43126602
	brca1_rev2	TGACAAGTACAAGCGCGCAC	chr17:(-) 43125951
	brca1_rev3	ACGACCACCCCATTGACTGG	chr17:(-) 43126268
	brca1_midRev_exon16	ATGATATAGGACTTTCGAAT	chr17:(-) 43070656
	brca1_midFwd_exon16	GTTGTTAAGTCTTAGTCATT	chr17:(+) 43070881
chr5 GM12878 70kb deletion	chr5_gm12878_del_fwd1	AGGAGAGTCCACTACCTAGG	chr5:(+) 105091232
	chr5_gm12878_del_rev1	ATGAGGCTTAATGACTGTGA	chr5:(-) 105172855
chr6 GM12878 70kb deletion	chr6_gm12878_del_fwd1	CTCTTAATAGTCTCGGACAG	chr6:(+) 78251982
	chr6_gm12878_del_rev1	CGGTAAGTCAACAGATGGTA	chr6:(-) 78331484
chr8 GM12878 150kb deletion	chr8_gm12878_del_fwd1	CCTTGAGGTGTAATTACATT	chr8:(+) 39369484
	chr8_gm12878_del_rev1	GAGATGTTATCTTCCTAACT	chr8:(-) 39535051

Table 3.8: Breast Cancer GuideRNAs Sequences, binding locations(Hg38) and directionality for guideRNAs used in breast cancer studies

References

Eberle, Michael A, Epameinondas Fritzilas, Peter Krusche, Morten Källberg, Benjamin L Moore, Mitchell A Bekritsky, Zamin Iqbal, Han-Yu Chuang, Sean J Humphray, Aaron L Halpern, Semyon Kruglyak, Elliott H Margulies, Gil McVean, and David R Bentley (2016). "A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree". In: *Genome Res.*

Zook, Justin M, David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng, Yuling Liu, Christopher E Mason, Noah Alexander, Elizabeth Henaff, Alexa B R McIntyre, Dhruva Chandramohan, Feng Chen, Erich Jaeger, Ali Moshrefi, Khoa Pham, William Stedman, Tiffany Liang, Michael Saghbini, Zeljko Dzakula, Alex Hastie, Han Cao, Gintaras Deikus, Eric Schadt, Robert Sebra, Ali Bashir, Rebecca M Truty, Christopher C Chang, Natali Gulbahce, Keyan Zhao, Srinka Ghosh, Fiona Hyland, Yutao Fu, Mark Chaisson, Chunlin Xiao, Jonathan Trow, Stephen T Sherry, Alexander W Zaranek, Madeleine Ball, Jason Bobe, Preston Estep, George M Church, Patrick Marks, Sofia Kyriazopoulou-Panagiotopoulou, Grace X Y Zheng, Michael Schnall-Levin, Heather S Ordonez, Patrice A Mudivarti, Kristina Giorda, Ying Sheng, Karoline Bjarnesdatter Rypdal, and Marc Salit (2016). "Extensive sequencing of seven human genomes to characterize benchmark reference materials". en. In: *Sci Data* 3, p. 160025.

ENCODE Project Consortium (2012). "An integrated encyclopedia of DNA elements in the human genome". en. In: *Nature* 489.7414, pp. 57–74.

Forbes, Simon A, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, Raymund Stefancsik, Bhavana Harsha, Chai Yin Kok, Mingming Jia, Harry Jubb, Zbyslaw Sondka, Sam Thompson, Tisham De, and Peter J Campbell (2017). "COSMIC: somatic cancer genetics at high-resolution". en. In: *Nucleic Acids Res.* 45.D1, pp. D777–D783.

Lee, Isac, Roham Razaghi, Timothy Gilpatrick, Norah Sadowski, Fritz Sedlazeck, and Winston Timp (2018). "Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing". en.

Welcsh, P L and M C King (2001). "BRCA1 and BRCA2 and the genetics of breast and ovarian cancer". en. In: *Hum. Mol. Genet.* 10.7, pp. 705–713.

Li, Heng (2011). "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". en. In: *Bioinformatics* 27.21, pp. 2987–2993.

Luo, Ruibang, Fritz J Sedlazeck, Tak-Wah Lam, and Michael C Schatz (2019). "A multi-task convolutional deep neural network for variant calling in single molecule sequencing". en. In: *Nat. Commun.* 10.1, p. 998.

Simpson, Jared T, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi, and Winston Timp (2017). "Detecting DNA cytosine methylation using nanopore sequencing". en. In: *Nat. Methods* 14.4, pp. 407–410.

Patterson, Murray, Tobias Marschall, Nadia Pisanti, Leo van Iersel, Leen Stougie, Gunnar W Klau, and Alexander Schönhuth (2015). "WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads". en. In: *J. Comput. Biol.* 22.6, pp. 498–509.

Tate, John G, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes (2019). "COSMIC: the Catalogue Of Somatic Mutations In Cancer". en. In: *Nucleic Acids Res.* 47.D1, pp. D941–D947.

Messier, Terri L, Jonathan A R Gordon, Joseph R Boyd, Coralee E Tye, Gillian Browne, Janet L Stein, Jane B Lian, and Gary S Stein (2016). "Histone H3 lysine 4 acetylation and methylation dynamics define breast cancer subtypes". en. In: *Oncotarget* 7.5, pp. 5094–5109.

Kabir, Nuzhat N, Lars Rönnstrand, and Julhash U Kazi (2014). "Keratin 19 expression correlates with poor prognosis in breast cancer". en. In: *Mol. Biol. Rep.* 41.12, pp. 7729–7735.

Martignano, Filippo, Giorgia Gurioli, Samanta Salvi, Daniele Calistri, Matteo Costantini, Roberta Gunelli, Ugo De Giorgi, Flavia Foca, and Valentina Casadio (2016). "GSTP1 Methylation and Protein Expression in Prostate Cancer: Diagnostic Implications". en. In: *Dis. Markers* 2016, p. 4358292.

Wang, Xi-Mei, Zhen Zhang, Li-Hui Pan, Xu-Chen Cao, and Chunhua Xiao (2018). "KRT19 and CEACAM5 mRNA-marked circulated tumor cells indicate unfavorable prognosis of breast cancer patients". en. In: *Breast Cancer Res. Treat.*

Noguchi, S, T Aihara, K Motomura, H Inaji, S Imaoka, and H Koyama (1996). "Detection of breast cancer micrometastases in axillary lymph nodes by means of reverse transcriptase-polymerase chain reaction. Comparison between MUC1 mRNA and keratin 19 mRNA amplification". en. In: *Am. J. Pathol.* 148.2, pp. 649–656.

Sedlazeck, Fritz J, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C Schatz (2018). "Accurate detection of complex structural variations using single-molecule sequencing". en. In: *Nat. Methods* 15.6, pp. 461–468.

Martin, Marcel, Murray Patterson, Shilpa Garg, Sarah Fischer, Nadia Pisanti, Gunnar W Klau, Alexander Schoenhuth, and Tobias Marschall. *Whatshap: fast and accurate read-based phasing. bioRxiv. 2016*.

Deininger, Prescott (2011). "Alu elements: know the SINEs". en. In: *Genome Biol.* 12.12, p. 236.

Kolmogorov, Mikhail, Jeffrey Yuan, Yu Lin, and Pavel A Pevzner (2019). "Assembly of long, error-prone reads using repeat graphs". en. In: *Nat. Biotechnol.* 37.5, pp. 540–546.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić (2017). "Fast and accurate de novo genome assembly from long uncorrected reads". en. In: *Genome Res.* 27.5, pp. 737–746.

Smit, Afa, R Hubley, and P Green. *RepeatMasker Open-4.0*. http://www.repeatmasker.org.

Li, Heng (2018). "Minimap2: pairwise alignment for nucleotide sequences". en. In: *Bioinformatics* 34.18, pp. 3094–3100.

Audano, Peter A, Arvis Sulovari, Tina A Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, Annemarie E Welch, Max L Dougherty, Bradley J Nelson, Ankeeta Shah, Susan K Dutcher, Wesley C Warren, Vincent Magrini, Sean D McGrath, Yang I Li, Richard K Wilson, and Evan E Eichler (2019). "Characterizing the Major Structural Variant Alleles of the Human Genome". en. In: *Cell* 176.3, 663–675.e19.

Chaisson, Mark J P, Ashley D Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J Gardner, Oscar Rodriguez, Li Guo, Ryan L Collins, Xian Fan, Jia Wen, Robert E Handsaker, Susan Fairley, Zev N Kronenberg, Xiangmeng Kong, Fereydoun Hormozdiari, Dillon Lee,

Aaron M Wenger, Alex Hastie, Danny Antaki, Peter Audano, Harrison Brand, Stuart Cantsilieris, Han Cao, Eliza Cerveira, Chong Chen, Xintong Chen, Chen-Shan Chin, Zechen Chong, Nelson T Chuang, Christine C Lambert, Deanna M Church, Laura Clarke, Andrew Farrell, Joey Flores, Timur Galeev, David Gorkin, Madhusudan Gujral, Victor Guryev, William Haynes Heaton, Jonas Korlach, Sushant Kumar, Jee Young Kwon, Jong Eun Lee, Joyce Lee, Wan-Ping Lee, Sau Peng Lee, Shantao Li, Patrick Marks, Karine Viaud-Martinez, Sascha Meiers, Katherine M Munson, Fabio Navarro, Bradley J Nelson, Conor Nodzak, Amina Noor, Sofia Kyriazopoulou-Panagiotopoulou, Andy Pang, Yunjiang Qiu, Gabriel Rosanio, Mallory Ryan, Adrian Stütz, Diana C J Spierings, Alistair Ward, Annemarie E Welch, Ming Xiao, Wei Xu, Chengsheng Zhang, Qihui Zhu, Xiangqun Zheng-Bradley, Ernesto Lowy, Sergei Yakneen, Steven McCarroll, Goo Jun, Li Ding, Chong Lek Koh, Bing Ren, Paul Flicek, Ken Chen, Mark B Gerstein, Pui-Yan Kwok, Peter M Lansdorp, Gabor Marth, Jonathan Sebat, Xinghua Shi, Ali Bashir, Kai Ye, Scott E Devine, Michael Talkowski, Ryan E Mills, Tobias Marschall, Jan O Korbel, Evan E Eichler, and Charles Lee (2018). "Multi-platform discovery of haplotype-resolved structural variation in human genomes". en.

Dixon, Jesse R, Jie Xu, Vishnu Dileep, Ye Zhan, Fan Song, Victoria T Le, Galip Gürkan Yardımcı, Abhijit Chakraborty, Darrin V Bann, Yanli Wang, Royden Clark, Lijun Zhang, Hongbo Yang, Tingting Liu, Sriranga Iyyanki, Lin An, Christopher Pool, Takayo Sasaki, Juan Carlos Rivera-Mulia, Hakan Ozadam, Bryan R Lajoie, Rajinder Kaul, Michael Buckley, Kristen Lee, Morgan Diegel, Dubravka Pezic, Christina Ernst, Suzana Hadjur, Duncan T Odom, John A Stamatoyannopoulos, James R Broach, Ross C Hardison, Ferhat Ay, William Stafford Noble, Job Dekker, David M Gilbert, and Feng Yue (2018). "Integrative detection and analysis of structural variation in cancer genomes". en. In: *Nat. Genet.* 50.10, pp. 1388–1398.

Timp, Winston and Andrew P Feinberg (2013). "Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host". eng. In: *Nat. Rev. Cancer* 13.7, pp. 497–510.

Jeffares, Daniel C, Clemency Jolly, Mimoza Hoti, Doug Speed, Liam Shaw, Charalampos Rallis, Francois Balloux, Christophe Dessimoz, Jürg Bähler, and Fritz J Sedlazeck (2017). "Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast". en. In: *Nat. Commun.* 8, p. 14061.

medaka: Sequence correction provided by ONT Research. https://github.com/nanoporetech/medaka.

Krueger, Felix and Simon R Andrews (2011). "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications". en. In: *Bioinformatics* 27.11, pp. 1571–1572.

Hansen, Kasper D, Benjamin Langmead, and Rafael A Irizarry (2012). "BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions". en. In: *Genome Biol.* 13.10, R83.

