Genomic mapping of single DNA molecules from *Saccharomyces cerevisiae* (brewer's yeast) by partial denaturation in nanochannels

Robert L. Welch

Master of Science

Department of Physics

McGill University

Montreal, Quebec

August 15th, 2012

A thesis submitted to the Faculty of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science

DEDICATION

This document is dedicated to Mary Koronyi.

ACKNOWLEDGEMENTS

Thank you to my parents, sister, family, friends, Laurel, my colleagues; especially Alex Klotz, Ilja Czolkos, Jaan Altosaar, Ahmed Khorshid, Yuning Zhang, Mathieu Massicotte and James Hedberg at McGill, Peter Østergaard, Rodolphe Marie, Anders Kristensen and Henrik Flyvbjerg at DTU-Nano, and Robert Sladek, Ken Dewar and Said Attiya at Génome Québec; and my supervisor Walter Reisner, for your help and support.

ABSTRACT

Optical mapping of DNA provides large-scale genomic information that could be used to diagnose disease, detect pathogens, and study the re-arrangements between single cells crucial to the development of cancer. A recent optical mapping technique called denaturation mapping has the unique advantage of applying physical principles rather than the action of enzymes to probe genomic structure. Denaturation mapping uses fluorescence microscopy to image the pattern of partial melting along a DNA molecule extended in a channel of cross-section 150nm at the heart of a nanofluidic device. We used denaturation mapping to locate single DNA molecules on the genome of *Saccharomyces cerevisiae* (12.1Mbp). We compared melting patterns to a computationally predicted melting pattern for the entire genome sequence and performed a statistical analysis to show location results were significant. By imaging 84 DNA molecules we assembled an optical map of the yeast genome with >50% coverage. Our results demonstrate the potential of denaturation mapping as single-molecule probe for long-range structure of a eukaryotic, megabase-pair sized genome.

ABRÉGÉ

La cartographie optique de l'ADN fournie de l'information génomique à grande-échelle qui pourrait être utilisée pour diagnostiquer les maladies, détecter les pathogènes, et étudier les réarrangements entre cellules individuelles qui sont cruciaux pour le développement du cancer. Une récente technique de cartographie optique appelée cartographie par dénaturation possède l'avantage unique d'appliquer des principes physique plutôt que l'action d'enzymes pour sonder la structure génomique. La cartographie par dénaturation utilise la microcopie à fluorescence pour imager les motifs créés par la fonte partiel de la molécule d'ADN allongé dans un canal d'un diamètre de 150nm au cœur d'un dispositif nanofluidique. Nous avons utilisé la cartographie par dénaturation pour localiser les molécules d'ADN individuelles dans le génome de Saccharomyces cerevisiae (taille de 12.1 Mbp). Nous avons comparé les motifs de fonte à ceux calculés par ordinateur pour la séquence génomique entière et effectué une analyse statistique pour montrer que les résultats de positionnement étaient significatifs. En imageant 84 molécules d'ADN, nous avons assemblé une carte optique du génome de la levure avec une couverture >50%. Nos résultats démontrent le potentiel de la cartographie par dénaturation comme sonde moléculaire pour des structures à longue portée d'un génome eucaryote ayant plusieurs millions de paires de base.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Genomic mapping

DNA sequencing is a pillar of modern biology. In addition to providing the ultimate basis for the study of genetics, the ability to read the base-pair sequence of complete genomes has profound consequences for medicine. Use of the human genome sequence, and those of pathogens, is now commonplace in the diagnosis of disease and the development of drugs and vaccines.[19] Sequencing technology has seen dramatic improvements in speed and efficiency for three decades: the cost of sequencing has dropped from \$0.50 per kbp (for the parallelized Sanger sequencing employed in the Human Genome Project), to below \$1.00 per Mbp (for next-generation sequencing techniques), leading to speculation that a \$1000 human genome sequence may be obtained in the near future.[20, 49]

Despite such intense development, sequencing has important limitations. The classic Sanger technology produces a sequence that reflects a large number of target DNA molecules, which makes it is incapable of observing genomic differences between DNA from different cells.[49] As well, because the read length of each sequence is only $\sim$ 1000bp the Sanger method alone cannot detect long-range structure.[20] Next-generation techniques have even smaller read lengths (<100bp) and retain the classic technology's inability to distinguish single DNA molecules. [20]

The task of assembling short reads into DNA sequences has lead to the development of complementary technology for *genomic mapping*. [36] A genomic map is a large-scale description of sequence structure. In the context of sequence assembly, a genomic map describes the relative arrangement of cloned DNA fragments. The strategy of mapping a clone library before sequencing greatly reduces the number of clones required to assemble a complete genomic sequence.[40] In addition to its role in sequencing, the information provided by a genomic map can be valuable on its own. Common large-scale genomic changes due to chromosomal re-arrangements (deletions, duplications, inversions, translocations) that are difficult to probe by sequencing are accessible to mapping technology.[17] These re-arrangements are the basis for important genetic diseases [41] and the development of cancer.[43]

Well-established genomic mapping technologies include PCR-based screening of known sequence tagged sites (STS) in clones, [1] sequencing of only small sections of clones, [40] and restriction mapping, which measures the pattern of fragment sizes produced by the action of sequence-specific restriction enzymes.[5] Like the sequencing assays they were developed to support, the use of cloning and amplification steps limits these mapping technologies to probing an average signal over an ensemble of many DNA molecules.

1.2 Optical mapping in nanofluidic systems

During the last decade a tremendous effort has been made to apply fabrication tools from the semiconductor industry to micro- and nanometer-scale fluidic systems for DNA assays.[22] The miniaturization of polymerase chain reaction (PCR) was an early milestone, and PCR-centric devices remain ubiquitous in microfluidic DNA

analysis.[10] In addition to PCR, the most important conventional tools for analyzing PCR products have all been successfully miniaturized and integrated on a PCR-based microfluidic chip; capillary gel electrophoresis [54], hybridization probes[47] and immunoassays.[50]

More recently however, a new generation of nanofluidic devices has emerged to perform entirely new types of DNA mapping assays. Nanofluidic devices exploit the small feature sizes uniquely accessible to microfabrication in order to alter the physical conformation of DNA. Confining a DNA molecule to a nanochannel of dimensions comparable to its radius of gyration ($\sim$100nm) forces the molecule's contour to unscroll into a linear shape where sequence position is linearly related to channel position.[44] In this arrangement, fluorescently labeling the DNA molecule in a sequence-dependent way will produce a gray-scale pattern of intensity with respect to position. This pattern of fluorescence, referred to as a *DNA barcode*, is a unique, coarse-grained description of the sequence.[27] The DNA barcodes recorded using a nanochannel devices are optical maps of the sequence of a single DNA molecule. This strategy can achieve a resolution, limited by diffraction of the barcode signal, as fine as $\sim$1000kbp.

Nanochannel devices are an elegant plaform for optical mapping. Nanochannel structures linearize DNA passively, without the need for an external apparatus to actively apply pressure or electric fields. The protocol of confining and imaging DNA molecules can be easily parallelized and automated in order to collect statistics

on many molecules or map large sequences. As well because DNA remains in solution during confinement, a molecule's shape represents its equilibrium conformation, rather than a distorted shape that can occur when it is deposited on a surface.

Nanochannel devices could offer an improved complementary technology to DNA sequencing. The bioinformatic task of assembling sequencing reads remains a major obstacle to de novo sequencing. [39] Nanochannel-based optical maps could expedite sequence assembly by providing a large-scale scaffolding over which to assemble contigs quickly and more inexpensively than traditional mapping methods.

Nanochannel optical mapping also offers important advantages over sequencing assays. The power to confine and image large molecules (>100kbp) allows nanofludic optical mapping methods to probe large large-scale genomic structure more quickly than ever before, and to probe repetitive regions that remain inaccessible with current sequencing tools.[20] An especially promising application of nanofluidic devices is the miniaturization of DNA assays to single molecule scale. The fact that nanofluidic optical mapping occurs without amplification or cloning steps makes it able tp detect differences between single DNA molecules. It has been suggested that improved nanochannel devices may one day map the entire genome of a single cell, overcoming the fundamental limitation of conventional assays to ensemble averages of many cells.[27]

The principle of optical mapping in nanochannels has been demonstrated with several labeling schemes. Restriction digesting of non-specifically dyed molecules in nanochannels will produce a sequence-dependent fluorescence pattern when gaps open up between digested fragments.[35] Enzymatic nicking followed by hybridization

of sequence-specific probes will also produce a signal that can form an optical map. [11] It should be noted that both of these approaches rely on the action of enzymes while DNA is confined in the nanofluidic device.

1.3 Denaturation mapping

A recent optical mapping technique developed by Reisner and Austin, called *denaturation mapping*, relies on the physical principle of partial melting.[33] The probability that DNA will denature (melt from double to single strands) is sequence-dependent because G-C base pairs melt at a lower temperature than A-T pairs. DNA is uniformly stained with an intercalating dye that unbinds from single-stranded regions and then partially melted under confinement in a nanochannel. The resulting pattern of fluorescence, where unmelted regions appear bright and melted regions appear dark, is a DNA barcode that reflects the degree of meltedness as a function of sequence position. Figure 1–1 depicts this strategy schematically. Note that the melting pattern detected optically cannot resolve features at the scale of single base-pairs: the example sequence is given only to illustrate the principle that melting is sequence dependent.

Among nanochannel optical mapping strategies denaturation has the unique advantage of occurring without enzymatic labeling or reaction steps, requiring only a uniform dye applied outside of the device. It applies physical, rather than chemical, principles to achieve sequence specificity. The simplicity of its protocol makes denaturation mapping especially attractive as a cost-effective and scaleable genomic mapping tool. As well, denaturation mapping is the only nanochannel-based optical mapping strategy that does not permanently modify the structure of DNA: melting

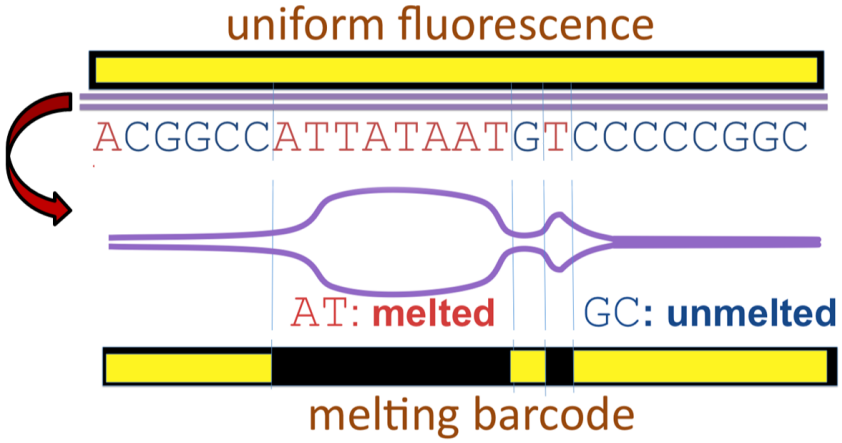

Figure 1–1: Schematic of denaturation mapping

is reversible. This makes it feasible to integrate denaturation mapping components in a larger microfluidic system where other assays follow. For example, it may be possible to selectively sequence part of a genome by identifying DNA fragments using denaturation mapping, and then choosing which enter a sequencing chamber downstream.

Here we used denaturation mapping to locate single DNA molecules globally on a genome of size 12.1Mbp (yeast). The successful genomic location of single molecules, rather than an average of an ensemble of molecules, is a significant step in establishing the denaturation approach as a optical genomic mapping technology. We located 84 DNA molecules to achieve an optical map with more than 50% coverage of the yeast genome. The ability to consistently record and genomically locate 100kbp+ optical maps establishes our method as a powerful probe of large-scale genomic structure.

This thesis is structured as follows. Chapters 2 and 3 are overviews of the physical theory that describes behaviour of DNA in denaturation mapping experiments. Chapter 2 concerns the physics of polymers in solution and in confinement. The models pertaining to the different physical regimes of polymer mechanics are reviewed and used to derive the scaling of mechanical properties relevant to the experiment with the parameters of the DNA and device. At the end of the chapter the scaling laws are summarized and used to make predictions of the behaviour of DNA in our nanochannel devices. Chapter 3 concerns the physics of DNA melting. The conceptual basis of the algorithms used to predict melting barcodes of the yeast genome in this work is outlined, and the experiments to measure the physical parameters used in our calculations are briefly described. Chapter 4 and 5 explain the methods used here to record and analyze melting barcodes. Chapter 4 gives details on the design and fabrication of the nanofluidic device, and the protocol of performing denaturation experiments. A few contributions of the author to the refinement of the denaturation method are described: preliminary work on a new plastic-based device and an improved chuck, and a protocol for cleaning devices for repeated use. Chapter 5 concerns the suite of computer algorithms used to produce and analyze melting barcodes. Chapter 6 summarizes the results of our experiments to map single molecules of yeast DNA. Chapter 7 is a critical discussion of the results given in the previous chapter. The limitations of the data are explained, specific challenges to the method are summarized and their impact on the data is accounted for, and directions to refine the protocol of denaturation mapping are suggested. Finally,

Chapter 8 is a summary of the findings of this work, their significance, and next steps for the technique of denaturation mapping.

CHAPTER 2
Physics of DNA in confinement

2.1 Polymers in solution

A polymer is a macromolecule made of repeating, connected monomers. This definition forms a general class of molecules that includes topologies with branches and rings. Specifically, DNA is a linear polymer, in which the monomers are connected in linear sequence. The curve through space that the sequence of monomers occupies is called the contour of a polymer, and the arrangement of that contour is called its conformation. The simplest arrangement in which to consider a polymer's conformation is in solution, where it exists free from external forces in all three dimensions.

The behaviour of a polymer in solution can be described using models that vary in scope and complexity. Two important features of polymer physics are *semiflexibility*, the ability of a polymer to bend over a finite length scale, and *self-avoidance*, the tendency of a part of a polymer to avoid other parts. This section reviews three models: the *ideal chain model* which accounts for neither, the *wormlike chain model* which accounts for semiflexibility, and the *Flory model* which accounts for self-avoidance. Each model predicts the scaling of two essential mechanical properties with the size of a polymer: end-to-end distance R and free energy F. These descriptions of polymer physics in bulk will become necessary later in this chapter to model the behaviour of polymers in confinement.

2.1.1 An ideal chain

Consider the space a polymer in a solvent occupies as a lattice in three dimensions. The contour of a polymer can be described as a series of steps between lattice sites. Each step is of equal length, the directions in which a step can be taken have equal statistical weight, and the directions of each step are independent from one another. This description lacks semiflexibility because the polymer can turn in any direction at each step and lacks self-avoidance because the polymer can cross itself. The physical parameters of the polymer's contour are the step length a and the number of steps N. This model is referred to as the ideal chain, or freely jointed chain (FJC) where a is called the Kuhn length. The description of the model reviewed here and the scalings that result from it are due to de Gennes.[12]

The *end-to-end distance* R of a polymer is the distance that separates the start of the first step and end of the last step on the lattice. If we consider the end-to-end vector $\mathbf{R}$ as the sum of vectors $\mathbf{a}_n$ describing each step, we can write a simple expression for the average of the squared end-to-end distance:

$$\langle \mathbf{R}^2 \rangle = \langle (\sum_{n=1}^{N} \mathbf{a}_n) \bullet (\sum_{m=1}^{N} \mathbf{a}_m) \rangle = \langle \sum_{n,m=1}^{N} \mathbf{a}_n \bullet \mathbf{a}_m \rangle = \langle \sum_{n=1}^{N} \mathbf{a}_n^2 \rangle = Na^2 \qquad (2.1)$$

where the second equality is possible because the steps are statistically independent and the third because non-orthogonal steps disappear from the dot product. We can use the relation in 2.1 to approximate the scaling of R as:

$$R_{idealchain} \sim \sqrt{\langle \mathbf{R}^2 \rangle} \sim N^{1/2} a. \qquad (2.2)$$

The free energy F can be computed by considering the entropy S of the polymer. From the thermodynamic definition $F = E - TS$, if we consider the internal energy E to be constant with respect to the length of the polymer N, free energy will take the form $F(R) = F_0 + TS(R)$, where F_0 is independent of R. Entropy is proportional to the logarithm of the number of paths of that can form the end-to-end distance $\mathcal{N}_N(R)$: $S(R) \sim ln(\mathcal{N}_N(R))$. Since each path is composed of a large number of statistically independent steps, it can be shown by the central limit theorem that the probability $p_N(R)$ of a path of N subunits reaching an end-to-end distance R is a Gaussian function of the form: [12]

$$p_N(R) \sim N^{-3/2} \exp\left(-\frac{3}{2}\frac{R^2}{\langle \mathbf{R}^2 \rangle}\right). \tag{2.3}$$

The fact that the number of paths with R end-to-end distance $\mathcal{N}_N(R) \sim p_N(R)$, the probability that a walk will achieve that distance, allows the entropy to be written as

$$S(R) \sim ln(\mathcal{N}_N(R)) \sim ln(p_N(R)) \sim S_0 - \frac{3}{2}\frac{R^2}{\langle \mathbf{R}^2 \rangle} \tag{2.4}$$

where S_0 is independent of R. Finally, using the result that $\langle \mathbf{R}^2 \rangle = Na^2$, we arrive at a scaling for free energy of an ideal chain:

$$F_{idealchain} \sim \frac{3}{2}\frac{R^2}{Na^2}. \tag{2.5}$$

2.1.2 A semiflexible chain

While an ideal chain is able to bend in all directions at each monomer, in a real polymer the directions available to the contour are limited. Interactions between monomers lead to a bending energy that make some bending angles unfavourable

or even impossible. Modeling a polymer as semiflexible takes into account the finite length scale along the contour at which a polymer can bend back on itself. The length scale of bending is quantified as the *persistence length* as follows. Consider a polymer as a continuous curve of length L. (L is referred to as the *contour length* and corresponds to $L = Na$ of an ideal chain.) Define a vector $\mathbf{f}(s)$ tangent to the contour of the polymer at a position s along the contour. If the polymer is randomly oriented, the average correlation between tangent vectors will decay exponentially:

$$\langle \mathbf{f}(s) \bullet \mathbf{f}(s') \rangle = \exp\left(-\frac{(s-s')}{P}\right). \tag{2.6}$$

The length scale of this decay P is the persistence length. P can be thought of as the distance along which the contour changes direction. This model is sometimes referred to as the Kratky-Porod, or wormlike chain (WLC) model. The account of the model reviewed here is from the textbook of Doi and Edwards.[13]

The concept of the tangent vector $\mathbf{f}(s)$ also gives an estimate for the end-to-end length. If the tangent vector is given the magnitude of an increment along the contour ds, we can formulate $\langle \mathbf{R}^2 \rangle$ by integration as follows:

$$R^2_{semiflexible} \approx \langle \mathbf{R}^2 \rangle = \int_{s=0}^{L} ds \int_{s'=0}^{L} ds' \langle \mathbf{f}(s) \bullet \mathbf{f}(s') \rangle = 2PL\left(1 + \frac{P}{L}[\exp\left(-\frac{L}{P}\right) - 1]\right) \tag{2.7}$$

solving the integral by substituting relation 2.6. In the limit of a short polymer where $L \ll P$, 2.7 gives the end-to-end distance of a rigid rod: $R^2_{semiflexible} \approx L^2$.

In the limit of a long polymer where $P \ll L$, 2.7 gives $R^2_{semiflexible} \approx 2PL$. Recall that for an ideal chain $R_{ideal} = Na^2$. R_{ideal} and $R_{semiflexible}$ are equivalent if $L = Na$ (by definition) and $P = a/2$. Given this equivalence, the same argument

can be applied as with the ideal chain model to compute free energy, and the same scaling relation for free energy is obtained. We can conclude that semiflexibility does not affect the mechanical properties of a long polymer in solution. In confinement, however, the distinction between ideal and semiflexible chains will become important.

2.1.3 A self-avoiding chain

A polymer that is self-avoiding occupies finite volume and cannot pass through itself. In order to account for the first property it is necessary to define a length scale w, as the *effective width* of the polymer, the extent perpendicular to the direction of contour that other portions of contour cannot enter. In real polymers w is determined by electrostatic interactions, which means the effects of self-avoidance are dependent on salt concentration.[42] For DNA self-avoidance is more dramatic at low salt concentrations where the screening of electrostatic effects is weak. Generally self-avoidance is more important for long polymers because a longer contour is more likely to encounter itself. Flory borrowed ideas from a model of non-ideal gases to describe the behaviour of a self-avoiding chain. The description of Flory's model and its scaling arguments reviewed here are due to de Gennes.[12]

A polymer in solution has the conformation of a randomly oriented coil. Within the coil the polymer can be considered as a "gas" of solid particles, each of which occupies an *excluded volume* inaccessible to the others. Each particle has a length l and width of the effective width w. The excluded volume of each particle is $\chi \approx wl^2$. The total extent of this coil, a quantity called the *radius of gyration* R_g, scales with the end-to-end distance $R \approx \sqrt{\langle \mathbf{R}^2 \rangle}$, and so the volume of the coil scales with R^3.

