## Title

Towards single-cell chromosome-specific single-base measurement of telomeres with nanopores

## Permalink

https://escholarship.org/uc/item/5ht768dw

## Author

Luong, Norman

## Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

# Towards single-cell chromosome-specific single-base measurement of telomeres with nanopores

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioengineering

by

Norman Luong

Committee in charge:

Professor Xiaohua Huang, Chair
Professor Gert Cauwenberghs
Professor Jan Karlseder
Professor Jon Pokorski
Professor Kun Zhang

2022

The Dissertation of Norman Luong is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

# DEDICATION

For my family

To my parents – Thank you for making me who I am

To my brother – Let this be proof that we are more capable than we could imagine

To my wife – Thank you for tagging along with me on this journey

To my daughter – May I be a model for reasoning and critical thinking

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

I would like to thank Professor Xiaohua Huang for providing the environment in which I could grow and be challenged as a graduate student. His forward thinking proposals for solving cutting edge problems and daily enthusiasm was truly inspiring. The creative freedom provided and diversity of disciplines involved with this project was invaluable to my training as a scientist and engineer.

I would like to thank my other committee members, Professor Jan Karlseder, Professor Gert Cauwenberghs, Professor Jon Pokorski, and Professor Kun Zhang, for their time and advice.

I am grateful for the other members of the Huang lab. Dr. Eric Chu trained me on photolithography, which gave me the confidence to later develop the process for solid state nanopore fabrication. His training on microfluidic fabrication and usage also allowed me to design my own devices for the applications described herein. I would like to thank Sylvia Liang for her expertise and work on lipid bilayer experiments, and Parker Dow for sharing his experience regarding molecular cloning techniques.

I would like to thank my wife, Dr. Lian Elizabeth Wong, for providing a supportive home environment during times of difficulty.

Chapter 2 contains unpublished material co-authored with Huang, Xiaohua. The dissertation author was the primary author of this chapter.

Chapter 3 contains unpublished material co-authored with Huang, Xiaohua. The dissertation author was the primary author of this chapter.

Chapter 4 contains unpublished material co-authored with Huang, Xiaohua. The dissertation author was the primary author of this chapter.

# VITA

| | |
|---|---|
| 2013 | Bachelor of Science in Bioengineering, University of California Merced |
| 2014–2016 | Scientist I/II, Roche Molecular Systems, Pleasanton, CA |
| 2019 | Master of Science in Bioengineering, University of California San Diego |
| 2022 | Doctor of Philosophy in Bioengineering, University of California San Diego |

# ABSTRACT OF THE DISSERTATION

Towards single-cell chromosome-specific single-base measurement of telomeres with nanopores

by

Norman Luong

Doctor of Philosophy in Bioengineering

University of California San Diego, 2022

Professor Xiaohua Huang, Chair

Telomeres protect the ends of chromosomes from DNA repair processes. Somatic cells reach senescence as a protective mechanism when telomeres become critically short. Under certain conditions, a small subset of cells can continue dividing to the point where telomeres are no longer adequately protected, leading to chromosome instability or crisis, at which point the cell is fated to either apoptosis or carcinogenesis. Senescence can be triggered by as few as a single telomere if it is sufficiently short. Chromosomes have been shown to have heritable telomere lengths and telomere length regulatory factors. Studies on the influence and dynamics of these factors provide insight that is limited by the resolution of tools currently available. Single-cell chromosome-specific techniques are time-consuming, while scalable single-cell methods can only probe the telomere length averaged across all chromosomes. Nanopores have the capability to measure single-molecule telomere lengths with high throughput, and we have developed methods and tools to bring us closer to realizing their potential for this application.

# 1 INTRODUCTION

## 1.1 Telomere structure and function

The distal regions of linear chromosomes are comprised of telomeres which associate with telomere-binding proteins to protect the ends of chromosomes against DNA damage response (DDR), non-homologous end-joining (NHEJ), and homologous recombination (HR) (*1*). In humans, the telomere sequence consists of a repeating double stranded DNA (dsDNA) telomeric repeat (TR) that reads 5'-TTAGGG-3'. The double stranded region generally spans a few kb to several tens of kb terminates with a single stranded DNA (ssDNA) TR overhang of 130-210 bases on average (*2*–*4*).

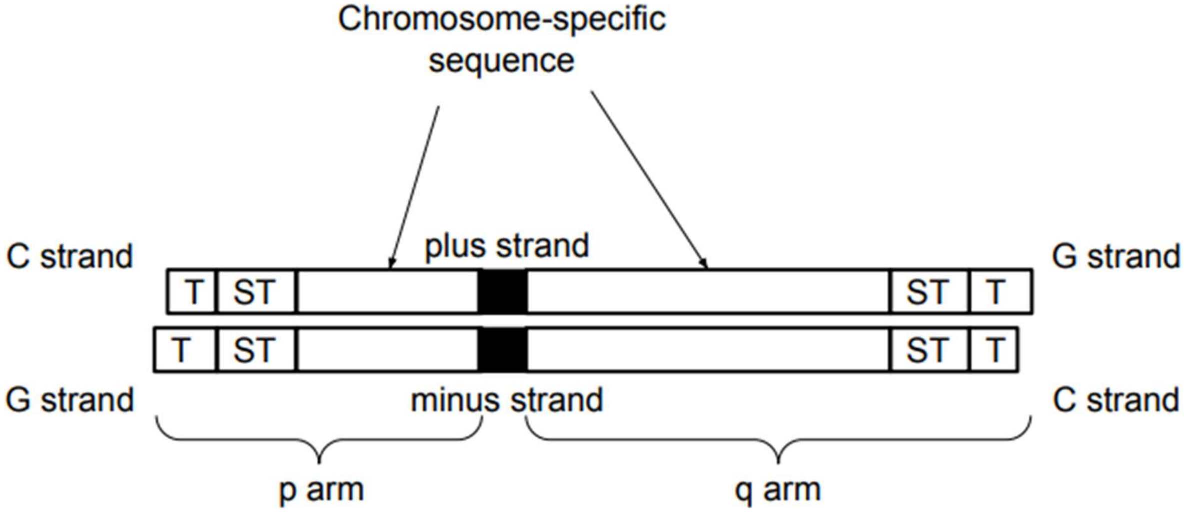


Figure 1.1: Diagram of chromosome layout. T: telomere, ST: subtelomere.

A six-protein complex known as shelterin coats the telomeric DNA through its DNA recognition proteins TRF1 (Telomere Repeat-binding Factor 1), TRF2 (Telomere Repeat-binding Factor 2), and POT1 (Protection of Telomeres 1) (*5*–*7*). TRF1 and TRF2 are specific to double stranded TRs through their Myb domains, while POT1 binds to the single stranded region. These will associate with the other three shelterin proteins, RAP1, TPP1, and TIN1, which together aid in telomere protection. Shelterin remodels the telomeres by facilitating displacement (D-loop) of a portion of the double stranded region by the single stranded overhang (Fig 1.2). This forms a telomere loop (T-loop) structure which protects chromosome ends by preventing misrecognition as a double-stranded break (DSB) by DNA repair mechanisms. Loss of shelterin leads to loss of the T-loop and can trigger DNA repair mechanisms such as DNA Damage Response (DDR), Homologous Recombination (HR), and Non-homologous End Joining (NHEJ) (*1*).

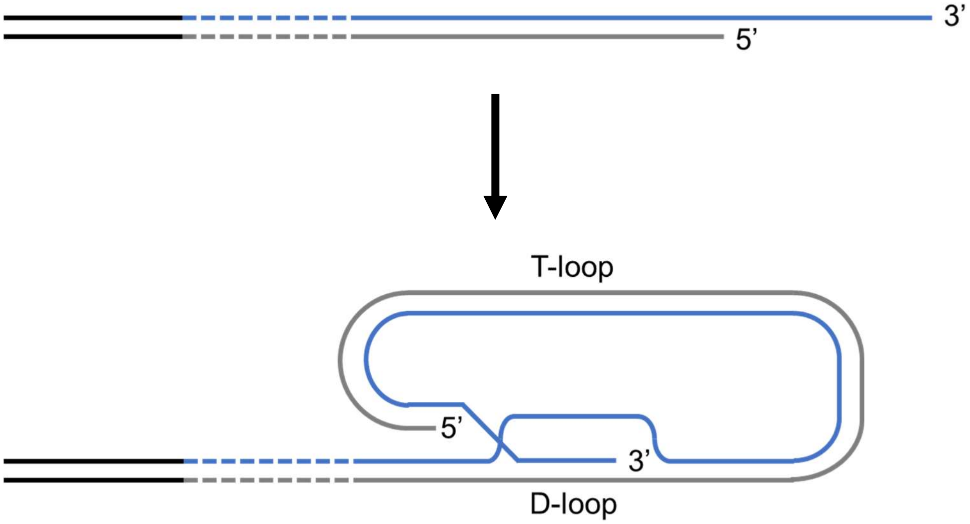


Figure 1.2: T-loop and D-loop formation in a telomere.

### 1.2 Telomere length dynamics and cancer

The T-loop unwinds during the Synthesis (S) and Gap 2 (G2) phases to allow telomere elongation with telomerase (*1*). Telomerase recruitment is inhibited for longer telomeres so it preferentially binds and elongates shorter telomeres (*8*). Intramolecular G-quadruplexes, which can be formed by the telomere overhang, are also implicated in the recruitment of telomerase (*9*). Upon binding, telomerase adds a single telomere repeat by reverse transcription using an RNA template known as telomerase RNA component (TERC). Progressive extension of telomeres requires multiple binding and extension cycles with telomerase. The catalytic domain, TERT (Telomerase Reverse Transcriptase) is only expressed in embryonic stem cells and in some somatic cell types, including male sperm cells (*10*), epidermal cells (*11*), lymphocytes (*12*, *13*), and some adult stem cells (*14*).

The replication of the majority of the telomere is done by replication forks that originate from the subtelomeres (*15*–*17*). As replisomes progress through the telomeres, they depend on several pathways which help overcome replicative challenges associated with G-quadruplex (G4) formation in G-rich ssDNA strands exposed by the replisome (*18*). Even if these pathways are functioning properly, telomeres in somatic cells shorten by 50-100 bp per population doubling due to the nature in which conventional DNA replication synthesizes lagging strands (*19*). This is known as the end-replication problem (*20*). Telomere shortening can also be caused by rapid, stochastic shortening events due to replication-fork collapse, t-loop excision or oxidative stress (*8*, *21*). Shortened telomeres recruit less shelterin and are unable to protect chromosome ends from activation of DDR. Once this occurs, p53 binding protein 1 (53BP1) and several other proteins are recruited to the telomere in what is called a telomere dysfunction-induced foci (TIF) (*21*). Accumulation of at least five dysfunctional telomeres leads to activation of the p53 tumor suppressor pathway and cell senescence (*2*, *22*). If tumor suppressor p53 and retinoblastoma protein (Rb) are mutated, cells can continue dividing until telomeres become critically short and either enter crisis or circumvent telomere shortening to become tumorigenic (*21*).

In crisis, telomeres are too short to bind shelterin proteins and undergo fusion through non-homologous end joining (NHEJ) (*21*). Failure to segregate chromosomes causes mitotic arrest, leading to amplification of the DNA-damage response and subsequent cell death (*23*, *24*). However, a small subset of cells can escape crisis by telomerase reactivation or alternative telomere-maintenance pathways (*21*, *25*). Such cells have genome instability and this is evident in the presence of tumors with sub-tetraploid karyotypes (*26*). In approximately 80-90% of human cancers, telomerase is upregulated which allows continual division (*1*, *27*). Other cancer types that do not rely on telomerase may instead use a homologous recombination-based pathway called alternative lengthening of telomeres (ALT) (*28*).

Short telomeres represent risk factors for tumor development because they can lead to structural remodeling and subsequent structural aberrations such as loss of heterozygosity and chromosomal deletions. Loss of heterozygosity (LOH) and chromosomal deletions in 11q, 13q, 17p, and 22q are associated with chronic lymphocytic leukemia (CLL) (*29*), which is the most common form of leukemia in adults (*30*). Only in recent years has LOH shown to be a common biomarker for a wide set of tumor types (*31*). Chromosome 17p in particular was found to have shorter telomere lengths, which could lead to chromosome instability (*32*). 17p is known to contain *p53 (33)*, and 17p deletion is associated with multiple cancers and their treatment outcomes (*34*, *35*).

The heritability and dynamics behind 17p telomere length are not yet fully understood. Individuals display different distributions of telomere length (*36*), and this can be passed down through generations (*37–40*). Homologous chromosomes can also display different telomere lengths, which stay consistent over multiple population doublings (*41*). The heritability of TL has so far only been studied at the cohort level with qPCR-based TL measurement (*40*). Although highly scalable, a major limitation to this approach is that it measures the average TL across multiple cells (leukocytes from whole blood) and across all chromosomes. As a result, we can only ascertain the strength of correlation in TL between familial pairs (e.g. sibling-sibling, mother-offspring, father-offspring). Perhaps due to this limitation, such studies provide conflicting reports as to whether maternal or paternal TL is more correlated with offspring TL (*37*, *39*, *42*). The influence of specific genes on telomere length has so far only been studied by pairing qPCR TL with microarray-based SNP analysis (*43*). It is not clear what mechanisms drive specific chromosome arms, such as 17p, to have shorter telomeres. Several short telomere syndromes (STS), which affect whole-cell telomere length, have been identified and linked to telomerase-specific and shelterin-specific genes (*44*, *45*). Only recently had subtelomere-specific factors been implicated in chromosome-specific telomere length regulation (*46*, *47*). To better understand how telomere lengths are regulated at the chromosome-specific level and how this can factor into tumorigenesis, there is a need for a single cell chromosome-specific telomere length measurement.

Telomere composition can vary, leading to a possibility of altered functionality. Telomeres can contain degenerate or variant telomere repeats of the form (TGAGGG), (TTGGGG), or (TCAGGG). These variant repeats are present in different proportions across several different cancer cell lines as well as across different chromosomes within the same cell line (*48*). The frequency and distribution of these variants across telomeres can affect binding of TRF1, TRF2, and other DNA binding proteins. This in turn can affect shelterin recruitment, and could also provide paths for escapement from crisis by ALT. However, the impact and pathways by which this can occur are not yet well understood (*48*). Aside from variant repeats, telomere composition can also be altered by oxidative damage. Guanine (G) is particularly susceptible to oxidation to form 8-oxo-7,8-dihydroguanine (OG). Furthermore, formation of the G-quadruplex enhances oxidation rate compared to duplex DNA due to increased accessibility of G (*49*). Presence of OG disrupts recognition by TRF1 and TRF2 (*50*), which in turn can lead to loss of shelterin and premature senescence.

### 1.3 Current methods for telomere length measurement

Single cell TL measurement enables the study of TL heterogeneity on a cell-by-cell basis, an issue of fundamental importance for studies on aging and carcinogenesis. Several methods exist for TL measurement and new methods are reported nearly every year. The “gold standard” method is terminal restriction fragmentation (TRF).

In TRF, genomic DNA is digested with a cocktail of restriction enzymes that leave telomeres uncut. The digest is then resolved by size on a gel and telomere-containing fragments are revealed by hybridizing with radiolabeled oligonucleotides (*19*). This method is amplification-free, but requires micrograms of input DNA. The resulting Southern Blot therefore gives a distribution of all telomere lengths across a large number of extracted cells.

FISH-based methods rely on a fluorescent $(CCCTAA)_3$ peptide nucleic acid (PNA) probe to hybridize with the telomeric repeats. PNA is used instead of DNA due to its higher hybridization affinity. In Flow-FISH, flow cytometry is used to measure the fluorescence signal from individual cells. This can then be used to estimate the average telomere length for each cell. In metaphase quantitative FISH, cells are arrested during metaphase and stained using the PNA probe. With enough imaging resolution, chromosomes can be identified based on their relative shape and size. This provides single cell chromosome specific TL, but is labor intensive and has limited length resolution.

PCR based methods are well suited for cohort studies, but also suffer from low accuracy. In general, telomeres and a reference, either a short interspersed nuclear element or a single copy gene, are both amplified. The ratio of telomere to reference is then compared with samples with a known TL to estimate the average TL of the sample. Like FISH, this does not provide an absolute TL since it relies on a reference, but it can be extended to single cells. Universal Single Telomere Length Analysis (U-STELA) can detect TL from each chromosome, but is not efficient in detecting TL over 8 kb (*51*). Whole genome sequencing can be used to measure TL, but only provide average TL and does not correlate well with TRF (*52*). Telomere Shortest Length Assay (TeSLA) is able to measure TL from <1 kb to 18 kb, but requires multiple ligation steps and multiple PCRs per sample to obtain a reliable result. It is also not a single-cell method. Currently, single cell TL measurement is PCR based and gives an average TL across all chromosomes for each cell (*53*).

Single telomere absolute-length rapid (STAR) assay uses digital PCR to measure telomere lengths of individual fragments from samples <1 ng (*54*). This provides a telomere length distribution or profile, but no information about which length corresponds to which chromosome arm is recovered.

## 1.4 Applicable technologies not yet implemented for telomere analysis

Poly(dimethylsiloxane) (PDMS) based microfluidic systems are ubiquitous and provide the means for single cell manipulation. Pneumatic valves enable precise nanoliter-scale delivery of reagents and controlled cell lysis (*55*). The micrometer dimensions are on the same size scale as a single cell and minimize sample dilution (*56*). Microfluidics are well-suited for amplification, as use of sub-microliter volumes of reagents and the potential for parallelization present significant advantages over tube-based formats.

Copolymers of polyethylene glycol and polyacrylamide are semi-permeable and allow passage of ions but not larger macromolecules (*57*). These polymer barriers can be formed in microfluidic devices and can be used for capture and immobilization of cells and charged analytes such as genomic DNA. Immobilization facilitates lysis, denaturation, and rinsing without loss of analytes, which is critical for single cell work.

Solid state nanopores have proved robustness for DNA length measurements and can be fabricated at scale. Protein nanopores can deliver single base resolution, but are usually embedded in lipid bilayers which are inherently unstable. Commercial nanopore solutions exist (e.g. Oxford Nanopore Technologies), and others have demonstrated embedding of protein nanopores into solid state pores to form a hybrid nanopore, which can increase stability (*58*).

A handful of developed telomere amplification methods exist, of which the length-assay (LA) family of methods (TeSLA, STELA, U-STELA) may be amenable to sequencing. However, these methods rely on double overnight ligation followed by PCR, both of which may be challenging to implement in a microfluidic format.

Furthermore, visualization of telomeres requires Southern blotting, implying that a large portion of amplicons are nonspecific. Isothermal amplification methods have been demonstrated to generate long (10 kbp) amplicons (*59*, *60*), but have not yet been applied to telomeres.

### 1.5 Scope of the dissertation

The objective of this dissertation work was to develop and integrate the technologies described in the previous section in order to get closer to the goal of probing the telomeres of specific chromosomes at the single cell level.

To this end, we developed a process for fabricating solid state pores at scale, as well as expression and purification of a protein nanopore. We have shown both to be capable of DNA length measurement.

A method was developed to enrich telomeres from genomic DNA extract for nanopore sequencing, which enables single-base resolution of telomere length as well as increased coverage of subtelomeres. The former from which single cell telomere length measurements could be referenced against, and the latter forming the reference from which cell-line specific primers can be designed.

Isothermal methods for telomere amplification were investigated, and microfluidic devices were designed for single cell lysis and DNA extraction.

# 2 Creation and use of nanopores for DNA analysis

## 2.1 Introduction

### 2.2.1 Solid state nanopore utility

A solid state nanopore (or pore) is generally formed from a thin (<100 nm) inorganic (e.g. SiO, SiN, $HfO_2$, $TiO_2$) film square with edges that are 10 to 100 microns in length. The film is generally suspended by a silicon frame. The pore itself has a diameter that is within an order of magnitude of the diameter of the analyte of interest or depending on the method of fabrication. When a voltage is applied across the nanopore in an electrolytic solution, the measured ionic current is dependent on the cross-sectional area and length of the nanopore. The passage of an analyte through the nanopore causes blockade of the ionic current, which is the measured signal. The length for which the ionic current is blocked is the translocation or dwell time of the analyte. In some cases where the analyte is able to translocate through the pore while folded, the event charge deficit (ecd), or the integral of the current blockade with respect to time, is representative of the analyte's unfolded length (*61*). Solid-state nanopores can be used to measure the length of dsDNA from 30 bp up to 97 kbp (*62*, *63*), as well as reliably measure the relative abundance of analytes (*64*).

Common methods for solid state nanopore fabrication are focused-ion beam drilling followed by shrinking with transmission electron microscopy (TEM) or controlled dielectric breakdown. These methods offer precise control of nanopore size at the sub-10 nm level, but have limited throughput since they are serial methods. Verschueren et al. has recently demonstrated a method for wafer-scale production of solid state nanopores using e-beam lithography (EBL) with reactive ion etching (RIE) (*65*). Using this method, nanopores as small as 16 nm can be fabricated and it is speculated that smaller pores can be fabricated by using a smaller electron-beam spot size.

The utility of solid state pores to this work is manyfold. Isolation of telomeres, whether by digest or by amplification, would produce dsDNA in exactly the length range that is measurable by solid state nanopores. Translocation of such molecules would then provide both the telomere length and their relative abundance, similar to telomere length profiles provided by telomere length assays such as TeSLA and U-STELA. For higher resolution and without sacrificing device stability, protein nanopores can be embedded into a solid state pore, allowing nearly single base resolution. The wide applicability of this technology compelled us to develop a fabrication process.

### 2.2.1 Protein nanopore utility

By their nature, protein nanopores provide higher spatial resolution than solid state nanopores. Solid state nanopores have been fabricated in 2D materials such as graphene and molybdenum disulfide ($MoS_2$) but are limited to diameters of 8-10 nm and are plagued with issues which as high noise, low yield from fabrication, and difficulty wetting (*65*, *68*, *69*). On the other hand, protein nanopores have a hydrophilic lumen and their formation by protein folding means very low pore to pore variation. Also, insertion of protein nanopores into solid state nanopores has previously been demonstrated. Doing the same here would provide the resolution of a protein nanopore with the robustness of a solid state nanopore. For its demonstrated applicability in detecting single stranded DNA and its robustness in a wide variety of denaturing conditions, we selected a variant of MspA (M2N) for protein cloning.

## 2.2 Methods Development and Results

### 2.2.1 Fabrication of solid state nanopores on silicon nitride

Solid state nanopore fabrication was based on the method from Verschueren et al. 2018 with some modifications (*65*). The overall process is illustrated in Fig 2.1.1. Double-side polished P-doped (100) silicon wafers with wet thermal oxide (SiO) and low stress low pressure chemical vapor deposition (LPCVD) silicon nitride (SiNx) were purchased from University Wafer, Inc. For consistency of orientation referencing, we found that we must specify the primary flat to be on the <110> plane. Although this is standard, omitting this specification led to receipt of wafers with flats that were 45° to <110>. Initially, we developed the process on 525 μm thick wafers with 100 nm SiO and 20 nm SiNx because of the ease of handling. Later on, we purchased a batch of 200 μm thick wafers with 100 nm SiO and 12 nm SiNx and transferred our method to that format as well. Another modification is that we added 20 μm gold squares on the same face as the membrane to allow precise location of the 10-40 μm square membrane during electron beam lithography (EBL). Using gold squares as alignment marks rather than sacrificial membranes allowed us to increase the yield per wafer for our process. Also, this method is amenable to custom patterning of metals on the membrane plane, allowing for the future possibility of coplanar electrode design for enhanced capture of analytes.

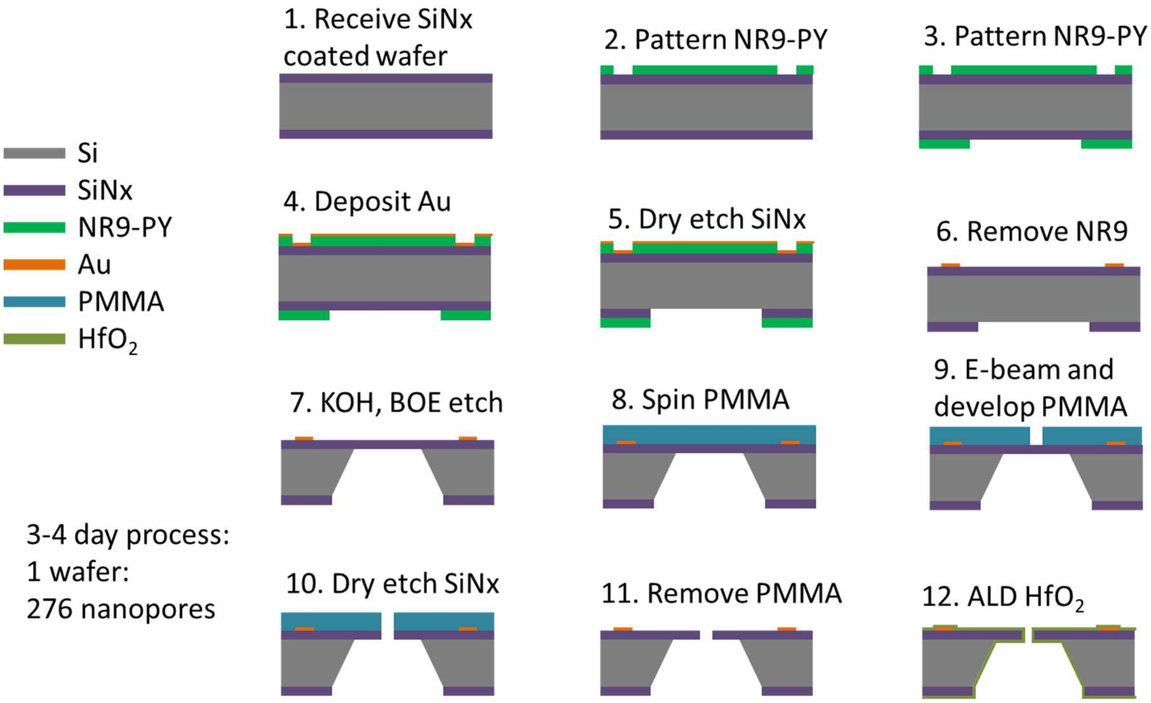


Figure 2.1: Schematic of solid state nanopore fabrication.

We define the top side of the wafer to be coplanar with the free-standing membranes that are left after wet etching, also known as windows. All baking steps were done on a hotplate. Before any processing, the wafer was dehydrated by baking at 150°C for 5 minutes. The top side is patterned first with a ~1 µm layer of NR9-1500PY by spin-coating at 500 rpm for 10s followed by 4000 rpm for 40s with an acceleration of 12000 rpm/s. The NR9 was soft-baked at 150°C for 1 min, and deliberately underexposed with a Karl Suss MA6 using vacuum contact exposure for 15.0 sec with an intensity of 11 mW/cm$^2$. The top side mask (Fig 2.1.1) was designed in AutoCAD (AutoDesk) and ordered as a chrome-on-glass mask from Front Range Photomask, LLC. Post exposure bake was 100°C for 1 min. Resist development was done by immersing in RD6 in a PTFE container for 12s, followed by immersion in DI water in another PTFE container to stop development. After further rinsing with DI water and drying with pressurized nitrogen gas, the top side was treated with oxygen plasma using a Tepla Asher. Plasma treatment was done at 150W with 120 sccm $O_2$ for 60s. The top side is then coated with 5 nm of Cr at 0.5 nm/s and 50 nm of Au at 0.8 nm/s using a Temescal e-beam evaporator. Throughout this process, the bottom side is bare and the wafer is only handled when necessary using PVDF (polyvinylidene fluoride)-tipped tweezers to protect the bottom side. After metal deposition, the bottom side is coated with a 3 μm layer of NR9 with a pattern for wet etching (Figure 2.2). This was done by spin-coating NR9-3000PY at 500 rpm for 10s followed by 800 rpm for 40s with an acceleration of 2400 rpm/s. This layer was soft baked at 150°C for 1 min, exposed by vacuum contact for 52.8 sec with back-side alignment, and post-exposure baked at 100°C for 1 min. The bottom side mask (Fig 2.1.1) was printed on transparency by CAD/Art Services, Inc and fixed to a glass plate with Kapton tape. Development was done by immersion in RD6 for 24 sec, followed by immersion and rinsing in DI water. The bottom side is then dry etched in a Trion Minilock Etcher at 150W, 40°C, 10 mTorr, 60 sccm $CF_4$, and 6 sccm $O_2$ for 4 min. Removal of NR9 from both sides was done by submersion in RR41 heated to 80°C for 3 hours or overnight. The wafer was then rinsed with RR41 for 2 min followed by rinsing with DI water. The next step involved wet etching, so complete dryness was not necessary.

