UNIVERSITY OF CALIFORNIA, SAN DIEGO

Single Molecule FRET Sequencing on Consumable Microfluidic Devices

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Electrical Engineering (Photonics)

by

Jeremy Yuenchen Huang

Committee in charge:

Professor Yu-Hwa Lo, Chair
Professor Joseph R. Ecker
Professor Drew A. Hall

2015

The Thesis of Jeremy Yuenchen Huang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2015

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Joseph Ecker for his guidance and support during my two years in the Plant Biology Laboratory at The Salk Institute for Biological Studies. His guidance and support for my journey has been invaluable.

I would like to thank Professor Yu-Hwa Lo for being the chair of my committee and the assistance he and his group had as our collaborators.

I would like to thank Professor Drew A. Hall for his helpful comments and suggestions for this thesis.

I would like to thank Dr. Chongyuan Luo for all the help he has given me on this project over the past two years. His support and advice have been truly stellar.

I would like to thank Dr. Joe Beecham, Theo Nikorelev and Dr. Xin Wang for being our collaborators: Theo for providing us with the Starlight sequencing chemistry and Xin for helping us with data analysis.

Chapters 1 – 3, in part, are currently being prepared for submission for publication of the material. Huang, Jeremy; Huang, Jeremy Y. The thesis author was the primary investigator and author of this material.

ABSTRACT OF THE THESIS

Single Molecule FRET Sequencing on Consumable Microfluidic Devices

by

Jeremy Yuenchen Huang

Masters of Science in Electrical Engineering (Photonics)

University of California, San Diego, 2015

Professor Yu-Hwa Lo, Chair

We present a technique based on Fluorescent Resonance Energy Transfer (FRET) mechanism that enables single-molecule sequencing with a minute amount of starting cells or tissues. During DNA polymerase catalyzed nucleotide incorporation, a donor fluorophore conjugated onto the polymerase acts as a transition state for a highly localized transfer of energy to a recipient fluorophore on a target deoxyribonucleotide

triphosphate (dNTP). This enables local signal amplification in four target wavelengths eliminating the need for DNA amplification through PCR or waveguide-mediated signal boosting. The donor particle is stimulated using a 532nm laser in a total internal reflection microscope (TIRFM) setup that reduces background noise by using evanescently propagating waves to interact only within 100nm of the surface of the glass substrate. The sample can be prepped using fluidics in low concentrations (< 100pM) and is attached onto a biotin-streptavidin functionalized surface. Since the fluorescent dNTPs are not end-terminated, all templates exposed to the dNTP solution will sequence, regardless if they are being observed. Our solution is a microfluidic polydimethylsiloxane (PDMS) device that is capable of individually addressing isolated sequencing sites. We demonstrate the viability of the complete system to provide accurate short template sequencing and provide a proof of principle for DNA methylation detection and long-template sequencing on our next-generation protocols and devices.

CHAPTER ONE

1.1 Introduction

The availability of next generation sequencing (NGS) tools has enabled rapid advances in genomic science within the past five years[1, 2]. Current high-throughput techniques such as Illumina/Solexa's sequencing-by-synthesis (SBS), Roche's 454, and ABI's SOLID are capable of sequencing an entire human genome within several days[3]. The flexibility of systems like Illumina's SBS allows for protocol modifications such as RNA-seq[4], MethylC-Seq[5], and CHIP-Seq[6] to analyze the transcriptome, methylome, and DNA-protein interactions. However, these high-throughput techniques are limited to sequencing short read-lengths which greatly restricts the amount of structural variation in the genome that can be detected[7, 8]. Furthermore, these methods have complicated library preparation steps and necessitate PCR amplification which introduces PCR-specific biases[9] and phasing issues[10].

Current approaches that attempt to achieve accurate high-throughput long-template sequencing include: fluorescent observation of single-nucleotide incorporation events in zero-mode waveguides[11] and measurements of conductance fluctuations across nanopores[12, 13]. In particular, Pacific Bioscience's single-molecule real-time (SMRT) sequencing platform recently demonstrated near 85% single-base accuracy over lengths of up to 15kb[14,15]. However, these methods still need a significant amount of starting DNA to sequence, have lengthy library preparation steps, and fabrication challenges which restrict their ability for introduction into the sequencing market.

An integrated single-molecule sequencing process that is capable of sequencing kilo-base or mega-base long templates with simple library prep, picomolar input DNA

concentration, and without necessitating DNA amplification has yet to be demonstrated. Such a tool will enable better understanding of structural variations, repeats, and possibly open the door for single-cell amplification-less sequencing.

In order to address this need we introduce a novel high-throughput FRET-based sequencing platform that utilizes resonance properties between two fluorophores. This FRET-based solution utilizes one excitation source and is able to accurately sequence DNA strands at pico-molar concentrations. Individual single-stranded DNA templates are able to be anchored onto an un-patterned surface and spaced out by dilution only, simplifying surface prep and protocol. Individually toggled microfluidic reaction sites provide potential for an entire genome being sequenced on a single consumable PDMS chip. Simplicity of consumable fabrication through soft-lithographic techniques boosts out platform's accessibility for laboratory and high-throughput commercial and laboratory needs. We provide proof-of-principle for a high-throughput short-template sequencing and give a proof-of-concept for long-template sequencing using a parallelized short-template sequencing technique.

1.2 Overview of FRET Mechanisms

Fluorescence resonance energy transfer is a phenomenon that describes a distance-dependent transfer of energy between light-sensitive particles[16, 17]. Its use in detecting interactions at the nanoscale has been well documented[18, 19]. In this phenomena, electrons in the donor particle are first stimulated to an excited state by an external light source. Through dipole-dipole coupling, energy is transferred to an acceptor particle and the subsequent relaxation and energy release gives off an emission in a lower energy band. Response time for an individual excitation and emission event is on the order of

nanoseconds and the efficiency of this system is inversely proportional to the distance (Figure 1) between the donor and acceptor particle.

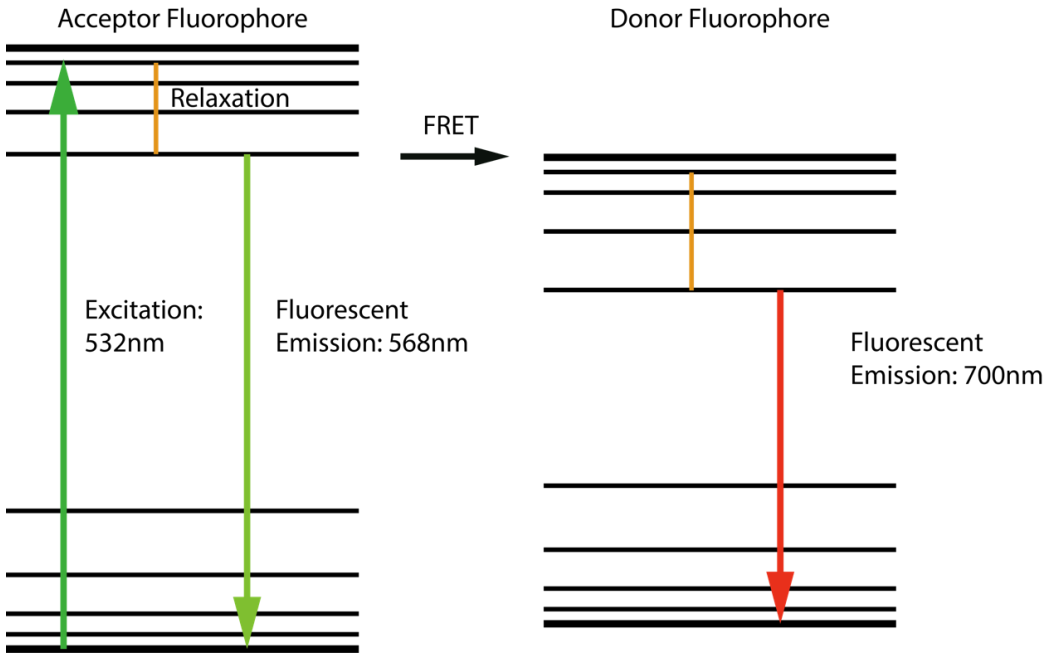

Figure 1: Jablonski diagram of FRET resonant energy transfer between donor and acceptor fluorophores. In this case, coupling is demonstrated between Alexa Fluor 555 dye and Alexa Fluor 700.

Efficiency of FRET is a function of the inverse-sixth-power of distance, R, between the donor and acceptor fluorophores[20]. This is expressed in the following equation:

$$E = \frac{1}{1 + \left[\frac{R}{R_0}\right]^6} \qquad (1)$$

Where R is the distance between the donor and acceptor particle in media, $R_0$is the constant where energy transfer E is 50% efficient. R_0 is defined by the following equation[20]:

$$R_0 = 0.211[J(\lambda)\,\phi_d n^{-4}\kappa^2]^{-1/6} \quad (in\ \text{Å}) \qquad (2)$$

J is the spectral overlap between the donor and acceptor (figure) and ϕ_d is the donor quantum yield, n is the index of refraction of the medium and κ^2 accounts for the orientation of the dipoles (in most systems κ^2 is assumed to be 2/3). Spectral overlap is determined by Equation 3:

$$J = \int_0^\infty F_D(\lambda)\varepsilon(\lambda)\lambda^4 d\lambda \quad (3)$$

where ε is the extinction coefficient of the acceptor, F_D is the emission spectra of the donor, and lambda is the wavelength.

1.2.1 The Starlight FRET System

By conjugating a donor fluorophore onto a DNA polymerase and acceptor fluorophores onto dNTPs, we can reliably discern single-molecule FRET interactions during DNA synthesis. During incorporation of a single nucleotide in our sequencing by synthesis approach, the acceptor nucleotide is within 4nm of a field of up to 15 donor fluorophores which are located near the surface of the DNA polymerase. Our sequencing chemistry utilizes Alexa Fluor 555 as our donor particle and four different acceptor fluorophores DY 634, AF 647, AF 676, and AF 700, attached to the four different nucleotides. The long emission tail of the Alexa Fluor allows multiple fluorophores to couple with a single donor (Figure 2). As shown in Equation 2, FRET efficiency is proportional to spectral overlap and thus we selected four dyes which excitation spectra interact efficiently with our Alexa Fluor 555's emission.

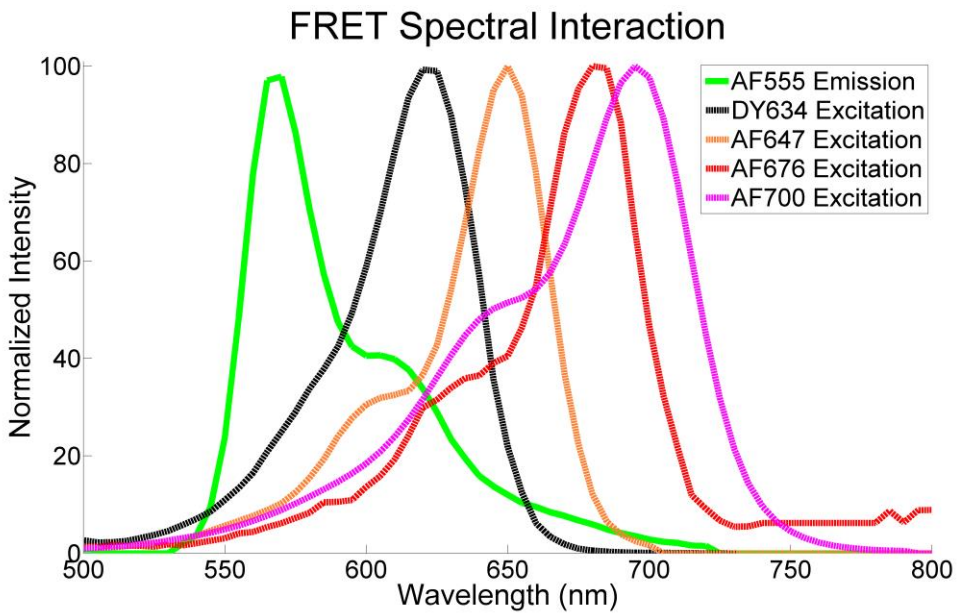

Figure 2: Emission-Excitation spectra of tested fluorophores for Starlight Chemistry. Long emission tail of Alexa Fluor 555 allows for FRET interaction with multiple fluorophores.

Acceptor fluorophores are also carefully selected to minimize emission overlap. However, single-molecule events are difficult to resolve even with the sensitivity of an Electron Multiplying Charge Coupled Device (EMCCD) and thus we need broadband filters to capture as much light as possible, while still being able to individually distinguish each fluorophore. Ratios of signals from different channels, caused by filter bleed-through is utilized to determine the identity of the incorporating nucleotide. Figure 3 demonstrates the emission spectra of our four fluorophores and the creation of three channels utilizing broad-band filters. The filter set was designed such that there was minimal interaction with the strong emission signature of the Alexa Fluor 555 molecule and to facilitate base-calling utilizing our combination of distinct signals from three channels (Figure 4). During data acquisition our camera is optically sectioned into three different regions, each corresponding to the same field-of-view on the chip but containing a different filtered spectrum.