If there are N particles the concentration of particles in the coil is $c \sim N/R^3$. From these quantities we can compute F as a function of R.

According to the equipartition theorem each quadratic degree of freedom in the gas contributes $\frac{1}{2}k_BT$ to the free energy of the polymer. We can approximate the number of these degrees of freedom by borrowing an idea from the van der Waals model: count each possible excluded volume interaction between two particles as a degree of freedom. The total number of these interactions in the gas the product of the excluded volume per particle χ, the squared of the concentration of particle c^2 (since two particles must interact), and the total volume of the gas where those interactions can occur R^3. Substituting these quantities and simplifying gives the free energy due to excluded volume interactions:

$$F_{excl.volume} = (\frac{1}{2}k_BT)\frac{wL}{R^3}. \tag{2.8}$$

The total free energy of a self-avoiding polymer is a sum of this free energy and the free energy in the absence of excluded volume interactions. The second term, due to the entropy of the coil, is simply the free energy of an ideal chain. Setting $a = l$ and $N = L/l$ in the expression for free energy of an ideal chain $F_{ideal} \sim \frac{3}{2}\frac{R^2}{Na^2}$ gives a scaling for the free energy due to entropy:

$$F_{entropy} \sim \frac{3}{2}\frac{1}{lL}R^2. \tag{2.9}$$

Finally, using these two results for free energy we can determine the scaling of the end-to-end distance R in the Flory model. Writing the total free energy $F = F_{excl.volume} +$

$F_{entropy}$ and minimizing with respect to R gives:

$$R_{Flory} \sim (wl)^{1/5} L^{3/5}. \tag{2.10}$$

This scaling represents the end-to-end distance R of the conformation that the polymer adopts at thermal equilibrium.

2.2 Polymers in confinement

The randomly coiled conformation of a polymer in solution is difficult to follow experimentally. Even if a microscope could resolve a polymer's unpredictable twists and turns, thermal fluctuations quickly re-arrange its contour to a new conformation. The task of distinguishing parts of the contour is important to study DNA because its function in nature revolves around the spatial ordering of base-pairs in the contour. In order to track position along a DNA molecule's contour there needs to be a one-to-one correspondence between position in space and position in the DNA sequence. A way of establishing this correspondence is to confine the polymer to a channel with walls it cannot pass through. Here a channel refers to a pseudo-one-dimensional geometry where the polymer is unconfined in one direction, but confined to an extent of length D in the other two. (This channel can have a circular or square cross section, or any number of other shapes, and the essential physics remains the same). Figure 2–1 indicates this geometry schematically, where the top wall of the channel is not depicted. By restricting the conformations available to the polymer, confinement in a channel will force the polymer to "stretch out" into a linearized conformation where position along the channel is related to position along the contour.[27] Figure 2–2 is an image of a fluorescently stained DNA molecule with a coiled conformation

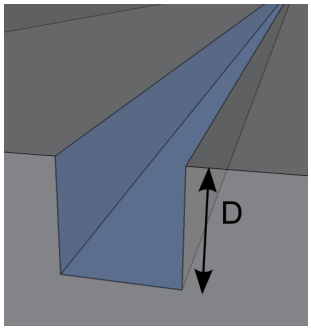

Figure 2–1: Schematic of a channel geometry

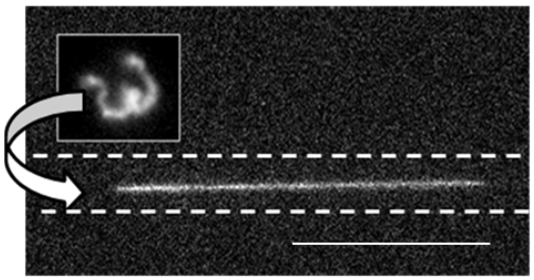

Figure 2–2: A DNA molecule confined in a channel of width ∼120nm

in solution (inset) and linearized conformation while confined in a channel (dotted lines). (Scale bar 10 µm)

Two quantities important to the design a device that controls a polymer by confinement in a channel are *extension* and *relaxation time*. Extension is the spatial extent of a polymer in the direction parallel to the channel. Because extension scales with the end-to-end distance R of this chapter, R will be used to denote extension as well. Extension determines the degree to which a polymer is linearized compared to the total contour length L. As a result, the extension of a DNA molecule will determine the resolution in base-pairs of an optical mapping technique that is limited in spatial resolution by diffraction. Relaxation time, denoted τ, is the time scale

over which a polymer will reach thermal equilibrium. Measurements of a polymer's conformation must occur over a time scale longer than the relaxation time in order to be considered averages of thermal fluctuations. Conversely, measurements that occur more quickly than the relaxation time cannot be considered to measure an equilibrium conformation.

This chapter reviews the scaling of R and τ with the confinement dimension D. Confinement first affects the conformation of a polymer once D approaches the extent of the coil in solution R_g. For $D < R_g$ there exist two broad scaling regimes: the *de Gennes regime* where $D \approx R_g$ and excluded volume effects are most important, and the *Odijk regime* at greater confinement where $R < P$ and bending energy dominates. Furthermore between these regimes where $2P < D$ there exists an intermediate regime called the *extended de Gennes regime* with its own scaling law for free energy. In this section the physical factors for each regime leading to the scaling of R and F are briefly reviewed, and in the following section the scalings of R and F each regime are used to determine an expression for τ.

2.2.1 The de Gennes regime

Consider a polymer confined to a channel of dimension close to its extent in solution $D \approx R_g$. de Gennes argued that the polymer's contour can be thought of as a series of symmetrical *blobs* of size D along the length of the channel. The description of de Genne's mode reviewed below is from his 1979 text.[12]

The blobs repel each other like hard spheres due to self-avoidance. Within a blob, however, the polymer behaves as a self-avoiding chain in solution. Namely, there is a length scale along the contour L_b below which the polymer does not interact

with the walls of the channel. At the scale of L_b the extent of the blob R_b obeys the Flory relation for self-avoiding chain 2.10: $R_b \sim (wl)^{1/5} L_b^{3/5}$. The fact that R_b must also scale with the channel dimension D allows this scaling to be re-arranged as $L_b \sim D^{5/3}(wl)^{1/3}$. In turn, because the number of blobs is the total contour divided by the contour per blob L/L_b and the extension of each blob scales with D, this scaling for L_b gives a prediction for the extension of the polymer $R_{deGennes}$:

$$R_{deGennes} \sim \frac{L}{L_b} D \sim L \frac{(wl)^{1/3}}{D^{2/3}}. \tag{2.11}$$

The scaling of the free energy due to confinement found simply by assigning $k_B T$ to each blob and counting the number of blobs L/L_b that the polymer forms in the channel:

$$\Delta F_{deGennes} \sim k_B T \frac{L}{L_b} \sim k_B T L \frac{(wl)^{1/3}}{D^{5/3}}. \tag{2.12}$$

This free energy is denoted ΔF to distinguish it from free energy in the absence of confinement.

The extended de Gennes regime

Odijk determined that if the channel dimension is decreased from the de Gennes regime, there is a regime where de Gennes' description of a polymer as a series of blobs continues to hold but the blobs lose their symmetry.[29] As well as a dimension perpendicular to the channel D the blobs take on a different dimension parallel to the channel H. This regime is called the *extended de Gennes regime* because the conceptual description is similar to de Gennes' but the asymmetry of blobs leads to a different scaling of free energy. The arguments for this scaling reviewed below follow the account of Doi and Edwards. [13]

The free energy scaling of the de Gennes regime requires the contour within a blob to behave as a self-avoiding chain. Recall that the effects of self-avoidance are more important for longer contours. The extended de Gennes regime occurs when D is sufficiently small that the contour in a blob L_b is too short to follow Flory's model of self-avoidance. The critical length of contour L_c at which the Flory theory breaks down is given by a more general version of Flory's theory, noted by Doi and Edwards, as $L_c \approx l^3/w^2$. [13] Making use of the de Gennes scaling for the contour in a blob $L_b \sim D^{5/3}/(wl)^{1/3}$ and simplifying gives the confinement dimension $D_c \approx l^2/w$ at which the Flory theory breaks down. Therefore the extended de Gennes regime lies in the range of channel dimensions $D < D_c$.

A blob whose contour $L_b \approx L_c$ is at the border of ideal chain and self-avoiding chain models. We can determine how the blob width H scales with the parameters of the channel and polymer by applying both kinds of statistics. Firstly, recall that the scaling of R for an ideal chain in solution is $R_{ideal} \sim N^{1/2}a$. We can write this in the language of a Flory chain by setting $N = L/l$ and $a = l$ to give $R_{ideal} \sim (lL)^{1/2}$. For an asymmetric blob the end-to-end distance R scales with the blob width H rather than D, and so we can write a scaling for H from ideal chain statistics:

$$H_{ideal} \sim (lL_b)^{1/2}. \tag{2.13}$$

We can obtain a second scaling for H using the fact that self avoidance still acts to separate blobs. Therefore the excluded volume due to a blob's contour $L_b^2 w$ should scale with the total volume of the blob HD^2, which gives:

$$H_{excl.vol} \approx \frac{L_b^2 w}{D^2}. \tag{2.14}$$

Considering $H_{excl.vol}$ and H_{ideal} as equivalent and solving relations 2.13 and 2.14 gives scalings of the blob dimensions L_b and H in terms of only the physical parameters of the polymer and channel:

$$H \sim \frac{(Dl)^{2/3}}{w^{1/3}}, L_b \sim \frac{l^{1/3}D^{4/3}}{w^{2/3}}. \tag{2.15}$$

Finally, these scalings of L_b and H can be used to predict a scaling for the polymer's free energy. The relations in 2.15 can be substituted into a generalized expression for free energy $F(R, H, L_b)$ for a chain of asymmetric blobs given by Doi and Edwards (whose derivation is omitted here for brevity).[13] The resulting expression for the free energy of confinement is:

$$\Delta F_{ex.deGennes} \sim k_B T(\frac{R^2}{lL} + \frac{L^2 w}{RD^2}). \tag{2.16}$$

This scaling is completely different from the scaling $\Delta F_{deGennes} \sim L\frac{(wl)^{1/3}}{D^{5/3}}$ of the de Gennes regime.

Perhaps surprisingly, the scaling of R with confinement dimension does *not* change from the de Gennes regime. The reason is simple: the interaction between blobs is distinct from the interactions within a blob. Even if the contour within a blob acts as an ideal chain, the polymer on the scale of blobs continues to experience self-avoidance. The blobs themselves continue to repel each other as hard objects. The same description of extension as the sum of the lengths of blobs continues to apply, and so $R_{ex.deGennes} \sim 1/D^{2/3}$.

The existence of the extended de Gennes regime is remarkable because it demonstrates that confinement can make self-avoidance act on a polymer free from self-avoidance effects in solution. A polymer in solution with contour length $L < L_c$ behaves as an ideal chain. However such a polymer still may have sufficient contour $L > L_b$ to form blobs while in confinement, and the blobs will interact with each other by self-avoidance. In this example confinement alters the fundamental statistics of the polymer's free energy. This is not the case with a polymer in the de Gennes regime, where $L < L_c$ and so self-avoidance effects are important in both solution and confinement.

2.2.2 The Odijk regime

Consider a polymer is confined to a channel of dimension $D < P$. At this length scale, bending energy prevents the polymer from turning back on itself to form coils, so it is no longer possible to describe the polymer using blobs. Odijk proposed a model of this regime where the contour within the channel is described as a rigid rod. The polymer interacts with the wall by deflecting to change its direction, and so the contour can be thought of as a series of rigid segments separated by deflections.[28]

Odijk defined the *deflection length* λ as the average length of contour that separates deflection events. He argued that because the volume occupied by a rigid segment touching both walls scales with PD^2, λ obeys the cube root of this scaling $\lambda \sim (PD^2)^{1/3}$. A brief review of Odijk's argument of the scaling of R and ΔF from this fact follows.

The total extension of a polymer R can be written as the product of the number of rigid segments L/λ and the average extension of each segment parallel to the channel. The average extension of a segment is $\lambda cos(\Theta)$ where Θ is defined as the average angle of deflection, which gives $R = \frac{L}{\lambda}\lambda cos(\Theta) = Lcos(\Theta)$. By approximating the angle Θ as small this expression becomes $R = L(1 - \frac{1}{2}(\frac{D}{\lambda})^2)$. Finally, by substituting the scaling of $\lambda \sim (PD^2)^{1/3}$ we arrive at the scaling for extension:

$$R_{Odijk} \sim L(1 - A(\frac{D}{P})^{2/3}) \tag{2.17}$$

where A is a numerical prefactor.

As in the other regimes of confinement we can estimate ΔF by applying the equipartition theorem and assigning $\frac{1}{2}k_BT$ to each quadratic degree of freedom. Here the number of degrees of freedom is simply the number of rigid segments L/λ. Substituting the scaling of $\lambda \sim (PD^2)^{1/3}$ gives the result:

$$\Delta F_{Odijk} \sim k_BT\frac{L}{P^{1/3}D^{2/3}}. \tag{2.18}$$

2.2.3 Relaxation time

To predict the relaxation time of a polymer in confinement it is useful to consider the polymer as a single, uniform Hookean spring. While a real polymer accommodates a large number of vibrational modes, the scaling of the relaxation time of its fundamental mode is a good starting point to consider thermal fluctuations in general. As a Hookean spring the polymer obeys the equation of motion:

$$\xi\frac{d}{dt}\delta R(t) = -k_{eff}\delta R(t) + B(t) \tag{2.19}$$

where $\delta R(t)$ is the fluctuation in extension R, ξ is a friction factor for the polymer, k_{eff} is the effective spring constant of the polymer, and B is a Brownian noise term formulated to obey the Einstein relation for the diffusion constant $D = k_B T/\xi$ when $k_{eff} = 0$. Over time scales sufficiently large that the system appears overdamped ($(t-t') \gg \xi/k_{eff}$), it can be shown that the solution to equation 2.19 has a correlation that decays exponentially with time:

$$\langle \delta R(t) \delta R(t') \rangle \approx \frac{k_B T}{k_{eff}} \exp\left(-\frac{k_{eff}}{\xi}(t-t')\right). \tag{2.20}$$

The time scale of this decay $\tau = \xi/k_{eff}$ is the relaxation time. The approach used here to defining τ in terms of correlations in R is consistent with the intuitive idea that τ is the time scale over which thermal fluctuations can be averaged.

The definition $\tau = \xi/k_{eff}$ reduces the task of predicting the relaxation time to finding the effective spring constant k_{eff} and friction factor ξ of a polymer. A general strategy for finding k_{eff} in a regime of confinement, given knowledge of that regime's scaling F with respect to R, is to apply the fact that $k_{eff} = \frac{d^2F}{dR^2}$ for free energy due to elastic energy in a Hookean spring. ξ, in turn, is determined by calculating the Newtonian drag constant of the geometry that models the conformation of the polymer in that regime. A brief review is given below of how k_{eff} and ξ give an estimate of τ for the three regimes of confinement outlined in this chapter.

Relaxation time in the de Gennes and extended de Gennes regimes

The contour is described as a series of hard blobs in both the de Gennes and extended de Gennes regime. Regardless of a blob's extent parallel to the channel, the cross-section of the channel occupied by the blob is $\approx D^2$ in both regimes. The

classical result for the drag constant of a sphere with cross-sectional area D^2 in a Stokes flow is simply $6\pi\eta D$. [23]. The drag constant of the polymer is the sum of drag constants for each blob. Recalling that there are (L/L_b) blobs in the chain, and that L_b scales as $L_b \sim D^{5/3}/(wl)^{1/3}$ in both regimes, the drag constant of the polymer can be written:

$$\xi^{deGennes} \approx \xi^{ex.deGennes} \approx 6\pi\eta L \frac{(Pw)^{1/3}}{D^{2/3}}. \tag{2.21}$$

The spring constant, as well, can be estimated with a similar argument for both regimes. Start from a generalized version of the free energy of a Flory chain, given in Doi and Edwards and mentioned in the derivation of $F_{ex.deGennes}$:[13]

$$F_{Flory} \approx k_B T[\frac{R^2}{H^2}(L/L_b)^{-1} + \frac{H}{R}(L/L_b)^2] \tag{2.22}$$

Differentiating twice with respect to R twice gives:

$$\frac{dF_{Flory}}{dR^2} \approx 2k_B T[\frac{1}{H^2}(L/L_b)^{-1} + \frac{2H}{R^3}(L/L_b)^2]. \tag{2.23}$$

For small changes in R the second term will be negligible compared to the first, and we can reduce this expression to $\frac{dF_{Flory}}{dR^2} \approx 2k_B T \frac{1}{H^2}(L/L_b)^{-1}$. From this expression we can eliminate two variables using scalings from the discussion of the extended de Gennes regime. Substituting the de Gennes scaling $L_b \sim D^{5/3}/(wl)^1/3$ gives:

$$\frac{dF_{Flory}}{dR^2} \approx 2k_B T \frac{1}{L}\frac{1}{H^2}\frac{D^{5/3}}{(wl)^{1/3}}. \tag{2.24}$$

Distinguishing the de Gennes and extended de Gennes regimes is now a matter of setting the blob width H. In the de Gennes regime, the blobs are symmetrical so

$H = D$. In the extended de Gennes regime, the scaling $H \sim \frac{(Dl)^{2/3}}{w^{1/3}}$ from relation 2.15 applies. The two choices produce the following estimates for the spring constant:

$$k_{eff}^{deGennes} \approx \frac{k_B T}{L} \frac{1}{(Dwl)^{1/3}}, k_{eff}^{ex.deGennes} \approx \frac{k_B T}{L} \frac{1}{l}. \tag{2.25}$$

Finally, the results in relation 2.25 for k_{eff} can be substituted into the ratio $\tau = \xi / k_{eff}$ to give an estimate for the relaxation time of a polymer in the de Gennes and extended de Gennes regimes:

$$\tau^{deGennes} \approx \frac{\eta}{k_B T} L^2 \frac{(lw)^{2/3}}{D^{1/3}}, \tau^{ex.deGennes} \approx \frac{\eta}{k_B T} L^2 \frac{l(lw)^{1/3}}{D^{2/3}}. \tag{2.26}$$

Note that in both regimes the τ increases as D decreases (although with different scalings), and that both regimes have the same dependence on contour length L.

Relaxation time in the Odijk regime

Rearranging $R_{Odijk} \sim L(1 - A(\frac{D}{P})^{2/3})$ (relation 2.17) in terms of D, substituting for D in the free energy scaling $\Delta F_{Odijk} \sim k_B T \frac{L}{P^{1/3} D^{2/3}}$ (relation 2.18), and then taking the second derivative of F with respect to R gives an estimate for the spring constant:

$$k_{eff}^{Odijk} \approx \frac{k_B T}{PL} (\frac{P}{D})^2. \tag{2.27}$$

In the Odijk regime the contour is described as a series of rigid rods. The drag constant on the rods as they move through the channel can be approximated as the drag on a cylinder moving inside a larger cylinder through a Stokes flow. Using the result of classical fluid mechanics for this geometry gives the estimate: [23]

$$\xi^{Odijk} \approx \frac{2\pi\eta L}{lnD/w} \tag{2.28}$$

where η is the kinematic viscosity of the solvent.

Combining relations 2.27 and 2.28 gives an estimate for the time constant:

$$\tau^{Odijk} \approx \frac{2\pi\eta}{k_B T P} \frac{D^2}{ln(D/w)L^2}. \tag{2.29}$$

Note that in the Odijk regime τ *decreases* as $D \to 0$, in contrast to the de Gennes and extended de Gennes regimes where τ increases as $D \to 0$.