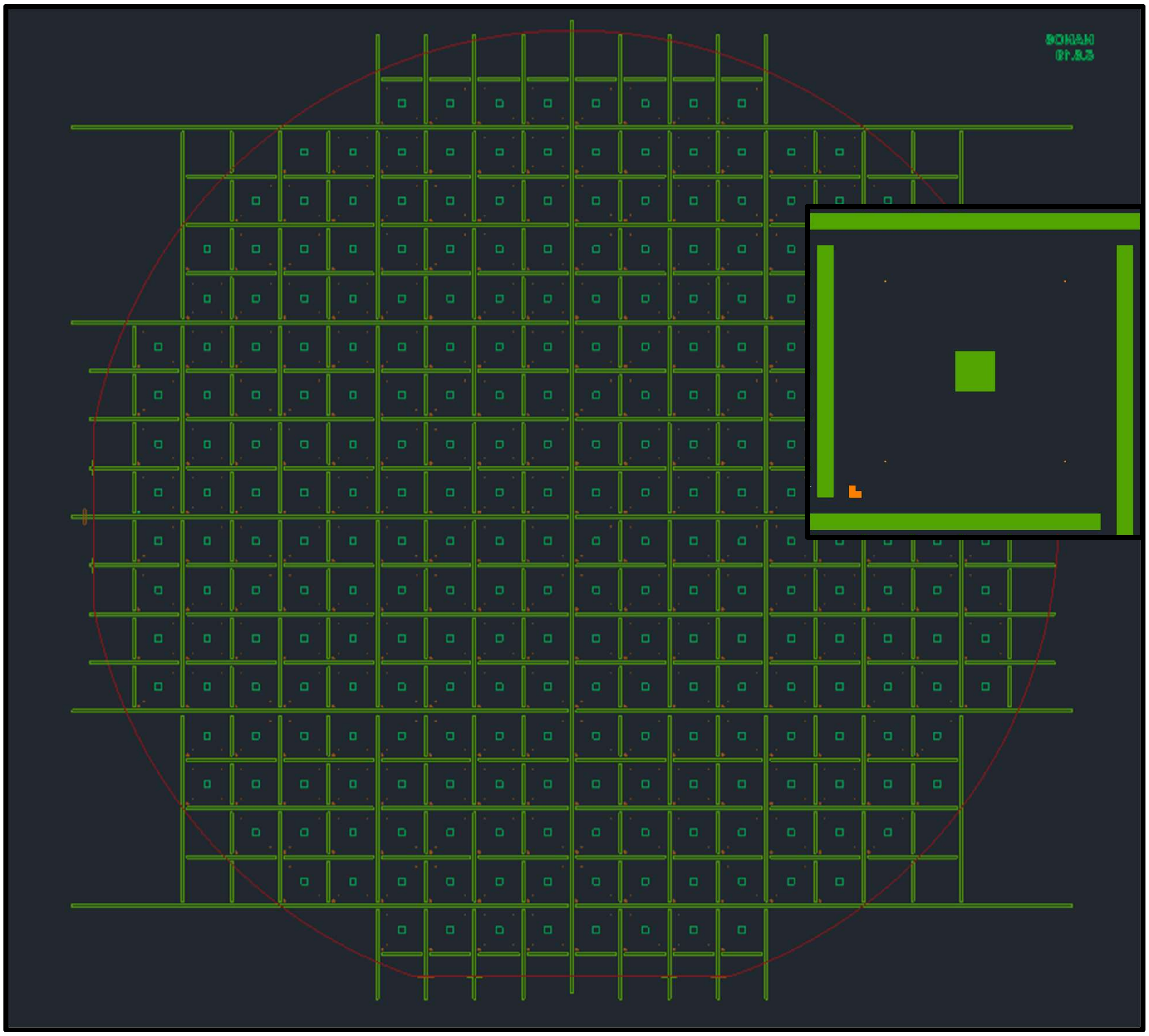


Figure 2.2: AutoCAD schematic of 100 mm wafer (in red) with top (orange) and bottom (green) masks overlaid. Horizontal markers near the bottom of the mask were used for fine rotational alignment with the wafer primary flat or <110> direction. Vertical markers at the middle left and middle right of the mask were used for centering the mask with the wafer. (Inset, upper right) A single 5 mm x 5 mm section with 661 μm square etch opening, 20 μm square alignment marks, and “L”-shaped orientation reference.

During method development, we found that patterning both sides with NR9 before metal deposition would lead to incomplete lift-off and reduce alignment mark quality, presumably due to rearrangement of the topside NR9 during soft baking of the bottomside NR9. To be functional for EBL, alignment marks must be a complete square (no missing sections), and be free of extraneous metal flakes. Four alignment marks are placed in four quadrants in the vicinity of each membrane, and a minimum of three are necessary for automatic EBL alignment. By depositing the metal layer before patterning the bottomside NR9, we saw an improvement from at least three adequate alignment marks for ~70% of membranes to four adequate alignment marks for ~95% of membranes. We also found that using Ti as an adhesion layer led to loss of some alignment marks during KOH etching, whereas no KOH-associated loss was seen when using Cr as an adhesion layer.

After lift-off, the wafer is wet etched bottom side facing up in 35% w/v KOH at 80°C for 5-7 hours. During wet etching, the primary flat of the wafer is propped up against a small (~3 cm D x 3 cm L) PTFE cylinder placed in the container. This prevents the top side from contacting the bottom of the glass container and allows any bubbles formed during etching to shed from the top side. The final etch time for each wafer could vary depending on the etch temperature, target window size, wafer thickness, and doping content. Throughout the process, the etching was paused by transferring the wafer to a water bath for rinsing, followed by gentle drying by holding the wafer upright on a clean dry wipe. The etch depth on the bottom side could then be measured on a microscope. The etch was allowed to proceed for 2 hours followed by hourly checking to estimate the final etch time. The estimation was a linear approximation and was generally an underestimate, so the etch progress was then checked every thirty minutes towards the end of the process. KOH etching was complete when a majority of windows were at or exceeded the target window size. Although KOH etching can remove SiO as well (*70*), we found this to be inadequate and followed up with submersion in Buffered Oxide Etch for 1 min followed by rinsing with DI water .

The KOH etch rate of Si is anisotropic with respect to crystal plane orientation and ideally reveals (111) planes which are etched significantly slower than others (*71*). The (111) plane is 54.7-° to the wafer (100) plane. In theory, the plane revealed after etching should follow that angle. In practice, we found that the windows left after etching were 10s of μm larger than expected. Misalignment of the bottom side etch opening with the <110> direction can also generate larger than expected windows, but this would require a misalignment of 8° to reach sizes similar to what we observed. During rotational alignment of a mask to the wafer flats, horizontal alignment marks are separated by 30000 μm and Y-axis misalignment could at worst be 50 μm, resulting in a maximum rotation error of 0.1°. Nonetheless, the etch angle was confirmed to be 58° by measuring the entrance size and the window size after complete etching. The cause for this discrepancy was not determined, but it was consistent for multiple wafers from separate wafer batches. We compensated for this effect by shrinking the etch window size. Another compensation factor was the slow enlargement of the etch opening. We found that the bottom side window enlarges by roughly 10 μm/hr during KOH etching. To generate 10 μm windows for our 525 μm and 200 μm thick wafers, we used a bottom side etch opening of 661 μm and 331 μm, respectively.

In our design, we aim to place the window in the center of four alignment marks that are spaced 3000 μm apart. Due to backside alignment error, the membrane may be off-center by as much as 15 μm in the X and Y directions. In cases where windows are 10 μm, this needs to be accounted for as not doing so would cause an EBL single pore exposure to miss the membrane. To do this, we use a brightfield microscope to take wide-field images encompassing the membrane and an alignment mark at 5X magnification. For each multichip, a minimum of one set of alignment marks are needed for translational error, but we generally use two sets of alignment marks to account for wafer-scale error drift. The images are then overlaid, rotated, and scaled in our CAD file to obtain the actual location of the membrane with respect to the alignment marks. The CAD file is then updated with new alignment mark positions before conversion for EBL.

After wet etching, the wafer was split into multi-chips (generally 4x4 or 4x2) to optimize EBL, dry etch, and atomic layer deposition (ALD) conditions. Multi-chips were further split into smaller sections when necessary. To optimize EBL conditions, we looked at PMMA thickness, EBL exposure type, and dosage. For all EBL, we used PMMA-950K A2 (MicroChem) as our photoresist. Spin coating thickness was measured by scratching the soft baked resist and measuring the depth of the scratch using a Dektak 150 or Dektak XT profilometer.

To optimize dosage and exposure type, we first spin-coated a single chip with a 100 μm window with 100 nm 950K PMMA. To spin-coat small pieces, a 4-inch wafer with double sided tape was used as the multichip holder. After spin-coating, samples were removed from the holder and soft-baked at 180°C for 1 minute. Samples were then submitted for EBL writing, where multiple 4x4 spot arrays were patterned on the window. Each spot array consisted of pores spaced 200 nm apart and corresponded to a specific exposure type and dosage. Exposure types were single-shot and sequence, and doses ranged from 2600 $\mu C/cm^2$ to 5000 $\mu C/cm^2$ and later 500 $\mu C/cm^2$ to 3000 $\mu C/cm^2$ (Figure 2.1.4). Developer was prepared by mixing four parts ethanol with one part DI water and chilling to 5°C. Chips were developed for 30 seconds, followed by rinsing in ethanol for 45 seconds. Before dry etching the nanopore patterned chips, we first measured etch rate in case of instrument variability. To measure etch rate, individual chips from a separate uncoated multichip were used as test samples. Test chips were dry etched in a Trion Minilock Etcher at 50W, 40°C, 50 sccm $CF_4$, 10 mTorr for 10 sec and 20 sec. The surface SiNx thickness was measured using a J.A. Woollam Ellipsometer, and the etch rate was estimated by linear approximation. We define the time required to etch the thickness of the SiNx layer as 100% etch. The nanopore patterned chips were dry etched under the same conditions to 130% etch. PMMA was removed by rinsing with acetone, isopropanol, and water for 30 sec each, followed by submersion in 4 parts $H_2SO_4$ to 1 part 30% $H_2O_2$ (piranha etch) at 80°C for 1 hour. Chips were left in piranha etch overnight to cool, followed by rinsing with DI water for 2 minutes and gentle drying with pressurized $N_2$ for 1 minute. Chips were then imaged by scanning electron microscopy (SEM) using a Zeiss Sigma 500 at 1.0-1.5 kV with In-Lens detection and a working distance of 2.0-3.0 mm. These SEM settings yielded sufficient resolution for pores with diameters greater than 10 nm.

A custom matlab script was written to measure pore diameters. In brief, this script does the following for each image: estimate initial pore positions by finding local minima of 1-dimensional standard deviation operations of the image, fine-tune pore positions by high climbing optimization, gather intensity profiles for multiple angles centered on the pore, and find the full width at half maximum (FWHM) of the averaged pore profile. Some images could not be processed automatically due to poor sample cleanliness or poor image quality. In these cases, ImageJ was used to obtain radial profiles and the pore diameter was defined as the FWHM of the radial profile. The diameter was scaled from pixels to nanometers using the predefined 200 nm spacing between pores.

To investigate the effect of PMMA thickness, RIE time, and EBL dose on pore sizes, we started with a 2x2 multichip with 40 μm square windows. The multichip was split in half (1x2) and each half was spin-coated with PMMA-950K to 60 nm and 140 nm. PMMA A2 was thinned with A Thinner (anisole) to obtain a 60 nm thickness. A dose series with 4x5 arrays of pores was patterned on each membrane, and the multichips were split again into individual chips. Chips were dry-etched as described previously, but at 1.3X and 1.6X etch. PMMA was then stripped and the chips were imaged by SEM.

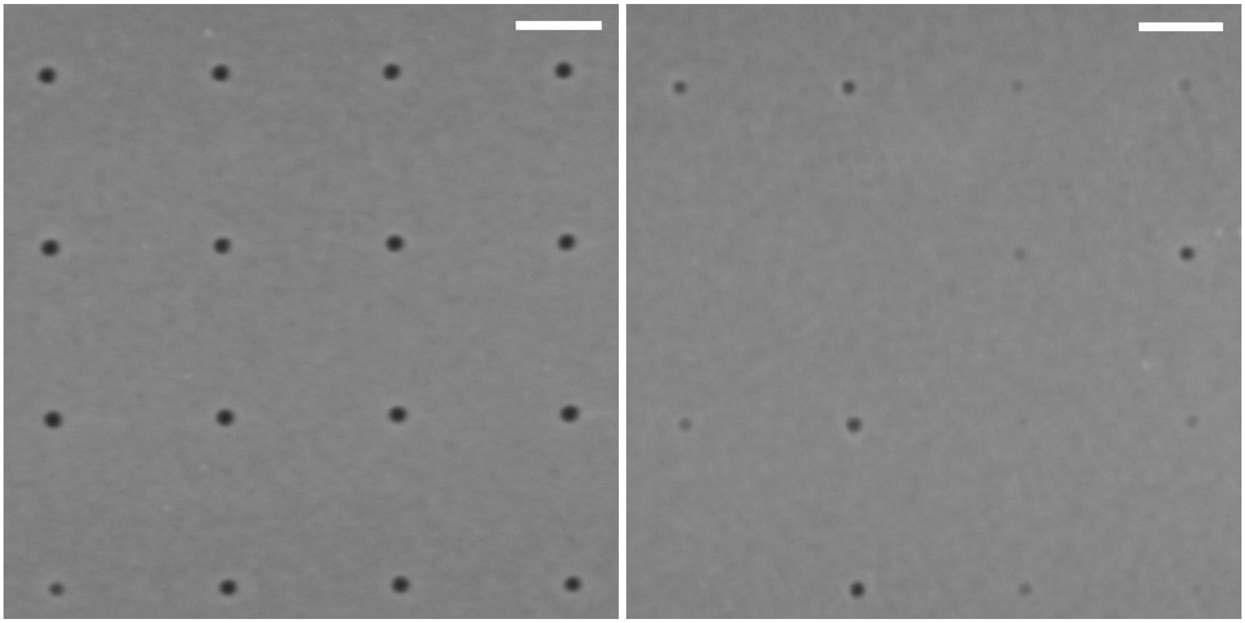


Figure 2.3: Representative SEM images of solid state nanopores. Scale bar: 100 nm. (left) 4x4 array of pores, all present. (right) 4x4 array of pores with some missing and some incompletely formed due to lower EBL dose. Images were taken using an EHT voltage of 1.00 kV, 2.6 mm working distance, and In-Lens detection.

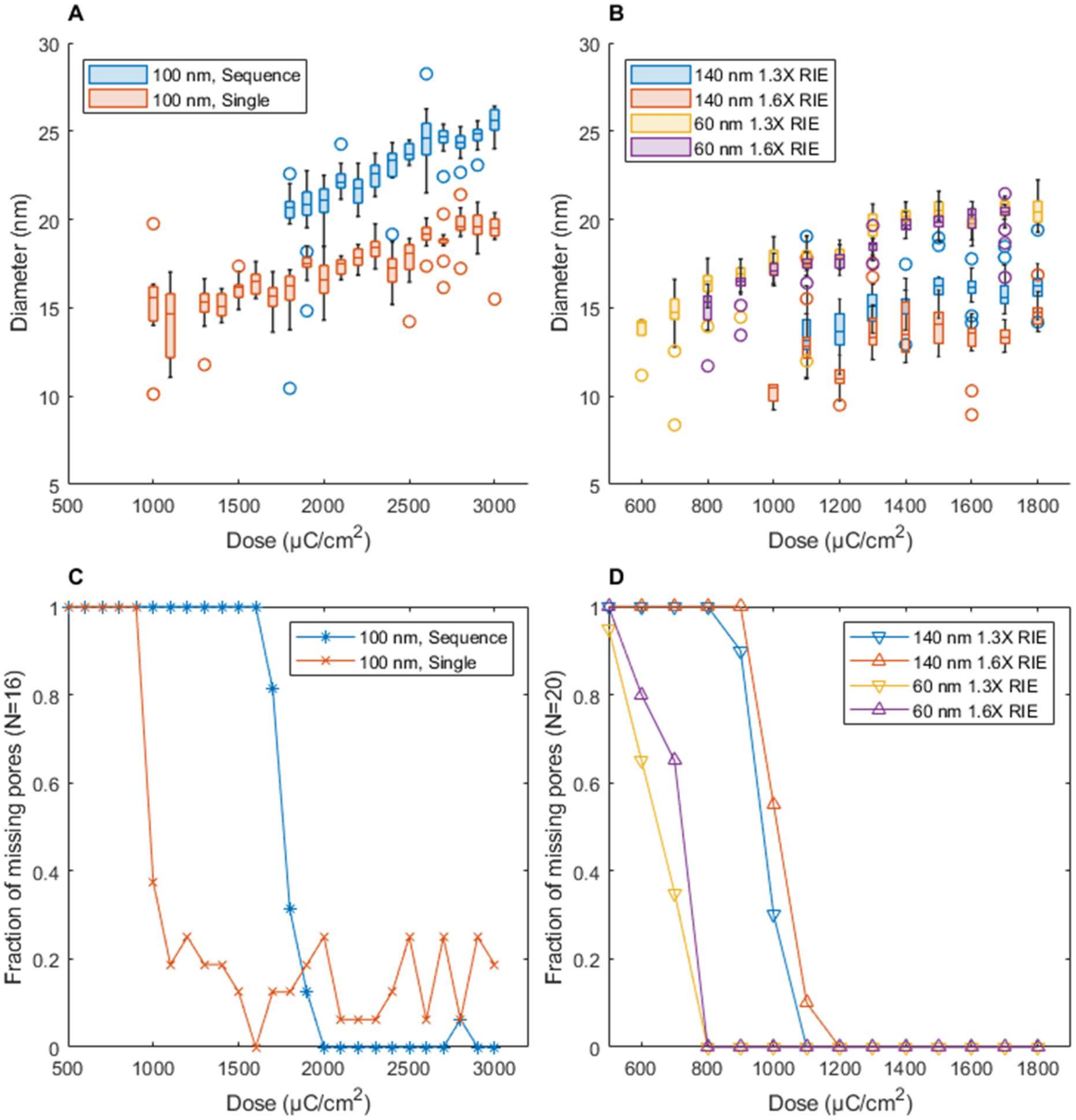


Figure 2.4: Effect of dose, PMMA thickness, and etch time on nanopore diameter and presence.

Chips that were further processed with atomic layer deposition (ALD) were taken directly after overnight piranha and thorough DI water rinsing. Without this step, chip coating was non-uniform. ALD was done in a BeneQ TFS 200 using predefined recipes for alumina ($Al_2O_3$) or hafnia ($HfO_2$). Before coating chips, the deposited thickness per cycle was determined by coating a glass slide with 100 cycles and measuring the thickness of the deposited layer by ellipsometry. After ALD, chips were stored in equal parts ethanol and DI water until use.

### 2.2.2 Machining of TEM chip holder and TEM imaging

The use of ALD necessitated higher resolution imaging for accurate pore diameter measurement. Generally, transmission electron microscopes (TEM) are well-suited for liquid samples or dry samples that can be resuspended in liquid. This allows for sample embedding on a 3 mm diameter mesh disc which is analogous to the glass slide used on a light microscope. This was incompatible with our nanopore-on-membrane samples, which would surely fracture or tear if transfer from the 5 mm silicon frame was attempted. Available to us through the UCSD Electron Microscopy Core Facility is a JEOL 1400 plus transmission electron microscope (TEM) equipped with a JEOL Common Specimen Holder (JEOL EM-21010 SCSH). To our advantage, this specimen holder is designed to have a quick-release swappable tip (JEOL EM-11610 QR1), on which one would mount TEM discs. Relevant measurements from the specimen tip were used to design the chip holder in Autocad Fusion 360 (Figure 2.5). The piece was then milled using a 3-axis Tormach PCNC 1100. Copper was chosen as the stock material for its high thermal conductivity. The tip was machined out of a copper blank mounted to a sacrificial aluminum plate using cyanoacrylate adhesive. This allowed the piece to be held securely while giving the mill complete access to five sides of the piece. The piece was released from the sacrificial plate by heating over a bunsen burner. After cleaning, the chip holder was fashioned with a rotating spring clip (Figure 2.5). Chips were loaded window side down in the chip holder before loading into the TEM. Images were taken with a high tension voltage of 80 kV.

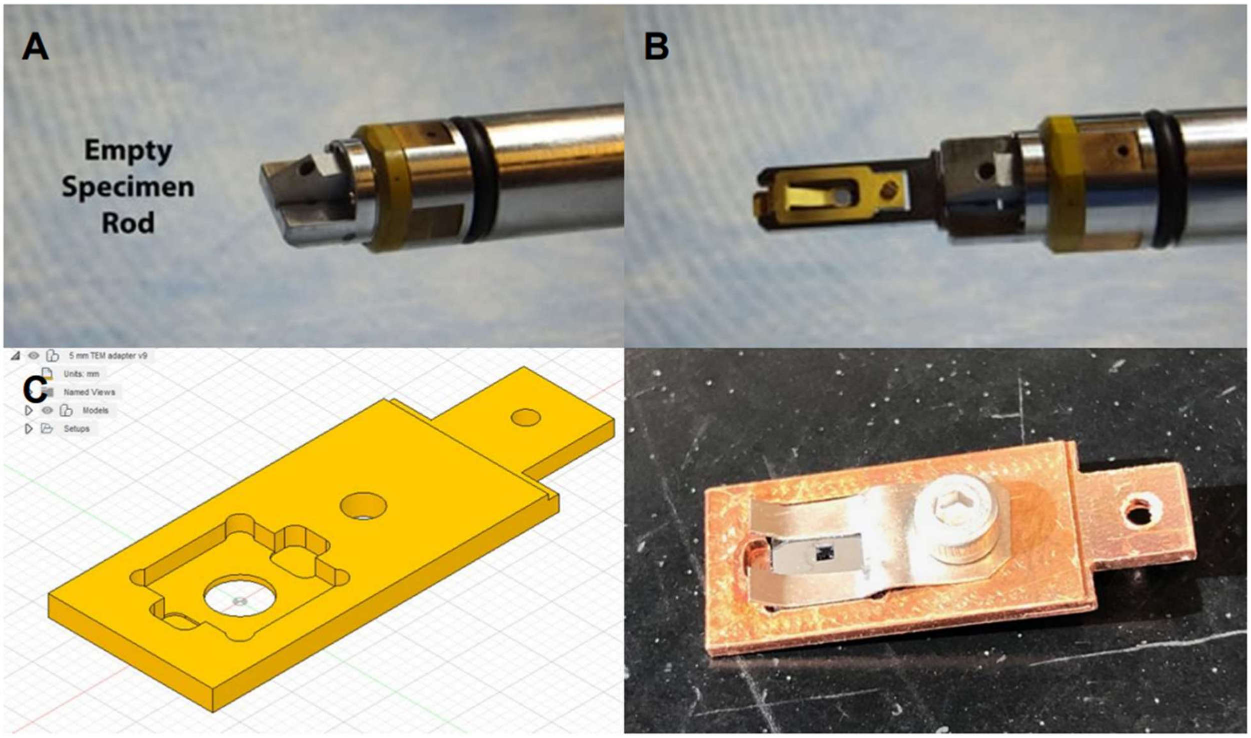


Figure 2.5: (A) JEOL Common Specimen Holder, also known as sample rod, with no attachment loaded. (B) Sample rod with standard quick release tip loaded (C) CAD drawing of custom chip holder (D) Finished chip holder with chip loaded window side down. Images A and B courtesy of IU Bloomington EM Center.

Nanopores coated with $HfO_2$ had a darker interior lining under TEM, a result of atomic number (Z) contrast (Figure 2.6). The atomic numbers of Si, Hf, and Al are Z = 14, 72, and 13, respectively. Although the planar surface of the membrane is also coated, high contrast is localized to the circumference of the pore due to the higher %Hf taken through the membrane’s thickness. To the scale that is resolvable, this suggests that the pore has the geometry of a cylinder rather than a truncated cone, which would have resulted in a contrast gradient. Pores coated with $Al_2O_3$ have a lighter interior lining than the surrounding membrane, as predicted by Z contrast.

### 2.2.3 Solid state nanopore conductivity measurements and translocation of DNA

Functional diameter may differ from TEM diameter due to surface hydrophobicity and the presence of trapped nanobubbles. We can model the resistance of the pore as a conducting cylinder, or $R_{\text{channel}} = 4\rho l/d^2$ to obtain the relationship between a pore's functional diameter and its I-V curve (*72*). Here, *ρ* is the resistivity of the medium within the pore, *l* is the pore length, and *d* is the pore diameter. We also include a term for access resistance $R_{\text{access}} = \rho/2d$, which represents the spatial restriction due to the size of the pore entrance. The resistance of the pore is therefore:

$$R = R_{\text{channel}} + 2R_{\text{access}}$$

And the conductance is

$$G = \sigma\left(\frac{4l}{\pi d^2} + \frac{1}{d}\right)^{-1}$$

Where *σ* is the conductivity of the buffer. Given an I-V curve, we can do a least-squares fit to obtain *G* ($G = I/V$), and the equation above gives a single positive real solution when solving for *d*.

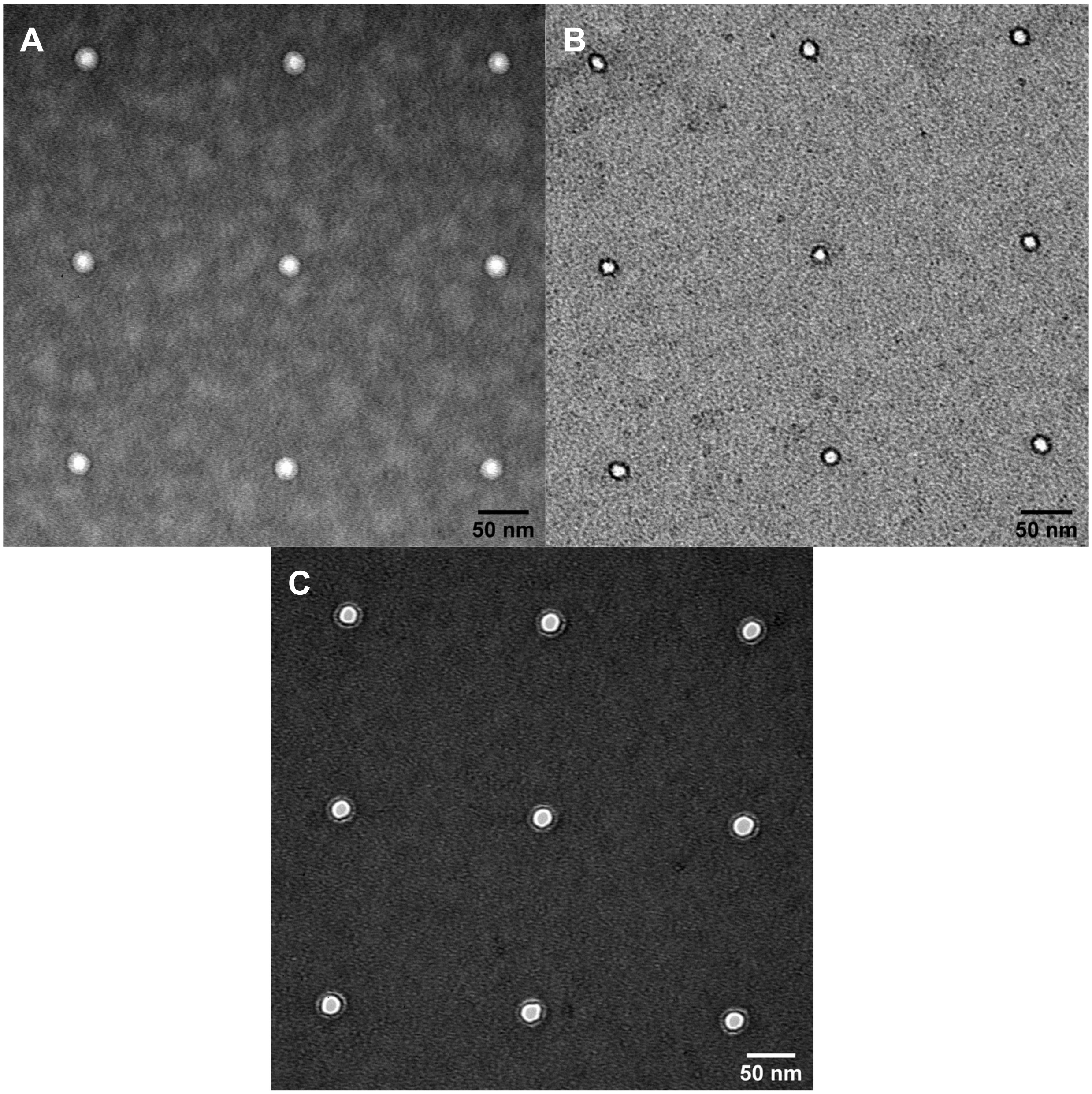


Figure 2.6: (A) Uncoated nanopores in SiN. (B) Nanopores after $HfO_2$ coating. (C) Nanopores after $Al_2O_3$ coating.