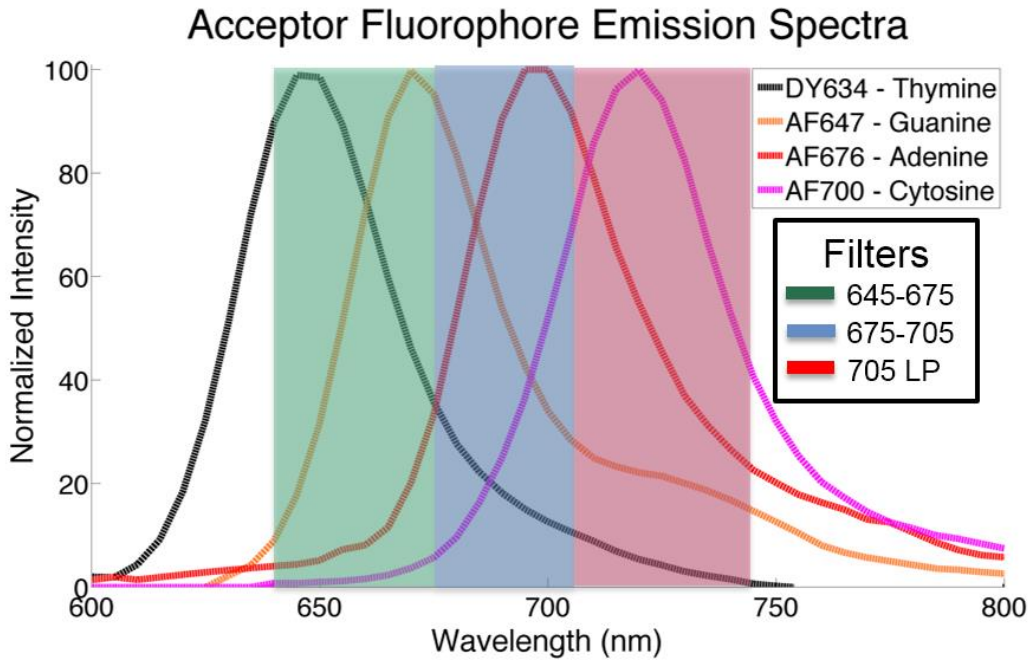

Figure 3: Emission spectra of acceptor fluorophores. Filters were designed so that there was minimal interaction with the bright fluorescent emission of AF555. Signals are distinguished in post-processing as the combination of signals from three channels.

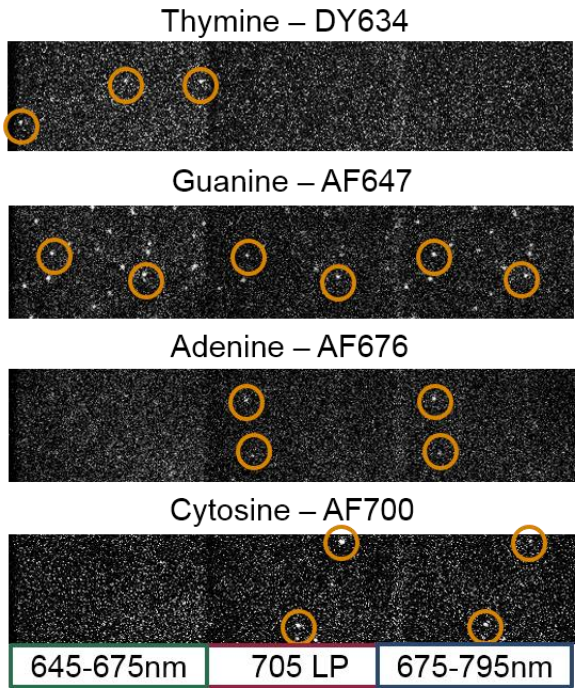

Figure 4: Single-base Starlight sequencing example. Base calls are determined by combination of signals from three channels. Each channel observes the same field-of-view but is filtered in a different spectrum. Corresponding circles highlight a single sequencing site (a single dot), consisting of one template and one DNA polymerase. Each of the colors in the signals on the left correspond to a filter mentioned in Figure 3.

We utilize a custom two-camera TIRF setup (Figure 5) to observe our multi-channel signal. A 532nm laser is coupled to our Olympus IX81 inverted microscope as our excitation source. A 100x oil immersion TIRF objective interfaces with our mounted microfluidic device Free-space dichroic cubes, mirrors, and filters localize the three different channels spatially on our primary EMCCD. A secondary EMCCD takes the emission from the Alexa Fluor 555 fluorophore to use as a reference data set for DNA site alignment.

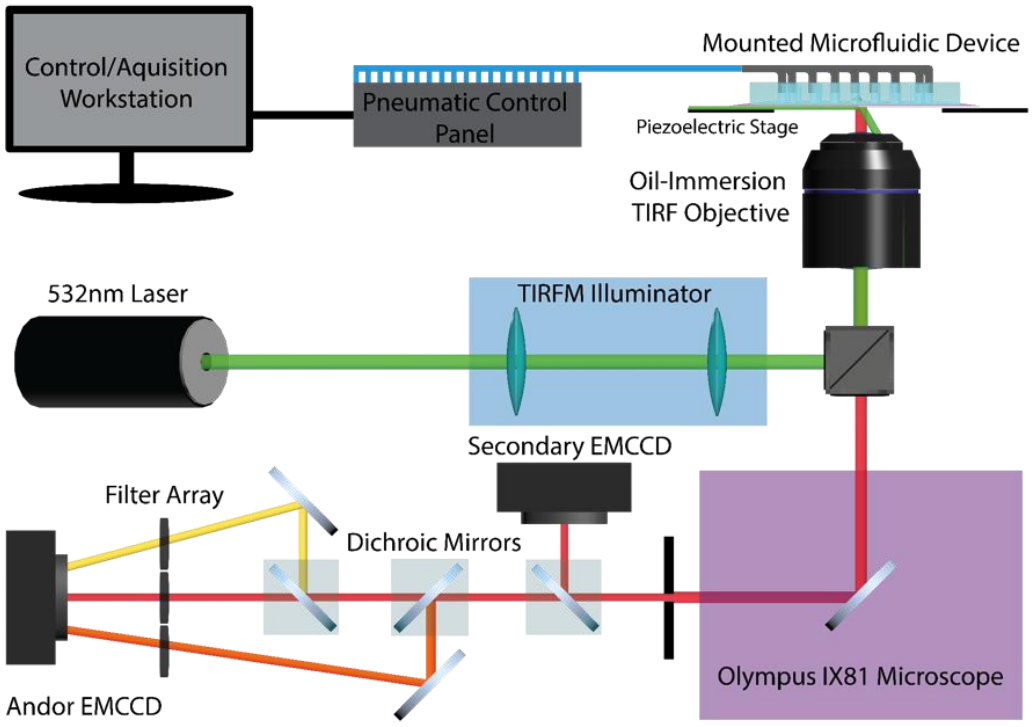

Figure 5: Schematic of integrated TIRFM setup. Microfluidic device is mounted on a remotely controlled piezo-electric stage and FRET sequencing sites are individually toggled and excited through coupling of a 532nm laser source.

Due to the high concentration of nucleotides in solution, there is an abundant number of non-specific interactions with the donor dye contributing to an increase in background noise. Any acceptor fluorophore that is within the FRET criteria will interact

with the donor fluorophore. However, the time-averaged interaction of an unbound fluorescent nucleotide in solution within an ample efficiency radius is measured in the nanoseconds because they are not anchored and quickly carried away by Brownian motion and fluid flow. Incorporation frequency is measured on hundred millisecond and second timescales and the integration time of the EMCCD is also in the milliseconds, reducing floating dNTP interactions to slightly increasing the background noise. We are able to not only confine our sequencing interaction to within several nanometers, but also limit the amount of background auto-fluorescent noise measured using TIRF microscopy.

1.3 Overview of Total Internal Reflection Microscopy

Though we are able to localize the majority of our excitation energy in a small window utilizing FRET, background fluorescence and scattering in the media and the device still significantly impede the signal to noise ratio (SNR). In order to maximize SNR, we confine donor fluorophore excitation energy to within 100nm of the substrate by utilizing total internal reflection microscopy (TIRFM) [21, 22]. In total internal reflection, a propagating wave that strikes an index of refraction transition boundary at greater or equal to a critical angle will reflect. At angles slightly greater than the critical angle, there is a limited amount of evanescent wave energy that penetrates into the media. This energy dies down exponentially as a function of distance into the media and thus is a prime source of localized excitation energy for our near-substrate FRET interactions.

The interaction distance intensity (localization of the intensity) is based on a set of commonly known equations. We first have the incident total internal reflection angle based on the equation for total internal reflectance given by deriving Snell's Law.

$$\theta_c = sin^{-1}(\frac{n_1}{n_3}) \quad (4)$$

where θ_c is the critical angle, $n_1$is the refractive index of the media before the interface, and n_3 is the refractive index of the media after the interface.

Intensity decays as a function of distance z from the interface[22] and is given as:

$$I(z) = I(0)e^{-z/d} \quad (5)$$

where the penetration depth d is:

$$d = \frac{\lambda_0}{\pi}[(n_3)^2 sin^2(\theta) - (n_1)^2]^{-1/2} \quad (6)$$

and λ_0 is the wavelength of the incident light and θ is an angle greater than the critical angle θ_c.

1.4 Single-Molecule Sequencing Process

We combine both TIRF and FRET in our DNA sequencing by synthesis approach (Figure 6). A polyethylene glycol-biotin functionalized surface to capture in-media floating DNA templates. These templates are terminated at either end with biotin attached to an extended adenosine chain and also ligated with known primers to facilitate DNA polymerase incorporation. We treat the functionalized biotin surface with free-floating streptavidin proteins in buffer solution to cover all free biotin strands. Low-concentration of free-floating templates are introduced and through experimentation, approximately 1500 per 100 um^2 are allowed to bind to the surface. Mutant phage B103 DNA polymerase is conjugated with several donor fluorescent particles and introduced in the media, binding onto the primer on the parent strand. This B103 polymerase is engineered to slow down incorporation events from tens of milliseconds to seconds to increase signal integration time for capture on EMCCD cameras.

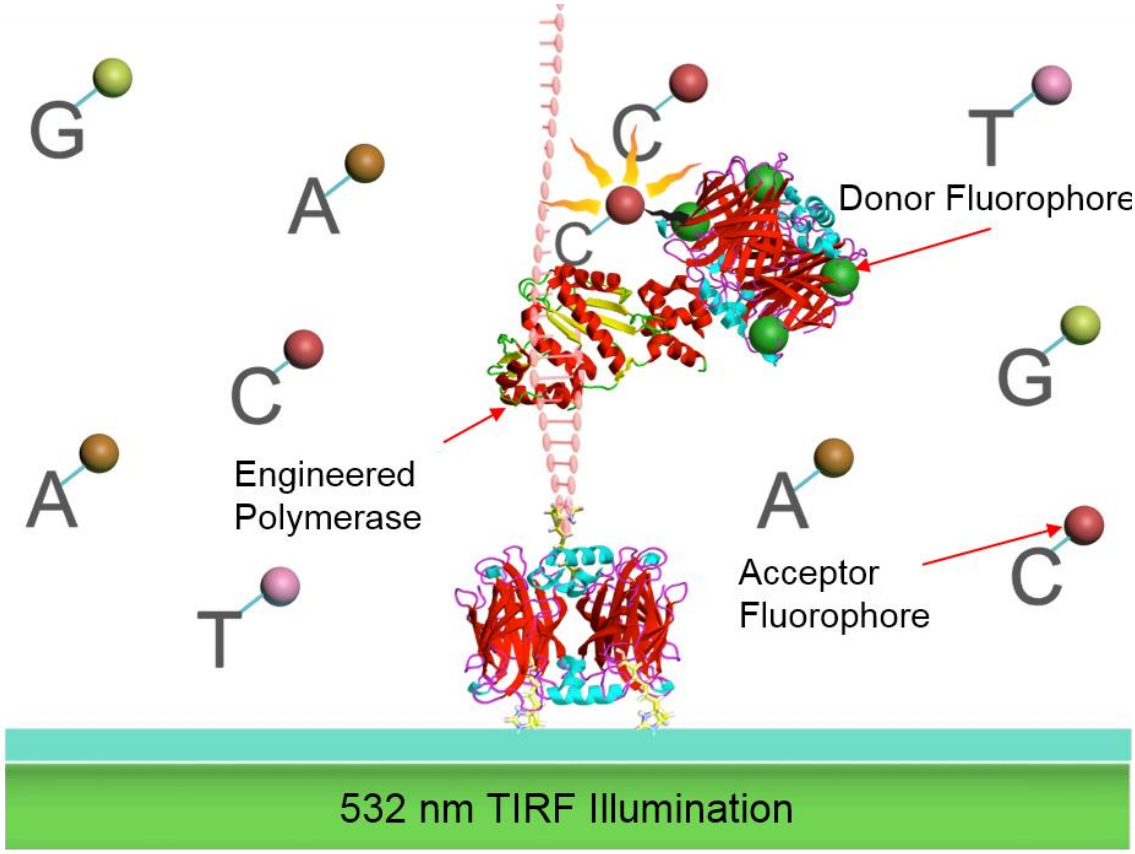

Figure 6: Schematic of DNA sequencing mechanism. Streptavidin[23] on the 3' end is conjugated with a biotin ligated parent template and anchors to the biotin[24] on the surface. DNA polymerase[25] (in our case B103) is conjugated with a streptavidin that is coated with multiple AF 555 fluorophores. TIRF illumination excites the donor fluorophore which in turn resonates with the acceptor fluorophore conjugated on the cytosine.

Our DNA sequencing reaction is initiated by filling the reaction site with fluorescently labeled dNTPs in oxygen scavenging buffer to assist in prolonging polymerase lifetime. These dNTPs have additional phosphate groups attached to the 5' triphosphate to enable conjugation of a fluorescent acceptor particle. During a specific incorporation event (Figure 7), the acceptor fluorophore receives energy from the donor fluorophores located on the outside of the B103 polymerase and individual spectra are discernable for different nucleotides. The polymerase proceeds to incorporate nucleotides until the entire strand is synthesized.