2.3 Summary of scaling relations

Model	F	R	
Ideal chain	$\sim \frac{3}{2}\frac{R^2}{Na^2}$	$\approx N^{1/2}a$	
Semiflexible chain	$\sim \frac{3}{2}\frac{R^2}{(2PL)^2}$ (for $P \ll L$)	$\approx \sqrt{2PL}$ (for $P \ll L$)	
Self-avoiding chain	$\sim \frac{3}{2}\frac{1}{lL}R^2 + (\frac{1}{2}k_BT)\frac{wL}{R^3}$	$\sim (wl)^{1/5}L^{3/5}$	
Confinement regime	ΔF	R	τ
de Gennes	$\sim k_BTL\frac{(wl)^{1/3}}{D^{5/3}}$	$\sim L\frac{(wl)^{1/3}}{D^{2/3}}$	$\approx \frac{\eta}{k_BT}L^2\frac{(lw)^{2/3}}{D^{1/3}}$
Extended de Gennes	$\sim k_BT(\frac{R^2}{lL} + \frac{L^2w}{RD^2})$	$\sim L\frac{(wl)^{1/3}}{D^{2/3}}$	$\approx \frac{\eta}{k_BT}L^2\frac{l(lw)^{1/3}}{D^{2/3}}$
Odijk	$\sim k_BT\frac{L}{P^{1/3}D^{2/3}}$	$\sim L(1 - A(\frac{D}{P})^{2/3})$	$\approx \frac{2\pi\eta}{k_BTP}\frac{D^2}{ln(D/w)L^2}$

2.4 DNA as a polymer

The classic Crick and Watson picture of DNA is of two helices, connected by hydrogen bonding between base-pairs, wound about the same axis. The helices have a pitch of 36° and advance in the axial direction 0.34nm per base pair.[52] This arrangement, referred to as β-DNA, is essentially correct as a description of the way DNA is structured at the finest scale in the nucleus. In eukaryotes the β-DNA structure is one of a hierarchy of structures that DNA forms over many length scales.

β-DNA wraps around proteins called histones to form complexes called nucleosomes $\approx$ 10nm in size, which bundle into fibers $\approx$ 30nm in size, which are themselves folded by scaffolding proteins to a highly condensed complex called chromatin during anaphase.[53]. The experiments in this work made use of yeast DNA that was previously treated to break down higher order structure and leave purified β-DNA.

Although β-DNA has a helical fine structure, for a large number of base-pairs a DNA molecule obeys the same theories of polymer physics that apply to linear inorganic polymers such as polystyrene. Optical tweezers experiments pulling single DNA molecules have measured a force-extension relationship that is well described by the semiflexible chain model with a persistence length of $P = 53nm$.[8, 25] The bare width of β-DNA in the Crick-Watson model is 2nm, but the effective width w of a DNA molecule in solution is greater due to electrostatic interactions. A theory due to Stigter predicts w for different salt concentrations.[42] Stigter's model gives an estimate of $w \approx 10nm$ for the salt concentration of 10mM used in experiments in this work.

Informed by the physical parameters of DNA given above, the scaling laws summarized in section 2.3 can be used to predict the behaviour of DNA in a real nanofluidic device. To estimate the extension R and relaxation time τ of DNA molecules confined to nanochannels in our denaturation experiments we use the following parameters: confinement dimension $D = 150$ (since the nanochannels are between 120nm and 150nm in width in our devices); segment length $l = P = 53nm$ (from [8], considering DNA to be both self-avoiding and semiflexible); effective width $w = 10nm$ (using the prediction of [42] for our salt concentration). We'll

consider a DNA molecule of contour length $L = 5600nm$ which corresponds to 166kbp, the mean size observed in our denaturation mapping experiments (assuming each base pair contributes 0.34nm to the contour: see chapter 6 for details on the length of molecules observed). We'll use physical constants of viscosity of buffer $\eta = 1mPa \bullet s = 10^{-9} \frac{pN \bullet s}{(nm)^2}$ (assuming buffer has a viscosity close to that of water at 25°C), and $k_B T = 4.11pN \bullet nm$ (also assuming a temperature of 25°C).

The DNA molecule described by these parameters is in the extended de Gennes regime in our nanochannels. Recalling that the Odijk regime begins below $D < P$ and the extended de Gennes below D_c, we find D lies comfortably within the extended de Gennes range.

$$P = 53nm < D \approx 150nm < D_c \approx l^2/w \approx 280nm. \tag{2.30}$$

Applying the extended de Gennes relation for R, we have the estimate:

$$R_{ex.deGennes} \sim L\frac{(wl)^{1/3}}{D^{2/3}} \approx 16\,\mu\text{m}. \tag{2.31}$$

Since the scaling relation is not exact, this should be viewed as an order of magnitude estimate of extension. It is an underestimate because YOYO-1 is known to increase the persistence length P of a DNA molecule.[34] At the very least, this result confirms that the DNA molecule is extended sufficiently that a fluorescence microscope limited by diffraction to a resolution $\sim$ 200nm can resolve features along the length of the molecule.

Applying the exteded de Gennes relation for τ, we have the estimate:

$$\tau^{ex.deGennes} \approx \frac{\eta}{k_B T} L^2 \frac{l(lw)^{1/3}}{D^{2/3}} \approx 0.218s. \tag{2.32}$$

Once more this result should be considered approximate. However it does indicate that the time of an exposure during our experiments (0.100s) is probably shorter than the relaxation time of the DNA molecule. This means that many exposures are necessary to detect an equilibrium conformation averaged over thermal fluctuations. The fact that our duration of measurement (5s for 50 frames) is an order of magnitude larger than τ confirms that we observe DNA molecules over a sufficiently long time that an average of measurements can approximate an equilibrium conformation well.

CHAPTER 3
Physics of DNA melting

In eukaryotes genomic DNA is stored in its double stranded state (dsDNA). However at higher temperatures dsDNA will become unstable and undergo a transition to single-stranded DNA (ssDNA) which exists as a random coil. This transition is referred to as thermal denaturation or melting. The coils of ssDNA have greater conformational freedom and thus greater entropy than the dsDNA helix, and so DNA melting is an order-disorder phase transition.[31]

Chemical agents can also induce a melting transition in DNA by disrupting the hydrogen bonds between base-pairs. Formamide (CH_3NO) is a common tool for encouraging DNA melting in molecular assays. Melting experiments for several different genomes give the empirical relationship that melting temperature is lowered by 0.62°C for each percent of formamide by volume in solution.[37]

Denaturation mapping relies on the fact that DNA melting is sequence-dependent. Most generally speaking, GC-rich regions are more stable than AT-rich regions. For non-uniform sequences like the ones found in nature, DNA melting occurs both heterogeneously and in a co-operative fashion. [16] Computing the spatial pattern and probability with which a DNA strand will melt as a function of temperature and sequence is a therefore a subtle problem. The solution is valuable to inform basic tasks of molecular biology, such as designing primers for PCR reactions.[24]

In 1966 Poland and Scheraga captured the essential features of DNA melting using a one-dimensional Ising model. In their model each base-pair exists in either a melted (dsDNA) or unmelted (ssDNA) state and is subject to nearest-neighbour interactions. While 1D Ising chains do not undergo phase transitions, Poland and Scheraga argued that such a chain could in the presence of long-range interactions.[31] Poland later extended this model to a recursive algorithm to compute the probability of melting of each base-pair for a given DNA sequence.[30] He accounted for the main forces that govern the stability of a dsDNA helix as follows: hydrogen bonding between the nucleotides gives the stability of a single base pair, $\pi - \pi$ stacking bonds between adjacent base-pairs give nearest-neighbour interactions, and the increased entropy of coiled ssDNA regions gives long-range interactions. These thermodynamic quantities must be measured in order to model the melting of a real DNA molecule using the Poland-Scheraga model.

3.1 Empirical measurements: enthalpy of base-pair duplexes

The close-range forces of Poland's model, hydrogen bonding and stacking interactions, both contribute to the stability of neighbouring base-pairs. The combination of four nucleotides that composes two neighbouring base-pairs is referred to as a *duplex*. If the neighbouring nucleotides N_1 and N_2 on a DNA strand are denoted $N_1 \bullet N_2$, and the two complimentary DNA strands i and j are denoted $\frac{i}{j}$, where the strand i is written 5' to 3' and the strand j is written 3' to 5', the ten possible duplexes are as follows: $\frac{T \bullet A}{A \bullet T}$, $\frac{A \bullet T}{A \bullet T}$, $\frac{A \bullet T}{T \bullet A}$, $\frac{G \bullet C}{A \bullet T}$, $\frac{A \bullet T}{G \bullet C}$, $\frac{G \bullet C}{A \bullet T}$, $\frac{C \bullet G}{G \bullet C}$, $\frac{G \bullet C}{G \bullet C}$, $\frac{A \bullet T}{C \bullet G}$ and $\frac{G \bullet C}{C \bullet G}$.

The duplex formed by pairing neighbouring nucelotides i to complimentary neighbouring nucleotides j has a unique enthalpy ΔH_{ij} and melting temperature

T_{ij}. Measuring these quantities provides physical constants for the hydrogen bonding and stacking interactions in Poland's algorithm. Specifically, the statistical weight given to the melting probability of a duplex in Poland's model has the form $\exp^{-\Delta G_{ij}} = \exp^{-\frac{\Delta H_{ij} - T_{ij}\Delta S_{ij}}{RT}}$.

Blake and Delcourt obtained measurements of ΔH_{ij} and T_{ij} for all 10 duplexes by applying a clever statistical approach to conventional melting experiments. [3] DNA melting in bulk has been analyzed extensively using UV absorption measurements. Because light with a wavelength of 270nm is absorbed by ssDNA more greatly than by dsDNA, the absorbance of a DNA sample at 270nm A_{270} is a measure of its degree of meltedness. The trace of the differential dA_{270}/dT with respect to T is called a *melting curve*. The melting curve is unique for each DNA molecule and can be used to measure thermodynamic quantities of the melting process. [16] Blake and Delcourt recorded melting curves for DNA from a collection of plasmids with inserts of tandemly repeating sequences. They chose these sequences to be pseudorandom and AT-rich, so that they melted a temperature lower than the rest of the plasmid and with high co-operatively, giving a sharp helix-coil transition that could be well approximated by a two-state equilibrium. The plasmids were linearized and melted. By adjusting melting curves from plasmids with an insert by subtracting curves from plasmids without the insert, the features of the melting curve pertaining to the inserted sequences were isolated.

The resulting melting curves of the insert sequences were fit to a form of the van't Hoff equation for a two-state equilibrium to measure the total enthalpy ΔH_{total} of the insert sequences' helix-coil transition:

$$\frac{dA_{270}}{dT} = \frac{\Delta A_{270}\Delta H_{total}/4RT^2}{cosh^2[(\Delta H_{total}/2RTT_m)(T - T_m)]} \tag{3.1}$$

where T_m is the melting temperature of the insert, measured as the temperature at which the local maximum of dA_{270}/dT occurs. Blake had applied this method previously to quantify the destabilizing effects of formamide. [2] From results derived from the fits of van't Hoff equation to melting curves, an array of equations was assembled relating ΔH_{total} for each insert as a sum of ΔH_{ij} for all the duplexes composing the insert sequence. The set of equations was solved by regression analysis to estimate the 10 duplex enthalpies ΔH_{ij}. As well the authors measured the melting temperatures of each duplex T_{ij} by plotting T_m from many inserts against f_{ij}, the fraction of a given duplex in the sequence of that insert, and extrapolating a fit line to $f_{ij} = 1$. Their results at 0.745M NaCl are given below.

Duplex $\frac{i(5'-3')}{j(3'-5')}$	$\frac{T\bullet A}{A\bullet T}$	$\frac{A\bullet T}{A\bullet T}$	$\frac{A\bullet T}{T\bullet A}$	$\frac{G\bullet C}{A\bullet T}$	$\frac{A\bullet T}{G\bullet C}$	$\frac{G\bullet C}{A\bullet T}$	$\frac{C\bullet G}{G\bullet C}$	$\frac{G\bullet C}{G\bullet C}$	$\frac{A\bullet T}{C\bullet G}$	$\frac{G\bullet C}{C\bullet G}$
$\Delta H_{ij}(\frac{cal}{mol})$	7.81	8.45	8.50	8.51	9.10	9.47	9.53	10.34	10.51	11.91
$T_{ij}(^\circ C)$	54.58	66.77	67.75	67.95	78.98	85.98	87.23	102.50	105.71	132.22
ΔS_{ij}	23.83	24.86	24.94	24.96	25.83	26.36	26.45	27.52	27.73	29.37

These values confirm the observation that GC-rich regions are more thermally stable than AT-rich regions. Performing the regression approach on a large number of pseudorandom inserts allowed the authors to measure these transition enthalpies very accurately, with an experimental uncertainty of ±4%.

3.2 Empirical measurements: looping entropy

The long-range interactions in Poland's model are due to the increased entropy of melted regions of DNA. Because ssDNA has a much smaller persistence length than dsDNA, contour that melts to the coiled state has greater configurational freedom.[6] However the entropic incentive to melting differs between loops, which form between two unmelted regions, and free ends, which occur at the end of a molecule. The entropy of a loop is lower than a free end because its configurations are constrained at two ends rather than one.

In Poland's algorithm the entropic incentive of forming a loop is determined by counting the number configurations available to a closed random walk of the same contour length as the loop.[30] The number of configurations L_l of a loop of contour length l (for long loops where $l \to \infty$) obeys the scaling relation:

$$L_l \sim \sigma \frac{\mu^l}{l^c}.$$

In the Poland algorithm this factor $\sigma \frac{\mu^l}{l^c}$ is used as the statistical weight of melting probability given to base-pairs participating in a ssDNA loop. μ is a geometric factor related to the size of the bubble and stiffness of ssDNA, and σ and c are called the *co-operativity* and *loop exposure* parameters respectively. The prefactor σ is used to compare the entropy of a loop to dsDNA in order to quantify the absolute probability of a loop opening. It reflects the tendency of a loop to form co-operatively. A value of $\sigma \ll 1$ means that melting to form a loop is less likely because a large degree of co-operativity is required, whereas $\sigma \approx 1$ means that loop formation is favourable because small loops open more easily. The scaling exponent c is a universal quantity

related to the geometry of a loop. The scaling of entropy with l^{-c} does not apply to free ends, whose entropy is given by an open random walk. The parameters σ and c must be experimentally determined in order for the Poland model to correctly account for long-range entropic interactions. The scaling exponent c provides a weighting factor for the stability constant that distinguishes a loop and a free end.

Considering a loop as a closed random walk leads to a numerical estimate for the exposure factor $c \approx 1.75$. In their 1998 work measuring the transition enthalpies ΔH_{ij}, Blake and Delcourt used the value of $c = 1.75$ to measure $\sigma = 1.26 \times 10^{-5}$. [3]. However, it has been suggested that a closed random walk underestimates c in a DNA strand.[4] While a closed random walk has entropy equivalent to an *isolated loop*, the number of configurations available to a loop embedded between dsDNA segments is limited by the presence of the rest of the dsDNA chain. Fortunately the scaling of $L_l \sim l^{-c}$ continues to hold in this case, but requires a higher value of c. Both analytical and Monte Carlo approaches that take into account interaction of the loop with the dsDNA chain give $c \approx 2.15$.[21][18].

Blossey and Carlon analyzed the melting data of Blake and Delcourt using the corrected exposure factor of $c = 2.15$. They measured a new co-operativity factor of, $\sigma = 1.75 \times 10^{-4}$, an order of magnitude higher than measured previously. They showed that the Poland algorithm predicted melting curves that agreed well with the Blake and Delcourt data for the adjusted values of c and σ for loops. They confirmed these values were realistic by showing that the difference in melting temperature between loops and free ends ΔT_m scaled with contour length l in a manner consistent with the choice of $c = 2.15$ for many different insert sequences. [4]

3.3 Computational modeling of DNA melting

Once informed by experimentally measured parameters governing the stability of a DNA molecule; the duplex melting quantities ΔH_{ij} and T_{ij}, and co-operativity factor σ and loop exposure exponent c; the Poland algorithm provides a way to employ the Poland-Scherega model to make predictions of the melting of a DNA molecule. As well as accurately predicting *melting curves* of $\frac{dA_{270}}{dT}$ vs. temperature, as demonstrated by Blossey and Carlon, Poland's method can make a more detailed prediction of the spatial probability of melting: the probability of melting as a function of sequence position in base-pairs (sometimes referred to as a *denaturation profile*). [30] It should be noted that Poland-Scherega model, and consequently any algorithm that makes predictions from it, does not account for the effects of confinement on a denaturation profile, or the effect of a bound fluorescent stain.

In 2003 Tøstesen *et al.* wrote an improved algorithm to implement the Poland-Scherega model. As well as increasing the accuracy of melting predictions, the Tøstesen algorithm reduced the computing time required to analyze a genome of N base pairs from $O(N^2)$ to $O(N)$, making it more feasible to predict the melting of entire genome sequences.[45] This algorithm was made publicly available as a tool for computationally modeling DNA melting on the web server Stitchprofiles.uio.no in 2005.[46] This server uses Blake and Delcourt's duplex enthalpies and Blossey and Carlon's loop entropy weights as experimental inputs. It allows the user to input a DNA sequence and can return melting curves, denaturation profiles, and also *stitch profiles* which describe the alternate spatial patterns of melting available to a DNA

molecule. For the work composing this thesis Tøstesen generously provided the authors with software package in Python that makes use of the same algorithms as Stitchprofiles.uio.no, called Bubblyhelix, which can executed locally on arbitrarily long DNA sequences.

CHAPTER 4
Experimental methods

4.1 The nanofluidic device

The nanofluidic device for denaturation mapping in this work is a refinement of the design used in the first denaturation mapping experiments by Reisner *et al.*[33] Figure 4–1 is a schematic of the device design. At the centre of the device is an array of nanochannels where DNA is confined and adopts a linear conformation. The nanochannels are of width and depth 120-150nm, and length 200 µm, spaced 2 µm apart in the array. While confined in these channels DNA is partially melted, by using an appropriate ambient temperature and buffer, to form series of double-stranded segments (depicted schematically as red) and single-stranded bubbles (brown). DNA is brought into the nanochannels from larger loading microchannels. The microchannels are 50 µm wide and, in two different design iterations either 1 µm or 1 µm deep. DNA is made to enter the nanochannels from microchannels by applying a pressure gradient to overcome the entropic barrier to confinement. The microchannels terminate in macro-scale loading ports of radius $\sim$ 1mm. DNA is loaded into the nanofluidic system by micropipetting into these ports.

4.2 Device fabrication

The nanofluidic devices are fabricated in fused silica in a cleanroom fabrication facility. The devices used in this work were fabricated in three facilities: Danchip at DTU-Nano in Denmark, Nanotools at McGill University and the Laboratory of

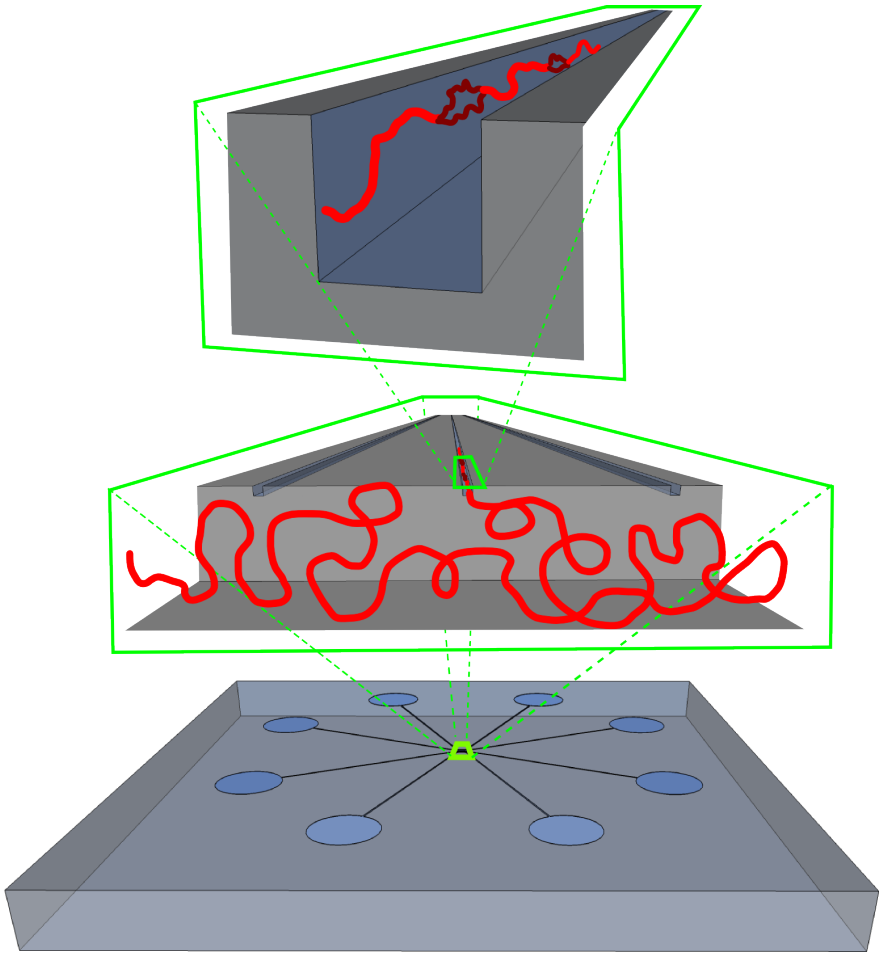

Figure 4–1: Schematic of the nanofluidic device for denaturation mapping

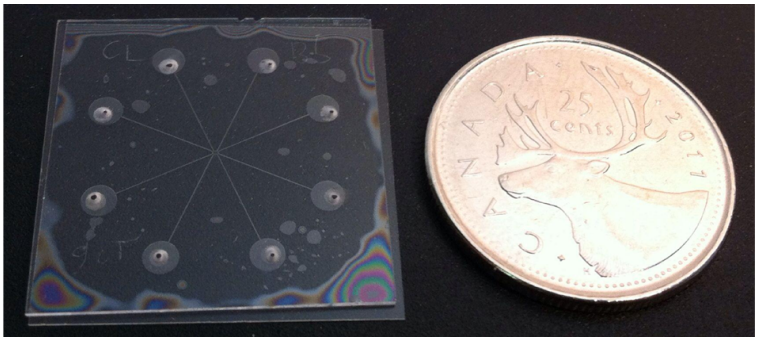

Figure 4–2: A fused silica nanofluidic device

Micro and Nanofabrication (LMN) at INRS Varennes in Montreal. Many melting barcodes of yeast DNA were obtained using devices fabricated by Walter Reisner at Danchip. Melting barcodes were also obtained using devices made by the author using a similar process flow at LMN and Nanotools. 4–2 is a photograph of a device fabricated in Montreal.

