Leveraging New Technologies to Explore Regulatory and Structural Elements of the Human Genome

by

Torrin McDonald

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics)
in The University of Michigan
2021

Doctoral Committee:

Associate Professor Alan Boyle, Chair
Professor Sally Camper
Associate Professor Peter Freddolino
Assistant Professor Sue Hammoud
Professor Trisha Wittkopp

Torrin McDonald
torrin@umich.edu
ORCID iD: 0000-0001-5670-9375

To my family and friends

ACKNOWLEDGEMENTS

The work in this dissertation could not be possible without the collective support of many incredible humans. To all of the members of the Boyle lab who have made my graduate experience enjoyable, I cannot thank you enough. Your upbeat attitudes and enthusiasm for research and learning have helped keep me positive throughout the years. A special shout out to my demi-decadal partners in pipetting, Jessica, Sierra, and Mel, who have helped me in so many ways, and have made the wetlab feel like a home away from homes. I will profoundly miss the coffee walks, happy hours, and late nights in the lab.

To my McDonald Clan, Ian, Katie, and Emily McDonald, I could not have finished this work, or many things in my life, without your unwavering support. Your encouragement, kindness, and family comradery have always lifted my spirits on the darkest of days and I will forever cherish you. To my family and friends, you have helped me surmount so many obstacles, large and small. To my father and mother, thank you for nurturing my burgeoning interests in biology from childhood, and for your enduring support all these years. To my friends Becca, Tyler, David, Adam, and Deborah, thank you for the scientific discussions, laughs, adventures, and late night video games. To Karen and John Van Eck who have been unendingly generous and kind, this work would have been impossible without you both.

To my many mentors at the University of Michigan and in the Department of Human Genetics, thank you for your guidance and instruction. To Dr. Peter Freddolino, thank you for your insights into molecular biology, and for sharing your lab and resources for my projects. To Dr. John Moran, Dr. Anthony Antonellis, and the Genetics Training Program for their careful instruction about research and critical thinking. To my thesis committee for their immense support and guidance over the years. Thank you to the Department of Human Genetics administrative staff, Molly Martin, Sue Kellogg, Kim White, and all others, whose ongoing efforts to maintain the department and allow the students to focus on their education.

Lastly, a special thanks to my thesis advisor, Alan, for helping to channel our shared zeal for new technology and methods into exciting projects. It has been the experience of a lifetime to grow with the lab and make lifelong friends. Thank you for close mentorship and building a lab culture that is friendly, engaging, and fun. I will deeply miss participating in lab parties, inflatable gladiator competitions, escape rooms, VR dance-offs, and potlucks. I am proud to have been a student of your lab.

TABLE OF CONTENTS

DEDICATION . . . . **ii**

ACKNOWLEDGEMENTS . . . . **iii**

LIST OF FIGURES . . . . **viii**

LIST OF TABLES . . . . **x**

ABSTRACT . . . . **xi**

CHAPTER

I. Introduction 1

1.1 Basic Genome Organization: A Brief History 2
1.1.1 Nucleosome Positioning and Chromatin States 4
1.1.2 Transcription Factor Binding and Gene Regulation 6
1.1.3 Transcription Factors and Disease Associations 7
1.2 Approaches to Characterizing Genomic Features 8
1.2.1 Methods for Characterizing Open Chromatin 8
1.2.2 Methods to Measure Transcription Factor Binding 10
1.2.3 Transcription Factor Footprinting 11
1.2.4 Limitations to Characterizing Gene Regulatory Elements 12
1.3 Repetitive Sequence Composition of the Human Genome 13
1.3.1 Transposable Elements 13
1.3.2 Transposable Elements and Disease 15
1.3.3 Limitations to Characterizing Mobile Element Insertions 16
1.3.4 Polynucleotide Repeats and Associated Repeat Expansion Disorders 17
1.3.5 Limitations to Characterizing Structural Variation 18
1.4 Conclusion 18

II. Isolating and Sequencing Protein-Occupied Open Chromatin Regions . . 20

2.1 Abstract 20
2.2 Introduction 20
2.3 Methods 23
2.3.1 Isolation of Crosslinked K562 Nuclei 23
2.3.2 Lysis and Nuclease Digestion of Crosslinked K562 Nuclei 24
2.3.3 Nitrocellulose Filter Binding of K562 Nuclear Lysate 24
2.3.4 Pan-histone Immunodepletion of Nucleosomal Complexes 25
2.3.5 Reverse Crosslinking and Purification of DNA 26
2.4 Results 27
2.4.1 Simultaneous Isolation and Formaldehyde Crosslinking of K562 Nuclei 27

2.4.2 Recovery and Sequencing of DNA from Filter Bound, Crosslinked K562 Nuclear Lysates Shows Subnucleosomal-Sized Fragments . . . 30
2.4.3 Isoelectric Point Analysis of Transcription Factors Revealed a Subset of Nitrocellulose Incompatible Transcription Factors 31
2.4.4 Acidic Nitrocellulose Filter Binding of Crosslinked K562 Nuclear Lysates Reveals Enhanced Binding to Membrane 34
2.4.5 Pan Histone Immunodepletion Depletes Mononucleosomal Fragments and Enriches for Low Molecular Weight DNA 35
2.4.6 Sequencing of Pan-histone Immunodepleted Digested Luclear Lysates 36
2.5 Discussion 38
2.6 Notes and Acknowledgements 40

III. Cas9 Targeted Enrichment of Mobile Elements Using Nanopore Sequencing 41

3.1 Abstract 41
3.2 Introduction 42
3.3 Results 45
3.3.1 Cas9 Targeted Enrichment Strategy for Mobile Elements Using Nanopore Sequencing 45
3.3.2 Cas9 Targeted Enrichment Efficiently Captures Mobile Element Signals in Nanopore Reads 52
3.3.3 Cas9 Enrichment and Nanopore Sequencing Rapidly Saturates Reference and Non-Reference MEIs 55
3.3.4 Detectable Transmission of Non-reference L1Hs Within a Trio . . . 62
3.3.5 Cas9 Enrichment and Nanopore Sequencing Captures Non-reference Mobile Elements in Complex Genomic Regions 64
3.4 Discussion 70
3.5 Methods 76
3.5.1 Cell Culture, Counting, and Genomic DNA Isolation 76
3.5.2 Design of Unique Guide RNAs for L1Hs, AluYb, AluYb, SVA_F, and SVA_E 77
3.5.3 On-target Boundary Calculations for MEIs 78
3.5.4 In Vitro Transcription of Guide RNA and Cas9 Ribonucleoprotein Formation 78
3.5.5 Cas9 Enrichment for L1Hs on a MinION Flow Cell 79
3.5.6 Pooled Cas9 Enrichment for L1Hs, AluYb, AluYa, SVA_F, and SVA_E in GM12878 (MinION) 81
3.5.7 Cas9 Enrichment for Single MEI Subfamily on a Flongle Flow Cell 82
3.5.8 Cas9 Enrichment for L1Hs in Trio (MinION) 83
3.5.9 Nanopore Flow Cell Preparation, Sequencing, Base-calling, and Cleavage-site Analysis 84
3.5.10 Nano-Pal for Detection and Refinement of MEIs from Nanopore Cas9 Enrichment 85
3.5.11 GM12878 Trio Data, Reference Genome, and Reference MEI Information 86
3.5.12 Enhanced PALMER for Resolving Non-Reference MEIs from Whole-genome Long-read Sequencing 87
3.5.13 MEI Callsets in Orthogonal Short-read and Long-read Data 88
3.5.14 Inspection and Validation of Nanopore-specific Non-reference MEIs 88
3.5.15 Non-reference MEIs Captured by Nanopore Cas9 Enrichment Sequencing in GM12878 90
3.5.16 Analysis of L1Hs CpG Methylation 90

3.5.17 Data Availability 91
3.5.18 Code Availability 91
3.6 Notes and Acknowledgements 91
3.7 Author Contributions 92

IV. Enrichment and Sequencing of Polynucleotide Repeat Expansions 93

4.1 Abstract 93
4.2 Introduction 93
4.3 Methods 98
4.3.1 dCas9 Pulldown of L1Hs and CGG Trinucleotide Repeats 98
4.3.2 Paralleled Cas9 Enrichment of Multiple Repeat Loci 99
4.3.3 In Vitro Transcription of Guide RNAs 100
4.3.4 Cell Culture and DNA Extraction 100
4.4 Results 101
4.4.1 dCas9 Efficiently Enriches for a Triplet CGG Trinucleotide Repeat and L1Hs in GM12878 DNA 101
4.4.2 Cas9 Enrichment and Nanopore Sequencing on a Flongle Captures Repeat Associated Disease Loci in GM12878 102
4.5 Discussion 105
4.6 Notes and Acknowledgements 109

V. Conclusions and Future Directions 110

5.1 Conclusions and Future Directions 110
5.1.1 Improving unbiased isolation of transcription factor bound sequences 111
5.1.2 Characterizing Repetitive Elements Using Targeted Enrichment and Nanopore Sequencing 112
5.1.3 Concluding remarks 115

BIBLIOGRAPHY 116

LIST OF FIGURES

Figure

1.1	Levels of eukaryotic chromatin organization.	3
1.2	Euchromatin and heterochromatin.	5
1.3	Basic mechanism of retrotransposition.	15
2.1	Protocol for isolating transcription factor bound sequences.	22
2.2	Histogram of cell and nuclei size.	28
2.3	A brightfield/DAPI overlay of live cell and isolated nuclei mix.	29
2.4	Tapestation of DNA fragment size distribution isolated from basic nitrocellulose filter binding.	31
2.5	Heatmap of sequencing data from basic nitrocellulose filter binding.	32
2.6	Histogram of human transcription factor isoelectric points.	33
2.7	Heatmap of reads from DNA isolated from acidic nitrocellulose filter binding.	34
2.8	Tapestation of DNA fragment size distribution isolated from histone immunodepleted lysates.	36
2.9	Heatmap of reads from DNA isolated from supernatant of histone immunodepleted fraction.	37
3.1	A schematic Cas9 targeted enrichment and nano-pal pipeline for mobile elements using nanopore sequencing.	46
3.2	Guide RNA design for MEIs and guide RNA cleavage-site distribution.	48
3.3	Five final guide RNAs for MEIs	49
3.4	Distributions of guide RNA cleavage-Site for five MEI subfamilies.	50
3.5	Read length distributions for MEI categories (L1Hs, AluY, and SVA).	52
3.6	Number of supporting reads for three categories of non reference MEIs from the Cas9 targeted nanopore sequencing and the whole-genome nanopore sequencing by Ewing et al. 2020.	54
3.7	Venn diagram of the PALMER callset and the PAV callset for non reference MEIs in NA12878 genomes.	56
3.8	Systematic evaluation of known MEIs captured by nanopore Cas9 enrichment approach in different flow cells.	59
3.9	Summary of recovered known reference and non-reference MEIs.	60
3.10	Non-reference MEIs captured by nanopore Cas9 enrichment approach.	65
3.11	MEI distributions in various number of flow cells.	66
3.12	Trio transmission of 194 non-reference L1Hs captured by nanopore in GM12878 sample	68
3.13	Seventeen nanopore-specific MEIs have Hallmarks of retrotransposition consistent with bona fide insertions.	69
3.14	Examining CpG methylation of captured L1Hs reads.	73
4.1	Molecular workflow for dCas9-mediated capture of polynucleotide repeat regions.	96
4.2	dCas9 captures putative CGG repeats for nanopore sequencing.	102
4.3	Flanked Cas9 enrichment for 48 disease-associated repeat loci in GM12878.	103
4.4	Read structure and level of enrichment for Fuchs endothelial corneal dystrophy (FECD) associated repeat.	104

4.5 Read structure and level of enrichment for hand-foot-genital syndrome (HFGS) associated repeat. 104
4.6 Asymmetric Cas9 enrichment for 48 disease-associated repeat loci in GM12878: experiment one. 105
4.7 Asymmetric Cas9 enrichment for 48 disease-associated repeat loci in GM12878: experiment two. 106

LIST OF TABLES

Table

3.1 Summary of seven representative flow cells. 53
3.2 Known MEIs captured by nanopore Cas9 enrichment approach in different flow cells based on different boundaries. 57
3.3 Enrichment of mobile element signals in nanopore reads from GM12878 trio L1Hs experiments. 63
4.1 Table of targeted disease-associated repeat expansion diseases. 97

ABSTRACT

Advances in characterizing regulatory and structural regions of the genome have provided important mechanistic insights into how they influence gene regulation. Similarly, modern genetics better comprehends how these regions influence misregulation and manifest in disease states. Despite the unprecedented leaps in technology and genome annotation, particular aspects of genomic regulatory and structural elements continue to elude us. My dissertation is focused on developing and leveraging novel technologies to expand our understanding of these regulatory and structural elements, which collectively make up 50% of the human genome.

In chapter 1, I discuss the basics of genomic organization, composition, and regulation, as well as the various techniques developed to measure them. In chapter 2, I explore development of a novel open chromatin assay to measure transcription factor bound open chromatin regions. I developed novel nuclei isolation, crosslinking, lysis, and nucleosome depletion techniques to attempt to isolate transcription factor bound DNA. I show that nucleosomes can be robustly depleted from digested nuclear lysates, but further work is necessary to specifically isolate transcription factor bound DNA.

In chapter 3, I applied a novel adaptation of a Cas9-based enrichment and nanopore sequencing technique to capture transposable elements in the human genome. I showed that transposable elements are readily captured using this technology. These polymorphic and reference transposable elements were almost fully saturated for in

single experiments. In addition, a subset of these transposable element insertions were unique to our experiments, and were neither annotated in the reference or anticipated polymorphisms.

In chapter 4, I developed two separate nuclease-based enrichment strategies for polynucleotide repeats. The first was a biotinylated dCas9 approach that directly targeted triplet CGG repeats and efficiently precipitated repetitive DNA regions for nanopore sequencing. In addition, I adapted the Cas9 enrichment method from chapter 3 to efficiently capture 48 distinct, disease-associated polynucleotide repeat expansion loci throughout the genome. These experiments establish a new approach for interrogating clinically relevant disease-associated repeat expansions. Together, my dissertation expands our understanding of human genome structure, and adds new methods to the tool belt of modern genomics research.

CHAPTER I

Introduction

Gene regulation in humans is a complex interplay between transcription factors and sequence elements that control the spatio-temporal patterning of gene expression. This regulatory control relies on proximal transcription factor binding as well as distal chromatin interactions driven through 3-dimensional conformation of the genome. Illuminating gene regulatory controls is fundamental to our interpretation of their dysregulation in human disease. Gene dysregulation arising through variation in the human population may manifest in the form of single nucleotide variants, or through large structural variation. In this thesis, I will explore new technologies to both characterize and map functional variation: from regulatory control of genes to structural changes that can drive dysregulation within an individual and across populations.

To better map transcription factors and their requisite binding within regulatory regions, I first developed an orthogonal open chromatin assay to enrich for protein-occupied open chromatin regions (chapter 2). In this chapter, I explored a variety of approaches to enrich for transcription factor-bound DNA by the depletion of nucleosomal DNA. I leveraged filter binding assays, nuclease digestions, and pan-histone immunodepletions to efficiently deplete nucleosomal DNA, while preserving DNA fractions that would represent genome-wide transcription factor binding. In addition, I aimed to develop techniques to better map changes to the physical structure of the genome from mobile element insertions (chapter 3) and short tandem repeat expansions (chapter 4). As discussed below, these changes to the genome can have dramatic effects on gene regulation and human health. Due to their highly repetitive nature and abundance in the human genome, mapping and characterizing the full spectrum of these classes of structural variation has been technologically challenging. I implement novel enrichment strategies paired with long read sequencing to fully saturate subsets of mobile element insertions and disease-associated polynucleotide repeat regions. These projects attempt to improve our understanding of gene regulation both at the level of individual regulatory elements as well as through larger scale genomic changes. The work presented here leverages novel enrichment approaches to sequence regions of the genome known to influence gene regulation

1.1 Basic Genome Organization: A Brief History

The diploid human genome is approximately 2 meters in length, and yet it is compacted into the cell's nucleus, an organelle with an average diameter of $6x10^{-6}$ meters. This extensive compaction is accomplished by double stranded DNA wrapping around an octamer of histone proteins (H2A, H2B, H3, and H4), which are further wound into coiled fibers (Figure 1.1). The fundamental unit of chromatin, originally referred to as "PS particles" or "v-bodies", is now known as the nucleosome. To package the genome, 147 bp of DNA wraps around an octamer of the core histone proteins in 1.7 superhelical twists [1]. A repeated pattern is achieved when individual nucleosomes are connected by a stretch of linker DNA, giving rise to the "beads on a string" appearance. Linker DNA is frequently bound by the fifth histone protein, linker histone 1 (H1), which is an essential component of stabilizing nucleosomes and higher order chromatin structures [2]. These early studies endeavored to understand bulk chromatin organization in the nucleus. The nucleosome is now well understood to be an indispensable component of genome organization and regulation.

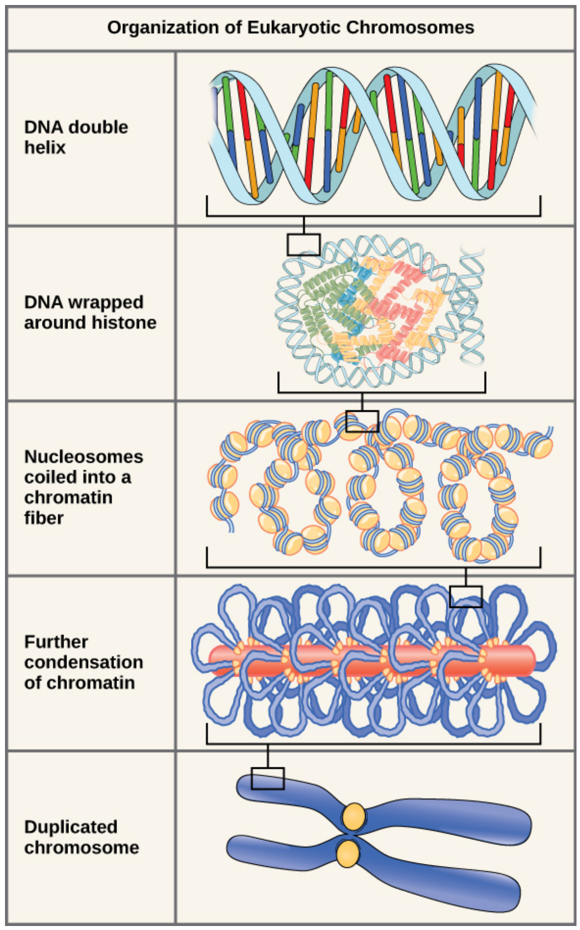

Figure 1.1: Levels of eukaryotic chromatin organization. The most basic structure is the double stranded DNA helix (Top). DNA is further packaged by wrapping around an octamer of histone proteins (Middle Top). Repeating nucleosomes structures are further coiled into chromatin fibers (Middle), which are condensed into higher order chromatin fibers (Middle bottom). Chromatin condensation and formation of higher order structures contributes to chromatid formation (Bottom). Figure borrowed from `https://commons.wikimedia.org/wiki/File:Figure_10_01_03.jpg.` under creative commons license CC BY-SA 4.0.

Through a combination of sedimentation and electric dichroism experiments on nuclease-resistant calf thymus nuclear extracts, Rill and Van Holde were among the first to observe coiled structures of native chromatin [3]. They also observed singlets and doublets of nuclease-resistant “PS-particles” that resembled in conformation and diameter the “v-bodies” described in work from Olins and Olins on micrographs of hypoosmotically ruptured nuclei [4, 5]. Shortly after these early observations, Van Holde and colleagues proposed a working model whereby chromatin was made up repeating, fundamental units of DNA coiled around a core of histone proteins, which were linked together to resemble “beads on a string”[4]. As these formative studies in basic chromatin biology were underway, other groups were attempting to illuminate genome structure complexity well before the sequence of the human genome was established.

1.1.1 Nucleosome Positioning and Chromatin States

In addition to playing essential roles in genome packaging, the presence of nucleosomes can influence the state of chromatin. Nucleosomes act as physical barriers to the transcription machinery necessary to initiate gene regulation. In this particularly dense form, the chromatin state is referred to as heterochromatin and is associated with very large chromatin structures that render the DNA inaccessible to most regulatory factors. This results in the genomic region remaining transcriptionally silent. In addition, histone proteins of nucleosomes in this heterochromatin commonly contain post translational modifications that are associated with this “off” or repressed chromatin state and transcriptional silencing.

However, other regions of the genome are not so densely packed. These regions are often actively transcribed and are referred to as Euchromatin. Euchromatic regions of the genome are depleted for nucleosomes because active transcription displaces nucleosomes. Current evidence suggests that this active removal of nucleosomes is the result of a few different processes. Some transcription factors, known as pioneer factors, may interact with occluded binding sites to destabilize nucleosomes and make genes accessible to transcription machinery [6]. In addition, histone post translational modifications such as acetylation and methylation marks, and chromatin remodeling proteins play a role in displacing nucleosomes from transcriptionally active regions of the genome [7](Figure 1.2).

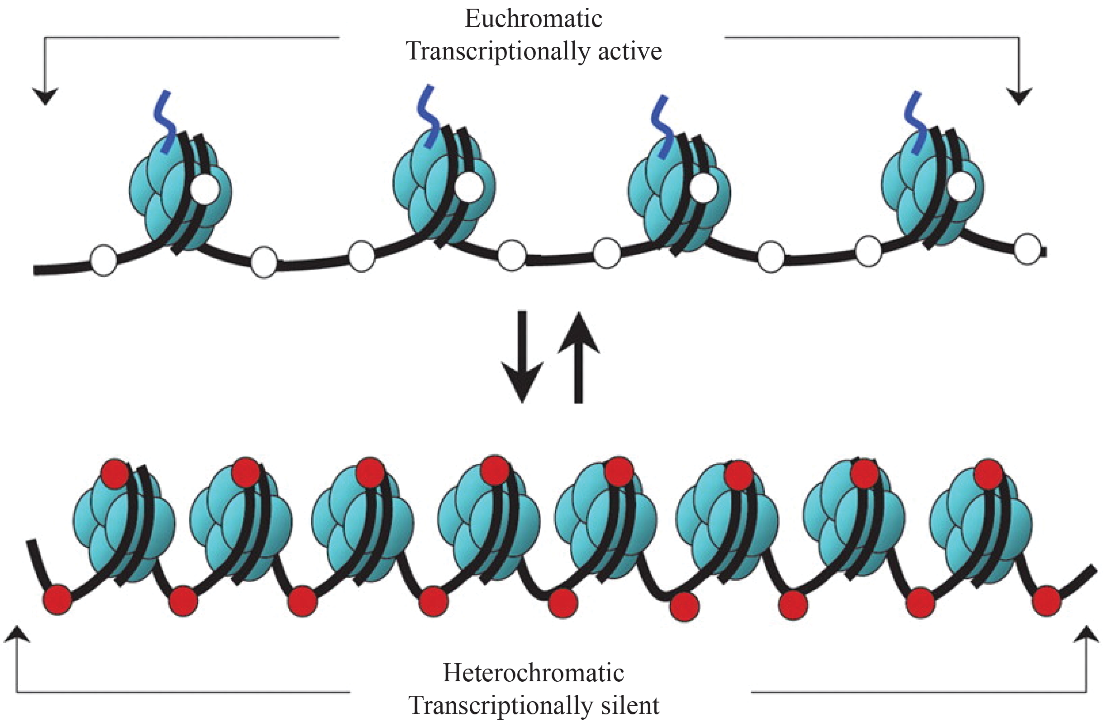

Figure 1.2: Euchromatin and heterochromatin. Heterochromatin (Bottom) is compacted, nucleosome-dense, transcriptionally repressed chromatin with methylated cytosines (red circles). Euchromatin (Top) contains fewer nucleosomes with unmethylated cytosines (white circles) and is transcriptionally active. Figure adapted from Luong, LD. Basic Principles of Genetics 2009.

Active promoters are examples of regions which are depleted for nucleosomes. This has been observed in eukaryotic organisms such as yeast, where decreased nucleosome occupancy was measured at actively transcribing promoters following heat shock or a switch in nutrient source [8]. Furthermore, promoters that were repressed were found to have increased nucleosome occupancy. As a result of the depletion of nucleosomes at promoters, transcription factors and transcription machinery such as RNA polymerase II, have access to the DNA sequences necessary to facilitate transcription. This is consistent with the prevailing model that nucleosomes actively prevent many, but not all, transcription factors from interacting with their respective binding sites [9]. These genomic regions that are depleted for nucleosomes are “open” to the critical regulatory factors and are generally referred to as “open chromatin regions”. Understanding precise chromatin states genome-wide can enable rapid identification of active regulatory regions.

1.1.2 Transcription Factor Binding and Gene Regulation

Transcription factors are DNA binding proteins that are essential to control the expression of genes in all of life. Currently, it is unclear how many transcription factors are encoded by the human genome. TFCat, a manually curated catalog of human and murine transcription factors, indicates there are approximately 500 factors that have both DNA binding and gene regulatory characteristics [10]. Vaquerizas et al. have compiled a list of 1,391 transcription factors based on known DNA binding domains and their probable activities in transcription [11]. While the complete repertoire of transcription factors in the human genome remains a mystery, it is clear they play irreplaceable roles in gene regulation.

The fundamental process of transcriptional regulation begins with the binding of a transcription factor to its requisite binding sequence at a regulatory element. In a cooperation between transcription factors, ATP-dependent chromatin remodelers, and histone acetyl- and methyltransferases, nucleosomes occluding regulatory sequence are displaced to expose the bare DNA [12]. Once the regulatory sequence is exposed, recruitment of RNA polymerase II is mediated directly by the DNA binding transcription factor, or through the association of accessory non-DNA binding transcription factors.

To achieve complex regulatory control across different cell types, combinations of transcription factors coordinate to precisely modulate gene expression. This is accomplished by the interplay of transcription factors that associate proximally and distally to the gene. Some transcription factors only regulate in close proximity to genes, such as sp1 transcription factor known to bind GC rich regions of multiple promoters [13, 14]. Conversely, some transcription factors interact at enhancers, which are regulatory regions distal to genes. Together, transcription factors can collaborate across distant genetic loci to exert specific gene expression patterns [15].

1.1.3 Transcription Factors and Disease Associations

The importance of transcription factors is emphasized by the close association of transcriptional dysregulation to disease. Mutations both in and around transcription factor genes, as well as within transcription factor binding sites have been linked with disease [16, 17]. In nearly 40-60 percent of recorded T-cell acute lymphoblastic leukemia (T-ALL), the transcription factor TAL1 is overexpressed. While several different genetic perturbations have been discovered to contribute to TAL1 misregulation, the upregulation of TAL1 establishes a positive feedback loop for TAL1 expression and ultimately maintains expression of a downstream target essential for the survival of T-ALL cells. Beyond expression abnormalities, direct transcription factor fusion gene events from translocations have been observed frequently in acute myelogenous leukemia (AML) [18]. These events overwhelmingly disrupt the wild-type function of the genes involved in the translocation.

Another means by which transcription factors are associated with disease are through their binding sites. One of the earliest observations was in an individual with beta thalassemia, a blood disorder, in which the authors discovered a C to G substitution -87bp relative to the start of the human beta globin gene [19]. It was later discovered that mutation was in the promoter element of the beta globin gene, and was in the binding site of the Erythroid Kruppel like factor (EKLF) [20]. The -87bp substitution that altered this binding site from ACACCC to ACAGCC was associated with a reduction of mRNA from the beta globin gene [21].

Transcription factor binding site mutations within active regulatory regions are often more subtle in their effects on gene regulation than direct mutations to the gene itself. By virtue of their subtlety, capturing the extent to which non-coding, regulatory variation contributes to human traits and disease has been a challenging endeavor. Genome wide association studies (GWAS), a statistical process by which potentially causal common variation for a trait or disease can be disambiguated and prioritized by comparing groups of people, has been useful in discovering loci associated with many complex human disorders such as anorexia, cancer, and diabetes [22, 23, 24]. However, the approach has been criticized in its ability to pinpoint a single causal variant and avoid spurious associations [25]. Recent work has leveraged functional genomics data to predict the effects of regulatory variants on transcription factor binding [26, 27]. Future strategies for elucidating the effects of non-coding variation will likely require a two-pronged approach: implementing computational variant prioritization by integrating functional genomics data, as well as high-throughput molecular validation.

1.2 Approaches to Characterizing Genomic Features

1.2.1 Methods for Characterizing Open Chromatin

The completion of the first draft of the human genome unlocked avenues for research characterizing features in a genome-wide perspective. With a public reference genome available, the past two decades have given rise to several experimental approaches developed to characterize gene regulatory regions across the genome. These techniques, collectively referred to as open chromatin assays, have been transformative in our understanding of genome regulation and organization.

Due to the involvement of chromatin state in gene regulation, developing approaches to measure it has been a focus of modern genomics. One approach to determining the openness of chromatin is to expose native nuclei to low concentrations of nuclease and sequence the hydrolyzed fragments. This technique, known as DNase-seq, leverages the DNase I endonuclease at dilute concentrations to preferentially cut at regions of open chromatin [28, 29, 30]. Because transcriptionally active regions of the genome are depleted for nucleosomes, the DNA of active regulatory regions is exposed and susceptible to nucleolytic attack. As a result, DNase I digests more frequently at regulatory regions and the sequencing reads accumulate in peaks that correspond to actively regulating loci. Another commonly used enzymatic-based approach to probing native chromatin states is the Assay for Transposase Accessible Chromatin (ATAC-seq) [31]. This method incubates native nuclei with hyperactive Tn5 transposase loaded with high-throughput sequencing adapters. Similar to DNase-seq, ATAC-seq preferentially transposes in open chromatin regions and correlates strongly with active regulatory regions. Both DNase and ATAC-seq datasets substantially overlap in their enrichment for regulatory regions [32].

Another relevant, but less used method is formaldehyde assisted isolation of regulatory elements followed by sequencing (FAIRE-seq). This is a chemical method of isolating regions of active chromatin [33]. Cells are formaldehyde crosslinked in vivo and sheared by sonication. The formaldehyde-fixed cell lysate is emulsified in a phenol-chloroform solution and centrifuged. The resultant, biphasic sample contains nucleic acids in the aqueous layer and proteins in the organic layer. Recovering the aqueous layer enriches for DNA fragments that contain little to no protein.

Since the sample is formaldehyde crosslinked, nucleosome depleted regulatory regions phase into the aqueous layer and can be purified and prepared for sequencing. FAIRE-seq differs from DNase-seq and ATAC-seq as it lacks an enzymatic component and instead depends on the physical properties of proteins and nucleic acids in a phenol/chloroform emulsion.

Despite their prevalent usage in functional genomics, modern open chromatin assays are not without limitations. FAIRE-seq assumes that active regulatory loci will be depleted for protein and solubilize in the aqueous layer. However, highly active promoters and other regulatory elements may have protein occupancy profiles favor association in the organic phase. Likewise, DNase-seq and ATAC-seq utilize enzymes at specific concentrations that have been observed to have sequence biases [34]. In addition, the aforementioned approaches are limited to only annotating active regulatory loci, and are unable to assign specific regulatory function [35]. Therefore, it is critical to explore additional molecular approaches to understand and assign function to regulatory regions.

1.2.2 Methods to Measure Transcription Factor Binding

To better assign function to open chromatin regions, several methods have been established that directly assess the identity of transcription factor binding. Chromatin immunoprecipitation and sequencing (ChIP-seq) was originally leveraged to measure histone post translational modifications at the genome level, but has been instrumental in characterizing transcription factor binding sites [36]. In short, cells are fixed with formaldehyde, nuclear lysate is extracted and heavily fragmented, and antibodies targeted to transcription factor protein-DNA complexes are precipitated out of the lysate. Following precipitation, the DNA is purified, sequenced, and aligned to the genome. Alternative to identifying open chromatin, ChIP-seq can indicate where a specific transcription factor binds throughout the genome, and based on the identity of the factor, ChIP-seq can also provide an indication of the activity of the bound loci. However, fragment sizes from ChIP-seq experiments are longer than the transcription factor bound DNA, which limits the resolution of the precise binding site. ChIP-exo, which integrates a Lambda exonuclease digestion to enzymatically remove DNA flanking a bound transcription factor, has been used to increase the resolution of sequencing data from ChIP experiments [37].

1.2.3 Transcription Factor Footprinting

While measuring open chromatin and transcription factor binding sites via ChIP approaches allow for a general idea of the location of a regulatory element, it is often beneficial to know the precise binding site of a factor and this can be accomplished through DNase footprinting assays. As transcription factor proteins associate with their respective binding sites to promote gene regulation the specific DNA motif is physically protected from enzyme activity, as observed in early DNase experiments [38]. They showed that the binding site of the E. coli lac repressor can be determined by autoradiograph display of limited, in vitro digestions of lac repressor-bound oligonucleotides by DNase I. The nucleotides that were protected from digestion were referred to as “footprints” of the factors that were once bound.

With the advent of high-throughput sequencing, experimental approaches that rapidly profiled regulatory regions (DNase-seq and ATAC-seq) were found to also function as genome-wide footprint experiments, due to transcription factor occupancy blocking enzymatic activity. In high throughput sequencing data, the footprints of transcription factors are represented as a depletion of coverage (trough) between two peaks of hypersensitivity to DNase. Computational models trained on these sequence features have been designed to discern specific binding motifs of transcription factors. These models approach transcription factor footprinting from two perspectives: de novo and motif-centric. Each are limited in their ability to discern transcription factor identity from the sequence context. Motif-centric models generally require predetermined transcription factor binding sequences on which to generate potential binding sites across the genome and evaluate if they overlap footprint signal [39]. However, these are unable to discover candidate binding sites, and often require ChIP-seq data input. Alternatively, the de novo models assume that the data harbor patterns at footprint sites, and footprints are queried for known transcription factor binding sequences [40]. These approaches are limited in their potential to ascertain binding sites of interest. A recent method, known as TRACE, have overcome this by implementing an unsupervised hidden Markov model framework which outperforms other footprinting methods and does not require ChIP-seq or sequence motif inputs [41].

1.2.4 Limitations to Characterizing Gene Regulatory Elements

Despite the variety of methods to characterize and interpret gene regulatory elements, the current strategies are still limited. Conventional ChIP-seq produces enrichment peaks that are large (200-500bp), which renders precise binding difficult to resolve. Recent modifications of ChIP-seq have increased resolution, but still require antibodies [42, 37]. This represents one of the largest disadvantages of ChIP-related technologies. As indicated previously, there are likely upwards of 1000 transcription factors encoded in the human genome [11]. While a comprehensive analysis of the number of antibodies available has not been completed, the most expansive ChIP approaches to characterize human transcription factor binding profiles have assayed 681 unique proteins [43]. However, the antibodies used were generated from the Protein Capture Reagents Program (PCRP) and lacked robust validation.

Even well-tested antibodies have been shown to poorly reproduce, with a 20 percent failure rate in some instances [44, 45]. Therefore, illuminating every transcription factor-bound regulatory region with ChIP-seq is marred by the reproducibility of antibodies and scalability to all transcription factors.

Although open chromatin assays and footprinting algorithms offer an alternative, antibody-free method for gaining insights into gene regulatory regions, they too have disadvantages. For example, in ATAC-seq a Tn5 transposase is loaded with adapters randomly, which can result in large fractions of the sample unusable [46]. Also, depending on the cell type, ATAC-seq preparations may be heavily contaminated with mitochondrial reads [46, 47]. Furthermore, both ATAC-seq and DNase-seq leverage enzymes with cleavage biases. No consensus or best practices exist on a bias correction approach, and the unique activities of theses enzymes can produce different footprints for the same transcription factor [32]. Despite improvements in footprinting algorithms, footprinting fundamentally relies on a negative enrichment strategy, whereby transcription factors occlude the acquisition of sequencing coverage. Collectively, these limitations encourage molecular innovation in assays to understand gene regulatory regions. New technologies that discard mainstream antibody and footprinting workflows may fill a much needed gap in the molecular disambiguation of regulatory regions.

1.3 Repetitive Sequence Composition of the Human Genome

1.3.1 Transposable Elements

Almost 30 years prior to the first draft of the human genome, researchers were attempting to understand genome complexity and composition from different organisms using C0t analysis on single stranded DNA reassociation rates. A striking finding was that substantial fractions of higher order organisms' genomes, humans included, reassociated rapidly [48, 49]. This indicated that repetitive sequences constituted a large portion of these genomes. The first sequencing draft of the human genome deepened our understanding of these repeated DNA sequences, showing that transposable elements represented the largest fraction of repetitive DNA at nearly 50 percent of the human genome [50]. Transposable elements are mobile DNA sequences that mobilize and disperse themselves throughout their host genome [51]. There are two main types of mobile elements: DNA transposons and retrotransposons. DNA transposons mobilize in their host genomes by a “cut and paste” mechanism. Only about 3 percent of the human genome is made of DNA transposons, and they have been immobile in the primates for millions of years, likely due to an accumulation of deleterious mutations [50, 52, 53]. On the other hand, retrotransposons are the largest class of structural variation, collectively accounting for about 43 percent of the human genome, with subsets of them actively dispersing into new loci [50]. Instead of a “cut and paste” mechanism, retrotransposons mobilize through a “copy and paste” strategy that depends on an RNA stage preceding a new insertion event (Figure 1.3).

Of the retrotransposable elements that remain active in the human genome, non-long terminal repeat (LTR) containing retrotransposons such as Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs) make up the largest fraction at approximately 34 percent of the human genome. LINEs contain 2 open reading frames (ORFs). ORF1 encodes a protein with apparent nucleic acid binding and chaperoning properties and is critical for retrotransposition. Likewise, ORF2 is essential to retrotransposition and contains reverse transcription and endonuclease functions [53]. Together, with an internal promoter, LINEs are capable of autonomous retrotransposition. Since they lack ORFs, SINEs are non-

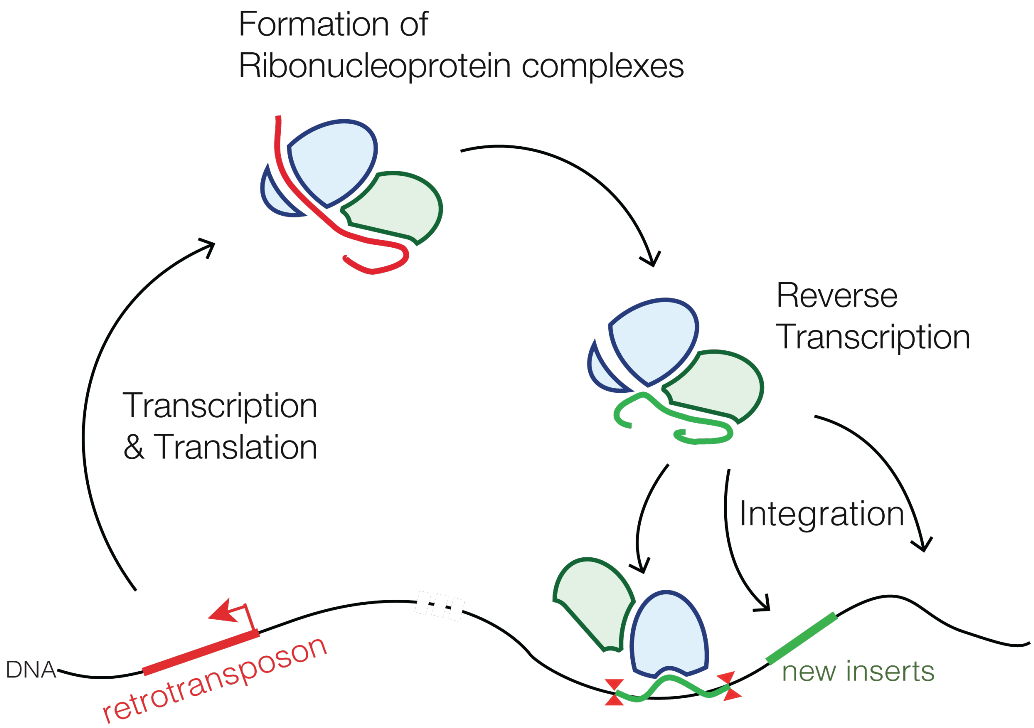

Figure 1.3: Basic mechanism of retrotransposition. Retrotransposition begins with transcription of retrotransposon in the genome. Open reading frames within the transcript are translated into proteins. Proteins form a ribonucleoprotein (RNP) complex with untranslated retrotransposon RNA. RNP facilitates reverse transcription of the RNP into cDNA, and subsequent insertion into a new genomic locus. Figure borrowed from `https://commons.wikimedia.org/wiki/File:Retrotransposons.png` under creative commons license CC BY-SA 4.0

autonomous retrotransposons and depend on expression products from active LINEs to facilitate their mobility. As a result of their activity, LINEs and SINEs continue to shape human genome structure and function.