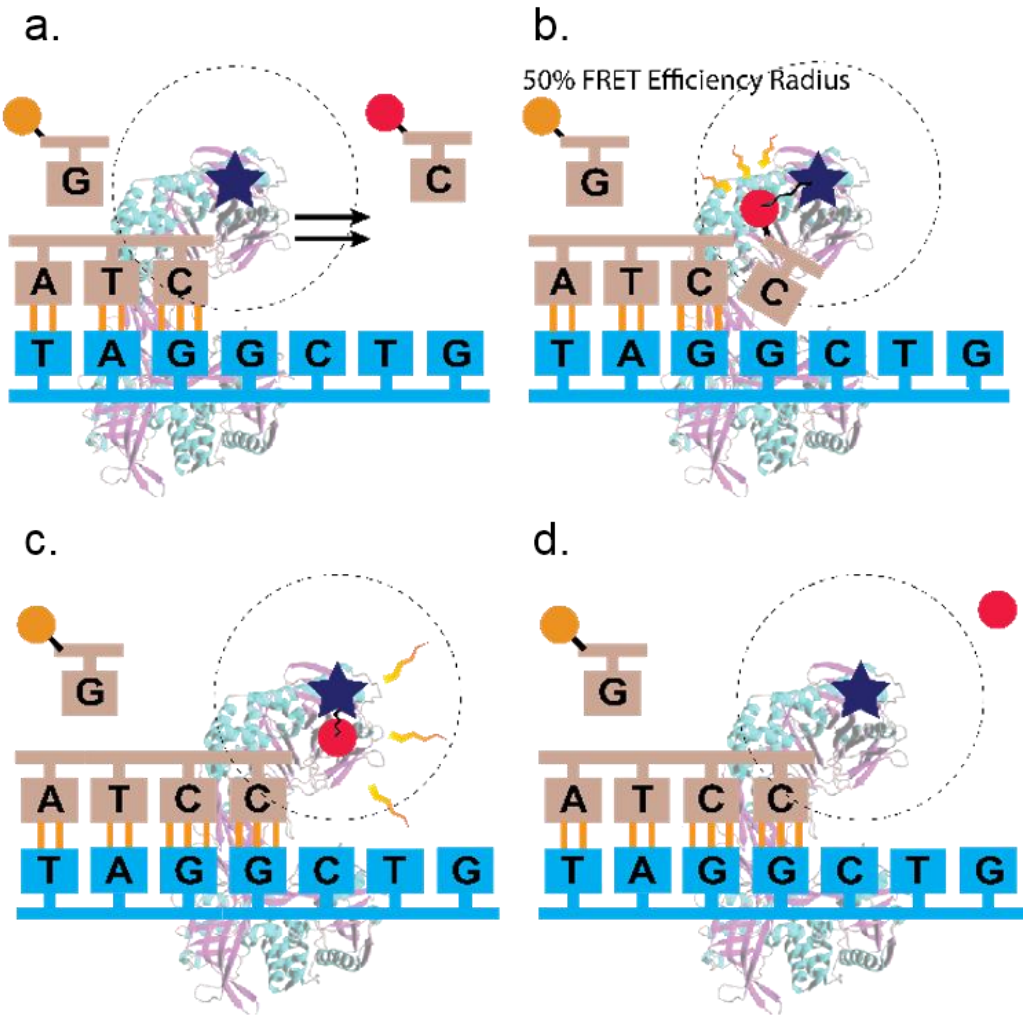

Figure 7: Process of an individual sequencing reaction. (a) Polymerase advances towards open template base. (b) Complimentary base-pair attaches, a reaction that takes approximately 300 milliseconds for a T and up to several seconds for a G. (c, d) When incorporation is complete, the fluorophore is cleaved off and Brownian motion or fluid flow carriers the particle outside of FRET range.

It is clear that the Starlight sequencing chemistry, in conjunction with a TIRFM setup, can help discern individual sequencing events on a single strand of DNA. However, in order to determine the genetic code of an organism in a cost-effective manner, thousands of sequencing reactions need to be observed in parallel. In the following two chapters, a high-throughput short-template sequencing platform utilizing the Starlight FRET sequencing chemistry is demonstrated.

CHAPTER TWO

2.1 Consumable High-Density Microfluidic PDMS Device

As a proof of principle for the integration of Starlight chemistry on a high-throughput platform, we fabricated a custom microfluidic device that has the capability of monitoring thirty-two individually addressable sequencing sites. To enable high-throughput sequencing on-chip, isolation of individual reaction sites is paramount. Once a dNTP is incorporated onto the parent strand, the DNA polymerase continues to incorporate additional nucleotides until it reaches the end of the template. Unlike current high-throughput amplified sequencing techniques like Illumina's sequencing by synthesis, incorporated nucleotides are not end terminated and retain their 3' OH group. This presents a significant challenge; namely, any DNA template exposed to nucleotide solution will automatically begin DNA synthesis even when not under direct observation. Due to the narrow field-of-view of our microscope setup, we can only directly image up to a 100um x 100um area of which there can be 1500 individually observable short-template strands. In order to approach whole-genome sequencing on one or several consumable devices, it is clear that we need to create isolated sequencing sites.

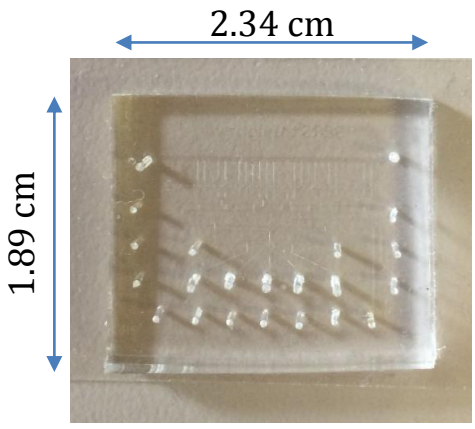

Figure 8: Photograph of our PDMS device.

Our solution is a high site density PDMS microfluidic device that utilizes pneumatically actuated valves (Figure 9) to sequentially address individually isolated testing sites. Fabrication of these devices utilizes a two-mask soft-lithographic process. In AutoCAD, we created mask schematics for two different layers, the flow layer, which provides the path for fluid, and the control layer, which pushes down upon the flow layer to pinch off fluid flow. By utilizing a 10 to 1 width to height design parameter, we are able to ensure that the push-down valve completely seals off any fluid flow, as demonstrated by the Quake group[26].

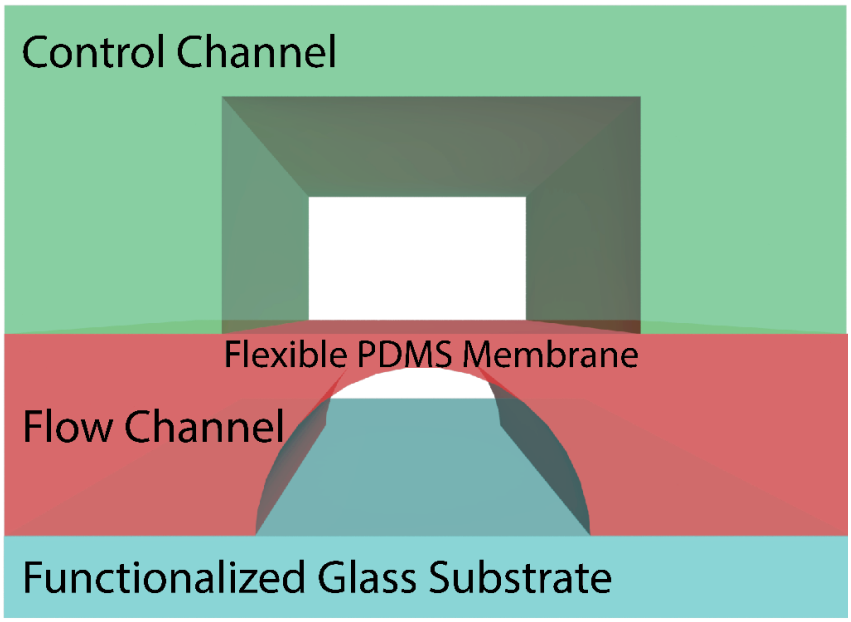

Figure 9: Diagram of push-down PDMS valves for isolation. Flow channel is where reagents, including DNA and nucleotides, are introduced. The control channel is primed with deionized water. When the control channel is toggled pneumatically, the PDMS membrane pinches down and shuts the flow channel.

Silicon masks were fabricated using standard lithographic process and the wafer mold of the flow layer was reflowed to create sloped walls for the flow channel. PDMS is poured onto the surface of the mold and lifted off for each layer. The two PDMS layers are bound together using oxygen plasma after they are aligned. The finished multi-layer

PDMS device is then subsequently bonded onto a functionalized glass coverslip by baking. Typical PDMS-to-glass bonding involves oxygen-plasma bonding to strengthen adhesion onto the surface of the PDMS[27]. In our device, however, plasma bonding would render the polyethylene-glycol to biotin surface unusable. We tested several methods of binding onto the glass coverslip and found that heat curing the device is the quickest and most efficient way of providing good PDMS-to-glass adhesion. The maximum amount of rated input pressure the device control channels can sustain utilizing this method is 20 psi. At pressures above this point, we observed liftoff of the PDMS from the glass surface, resulting in rapid PDMS to glass delamination. Any delamination in a port will expand in size with use and render the entire device unusable. A 3DS Max rendering of the device is shown in Figure 10.

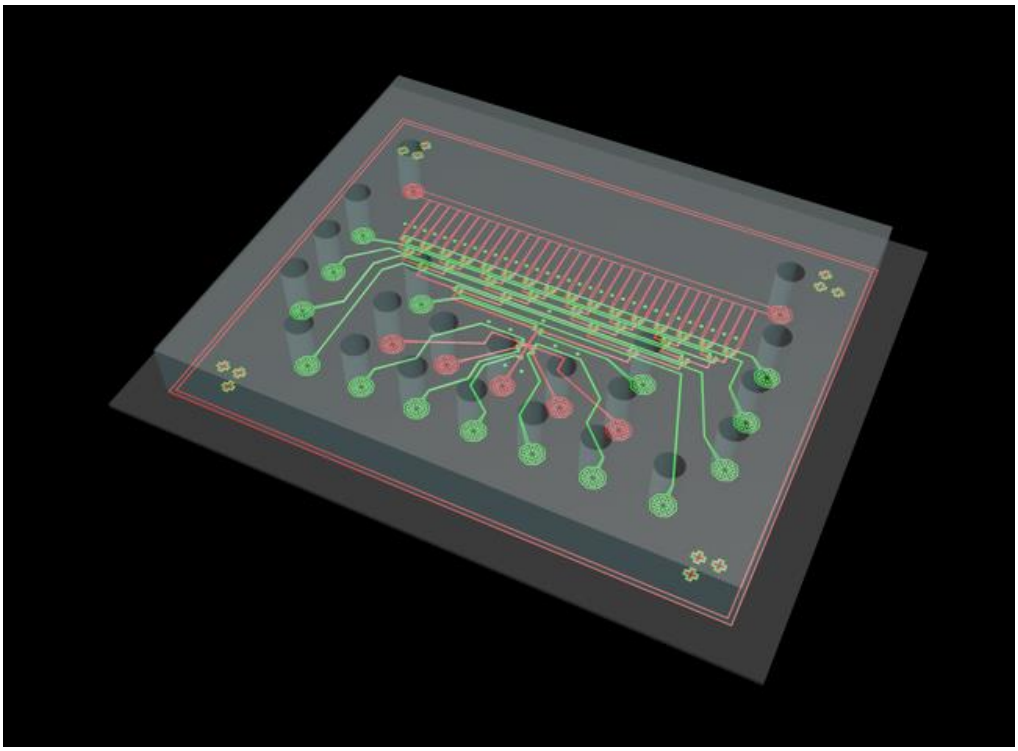

Figure 10: 3DS Max rendering of microfluidic PDMS device. Green channels represent the pneumatic control valves and the red channels represent the reagent flow channels. Thirty-two individual reaction sites are integrated on this chip iteration.

The design of the flow layer itself took into consideration two criteria: the need for isolated and selectable testing sites and potential for scalability. In order to solve the first these aims, we utilized a combination of binary tree design and a flow-based isolation method based on hydrodynamic focusing we call “purging flow” (Figure 11). To individually select sites we utilized pairs of valves in a binary tree format as individual bits. Thus, the number of control valve pairs increase linearly while the number of individually addressable sites scales exponentially. However, individual sites must be at least 100µm x 100µm in dimension and be separated by at least 200µm to allow for efficient flow layer access. Additionally, the channels and system were deliberately designed to ensure reagent consumption efficiency is high and the entire flow volume is measured in nanoliters.