In addition to fused silica devices, the author fabricated a test batch of plastic devices using a preliminary process flow based on injection molding at DTU-Nano.

4.2.1 Fused silica devices

Fused silica is a common substrate for nanofluidic devices because is well-suited to the manipulation and fluorescent imaging of DNA. It is hydrophilic, transparent, has low autofluorescence, can be patterned using standard microfabrication tools, and becomes negatively charged in typical buffer conditions, which discourages DNA from sticking electrostatically.[22].

The process flow to fabricate nanochannel devices in fused silica was only slightly modified from the one reported by Reisner.[33] A brief summary is given here. The features of the device were fabricated in multiple steps. Firstly, the nanochannel

array was defined in ZEP520A resist using electron-beam lithography (JEOL) and etched into the silica using CF_4:CHF_3 reactive ion etching (RIE). Secondly, the loading microchannels were defined using contact UV photolithography and etched into the silica using RIE, once more with a CF_4:CHF_3 chemistry. Next the ports were sandblasted manually through the device at the termini of the microchannels, using wax to protect the features at the interior of the device. Finally, the channels were sealed by direct silica-silica bonding on one side to a 150 µm thick fused silica cover glass (Valley Design).

Some devices used in this work made by Walter Reisner were fabricated with an additional feature, a nanoslit, as described in the process flow of [33]. The nanoslit is a shallow depression etched across the middle of the nanochannel array that provides a space above the channels through which buffer can flow but DNA cannot escape. In these devices the slit is 350 µm long in the direction perpendicular to the channels and 50 µm in the direction parallel, and etched to a depth of 30nm. The slit is fabricated in a separate UV contact lithography step followed by CF_4:CHF_3 reactive ion etching before the loading microchannels are patterned. The use of the slit to exchange buffer in the nanochannels was not needed for the experiments reported here. For the purposes of this work the presence of a slit was mostly unimportant: its effect was only to reduce the extension of DNA molecules slightly by giving a larger dimension of confinement in one direction.

4.2.2 Preliminary work: injection molded devices

Although fused silica an established substrate for nanofluidic devices it has drawbacks in practice. Fused silica devices are expensive. Including materials, tool allocation time and technician labour, the devices used here cost more than $100CDN each to make. They are also labour intensive to make and use: each wafer takes days to make in multiple steps, and in order to be used in an experiment they must be mounted on a chuck that is custom-built to the device design. As well they are fragile. They can be broken easily by handling both during fabrication and use. Their costliness and fragility makes fused silica devices poorly suited for prototyping, and puts their clinical use in question.

Nanofluidic devices have been fabricated using a plastic substrate that overcomes these limitations. Utko *et al.* at DTU-Nano patterned nanochannels of dimension $\sim 150\,\text{nm}$ in the polymer substrate Topas 5013 (Topas Advanced Polymers, Inc) using injection molding.[48] The injection molding process can produce many devices quickly from a single nanofabricated mold. The injection molded devices cost only dollars each to produce, which makes it feasible to treat them as disposable. Compared to fused silica the plastic devices are extremely durable. It is virtually impossible to break them by hand. Utko's plastic devices are also easier to employ in an experiment because they are fabricated with on-device luer lock ports that interface loading ports to pressure lines. Therefore the device does not need to be mounted on a chuck in order to control DNA with applied pressure. A plastic device can be loaded and set up for a simple experiment, ready to record fluorescence images of DNA, in less than 5 minutes (see figure 4–3).

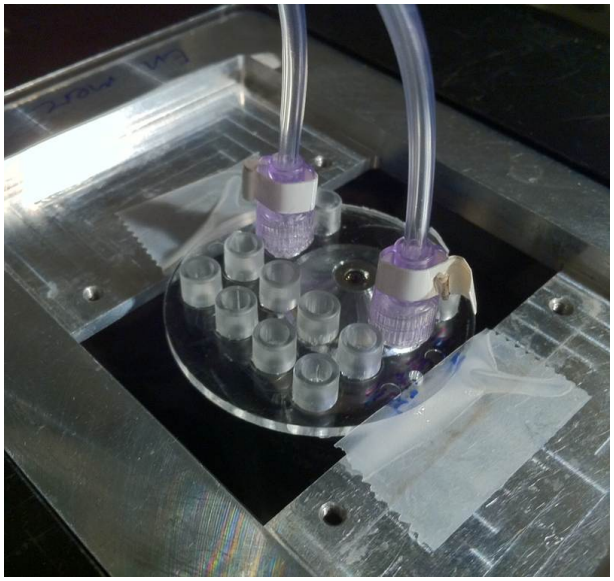

Figure 4–3: An injection molded nanofluidic device

A test batch of plastic devices for denaturation mapping was fabricated during an exchange at DTU-Nano in Danchip, with the help of Peter Friis Østergaard. 90 plastic devices in total were made from one master mold. 4–3 is a photograph of an injection molded nanochannel device set up for a DNA confinement experiment.

Fabrication

Plastic devices were fabricated with minor changes from the protocol published by Utko.[48] The process flow occurs in three stages: patterning a silicon master, producing a nickel shim from the silicon master, and making injection molded devices using the shim as a mold. Hundreds of injection molded devices can be made from one shim and master. The design of the device used here was similar to that of the fused silica devices presented in this chapter: an array of nanochannels connected to larger loading microchannels, which terminate in loading ports.

The silicon master was patterned in 6 inch n-type <100> single side polished silicon wafers. The micro- and nano-fluidic features present in the final device were patterned in two steps into the silicon master. First the nanochannel array was formed. A layer of $\sim$ 100 nm SiO_2 was deposited using oxide deposition. The nanochannels were patterned in ZEP520A resist using electron beam lithography (JEOL) and etched using RIE with a CF_4:CHF_3 chemistry. The microchannels and loading ports were patterned using contact UV lithography and etched once more using CF_4:CHF_3 RIE. Once the device features were patterned into the silicon master, a $\approx$ 130 nm layer of Ni/V was sputtered on the master to facilitate removing the shim, and then the shim was formed by electroplating a $\approx$ 300 µm layer of nickel onto the Ni/V layer. The shim and the silicon master were separated by hand. The plastic devices were formed by injection molding Topas 5013 polymer into a mold formed by the shim and a back-plate containing molds for luer locks in positions that match the location of microfluidic loading ports on the device. Finally, the devices were sealed by bonding a 100 µm Topas 5013 sheet. Just before bonding the inside surfaces of the sheet and device were exposed to UV light. Bonding was done at 10kN pressure at 120°C for 5 minutes. These bonding parameters were optimized from Utko's to discourage unbonding of the Topas sheet while applying high pressure to the luer locks during experiments. Devices made using these parameters were able to withstand pressures up to 1 bar before unbonding.

4–4 is an SEM micrograph of a plastic device before bonding. The loading microchannels were patterned with good fidelity in the device. The nanochannels were inspected at different places along the array and found to be unobstructed and

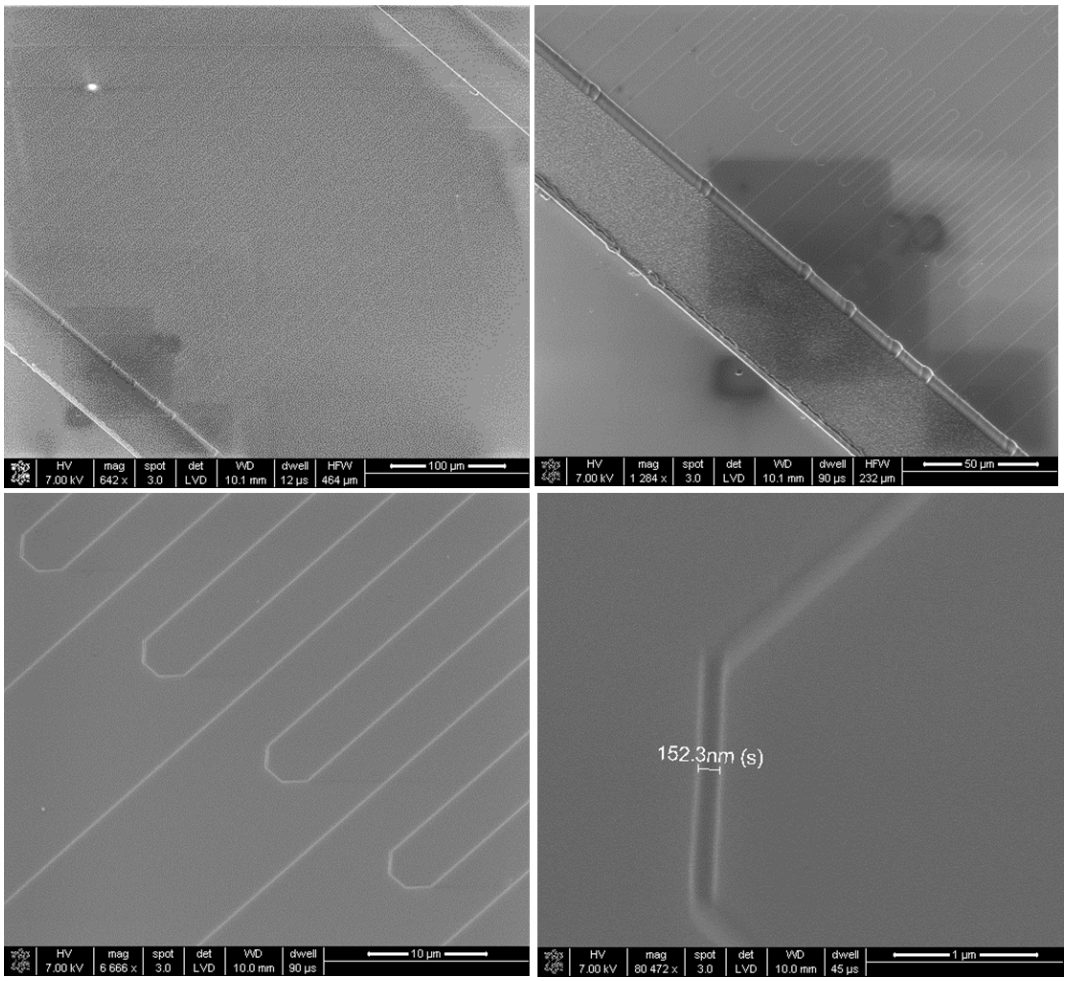

Figure 4–4: SEM micrographs of an injection molded device

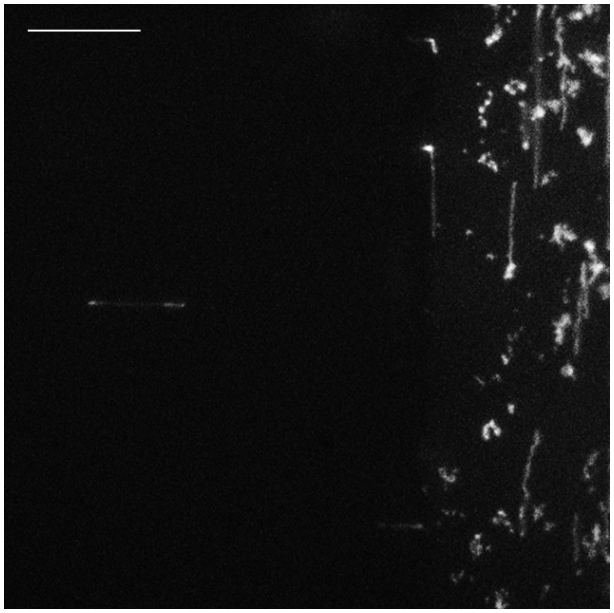

Figure 4–5: DNA confined in an injection molded device

of diameter close very close to the nanochannel features in the silicon mold (152 nm compared to 150 nm.

Testing

A simple DNA confinement experiment was performed to test the plastic devices. DNA from the λ -phage virus was dissolved in 1xTBE and stained with YOYO-1 outside the device. Although the plastic devices are easy to load, the fact that Topas 5013 is slightly hydrophobic requires a short step not necessary to load fused silica devices. Before loading the DNA solution the device is wet with ethanol. Once the microfluidic components are wet the DNA solution is loaded in the reservoirs, and then the DNA solution is flushed through the system to replace the ethanol. Some pressure must be applied to introduce ethanol (of the order $\sim$100mbar), but

the pressure required to introduce DNA solution into a dry device is more than the device can withstand (>1 bar) before unbonding.

DNA was successfully introduced into the loading microchannels. Most nanochannels could be wet with buffer, although a few could not be and remained dry. There were problems loading the DNA into nanochannels from the microchannels. At pressures typical for loading nanochannels in fused silica devices (∼100mbar), DNA would not enter the nanochannels. DNA was only introduced into the nanochannels after applying pressure so great that the device unbonded. Figure 4–5 is a fluorescence image of a single DNA molecule confined to a nanochannel while many are stuck outside, seconds before the device unbonded and failed (scale bar 10 µm).

Closer inspection by SEM revealed why DNA could not be loaded. 4–6 is a collection of SEM images of the entrance to different nanochannels from loading microchannels. The images show that the entrance to every nanochannel in the array was at least partially blocked. Some were completely blocked. The most open entrance observed had a diameter of 30 nm. This degree of blockage is consistent with the fact that some channels could be wet by buffer but that confining DNA was possible under only very great pressure. The rough features that block the nanochannel entrances were not observed SEM images of the silicon master or shim.

The nanochannels appear to be blocked by debris from the wall of the loading microchannel. We suggest that the anisotropic nature of the RIE etch to form microchannels in the silicon master causes walls of the loading microchannel to be slightly angled. These angled features could drag hot polymer over the walls of the microchannel during the injection molding process, as the shim is removed from the

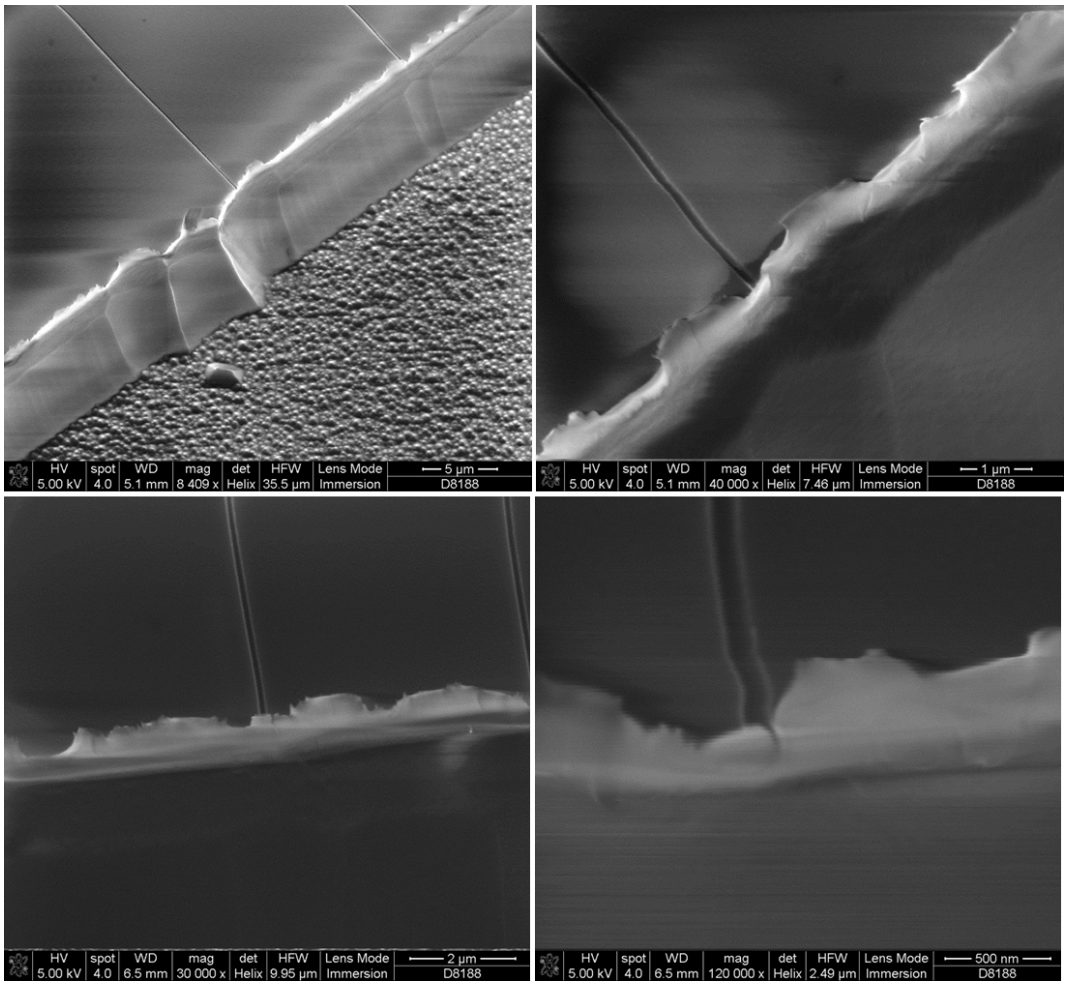

Figure 4–6: SEM micrographs of nanochannel entrances in an injection molded device

plastic device. The process of dragging hot polymer is consistent with the fact that the blockage is present in the injection molded device but not the shim, and is consistent with the irregular shape of the blockage. The problem of blocked nanochannels could be addressed by changing etching methods to produce walls that are angled differently, or by etching shallower microchannels so the effect of the wall angle is less dramatic.

4.3 DNA preparation

4.3.1 SCODA

DNA for denaturation experiments was obtained from whole chromosomes of the yeast strain YPH80 embedded in an agarose gel plug (chromosome sizes 225-1900kbp, New England BioLabs). The chromosomes were extracted to buffer using a recent Canadian-developed techology called SCODA (synchronous coefficient of drag alteration). SCODA is a pulsed-field electrophoresis technique that exploits nonlinearity in the dependence of mobility on electric field in order to selectively migrate DNA molecules to the centre of a centimeter-scale agarose gel.[26] SCODA has been demonstrated to recover very long intact DNA molecules DNA (100kbp+) with high purity.[14] Impressively, SCODA has been used to extract DNA from lysate mixed with whole blood and tar sands, and to recover DNA from starting quantities as few as 172 molecules.[7]

We concentrated DNA from gel plugs to buffer using a development unit supplied by Boreal Genomics. The SCODA gel was made with 0.1g 1% LMP agarose dissolved in 10mL 0.25XTBE. YPH80 gel plugs were embedded directly in the SCODA gel by cutting slices of the plug and pouring molten agarose around them in the SCODA gel

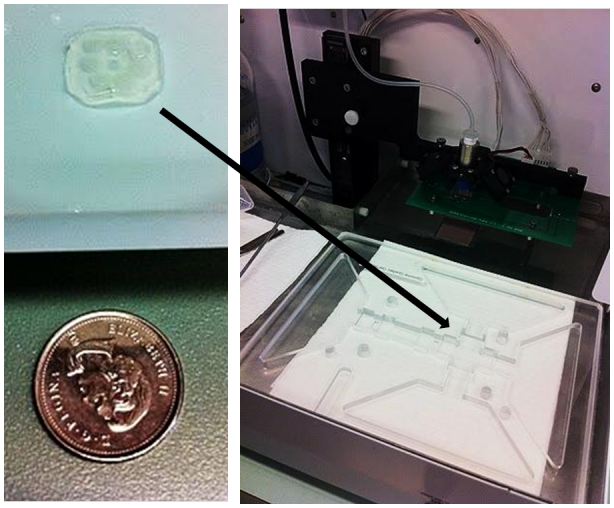

Figure 4–7: A SCODA gel with embedded gel plugs

mold. DNA was collected in 1XTBE a 50 µm reservoir in the centre of the gel. Figure 4–7 is a photograph of a SCODA gel with embedded gel plugs and a reservoir hole in the centre (left) and the SCODA gel boat where electrophoresis occurs (right). We used a custom SCODA protocol optimized to extract very long DNA molecules, with runs of duration 20 hours, voltage 12%, cycle length 2880s, and no bias voltage. The help of Said Attiya was invaluble in developing the SCODA protocol.