Chapter 4

TERT Sequencing in Thyroid Cancer

4.1 Introduction

'Telomerase' refers to the ribonucleoprotein complex that functions to maintains the ends of chromosomes (telomeres). Telomeres become shortened during DNA replication due to the lack of a priming site for the lagging strand (Muraki et al., 2012), a conundrum for all linear chromosomes that has been termed the "end replication problem". Telomerase activity is maintained in human stem cells and germ line cells, but expression in healthy somatic cells is normally absent or occurring at very low levels (Muraki et al., 2012). Conversely, reactivation of telomerase is a very common feature of neoplasia, and occurs in over 90% of human cancers (Koziel et al., 2011). This re-activation of telomerase results in maintained telomere length through cell division, sustaining the cell proliferation of malignant cells. The core protein component of telomerase is the telomerase reverse transcriptase (TERT) catalytic subunit.

Expression of TERT has been found to be a good proxy for studying telomerase levels and activity (Yi, Shay, and Wright, 2001). Relevantly, changes in TERT expression have been linked to alterations in the epigenetic milieu around the TERT promoter, to include changes in chromatin accessibility, and in CpG methylation patterns (Castelo-Branco et al., 2013). Regulation of TERT by CpG methylation is somewhat unique in that methylation of specific loci within the TERT promoter region is associated with increased transcriptional activity (Devereux et al., 1999). This non-conventional change in methylation may partly be explained that normal cells are devoid of CpG methylation at the TERT locus (indicating that the gene is suppressed though alternative mechanisms). In cancer transformation, cells often alter patterns of methylation around the TERT promoter, presumably in response to changes in the signaling pathways and changes in the transcription factors occupying the TERT promoter. In cases of mutations on only one single TERT allele, there is evidence of increased widespread CpG methylation and recruitment of epigenetic silencing machinery (PRC2) on the inactive (non-mutated) allele (Stern et al., 2017).

We studied changes in CpG methylation at the TERT promoter specifically in the context of thyroid cancer, using normal thyroid tissue and thyroid-carcinoma-derived cell lines (McKelvey et al., 2020; Avin et al., 2019). TERT promoter mutations have been described in many human cancers, including thyroid cancer (Liu et al., 2013). The mutation profile at the TERT promoter has been characterized for the thyroid cell lines used in these studies (**Table 4.1**) . These mutations alter transcription factor binding which can directly

lead to this dysregulation of TERT repression. Some distinct mutations that alter transcription factor binding have been characterized: two especially common TERT promoter mutations, (C228T and C250T, named based on Hg19 coordinates) are positioned 124nt and 146 nt upstream of the start codon (Liu et al., 2013). Both of these mutations give rise to an 11-mer consensus binding sequence for GABPA/B1 (ETS transcription factor family). In vitro transcription experiments have shown that this leads to increased transcription levels (Huang et al., 2013), and is linked to changes in the histone signature to reflect a more active state (Stern et al., 2017). We studied specifically the effects of TERT promoter mutations in thyroid cancer, examining how these correlated with changes in CpG methylation patterns and transcriptional activity. The long reads of targeted nanopore sequencing bring a unique advantage, allowing us to phase point mutations to methylation patterns at sites tens of kilobases away (including at other regulatory regions). Anther advantage of nanopore sequencing for this application, is that the TERT region is high-GC density and low-complexity, making it challenging to map short sequencing reads, especially after bisulfite conversion. Bisulfite conversion is especially confounding for this C>T mutation in that it not only (1) reduces genome complexity, but also (2) obliterates the allele-specific C>T mutation, hindering allele-specific study of methylation.

4.2 Results

4.2.1 CpG Methylation Studies

For initial studies of methylation patterns of the TERT promoter, we used bisulfite amplicons (**Supplementary Table** 4.2). In order to query the entire minimal TERT promoter (1100nt) three tiled amplicons were used to query the region (**Figure** 4.1). This system was used to compare methylation patterns in normal thyroid tissue versus two papillary thyroid carcinoma (PTC) cell lines and two follicular thyroid carcinoma cell lines (FTC). For PTC cell lines, we used the BCPAP and TPC-1 cell lines. (TPC-1 is considered a well-differentiated PTC, whereas BCPAP is considered a model poorly-differentiated PTC) (Saiselet et al., 2012).

Our results were in in agreement line with previous observations that transformed cells often carry foci of increased methylation of the TERT promoter. The increase in methylation versus the normal control was most evident in the proximal gene body, adjacent to a CTCF binding motif. Additionally, for the TPC-1 cell line, the region just upstream of the TSS showed a very dramatic bimodal spike in methylation versus the normal line massive increase in methylation, the poorly differentiated BCPAP showed much more muted

Cell Lines	Thyroid Cancer Subtype	*TERT*
TPC-1	Papillary (Well differentiated)	-124 C>T heterozygous mutant
BCPAP	Papillary (Poorly differentiated)	-124/-125 CC>TT heterozygous mutant
FTC-133	Follicular (local LN metastasis)	-124 C>T homozygous mutant
FTC-238	Follicular (distant metastasis)	-124 C>T heterozygous mutant

Table 4.1: TERT mutations in cell lines TERT Mutation status and cellular subtype/description in four thyroid cell lines

changes in methylation in this region

We also applied this strategy to examine methylation patterns in the follicular thyroid carcinoma (FTC) cell lines; FTC-133 and FTC-238. Interestingly, both of these cell lines originate from the same thyroid cancer patient. The FTC-133 cell line was isolated from an early regional LN metastasis and the FTC-238 cell line from a late distal metastasis. Both of these FTC lines showed methylation increases within the proximal gene body (as was seen with the PTC lines). Conversely, the FTC lines showed no increase in methylation in the region just proximal to the TSS, but rather maintained low levels of methylation in that region similar to the levels in the normal samples (**Figure 4.1**).

We next followed up these studies by evaluating the methylation pattern in these cell lines using the nCATS targeted enrichment strategy described in the previous chapters. This allowed us to study methylation patterns for the entire proximal regulatory region on both the mutant and wildtype alleles (**Figure 4.2A**). In cell lines with heterozygous mutations, we observed the TERT TSS was demethylated specifically on the mutant allele. Another common feature of the heterozygous cell lines was a demethylation of the upstream MYC binding site on the mutated allele. We also note that at the site of the mutation (the GABPA binding site) that there are only slight differences in methylation, with a slight increase in methylation on the wildtype allele for the papillary cell lines (BCPAP and TPC-1) at this site. As mentioned in the previous chapter, this data can also be visualized at the single molecule level, which makes it possible to appreciate the heterogeneity of methylation

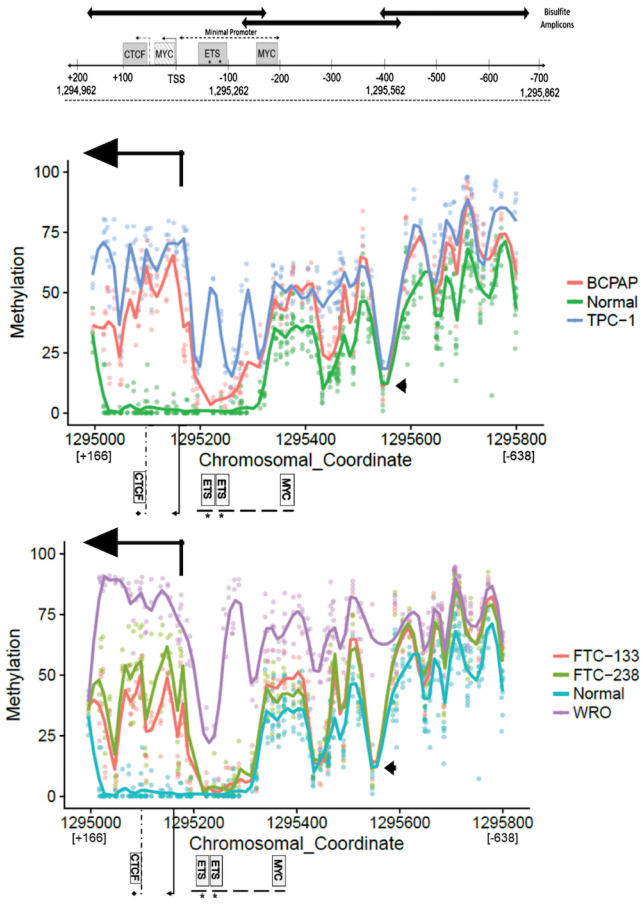


Figure 4.1: Tert Methylation with Bisulfite Amplicons Comparison of methylation patterns in thyroid cell lines and normal primary thyroid tissue. (Top) Methylation of normal thyroid tissue compared with two papillary thyroid carcinoma cell lines: BCPAP and TPC-1. (Bottom) Methylation of normal thyroid tissue compared with two follicular thyroid carcinoma cell lines: FTC-133 and FTC-238. The black arrow indicates the TERT transcriptional start site. Binding sites for CTCF, ETS, and MYC are shown along the X-axis of the methylation plots.

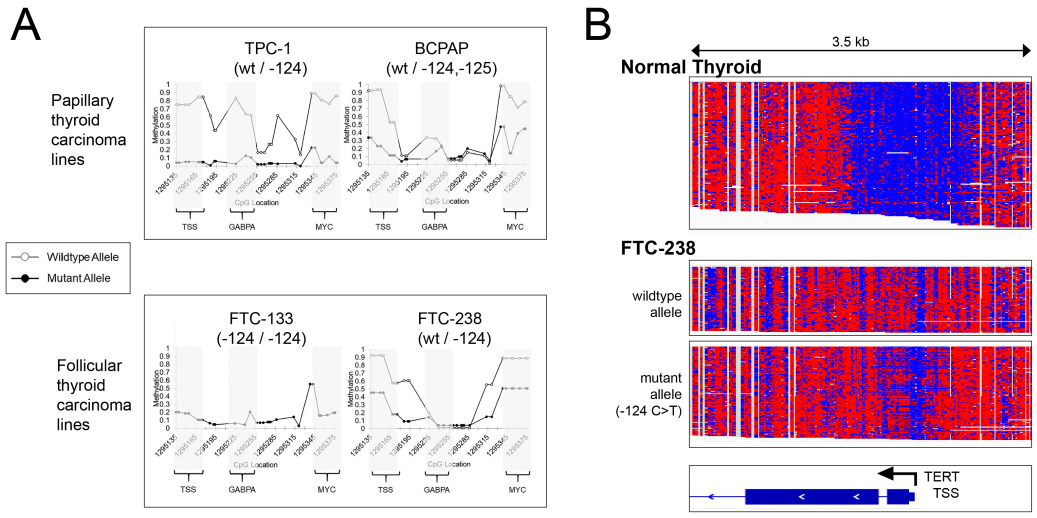


Figure 4.2: TERT methylation with nCATS Data (A) Methylation patterns on each of the alleles for (top) papillary thyroid carcinoma and (bottom) follicular thyroid carcinoma cell lines. (B) Read-level methylation plots of nCATS data from both normal thyroid tissue and the FTC-238 cell line.

patterns. For example in the FTC-238 cell line we see some mutant alleles with wide regions of demethylation around the TSS, while other mutant alleles showing higher levels of methylation and patterns that more closely mirror that seen on the wildtype allele (**Figure 4.2B**).

4.2.2 Chromatin Immunoprecipitation

We continued this investigation of TERT biology with chromatin immunoprecipitation (ChIP) studies to see if DNA methylated changes were correlated with increased presence of DNA binding proteins (namely, GABPA, CTCF, and MYC) in these transformed cell lines. We observed increases in binding of GABPA as well as CTCF binding relative to normal thyroid tissue (**Figure 4.3A**). Conversely, Myc appears to be associated with the upstream promoter

at decreased levels in thyroid cancer cell lines relative to normal thyroid tissue, which is at least partially explained by the increased methylation at the Myc binding site in the cancer cell lines. In order to interrogate which allele these DNA-binding proteins were associated to in heterozygous cell lines, we followed up the initial ChIP-qPCR studies with ChIP-Sanger sequencing. Here I highlight the results for the two cell lines with heteozygous single base changes: TPC-1 and FTC238. For both of these cell lines we observe that both Myc as well as GABPA appear to be binding nearly exclusively to the mutated allele (**Figure 4.3B**). Following up these studies with ChIP using antibodies directed against modified histones, we observed the repressive chromatin mark of lysine 27 trimethylation (H3K27me3) to be enriched at the wildtype TERT allele, contrasting with the H3K4me3 activation histone mod which was much more enriched in the mutated allele (**Supplementary Figure 4.5**).