1.3.2 Transposable Elements and Disease

The mobilization of transposable elements can have widespread, often deleterious, regulatory consequences across the genome. For instance, the internal L1 antisense promoter has been found to initiate antisense transcription to create chimeric mRNAs from several human genes [54]. Moreover, the L1 antisense promoter can act as an alternative transcriptional start site for these genes to increase the flexibility and diversity of expression. Insertions of retrotransposons have also been linked to sporadic disease. As many as 124 diseases have been reported from mutagenic retrotransposon insertions [55]. For instance, fukuyama muscular dystrophy is caused by an SVA (a type of SINE) insertion into the fukutin (FKTN) gene, which results in an alternative splicing of the mRNA and ultimately a mislocalization of the protein product [56, 57]. Further, emerging evidence suggests that reactivation of retrotransposition in early embryogenesis results in mosaic insertion patterns in the brain, which may influence an array of psychiatric disorders [58].

1.3.3 Limitations to Characterizing Mobile Element Insertions

Due to the dramatic effects mobile element insertions on structure and function of the human genome, characterizing insertions from ongoing retrotransposition has been a focus of modern genetics. Historically, molecular techniques implementing a high throughput short read sequencing strategy have been leveraged to discover insertion events that are polymorphic from the human genome reference assembly. These include hybridization arrays, paired-end fosmid sequencing, and sequencing of L1/genomic flank amplicons [53]. However, these methods suffer from limitations. For instance, paired-end sequencing of fosmids selects for the addition of 6kb of LINE sequence in a large 40kb fosmid. Due to the size selection of fosmids with 6kb gnomic inserts, substantially smaller, yet human disease-relevant mobile element insertions (i.e SINEs) may not be resolvable through this approach. In addition, this approach is labor-intensive and may miss highly repetitive or partially truncated insertions. However, a key limitation shared among these techniques is their collective ability to capture sequence information of the element. Short read sequencing limits mapping to reads that span the insertion flanks of mobile elements. As a result, the larger sequence context outside of the insertion is missed. In addition, many of these approaches are low throughput and laborious [59]. Therefore, an unmet need in retrotransposon discovery is rapid, genome-wide, sequence level characterization approach for polymorphic insertions.

1.3.4 Polynucleotide Repeats and Associated Repeat Expansion Disorders

As mentioned above, transposable elements are the largest class of structural variation in the human genome, but not the only class. The second largest class of structural variation found in the human genome is polynucleotide repeats. These elements are tandemly arranged sets of repeating nucleotides that can occur as few as sets of 3 (trinucleotide repeat) to as many as 12 (dodecanucleotide repeat) [60]. Approximately 3 percent of the human genome is made up of repeating nucleotides [50]. The origin of polynucleotide repeats in the human genome is unclear, however it has been hypothesized that polynucleotide repeats may have played a role in early eukaryotic evolution. Specifically, nucleotide repeats within codons may have undergone substitutional mutations which acted as a diversifying mechanism for ancestral eukaryotic genomes and proteomes [61, 62].

In humans, some polynucleotide repeats are known to undergo expansion and cause disease. A classic example of a repeat expansion disorder is fragile x syndrome (FXS), whereby a CGG trinucleotide repeat expands in the 5' untranslated region of the fragile x mental retardation gene 1 (FMR1). The average unaffected individual usually contains an FMR1 locus with 6 to 54 CGG repeats [63, 64]. With expansion to 50-200, this is referred to as the premutation and may clinically present as fragile X-related primary ovarian insufficiency (FXPOI) in women, and/or fragile X-associated tremor/ataxia syndrome (FXTAS) in both sexes, with onset and penetrance generally correlating with the extent of CGG expansion as well as age [65, 63]. The premutation allele results in an upregulation of transcription from the FMR1 gene, yet a reduction of the fragile x mental retardation protein (FMRP). The full mutation causes FXS and occurs when the 5' UTR accumulates greater than 200 CGG repeats. Unlike the premutation, an FMR1 allele with the full, the promoter and 5'UTR accumulate DNA methylation which results in a transcriptional silencing of the gene. FXS is but one example of repeat expansion associated disorders. Expansions of polynucleotide repeats are the culprits of upwards of 40 different human genetic disorders, and likely many more yet undiscovered [60].

1.3.5 Limitations to Characterizing Structural Variation

Despite their discovery 30 years ago, repeat expansion disorders have been challenging to characterize due to their repetitive structure, diverse clinical manifestations, and molecular mechanisms [60, 66]. Direct clinical testing of candidate genes using polymerase chain reaction (PCR) can some elucidate some repeat expansions. However, this is low throughput and is increasingly less scalable as novel repeat expansion diseases are discovered. In addition, some polymerases have been characterized to make errors when amplifying repeat DNA, which may complicate precise repeat characterization through PCR[67]. High throughput sequencing has aided in identification of repeat expansion disorders in patient by screening for many different candidate affected genes at once, but the short read length poses computational challenges that complicate precise mapping and characterization [68]. These obstacles emphasize the need for methods with improved detection and characterization of disease-associated polynucleotide repeats.

1.4 Conclusion

In the span of nearly half of a century, the field of genetics has advanced from understanding the basic structures of chromatin and genome compositions in the 1970s, to genome-wide elucidation of regulatory and structural elements across many cell types. This breakneck pace is, in part, afforded by advancements in molecular methodologies. However, even the current technologies have unavoidable limitations that require further innovation. To better understand gene regulatory and structural elements of the genome, and how these elements are involved in human disease and suffering, new methods require development to overcome the limitations outlined earlier in this introduction. In this thesis, I will focus on developing and leveraging new technologies that can surmount the intractable shortcomings of earlier approaches, and demonstrate that these novel implementations present practical solutions to modern genomics.

CHAPTER II

Isolating and Sequencing Protein-Occupied Open Chromatin Regions

2.1 Abstract

Open chromatin plays a critical role in transcriptional regulation of genes. A variety of methods have been developed to probe regions of open chromatin in the genome. Collectively, these assays have been crucial in defining the regulatory landscape of the human genome. However, certain features of open chromatin remain difficult to characterize. For instance, the precise binding events of transcription factors within open chromatin to impose a regulatory potential have been challenging to precisely map. Footprinting analysis, which attempts to infer transcription factor binding from troughs in high throughput sequencing data from open chromatin assays, is limited in its ability to capture many transcription factor binding events. Here, we explore a novel open chromatin assay to capture specifically transcription factor bound sequences.

2.2 Introduction

The Encyclopedia of DNA Elements (ENCODE) Project indicates that, in any given cell type, only about 1 percent of the genome is open chromatin as defined by hypersensitivity to DNase I. Despite this small fraction of genome that is involved in active regulation, as much as 95 percent of transcription factor binding data (ChIP-seq peaks in K562) overlap these hypersensitive regions [69]. Moreover, 12 percent of disease-associated variation determined from GWAS lies within these regulatory regions and transcription factor binding sites and 36 percent occur in hypsersensitive sites [69]. Therefore, further understanding the relationship between transcription factors and the open chromatin regions they occupy is instrumental to further understanding how disruptions in gene regulation contribute to disease.

However, while a handful of experimental techniques have been instrumental in annotating regions of open chromatin, they are limited in their ability to understand the detailed interplay of transcription factors at these regions. DNase-seq and ATAC-seq are two examples of open chromatin assays that leverage DNase I and Tn5, respectively, to fragment regions of exposed DNA in a native chromatin context [29, 31]. Since transcription factors occlude their bound sequence from the activities of these enzymes, they leave behind low coverage "troughs" in high throughput sequencing data, which are referred to as transcription factor "footprints". De novo and motif-centric computational approaches (reviewed in chapter 1) have been developed to extract information about the identity of the transcription factor based on its footprint, however intrinsic experimental bias may confound these algorithms' ability to interpret true transcription factor binding events. For instance, compressed minor grooves proximal to CpG methylation appear to sustain higher rates of DNase I activity compared to cytosine methylation [70]. Recent work has suggested that transcription factor footprints identified in embryonic stem (ES) cells may represent artifacts of overlooked DNase I cleavage bias [34, 71].

An alternative approach to open chromatin assays and footprinting for understanding transcription factor binding is ChIP-seq. ChIP-seq was first applied to

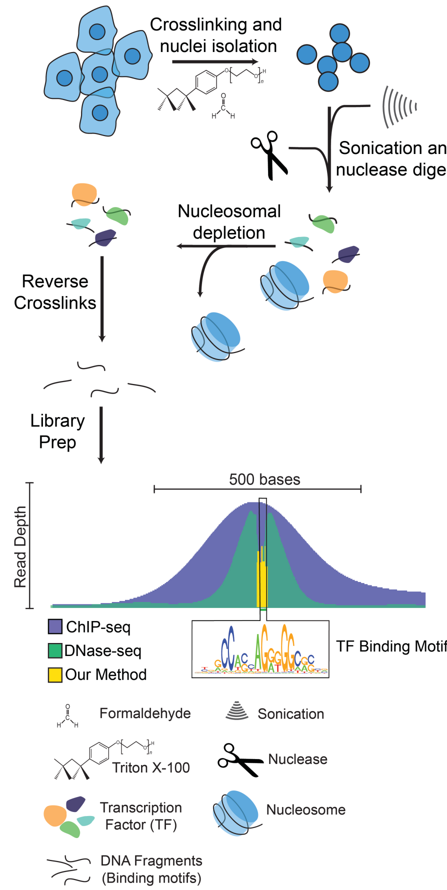

Figure 2.1: Protocol for isolating transcription factor bound sequences. Fresh K562 cells are incubated in a crosslinking nuclei isolation buffer containing formaldehyde and Triton X-100. Crosslinked nuclei are lysed using sonication and heavily nuclease digested. Digested lysates are depleted for nucleosomal structures and nucleosomal DNA. After nucleosomal depletion, the residual lysate contains transcription factor-DNA complexes. These complexes are reverse crosslinked and the DNA is purified and prepared for high-throughput sequencing. Our method (yellow) is overlaid with example functional genomics data, including DNase-seq (Green) and ChIP-seq (Blue).

profiling histone post translational modifications genome-wide, but widely used for transcription factors as well [36]. Since ChIP-seq uses antibodies to recognize specific protein epitopes, it is able to specifically isolate transcription factors and their bound DNA fragments. While ChIP-seq has been pivotal in mapping over 100 transcription factors, it is not without limitations [69]. For example, the availability of antibodies for known transcription factors bottlenecks discoverable binding profiles. With estimates of anywhere between 500 and 1400 DNA binding transcription factors in human, developing antibodies for each target is difficult to scale [11, 10]. Moreover, antibodies have been observed to vary in reproducibility between production lots [72].

Altogether, these limitations inspired us to develop an orthogonal molecular approach to characterize protein occupancy within open chromatin. This approach aims to unbiasedly discover protein-occupied open chromatin while avoiding the footprinting and cleavage bias in DNase-seq, and antibody limitations of ChIP-seq. Similar to ChIP-seq, the first step involves isolating formaldehyde crosslinked nuclei to preserve protein DNA interactions and eliminate cytosolic protein and mitochondrial DNA (Figure 2.1). Next, isolated nuclei are sonicated and the lysate digested with a nuclease to reduce DNA fragment size. After nucleosomes are depleted, the residual lysate containing non nucleosomal DNA binding proteins is reverse crosslinked to recover DNA. Purified DNA is prepared for high throughput short read sequencing.

2.3 Methods

2.3.1 Isolation of Crosslinked K562 Nuclei

Crosslinking buffer was formulated to contain 1 percent formaldehyde, 20 mM Sodium Phosphate Dibasic, and protease inhibitors. To solubilize the cell membrane, the crosslinking buffer was supplemented with Triton X100 at the minimum effective concentration (0.3322uL/mL) [73]. Phosphate buffered saline (PBS) washed K562 pellets of 4x107 cells were resuspended in 40 milliliters of the crosslinking buffer and incubated at room temperature with gentle rocking for 10 minutes. Crosslinking was quenched by the addition of 1M Tris-HCl pH 8.0 to 250mM. After quenching, the suspension was evenly distributed into 15mL conical tubes and centrifuged at 500xg for 10 minutes at 4 degrees Celsius. The nuclei pellet from each tube was consolidated into a single 15mL falcon tube and washed in 20mM sodium phosphate dibasic with protease inhibitor cocktail to 1x.

2.3.2 Lysis and Nuclease Digestion of Crosslinked K562 Nuclei

Freshly isolated and formaldehyde crosslinked K562 nuclei were resuspended in 300uL 20mM Sodium Phosphate buffer. For immunodepletion (see **Pan-histone immunodepletion**), SDS was added to 1 percent, otherwise detergents were omitted from lysis. Resuspended K562 nuclei were sonicated on ice at 15 percent amplitude for 5 second on/off cycles for a total of 30 seconds using a probe sonicator. Sonicated nuclear lysates were supplemented with CaCl2 or MnCl2 for enzyme activity and digested with 10 units of MNase or DNase I for varying time increments. Nuclease activity was mitigated by chelating Ca2+ with EGTA to 10mM and incubating on ice.

2.3.3 Nitrocellulose Filter Binding of K562 Nuclear Lysate

Digested K562 nuclear lysate prepared from nuclei resuspensions in 300uL of 20mM Sodium Phosphate without SDS. Total sample protein was estimated and the necessary size of nitrocellulose membrane was determined based on protein binding capacity (80ug/cm2). Membranes were prepared by a 10 minute incubation in 400mM KOH in a petri/culture dish. Membranes were washed in pure water for 5 minutes. Water washes were repeated 2 more times to remove KOH. Membranes were left to equilibrate in 1x Nitrocellulose binding buffer (500M KCl, 5 mM EDTA, and 20mM Tris HCl pH 7.5) for 1 hour. Sample was mixed with 1 volume of 2x nitrocellulose binding buffer and applied dropwise onto the prepared nitrocellulose membrane and dried at 37 degrees with gentle agitation. 1x nitrocellulose buffer was added dropwise until the filter was covered and the membrane was incubated at room temperature for 5 minutes. Membrane was transferred to a clean culture/petri dish and 100mM Tris HCl 7.5 pH was added dropwise to cover the membrane. After

5 minutes, membrane was washed with 1x TE 2 times, the first wash for 5 minutes, followed by a second wash for 30 minutes. Residual TE was removed and the membrane was transferred to a 50mL conical. 1.8mL of TE was added and the membrane was rotated and reverse crosslinked in a hybridization oven at 65 degrees overnight. Eluted samples from the overnight membrane reverse crosslinking RNase (fermentas) treated at 37 for 2 hours. RNase treated samples were volume reduced using sec butanol and ethanol precipitated (See **Reverse Crosslinking and Purification of DNA**).

2.3.4 Pan-histone Immunodepletion of Nucleosomal Complexes

Following the lysis, the first step was to neutralize the SDS. Previous work has indicated that the denaturing effects of SDS can be overcome by adding 8 fold higher concentrations of non ionic detergents [74, 75]. This was accomplished by diluting the 300uL of nuclei lysate with 730uL of a SDS Neutralization Buffer (1.6 percent Triton X100, 7mM MnCl2, 7mM CaCl2). At this step the sample was DNase digested (see above). To prepare for the immunodepletion after the DNase digestion, the sample was salt balanced by adding 40uL of 5M NaCl and 40uL of 500mM Tris HCl pH 8.0 to a final concentration of 167mM NaCl and 16.7mM Tris HCl. These salts were omitted from the SDS neutralization buffer as DNase I retains the highest activity at low salt concentrations. To the salt balanced lysate, 100uL of protein G magnetic beads (88847, NEB) crosslinked with anti-histone antibodies (MAB3422, Millipore) were added and incubated at 4 degrees overnight with rotation. Since the supernatant from the overnight immunoprecipitation represents the depleted fraction, and theoretically contains the non nucleosomal DNA-protein complexes, the fraction was saved and the DNA purified (See Reverse crosslinking and purification of DNA). The beads were magnet immobilized and washed two times in 200uL of

RIPA buffer (0.1 percent Sodium Deoxycholate, 0.1 percent SDS, 1 percent Triton X-100, 10mM Tris HCl pH 8.0, 140mM NaCl, and 1mM EDTA). Next, the beads were washed 2 times with 200uL of High Salt RIPA Buffer (0.1 percent Sodium Deoxycholate, 0.1 percent SDS, 1 percent Triton X-100, 10mM Tris HCl pH 8.0, 360mM NaCl, and 1mM EDTA). Lastly, the beads were washed 2 times in 200uL of a LiCl wash buffer (0.5 percent NP40, 0.5 percent of Sodium Deoxycholate, 1mM EDTA, 10mM Tris HCl pH 8, and 250mM LiCl). The immunoprecipitated complexes were eluted by resuspending the beads in 1x Elution buffer (1 percent SDS, 50mM Tris HCl pH 8, and 10mM EDTA). The amount of beads and antibody ratio and crosslinking protocol were provided by the bead manufacturer (NEB). Antibody coated beads were added in excess to estimated histone content.

2.3.5 Reverse Crosslinking and Purification of DNA

To reverse crosslinks, 4x Elution buffer (200mM Tris HCl pH 8, 40mM EDTA, 4 percent SDS) was added to lysate fractions to a final concentration of 1x. 5uL of proteinase K was added to each sample, which was mixed thoroughly and distributed equally in PCR tubes and incubated overnight (16 hours) at 65 degrees Celsius. After reverse crosslinking, samples were consolidated into 750uL aliquots and mixed vigorously with 1 volume of phenol: chloroform: isoamyl alcohol. The emulsion was centrifuged at 20000xg for 5 minutes and the aqueous layer recovered and rextracted with 1 volume of chloroform. After the second extraction, the recovered aqueous layer was prepared for ethanol precipitation by adding 1 volume of ethanol, sodium acetate to 300mM, and 1uL of GlycoBlue coprecipitant. The sample was incubated on ice for 30 mintues, followed by a 30 minute to overnight incubation at -20. After incubations, the sample was precipitated by centrifuging at 20000xg for 30 minutes at 4 degrees Celsius and carefully washed with 95 percent ethanol avoid solubilizing small fragments. Pellets were washed 2 times with 95 percent ethanol and centrifuged at 20000xg for 10 minutes with each wash. The final precipitates were air dried and stored in 1x TE buffer (10mM Tris HCl pH 8, 1mM EDTA) at -20 degrees C. Purified DNA was quantified using Qubit dsDNA high sensitivity kit.

2.4 Results

2.4.1 Simultaneous Isolation and Formaldehyde Crosslinking of K562 Nuclei

To ensure that transcription factors remain bound to the DNA, it is common to formaldehyde crosslink. However, formaldehyde promiscuously forms crosslinks with cell components rendering them resistant to lysis and solubilization, even in strong ionic detergents. We observed that when crosslinked at 1 percent formaldehyde, K562 cells resisted conventional lysis techniques used in ChIP-seq workflows. As a result, we developed a simultaneous crosslinking and nuclei isolation method for suspension cells (see methods). For downstream applications, we aimed to minimize the amount of non-nuclear proteins. To do this, we formulated a minimal hypotonic crosslinking buffer containing 20mM Sodium Phosphate, minimal effective concentration of Triton X-100, formaldehyde to 1 percent, and protease inhibitors. The rationale behind the formulation was to 1) induce hypoosmotic swelling in the cells and 2) perforate the cell membrane using the minimum effective concentration of Triton X100, liberating nuclei from cytoplasmic components while 3) crosslinking in formaldehyde. Using the automated cell counter, we measured the diameter of cells prior to and after the nuclei isolation (Fig. 2A and 2B). We observed that the diameter of the average K562 cell from culture to be approximately 17.6 uM (Figure 2.2A). Following the crosslinking and nuclei isolation protocol, the average size of the cell was reduced by nearly 50 percent, with an average size about 9uM (Figure 2.2B). This suggested that our approach was efficiently isolating subcellular-sized compartments.

Figure 2.2

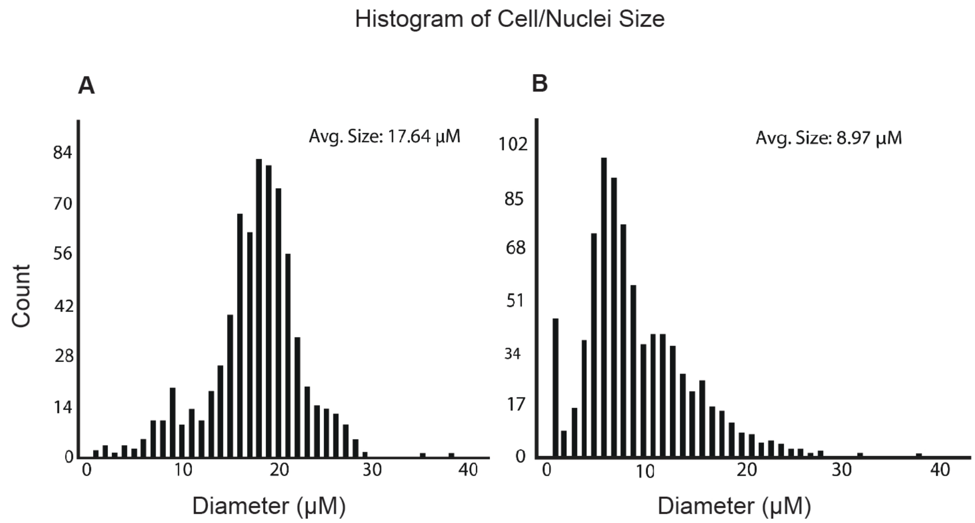

Figure 2.2: Histogram of cell and nuclei size. Distribution of 2A) cell sizes and 2B) nuclei sizes as determined by Countess II. X-axis is diameter measured in micrometers and Y-axis is number of counts.

To confirm that the isolated cellular material were indeed nuclei, we performed comparative DAPI staining between mixes of isolated nuclei and live cells from culture. DAPI stain is considered a vital dye because it is unable to cross the plasma membrane of viable cells. Only dead cells with compromised cell membranes will be stained with DAPI and fluorescent as a result [76]. We imaged DAPI stained mixes of live cells from culture and isolated nuclei on brightfield/DAPI overlays (Figure 2.3) Together, these data suggested that our developed approach efficiently isolates crosslinked nuclei from K562 cells

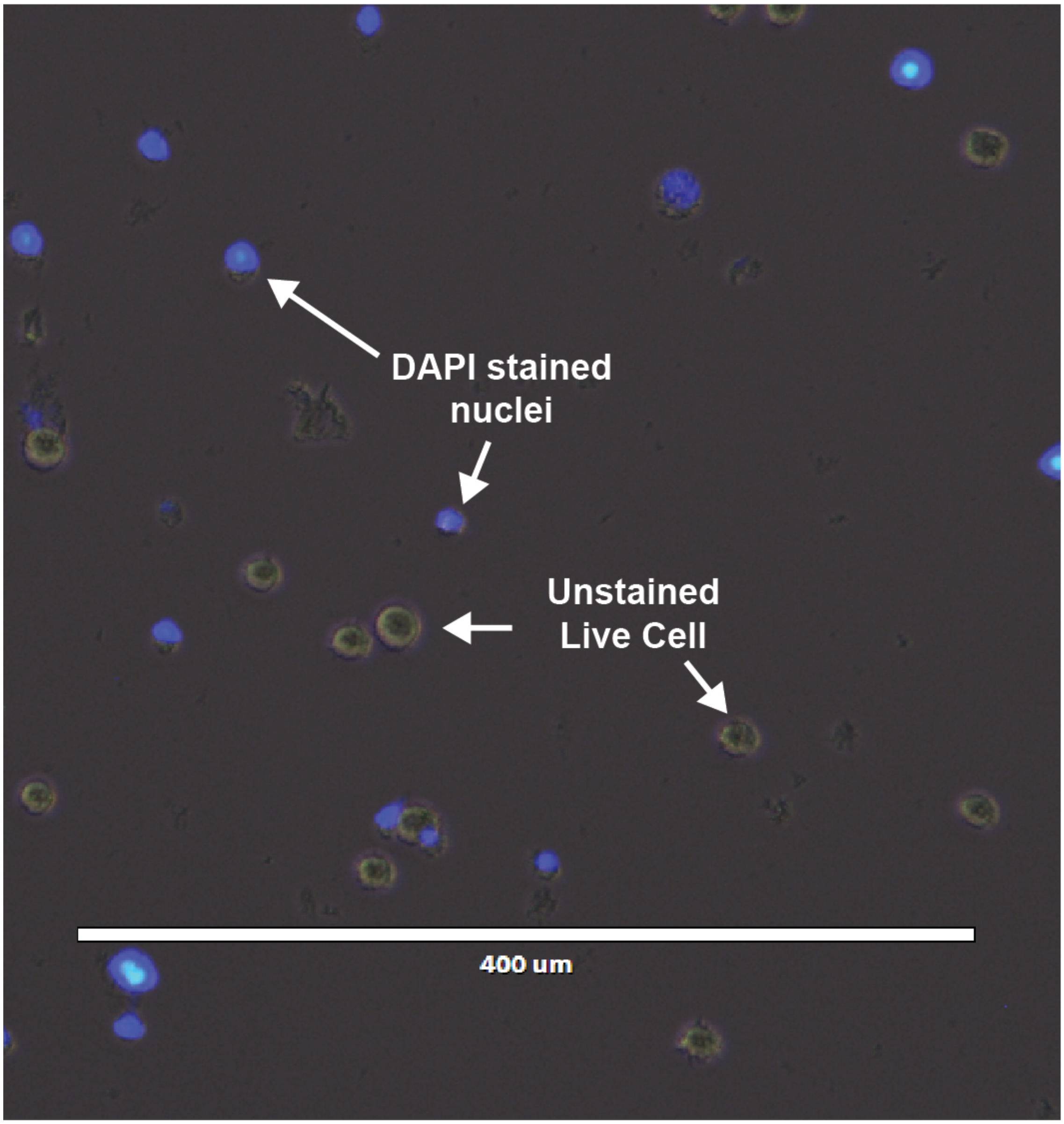

Figure 2.3: A brightfield/DAPI overlay of live cell and isolated nuclei mix. Scale bar is 400 micrometers.

2.4.2 Recovery and Sequencing of DNA from Filter Bound, Crosslinked K562 Nuclear Lysates Shows Subnucleosomal-Sized Fragments

To immobilize non nucleosomal DNA binding proteins, we performed a modified filter binding assay (See methods) of crosslinked, micrococcal nuclease digested K562 nuclear lysate to nitrocellulose. Purified DNA isolated from the filter binding experiment was quantified and a small quantity was run on tapestation to observe fragment size distribution (Figure 2.4) Compared to the input (no nitrocellulose filter binding), there was a noticeable shift in fragment size distribution. While the input fraction contained a large peak at 159bp, the nitrocellulose extracted fraction contained a proportional enrichment of subnucleosomal sized fragments with a peak of 59bp. Considering that the nucleosome is 147bp of DNA wrapped around a histone octamer, the 159 bp peak in both the input and the nitrocellulose extracted fraction was anticipated to be MNase digested mononucleosomes. We next prepared an Illumina sequencing library out of both the input and nitrocellulose experiments. We evaluated the extent of enrichment over all annotated DNase hypersensitive sites in K562 for both the nitrocellulose and the input sequencing (Figure 2.5). We observed that there was a depletion of sequencing reads over the relative center of open chromatin regions, extending approximately 300bp out from the center. This depletion was relative to sequence immediately flanking the center of open chromatin sites. In addition, these regions contained a fragment size distribution with 10bp periodicity. Together, these data suggested that MNase digested mononucleosomes were binding nitrocellulose and histone bound DNA was efficiently being eluted.

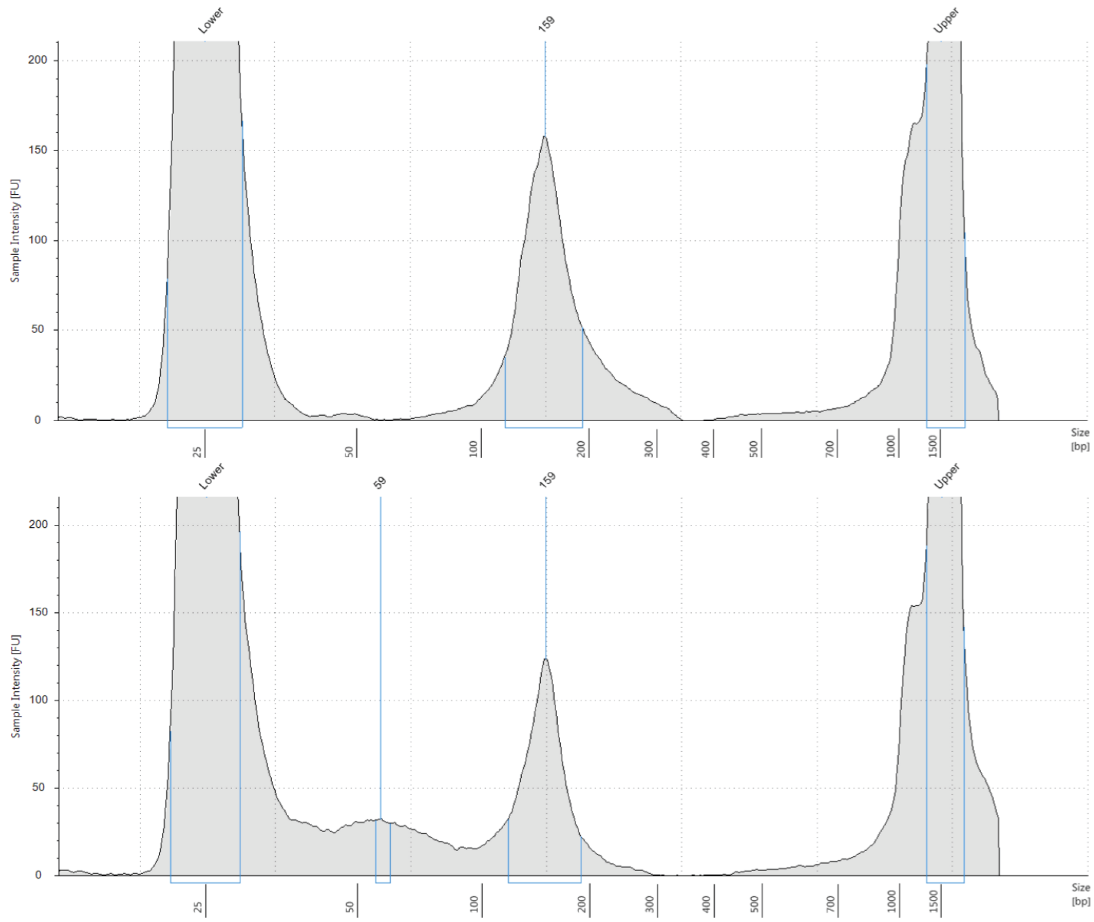

Figure 2.4: Tapestation of DNA fragment size distribution isolated from basic nitrocellulose filter binding. The fragment size distribution of DNA from input (Top) and after basic nitrocellulose filter binding (Bottom). X-axis is size of DNA fragments in bp, with upper and lower molecular weight loading standards at 1500bp and 25bp, respectively.

2.4.3 Isoelectric Point Analysis of Transcription Factors Revealed a Subset of Nitrocellulose Incompatible Transcription Factors

Nitrocellulose has been commonly used in filter binding approaches to immobilize DNA-protein complexes. The specific mechanism between protein and nitrocellulose interactions is unclear, however the isoelectric points of proteins and their charges relative to the nitrocellulose membranes play a role. To evaluate if our nitrocellulose filter binding experiments were potentially missing subsets of transcription factors that would have a neutral or negative charge, we computationally summed charged

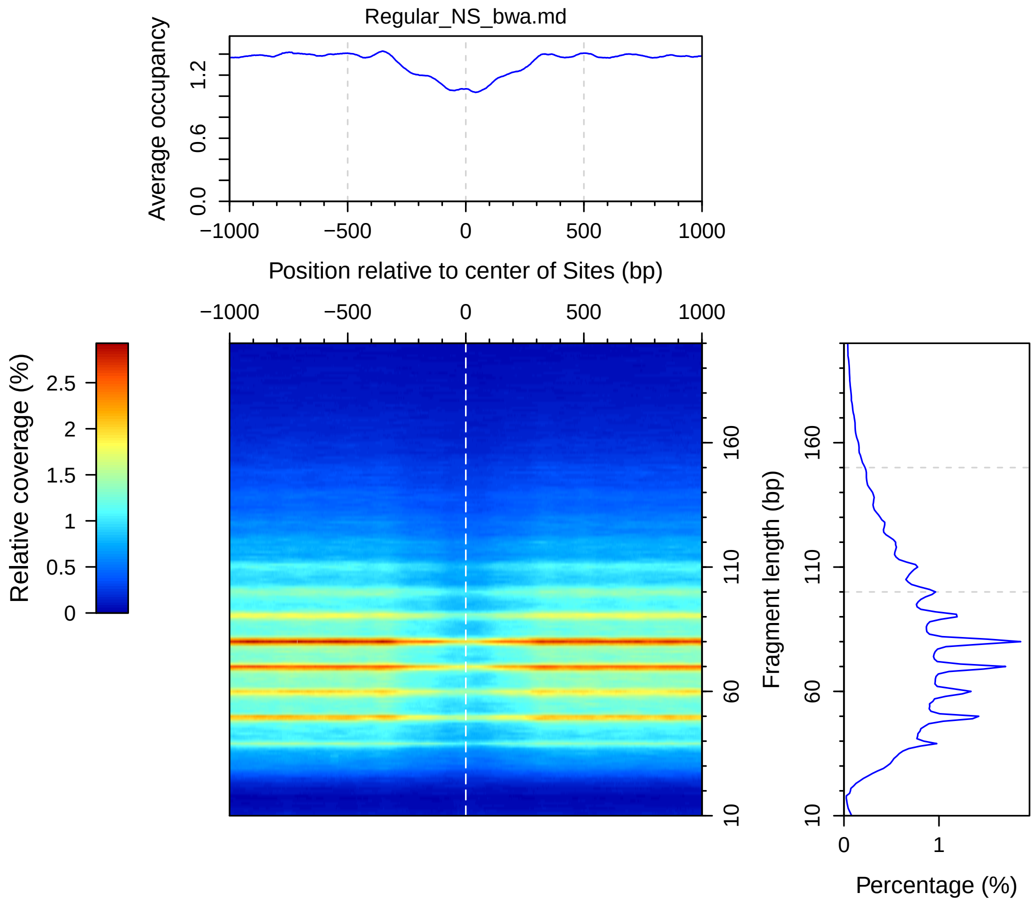

Figure 2.5: Heatmap of sequencing data from basic nitrocellulose filter binding. Figure displays a heatmap of reads from DNA isolated from basic nitrocellulose filter binding experiment (Middle), with x-axis as distance (base pairs) from the averaged center DNase Hypersensitive sites in K562. Y-axis represents fragment length in base pairs. Right of middle represents percentage profile enrichment of the heatmap. Top (above middle) represents enrichment profile over hypersensititve sites.

amino acid residues from each predicted transcription factor in human to calculate isoelectric point. The isoelectric point of a molecule is the pH at which the net charge is zero. Using python and biopython, we queried a list of isoelectric points of human transcription factors which were plotted as a histogram (Figure 2.6.). Consistent with multimodal isoelectric point distributions of proteomes from other organisms, we observed a defined bimodal distribution for human transcription factors. Figure 2 indicates that a substantial fraction of human transcription factors have isoelectric points below a pH of 8. Our filter binding experiments used binding buffers with a slightly basic pH of 8. This indicates that transcription factors with isoelectric points below 8 contained a net negative charge in solution and were likely not bound to the nitrocellulose.

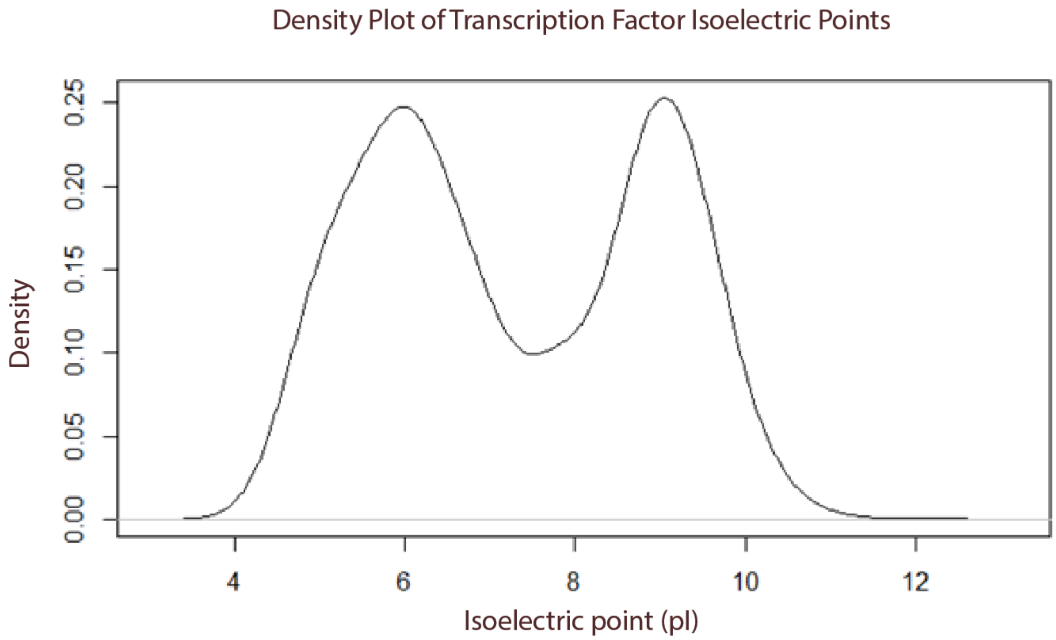

Figure 2.6: Histogram of human transcription factor isoelectric points. X-axis is the isoelectric point between 4 and 12. The y-axis is the proportion of factors.