SCODA allowed us to obtain DNA of high molecular weight in solution. Figure 4–8 is a photograph of a polyacrylamide gel, after electrophoresis, of yeast DNA obtained from SCODA (right lane) and a size ladder (left lane). The largest fragments on the size ladder are of length 1000kbp. The presence of a band in the right lane higher than the top band on the left lane indicates that all the fragments obtained from SCODA are much longer than 1000kbp.

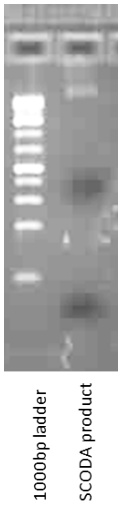

Figure 4–8: Polyacrylamide gel electrophoresis of SCODA products

The chief disadvantage of the SCODA protocol is the low quantity of its DNA product. Using a Quant-iT PicoGreen quantification assay (Invitrogen) we measured 11 different 20-hour SCODA runs to produce an average of 0.86 µg of DNA at a concentration of 14.2 µg/mL. 14 20-hour SCODA runs were necessary to produce sufficient DNA for a Roche 454 sequencing assay.

4.3.2 Buffer chemistry

DNA was prepared for denaturation mapping experiments using a running buffer described in ref. [33]. Its contents are listed here briefly. DNA was stained with the intercalating dye YOYO-1 (Invitrogen) at a ratio of 1 dye molecule to 10 base pairs. The stained DNA was added to a running buffer containing 0.05XTBE (4.5mM Tris, 4.5mM boric acid, 0.1mM EDTA), 10mM NaCl, and formamide at a ratio of 50% by volume (Sigma). The running buffer also contained a system to discourage photo-nicking, added before mixing with formamide, composed of β-mercaptoethanol at

3% per volume, as well as a set of oxygen-scavenging reagents including 0.2mg/mL glucose oxidase, 0.04mg/mL catalase and 4mg/mL β-D-glucose.

4.4 Experimental setup

Experiments with the nanofluidic device were carried out using a custom-designed chuck machined in polyetheretherketone (PEEK). The chuck contains pressure lines that connect to loading ports on the device by an o-ring seal and terminate in luer locks. Pneumatic pressure is applied to these lines to control the flow of DNA in the device. The chuck also contains a port that holds a resistive heating element and thermocouple in contact with the device in order to control temperature in the nanochannels. Lastly, the chuck fits to a custom-designed holder that positions the device above the objective of an inverted microscope for imaging. Fluorescence images were recorded using a Nikon Eclipse TE2000 inverted microscope with a 100x N.A. 1.4 immersion objective and electron multiplying CCD camera (Andor iXon).

Figure 4–9 shows how the chuck and device form the experimental setup for denaturation mapping. The top photograph shows the underside of the chuck with the device mounted on the centre of the face. The device is held in place using a square metal bracket fastened by four screws. The heating element in contact with the device is visible by the circle of light gray thermal grease in the centre of the device. The bottom photograph shows a chuck set up during an experiment. This view is of the side opposite the first photograph. The inset shows the objective lens of the fluorescence microscope in contact with the device on the underside of the chuck.

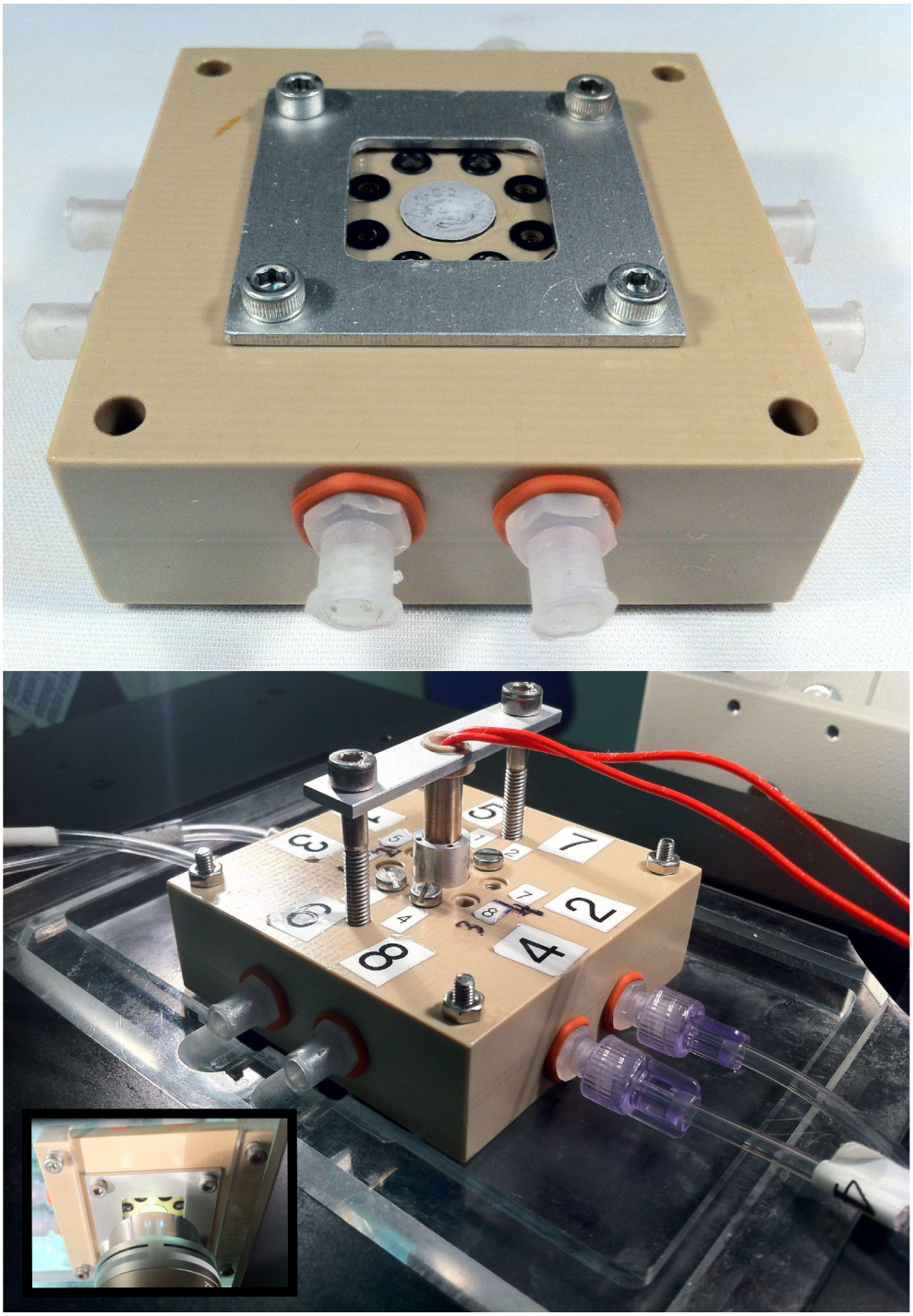

Figure 4–9: Experimental setup: the chuck and device

A brief description of a typical experiment follows. The DNA-containing running buffer is pipetted into the loading ports on the device and allowed to wet the nanofluidic circuit. The device is mounted on the chuck, aligning the device's loading ports with the corresponding pressure lines, and fastened to the chuck using a metal bracket to form a seal around the o-rings. The resistive heating element is mounted on the chuck with thermal grease applied to the tip to create good thermal contact with the device. The chuck, in turn, is mounted on the sample positioning stage of the inverted microscope and the objective lens is brought into oil immersion contact with the device. The device is positioned so that the interface between the nanochannel array and a loading microchannel is in the field of view. The heating element is set to a temperature of 29°C and allowed to reach thermal equilibrium with the device for a few minutes. A pressure differential is applied across the loading microchannel to bring DNA molecules to the nanochannel array, and then a pressure differential is applied across the nanochannel array in order to load DNA molecules into the nanochannels. Finally, a confined DNA molecule is imaged in the nanochannel array and a movie is recorded of 50 images at 10 frames per second. The procedure of loading and imaging DNA molecules is iterated to collect melting barcodes.

Figure 4–10 is a DNA barcode from a typical denaturation mapping experiment (scale bar 10 µm). *(a)* is a fluorescence image of a confined, partially melted DNA molecule dyed with YOYO-1 that exhibits a characteristic barcode pattern of fluorescence intensity. A time series of 100 images is recorded *(b)* and processed to remove distortions due to thermal fluctuation and produce a barcode that represents

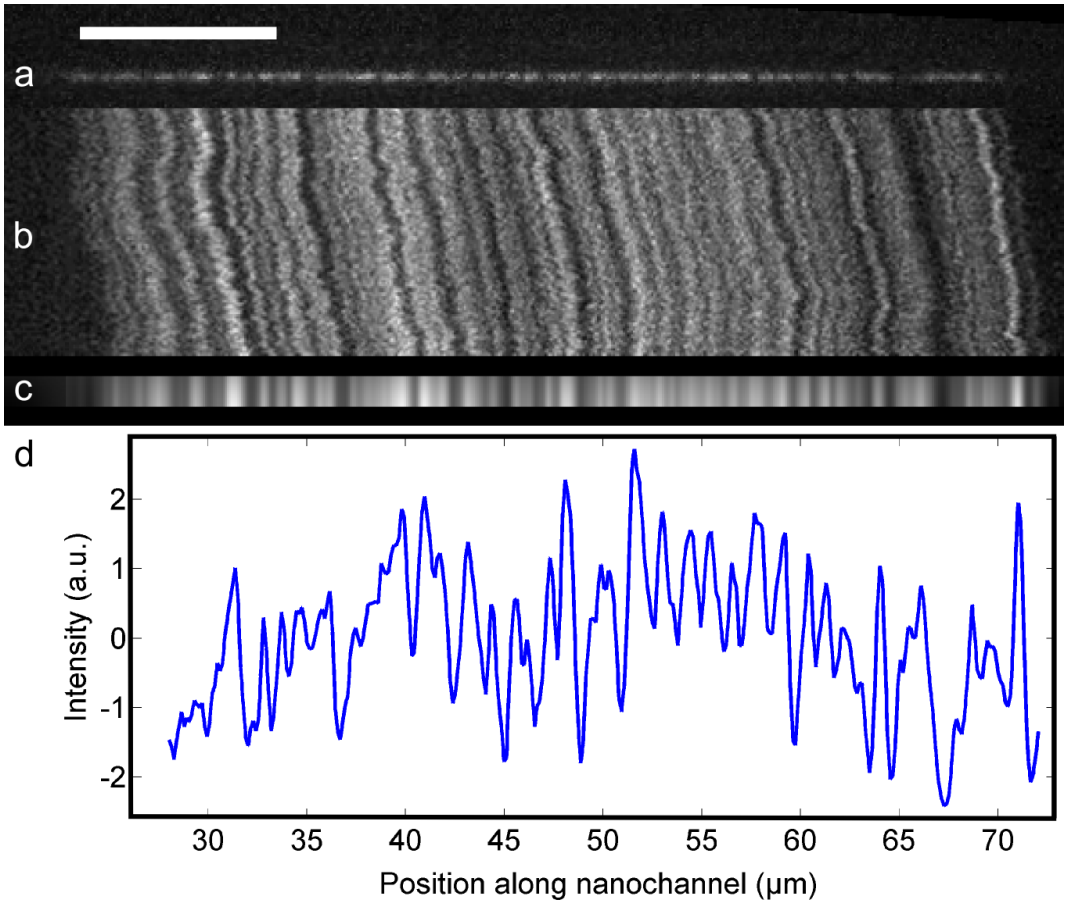

Figure 4–10: A DNA barcode from a denaturation experiment

the molecule's equilibrium conformation *(c) and (d)*. (See Chapter 5 for details on the image processing algorithms.)

4.4.1 A device cleaning protocol

The fact that fused silica devices are expensive and time-intensive to fabricate is a strong incentive to re-use devices for multiple experiments. However, there are a few reasons it is not good practice to simply store and re-use devices after denaturation mapping. If the buffer in the device allowed to dry out, the salts it contains will be deposited inside the device and contaminate the buffer of the next experiment. Because the mechanics of DNA melting are sensitive to the strength of electrostatic interactions, deviations from the intended salt concentration can prevent a melting barcode from forming correctly. Conversely, if the buffer is kept inside the device, the components of buffer intended to discourage photonicking (β-mercaptoethanol and the oxygen-scavenging enzymatic components) will become inactive over the time between experiments and affect the stability of DNA. As well, fragments of DNA left in the device could block nanochannels and affect the conformation of confined DNA. Therefore it would be useful to have a method of cleaning devices between experiments that would allow them to be stored indefinitely and re-used. A simple protocol is given here.

After an experiment the device is loaded with DI and flushed for 20 minutes. This dilutes the salts in the buffer and dislodges loose DNA fragments from the nanochannels. Next the device is baked on a fused silica wafer in an oven at 500°C for 30 minutes. Exposure to high temperature breaks down organic molecules in the device (DNA, dye and enzymes from the oxygen scavenging system) volatilizing some

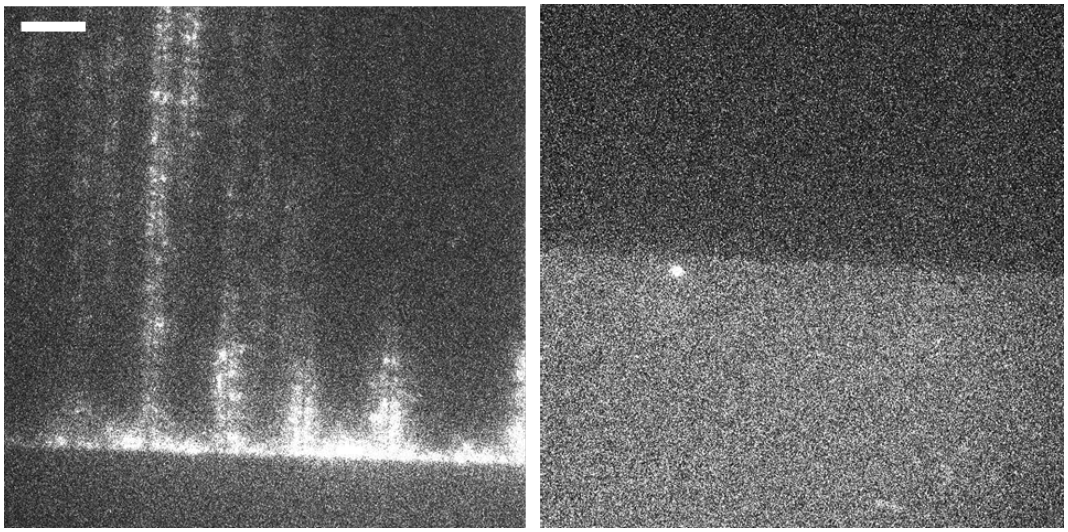

Figure 4–11: Background fluorescence in a device before and after cleaning

of them, and leaves the device dry. This step takes 2 hours including warming and cooling of the oven. Lastly, the device is submerged in an excess volume of ethanol (in a small dish, so that the device is covered to a depth $\approx$ 5mm) and degassed under vacuum until the ethanol completely evaporates. During this step ethanol first enters the nanofluidic system, and then leaves as the covering layer evaporates. This step washes the device once more to remove organic molecules broken down in the baking steps, and leaves the device dry so that it can be stored. Ethanol specifically is useful for washing because it evaporates quickly: this step takes about 2 hours. In total the protocol takes $\approx$2.5 hours.

Figure 4–11 is a fluorescence image of a nanochannel array before (left) and after (right) the cleaning protocol outlined above (scale bar 10 µm). Cleaning dramatically reduced the background fluorescence in the nanochannels.

4.4.2 Preliminary work: an improved chuck

At the time of writing a new chuck is being machined that makes some improvements over the one used for this work. The motivating idea behind its design was to reduce the fragmentation of DNA in the device in order to map longer intact DNA molecules. Currently the step of loading DNA into the device is a major source of fragmentation. Hydrodynamic shear forces from pipetting break Mbp-sized DNA molecules into smaller fragments before they ever reach the nanochannels. A way to avoid pipetting is to prepare DNA on the chuck. We aim to extract DNA from a gel plug by heating the gel plug in a reservoir on the chuck, and then applying an electric field across the reservoir to bring DNA into the device from the melted plug by electrophoresis.

This goal necessitates several new structures. First, to melt the plug we enlarged two loading ports to be large enough to accommodate a gel plug. We envision a simple heating element and electrode, positioned atop the melted plug in the reservoir, to be held in place by a custom-machined screw that fits the enlarged loading ports. Secondly, to deliver DNA to the channels stained in the appropriate buffer there must be structures on the device to mix incoming DNA with a running buffer. We envision an on-device diffusive mixer in a future device design that takes as input DNA from the plug and also buffer from another loading port. In order to service the extra loading ports for the running buffer and stain entering the diffusive mixer the chuck must have extra reservoirs and pressure lines as well. The new device, in total, would have 12 loading ports. The new loading ports are arranged so that the

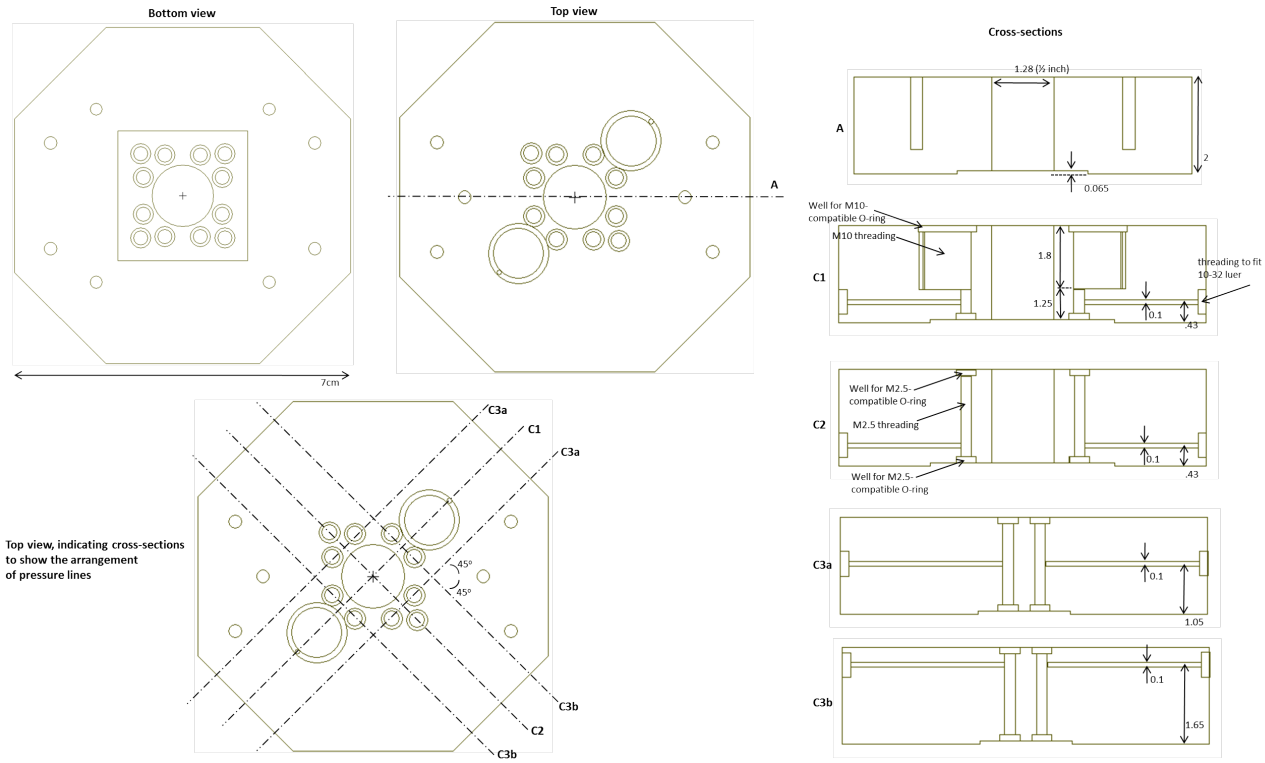

Figure 4–12: Design drawings of an improved chuck

chuck remains backwards compatible: the 8 loading ports of the older device design can still be serviced by pressure lines on the chuck.