4.2.3 TERT Transcriptional Analysis

We also compared levels of the TERT transcript in both the cell lines as well as healthy normal thyroid tissue. As anticipated, we found that transcript levels of TERT were elevated in all of the transformed lines (**Figure 4.4**). Levels for TERT appeared to be higher in the more aggressive cell lines. This was true in the case of both papillary thyroid carcinoma (the poorly differentiated BCPAP versus the well-differentiated TPC-1), as well as follicular thyroid carcinoma (the locally metastatic FTC-133 versus hematogenously spread FTC238). To show a direct link between the mutation, protein-binding, and transcription, we identified variants within the coding region for the heterozygous cell

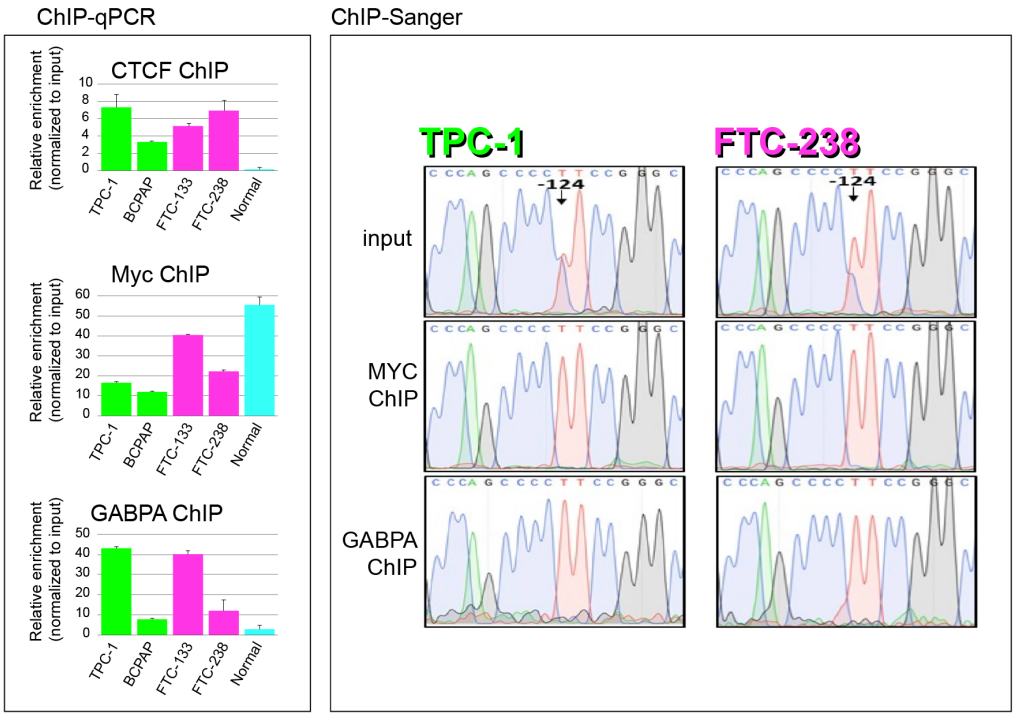


Figure 4.3: Chromatin Immunoprecipitation in Thyroid Cell lines (Left) ChIP-qPCR to comparing total binding of CTCF, Myc, and GABPA in thyroid cancer cell lines and normal thyroid. Enrichment is given relative to 5% input. (Right) ChIP-Sanger for Myc and GAPBA in the thyroid cell lines with a heterozygous single nucleotide change (TPC-1 and FTC-238)

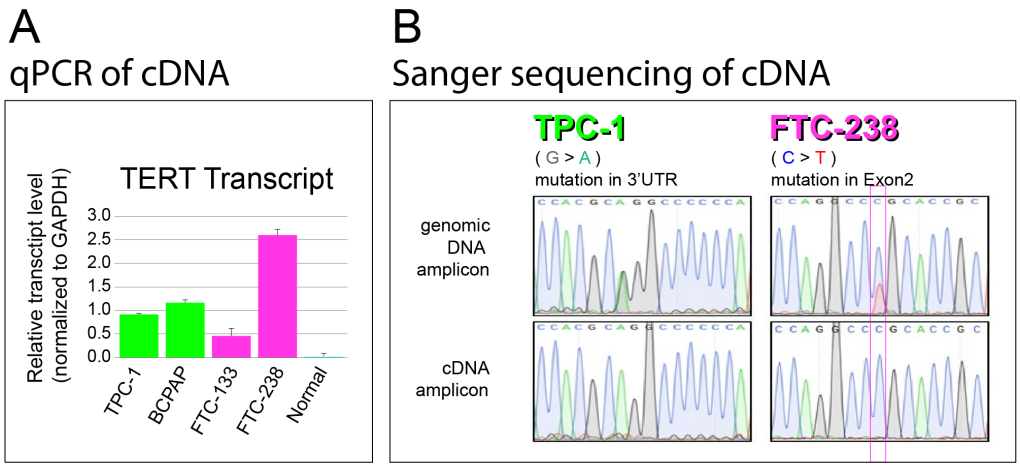


Figure 4.4: TERT Transcript Analysis (A) TERT relative transcript levels measured by qPCR. Values normalized to GAPDH. (B) Sanger sequencing of genomic DNA and transcript-derived cDNA in the TPC-1 and FTC-238 cell lines.

lines. For TPC-1, the variant used is a G>A mutation in Exon2. Both FTC-238 as well as the BCPAP cell lines contained a mutation in the 3' UTR of C>T. Sanger sequencing of amplicons generated either from genomic DNA or transcript-derived cDNA showed that only one of the alleles is giving near exclusive rise of the transcripts. This is shown for TPC-1 and FTC-238 cell lines in (**Figure** 4.4) and the data for BCPAP is shown in (**Supplementary Figure** 4.6). Further, because we have the long read data, we were able to phase the transcript mutation with the distal promoter mutations. This confirmed that the mutated promoter to be present on the same allele giving rise to the more abundant transcript.

4.3 Discussion

For this work we characterized changes in transcription factor binding at the TERT promoter in thyroid cancer and linked those changes to mutation, methylation patterns and transcriptional activity. Features of the TERT promoter methylome have already been used in clinical studies, as increased upstream methylation is linked to poor clinical outcomes and is used as a biomarker for cancer diagnosis in melanoma as well as some brain cancers (Lee, Borah, and Bahrami, 2017). It was previously understood that select methylation of the TERT promoter is correlated with increased transcription (Lopatina et al., 2003), and this work helps to highlight the complexity of the relationship between DNA methylation and TERT expression, and provide some potential mechanistic insight.

For instance, Comparing for instance the early and late stage cell lines FTC-133 and FTC-238, we see that the later stage has higher levels of TERT expression associated with increased methylation, specifically on the non-mutated allele. This implies that the cell may have activated suppresive signaling (i.e. DNA methylation) in response to TERT overexpression as an attempt to quell the aberration and return the cell to homeostasis.

The ChIP analysis, paired with transcriptional studies and allele specific methylation help to paint a clearer picture of how dysregulation may occur in these cell lines. The TPC-1, FTC-238, and FTC-133 cell lines (having mutations that create the consensus sequence for GAPBA binding), all showed dramatically increased amounts GAPBA binding. BCPAP demonstrated a much more modest increase in GAPBA binding, which can be attributed to

the unique mutation in this cell line which does not create the perfect consensus sequence and therefore a a decreased affinity for this transcription factor is anticipated. It is also possible that the poorly differentiated BCPAP cell line has developed alternative activation mechanisms for TERT and is no longer GAPBA-dependent. Interestingly, the observed increase in TERT expression appears to be inversely correlated with the amount of GABPA binding, implying that this might be an early mechanism of initiation of TERT activity.

This work confirmed that there is increased binding of GABPA to the mutant TERT promoter in thyroid cancer cell lines, as seen previously with other cancer types (Stern et al., 2017). This work added the new layer of information that Myc also can bind in an allele-specific manner. It had previously been understood that Myc was an activator of TERT (Wu et al., 1999), but its binding activity in the context of heterozygous mutations had not been observed previously. In fact, this work contradicts some previous work which claimed that Myc bound to the TERT promoter in a mutation-agnostic fashion (Liu et al., 2018). We believe this discrepancy to be the result of their experimental system which employed MYC knockdown and therefore has the potential to be confounded by changing non-local activity of this master transcription factor.

Finally, we linked the study of TERT regulation together with the transcriptional studies. By employing coding mutations we identified mono-allelic expression, and by connecting this with long-read data we confirmed the expression to be occurring from the mutated allele, providing direct evidence of

the TERT transcripts' allele-of-origin. Together this offers new insight into the mechanisms that regulate TERT expression in transformed cells, suggesting that the cells tend towards a state of monoallelic expression in the instance heterozygously mutated cell lines.

4.4 Methods

4.4.1 Cell lines, culture conditions, and TERT mutation status

Papillary [TPC-1, BCPAP] and follicular thyroid cancer cell lines [FTC-133, FTC-238] were kindly provided to the Umbricht lab by Dr. Motoyasu Saji (The Ohio State University Wexner Medical Center, Columbus, OH). All cell lines were authenticated using short tandem repeat profiling. The BCPAP cell line was grown in Roswell Park Memorial Institute-1640 (RPMI-1640; Sigma-Aldrich, St. Louis, Missouri) medium with 10% heat-inactivated fetal bovine serum (FBS) (GE Healthcare Life Sciences, Marlborough, Massachusetts), while TPC-1, FTC-133, and FTC-238 were grown in HyClone Dulbecco's Modified Eagle Medium (DMEM) (GE Healthcare Life Sciences) with 10% FBS. All cultures were supplemented with 1x antibiotic antimycotic solution (Sigma-Aldrich), 2 mM L-glutamine (Life Technologies, Carlsbad, California), and 1x MEM-Non-Essential Amino Acids (Quality Biological, Gaithersburg, Maryland) and maintained at 37°C and 5% CO2. For primary patient tissue (true normal fresh tissue), six frozen thyroid tissue samples were selected from our thyroid tissue bank, which has been approved by Institutional Review Board at Johns Hopkins Medical Institutions. Samples were immediately

placed on ice, resected and reviewed in the Department of Pathology, snap frozen in liquid nitrogen, and stored at -80degC until use. Five micrometre Cryostat H&E sections were obtained to dissect pathology from adjacent normal thyroid tissue.

4.4.2 DNA Isolation and Bisulfite Modification

Genomic DNA from cell lines and patient tissue was isolated by utilizing Proteinase K digestion, phenol/chloroform extraction, and ethanol precipitation. DNA was diluted in 1x Low-EDTA TE pH 8 (Quality Biological). Isolated DNA (100 ng) was used for bisulfite treatment using the EZ DNA Methylation-Lightning Kit, according to manufacturer's guidelines (Zymo Research, Irvine, California).

4.4.3 Bisulfite Sequencing PCR of TERT Promoter

Bisulfite modified DNA was amplified with three separate tiled primer sets listed in (**Supplementary Table 4.2**), amplifying the region -662 to +174 relative to the TERT TSS. The 50 uL PCR amplification reaction contained 5uL of bisulfite-treated DNA, 300nM of forward and reverse tile primers, and 25uL KAPA HiFi HotStart Uracil+ ReadyMix (2X) (KAPA Biosystems, Wilmington, Massachusetts) using the ProFlex PCR System (Applied Biosystems, Foster City, California) [2 minutes 95degC, 35 cycles × (20 seconds 98degC, 20 seconds (T1 58degC/T2 54degC/T3 54degC), 40 seconds 72degC), 2 minutes 72degC].