2.4.4 Acidic Nitrocellulose Filter Binding of Crosslinked K562 Nuclear Lysates Reveals Enhanced Binding to Membrane

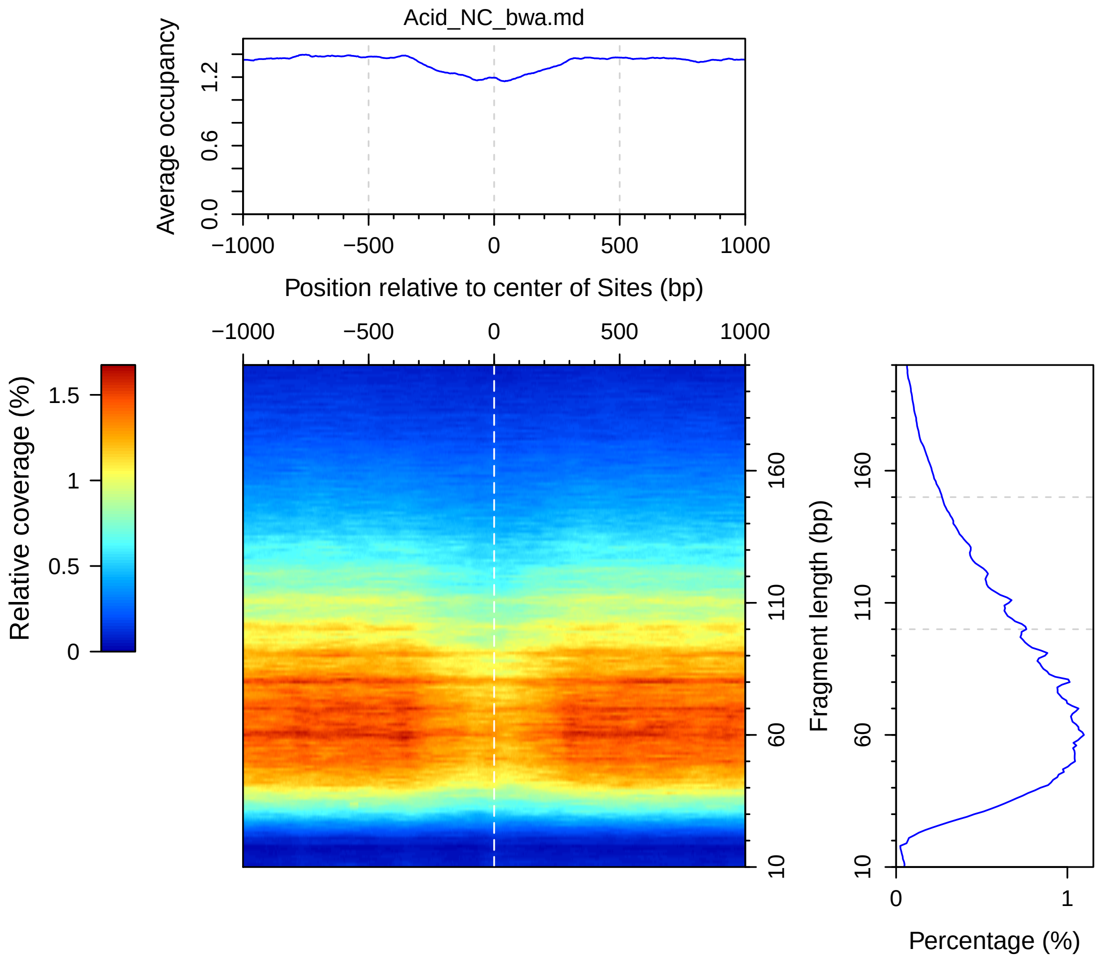

Figure 2.7: Heatmap of reads from DNA isolated from acidic nitrocellulose filter binding. Sequencing results from acidic nitrocellulose filter binding (Middle), with x-axis as distance (base pairs) from the averaged center DNase Hypersensitive sites in K562. Y-axis represents fragment length in base pairs. Right of middle represents percentage profile enrichment of the heatmap. Top (above middle) represents enrichment profile over hypersensititve sites.

To capture the full repertoire of human transcription factors with a variety of isoelectric points, we substituted Tris pH balanced a pH of 8 with hydrochloric acid for a sodium citrate salt balanced to a pH of 4 with citric acid. With a final pH of 4, the sodium citrate buffer would confer a net positive charge to all proteins with an isoelectric point of 4 or higher. We repeated our filter binding experiment with MNase digested K562 nuclear lysate in acidic conditions. DNA eluted from the acidic filter binding experiment was prepared into an Illumina library and sequenced. We evaluated the extent of enrichment over all annotated DNase hypersensitive sites in K562 for both the acidic nitrocellulose and the input sequencing (Figure 2.7). When compared to the normal binding conditions (Figure 2.6) we observe a reduced depletion of reads overlapping hypersensitive sites in the acidic binding. This suggests that a larger fraction of protein-DNA complexes were immobilized onto nitrocellulose under acidic conditions versus basic conditions. However, while the fragments are mostly subnucleosomal in size, they largely appear to overlap proximally to open chromatin. In addition, as seen in the basic nitrocellulose filter binding experiments, we also observed periodicity in the heatmap that corresponds to previous observations of MNase digesting DNA wrapped around the nucleosome [77].

2.4.5 Pan Histone Immunodepletion Depletes Mononucleosomal Fragments and Enriches for Low Molecular Weight DNA

Due to the highly similar result between both filter binding experiments, and the abundance of nucleosomal sequences, we performed a pan histone immunodepletion on crosslinked and digested K562 nuclear lysate. In addition, we used DNase I instead of MNase to remove protein-free DNA and avoid extensively digesting nucleosomal-bound DNA. By targeting the core histone H3, we efficiently depleted mononucleosomal-sized DNA (Figure 2.8). We observed that the nucleosome depleted fraction contained trace amounts of DNA fragments that appeared to be low molecular weight, and strongly depleted for mononucleosomal DNA fragments (Figure 2.8, top). Likewise, in the undepleted fraction (Figure 2.8, bottom), we observed that the overwhelming fraction of DNA was approximately mononucleosomal, with little to no representation of low molecular weight candidate transcription factor bound sequences.

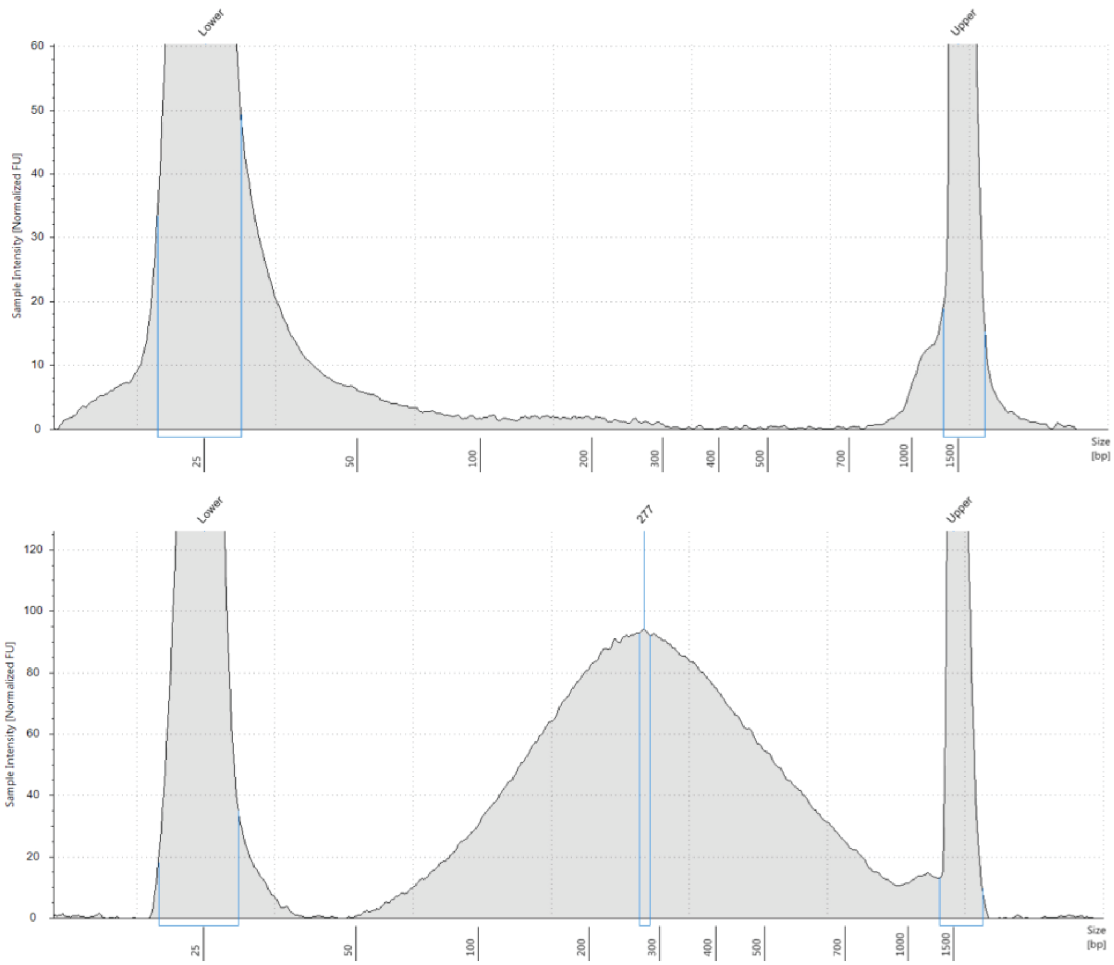

Figure 2.8: Tapestation of DNA fragment size distribution isolated from histone immunodepleted lysates. The fragment size distribution of DNA from input (Top) and after histone immunodepletion (Bottom). X-axis is size of DNA fragments in bp, with upper and lower molecular weight loading standards at 1500bp and 25bp, respectively.

2.4.6 Sequencing of Pan-histone Immunodepleted Digested Luclear Lysates

Optimistic that the nucleosome immunodepletion was successful, we prepared the supernatant fraction for high throughput Illumina sequencing Figure 2.9. We observed that the supernatant fraction had an unexpectedly high fragment size distribution. The majority of the fragment sizes were approximately mononucleosomal in length. Consistent with the change from MNase to DNase, we noticed that the 10bp periodicity seen previously was prevented. In addition, we observed very slight enrichment centered over DNase hypersensitive sites. However, this observation occurred in fragment sizes nearly 3 fold larger than we anticipated would constitute candidate transcription factor bound sequences. No signal of low molecular weight fragments (>60bp) was observed in the heatmap. Altogether, these results suggested that our samples largely contained intact or fragmented nucleosomal DNA.

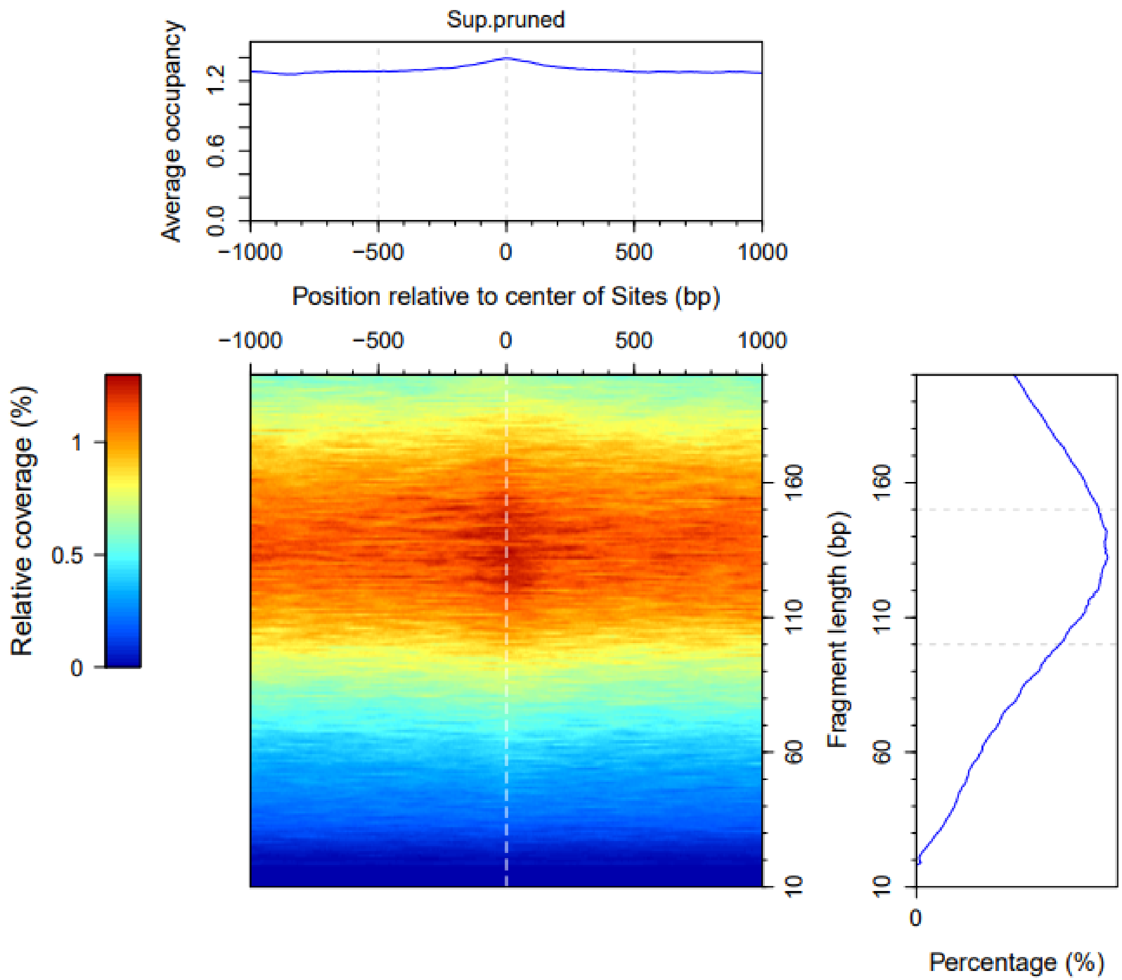

Figure 2.9: Heatmap of reads from DNA isolated from supernatant of histone immunodepleted fraction. Sequencing data from supernatant of histone immunodepleted fraction (Middle), with x-axis as distance (base pairs) from the averaged center DNase Hypersensitive sites in K562. Y-axis represents fragment length in base pairs. Right of middle represents percentage profile enrichment of the heatmap. Top (above middle) represents enrichment profile over hypersensititve sites.

2.5 Discussion

The work presented here was a molecular exploration into developing an auxiliary open chromatin assay that specifically captures transcription factor-bound DNA. If successful, this approach would directly recover non histone bound DNA fragments, opposite of modern open chromatin assays. This would eschew the need for footprinting analyses by providing direct sequencing of DNA protected from nuclease digestion by non nucleosomal proteins. We anticipated that depleting nucleosomes through a variety of chemical methods would leave behind trace amounts of DNA that was protein bound and would map to active regulatory regions where transcription factors were bound. Our efforts showed that, in all cases, the open chromatin regions were not enriched for DNA relative to the regions immediately flanking. Instead, we observed that an enrichment for nucleosomal-bound sequence to varying degrees. This is consistent with a few lines of evidence: 1) DNA fragments that align flanking open chromatin, 2) subnucleosomal fragments with 10bp laddering periodicity, and 3) the highest density of fragments at 70 bp, much larger than transcription factor footprints. Similar digestion patterns of nucleosomes have previously been reported from concentrated MNase digestions on native chromatin, with the most regular cut on the dyad axis [77]. Even when MNase was switched for DNase I, which does not cut on nucleosomal bound DNA, we did not find enrichments within open chromatin (Figure 2.9).

A few circumstances may account for a lack of enrichment for open chromatin regions. The first could be the ratio of histone protein to non-histone DNA binding protein. As the essential eukaryotic packaging protein for the genome, histones comprise a large fraction of nuclear protein. Even our heavily nucleosome-depleted samples (Figure 2.8) likely contain enough nucleosomal sequence so that it can competitively amplify in library preparation. Another possibility is overfixation from formaldehyde crosslinking. Formaldehyde forms covalent bonds with nucleophilic groups on either DNA or amino acids, which is converted to a Schiff base. This Schiff base can further be stabilized by another macromolecule, forming a methylene bridge between the two molecules. While formaldehyde crosslinks colocalizing macromolecules 2Å apart, there is evidence that extended incubations can induce artefactual macromolecular associations and bias ChiP-seq data [78]. Unoptimized formaldehyde crosslinking may be a source of noise in our experimental data, whereby any potential signal in open chromatin is lost due to spurious crosslinking to adjacent nucleosomes. Our approach is likely confounded by a combination of both of the aforementioned circumstances.

To meaningfully address these problems, we have considered alternative approaches in the method's development. The first would be to substitute for an alternative crosslinking step. Recent evidence suggests that high intensity ultraviolet lasers have many advantages in crosslinking over widely-used formaldehyde crosslinking approaches [79]. Specifically, this may mitigate crosslinking artifacts observed in extended formaldehyde treated cells. It is possible that our formaldehyde crosslinking conditions promote transcription factors or regulatory proteins binding in open chromatin regions to crosslink to flanking nucleosomes. This is probable given that our nuclei isolation and crosslinking is simultaneous, meaning that formaldehyde will diffuse directly into nuclei without first traveling through the cytoplasm. In addition to changes in crosslinking, future studies would leverage immunoprecipitations and co immunoprecipitations to assess if known transcription factors or regulatory proteins were depleted alongside the nucleosomal fractions, or likewise, if transcription factors were present in the residual lysate. Optimizing crosslinking and establishing that known transcription factors are not artifactually fixed to nucleosomes would provide critical evidence to proceed in developing this approach. Recent work in bacteria emphasizes the merit of this concept. Freddolino et al. have shown that fixing Escherichia coli and isolating nuclease digested DNA-protein complexes through phase extraction yields transcription factor occupancy peaks at regulatory loci [80]. Developing a similar approach for eukaryotic systems would be widely beneficial to functional genomics annotation.

2.6 Notes and Acknowledgements

The work in this chapter would not have been possible without the indispensable contributions of my collaborators from Dr. Peter Freddolino's lab. I developed the nuclei isolation and immunodepletions, and executed all experimental work. Nitrocellulose filter binding protocol (Basic and acidic) was used with permission from Dr. Peter Freddolino and executed with the assistance of Dr. Rebecca Hurto. Dr. Alan Boyle performed analysis of the sequencing data from the nitrocellulose and immunodepletion experiments. The work in this chapter remains unpublished.

CHAPTER III

Cas9 Targeted Enrichment of Mobile Elements Using Nanopore Sequencing

3.1 Abstract

Mobile element insertions (MEIs) are repetitive genomic sequences that contribute to genetic variation and can lead to genetic disorders. Targeted and whole-genome approaches using short-read sequencing have been developed to identify reference and non-reference MEIs; however, the read length hampers detection of these elements in complex genomic regions. Here, we pair Cas9-targeted nanopore sequencing with computational methodologies to capture active MEIs in human genomes. We demonstrate parallel enrichment for distinct classes of MEIs, averaging 44% of reads on-targeted signals and exhibiting a 13.4-54x enrichment over whole-genome approaches. We show an individual flow cell can recover most MEIs (97% L1Hs, 93% AluYb, 51% AluYa, 99% SVA_F, and 65% SVA_E). We identify seventeen non-reference MEIs in GM12878 overlooked by modern, long-read analysis pipelines, primarily in repetitive genomic regions. This work introduces the utility of nanopore sequencing for MEI enrichment and lays the foundation for rapid discovery of elusive, repetitive genetic elements.

3.2 Introduction

At least 45% of the human genome is composed of transposable element (TE)-derived sequences[81]. TEs can be subdivided into four major categories: (i) DNA transposons; (ii) long terminal repeat (LTR) retrotransposons; (iii) long interspersed elements (LINEs); and (iv) short interspersed elements (SINEs). L1 (or, LINE-1) represents a subclass of LINEs and L1-derived sequences comprise approximately 17% of the human genome[81, 82]. Alu elements, a subclass of SINEs, are ancestrally derived from a dimerization of the 7SL RNA gene and make up 11% of the human genome, spread out over 1 million copies[83]. SVA (SINE-VNTR-Alu) elements are active chimeric elements that have recently evolved and are derived from a SINE-R sequence coupled with a VNTR (variable number of tandem repeats) region and an Alu-like sequence[84]. An average human genome contains approximately 80-100 active full-length human-specific L1s (L1Hs)[85, 86, 87] and a small number of highly active, or "hot," L1Hs sequences that are responsible for the bulk of human retrotransposition activity[85, 86, 88, 89]. This includes the mobilization of Alus and SVAs which require trans-acting factors from L1s to transpose[83]. Collectively, the result of such recent mobilization events are referred to as mobile element insertions (MEIs).

Regions harboring these repetitive elements have long been considered part of the 'dark matter' of the genome with no expected impact on human phenotypes. However, recent studies indicate that at least some recent insertions indeed play a functional role in various aspects of the cell. L1-mediated retrotransposition events can be mutagenic, and germline retrotransposition events within the exons or introns of genes can result in null or hypomorphic expression alleles, leading to sporadic cases of human disease[90]. In addition, Lubelsky and Ulitsky demonstrated that sequences enriched in Alu repeats can drive nuclear localization of long RNAs in human cells[91]. An SVA element insertion was recently reported in an intron of TAF1 that ablated expression through aberrant splicing, and is a driving mutation in X-linked Dystonia-Parkinsonism[92]. Another study showed that a recurrent intronic deletion results in the exonization of an Alu element that is found in 6% of families with mild hemophilia A in France[93]. Somatic L1 retrotransposition can occur in neuronal progenitor cells[94, 95, 96, 97, 98], indicating a possible role for L1s in the etiology of neuropsychiatric diseases[99]. In addition, a mutagenic L1 insertion that disrupted the 16th exon of the APC gene has been shown to instigate colorectal tumor development[89]. Beyond a widespread repertoire of disease associations, mobile elements also influence large scale genome structure. Recent work has demonstrated that transposition events are associated with three dimensional genome organization and the evolution of chromatin structure in human and mouse[100, 101, 102].

A tremendous effort has been made to understand the varied functional outcomes of active MEIs. Similar efforts are underway to capture and resolve MEIs to discover additional avenues of genetic pathogenesis[88, 103, 104, 105, 106, 107, 108]. While transformative, these studies were confounded by the shortcomings of existing sequencing methodologies and bioinformatics pipelines, and limited in their ability to access a large (50%), highly repetitive proportion of the genome[109]. The difficulty in uniquely aligning short-read sequences to repetitive genomic regions likely leads to an under-representation of MEIs that have inserted within these regions. Several tools that have been developed to identify non-reference MEIs from whole genome short-read data, including the Mobile Element Locator Tool (MELT)[107], Mobster[110], Tangram[111], TEA[112], and others, are further restricted by the short read length and repetitive nature of mobile elements when resolving longer, non-reference insertions, such as L1Hs and SVA[109, 113]. Experimental approaches from paired-end fosmid sequencing[88, 114] to PCR capture-based approaches[115, 103, 116, 117, 105, 118] have been developed to capture MEIs, but they have the disadvantage of being low throughput. Recent methods[106, 109, 119, 120] combining short-read sequencing and MEI 3' end capture techniques provide a more reliable way for the MEI discovery. Such approaches can be used in the investigation of single-cell MEI profiles[106, 109], yet they too are hindered by the aforementioned disadvantages due to the short-read dependence.

The advent of long-read sequencing technologies provides a powerful tool for characterizing repeat-rich genomic regions by providing substantially longer sequence reads compared to traditional short-read platforms[121, 122]. We have recently applied these technologies to demonstrate that there are at least 2-fold more polymorphic L1Hs sequences in human populations than previously thought[109]. Several existing tools and pipelines have the ability to resolve reference and non-reference MEIs in the human genomes; however, most require a whole-genome pipeline for haplotype assembly, local assembly, or cross-platform support[123, 124, 125, 126]. This often necessitates whole genome long-read sequencing, which is currently cost-prohibitive at scale and precludes an in-depth exploration into the impact of MEIs on human biology and disease. One solution to these barriers is the application of Cas9 targeted sequence capture with long read sequencing that allows for alignment to unique flanking genomic regions[127]. This approach significantly lowers costs and enables a focused and efficient computational analysis for MEI discovery. Here, we demonstrate the utility of an in vitro Cas9 enrichment of targeted sequence elements combined with Oxford Nanopore long-read sequencing and established computational methodologies to identify a set of MEIs (L1s, AluYs, and SVAs)[127] that account for over 80% of currently active mobile elements in the human genome[128, 129, 130]. This technology has been previously utilized to resolve a variety of genomic structural variants, including diseases associated with polynucleotide repeats and oncogenic translocation events[131, 132, 133, 127]. By targeting the Cas9 to subfamily-specific sequences within each element and developing a computational pipeline (Nano-Pal) for analysis of Cas9-enriched nanopore sequencing data, we demonstrate enrichment of mobile elements across the genome that are both annotated and unannotated in the GRCh38 reference (reference and non-reference MEIs, respectively).

3.3 Results

3.3.1 Cas9 Targeted Enrichment Strategy for Mobile Elements Using Nanopore Sequencing

Cas9 targeted enrichment strategy for mobile elements using nanopore sequencing We chose GM12878 (NA12878), a member of the CEPH pedigree number 1463 (GM12878, GM12891, GM12892)[134], as the benchmark genome in this study. GM12878 is one of the most thoroughly investigated human genetic control samples and has been used in many large-scale genomic projects, such as HapMap[135], 1000 Genomes Project[136, 137, 114], the Human Genome Structural Variation Consortium[113, 126], Genome In A Bottle[138, 139], and reference genome improvement projects[140]. To precisely capture MEIs of interest, we applied Cas9 targeted nanopore sequencing[131, 132, 133, 127] to enrich for five active subfamilies of MEIs (L1Hs, AluYb, AluYa, SVA_F, and SVA_E) in GM12878 (Figure 3.1) as well as L1Hs in the corresponding parental samples.

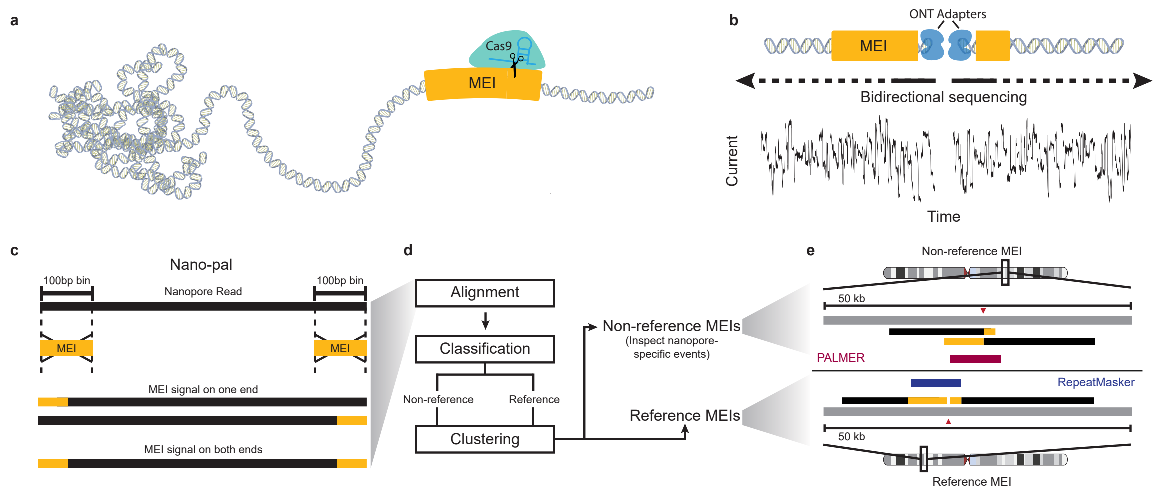

Figure 3.1: A schematic Cas9 targeted enrichment and Nano-Pal pipeline for mobile elements using nanopore sequencing. a, Purified genomic DNA (gDNA) is isolated by salting out and then extensively dephosphorylated. Dephosphorylated gDNA is incubated with the Cas9 ribonucleoprotein which is targeted to MEI subfamily-specific sequences near the 3' end of the element. Taq polymerase (not shown), and dATPs (not shown) monoadenylate DNA ends. b, Cas9 cleaved sites are ligated with Oxford Nanopore Technologies (ONT) sequencing adapters and sequenced on a flow cell. Sequencing is bi-directional from the cleavage site. c, Nano-Pal scans the nanopore sequencing reads (black bars) after Cas9 enrichment for MEI signal on one or both ends. The yellow bar represents MEI consensus sequence or MEI signals in pairwise comparison of Nano-Pal. d, All reads with or without annotated MEI signal are imported into the downstream pipeline. Alignment, classification, and clustering processes are sequentially conducted. Nano-Pal identifies reference and non-reference MEIs followed by the inspection of nanopore-specific non-reference MEIs (see Methods). e, Examples illustrating capture and alignment of reads containing non-reference L1Hs signal (top) and reference L1Hs signal (bottom). Aligned reads display a non-reference insertion (top) with L1Hs signal (yellow bar) and flanking genomic sequence (black bar). MEI components of reads in non-reference insertions are displayed as overlapping (soft clipping) due to lack of reference genome MEI annotation (grey bar). Aligned reads display annotated reference L1Hs (bottom, yellow bar), flanked by surrounding genomic sequence (black bar), separated by the Cas9 cleavage site (red triangle). PALMER and RepeatMasker tracks are illustrated in red and blue, respectively.

We designed guide RNAs using unique subfamily-specific sequences within each element to maximize the specificity of Cas9 targeting (Figure 3.2, see Methods). Using this approach, we generated a list of candidate 20bp guide RNAs for each MEI category (Figure 3.2a-c). We selected an L1Hs candidate guide RNA with the 'ACA' motif at 5929bp of the L1Hs consensus sequence, which distinguishes the L1Hs subfamily from other L1 subfamilies (e.g. L1PA)[130, 103, 119]. For AluY and SVA, 18 unique guide RNA candidates were obtained (one for AluYb, three for AluYa, seven for SVA_F, and six for SVA_E) (see Methods, Figures Supplementary Figure 1*, Supplementary Figure 2*, and 3.3 full list not shown). Candidates were further prioritized to those with the largest number of subfamily-specific bases and proximity to the 3' end of the MEI sequence, near the polyA tail, which is an obligate component of the TPRT mechanism of retrotransposition[141, 142]. From the pool of candidates, a single guide RNA for each MEI subfamily was selected for downstream enrichment experiments (Figure 3.3). After Cas9 enrichment and read processing, we assessed the cleavage sites of all five guide RNAs (Figure 3.2d and Figure 3.4). The resulting distribution showed a vast majority of the forward-strand reads start at the third or fourth base-distance from the 'NGG' PAM site, and reverse-strand reads begin at the seventh base. This is consistent with previous characterization studies of Cas9 cleavage activity[143]. Furthermore, we observed strand bias with approximately 4.6-fold more reads on the forward strand compared to the reverse strand, which has been hypothesized to be caused by Cas9 remaining bound after cutting and obstructing adapter ligation and sequencing[144, 127]. We detected directional sequencing biases within different MEI subfamilies and enrichment runs (Supplementary Data 2*, Figure 3.4).

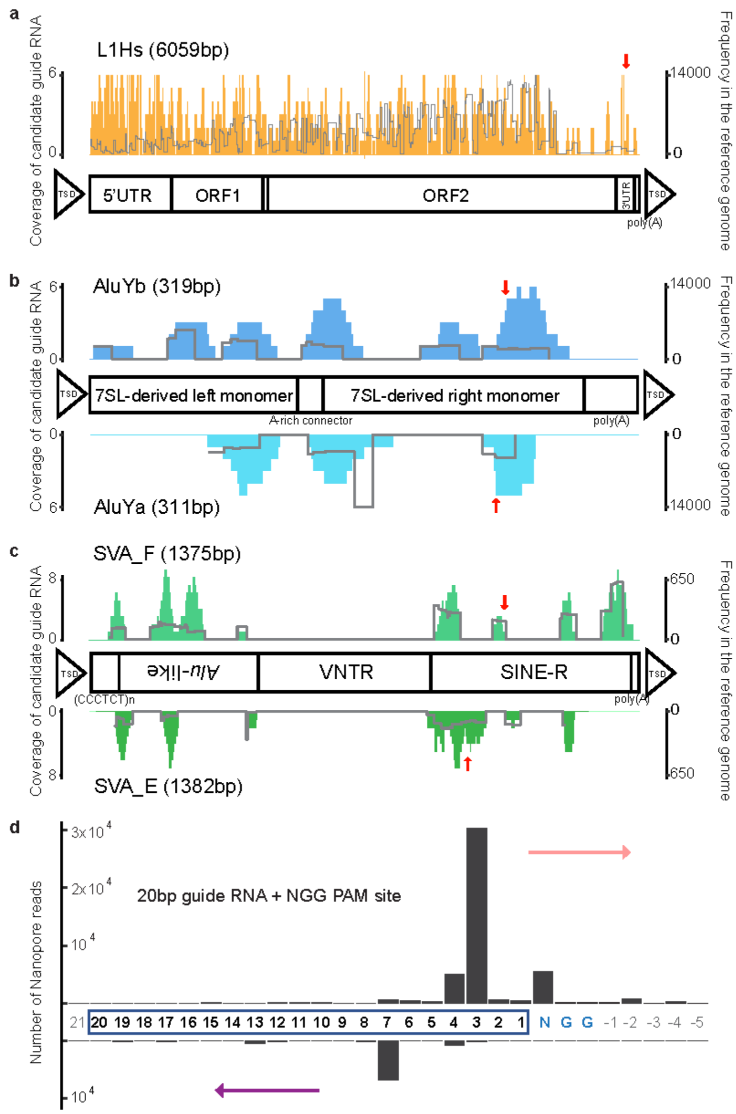

Figure 3.2: Guide RNA design for MEIs and guide RNA cleavage-site distribution. a, Distributions of candidate guide RNAs (left Y-axis and the histogram) in the L1Hs consensus sequence and structure information. The right Y-axis and the line indicate frequency of corresponding candidates in the reference genome sequence. b, Upper panel shows the distribution for AluYb and the lower panel for AluYa. c, Upper panel shows the distribution for SVA_F and the lower panel for SVA_E. Red arrows in a, b, and c indicate where the selected guide is. d, Cleavage-site distribution of all guide RNAs in this project. The x-axis indicates the position where the read ends or begins, with the number depicting the base distance from the PAM site (NGG). The PAM site (NGG) is colored blue and guide RNA bases are highlighted by a rectangle. Bases outside of the guide RNA or the PAM site are colored grey. The y-axis is the number of nanopore reads counted. The upper bar represents reads with forward strand sequencing outward from the 3' end of the guide RNA (rose arrow) and the lower bar represents reads with reverse strand sequencing outward from the 5' end of the guide RNA (purple arrow).

L1Hs	5′-CTAATGTGTCATCTAGCATT-AGG-3′
AluYb	5′-CGCCACTGCAGTCCGCAGTC-CGG-3′
AluYa	5′-CCAGGCTGGAGTGCAGTGGC-GGG-3′
SVA_F	5′-TCAACAGGATCCCAAGGCAG-AGG-3′
SVA_E	5′-TGAGAAATCGGATGGTTGCC-GGG-3′

Figure 3.3: Five final guide RNAs MEIs. Figure shows the composition of forward/reverse strand reads based on these guide RNA sequence.

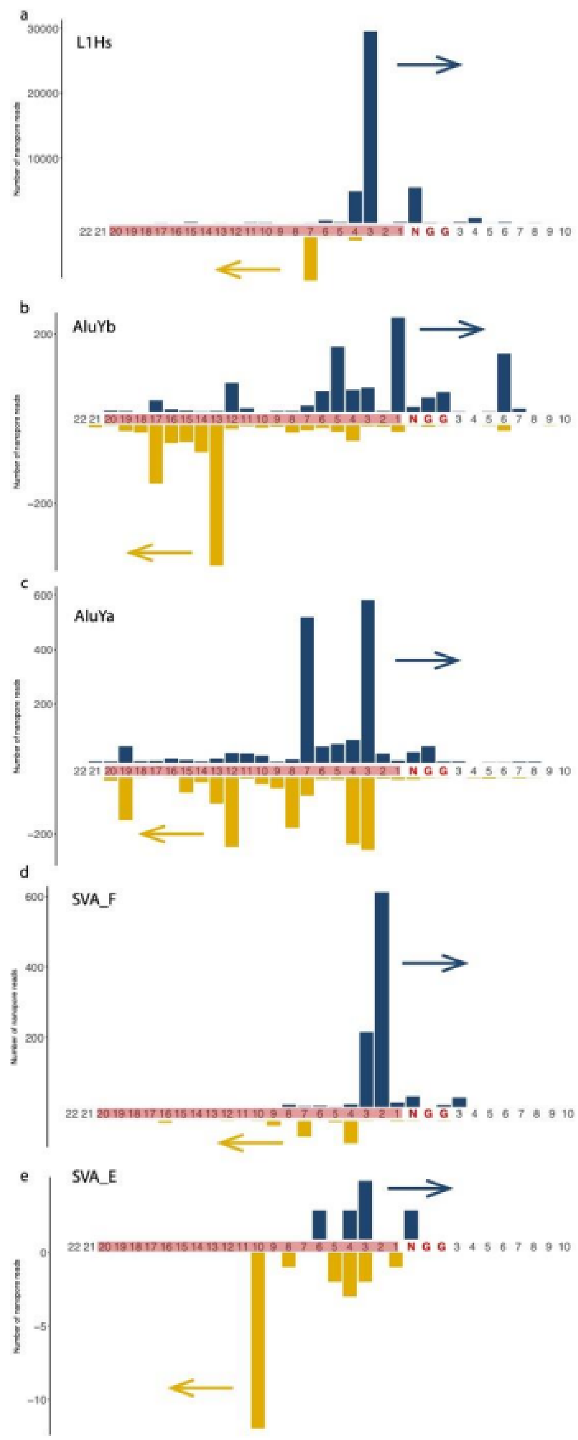

Figure 3.4: Distributions of guide RNA cleavage-site for five MEI subfamilies. a, Cleavage-site distribution of L1Hs guide RNA. X-axis shows the position where the read ends or begins with the number indicating the distance from the ‘N’ of the PAM site (NGG). The PAM site (NGG) was colored red and guide RNA bases were highlighted by red background. Y-axis is the number of nanopore reads counted. The upper blue bar represents the reads with forward strand sequencing outward the 3’ end of guide RNA and the lower yellow bar represents the reads with reverse strand sequencing outward the 5’ end of guide RNA. b, Cleavage-site distribution of AluYb guide RNA. c, Cleavage-site distribution of AluYa guide RNA. d, Cleavage-site distribution of SVA_F guide RNA. e, Cut-site distribution of SVA_E guide RNA

We developed a computational pipeline, Nano-Pal, to analyze captured long reads after base-calling and trimming, estimate the on-target rate of Cas9 enrichment from MEI signals on the ends of reads, and identify reference and non-reference MEIs (Figure 3.1b,c). Due to the frequency of targeted MEIs in the genome, an individual nanopore read may harbor a MEI signal on one or both ends. Reads with a single-end MEI signal had similar read lengths within all MEI experiments, yet were significantly larger than reads with two-end MEI signals. This was especially true in L1Hs experiments (L1Hs 1.9-fold, AluY 1.1-fold, SVA 1.4-fold, Figure 3.5). To better distinguish non-reference MEI signals from those present in the reference, particularly where non-reference MEIs are embedded into reference MEIs[109], the pre-masking module from an enhanced version of our long-read non-reference MEI caller, PALMER[109], was implemented into Nano-Pal (Figure 3.1d). This enabled identification and masking of reference MEIs in individual long-reads, enhancing detection of non-reference MEI signals within the remaining unmasked portion[109]. Non-reference and reference MEIs were then summarized by clustering nearby nanopore reads.