Figure 4–12 is a design drawing of the new chuck (all dimensions given in centimetres). It is octagonal to accommodate the extra luer locks required to interface pressure lines to 12 ports. The positioning of 12 non-intersecting pressure lines to the edge of the chuck was a design challenge. The new chuck also has new versions of the accompanying brackets that hold the device and heater in place (not pictured).

CHAPTER 5
Computational methods

We analyzed melting barcodes using a suite of custom MATLAB programs and a handful of other tools. The ultimate goal of the software is to *align* melting barcodes to the genome. Alignment refers to the process of identifying the location of a DNA molecule on the genome by the features of its melting barcode. The process of alignment can be broken into three tasks: processing fluorescence images of a DNA molecule from a denaturation experiment to produce a melting barcode, predicting a theoretical melting barcode for the genome, and comparing the experimentally obtained melting barcode to the predicted melting barcode of the genome in order to identify the location of the molecule.

The programs used in this work and their role in the alignment workflow are summarized below. Unless otherwise noted the programs are written in MATLAB by the author.

Preparing melting barcodes from images *once per experiment*

- Rob_batch_rotate (ImageJ macro)

  Takes .nd2 images of DNA molecules from fluorescence microscopy. Rotates the images so the molecules are vertically oriented and saves them as .TIF images.

- Rob_Batch_tif2barcode_v2

  Takes .TIF images of DNA molecules.

  - *calls* Calculate_Time_Trace *written by Walter Reisner*

    Uses edge detection to find fluorescence traces of DNA molecules in .TIF images. Gives a time series of fluorescence intensity as a function of pixel position along a DNA molecule.

  - *calls* Normalize_Time_Trace *written by Walter Reisner*

    Performs a multiple segment rescale on the time series (see section 5.1). Gives melting barcodes.

- RobScript_trim_exp_barcode

  Removes leading or trailing regions of noise (intensity less than 10% the maximum) from a melting barcode.

Predicting a melting barcode for the genome *once per genome*

- Bubblyhelix 0.9.3 (Perl, C) *provided by Eivind Tøstesen*

  Takes a genomic base-pair sequence. Computes melting probability, as a function of position in base pairs, for a given temperature and salt concentration.

- RobScript_yeast_theory_barcode

  Takes the output of Bubblyhelix. For each yeast chromosome, calls Rob_predict_barcode.

  - Rob_predict_barcode

    Produces a helicity interpolation matrix, composed of predicted melting barcodes for different degrees of meltedness (see section 5.2).

- RobScript_yeast_bp_lookup

  Takes the output of Bubblyhelix. For each yeast chromosome, calls Rob_bp_lookup_matrix.

  - Rob_bp_lookup_matrix

Produces a *base-pair lookup matrix*, which maps position in nanometres to position in base pairs for different degrees of meltedness. These matrices are used later during alignment.

Aligning melting barcodes to the genome *once per experiment*

- Rob_batch_align_barcode

  Takes experimental melting barcodes. For each barcode, calls Rob_align_barcode.

  - Rob_align_barcode

    Compares melting barcodes to the predicted barcode for a given genome. Identifies the sequence position where the DNA molecule is located. Saves plots and results (see section 5.3).

    * *calls* Rob_estimator_calc

      Calculates the least-squares estimator for all the search parameters of the alignment procedure (see section 5.3).

    * *calls* Rob_estimator_analyze

      Analyzes the distribution of estimator values to assess the statistical significance of the alignment result (see section 5.3).

    * *calls* Rob_nm2bp to plot alignment results as a function of position in base pairs, using the output of Rob_bp_lookup_matrix.

- Rob_assemble_results

  Assembles the results of Rob_batch_align_barcode, including the identified sequence position and statistics of the estimator distribution for each melting barcode.

- Rob_alignment_coverage

Takes the output of Rob_assemble_results. For a set of alignment results, and a given criteria of statistical significance for those alignments, calculates the percent coverage of the genome that is optically mapped by melting barcodes.

This remainder of this chapter details the algorithms and strategy for data analysis that is executed in these programs.

5.1 Preparing melting barcodes from images

A DNA molecule in a nanochannel experiences Brownian fluctuations that create time-dependent spatial distortions in its fluorescence profile. We remove these distortions from movies of confined DNA, following the method of [33] as follows. The first frame of the movie is taken as a template. Firstly, centre-of-mass diffusion is removed by globally translating each of the remaining frames to maximize cross-correlation with the template. Secondly, local variations in contour density are also corrected for by locally rescaling the resulting frames in segments. We refer to this process as *mutli-segment rescaling*. Each frame i contains a fluorescence profile $P_i(x_j)$ with respect to position in pixels x_j. We divide the profiles after the template by position into a series of segments. The length of each segment k is then dilated linearly by a factor d_k. The set of dilation factors for each frame is chosen to minimize the least squared difference δ between the dilated profile $P_i'(x_j, d_k)$ and the template $P_1(x_j)$:

$$\delta = \sum_{j=1}^{N} [P_i'(x_j, d_k) - P_1(x_j)]^2. \tag{5.1}$$

This choice of dilation factors represents a map $M(d_k)$ that rescales each frame to match the local variation in contour density of the template. Applying M to every

frame and averaging the resulting profiles over time produces a consensus profile that has the instantaneous conformation of the template. Taking the average map $\langle M(d_k)\rangle$ over all the frames gives the average transformation that produces this conformation. However, because the template is recorded over a time scale less than the relaxation time of a DNA molecule (see section 2.40), its conformation does not represent a thermal equilibrium. By applying the inverse of the map $\langle M(d_k)\rangle^{-1}$ to the consensus profile we arrive finally at a melting barcode that represents the DNA molecule's conformation at thermal equilibrium. We used multi-segment rescaling with 50 local segments to produce a melting barcode from images of each molecule imaged during our experiments.

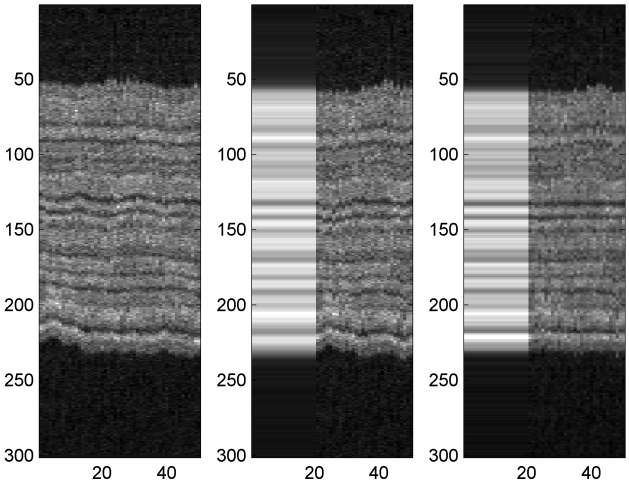

Figure 5–1: Preparing a melting barcode from a time-series of fluorescence images

Figure 5–2 depicts the process of preparing a melting barcode from fluorescence images. The vertical axes are position (in pixels) along the nanochannel and each row of pixels is from a different frame of the time series. From left to right, the

images show: an unadjusted time series from a denaturation experiment, a time series adjusted for centre-of-mass diffusion, and a time series after both adjusting for centre-of-mass diffusion and multi-segment rescaling. The first column in the middle and right images is an average over the entire time series, that has been repeated over several columns for visibility.

5.2 Predicting a melting barcode for the genome

We located DNA molecules on the genome by comparing experimentally obtained melting barcodes to a theoretically predicted barcode of the complete genomic sequence. The genomic barcode was predicted in three steps: by generating a melting probability map from the nucleotide sequence; predicting a barcode from the melting probability map; and then interpolating between intensity maps predicted for a range of temperatures.

The reference yeast genome nucleotide sequence was downloaded from NCBI Genbank (RefSeq numbers NC001133 to NC001148). The melting probability map was calculated from this sequence using version 0.9.3 of the software package bubblyhelix provided by Eivind Tøstesen.[45] Bubblyhelix uses the Poland-Scheraga model to compute the probability of remaining double-stranded p_{ds} as a function of base pair position for a given temperature and salt concentration (see section 3.3 for details). It takes as inputs the respective enthalpies and entropies of the 10 possible base-pair interactions and takes into account the energetics of bubble formation, and distinguishes between a bubble and the melted end of a molecule.[51]

The genomic barcode is predicted from the melting probability map by accounting for two factors: local variation in DNA extension due to partial melting and the

point-spread function of the objective lens (the diffraction limited profile of a point-source imaged with our microscope system). The intensity profile $P(s) \propto p_{ds}(s)$ and extension $r(s)$ as functions of sequence position s (in bp) is determined by:

$$r(s) = r_{ds}p_{ds}(s) + r_{ss} * [1 - p_{ds}(s)]. \tag{5.2}$$

where the value of r_{ds} defines the extension of an unmelted base-pair along the nanochannel, and the ratio r_{ss}/r_{ds} determines the relative extension of a melted base-pair. From observations of λ-phage DNA molecules in similar nanochannel devices we find that values of $r_{ds} = 0.186$ and $r_{ss}/r_{ds} = 0.85$ are reasonable.[33] The series of extensions $r(s)$ of each base-pair forms a position co-ordinate in nanometers $x(s) = \sum_{s'=1}^{s} r(s')$ that describes the position of each base-pair in DNA molecule along the length of the nanochannel. We simulate the effect of diffraction due to the microscope objective by convolving the expected intensity profile $(x(s), P(s))$ with a point-spread function parameterized by a Gaussian of standard deviation $\sigma = 200\,\text{nm}$. The result is a simulated profile of fluorescence intensity with respect to position, or predicted melting barcode, of each of the chromosomes of the yeast genome at a given temperature.

It is difficult to know the precise temperature of a DNA molecule in the nanochannel device. Consequently it is useful to compare experimental melting barcodes not to a single genomic barcode, but to a matrix that interpolates between genomic barcodes predicted for a range of temperatures. We use helicity, measured as the average probability of remaining double-stranded of all the base-pairs of a DNA

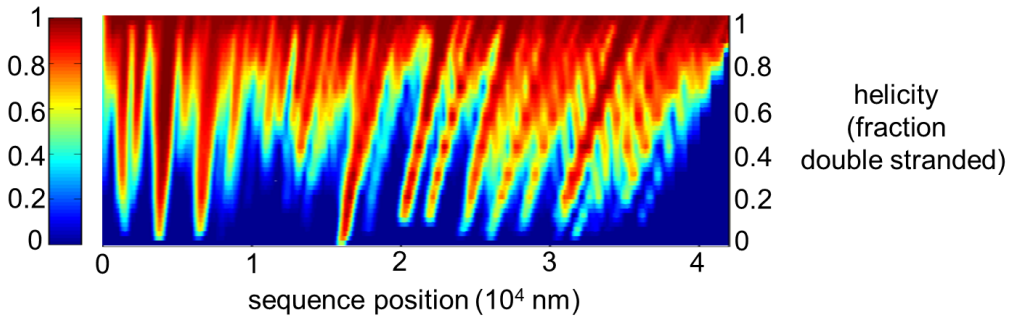

Figure 5–2: The predicted melting barcode of yeast chromosome 1

molecule $h = \langle p_{ds}(s) \rangle$, as the co-ordinate to interpolate between barcodes. Interpolation was performed using the MATLAB function *interp*2 to produce a matrix of form (x, P, h), where for a given helicity co-ordinate h the matrix (x, P) is a barcode. We formed a helicity interpolation matrix for the yeast genome by interpolating 20 genomic barcodes, predicted for temperatures between 50°C and 70°C, over 50 helicities from $h = 0.001$ to 0.990. Figure 5–2 is the predicted helicity interpolation matrix of the yeast chromosome 1 (where the colour axis indicates fluorescence intensity). Note that the chromosome becomes shorter as it melts due to the greater extension of double-stranded segments. Regions of the helicity interpolation matrix that are entirely zero are removed in the alignment step.

5.3 Aligning melting barcodes to the genome

We identified the position of single DNA molecules on the yeast genome by comparing experimentally obtained melting maps to the calculated genomic barcode helicity interpolation matrix. Alignment was performed by finding the genomic position that minimized the least-squared difference between them. We compared an experimental melting barcode P_{exp} to a segment of the same length from position i

on the genomic barcode P_g. From these we subtracted the mean and divided by the local standard deviation to obtain the adjusted barcodes:

$$\delta P_{exp} = \frac{P_{exp} - \langle P_{exp} \rangle}{\langle (P_{exp} - \langle P_{exp} \rangle)^2 \rangle^{1/2}},$$

$$\delta P_g = \frac{P_g - \langle P_g \rangle_{i,N}}{\langle (P_g - \langle P_g \rangle_{i,N})^2 \rangle_{i,N}^{1/2}}$$

where i is the starting position of P_g on the genomic barcode and N is the length of both fragments. Using these definitions, for each position i we define the least-squares estimator:

$$\Delta(i) = \frac{1}{2N} \sum_{j=1}^{N} [\delta P_g(x_{i+j}) - \delta P_{exp}(x_j)]^2.$$

To determine the global minimum of the estimator we perform a search across four parameters. We vary the helicity on the genomic interpolation matrix from which P_g is taken, the sequence position i, and the relative orientation of the experimental barcode (forward or backward). As well we perform a global dilation to the experimental barcode, attempting alignment after elongation by a range of factors between 0.8 and 1.2 in order to allow for some variation in the dimensions of the nanochannels. We perform a coarse search varying these parameters widely, and then perform a finer search to find the global minimum of the estimator.

Criteria for statistical significance of alignment results

We assess the statistical significance of a location result by analyzing the distribution of estimators at different sequence positions i with the other search parameters held constant. Given the form of the adjusted barcodes δP_{exp} and δP_g we can write

the estimator as

$$\Delta(i) = 1 - \frac{1}{N}\sum_{j=1}^{N} \delta P_g(x_{i+j})\delta P_{exp}(x_j).$$

If δP_{exp} and δP_g are independent random variables, the set of values $\frac{1}{N}\sum_{j=1}^{N} \delta P_g(x_{i+j})\delta P_{exp}(x_j)$ for each position i will also be a sequence of random numbers with a mean of zero. It arises from the central limit theorem that the distribution of estimators for all values of i where an experimental barcode cannot be successfully located will be a Gaussian centred about 1. Due to the finite length of the barcodes we add a quartic correction to obtain an expected distribution of estimators

$$P(\Delta) = Ae^{(-\frac{(\Delta-\Delta_0)^2}{2\sigma^2} - \frac{(\Delta-\Delta_0)^4}{24\eta^4})},$$

where the mean value $\Delta_0 = 1$. During a location procedure we fit the distribution of estimators with the function $P(\Delta)$. Integrating $P(\Delta)$ from $\Delta = 0$ to the global minimum Δ_f gives an estimate n_L of the number of locations more significant than the one obtained that would be expected to occur between uncorrelated DNA sequences.

To judge an alignment result as statistically significant we require two conditions: that $n_L \ll 1$, and that the function $P(\Delta)$ used to determine n_L fits the data well. A result of $n_L \ll 1$ occurs when the global minimum Δ_f is much smaller than other values of the estimator. Specifically, $n_L \ll 1$ occurs when Δ_f lies outside the distribution of other estimator values $\Delta(i)$. Δ_f far to the left of the distribution implies that an estimator value as small as Δ_f would not be expected to occur by chance if the melting barcode did not map to a unique location on the genome: the location where Δ_f occurs is a unique alignment result.

However, the value of n_L is only meaningful if the fit of $P(\Delta)$ to the estimator histogram is good. We assess the goodness of fit of $P(\Delta)$ by performing Pearson's chi-squared test on the bins of the estimator histogram between the global minimum Δ_f and $\Delta = 1$. The chi-squared test determines the probability of observing a set of values under the assumption that the data have a distribution given by a null hypothesis. Specifically, it gives a *p-value*, the probability that a *chi-squared test statistic* $\chi^2 = \sum_{j=0}^{N} \frac{(O(j)-E(j))^2}{E(j)}$ (where O is the observed distribution, E is the expected (null) distribution, and the index j denotes the bins of a histogram) will be observed assuming the test statistic obeys the canonical χ^2 probability distribution. We used $P(\Delta)$ as the null hypothesis E to determine the probability the distribution O of observed estimators $\Delta(i)$ could originate from it. A high p-value $p \sim 1$ indicates $\Delta(i)$ follows $P(\Delta)$ exactly, where a low value $p \approx 0$ indicates $\Delta(i)$ does not arise from $P(\Delta)$.

Here we judge a location as successful whose estimator distribution can be fit by $P(\Delta)$ with $n_L < 0.02$ that passes a chi-squared test of $p > 0.01$.

Figure 5–3 is a trace of the estimator $\Delta(i)$ with respect to sequence position for a typical melting barcode (top); and a histogram showing the distribution of estimator values (below). The red circle indicates the global minimum of the estimator Δ_f. In the histogram the blue line is a trace of the fit function $P(\Delta)$, and green circles are the histogram of the estimator trace $\Delta(i)$ above. Note that in the estimator trace the horizontal axis is the index of the alignment search, which corresponds to increments along the nanochannel of 200nm (the point spread function of the objective lens). The fact that Δ_f lies to the left of the distribution of other estimator values indicates

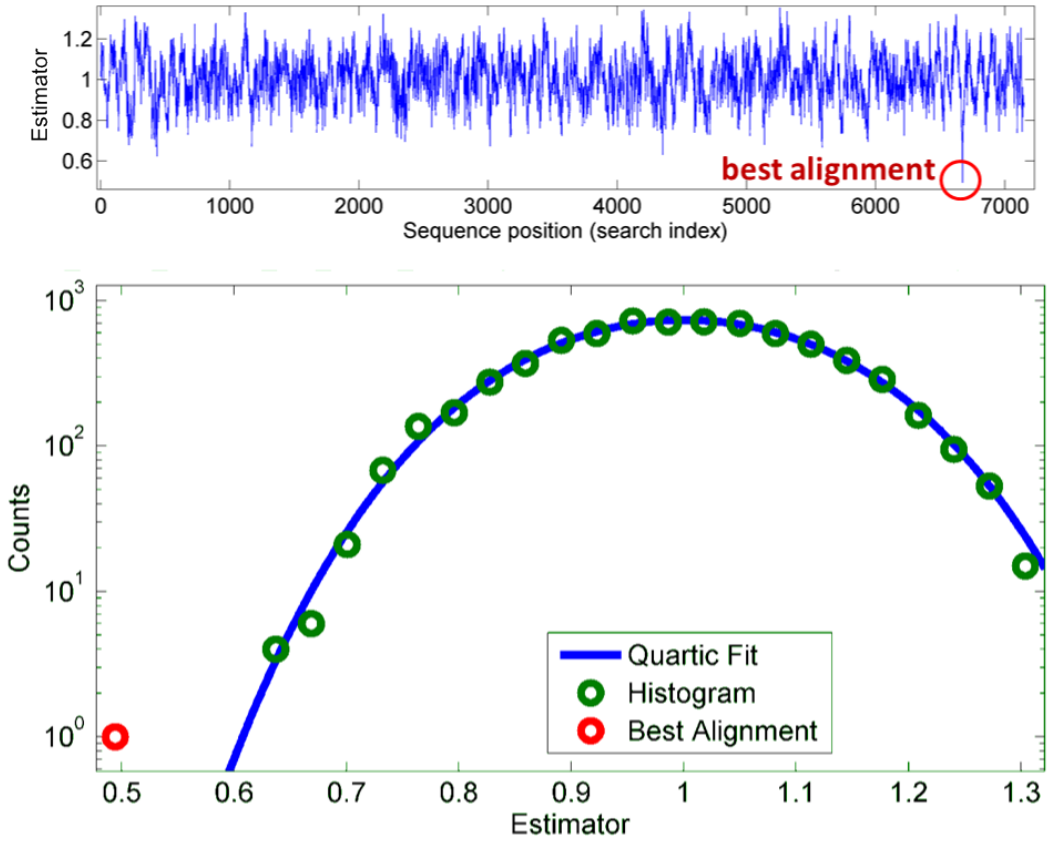

Figure 5–3: Assessing the significance of a location result by analysis of least-squared estimator values

that the correspondence between experimental and predicted barcodes is unusually good at the location where Δ_f occurs, which indicates that the alignment result is unique. This alignment has $n_L = 1.70 \times 10^{-5}$ with a p-value of 0.196, and places the DNA molecule on yeast chromosome 16.