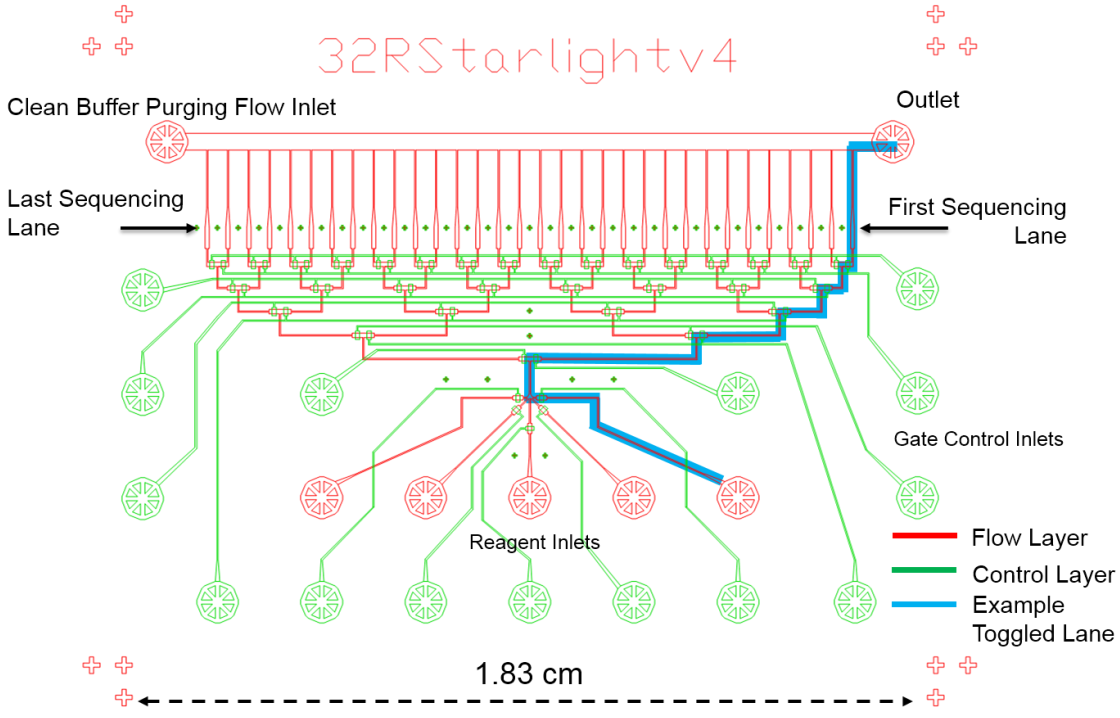

Figure 11: AutoCAD schematic of chip design and purging flow. Individual testing sites are located just before the outlet ports. Flow lanes are shown in red and control lanes are shown in green.

2.2 Hydrodynamic Focusing for Nano-Sized Particle Isolation

Simple methods to guarantee against backflow include incorporating a parallel set of control valves, individual outlets, or extremely long outlet ports. Each of these methods decreases the effective testing site density in the device. We devised a flow-based system based on hydrodynamic focusing[28] that wicks away any particulates from contaminating subsequent testing sites. Inspired by hydrodynamic flow confinement in high-throughput FACS machines[29,30], our "purging flow" system creates a high-velocity transverse flow to minimize backflow into currently unobserved and un-sequenced areas of the chip. To examine such interactions at the micron-level we utilized COMSOL to model compressible flow in non-slip conditions for a T-junction interface (Figure 12) where the baseline COMSOL Navier-Stokes equation is below.

$$\rho\left(\frac{du}{dt} + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\right) = -\nabla \mathrm{p} + \nabla\cdot\left(\mu\left(\nabla\mathbf{u} + (\nabla\mathbf{u})\right)^{T} - \frac{2}{3}\mu(\nabla\cdot\mathbf{u})\mathbf{I}\right) + \mathbf{F} \qquad (7)$$

Where u is fluid velocity, p is fluid pressure, ρ is the fluid density, μ is the fluid viscosity, and F accounts for external forces applied to the fluid. This is solved in conjunction with the continuity equation.

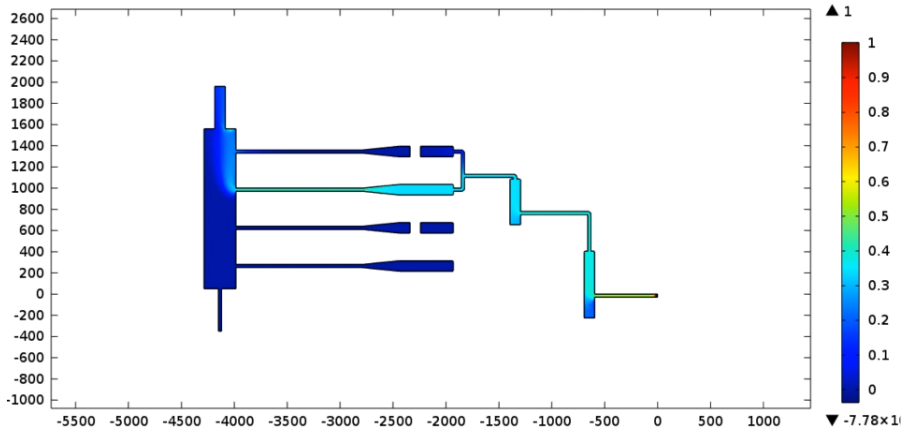

Figure 12: Simulation of T-junction purging flow. 100nm particulates are introduced and monitored over time for contamination in subsequent testing sites

$$\frac{dp}{dt} + \nabla \cdot (\rho \boldsymbol{u}) = 0 \qquad (8)$$

COMSOL modeling software was utilized to simulate isolation effectiveness in a complex volume. The layout was created in AutoCAD and integrated into the simulation system. On-chip testing on a non-biotin functionalized surface utilizing micro-molar concentrations of streptavidin labeled with Alexa Fluor 555 was completed in order to capture real-time data on the device isolation functionality (Figure 13).

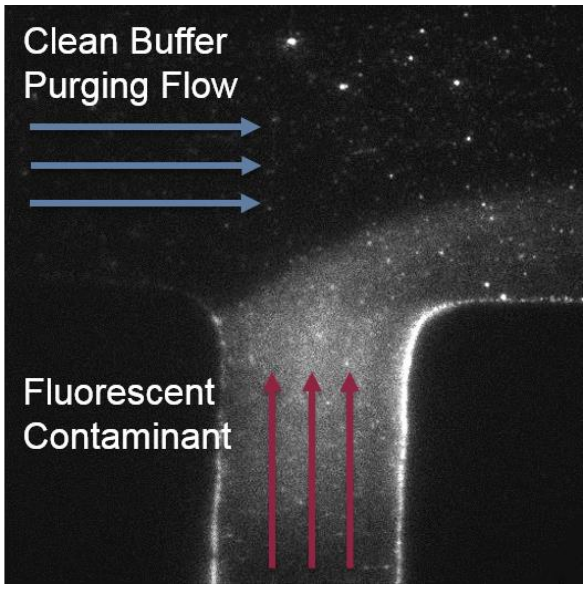

Figure 13: On-chip demonstration of purging flow utilizing high concentration of streptavidin labeled with Alexa Fluor 555. High-velocity transverse flow limits particle interaction at the interface and prevents backflow of contaminants into subsequent "clean" testing sites

However, in a real sequencing scenario, we introduce nucleotides in a much smaller concentration and in order to test the validity of our system in long timeframes we introduced 400 pico-molars of adenosine nucleotide tagged with Alexa Fluor 555 into a chip with a biotin-functionalized surface. Contaminants would thus not only have a chance to bind onto the surface, but also float in the channels, contributing to background noise. We took average observed fluorescent intensity measurements over a period of thirty minutes for three individual testing sites in succession (Figure 14). As predicted,

diffusion has a significant impact on site contamination over minute time scales while when purging flow was toggled on, observed increase in intensity over time is minimized. This result demonstrates our device's capability to isolate multiple testing sites, enabling hands-off sequencing tests.

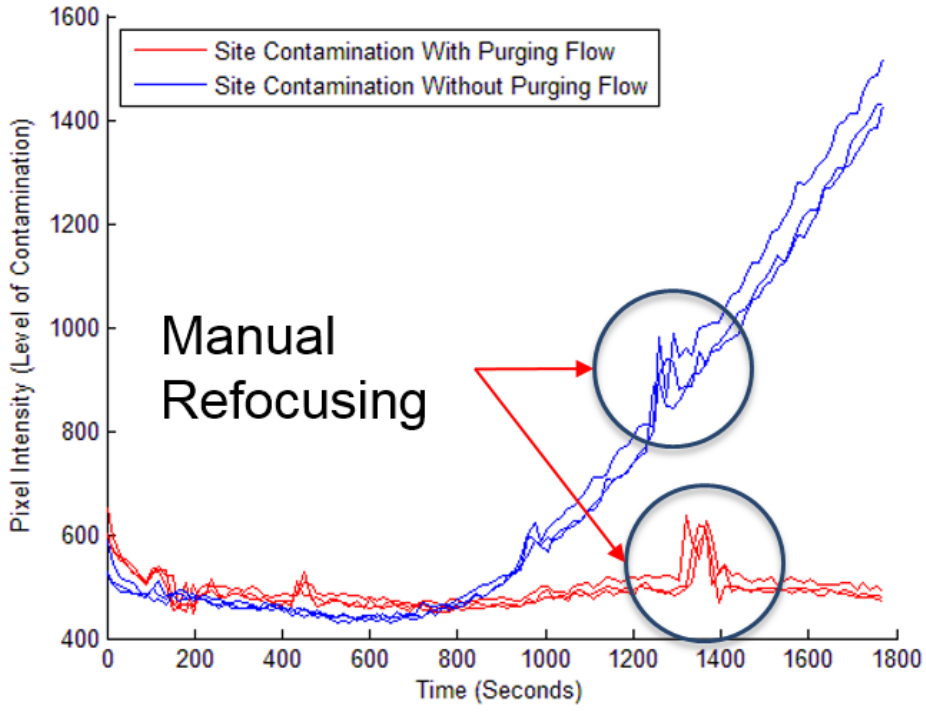

Figure 14: Comparison of contamination when purging flow is off/on. Three individual sites were tested. Overall dip in intensity from 0 to 800 seconds can be explained by photobleaching of background contaminants

2.3 Device Control and Operation

Sequential toggling of target sites while simultaneously limiting reagent flow to an observable area requires pneumatic and piezo-electric control. In order to precisely actuate our control valves we utilize Festo solenoid valves attached to a programmable WAGO fieldbus controller (Figure 15). A combination of C packages and MATLAB output the binary step-wise outputs we need to create our binary tree fluidic isolation

system on our device. Baseline MATLAB interface package with the WAGO system was written by Rafael Gómez-Sjöberg from Lawrence Berkeley National Laboratory's Microfluidics Lab. The solenoid valves are attached to a regulator fed by house air and feed 20 psi of pressure into the control lanes when toggled. We utilize a piezoelectric stage to precisely move from site to site during the sequencing process and the locations are stored so that each individual run (lasting for more than an hour) can be completed in an automated fashion utilizing a custom program (Figure 16).

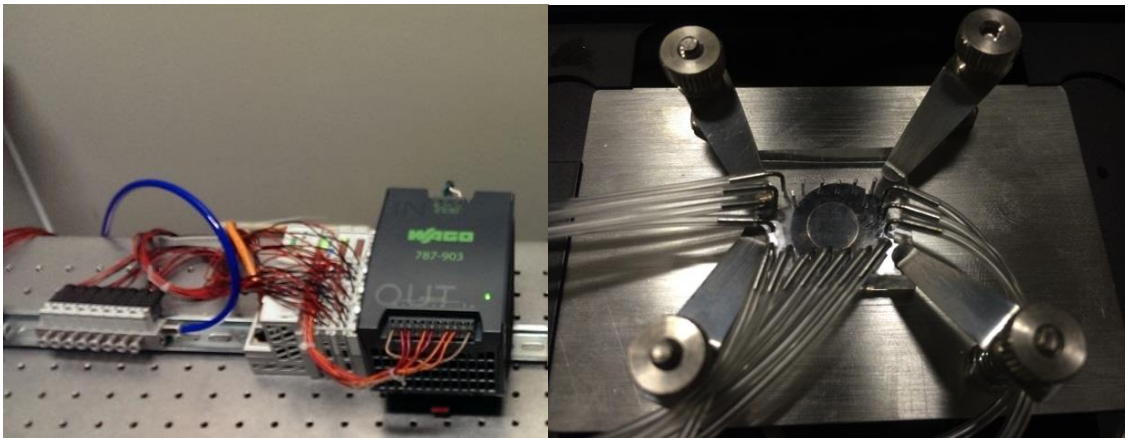

Figure 15: Overview of control elements. (a) Computer controlled pneumatic system including Festo solenoid valves, WAGO Fieldbus Controller and Power supplies. (b) L-shaped metal connectors are inserted into control layer access-holes punched into the PDMS.

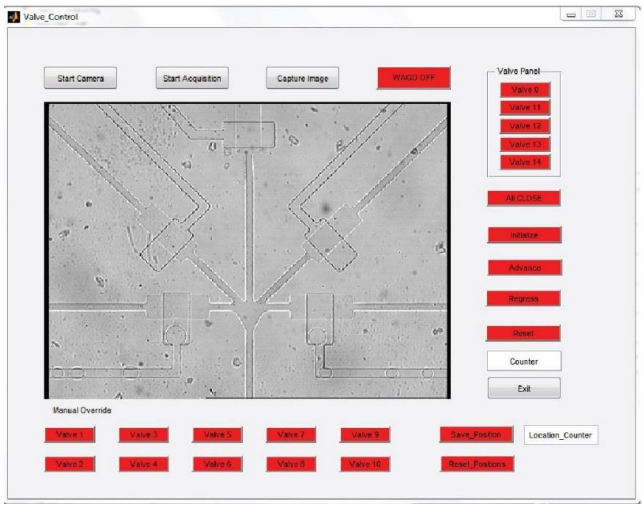

Figure 16: MATLAB automation allows for hands-off reagent delivery, sequencing, and data acquisition.