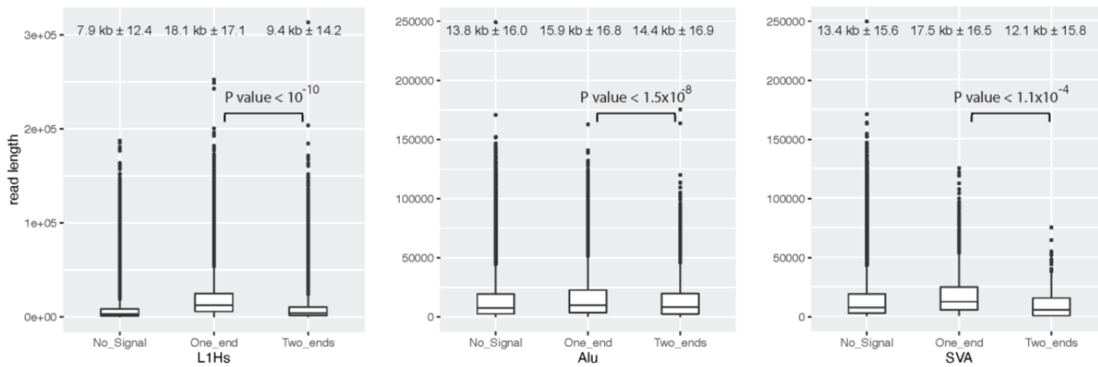

Figure 3.5: Read length distributions for MEI categories (L1Hs, AluY, and SVA). All reads were identified into three categories: reads with on-target reference MEI signals, ones with on-target non-reference MEI signals, and ones with no signals (off-target). The read length (mean standard deviation) are shown above the boxplots. And the P-value (student's T-test, two-tailed) are shown between the boxplots of one-end reads and two-end reads. Error bars range from Q1−1.5IQR to Q3+1.5IQR (IQR, interquartile range).

3.3.2 Cas9 Targeted Enrichment Efficiently Captures Mobile Element Signals in Nanopore Reads

To estimate the enrichment efficiency for different flow cells and MEI subfamilies, all passed reads were classified into three categories: on-target, close-target, and off-target (see Methods). The on-target rate for nanopore reads from a single L1Hs experiment on a Flongle flow cell, including both reference and non-reference MEIs, was 56.9%. Relatively lower on-target rates were observed for the other MEIs on the Flongle flow cell: 46.7% for AluYb, 23.8% for AluYa, 5.8% for SVA_F, and 2.3% for SVA_E (Table 3.1). When an L1Hs enrichment experiment was sequenced on a MinION flow cell, the on-target rate was approximately 35.0% (FAL11389) and 23.3% for a pooled MEI run (FAO84736). Compared to earlier studies (2.09% in Flongle and 4.61% in MinION)[127], these results show substantially improved enrichment, with a 1- to 25-fold increase relative to the Flongle flow cell, and over 5-fold enrichment relative to the MinION flow cell. These enrichment increases are likely due to the frequency of the targets in the genome. Overall, our approach reaches an average of 44% of nanopore sequencing reads with target MEI signal from these seven flow cell runs.

MEI	Run	Flow cell	Read number	On-target		Close-target
				Reference	Non-reference	Reference
Individual						
L1Hs	ABB607	Flongle	4,102	49.6%	7.3%	16.8%
AluYb	ACK645		2,271	40.2%	6.5%	1.0%
AluYa	ACK655		12,513	18.0%	5.8%	10.3%
SVA_F	ACK629		14,106	3.7%	2.1%	3.7%
SVA_E	ABO395		7,297	1.7%	0.6%	0.2%
L1Hs	FAL11389	MinION	110,029	30.7%	4.3%	37.9%
Pooled	FAO84736	MinION	105,410	20.1%	3.2%	38.4%
L1Hs				8.9%	1.6%	33.9%
AluYb				7.0%	1.2%	3.2%
AluYa				2.8%		
SVA_F				1.2%	0.4%	1.3%
SVA_E				0.2%		

Table 3.1: Summary of seven representative flow cells. Five individual Flongle flow cells for L1Hs (ABB607), AluYb (ACK645), AluYa (ACK655), SVA_F (ACK629), and SVA_E (ABO395) each, one individual MinION flow cell for L1Hs (FAL11389), and one pooled MinION flow cell for five MEIs (FAO84736)

To further assess the improved enrichment of MEI subfamilies, the extent of representation of MEI targets with high sequence identity were examined within the data. A portion of the enrichment data contained 'close-target' reads that resemble related subfamilies of the intended targets and can be explained by the base mismatch tolerance between the guide RNA sequence and the targeted MEI sequence[145, 146]. For L1Hs, a rate of 16.8% on the Flongle and 33.9% to 37.9% on the MinION flow cell was observed, with close-target reads mapping to reference L1PA regions. Flongle sequencing of AluYa had a rate of 10.3% of close-target reads to other reference AluY elements, in contrast to AluYb where a dramatically reduced 'close-target' rate of 1.0% was observed (Table 3.1). This enhanced specificity may be explained by a specific insertion sequence within AluYb (5'-CAGTCCG-3') that was included in the guide RNA and is unique to the youngest Alu elements (AluYb)[129] of the genome. For the SVA Flongle sequencing, 'close-target' rates of 3.7% and 0.2% to the other reference SVAs were observed in the SVA_F and SVA_E enrichment, respectively.

We next assessed the efficiency of our target enrichment of MEIs compared to whole-genome sequencing (WGS) approaches. A recent, related study used a whole-genome nanopore sequencing approach[125] to study MEIs and methylation and provides an excellent benchmark to which we may compare our results. When taking total sequenced reads into account, our targeted approach exhibited between a 13.4 to 54 fold increase in the average number of reads per MEI compared to WGS (Figure 3.6). Furthermore, our read length N50 ranged from 14.9Kbp to 32.3Kbp compared to 5.14Kbp to 10.57Kbp reported in Ewing et al, suggesting that our targeted approach also results in a higher number of MEI spanning reads. Overall, these comparisons indicate that on the basis of per-base sequenced, MEI target capture exhibits significant enrichment advantages over whole genome approaches.

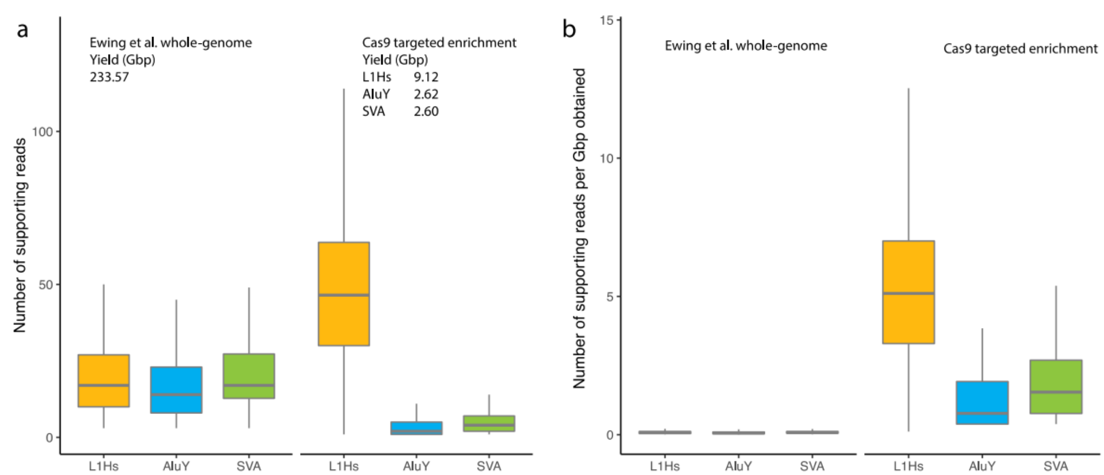

Figure 3.6: Number of supporting reads for three categories of non reference MEIs from the Cas9 targeted nanopore sequencing and the whole-genome nanopore sequencing by Ewing et al. 2020. a, Number of supporting reads for non-reference L1Hs, AluY, and SVA in the whole-genome nanopore sequencing (from five PromethION flow cells) and the Cas9 targeted nanopore sequencing (L1Hs from 11 MinION/Flongle flow cells, AluY from four from 4 MinION/Flongle flow cells, and SVA from 4 MinION/Flongle flow cells). b, Number of supporting reads of non-reference MEIs normalized by the total yield base pairs from flow cells in two studies.

3.3.3 Cas9 Enrichment and Nanopore Sequencing Rapidly Saturates Reference and Non-Reference MEIs

Due to the possibility that the guide RNA may bind to off-target sites and direct Cas9 to cut in MEIs that are not perfectly matched to the guide sequence[145, 146], we established expectation thresholds to evaluate the number of captured reference and non-reference MEIs. The reference MEI sets were obtained from RepeatMasker[147]. A 'PacBio-MEI' callset in GM12878 was generated by comprising a mapping-based callset and an assembly-based callset as a comprehensive gold standard set for non-reference MEIs (see Methods). The PacBio-MEI includes 215 L1Hs, 362 AluYb and 593 AluYa in 1404 Alus, and 33 SVA_F and 24 SVA_E in 72 SVAs (Figure 3.7 and Supplementary Data 4*). Three categories (lower, intermediate, and upper) were defined that contain a number of reference and non-reference MEIs from each subfamily (Table 3.2). Each threshold classifies MEIs depending on the extent of allowed mismatches between the guide and the sequence. The lower-bound is the most stringent and requires a perfect match between the guide sequence and the MEI. The intermediate-bound is less stringent and can tolerate three or fewer mismatches. We consider this to be the closest estimation to the number of MEIs that a guide RNA could reasonably capture among these three boundaries. Finally, the upper-bound is the least stringent and most inflated, requiring that at least 60% of the guide sequence matches the MEI (see Methods).

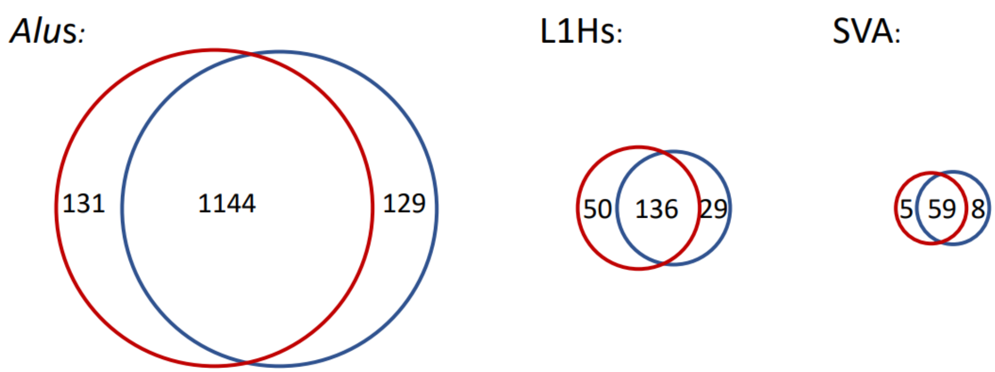

Figure 3.7: Venn diagram of the PALMER callset and the PAV callset for non reference MEIs in NA12878 genomes. PALMER callset (red circle) is from PacBio raw sub-reads, and PAV (the Phased Assembly Variant caller, https://github.com/EichlerLab/pav, see Methods) callset (blue circle) is from PacBio assembly-based pipeline. The circles are depicted by the scale of the numbers showed inside. The union of two sets is generated as 'PacBio-MEI' to be the gold standard set to compare with the calls from nanopore data.

Upper-bound

Reference_L1Hs	Reference_L1PA	Other_reference_L1		Non_reference_L1Ta	Non_reference_L1PreTa	Non_reference_L1Ambig
1,139	101,631	496,786		144	7	64
64.5303%	0.4133%	0.0002%		61.8056%	71.4286%	31.2500%
79.7191%	7.2980%	0.0046%		95.8333%	85.7143%	43.7500%
76.2950%	7.0825%	0.0054%		92.3611%	71.4286%	46.8750%
Reference_AluYb	Reference_AluYa	Other_reference_AluY	Other_reference_Alu	Non_reference_AluYb	Non_reference_AluYa	
3,056	4,310	135,419	964,968	362	593	
22.9058%	0.0000%	0.0140%	0.0127%	16.0221%	0.0000%	
0.2291%	36.1485%	0.6912%	0.4363%	1.1050%	29.0051%	
68.8154%	42.9002%	1.9133%	1.4522%	60.4972%	36.0877%	
Reference_SVA_F	Reference_SVA_E	Other_reference_SVA		Non_reference_SVA_F	Non_reference_SVA_E	
393	147	2,440		33	24	
39.9491%	15.6463%	11.3525%		57.5758%	8.3333%	
0.5089%	52.3810%	0.4918%		0.0000%	33.3333%	
68.9567%	61.2245%	35.5738%		84.8485%	41.6667%	

Intermediate

Reference_L1Hs		Non_reference_L1Ta	Non_reference_L1PreTa	Non_reference_L1Ambig
905		134	6	21
81.2155%		66.4179%	83.3333%	95.2381%
100.0000%		100.0000%	100.0000%	100.0000%
96.0221%		100.0000%	83.3333%	100.0000%
Reference_AluYb	Reference_AluYa	Non_reference_AluYb	Non_reference_AluYa	
2,240	3,506	249	510	
31.2500%	0.0000%	23.2932%	0.0000%	
0.3125%	44.4381%	1.6064%	33.7255%	
93.8839%	52.7382%	87.9518%	41.9608%	
Reference_SVA_F	Reference_SVA_E	Non_reference_SVA_F	Non_reference_SVA_E	
221	131	29	24	
71.0407%	17.5573%	65.5172%	8.3333%	
0.9050%	58.7786%	0.0000%	33.3333%	
100.0000%	68.7023%	96.5517%	41.6667%	

Lower-bound

Reference_L1Hs	
479	
100.0000%	
100.0000%	
100.0000%	
Reference_AluYb	Reference_AluYa
1,859	3,177
37.6547%	0.0000%
0.3765%	49.0400%
100.0000%	58.1996%
Reference_SVA_F	Reference_SVA_E
192	92
81.7708%	25.0000%
1.0417%	83.6957%
100.0000%	97.8261%

Table 3.2: Known MEIs captured by nanopore Cas9 enrichment approach in different flow cells based on different boundaries. Upper-bound, intermediate, and lower-bound values of different categories of MEIs are included regarding background (number) and seven representative flow cells (percentage).

Upon comparing our MEI enrichment data to the aforementioned intermediate value estimates (Figure 3.8a,b, Figure 3.9, and Table 3.2, Supplementary Data 6*), the individual and the pooled MinION flow cell captured 100% (35.8 mean coverage) and 96.0% (10.5 mean coverage) of known reference L1Hs with on-target reads, respectively. The individual Flongle captured 81.2% (2.7 reads mean coverage) of known reference L1Hs. For the non-reference L1Hs, the individual and pooled MinION flow cell captured 100% and 99.4% for all the L1Hs subfamilies, respectively. The individual Flongle captured 66.4%, 83.3%, and 95.2% for non-reference L1Ta, L1PreTa, and L1Ambig, respectively. Our results showed that only one of the MinION flow cells (FAL11389 or FAO84736) was necessary to capture most of the known reference and estimated non-reference L1Hs subfamilies in the genome when considering intermediate values, indicating a very high sensitivity of guide RNA targeting in the experiments.

Compared to the least stringent upper-bound estimates, 64.5% and 79.7% of known reference L1Hs were captured using the individual Flongle and MinION flow cell, respectively. Non-reference L1Hs capture ranged from 53.0% to 80.0%, compared to 4.1% to 7.3% of close-target L1PA elements, and less than 0.01% of off-targeting to other L1 elements (Figure 3.8a,b, Figure 3.9, and Table 3.2, and Supplementary Data 6*). The high percentage of elements captured that were on-target versus the other categories, including off-target, indicates the high specificity of the guide RNA to L1Hs in the enrichment. The read depth of the reference and non-reference L1Hs elements observed in these MinION flow cells has an approximate ratio of 2:1 (Figure 3.8b), consistent with the expectation that reference MEIs are homozygous, and a considerable portion of non-reference MEIs are heterozygous[148].

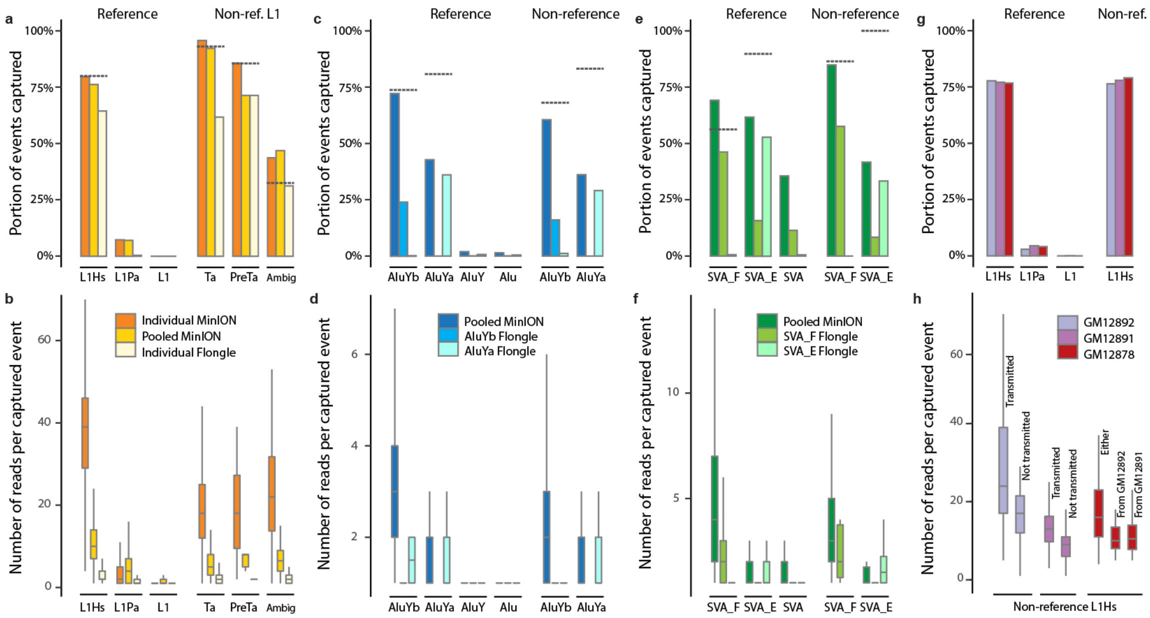

Figure 3.8: Systematic evaluation of known MEIs captured by nanopore cas9 enrichment approach in different flow cells. a, Known L1Hs in GM12878 recovered by Cas9 targeted enrichment from the individual MinION flow cell (FAL11389), pooled-MEI MinION flow cell (FAO84736), and individual Flongle flow cell (ABB607), displayed as a proportion of the upper-bound known reference L1Hs, L1Pa, and other L1 as well as non-reference (non-ref.) L1Hs from the PacBio-MEI set. Non-reference L1Hs were divided into different subfamilies (L1Ta, L1PreTa, and L1Hs with ambiguous subfamilies). Dotted-grey line represents the intermediate values (as proportion) of MEIs that the guide RNA binds when allowing a $\leq$3bp mismatch or gap. b, Number of supporting reads of each captured L1 in the context of a. c, Known AluY elements in GM12878 recovered by Cas9 enrichment on one pooled MinION flow cell (FAO84736), one individual AluYb Flongle flow cell (ACK645), and one individual AluYa Flongle flow cell (ACK655). d, The number of supporting reads of each captured Alu element in the context of c. e, Known SVA elements in GM12878 recovered by Cas9 enrichment on one pooled MinION flow cell (FAO84736), one individual SVA_F Flongle flow cell (ACK629), and one individual SVA_E Flongle flow cell (ACK395). f, The number of supporting reads of each captured SVA element in the context of e. g, Known L1Hs captured in the GM12878 trio by Cas9 enrichment on one pooled MinION flow cell (FAL15177). h, The number of supporting reads of each captured non-reference L1Hs based on transmission in the GM12878 trio. The non-reference L1Hs in the parents (GM12892 and GM12891) were categorized into transmitted and not-transmitted. The non-reference L1Hs in the child (GM12878) were categorized as insertions inherited from GM12892 or GM12891, and from either parents (unknown parental lineage). In b, d, f, h, the numbers of captured MEI subfamily can be found in Supplementary Data 6 with information of mean and standard deviation; The error bars of boxplot range from Q1−1.5IQR to Q3+1.5IQR (IQR, interquartile range) and outliers are not shown.

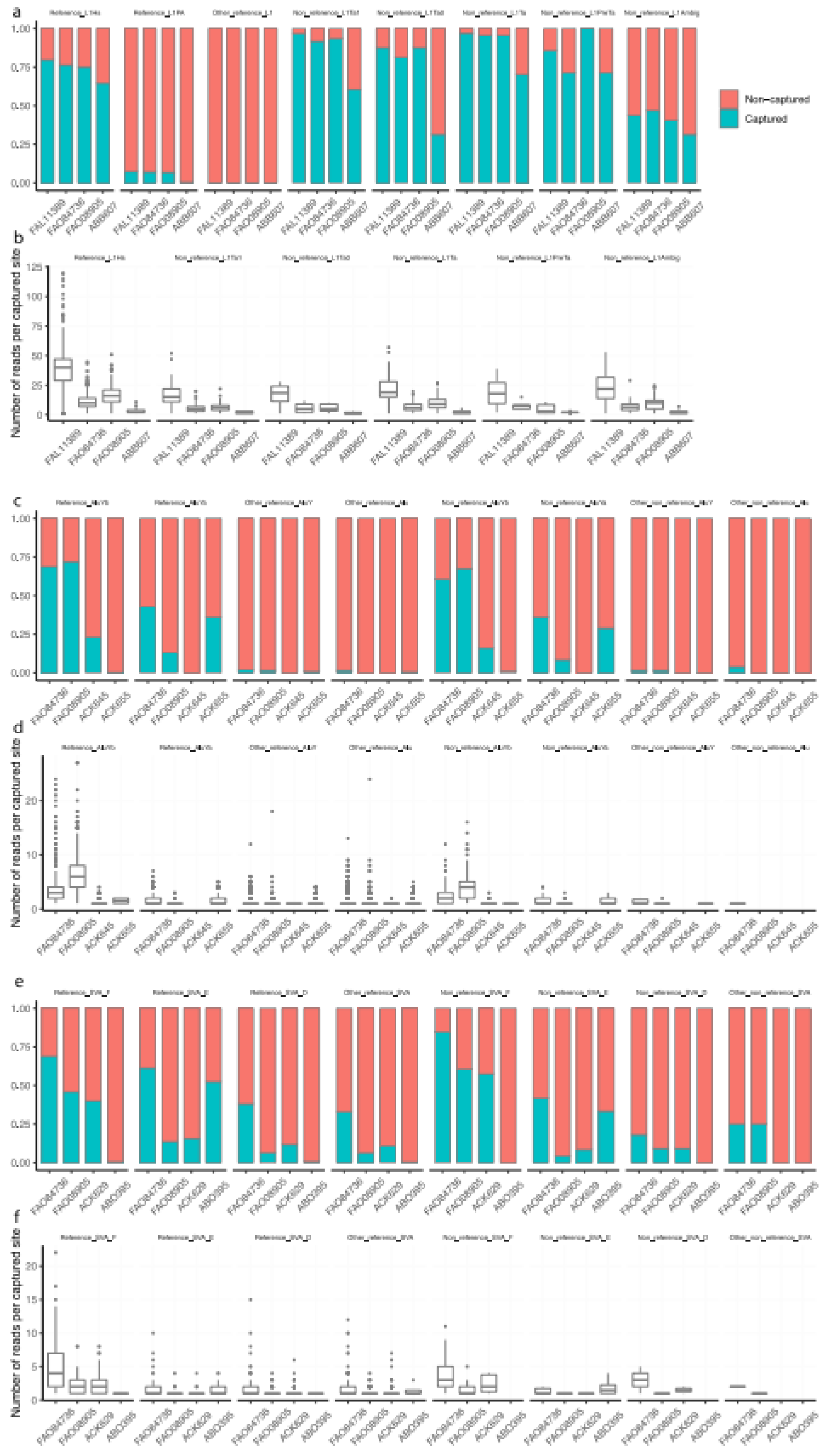

Figure 3.9: Summary of recovered known reference and non-reference MEIs. a, Known L1Hs in GM12878 recovered by Cas9 targeted enrichment from the individual MinION flow cell (FAL11389), pooled-MEI MinION flow cell (FAO84736), and individual Flongle flow cell (ABB607), displayed in a way of proportion of the upper-bound known reference L1Hs, L1Pa, and other L1 as well as non-reference (non-ref.) subfamilies (L1Ta1, L1Ta0, L1Ta, L1PreTa, and L1Hs with ambiguous subfamilies) of L1Hs from PacBio-MEI set. b, The number of supporting reads in each captured L1 in the context of a. c, Known AluY elements in GM12878 recovered by Cas9 enrichment in two pooled MinION flow cells (FAO84736 and FAO08905), one individual AluYb Flongle flow cell (ACK645), and one individual AluYa Flongle flow cell (ACK655). d, The number of supporting reads in each captured Alu element in the context of c. e, Known SVA elements in GM12878 recovered by Cas9 enrichment in two pooled MinION flow cells (FAO84736 and FAO08905), one individual SVA_F Flongle flow cell (ACK629), and one individual SVA_E Flongle flow cell (ACK395). f, The number of supporting reads in each captured Alu element in the context of e. Error bars of boxplot range from Q1−1.5IQR to Q3+1.5IQR (IQR, interquartile range).

For AluY subfamilies, individual Flongles were utilized for separate runs of AluYb (ACK645) and AluYa (ACK655), and one pooled MinION flow cell (FAO84736) (Figure 3.8c,d, Figure 3.9, Table 3.2, and Supplementary Data 6*). 93.9% (3.5 mean coverage) reference and 88.0% (2.4 mean coverage) non-reference AluYbs were captured from the pooled MinION flow cell run, based on intermediate values. Similar to the L1Hs, the MinION flow cell was able to capture a vast majority of AluYb elements when considering intermediate values, indicating high sensitivity performance of the AluYb enrichment. A relatively lower rate of capture for reference (52.7%, 1.6 mean coverage) and non-reference (42.0%, 1.4 mean coverage) AluYa enrichment was observed in the pooled MinION flow cell run based on intermediate values. Cross-capture rate from the two individual Flongle flow cells was less than 0.1%, and close- and off-target reference elements were <0.1% and 2% for the Flongles and MinION, respectively, indicating a high specificity of the guide RNA for each AluY subfamily (Figure 3.8c, Table 3.2, and Supplementary Data 6*).

Similar enrichment rates were obtained from the two individual Flongle flow cells for SVA_F (ACK629) and SVA_E (ACK395), and in the pooled MinION flow cell (FAO84736) (Figure 3.8e,f, Figure 3.9, Table 3.2, and Supplementary Data 6*). 100% (4.6 mean coverage) reference and 96.6% (3.9 mean coverage) non-reference SVA_F were captured from the pooled MinION flow cell run based on intermediate values. A relatively lower rate of capture for reference (68.7%, 1.9 mean coverage) and non-reference (41.7%, 1.5 mean coverage) SVA_E enrichment was observed in the pooled MinION flow cell run based on intermediate values. The close-target reference SVAs capture rate was relatively high in two of the runs (35.6% in MinION flow cell and 11.4% in ACK629 Flongle flow cell). This could be due to SVAs sharing less base pair substitutions among their subfamilies compared to the other MEI families, as it is the youngest retrotransposon family found in the hominid lineage[128].

Our results indicate that an individual MinION flow cell (FAL11389) is able to completely (100%) capture reference and non-reference instances of a single MEI subfamily (L1Hs) compared to sequencing on the smaller Flongle flow cells. More importantly, a pooled run of an unbarcoded, five MEI subfamily enrichment experiment can recover the vast majority of known reference and non-reference MEIs (96.5% L1Hs, 93.3% AluYb, 51.4% AluYa, 99.6% SVA_F, and 64.5% SVA_E) in the genome when considering elements with a $\leq$ 3bp mismatch to the guide RNA. Such an approach outperformed individual Flongle flow cells and approached the same capture level as the single MEI subfamily MinION run (Figure 3.8, Figure 3.9, and Supplementary Data 3*, Table 3.2, Supplementary Data 6*). This suggests that the MinION flow cell has ample sequencing capacity to accommodate each experiment with negligible competition between samples, despite sequencing multiple MEI enrichments on one platform. Finally, a substantial fraction of reference and non-reference MEI events can be captured in a single MinION sequencing experiment with multiple supporting reads. With further optimization of the enrichment and sequencing methodologies, it is plausible to fully saturate reference and non-reference MEIs in a single experimental iteration.

3.3.4 Detectable Transmission of Non-reference L1Hs Within a Trio

To trace the transmission of non-reference L1Hs in GM12878 from the parents, another enrichment experiment of L1Hs elements was performed in GM12878, GM12891, and GM12892 (Figure 3.8g,h and Supplementary Data 3*,6*). The on-target rate for L1Hs ranged from 34.3% to 40.4% in the individual Flongles for GM12891 and GM12892 (parents), which mirrors the on-target rate of GM12878 (child) (Table 3.3).

Sample	Run	Flow cell	Read number	On-target		Close-target	Off-target
				Reference	Non-reference	Reference	
Individual							
GM12878	ABG188	Flongle	3,537	35.1%	5.8%	21.7%	37.5%
GM12891	ABO515		3,596	30.5%	3.8%	22.4%	43.4%
GM12892	ABN780		3,204	35.0%	5.4%	28.2%	31.4%
Pooled	FAL15177	MinION					
GM12878			74,171	30.4%	4.1%	16.9%	48.5%
GM12891			160,641	13.5%	1.9%	8.9%	75.8%
GM12892			101,294	37.8%	5.8%	21.2%	35.2%

Table 3.3: Enrichment of mobile element signals in nanopore reads from GM12878 trio L1Hs experiments. Four flow cells were carried out for trio experiments in the project: three individual Flongle flow cells for GM12878 (ABG188), GM12891 (ABO515), and GM12892 (ABN780) each, and one MinION flow cell for pooled three samples (FAL15177).

Transmission of non-reference L1Hs to GM12878 from the parental genomes was further examined using available GM12891 and GM12892 sequencing data. As the parental genomes lack sufficient long-read sequencing data, we utilized MELT to resolve non-reference L1Hs callsets from high-coverage Illumina short-read sequencing data. This analysis yielded 123 and 118 high confidence, non-reference L1Hs in GM12892 and GM12891, respectively. The number of MEIs identified in the MELT call sets are relatively lower than the number (n=205) detected in the 'PacBio-MEI' set for GM12878, consistent with previous observations that long-reads are more sensitive for MEI discovery[109, 113, 126] (Supplementary Data 4*). Additional evidence of transmission can be derived by the enrichment of reads from non-reference MEIs in GM12878. We expect MEIs that can be transmitted from either parent will have higher read coverage due to a portion of these being homozygous, and single parent transmitted MEIs will be heterozygous in GM12878. As predicted, an enrichment of approximately 1.52-fold (17.6 vs 11.6 mean coverage) was observed for these reads (Figure 3.8h and Supplementary Data 6*). Similarly, the 'not transmitted to child' non-reference L1Hs in parent samples should be heterozygous and were observed to be depleted by approximately 0.56 0.72-fold (17.4 vs. 31.3 mean coverage in GM12892 and 10.7 vs. 14.8 mean coverage in GM12891) of the nanopore reads that have been transmitted to the child (Figure 3.8h). These observations showed an expected supporting-read distribution of non-reference L1Hs, supporting the efficient nanopore Cas9 targeted enrichment in the pooled trio samples.

3.3.5 Cas9 Enrichment and Nanopore Sequencing Captures Non-reference Mobile Elements in Complex Genomic Regions

To estimate the efficacy of enrichment for non-reference MEIs with different sequencing coverage, we manually inspected each non-reference MEI reported by NanoPal and performed subsequent saturation analysis for all flow cells (Figure 3.10, see Methods). We find few additional L1Hs insertions by including additional on-target reads beyond approximately 30,000, using a cutoff of 15 supporting reads (Figure 3.10a). This is consistent with the observation that the MinION (individual or pooled, usually with >100k passed reads) has the ability to capture most non-reference L1Hs. In addition, there was no observable enrichment bias of MEI subfamilies from different flow cells (Figure 3.11).

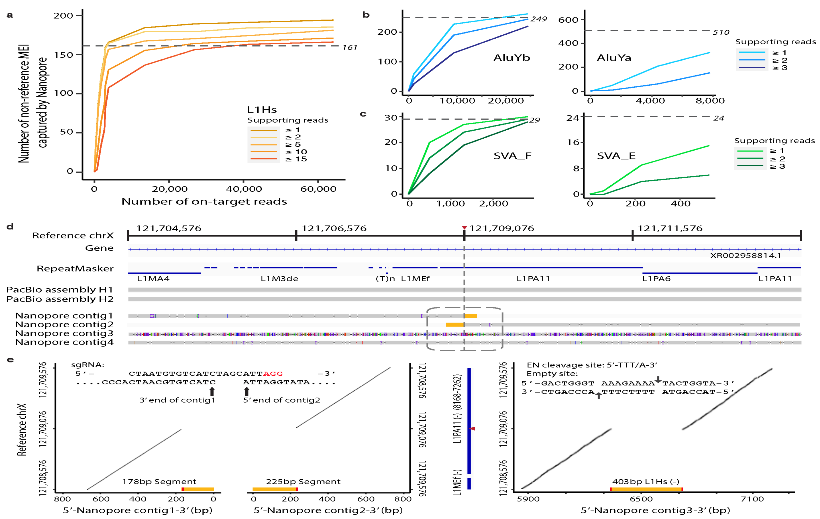

Figure 3.10: Non-reference MEIs captured by nanopore Cas9 enrichment approach. a, Number of non-reference L1Hs captured by nanopore Cas9 enrichment at different on-target read coverages for different supporting read cutoffs. The dotted-grey line with italic number represents the theoretical number of MEIs that the guide RNA binds when allowing a ≤3bp mismatch or gap in the PacBio-MEI set b, c, Number of non-reference AluYb, AluYa, SVA_F, and SVA_E, respectively, captured by nanopore Cas9 enrichment at different on-target read coverages. Axis labels and theoretic guide number as in a. d, An example of non-reference L1Hs specifically captured by nanopore sequencing at chrX:121,709,076. The tracks from top to bottom are as follows: reference coordinates with a red triangle represent the insertion site, gene track, RepeatMasker track (blue bars) with reference element annotation, PacBio contigs assembly for two haplotypes, four nanopore local-assembled contigs by CANU from different classifications of nanopore reads based on insertion signals (contig1, signal on 3' end; contig2, signal on 5' end; contig3, signal in the middle of the read; and contig4, no signal). e, Recurrence (dot) plots for nanopore contigs versus the reference region chrX:121,708,576−121,7089,576 sequence. Left panel shows the most 3' end of contig1 and the most 5' end of contig2 versus the reference sequence. Yellow bar represents the non-reference L1Hs sequence contained in the contig. The red bar represents one side of the target site duplication motif for the non-reference L1Hs contained in the contig. The upper part of this panel demonstrates sequences at the end of two contigs regarding the cleavage site when aligning to the guide RNA sequence. Blue bars in the middle panel represent the RepeatMasker track with reference L1 information annotated and the red triangle represents the insertion site in the reference L1 region. The right panel shows contig3 versus the reference sequence. Details of this non-reference L1Hs are detailed in the panel, including length, strand, empty site, and endonuclease (EN) cleavage site sequence.

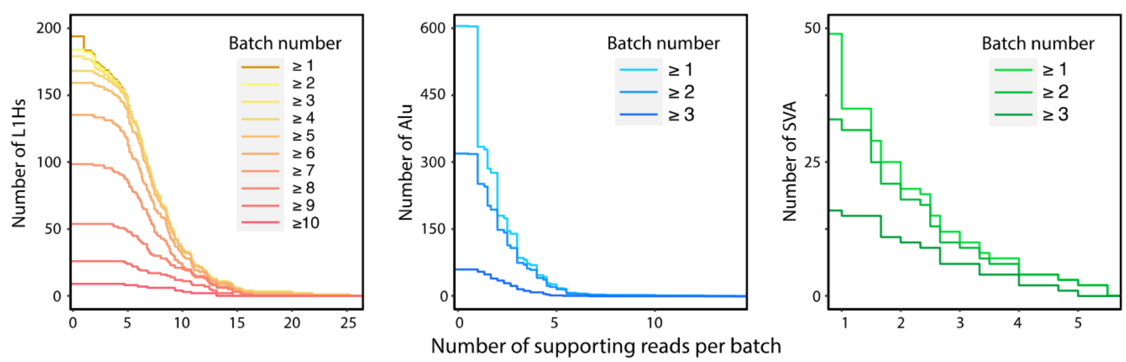

Figure 3.11: MEI distributions in various number of flow cells.The number of MEIs (L1Hs, yellow; AluY, blue; SVA, green) can be captured by nanopore Cas9 enrichment regarding different numbers of flow cells and cutoffs of supporting reads.

We examined the 182 non-reference L1Hs in GM12878 that overlapped with the PacBio-MEI set. Of these 175 (96.2%) could be accounted for by the parental (GM12891 and GM12892) sequencing data (Figure 3.12), and three overlapped known polymorphic insertions[126]. The remaining four non-reference L1Hs are located within centromeric regions, which could be missed in the parental samples due to lack of supporting reads. In addition, we observed 601 non-reference AluY (including 323 AluYb and 263 AluYa) and 49 non-reference SVA (including 30 SVA_F and 15 SVA_E) that overlapped with the PacBio-MEI set. We further examined the set of MEIs that were captured exclusively by Cas9 targeted enrichment and nanopore sequencing, but not found in the PacBio-MEI intersection (Supplementary Data 8*). We identified 12 additional L1Hs insertions as nanopore specific with ≥ 4 supporting reads that had been missed by the PacBio-MEI set with valid hallmarks, including target site duplication motifs, poly(A), EN Cleavage site, and empty site sequences, indicating a retrotransposition event induced by target-primed reverse transcription mechanism (TRPT) (Supplementary Data 9*, Figure 3.13). In addition, we detected 5 AluY elements that were specifically captured by nanopore reads in the GM12878 genome (Supplementary Data 9*). After refinement and inspection, we generated a full set of non-reference MEIs (194 L1Hs, 606 AluY, and 49 SVA) captured by Cas9 enrichment and nanopore sequencing in the GM12878 genome (Supplementary Data 10*). Of note, all intermediate-value calls of PacBio-MEI for L1Hs, AluYb, and SVA_F were recovered in this study (Table 3.2, Supplementary Data 9*and 10*). Additionally, 46 calls with L1Hs sequence and 14 with AluY sequence were captured by Cas9 target nanopore sequencing yet not included in the final non-reference callset (Supplementary Data 8*). Though lacking the support of TPRT hallmarks indicating a retrotransposition event, they may represent polymorphic duplicated sequences harboring an existing L1Hs or AluY element.

