4-2013

# IonSeq Genome Sequencing

Kendrick Chow
*University of Pennsylvania*

Tushmit Hasan
*University of Pennsylvania*

Gawain Lau
*University of Pennsylvania*

Joan Liu
*University of Pennsylvania*

# IonSeq Genome Sequencing

**Abstract**
The emergence of advanced DNA sequencing methods has presented disruptive opportunities in biotechnology, establishing the foundation for the personalized medicine industry. Since the completion of the Human Genome Project, the number of genomes sequenced has grown exponentially and the sequencing price has dropped precipitously. To make personalized medicine a reality, there is a need for a large collection of sequenced genomes in order to link specific genes to diseases. IonSeq seeks to be the leading DNA sequencing service, employing new semiconductor- based sequencing technology offered by Ion Torrent, to help pharmaceutical companies generate these libraries of genomes for their drug-development processes. To support sequencing reliability and throughput, IonSeq will explore such technical details such as chip configuration, insertion kinetics, signal generation, base-calling methods, and accuracy metrics. IonSeq will prove a 40 genome/day output, made possible by the massively parallel procedure employed by the sequencers. IonSeq will sequence each genome at a price of $2,000 while the cost of 'manufacture' will only be $645. Series A will consist of a $3,682,886 investment and will yield the investors a MIRR of 102.98% over four years. The Series B investment will total $4,510,491 and result in a 93.43% MIRR over a three period. The NPV by the time of liquidation or acquisition event will be $39,322,347, at a conservative projected growth rate of 5%.

**Disciplines**
Biochemical and Biomolecular Engineering | Chemical Engineering | Engineering

# IonSeq Genome Sequencing

Kendrick Chow

Tushmit Hasan

Gawain Lau

Joan Liu

*Project Advisor: Dr. John C. Crocker*

Professor Leonard A. Fabiano

Department of Chemical and Biomolecular Engineering

University of Pennsylvania

April 9, 2013

Kendrick Chow
Tushmit Hasan
Gawain Lau
Joan Liu

3300 Walnut St
Philadelphia, PA
April 9, 2013

Dr. Leonard Fabiano
Dr. Warren Sieder
Dr. John Crocker
University of Pennsylvania, Chemical and Biomolecular Engineering

Room 311A Towne Building
220 South 33rd Street
University of Pennsylvania
Philadelphia, PA 19104-6393

Dear Sirs,

The following enclosure is a detailed design of our solution to the proposed problem: "Scalability of DNA Sequencing with Moore's Law." As a solution to this problem we have designed a biotechnology start-up company, IonSeq, which sequences whole human genomes as a service.

The enclosed report documents our entire design process and decisions that were made during the formulation of this process. The report also contains a competitive analysis, market analysis, and financial models of the start-up company.

This report contains all necessary information for hypothetical start-up scientists and engineers, whom could, with the use of this report, design a working prototype for IonSeq as per this project's specifications. This report also provides all necessary information that a potential investor in IonSeq would require in order to make an investment decision for the company's future.

We submit this report for your review with our strongest recommendations of the success of IonSeq.

Sincerely,

_____ _____

Kendrick Chow Gawain Lau

_____ _____

Tushmit Hasan Joan Liu

# TABLE OF CONTENTS

Tables

Figures

# 1. ABSTRACT

The emergence of advanced DNA sequencing methods has presented disruptive opportunities in biotechnology, establishing the foundation for the personalized medicine industry. Since the completion of the Human Genome Project, the number of genomes sequenced has grown exponentially and the sequencing price has dropped precipitously. To make personalized medicine a reality, there is a need for a large collection of sequenced genomes in order to link specific genes to diseases. IonSeq seeks to be the leading DNA sequencing service, employing new semiconductor-based sequencing technology offered by Ion Torrent, to help pharmaceutical companies generate these libraries of genomes for their drug-development processes. To support sequencing reliability and throughput, IonSeq will explore such technical details such as chip configuration, insertion kinetics, signal generation, base-calling methods, and accuracy metrics. IonSeq will prove a 40 genome/day output, made possible by the massively parallel procedure employed by the sequencers. IonSeq will sequence each genome at a price of $2,000 while the cost of 'manufacture' will only be $645. Series A will consist of a $3,682,886 investment and will yield the investors a MIRR of 102.98% over four years. The Series B investment will total $4,510,491 and result in a 93.43% MIRR over a three period. The NPV by the time of liquidation or acquisition event will be $39,322,347, at a conservative projected growth rate of 5%.