In order to prevent the PDMS-Glass or PDMS-PDMS interfaces from delaminating, we designed a custom pressure-control system for priming our control valves and loading our reagents into the flow channels. Each control valve must be pre-primed with water or oil in order to guarantee the absence of air transfer from the control layer to flow layer. PDMS is a porous material and because the valve barrier is extremely thin (several hundred microns), air is semi-permeable through the membrane, causing the creation of micro-bubbles in our reagent channels. Because of our device's pressure limitations, stemming from the non-plasma bonding properties, we cannot use traditional syringe pumps to load our control channels. Instead, our high-throughput priming system couples individual actionable Festo pneumatic valves with a pressure regulator and several 2 ml test-tubes.

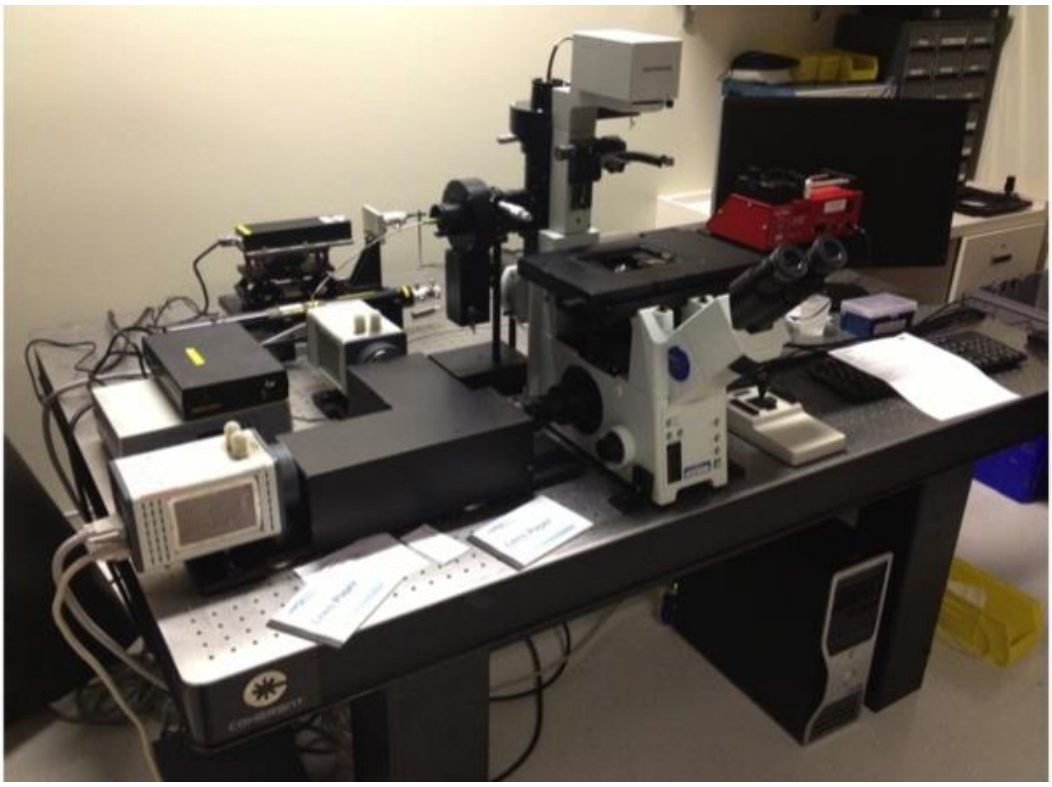

Figure 17: Photograph of integrated Starlight sequencing and control setup.

CHAPTER 3

3.1 Characterizing Data Output

To parse the data received by the EMCCD and perform correct sequencing base calls requires robust computer vision and signal analysis tools. For ease of integration, our entire post-processing pipeline is scripted in MATLAB. TIFF video outputs are aligned and interpreted to create hundreds of individual strand-specific time series traces. These traces can then be interpreted to produce base-calls and collated to produce a sequencing read.

3.1.1 Raw Output Signal Processing

As mentioned in Section 1.2.1, the overlapping emission spectra of our acceptor fluorophores necessitates a base-call to be determined from the combination of signals from three channels. The spatial location of an individual DNA template is fixed and marked by the fluorescent emission signature emanating from the Alexa Fluor 555. Because our signal is split utilizing free-space optics, each channel needs to be spatially aligned to the reference signal (Figure 18). We utilized the MATLAB Computer Vision System Toolbox to align relevant sequencing sites utilizing the Blob Analysis algorithm (Figure 19). The algorithm analyzes high intensity features in the field, which corresponds to a single tethered DNA parent strand and accompanying polymerase, and distinguishes it from the background. Once we have a feature map for each of the three channels, we compare each individual channel to the reference channel utilizing a Geometric Transform Estimator function to predict spatial misalignment and scaling.

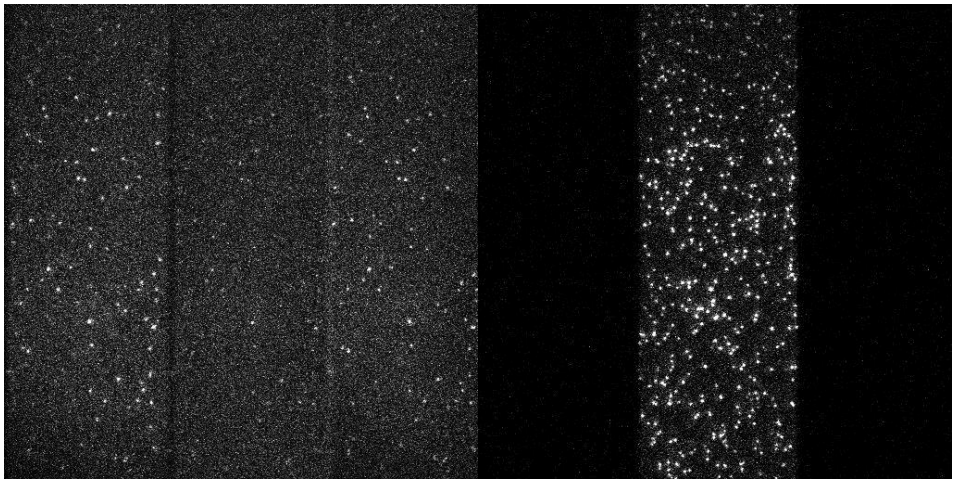

Figure 18: Starlight sequencing raw data. The left three channels are comprise of signals from the following spectra: 645-675nm, 705LP, and 675-705nm. The right channel is the AF555 reference signal.

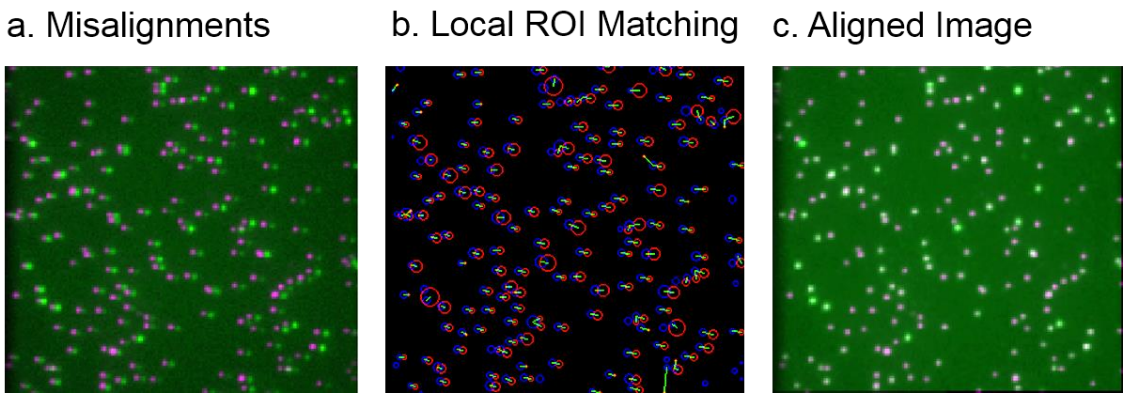

Figure 19: Channel alignment to reference signal. (a) Misalignments caused by local and global distortions are apparent. (b) Region of Interest matching using Blob Analysis is performed. (c) Channel is geometrically transformed to match reference signal.

Once we successfully aligned the output of our three channels to the reference, we gather the three time-series intensity traces, collate them, and compare the time-series data to determine our sequencing base calls (Figure 20). To insure that each trace has a functioning polymerase, a mean intensity integration over the first two-hundred frames recognizes sequencing sites that have non-functional polymerases. Once these regions are identified, their traces are not processed, saving computational power.

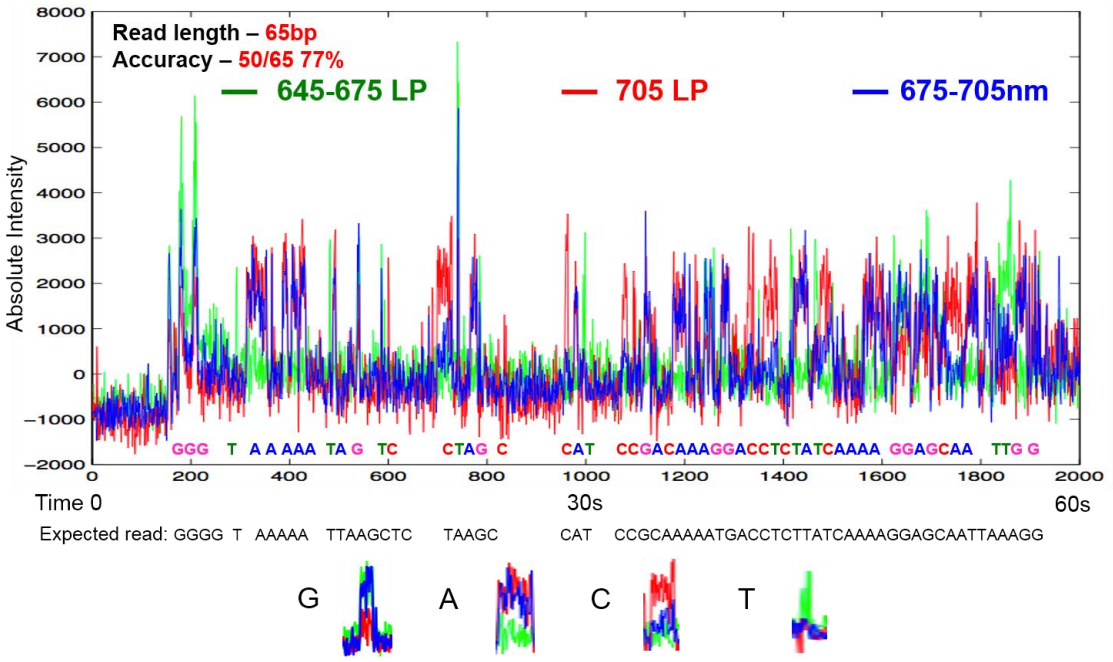

Figure 20: Example output trace from a sequencing read. Each peak represents a single base and the combination of three signals, green, red, and blue, indicate the specific nucleotide.

3.1.2 Manual Read-Length and Accuracy Distributions.

Read-length distribution and sequencing accuracy of the Starlight system sequencing chemistry was determined utilizing a known oligo 64 base-pairs in length with the following sequence: CCTTTAATTGCTCCTTTTGATAAGAGGTCATT TTTGCGGATGGCTTAGAGCTTAATTTTTACCCCTTCCTCCTCCTCCGTTTCGTG TAGGGAAAGAGTGT/3bio/. After acquiring the traces from individual polymerases, we polled 387 active sequencing sites (1 base call or above) (Figure 20). By comparing the known sequence to the traces, we were able to determine the read-length distribution for the 65 bp oligo. The average read length was found to be 17.44 bases. Of that subset of data, we looked in-depth at sequencing reads above 15 base-pairs (Figure 21) and found a mean single-base sequencing accuracy of 77.94% by comparing manual traces to the known oligo sequence with a Smith-Waterman alignment algorithm[31] (Figure 22).

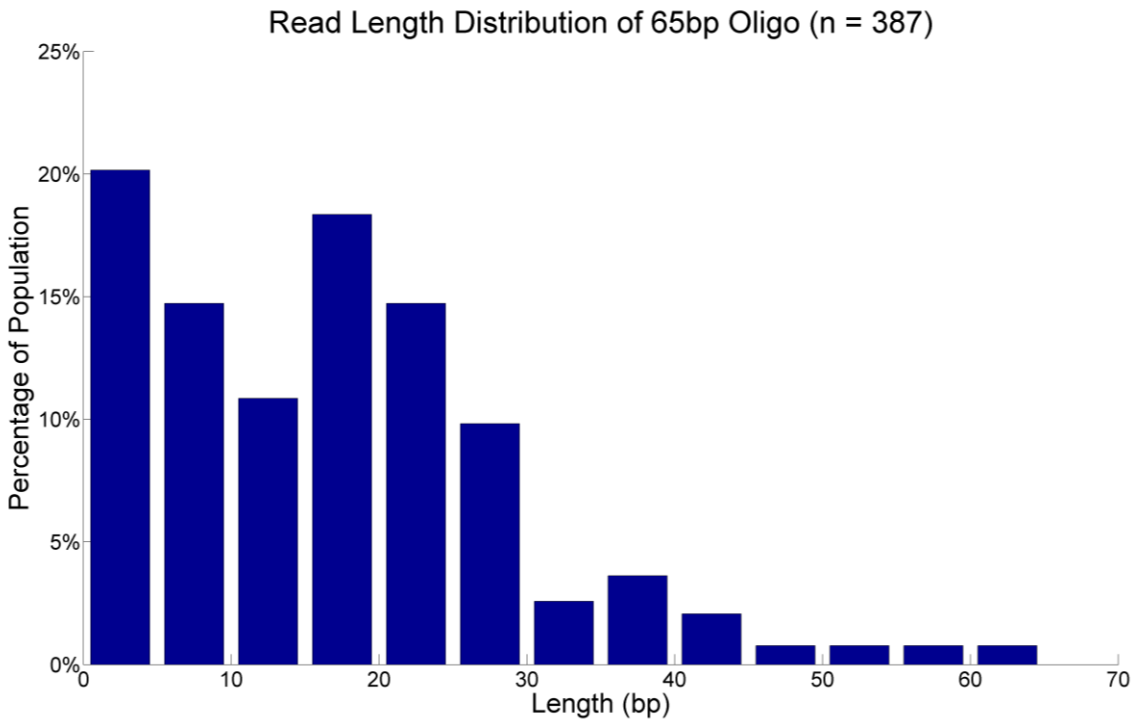

Figure 20: Histogram of read length distribution for traces with bp>0. Average read length is 17.44 bases.