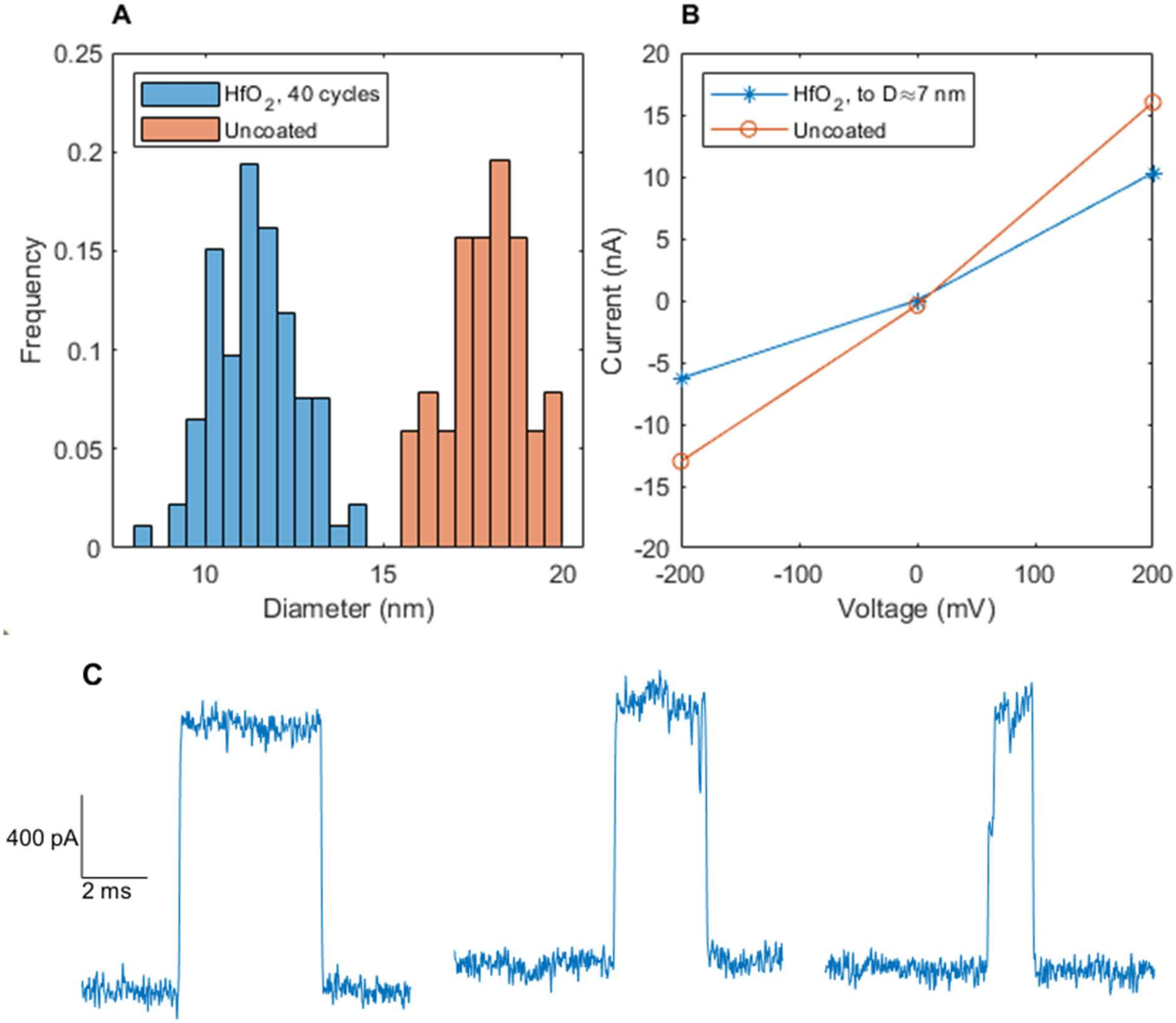


Figure 2.7: (A) Diameters of uncoated nanopores and nanopores coated with 40 ALD cycles of $HfO_2$, based on TEM images. (B) I-V curve of representative uncoated and coated nanopores. Based on conductance, the $HfO_2$ coated sample has a diameter of 8.7 nm and the uncoated pore has a diameter of 14 nm. (C) Representative translocations of lambda DNA through $HfO_2$ coated pore.

### 2.2.4 TEM of Focused Ion Beam Lithography (FBL) samples

Some preliminary work was done to investigate the size of pores generated through focused ion beam lithography (FBL), which could potentially cut out multiple fabrication steps associated with EBL. To this end, we imaged pores that were generated by FBL. Chips were written using Si and Au ions at varying doses. TEM images were manually processed using ImageJ due to the variable eccentricity of the pores. Si ions gave smaller pores than Au ions, and the smallest pores achieved had a major diameter of 8.3 ± 1.7 nm, a minor diameter of 7.6 ± 1.2 nm, and an eccentricity of 0.40. Below a specific dose, pores did not appear by TEM to be drilled entirely through the membrane. The wettability and I-V characteristics of these pores have not yet been investigated.

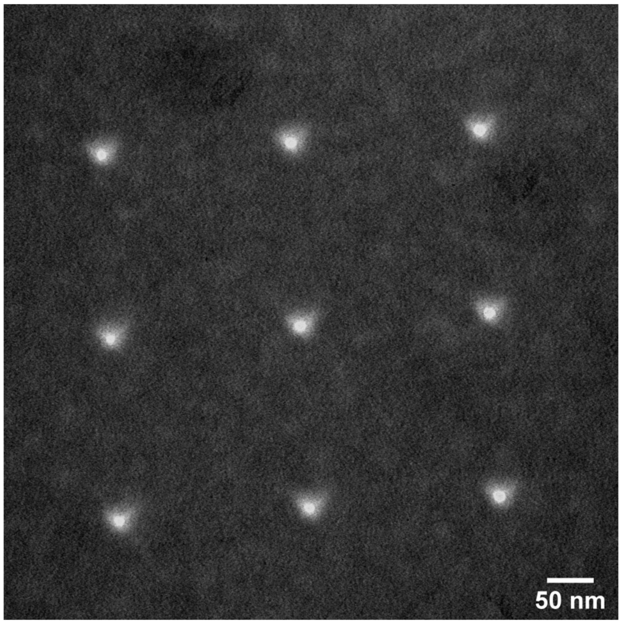


Figure 2.8: TEM images of FBL pores

### 2.2.5 Method development for M2N purification

The gene encoding M2N was custom synthesized as a gBlock (IDT). Linearized plasmid backbone and overlapping gene insert were generated by PCR with Q5 High-Fidelity Master Mix (NEB). Plasmid primers were FWP2 and RVP2, and insert primers were FWP1 and FWP2. Genes were inserted into plasmids using the NEBuilder HiFi Assembly kit (NEB). The HiFi assembled product was heat-shock transformed into NEB 5-alpha competent *E. coli*, which were grown on Luria-Bertani (LB) agar plates with 30 mg/mL kanamycin. Colonies were picked for PCR and sequencing screening of the insert using FWP1 and RVP1. The selected colony was grown to $OD_{600}$=1.0 in 5 mL of LB medium with 30 mg/mL kanamycin. Insert-containing plasmid was extracted from 4.5 mL of cells using a Monarch Plasmid Miniprep Kit (NEB), and the remaining 0.5 mL was combined with 0.5 mL of 30% autoclaved glycerol for freezing at -80°C. Extracted plasmid was quantified using a Nanodrop UV-vis Spectrophotometer, and was subsequently transformed into BL21(DE3) competent *E. coli*. Transformants were grown on LB agar plates with 30 mg/mL kanamycin.

Table 2.1: Oligonucleotides and protein mutants used in this study.

| | |
|---|---|
| FWP1 | GCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCA |
| RVP1 | ATCCGGATATAGTTCCTCCTTTCAGCAAAAAACCCCTCAAGACCC |
| FWP2 | GGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT |
| RVP2 | TGAGTCGTATTAATTTCGCGGGATCGAGATCTCGATCCTCTACGC |
| MspA-M2N | D90N/D91N/D93N/D118R/D134R/E139K |

M2N has been expressed before in its native host *Mycobacterium smegmatis,* but doing so in BL21(DE3) *E. coli* presented some challenges even though it had been demonstrated before (*73*). Early attempts at detergent-based extraction were hindered by loss of product during centrifugation and low presence of product. Since we anticipated that column purification (whether FPLC or HPLC) would be necessary downstream of extraction, we wanted to prevent loss of protein due to centrifugation or sample filtering. Such losses are generally associated with formation of insoluble protein aggregates, also known as inclusion bodies, which can be remedied with the addition of urea.

To examine the effect of urea on lysis, we grew transformed BL21(DE3) from a single colony in 20 mL LB with 30 mg/mL kanamycin at 37°C, 175 rpm. At $OD_{600}$=0.7, we added isopropyl ß-D-1-thiogalactopyranoside (IPTG) to 1 mM and decreased the temperature to 16°C. The cells were incubated at this temperature overnight at 175 rpm. The cells were pelleted by centrifugation, resuspended with 1 mL of DI water and aliquot to 500 µL before pelleting again at 16100 g for 5 min at 4°C. One aliquot was resuspended with 1 mL of Phosphate-Genapol Lysis Buffer (PGLB: 100 mM sodium phosphate, 0.1 mM EDTA, 150 mM sodium chloride, and 0.5% w/v Genapol X-080, pH 6.5) at on a shaker set to 800 rpm and 60°C for 10 minutes. The other aliquot was was resuspended with 1 mL of Phosphate-Genapol-Urea Lysis Buffer (PGULB: 100 mM sodium phosphate, 0.1 mM EDTA, 150 mM sodium chloride, 0.5% w/v Genapol, and 8 M urea, pH 6.5). This was incubated at 30°C for 1 hr at 800 rpm. To assess the effects of downstream processing, aliquots were subjected to centrifugation at 16100 g for 5 minutes or 40 minutes, filtration through a 0.22 µm PES filter, or a combination of centrifugation and filtering.

Samples were analyzed on Mini-PROTEAN TGX gels (Bio-Rad) with Color Prestained Protein Standard, Broad Range (NEB) as a molecular weight marker. Gels were stained with ReadyBlue Protein Gel Stain (Sigma-Aldrich) and destained with DI water for 1 hr.

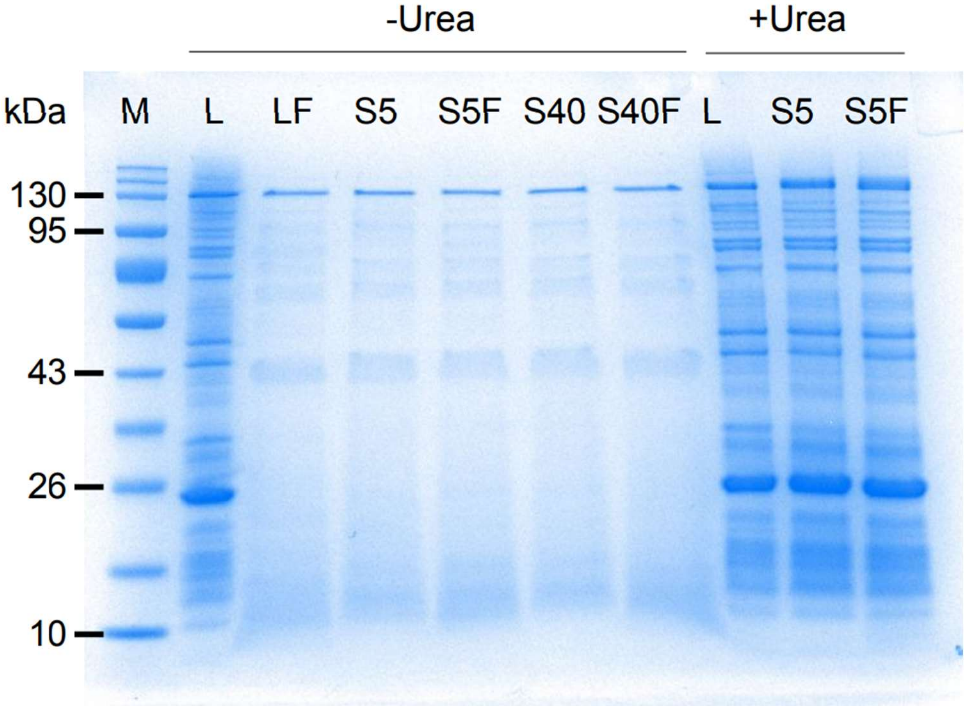


Figure 2.9: Lysate aliquots on Any kD TGX gel. M: marker, L: lysate, F: filtered, S5: centrifuged 5 min, S40: centrifuged 40 min.

Based on gel analysis, lysates processed by filtration or centrifugation without urea had significant loss of protein, except for a band at 130 kDa. Addition of urea to the lysis buffer gave full recovery of all protein that was present in lysate. Lysates without urea retained the 130 kDa band after centrifugation and filtration. Initially this appeared unremarkable, as we had expected the MspA oligomer to migrate near 100 kDa. However, we later found that this band may be correlated to the MspA oligomer and had migrated at a higher MW, possibly due to a difference in gel composition between any kD. We did not investigate this further as we used 4-20% TGX gels moving forward.

We proceeded with the lysate from PGULB because we were mainly interested in maximizing recovery of M2N in order to guarantee sufficient material for column purification. We anticipated that a two-step purification may be necessary, which would inevitably come at the cost of lower yields. Since proteins were solubilized by urea, we wanted to confirm whether subsequent removal or urea would lead to precipitation. To assess this, we purified the urea-containing lysate with NEBExpress Ni Spin Columns using a process with and without urea. For the purification without urea, the wash buffer contained 20 mM sodium phosphate, 300 mM sodium chloride, 5 mM imidazole, and 0.5% w/v Genapol X-080, pH 6.5. Buffer E1 and E2 were similar to the wash buffer, but with 200 mM and 500 mM imidazole, respectively. Purification with urea had all of the same buffers, but with 8 M urea added. Urea-containing lysate was purified according to manufacturer instructions but with the buffers described instead of manufacturer provided buffers. Each fraction was then analyzed by denaturing PAGE using a 4-20% TGX gel (Figure 2.10).

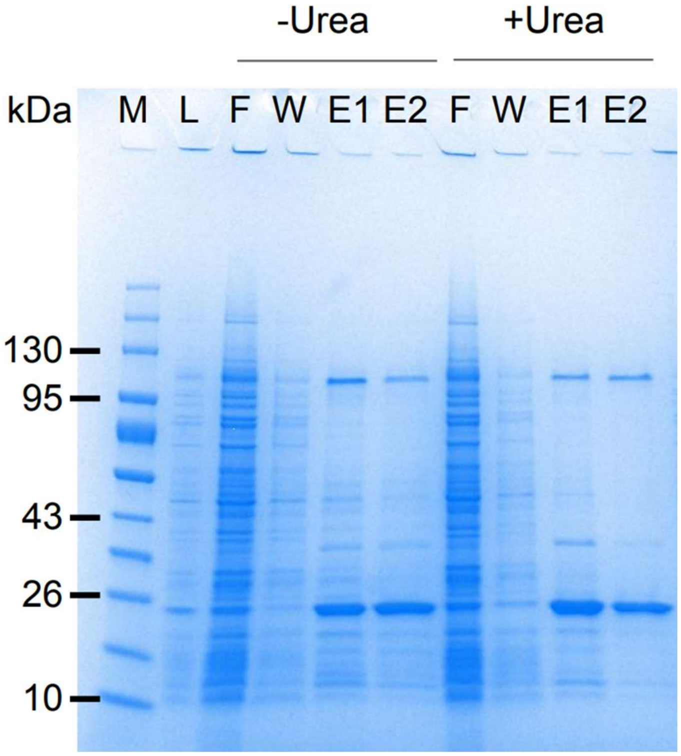


Figure 2.10: Denaturing PAGE of Ni-NTA purification process with and without 8 M urea. M: marker, L: lysate, F: unbound flow through, W: wash fraction, E1 and E2: elution fractions.

Purification in this manner revealed that, despite stepwise elution with imidazole, the monomer and oligomer coelute under the provided conditions. We anticipated that subsequent purification with size exclusion chromatography (SEC) presents a straightforward path to isolation of the oligomer. To generate enough material for this, we performed induction on 200 mL of cells and lysed this with 10 mL of PGULB at 30°C for 1 hour. The lysate was then pelleted and the supernatant passed through a 0.2 μm syringe filter. We then scaled up Ni purification by using a 1 mL HisPur Ni-NTA column (Fisher Scientific). The lysate was loaded onto the column and eluted with a linear gradient from Buffer A to B. Buffer A was 100 mM sodium phosphate, 300 mM NaCl, 0.1%w/v Genapol X-080, and 5 mM imidazole, pH 6.5. Buffer B had 500 mM imidazole instead of 5 mM imidazole. Fractions were collected during the gradient and analyzed by denaturing PAGE using a 4-20% TGX gel (Figure 2.11).

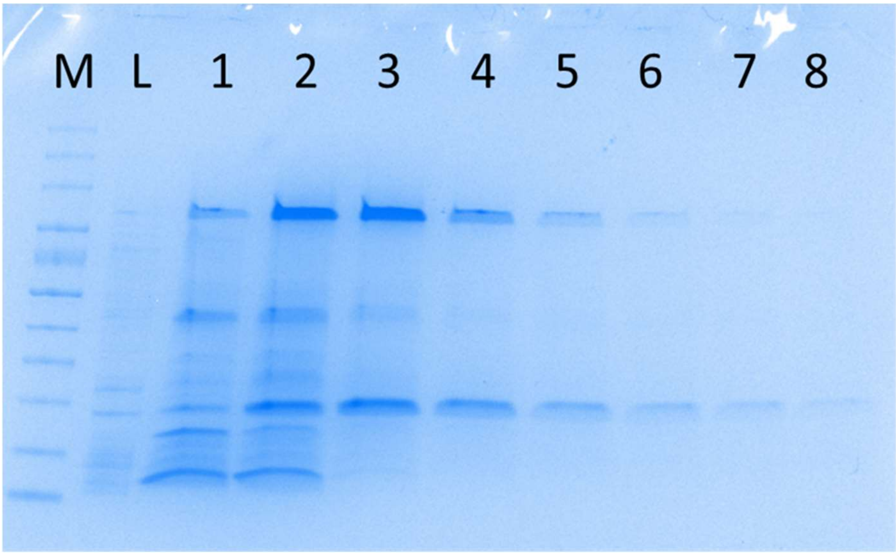


Figure 2.11: Denaturing PAGE of M2N fractions after Ni-NTA column chromatography.

PAGE analysis revealed that fractions 1 and 2 still had a significant proportion of off-target proteins, which overlapped with oligomer elution. To guarantee purity, we pooled fractions 3 through 8 to proceed with size-based purification. SEC was done on an Agilent 1200 HPLC controlled with ChemStation. The column used was an Agilent Bio-SEC 3 with 150Å pore size. We hypothesized that decreasing the concentration of Genapol X-080 below its critical micelle concentration (CMC) would lead to better separation, as this may have been reducing specificity of interaction to the Ni-NTA column. Too high of a concentration and Genapol X-080 could form micellar-M2N agglomerates. Genapol X-080 has a CMC of 0.05–0.35 mM (*74*), or 0.28% - 1.9% w/v. Too low of a detergent concentration could cause the oligomers to aggregate and form precipitate. To optimize the buffer for M2N separation on the SEC column, we first did small test injections of pooled imidazole-eluted fractions while running a parameter sweep with the quaternary pump. The following mobile phases were used as pump inputs: (A) 200 mM sodium phosphate, pH 7.5, (B) 10% w/v Genapol X-080, (C) 5 M sodium chloride, (D) 9:1 (v) MilliQ purified water:ethanol. All mobile phases were vacuum filtered using a 0.22 μm filter. Absorbance was recorded at 214 nm and 280 nm. Test injections involved the following buffer compositions (Table 2.2).

Table 2.2: Buffer concentrations used for optimization of SEC separation

| Mix# | Sodium phosphate (mM) | Genapol X-080 (w/v) | NaCl (mM) |
|---|---|---|---|
| 1 | 25 | 0.50% | 0 |
| 2 | 25 | 0.50% | 150 |
| 3 | 25 | 0.50% | 500 |
| 4 | 100 | 0.50% | 0 |
| 5 | 100 | 0.50% | 150 |
| 6 | 100 | 0.50% | 500 |
| 7 | 25 | 0.10% | 150 |
| 8 | 25 | 0.05% | 150 |
| 9 | 25 | 0.05% | 0 |
| 10 | 25 | 0.10% | 0 |
| 11 | 25 | 0.10% | 500 |
| 12 | 25 | 0.25% | 0 |
| 13 | 25 | 0.25% | 150 |
| 14 | 25 | 0.25% | 500 |

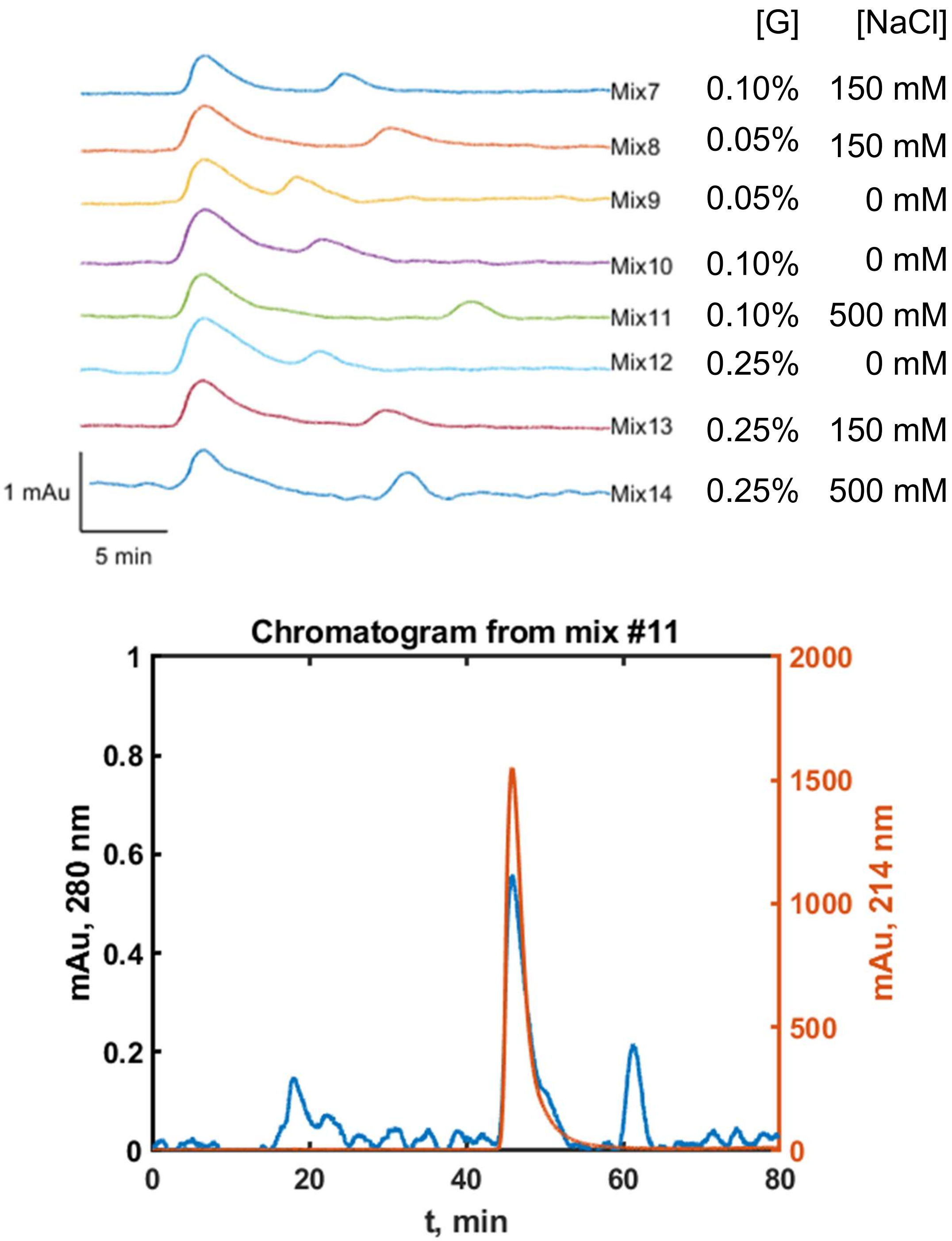


Figure 2.12: (top) SEC chromatograms from test injection of M2N. [G] is %w\v of Genapol X-080. (bottom) Overlay of chromatograms at 280 nm and 214 nm, showing a peak near 60 min that is only visible at 280 nm.

All chromatograms at 214 nm had a single peak. No buffer condition tested was able to resolve additional peaks at this wavelength. In contrast, chromatograms at 280 nm had a peak that corresponded to the sole 214 nm peak in addition to a later-eluting peak. In all buffer conditions with 0.5% w/v Genapol X-080 (Mix#1 through Mix#6), no peaks were visible at 280 nm, possibly due to . Hence, the optimal buffer was chosen from Mix#7 through Mix#14.

Comparing the chromatograms at 280 nm, we see that increasing sodium chloride concentration increases resolution for all concentrations of Genapol X-080 tested between 0.05% w/v to 0.25% w/v. Conversely, holding sodium chloride concentration constant and comparing different Genapol X-080 concentrations shows a different relationship. At 500 mM sodium chloride, resolution is better at 0.10% w/v than 0.25% w/v Genapol X-080. However, with 150 mM sodium chloride, resolution is better at 0.05% w/v and 0.25% w/v than at 0.10% w/v. This seems to suggest that Genapol X-080 has a negative effect on resolution that is dependent on ionic strength. The effect of sodium phosphate concentration was not examined since we obtained sufficient resolution with #11.

The pooled sample was concentrated to <1 mL using an Amicon 3 kDa centrifugal filter. After column equilibration with 30 mL, the sample was injected and eluted with 30 mL of buffer. Fractions from both peaks were collected and analyzed on a 4-20% TGX gel.

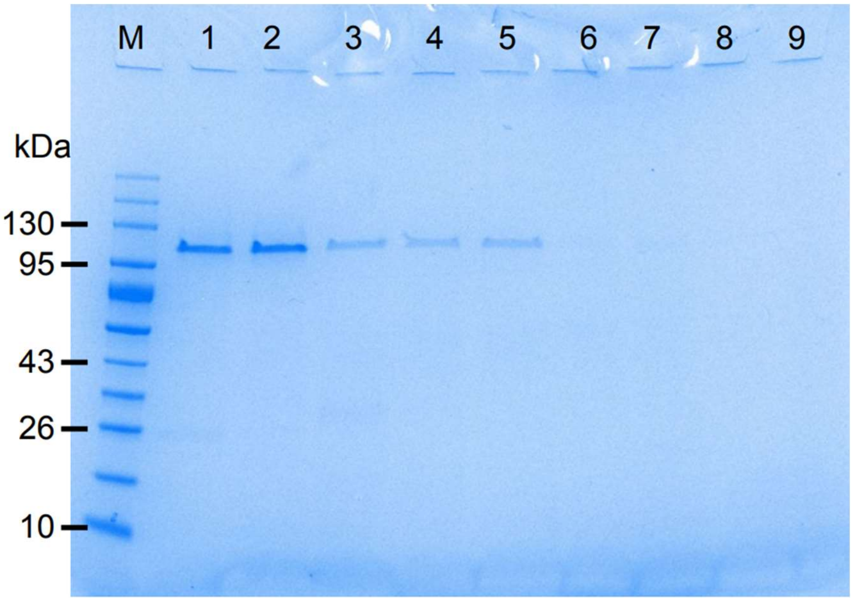


Figure 2.13: Denaturing PAGE of SEC fractions. Lanes 1 through 7 correspond to the first peak at 280 nm. Lanes 8 and 9 correspond to the second 280 nm peak

We initially expected the second peak to consist primarily of the monomer because smaller molecules have more interactions in an SEC column and are retained longer. However, we found no visible protein in the second peak using PAGE. It may be possible that the monomer was strongly retained by the column. However, we did not collect any fractions when we performed stringent wash of the column. M2N mass concentration was determined using a Nanodrop UV-vis spectrophotometer and the coextinction coefficient calculated from its monomer sequence.

### 2.2.6 Protein nanopore characterization by open pore current and translocation experiments

50-100 µm diameter apertures were prepared on PTFE films by indenting the film with a sharp object and applying 7-9 sparks at 1 Hz using a Daedelon Universal Spark Generator (Science First). PTFE films were sandwiched between custom PTFE flow cells that were designed in Fusion 360 and milled using a Tormach PCNC 1100. A hexadecane annulus was formed on the apertures by painting with 10% hexadecane v/v in pentane and allowing the pentane to evaporate. Both chambers of the flow cell were filled with 1 M KCl, 10 mM HEPES pH 7.5, and a droplet of 10 mg/mL DPhPC in pentane is added to both sides. After allowing the pentane to evaporate, a bilayer was formed by slowly lowering and raising the solution past the aperture level using a pipette. Bilayer sealing was confirmed by near-zero current when applying a DC voltage, and bilayer capacitance was measured by applying a ramp-wave voltage. Current-voltage measurements were taken using an Axopatch 200B Amplifier and Axon Digidata 1550B with Clampex software. Ramp-wave voltage was applied using an Agilent function generator connected to the Axopatch 200B Amplifier front-panel switched external command port.  To perform multiple insertion, M2N was diluted in 1 M KCl, 10 mM HEPES pH 7.5 and added to one chamber.