4.4.4 Bisulfite Library Preparation and Sequencing

PCR products from the bisulfite-treated TERT tiles were purified with Agencourt AMPure XP beads, using 1.8x beads (Beckman Coulter Life Sciences, Indianapolis, Indiana) and quantified using the Qubit dsDNA HS Assay Kit (ThermoFisher Scientific, Waltham, Massachusetts). The NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, Massachusetts) protocol was followed for library preparation. Because of the variance in on-target fraction in the PCR products, the amplicons were pooled with DNA input amounts of 20, 40 and 200 ng of, T1, T2, and T3, respectively. Library concentration and size distribution were quantified with the High Sensitivity DNA kit on the 2100 Bioanalyzer (Agilent, Santa Clara, California) and KAPA Library Quantification Kit for Illumina Platforms (KAPA). A 12 pM sample library for promoter bisulfite sequencing with 30% PhiX control was paired-end sequenced on the MiSeq v3 600 PE system (Illumina, San Diego, California) for a targeted minimum of 10,000 reads per amplicon.

4.4.5 Data Processing for Methylation

TERT promoter methylation status was analyzed with the Bismark (Krueger and Andrews, 2011) software package. Bisulfite conversion and mapping to the Human Genome Reference Consortium build 37 (GRCh37) was performed on trimmed reads (Trim Galore) (Krueger, 2015). Sequencing coverage for each CpG was analyzed with bsseq (Hansen, Langmead, and Irizarry, 2012). A CpG site was included in the analysis if the sequencing coverage was over 60 and present in more than two cell lines.

4.4.6 Chromatin Immunoprecipitation (ChIP) Analysis

For ChIP analysis, DTC cells were grown to 85% confluency in 15 cm plates. For ChIP analysis in normal thyroid tissue, the Novus Biologicals protocol was followed (Cat NBP1-71709). Cells were fixed for 8 minutes in 1% formaldehyde. Chromatin was sheared by sonication for 2x [4x 20s on/20s off] using the Bioruptor Pico (Diagenode, Denville, NJ). The following antibodies were used (10 ug per ChIP): anti-MYC (No. 9402; Cell Signaling Technology, Danvers, MA), anti-GABPA (No. 27795; Thermo Fisher, Waltham, MA), anti-H3K4me3 (No. 9727; Cell Signaling Technology), anti-H3K27me3 (No. 9733; Cell Signaling Technology). Antibodies were bound to proteinG Dynabeads (Thermo, 10003D). DNA was purified by MinElute PCR Purification Kit (Qiagen). ChIP and input DNA were analyzed by qPCR with SYBR Green and melt curve analysis. qPCR reactions were carried out in triplicate, with two biological replicates, and positive and negative control regions for each antibody. Factor binding was determined by the percentage of input normalization method, normalizing binding at locations surrounding the TERT promoter mutation and transcription start site (TSS). For Sanger analysis immunoprecipitated DNA and input DNA were amplified by PCR surrounding the TERT -124 C > T mutation and Sanger sequenced. Primer details are included in **Supplementary Table 4.2**.

4.4.7 TERT Expression qRT-PCR

RNA from the cell lines was extracted by Trizol isolation (ThermoFisher, 15596026). RNA from normal thyroid tissue was extracted using the Highpure

RNA kit (Roche, Basel, Switzerland) following the manufacturer's protocol. RNA was reverse transcribed to cDNA using SuperScript III Reverse Transcriptase (ThermoFisher) according to manufacturer instructions. Quantitative PCR was performed with SYBR Green PCR Master Mix (ThermoFisher), with primers designed to detect the full-length TERT transcript levels (Rao et al., 2005). The expression level of TERT was normalized to GAPDH .

4.4.8 Nanopore Cas9 targeted sequencing

Nanopore Cas9 targeted sequencing (nCATS) was conducted as described in the previous chapter. The target region had a size of 7.8 kb surrounding the TERT promoter at chr5:1,288,699–1,296,505. The guideRNA sequences used were: Forward "AAGGCTTAGGGATCACTAAG" and Reverse "AGCGGCAGGTGCCCAGAATA." Each sample was sequenced on a nanopore flow cell (version 9.4.1) using the GridION sequencer (Oxford Nanopore Technologies, Oxford, UK).

4.4.9 Data processing for methylation from nanopore data

Base calling to generate FASTQ reads was performed by the GUPPY algorithm. The resulting reads were aligned to the human genome, hg19, by Minimap2. CpG methylation calling was conducted using nanopolish. Reads were phased into wild-type or mutant allele by identifying the promoter motif in FASTQ reads.

4.4.10 Allele-specific transcription characterization

SNPs in the thyroid cancer cell lines were identified by Sanger sequencing of genomic DNA in the TERT 3′ untranslated region (UTR) and by integrative genome viewer examination of the nCATS data in the TERT 5′ UTR and exons 1 and 2. Isolated RNA from the cell lines was DNase I treated and reverse transcribed as above, then amplified at the SNP location (**Supplementary Table** 4.2) and Sanger sequenced utilizing the forward primer to determine the genotype.

4.5 Supplementary Materials

Tiled primers for bisulfite sequencing of the TERT promoter

F 5' GGGTTTGTGTTAAGGAGTTTAAGT 3' R 5' CATAATATAAAAACCCTAAAAACAAAT 3'
F 5' TTGTTTTTAGGGTTTTTATATTATGGT 3' R 5' AAACTAAAAAATAAAAAAACAAAAC 3'
F 5' GGTATTYGTTTTGTTTTTTTATTTTT 3' R 5' CACCAACCRCCAACCCTAAA 3'

Primers for ChIP-qPCR and ChIP-Sanger Analysis of TERT

F: CACCCGTCCTGCCCCTTCA R: CTGCCTGAAACTCGCGCC

Primers for Allele-specific mRNA Analysis

Exon2	F: CGTGGTTTCTGTGTGGTGTC R: CCTTGTCGCCTGAGGAGTAG
3'UTR	F: CACTGCCCTCAGACTTCAA R: GGGATGGACTATTCCTATGTGG

Table 4.2: Primers for TERT Analysis Sequences for the primers used for ChIP-qPCR, ChIP-Sanger, and cDNA-Sanger sequencing of TERT

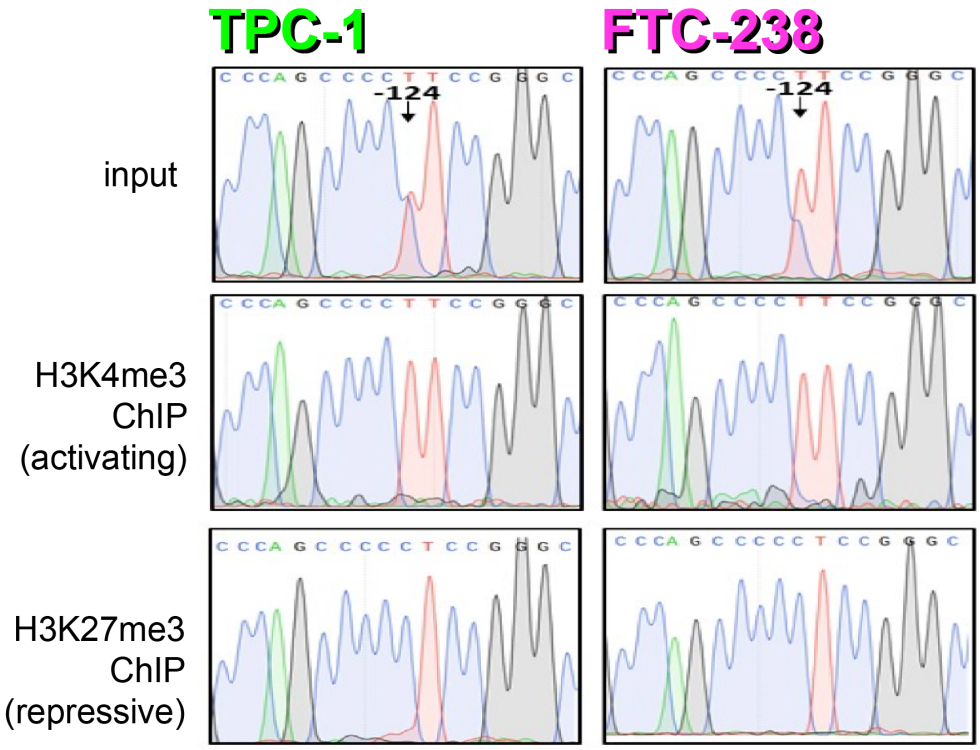


Figure 4.5: ChIP-Sanger for Histone Modifications ChIP-Sanger for H3K4me3 and H3K27me3 in the thyroid cell lines with a heterozygous single nucleotide change (TPC-1 and FTC-238).

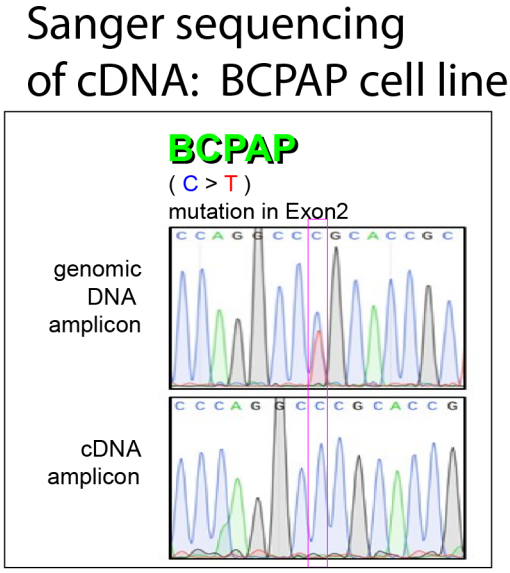


Figure 4.6: Sanger sequencing of BCPAP cDNA Exon 2 sequencing of TERT cDNA, containing a heterozygous mutation in BCPAP.

References

Muraki, Keiko, Kristine Nyhan, Limei Han, and John P Murnane (2012). "Mechanisms of telomere loss and their consequences for chromosome instability". en. In: *Front. Oncol.* 2, p. 135.

Koziel, Jillian E, Melanie J Fox, Catherine E Steding, Alyssa A Sprouse, and Brittney-Shea Herbert (2011). "Medical genetics and epigenetics of telomerase". en. In: *J. Cell. Mol. Med.* 15.3, pp. 457–467.

Yi, X, J W Shay, and W E Wright (2001). "Quantitation of telomerase components and hTERT mRNA splicing patterns in immortal human cells". en. In: *Nucleic Acids Res.* 29.23, pp. 4818–4825.

Castelo-Branco, Pedro, Sanaa Choufani, Stephen Mack, Denis Gallagher, Cindy Zhang, Tatiana Lipman, Nataliya Zhukova, Erin J Walker, Dianna Martin, Diana Merino, Jonathan D Wasserman, Cynthia Elizabeth, Noa Alon, Libo Zhang, Volker Hovestadt, Marcel Kool, David T W Jones, Gelareh Zadeh, Sidney Croul, Cynthia Hawkins, Johann Hitzler, Jean C Y Wang, Sylvain Baruchel, Peter B Dirks, David Malkin, Stefan Pfister, Michael D Taylor, Rosanna Weksberg, and Uri Tabori (2013). "Methylation of the TERT promoter and risk stratification of childhood brain tumours: an integrative genomic and molecular study". en. In: *Lancet Oncol.* 14.6, pp. 534–542.

Devereux, T R, I Horikawa, C H Anna, L A Annab, C A Afshari, and J C Barrett (1999). "DNA methylation analysis of the promoter region of the human telomerase reverse transcriptase (hTERT) gene". en. In: *Cancer Res.* 59.24, pp. 6087–6090.

Stern, Josh Lewis, Richard D Paucek, Franklin W Huang, Mahmoud Ghandi, Ronald Nwumeh, James C Costello, and Thomas R Cech (2017). "Allele-Specific DNA Methylation and Its Interplay with Repressive Histone Marks at Promoter-Mutant TERT Genes". en. In: *Cell Rep.* 21.13, pp. 3700–3707.

McKelvey, Brittany A, Timothy Gilpatrick, Yongchun Wang, Winston Timp, Christopher B Umbricht, and Martha A Zeiger (2020). "Characterization of

Allele-Specific Regulation of Telomerase Reverse Transcriptase in Promoter Mutant Thyroid Cancer Cell Lines". en. In: *Thyroid*.