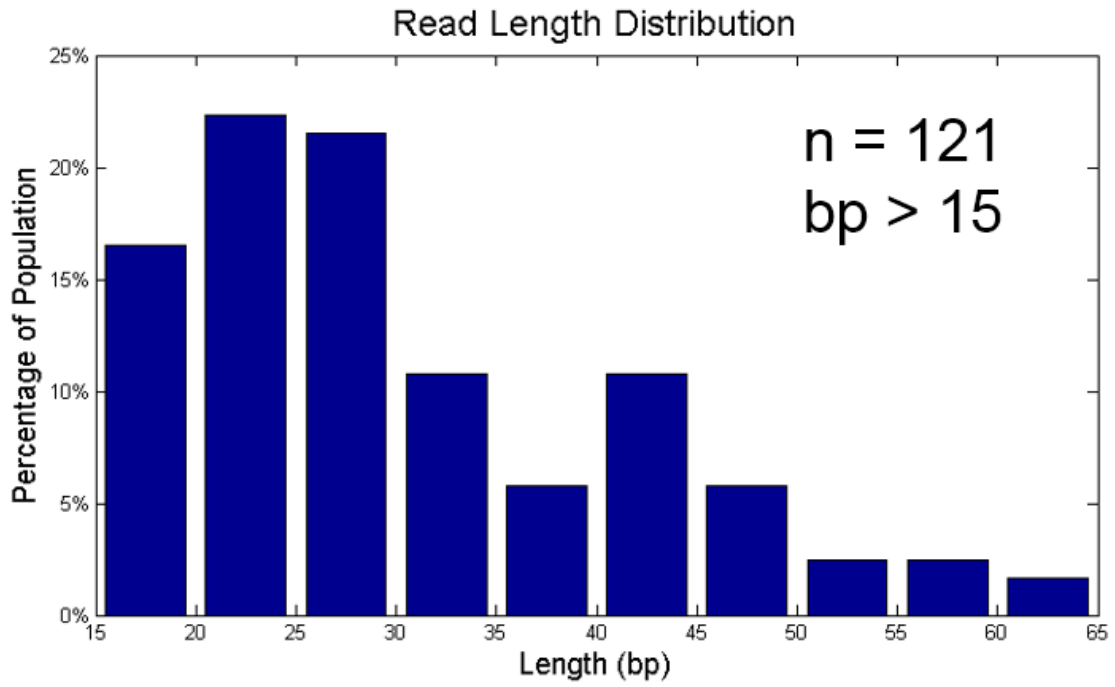

Figure 21: Histogram of read length distribution for traces with bp>15.

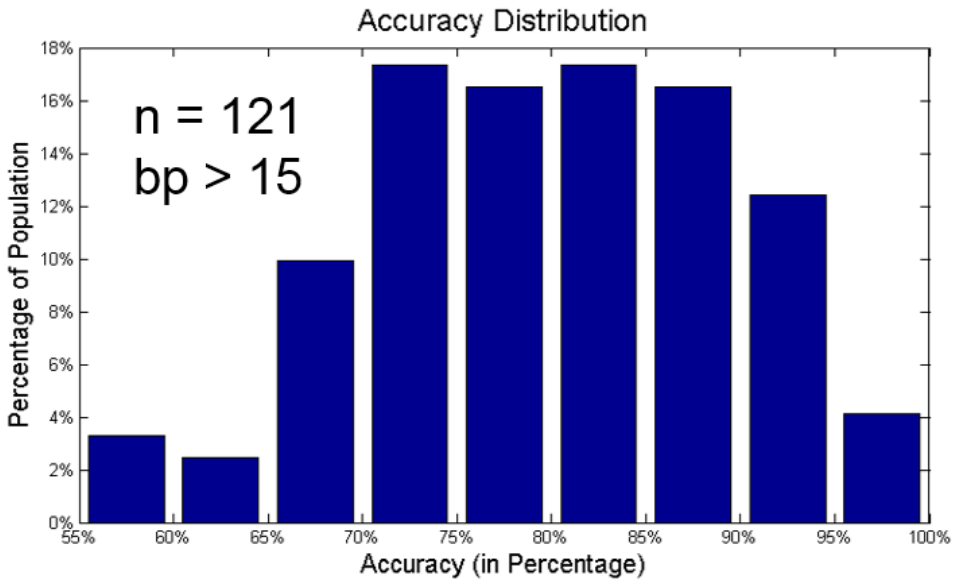

Figure 22: Histogram of single-base accuracy with bp>15. Mean single-base accuracy was 77.94%.

3.2 Automated Base-Calling Algorithms

As manually curating sequence reads is ultimately untenable, we are developing algorithms to clear up the high frequency background noise which is pervasive in our reads. To achieve this we individually look at the absolute intensity histograms for all three channels an individual trace and fit the histogram to a bimodal Gaussian distribution (Figure 23). To create these intensity histograms, we poll the absolute intensity at each point in our time series output trace. The bimodal characteristic of the intensity profile indicates the normalized distribution of the noise and of our signal. We can also utilize the bimodal Gaussian distribution to estimate the SNR of our traces by utilizing the mean of the signal and noise curves. In testing the traces of 6 sequencing sites, we found that the average SNR was 3.74dB, the high was 6.3dB the low was 2.32dB. We noticed, however, that sites with lower SNR had much fewer active traces and thus in the future the SNR estimate may be utilized as a site-to-site quality control mechanism. Code to determine SNR is included in the linked supplementary repository.

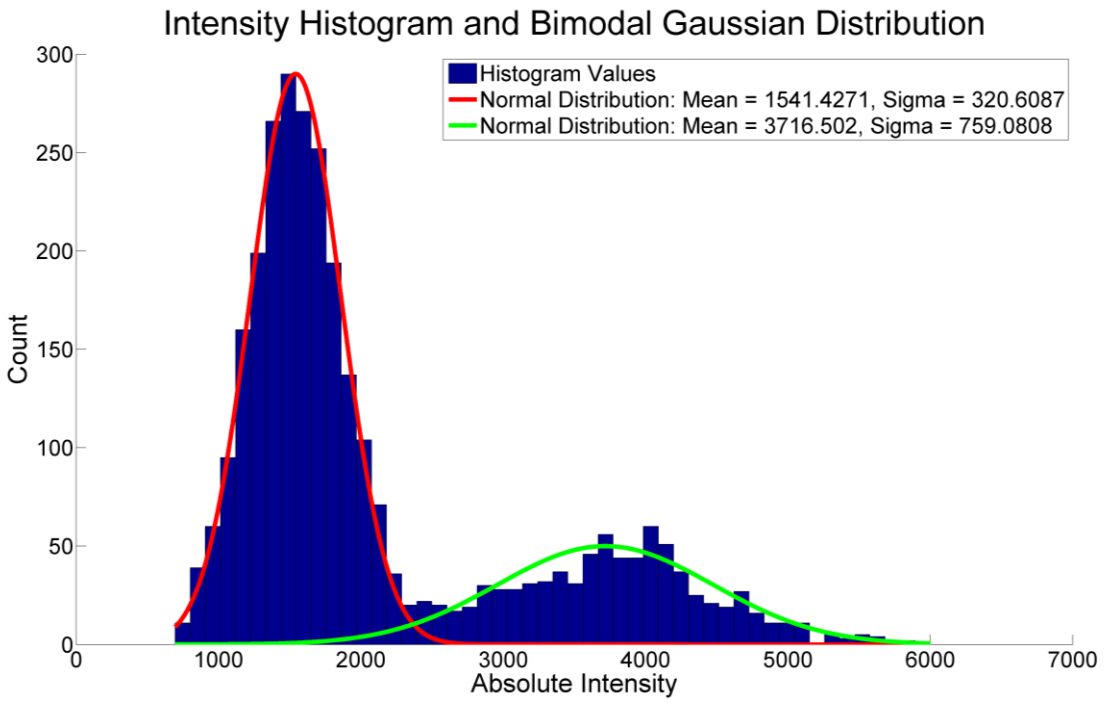

Figure 23: Demonstration of Bimodal Gaussian curve fitting. Utilized to determine and suppress noise distribution.

Once the noise mean and standard deviation are determined for each channel, we create a table of p-values utilizing a normalized probability density function. A Benjamini Hochberg[32, 33] False Discovery Rate (FDR) method then analyzes the signal with regard to the p-value table created and identifies and highlights true signals while suppressing the background noise (Figure 24). This technique is particularly powerful because it also normalizes variation in overall signal intensity between sequencing traces, simplifying subsequent thresholding-based base-calling determination algorithm reads.

We are currently investigating a number of methods to utilize our cleaned signal to provide accurate base-calling. Ideal ratios of thresholding for a particular sequencing site can be determined through iterative recursion of returned sequencing scores from a local alignment algorithm by Smith-Waterman[31] that we call from the MATLAB

Bioinformatics Toolbox. Recognition of individual base-pairs are heavily dependent on observed intensity ratios and thus is susceptible to non-uniform background variations. Thus, each dataset is also pre-processed utilizing a rolling ball radius normalization in ImageJ. Our current algorithms are limited to iterative thresholding approaches; however, machine learning and Hidden Markov Model based algorithms are currently in development. An example output dataset of our automated threshold-based base-calling algorithm is demonstrated in Table 1. An example program will be uploaded onto the code repository listed in the supplementary.

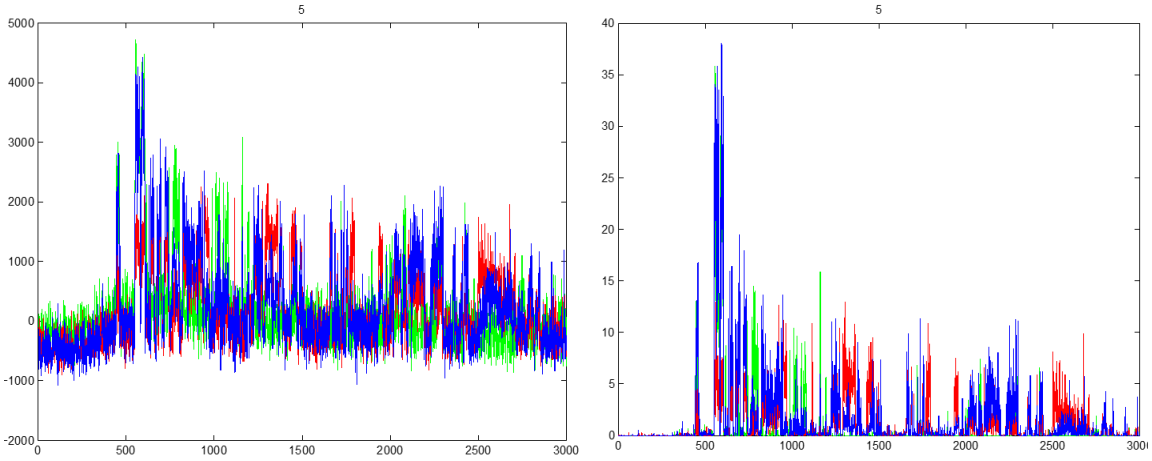

Figure 24: Comparison of data before and after Benjamini Hochberg FDR processing.

Table 1: Example processing of automated base-calling outputs with Benjamini Hochberg FDR pre-processing and thresholding approach. Smith-Waterman alignment is used to determine single-base accuracy

Output from threshold-based algorithm	**Smith-Waterman Processing**		**Accuracy**
23bp Read: GGGGGTAAAAATTAA GATCAGCTT	Reference: Matches: Output:	GGGGTAAAAATTAAGCTCTAAGC |||||||||||||||| || ||| GGGGTAAAAATTAAGATC--AGC	20/23 = 87%
26bp Read: GGGGGAAAAATAGCC CCCCCCCCCCC	Reference: Matches: Output:	GGGGTAAAAATTAAGCTCTAAGCCATCC |||| ||||| | ||| | || || GGGGGAAAAA-T-AGCCCCCCCCCCCCC	18/26 = 69%

Chapter 4

4.1 Discussion

Our demonstration of a short template DNA sequencing provides the basis for the continued development of the Starlight sequencing system. Continued modifications to protocols, methods, and devices will take aim at the emerging needs of next-generation sequencing platforms including: long-template sequencing, Giga-base yields, and detection of DNA base-pair modifications. The following sections document our plans towards solving the various challenges that still remain.