CHAPTER 6
Results

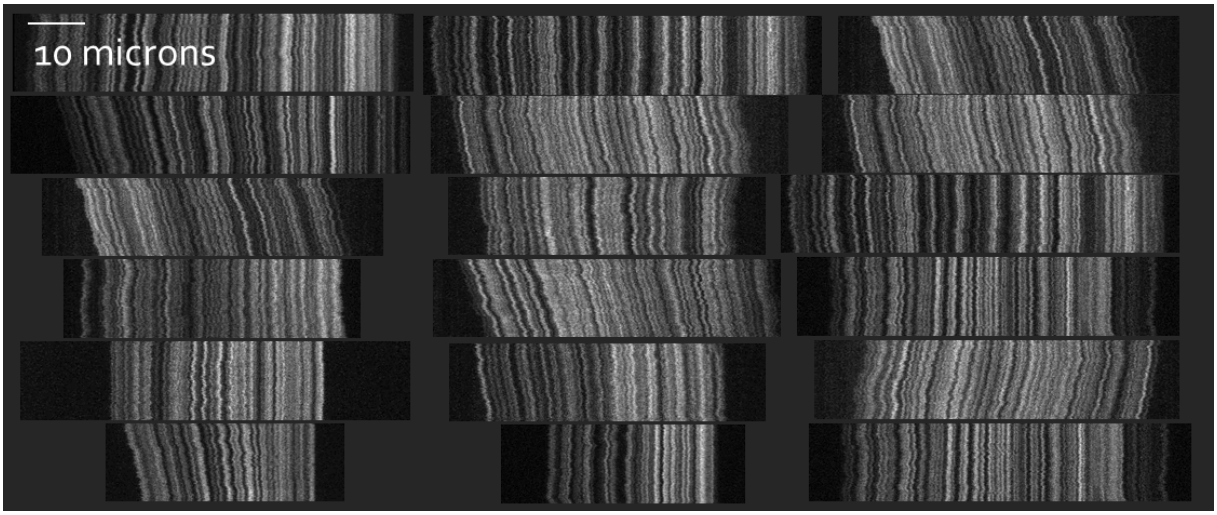

Figure 6–1: Denaturation profiles of yeast DNA

We recorded melting barcodes of more than 130 yeast DNA molecules. Figure 6–1 is a sample of typical fluorescence patterns of partially denatured yeast DNA molecules confined in a nanochannel. The images are time traces where each row is set of pixels, along a nanochannel, from a different frame in a movie. The strategy of preparing DNA by SCODA from gel plugs allowed us to consistently confine and image very long DNA molecules. Some molecules extended beyond the field of view ($\approx 80\,\text{nm}$) when confined.

We aligned melting barcodes from single DNA molecules to the yeast genome, thereby identifying the location of the molecules on the genome, and showed the

location alignment result was statistically significant. (See chapter 5 for details on the alignment algorithms and criteria of statistical significance). This is the first time denaturation mapping has been used to identify a *single* DNA molecule on a genome of this size. Previously, only ensemble averages of melting barcodes were mapped to the λ-phage (48kbp), T4 virus (169kbp) and human chromosome 12 (143Mbp) genomes, and a single melting barcode was mapped only to the short λ-phage genome. [33]. Our result demonstrates that partial denaturation can map single molecules to a genome two orders of magnitude larger than previously shown.

Figure 6–2 shows three typical melting barcodes that have been aligned via comparison to a theoretical barcode computed for the entire genomic sequence. The three molecules, of length 200-250kbp, are located on chromosome 10 *(A)*, chromosome 16 *(B)* and chromosome 12 *(C)* respectively. *(a)* is a comparison of the intensity profiles with respect to position in kbp between experimental melting barcodes (blue) and predicted genomic barcodes (green) at the sequence position where we locate the barcode. The blue trace arises from a time series of fluorescence images of a single DNA molecule (inset, scale bar 10 µm). The traces have been processed using a custom algorithm to remove background and thermal distortions to produce an adjusted barcode dP (see the chapter 5 for details). The profiles are first compared using a position scale of nanometres, and then plotted as seen here on a position scale of base-pairs inferred using the predicted genomic trace. *(b)* is a histogram of the distribution of least-squares estimators Δ calculated at each search position along the genomic sequence (green). The blue curve is a fit of the expected quartic distribution $P(\Delta)$. The red circle is the global minimum of the estimator, taken at

the sequence position where we locate the molecule. The fact that the red point lies to the left of $P(\Delta)$ means the global minimum estimator is lower than would be expected by chance for uncorrelated experimental and genomic barcodes, and therefore that the location result is statistically significant (once more, see chapter 5 for details of the criteria for statistical significance). *(c)* displays the estimator calculated at each search position. Note that the search algorithm excludes genomic barcodes from chromosomes shorter than the experimental barcode. As a result long molecules are located by searching over fewer sequence positions in total, as in *(B)*.

In total we aligned 84 DNA molecules with statistical significance. These 84 melting barcodes form an optical map that covers 56.3% of the yeast genome sequence, about 6.8Mbp in total. Figure 6–3 shows the genomic positions mapped by our experiments. Blue lines represent the genomic sequence of each yeast chromosome on a position scale of 100kbp. Overlapping red lines represent the set of sequence positions on that chromosome where a melting barcode has been located. The shorter red lines below represent the sequence positions occupied by individual DNA molecules. A handful of chromosomes are especially well covered. We mapped 96% of chromosome 3 (total length 317kbp), 92% of chromosome 7 (total length 1.08Mbp) and 91% of chromosome 2 (total length 813kbp). The successful mapping of $> 50\%$ of a Mbp-size eukaryotic genome at the scale of single molecules is an important accomplishment for denaturation mapping as a single molecule optical mapping technique.

Using SCODA to prepare DNA samples overcomes fragmentation challenges and allows the recording of very long optical maps. We observe melting barcodes from

DNA fragments on average 167kbp in length, including three longer than 300kbp. Our barcodes represent large fractions of chromosomes, on average 24% of the entire contour and often more in the case of shorter chromosomes. One 204kbp fragment occupied 64% of chromosome 3. The ability to consistently map such long molecules demonstrates the potential of denaturation mapping as a probe of long-range genomic structure.

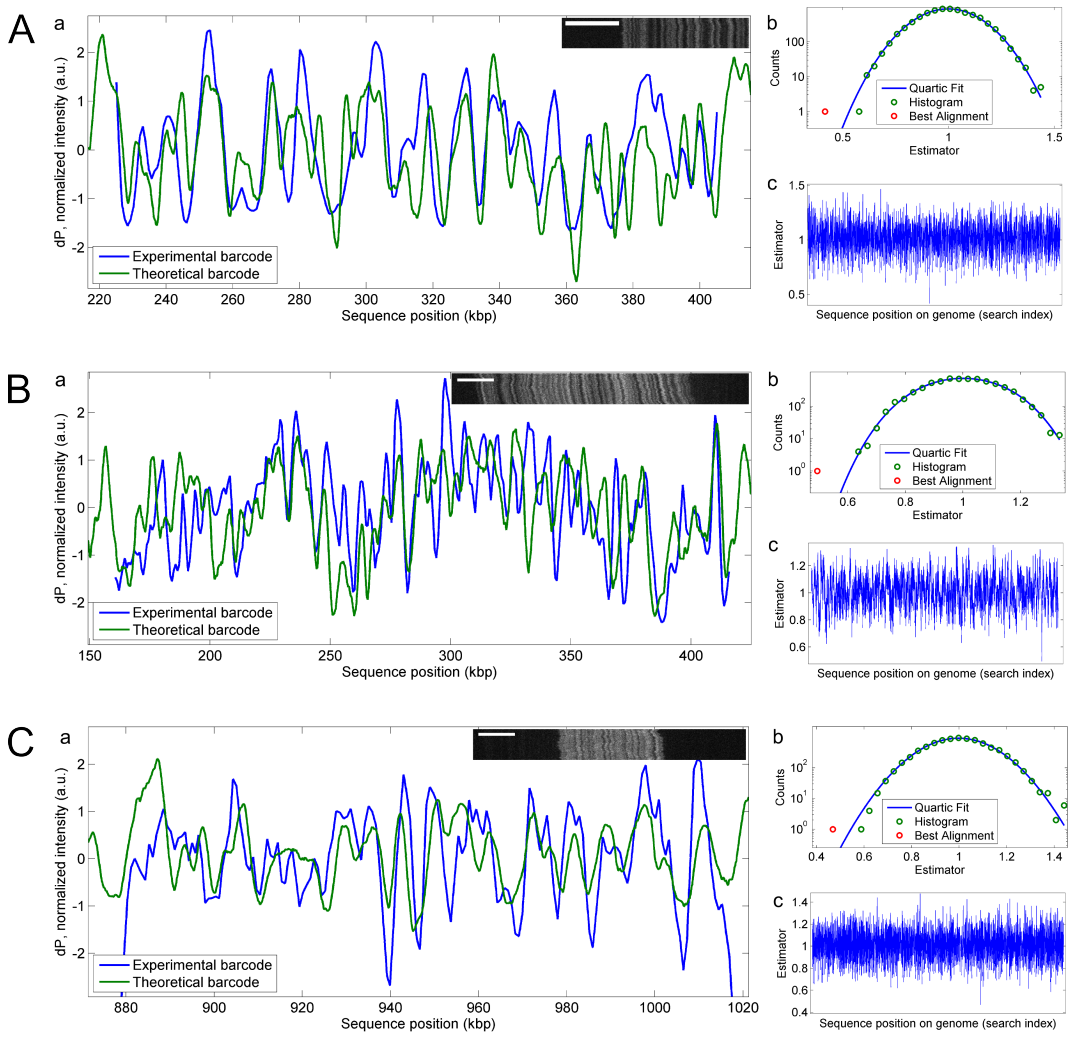

Figure 6–2: Alignment of melting barcodes from single DNA molecules to the yeast genome

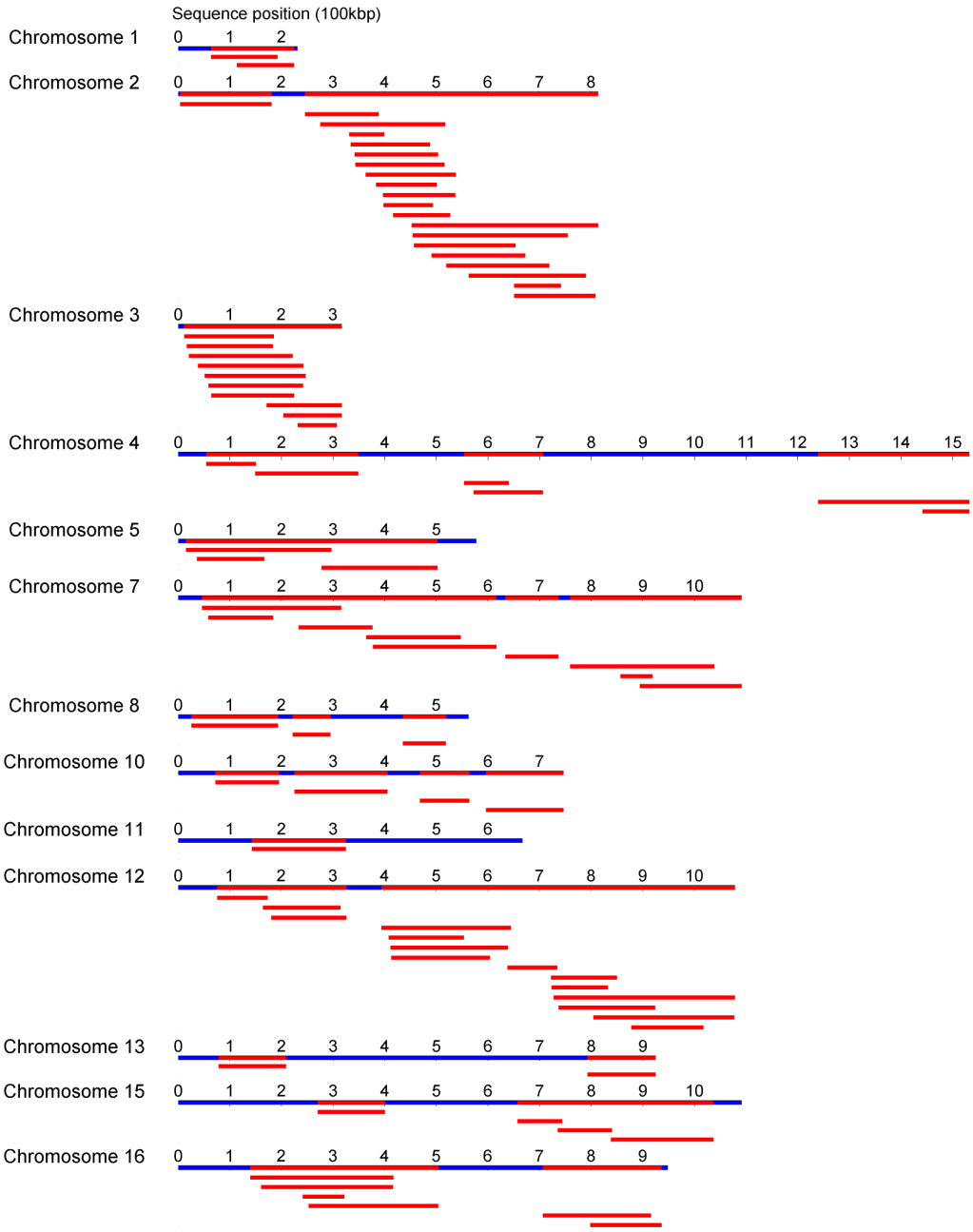

Figure 6–3: Mapping coverage of the yeast genome by alignment of melting barcodes

CHAPTER 7
Discussion

While 84 DNA molecules could be aligned to the genome with statistical significance, of the total of 134 molecules we observed there were 50 molecules that could not be located. These had distributions of the least squares estimator Δ that did not meet our criteria of significance; either they gave a value of $n_L < 0.02$ (3 molecules) or, much more frequently, had a histogram which could not be well-fit to a quartic function $P(\Delta)$ and so n_L could not be meaningfully calculated (the remainder of cases). What factors might prevent the alignment of a melting barcode with statistical significance? It is important to distinguish between molecules which could not be aligned for experimental reasons, because a melting barcode was not correctly formed, and those which could not be aligned for more fundamental reasons. A brief discussion of the relative importance of these factors follows.

7.1 Challenges of producing melting barcodes

There are several reasons that images from a denaturation mapping experiment might fail to correctly reflect a melting pattern. These experimental defects have distinct visual signatures that can be diagnosed by inspecting barcodes. We observed DNA molecules that tied into a knot before confinement (figure 7–1), entered a nanochannel while folded (figure 7–2), failed to melt before an image was recorded, appeared faint due to photobleaching, and broke into multiple fragments (figure 7–3). The statistical test of alignment provided a rigorous way to to identify these

problematic DNA molecules and prevent them from introducing errors in the optical map of the yeast genome. Of the poorly formed barcodes, 10 could be aligned by manually removing portions containing knots, folds, or unmelted contour. These 10 corrected barcodes have been included in the reported final count of 84 molecules.

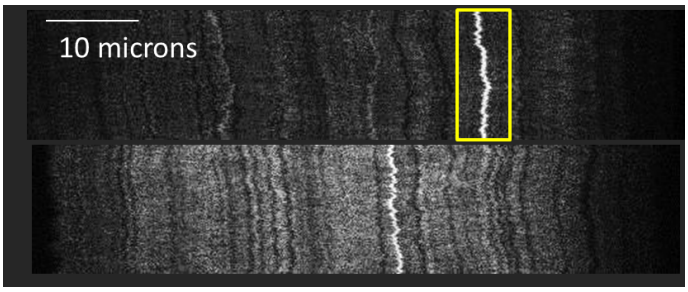

Figure 7–1: DNA molecules with knots

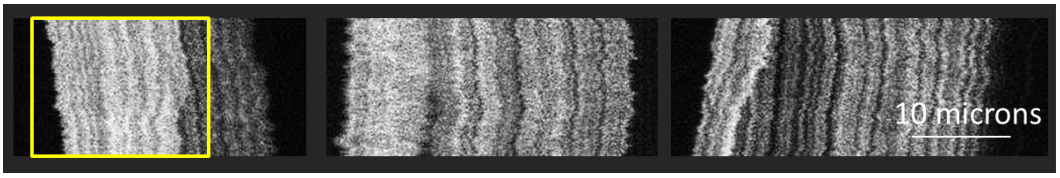

Figure 7–2: DNA molecules with folded contour

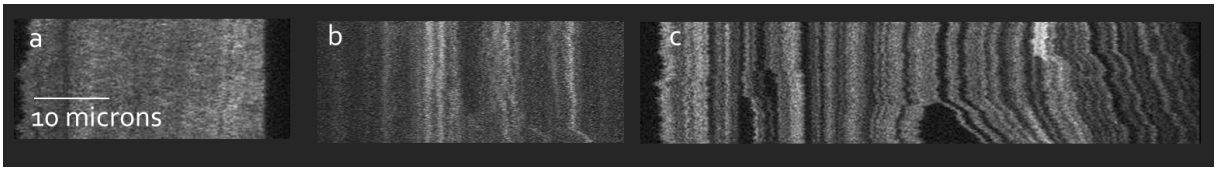

Figure 7–3: DNA molecules with low contrast due to failure to melt completely, low fluorescence intensity due to bleaching, and fragmentation due to photonicking

The numbers of DNA molecules that could not be identified for each these reasons are summarized in the table below.

Experimental defect	# molecules (# adjusted to be aligned successfully)
Knot	7 (4)
Fold	9 (3)
Failed to melt	10 (3)
Photobleached	3
Fragmented	3
Total	32 (10)

7.2 Challenges of aligning melting barcodes to the genome

Of the 50 barcodes that could not be aligned with statistical significance, 18 remain which have no obvious experimental defect. This represents an alignment rate of 82% among melting barcodes which have been correctly formed.

To investigate whether there are reasons intrinsic to a genome sequence that a melting barcode might fail to align, we performed a simple simulation. We attempted to locate samples from the predicted genomic barcode to itself. Here we refer to these as *simulated* melting barcodes. We took 200kbp fragments from the genomic barcode at intervals of 200kbp, so that the entire genome was covered at least once, and added 200kbp fragments from the ends of the chromosomes. In total we took 68 simulated barcodes. Our alignment algorithms correctly located all of the fragments with a least-squares estimator value of 0, meaning a perfect agreement between the fragment and genomic barcode at the location where it was identified. However, not all of these alignments met our statistical criteria of significance. The alignments of 48 fragments were statistically significant. However, 20 of the 68 simulated fragments had distributions of the estimator Δ whose fit with a quartic function $P(\Delta)$ failed a

chi-squared test with a p-value threshold of 0.01. Figure 7–4 shows the estimator distribution of a simulated barcode that cannot be aligned because it fails the chi-squared test.

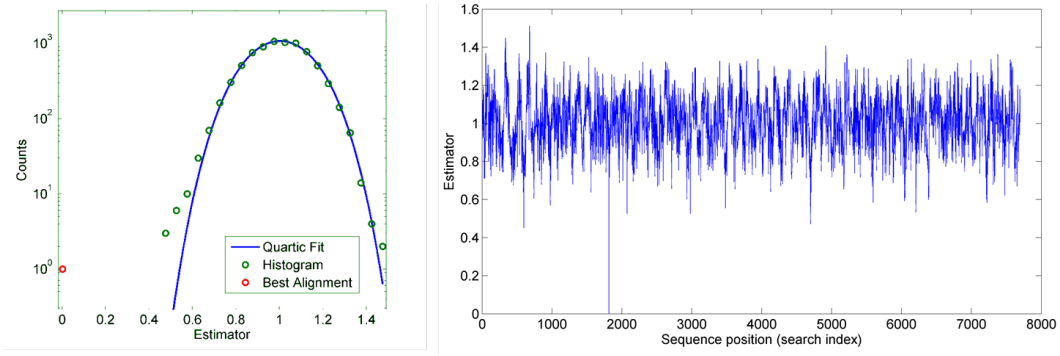

Figure 7–4: Estimator analysis of a simulated melting barcode

Figure 7–5 shows the genomic locations where 200kbp simulated barcodes could and could not be aligned to the yeast genome. Because some fragments could not be aligned with statistical significance, the optical map formed by 200kbp non-overlapping simulated barcodes had only 74% coverage of the yeast genome. The number of alignments which had statistical significance was 80% of the total.