Avin, Brittany A, Yongchun Wang, Timothy Gilpatrick, Rachael E Workman, Isac Lee, Winston Timp, Christopher B Umbricht, and Martha A Zeiger (2019). "Characterization of human telomerase reverse transcriptase promoter methylation and transcription factor binding in differentiated thyroid cancer cell lines". en. In: *Genes Chromosomes Cancer* 58.8, pp. 530–540.

Liu, Xiaoli, Justin Bishop, Yuan Shan, Sara Pai, Dingxie Liu, Avaniyapuram Kannan Murugan, Hui Sun, Adel K El-Naggar, and Mingzhao Xing (2013). "Highly prevalent TERT promoter mutations in aggressive thyroid cancers". en. In: *Endocr. Relat. Cancer* 20.4, pp. 603–610.

Huang, Franklin W, Eran Hodis, Mary Jue Xu, Gregory V Kryukov, Lynda Chin, and Levi A Garraway (2013). "Highly recurrent TERT promoter mutations in human melanoma". en. In: *Science* 339.6122, pp. 957–959.

Saiselet, Manuel, Sébastien Floor, Maxime Tarabichi, Geneviève Dom, Aline Hébrant, Wilma C G van Staveren, and Carine Maenhaut (2012). "Thyroid cancer cell lines: an overview". en. In: *Front. Endocrinol.* 3, p. 133.

Lee, Seungjae, Sumit Borah, and Armita Bahrami (2017). "Detection of Aberrant TERT Promoter Methylation by Combined Bisulfite Restriction Enzyme Analysis for Cancer Diagnosis". en. In: *J. Mol. Diagn.* 19.3, pp. 378–386.

Lopatina, Nadejda G, Joseph C Poole, Sabita N Saldanha, Nathaniel J Hansen, Jason S Key, Mark A Pita, Lucy G Andrews, and Trygve O Tollefsbol (2003). "Control mechanisms in the regulation of telomerase reverse transcriptase expression in differentiating human teratocarcinoma cells". en. In: *Biochem. Biophys. Res. Commun.* 306.3, pp. 650–659.

Wu, K J, C Grandori, M Amacker, N Simon-Vermot, A Polack, J Lingner, and R Dalla-Favera (1999). "Direct activation of TERT transcription by c-MYC". en. In: *Nat. Genet.* 21.2, pp. 220–224.

Liu, Rengyun, Tao Zhang, Guangwu Zhu, and Mingzhao Xing (2018). "Regulation of mutant TERT by BRAF V600E/MAP kinase pathway through FOS/GABP in human cancer". en. In: *Nat. Commun.* 9.1, p. 579.

Krueger, Felix and Simon R Andrews (2011). "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications". en. In: *Bioinformatics* 27.11, pp. 1571–1572.

Krueger, Felix (2015). "Trim galore". In: *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files* 516, p. 517.

Hansen, Kasper D, Benjamin Langmead, and Rafael A Irizarry (2012). "BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions". en. In: *Genome Biol.* 13.10, R83.

Rao, Annavarapu Srinivas, Peter E Goretzki, Josef Köhrle, and Georg Brabant (2005). "Letter Re: Id1 gene expression in hyperplastic and neoplastic thyroid tissues". en. In: *J. Clin. Endocrinol. Metab.* 90.10, p. 5906.

Chapter 5

Additional Applications and Future Directions of Targeted Sequencing

5.1 Introduction

There are additional ways that strategies using Cas9 targeted sequencing can be applied to explore features of DNA. This chapter will discuss several avenues of additional study that we have begun to apply. This will focus on three extensions of this method: (1) locating sites of random gene integration (2) generating targeted chromatin accessibility data and (3)using panels of tiled guideRNAs to query large genomic regions for structural variant discovery. It is worth noting that the 2020 outbreak of coronavirus led to a limitation in our ability to complete some benchwork experiments, and therefore all of the projects in this chapter have future planned work to be carried forward at a later time either by myself or others members of the Timp lab.

5.2 Gene Localization

5.2.1 Background

The insertion of genes into expression systems is a commonly used technique in the biopharmaceutical industry (Romanova and Noll, 2018). When proteins are produced in a mammalian system, the expression vector of choice is largely the Chinese Hampster Ovary (CHO) epithelial cell (Stolfa et al., 2018). Conventionally, an expression vector containing the transgene and associated promoter elements is delivered to the cells and randomly integrates into the CHO genome. Single-cell clones are selected from this transfection, followed by a largely empirical process to identify clones that are expressing the protein of interest (Stolfa et al., 2018). Often it can be difficult to identify the precise location of these inserts, making it challenging to characterize the epigenomic landscape of the gene insertion. This is further complicated by the potential for the transgene to concatamerize during integration, leading to tangential repeats of the insert. Often the insertions can be unstable, leading to inconsistent gene expression and loss of productivity over time. The ability to characterize the insert would be valuable in understanding this variation and the propensity of inserts to inactivate.

This project was carried out with industry collaborators to express antibodies in the CHO cell line. Due to proprietary limitations, some details of the data can not be publicly divulged. For this reason, locations of the inserted genes are censored from this thesis. Our industry collaborators also declined to offer us the the full sequence of the expression vector, and we therefore had

to decipher this on our own. What we were provided with was the sequence of the light chain and heavy chain genes for this antibody. We received DNA from clonal CHO cell lines with varying expression levels, with randomly inserted plasmid(s) containing the light chain and heavy chain genes.

5.2.2 Results

As before, guideRNAs were designed to create cuts in the sequence of interest. Only this time, the guides were directed "outwards" to generate sequencing information out into the plasmid sequence, with the hopes that some reads would further extend into the CHO genomic sequence (**Figure** 5.1). This was performed on four cell lines. Two of these clones were characterized as 'stable' in their expression the transgene, and two others were described as 'unstable'. The four cell lines were sequencing in a single MinION flow cell run, each sample barcoded using Oxford Nanopore's native barcoding kit. This resulted in 2000-4000 total sequencing reads per sample (**Table** 5.1).

CHO Clone Name	Total reads per sample
Stable2	2238
Stable3	4052
Unstable2	4108
Unstable3	4901

Table 5.1: Read Counts for CHO insertions CHO clone identifier and total sequencing reads for the corresponding sample in a multiplexed MinION sequencing run

The first challenge was to reconstruct the sequence of the plasmid using

the targeted sequencing data. If, as anticipated the plasmid contatamerized during insertion into the the CHO genome, it is expected that many sequencing reads containing only vector (**Figure** 5.1). Also, because the plasmid can be orientated either forwards or backwards during this process, smaller fragments that contain only parietal sequences can result from Cas9 cleavage. To determine the vector sequence, we first pulled out only the reads which contained the sequence of the known gene insert. For this step, we used the data from the "Stable3" clone, where we identified 407 reads (10.0%) containing the transgene sequence. Looking at the read length distribution lengths showed a main peak around 12kb and other smaller peaks at 7kb and 20kb. We fed the data to the Flye assembly tool (Kolmogorov et al., 2019), resulting in multiple assemblies of varying length. We evaluated the different assemblies for completeness of the two transgenes (heavy chain and light chain). Because of plasmid discontiguity or incomplete/duplicated gene inserts, we identified the assembly most likely to represent the original vector. By adding back in the missing piece of the heavy chain gene, we arrived at an assembly 11.5 kb in size. Additional software tools Racon (Vaser et al., 2017) and Medaka (ONT) were further applied to polish this draft of the assembly sequence (**Figure** 5.2) (see Methods).

Once we had the sequence for the plasmid in hand, the next steps were to identify locations where this gene insertion had occurred. To identify reads potentially informative about insert location, we mapped the plasmid against the 'on-target' reads, and identified reads which contained non-plasmid sequence at their 3' end (we set a cut-off of at least 50nt of non-vector sequence

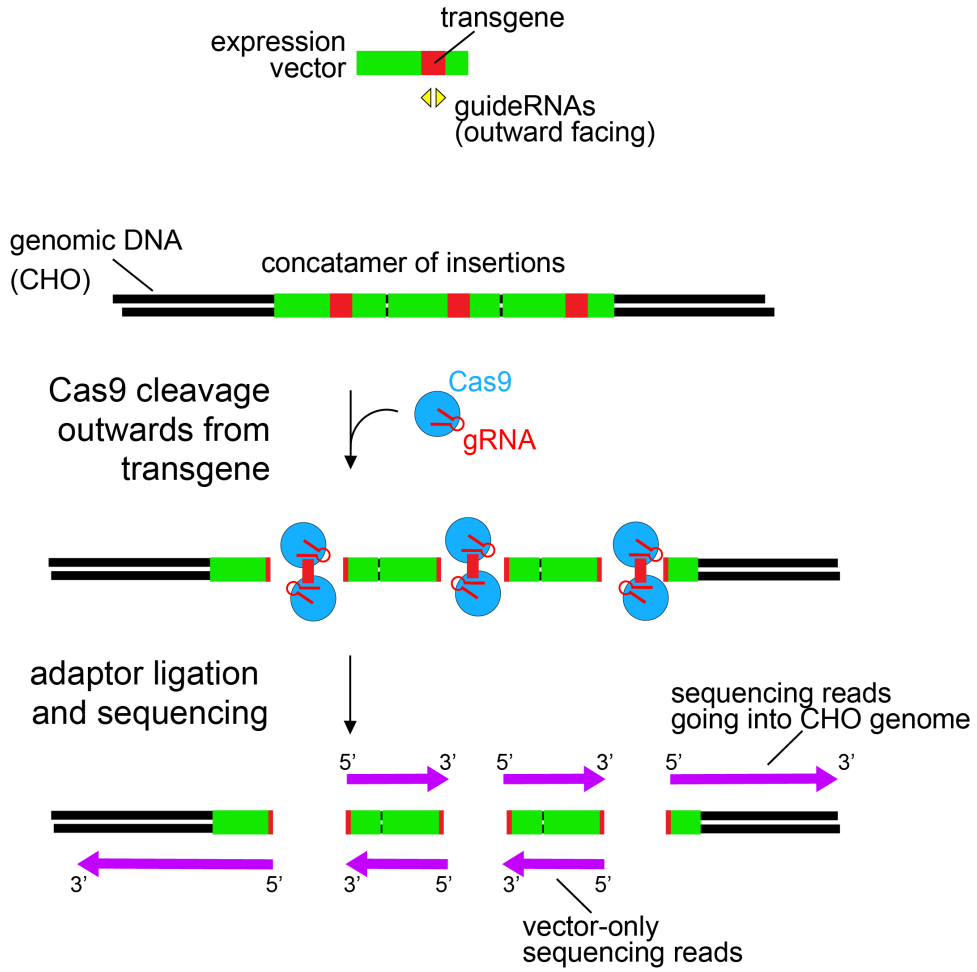


Figure 5.1: Cas9 Targeted Sequencing for Insert Localization Cartoon showing the strategy and resulting reads using targeted cleavage to locate insertion points