4.1.1 Increasing Read-Length and Long Template Sequencing

The advantages of our FRET-based sequencing system is theoretical capability of a polymerase that sequence thousands or even tens of thousands of base-pairs. Current derivatives of phi-29 polymerase are known to sequence up to 15kb without degradation and directed evolution of such enzymes utilizing the Kapa Biosystems platform promises for even longer read lengths. Our enzyme was primarily evolved to increase incorporation time and not for resistance to photodamage. The latest unpublished data for Starlight polymerases indicate an increase of accuracy from ~78% to ~85% and an increase in average read-length to ~ 200 base pairs. In anticipation of continued advances in polymerase evolution, we designed a microfluidic device that utilizes high-velocity fluid flow to stretch long-template DNA (Figure 25). Since we utilize a TIRF setup to excite our donor fluorophore, as mentioned in Chapter 1, it is necessary to ensure that the polymerase remains within 100nm of the glass substrate as the evanescent wave energy decreases as a function of distance, as mentioned in Equation 5. By stretching out the DNA in the x-y plane, we keep the polymerase within range for excitation.

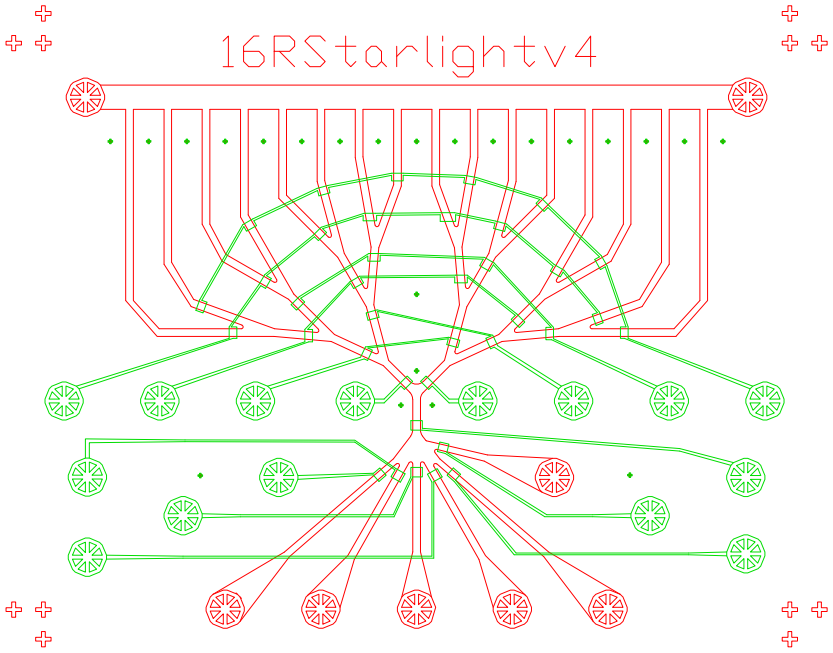

Figure 25: Design schematic for flow-based long-template sequencing device. Minimizes turbulent fluid interaction by reducing right angle corners.

This new design that optimizes the flow velocity by minimizing sharp corners and designing uninterrupted widths for the flow channel, giving a device density tradeoff. Through experimentation and scaling, we discovered we needed to achieve at least a two-hundred centimeter per second flow rate to fluidically stretch our lambda phage DNA template (Figure 26).

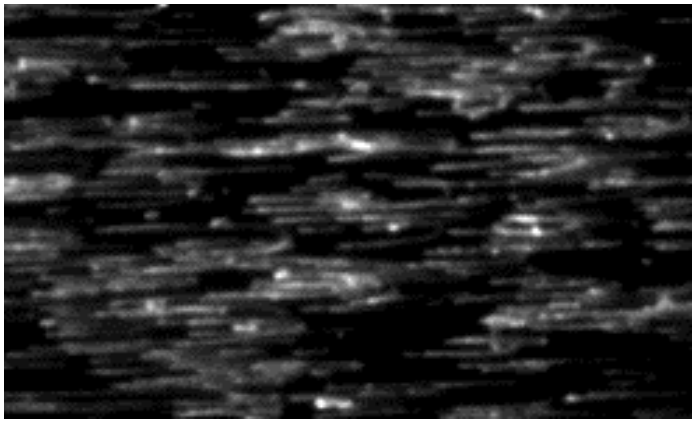

Figure 26: Photograph of long-template stretching tests on large sequencing device.

During our stretching tests we observed that a majority of templates were looped onto the surface (Figure 27). In order to resolve this issue, we propose to incorporate our parent template biotin with a fifty percent probability random with the 3’ and 5’ poly-A tail. Thus, there is a 25% chance of no biotin ends, a 50% chance of one-end, and a 25% chance of both ends, resulting in two-thirds of attached DNA on the substrate being Brief flow stretching would be initiated after free floating DNA ends are ligated with a biotin, potentially anchoring the DNA onto the surface.

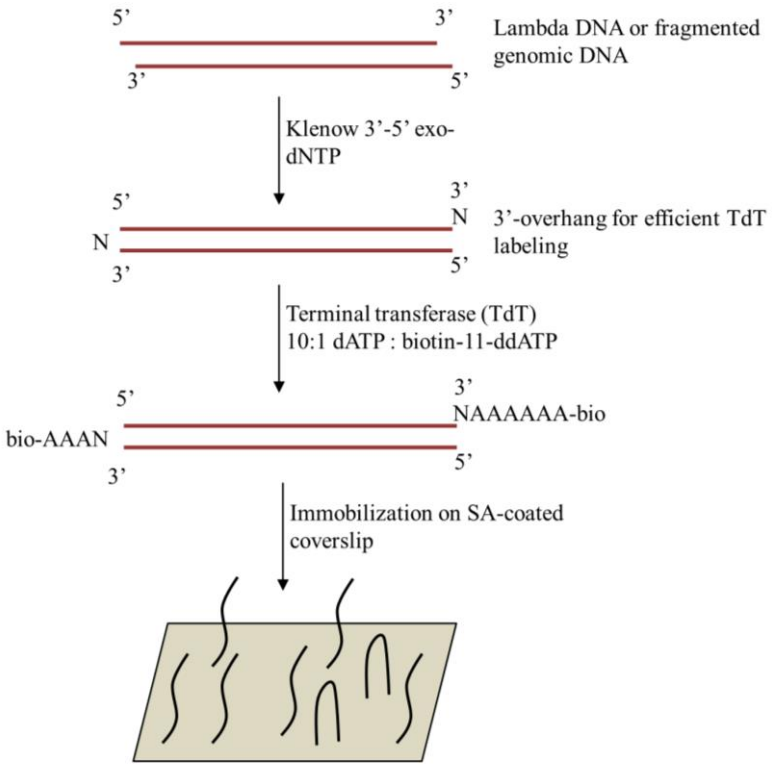

Figure 27: Proposed protocol for loop-less long-template stretching.

The on-template characteristics of this sequencer also open the possibility of parallel short-read sequencing using hundreds of polymerases attached onto the surface of a long template (Figure 28). This scheme would utilize restriction nickase[34] to create a gap in the daughter strand of a DNA template.

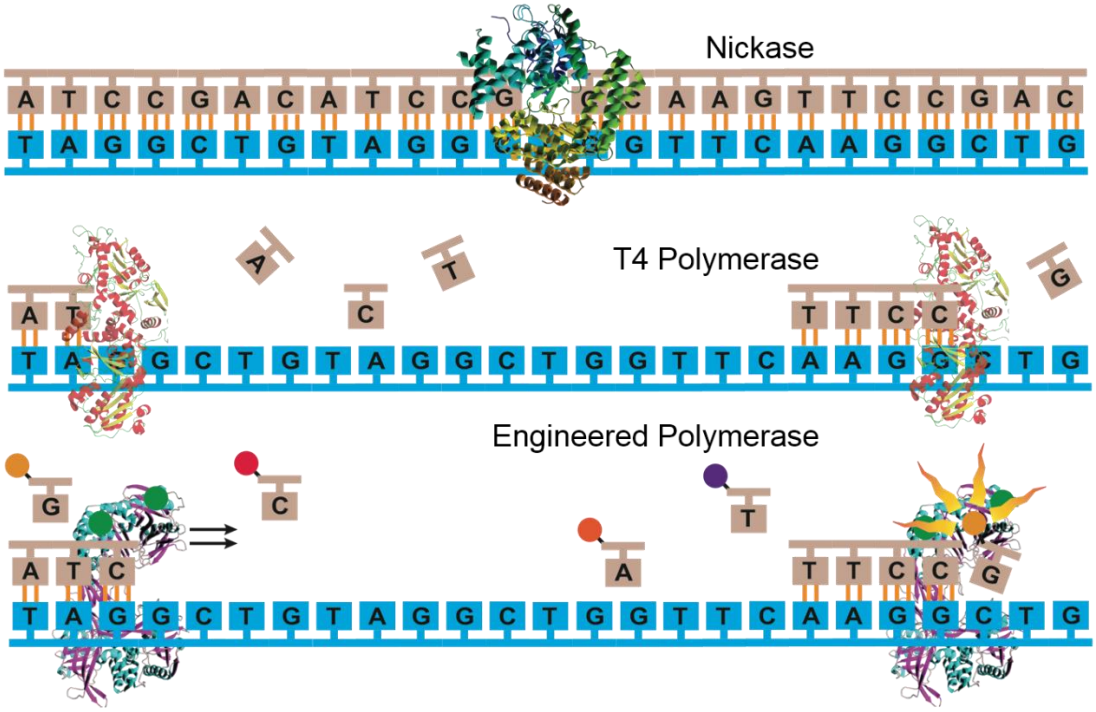

Figure 28: Method to integrate hundreds of short read-length sequencers onto a long-template for parallel sequencing. This procedure could also open up the window for resequencing of the same strand increasing single-strand coverage.

For a proof-of-concept long-template sequencing of lambda phage DNA, we will utilize a combination of Nicking Endonucleases Nt.A1wI[35], Nt.BsmAI[36], and Nt.BstNBI[37] which recognize the 5 bp DNA sequences GGATC, GTCTC, and GAGTC respectively. By utilizing these nickases, cuts in a single will be created at target sites strand (known locations when mapped back to the reference genome) in double stranded lambda phage DNA. T4 DNA polymerase[38] will then be introduced with 3' to 5' endonuclease activity to selectively cleave off the daughter strand under specific time and temperature restraints. Our mutant B103 DNA polymerases will then be introduced into the medium where FRET sequencing could occur. The double-stranded DNA will be stretched and anchored onto the substrate surface, thus enabling sequencing advancement in the x-y plane, rather than in the z-plane (as demonstrated in short-read sequencing). In

this fashion, hundreds of small stretches will be opened for simultaneous short-read sequencing on a long template. Furthermore, by introducing the nickase solution repeatedly, a single strand could be resequenced and thus provide a method to provide multiple resquencing runs without neccessitating amplification.

4.1.2 Maximizing On-Chip Sequencing Density

By utilizing this sequencing chemistry, we are able to create a device without need for on-substrate physical modifications or patterned surfaces. Instead, by limiting the concentration of the DNA needs to be analyzed, we are able to get enough localization of testing sites to resolve individual DNA templates. In a short-template two-camera setup we are able to resolve up to 500 anchored DNA strands of up to 65bp in length. Utilizing a four-camera setup we will be able to take full advantage of our microscopes field-of-view and increase the number of observable DNA strands to 1500 per sequencing site. Each 32-site device can theoretically sequence up to 3.12 million base-pairs (1500 * 65 * 32) and each site can be sequenced in two minutes. However, this estimate is below the theoretical limit for a sequencing setup utilizing the short read-length starlight system. Our PDMS chip form-factor is approximately 4.4 cm^2 and an appropriately scaled system, up to the size of flow cells commonly used in high-throughput sequencing like Illumina's HiSeq as a template, could accommodate up to four parallel lanes of one-hundred fifty testing sites each for a total of 600 sequencing sites. This brings the total amount of base pairs sequenced per 20 hour run to 58.5 million. Total sequencing throughput utilizing these devices in a short read-length sequencing approach would max out at 491.4 million base pairs sequenced per week.

However, with the integration of our proposed long-template sequencing protocol we would be able to image and entire field of stretched-out long templates in the same time it would sequence a field of short-templates. If the polymerases are adequately spaced, they will collectively sequence using all possible EMCCD real-estate in parallel. Preliminary estimates utilizing lambda phage DNA indicate that we may be able to fit up to 120 50kb DNAs per sequencing site, equivocating to 6 million readable bases. If we are able to implement the parallel short-sequencing method outlined in Section 4.1.1, we can read 6 million bases every two minutes. Utilizing our proposed scaled 600 sequencing site chip, we can effectively sequence up to 3.6Gb per run. A comparison of an optimized Starlight short-read system, proposed Starlight long-read system, and next generation sequencing (NGS) platforms[39, 40] is shown in Table 1.

4.1.3 DNA Methylation Tests

With the recent interest in epigenetic markers and mapping the epigenome, the next generation of sequencing technologies attempt to detect methyl group additions on the nucleotides themselves. The presence of methyl groups is exceedingly high in the human genome and is an important factor in determining gene expression. It has been demonstrated that single-molecule methods can detect the presence of methylation; however, the coverage needed to detect these perturbations is exceedingly high (up to 504x) to resolve individual methylated bases[11]. Detection of methyl bases utilizes the observation that dissociation timing of complementary methylated nucleotides exceeds that of non-methylated bases. However, the timing delay is not absolutely quantifiable and is determined by an additional number of unknown factors.

* Estimated from cost of PacBio RSI.
** Consumable PDMS devices are currently fabricated in institutional nanolab.