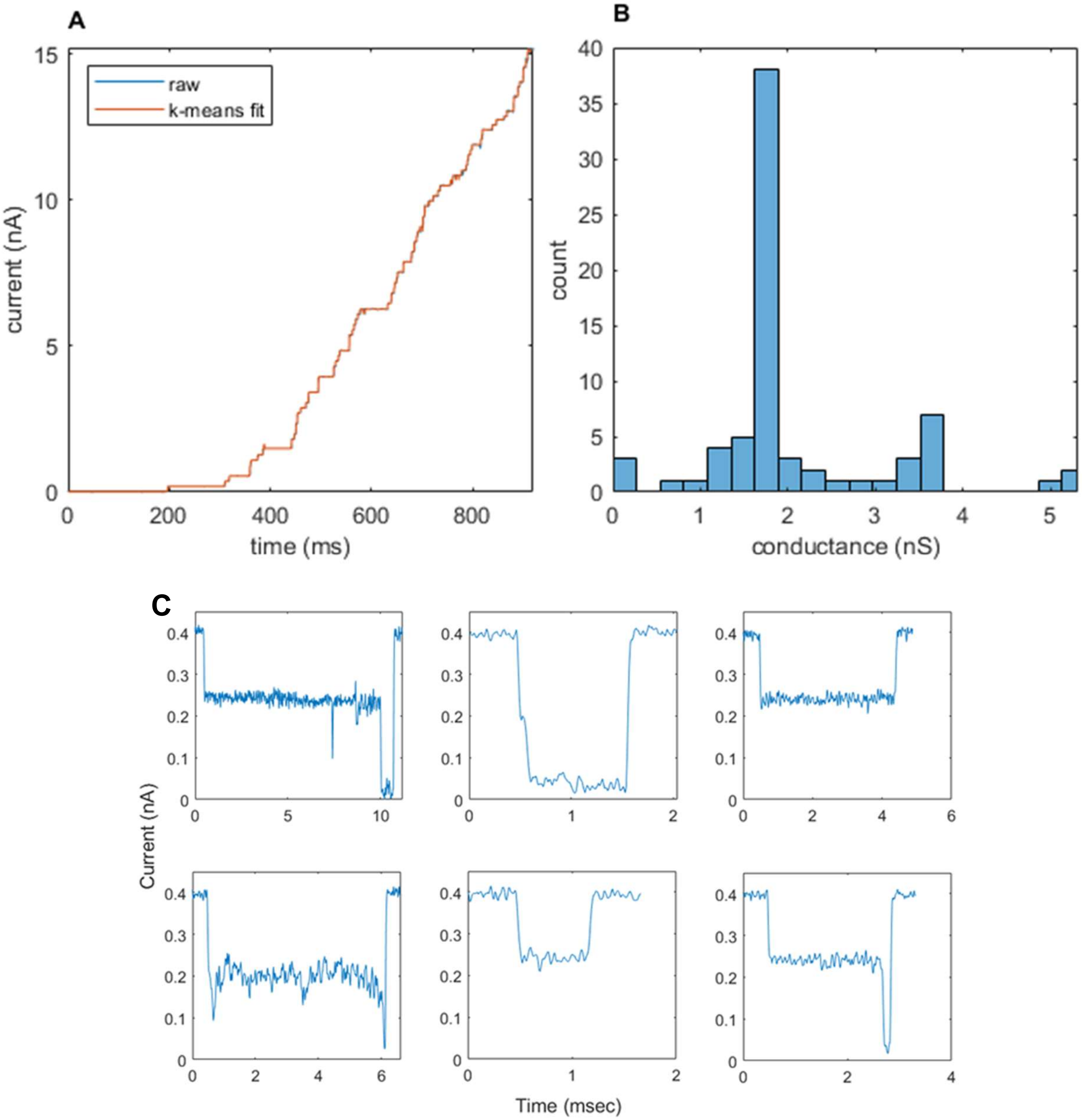


Figure 2.14: (A) Current trace starting at the beginning of multiple insertion of M2N in a DPhPC lipid bilayer. (B) Histogram of M2N conductance from multiple insertion trace. Applied voltage was 100 mV. The bin with the highest frequency corresponds to 1.76 nS. (C) Representative traces from translocation of 120 nt ssDNA at 150 mV.

Current traces from multiple insertion experiments were processed using MATLAB (Mathworks). A median filter was applied to the trace and current levels were partitioned by *k*-means clustering. We chose *k* by counting spikes in the first-order difference of the trace. Pore conductance was calculated based on the differences in current levels divided by the applied voltage (100 mV). The measured pore conductance of 1.76 nS is similar to what was previously seen in literature for this mutant (*75*). Addition of 120 nt long ssDNA (GAAGCAGCCAAAGCCGCAGCA GAGGCACAGAAAAAAGCCGAGGCAGCAGCGGCAGCACTGAAAAAAAAAGCAGAG GCTGCAGAAGCAGCTGCAGCAGAAGCCCGTAAAAAAGCAGCAACCGAA) to one side and applying a voltage of 150 mV gave measurable translocations.

Several attempts were made to insert MspA into solid state pores that had been coated with $HfO_2$ to a diameter of 7-8 nm, but none were successful.

### 2.3 Conclusion

We generated solid state nanopores and protein nanopores with the aim of integrating them to form a hybrid nanopore. Solid state nanopores could be reliably fabricated with average diameters as low as 13 nm using an EBL-RIE process, and this could be further reduced to 7-8 nm using ALD. Although the diameter of the pore could be reduced further, we risked additional difficulty in using the pore due to wettability issues that may arise from the high aspect ratio of the pore. Based on the 3D structure of most protein nanopores, the size that we reliably achieved with ALD appeared sufficient.

We also developed a method for isolation of the MspA oligomer from an induction that was difficult to solubilize. We worked around this by forcing solubilization with urea during lysis, followed by removal of urea the Ni affinity chromatography and subsequent isolation of the oligomer by SEC. The resulting nanopores were functionally tested by measuring conductance and ssDNA translocation, confirming that the initial presence of urea did not have a deleterious effect for this variant. Attempts to insert

MspA into the solid state nanopore were unsuccessful. Strategies for overcoming this may include attachment of an oligonucleotide to the stem portion of the protein nanopore, or insertion using vesicle-embedded MspA.

## 2.4 Acknowledgements

We thank Sylvia Liang, who quickly picked up training on cleanroom fabrication methods and collaborated on nanopore fabrication for 200 µm wafers. Sylvia Liang also performed all lipid bilayer experiments for this study.

We thank Dr. Maribel Montero of the Nano3 facility for her work in EBL writing and PMMA development. We thank Dr. Xuekun Lu for his suggestions on troubleshooting ALD uniformity. We thank Timo Meerloo, who formerly staffed the TEM, for lending us the spare specimen holder and tip so that we could design our custom chip holder, and for instruction on TEM usage. We thank Chawina De-Eknamkul for discussions around fabrication such as resist selection, lift-off and stripping procedures. We thank Dr. Wayne Yang for discussions on silicon nitride windows as alignment markers.

We thank Dr. Dan Read, the Raith Velion operator at University of California, Santa Barbara, for writing pores on multichips that we sent.

This work was supported in part by the NIH (5R21CA223727-02, 5R01GM126013-04) and was performed in part at the San Diego Nanotechnology Infrastructure of the University of California, San Diego, a member of the National Nanotechnology Coordinated Infrastructure, which is supported by National Science Foundation Grant ECCS-1542148.

Chapter 2 contains unpublished material co-authored with Huang, Xiaohua. The dissertation author was the primary author of this chapter.

# 3 Telomere enrichment by magnetic beads

## 3.1 Introduction

Specific amplification and identification of the telomeres requires prior knowledge of the subtelomeres. However, high quality subtelomere assemblies are only available for a handful of cell lines. In order for telomere amplification to be transferable to multiple cell lines, either of two conditions must be satisfied: 1) there needs to be enough homology between cells such that the location of the priming site is conserved, or 2) subtelomere sequences for cell lines must be known. We’ve seen from previous studies that (1) may only be true for cells from the same super-population (e.g. Africans, Americans, East Asians, Europeans, South Asians) (*76*). High quality subtelomere-containing assemblies are not readily available. Only recently has a complete telomere-to-telomere assembly been produced (*77*), and even the latest version of the human genome reference (hg38) has gaps in several subtelomeres (*78*).

Widespread application of subtelomere amplification therefore requires subtelomere sequences for cell lines from multiple super-populations. In other words, there exists a need for a method that facilitates gathering of subtelomere sequences. Of the sequencing methods available, nanopore sequencing was chosen because of its ability to produce long reads (>50 kb) which would facilitate subtelomere assembly. However, nanopore sequencing is costly. We hypothesized that enrichment by physical/chemical means can reduce the sequencing effort needed to obtain an actionable subtelomere sequence. We planned to do this by capturing telomere-containing fragments with biotinylated oligonucleotides on streptavidin-coated magnetic beads. We chose to investigate the feasibility of this method on GM12878 because successful enrichment could be confirmed by existing sequence data.

## 3.2 Experimental design

### 3.2.1 Enrichment constraints due to yield and fragment length

The success of enrichment depends on several factors that require balancing. Directly related to the preceeding step is the input required for nanopore sequencing by adapter ligation, which is stated as 1000 ng of high molecular weight (HMW) genomic DNA or 100-200 fmol. Using that mass to mole relation, ONT defines HMW dsDNA as having an average length between 7.7 kbp to 15 kbp. For subtelomere mapping, we may require average lengths of 50 kbp or higher, so we will set the mol quantity as a requirement.

Higher fragment sizes are ideal for mapping, but as fragment size increases, the number of fragments decreases. One would imagine that this is not relevant, since the number of telomere-containing fragments (targets) only depends on the input amount. However, the ONT sequencing platform has low tolerance for sample loads outside of the specification. Below the limit, the base output rate appears to be non-linear with respect to sample load, dropping significantly when load is below the recommended lower limit (data not shown). Enrichment is likely to be non-ideal, in which background molecules will still be present. Given this, a larger fragment size is likely to reduce sequencing yield by way of underloading.

This can be compensated by increasing the starting amount at the beginning of pulldown, but this would cause issues due to the non-Newtonian properties of concentrated genomic DNA. Successful enrichment depends on uniform dispersion and reconcentration of magnetic beads. High viscosity can hinder these processes, leading to difficulties in completing the procedure.

This also can be compensated by increasing the volume of the enrichment process, but again this causes issues due to binding kinetics of long biotinylated molecules and streptavidin beads. Finally, kinetics can be dealt with by using a higher concentration of beads, but this adds to the cost of an enrichment. With all these requirements in mind, the parameter space becomes tightly constrained.

Table 2.5 gives examples of enrichment outcomes given various fragment sizes and enrichment factors. Here, the enrichment factor is defined as the fold increase in fraction of targets. For example, an enrichment factor of 5 represents an increase from 1/10000 to 5/10000. In this case, we also assume ideal recovery, meaning no depletion of targets. Target depletion can occur when targets are not captured by beads or when experimental conditions cause targets to unbind and get washed away. With this assumption, we define the total fragment yield as the sum of the moles of target contained in the starting quantity and the moles of background fragments remaining after depletion. We desire fragments to be 50 kbp or greater, so already with an input of 100 µg gDNA, enrichment factors above 10 become incompatible with nanopore sequencing.

Table 3.1: Scenarios of elution composition and quantity after enrichment. Scenarios with yields greater than 100 fmol in bold. Ideal recovery is assumed.

| Input mass | Average fragment length | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 100 µg | 5 kbp | | 10 kbp | | 50 kbp | | 100 kbp | |
| Enrichment factor | mol ratio of telomere fragments, mol of total fragments | | | | | | | |
| 1 | **0.01%** | **31 pmol** | **0.01%** | **15 pmol** | **0.07%** | **3.1 pmol** | **0.14%** | **1.5 pmol** |
| 10 | **0.07%** | **3.1 pmol** | **0.14%** | **1.5 pmol** | **0.72%** | **310 fmol** | **1.4%** | **150 fmol** |
| 100 | **0.72%** | **310 fmol** | **1.4%** | **150 fmol** | 7.2% | 31 fmol | 14% | 15 fmol |
| 1000 | 7.2% | 31 fmol | 14% | 15 fmol | 72% | 3.1 fmol | 140% | 1.5 fmol |

### 3.2.2 Kinetics model of enrichment by bead capture

Of relevance is the biotin-streptavidin binding timescale when biotin is tethered to ultralong dsDNA and streptavidin is anchored to microbeads. Whether or not binding occurs on a practical timescale determines what concentrations are acceptable for attempting enrichment.

A set of binding constants and forward rate constants were previously generated by Huang et al for biotin-tethered dsDNA ranging from 100 to 5000 bp and streptavidin coated polystyrene (PS) beads ranging from 90 to 944 nm (*79*). Near 10-fold increase in binding constant and 5-fold increase in forward rate constant is seen when bead size is decreased from 944 nm to 90 nm. However, as the bead size decreases, so does the magnetic force on the bead since it is proportional to the volume of the bead (*80*). While the Stokes' drag would also decrease for smaller beads, the decrease in magnetic force will be greater because of the difference in diameter dependence ($d$ versus $d^3$). For our beads, we chose Dynabeads MyOne Streptavidin C1 (C1), which are 1 µm in diameter.

Given the potential of high viscosity for concentrated gDNA samples, we anticipated that using larger beads would provide a greater magnetic force and facilitate the completion of pulldown.

We used the data and model from Huang et al. to estimate the binding constant for our system. Equation 1 gives the rate law for irreversible binding of biotinylated dsDNA *[B]* with binding sites *[S]*, where $k_f$ is the forward rate constant.

$$\frac{d[SB]}{dt} = k_f[S][B] \qquad [1]$$

Solving the differential equation yields:

$$\ln\frac{([S]_0 - [SB])[B]_0}{[S]_0([B]_0 - [SB])} = ([S]_0 - [B]_0)k_f t \qquad [2]$$

We can then define a characteristic time, $t = \tau$, such that half of all biotinylated dsDNA is bound, or $[SB] = \frac{1}{2}[B]_0$. Solving for $\tau$ yields:

$$\tau = \frac{\ln(2 - [B]_0/[S]_0}{k_f([S]_0 - [B]_0)} \qquad [3]$$

Fitting the kinetics data from Huang et al to multiple models using the Matlab Curve Fitting Toolbox revealed that a two-term exponential ($f(x) = ae^{bx}+ce^{dx}$) gave the best fit ($R^2$=0.9999, Figure 2.X). This was not pointed out by the original authors and, if representative, suggests that two independent factors contribute to the forward rate constant. Setting the terms to be equal and solving for length *x* reveals that the both terms contribute equally at 231 bp. Above this, the second term dominates. Whether or not this is related to the RD effect described previously is outside the scope of this work.

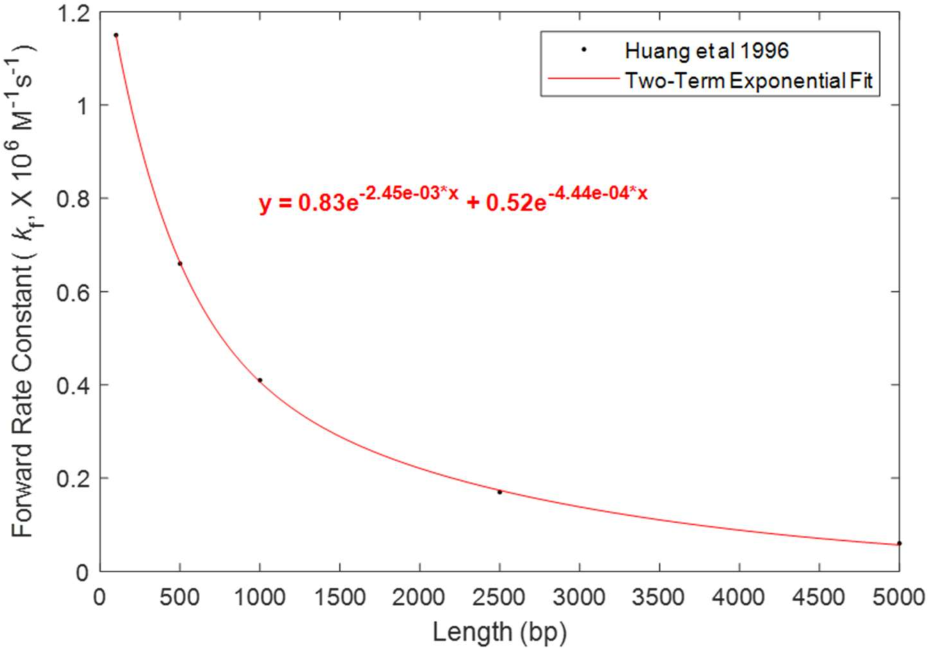


Figure 3.2: Two-term exponential fit of binding kinetics data from Huang et al.

Nonetheless, using this model to extrapolate kinetics constants shows that above 10 kbp, binding may not occur on a practical timescale. For example, we can estimate the biotin and streptavidin concentrations as follows. In a high-volume scenario, we may have 500 µg of gDNA in 15 mL of buffer. With this much gDNA, approximately 83 million diploid genomes or 850 fM telomere-containing fragments are present. For simplicity, we assume that all telomeres are hybridized to a biotin probe and therefore 850 fM of biotinylated dsDNA are present. C1 beads have a free biotin binding capacity of 2500 pmol/mg. We use the free biotin binding capacity instead of the dsDNA binding capacity because steric effects are already accounted for by the kinetics model. The bead stock concentration is 10 mg/mL which is equivalent to a binding site concentration of 25 µM. If we use 150 µL of beads in this high-volume scenario, this gives a binding site concentration of 250 nM. Using this concentration with a biotinylated dsDNA concentration of 850 fM and the extrapolated rate constants for longer DNA, the characteristic times are 48 seconds, 7.5 minutes, and 10 hours for 5 kbp, 10 kbp, and 20 kbp targets, respectively (Table 3.2). We can also examine a 200 µL scenario with 100 µg of gDNA. Here, the telomere concentration is 6.4 pM. Starting with a bead volume of 50 µL leaves us with a final concentration of 6.25 µM binding sites. At these concentrations, we estimate a characteristic time of 25 min for 20 kbp, which is experimentally feasible. However, difficulties due to sample viscosity may be present in this regime.

Table 3.2: Characteristic time of binding for selected biotinylated dsDNA and binding site concentrations.

| Reaction Volume | Analyte length (bp) | | 1,000 | 5,000 | 10,000 | 20,000 | Bead Cost* (USD) |
|---|---|---|---|---|---|---|---|
| | $k_f$, $M^{-1}$ $s^{-1}$ | | $4.07×10^5$ | $5.67×10^4$ | $6.14×10^3$ | $7.22×10^1$ | |
| | [B]_0 | [S]_0 | Characteristic time τ | | | | |
| 15 mL | 850 fM | 125 nM | 13 s | 97 s | 15 min | 21 hr | $19 |
| | 850 fM | 250 nM | 6.8 s | 48 s | 7.5 min | 10 hr | $38 |
| | 850 fM | 500 nM | 3.4 s | 24 s | 3.7 min | 5 hr | $75 |
| 200 µL | 6.4 pM | 500 nM | 3.4 s | 24 s | 3.7 min | 5 hr | $1 |
| | 6.4 pM | 3.13 µM | 0.54 s | 3.9 s | 36 s | 51 min | $6 |
| | 6.4 pM | 6.25 µM | 0.27 s | 1.9 s | 18 s | 25 min | $13 |
| | 6.4 pM | 12.5 µM | 0.13 s | 0.97 s | 9 s | 12 min | $25 |

*based on a bead cost of $250 per mL.

Another consideration is the order in which the bead-probe-target complex is assembled. If separated into individual steps, we have the options of 1) free-solution hybridization of probes and targets followed by immobilization into beads versus 2) immobilization followed by hybridization. There is also the third option of mixing all components at the same time, but we lose out on potential advantages that a stepwise procedure may present. The mechanics behind the immobilization-first approach are described differently than in Huang et al. Rather, the kinetics of hybridization of DNA to surface-immobilized probes can be described by a combination of: 1) direct 3-dimensional (3D) diffusion of targets to the immobilized probe and 2) nonspecific adsorption of target to the surface followed by 2-dimensional diffusion to the probe (*81*). The latter is also known as reduction of dimensionality (RD) and can enhance the overall reaction rate. The relative contribution of the two factors depends on the target length and probe length. Although 3D diffusion appears to dominate for longer targets, this varies depending on the desorption rate (*81*). In the hybridization-first scheme, this two-factor dependence is likely to also be present since the association reaction is of biotin and streptavidin rather than two complementary DNA strands, but it has not been as well characterized.

The potential for the telomere overhang to form G-quadruplexes may complicate things further. Near physiological conditions, the telomere overhang may switch between an open ssDNA state and a folded G-quadruplex state, of which only the open state can be captured by probes. However, FRET hybridization studies by Ying et al have shown that the rate-determining step is hybridization rather than unfolding of the quadruplex (*82*). In addition, they report that hybridization of C and G strands can occur on the order of minutes at room temperature.

We decided to proceed with the immobilization-first approach because we anticipated that it would be more economical. Given 100 µg of gDNA, we have approximately 1.3 fmol of telomere ends. To guarantee hybridization, we planned to use a molar excess of probes (e.g. 400 pmol). In either scheme, we aim to saturate our beads since these are the costliest component. However, In the hybridization-first approach, it seemed likely that beads would become saturated with non-hybridized probes since: 1) they are in excess, 2) they are smaller than hybridized duplexes and thus able to access more binding sites, and 3) they have a higher diffusion constant. This may not seem problematic since those immobilized probes are still able to hybridize to free targets. Any targets in solution would have already been hybridized to free probes and are unlikely to be displaced by an immobilized probe. In the immobilization-first scheme, the matter of using beads to their full capacity is solved by pre-saturating with probes. The risk of generating free-solution probe-target duplexes is eliminated by rinsing away excess probes before adding targets. It may take longer for hybridization to occur since probes are spatially confined to beads, but there is also the possibility of rate enhancement from the RD effect.

## 3.3 Methods

### 3.3.1 Cell culture and HMW gDNA extraction

GM12878 was grown to 50% confluence in Roswell Park Memorial Institute Medium 1640 supplemented with 15% fetal bovine serum. Between 30–50 mL of cell suspension was resuspended with 10 mL of phosphate buffered saline (PBS) and then resuspended with 100 µL of PBS in a 50 mL centrifuge tube.

Lysis and extraction of gDNA was based on the protocol for HMW extraction from suspension cultures from Sambrook and Russell (*83*). 10 mL of Tris Lysis Buffer (Tris, NaCl, EDTA, SDS) was added to 100 µL of cell-PBS suspension, followed by vortexing for 5 s and incubating at 37°C for 1 hr. 50 µL of 20 mg/mL Proteinase K (Roche) was added, and the tube was slowly inverted 10 times. Tubes were incubated at 50°C for 3 hrs and mixed every hour by inverting 10 times. After cooling to room temperature, 10 mL of Tris-saturated Phenol was added and inverted for 10 minutes to form an emulsion. The tubes were centrifuged at 4000 rpm for 15 minutes at 22°C, and the aqueous phase was slowly pipetted off using a serological pipet. To the aqueous phase, 5 mL of Tris-saturated phenol and 5 mL chloroform:isoamyl alcohol 24:1 was added. The inversion and centrifugation was repeated, followed by slowly pipetting off the aqueous phase and transferring it to a new tube.

180 mL of GM12878 at 50% confluence was split between six 50 mL tubes and lysed using the protocol described previously. After extraction with phenol:chloroform, 4 mL of cold 5 M Ammonium acetate was added, followed by 30 mL of ethanol chilled to -20°C. Tubes were left at -20°C overnight. Three out of six tubes had visible threadlike precipitate, and the other three were mixed by inverting and rolling until a precipitate formed. Glass hooks were formed by heating a glass pipette over a bunsen burner and sterilized before usage by briefly passing the hook over the flame. Hooks were used to pull precipitate out of the tube, while letting liquid droplets shed off of the hooked strands. Precipitate was collected in 10 mL of 70% ethanol, and centrifuged at 4000 rpm for 5 minutes. The supernatant was decanted, and the pellet was resuspended with another 10 mL of 70% ethanol. Samples were centrifuged at 4000 rpm for 5 minutes, followed by decanting and leaving in vacuum to dry. After drying, the extract was wet and gel-like but was not pourable. 150 µL of TE was added to each tube and left at 4°C overnight. The DNA was quantified using a Nanodrop spectrophotometer. 1000 ng of extract was taken from this sample for nanopore sequencing by ligation (Oxford Nanopore, SQK-LSK109) on a MinION flow cell (r9.4.1). The sequencing data from this run would be used as a baseline for telomere and subtelomere enrichment.

To probe fragment length with solid state nanopores, DNA extract was resuspended to 50 ng/µL in 4 M LiCl, 10 mM Tris-HCl pH 7.5, and 1 mM EDTA. Due to its viscosity, the mixture was left at 4°C overnight to resuspend. This solution was loaded on the top-side (*cis*) of the membrane, with 4 M LiCl, 10 mM Tris-HCl pH 7.5, and 1 mM EDTA on the bottom-side. A positive bias of 200 mV was applied to the *trans* side using an Axon Axopatch 200B. Attempts to quantify fragment length by nanopore translocation were hindered by permanent blockade of the ionic current. In the absence of DNA-binding proteins, this clogging effect was likely due to the tendency of HMW DNA to form knots (*84*, *85*).

### 3.3.2 Telomere enrichment from TE-resuspended DNA extract (TPD1)

50 µL of Dynabeads MyOne Streptavidin C1 and 50 µL of Wash buffer was added to a new tube. The beads were vortexed, followed by placing the tube on a magnet and decanting the supernatant. The beads were then resuspended with 25 µL of 16 µM Teloprobe6 in Wash buffer. This was placed on an inversion mixer for 5 minutes, followed by placing on a magnet and decanting the supernatant. The beads were resuspended with 50 µL of Wash buffer, which was then decanted after applying a magnet. This wash step was repeated, followed by resuspending again in 50 µL of Wash buffer. 100 µg of gDNA was brought to 200 µL with Wash buffer and left at 65°C for 5 min followed by 20°C for 3 min. The DNA solution was combined with the bead solution and placed on an inversion mixer.

At this point we experienced difficulties in continuing the process. Upon resuspension of the beads, the beads formed a coating around a mucousy aggregate, presumably the DNA. To avoid shearing the DNA, the tube was placed on an inversion mixer to resuspend the beads. However, after 18 hours of inversion at room temperature, no change in solution homogeneity was observed. Application of a magnet would partially displace the beads, but they would immediately return to their original position after removal of the magnet. We did find that the beads could be aggregated by centrifugation at 16100 g for 15 minutes, after which the beads could remain held in place during buffer exchange while a magnet was applied. As much buffer was removed as possible without disturbing the beads, and the beads were resuspended with 100 µL of Wash buffer. This was repeated once with 100 µL Wash buffer, twice with 100 µL ice cold LS buffer, and twice with 100 µL ice cold nuclease-free water. To elute, the beads were mixed with 25 µL 1X TE at 37°C and incubated for 10 minutes at 37°C. This was repeated again and the elutions were combined. The elution was quantified by using a Nanodrop UV-vis spectrophotometer, and 1000 ng of product was sequenced on a MinION (r9.4.1) for 16 hours using the Oxford Nanopore Ligation Kit (SQK-LSK109).

### 3.3.3 Telomere enrichment directly after phenol:chloroform extraction (TPD2, TPD3)

Execution of the first telomere pulldown (TPD1) was severely hindered by the viscosity of the sample. Based on earlier kinetics estimations, we anticipated that increasing the reaction volume would still allow a reasonable capture time.

150 mL of GM12878 at 50% confluence was resuspended in 200 µL PBS. From this, extraction was performed as described in Section 3.3 using 10 mL TLB, 10 mL TE-saturated phenol, and 10 mL phenol:chloroform:isoamyl alcohol (25:24:1). The aqueous phase was transferred to a new 50 mL centrifuge tube. From here, we proceeded with our second enrichment (TPD2). To this, we added 3 mL of 2 M NaCl, 20 µL of 20 µM TeloProbe6 (to 400 pmol), and 200 µL of 10 mg/mL Dynabeads MyOne Streptavidin C1 (ThermoFisher). The tube was inverted for 30 minutes at room temperature. A uniform bead suspension was observed at 6 minutes. Following incubation, a 5 mm x 5 mm x 10 mm neodymium magnet was taped to the bottom of the tube and left at 4°C for 1 hr and 40 minutes. The solution was slowly decanted using a serological pipet. Approximately ~1 mL of solution could not be decanted without disturbing the pellet. The beads were

rinsed twice with 5 mL of LS buffer. During these rinses, the beads were dispersed after 4 minutes. The first rinse required 20 min to re-pellet the bead while the second rinse required 5 min. The beads were rinsed twice with 5 mL of nuclease free water without resuspending or disturbing the beads. We then resuspended the beads with 200 µL of heated TE, followed by incubation at room temperature for 5 minutes. The beads were re-pelleted and the eluate was collected. The eluate concentration was measured as 111 ng/µL using a Nanodrop Spectrophotometer.

From the eluate, libraries were prepared using the Oxford Nanopore Ligation Kit (SQK-LSK109) according to manufacturer instructions and quantified using the Qubit 1X dsDNA HS Assay (Thermofisher). Due to major losses in Qubit yield after cleanup step (near 1%), several libraries were prepared with slightly different input or processing parameters (Table 2.5.1).

To compensate for low yields, enrichment was repeated this way but with a starting volume of 300 mL cells at 50% confluence, followed by pooling the aqueous phases into a single 50 mL tube (TPD3). However, due to the high concentration of DNA present, the beads could not be easily dispersed or concentrated. After initial incubation with Teloprobe6 in Wash buffer, a magnet was taped to the tube and left for two hours at 4°C. Following this, the beads were only partially aggregated, occupying roughly the lower ½ of the total volume. The upper volume was carefully aspirated and replaced with LS buffer, and only about half of the total volume could be replaced. To aggregate the beads, the magnet was left on the tube for 72 hours at 4°C. The supernatant was aspirated off, and the beads were rinsed twice with water without resuspension. After adding TE, we found that the beads were highly aggregated and could not be dispersed with vortexing or pipetting. We therefore had little confidence in recovering any DNA by resuspension. We decided to use the DNA from the $H_2O$ elutions for sequencing.

Initial attempts to prepare libraries by ligation gave poor recovery (1.5%) similar to previous experiments. We decided to make use of the high concentration of this sample by subjecting it to the transposase-based rapid library kit, which has been shown to provide the longest read lengths and not require any in-process cleanup steps (*86*).