# 2. INTRODUCTION

The market for personalized medicine is booming, and faster, more efficient sequencing methods are at the forefront of this revolution in genetics. Ion Torrent, a subsidiary of Life Technologies, has developed DNA genome sequencing machines that are capable of providing quick and relatively inexpensive sequences of large genomes. With these machines gaining popularity among research labs, pharmaceutical companies, and clinical medicine, there is a greater demand for not only faster but more accurate sequencing and for the development of clinical data generated from genome sequencing.

To help fulfill this growing demand, a service organization that performs sequencing, led by experts in the field, may prove to be a highly profitable venture. In this light, IonSeq is proposed as a full-service arm of Ion Torrent, aimed at providing DNA genome sequencing services for pharmaceutical companies and finding methods to improve genome throughput. This chapter will detail the context within which IonSeq will be starting its venture.

## 2. A. PURPOSE OF GENOME SEQUENCING

The major goal of the Human Genome Project, and of all major genomic sequencing research, has been to gain a fundamental knowledge of the human body. Before the Human Genome Project was even completed, small companies such as Myriad Genetics began to offer an easy way to administer genetic tests that can identify certain diseases such as cancer or degenerative disorders.

Utilizing genetic information, it is speculated that health care professionals will eventually be able to predict an individual's predisposition for certain diseases, potentially creating opportunities for early intervention to either minimize the impact of the disease or avoid it completely. All disease related genetic variants will be detected, and this will enable the development of rapidly emerging medical fields like personalized and predictive medicine, new fields that allow for a completely new level of precision to determine what medical treatments are appropriate for particular individuals. This is a revolution in biomolecular sciences, allowing for the development of a "'new taxonomy' that defines disease based on underlying molecular and environmental causes, rather than on physical signs and symptoms"[1]

The National Academy of Sciences continues to stress the importance of data availability in future medical research. While part of this research will come from a greater understanding of biomolecular reactions within the body, a larger part relies on the ability to see the underlying genetic code that relates to certain diseases. Personalized medicine will also exist in the form of pharmacogenomics, where genetic information will be used to select the most appropriate drugs to prescribe to a patient to minimize hazardous side effects and maximize beneficial effects.

[1] National Academy of Sciences. Division on Earth & Life Studies, Board on Life Sciences. (2013). *Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease* . Retrieved from National Academy of Sciences website: http://dels.nas.edu/Report/Toward-Precision-Medicine-Building-Knowledge/13284

Development of this knowledge is already underway. As of October 2012, there are over 150 drugs in development targeting or paired with certain genes[2].

## 2. B. Basics of Genome Sequencing

DNA sequencing can involve sequencing of a whole genome or portions of a genome, such as the exome, all exons of the genome, or the transcriptome, all protein coding regions of the genome. Regardless, there are three general steps in all DNA sequencing methods – sample preparation, physical sequencing, and reassembly. During the sample preparation stage, a large sample of DNA is broken up into small fragments; each fragment is clonally amplified hundreds of thousands of times and then processed for sequencing. The fragmentation process is random, resulting in many overlapping fragments. In the sequencing phase, the individual bases in each fragment are identified by using the fragment as a template to sequence its complementary strand. The number of bases identified on a single template is defined as the read-length. During reassembly, bioinformatics software is used to align the overlapping reads, which allows the original genome to be assembled as a continuous sequence. Often times, the reassembly is accomplished by aligning the fragments to a reference genome if one exists. Longer the read lengths result in easier reassembly. More overlap between the fragments results in greater coverage, defined as the average number of times an individual base has been sequenced. This will ultimately improve the accuracy of the final aligned genome.