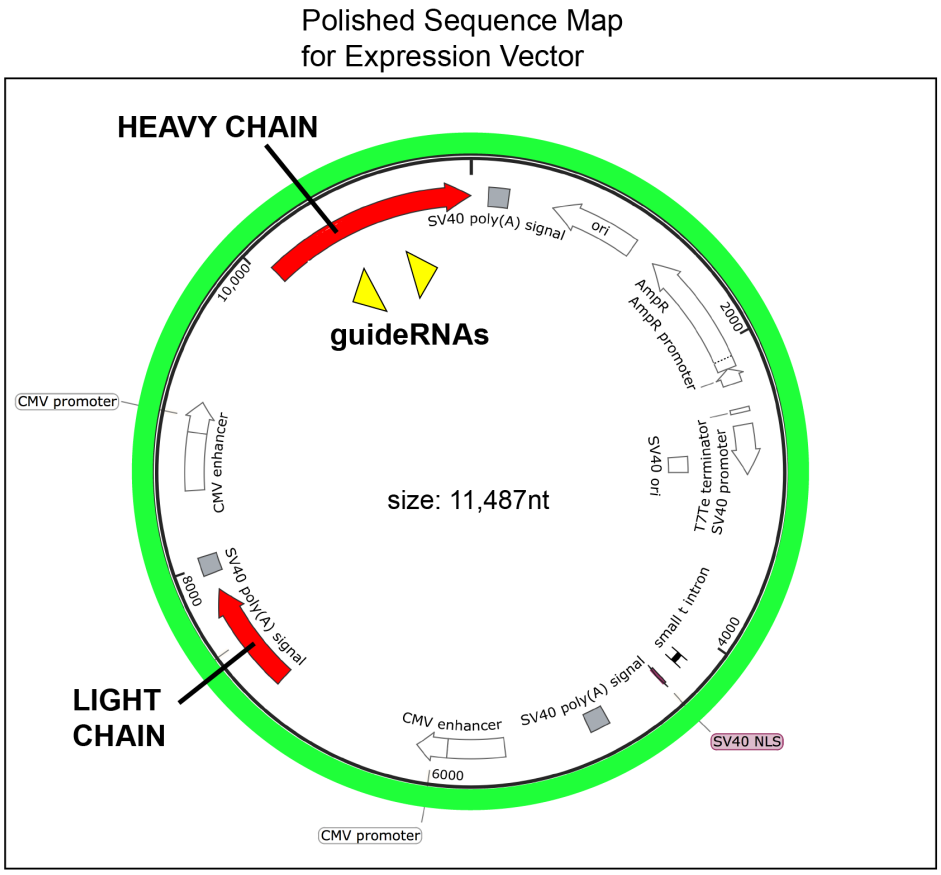


Figure 5.2: Expression Vector Map Map showing the genes and regulatory elements in the assembled expression vector sequence

in the read). This led to a substantial attrition in the sequencing reads, leading to potentially informative read counts of 6 (Stable2), 5 (Stable3), 4 (Unstable2) and 2(Unstable3). The non-plasmid sequence from these reads was then aligned back to the CHO genome. These alignments were few enough that we were able to manually inspect each mapping site. We identified that low-quality mapping to the CHO genome often reflecting artefacts of sequence homology and were able to eliminate those. This resulted in the identification of two sites of genomic integration into the CHO genome. Interestingly, the two insertion sites were common between all four clones. As an additional complication, the insertion sites were identified to be highly repetitive regions. Fortunately, we had enough reads extending into less repetitive regions, allowing unambiguous identification of insertion sites.

5.2.3 Discussion

The positioning of the genes within these low-complexity regions, underscores the challenging task of finding insertion locations with short-read sequencing. With conventional sequencing methods, the rare short reads that would contain both plasmid and genomic sequence would not extend far enough out into the genome to enable precise mapping of the insert location.

Now that we have identified sites of gene insertion, next steps for this project are to sequence inwards from the insertion points. The goal is that this will (1) enable us to get information about the insertion concatemer, full characterizing its structure and (2) allow us to compare methylation patterns both between the insertion sites and between different copies of the gene at

one insertion site. This will help us to understand the role that methylation plays in gene insertion, and could prove a valuable tool in the biopharma industry for evaluating and selecting transfected CHO clones.

5.3 Targeted NanoNOME

5.3.1 Background

As our understanding of DNA regulation develops, many approaches have been created that explore how a chromosomes are organized in the nucleus. Existing assays look at many features of the DNA including nucleosome occupancy, chromatin accessibility, CpG methylation, and transcription factor binding. One such method, NOME-seq (Kelly et al., 2012), uses an exogenous GpC methyltransferase (M. CviPI) to label regions of the genome with a modification not normally present in mammallian cells. When paired with bisulfite conversion, this permits detection of nucleosome-free regions of the genome, as only accessible regions will be GpC labeled. My colleague Isac Lee performed extensive work to adapt the NOME-seq work flow to the nanopore sequencing platform, a strategy that we now call "nanoNOME" (**Figure 5.3**) (Lee et al., 2019). There were a number of technical hurtles to overcome for this analysis. Importantly, although detection of cytosine methylation had already been validated by our lab in the CpG context, a model was needed for validation of the GpC context. The building of this model this led to the development of a kernel-smoothing method, which helped to deal with the noisy GpC methylation data, making it much possible for the first time to identify nucleosomal occupancy directly on nanopore sequencing reads(Lee

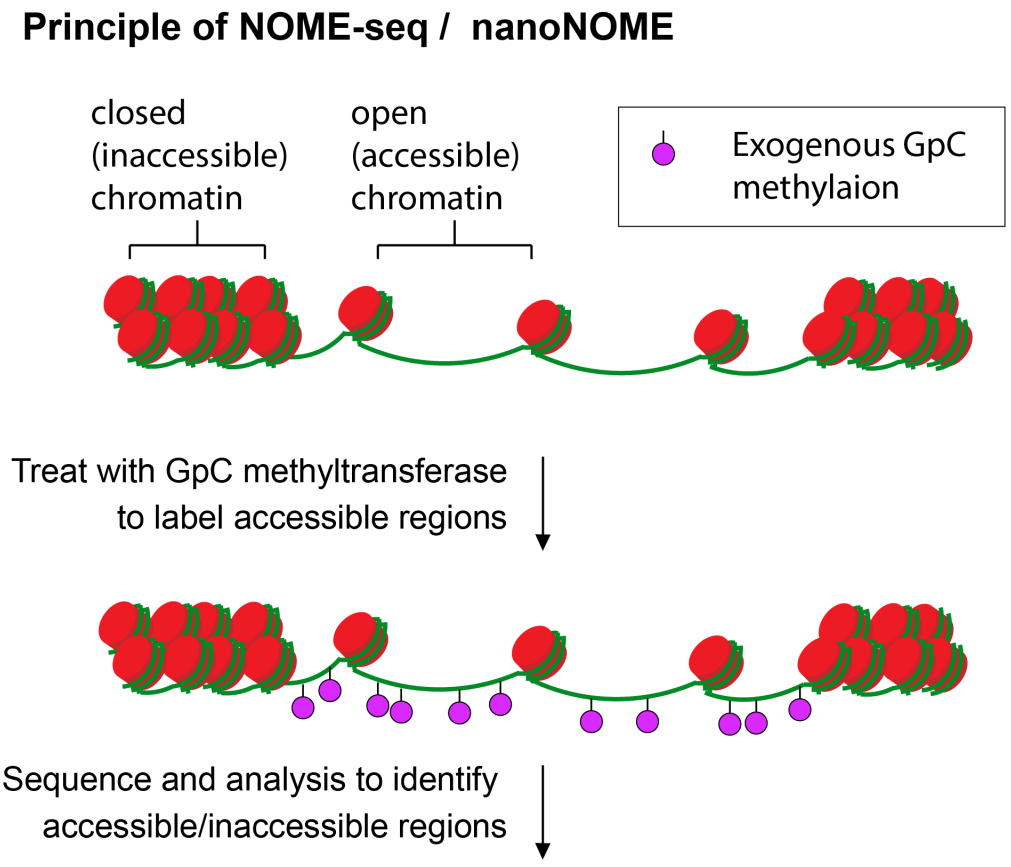


Figure 5.3: NanoNOME Schematic Acccessible regions of the DNA are labeled with exogenous modification (methyl-GpC) when incubated with a methyltransferase (M. CviPI) and methyl donor group (SAM)

et al., 2019).

This avenue offers another useful application for targeted sequencing studies, as there is currently a paucity of methods for studying chromatin accessibility at select loci. To perform initial testing and validation of the Cas9 targeted sequencing approach in conjunction with the nanoNOME pipeline, we applied the same set of guideRNAs that had been used for the breast cancer studies. We then compared methylation and accessibility patterns with

whole-genome NanoNOME data, and explored the interplay between CpG methylation and DNA accessibility at enriched genes.

5.3.2 Results

The nanoNOME protocol is sensitive to input DNA amounts. To account for this, we used only 1ug of input DNA, less than the 3ug usually used for Cas9 targeted nanopore sequencing. Despite this reduction in input DNA amounts, we found that we were still able to generate significant enrichment over background DNA using this method. Through this method we achieved average on-target max coverage of 52X. This corresponded to an 'on-target' rate of 9.0% from the 10,000 reads achieved (**Table** 5.2).

Total read: 10291
On-target reads: 931 (9.0%)

Region	Max local coverage with targeted NanoNOME
GPX1	30
SLC12A4	87
KRT19	63
GSTP1	45
TPM2	75
ch 5 del	50
ch7 del	35
BRAF	33
KRAS	47
TP53	59

Table 5.2: Targeted NanoNOME, regional coverage The max coverage achieved within the targeted for the respective gene

To evaluate whether the methylation calls were maintained during this

targeting study, we compared the calls for methylation and accessibility to whole genome data nanoNOME data for the GM12878 cell line (Lee et al., 2019). We found that the accessibility calls as well as the CpG methylation calls agreed strikingly well between targeted nanoNOME data and whole-genome nanoNOME data. Line plots for all regions are shown in (**Supplementary Figure** 5.7). Some regions queried demonstrated a clear inverse relationship between the CpG methylation calls and chromatin accessibility. Two example genes demonstrating this phenomena are shown in (**Figure** 5.4), where we see increased accessibility and decreased CpG methylation at the transcription start site for KRAS and GPX1.

GpC accessibility is inherently noisy, a feature which is further exacerbated by the error profile of nanopore sequencing data. In order to better visualize the accessibility data, we applied the kernel smoothing function developed for nanoNOME data (Lee et al., 2019). This greatly aids in our ability to appreciate the data at the single molecule level while evaluating for protein occupancy. We specifically applied this to the enrichment data at the KRAS locus (**Figure** 5.5), where we see evidence of nucleosome depletion at the highly accessible transcriptional start site, in addition to being able to assess the diversity of accessibility patterns on the sequencing reads.

5.3.3 Discussion

This work combining accessibility studies with targeted sequencing provides a range of new opportunities to evaluate accessibility with high coverage. This provides a facile way to simultaneously study DNA accessibility and CpG

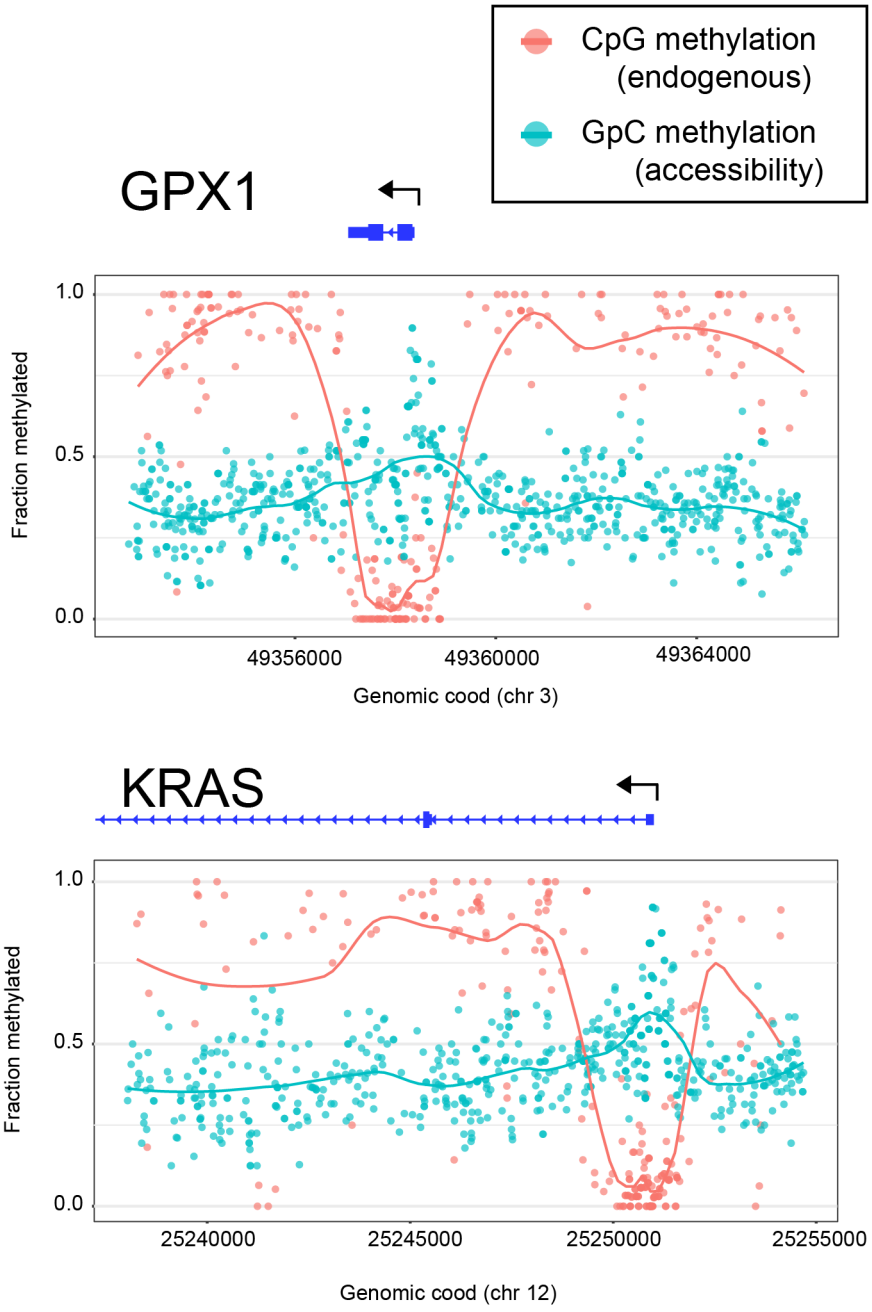


Figure 5.4: NanoNOME Data at KRAS and GPX1 CpG methylation and accessibility (nanoNOME) at the KRAS and GPX1 genes in the GM12878 cell line