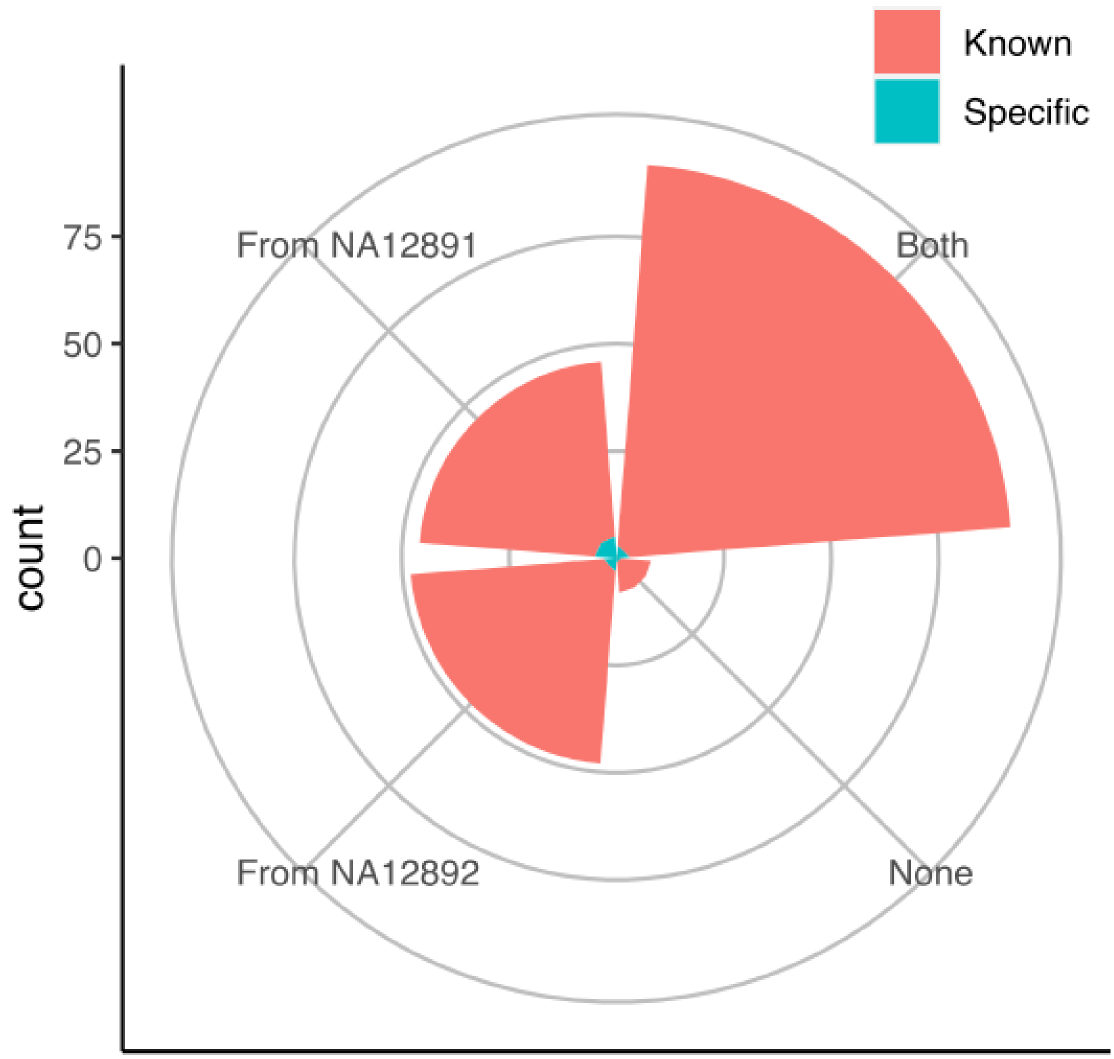

Figure 3.12: Trio transmission of 194 non-reference L1Hs captured by nanopore in GM12878 sample. The intersections with 'PacBio-MEI' were shown by red and the nanopore-specific non-reference L1Hs that were missed by 'PacBio-MEI' were shown by green.

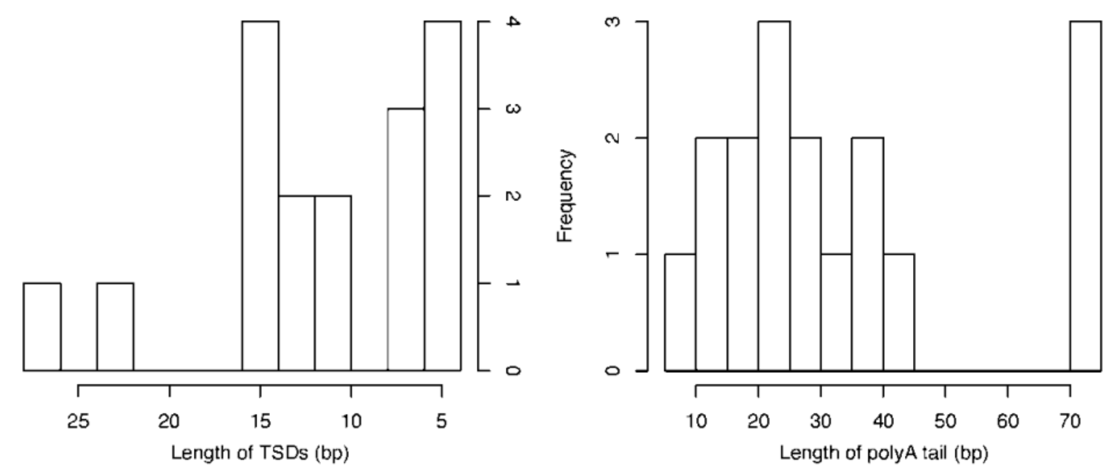

Figure 3.13: Seventeen nanopore-specific MEIs have Hallmarks of retrotransposition consistent with bona fide insertions. Left, length distribution of TSD motifs. Right, length distribution of poly(A) tails.

One non-reference L1Hs insertion at chrX:121,709,076 was particularly intriguing. The PacBio genome assembly-based approach overlooked this insertion, as it fell within a 'reference L1 rich' region (Figure 3.10d). Upon further inspection, this event was supported as a 403bp heterozygous L1Hs insertion by the existence of significant retrotransposition hallmarks, as well as recurrence (dot) plots[109, 149] (Figure 3.10e, see Methods). This insertion also shares a high sequence identity with a nearby reference L1PA11 element (Supplementary Data 8*). The decrease in efficacy of the PacBio assembly-based approach in this region could be explained by the intricate nested 'L1 in L1' structure and observed heterozygosity (Figure 3.10d). Likewise, the AluY nanopore Cas9 enrichments captured interesting non-reference AluY instances: A homozygous AluYb8 insertion at chr19:52384635, an exonic region within the ZNF880 gene, was reported to alter RNA expression due to the Alu element's effects on the RNA secondary structure[150]. Another heterozygous AluYa5 insertion at chr16:69157709 was located within a reference AluJr region, indicating a potential nested 'Alu in Alu' structure that could hinder non-reference AluY discovery. These observations demonstrate the high sensitivity of nanopore Cas9 enrichment, suggesting its feasibility for MEI discovery in complex genomic regions.

3.4 Discussion

Here we describe our design and implementation of Cas9 targeted nanopore sequencing to enrich for retrotransposition competent, repetitive mobile elements in the human genome[127]. After carefully designing guide RNAs to each MEI subfamily and coupling the enrichment with an established computational pipeline, our approach reaches an average of 44% nanopore sequencing reads with target MEI signals. We recovered a vast majority of reference and known non-reference mobile elements (96.5% L1Hs, 93.3% AluYb, 51.4% AluYa, 99.6% SVA_F, and 64.5% SVA_E) in the genome using only a single MinION flow cell. In addition, we discovered 21 non-reference MEIs within the GM12878 genome that were previously missed by other orthogonal long-read pipelines. Our data suggest that a MinION flow cell is ideal for a pooled, multiple-element enrichment experiment, as a prohibitively reduced enrichment or extensive cross-capturing of subfamilies was not observed. However, we observe that some of our MEIs targets (SVAs and Alus) have reduced enrichment compared to L1Hs. While we find relatively stable on-target rates for a specific guide target, differences in guide target rates can be expected due to different numbers of genomic integration sites, high numbers of similar target sequences, or guide RNA efficiency variation. For example, the low number of SVAs in the genome resulted in rather low on-target reads, but sufficient coverage to identify most of these elements in the genome. Similarly, the high number of Alus in the genome increases the set of near-matched guides and so we have a high background of other Alus enriched. The work presented here highlights the potential of targeted enrichment and nanopore sequencing to rapidly discover distinct MEIs, and cements an experimental foundation to probe even the most elusive mobile element insertion events.

Cas9 targeted enrichment paired with nanopore sequencing has the potential for resolving complex structural variation, previously obfuscated by sequencing and computational limitations. We leveraged the nanopore Cas9 targeted sequencing[127] method to target active retrotransposons in the human genome in a discovery-based approach. To our knowledge, this is the first application of this method for repetitive mobile element detection. Our experiments indicate that by utilizing guides targeted to specific subfamily sequences, both reference and non-reference insertions can be efficiently enriched and mapped with multiple supporting reads on even the smallest of nanopore sequencing flow cells. Moreover, we demonstrate that individual sequencing experiments readily capture a majority of reference and non-reference elements. In both pooled and single element experiments, MEIs of five subfamilies are robustly enriched, suggesting that this method is widely applicable across mobile elements, and most suitable for high copy genomic elements. In addition, nanopore sequencing offers the ability to detect DNA modifications like 5mC. Even though we obtained nanopore reads with two directions, of the reads belonging to captured full-length L1Hs, 65% extend across the consensus L1 sequence and beyond the L1Hs promoter regions in the 3'−5' direction, indicating the ability to investigate L1Hs promoter methylation in Cas9 targeted nanopore sequencing experiments. We further examine CpG methylation profiles of full-length reference and non-reference MEIs (Figure 3.14), showing consistent results with a prior study[125]. Guide RNA design process is straightforward, and targeting elements based on subfamily nucleotide differences captures both reference and non-reference elements, with negligible loss of sequencing to related close subfamilies or off-targeting. Unlike other Cas9 targeted enrichment experiments, on-target (reference and non-reference) rates for our method are comparatively higher, exceeding 50% in some cases. While this is likely a consequence of the number of genomic copies of the targeted element and, to a lesser extent, the fidelity of the guide sequence, it reiterates this method is particularly suitable for MEI discovery.

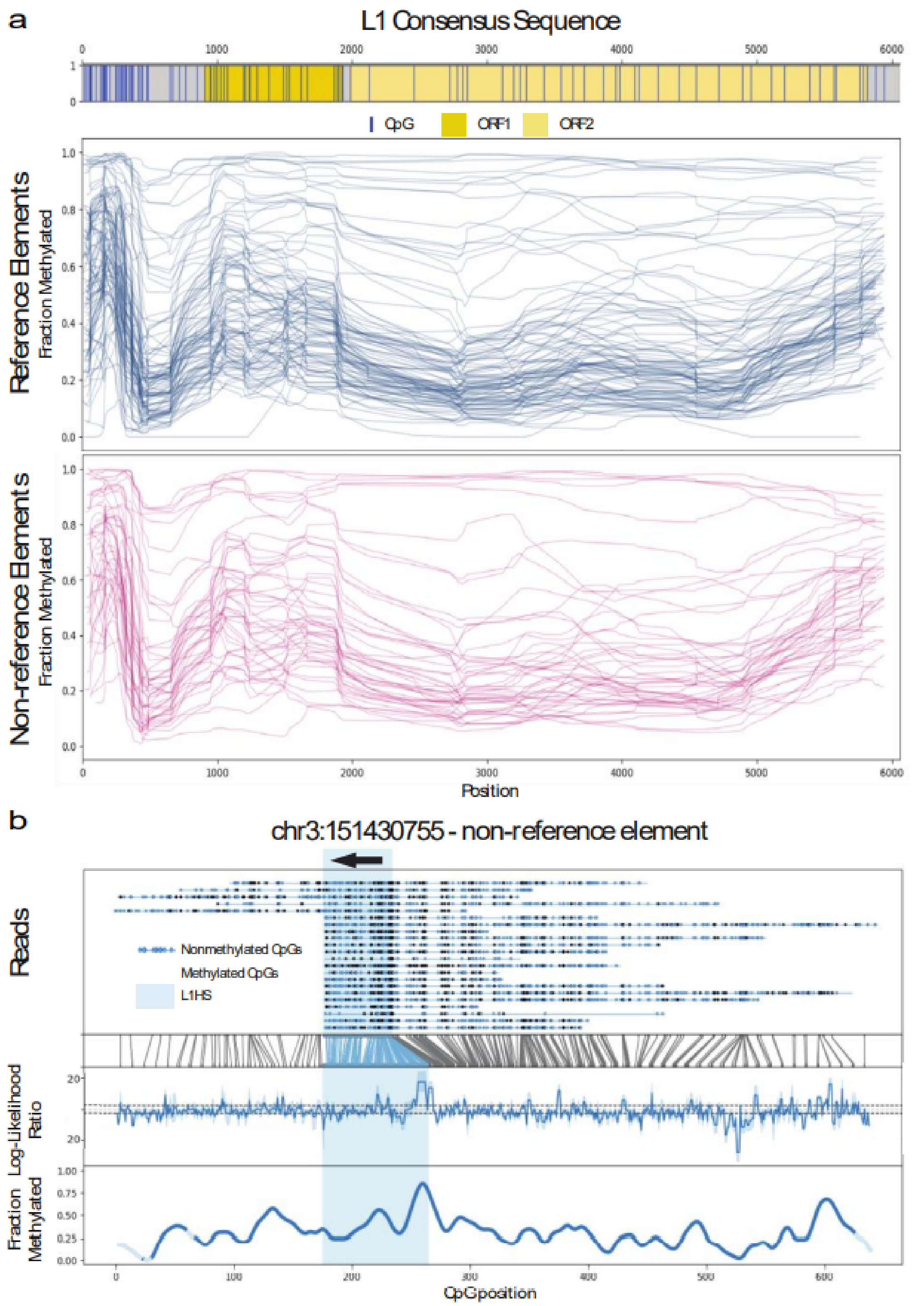

Figure 3.14: Examining CpG methylation of captured L1Hs reads. a, L1Hs methylation profile over consensus L1 sequence in reference (blue) and non-reference elements (magenta). b, An example of methylation profile at chr3:151430755 non-reference L1Hs (black arrow 3'->5' orientation). Individual reads with methylation profiles are shown (top) along with their aggregate methylation profile (bottom).

The preponderance of uncharacterized MEIs, taken together with their potential contribution to genomic variation and disease, emphasizes the critical need for efficient mobile element detection strategies. Our experiments using Cas9 enrichment and nanopore sequencing can quickly map active mobile elements in the human genome, as well as their larger genomic context. As we expected, in the nanopore-specific MEIs captured by our experiments, 15 out of 17 have reference Alu, L1, or LTR regions flanked by the insertion site. The assembly method or PacBio subread mapping (usually shorter than nanopore reads) could find difficulties in these reference repeat regions, where the Cas9 target method with nanopore sequencing could overcome the obstacles (with read length N50 ranged from 14.9Kbp to 32.3Kbp in the 17 flow cells of our experiments, Supplementary Data 3*). Although this is only a handful of overlooked insertions, it is a surprising result from such an extensively sequenced genome as GM12878. As we have shown previously, upwards of 50% of MEIs are missed in data generated from short read sequencing approaches[109]. Taken together with our results showing near complete saturation for MEIs in a single enrichment experiment, we anticipate that this may be highly effective at mapping patient samples, where a substantially larger proportion of MEI sites may differ from the GM12878 reference. In addition, larger nanopore sequencing platforms and multiplexed patient samples for pooled enrichments would streamline processing and maximize cost efficiency.

Since Cas9 enrichment has been used to target rare rearrangement events[151], implementation of this approach to detect unique, de novo, or somatic mobile element insertions is a feasible endeavor. This is emphasized by mounting evidence of MEIs escaping repression in embryogenesis, such as 5' truncated LINE-1s dodging repression by YY1 in neurons[152]. Nanopore-based identification of particularly rare insertion events, such as mosaic MEIs, is plausible. Recent work using whole genome sequencing of neuronal cells has demonstrated lineage tracing of retrotransposition events in early embryogenesis[153]. The application of targeted enrichment approaches to genetic mosaicism may improve the rate and depth at which this type of variation is sequenced. However, efficiently detecting genetic mosaicism through Cas9 enrichment may require larger sequencing platforms or methodological innovations, as the rarity of target sites in a sample increases.

While we surmise that the vast majority of MEIs can be captured using this approach, it is important to recognize that some genomic locations may persistently conceal recently transposed elements. Centromeric regions and long palindromic repeats are examples of complex genomic features that could be recalcitrant to MEI discovery[154]. With N50s of more than 25kb, we observed some MEI signals in centromeric and highly repetitive, palindromic regions from our nanopore sequencing reads. However, these regions still complicate mapping, requiring substantially longer sequencing and comprehensive analysis to confidently pinpoint elusive insertions[155]. Merging the enrichment experiments discussed here with improved commercial kits and extremely high molecular weight genomic DNA, may be critical for preserving the extremely long fragments necessary to map MEIs in complex genomic landscapes. The importance of mobile element activity in shaping the genomes they inhabit cannot be overstated. Even beyond the scope of the human genome, mobile element activity plays an intricate role in evolution across many organisms[156, 157]. Accelerated discovery of active mobile elements and other repetitive genetic elements will expand our understanding of their contributions to phenotypic diversity in genomes from every form of life.

3.5 Methods

3.5.1 Cell Culture, Counting, and Genomic DNA Isolation

The following cell lines/DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: GM12878, GM12891, GM12892. Each cell line was cultured at 37ºC, 5% CO2 in RPMI 1640 media (ThermoFisher, 11875093) with L-glutamine, and supplemented with 15% fetal bovine serum (ThermoFisher, 10437028) and 1x antimycotic-antibiotic (ThermoFisher, 15240112). Cells were regularly passed and the media replenished every 3 days.

High molecular weight genomic DNA was isolated from GM12878, GM12891, and GM12892 using a "salting out" method[158] with the following modifications. Lymphoblastoid cells were harvested from culture and centrifuged at 500xg for 10 minutes at 4ºC. Each pellet was washed in 4ºC 1X phosphate buffered saline (PBS) and cell number was counted using the Countess II. Washed cell pellets were resuspended in 3 mL of nuclei lysis buffer (10mM Tris-HCl pH 8.2, 400mM NaCl, and 2mM EDTA pH 8.2). 200µL of 10% SDS was added to the resuspension and inverted until evenly mixed. 50µL of RNase A (10mg/mL) was added and the lysate was rotated at 37ºC for 30 minutes followed by addition of 50µL of proteinase K (10mg/mL) and rotation at 37ºC overnight. 1 mL of saturated NaCl solution was added to the lysate and mixed by handheld shaking until evenly mixed. The sample was then centrifuged at 4000xg for 15 minutes at room temperature. The supernatant was transferred into a new 1.5mL microcentrifuge tube. Two volumes of 100% ethanol were added to the supernatant and the tube was inverted approximately 20 times, or until the precipitate coalesced. The precipitate was isolated via spooling with a sterile p10 pipette tip and resuspend in a sufficient volume of 1X TE buffer (250-500µL, depending on starting amount of cell material). Genomic DNA was passed through a 27G needle 3 times and stored at 4ºC. The DNA concentration was measured using a Qubit 3 Fluorimeter and the dsDNA Broad Range Assay kit (ThermoFisher, Q32850)

3.5.2 Design of Unique Guide RNAs for L1Hs, AluYb, AluYb, SVA_F, and SVA_E

To maximize the enrichment performance for each MEI subfamily, the guide RNA (gRNA) candidates were designed to bind to the unique sequences within each subfamily. A pairwise comparison was conducted for the target MEIs with other, non target subfamilies. The consensus sequences for each target MEI subfamily were obtained from Repbase[159], namely L1Hs in the L1 family, AluYa5 and AluYb8 in the Alu element family representing the AluYa and AluYb subfamily, and SVA_E and SVA_F in the SVA family. These subfamilies account for over 80% of currently active mobile elements in the human genome[128, 129, 130]. The consensus sequences of L1PA2, AluY, primate Alu, and SVA_D were retrieved from Repbase and included as outgroups in the comparison analysis. Furthermore, AluYa5 was added as an outgroup in the design of the gRNA for AluYb, AluYb8 for AluYa, SVA_E for SVA_F, and SVA_F for SVA_E to avoid enrichment across target MEIs. Guide RNA target sites (20bp sgRNA + 3bp NGG PAM site) for S. pyrogenes Cas9 were identified that are within unique MEI regions to obtain optimal guide candidates. Jellyfish2.0[160] was utilized to create a k-mer (k = 23) index for the sequences of these unique regions, and 23mers with a 5' 'CC' or 3' 'GG' were selected as gRNA candidates. The frequency of gRNA candidates and the three base substitution options in the 'NGG' PAM site for each candidate in the reference genome was calculated to confirm that the number of unique guide sequences is similar to the genomic reference MEI sequence frequency (Table 3.3* and Figure Supplementary Figure 1* and Supplementary Figure 2*). The guide RNA candidates of AluY and SVA with unique sequences into were categorized into different tiers: Tier0, sequence has subfamily-specific bases in GG/CC of the PAM site; Tier1, the frequency of 23mer falls into a reasonable range (<2-fold of target MEI frequency) in reference genome; Tier2, the sequence has only subfamily-specific bases at the N site of the PAM or the frequency of the 23mer falls out of a reasonable range. A gRNA sequence falling in Tier0 was considered an ideal candidate.

3.5.3 On-target Boundary Calculations for MEIs

Using the final gRNA selection as a reference, an upper-bound, lower-bound, and intermediate value of the theoretical numbers of target MEIs could be estimated. The lower-bound for target MEIs was defined as a MEI sequence that contains the sequence the gRNA binds to with 100% (or 23bp) matched sequence, the intermediate bound allows for $\leq$3bp mismatch or gap between the gRNA sequence and the matched MEI sequence, and the upper bound is a gRNA that aligns with more than 60% matched sequence (or $\geq$14bp) to the MEI.

3.5.4 In Vitro Transcription of Guide RNA and Cas9 Ribonucleoprotein Formation

Single stranded DNA oligos were designed using the EnGen sgRNA Designer tool (https://sgrna.neb.com/#!/sgrna, New England Biolabs) and purchased from IDT (Integrated DNA Technologies) to be used in the EnGen sgRNA Synthesis Kit (New England Biolabs, E3322S). Lyophilized oligos were resuspended in molecular biology grade water to a concentration of 100 uM, and 1:10 dilutions were made for working stocks. Each reaction was set up containing 10µL of EnGen 2X sgRNA Reaction Mix (S. pyrogenes), 0.5µL of 100mM DTT, 2.5µL of 10uM oligonucleotide, 2µL of EnGen sgRNA Enzyme Mix, and brought to 20µL total with PCR grade water. The reactions were incubated at 37ºC for 30 minutes to 1 hour. To degrade leftover

DNA oligonucleotides, the reaction volume was adjusted to 50µL using PCR grade water, and 2µL of DNase (New England Biolabs, E3322S) was added to the sample and incubated for 15 minutes at 37ºC. The sgRNA was purified by adding 200µL of Trizol and 50µL of chloroform to the sample, vortexed to mix and centrifuged at 20,000xg at room temperature. The aqueous layer was removed and placed into a new 1.5mL microcentrifuge tube and extracted again using 50µL of chloroform. The aqueous layer was removed and placed into a new 1.5mL microcentrifuge tube and ethanol precipitated in 2 volumes of 100% ethanol, and sodium acetate was added to a final concentration of 0.3M. The sample was centrifuged at max speed at 4ºC for 30 minutes. The RNA pellet was washed with 70% ethanol, air dried, and resuspended in 10µL of PCR grade water. RNA concentration was measured using the Qubit RNA BR Assay Kit (ThermoFisher, Q10211). Fresh guide RNA was transcribed for every experiment, and prepared no more than a day in advance. The Cas9 ribonucleoprotein (RNP) was formed by combining 850ng of in vitro transcribed guide RNA, 1µL of a 1:5 dilution of Alt-R S.p.Cas9 Nuclease V3 (Integrated DNA Technologies, 1081058), and 1X Cutsmart buffer (New England Biolabs, B7204S) in a total of 30µL. To allow for sufficient RNP formation, the reaction was incubated at room temperature for 20 minutes.

3.5.5 Cas9 Enrichment for L1Hs on a MinION Flow Cell

To perform a Cas9 sequencing enrichment for L1Hs, a modified Cas9 enrichment experiment was performed[127]. Three identical aliquots of 10ug of GM12878 genomic DNA were exhaustively dephosphorylated in a total volume of 40µL, with 1X Cutsmart buffer, 6µL of Quick CIP (New England Biolabs, M0525S), 10ug of gDNA, and H2O for 30 minutes at 37ºC. The Quick CIP was heat inactivated at 80ºC for 20 minutes. 20µL of RNP (Cas9 + gRNA) was added to the reaction along with 2µL of Taq polymerase (New England Biolabs, M0273L) and 1.5µL of 10mM dATP. The reaction was mixed by tapping and incubated at 37ºC for 30 minutes for Cas9 cleavage, and 72ºC for 10 minutes for monoadenylation. Following monoadenylation, each reaction was combined with 50µL of ligation mix: 25µL Ligation Buffer (LNB; Oxford Nanopore Technologies, SQK-LSK109), 5µL of Adapter Mix X (AMX; Oxford Nanopore Technologies, SQK-LSK109), 12.5µL of T4 DNA ligase (New England Biolabs, M0202M), and 5µL of nuclease-free water. The nanopore adapters were ligated to the genomic DNA at room temperature for 30 minutes on a tube rotator. Once completed, the ligations were diluted with 1 volume of 1X TE buffer (100µL). 60µL of SPRI beads (Beckman Coulter, B23317) were added to the adapterr ligated samples and incubated at room temperature for 10 minutes with rotation and for another 5 minutes without rotation. Beads were immobilized using a magnet and the supernatant was removed. Immobilized beads were resuspended with 200ul of room temperature L fragment buffer (LFB; Oxford Nanopore Technologies, SQK-LSK109. At this step, the resuspendend beads from the three samples were pooled into one Eppendorf tube. The magnet was applied again to immobilize the beads and remove the supernatant and the wash was repeated. Washed samples were pulse spun on a tabletop centrifuge for 1 second to collect beads at the bottom. Residual LFB was aspirated with a pipette. Beads were resuspended in 16.8µL of Elution Buffer (EB; Oxford Nanopore Technologies, SQK-LSK109) and incubated at room temperature for 10 minutes. Following the elution, the magnet was applied and the supernatant was collected and placed into a sterile Eppendorf tube. In some sample preparations, the adapter ligated library eluted in the last step may be viscous and the beads will resist immobilization on the magnet. A maximum speed centrifugation step prior to applying the magnet will help to immobilize the beads.

Once the supernatant was separated from the beads into a sterile Eppendorf tube, 26µL of Sequencing Buffer (SQB; Oxford Nanopore Technologies, SQK-LSK109) was added and placed on ice until the sequencing flow cell was prepared. Immediately prior to loading of the sample, 0.5µL of Sequencing Tether (SQT, Oxford Nanopore Technologies, SQK-LSK109) was added along with 9.5 µL of Loading Beads (LB; Oxford Nanopore Technologies, SQK-LSK109). The sample was mixed evenly by pipetting with a p20 and loaded onto the sequencing platform.

3.5.6 Pooled Cas9 Enrichment for L1Hs, AluYb, AluYa, SVA_F, and SVA_E in GM12878 (MinION)

Five parallel Cas9 enrichment experiments were performed for the five MEI subfamilies in GM12878 for a pooled sequencing run. Five separate aliquots of 10ug of genomic DNA were dephosphorylated in 40µL total (30µL of gDNA, 4µL of 10X CutSmart, 6µL of Quick CIP) for 25 minutes at 37ºC then heat inactivated at 80ºC for 5 minutes. Following the heat inactivation, each dephosphorlyated genomic DNA sample was combined with 20µL of Cas9 RNP, 1µL of Taq polymerase, and 1µL of 10mM dATP. After briefly mixing by tapping, the reaction was incubated at 37ºC for 30 minutes to enable Cas9 cleavage, then incubated to 75ºC for monoadenylation by Taq polymerase. The Cas9 digested and monoadenylated samples were pooled into the ligation reaction (164µL of Custom LNB, 10µL of AMX , 20µL of T4 DNA ligase, and 164µL of nuclease-free water) and rotated at room temperature for 30 minutes. One volume of 1X TE buffer was added to the ligation and mixed by inversion approximately 10 times, or until evenly mixed. 0.3X sample volume of SPRI beads (394.8µL) was added and incubated at room temperature with rotation for 5 minutes. The beads were immobilized using a magnet and washed twice with 100µL of room temperature LFB. After the final wash, the beads were pulse spun for 1 second in a table top centrifuge, immobilized on a magnet, and residual LFB was removed. The washed beads were eluted in 13µL of EB for 10 minutes at room temperature and removed using a magnet. The supernatant was collected and combined with 26µL of SQB. The library was incubated on ice until the flow cell was prepared. 0.5µL of SQT and 9.5µL of LB were added to the library before the loading onto the flow cell.

3.5.7 Cas9 Enrichment for Single MEI Subfamily on a Flongle Flow Cell

10ug of purified genomic DNA was dephosphorylated using 4µL of Quick CIP in 1X Cutsmart buffer and brought to a total reaction volume of 40µL, then incubated for 30 minutes at 37ºC. The sample was then incubated at 80ºC for 5 minutes to inactivate the Quick CIP. 1µL of Taq polymerase, 1µL of 10mM dATP, and 20µL of the corresponding Cas9 RNP (targeting L1Hs, AluYb, AluYb, SVA_F, or SVA_E) , was added to the dephosphorlyated genomic DNA, gently mixed, and incubated at 37ºC for 30 minutes, followed by a 10 minute incubation at 75ºC. The sample was added to the ligation solution (25µL of custom LNB, 6µL of T4 DNA ligase, 5µL of AMX, and nuclease-free water to 100µL total), and incubated at room temperature for 20 minutes with rotation. The ligation was mixed with 1 volume (100µL) of 1X TE buffer and mixed by inversion approximately 10 times, or until evenly mixed. SPRI beads were added to a final 0.3X (60µL) to the sample volume (200µL) and the sample was rotated at room temperature for 5 minutes. The SPRI beads were immobilized on a magnet and washed twice with 100µL of room temperature LFB. After the final wash, the beads were pulse spun for 1 second on tabletop centrifuge and residual LFB was removed. The beads were resuspended in 9µL of EB and incubated at room temperature for 10 minutes. After the elution, the beads were immobilized on a magnet and the supernatant was transferred to a new 1.5mL microcentrifuge tube.

13µL of SQB was added to the supernatant and this library was placed on ice until the flow cell was prepared. Before loading the sample onto the flow cell, 0.5µL of SQT and 9.5µL of LB were added and mixed by gentle tapping.

3.5.8 Cas9 Enrichment for L1Hs in Trio (MinION)

To detect L1Hs in the lymphoblastoid trio cells (GM12878/91/92), a modified Cas9 enrichment assay, originally described by Gilpatrick et al. 2020, was performed. 10ug of genomic DNA for each genome (30ug total) was exhaustively dephosphorylated using 6µL Quick CIP for 45 minutes at 37ºC, and heat inactivated at 80ºC for 20 minutes. 20µL of the RNP (Cas9 and sgRNA) was added to the dephosphorylated genomic DNA along with 2µL of Taq DNA polymerase and 1.5µL of 10mM dATP. The reaction was incubated at 37ºC for 30 minutes, then 72ºC for 10 minutes. Each genomic DNA reaction was combined with an equal volume (50µL) of ligation mix for nanopore adapter ligation: 25µL Ligation Buffer (Oxford Nanopore Technologies, EXP-NBD104), 5µL of AMII (Adapter Mix II; Oxford Nanopore Technologies, EXP-NBD104), 12.5µL of T4 DNA ligase, 2.5µL of the barcode (NB01/02/03; Oxford Nanopore Technologies, EXP-NBD104), and 2.5µL of nuclease-free water, and incubated at room temperature for 30 minutes on a tube rotator. After adapter ligation, an equal volume of 1X TE buffer was added to the reaction, and SPRI beads were added to a final 0.3X (60µL). The library was incubated at RT for 10 minutes with rotation, and 10 minutes without rotation for a total of 20 minutes to allow for DNA binding to the SPRI beads. The beads were washed twice with 200µL of L Fragment Buffer (LFB; Oxford Nanopore Technologies, EXP-NBD104). The uniquely barcoded samples were pooled by combining the resuspended beads in the first wash, then washed again. The washed beads were resuspended in 16.8µL of the Elution Buffer (EB; Oxford Nanopore Technologies, EXP-NBD104). Resuspended beads were incubated at room temperature for 10 minutes. Following the incubation, the beads were collected with a magnet and the supernatant collected into a separate 1.5mL microcentrifuge tube and placed on ice. The sample was prepared for sequencing by adding 26µL of Sequencing Buffer (SQB; Oxford Nanopore Technologies, EXP-NBD104) and kept on ice while the flow cell was primed. 0.5µL of Sequencing Tether (SQT; Oxford Nanopore Technologies, EXP-NBD104) and 9.5µL of Loading Beads (LB; Oxford Nanopore Technologies, EXP-NBD104) were added after flow cell priming, before the sample was loaded onto the flow cell.

3.5.9 Nanopore Flow Cell Preparation, Sequencing, Base-calling, and Cleavage-site Analysis

A MinION flow cell was purchased from Oxford Nanopore Technologies and stored at 4ºC per manufacturer's instructions. The Ligation Sequencing Kit (SQK-LSK109) and Native Barcoding Kit (EXP-NBD104) were used to prepare the pooled libraries. Upon arrival and prior to usage, MinION flow cell QC was performed using the MinKNOW software. No appreciable loss of active pores was noted during storage. Prior to loading the library, the MinION was flushed with 800µL of FLB in the priming port, followed by priming of the flow cell with 200µL of 0.5x SQB diluted with water. Flongle flow cells were purchased from Oxford Nanopore Technologies in batches and stored at 4ºC. Flow cells were QC'd upon arrival and the number of active pores was noted. Flongles to be used in sequencing were QC'd immediately before use to assess pore loss during the storage period. Base-calling was processed by Guppy 4.0.15 (Oxford Nanopore Technologies) using the high accuracy, modified base model (dna_r9.4.1_450bps_modbases_dam-dcm-cpg_hac.cfg). Porechop[161] was used to trim nanopore adapters and barcodes from the reads with Q-score > 7, as well as demultiplex the reads in the pooled sample MinION run. To determine cut site preferences, reads were aligned to the consensus sequence of each mobile element class that were investigated; L1Hs, AluYa5, AluYb8, SVA_E and SVA_F. Only the first 80 base pairs of each read were used for alignment to focus on the cut site region and to ensure that the beginning of the reads aligned correctly. Pairwise alignments were performed using the Biopython Bio.Align package[162] with FASTA files as input. Each read was aligned to the mobile element consensus sequence as well as the reverse complement to determine sequence orientation. To obtain high-confidence alignments, strict gap penalties were enforced (open gap:-10, extend gap:-5). In order for an alignment to be considered for cut-site analysis, it had to meet two criteria; the alignment had to start at the very first base of the read, and needed to have an alignment score of at least 100. The 5' ends of reads meeting this criteria were then used to estimate cleavage site location.

3.5.10 Nano-Pal for Detection and Refinement of MEIs from Nanopore Cas9 Enrichment

To resolve both the reference and non-reference MEI signals from nanopore Cas9 enrichment, we developed a computational pipeline, Nano-Pal, to analyze the nanopore reads and customized it for different MEI subfamilies (Fig.1c,d). Information of the potential targeting MEI signals was obtained by Nano-Pal scanning through both sides (100bp bin size) of all quality-passed nanopore raw reads using BLASTn[163, 164]. Next, it aligned the reads to the reference genome (GRCh38) using minimap2[165] and discarded reads with low mapping quality (MAPQ<10). The aligned reads were screened by RepeatMasker[147] and the pre-masking module in PALMER[109] to bin them into different categories: reads with reference MEI signals, reads with non-reference MEI signals, and off-target reads. All reads that were reported by the PALMER pre-masking module fell into the on-target non-reference MEI category. If reads were not reported by PALMER, but were annotated by RepeatMasker, they fell into the reference MEI category. They will be further classified as on-target, close-target, and off-target reads depending on where they mapped to the reference regions. For L1Hs experiments, the reads mapped to the reference L1PA are considered as close-target and the ones mapped to other reference L1 are considered as off-target. For AluY experiments, the close-target reads are those that mapped to other reference AluY besides AluYb and AluYa and the off-target reads are defined when mapped to other reference Alu elements besides the reference AluY. For SVA experiments, the close-target reads are the ones mapped to other reference SVAs besides SVA_F and SVA_E and no off-target reads were defined when they have the MEI signals. Any remaining reads with no MEI signals were classified as off-target reads as well. The reads in the first and second category were then clustered into non-reference MEIs and reference MEIs, respectively. Nano-Pal was performed for each Flongle and MinION flow cell separately for GM12878, GM12891, and GM12892.

3.5.11 GM12878 Trio Data, Reference Genome, and Reference MEI Information

We obtained Pacific Bioscience (PacBio) long-read CLR sequencing data from Audano et al. 2019[140] for the GM12878 genome (50x coverage). The 30x Illumina NovaSeq sequencing data for GM12878 and the related samples (GM12891 and GM12892) were obtained from the 1000 Genomes project phase 3 sample set, which were generated at the New York Genome Center (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/`)[137, 166]. All analyses in this project were carried out using the GRCh38 (GRCh38+decoy) reference genome obtained from the 1000 Genomes Project (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/`). Information of reference MEIs, including the five target subfamilies, were obtained from

RepeatMasker[147].

3.5.12 Enhanced PALMER for Resolving Non-Reference MEIs from Whole-genome Long-read Sequencing

We developed an enhanced version of PALMER (Pre-mAsking Long reads for Mobile Element inseRtion)[109] in this study to detect non-reference MEIs across the long-read sequenced genomes (https://github.com/mills-lab/PALMER). Reference-aligned BAM files from long-read technology were used as input. Known reference repetitive sequences (L1s, Alus or SVAs) were used to pre-mask portions of individual reads that aligned to these repeats and also utilized in the Nano-Pal pipeline. After the pre-masking process, PALMER searched subreads against a library of consensus mobile element sequences within the remaining unmasked sequences and identified reads with a putative insertion signal (including 5' inverted L1 sequences, if available) as supporting read candidates. PALMER opens bins 5' upstream and 3' downstream of the putative insertion sequence for each read and identifies hallmarks of mobile elements, such as target site duplication (TSD) motifs, transductions, and poly(A) tract sequences. All supporting reads are clustered at each locus and those with a minimum number of supporting events are reported as putative insertions. To improve the accuracy of non-reference MEI sequences derived from individual subreads, which tend to have lower per-read base-pair accuracy, local sequence alignments and error correction strategies were performed. Error correction was conducted by CANU[167] (ver2.2) using default parameters on the subreads with MEI signals reported by PALMER, allowing the generation of error-corrected reads that served as inputs for local realignment using minimap2. A second-pass of the PALMER pipeline then was executed using these locally aligned error-corrected reads to generate a high-confidence call set of germline non-reference MEIs. CAP3[168] was used with default parameters to assemble all MEI sequences reported by the second-pass of the PALMER pipeline to generate a high-confidence consensus contig for each non-reference MEI event.

3.5.13 MEI Callsets in Orthogonal Short-read and Long-read Data

As there is no public long-read data available for GM12891 and GM12892, we used the Mobile Element Locator Tool (MELT) to identify non-reference MEIs in the short-read Illumina sequencing data for the GM12878 trio[107, 148] as a benchmark set in the trio analysis. We applied the enhanced version of PALMER and carried out the non-reference MEI calling in GM12878. To generate a more comprehensive callset of non-reference MEIs in GM12878, the Phased Assembly Variant (PAV) caller was included (https://github.com/EichlerLab/pav), which can discover genetic variants based on a direct comparison between two sequence-assembled haplotypes and the human reference genome[126]. The callset by PAV for GM12878 was generated from the PacBio HIFI sequencing data after haplotype-assembly. A 'PacBio-MEI' callset in GM12878 was generated by applying the union set of the mapping-based PALMER callset and the assembly-based PAV callset, both of which resolved the MEIs from PacBio long-read sequencing data. The details of MEI merging and subfamily defining strategy for the two approaches are described in a prior study[126].