## 3.4 Results and Discussion

### 3.4.1 Sequencing yields

Sequencing with material from TPD1 and TPD3 proceeded without any major issues, as the amount of starting material was more than sufficient. Here, we will focus on the throughput of individual Flongle runs using material from TPD2, and the effect that specific sample treatments had.

Library preparation with TPD2 was consistently marked by a sharp decrease in yield. In the first run, despite 500 ng of input, the output was below the detection limit of the Qubit. During sequencing, we observed low pore occupancy (1 in 9 active pores sequencing) and a sharp decrease in active pores after 9 hours. Reductions in active pores can result from underloading (manufacturer communication). A majority of reads mapped to Lambda and gave partial coverage, indicating sequencing of the calibration strand.

Scaling to 2500 ng in the second run, the output was 20 ng, or slightly under 1% recovery. This also resulted in similar sequencing yield to the first run. A conservative estimate for the recovery of each cleanup step would be 50% so the yield after library preparation should be at least 25% since it entails two cleanups. Despite the use of two different means of quantification (Nanodrop for input, Qubit for output), such a discrepancy is outside the variation between the two measurements (*87*). It then seemed that either the beads were nonfunctional or that the enriched sample was not amenable to cleanup. We confirmed that the beads were capable of recovery using human genomic DNA (Promega) and various conditions to simulate cleanup after library preparation steps (data not shown).

To see if there was an issue with the sample, we re-did library preparation a fourth time but the addition of 100 ng Lambda DNA. Pore occupancy was improved in this sequencing run, staying above 2 out of 3 active pores. In contrast to earlier runs, the number of active pores steadily declined throughout the 24 hours, and the flowcell yielded nearly 10 times more reads. However, as evidenced by mapping, 89% of reads mapped to Lambda and only 1% mapped to CHM13. In addition, mapping to Lambda gave complete coverage, as opposed to the partial coverage in other runs that results from sequencing of calibration strands. This indicates that the improved sequencing yield was due to addition of Lambda DNA, but it was disproportionate to the amount of Lambda DNA included, even after accounting for MW/molarity differences.

To improve sequencing yield in the fifth run, we scaled the library preparation reaction up by using protocols for MinION rather than Flongle, and we used 5120 ng of enriched sample. To prevent incomplete ligation, we also increased end repair/end-prep cycle times from 5 minutes to 30 minutes, and increased Ligation time from 10 minutes to 45 minutes. After cleanup, the yield was 64 ng, which is consistent with the ~1%

recovery seen in previous runs. The pore occupancy was better than most other runs, at 1 in 4. In the sixth run, 2600 ng of DNA was sheared by submitting the solution to 5 aspiration and dispense cycles through a 28G needle (I.D. 0.184 mm) before starting library preparation. This run gave the highest yield in bases mapping to CHM13, as well as the highest proportion of reads mapping to CHM13 as opposed to Lambda. Despite shearing, average read lengths were comparable to previous runs.

As evidenced by Qubit yields and high proportion of reads mapping to Lambda, the samples enriched in this manner were not amenable to Ampure XP cleanup. Despite enrichment by at least 10X, the incompatibility with cleanup offsets any gains in subtelomere coverage. A total of 6.3 Mb were mapped to the subtelomere regions, representing a sequencing depth of 0.69X.

### 3.4.2 Basecalling, correction, mapping, and analysis of enrichment

Initially, all reads (FAST5 files) were basecalled using Guppy (v6.0.1) in high-accuracy mode. Visual inspection of reads showed that some contained highly repetitive non-telomeric repeats such as $(TTAAAA)_n$. Because we expect telomeric repeats to be present, there is a possibility that these repeats are miscalled. This could be due to a fundamental aspect of nanopore sequencing–current blockades are proportional to analyte size. Translocation of purines (A, G) should have similar blockades, and the same goes for pyrimidines (T, C). Tan et al identified the miscall frequency and types of miscalls that can occur, in addition to providing a workflow for identifying and correcting such reads (*88*). In brief, error-prone reads are identified based on the frequency of common miscalls. Those reads are then extracted from the original FAST5 file and basecalled using Bonito (v0.3.5), a developmental-phase basecaller. Unlike Guppy, Bonito can be trained by end-users. Tan et al found that training Bonito with a set of ground truth telomeric sequences extracted from the CHM13 reference genome provides an improvement in telomere repeat calling accuracy. We also found that using Guppy with super-accurate (sup) mode partially reduced miscalling errors. In our final workflow, we ran Guppy-sup for first-round basecalling, then used the error-correcting workflow to identify and re-basecall reads that were error prone. All reads were then mapped to the CHM13 reference with *minimap2* using the nanopore to reference preset. Reads were also mapped to the Lambda DNA sequence to check Lambda coverage.

For estimation of subtelomere enrichment, we defined the telomere region to be the distal 20 kb of each chromosome, the subtelomere to be the distal 20 kb to 200 kb, and the remainder to be the regions between subtelomeres plus mitochondrial DNA. Coverage and depth per position were calculated using samtools and bedtools. The subtelomere to remainder (ST:R) ratio is the number of bases mapped to the subtelomeres divided by the number of bases mapped to the remainder. The ST enrichment ratio is then defined as the fold increase in ST:R compared to the ratio of the region sizes (2.66E-03).

Table 3.2: Summary of results from telomere enrichment directly after phenol:chloroform extraction (TPD2).

| Sequencing run | TPD2-1 | TPD2-2 | TPD2-4 | TPD2-5 | TPD-6 |
|---|---|---|---|---|---|
| Input (ng) | 500 | 2500 | 500 | 5120 | 2600 |
| Protocol modification | none | none | +100 ng Lambda | extended reactions | sheared input |
| Ligation protocol | Flongle | Flongle | Flongle | MinION | Flongle |
| Flowcell | Flongle | Flongle | Flongle | Flongle | Flongle |
| Runtime (hrs) | 24 | 24 | 24 | 24 | 17 |
| Output (ng) | under | 20 | n.a | 64 | 16 |
| Total # of reads | 14752 | 8749 | 125017 | 49330 | 31384 |
| %Reads mapped to CHM13 | 5% | 23% | 1% | 19% | 26% |
| %Reads mapped to Lambda | 80% | 55% | 89% | 74% | 70% |
| %Coverage on Lambda | 15.39% | 7.92% | 100% | 14.36% | 12.08% |
| Bases, distal 200kb | 301,987 | 984,844 | 2,650 | 2,134,527 | 2,916,156 |
| Distal 20kb | 52,287 | 94,326 | 441 | 156,451 | 329,108 |
| Remainder | 6,736,304 | 28,683,083 | 2,435,905 | 62,205,661 | 77,992,535 |
| ST:R | 0.0448 | 0.0343 | 0.0011 | 0.0343 | 0.0374 |
| ST enrichment | 14.4 | 11.0 | 0.35 | 11.0 | 12.0 |
| Max length | 85793 | 149327 | 114287 | 161878 | 92569 |
| Median length | 3476 | 3391 | 3514 | 3484 | 3463 |
| Average length | 3311 | 6820 | 5543 | 4339 | 5587 |
| N50 | 3524 | 31253 | 8363 | 3538 | 11026 |

### 3.4.3 Telomere length measurement

After correction with the tuned guppy basecaller, telomere-containing reads were identified using a custom Matlab script. This was done by finding all instances of “GGG” or “CCC” within a read and assigning a value of one to each location per instance. The resulting signals were then filtered with a moving average window of 90 points or 15 6-base repeats, and rescaled by multiplying by 6, the expected size of a repeat. The transformed signals are such that a value of 1 corresponds to continuous tract of “NNNGGG” or “NNNCCC” repeats. We applied this transformation to all reads to obtain signal pairs (one for “GGG” and one for “CCC”) for each read. Reads were then filtered out if both signals had less than 500 positions corresponding to a repeat. For reads with high “GGG” content, the telomere boundary was defined as the furthest (towards 3’) position at which the signal crosses 0.5. This boundary could not be found for a small fraction of reads because they were entirely telomeric. For reads that did have a boundary, the telomere length was appended to the sequence header in the fasta file and the read was truncated so that only the subtelomeric portion remained. These truncated reads would form the query for local alignment as described later.

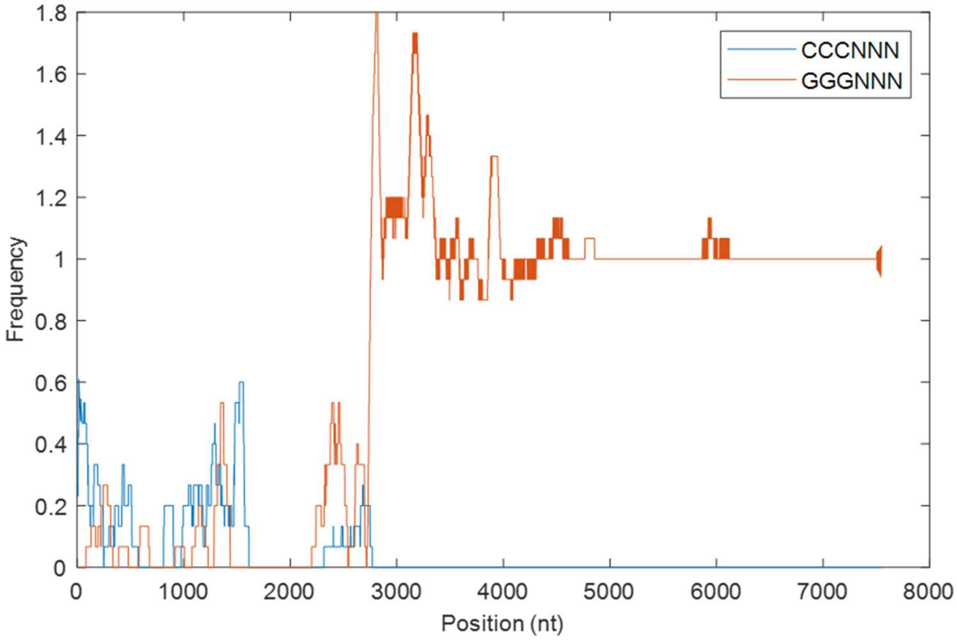


Figure 3.3: Representative trace of read data after telomere signal transform. A value of 1 corresponds to a telomere-like sequence.

Interestingly, no telomere C strands were found. This is likely due to the position at which sequence adapters are attached, regardless of library preparation method. For any telomere-containing fragment, the telomere overhang is likely to remain unmodified during the end repair and end preparation reaction prior to ligation. Therefore, the overhang remains incompatible with T/A ligation and this fragment would only be ligatable from the centromere-facing end. This exclusively places the motor protein on the G strand. Transposase-based library preparation for ultralong reads would produce similar results. In this setting, random fragmentation of HMW genomic DNA by transposase only places sequencing adapters within the sequence; the transposase selected for library preparation cannot act on ssDNA overhangs unless they form hairpin loops (*89*). This may also explain the lower than expected frequency of telomere-containing reads in WGS data. However, there are likely other factors at play since suppression of all C strands would only decrease representation of telomeres by half. In an unenriched WGS run, we observed telomere representation in 3 in 539,665 reads which is nearly ten times less than what we would expect if reads were randomly sampled from the genome (Table 3.2). The WGS effort by M. Jain et al also gave a similar proportion of telomere-containing reads, 140 out of 14 million.

Sequencing runs with enrichment had higher proportions of telomere reads compared to unenriched runs. TPD1 and TPD2 had greater telomere enrichment than TPD3 (18.0-20.9 vs. 5.74). This may be due to the efficacy of rinsing with the two earlier enrichments compared to the latter.

Table 3.3: Summary of subtelomere and telomere enrichment from sequencing data.

| Experiment | WGS | TPD1 | TPD2 | TPD3 | M. Jain et al 2018 |
|---|---|---|---|---|---|
| Flowcells | 1 MinION 24h* | 1 MinION 24h | 5 Flongles | 1 MinION 72h | 39 MinIONs |
| Library method | Ligation | Ligation | Ligation | Ultralong | Mixed |
| Total reads | 539,665 | 765,286 | 225,022 | 1,107,629 | 14,183,584 |
| Mean read length | 2684 | 4468 | 5240 | 9037 | 7214 |
| ST mapped bases | 4.69 Mb | 7.72 Mb | 5.99 Mb | 37.0 Mb | - |
| Remainder mapped bases | 1.50 Gb | 2.87 Gb | 195 Mb | 9.53 Gb | - |
| ST:R | $3.12\times10^{-3}$ | $2.69\times10^{-3}$ | $3.08\times10^{-2}$ | $3.88\times10^{-3}$ | - |
| Normalized ST enrichment | 1.0 | 0.86 | 9.9 | 1.25 | - |
| ST coverage | 0.5X | 0.8X | 0.65X | 4.0X | - |
| Approximate genome coverage | 0.465X | 1.10X | 0.378X | 3.21X | 35X |
| Expected telomere reads | 21.4 | 50.5 | 17.4 | 148 | 1610 |
| Actual telomere reads | 3 | 148 | 44 | 119 | 140 |
| Actual/Expected | 0.1 | 2.93 | 2.53 | 0.806 | 0.0870 |
| Normalized Telomere Enrichment | 1.0 | 20.9 | 18.0 | 5.74 | 0.9 |

*The MinION is advertised to run up to 72h before throughput drops to near zero.

For ease of verification, we performed local alignment between reads and subtelomeres using the Smith-Waterman algorithm built into Matlab. To generate our subtelomere dataset, we first mapped the 35X coverage contigs from GM12878 (*86*) to a complete reference, T2T-CHM13v2.0 (*77*), using *minimap2* with assembly to reference presets (*-x asm5*). *bcftools* was used to call variants and form a consensus. We then applied the telomere signal transform on the distal 20 kbp of each chromosome of the consensus to find the subtelomere-telomere boundary. The distal 1 kbp of each subtelomere was extracted from the boundary towards the centromere to form our targets for local alignment. Each read was aligned with each truncated subtelomere, and the optimal alignment was chosen based on the highest alignment score. A score cutoff of 2500 was determined by visually inspecting the alignment strings of each optimal alignment and because it also captured a cluster of high-scoring alignments (Figure 3.3). Out of 192 telomere reads (TPD1 and TPD2), 62 reads could be mapped this way.

We report several additional telomere lengths for specific chromosome arms this way. Surprisingly, there was not substantial overlap in reported chromosomes between what we report here and the ultralong sequencing work by M. Jain et al. Out of 14 previously reported arms, we found telomere lengths for 5. We also found telomere lengths for 12 arms not previously reported. We found agreement in telomere lengths on 5q, 9p, 18p, and 19q, in addition to some heterogeneity in 10p, 19q, and others (Figure 3.3). This is likely to be due to a combination of cell to cell variation in telomere length in addition to variation between homologous chromosomes.

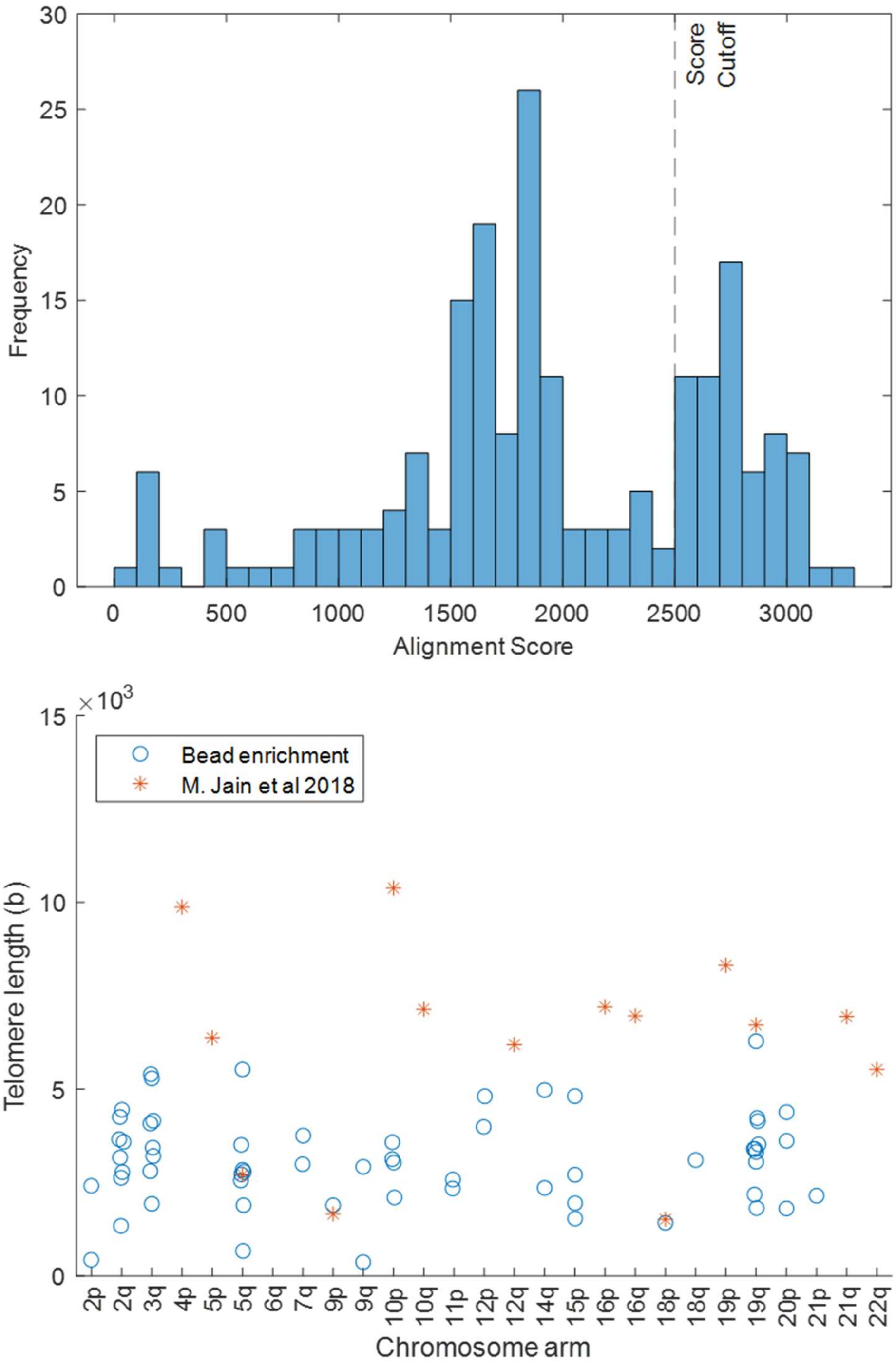


Figure 3.3: (Top) Alignment score distribution of telomere reads vs. subtelomeres. (Bottom) Telomere lengths reported after filtering by alignment score overlaid with telomere lengths reported by M. Jain et al.

In conclusion, we report a method for enrichment of telomere and subtelomere reads using biotinylated oligonucleotide probes and paramagnetic streptavidin microbeads. We investigated the feasibility and effectiveness of enrichment using an immobilization-first approach, where biotin probes are irreversibly bound to streptavidin beads before proceeding with hybridization to targets. We did so with three different sets of volume-concentration parameters, and found that the high volume-dilute gDNA trial was both the most feasible to complete as well as highly effective in telomere and subtelomere enrichment. However, there remains room for improvement as proceeding this way also gave reduced sequencing yields. We accumulated an additional 55.4 Mb on our ST region spanning 8.28 Mb, representing a theoretical coverage of 6.69X, and we found additional chromosome-specific telomere lengths for the cell line GM12878.

### 3.5 Acknowledgements

This work was supported in part by the NCI (5R21CA223727-02).

Chapter 3 contains unpublished material co-authored with Huang, Xiaohua. The dissertation author was the primary author of this chapter.

# 4 Amplification of telomeres for single cell sequencing

## 4.1 Introduction

As discussed previously, single cell methods for telomere length measurement have their limitations. Current methods for measuring at the single cell level sacrifice single base resolution and chromosome resolution. We hypothesize that single cell processing, telomere amplification, and long read sequencing would enable us to regain these. Microfluidic single cell processors have matured thoroughly since 2010 (*90*), but there is a lack of telomere amplification methods that would be transferable to a microfluidic format. An ideal amplification would generate whole copies of telomeres along with a length of subtelomere that is sufficient for chromosome identification.

The subtelomere is the region between telomeres and chromosome-specific DNA (Figure 1.2). It spans roughly 10-300 kb and is described as a mosaic or patchwork of sequences that can be found on multiple chromosome arms, also known as paralogous blocks (*91–93*). Copies of a paralogous block can have 88-99.9% identity between chromosomes (*94*). The presence of common paralogous blocks occupying large portions of the subtelomeres makes it challenging to find chromosome-specific subtelomere sequences. In addition to paralogous blocks, subtelomeres contain degenerate telomere repeats and regions with a high density of 5’-CG-3’ repeats, also known as CpG islands.

Subtelomere variation can also be caused by recurrent inter-chromosomal interactions. 4q and 10q share highly similar repeat arrays, and these subtelomeres were found to be swapped in roughly 20% of a Dutch population (*95*). Some individuals are mosaic for 4q/10q subtelomeric translocations, indicating that subtelomeric sequences can interchange in somatic cells (*96*).

Specific amplification of a sequence requires prior knowledge of the target. To investigate the feasibility of amplification, we chose the cell line GM12878 as our target. Others have previously performed several sequencing studies on this cell line (*86*), providing a high confidence assembly from which we can design our primers.

An amplification method that would be applicable to single cells would ideally be isothermal, as this would facilitate the transfer of a method developed from a tube format to a microfluidic format. Although solutions exist to thermocycle samples in microfluidic devices, these would present additional engineering challenges when combined with a single-cell capturing device. Even in a scenario where a single telomere is targeted, denaturation at high temperature would reveal the entirety of the telomere length on which the C-strand primer can hybridize. This can be overcome by ligating a priming region to the 3’ end of the G-strand, as in TeSLA. However, this process is time consuming, requiring two 12-16 hr ligations in the case of TeSLA and U-STELA (*97*).  Unless there is a more efficient way to anchor the C-strand primer to the 3’ end of the telomere G-strand, we believe the probability of PCR being applicable to this task to be low. Thus, we chose to investigate isothermal methods.

### 4.2 Design of microfluidic single cell processor with integrated nanopore chip

The basic requirements of a single cell processor are modules for: 1) single cell isolation from a cell suspension, 2) cell lysis and DNA cleanup, and 3) amplification. We adapted elements from microfluidic devices previously designed in our laboratory (*57*, *98*, *99*). Figure 4.1 shows a schematic of the center of such a device. Outside of this region, fluidic and valve channels splay out and terminate at 0.75 mm punch ports.

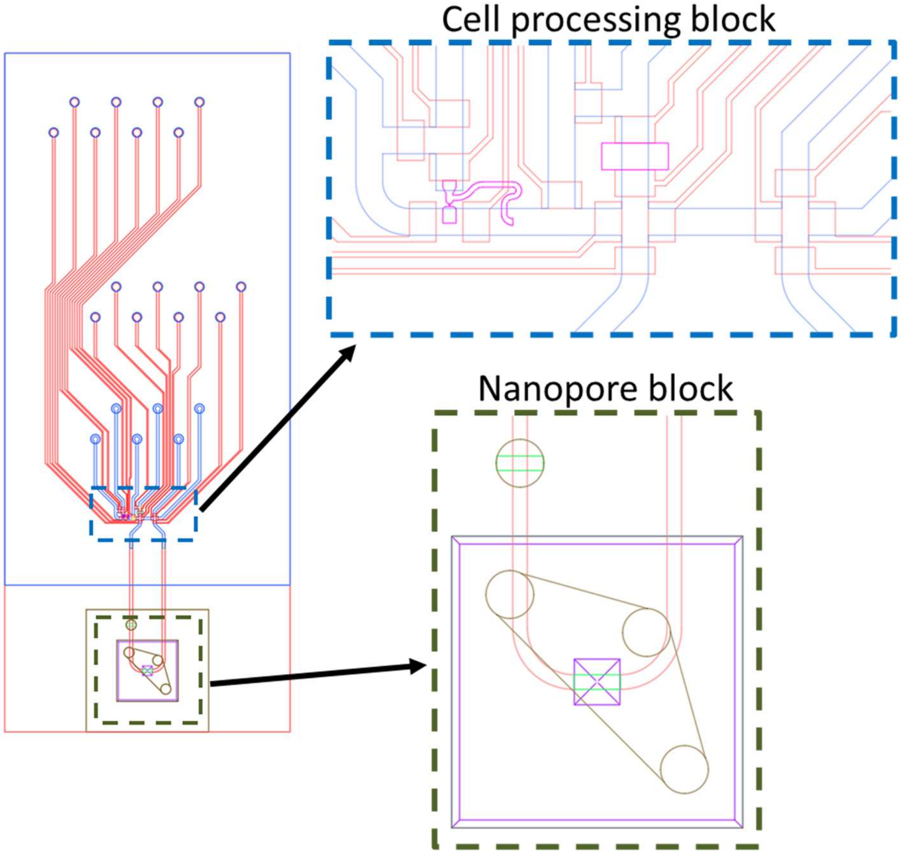


Figure 4.1: Center of activity of microfluidic processor. (red) Valve layer, (blue) ~20 µm tall rounded fluidic channel (magenta) 40 µm tall fluidic channel with rectangular cross section, (green) 80 µm tall rectangular fluidic channel.

For single cell capture, we use a sieve. The dimensions of the sieve are such that pressure driven flow preferentially convects through the constriction when no cell is present. When a cell is captured, it acts as a check valve and prevents fluid flow through the constriction. Washing with phosphate buffered saline (PBS) removes excess cells in the fluid path by pushing them through the alternate path towards a dedicated waste line, while the captured cell is held in place due to the pressure gradient across the constriction.

For cell lysis and DNA cleanup, we use a polyethylene glycol-polyacrylamide copolymer barrier. The composition of this copolymer can be tuned such that it is permeable to water, ions, and small molecules, but not larger dsDNA fragments. After washing out excess cells, the single captured cell can be released from the sieve by applying flow in the opposite direction. Convective transport to the barrier vicinity would then be followed by electrophoretic capture by the polymer barrier. This enables solution exchange without loss of sample. This way, we can proceed with cell lysis with detergent-containing buffers and follow up with Proteinase K digestion and restriction digestion if necessary. Compounds that may interfere with amplification such as detergents, denaturants, and chaotropes can be washed away without dilution of sample.

Lastly, we have a dedicated reaction chamber for amplification. Amplification in a confined nanoscale volume is especially reagent-limited. Designing a larger volume chamber for amplification allows us to tune the chamber volume to meet the mole requirements for amplification. We also incorporate in-line mixing by peristaltic action to enable uniform distribution of the sample when diluting with amplification mastermix.

Fabrication of the device would follow established procedures, where polydimethylsiloxane (PDMS) is cast using molds patterned by photolithography. PDMS layers are then aligned, bonded, and the stack bonded to glass.

We envision that an easy to fabricate all-in-one system would also include a separate PDMS block with a nanopore chip housed inside. The nanopore block is separate from the cell processing block because the mold fabrication techniques for one may not be compatible with the other. For instance, the channel heights in the cell processing block are at most 40 μm, which photolithography with SU-8 is well-suited for. On the other hand, the channel height required to house a nanopore chip is at least the thickness of the chip itself or 200 μm. This may be doable with multiple SU-8 layers, but we risk poor sealing since the height of tall structures patterned by viscous SU-8 is generally nonuniform. A more appropriate method for mold fabrication at this size range would be computer-numerical-controlled micro-milling.

In the two-block design, vias patterned in the valve layer would allow transport of the amplified material to the nanopore chip. We aim to minimize the distance between the amplification chamber and the nanopore so that the amplified material is diluted less before it arrives at the nanopore. In an ideal case, we would superimpose the nanopore with the reaction chamber, but this presents microfluidic design challenges. In the case where the nanopore is situated below the reaction chamber and the valve layer, we would need to replace at least that area of glass slide (5 mm x 5 mm) with the nanopore chip. This would present potential for leakage at the perimeter of the nanopore chip but may be solved by including an additional PDMS layer between the glass-silicon layer and the valve layer. Another way to circumvent the leakage problem is to replace the entire glass slide with the silicon frame, but this would present additional costs in terms of materials and fabrication time. We do not envision that the nanopore can be superimposed above the reaction chamber because the chip itself would exclude the volume of the fluidic channel layer. When we migrate the chip away from the reaction chamber, we eventually come back to the two-block design described above.

### 4.2.1 Projected performance of telomere length measurement device

The initial quantity of telomeres for a single diploid cell is 96 molecules. Based on the T2T-CHM13 sequence, we calculated the restriction site frequency of BspQI to be 0.18 pmol/µg using a custom Matlab script. In the case of a single diploid cell, we expect roughly 6 pg of DNA, which would then turn into 650,000 fragments after BspQI digest. Therefore, our relative abundance of telomeres is 1 per 6,800 fragments.

We imagine two scenarios for dilution of the amplified DNA since the microfluidic design challenges required for superimposing the nanopore with the reaction chamber are yet to be solved. In the two-block design described previously, the amplification product is diluted by the dead volume between the amplification chamber and the nanopore. Based on the cross-sectional geometry and the length of the channels, we calculate this to be 40 nL. The amplification chamber itself is 1.3 nL so the product is diluted by a factor of 31 when it arrives at the nanopore, assuming uniform dilution. In the super-imposed design, no dilution takes place.