[2] *The pharmacogenomics knowledgebase*. (n.d.). Retrieved from http://www.pharmgkb.org/

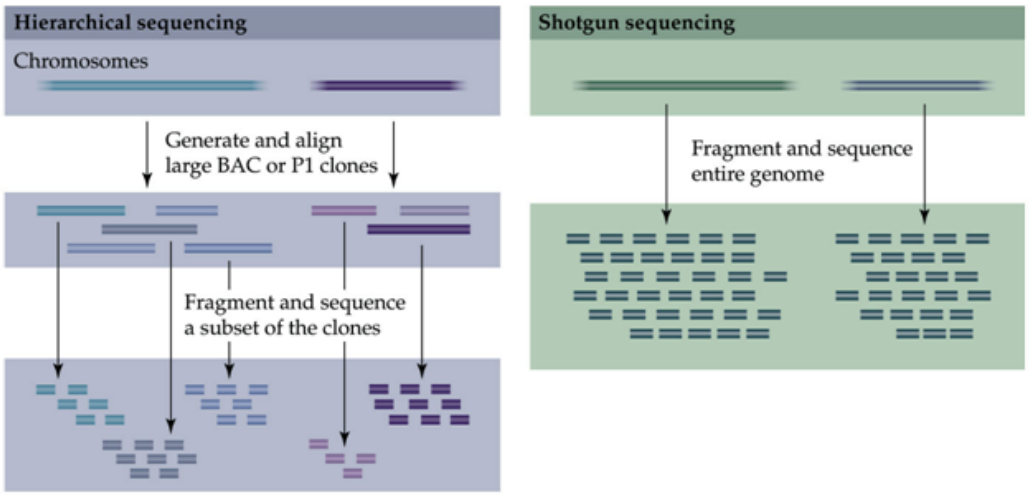



**Figure 1: Hierarchical sequencing is a long process that allows for more accurate DNA mapping. Shotgun sequencing allows for higher throughput but at the expense of ease of mapping.** [3]

There are two primary strategies for genome sequencing: hierarchical sequencing and whole-genome shotgun sequencing as shown in Figure 1. In hierarchical sequencing, the genome is broken up into segments, approximately 200,000 bases in length. These fragments are ligated into bacterial vectors and cultured to form libraries. The fragments are organized to form a low resolution physical map of the genome. Since there is significant overlap in the segments, only those that form the minimum tiling path may be selected to be sequenced. These segments are now broken up into even smaller fragments, sequenced, and then assembled to form continuous stretches of DNA. The fragments are then reassembled to give the sequences of the entire genome. Hierarchical sequencing is an extremely long and laborious process, but it is the most accurate method of mapping sequenced DNA fragments to portions of the genome. This method was employed in the Human Genome Project launched in 1990.

The shotgun approach skips the vector library creation steps and directly breaks up the entire genome into short sequenceable fragments about 500 base pairs (bp) in length. This makes reassembly more difficult and powerful computer algorithms are required for reassembly once the

[3] Gibson G, Muse SV. A Primer of Genome Science. Third Edition. 2009. Sinauer Associates, Inc. Publishers. Sunderland, Massachusetts, USA.

sequences of each fragment are obtained. However, this approach yields considerably higher throughput as the time consuming steps present in hierarchical sequencing are avoided. Therefore, the shotgun approach is adopted by most genome projects in research centers around the world. However, hierarchical sequencing has its merits, especially when sequencing the genome of an organism for the first time because no reference genome exists. But in all other applications, such as the one described in this report, the shotgun approach is employed because it is faster, less expensive, and more efficient.

## 2. C. HISTORY OF GENOME SEQUENCING

The history of DNA sequencing has been built with the contribution of many minds. Today, IonSeq is able to base its fundamental technologies off the work of many its predecessors. IonSeq is motivated by the same goal: seeking to further the knowledge of the human species and advance medical treatments to improve the quality of life for the billions of lives on this planet.

While the structure of DNA as a double helix was established in 1953, the earliest form of genome sequencing technology would not take shape until decades later. In 1972, Walter Fiers, from the University of Ghent, sequenced a single RNA gene of a virus, named Bacteriophage MS2[4], and through the rest of the 1970s, more progress led to the development of rapid DNA sequencing technology. The DNA sequencing movement was further accelerated by the development of recombinant DNA technology, allowing DNA to be extracted from non-viral sources. At this time, two technologies came to surface: Sanger sequencing and Maxam-Gilbert sequencing.