Table 2: Comparison of Starlight Sequencing with Current and Next Generation Sequencing Platforms

Method	Illumina HiSeq 2500	Life Technologies Ion Torrent PGM	PacBio RSII	Optimized Starlight Short-Read	Proposed Starlight Long-Read
Instrument Cost	$750 K	$80 K	$695 K*	$120 K	$120 K
Reported Accuracy	Q30	Q20	<Q10	<Q10	<Q10
Raw Error Rate	0.80%	1.71%	12.86%	22.06%	~15%
Run Time	12 days	4 h	2 days	20 hr	20 hr
Output per Run	600 Gb	1.5 – 2 Gb	5 Gb	58.5 Mb	3.6 Gb
Sequencing cost per Gb	$41	$1000	$2000*	$5000**	~$100
Average Read Length	2 x 100	200-400	50% > 10kb	20bp	200bp x 200
DNA Requirements	50-1000ng	100-1000 ng	~ 1ug	< 1ng	10 – 100 ng

In our methylation testing protocol, complimentary methyl cytosine is incorporated onto the sequenceable oligo. We utilized the following sequence: TTGATAAGAGGTCATTTTTGCGGATGGCTTAGAGCTTAATTTTT/i5HydMe-dC/GA/i5HydMedC/GAACCCCTTCCTCCTCCTCCGTTTCGTGTAGGGAAAGAGTGT/3bio/. Preliminary results may indicate the similar of polymerase kinetic delays as similar to those observed by Pacific Biosciences SMRT platform[11].

4.2 Conclusions

We have demonstrated a single-molecule FRET sequencing device and protocol for short-template DNA sequencing. A high-throughput microfluidic PDMS device was designed, fabricated, and tested. An automated sequencing setup, including piezoelectric stages and pneumatic controls augmented existing Starlight free-space optics and sequencing procedures was developed and implemented. Methods for on-chip long-template stretching and sequencing and DNA methylation studies were also discussed. Our platform allows for efficient and precise single molecule sequencing of short-template DNA, potential for sequencing long-template DNA, and will enable and accelerate discoveries in biological science.

4.3 Supplementary Materials

4.3.1 Sequencing Protocols

Regular Sequencing Loading Protocol

1. Wet all 8 sequencing lanes with 100uL 1X T/B purging buffer.
2. Introduce 45uL of 5nM SA for all sequencing lanes.
3. Static wait time of 15 minutes followed by 4 individual T/B buffer flushes for each lane. Two flushes for each orientation (up-down and down-up).
4. Introduce 100uL of 10pM Life Template for all of the lanes.
5. Static wait time of 15 minutes followed by 4 individual T/B buffer flushes for each lane. Two flushes for each orientation.
6. Introduce 45uL of 400-fold dilution fluorescently-labeled polymerase enzyme to all sequencing lanes.
7. Static wait time of 3 minutes followed by 4 individual O2-removing buffer flushes for each lane. Two flushes for each orientation.
8. Introduce fluorescently labeled dNTP reaction reagent to begin sequencing.

Methylated DNA Sequencing Loading Protocol

1. Wet all 8 sequencing lanes with 100uL 1X T/B purging buffer.
2. Introduce 45uL of 5nM SA for all sequencing lanes.
3. Static wait time of 15 minutes followed by 4 individual T/B buffer flushes for each lane. Two flushes for each orientation (up-down and down-up).
4. Introduce 100uL of 10pM Control template for half of the lanes and 100uL of 10pM Methylated template for the remaining sequencing lanes.

5. Static wait time of 15 minutes followed by 4 individual T/B buffer flushes for each lane. Two flushes for each orientation.
6. Introduce 45uL of 400-fold dilution fluorescently-labeled polymerase enzyme to all sequencing lanes.
7. Static wait time of 3 minutes followed by 4 individual O2-removing buffer flushes for each lane. Two flushes for each orientation.
8. Introduce fluorescently labeled dNTP reaction reagent to begin sequencing.

4.3.2 MATLAB Code and Sequencing Video Repository Link

https://sourceforge.net/p/salkstarlightcoderepo/

Please read navigation.txt file for file definitions. Large raw TIFF files will be provided through sftp or ssh by request. Email jyhuang@g.harvard.edu for formal file requests.

4.3.3 Journal Content Notice

Chapters 1 – 3, in part, are currently being prepared for submission for publication of the material. Huang, Jeremy; Huang, Jeremy Y. The thesis author was the primary investigator and author of this material.

REFERENCES

1 Mardis, Elaine R. "The impact of next-generation sequencing technology on genetics." *Trends in genetics* 24.3 (2008): 133-141.

2 Schuster, Stephan C. "Next-generation sequencing transforms today's biology." *Nature methods* 5.1 (2008): 16-18.

3 Ansorge, Wilhelm J. "Next-generation DNA sequencing techniques." *New biotechnology* 25.4 (2009): 195-203.

4 Wang, Zhong, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics." *Nature Reviews Genetics* 10.1 (2009): 57-63.

5 Lister, Ryan, Ronan C. O'Malley, Julian Tonti-Filippini, Brian D. Gregory, Charles C. Berry, A. Harvey Millar, and Joseph R. Ecker. "Highly integrated single-base resolution maps of the epigenome in Arabidopsis." *Cell* 133.3 (2008): 523-536.

6 Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold. "Genome-wide mapping of in vivo protein-DNA interactions." *Science* 316.5830 (2007): 1497-1502.

7 Alkan, Can, Saba Sajjadian, and Evan E. Eichler. "Limitations of next-generation genome sequence assembly." *Nature methods* 8.1 (2011): 61-65.

8 Alkan, Can, Bradley P. Coe, and Evan E. Eichler. "Genome structural variation discovery and genotyping." *Nature Reviews Genetics* 12.5 (2011): 363-376.

9 Aird, Daniel, et al. "Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries." *Genome Biol* 12.2 (2011): R18.

10 Whiteford, Nava, Tom Skelly, Christina Curtis, Matt E. Ritchie, Andrea Löhr, Alexander Wait Zaranek, Irina Abnizova, and Clive Brown, "Swift: primary data analysis for the Illumina Solexa sequencing platform." *Bioinformatics* 25.17 (2009): 2194-2199.

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

12 Branton, Daniel, David W. Deamer, Andre Marziali, Hagan Bayley, Steven A. Benner, Thomas Butler, Massimiliano Di Ventra, Slaven Garaj, Andrew Hibbs, Xiaohua Huang, Stevan B Jovanovich, Predrag S Krstic, Stuart Lindsay, Xinsheng Sean Ling, Carlos H Mastrangelo, Amit Meller, John S Oliver, Yuriy V Pershin, J

Michael Ramsey, Robert Riehn, Gautam V Soni, Vincent Tabard-Cossa, Meni Wanunu, Matthew Wiggin, and Jeffery A Schloss. "The potential and challenges of nanopore sequencing." *Nature biotechnology* 26.10 (2008): 1146-1153.

13 Rhee, Minsoung, and Mark A. Burns. "Nanopore sequencing technology: research trends and applications." *Trends in biotechnology* 24.12 (2006): 580-586.

14 Levene, Michael J., Jonas Korlach, Stephen W. Turner, Mathieu Foquet, Harold G. Craighead, and Watt W. Webb. "Zero-mode waveguides for single-molecule analysis at high concentrations." *Science* 299.5607 (2003): 682-686.

15 Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex deWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, Stephen Turner. "Real-time DNA sequencing from single polymerase molecules." *Science* 323.5910 (2009): 133-138.

16 Förster, Th. "Zwischenmolekulare energiewanderung und fluoreszenz." *Annalen der physik* 437.1-2 (1948): 55-75.

17 Chirio-Lebrun, Maria-Chantal, and Michel Prats. "Fluorescence resonance energy transfer (FRET): theory and experiments." *Biochemical Education* 26.4 (1998): 320-323.

18 Morgner, Frank, Daniel Geißler, Stefan Stufler, Nathaniel G. Butlin, Hans-Gerd Löhmannsröben, and Niko Hildebrandt. "A Quantum-Dot-Based Molecular Ruler for Multiplexed Optical Analysis." *Angewandte Chemie International Edition* 49.41 (2010): 7570-7574.

19 Stein, Ingo H., Verena Schüller, Philip Böhm, Philip Tinnefeld, and Tim Liedl. "Single-Molecule FRET Ruler Based on Rigid DNA Origami Blocks." *ChemPhysChem* 12.3 (2011): 689-695.

20 Zimmermann, Timo, Jens Rietdorf, Andreas Girod, Virginie Georget, and Rainer Pepperkok. "Spectral imaging and linear un-mixing enables improved FRET efficiency with a novel GFP2–YFP FRET pair." *FEBS letters* 531.2 (2002): 245-249.

21 Axelrod, Daniel, Thomas P. Burghardt, and Nancy L. Thompson. "Total internal reflection fluorescence." *Annual review of biophysics and bioengineering* 13.1 (1984): 247-268.

22 Axelrod, Daniel. "Total internal reflection fluorescence microscopy in cell biology." *Traffic* 2.11 (2001): 764-774.

23 PDB ID: 4JO6
Barrette-Ng, Isabelle H., S-C. Wu, W-M. Tjia, S-L. Wong, and Kenneth KS Ng. "The structure of the SBP-Tag-streptavidin complex reveals a novel helical scaffold bridging binding pockets on separate subunits." *Acta Crystallographica Section D: Biological Crystallography* 69.5 (2013): 879-887.

24 PDB ID: BTN
Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. "The protein data bank." *Nucleic acids research* 28.1 (2000): 235-242.

25 PDB ID: 1BPE
Sawaya, Michael R., Huguette Pelletier, Amalendra Kumar, Samuel H. Wilson, and Joseph Kraut. "Crystal structure of rat DNA polymerase beta: evidence for a common polymerase mechanism." *Science* 264.5167 (1994): 1930-1935.

26 Melin, Jessica, and Stephen R. Quake. "Microfluidic large-scale integration: the evolution of design rules for biological automation." *Annu. Rev. Biophys. Biomol. Struct.* 36 (2007): 213-231.

27 Eddings, Mark A., Michael A. Johnson, and Bruce K. Gale. "Determining the optimal PDMS–PDMS bonding technique for microfluidic devices." *Journal of Micromechanics and Microengineering* 18.6 (2008): 067001.

28 Knight, James B., Ashvin Vishwanath, James P. Brody, and Robert H. Austin. "Hydrodynamic focusing on a silicon chip: mixing nanoliters in microseconds." *Physical Review Letters* 80.17 (1998): 3863.

29 Lee, Gwo-Bin, Chih-Chang Chang, Sung-Bin Huang, and Ruey-Jen Yang. "The hydrodynamic focusing effect inside rectangular microchannels." *Journal of Micromechanics and Microengineering* 16.5 (2006): 1024.

30 Chen, Chun H., Sung Hwan Cho, Frank Tsai, Ahmet Erten, and Yu-Hwa Lo. "Microfluidic cell sorter with integrated piezoelectric actuator." *Biomedical microdevices* 11.6 (2009): 1223-1231.

31. Smith, Temple F., and Michael S. Waterman. "Identification of common molecular subsequences." *Journal of molecular biology* 147.1 (1981): 195-197.

32 Benjamini, Yoav, and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B (Methodological)* (1995): 289-300.

33 Hochberg, Yosef, and Yoav Benjamini. "More powerful procedures for multiple significance testing." *Statistics in medicine* 9.7 (1990): 811-818.

34 PDB ID: 2EWF
Kachalova, Galina S., Eugeny A. Rogulin, Alfiya K. Yunusova, Rimma I. Artyukh, Tatyana A. Perevyazova, Nickolay I. Matvienko, Ludmila A. Zheleznaya, and Hans D. Bartunik. "Structural analysis of the heterodimeric type IIS restriction endonuclease R. BspD6I acting as a complex between a monomeric site-specific nickase and a catalytic subunit." *Journal of molecular biology*384.2 (2008): 489-502.

35 Luo, Fukang, Guimin Xiang, Xiaoyun Pu, Juanchun Yu, Ming Chen, and Guohui Chen. "A Novel Ultrasensitive ECL Sensor for DNA Detection Based on Nicking Endonuclease-Assisted Target Recycling Amplification, Rolling Circle Amplification and Hemin/G-Quadruplex." *Sensors* 15.2 (2015): 2629-2643.

36 Zhu, Zhenyu, James C. Samuelson, Jing Zhou, Andrew Dore, and Shuang-yong Xu. "Engineering strand-specific DNA nicking enzymes from the type IIS restriction endonucleases BsaI, BsmBI, and BsmAI." *Journal of molecular biology* 337.3 (2004): 573-583.

37 Morgan, Richard D., Celine Calvet, Matthew Demeter, Rafael Agra, and Huimin Kong. "Characterization of the specific DNA nicking activity of restriction endonuclease N. BstNBI." *Biological chemistry* 381.11 (2000): 1123-1125.

38 PDB ID: 1NOY
Wang, J., P. Yu, T. C. Lin, W. H. Konigsberg, and T. A. Steitz. "Crystal structures of an NH2-terminal fragment of T4 DNA polymerase and its complexes with single-stranded DNA and with divalent metal ions." *Biochemistry* 35.25 (1996): 8110-8119.

39 Quail, Michael A., Miriam Smith, Paul Coupland, Thomas D. Otto, Simon R. Harris, Thomas R. Connor, Anna Bertoni, Harold P. Swerdlow, and Yong Gu. "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers." *BMC genomics* 13.1 (2012): 341.

40 Buermans, H. P. J., and J. T. Den Dunnen. "Next generation sequencing technology: advances and applications." *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1842.10 (2014): 1932-1941.