We can estimate the time required for sufficient amplification as follows. The nanopore capture rate is a function of pore size, analyte concentration, analyte size, and applied voltage. For MspA, others have determined that the capture rate for 500 nM 120 nt ssDNA to be 20 Hz (108). The analyte size dependence has not been investigated for MspA, but others report a 3-fold increase in capture rate for dsDNA in a solid state pore experiment (68). Since ssDNA has a low persistence length compared to dsDNA, we assume the length-based enhancement of capture rate for ssDNA to be negligible. We also assume that translocation takes place under denaturing conditions, meaning all dsDNA fragments are denatured to ssDNA and effectively doubles the concentration of all fragments. With the starting quantities and dilution volumes mentioned previously, we can calculate the initial concentration of total (background and target) fragments and the gain required to reach 500 nM. To simplify the analysis, we assume non-ideal amplification specificity and, conservatively, set it to be 10:1 per cycle. This means that for every 10 telomeres amplified, 1 background fragment is amplified. We also assume that amplification proceeds using NEM-SDA$^2$ described in Section 4.4, meaning quadratic amplification with respect to time and where the nicking cycle occurs every minute (59). Under these assumptions, it takes 1 hour of amplification to reach 500 nM in the super-imposed design and 5 hours of amplification to reach 500 nM after dilution in the two-block design. Given the conservative estimate for specificity, the relative abundance of telomeres reaches a plateau on the order of tens of minutes. In either case, the relative abundance of telomeres was estimated to be 1 in 340 fragments.

The extent of amplification determines the required reagent capacity. In the super-imposed design, we need to generate 700 amol of product, whereas we need 20 fmol for the two-block design. Because primers are not consumed, but rather reused in the NEM-SDA$^2$ scheme, the limiting reagent is dNTPs. If we assume an average fragment length of 5 kbp, then we require a minimum of 7 pmol and 200 pmol of dNTPs for the super-imposed and two-block designs, respectively. However, we generally require a molar excess to drive the polymerization forward, so we assume a practical minimum to be doubled to 14 pmol and 400 pmol, respectively. With these assumptions, the minimum dNTP concentration in a 1.3 nL reaction chamber would be 10 mM for the super-imposed design, which is easily achievable in a mastermix formulation. However, the required dNTP concentration for the two-block design is higher than what is commercially available, at 310 mM. Therefore, moving forward with a two-block design would first require enlargement of the reaction chamber.

For practicality, we constrain ourselves to a 1-hour measurement time after amplification, meaning that we can obtain approximately 72,000 translocations at 500 nM. However, because the average fragment length is relatively long, the pore may be in an occupied state for the majority of the time and bottleneck the throughput. Based on previous experiments, translocation of 5 knt ssDNA takes approximately 80 ms. Therefore, the effective translocation throughput is 13 Hz, and we obtain 47,000 translocations in 1 hour. Based on the relative abundance estimated earlier, 210 of those translocations would pertain to telomeres. Of 46 possible subtelomeres to map to, we would obtain a depth of approximately 4.6X. Based on the subtelomere dissimilarity described in Section 4.4.2, one read is sufficient in most cases for identification of the non-homologous chromosome. For the homologous chromosomes, we can assume random sampling and estimate a 96% probability of sampling both chromosomes with the given depth. Certainly, if this depth is not adequate, we can double or quadruple the measurement duration as needed.

## 4.3 Isothermal amplification of telomeres by selective whole genome amplification (SWGA)

### 4.3.1 Rationale

Selective whole genome amplification (SWGA) is a branch of the multiple displacement amplification (MDA) method. MDA makes use of a strand displacing polymerase (usually Phi29) in combination with random hexamer oligonucleotides. In principle, the hexamers prime throughout the genome and become extended. As extension proceeds, 5’ ends of the original hexamers are displaced by the growing 3’ end. The displaced ssDNA product is then also randomly primed, leading to an amplification cascade that can generate micrograms of product from a single cell. In SWGA, selective primers are used instead of hexamers. These primers are chosen to have higher incidence in a target genome compared to a background genome. In doing so, SWGA can enrich a target genome beyond its background, and has demonstrated success in applications involving the detection of pathogens. MDA has an average product length of >10 kb (*100*). It follows that SWGA should generate products with lengths similar to or greater than MDA, owing to the reduced binding frequency of specific primers in comparison to random hexamers.

To examine how effective this method is for subtelomere amplification, we started by using an existing *in silico* toolkit, called swga, for SWGA primer design (*101*). We then used those primer sets to amplify and sequence DNA extracted from GM12878. Reads were then mapped to T2T CHM13 and coverage statistics were used to calculate subtelomere and telomere enrichment.

### 4.3.2 Primer design

The toolkit swga finds primer sets by counting *k*-mers in a target and background sequence set. *K*-mers are substrings of length *k* contained within a sequence. In the context of amplification, a *k*-mer represents a potential priming site, where *k* is the primer length and the *k*-mer frequency is the number of priming sites on a sequence. Larger *k* would be more specific to targets, whereas lower *k* would give more even target coverage. For the purpose of SWGA, *k* is constrained by the reaction temperature of Phi29, which is optimal at 30°C (*102*). In our use of swga, we searched for *k*-mers with lengths from 5-12 bases. We defined our target as the distal 200 kb of each chromosome from GRCh38, and the background as the remaining interior sequence. The binding frequency on the target and background is then used to score and filter potential primers.

A limitation to this search approach is the fact that SWGA was developed to target microbial species. While bacteria generally contain a single circular genome, human subtelomeres are linear and segregated. In other words, the approach of maximizing binding frequency on a bacterial genome is likely to succeed, but we risk missing several subtelomere arms if we rely on binding frequency alone. In an extreme case, a k-mer with high target binding may only be attributed to one subtelomere out of the 46 available. For this reason, we needed to introduce our own metrics to incorporate into the selection processes for primers and primer sets. In order to calculate the gini coefficient of a primer set, swga records the binding locations of each *k*-mer. For each primer set generated, swga can then output a BED file containing this information. We generated a script in MATLAB to parse this data at scale and generate the following metrics: number of binding sites or hits for each telomere arm, number of arms not hit, and the standard deviation of number of hits. Similar to how the swga uses the gini coefficient to measure uniformity of binding site distribution, we use the standard deviation of hits sampled from each telomere arm to measure uniformity.

We then proceed with the swga pipeline, but we modify primer set filtering parameters to output thousands of top-scoring primer sets rather than a handful. We pass the corresponding BED files into our MATLAB script, and use the resulting metrics to rank the sets firstly by number of subtelomeres covered and secondly by swga score.

Table 4.1: Primer sets used for SWGA.

| Set # | 1 | 2 | 3 |
|---|---|---|---|
| Number of sets evaluated | 3,000,000 | 6,000,000 | 6,000,000 |
| Mean background distance (bp) | 241142 | 845826 | 956254 |
| Mean foreground distance (bp) | 56470 | 73619 | 62827 |
| Gini coefficient | 0.620 | 0.616 | 0.652 |
| Score (gini×fg/bg) | 0.145 | 0.054 | 0.043 |
| Primers | AGTCTGCATT<br>CAACCTTTAGA<br>CTTTAGAGTCTG<br>GTTAGGGTTAG | AGAGCATACTAT<br>CGGACTCTAA<br>CTCTCTATCTGA<br>GAAATCGTGTT<br>GACTCTAAACG<br>TAAACCCTAAC<br>TAGATGTCTAAA<br>TCCAATACTAAT<br>TCGCTGTAATA | ATCCTAACCCTAA<br>CAAACACGATTTC<br>CAGACTCTAAGG<br>CTTAACCCTAAC<br>GAGTCTCTATTG<br>GTGAGTTTATAC<br>TCGCTTCCAA<br>TGACCCTAACC |
| Telomere arms hit | 33 out of 46 | 31 out of 46 | 39 out of 46 |

### 4.3.3 Methods

Each 25 µL reaction contained 1X Phi29 Buffer, BSA, dNTPs, 0.5-2.5 µM each primer, 15 U Phi29, and between 5 pg to 50 ng genomic DNA. Reactions were incubated with the following temperature program: Ramp from 35°C to 30°C over 1 hour, 30°C for 16 hours, 65C for 15 min. Sequencing was done using the Ligation Kit (SQK-LSK109), and in some cases we preceded ligation with debranching by T7 endonuclease I as recommended by Oxford Nanopore when working with MDA products.

Basecalling was initially done using guppy in high accuracy (hac) mode. Later, guppy was released with the addition of a super accurate (sup) mode, and we switched to this method for all subsequent basecalling. Reads were mapped to the T2T-CHM13 reference. We did not see a large change (<1% difference) in coverage when switching between the T2T-CHM13 and the GM12878 references.

### 4.3.4 Results

Using a higher accuracy basecaller gave improved mapping to GM12878. For an SWGA reaction with Set #1, high accuracy (hac) basecalling with guppy gave 80.95% mapped reads, whereas super accurate (sup) basecalling gave 93.2% mapped reads.

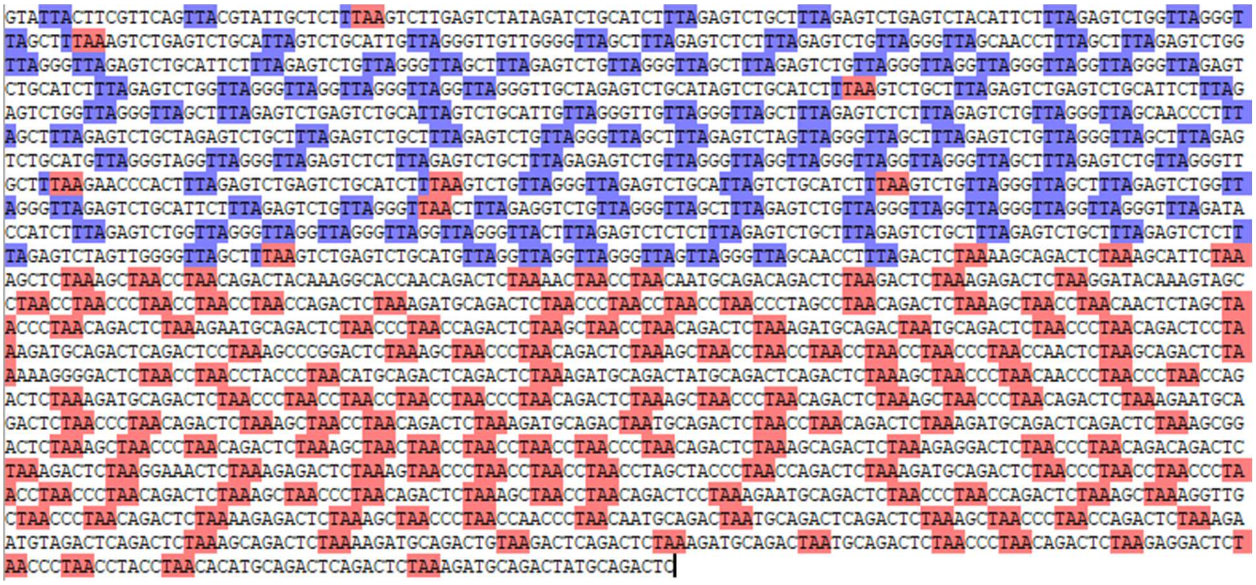
Of the three primer sets generated, only Set #1 gave increased coverage of subtelomeric regions (Table 4.2). The other sets had lower than WGS levels of subtelomere and telomeric reads. Inspection of chromosome coverage confirmed that specific off-target regions had been amplified.

Table 4.2: Sequencing and enrichment statistics from SWGA using generated primer sets.

| Set # | 1 | 2 | 3 |
|---|---|---|---|
| %Bases mapped | 30% | 82% | 77% |
| Bases on distal 20k | 8,578,156 | 12,712 | 23,721 |
| Bases on distal 20k - 200k | 1,421,216 | 38,016 | 17,893 |
| Bases on remainder | 38,118,197 | 68,155,280 | 23,323,297 |
| Subtelomere gain | 12.6 | 0.19 | 0.26 |
| Telomere gain | 602 | 0.63 | 3.4 |
| Outcomes | Increased ST gain<br>Chimeric telomere reads | Amplifies non-telomeric regions | Amplifies non-telomeric regions |

In addition to subtelomeres, SWGA with primer set 1 gave enhanced coverage in telomeres. However, inspection of these reads revealed that they did not have the typical structure of a telomere-containing read. Instead, many contained TTAGGG-like (G-strand) repeats near the beginning of the read, followed by either non-repetitive sequence or C-strand-like telomeric sequence. The telomeric repeats found were distinct from variant repeats typically found in telomeres in that they had far less similarity to the standard telomere repeats as well as highly variable repeat lengths. Virtually no reads contained true telomeric sequence, but many contained this G-strand to C-strand pattern that is indicative of chimeras formed by inverted sequences *(103)*. This suggests that the SWGA process is prone to chimera formation, similar to MDA

(103), specifically chimeras with inverted sequences. Of the read data generated by primer set 1, only 30% of bases could be mapped to GM12878.



```
GTATTACTTCGTTCAGTTACGTATTGCTCTTTAAGTCTTGAGTCTATAGATCTGCATCTTTAGAGTCTGCTTTAGAGTCTGAGTCTACATTCTTTAGAGTCTGGTTAGGGT
TAGCTTTAAAGTCTGAGTCTGCATTAGTCTGCATTGTTAGGGTTGTTGGGGTTAGCTTTAGAGTCTCTTTAGAGTCTGTTAGGGTTAGCAACCTTTAGCTTTAGAGTCTGG
TTAGGGTTAGAGTCTGCATTCTTTAGAGTCTGTTAGGGTTAGCTTTAGAGTCTGTTAGGGTTAGCTTTAGAGTCTGTTAGGGTTAGGTTAGGGTTAGGTTAGGGTTAGAGT
CTGCATCTTTAGAGTCTGGTTAGGGTTAGGTTAGGGTTAGGTTAGGGTTGCTAGAGTCTGCATAGTCTGCATCTTTAAGTCTGCTTTAGAGTCTGAGTCTGCATTCTTTAG
AGTCTGGTTAGGGTTAGCTTTAGAGTCTGAGTCTGCATTAGTCTGCATTGTTAGGGTTGTTAGGGTTAGCTTTAGAGTCTCTTTAGAGTCTGTTAGGGTTAGCAACCCTTT
AGCTTTAGAGTCTGCTAGAGTCTGCTTTAGAGTCTGCTTTAGAGTCTGTTAGGGTTAGCTTTAGAGTCTAGTTAGGGTTAGCTTTAGAGTCTGTTAGGGTTAGCTTTAGAG
TCTGCATGTTAGGGTAGGTTAGGGTTAGAGTCTCTTTAGAGTCTGCTTTAGAGAGTCTGTTAGGGTTAGGTTAGGGTTAGGTTAGGGTTAGCTTTAGAGTCTGTTAGGGTT
GCTTTAAGAACCCACTTTAGAGTCTGAGTCTGCATCTTTAAGTCTGTTAGGGTTAGAGTCTGCATTAGTCTGCATCTTTAAGTCTGTTAGGGTTAGCTTTAGAGTCTGGTT
AGGGTTAGAGTCTGCATTCTTTAGAGTCTGTTAGGGTTAACTTTAGAGGTCTGTTAGGGTTAGCTTTAGAGTCTGTTAGGGTTAGGTTAGGGTTAGGTTAGGGTTTAGATA
CCATCTTTAGAGTCTGGTTAGGGTTAGGTTAGGGTTAGGTTAGGGTTACTTTAGAGTCTCTCTTTAGAGTCTGCTTTAGAGTCTGCTTTAGAGTCTGCTTTAGAGTCTCTT
TAGAGTCTAGTTGGGGTTAGCTTTAAGTCTGAGTCTGCATGTTAGGTTAGGTTAGGGTTAGTTAGGGTTAGCAACCTTTAGACTCTAAAAGCAGACTCTAAAGCATTCTAA
AGCTCTAAAGCTAACCTAACAGACTACAAAGGCACCAACAGACTCTAAAACTAACCTAACAATGCAGACAGACTCTAAGACTCTAAAGAGACTCTAAGGATACAAAGTAGC
CTAACCTAACCCTAACCTAACCTAACCAGACTCTAAAGATGCAGACTCTAACCCTAACCTAACCTAACCCTAGCCTAACAGACTCTAAAGCTAACCTAACAACTCTAGCTA
ACCCTAACAGACTCTAAAGAATGCAGACTCTAACCCTAACCAGACTCTAAGCTAACCTAACAGACTCTAAAGATGCAGACTAATGCAGACTCTAACCCTAACAGACTCCTA
AAGATGCAGACTCAGACTCCTAAAGCCCGGACTCTAAAGCTAACCCTAACAGACTCTAAAGCTAACCTAACCTAACCTAACCTAACCCTAACCAACTCTAAGCAGACTCTA
AAAAGGGGACTCTAACCTAACCTACCCTAACATGCAGACTCAGACTCTAAAGATGCAGACTATGCAGACTCAGACTCTAAAGCTAACCCTAACAACCCTAACCCTAACCAG
ACTCTAAAGATGCAGACTCTAACCCTAACCTAACCTAACCTAACCCTAACAGACTCTAAAGCTAACCCTAACAGACTCTAAAGCTAACCCTAACAGACTCTAAAGAATGCA
GACTCTAACCCTAACAGACTCTAAAGCTAACCTAACAGACTCTAAAGATGCAGACTAATGCAGACTCTAACCTAACAGACTCTAAAGATGCAGACTCAGACTCTAAAGCGG
ACTCTAAAGCTAACCCTAACAGACTCTAAAGCTAACTAACCTAACCTAACCTAACCCTAACAGACTCTAAAGCAGACTCTAAAGAGGACTCTAACCCTAACAGACAGACTC
TAAAGACTCTAAGGAAACTCTAAAGAGACTCTAAAGTAACCCTAACCTAACCTAACCTAGCTACCCTAACCAGACTCTAAAGATGCAGACTCTAACCCTAACCTAACCCTA
ACCTAACCCTAACAGACTCTAAAGCTAACCCTAACAGACTCTAAAGCTAACCTAACAGACTCCTAAAGAATGCAGACTCTAACCCTAACCAGACTCTAAAGCTAAAGGTTG
CTAACCCTAACAGACTCTAAAAGAGACTCTAAAGCTAACCCTAACCAACCCTAACAATGCAGACTAATGCAGACTCAGACTCTAAAGCTAACCCTAACCAGACTCTAAAGA
ATGTAGACTCAGACTCTAAAGCAGACTCTAAAAGATGCAGACTGTAAGACTCAGACTCTAAAGATGCAGACTAATGCAGACTCTAACCCTAACAGACTCTAAGAGGACTCT
AACCCTAACCTACCTAACACATGCAGACTCAGACTCTAAAGATGCAGACTATGCAGACTC
```

Figure 4.2: Representative chimeric read generated by SWGA using primer set 1. (blue) TTA motif associated with G strand telomere repeats, (red) TAA motif associated with C strand telomere repeats.

SWGA with primer sets 2 and 3 gave fewer reads that mapped to telomeres or subtelomeres, compared to unamplified material. Close inspection of mapping results showed that specific regions or loci within the chromatin regions were preferentially amplified, leading to depletion of the telomere and subtelomeric content.

### 4.3.5 Discussion

As it stands, SWGA performed with set #1 amplifies gDNA from as low as 5 pg to 1000s of ng, an amount which is amenable to nanopore sequencing by ligation. Subtelomere sequencing depth was enriched by 33X, but no true telomere-containing reads could be obtained due to high presence of chimeric reads.

While SWGA has the capability of enriching subtelomeres, acquisition of whole-telomere amplicons by SWGA would require additional modifications to the scheme.

This is due to the open telomere problem, where complementary primers can hybridize anywhere along the G strand and form truncated telomere products. Workarounds may require a method of anchoring primers to the telomere end, whether by terminal transferase, ligase, or other means. This is further complicated by the hyperbranched nature of WGA amplicons, which may not be amenable to T/A ligation for sequencing.

Additionally, SWGA for telomeres and subtelomeres may benefit from a more tailored approach to primer design with a key requirement being awareness of chromosome structure. Investigation into computational methods or mathematical models that can account for this while still maintaining computational efficiency are beyond the scope of this work. Due to the possible presence of chimeric reads, data processing with SWGA reads may also benefit from methods for chimera correction prior to mapping or chimera-aware mapping. An improved and perhaps more stringent workflow might include verification of binding locations by blast alignment, as well as consideration of binding locations with partial mismatches.

Nonetheless, SWGA has been demonstrated to amplify the region that is proximal to telomeres, representing a preliminary step towards the goal of isothermal telomere amplification from single cells.

## 4.4 Isothermal amplification of telomeres by dual nicking endonuclease mediated strand displacement amplification (NEM-SDA$^2$)

### 4.4.1 Rationale

The long read sequencing effort on GM12878 produced a set of 1172 contigs representing 35X coverage of the genome (*86*), but attempting to map this to the latest human genome reference (hg38) leaves us with a gapped result. The reference hg38 has gaps not only in the acrocentric arms (13p, 14p, 15p, 21p, 22p), but also on the distal (farthest from the centromere) portions of several subtelomeres. With the advent of a complete telomere to telomere (T2T-CHM13) reference (*77*), we are now able to map those contigs and form a gapless consensus as a reference.

With a complete reference available, we are now able to design primers specific to our cell line. Using the G strand as our frame of reference, the forward primer is to be located within the subtelomere and the reverse primer on the telomere overhang. The actual location of the forward primer on the subtelomere with respect to the telomere boundary is critical. Being too far from the telomere may lead to problems with amplification due to product length. If it is too close then there is sequence data available for accurate mapping. As for the reverse primer, initiation of amplification must be anchored to the overhang to avoid forming truncated G strands. While others have demonstrated anchoring by use of ligase (*97*) or terminal transferase with ligase (*104*), we wanted to avoid the use of ligase due to its long processing time. The telomere overhang presents itself as a small (100-300 bp) window on which polymerization can initiate. We designed our scheme in a way that primes this window in an early phase only, from which the resulting amplicon serves as a seed for subsequent amplicons (Figure4.3). There is some loss of precision due to the size of the overhang, but this is small relative to the length of the telomere (1 kbp - 20 kbp). We may also be able to compensate by selecting the reads with the longest telomeres from if they form a distribution that is the width of the overhang length.

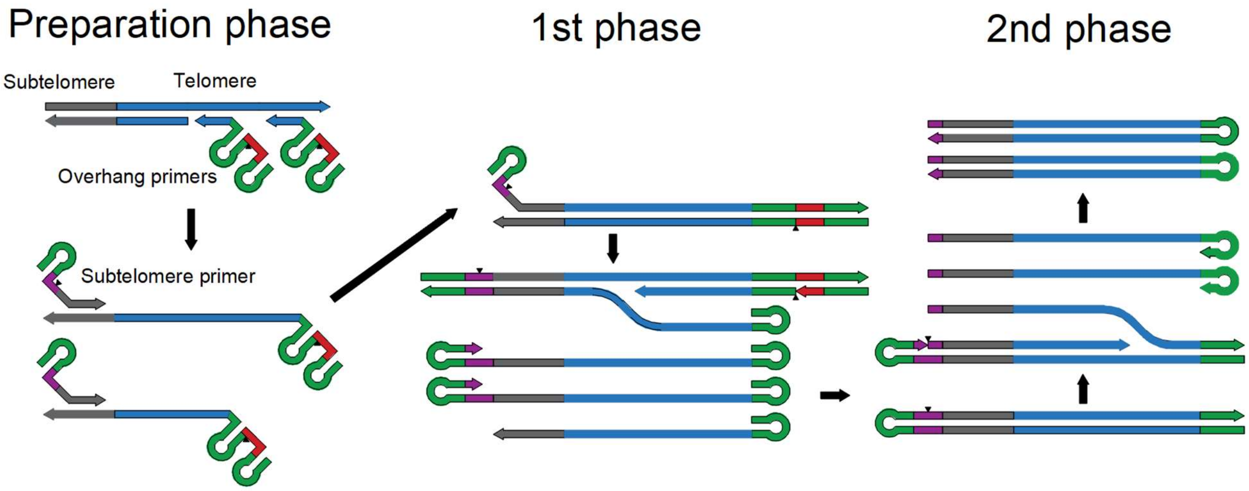


Figure 4.3: Amplification of telomeres with NEM-SDA[2].

First, the genomic DNA is cut using a restriction enzyme with the same recognition site as the overhang primer nicking site. This way, all such restriction sites are cut with blunt ends and do not form potential initiation sites for SDA. For reasons that will be discussed later, we chose BspQI as the restriction enzyme.

Second, a primer with a double hairpin and a nicking site proximal to the internal hairpin hybridizes to the overhang. We chose our overhang complement to have 4.5 TR. This way, the hybridization of the overhang to the primer is favorable over folding into a G-quadruplex. Given the variable length of the overhang, we estimate that 3-10 primers will hybridize to each telomere. After addition of a strand displacing polymerase, a corresponding number of C strands are released from the template.

Third, treatment with lambda exonuclease leads to degradation of dsDNA to ssDNA due to its 5’-3’ exonuclease activity. Conversion of background (non-telomeric) genomic fragments to ssDNA makes them incompatible with ligation to sequencing adapters. Incorporation of 5’ phosphorothioate linkages in the overhang primer prevents degradation by lambda exonuclease (*105*)). In an ideal digestion by lambda exonuclease, only the C strands initiated from the overhang primer remain, along with ssDNA of BspQI-digested genomic fragments.

Fourth, we initiate double NEM-SDA by addition of the subtelomere primer, a strand displacing polymerase, and two nicking enzymes. Two processes occur simultaneously, which we will call primary and secondary amplification. At the start of primary amplification, the forward or subtelomere primer is extended. Complete extension unfolds the double hairpin and forms the nicking enzyme recognition site on the overhang side. This allows primary amplification to occur, where NEM-SDA proceeds from the overhang towards the subtelomere. This generates C strands at a linear rate, and displaced C strands form hairpins at the 3' and 5' end since the template has hairpins. The 3' hairpin acts as a forward priming site. A polymerization initiates from this hairpin, forming a whole-telomere dsDNA molecule that is joined at the subtelomere side. Since nicking enzymes leave a 5'-phosphate and Sequenase 2.0 leaves a 3'-A overhang, the resulting amplicon is T/A ligatable on the overhang side. These molecules also form the template for secondary amplification.

In secondary amplification, NEM-SDA proceeds from the subtelomere side using the subtelomere primer nicking site. This generates G strands at a linear rate for each primary amplicon. Since primary amplicons are also generated at a linear rate, this leads to quadratic amplification. Free G strands form a 3' hairpin since the template includes the internal hairpin from the overhang primer. These self-priming molecules then form whole-telomere dsDNA that is joined at the overhang side and T/A ligatable on the subtelomere side.

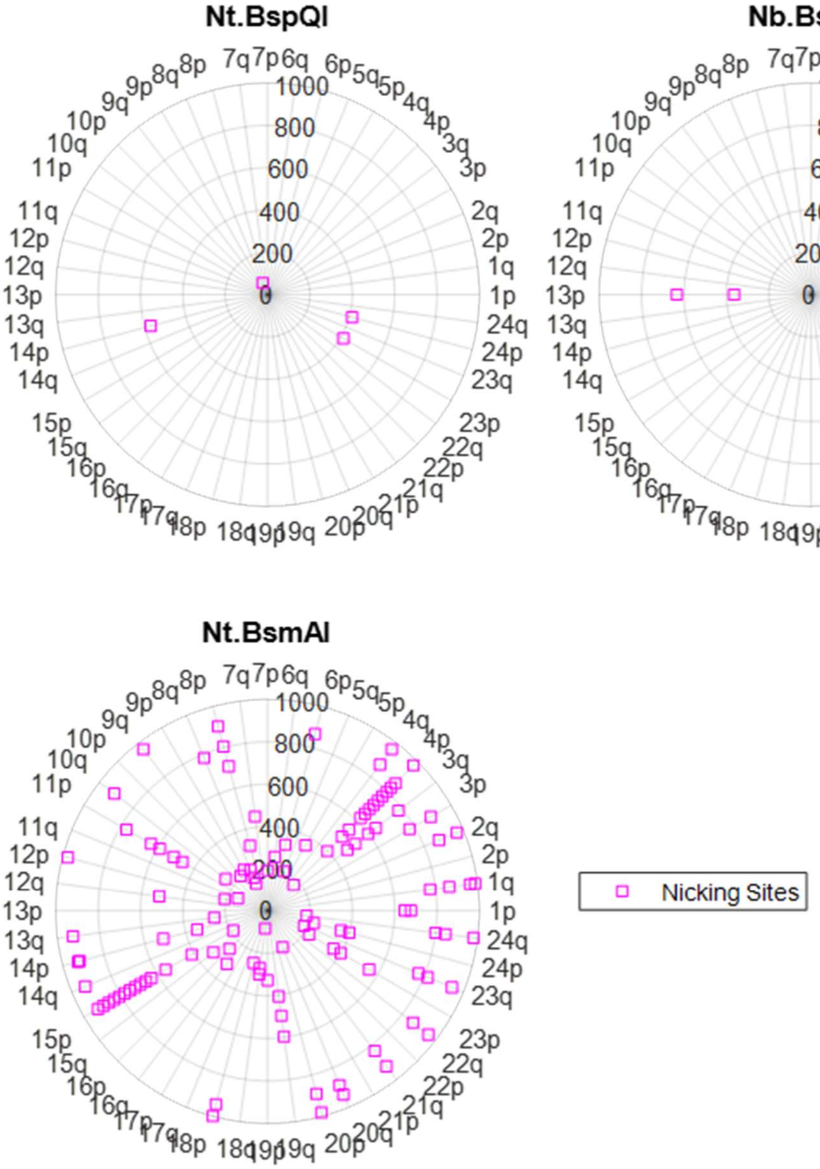
Since T/A ligation is involved, we preferred that our final product have 3’-dA overhangs. That way we can skip the end preparation step that is typically done prior to ligation. This means we could not use a polymerase with 3’-5’ exonuclease activity since these do not leave a 3’-dA. In addition, the polymerase must be compatible with nicking enzyme mediated strand displacement amplification. Such polymerases include Sequenase 2.0 and *Bst* 3.0 polymerase.