3.5.14 Inspection and Validation of Nanopore-specific Non-reference MEIs

All non-reference MEI calls were further intersected with the PacBio-MEI set and classified as known non-reference calls and potential nanopore-specific non-reference calls. A filtering module, an empirical curation of read-depth (<2-fold difference) within the 500bp bin of the insertion site from public data[140], and manual inspection were applied to exclude false-positive (FP) signals in the nanopore-specific non-reference MEIs. All potential nanopore-specific non-reference MEIs were classified into three categories: a) true positive (TP) non-reference event, b) FP non-reference event, and c) an ambiguous event (Supplementary Data 8*). The category was further defined as TP missed by the PacBio sequencing data, TP missed by the PacBio mapped-based and assembly-based pipelines but with PacBio read signals, or TP redundant with the called non-reference one. Category (b) was broken down into three subcategories; FP redundant with the called reference event, FP targeting on the other off-target reference repeat, or ambiguous. All subcategories in (a), and the first (b) subcategory, were on-target reads with or without correct annotations. For example, one FP nanopore-specific non-reference call originated from reads targeted to reference MEIs, yet was categorized as non-reference due to a mapping error introduced by flanking structural variations (deletions, duplications, or inversions) (Supplementary Data 8*). A further in-depth inspection was employed for the non-reference calls categories as nanopore-specific. Each call has been inspected by two sections: a) general information including TRPT hallmarks (TSD motifs, poly(A), EN Cleavage site, and empty site sequence), length, strand, genotype, and population frequency in 32 genomes reported by Ebert et al. [126], and b) IGV screenshot in a range of genomic region with genomic content annotation. The TRPT hallmarks were identified by Cas9 target nanopore reads or from PacBio assembled contigs (Supplementary Data 9*). A recurrence plot analysis was employed for further in-depth validation as well. For the recurrence plot analysis, a region of one sequence (X-axis) is compared to another sequence (Y-axis) and small (i.e. 10 bp) segments that are identical between the two sequences are denoted with a plotted point. Thus, a continuous diagonal line comprising multiple points indicates portions of the compared sequences that are identical. By comparison, gaps and shifts from the diagonal denote an insertion or deletion in one sequence relative to the other.

3.5.15 Non-reference MEIs Captured by Nanopore Cas9 Enrichment Sequencing in GM12878

The flow cell runs for each MEI subfamily were merged to investigate the read coverage enrichment performance and generate the final non-reference MEIs in GM12878. For the saturation analysis, the flow cells were ranked by the number of on-target reads. Reads from the flow cells were then added and merged, one flow cell at a time, based on the above ranking. Non-reference calls for L1Hs, AluYb, AluYa, SVA_F, and SVA_E were resolved by Nano-Pal and our validation process after every merging instance. By merging all batches for each subfamily, a final non-reference MEI callset in GM12878 captured by nanopore Cas9 enrichment approach was produced.

3.5.16 Analysis of L1Hs CpG Methylation

Nanopolish[169] was used to call methylation on pooled GM12878 MinION and Flongle runs. Reference L1Hs methylation profiles were generated using methylartist (https://github.com/adamewing/methylartist), where methylation was aggregated across L1HS intervals from RepeatMasker[147]. To generate profiles for non-reference elements, we built contigs using reads supporting full-length L1Hs and ±100kb of flanking sequences. Reads within 500bp of insertion sites were then extracted from the merged data and aligned to the constructed contigs using minimap2[165]. These alignments and reads were then used for methylation calling with nanopolish. The data from nanopolish was aggregated across L1Hs sequence in the constructed contigs and used for locus and consensus plotting using methylartist.

3.5.17 Data Availability

The nanopore sequencing data for the Cas9 targeted enrichment of MEIs in this study are available in the SRA repository under BioProject accession PRJNA699027 (https://www.ncbi.nlm.nih.gov/sra/PRJNA699027).

3.5.18 Code Availability

All scripts and pipelines in this publication, including Nano-Pal, are available on GitHub: https://github.com/Boyle-Lab/NanoPal-and-Cas9-targeted-enrichement-pipelines [170]. The enhanced version of PALMER is available at https://github.com/mills-lab/PALMER [171].

3.6 Notes and Acknowledgements

We thank Dr. Scott Devine, Dr. Qihui Zhu, and Dr. Charles Lee for providing the MELT callset for GM12892 and GM12891. We thank Jixin Guan for their help with improving the performance of the PALMER software. This research was supported by the National Institutes for Health (NIH) under award no. R21HG011493 to A.P.B. and R.E.M.. T.M. was supported by T32GM007544. C.C. was supported by the University of Michigan Rackham Merit Fellowship and the Training Program in Bioinformatics (T32GM070449). Figures and Data labeled with asterisks (*) in main text are of excessive length or not conducive to formatting. Please visit our publication below for detailed supplementary information.

I implemented nanopore sequencing in the lab, optimized loading protocols, reconstituted proprietary reagents, streamlined guideRNA preparation, established and improved Cas9 enrichment, performed all Cas9 enrichment experiments for data generation, and wrote this manuscript.

Alan Boyle, Ryan Mills, Weichen Zhou, and I conceived of this project. Thank you

to Alan Boyle and Ryna Mills for critical thinking, data interpretation, and project design. Thank you to Jessica Switzenberg for cell culture, gDNA isolation, and assistance in performing experiments. Thank you to Arthur (Weichen) Zhou for his indispensable work on mobile element enrichment analysis, Nano-pal pipeline design and implementation, interpretation of results, and extensive figure and table design. Thank you to Camille Mumm for computational efforts, including CpG methylation analysis and the automated raw data processing pipeline. Thank you to Christopher Castro for guide RNA cleavage site analysis and computational support.

This work was published in Nature Communications McDonald, T.L., Zhou, W., Castro, C.P. et al. Cas9 targeted enrichment of mobile elements using nanopore sequencing. Nat Commun 12, 3586 (2021). https://doi.org/10.1038/s41467-021-23918-y

3.7 Author Contributions

A.P.B., R.E.M., W.Z., and T.M. conceived the project. J.S. established and cultured cell lines and isolated gDNA. T.M. performed the Cas9 targeted enrichment and nanopore sequencing. W.Z. developed the NanoPal pipeline and PALMER software. W.Z., C.C., and C.M. performed computational analysis. All authors guided the data analysis strategy. A.P.B., R.E.M., W.Z., and T.M. wrote the manuscript. All authors edited the manuscript. All authors read and approved the final manuscript.

CHAPTER IV

Enrichment and Sequencing of Polynucleotide Repeat Expansions

4.1 Abstract

Polynucleotide repeat expansions are causative in upwards of 40 human genetic diseases. Identifying and characterizing repeat expansions and the diseases has been challenging and the field currently lacks efficient methods to probe direct repeat lengths. Here we explore using nanopore sequencing paired with different enrichment approaches to attempt to capture polynucleotide repeats. We explore two orthogonal approaches to enrich for polynucleotide repeat regions. We demonstrate that each method captures targeted repeat regions, and that as many as half of the targeted repeats are captured with nanopore sequencing. Overall, this work introduces proof of principle for characterizing disease-associated polynucleotide repeats.

4.2 Introduction

Polynucleotide repeats are a normal type of structural variation in the human genome. However, in some cases repeats can experience an abnormal expansion and manifest disease. In fact, due to their unstable nature, it is not uncommon for repeats to change in size in transmission across generations. As many as 50 human genetic diseases have been documented to be caused by expansions in polynucleotide repeats in the genome [66]. While heterogeneous in respect to the genomic context (coding vs non coding), nucleotide content (CGG, CAG, etc), and repeat pattern (tri-, tetra-, hexa-,etc), repeat expansions tend to inversely correlate with age of onset of disease, as well as correlate to the severity of the disease. This phenomenon is also referred to as anticipation [60].

The precise mechanism and timing of the expansion of repeats is unclear. Instead, evidence supports a few different pathways for expansion of polynucleotide repeats. For instance, in almost exclusively maternally transmitted non coding trinucleotide repeat expansions such as in FXS and myotonic dystrophy type 1 (DM1), evidence supports a replication independent mechanism in quiescent oocytes. This suggests that DNA repair mechanisms maintaining the integrity of arrested oocytes are involved in expansion events. Indeed, base excision repair pathways on oxidized DNA bases appears to play a possible role in expanding trinucleotide repeats. It has been observed in FXS mouse models that the strongly oxidative chemical potassium bromide induces CGG expansions in oocytes [172]. However, not all glycosylases (proteins involved in base excision repair) that target 8-oxoG (oxidized guanine) are implicated in the expansion of trinucleotide repeats [173]. Another mechanistic source for repeat expansion may be replication restart. When polymerase encounters a trinucleotide repeat of CAA, CAG, or CGG, the polymerase stalls and can utilize the opposite strand to finish synthesis through the repeat [173]. Structures consistent with this model have been directly observed in electron micrographs, and DNA polymerase inhibitors can prevent expansion, suggesting that these expansions may occur through repair independent mechanisms as well [173, 174, 175]

Since the discovery of the first repeat expansions 1991, at least 50 repeat expansion associated disorders have been characterized. However, their discovery has been challenging and irregular, with 17 of them only characterized in the past 3 years [66]. Due to the genetic structure of repeat expansions, and the historic molecular and bioinformatic limitations, regular characterization of disease associated repeat expansions as well as discovering novel expansions has been limited [68]. Here, we explore using a combination of nuclease enrichment strategies and nanopore sequencing to specifically target and sequence 48 repeat expansion disease associated loci. We first explore using a biotinylated dCas9 to directly bind and pulldown repetitive DNA fragments (Figure 4.1). Second, we pilot paralleled, single target Cas9 enrichment and sequencing of known disease associated repeat loci. Previous work has established using Cas9 targeted nanopore sequencing for single locus events [176]. However, their gRNA preparation and pooling strategy limits scaling and highly parallel targeting. We leverage a pooled in vitro transcription of 96 guide RNAs that directs Cas9 upstream and downstream of each of the 48 loci (Table 4.1). Our data indicates that it is possible to capture large swaths of repeat expansions with targeted long read nanopore sequencing. In addition, we show that this targeting strategy works with pooled of in vitro transcribed gRNAs, indicating that this method will efficiently scale as more disease-associated repeat expansion loci are discovered.

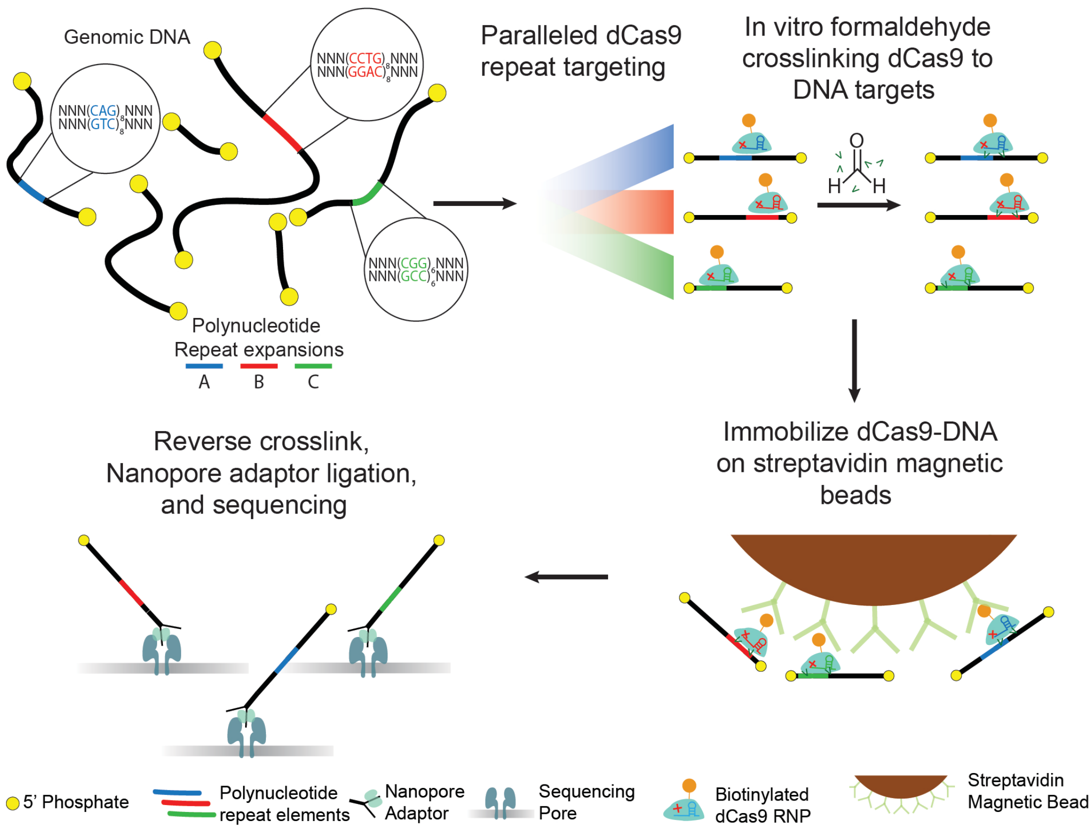

Figure 4.1: Molecular workflow for dCas9-mediated capture of polynucleotide repeat regions. First, purified genomic DNA containing polynucleotide repeat sequence (top left) is incubated with biotinylated dCas9 loaded with repeat targeting guide RNAs. After incubation with dCas9 RNP, protein-DNA interactions are fixed with low concentrations of formaldehyde (top right). Fixed protein-DNA complexes are immobilized on streptavidin magnetic beads and washed thoroughly (bottom right). Immobilized complexes are reverse crosslinked and DNA is purified and prepared for nanopore sequencing (bottom left.

Disease	Coordinates	Abbreviation
Autism spectrum disorder, associated with fragile site FRA7A	chr7:55887601-55887934	FRA7A
Baratela-Scott syndrome	chr16:17470890-17470930	BSS
Blepharophimosis, ptosis and epicanthus inversus syndrome	chr3:138946021-138946062	BPES
C9ORF72 amyotrophic lateral sclerosis frontotemporal dementia	chr9:27573485-27573546	C9 ALS/FTD
Cerebellar ataxia, neuropathy, vestibular areflexia syndrome	chr4:39348425-39348479	CANVAS
Cleidocranial dysplasia	chr6:45422751-45422801	CCD
Congenital central hypoventilation syndrome	chr4:41745972-41746031	CCHS
Dentatorubral-pallidoluysian atrophy	chr12:6936717-6936775	DRPLA
Familial adult myoclonic epilepsy 3	chr5:10398839-10398849	FAME3
Familial adult myoclonic epilepsy 6	chr16:24646096-24646106	FAME6
Familial adult myoclonic epilepsy 7	chr4:57146650-57146660	FAME7
Fragile X syndrome	chrX:147912037-147912111	FXPOI
Fragile XE syndrome	chrX:148500602-148500743	FRAXE
Fridreich's ataxia	chr9:69037287-69037305	FA
Fuchs endothelial corneal dystrophy	chr18:55586154-55586229	FECD
Glutaminase deficiency	chr2:190880874-190880919	GAD
Hand-foot-genital syndrome	chr7:27199925-27199966	HFGS
Holoprosencephaly 5	chr13:99985449-99985493	HPE
Huntington's disease	chr4:3074877-3074940	HD
Huntington's disease-like 2	chr16:87604283-87604329	HDL2
Intelleclual disability associated with fragile site FRA12A	chr12:50505002-50505053	FRA12A
Intellectual disability associated with fragile site FRA2A	chr2:100104620-100104860	FRA2A
Jacobsen syndrome	chr11:119206290-119206323	FRA11B
Myotonic dystrophy type 2	chr3:129172577-129172659	DM2
Myotonic dystrophy type I	chr19:45770205-45770264	DM1
Neuronal intranuclear inclusion disease	chr1:149390802-149390842	NIID
Oculopharyngeal muscular dystrophy	chr14:23321464-23321543	OPMD
Oculopharyngeal myopathy leukoencephalopathy	chr10:79826315-79826404	OPML
Oculopharyngodistal myopathy 2	chr19:14496042-14496085	OPDM2
Oculopharyngodistal myopathy I	chr8:104489495-104489528	OPDM1
Pseudoachondmplasia and multiple epiphyseal dysplasia	chr19:18786010-18786050	PSACH/MED
Spinal and bubar muscular atrophy	chrX:67545317-67545419	SBMA
Spinocerebellar ataxia typa 12	chr5:146878729-146878758	SCA12
Spinocerebellar ataxia type 1	chr6:16327634-16327724	SCA1
Spinocerebellar ataxia type 10	chr22:45795355-45795425	SCA10
Spinocerebellar ataxia type 17	chr6:170561907-170562017	SCA17
Spinocerebellar ataxia type 2	chr12:111598951-111599019	SCA2
Spinocerebellar ataxia type 3	chr14:92071011-92071052	SCA3
Spinocerebellar ataxia type 31	chr16:66521961-66521971	SCA31
Spinocerebellar ataxia type 36	chr20:2652733-2652775	SCA36
Spinocerebellar ataxia type 6	chr19:13207859-13207897	SCA6
Spinocerebellar ataxia type 7	chr3:63912686-63912715	SCA7
Spinocerebellar ataxia type 8	chr13:70139384-70139429	SCA8
Synpolydactyly I	chr2:176093059-176093103	SPD
Unverricht Lundborg disease	chr21:43776444-43776479	EPM1
X-linked dystonia-parkinsonism	chrX:71453055-71453129	XDP
X-linked hypopituitarism	chrX:140504317-140504381	XH
X-linked intellectual disability	chrX:25013654-25013697	XLID

Table 4.1: Table of targeted disease-associated repeat expansion diseases. List of 48 targeted repeat expansion diseases (Left), their known coordinates (Middle), and their abbreviation (Right).

4.3 Methods

4.3.1 dCas9 Pulldown of L1Hs and CGG Trinucleotide Repeats

First the gDNA was end-repaired by mixing 10µL of Cutsmart, 5µL of Klenow, 3.3µL of 2mM dNTP with approximately 77µL of gDNA (for a final volume of 100µL). The final mix was split into two 50µL reactions and incubated for 30 minutes at 25 degrees C and then 20 minutes at 75 degrees. For each guide RNA, the RNP was formed by combining 3µL of Cutsmart, 4.31µL of dCas9-biotin (Sigma), 850ng of guide RNA, H2O to 30µL, and incubating at room temperature for 20 minutes. The end repaired genomic DNA was pooled with the RNP (130µL total) and incubated at 37 for 30 minutes. To fix the dCas9 complexes to their targets, we added 7.5µL of 16% formaldehyde to the sample and 32.5µL of 1x cutsmart. Formaldehyde crosslinking reaction was incubated at 37 degrees for 15 minutes, and neutralized by adding Tris HCl pH 8.4 to 200mM and NaCl to 300 mM in 1mL total. Hydrophillic streptavidin beads (300µL) were washed 2 times in 1x phosphate buffered saline. The quenched sample was added to the washed streptavidin bead pellet and rotated at 4 degrees overnight. After the overnight binding, the beads were washed 3 times in 750µL of bead wash buffer (BWB; 2M NaCl , 5mM Tris HCl pH 7.5). Beads were washed in 1x Cutsmart buffer and resuspended in 100µL of cutsmart. 2µL of 1M Tris HCl pH, 2µL of Taq Polymerase, and 2µL of 10mM dATP were added to the sample and incubated at 72 degrees for 30 minutes and then reverse crosslinked at 65 degrees for 16 hours. After the reverse crosslinking, the beads were magnetically immobilized and the supernatant was recovered and added to ligation buffer (50µL LNB, 2µL of AMX, 7µL H2O, and 6µL of T4 DNA ligase HC) and incubated at room temp for 20 mins. 200µL of 1x TE buffer was added to the reaction and the DNA was immobilized on SPRI beads at 0.3x the volume of the sample. After 2 washes with

100µL of LFB, the sample was eluted and prepared for sequencing on nanopore.

4.3.2 Paralleled Cas9 Enrichment of Multiple Repeat Loci

Genomic DNA was dephosphorylated in 40µL total with 6µL of QuickCip and 4µL of Cutsmart buffer at 37 degrees for 30 minutes. The QuickCip was heat inactivated at 80 degrees C for 2 minutes. During the dephosphorylation reaction, the Cas9 ribonucleoprotein (RNP) was prepared by combining 1µL of a 1:5 dilution of the Cas9 enzyme (IDT) with 850ng total guide RNA and 2µL of Cutsmart buffer in a total volume of 20µL. Components of the RNP were mixed by tapping, pµLse spun, and incubated at room temperature for 20 minutes, then incubated on ice. To perform the Cas9 enrichment, the dephosphorylated gDNA was combined with 20µL of the RNP, 1.5µL of Taq Polymerase, and 1µL of 10mM dATP. The reaction was mixed and incubated at 37 degrees for 30 minutes for Cas9 cutting and then at 75 degrees C for 10 minutes for Taq monoadenylation. To ligate the nanopore adapters, 25µL of ligation buffer (40% Polyethylene glycol, 4x T4 DNA ligase buffer), 60µL of sample, 5µL of AMX, and 5µL of T4 DNA ligase were mixed together and rotated for 30 minutes at room temperature. One volume of 1xTE buffer was added to the ligation and mixed by tapping. SPRI beads were added at 0.3x the volume of the ligation and incubated at room temperature with rotation for 5 minutes. To wash the sample, 200µL of room temperature large fragment buffer (LFB) was added and the beads were resuspended by tapping. Cas9 enrichment experiments were pooled at this step. A second wash was performed after applying the magnet and discarding the supernatant from the pelleted beads. The washed beads were pµLse spun and immobilized on the magnet, and the residual LFB was aspirated with a pipette. The sample was eluted from the beads in elution buffer for 10 minutes are room temperature (EB; 6µL for Flongle, 10µL for Minion). The sequencing library was prepared by the addition of sequencing buffer (SQB; 16µL for Flongle, 26 for MinION), sequencing tether (SQT; 0.5µL) and loading beads (LB; 9.5µL).

4.3.3 In Vitro Transcription of Guide RNAs

Oligonucleotides were designed according to specifications for the EnGen In Vitro transcription kit and synthesized by IDT. Guide plates were resuspended to 10uM in PCR grade H2O. Per plate, oligos were pooled together according to upstream or downstream targeting groups for a total of two pools (48 upstream, 48 downstream) per plate. 2.5µL of 10uM oligos were combined with EnGen In Vitro transcription kit components and incubated at 37 degrees C for 45 minutes. After a DNase I treatment for 15 minutes at 37, guide RNAs were extracted in 200µL of Trizol and 50µL of chloroform. The aqueous layer was collected and reextracted with 50µL of chloroform to remove residual phenol. The subsequent aqueous layer was combined with 2 volumes of 100% ethanol and Sodium acetate to 0.3M and precipitated in a 20,000xg spin at 4 degrees C. RNA pellets were washed with 70% ethanol, air dried, and resuspended in pure water. All guide RNAs were prepared no earlier than a day in advance of the sequencing experiments.

4.3.4 Cell Culture and DNA Extraction

GM12878 cells were obtained from the Coriell Institute for Medical Research. Cells were cultured in RPMI 1640 supplemented with 1x anti/anti and 15% non heat inactivated fetal bovine serum and incubated at 37 degrees C with 5% CO2. Genomic DNA was isolated by the salting out method described in Chapter 3. First cells were harvested from culture and centrifuged at 1000xg for 5 minutes. Supernatant was aspirated off and the cell pellet was resuspended in 3mL of cell lysis buffer (400mM NaCl, 10mM Tris HCl pH 8.0, and 1mM EDTA). While gently vortexing, 200µL of 10% SDS was pipetted dropwise into the resuspension to lyse the cells. 50µL of RNase (10mg/mL) was added and the lysate was incubated at 37 degrees for 30 minutes with rotation. To remove protein, 200µL of proteinase K (10mg/mL) was added and the lysate was incubated at 45 degrees C with agitation for 16-20 hours. After this incubation, 1mL of saturated NaCl was pipetted into the lysate and the tube was shaken vigorously and subsequently centrifuged for 15 minutes at 4000xg. The supernatant was gently poured into a 15mL conical tube containing 2 volumes (8mL) of 100% ethanol. The tube was inverted several times until the DNA was precipitated out. After spooling DNA out using a p10 filter tip, the DNA was added to a 1.5mL tube and resuspended in 300µL of 1x TE buffer (10mM Tris HCl pH 8.0, 1mM EDTA). Isolated gDNA was stored at 4 degrees.

4.4 Results

4.4.1 dCas9 Efficiently Enriches for a Triplet CGG Trinucleotide Repeat and L1Hs in GM12878 DNA

Previous work has shown that targeting CGG polynucleotide repeat expansions in vivo with dCas9 systems activates otherwise transcriptionally silent loci [177]. We reasoned that dCas9 should also efficiently target CGG repeats within purified genomic DNA in vitro, and that with a selectable modification, it would be possible to physically enrich for dCas9 targets. We employed biotinylated dCas9 in a custom enrichment pipeline (See Methods 4.3) on purified GM12878 genomic DNA to capture CGG repeat regions over background sequence (Figure 4.2). In addition, we included an in vitro transcribed guide that targets L1Hs to assess performance of this method with the Cas9 enrichment (See Chapter 3 methods and results). In total, this experiment produced 12,062 passing reads with and N50 of 10.8 kb

To assess the number of reads that contained putative L1Hs signal, we scanned each passed read for a perfect match to the guide sequence used. Because this approach uses a dCas9, there is no cut site during the enrichment. Therefore, L1Hs signal will not appear at either end of the read as seen previously (See Chapter 3 methods and results), and any read containing a match to the guide RNA is considered enriched. To determine CGG enrichment, we identified all reads that contained at least 3 consecutive CGG trinucleotide repeats (CGG3). L1Hs signal was found in 12% of the total reads (1,455/12,062) and CGG3 was found in 2% of the total reads (239/12,062). One example of overlapping reads containing CGG3 was found at a ribosomal RNA (rRNA) cluster on the p arm of chromosome 21. Three long reads directly overlap a CGG3 motif within the RNA45SN1 gene (Figure 4.2). In addition, multiple reads aligned proximal to the CGG3 but did not overlap an annotated CGG3 repeat.

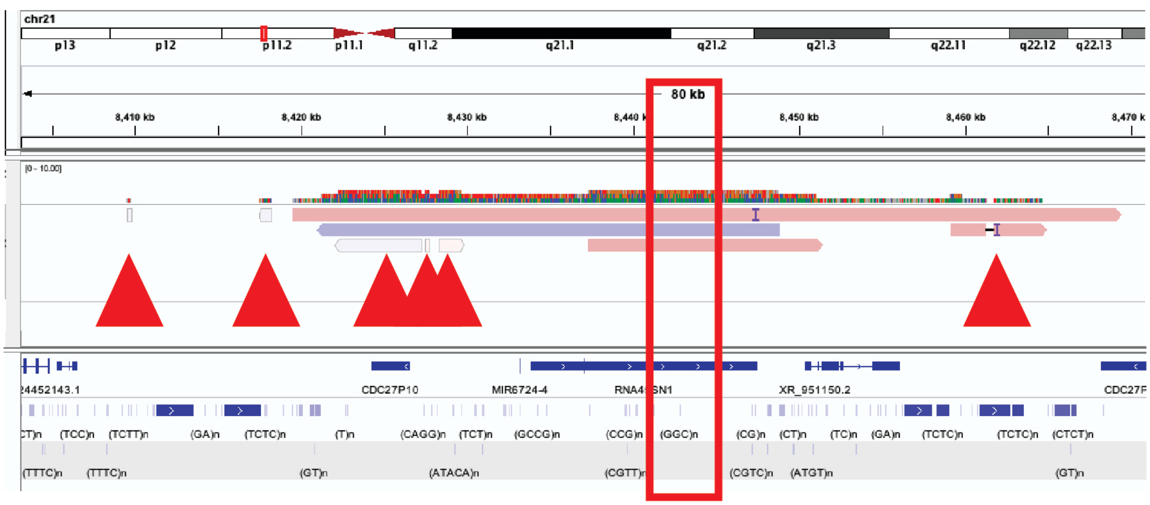

Figure 4.2: dCas9 captures putative CGG repeats for nanopore sequencing. Integrated Genome Viewer snapshot of nanopore reads from dCas9 enrichment for CGG

4.4.2 Cas9 Enrichment and Nanopore Sequencing on a Flongle Captures Repeat Associated Disease Loci in GM12878

Cas9 has been developed and used previously to capture individual loci [176]. We selected 48 known (Table 4.1) disease-associated polynucleotide repeat regions in the genome and designed oligonucleotide pairs flanking each repeat, for a total of 96 oligonucleotides. First, we in vitro transcribed equimolar pools of the 96 oligonucleotides into guide RNAs and performed Cas9 enrichment and nanopore sequencing experiments on a Flongle to capture 46 different disease associated repeats (Figure 4.3). We obtained 20,786 passed reads with an approximate on target rate of 0.4% and a N50 of 20.5kb. Slightly over half of the target regions (25/48) were captured by at least one read. The Fuchs endothelial corneal dystrophy (FECD) gene had the highest fraction of on-target reads of any of the selected regions (Figure 4.4). To further understand the specific structure of read alignments over target sites, we manually examined two loci, the FECD (an alias of TCF4) and a polynucleotide repeat at the HOXA13 gene responsible for hand-foot-genital syndrome (HFGS). We found that reads overlapping the FECD locus were almost uniform in length (1.5kb) in length (Figure 4.4). In contrast, reads overlapping the HFGS (HOXA13) locus were substantially longer (>6kb) (Figure 4.5.)

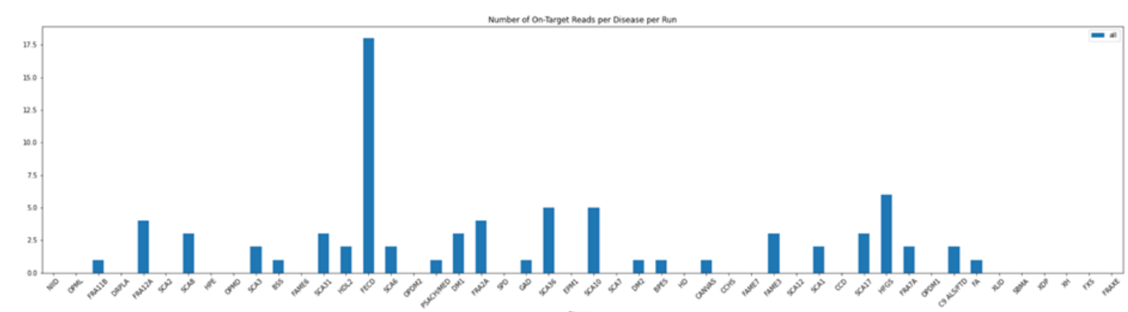

Figure 4.3: Flanked Cas9 enrichment for 48 disease-associated repeat loci in GM12878. X-axis represents repeat expansion associated diseases. Y-axis represents number of reads.

The difference in read structure between both the FECD and HFGS targets is attributable to a difference in which guide targets resulted in cuts. In FECD (Figure 4.4), the shorter 1.5kb reads resulted from both target sites being cut by Cas9. Oppositely, the longer reads (>6kb) at both the FECD and HFGS loci represent

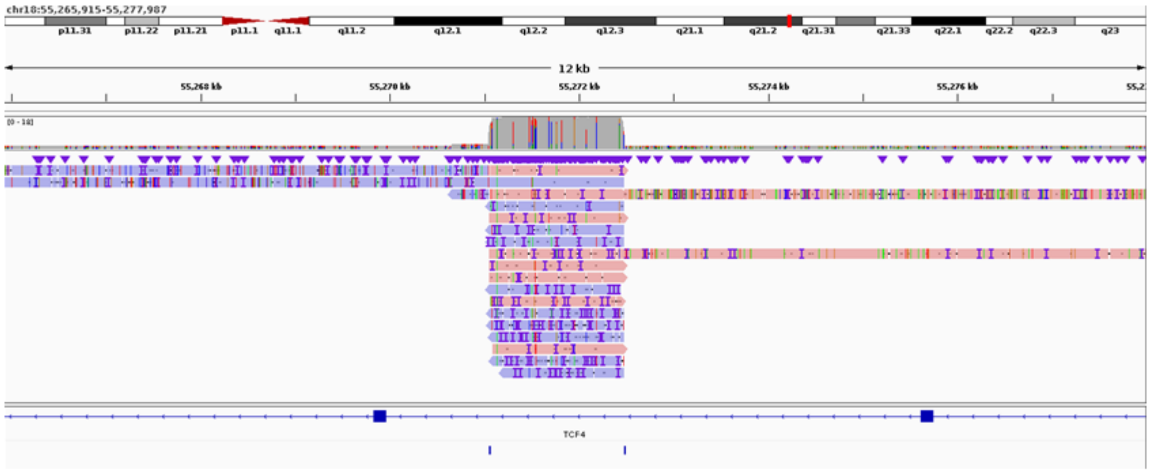

Figure 4.4: Read structure and level of enrichment for Fuchs endothelial corneal dystrophy (FECD) associated repeat. Integrated Genome Viewer snapshot of nanopore reads overlapping the FECD locus on Chromosome 18. Window size is 12kb. Gene track (bottom) and guide RNA cut positions (blue bars aligned with reads) are centered on repeat.

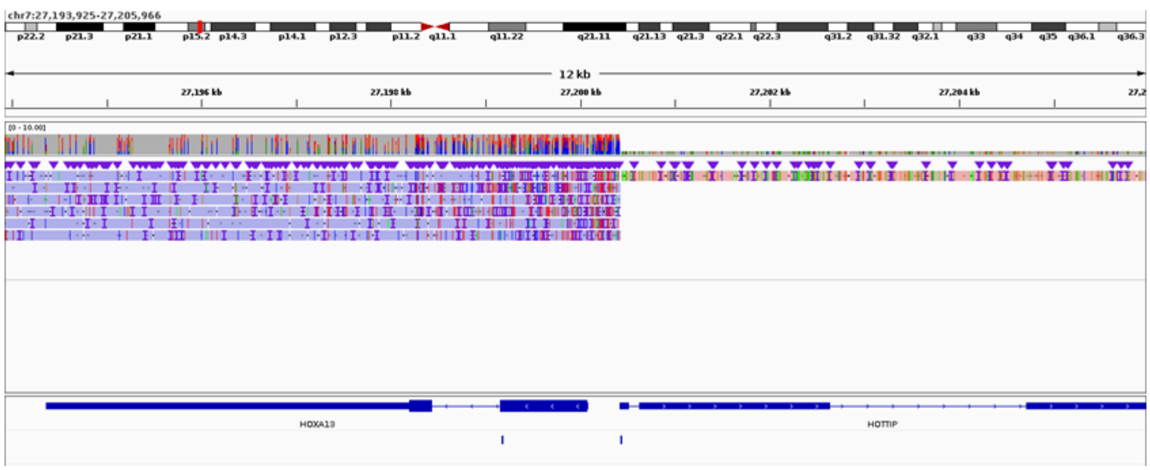

Figure 4.5: Read structure and level of enrichment for hand-foot-genital syndrome (HFGS) associated repeat. Integrated Genome Viewer snapshot of nanopore reads overlapping the HFGS locus on Chromosome 18. Window size is 12kb. Gene track (bottom) and guide RNA cut positions (blue bars aligned with reads) are centered on repeat.

a targeting of a single Cas9 (Figure 4.5. To maximize read lengths per region, we modified our targeting strategy and separated our guide RNAs into two groups: 1) upstream to the target and 2) downstream to the target. Each guide RNA group was separately transcribed in vitro and used in a Cas9 enrichment reaction and sequenced on a Flongle. To test the performance of this targeting strategy, we executed two enrichment and sequence experiments in total. The first yielded a total of 4,256 passed reads with an on-target rate of 0.41%. The total number of reads was substantially reduced compared to previous experiments. This is likely due to spontaneous sequencing pore death upon sample loading, leading to an overall reduction in sequencing capacity. The total number of captured regions was also reduced, with only 12 of 48 regions represented in the data (Figure 4.6.). The second experiment yielded 7,680 reads total with an on-target rate of 0.3%. Out of 48 targets, only 15 were captured by the enrichment and sequencing. Between the first and the second experiments, only 3 of the same target regions were captured in both (Figure 4.7).

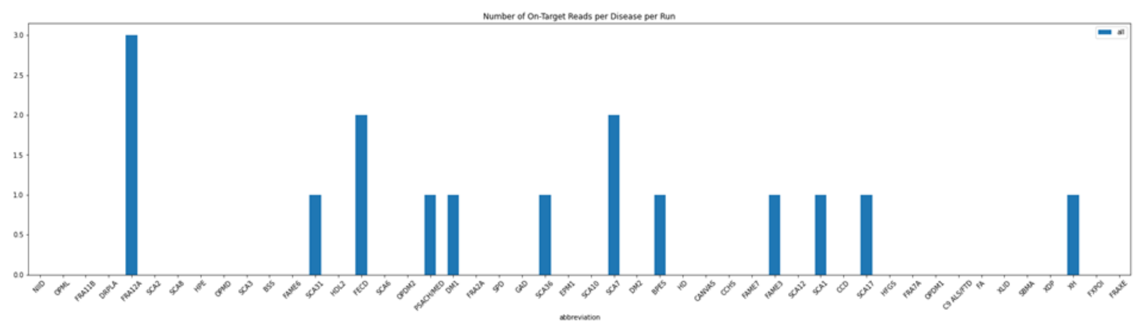

Figure 4.6: Asymmetric Cas9 enrichment for 48 disease-associated repeat loci in GM12878: experiment one. X-axis represents repeat expansion associated diseases. Y-axis represents number of reads.

The second experiment yielded 7,680 reads total with an on-target rate of 0.3%. Out of 48 targets, only 15 were captured by the enrichment and sequencing. Between the first and the second experiments, only 3 of the target regions were captured in both Figure.

4.5 Discussion

Here we present the framework for experiments to efficiently capture short tandem disease-associated repeats for nanopore sequencing. We explored two main approaches for capturing polynucleotide repeat regions: a dCas9 and a Cas9 method.

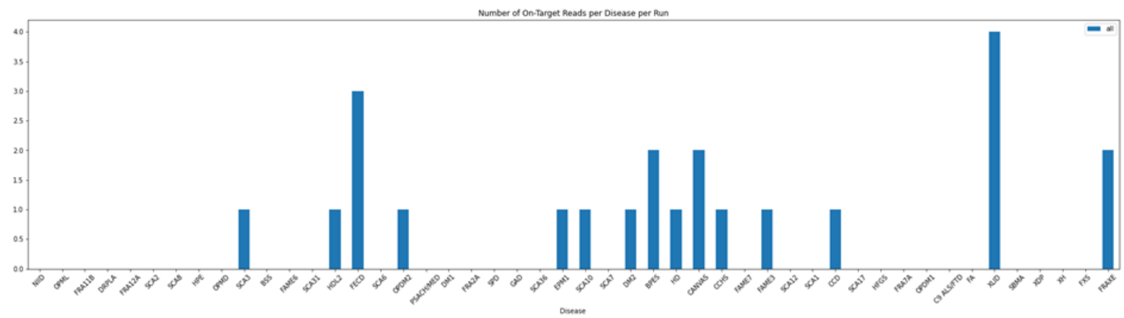

Figure 4.7: Asymmetric Cas9 enrichment for 48 disease-associated repeat loci in GM12878: experiment two. X-axis represents repeat expansion associated diseases. Y-axis represents number of reads.