To investigate whether the inability to align simulated barcodes was a function of the fragment size chosen we performed a second simulation. We took simulated barcodes of length 150kbp, which is closer than 200kbp to the mean fragment size 166kbp from our experiments. This time fragments were chosen to overlap by 75kbp so that the simulated barcodes represent at least a two-fold sampling of the genome. Out of a total 131 fragments, 118 could be aligned with statistical significance (90% of the total) and the optical map formed by significant alignments had 93% coverage. The fact that more could be aligned than in the first simulation indicates that

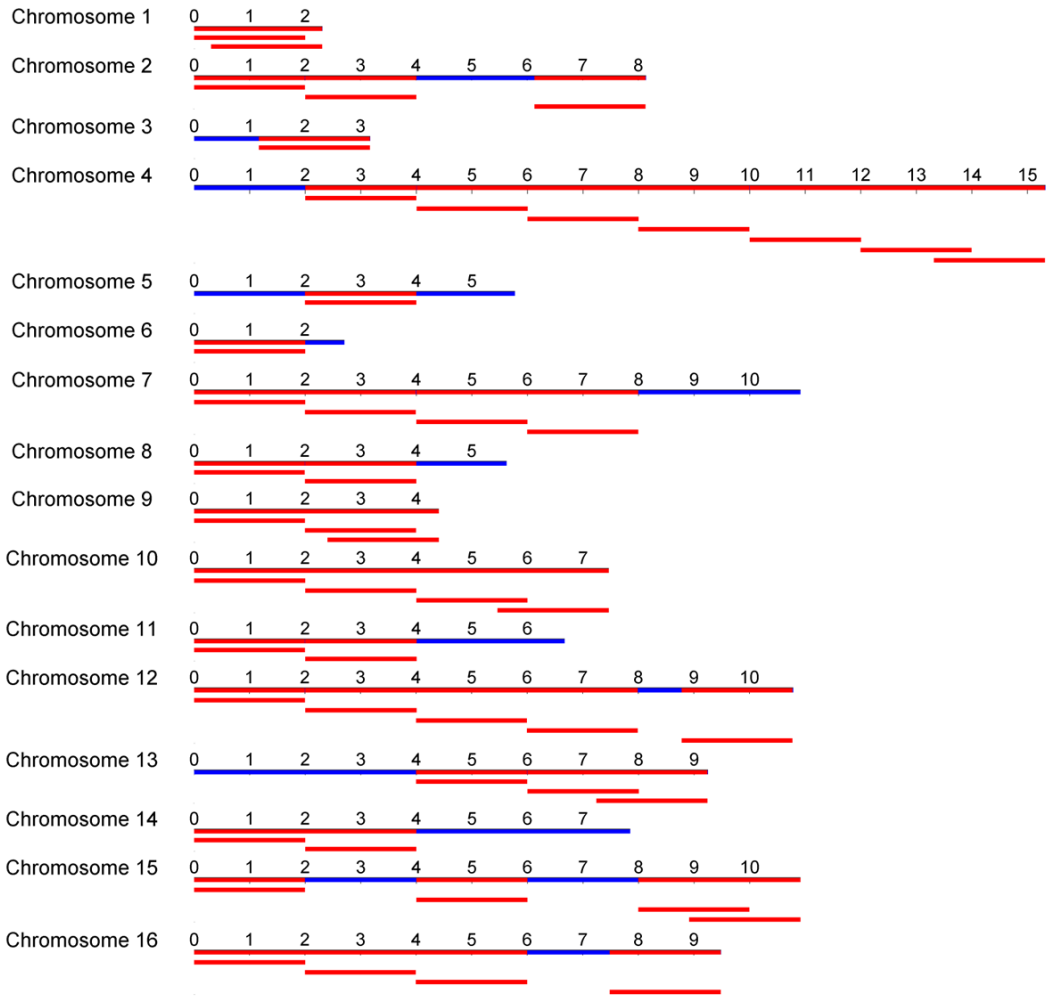

Figure 7–5: Coverage achieved by location of simulated melting barcodes

fragment size affects the statistical significance of alignment results. The effect is due to the structure of the genome and should vary between DNA fragments from different genomes.

Why might the fit of $P(\Delta)$ to an estimator distribution fail a chi-squared test? A distribution of the estimator would fail to follow a gaussian with a quartic correction if a genome's sequence of base-pairs was non-random: correlations between regions of the sequence would lead an estimator distribution to deviate from the predictions of central limit theorem. Real genomes are certainly not random sequences of base-pairs. As a result there exist regions of a genome that display melting barcodes similar to those of other regions. So for each genome the extent of correlations in the sequence gives some intrinsic limit on the success rate of locating barcodes produced by denaturation mapping. The fact in our simulations, only 80-90% of the fragments taken from yeast's predicted genomic barcode had a estimator histogram well-fit with a quartic function, gives some idea of the optimal rate at which of barcodes can be located from our experiments using our criteria of statistical significance. However, the higher optimal rate 90% for fragments closer in size to the experimental barcodes remains larger than our observed 82%. We suggest two more reasons a well-formed barcode may not be aligned to the genome with statistical significance.

Many DNA molecules could be located on the genome with statistical significance despite discrepancies between melting barcodes and the predicted genomic barcode. Occasionally discrepancies of about 5kbp in size may appear between the two profiles that are in otherwise strong agreement. These details may reflect stochasticity in the melting process. While the computed genomic barcode represents the

most probable melting pattern, the process of melting is co-operative and can occur in many trajectories that differ depending on the order in which the features of the melting pattern form.[16] In previous work where barcodes were averaged over an ensemble of molecules these stochastic features of the pattern are "averaged out."[33] We believe that even given small discrepancies the large-scale features of the melting barcodes and predicted barcodes are in agreement. The confirmation that Δ_f lies outside the distribution of other values of Δ across the sequence validates the strength of our location results.

Finally, it may be the case that parts of the yeast genome are difficult or impossible to map due to real differences in sequence between the strain of the reference sequence (S288c) and the strain used for these experiments (YPH80). Large-scale genomic changes, such as insertions, deletions and chromosomal re-arrangements, typically differentiate yeast strains. [9]. These changes are large enough to detect using our device and may contribute both to the number of observed DNA molecules which we cannot locate and the fraction of the genome we have not yet mapped. We are currently sequencing the yeast strain YPH80 in a Roche 454 assay at Génome Québec to improve the coverage of our optical map. We believe that in light of possible genomic differences between YPH80 and S288c our current success in mapping more than half of the genome is a strong demonstration the utility of denaturation mapping.

7.3 Directions for improvement of the denaturation mapping method

The number of molecules which cannot be located due to melting barcodes that are poorly formed can be reduced by refinements in protocol and device design.

Knotting and folding could be discouraged by introducing a post array before the nanochannels to unscroll DNA gradually. Post arrays were used in a recent device to confine Mbp-length DNA molecules in channels without knots or folds. [15] Nicking and photobleaching might be discouraged by optimizing buffer conditions so that oxygen radicals from fluorescence of YOYO-1 are scavenged more effectively. Lastly, DNA molecules could be melted more consistently by using heating apparatus a more precise than the simple resistive heater from this setup. More precise control of temperature during an experiment could also greatly simplify the analysis of melting barcodes. If the helicity of partially melted DNA molecules was the same between experiments it would no longer be necessary to use a helicity interpolation matrix as a predicted genomic barcode. Instead, a genomic barcode could be predicted for just one temperature, and helicity could be removed as a search parameter in the alignment algorithms. Searching over fewer parameters would decrease the time required to align barcodes, and could improve the fraction of alignment results that are statistically significant.

To map larger genomes such as the human genome ($\approx$ 3,100Mbp) it will be necessary to record orders of magnitude more molecules than were used to map the yeast genome in this work. The number of molecules mapped during a single experiment could be improved by making loading and imaging DNA parallel and automated. Several nanochannels are in the field of view at once. If the concentration of input DNA was sufficiently high, the channels could be all loaded and imaged simultaneously. If pumps controlling flow in the device were controlled by computer, and edge

detection algorithms were used to detect the presence of a molecule in a nanochannel, a feedback loop could iterate loading and imaging molecules continuously. Since recording images to produce a melting barcode takes only 5 seconds (50 frames at 10 frames/second), and molecules are on average 166kbp long, if loading new DNA is done in 25 seconds and 10 channels are in the field of view at once a full haploid human genome of 3,100Mbp of DNA could be mapped in just over 15 hours.

The time required to map a genome could also be improved by melting longer molecules. Nanofluidic experiments have been performed with Mbp-sized DNA molecules.[15, 32] The chromosomes of the YPH80 yeast genome are of this size range (225-1900kbp). There are 16 in total: if the intact chromosomes could be loaded into nanochannels, the genome of YPH80 could be mapped in just 16 molecules. As mentioned in section 4.4.2, the step of loading DNA by micropipet creates hydrodynamic shear that breaks molecules into fragments. A new denaturation mapping device design could avoid this shearing by extracting DNA either on the device, as in [32] where chromosomes are digested in a fluidic trap, or on the chuck, as proposed in the preliminary design of 4.4.2. Obtaining long melting barcodes may be crucial to map larger genomes, where the search space for alignment is much larger, in order for a single barcode to have sufficient information that an alignment result is statistically significant.

CHAPTER 8
Conclusion

We used denaturation mapping, the technique of measuring the pattern of partial melting in a DNA molecule confined to a nanochannel, to map single DNA molecules to the genome of *Saccharomyces cerevisiae* (brewer's yeast). In total we located 84 DNA molecules on the genome to form an optical map with more than 50% coverage. This is a dramatic increase in the size of genome that has been mapped by denaturation at the scale of single molecules (12.1Mbp, compared to 48kbp in previous work).[33] Our results are a milestone in demonstrating the potential of denaturation mapping as a probe of long-range genomic structure at a scale of single DNA molecules.

We made several refinements to the protocol of denaturation mapping. We analyzed a large number of DNA molecules quickly using a new suite of MATLAB programs capable of batch processing measurements of many molecules and determining the genomic coverage achieved by the resulting optical map. Overcoming challenges of DNA fragmentation was crucial to identifying the genomic location of a single molecule. We developed a SCODA-based DNA preparation procedure that allowed us to obtain DNA molecules as long as 300kbp. Preliminary work was done to develop a robust and disposable plastic nanochannel device, and an improved chuck to extract DNA from gel plugs directly in order to avoid fragmentation due to pipetting.

Mapping a Mbp-sized eukaryotic genome by denaturation sets the stage for mapping Gbp-sized genomes, such as the human genome. Mapping single human DNA molecules would make denaturation mapping of clinical value as a fast, inexpensive probe of gross chromosomal re-arrangements that lead to genetic disease.[41] Furthermore, identifying single yeast DNA molecules suggests an application to pathogen detection. Melting patterns of a clinical DNA sample could be aligned to the genomes of known pathogens, such as infectious strains of yeast, and a unique location result to a genome would indicate the presence of that pathogen in the sample. Finally, an especially exciting next step from mapping single DNA molecules would be to map DNA from single cells. Distinguishing single cells will require a device capable of extracting DNA as well as mapping it. The ability to probe genomic heterogeneity between single cells would open crucial questions in the development of cancer to investigation for the first time.[38]

References

[1] H Aburatani, V P Stanton, and D E Housman. High-resolution physical mapping by combined Alu-hybridization/PCR screening: construction of a yeast artificial chromosome map covering 31 centimorgans in 3p21-p14. *Proceedings of the National Academy of Sciences of the United States of America*, 93(9):4474–9, April 1996.

[2] R D Blake and S G Delcourt. Thermodynamic effects of formamide on DNA stability. *Nucleic Acids Research*, 24(11):2095–103, June 1996.

[3] R D Blake and S G Delcourt. Thermal stability of DNA. *Nucleic Acids Research*, 26(14):3323–32, July 1998.

[4] R Blossey and E Carlon. Reparametrizing the loop entropy weights: Effect on DNA melting curves. *Physical Review E*, 68(6):1–8, December 2003.

[5] D Botstein, R L White, M Skolnick, and R W Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32(3):314–31, May 1980.

[6] C Bouchiat, M D Wang, J Allemand, T Strick, S M Block, and V Croquette. Estimating the persistence length of a worm-like chain molecule from force-extension measurements. *Biophysical Journal*, 76(1 Pt 1):409–13, January 1999.

[7] David J Broemeling, Joel Pel, Dylan C Gunn, Laura Mai, Jason D Thompson, Hiron Poon, and Andre Marziali. An instrument for automated purification of nucleic acids from contaminated forensic samples. *Journal of Laboratory Automation*, 13:40–48, 2008.

[8] C Bustamante, J F Marko, E D Siggia, and S Smith. Entropic Elasticity of Lambda-phage DNA. *Science*, 265(5178):1599–1600, September 9 1994.

[9] Laura Carreto, Maria F. Eiriz, Ana C. Gomes, Patricia M. Pereira, Dorit Schuller, and Manuel A. S. Santos. Comparative genomics of wild type yeast

strains unveils important genome diversity. *BMC Genomics*, 9, November 4 2008.

[10] Lin Chen, Andreas Manz, and Philip J. R. Day. Total nucleic acid analysis integrated on microfluidic devices. *Lab on a Chip*, 7:1413–1423, 2007.

[11] Somes K. Das, Michael D. Austin, Matthew C. Akana, Paru Deshpande, Han Cao, and Ming Xiao. Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Research*, 2010.

[12] P de Gennes. *Scaling Concepts in Polymer Physics*. Cornell University Press, 1979.

[13] M Doi and S F Edwards. *The Theory of Polymer Dynamics*. Oxford University Press, 1986.

[14] Katja Engel, Lee Pinnell, Jiujun Cheng, Trevor C. Charles, and Josh D. Neufeld. Nonlinear electrophoresis for purification of soil DNA for metagenomics. *Journal of Microbiological Methods*, 88(1):35 – 40, 2012.

[15] Camilla Freitag, Joachim Fritzsche, Fredrik Persson, U. Mir, Kalim, and Jonas O Tegenfeldt. Meandering nanochannels for imaging of ultra-long DNA molecules. In *microTAS*, Seattle, 2011.

[16] O Gotoh. Prediction of melting profiles and local helix stability for sequenced DNA. *Advances in Biophysics*, 16:1–52, 1983.

[17] AJF Griffiths, WM Gelbart, and JH Miller. Chromosomal Rearrangements. In *Modern Genetic Analysis*. W. H. Freeman, New York, 1999.

[18] Andreas Hanke and Ralf Metzler. Comment on Why is the DNA Denaturation Transition First Order?. *Physical Review Letters*, 90(15):159801, April 2003.

[19] Clyde a Hutchison. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research*, 35(18):6227–37, January 2007.

[20] Shendure J and Ji H. Next generation DNA sequencing. *Nat. Biotech.*, 26:1135–1145, 2008.

[21] Y. Kafri, D. Mukamel, and L. Peliti. Kafri, Mukamel, and Peliti Reply:. *Physical Review Letters*, 90(15):159802, April 2003.

[22] Stephen L. Levy and Harold G. Craighead. DNA manipulation, sorting, and mapping in nanofluidic systems. *Chemical Society Reviews*, 39:1133–1152, 2010.

[23] E Lifshitz and L Landau. *Fluid Mechanics*. Butterworth Heinmann, 1998.

[24] E Lyon. Mutation detection using fluorescent hybridization probes and melting curve analysis. *Expert Review of Molecular Diagnostics*, 1(1):92–101, May 2001.

[25] John F. Marko and Eric D. Siggia. Stretching DNA. *Macromolecules*, 28(26):8759–8770, December 1995.

[26] Andre Marziali, Joel Pel, Dan Bizzotto, and Lorne A Whitehead. Novel electrophoresis mechanism based on synchronous alternating drag perturbation. *Electrophoresis*, 26(1):82–90, 2005.

[27] Robert K. Neely, Jochem Deen, and Johan Hofkens. Optical mapping of DNA: Single-molecule-based methods for mapping genomes. *Biopolymers*, 95(5):298–311, 2011.

[28] Theo Odijk. On the statistics and dynamics of confined or entangled stiff polymers. *Macromolecules*, 1344(3):1340–1344, 1983.

[29] Theo Odijk. Scaling theory of DNA confined in nanochannels and nanoslits. *Physical Review E*, 77(6):060901, 2008.

[30] D Poland. Recursion relation generation of probability profiles for specific-sequence macromolecules with long-range correlations. *Biopolymers*, 13(9):1859–71, January 1974.

[31] Douglas Poland and Harold A Scheraga. Phase transitions in one dimension and the helixcoil transition in polyamino acids. *The Journal of Chemical Physics*, 45(5):1456, 1966.

[32] Kristian H. Rasmussen, Rodolphe Marie, Jacob M. Lange, Winnie E. Svendsen, Anders Kristensen, and Kalim U. Mir. A device for extraction, manipulation and stretching of DNA from single human chromosomes. *Lab on a Chip*, 11:1431–1433, 2011.

[33] Walter Reisner, Niels B. Larsen, Asli Silahtaroglu, Anders Kristensen, Niels Tommerup, Jonas O. Tegenfeldt, and Henrik Flyvbjerg. Single-molecule denaturation mapping of DNA in nanofluidic channels. *Proceedings of the National*

Academy of Sciences of the United States of America, 107(30):13294–13299, JUL 27 2010.

[34] Walter Reisner, Keith Morton, Robert Riehn, Yan Wang, Zhaoning Yu, Michael Rosen, James Sturm, Stephen Chou, Erwin Frey, and Robert Austin. Statics and dynamics of single DNA molecules confined in nanochannels. *Physical Review Letters*, 94(19):1–4, May 2005.

[35] Robert Riehn, Manchun Lu, Yan-Mei Wang, Shuang Fang Lim, Edward C. Cox, and Robert H. Austin. Restriction mapping in nanofluidic devices. *Proceedings of the National Academy of Sciences of the United States of America*, 102(29):10012–10016, 2005.

[36] Jared C Roach, Cecilie Boysen, I K A I Wang, and Leroy Hood. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics*, 353, 1995.

[37] Chanchal Sadhu and Santanu Dutta. Influence of formamide on the thermal stability of DNA. *Biophysics*, 6(6):817–821, 1984.

[38] Jesse J. Salk, Edward J. Fox, and Lawrence A. Loeb. Mutational heterogeneity in human cancers: origin and consequences. *Annual Review of Pathology-Mechanisms of Disease*, 5:51–75, 2010.

[39] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, OCT 2008.

[40] MW Smith, AL Holmsen, YH Wei, M Peterson, and GA Evans. Genomic sequence sampling - a strategy for high-resolution sequence-based physical mapping of complex genomes. *Nature Genetics*, 7(1):40–47, May 1994.

[41] Pawel Stankiewicz and James R. Lupski. Genome architecture, rearrangements and genomic disorders. *Trends in Genetics*, 18(2):74 – 82, 2002.

[42] Dirk Stigter. Interactions of highly charged colloidal cylinders with applications to double-stranded DNA. *Biopolymers*, 16(7):1435–1448, 1977.

[43] Michael R. Stratton, Peter J. Campbell, and Andrew P. Futreal. The cancer genome. *Nature*, 458(7239):719–724, April 2009.

[44] Jonas O. Tegenfeldt, Christelle Prinz, Han Cao, Steven Chou, Walter W. Reisner, Robert Riehn, Yan Mei Wang, Edward C. Cox, James C. Sturm, Pascal Silberzan, and Robert H. Austin. The dynamics of genomic-length DNA molecules in 100-nm channels. *Proceedings of the National Academy of Sciences of the United States of America*, 101(30):10979–10983, 2004.

[45] Eivind Tø stesen, Fang Liu, Tor-Kristian Jenssen, and Eivind Hovig. Speed-up of DNA melting algorithm with complete nearest neighbor properties. *Biopolymers*, 70(3):364–76, October 2003.

[46] E Tostesen, GI Jerstad, and E Hovig. Stitchprofiles.uio.no: analysis of partly melted DNA conformations using stitch profiles. *Nucleic Acids Research*, 33(2):W573–W576, JUL 1 2005.

[47] Dieter Trau, Thomas M. H. Lee, Alex I. K. Lao, Ralf Lenigk, I-Ming Hsing, Nancy Y. Ip, Maria C. Carles, and Nikolaus J. Sucher. Genotyping on a complementary metal oxide semiconductor silicon polymerase chain reaction chip with integrated DNA microarray. *Analytical Chemistry*, 74(13):3168–3173, 2002.

[48] Pawel Utko, Fredrik Persson, Anders Kristensen, and Niels B. Larsen. Injection molded nanofluidic chips: Fabrication method and functional tests using single-molecule DNA experiments. *Lab on a Chip*, 11:303–308, 2011.

[49] J. Craig et al. Venter. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[50] Jing Wang, Zongyuan Chen, Paul L. A. M. Corstjens, Michael G. Mauk, and Haim H. Bau. A disposable microfluidic cassette for DNA amplification and detection. *Lab on a Chip*, 6:46–53, 2006.

[51] Roger M. Wartell and Albert S. Benight. Thermal denaturation of DNA molecules: A comparison of theory with experiment. *Physics Reports*, 126(2):67 – 107, 1985.

[52] J.D. Watson. and Crick F.H.C. Molecular structure of nucleic acids - a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[53] Christopher L Woodcock and Stefan Dimitrov. Higher-order structure of chromatin and chromosomes. *Current Opinion in Genetics and Development*, 11(2):130 – 135, 2001.

[54] Adam T. Woolley, Dean Hadley, Phoebe Landre, Andrew J. deMello, Richard A. Mathies, and M. Allen Northrup. Functional integration of PCR amplification and capillary electrophoresis in a microfabricated DNA analysis device. *Analytical Chemistry*, 68(23):4081–4086, 1996.