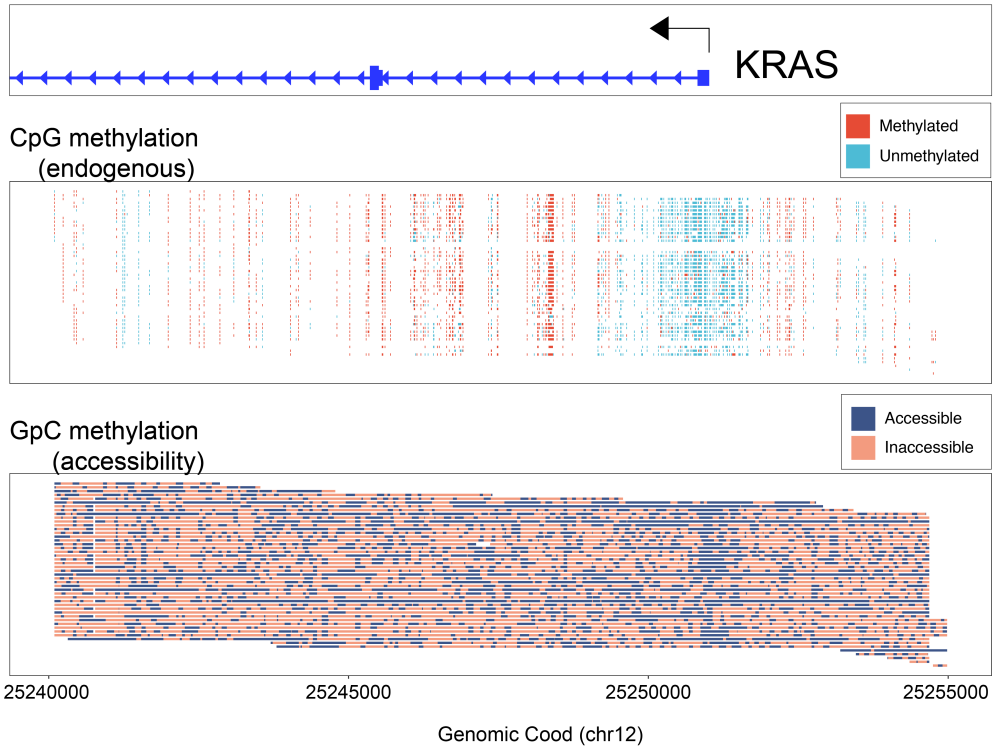


Figure 5.5: Kernel-Smoothed NanoNOME Read-level CpG methylation and kernel-smoothed accessibility data at the KRAS promoter

methylation, this has the potential to greatly reduce the cost for performing accessibility studies, which are usually performed genome-wide. Future applications of planned for this application is to monitor changes in accessibility during cell fate transition (e.g. differentiation) or during response to external stimuli (e.g. hormones or small molecules). By monitoring accessibility changes at the single-molecule level we can appreciate within-sample heterogeneity and monitor how accessibility changes over time, providing deeper insight into chromatin regulatory mechanisms.

5.4 In vitro transcribed guideRNAs

5.4.1 Background

In the study of breast cancer in earlier chapters, we used targeted sequencing to investigate for the presence of structural variants with *a priori* knowledge of deletion breakpoint locations. Often, when investigating structural variation, the goal is to discover breakpoint locations *de novo*. Commonly, these breakpoints will occur in locations known to be "hotspots" for chromosomal abnormalities. This is true in pancreatic ductal adenocarcinoma (PDAC), wherein we sought to investigate two tumor suppressor genes affected by this mechanism: p16 (Cdkn2a) (Caldas et al., 1994) and SMAD4 (Hahn et al., 1996). Commonly, these genes lose function in PDAC as a result of structural variation (deletions, insertions, translocations, inversions, etc). The ability to rapidly identify precise locations of patients' structural abnormalities makes it possible to design personalized assays that can surveil at high sensitivity for the rearrangement, and detect cancer recurrence. Pursuing that goal,

this section seeks to extend the targeting sequencing strategy to discovery of structural abnormalities in large genomic "hotspot" regions.

In PDAC, the locations of chromosomal breakpoints around p16 and SMAD4 is known to occur over many megabases. To increase the chances of having reads on both sides potential breakpoints, we enriched for an area with a total size of 14 megabases (9Mb flanking p16 and 5Mb flanking SMAD4). We tiled guideRNAs across the regions, spacing one guideRNA about every 10kb. To avoid the costs of synthesizing individual guideRNAs, we generated the guide RNAs using in vitro transcription (IVT). Through a collaboration with Agilent Technoloies, we designed a pool of 1100 guideRNA templates tiled across the 14Mb. After amplification of the guideRNA templates, in vitro transcription of the 1100 guideRNAs was performed in a single reaction for use with the Cas9 targeted nanopore sequencing assay.

5.4.2 Results

We performed initial experimentation with IVT-guideRNAs using the GM12878 cell line. After some optimization, an initial sequencing run using 5ug of DNA generated 670,000 reads, with greater than 200,000 'on-target' (30.7%). As discussed previously in chapter 3, guideRNAs have varying performance due to intrinsic sequence properties as well as uniqueness of the target sequence. This was evinced again in the data using in vitro transcribed guideRNAs, which showed highly variable coverage over the enriched areas. For example, in the 500kb window surrounding the target genes, on-target coverage ranged from a few reads to >700X coverage (**Figure 5.6**).

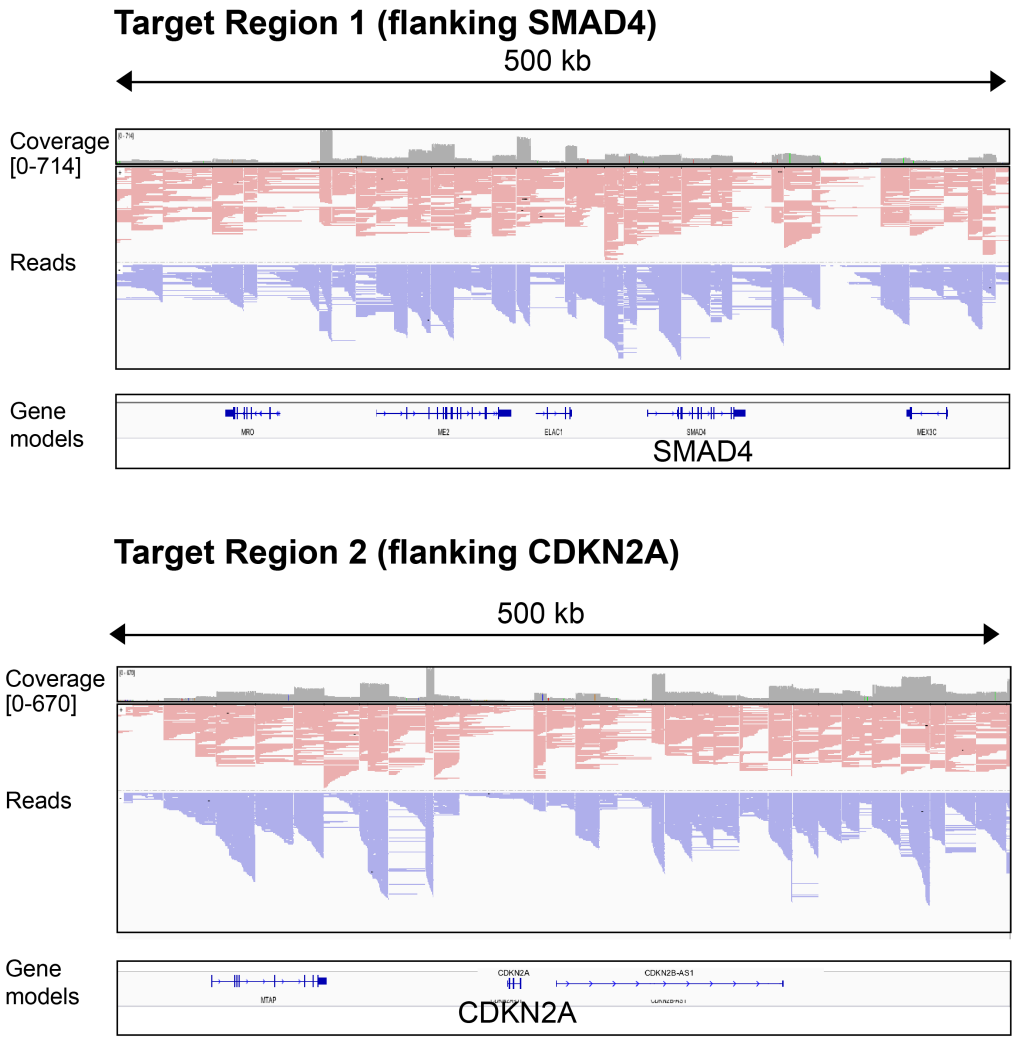


Figure 5.6: In vitro transcribed guideRNA panel Coverage and reads for a 500kb enriched region around genes commonly harboring deletions in pancreatic ductal adenocarcinoma (PDAC)

5.4.3 Discussion

Our studies with a panel of IVT guides demonstrate that with continued improvement this strategy could interrogate large regions for chromosomal aberrations. This IVT-guideRNA approach could also be adapted for a rapid in vitro screening test to evaluate guideRNA cutting efficiency. We note that in transformed PDAC cells annotated breakpoints frequently reside within repetitive regions. This again underscores the advantage provided by long-reads; which increase the chances that sequencing reads will extend into uniquely mapping genomic fragments. Future work in the Timp lab is implementation of this pipeline to with PDAC-derived DNA and development of analysis workflows for rapid structural variant identification.

5.5 Methods

5.5.1 Sequencing

Sequencing libraries were prepared as discussed in previous chapter with the following modifications. For the CHO gene-insertion sequencing, multiple samples were multiplexed using Oxford's native barcoding kit (ONT, NBD-104). For targeted nanoNOME studies the samples were first treated with GpC methyltransferase as detailed below. And for in vitro transcription studies, the guideRNAs were produced in house as described below. Samples were run on a MinION (ver 9.4.1) flow cell, using the MK1B or GridION sequencer.

5.5.2 Locating Gene Insertions

GuideRNAs were placed within the sequence of the antibody heavy-chain transgene. Reads were aligned to the transgene sequence using minimap2 (Li, 2018) and the paf format output was parsed using custom python scripts to identify reads containing vector sequence. The expression vector was was assembled using the Flye tool (Kolmogorov et al., 2019) and polished using Racon (Vaser et al., 2017) and Medaka (ONT), as described in the previous chapters. The plasmid sequence was next aligned to sequencing reads using minimap2 to identify reads that contained non-vector sequence at their 3′ end. This 3′ non-vector sequence was aligned to reference CHO genome using minimap2.