Targeting CGG repeats using dCas9 has previously been used to reactivate the FMR1 locus in FXS stem cells [177]. We surmised that dCas9 could also be used in a targeted enrichment approach for triplet CGG motifs occurring in the GM12878 genome. We demonstrated on target rates as high as 2% (239/12,062 reads), which is a substantial enrichment over CGG triplets occurrences in non CGG targeted nanopore data (0.03 to 0.07%). Since dCas9 does not cut the DNA, it is impossible to delineate dCas9-bound fragments, and it is probable that background, unenriched CGG triplets are contributing to this enrichment. Nevertheless, Figure 4.2 shows a region with 3 nanopore reads overlapping a ribosomal RNA gene. It is exceedingly unlikely that this pattern occurred by random chance. Furthermore, 6 nanopore reads flank proximally to the overlapping reads, which is most reasonably explained by random DNA shearing after dCas9 pulldown. In fact, it is likely that due to random shearing of high molecular weight DNA immobilized at a single point on a bead, the effective enrichment may be significantly higher than 2%. Despite the performance of the dCas9 approach, it is not without limitations. The protospacer adjacent motif (PAM) for Cas9 is NGG. Because we anticipate dCas9 to be targeting to the repeat itself, and not outside of it, this effectively limits targeting to small subset of repeats that are sequence compatible with the PAM [178]. While other

Cas proteins have been characterized with different PAMs, they are not necessarily compatible with direct targeting of disease associated repeat sequences, rendering eventual implementation unlikely [179].

To overcome the limitations of using a dCas9 system, we also explored using its wild type form, Cas9. We used Cas9 to target 48 individual disease-associated repeat loci. We showed that Cas9 directed to flanks of repeat regions captures as many as half of the targeted repeats in a sequencing run with at least one supporting read. Our experiments established a proof of concept that pooled in vitro transcription of 96 gRNAs for targeted Cas9 enrichment and nanopore sequencing. While it is likely that equimolar pooled guide oligos do not in vitro transcribe at the same efficiency, and consequent gRNA concentrations differ, this does not seem to adversely affect the data. In the replicate experiments where we separated upstream and downstream targets into distinct enrichment reactions (Figure 4.6 and 4.7), such that no region is cut more than once, we observe that they only overlap with 3 of the same regions captured. Even though the targeting and in vitro transcription is identical, the sequenced targets differ dramatically. Furthermore, when we implemented all 96 targets, with 2 on either side of each region, we achieved the highest coverage across the most regions. These results suggest that, likely in all cases, the Cas9 is able to cut the target sites, albeit at different efficiencies [180]. Instead, it is likely that cut fragments with nanopore adapters are unable to reach the immobilized nanopores for sequencing. This may explain the increased coverage in Figure 4.3, as the excised repeat regions are shorter fragments (500bp to 2kb) and more soluble compared to the high molecular weight background genomic DNA. This explanation is consistent with the results from our replicate asymmetric targeting experiments (Figure 4.6 and 4.7), where the regions are cut on a single side and the sequenceable fragment is not predetermined or selected for a specific size. Even though cut and adapter-ligated, the high molecular weight DNA may less efficiently navigate to the sequencing pores compared to their shorter counterparts.

Nevertheless, these experiments cement a proof of principle for efficiently targeting disease-associated repeats. Our Cas9 approach is more suitable for targeting candidate repeats than the dCas9, largely due to the flexibility of targeting afforded by the gRNA design and preparation. Despite a high on target rate for dCas9 (2%), further development is stagnated by the intrinsic specificity of the PAM and unlikely to be scalable, even when incorporating additional Cas proteins with alternative PAMs. Previous work has shown that targeting multiple loci with flanking gRNAs can produce between 0.4 and 4% on target enrichment [176]. While our best enrichments barely measure up to their worst performing experiments, modeling a similar experimental approach will likely prove fruitful for our targeting strategy. These changes will likely include: 1) maximizing molecular weight of DNA from gDNA extraction, 2) extending Cas9 incubation, 3) increasing sequencing throughput, and 4) mitigating targeting disparities by redundant gRNA design. Altogether, we feel that this a positive first step in the future direction of characterizing repeat expansions. Due to the portability of nanopore sequencing, this concept may be a promising candidate as a clinical screening platform. In addition, the fast turnaround time and suitability for multiplexing samples could equip clinicians and genetic counselors with the tools to rapidly diagnose or rule-out known repeat expansion diseases in patients with undiagnosed syndromes.

4.6 Notes and Acknowledgements

I performed all enrichment and sequencing experiments. I developed the entire dCas9 enrichment method and conceived of guide RNA design strategy for the disease associated loci. Special thanks to Kinsey Van Deynze for providing the enrichment figures and analysis for the capture of disease loci, as well as for the computational automation of precise guide RNA positioning and design. Thank you to Camille Mumm for automated data processing pipeline. And thank you to Alan Boyle for basic enrichment analysis for the dCas9 experiment. This is an ongoing collaboration with Dr. Peter Todd's lab. Thank you to Connor Maltby for providing the list of disease associated repeat regions and their genomic coordinates.

CHAPTER V

Conclusions and Future Directions

5.1 Conclusions and Future Directions

Achieving a more complete understanding of regulatory and structural elements of the human genome has been focus of modern genetics. Since the publication of the first human genome draft sequence, interest in these regions, which constitute upwards of 50% of the genome, has only been heightened [50]. We now understand that these regions, collectively, are composed of open chromatin regions and repetitive DNA sequences. In addition, these regions of the genome dramatically influence the way in which genes are expressed, and their misregulation is responsible for a vast array of human diseases. Continued advancements in molecular methodologies are essential for fully appreciating the ongoing impact that regulatory and structural elements in humans.

My dissertation intersects with the development of molecular methodologies in genomics. The work presented here implements new technological approaches for characterizing regulatory and structural elements of the human genome. This work is anticipated to provide a foundation and precedent for leveraging novel enrichment and sequencing approaches to characterize open chromatin and highly repetitive structures.

5.1.1 Improving unbiased isolation of transcription factor bound sequences

In chapter 2, I developed a molecular protocol aimed at specifically isolating transcription factor bound fragments. While many open chromatin assays and computational algorithms have been established to provide functional annotation to regulatory regions, a collective limitation is their ability to discern precise transcription factor binding sites [36, 31, 33, 29, 30, 32, 39, 34]. In this chapter, I expressed a number of molecular biological steps overcome some of the limitations observed in other techniques. Due to nuclei isolation difficulties from conventional protocols, I developed a method that simultaneously crosslinks and isolates nuclei [73]. This doubled to remove mitochondrial DNA that regularly contaminates ATAC-seq [47]. In addition, it eliminated cytoplasmic proteins that would contaminate the filter binding experiments. Exploring nucleosome depletion methods eschewed the requirement of antibodies in ChIP approaches [181, 36], which would enable unbiased recovery of transcription factor bound sequences. Extensive digestion with nucleases would effectively reduce fragment size of transcription factor-DNA complexes, providing higher resolution of the specific protein-occupied sequence. Despite these implementations, I was unable to demonstrate isolation of transcription factor bound sequence. The largest obstacle in this chapter was fully depleting nucleosomes to give way to signal from transcription factor bound sequences. In a given diploid genome, there are about 30 million nucleosomes [182]. It is not surprising that the most common DNA binding protein is the histone packaging proteins that make up the nucleosome. Considering that K562 is near triploid, the nucleosome content is expected to be substantially higher than a cell line with a normal karyotype [183]. This overabundance of nucleosomal DNA and histone protein likely contributed to the difficulties associated with a complete depletion. However, this does not suggest that the approach is without merit or feasibility. To achieve full depletion of nucleosomes, it is likely that a serial depletion approach is required. We found that the most efficient means to deplete nucleosomal sequence was by immunoprecipitating with a pan histone antibody. Repeated immunodepletions of the same sample may completely eliminate histone protein and associated DNA rapidly. Another worthwhile approach would be separating out nucleosomes from transcription factor-DNA complexes in classic sedimentation experiments, the same experiments used to first characterize their structure [3]. Furthermore, careful optimization of crosslinking, whether with formaldehyde or with alternative crosslinkers, will mitigate the occurrence of spurious crosslinking that has been observed in fixation-based assays [78, 79]. Recent work in bacterial systems have developed a similar approach that isolates protein-occupied regulatory sequences [80]. This reaffirms the value of understanding the specific sequences occupied in regulatory regions and how that relates to fine-tuned regulatory control. Despite the wide selection of assays available, eukaryotic genomics remains limited in its capacity to fully dissect transcription factor binding. Successful isolation of small molecular weight DNA fragments that are largely, if not exclusively, transcription factor bound sequences will unlock a new perspective of functional genomics.

5.1.2 Characterizing Repetitive Elements Using Targeted Enrichment and Nanopore Sequencing

The research in chapter 3 and 4 focuses on using Cas9-based enrichment strategies paired with nanopore sequencing to capture highly repetitive regions of the genome [176]. Chapter 3 focuses on selectively sequencing retrotransposable elements, which comprise upwards of 40% of the genome alone. Due to technological and experimental limitations, finding recent retrotransposition events has been challenging. In chapter

3, I aimed to discover polymorphic mobile element insertions of 5 active retrotransposons: L1Hs, AluYb8, AluYa5, SVA_F, and SVA_E. My Cas9 targeted enrichment and nanopore sequencing experiments efficiently captured mobile element insertions from a well-annotated cell line. Collectively, the enrichment experiments averaged a 44% on target rate for mobile element insertions. When compared to whole genome sequencing on nanopore, mobile elements from the Cas9 enrichment datasets were between 13 and 50 fold more enriched. In addition, both polymorphic and reference mobile element insertion events reached near saturation in a single sequencing run, demonstrating the capacity of this approach to capture transposable elements. Finally, 17 insertion events were discovered that were neither reference, nor anticipated to be non reference. In chapter 4, I explored another type of structural variation known as polynucleotide repeat expansions. While these repeat regions make up a considerably smaller portion of the genome than mobile elements, expansions in these repeats are causative of upwards of 40 human diseases. Due to their tandem repetitive nature, efficient nucleotide repeat characterization has remained an obstacle to modern genomics methods[66, 60]. I characterized polynucleotide repeats using both dCas9 and Cas9 based enrichments paired with nanopore sequencing. I first developed a novel approach that leverages a biotinylated dCas9 to directly bind CGG triplet repeats and pulldown on streptavidin magnetic beads. This technique resulted in a near 200 fold enrichment of triplet CGG over background sequence. It was able to capture multiple DNA fragments overlapping a putative triplet CGG within ribosomal RNA gene. In addition, I used Cas9 for site-specific enrichment of 48 disease-associated repeat loci. Through a series of enrichment experiments, I showed that a majority (25/48) regions could be specifically captured and sequenced. This approach was done using pooled in vitro transcription to generate guide RNAs, which did not appear to reduce targeting efficiency, indicating that the addition of more targeted regions is possible. As a result, this technique can likely scale to assay many more genomic sites than were tested here, and will be a useful approach for the future discovery and characterization of disease-associated repeats. Chapters 3 and 4 emphasize the importance of enrichment approaches that collaborate with nanopore sequencing to discovery structural variation in the genome. Whole genome sequencing to capture specific regions in the genome diverts a majority of the sequencing potential to non target regions. These chapters have shown that targeted enrichment approaches can dramatically shift the distribution of captured reads to elements of interest. This was most evident in chapter 3, where mobile elements were abundantly enriched over background. However, with decreasing abundance of target regions, Cas9 works less efficiently, as seen in chapter 4, capturing only a subset of all of the target regions. While repeated experiments would eventually saturate all of the targets, the low enrichment partly inspired developing the dCas9 approach. By directing a dCas9 with biotin tag to repeats of CGG, it was not necessary to predefine regions of the genome to target [184]. Instead, this offered a discovery-based enrichment approach, much like chapter 3. In addition, the CGG repeat regions were immobilized by a crosslinked dCas9 on a streptavidin bead, which afforded washes and buffer exchanges to remove unbound and unenriched DNA sequences. These experiments provide insight into promising avenues for enrichment strategies, but also indicate that further research is required to efficiently capture rare structural events. New techniques for capturing exceedingly rare variation, such as somatic mosaicism, will be especially in demand in the future of genomics research.

5.1.3 Concluding remarks

The work in this dissertation has been an exploration of new molecular methods and technologies to understand structural and regulatory elements in the human genome. While the work discussed here expands the envelope of knowledge, there is still substantial work to be done. Functional genomics has accelerated in the past 20 years, but important questions remain about detailed interactions between transcription factors and their regulatory regions. The limitations of contemporary methods in genomics require innovative solutions to enhance our understanding of these interactions. Regardless of the detailed molecular strategy, future efforts to dissect precise transcription factor and binding site interactions in an unbiased, genome-wide way, with special attention to predecessors' limitations, will likely be fruitful. The development of such an approach would provide an unprecedented view eukaryotic regulatory control, and would be invaluable to understanding the genetic etiology of many human diseases.

In addition, continued exploration of mobile elements, polynucleotide repeats, as well as other classes of structural variation, will be essential to understanding how these elements continue to influence genome regulation and evolution. Technological limitations have stagnated rapid, and widespread discovery of structural variation. However, the advent of sequencing platforms able to capture unprecedented read lengths is quickly proving a valuable force in the tide of research. While this sequencing technology continues to improve, inventing novel enrichment techniques to capture low abundance genomic targets will be an important complement to efficiently capture rare and mosaic structural variation.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome core particle at 2.8 a resolution," *Nature*, vol. 389, pp. 251–260, Sept. 1997.

[2] F. Thoma, T. Koller, and A. Klug, "Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin," *J. Cell Biol.*, vol. 83, pp. 403–427, Nov. 1979.

[3] R. Rill and K. E. Van Holde, "Properties of nuclease-resistant fragments of calf thymus chromatin," *J. Biol. Chem.*, vol. 248, pp. 1080–1083, Feb. 1973.

[4] K. E. Van Holde, C. G. Sahasrabuddhe, and B. R. Shaw, "A model for particulate structure in chromatin," *Nucleic Acids Res.*, vol. 1, pp. 1579–1586, Nov. 1974.

[5] A. L. Olins and D. E. Olins, "Spheroid chromatin units (v bodies)," *Science*, vol. 183, pp. 330–332, Jan. 1974.

[6] K. S. Zaret, "Pioneer transcription factors initiating gene network changes," *Annu. Rev. Genet.*, vol. 54, pp. 367–385, Nov. 2020.

[7] J. L. Workman, "Nucleosome displacement in transcription," *Genes Dev.*, vol. 20, pp. 2009–2017, Aug. 2006.

[8] C.-K. Lee, Y. Shibata, B. Rao, B. D. Strahl, and J. D. Lieb, "Evidence for nucleosome depletion at active regulatory regions genome-wide," *Nat. Genet.*, vol. 36, pp. 900–905, Aug. 2004.

[9] F. Zhu, L. Farnung, E. Kaasinen, B. Sahu, Y. Yin, B. Wei, S. O. Dodonova, K. R. Nitta, E. Morgunova, M. Taipale, P. Cramer, and J. Taipale, "The interaction landscape between transcription factors and the nucleosome," *Nature*, vol. 562, pp. 76–81, Oct. 2018.

[10] D. L. Fulton, S. Sundararajan, G. Badis, T. R. Hughes, W. W. Wasserman, J. C. Roach, and R. Sladek, "TFCat: the curated catalog of mouse and human transcription factors," *Genome Biol.*, vol. 10, p. R29, Mar. 2009.

[11] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe, "A census of human transcription factors: function, expression and evolution," *Nat. Rev. Genet.*, vol. 10, pp. 252–263, Apr. 2009.

[12] P. B. Becker and J. L. Workman, "Nucleosome remodeling and epigenetics," *Cold Spring Harb. Perspect. Biol.*, vol. 5, Sept. 2013.

[13] W. S. Dynan and R. Tjian, "The promoter-specific transcription factor sp1 binds to upstream sequences in the SV40 early promoter," *Cell*, vol. 35, pp. 79–87, Nov. 1983.

[14] V. Goffin, D. Demonté, C. Vanhulle, S. de Walque, Y. de Launoit, A. Burny, Y. Collette, and C. Van Lint, "Transcription factor binding sites in the pol gene intragenic regulatory region of HIV-1 are important for virus infectivity," *Nucleic Acids Res.*, vol. 33, pp. 4285–4310, Aug. 2005.

[15] R. Andersson and A. Sandelin, "Determinants of enhancer and promoter activities of regulatory elements," *Nat. Rev. Genet.*, vol. 21, pp. 71–87, Feb. 2020.

[16] T. I. Lee and R. A. Young, "Transcriptional regulation and its misregulation in disease," *Cell*, vol. 152, pp. 1237–1251, Mar. 2013.

[17] B. Deplancke, D. Alpern, and V. Gardeux, "The genetics of transcription factor DNA binding variation," *Cell*, vol. 166, pp. 538–554, July 2016.

[18] J. M. Scandura, P. Boccuni, J. Cammenga, and S. D. Nimer, "Transcription factor fusions in acute leukemia: variations on a theme," *Oncogene*, vol. 21, pp. 3422–3444, May 2002.

[19] S. H. Orkin, H. H. Kazazian, Jr, S. E. Antonarakis, S. C. Goff, C. D. Boehm, J. P. Sexton, P. G. Waber, and P. J. Giardina, "Linkage of beta-thalassaemia mutations and beta-globin gene polymorphisms with DNA polymorphisms in human beta-globin gene cluster," *Nature*, vol. 296, pp. 627–631, Apr. 1982.

[20] I. J. Miller and J. J. Bieker, "A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the krüppel family of nuclear proteins," *Mol. Cell. Biol.*, vol. 13, pp. 2776–2786, May 1993.

[21] S. H. Orkin, S. E. Antonarakis, and H. H. Kazazian, Jr, "Base substitution at position -88 in a beta-thalassemic globin gene. further evidence for the role of distal promoter element ACACCC," *J. Biol. Chem.*, vol. 259, pp. 8679–8681, July 1984.

[22] L. Duncan, Z. Yilmaz, H. Gaspar, R. Walters, J. Goldstein, V. Anttila, B. Bulik-Sullivan, S. Ripke, Eating Disorders Working Group of the Psychiatric Genomics Consortium, L. Thornton, A. Hinney, M. Daly, P. F. Sullivan, E. Zeggini, G. Breen, and C. M. Bulik, "Significant locus and metabolic genetic correlations revealed in Genome-Wide association study of anorexia nervosa," *Am. J. Psychiatry*, vol. 174, pp. 850–858, Sept. 2017.

[23] R. L. Milne, K. B. Kuchenbaecker, K. Michailidou, J. Beesley, S. Kar, S. Lindström, S. Hui, A. Lemaçon, P. Soucy, J. Dennis, X. Jiang, A. Rostamianfar, H. Finucane, M. K. Bolla, L. McGuffog, Q. Wang, C. M. Aalfs, ABCTB Investigators, M. Adams, J. Adlard, S. Agata, S. Ahmed, H. Ahsan, K. Aittomäki, F. Al-Ejeh, J. Allen, C. B. Ambrosone, C. I. Amos, I. L. Andrulis, H. Anton-Culver, N. N. Antonenkova, V. Arndt, N. Arnold, K. J. Aronson, B. Auber, P. L. Auer, M. G. E. M. Ausems, J. Azzollini, F. Bacot, J. Balmaña, M. Barile, L. Barjhoux, R. B. Barkardottir, M. Barrdahl, D. Barnes, D. Barrowdale, C. Baynes, M. W. Beckmann, J. Benitez, M. Bermisheva, L. Bernstein, Y.-J. Bignon, K. R. Blazer, M. J. Blok, C. Blomqvist, W. Blot, K. Bobolis, B. Boeckx, N. V. Bogdanova, A. Bojesen, S. E. Bojesen, B. Bonanni, A.-L. Børresen-Dale, A. Bozsik, A. R. Bradbury, J. S. Brand, H. Brauch, H. Brenner, B. Bressac-de Paillerets, C. Brewer, L. Brinton, P. Broberg, A. Brooks-Wilson, J. Brunet, T. Brüning, B. Burwinkel, S. S. Buys, J. Byun, Q. Cai, T. Caldés, M. A. Caligo, I. Campbell, F. Canzian, O. Caron, A. Carracedo, B. D. Carter, J. E. Castelao, L. Castera, V. Caux-Moncoutier, S. B. Chan, J. Chang-Claude, S. J. Chanock, X. Chen, T.-Y. D. Cheng, J. Chiquette, H. Christiansen, K. B. M. Claes, C. L. Clarke, T. Conner, D. M. Conroy, J. Cook, E. Cordina-Duverger, S. Cornelissen, I. Coupier, A. Cox, D. G. Cox, S. S. Cross, K. Cuk, J. M. Cunningham, K. Czene, M. B. Daly, F. Damiola, H. Darabi, R. Davidson, K. De Leeneer, P. Devilee, E. Dicks, O. Diez, Y. C. Ding, N. Ditsch, K. F. Doheny, S. M. Domchek, C. M. Dorfling, T. Dörk, I. Dos-Santos-Silva, S. Dubois, P.-A. Dugué, M. Dumont, A. M. Dunning, L. Durcan, M. Dwek, B. Dworniczak, D. Eccles, R. Eeles, H. Ehrencrona, U. Eilber, B. Ejlertsen, A. B. Ekici, A. H. Eliassen, EMBRACE, C. Engel, M. Eriksson, L. Fachal, L. Faivre, P. A. Fasching, U. Faust, J. Figueroa, D. Flesch-Janys, O. Fletcher, H. Flyger, W. D. Foulkes, E. Friedman, L. Fritschi, D. Frost, M. Gabrielson, P. Gaddam, M. D. Gammon, P. A. Ganz, S. M. Gapstur, J. Garber, V. Garcia-Barberan, J. A. García-Sáenz, M. M. Gaudet, M. Gauthier-Villars, A. Gehrig, GEMO Study Collaborators, V. Georgoulias, A.-M. Gerdes, G. G. Giles, G. Glendon, A. K. Godwin, M. S. Goldberg, D. E.

Goldgar, A. González-Neira, P. Goodfellow, M. H. Greene, G. I. G. Alnæs, M. Grip, J. Gronwald, A. Grundy, D. Gschwantler-Kaulich, P. Guénel, Q. Guo, L. Haeberle, E. Hahnen, C. A. Haiman, N. Håkansson, E. Hallberg, U. Hamann, N. Hamel, S. Hankinson, T. V. O. Hansen, P. Harrington, S. N. Hart, J. M. Hartikainen, C. S. Healey, HEBON, A. Hein, S. Helbig, A. Henderson, J. Heyworth, B. Hicks, P. Hillemanns, S. Hodgson, F. B. Hogervorst, A. Hollestelle, M. J. Hooning, B. Hoover, J. L. Hopper, C. Hu, G. Huang, P. J. Hulick, K. Humphreys, D. J. Hunter, E. N. Imyanitov, C. Isaacs, M. Iwasaki, L. Izatt, A. Jakubowska, P. James, R. Janavicius, W. Janni, U. B. Jensen, E. M. John, N. Johnson, K. Jones, M. Jones, A. Jukkola-Vuorinen, R. Kaaks, M. Kabisch, K. Kaczmarek, D. Kang, K. Kast, kConFab/AOCS Investigators, R. Keeman, M. J. Kerin, C. M. Kets, M. Keupers, S. Khan, E. Khusnutdinova, J. I. Kiiski, S.-W. Kim, J. A. Knight, I. Konstantopoulou, V.-M. Kosma, V. N. Kristensen, T. A. Kruse, A. Kwong, A.-V. Lænkholm, Y. Laitman, F. Lalloo, D. Lambrechts, K. Landsman, C. Lasset, C. Lazaro, L. Le Marchand, J. Lecarpentier, A. Lee, E. Lee, J. W. Lee, M. H. Lee, F. Lejbkowicz, F. Lesueur, J. Li, J. Lilyquist, A. Lincoln, A. Lindblom, J. Lissowska, W.-Y. Lo, S. Loibl, J. Long, J. T. Loud, J. Lubinski, C. Luccarini, M. Lush, R. J. MacInnis, T. Maishman, E. Makalic, I. M. Kostovska, K. E. Malone, S. Manoukian, J. E. Manson, S. Margolin, J. W. M. Martens, M. E. Martinez, K. Matsuo, D. Mavroudis, S. Mazoyer, C. McLean, H. Meijers-Heijboer, P. Menéndez, J. Meyer, H. Miao, A. Miller, N. Miller, G. Mitchell, M. Montagna, K. Muir, A. M. Mulligan, C. Mulot, S. Nadesan, K. L. Nathanson, NBSC Collaborators, S. L. Neuhausen, H. Nevanlinna, I. Nevelsteen, D. Niederacher, S. F. Nielsen, B. G. Nordestgaard, A. Norman, R. L. Nussbaum, E. Olah, O. I. Olopade, J. E. Olson, C. Olswold, K.-R. Ong, J. C. Oosterwijk, N. Orr, A. Osorio, V. S. Pankratz, L. Papi, T.-W. Park-Simon, Y. Paulsson-Karlsson, R. Lloyd, I. S. Pedersen, B. Peissel, A. Peixoto, J. I. A. Perez, P. Peterlongo, J. Peto, G. Pfeiler, C. M. Phelan, M. Pinchev, D. Plaseska-Karanfilska, B. Poppe, M. E. Porteous, R. Prentice, N. Presneau, D. Prokofieva, E. Pugh, M. A. Pujana, K. Pylkäs, B. Rack, P. Radice, N. Rahman, J. Rantala, C. Rappaport-Fuerhauser, G. Rennert, H. S. Rennert, V. Rhenius, K. Rhiem, A. Richardson, G. C. Rodriguez, A. Romero, J. Romm, M. A. Rookus, A. Rudolph, T. Ruediger, E. Saloustros, J. Sanders, D. P. Sandler, S. Sangrajrang, E. J. Sawyer, D. F. Schmidt, M. J. Schoemaker, F. Schumacher, P. Schürmann, L. Schwentner, C. Scott, R. J. Scott, S. Seal, L. Senter, C. Seynaeve, M. Shah, P. Sharma, C.-Y. Shen, X. Sheng, H. Shimelis, M. J. Shrubsole, X.-O. Shu, L. E. Side, C. F. Singer, C. Sohn, M. C. Southey, J. J. Spinelli, A. B. Spurdle, C. Stegmaier, D. Stoppa-Lyonnet, G. Sukiennicki, H. Surowy, C. Sutter, A. Swerdlow, C. I. Szabo, R. M. Tamimi, Y. Y. Tan, J. A. Taylor, M.-I. Tejada, M. Tengström, S. H. Teo, M. B. Terry, D. C. Tessier, A. Teulé, K. Thöne, D. L. Thull, M. G. Tibiletti, L. Tihomirova, M. Tischkowitz, A. E. Toland, R. A. E. M. Tollenaar, I. Tomlinson, L. Tong, D. Torres, M. Tranchant, T. Truong, K. Tucker, N. Tung, J. Tyrer, H.-U. Ulmer, C. Vachon, C. J. van Asperen, D. Van Den Berg, A. M. W. van den Ouweland, E. J. van Rensburg, L. Varesco, R. Varon-Mateeva, A. Vega, A. Viel, J. Vijai, D. Vincent, J. Vollenweider, L. Walker, Z. Wang, S. Wang-Gohrke, B. Wappenschmidt, C. R. Weinberg, J. N. Weitzel, C. Wendt, J. Wesseling, A. S. Whittemore, J. T. Wijnen, W. Willett, R. Winqvist, A. Wolk, A. H. Wu, L. Xia, X. R. Yang, D. Yannoukakos, D. Zaffaroni, W. Zheng, B. Zhu, A. Ziogas, E. Ziv, K. K. Zorn, M. Gago-Dominguez, A. Mannermaa, H. Olsson, M. R. Teixeira, J. Stone, K. Offit, L. Ottini, S. K. Park, M. Thomassen, P. Hall, A. Meindl, R. K. Schmutzler, A. Droit, G. D. Bader, P. D. P. Pharoah, F. J. Couch, D. F. Easton, P. Kraft, G. Chenevix-Trench, M. García-Closas, M. K. Schmidt, A. C. Antoniou, and J. Simard, "Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer," *Nat. Genet.*, vol. 49, pp. 1767–1778, Dec. 2017.

[24] W. Zhao, A. Rasheed, E. Tikkanen, J.-J. Lee, A. S. Butterworth, J. M. M. Howson, T. L. Assimes, R. Chowdhury, M. Orho-Melander, S. Damrauer, A. Small, S. Asma, M. Imamura, T. Yamauch, J. C. Chambers, P. Chen, B. R. Sapkota, N. Shah, S. Jabeen, P. Surendran, Y. Lu, W. Zhang, A. Imran, S. Abbas, F. Majeed, K. Trindade, N. Qamar, N. H. Mallick, Z. Yaqoob, T. Saghir, S. N. H. Rizvi, A. Memon, S. Z. Rasheed, F.-U.-R. Memon,

K. Mehmood, N. Ahmed, I. H. Qureshi, Tanveer-Us-Salam, W. Iqbal, U. Malik, N. Mehra, J. Z. Kuo, W. H.-H. Sheu, X. Guo, C. A. Hsiung, J.-M. J. Juang, K. D. Taylor, Y.-J. Hung, W.-J. Lee, T. Quertermous, I.-T. Lee, C.-C. Hsu, E. P. Bottinger, S. Ralhan, Y. Y. Teo, T.-D. Wang, D. S. Alam, E. Di Angelantonio, S. Epstein, S. F. Nielsen, B. G. Nordestgaard, A. Tybjaerg-Hansen, R. Young, CHD Exome+ Consortium, M. Benn, R. Frikke-Schmidt, P. R. Kamstrup, EPIC-CVD Consortium, EPIC-Interact Consortium, Michigan Biobank, J. W. Jukema, N. Sattar, R. Smit, R.-H. Chung, K.-W. Liang, S. Anand, D. K. Sanghera, S. Ripatti, R. J. F. Loos, J. S. Kooner, E. S. Tai, J. I. Rotter, Y.-D. I. Chen, P. Frossard, S. Maeda, T. Kadowaki, M. Reilly, G. Pare, O. Melander, V. Salomaa, D. J. Rader, J. Danesh, B. F. Voight, and D. Saleheen, "Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease," *Nat. Genet.*, vol. 49, pp. 1450–1457, Oct. 2017.

[25] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre, "Benefits and limitations of genome-wide association studies," *Nat. Rev. Genet.*, vol. 20, pp. 467–484, Aug. 2019.

[26] S. Dong and A. P. Boyle, "Predicting functional variants in enhancer and promoter elements using RegulomeDB," *Hum. Mutat.*, vol. 40, pp. 1292–1298, Sept. 2019.

[27] S. S. Nishizaki, N. Ng, S. Dong, R. S. Porter, C. Morterud, C. Williams, C. Asman, J. A. Switzenberg, and A. P. Boyle, "Predicting the effects of SNPs on transcription factor binding affinity," *Bioinformatics*, vol. 36, pp. 364–372, Jan. 2020.

[28] G. E. Crawford, I. E. Holt, J. C. Mullikin, D. Tai, R. Blakesley, G. Bouffard, A. Young, C. Masiello, E. D. Green, T. G. Wolfsberg, F. S. Collins, and National Institutes Of Health Intramural Sequencing Center, "Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, pp. 992–997, Jan. 2004.

[29] G. E. Crawford, I. E. Holt, J. Whittle, B. D. Webb, D. Tai, S. Davis, E. H. Margulies, Y. Chen, J. A. Bernat, D. Ginsburg, D. Zhou, S. Luo, T. J. Vasicek, M. J. Daly, T. G. Wolfsberg, and F. S. Collins, "Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS)," *Genome Res.*, vol. 16, pp. 123–131, Jan. 2006.

[30] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford, "High-resolution mapping and characterization of open chromatin across the genome," *Cell*, vol. 132, pp. 311–322, Jan. 2008.

[31] J. D. Buenrostro, B. Wu, H. Y. Chang, and W. J. Greenleaf, "ATAC-seq: A method for assaying chromatin accessibility Genome-Wide," *Curr. Protoc. Mol. Biol.*, vol. 109, pp. 21.29.1–21.29.9, Jan. 2015.

[32] A. Karabacak Calviello, A. Hirsekorn, R. Wurmus, D. Yusuf, and U. Ohler, "Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling," *Genome Biol.*, vol. 20, p. 42, Feb. 2019.

[33] P. G. Giresi, J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb, "FAIRE (Formaldehyde-Assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin," *Genome Res.*, vol. 17, pp. 877–885, June 2007.

[34] H. H. He, C. A. Meyer, S. S. Hu, M.-W. Chen, C. Zang, Y. Liu, P. K. Rao, T. Fei, H. Xu, H. Long, X. S. Liu, and M. Brown, "Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification," *Nat. Methods*, vol. 11, pp. 73–78, Jan. 2014.

[35] L. Song, Z. Zhang, L. L. Grasfeder, A. P. Boyle, P. G. Giresi, B.-K. Lee, N. C. Sheffield, S. Gräf, M. Huss, D. Keefe, Z. Liu, D. London, R. M. McDaniell, Y. Shibata, K. A. Showers, J. M. Simon, T. Vales, T. Wang, D. Winter, Z. Zhang, N. D. Clarke, E. Birney, V. R. Iyer,

G. E. Crawford, J. D. Lieb, and T. S. Furey, "Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity," *Genome Res.*, vol. 21, pp. 1757–1767, Oct. 2011.

[36] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao, "High-resolution profiling of histone methylations in the human genome," *Cell*, vol. 129, pp. 823–837, May 2007.

[37] M. J. Rossi, W. K. M. Lai, and B. F. Pugh, "Simplified ChIP-exo assays," *Nat. Commun.*, vol. 9, p. 2842, July 2018.

[38] D. J. Galas and A. Schmitz, "DNAase footprinting a simple method for the detection of protein-DNA binding specificity," *Nucleic Acids Res.*, vol. 5, pp. 3157–3170, Sept. 1978.

[39] B. Quach and T. S. Furey, "DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter," *Bioinformatics*, vol. 33, pp. 956–963, Apr. 2017.

[40] A. P. Boyle, L. Song, B.-K. Lee, D. London, D. Keefe, E. Birney, V. R. Iyer, G. E. Crawford, and T. S. Furey, "High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells," *Genome Res.*, vol. 21, pp. 456–464, Mar. 2011.

[41] N. Ouyang and A. P. Boyle, "TRACE: transcription factor footprinting using chromatin accessibility data and DNA sequence," *Genome Res.*, vol. 30, pp. 1040–1046, July 2020.

[42] J. Gutin, R. Sadeh, N. Bodenheimer, D. Joseph-Strauss, A. Klein-Brill, A. Alajem, O. Ram, and N. Friedman, "Fine-Resolution mapping of TF binding and chromatin interactions," *Cell Rep.*, vol. 22, pp. 2797–2807, Mar. 2018.

[43] W. K. M. Lai, L. Mariani, G. Rothschild, E. R. Smith, B. J. Venters, T. R. Blanda, P. K. Kuntala, K. Bocklund, J. Mairose, S. N. Dweikat, K. Mistretta, M. J. Rossi, D. James, J. T. Anderson, S. K. Phanor, W. Zhang, Z. Zhao, A. P. Shah, K. Novitzky, E. McAnarney, M.-C. Keogh, A. Shilatifard, U. Basu, M. L. Bulyk, and B. F. Pugh, "A ChIP-exo screen of 887 protein capture reagents program transcription factor antibodies in human cells," *Genome Res.*, vol. 31, pp. 1663–1679, Sept. 2021.

[44] T. A. Egelhofer, A. Minoda, S. Klugman, K. Lee, P. Kolasinska-Zwierz, A. A. Alekseyenko, M.-S. Cheung, D. S. Day, S. Gadel, A. A. Gorchakov, T. Gu, P. V. Kharchenko, S. Kuan, I. Latorre, D. Linder-Basso, Y. Luu, Q. Ngo, M. Perry, A. Rechtsteiner, N. C. Riddle, Y. B. Schwartz, G. A. Shanower, A. Vielle, J. Ahringer, S. C. R. Elgin, M. I. Kuroda, V. Pirrotta, B. Ren, S. Strome, P. J. Park, G. H. Karpen, R. D. Hawkins, and J. D. Lieb, "An assessment of histone-modification antibody quality," *Nat. Struct. Mol. Biol.*, vol. 18, pp. 91–93, Jan. 2011.

[45] M. Baker, "Blame it on the antibodies," *Nature*, vol. 521, May 2015.

[46] Y. Sun, N. Miao, and T. Sun, "Detect accessible chromatin using ATAC-sequencing, from principle to applications," *Hereditas*, vol. 156, p. 29, Aug. 2019.

[47] M. R. Corces, A. E. Trevino, E. G. Hamilton, P. G. Greenside, N. A. Sinnott-Armstrong, S. Vesuna, A. T. Satpathy, A. J. Rubin, K. S. Montine, B. Wu, A. Kathiria, S. W. Cho, M. R. Mumbach, A. C. Carter, M. Kasowski, L. A. Orloff, V. I. Risca, A. Kundaje, P. A. Khavari, T. J. Montine, W. J. Greenleaf, and H. Y. Chang, "An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues," *Nat. Methods*, vol. 14, pp. 959–962, Oct. 2017.

[48] R. J. Britten and D. E. Kohne, "Repeated sequences in DNA," *Science*, vol. 161, pp. 529–540, Aug. 1968.

[49] D. E. Kohne, S. A. Levison, and M. J. Byers, "Room temperature method for increasing the rate of DNA reassociation by many thousandfold: the phenol emulsion reassociation technique," *Biochemistry*, vol. 16, pp. 5329–5341, Nov. 1977.

[50] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, Feb. 2001.

[51] B. McClintock, "The origin and behavior of mutable loci in maize," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 36, pp. 344–355, June 1950.

[52] J. K. Pace, 2nd and C. Feschotte, "The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage," *Genome Res.*, vol. 17, pp. 422–432, Apr. 2007.

[53] C. R. Beck, J. L. Garcia-Perez, R. M. Badge, and J. V. Moran, "LINE-1 elements in structural variation and disease," *Annu. Rev. Genomics Hum. Genet.*, vol. 12, pp. 187–215, 2011.

[54] K. Mätlik, K. Redik, and M. Speek, "L1 antisense promoter drives tissue-specific transcription of human genes," *J. Biomed. Biotechnol.*, vol. 2006, no. 1, p. 71753, 2006.

[55] D. C. Hancks and H. H. Kazazian, Jr, “Roles for retrotransposon insertions in human disease,” *Mob. DNA*, vol. 7, p. 9, May 2016.

[56] M. Taniguchi-Ikeda, K. Kobayashi, M. Kanagawa, C.-C. Yu, K. Mori, T. Oda, A. Kuga, H. Kurahashi, H. O. Akman, S. DiMauro, R. Kaji, T. Yokota, S. Takeda, and T. Toda, “Pathogenic exon-trapping by SVA retrotransposon and rescue in fukuyama muscular dystrophy,” *Nature*, vol. 478, pp. 127–131, Oct. 2011.

[57] M. Taniguchi-Ikeda, K. Kobayashi, M. Kanagawa, Y. C-C, K. Mori, T. Oda, A. Kuga, H. Kurahashi, H. O. Akman, S. DiMauro, R. Kaji, T. Yokota, S. Takeda, and T. Toda, “An ancient retrotransposal insertion causes fukuyama-type congenital muscular dystrophy.” `https://www.nature.com/articles/28653.pdf`, July 1998.

[58] S. R. Richardson, S. Morell, and G. J. Faulkner, “L1 retrotransposons and somatic mosaicism in the brain,” *Annu. Rev. Genet.*, vol. 48, pp. 1–27, July 2014.