Sequencing by ligation of amplicons with terminal hairpins has the advantage of obtaining reads with higher accuracy similar to the 2D, $1D^2$, and Duplex read technologies developed by Oxford Nanopore. Both the G and C strands are sequenced together, allowing sequencing errors from one strand to be corrected by the other and vice versa. Since the majority of product are secondary amplicons which are ligated on the subtelomere side, we can expect reads to contain the following order of sequence from 5’ to 3’: partial subtelomere primer nicking site, G strand subtelomere sequence, G strand telomere repeats, telomere overhang primer internal hairpin, C strand telomere repeats, C strand subtelomere sequence, and partial subtelomere primer nicking site.

### 4.4.2 Primer design

We mapped the GM12878 contigs to the T2T-CHM13 reference using minimap2 (2.24-r1122) in assembly to reference mode with ~0.1% sequence divergence. A consensus, hereby referred to as our reference, is formed using bcftools. We characterized this reference by counting SNPs with respect to the T2T reference and find that it has a SNP rate of 1 per 1.2 kb through the whole genome, and 1 per 0.5 kb in the distal 20 kb. The overall SNP rate is relatively low, but this is likely due to the similar ancestry of the two sources, both being described as originating from donors with European ancestry. The increased SNP rate in the subtelomeres is expected, due to subtelomeres being hotspots of shuffling in the context of an evolutionary timescale (*94*).

Custom scripts were written in Matlab to handle restriction site identification and primer design. We identified telomere boundaries by applying the telomere signal transform on the distal 20 kb of each chromosome of the reference. We then extracted the distal 1 kb of each subtelomere, counting from the telomere boundary towards the centromere. On each 1 kbp of subtelomere, we searched for nicking sites of 10 out of the 11 available nicking endonucleases from New England Biolabs (NEB) in the forward and reverse directions: Nt.BspQI, Nt.BstNBI, Nt.BsrDI, Nb.BtsI, Nt.AlwI, Nb.BbvCI, Nt.BbvCI, Nb.BsmI, Nb.BssSI, and Nt.BsmAI. We did not search for nicking sites for Nt.CviPII because of its short recognition sequence (CCD, where D is A, G, or T). Of the 10 that we examined, Nt.BspQI and Nb.BssSI had the least number of nicking sites in the distal subtelomere and Nt.BsmAI had the most nicking sites (Figure 4.4).


Nt.BspQI

Nb.BssSI

Nt.BsmAI

Figure 4.4: Nicking site locations of selected nicking endonucleases on the distal 1000 kb of each subtelomere. The radius represents the distance from the telomere boundary.

To identify primer candidates, we counted *k*-mers from the distal 1 kb of subtelomere. If a subtelomeric sequence corresponded to a *p* arm, we used the reverse complement so that we only searched along the G strand. We counted *k*-mers for *k* from 8 through 25, and *k*-mers were filtered out if they did not meet requirements for melting temperature ($T_m$), GC content, or lack of hairpin formation. We also removed *k*-

mers that contained “GGG” or “CCC”, as these resembled telomere repeats, which can sometimes be found in the subtelomere. Using a $T_m$ range of 37°C to 70°C and a %GC between 40% and 60%, we identified 16 *k*-mers that, individually, can cover at least 23 subtelomeres and 17,789 *k*-mers that were singletons or unique to a subtelomere. We then selected a *k*-mer that covers 31 subtelomeres (“multiST”: TCAGCACAGA) and verified our algorithm by using ncbi-blast (2.13.0+) to align multiST against the distal 1 kb of subtelomeres (Figure 4.5). From the alignment, all exact hits corresponded to subtelomeres from which we expected multiST to originate from. No partial hits (single mismatch) were found or reverse hits (C strand) were found. Comparison to the nicking site locations also showed no interference with multiST.

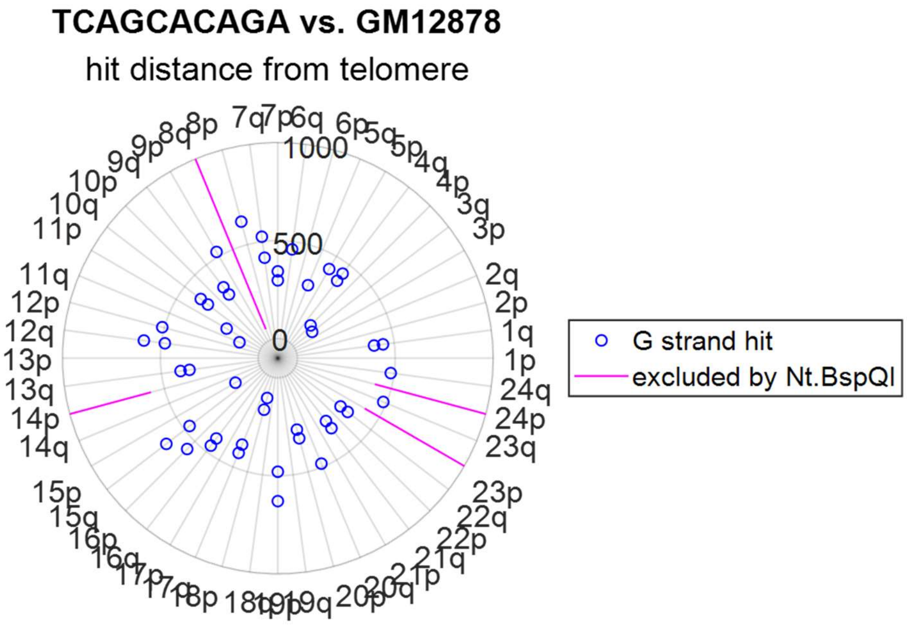


Figure 4.5: Alignment locations from multiST against subtelomeres generated by ncbi-blast. No hits were found in the region excluded by BspQI/Nt.BspQI.

From alignment results, we were also able to predict amplicon sequences by subsampling the distal 1 kbp from the hit location to the telomere boundary. This predicts amplicons with subtelomere lengths ranging from 100 to 700 bases, and we compared predicted amplicons by generating a distance tree (Figure 4.6). Based on the distance tree, most predicted amplicons were sufficiently dissimilar for identification by nanopore sequencing. Nanopore sequencing has a purported single read accuracy of at least 98%, meaning that there should be a high confidence of mapping if the percent identity between amplicons is less than that. Of all amplicons, a handful of pairs or groups had higher than 95% identity (9p-12p, Xq-Yq-20q, 1q-4q, 10q-13q, 5q-9q, and 21q-22q). For these groups, we imagine that we can resolve mapping with additional sequencing depth.

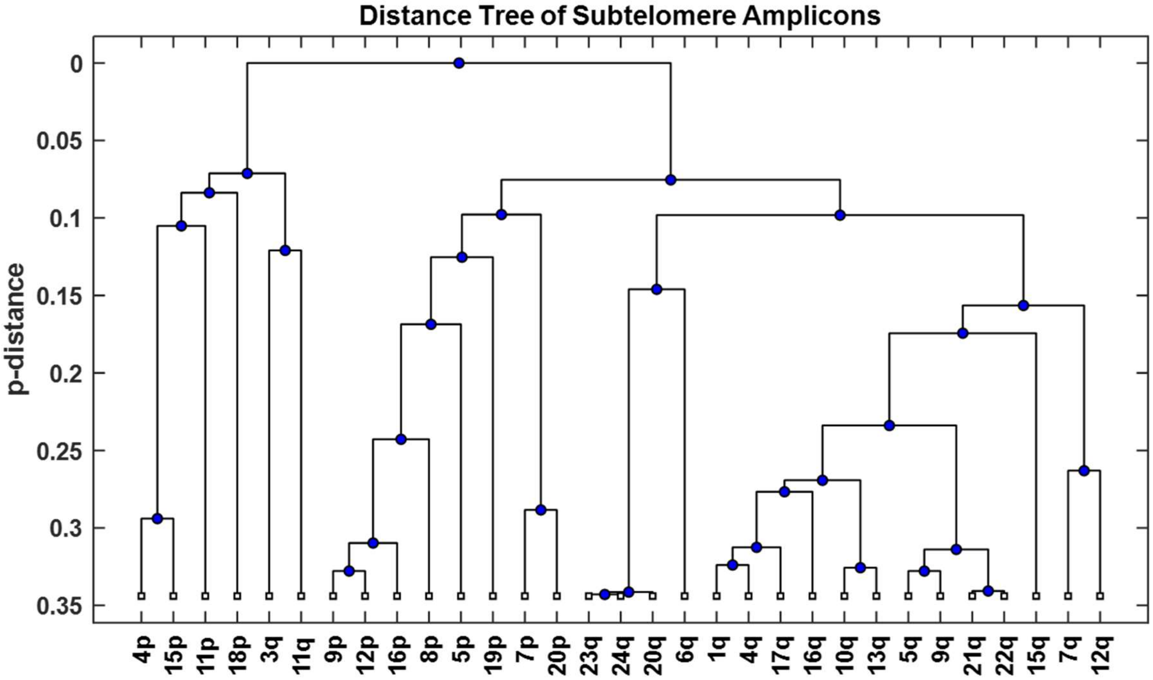


Figure 4.6: Subtelomere amplicon similarity represented by a distance tree. The p-distance is the sequence difference per length.

### 4.4.3 Methods

Genomic DNA was purified from GM12878 cells as described in section 3.3.1. In addition to phenol:chloroform extraction, DNA was also purified by ethanol precipitation. To digest genomic DNA, a 50 μL reaction contained 1X NEBuffer r3.1, 1500 ng DNA, and 10 U BspQI. The reaction was incubated at 37°C for 2 hours, followed by heat deactivation at 80°C for 20 minutes.

To gauge the effectiveness of amplification, we omitted the lambda digest and performed all phases of amplification in parallel. We also used a forward primer specific to 17p (“17p-BssSI”) instead of multiST to gauge specificity. A 100 µL reaction contained 1X NEBuffer r3.1, 100 ng of BspQI-digested DNA, 0.1 µM 17p-BssSI, 0.1 µM 4.5TR-BspQI-2HP, 500 µM dNTPs, 10 U of *Bst* 3.0 polymerase, 20 U of Nt.BspQI, and 20 U of Nb.BssSI. The reaction was incubated at 65°C for 2 minutes, followed by 55°C for 30 minutes and heat inactivation at 85°C for 15 minutes. 20 µL of reaction was analyzed by agarose gel electrophoresis along with 100 ng of BspQI-digested DNA and 100 ng of MassRuler DNA Ladder Mix, ready-to-use (Thermo Scientific).

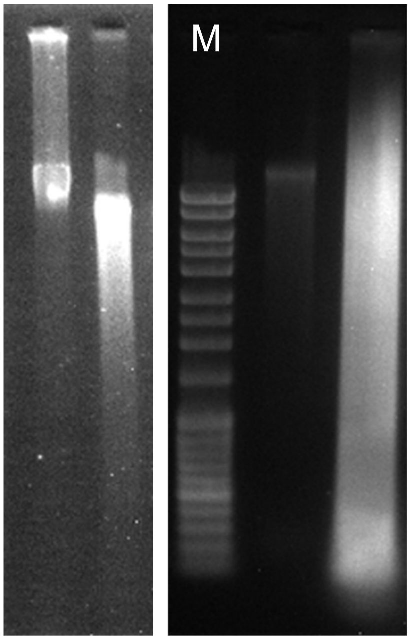


Figure 4.7: (left) Before and after BspQI digest of DNA extract. (right) Before and after amplification of digested DNA by NEM-SDA[2]. M: 100-10000 bp ladder.

Sequencing was done using the ligation kit (SQK-LSK109), following the Amplicons by Ligation protocol. By Qubit measurement with 1X HS dsDNA buffer, the reaction had a dsDNA concentration of 236 ng/µL. We omitted the FFPE repair / end prep step and started by purifying 20 µL of reaction with 40 µL of Ampure XP bead solution. Sample was incubated with beads for 5 minutes at room temperature, followed by two rinses with 70% ethanol without disturbing the beads. DNA was eluted by adding 61 µL of nuclease free $H_2O$ and incubating for 2 minutes at room temperature. The eluate was quantified by Qubit and we proceeded with ligation and cleanup using the entirety of the eluate according to manufacturer instructions. In the post-ligation cleanup, beads were rinsed with Long Fragment Buffer to deplete shorter fragments. The eluate was quantified by Qubit, and 178 ng of prepared library was loaded onto a MinION (r9.4.1) flow cell. The MinION was allowed to run for 16 hours before stopping the run. Reads were basecalled using guppy_basecaller (6.1.3) in super-accurate mode

(-c dna_r9.4.1_450bps_sup.cfg) and mapped to the GM12878 reference using minimap2 (2.24-r1122).

### 4.4.4 Results

Sequencing produced 19,341 reads with an average quality score of 12.8 and an average read length of 3243 bases. Initial attempts to identify telomere-containing reads by telomere signal transform proved difficult, as many reads had noise issues or long repetitive regions with poor quality score.

Instead, we searched for reads pertaining to amplicons generated by the reaction. To do this, we first performed Smith-Waterman alignment of primer-specific sequences to the reads. The subset of sequence used for Smith-Waterman alignment is underlined (17p-BssSI: AGCAATCCTTCGTTTTTCGAAGGATTGCTCGTGTTTTTCCTTTTGTGGTCTGTGCTTTTGGTG, 4.5TR-BspQI-2HP: CGATGCTAGCTCAGGTTTTTCCTGAGCTAGCATCGCTCTTCGCTCGCTGAGTCCTTTTTGGACTCAGCGAGCGATTTTCCCTAACCCTAACCCTAACCCTAACCC). We also performed the same alignment to a control sample of sequencing data generated from unamplified DNA or WGS reads (Figure 4.8). In the sequencing run with amplified material, we observed an additional group of reads that had high alignment score to the overhang primer (4.5TR-BspQI-2HP). This was not present in the WGS reads, confirming that such reads were derived from amplification. An identical match with 17p-BssSI would have given a score of ~60. We did not see a distinct group of reads clustering to this position, indicating that 17p-BssSI was not involved in the reaction. We isolated the subset of reads matching the overhang primer by *k*-means clustering. Amplicon reads had an average read length of 1756 bases, which was lower than the average background read length of 3355 bases.

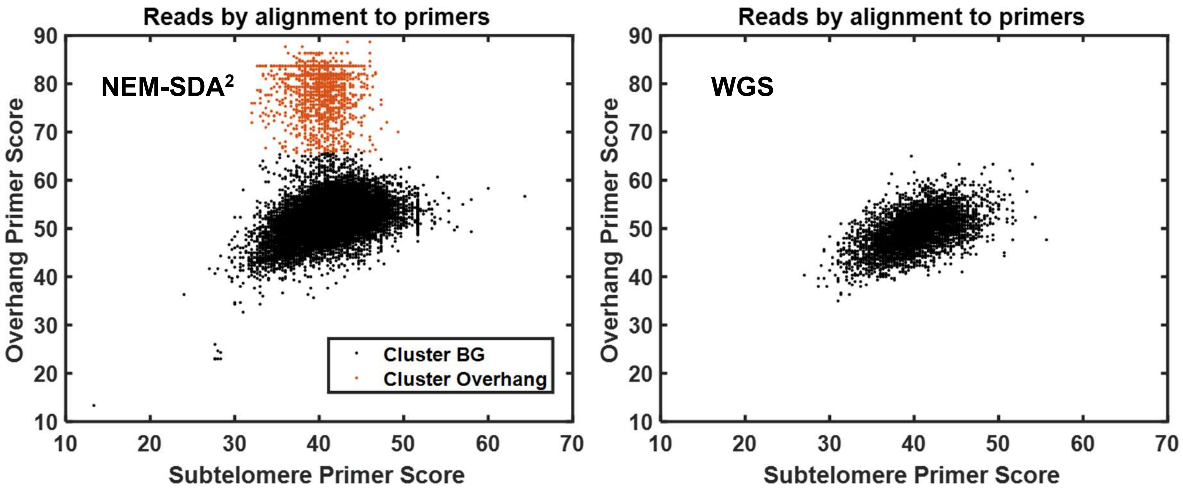


Figure 4.8: Smith-Waterman alignment scores for primer-specific sequences against reads. (left) Sequencing data from NEM-SDA$^2$ reaction, (right) data from unamplified WGS sample.

We then identified the location of the primer-specific sequence on the read and found that most alignments were found towards the beginning or 5' end of the read (Figure 4.9). This was true for both forward and reverse alignments. The localization of forward alignments to the beginning of the read was expected since this the natural result of extension from the overhang primer. Reverse complement alignments to the end of the read were surprisingly low, suggesting that complement strands (or G strand 5' ends) had poor ligation efficiency. Surprisingly, there was a high presence of reverse complement alignments near the beginning of the read. A possible cause is that the reaction is synthesizing reverse complements of the overhang primer, which themselves are able to prime DNA and form initiation sites for extension. Of 1357 reads, 997 alignments (73.5%) were forward and 360 alignments (26.5%) were reverse complemented. Overhang primer alignments to background reads were uniformly distributed, confirming that such reads were not a result of amplification.

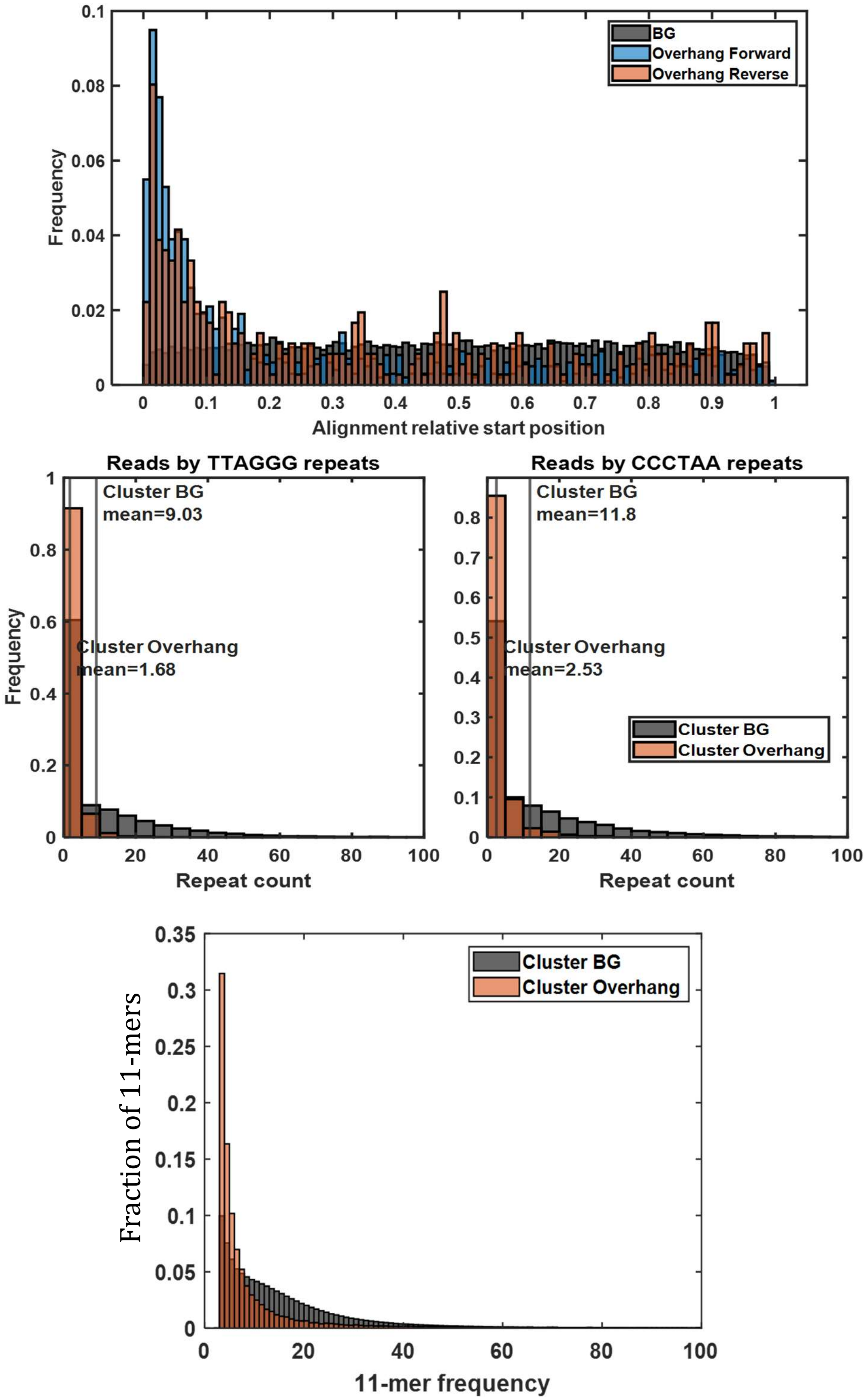


Figure 4.9: (top) Relative position of alignments of primer-specific sequence to reads. (middle) Distribution of telomere repeats within clustered reads. (bottom left) 11-mer spectrum of clusters.

We checked the TR count of amplicon reads and found that on average, they had significantly less TRs than background reads. This confirms nonspecific or off-target amplification. Amplicon reads had a very low mapping rate (7.88%) to the GM12878 reference, compared to background reads (62.36%). Local alignment of 20 randomly selected unmapped reads to the blast Nucleotide collection (nr/nt) showed some similarity (up to 20% of the read) to human references. Aside from other primate references, no other organisms were detected. Remarkably, the majority of hits had identical sequences and were matched with a motif unique to 17q (CATGTGCCGACCCACATGCTCGTGCCAGCGCTGGCA). This motif was found in 74% of amplicons. The amplicons also displayed low sequence variation. Based on an 11-mer spectrum (Figure 4.9), the sequence composition of amplicons was represented by fewer 11-mers compared to background reads. 1005 of amplicons had alignments to both the 17q motif and the overhang primer. All these factors suggest that off-target amplicons may have been derived from a common molecule that underwent runaway exponential amplification.

```
TGCTTCACTTCGTTCAGTTACGTGGCCCCGCGCTTTGCGAGAGCGGAGATGCGTTCTCT
GTAGCACAGACCCTAACAGACCCGGAGAGCATCGCGAGGGCGGAGCTGCGTTCTCCT
CTGCACAGACTTCGGGGGTACTGCGAAGGTGGAGCAGAGTACTCCTCAGCACAGACCA
GGCAGGCAGGCCCAGGGCACCGCGAGGAGCGGAGCTGCGTTCTGCTCAGCAGAGAC
CTAGGGGACTTCTTAAAGCGGACAGCATTCTCTTCACCACAAGTCATTGAAGAGGCAGT
GCCTCGCTGTGGACAACTCAGACGCAACGACAGTGAAGAAAATTTGCAGTTGCACCCT
GAATAATCAAGGTCAGAGAGCAGTTAGAAGGGTTCAGTGTGGAAAACGGAAAAGCAAAA
GCCCCTGTGAATCCTGTACACCGAGATGCTCCCAAGGAAGGCTTGGGGCTGCATTGCA
AGGTCCAACTGCAGGCTCGAATTTTTCAATCCCAGCCTTCTAATGCCTGCATGCTGCAA
AATGTGATATCACATTGCTCATGTAACAAGCACCTGTATGCTAATGCACTCCCTCAATAC
AAAATTGTTAATATAAGATGGGAGGCATAGGAGGGTCAGGGTCGGGGGGTCGGGGTCG
GGGTCGGGGGTCAGGGTCAGGGGGTCAGGGTTCGGGTTCAGGGTTAGGGTTAGGGTT
AGGGTTAGAGGGGTTAGGGGTTAGGGGTTGGGGGTTAGGGTTAGGGGTTGGGGGTTA
GGAGTTGGGGTTGGGGTAGGAGTTGGGGTGGGGGTTGGGGTTGGGGTTGGGGTTGG
GGTTGGGGTTGGGGTTGGGGGTTAGGGTTAAGGGTTGGGGTTGGGGTTGGGGTTGGG
GTTGGGGTTGGGGGATAGGGTTGGGGTTGGGGTTGGGGTTGGGGGTTAGGAGTTGGG
GTTGGGGGTTGGGGTTGGGGTTTAGGGTTTAGGGTTAGGGTTAGGGGTTAGGGGTTGG
GGTTGGGGTTGGGGTTGGGGTTAGGGGTTAGGGTTAGGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAGAGGGGTTGGGAGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTGGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGAGGTTAAGTTGGGG
TTAGGTGGGGTTAGGAGAAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTGGGGTTAGGG
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
GGGTTAGGGTTAGGGTTAGGGAAAATCGCTCGCTGAGTCCAAAAGGACTCAGCGAGCA
GCAGAAAAGTGACTCCTAACTCAGGGAAAAACCTA
```

Figure 4.10: Sole amplicon read that had telomeric repeats. Black: subtelomere, red: telomeric repeats, blue: sequence similar to overhang primer.

Of 1356 amplicons, 1 had definitive telomeric repeats (Figure 4.10). It also had the expected structure of a G strand amplicon: subtelomere, telomere, followed by reverse complement of the overhang primer. The presence of this read implies that the C strand was synthesized as well, since this read contains the reverse complement of the primer. Unlike the majority of amplicons, this was ligatable from the subtelomeric side which left the overhang primer sequence on the 3' end. As expected, the subtelomere-proximal portion of the telomere was marked by a larger proportion of variant repeats. This read mapped unambiguously to 5p with a mapping quality of 58 (scale from 0 to 60). 5p also has a similar pattern of variant repeats. No secondary or supplementary alignments were found with minimap2. However, the telomere length of this read was shorter than expected at 736 bases, whereas M. Jain et al report a telomere length of approximately 6.5 kb for 5p. Without additional telomeric reads, we could not determine whether this was a telomere truncated by the amplification process or a true whole telomere.

### 4.4.5 Conclusion

We developed a *k*-mer based primer design workflow for subtelomeres and identified nicking endonucleases that would be suitable for NEM-SDA$^2$ of telomeres. We can identify primers that can bind to multiple subtelomeres as well as single subtelomeres. Comparison of the subtelomeric region of expected amplicons showed that a majority of amplicons should be easily distinguishable with subtelomere coverages ranging from 100 to 600 bases.

Simultaneous NEM-SDA$^2$ with *Bst* 3.0 Polymerase gave mostly off-target amplification. From the sequencing data, we reason that a majority of off-target amplification was derived from the same or similar molecules based on their low sequence complexity. *Bst* 3.0 has been reported to generate repetitive amplicons from a single primer due a combination of its strand displacement and reverse transcription activities (*106*). In these cases, amplification was exponential and saturated within 2 hours, and a similar process may have occurred here. Amplification of telomeres in this manner would benefit greatly from strategies for mitigation of off-target amplification which may include studies on the effect of temperature, alternative polymerases, and SSB concentration.

Of 1357 reads, a single read had definitive telomeric repeats and was mapped unambiguously to 5p only, as predicted by comparison of subtelomeres from the reference.

### 4.5 Acknowledgements

This work was supported by the NIH (5R21CA223727-02).

Chapter 4 contains unpublished material co-authored with Huang, Xiaohua. The dissertation author was the primary author of this chapter.

# 5 Conclusion

## 5.1 Summary of work

### 5.1.1 Tools for DNA length analysis

We have developed a process for fabricating solid state nanopores on silicon nitride membranes with coplanar metal patterns. The inclusion of metal patterning to the process opens the possibility for device enhancement. We demonstrated precise tuning of nanopore diameters by ALD, which allows us to generate nanopores with diameters below the limit imposed by the EBL-RIE process. These nanopores enable length measurement for dsDNA in the kb range.

We also cloned and expressed MspA in *E. coli* and developed a process for purification of the oligomer. The process was not detrimental to the functionality of MspA, as demonstrated by the protein nanopore's conductance and translocation characteristics. These nanopores enable length measurement of ssDNA with even greater length resolution than solid state nanopores.

### 5.1.1 Telomere enrichment with magnetic beads

We have demonstrated a method for capture of telomere-containing fragments for telomere and subtelomere sequencing. Using streptavidin-coated magnetic beads and biotinylated telomere overhang probes, we found that DNA concentration was the most critical factor when performing enrichment. A sample volume of 10 mL containing DNA from 50 million cells was an ideal starting point. Concentrating the sample by reducing the volume by ethanol precipitation or by scaling up the number of cells led to difficulty in completing the process. In the optimal condition, we enriched telomere-containing reads by 20 times and increased the proportion of bases mapped to subtelomeres by 10 times. This enabled telomere length measurement for the cell line GM12878, which uncovered additional telomere lengths as well as provide telomere lengths in agreement with other sources.

### 5.1.3 Isothermal telomere amplification

We investigated the efficacy of two isothermal methods for amplification: SWGA with Phi29 and NEM-SDA$^2$ with *Bst* 3.0. One out of three SWGA primer sets examined was found to increase subtelomere and telomere coverage from inputs as low as 5 pg. However, telomere reads were chimeric and prevented accurate telomere length measurement. At best, we could generate a telomere length distribution for the sampled DNA.