5.5.3 GpC methyltransferase treatment for targeted nanoNOME

GM12878 suspension cells were snap frozen in 1.5mL tubes (roughly 1 million cells per tube). Cells were thawed into resuspension buffer (100 mM Tris-Cl, pH 7.4, 100 mM NaCl, 30 mM MgCl2) with 0.25 % NP-40 for 5 minutes on ice. Intact nuclei were collected by centrifugation for 5 minutes at 500xg at 4degC. The nuclei were treated with a solution of 1x M. CviPI Reaction Buffer (NEB), 300 mM sucrose, 96 uM S-adenosylmethionine (SAM; New England Biolabs, NEB), and 200 U M. CviPI (NEB) in 500 uL volume per 500,000 nuclei. The reaction mixture was incubated at 37degC with shaking on a thermomixer at 1,000 rpm for 15 minutes. S-adenosyl methionine (SAM) was replenished at 96 uM at 7.5 minutes into the reaction. The reaction was stopped by the addition of an equal volume of stop solution (20 mM Tris-Cl, pH 7.9, 600 mM NaCl,

1% SDS, 10 mM disodium EDTA). Samples were treated with proteinase K (NEB) at 55degC for > 2 hours, and DNA was extracted via phenol:chloroform extraction and ethanol precipitation.

5.5.4 Generating in vitro transcribed guideRNAs

The panel of guideRNAs was delivered as a DNA library with a T7 binding site for in vitro transcription. The library was amplified using Kapa Hifi ReadyMix (Roche, KK2501). The PCR product was cleaned up using a MinElute kit (Qiagen, 28004). The in vitro transcription was performed using a kit from New England Biolabs (NEB, E2050S). This reaction was cleaned up using the Monarch RNA CleanUp kit (NEB, T2040S). GuideRNAs were used immediately or frozen at -80degC.

5.6 Supplementary Material

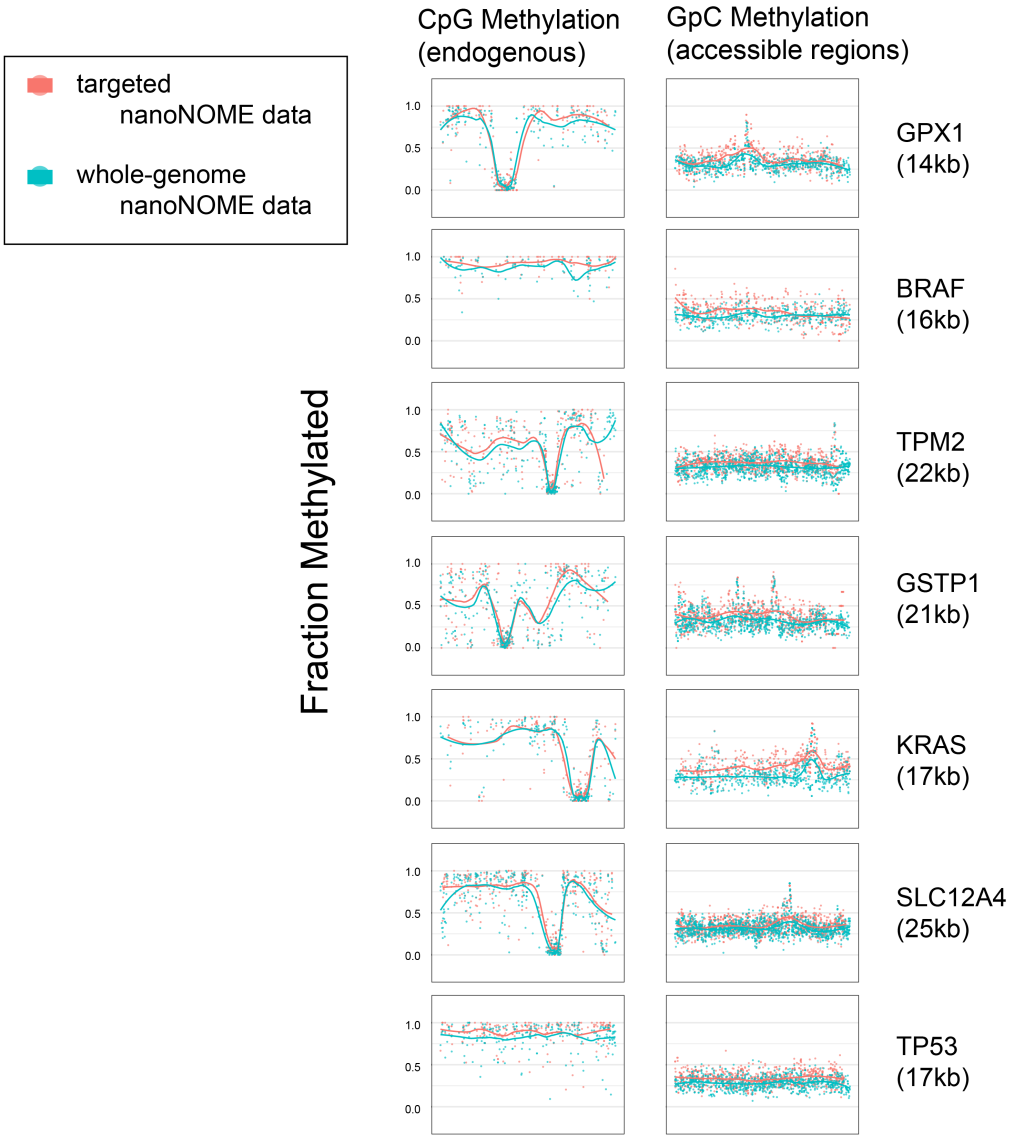


Figure 5.7: Targeted versus whole-genome nanoNOME Comparing CpG methylation and GpC methylation calls between targeted and whole-genome nanoNOME data

References

Romanova, Nadiya and Thomas Noll (2018). "Engineered and Natural Promoters and Chromatin-Modifying Elements for Recombinant Protein Expression in CHO Cells". en. In: *Biotechnol. J.* 13.3, e1700232.

Stolfa, Gino, Matthew T Smonskey, Ryan Boniface, Anna-Barbara Hachmann, Paul Gulde, Atul D Joshi, Anson P Pierce, Scott J Jacobia, and Andrew Campbell (2018). "CHO-Omics Review: The Impact of Current and Emerging Technologies on Chinese Hamster Ovary Based Bioproduction". en. In: *Biotechnol. J.* 13.3, e1700227.

Kolmogorov, Mikhail, Jeffrey Yuan, Yu Lin, and Pavel A Pevzner (2019). "Assembly of long, error-prone reads using repeat graphs". en. In: *Nat. Biotechnol.* 37.5, pp. 540–546.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić (2017). "Fast and accurate de novo genome assembly from long uncorrected reads". en. In: *Genome Res.* 27.5, pp. 737–746.

Kelly, Theresa K, Yaping Liu, Fides D Lay, Gangning Liang, Benjamin P Berman, and Peter A Jones (2012). "Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules". en. In: *Genome Res.* 22.12, pp. 2497–2506.

Lee, Isac, Roham Razaghi, Timothy Gilpatrick, Michael Molnar, Norah Sadowski, Jared T Simpson, Fritz Sedlazeck, and Winston Timp (2019). "Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing". en.

Caldas, C, S A Hahn, L T da Costa, M S Redston, M Schutte, A B Seymour, C L Weinstein, R H Hruban, C J Yeo, and S E Kern (1994). "Frequent somatic mutations and homozygous deletions of the p16 (MTS1) gene in pancreatic adenocarcinoma". en. In: *Nat. Genet.* 8.1, pp. 27–32.

Hahn, S A, M Schutte, A T Hoque, C A Moskaluk, L T da Costa, E Rozenblum, C L Weinstein, A Fischer, C J Yeo, R H Hruban, and S E Kern (1996). "DPC4,

a candidate tumor suppressor gene at human chromosome 18q21.1". en. In: *Science* 271.5247, pp. 350–353.
Li, Heng (2018). "Minimap2: pairwise alignment for nucleotide sequences". en. In: *Bioinformatics* 34.18, pp. 3094–3100.

Chapter 6

Discussion and Conclusion

This thesis summarizes much of the work done in the Timp Lab during my pre-doctoral training. This work has sought to explore new applications for long-read sequencing – finding ways that these tools can be applied to gain increased insight into molecular biology.

This work focused on the use of a targeting strategy with CRISPR/Cas9 (Gilpatrick et al., 2020) to achieve greater coverage at desired loci with native DNA nanopore sequencing. I described applications of this strategy applied to query loci relevant to neoplasia (Gilpatrick et al., 2020; McKelvey et al., 2020; Avin et al., 2019). In detailing that work, I discuss several features of the cancer genome that we explored – including CpG methylation, mutations, and structural variations. Notably, the work in breast cancer also provided full-length unamplified sequencing reads of the BRCA1 gene, which is commonly disrupted in breast cancer, and difficult to interrogate with short-read methods (Welcsh and King, 2001). The work in thyroid cancer was focused on the study of the TERT gene. Targeted genomic sequencing of TERT was paired with chromatin immunoprecipitation and transcript studies to show that the

alterations observed in the genome and epigenome of these cells was reflected directly into changes in protein binding and transcriptional activity. Both of these studies also unveiled the prevalence of allele-specific methylation and near-exclusive use of one allele in transformed cells.

I also discussed several future and ongoing applications of targeted long-read sequencing being done in the Timp Lab. The targeted sequencing approach has myriad applications to clinical medical, scientific industry and basic science research. I discussed how we identified insertion sites of a transgene in mammalian cells for biopharma production. This could be applied in the industry to streamline the identification and selection of clones for the production of biological therapeutics. I talked about combining targeted sequencing with the study of chromatin accessibility, adapting the nanoNOME strategy developed by my colleague Isac Lee (Lee et al., 2019). This offers a rapid protocol for generating targeted accessibility data, making it possible for researchers to explore chromatin state without generating whole-genome data. As the repertoire of modifications that can be distinguished from nanopore signal continues to grow, there could even be more potential future layers added on to this data (e.g. information about nuclear localization or protein binding). Having all these layers of information on a single sequencing read will help provide new mechanistic insight into how these features are working in concert to regulate chromatin biology.

Another future direction I discuss is the combination of large numbers of guideRNAs (>1100) to investigate wide regions for DNA rearrangements.

There remain a large number of regions that are difficult to assembly without long-read data (e.g. BRCA1 and MHC), and a method to generate reads throughout these regions would aid both investigators and clinicians in understanding the diversity of chromosomal aberrations present in such challenging and repetitive regions.

In summary, the ability to achieve high-sequencing depth (such as with the targeted methods described herein) is advantageous in the new genomic insight offered, and the new questions that can be asked. The high coverage data is especially useful when trying to identify rare events, build assemblies from noisy sequencing data, or identify mutations by consensus. The applications of these strategies are of course not limited to those described here, as future iterations and development will continue to expand both the depth of coverage and the number of the features that we are able to interrogate via targeted nanopore sequencing.

References

Gilpatrick, Timothy, Isac Lee, James E Graham, Etienne Raimondeau, Rebecca Bowen, Andrew Heron, Bradley Downs, Saraswati Sukumar, Fritz J Sedlazeck, and Winston Timp (2020). "Targeted nanopore sequencing with Cas9-guided adapter ligation". en. In: *Nat. Biotechnol.*

McKelvey, Brittany A, Timothy Gilpatrick, Yongchun Wang, Winston Timp, Christopher B Umbricht, and Martha A Zeiger (2020). "Characterization of Allele-Specific Regulation of Telomerase Reverse Transcriptase in Promoter Mutant Thyroid Cancer Cell Lines". en. In: *Thyroid*.

Avin, Brittany A, Yongchun Wang, Timothy Gilpatrick, Rachael E Workman, Isac Lee, Winston Timp, Christopher B Umbricht, and Martha A Zeiger (2019). "Characterization of human telomerase reverse transcriptase promoter methylation and transcription factor binding in differentiated thyroid cancer cell lines". en. In: *Genes Chromosomes Cancer* 58.8, pp. 530–540.

Welcsh, P L and M C King (2001). "BRCA1 and BRCA2 and the genetics of breast and ovarian cancer". en. In: *Hum. Mol. Genet.* 10.7, pp. 705–713.

Lee, Isac, Roham Razaghi, Timothy Gilpatrick, Michael Molnar, Norah Sadowski, Jared T Simpson, Fritz Sedlazeck, and Winston Timp (2019). "Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing". en.