[59] J. Xing, D. J. Witherspoon, and L. B. Jorde, “Mobile element biology: new possibilities with high-throughput sequencing,” *Trends Genet.*, vol. 29, pp. 280–289, May 2013.

[60] H. Paulson, “Repeat expansion diseases,” *Handb. Clin. Neurol.*, vol. 147, pp. 105–123, 2018.

[61] P. Djian, “Evolution of simple repeats in DNA and their relation to human disease,” *Cell*, vol. 94, pp. 155–160, 1998.

[62] Z. M. Frenkel and E. N. Trifonov, “Origin and evolution of genes and genomes. crucial role of triplet expansions,” *J. Biomol. Struct. Dyn.*, vol. 30, no. 2, pp. 201–210, 2012.

[63] M. R. Santoro, S. M. Bray, and S. T. Warren, “Molecular mechanisms of fragile X syndrome: a twenty-year perspective,” *Annu. Rev. Pathol.*, vol. 7, pp. 219–245, 2012.

[64] Y.-H. Fu, D. P. A. Kuhl, M. Pieretti, J. S. Sutcliffe, t. S. Richards, A. J. M. Verkerk, J. J. A. Holden, G. Fenwick, S. T. Warren, Jr, . B. A. Oostra, D. L. Nelson, and C. Thomas Caskey’, “Variation in the CGG repeat at the fragile X site results in genetic instability: Resolution of the sherman paradox,” *Cell*, vol. 67, pp. 1047–1056, 1991.

[65] S. Jacquemont, R. J. Hagerman, M. A. Leehey, D. A. Hall, R. A. Levine, J. A. Brunberg, L. Zhang, T. Jardini, L. W. Gane, S. W. Harris, K. Herman, J. Grigsby, C. M. Greco, E. Berry-Kravis, F. Tassone, and P. J. Hagerman, “Penetrance of the fragile x-associated tremor/ataxia syndrome in a premutation carrier population,” *JAMA*, vol. 291, pp. 460–469, Jan. 2004.

[66] C. Depienne and J.-L. Mandel, “30 years of repeat expansion disorders: What have we learned and what are the remaining challenges?,” *Am. J. Hum. Genet.*, vol. 108, pp. 764–785, May 2021.

[67] D. Shinde, Y. Lai, F. Sun, and N. Arnheim, “Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites,” *Nucleic Acids Res.*, vol. 31, pp. 974–980, Feb. 2003.

[68] T. J. Treangen and S. L. Salzberg, “Repetitive DNA and next-generation sequencing: computational challenges and solutions,” *Nat. Rev. Genet.*, vol. 13, pp. 36–46, Nov. 2011.

[69] ENCODE Project Consortium, “An integrated encyclopedia of DNA elements in the human genome,” *Nature*, vol. 489, pp. 57–74, Sept. 2012.

[70] A. Lazarovici, T. Zhou, A. Shafer, A. C. Dantas Machado, T. R. Riley, R. Sandstrom, P. J. Sabo, Y. Lu, R. Rohs, J. A. Stamatoyannopoulos, and H. J. Bussemaker, “Probing DNA shape and methylation state on a genomic scale with DNase I,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, pp. 6376–6381, Apr. 2013.

[71] S. Neph, J. Vierstra, A. B. Stergachis, A. P. Reynolds, E. Haugen, B. Vernot, R. E. Thurman, S. John, R. Sandstrom, A. K. Johnson, M. T. Maurano, R. Humbert, E. Rynes, H. Wang, S. Vong, K. Lee, D. Bates, M. Diegel, V. Roach, D. Dunn, J. Neri, A. Schafer, R. S. Hansen, T. Kutyavin, E. Giste, M. Weaver, T. Canfield, P. Sabo, M. Zhang, G. Balasundaram, R. Byron, M. J. MacCoss, J. M. Akey, M. A. Bender, M. Groudine, R. Kaul, and J. A. Stamatoyannopoulos, "An expansive human regulatory lexicon encoded in transcription factor footprints," *Nature*, vol. 489, pp. 83–90, Sept. 2012.

[72] T. Hattori, J. M. Taft, K. M. Swist, H. Luo, H. Witt, M. Slattery, A. Koide, A. J. Ruthenburg, K. Krajewski, B. D. Strahl, K. P. White, P. J. Farnham, Y. Zhao, and S. Koide, "Recombinant antibodies to histone post-translational modifications," *Nat. Methods*, vol. 10, pp. 992–995, Oct. 2013.

[73] A. L. van de Ven, K. Adler-Storthz, and R. Richards-Kortum, "Delivery of optical contrast agents using Triton-X100, part 1: reversible permeabilization of live cells for intracellular labeling," *J. Biomed. Opt.*, vol. 14, p. 021012, Mar. 2009.

[74] D. Goldenberger, I. Perschil, M. Ritzler, and M. Altwegg, "A simple "universal" DNA extraction procedure using SDS and proteinase K is compatible with direct PCR amplification," *PCR Methods Appl.*, vol. 4, pp. 368–370, June 1995.

[75] R. R. B. Russell, "Use of triton X-100 to overcome the inhibition of fructosyltransferase by SDS," 1979.

[76] O. Kepp, L. Galluzzi, M. Lipinski, J. Yuan, and G. Kroemer, "Cell death assays for drug discovery," *Nat. Rev. Drug Discov.*, vol. 10, pp. 221–237, Mar. 2011.

[77] H. A. Cole, F. Cui, J. Ocampo, T. L. Burke, T. Nikitina, V. Nagarajavel, N. Kotomura, V. B. Zhurkin, and D. J. Clark, "Novel nucleosomal particles containing core histones and linker DNA but no histone H1," *Nucleic Acids Res.*, vol. 44, pp. 573–581, Jan. 2016.

[78] L. Baranello, F. Kouzine, S. Sanford, and D. Levens, "ChIP bias as a function of cross-linking time," *Chromosome Res.*, vol. 24, pp. 175–181, May 2016.

[79] A. Steube, T. Schenk, A. Tretyakov, and H. P. Saluz, "High-intensity UV laser ChIP-seq for the study of protein-DNA interactions in living cells," *Nat. Commun.*, vol. 8, p. 1303, Nov. 2017.

[80] P. L. Freddolino, H. M. Amemiya, T. J. Goss, and S. Tavazoie, "Dynamic landscape of protein occupancy across the escherichia coli chromosome," *PLoS Biol.*, vol. 19, p. e3001306, June 2021.

[81] A. F. Smit, "Interspersed repeats and other mementos of transposable elements in mammalian genomes," *Curr. Opin. Genet. Dev.*, vol. 9, pp. 657–663, Dec. 1999.

[82] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning,

T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, Feb. 2001.

[83] P. Deininger, "Alu elements: know the SINEs," *Genome Biol.*, vol. 12, p. 236, Dec. 2011.

[84] E. M. Ostertag, J. L. Goodier, Y. Zhang, and H. H. Kazazian, Jr, "SVA elements are nonautonomous retrotransposons that cause disease in humans," *Am. J. Hum. Genet.*, vol. 73, pp. 1444–1451, Dec. 2003.

[85] H. H. Kazazian, Jr and J. V. Moran, "Mobile DNA in health and disease," *N. Engl. J. Med.*, vol. 377, pp. 361–370, July 2017.

[86] B. Brouha, J. Schustak, R. M. Badge, S. Lutz-Prigge, A. H. Farley, J. V. Moran, and H. H. Kazazian, Jr, "Hot l1s account for the bulk of retrotransposition in the human population," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, pp. 5280–5285, Apr. 2003.

[87] D. M. Sassaman, B. A. Dombroski, J. V. Moran, M. L. Kimberland, T. P. Naas, R. J. DeBerardinis, A. Gabriel, G. D. Swergold, and H. H. Kazazian, Jr, "Many human L1 elements are capable of retrotransposition," *Nat. Genet.*, vol. 16, pp. 37–43, May 1997.

[88] C. R. Beck, P. Collier, C. Macfarlane, M. Malig, J. M. Kidd, E. E. Eichler, R. M. Badge, and J. V. Moran, "LINE-1 retrotransposition activity in human genomes," *Cell*, vol. 141, pp. 1159–1170, June 2010.

[89] E. C. Scott, E. J. Gardner, A. Masood, N. T. Chuang, P. M. Vertino, and S. E. Devine, "A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer," *Genome Res.*, vol. 26, pp. 745–755, June 2016.

[90] H. H. Kazazian, Jr, C. Wong, H. Youssoufian, A. F. Scott, D. G. Phillips, and S. E. Antonarakis, "Haemophilia a resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man," *Nature*, vol. 332, pp. 164–166, Mar. 1988.

[91] Y. Lubelsky and I. Ulitsky, "Sequences enriched in alu repeats drive nuclear localization of long RNAs in human cells," *Nature*, vol. 555, pp. 107–111, Mar. 2018.

[92] T. Aneichyk, W. T. Hendriks, R. Yadav, D. Shin, D. Gao, C. A. Vaine, R. L. Collins, A. Domingo, B. Currall, A. Stortchevoi, T. Multhaupt-Buell, E. B. Penney, L. Cruz, J. Dhakal, H. Brand, C. Hanscom, C. Antolik, M. Dy, A. Ragavendran, J. Underwood, S. Cantsilieris, K. M. Munson, E. E. Eichler, P. Acuña, C. Go, R. D. G. Jamora, R. L. Rosales, D. M. Church, S. R. Williams, S. Garcia, C. Klein, U. Müller, K. C. Wilhelmsen, H. T. M. Timmers, Y. Sapir, B. J. Wainger, D. Henderson, N. Ito, N. Weisenfeld, D. Jaffe, N. Sharma, X. O. Breakefield, L. J. Ozelius, D. C. Bragg, and M. E. Talkowski, "Dissecting the causal mechanism of X-Linked Dystonia-Parkinsonism by integrating genome and transcriptome assembly," *Cell*, vol. 172, pp. 897–909.e21, Feb. 2018.

[93] Y. Jourdy, A. Janin, M. Fretigny, A. Lienhart, C. Négrier, D. Bozon, and C. Vinciguerra, "Reccurrent F8 intronic deletion found in mild hemophilia a causes alu exonization," *Am. J. Hum. Genet.*, vol. 102, pp. 199–206, Feb. 2018.

[94] G. D. Evrony, E. Lee, P. J. Park, and C. A. Walsh, "Resolving rates of mutation in the brain using single-neuron genomics," *Elife*, vol. 5, Feb. 2016.

[95] K. R. Upton, D. J. Gerhardt, J. S. Jesuadian, S. R. Richardson, F. J. Sánchez-Luque, G. O. Bodea, A. D. Ewing, C. Salvador-Palomeque, M. S. van der Knaap, P. M. Brennan, A. Vanderver, and G. J. Faulkner, "Ubiquitous L1 mosaicism in hippocampal neurons," *Cell*, vol. 161, pp. 228–239, Apr. 2015.

[96] N. G. Coufal, J. L. Garcia-Perez, G. E. Peng, G. W. Yeo, Y. Mu, M. T. Lovci, M. Morell, K. S. O'Shea, J. V. Moran, and F. H. Gage, "L1 retrotransposition in human neural progenitor cells," *Nature*, vol. 460, pp. 1127–1131, Aug. 2009.

[97] A. R. Muotri, V. T. Chu, M. C. N. Marchetto, W. Deng, J. V. Moran, and F. H. Gage, "Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition," *Nature*, vol. 435, pp. 903–910, June 2005.

[98] X. Zhu, B. Zhou, R. Pattni, K. Gleason, C. Tan, A. Kalinowski, S. Sloan, A.-S. Fiston-Lavier, J. Mariani, A. Abyzov, D. Petrov, B. A. Barres, H. Vogel, J. V. Moran, F. M. Vaccarino, C. A. Tamminga, D. F. Levinson, A. E. Urban, and Brain Somatic Mosaicism Network, "Machine learning reveals bilateral distribution of somatic L1 insertions in human neurons and glia."

[99] M. J. McConnell, J. V. Moran, A. Abyzov, S. Akbarian, T. Bae, I. Cortes-Ciriano, J. A. Erwin, L. Fasching, D. A. Flasch, D. Freed, J. Ganz, A. E. Jaffe, K. Y. Kwan, M. Kwon, M. A. Lodato, R. E. Mills, A. C. M. Paquola, R. E. Rodin, C. Rosenbluh, N. Sestan, M. A. Sherman, J. H. Shin, S. Song, R. E. Straub, J. Thorpe, D. R. Weinberger, A. E. Urban, B. Zhou, F. H. Gage, T. Lehner, G. Senthil, C. A. Walsh, A. Chess, E. Courchesne, J. G. Gleeson, J. M. Kidd, P. J. Park, J. Pevsner, F. M. Vaccarino, and Brain Somatic Mosaicism Network, "Intersection of diverse neuronal genomes and neuropsychiatric disease: The brain somatic mosaicism network," *Science*, vol. 356, Apr. 2017.

[100] A. G. Diehl, N. Ouyang, and A. P. Boyle, "Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes," *Nat. Commun.*, vol. 11, p. 1796, Apr. 2020.

[101] Y. Zhang, T. Li, S. Preissl, M. L. Amaral, J. D. Grinstein, E. N. Farah, E. Destici, Y. Qiu, R. Hu, A. Y. Lee, S. Chee, K. Ma, Z. Ye, Q. Zhu, H. Huang, R. Fang, L. Yu, J. C. Izpisua Belmonte, J. Wu, S. M. Evans, N. C. Chi, and B. Ren, "Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells," *Nat. Genet.*, vol. 51, pp. 1380–1388, Sept. 2019.

[102] M. N. K. Choudhary, R. Z. Friedman, J. T. Wang, H. S. Jang, X. Zhuo, and T. Wang, "Co-opted transposons help perpetuate conserved higher-order chromosomal structures."

[103] R. C. Iskow, M. T. McCabe, R. E. Mills, S. Torene, W. S. Pittard, A. F. Neuwald, E. G. Van Meir, P. M. Vertino, and S. E. Devine, "Natural mutagenesis of human genomes by endogenous retrotransposons," *Cell*, vol. 141, pp. 1253–1261, June 2010.

[104] J. P. Steranka, Z. Tang, M. Grivainis, C. R. L. Huang, L. M. Payer, F. O. R. Rego, T. L. A. Miller, P. A. F. Galante, S. Ramaswami, A. Heguy, D. Fenyö, J. D. Boeke, and K. H. Burns, "Transposon insertion profiling by sequencing (TIPseq) for mapping LINE-1 insertions in the human genome," *Mob. DNA*, vol. 10, p. 8, Mar. 2019.

[105] C. R. L. Huang, A. M. Schneider, Y. Lu, T. Niranjan, P. Shen, M. A. Robinson, J. P. Steranka, D. Valle, C. I. Civin, T. Wang, S. J. Wheelan, H. Ji, J. D. Boeke, and K. H. Burns, "Mobile interspersed repeats are major structural variants in the human genome," *Cell*, vol. 141, pp. 1171–1182, June 2010.

[106] J. A. Erwin, A. C. M. Paquola, T. Singer, I. Gallina, M. Novotny, C. Quayle, T. A. Bedrosian, F. I. A. Alves, C. R. Butcher, J. R. Herdy, A. Sarkar, R. S. Lasken, A. R. Muotri, and F. H. Gage, "L1-associated genomic regions are deleted in somatic cells of the healthy human brain," 2016.

[107] E. J. Gardner, V. K. Lam, D. N. Harris, N. T. Chuang, E. C. Scott, W. S. Pittard, R. E. Mills, 1000 Genomes Project Consortium, and S. E. Devine, "The mobile element locator tool (MELT): population-scale mobile element discovery and biology," *Genome Res.*, vol. 27, pp. 1916–1929, Nov. 2017.

[108] E. M. Kvikstad, P. Piazza, J. C. Taylor, and G. Lunter, "A high throughput screen for active human transposable elements," *BMC Genomics*, vol. 19, p. 115, Feb. 2018.

[109] W. Zhou, S. B. Emery, D. A. Flasch, Y. Wang, K. Y. Kwan, J. M. Kidd, J. V. Moran, and R. E. Mills, "Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology," *Nucleic Acids Res.*, vol. 48, pp. 1146–1163, Feb. 2020.

[110] D. T. Thung, J. de Ligt, L. E. M. Vissers, M. Steehouwer, M. Kroon, P. de Vries, E. P. Slagboom, K. Ye, J. A. Veltman, and J. Y. Hehir-Kwa, "Mobster: accurate detection of mobile element insertions in next generation sequencing data," *Genome Biol.*, vol. 15, no. 10, p. 488, 2014.

[111] J. Wu, W.-P. Lee, A. Ward, J. A. Walker, M. K. Konkel, M. A. Batzer, and G. T. Marth, "Tangram: a comprehensive toolbox for mobile element insertion detection," *BMC Genomics*, vol. 15, p. 795, Sept. 2014.

[112] E. Lee, R. Iskow, L. Yang, O. Gokcumen, P. Haseley, L. J. Luquette, 3rd, J. G. Lohr, C. C. Harris, L. Ding, R. K. Wilson, D. A. Wheeler, R. A. Gibbs, R. Kucherlapati, C. Lee, P. V. Kharchenko, P. J. Park, and Cancer Genome Atlas Research Network, "Landscape of somatic retrotransposition in human cancers," *Science*, vol. 337, pp. 967–971, Aug. 2012.

[113] M. J. P. Chaisson, A. D. Sanders, X. Zhao, A. Malhotra, D. Porubsky, T. Rausch, E. J. Gardner, O. L. Rodriguez, L. Guo, R. L. Collins, and Others, "Multi-platform discovery of haplotype-resolved structural variation in human genomes. nat commun 10: 1784," 2019.

[114] J. M. Kidd, T. Graves, T. L. Newman, R. Fulton, H. S. Hayden, M. Malig, J. Kallicki, R. Kaul, R. K. Wilson, and E. E. Eichler, "A human genome structural variation sequencing resource reveals insights into mutational mechanisms," *Cell*, vol. 143, pp. 837–847, Nov. 2010.

[115] C. R. Beck, J. L. Garcia-Perez, R. M. Badge, and J. V. Moran, "LINE-1 elements in structural variation and disease," *Annu. Rev. Genomics Hum. Genet.*, vol. 12, pp. 187–215, 2011.

[116] G. J. Faulkner and J. L. Garcia-Perez, "L1 mosaicism in mammals: Extent, effects, and evolution," *Trends Genet.*, vol. 33, pp. 802–816, Nov. 2017.

[117] I. Ovchinnikov, A. B. Troxel, and G. D. Swergold, "Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion," *Genome Res.*, vol. 11, pp. 2050–2058, Dec. 2001.

[118] R. M. Badge, R. S. Alisch, and J. V. Moran, "ATLAS: a system to selectively identify human-specific L1 insertions," *Am. J. Hum. Genet.*, vol. 72, pp. 823–838, Apr. 2003.

[119] D. A. Flasch, Á. Macia, L. Sánchez, M. Ljungman, S. R. Heras, J. L. García-Pérez, T. E. Wilson, and J. V. Moran, "Genome-wide de novo L1 retrotransposition connects endonuclease activity with replication," *Cell*, vol. 177, pp. 837–851.e28, May 2019.

[120] H. Ha, J. W. Loh, and J. Xing, "Identification of polymorphic SVA retrotransposons using a mobile element scanning method for SVA (ME-Scan-SVA)," *Mob. DNA*, vol. 7, p. 15, July 2016.

[121] C.-S. Chin, D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, S. W. Turner, and J. Korlach, "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data," *Nat. Methods*, vol. 10, pp. 563–569, June 2013.

[122] M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O'Grady, H. E. Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan, T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, and M. Loose, "Nanopore sequencing and assembly of a human genome with ultra-long reads," *Nat. Biotechnol.*, vol. 36, pp. 338–345, Apr. 2018.

[123] J. M. Zook, N. F. Hansen, N. D. Olson, L. Chapman, J. C. Mullikin, C. Xiao, S. Sherry, S. Koren, A. M. Phillippy, P. C. Boutros, S. M. E. Sahraeian, V. Huang, A. Rouette, N. Alexander, C. E. Mason, I. Hajirasouliha, C. Ricketts, J. Lee, R. Tearle, I. T. Fiddes, A. M. Barrio, J. Wala, A. Carroll, N. Ghaffari, O. L. Rodriguez, A. Bashir, S. Jackman, J. J. Farrell, A. M. Wenger, C. Alkan, A. Soylev, M. C. Schatz, S. Garg, G. Church, T. Marschall, K. Chen, X. Fan, A. C. English, J. A. Rosenfeld, W. Zhou, R. E. Mills, J. M. Sage, J. R. Davis, M. D. Kaiser, J. S. Oliver, A. P. Catalano, M. J. P. Chaisson, N. Spies, F. J. Sedlazeck, and M. Salit, "A robust benchmark for detection of germline large deletions and insertions," *Nat. Biotechnol.*, vol. 38, pp. 1347–1355, Nov. 2020.

[124] C. Chu, B. Zhao, P. J. Park, and E. A. Lee, "Identification and genotyping of transposable element insertions from genome sequencing data," *Curr. Protoc. Hum. Genet.*, vol. 107, p. e102, Sept. 2020.

[125] A. D. Ewing, N. Smits, F. J. Sanchez-Luque, J. Faivre, P. M. Brennan, S. W. Cheetham, and G. J. Faulkner, "Nanopore sequencing enables comprehensive transposable element epigenomic profiling."

[126] P. Ebert, P. A. Audano, Q. Zhu, B. Rodriguez-Martin, D. Porubsky, M. J. Bonder, A. Sulovari, J. Ebler, W. Zhou, R. Serra Mari, F. Yilmaz, X. Zhao, P. Hsieh, J. Lee, S. Kumar, J. Lin, T. Rausch, Y. Chen, J. Ren, M. Santamarina, W. Höps, H. Ashraf, N. T. Chuang, X. Yang, K. M. Munson, A. P. Lewis, S. Fairley, L. J. Tallon, W. E. Clarke, A. O. Basile, M. Byrska-Bishop, A. Corvelo, U. S. Evani, T.-Y. Lu, M. J. P. Chaisson, J. Chen, C. Li, H. Brand, A. M. Wenger, M. Ghareghani, W. T. Harvey, B. Raeder, P. Hasenfeld, A. A. Regier, H. J. Abel, I. M. Hall, P. Flicek, O. Stegle, M. B. Gerstein, J. M. C. Tubio, Z. Mu, Y. I. Li, X. Shi, A. R. Hastie, K. Ye, Z. Chong, A. D. Sanders, M. C. Zody, M. E. Talkowski, R. E. Mills, S. E. Devine, C. Lee, J. O. Korbel, T. Marschall, and E. E. Eichler, "Haplotype-resolved diverse human genomes and integrated analysis of structural variation," *Science*, vol. 372, Apr. 2021.

[127] T. Gilpatrick, I. Lee, J. E. Graham, E. Raimondeau, R. Bowen, A. Heron, B. Downs, S. Sukumar, F. J. Sedlazeck, and W. Timp, "Targeted nanopore sequencing with cas9-guided adapter ligation," *Nat. Biotechnol.*, vol. 38, pp. 433–438, Apr. 2020.

[128] H. Wang, J. Xing, D. Grover, D. J. Hedges, K. Han, J. A. Walker, and M. A. Batzer, "SVA elements: a hominid-specific retroposon family," *J. Mol. Biol.*, vol. 354, pp. 994–1007, Dec. 2005.

[129] E. A. Bennett, H. Keller, R. E. Mills, S. Schmidt, J. V. Moran, O. Weichenrieder, and S. E. Devine, "Active alu retrotransposons in the human genome," *Genome Res.*, vol. 18, pp. 1875–1883, Dec. 2008.

[130] S. Boissinot, P. Chevret, and A. V. Furano, "L1 (LINE-1) retrotransposon evolution and amplification in recent human history," *Mol. Biol. Evol.*, vol. 17, pp. 915–928, June 2000.

[131] T. Karamitros and G. Magiorkinis, "Multiplexed targeted sequencing for oxford nanopore MinION: A detailed library preparation procedure," *Methods Mol. Biol.*, vol. 1712, pp. 43–51, 2018.

[132] T. Gabrieli, H. Sharim, D. Fridman, N. Arbib, Y. Michaeli, and Y. Ebenstein, "Selective nanopore sequencing of human BRCA1 by cas9-assisted targeting of chromosome segments (CATCH)," *Nucleic Acids Res.*, vol. 46, p. e87, Aug. 2018.

[133] P. Giesselmann, B. Brändl, E. Raimondeau, R. Bowen, C. Rohrandt, R. Tandon, H. Kretzmer, G. Assum, C. Galonska, R. Siebert, O. Ammerpohl, A. Heron, S. A. Schneider, J. Ladewig, P. Koch, B. M. Schuldt, J. E. Graham, A. Meissner, and F.-J. Müller, "Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing," *Nat. Biotechnol.*, vol. 37, pp. 1478–1481, Dec. 2019.

[134] J. Dausset, H. Cann, D. Cohen, M. Lathrop, J. M. Lalouel, and R. White, "Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome," *Genomics*, vol. 6, pp. 575–577, Mar. 1990.

[135] International HapMap Consortium, "The international HapMap project," *Nature*, vol. 426, pp. 789–796, Dec. 2003.

[136] R. E. Mills, K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kural, H. Y. K. Lam, J. Leng, R. Li, Y. Li, C.-Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Stütz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. Ye, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll, J. O. Korbel, and 1000 Genomes Project, "Mapping copy number variation by population-scale genome sequencing," *Nature*, vol. 470, pp. 59–65, Feb. 2011.

[137] T. . G. P. Consortium and The 1000 Genomes Project Consortium, "A global reference for human genetic variation," 2015.

[138] J. M. Zook, D. Catoe, J. McDaniel, L. Vang, N. Spies, A. Sidow, Z. Weng, Y. Liu, C. E. Mason, N. Alexander, E. Henaff, A. B. R. McIntyre, D. Chandramohan, F. Chen, E. Jaeger, A. Moshrefi, K. Pham, W. Stedman, T. Liang, M. Saghbini, Z. Dzakula, A. Hastie, H. Cao, G. Deikus, E. Schadt, R. Sebra, A. Bashir, R. M. Truty, C. C. Chang, N. Gulbahce, K. Zhao, S. Ghosh, F. Hyland, Y. Fu, M. Chaisson, C. Xiao, J. Trow, S. T. Sherry, A. W. Zaranek, M. Ball, J. Bobe, P. Estep, G. M. Church, P. Marks, S. Kyriazopoulou-Panagiotopoulou, G. X. Y. Zheng, M. Schnall-Levin, H. S. Ordonez, P. A. Mudivarti, K. Giorda, Y. Sheng, K. B. Rypdal, and M. Salit, "Extensive sequencing of seven human genomes to characterize benchmark reference materials," *Sci Data*, vol. 3, p. 160025, June 2016.

[139] J. M. Zook, B. Chapman, J. Wang, D. Mittelman, O. Hofmann, W. Hide, and M. Salit, "Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls," *Nat. Biotechnol.*, vol. 32, pp. 246–251, Mar. 2014.

[140] P. A. Audano, A. Sulovari, T. A. Graves-Lindsay, S. Cantsilieris, M. Sorensen, A. E. Welch, M. L. Dougherty, B. J. Nelson, A. Shah, S. K. Dutcher, W. C. Warren, V. Magrini, S. D. McGrath, Y. I. Li, R. K. Wilson, and E. E. Eichler, "Characterizing the major structural variant alleles of the human genome," *Cell*, vol. 176, pp. 663–675.e19, Jan. 2019.

[141] R. E. Mills, E. A. Bennett, R. C. Iskow, and S. E. Devine, "Which transposable elements are active in the human genome?," *Trends Genet.*, vol. 23, pp. 183–191, Apr. 2007.

[142] J. V. Moran, S. E. Holmes, T. P. Naas, R. J. DeBerardinis, J. D. Boeke, and H. H. Kazazian, Jr, "High frequency retrotransposition in cultured mammalian cells," *Cell*, vol. 87, pp. 917–927, Nov. 1996.

[143] F. A. Ran, P. D. Hsu, J. Wright, V. Agarwala, D. A. Scott, and F. Zhang, "Genome engineering using the CRISPR-Cas9 system," *Nat. Protoc.*, vol. 8, pp. 2281–2308, Nov. 2013.

[144] S. H. Sternberg, S. Redding, M. Jinek, E. C. Greene, and J. A. Doudna, "DNA interrogation by the CRISPR RNA-guided endonuclease cas9," *Nature*, vol. 507, pp. 62–67, Mar. 2014.

[145] L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang, "Multiplex genome engineering using CRISPR/Cas systems," *Science*, vol. 339, pp. 819–823, Feb. 2013.

[146] P. D. Hsu, D. A. Scott, J. A. Weinstein, F. A. Ran, S. Konermann, V. Agarwala, Y. Li, E. J. Fine, X. Wu, O. Shalem, T. J. Cradick, L. A. Marraffini, G. Bao, and F. Zhang, "DNA targeting specificity of RNA-guided cas9 nucleases," *Nat. Biotechnol.*, vol. 31, pp. 827–832, Sept. 2013.

[147] A. F. A. Smit, R. Hubley, and P. Green, "2015 RepeatMasker open-4.0," 2013.

[148] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. P. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. K. Lam, X. J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E.-W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, 1000 Genomes Project Consortium, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, and J. O. Korbel, "An integrated map of structural variation in 2,504 human genomes," *Nature*, vol. 526, pp. 75–81, Oct. 2015.

[149] X. Zhao, A. M. Weber, and R. E. Mills, "A recurrence-based approach for validating structural variation using long-read sequencing technology," *Gigascience*, vol. 6, pp. 1–9, Aug. 2017.

[150] J. Y. Hehir-Kwa, T. Marschall, W. P. Kloosterman, L. C. Francioli, J. A. Baaijens, L. J. Dijkstra, A. Abdellaoui, V. Koval, D. T. Thung, R. Wardenaar, I. Renkens, B. P. Coe, P. Deelen, J. de Ligt, E.-W. Lameijer, F. van Dijk, F. Hormozdiari, Genome of the Netherlands Consortium, A. G. Uitterlinden, C. M. van Duijn, E. E. Eichler, P. I. W. de Bakker, M. A. Swertz, C. Wijmenga, G.-J. B. van Ommen, P. E. Slagboom, D. I. Boomsma, A. Schönhuth, K. Ye, and V. Guryev, "A high-quality human reference panel reveals the complexity and distribution of genomic structural variants," *Nat. Commun.*, vol. 7, p. 12989, Oct. 2016.

[151] C. Stangl, S. de Blank, I. Renkens, L. Westera, T. Verbeek, J. E. Valle-Inclan, R. C. González, A. G. Henssen, M. J. van Roosmalen, R. W. Stam, E. E. Voest, W. P. Kloosterman, G. van Haaften, and G. R. Monroe, "Partner independent fusion gene detection by multiplexed CRISPR-Cas9 enrichment and long read nanopore sequencing," *Nat. Commun.*, vol. 11, p. 2861, June 2020.

[152] F. J. Sanchez-Luque, M.-J. H. C. Kempen, P. Gerdes, D. B. Vargas-Landin, S. R. Richardson, R.-L. Troskie, J. S. Jesuadian, S. W. Cheetham, P. E. Carreira, C. Salvador-Palomeque, M. García-Cañadas, M. Muñoz-Lopez, L. Sanchez, M. Lundberg, A. Macia, S. R. Heras, P. M. Brennan, R. Lister, J. L. Garcia-Perez, A. D. Ewing, and G. J. Faulkner, "LINE-1 evasion of epigenetic repression in humans," *Mol. Cell*, vol. 75, pp. 590–604.e12, Aug. 2019.

[153] G. D. Evrony, E. Lee, B. K. Mehta, Y. Benjamini, R. M. Johnson, X. Cai, L. Yang, P. Haseley, H. S. Lehmann, P. J. Park, and C. A. Walsh, "Cell lineage analysis in human brain using endogenous retroelements," *Neuron*, vol. 85, pp. 49–59, Jan. 2015.

[154] Y. Niu, X. Teng, Y. Shi, Y. Li, Y. Tang, P. Zhang, H. Luo, and others, "Genome-wide analysis of mobile element insertions in human genomes," *bioRxiv*, 2021.

[155] K. H. Miga, S. Koren, A. Rhie, M. R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G. A. Logsdon, V. A. Schneider, T. Potapova, J. Wood, W. Chow, J. Armstrong, J. Fredrickson, E. Pak, K. Tigyi, M. Kremitzki, C. Markovic, V. Maduro, A. Dutra, G. G. Bouffard, A. M. Chang, N. F. Hansen, A. B. Wilfert, F. Thibaud-Nissen, A. D. Schmitt, J.-M. Belton, S. Selvaraj, M. Y. Dennis, D. C. Soto, R. Sahasrabudhe, G. Kaya, J. Quick, N. J. Loman, N. Holmes, M. Loose, U. Surti, R. A. Risques, T. A. Graves Lindsay, R. Fulton, I. Hall, B. Paten, K. Howe, W. Timp, A. Young, J. C. Mullikin, P. A. Pevzner, J. L. Gerton, B. A. Sullivan, E. E. Eichler, and A. M. Phillippy, "Telomere-to-telomere assembly of a complete human X chromosome," *Nature*, vol. 585, pp. 79–84, Sept. 2020.

[156] J. L. Bennetzen, "Transposable element contributions to plant gene and genome evolution," *Plant Mol. Biol.*, vol. 42, pp. 251–269, Jan. 2000.

[157] T. Yu, X. Huang, S. Dou, X. Tang, S. Luo, W. E. Theurkauf, J. Lu, and Z. Weng, "A benchmark and an algorithm for detecting germline transposon insertions and measuring de novo transposon insertion frequencies," *Nucleic Acids Res.*, Jan. 2021.

[158] S. A. Miller, D. D. Dykes, and H. F. Polesky, "A simple salting out procedure for extracting DNA from human nucleated cells," *Nucleic Acids Res.*, vol. 16, p. 1215, Feb. 1988.

[159] J. Jurka, "Repeats in genomic DNA: mining and meaning," *Curr. Opin. Struct. Biol.*, vol. 8, pp. 333–337, June 1998.

[160] G. Marçais and C. Kingsford, "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers," *Bioinformatics*, vol. 27, pp. 764–770, Mar. 2011.

[161] R. R. Wick, L. M. Judd, C. L. Gorrie, and K. E. Holt, "Completing bacterial genome assemblies with multiplex MinION sequencing," *Microb Genom*, vol. 3, p. e000132, Oct. 2017.

[162] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon, "Biopython: freely available python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, pp. 1422–1423, June 2009.

[163] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller, "A greedy algorithm for aligning DNA sequences," *J. Comput. Biol.*, vol. 7, pp. 203–214, Feb. 2000.

[164] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403–410, Oct. 1990.

[165] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, pp. 3094–3100, Sept. 2018.

[166] M. Byrska-Bishop, U. S. Evani, X. Zhao, A. O. Basile, H. J. Abel, A. A. Regier, A. Corvelo, W. E. Clarke, R. Musunuri, K. Nagulapalli, S. Fairley, A. Runnels, L. Winterkorn, E. Lowy-Gallego, P. Flicek, S. Germer, H. Brand, I. M. Hall, M. E. Talkowski, G. Narzisi, M. C. Zody,

and The Human Genome Structural Variation Consortium, "High coverage whole genome sequencing of the expanded 1000 genomes project cohort including 602 trios."

[167] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy, "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation," *Genome Res.*, vol. 27, pp. 722–736, May 2017.

[168] X. Huang and A. Madan, "CAP3: A DNA sequence assembly program," *Genome Res.*, vol. 9, pp. 868–877, Sept. 1999.

[169] J. T. Simpson, R. E. Workman, P. C. Zuzarte, M. David, L. J. Dursi, and W. Timp, "Detecting DNA cytosine methylation using nanopore sequencing," *Nat. Methods*, vol. 14, pp. 407–410, Apr. 2017.

[170] W. a. Zhou, C. Castro, and C. Mumm, "Boyle-Lab/NanoPal-and-Cas9-targeted-enrichment-pipelines: First release of NanoPal and cas9 targeted enrichment pipelines," 2021.

[171] W. a. Zhou, H. Guan, and R. Mills, "mills-lab/PALMER: Release version for cas9 targeted enrichment pipelines," May 2021.

[172] A. Entezam, A. R. Lokanga, W. Le, G. Hoffman, and K. Usdin, "Potassium bromate, a potent DNA oxidizing agent, exacerbates germline repeat expansion in a fragile X premutation mouse model," *Hum. Mutat.*, vol. 31, pp. 611–616, May 2010.

[173] C. T. McMurray, "Mechanisms of trinucleotide repeat instability during human development," *Nat. Rev. Genet.*, vol. 11, pp. 786–799, Nov. 2010.

[174] N. Fouché, S. Ozgür, D. Roy, and J. D. Griffith, "Replication fork regression in repetitive DNAs," *Nucleic Acids Res.*, vol. 34, pp. 6044–6050, Oct. 2006.

[175] Z. Yang, R. Lau, J. L. Marcadier, D. Chitayat, and C. E. Pearson, "Replication inhibitors modulate instability of an expanded trinucleotide repeat at the myotonic dystrophy type 1 disease locus in human cells," *Am. J. Hum. Genet.*, vol. 73, pp. 1092–1105, Nov. 2003.

[176] T. Gilpatrick, I. Lee, J. E. Graham, E. Raimondeau, R. Bowen, A. Heron, B. Downs, S. Sukumar, F. J. Sedlazeck, and W. Timp, "Targeted nanopore sequencing with cas9-guided adapter ligation," *Nat. Biotechnol.*, vol. 38, pp. 433–438, Apr. 2020.

[177] J. M. Haenfler, G. Skariah, C. M. Rodriguez, A. Monteiro da Rocha, J. M. Parent, G. D. Smith, and P. K. Todd, "Targeted reactivation of FMR1 transcription in fragile X syndrome embryonic stem cells," *Front. Mol. Neurosci.*, vol. 11, p. 282, Aug. 2018.

[178] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, "A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity," *Science*, vol. 337, pp. 816–821, Aug. 2012.

[179] L. Tang, "PAM-less is more," *Nat. Methods*, vol. 17, p. 559, June 2020.

[180] D. Wang, C. Zhang, B. Wang, B. Li, Q. Wang, D. Liu, H. Wang, Y. Zhou, L. Shi, F. Lan, and Y. Wang, "Optimized CRISPR guide RNA design for two high-fidelity cas9 variants by deep learning," *Nat. Commun.*, vol. 10, p. 4284, Sept. 2019.

[181] H. S. Rhee and B. F. Pugh, "ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy," *Curr. Protoc. Mol. Biol.*, vol. Chapter 21, p. Unit 21.24, Oct. 2012.

[182] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Chromosomal DNA and Its Packaging in the Chromatin Fiber.* Garland Science, 2002.

[183] B. Zhou, S. S. Ho, S. U. Greer, X. Zhu, J. M. Bell, J. G. Arthur, N. Spies, X. Zhang, S. Byeon, R. Pattni, N. Ben-Efraim, M. S. Haney, R. R. Haraksingh, G. Song, H. P. Ji, D. Perrin, W. H. Wong, A. Abyzov, and A. E. Urban, "Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562," *Genome Res.*, vol. 29, pp. 472–484, Mar. 2019.

[184] J. Lee, H. Lim, H. Jang, B. Hwang, J. H. Lee, J. Cho, J. H. Lee, and D. Bang, "CRISPR-Cap: multiplexed double-stranded DNA enrichment based on the CRISPR system," *Nucleic Acids Res.*, vol. 47, p. e1, Jan. 2019.