For NEM-SDA$^2$, a subtelomere primer design workflow was developed. Nicking sites for commonly available nicking endonucleases were identified on the subtelomeres, and we determined that Nt.BspQI and Nb.BssSI were ideal for this mode of amplification. Comparison of the predicted amplicon sequences showed that subtelomeres could be identified with as little as 100 bases of subtelomere sequence. Implementation of NEM-SDA$^2$ with *Bst* 3.0 and the two nicking enzymes showed nonspecific amplification, where a majority of amplicons appeared to have been generated from the same or similar molecules. Out of 1357 amplicons, a single read contained telomeric sequence. This read was mapped unambiguously to 5p, but gave a telomere length shorter than expected.

## 5.2 Future directions

The ability to identify telomere lengths for specific chromosomes at scale may provide insight into the relationships between chromosome-specific telomere length, telomere variant frequency, senescence, and oncogenesis. Probing this information at the single scale level level may provide further insight into these processes where somatic cells are mosaic for telomere length, such as in age-related cancers.

Telomere amplification is the last unmet component of a single-cell chromosome-specific telomere length measurement workflow. Addition of a primer-anchoring motif to the telomere overhang, whether by overhang-specific adapter ligation or polydA addition by terminal transferase, may improve the likelihood of obtaining whole-telomere amplicons by SWGA. Sophisticated methods for chimera detection and processing exist which may be applicable to reads generated by SWGA (*107*).

A majority of amplicons generated by NEM-SDA$^2$ with *Bst* 3.0 appeared to have been derived from the same or similar molecules. Others have determined that polymerases with both strand displacement and reverse transcription activity are capable of generating runaway amplification from a single primer (*106*). Evaluation of alternative polymerases with both strand displacement activity and nicking endonuclease compatibility (e.g. Sequenase 2.0, Vent (exo-) Polymerase) may be key to eliminating runaway amplification.

# References

1. A. Bernal, L. Tusell, Telomeres: implications for cancer development. *Int. J. Mol. Sci.* **19** (2018), doi:10.3390/ijms19010294.

2. A. J. Cesare, Z. Kaul, S. B. Cohen, C. E. Napier, H. A. Pickett, A. A. Neumann, R. R. Reddel, Spontaneous occurrence of telomeric DNA damage response in the absence of chromosome fusions. *Nat. Struct. Mol. Biol.* **16**, 1244–1251 (2009).

3. H. Vaziri, W. Dragowska, R. C. Allsopp, T. E. Thomas, C. B. Harley, P. M. Lansdorp, Evidence for a mitotic clock in human hematopoietic stem cells: loss of telomeric DNA with age. *Proc Natl Acad Sci USA*. **91**, 9857–9860 (1994).

4. R. C. Allsopp, H. Vaziri, C. Patterson, S. Goldstein, E. V. Younglai, A. B. Futcher, C. W. Greider, C. B. Harley, Telomere length predicts replicative capacity of human fibroblasts. *Proc Natl Acad Sci USA*. **89**, 10114–10118 (1992).

5. Z. Zhong, L. Shiue, S. Kaplan, T. de Lange, A mammalian factor that binds telomeric TTAGGG repeats in vitro. *Mol. Cell. Biol.* **12**, 4834–4843 (1992).

6. L. Chong, B. van Steensel, D. Broccoli, H. Erdjument-Bromage, J. Hanish, P. Tempst, T. de Lange, A human telomeric protein. *Science*. **270**, 1663–1667 (1995).

7. D. Broccoli, A. Smogorzewska, L. Chong, T. de Lange, Human telomeres contain two distinct Myb-related proteins, TRF1 and TRF2. *Nat. Genet.* **17**, 231–235 (1997).

8. M. Chang, Long telomeres: too much of a good thing. *Biomol. Concepts*. **3**, 387–393 (2012).

9. A. L. Moye, K. C. Porter, S. B. Cohen, T. Phan, K. G. Zyner, N. Sasaki, G. O. Lovrecz, J. L. Beck, T. M. Bryan, Telomeric G-quadruplexes are a substrate and site of localization for human telomerase. *Nat. Commun.* **6**, 7643 (2015).

10. H. E. Fice, B. Robaire, Telomere dynamics throughout spermatogenesis. *Genes (Basel)*. **10** (2019), doi:10.3390/genes10070525.

11. C. Härle-Bachor, P. Boukamp, Telomerase activity in the regenerative basal layer of the epidermis inhuman skin and in immortal and carcinoma-derived skin keratinocytes. *Proc Natl Acad Sci USA*. **93**, 6476–6481 (1996).

12. S. Bougel, S. Renaud, R. Braunschweig, D. Loukinov, H. C. Morse, F. T. Bosman, V. Lobanenkov, J. Benhattar, PAX5 activates the transcription of the human telomerase reverse transcriptase gene in B cells. *J. Pathol.* **220**, 87–96 (2010).

13. E. V. Barsov, Telomerase and primary T cells: biology and immortalization for adoptive immunotherapy. *Immunotherapy*. **3**, 407–421 (2011).

14. Y.-S. Cong, W. E. Wright, J. W. Shay, Human telomerase and its regulation. *Microbiol. Mol. Biol. Rev.* **66**, 407–25, table of contents (2002).

15. S. Makovets, I. Herskowitz, E. H. Blackburn, Anatomy and dynamics of DNA replication fork movement in yeast telomeric regions. *Mol. Cell. Biol.* **24**, 4019–4031 (2004).

16. M. Higa, M. Fujita, K. Yoshida, DNA replication origins and fork progression at mammalian telomeres. *Genes (Basel)*. **8** (2017), doi:10.3390/genes8040112.

17. W. C. Drosopoulos, S. T. Kosiyatrakul, Z. Yan, S. G. Calderano, C. L. Schildkraut, Human telomeres replicate using chromosome-specific, rather than universal, replication programs. *J. Cell Biol.* **197**, 253–266 (2012).

18. E. Bonnell, E. Pasquier, R. J. Wellinger, Telomere replication: solving multiple end replication problems. *Front. Cell Dev. Biol.* **9**, 668171 (2021).

19. C. B. Harley, A. B. Futcher, C. W. Greider, Telomeres shorten during ageing of human fibroblasts. *Nature*. **345**, 458–460 (1990).

20. D. Wynford-Thomas, D. Kipling, The end-replication problem. *Nature*. **389**, 551–551 (1997).

21. N. Arnoult, J. Karlseder, Complex interactions between the DNA-damage response and mammalian telomeres. *Nat. Struct. Mol. Biol.* **22**, 859–866 (2015).

22. Z. Kaul, A. J. Cesare, L. I. Huschtscha, A. A. Neumann, R. R. Reddel, Five dysfunctional telomeres predict onset of senescence in human cells. *EMBO Rep.* **13**, 52–59 (2011).

23. M. T. Hayashi, A. J. Cesare, J. A. J. Fitzpatrick, E. Lazzerini-Denchi, J. Karlseder, A telomere-dependent DNA damage checkpoint induced by prolonged mitotic arrest. *Nat. Struct. Mol. Biol.* **19**, 387–394 (2012).

24. M. T. Hayashi, A. J. Cesare, T. Rivera, J. Karlseder, Cell death during crisis is mediated by mitotic telomere deprotection. *Nature*. **522**, 492–496 (2015).

25. L. Chin, S. E. Artandi, Q. Shen, A. Tam, S. L. Lee, G. J. Gottlieb, C. W. Greider, R. A. DePinho, p53 deficiency rescues the adverse effects of telomere loss and cooperates with telomere dysfunction to accelerate carcinogenesis. *Cell*. **97**, 527–538 (1999).

26. T. Davoli, T. de Lange, Telomere-driven tetraploidization occurs in human cells undergoing crisis and promotes transformation of mouse cells. *Cancer Cell*. **21**, 765–776 (2012).

27. N. W. Kim, M. A. Piatyszek, K. R. Prowse, C. B. Harley, M. D. West, P. L. Ho, G. M. Coviello, W. E. Wright, S. L. Weinrich, J. W. Shay, Specific association of human telomerase activity with immortal cells and cancer. *Science*. **266**, 2011–2015 (1994).

28. J.-M. Zhang, L. Zou, Alternative lengthening of telomeres: from molecular mechanisms to therapeutic outlooks. *Cell Biosci.* **10**, 30 (2020).

29. L. A. Forsberg, D. Absher, J. P. Dumanski, Republished: Non-heritable genetics of human disease: spotlight on post-zygotic genetic variation acquired during lifetime. *Postgrad. Med. J.* **89**, 417–426 (2013).

30. M. Hallek, T. D. Shanafelt, B. Eichhorst, Chronic lymphocytic leukaemia. *Lancet*. **391**, 1524–1537 (2018).

31. C. A. Nichols, W. J. Gibson, M. S. Brown, J. A. Kosmicki, J. P. Busanovich, H. Wei, L. M. Urbanski, N. Curimjee, A. C. Berger, G. F. Gao, A. D. Cherniack, S. Dhe-Paganon, B. R. Paolella, R. Beroukhim, Loss of heterozygosity of essential genes represents a widespread class of potential cancer vulnerabilities. *Nat. Commun.* **11**, 2517 (2020).

32. U. M. Martens, J. M. J. M. Zijlmans, S. S. S. Poon, W. Dragowska, J. Yui, E. A. Chavez, R. K. Ward, P. M. Lansdorp, Short telomeres on human chromosome 17p. *Nat. Genet.* **18**, 76–80 (1998).

33. M. T. Hemann, M. A. Strong, L. Y. Hao, C. W. Greider, The shortest telomere, not average telomere length, is critical for cell viability and chromosome stability. *Cell*. **107**, 67–77 (2001).

34. H. Seifert, B. Mohr, C. Thiede, U. Oelschlägel, U. Schäkel, T. Illmer, S. Soucek, G. Ehninger, M. Schaich, Study Alliance Leukemia (SAL), The prognostic impact of 17p (p53) deletion in 2272 adults with acute myeloid leukemia. *Leukemia*. **23**, 656–663 (2009).

35. M. de L. L. F. Chauffaille, I. Zalcberg, W. G. Barreto, I. Bendit, Detection of somatic TP53 mutations and 17p deletions in patients with chronic lymphocytic leukemia: a review of the current methods. *Hematology, Transfusion and Cell Therapy*. **42**, 261–268 (2020).

36. K. Takubo, N. Izumiyama-Shimomura, N. Honma, M. Sawabe, T. Arai, M. Kato, M. Oshimura, K.-I. Nakamura, Telomere lengths are characteristic in each human individual. *Exp. Gerontol.* **37**, 523–531 (2002).

37. L. Broer, V. Codd, D. R. Nyholt, J. Deelen, M. Mangino, G. Willemsen, E. Albrecht, N. Amin, M. Beekman, E. J. C. de Geus, A. Henders, C. P. Nelson, C. J. Steves, M. J. Wright, A. J. M. de Craen, A. Isaacs, M. Matthews, A. Moayyeri, G. W. Montgomery, B. A. Oostra, D. I. Boomsma, Meta-analysis of telomere length in 19,713 subjects reveals high heritability, stronger maternal inheritance

and a paternal age effect. *Eur. J. Hum. Genet.* **21**, 1163–1168 (2013).

38. D. A. Delgado, C. Zhang, K. Gleason, K. Demanelis, L. S. Chen, J. Gao, S. Roy, J. Shinkle, M. Sabarinathan, M. Argos, L. Tong, A. Ahmed, T. Islam, M. Rakibuz-Zaman, G. Sarwar, H. Shahriar, M. Rahman, M. Yunus, J. A. Doherty, F. Jasmine, B. L. Pierce, The contribution of parent-to-offspring transmission of telomeres to the heritability of telomere length in humans. *Hum. Genet.* **138**, 49–60 (2019).

39. O. T. Njajou, R. M. Cawthon, C. M. Damcott, S.-H. Wu, S. Ott, M. J. Garant, E. H. Blackburn, B. D. Mitchell, A. R. Shuldiner, W.-C. Hsueh, Telomere length is paternally inherited and is associated with parental lifespan. *Proc Natl Acad Sci USA*. **104**, 12135–12139 (2007).

40. J.-H. Kim, C. M. Nam, D. Lee, H. Bang, J.-H. Ko, I. Lim, G. J. Kim, B. W. Koes, D.-C. Lee, Heritability of telomere length across three generations of Korean families. *Pediatr. Res.* **87**, 1060–1065 (2020).

41. J. A. Londoño-Vallejo, H. DerSarkissian, L. Cazes, G. Thomas, Differences in telomere length between homologous chromosomes in humans. *Nucleic Acids Res.* **29**, 3164–3171 (2001).

42. K. Nordfjäll, A. Larefalk, P. Lindgren, D. Holmberg, G. Roos, Telomere length and heredity: Indications of paternal inheritance. *Proc Natl Acad Sci USA*. **102**, 16374–16378 (2005).

43. L. Mirabello, K. Yu, P. Kraft, I. De Vivo, D. J. Hunter, J. Prescott, J. Y. Y. Wong, N. Chatterjee, R. B. Hayes, S. A. Savage, The association of telomere length and genetic variation in telomere biology genes. *Hum. Mutat.* **31**, 1050–1058 (2010).

44. M. Armanios, E. H. Blackburn, The telomere syndromes. *Nat. Rev. Genet.* **13**, 693–704 (2012).

45. A. A. Mangaonkar, M. M. Patnaik, Short telomere syndromes in clinical practice: bridging bench and bedside. *Mayo Clin. Proc.* **93**, 904–916 (2018).

46. M. Feretzaki, P. Renck Nunes, J. Lingner, Expression and differential regulation of human TERRA at several chromosome ends. *RNA*. **25**, 1470–1480 (2019).

47. R. Sharma, S. S. Sahoo, M. Honda, S. L. Granger, C. Goodings, L. Sanchez, A. Künstner, H. Busch, F. Beier, S. M. Pruett-Miller, M. B. Valentine, A. G. Fernandez, T.-C. Chang, V. Géli, D. Churikov, S. Hirschi, V. B. Pastor, M. Boerries, M. Lauten, C. Kelaidi, M. W. Wlodarski, Gain-of-function mutations in RPA1 cause a syndrome with short telomeres and somatic genetic rescue. *Blood*. **139**, 1039–1051 (2022).

48. A. S. N. Alhendi, N. J. Royle, The absence of (TCAGGG)n repeats in some

telomeres, combined with variable responses to NR2F2 depletion, suggest that this nuclear receptor plays an indirect role in the alternative lengthening of telomeres. *Sci. Rep.* **10**, 20597 (2020).

49. V. A. Szalai, H. H. Thorp, Electron transfer in tetrads:  adjacent guanines are not hole traps in G quartets. *J. Am. Chem. Soc.* **122**, 4524–4525 (2000).

50. P. L. Opresko, J. Fan, S. Danzy, D. M. Wilson, V. A. Bohr, Oxidative damage in telomeric DNA disrupts recognition by TRF1 and TRF2. *Nucleic Acids Res.* **33**, 1230–1239 (2005).

51. L. Bendix, P. B. Horn, U. B. Jensen, I. Rubelj, S. Kolvraa, The load of short telomeres, estimated by a new method, Universal STELA, correlates with number of senescent cells. *Aging Cell*. **9**, 383–397 (2010).

52. T.-P. Lai, W. E. Wright, J. W. Shay, Comparison of telomere length measurement methods. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373** (2018), doi:10.1098/rstb.2016.0451.

53. F. Wang, X. Pan, K. Kalmbach, M. L. Seth-Smith, X. Ye, D. M. F. Antumes, Y. Yin, L. Liu, D. L. Keefe, S. M. Weissman, Robust measurement of telomere length in single cells. *Proc Natl Acad Sci USA*. **110**, E1906-12 (2013).

54. Y. Luo, R. Viswanathan, M. P. Hande, A. H. P. Loh, L. F. Cheow, Massively parallel single-molecule telomere length measurement with digital real-time PCR. *Sci. Adv.* **6** (2020), doi:10.1126/sciadv.abb7944.

55. A. R. Wheeler, W. R. Throndset, R. J. Whelan, A. M. Leach, R. N. Zare, Y. H. Liao, K. Farrell, I. D. Manger, A. Daridon, Microfluidic device for single-cell analysis. *Anal. Chem.* **75**, 3581–3586 (2003).

56. R. N. Zare, S. Kim, Microfluidic platforms for single-cell analysis. *Annu. Rev. Biomed. Eng.* **12**, 187–201 (2010).

57. H. S. Lee, W. K. Chu, K. Zhang, X. Huang, Microfluidic devices with permeable polymer barriers for capture and transport of biomolecules and cells. *Lab Chip*. **13**, 3389–3397 (2013).

58. B. Cressiot, S. J. Greive, M. Mojtabavi, A. A. Antson, M. Wanunu, Thermostable virus portal proteins as reprogrammable adapters for solid-state nanopore sensors. *Nat. Commun.* **9**, 4652 (2018).

59. A. Joneja, X. Huang, Linear nicking endonuclease-mediated strand-displacement DNA amplification. *Anal. Biochem.* **414**, 58–69 (2011).

60. Y. Zhang, N. A. Tanner, Isothermal Amplification of Long, Discrete DNA Fragments Facilitated by Single-Stranded Binding Protein. *Sci. Rep.* **7**, 8497 (2017).

61. D. Fologea, E. Brandin, J. Uplinger, D. Branton, J. Li, DNA conformation and base number simultaneously determined in a nanopore. *Electrophoresis*. **28**, 3186–3192 (2007).

62. A. J. Storm, C. Storm, J. Chen, H. Zandbergen, J.-F. Joanny, C. Dekker, Fast DNA translocation through a solid-state nanopore. *Nano Lett.* **5**, 1193–1197 (2005).

63. S. Carson, J. Wilson, A. Aksimentiev, M. Wanunu, Smooth DNA transport through a narrowed pore geometry. *Biophys. J.* **107**, 2381–2393 (2014).

64. M. Charron, K. Briggs, S. King, M. Waugh, V. Tabard-Cossa, Precise DNA Concentration Measurements with Nanopores by Controlled Counting. *Anal. Chem.* **91**, 12228–12237 (2019).

65. D. V. Verschueren, W. Yang, C. Dekker, Lithography-based fabrication of nanopore arrays in freestanding SiN and graphene membranes. *Nanotechnology*. **29**, 145302 (2018).

66. C. Heinz, H. Engelhardt, M. Niederweis, The core of the tetrameric mycobacterial porin MspA is an extremely stable beta-sheet domain. *J. Biol. Chem.* **278**, 8678–8685 (2003).

67. I. M. Derrington, T. Z. Butler, M. D. Collins, E. Manrao, M. Pavlenok, M. Niederweis, J. H. Gundlach, Nanopore DNA sequencing with MspA. *Proc Natl Acad Sci USA*. **107**, 16060–16065 (2010).

68. M. Graf, M. Lihter, D. Altus, S. Marion, A. Radenovic, Transverse detection of DNA using a mos2 nanopore. *Nano Lett.* **19**, 9075–9083 (2019).

69. S. J. Heerema, G. F. Schneider, M. Rozemuller, L. Vicarelli, H. W. Zandbergen, C. Dekker, 1/f noise in graphene nanopores. *Nanotechnology*. **26**, 074001 (2015).

70. K. R. Williams, K. Gupta, M. Wasilik, Etch rates for micromachining processing-part II. *J. Microelectromech. Syst.* **12**, 761–778 (2003).

71. K. Sato, M. Shikida, Y. Matsushima, T. Yamashiro, K. Asaumi, Y. Iriye, M. Yamamoto, Characterization of orientation-dependent etching properties of single-crystal silicon: effects of KOH concentration. *Sensors and Actuators A: Physical*. **64**, 87–93 (1998).

72. S. W. Kowalczyk, A. Y. Grosberg, Y. Rabin, C. Dekker, Modeling the conductance and DNA blockade of solid-state nanopores. *Nanotechnology*. **22**, 315101 (2011).

73. J. Cao, S. Zhang, J. Zhang, S. Wang, W. Jia, S. Yan, Y. Wang, P. Zhang, H.-Y. Chen, S. Huang, A Single-Molecule Observation of Dichloroaurate(I) Binding to

an Engineered Mycobacterium smegmatis porin A (MspA) Nanopore. *Anal. Chem.* **93**, 1529–1536 (2021).

74. A. Chatzilazarou, E. Katsoyannos, O. Gortzi, S. Lalas, Y. Paraskevopoulos, E. Dourtoglou, J. Tsaknis, Removal of polyphenols from wine sludge using cloud point extraction. *J. Air Waste Manag. Assoc.* **60**, 454–459 (2010).

75. T. Z. Butler, M. Pavlenok, I. M. Derrington, M. Niederweis, J. H. Gundlach, Single-molecule DNA detection with an engineered MspA protein nanopore. *Proc Natl Acad Sci USA*. **105**, 20647–20652 (2008).

76. M. Levy-Sakin, S. Pastor, Y. Mostovoy, L. Li, A. K. Y. Leung, J. McCaffrey, E. Young, E. T. Lam, A. R. Hastie, K. H. Y. Wong, C. Y. L. Chung, W. Ma, J. Sibert, R. Rajagopalan, N. Jin, E. Y. C. Chow, C. Chu, A. Poon, C. Lin, A. Naguib, P.-Y. Kwok, Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat. Commun.* **10**, 1025 (2019).

77. S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S. J. Hoyt, M. Diekhans, G. A. Logsdon, M. Alonge, S. E. Antonarakis, M. Borchers, G. G. Bouffard, S. Y. Brooks, G. V. Caldas, A. M. Phillippy, The complete sequence of a human genome. *Science*. **376**, 44–53 (2022).

78. D. Domanska, C. Kanduri, B. Simovski, G. K. Sandve, Mind the gaps: overlooking inaccessible regions confounds statistical testing in genome analysis. *BMC Bioinformatics*. **19**, 481 (2018).

79. S. C. Huang, M. D. Stump, R. Weiss, K. D. Caldwell, Binding of biotinylated DNA to streptavidin-coated polystyrene latex: effects of chain length and particle size. *Anal. Biochem.* **237**, 115–122 (1996).

80. S. S. Shevkoplyas, A. C. Siegel, R. M. Westervelt, M. G. Prentiss, G. M. Whitesides, The force acting on a superparamagnetic bead due to an applied magnetic field. *Lab Chip*. **7**, 1294–1302 (2007).

81. V. Chan, D. J. Graves, S. E. McKenzie, The biophysics of DNA hybridization with immobilized oligonucleotide probes. *Biophys. J.* **69**, 2243–2255 (1995).

82. L. Ying, J. J. Green, H. Li, D. Klenerman, S. Balasubramanian, Studies on the structure and dynamics of the human telomeric G quadruplex by single-molecule fluorescence resonance energy transfer. *Proc Natl Acad Sci USA*. **100**, 14629–14634 (2003).

83. M. R. Green, J. Sambrook, Isolation of High-Molecular-Weight DNA from Suspension Cultures of Mammalian Cells Using Proteinase K and Phenol. *Cold Spring Harb. Protoc.* **2018** (2018), doi:10.1101/pdb.prot093476.

84. R. Kumar Sharma, I. Agrawal, L. Dai, P. S. Doyle, S. Garaj, Complex DNA knots

detected with a nanopore sensor. *Nat. Commun.* **10**, 4473 (2019).

85. C. Plesa, D. Verschueren, S. Pud, J. van der Torre, J. W. Ruitenberg, M. J. Witteveen, M. P. Jonsson, A. Y. Grosberg, Y. Rabin, C. Dekker, Direct observation of DNA knots using a solid-state nanopore. *Nat. Nanotechnol.* **11**, 1093–1097 (2016).

86. 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, M. Loose, Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).

87. K. Masago, S. Fujita, Y. Oya, Y. Takahashi, H. Matsushita, E. Sasaki, H. Kuroda, Comparison between Fluorimetry (Qubit) and Spectrophotometry (NanoDrop) in the Quantification of DNA and RNA Extracted from Frozen and FFPE Tissues from Lung Cancer Patients: A Real-World Use of Genomic Tests. *Medicina (Kaunas)*. **57** (2021), doi:10.3390/medicina57121375.

88. K.-T. Tan, M. K. Slevin, M. Meyerson, H. Li, Identifying and correcting repeat-calling errors in nanopore sequencing of telomeres. *BioRxiv* (2022), doi:10.1101/2022.01.11.475254.

89. M. T. Radukic, D. Brandt, M. Haak, K. M. Müller, J. Kalinowski, Nanopore sequencing of native adeno-associated virus (AAV) single-stranded DNA using a transposase-based rapid protocol. *NAR Genom. Bioinform.* **2**, lqaa074 (2020).

90. W.-M. Zhou, Y.-Y. Yan, Q.-R. Guo, H. Ji, H. Wang, T.-T. Xu, B. Makabel, C. Pilarsky, G. He, X.-Y. Yu, J.-Y. Zhang, Microfluidics applications for high-throughput single cell sequencing. *J. Nanobiotechnology*. **19**, 312 (2021).

91. N. Stong, Z. Deng, R. Gupta, S. Hu, S. Paul, A. K. Weiner, E. E. Eichler, T. Graves, C. C. Fronick, L. Courtney, R. K. Wilson, P. M. Lieberman, R. V. Davuluri, H. Riethman, Subtelomeric CTCF and cohesin binding site organization using improved subtelomere assemblies and a novel annotation pipeline. *Genome Res.* **24**, 1039–1050 (2014).

92. H. Riethman, Human subtelomeric copy number variations. *Cytogenet. Genome Res.* **123**, 244–252 (2008).

93. H. C. Mefford, B. J. Trask, The complex structure and dynamic evolution of human subtelomeres. *Nat. Rev. Genet.* **3**, 91–102 (2002).

94. E. V. Linardopoulou, E. M. Williams, Y. Fan, C. Friedman, J. M. Young, B. J. Trask, Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature*. **437**, 94–100 (2005).

95. J. C. van Deutekom, E. Bakker, R. J. Lemmers, M. J. van der Wielen, E. Bik, M.

H. Hofker, G. W. Padberg, R. R. Frants, Evidence for subtelomeric exchange of 3.3 kb tandemly repeated units between chromosomes 4q35 and 10q26: implications for genetic counselling and etiology of FSHD1. *Hum. Mol. Genet.* **5**, 1997–2003 (1996).

96. S. M. van der Maarel, G. Deidda, R. J. Lemmers, P. G. van Overveld, M. van der Wielen, J. E. Hewitt, L. Sandkuijl, B. Bakker, G. J. van Ommen, G. W. Padberg, R. R. Frants, De novo facioscapulohumeral muscular dystrophy: frequent somatic mosaicism, sex-dependent phenotype, and the role of mitotic transchromosomal repeat interaction between chromosomes 4 and 10. *Am. J. Hum. Genet.* **66**, 26–35 (2000).

97. T.-P. Lai, N. Zhang, J. Noh, I. Mender, E. Tedone, E. Huang, W. E. Wright, G. Danuser, J. W. Shay, A method for measuring the distribution of the shortest telomeres in cells and tissues. *Nat. Commun.* **8**, 1356 (2017).

98. W. K. Chu, P. Edge, H. S. Lee, V. Bansal, V. Bafna, X. Huang, K. Zhang, Ultraaccurate genome sequencing and haplotyping of single human cells. *Proc Natl Acad Sci USA*. **114**, 12512–12517 (2017).

99. M. T. Walsh, A. P. Hsiao, H. S. Lee, Z. Liu, X. Huang, Capture and enumeration of mRNA transcripts from single cells using a microfluidic device. *Lab Chip*. **15**, 2968–2980 (2015).

100. F. B. Dean, S. Hosono, L. Fang, X. Wu, A. F. Faruqi, P. Bray-Ward, Z. Sun, Q. Zong, Y. Du, J. Du, M. Driscoll, W. Song, S. F. Kingsmore, M. Egholm, R. S. Lasken, Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci USA*. **99**, 5261–5266 (2002).

101. E. L. Clarke, S. A. Sundararaman, S. N. Seifert, F. D. Bushman, B. H. Hahn, D. Brisson, swga: a primer design toolkit for selective whole genome amplification. *Bioinformatics*. **33**, 2071–2077 (2017).

102. O. Alsmadi, F. Alkayal, D. Monies, B. F. Meyer, Specific and complete human genome amplification with improved yield achieved by phi29 DNA polymerase and a novel primer at elevated temperature. *BMC Res. Notes*. **2**, 48 (2009).

103. R. S. Lasken, T. B. Stockwell, Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol.* **7**, 19 (2007).

104. S. L. Sholes, K. Karimian, A. Gershman, T. J. Kelly, W. Timp, C. W. Greider, Chromosome-specific telomere lengths and the minimal functional telomere revealed by nanopore sequencing. *Genome Res.* **32**, 616–628 (2022).

105. S. Spitzer, F. Eckstein, Inhibition of deoxyribonucleases by phosphorothioate groups in oligodeoxyribonucleotides. *Nucleic Acids Res.* **16**, 11691–11704 (1988).

106. G. Wang, X. Ding, J. Hu, W. Wu, J. Sun, Y. Mu, Unusual isothermal multimerization and amplification by the strand-displacing DNA polymerases with reverse transcription activities. *Sci. Rep.* **7**, 13928 (2017).

107. N. Lu, J. Li, C. Bi, J. Guo, Y. Tao, K. Luan, J. Tu, Z. Lu, Chimeraminer: an improved chimeric read detection pipeline and its application in single cell sequencing. *Int. J. Mol. Sci.* **20** (2019), doi:10.3390/ijms20081953.

108. I. C. Nova, I. M. Derrington, J. M. Craig, M. T. Noakes, B. I. Tickman, K. Doering, H. Higinbotham, A. H. Laszlo, J. H. Gundlach, Investigating asymmetric salt profiles for nanopore DNA sequencing with biological porin MspA. *PLoS ONE*. **12**, e0181599 (2017).