### *2.C.I. FIRST GENERATION ADVANCEMENTS IN GENOME SEQUENCING*

At the MRC Centre of Cambridge, Frederick Sanger published a method for DNA Sequencing with chain-terminating inhibitors, while Walter Gilbert and Allan Maxam at Harvard developed a

[4] Min, J. W., Haegeman, G., Ysebaert, M., & Fiers, W. (1972). Nucleotide sequence of the gene coding for the bacteriophage ms2 coat protein. *Nature*, *237*, 82-88. doi: 10.1038/237082a0

method for sequencing using chemical degradation. Sanger sequencing uses the original DNA strand as a template. A number of complementary strands are built starting from the 5' end, and are each terminated at different points on the 3' end. In regular DNA, the bases are referred to as deoxyribonucleotides (dNTP) or nucleotides. The addition of a dNTP allows the DNA chain to continue its growth in a regular manner. In Sanger sequencing, the template DNA is exposed to a low concentration of dideoxyribonucleotides (ddNTP), which has one less oxygen atom. When a ddNTP is incorporated, or added to the strand, the chain ends. Using a probabilistic model, chains that end at each base in the DNA strand are generated. The ddNTPs are radioactively labeled, and the strands of varying sizes are run through an electrophoresis gel to separate them by relative sizes. The radioactive signatures for each band on the gel can then be read to determine the sequence of the DNA.

The Maxam-Gilbert sequencing method breaks up the DNA into fragments. The DNA is pre-treated such that the breaks, induced by four separate reactions, would occur only after specific bases. The concentration of the modifying chemicals is controlled to induce on average one modification per DNA molecule. Similar to the Sanger method, the 5' end of each fragment was then radioactively labeled, and then the sequence was deduced by determining the sizes of the fragments using slab gel electrophoresis. The sequence was then derived from the sizes of the fragments. Since its development, the Maxam-Gilbert method has fallen out of favor due to the technical complexity prohibiting the production of standard molecular biology kits, the use of hazardous chemicals, and the sheer number of reactions necessary that impede proper scale-up. Sanger sequencing was adopted as the primary technology in the first generation of DNA sequencing given its higher accuracy and low radioactivity.

Throughout the 1980s, scientists performed sequencing manually. The process was labor-intensive and time consuming and had very low throughput. In the 1990s several advances were made that allowed automation of Sanger sequencing, which saw the birth of the first generation of

high throughput sequencing. The first such machines were introduced by Applied Biosystems in 1987[5]. The automation replaced slab gel electrophoresis with capillary electrophoresis and made use of fluorescently-labeled ddNTPs, which removed the need to read the gels manually. Now, "trace files" with four colored peaks indicating the position of each base could be generated, and the sequence could directly read from the file. The first machine, named AB370, was able to read 500,000 bases a day, on read lengths of 600 bases. The latest model from Applied Biosystems, AB3730xl, is able to read 2.88 million bases per day, and reads in lengths of 900 bases. However, there has been little advancement with this technology since 1995.

### *2.C.II. NEXT GENERATION GENOME SEQUENCING*

Even with automation and advancements in sequencing technology, Sanger sequencing using capillary electrophoresis still costs $30-$50 million to sequence a complex genome[6]. Unless the cost decreased, it would not be possible to make wide-scale use of genome sequencing in treatment of diseases. High-throughput technologies that employed massively parallel sequencing were developed to lower the cost of DNA sequencing beyond standard dye-terminator methods.

The leaders of this second generation of genome sequencers were the 454 Life Sciences platform, the Illumina Genome Analyzer, and the ABI SOLiD System. These next generation sequencing technologies primarily differ from the Sanger method in aspects of paralleling analyses of many small sequences, resulting in higher throughput and reduced cost. Over the past decade[7], these systems have achieved significant improvements in read length and accuracy and have also reduced the cost of genome sequencing while yielding increased throughput.

[5] Liu, L Yinhu Li, Siliang Li, et al. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, *2012*, Retrieved from http://www.hindawi.com/journals/bmri/2012/251364/
[6] Gibson, G., & Muse, S. (2009). *A primer of genome science*. (3 ed.). Sunderland, Massachusetts, USA: Sinauer Associates, Inc. Publishers.
[7] Liu, "Comparison", 2012

## *2.C.III. THE HUMAN GENOME PROJECT*

With the advancements in DNA technology, and a greater understanding and acceptance of DNA as the "code of life", it was natural that interest grew in sequencing human DNA. It was and still is anticipated that advanced knowledge of the human genome will provide new areas for progress in medicine and biotechnology. The earliest reports even state that "knowledge of the human is as necessary to the continuing progress of medicine and other health sciences as knowledge of human anatomy has been for the present state of medicine"[8].

In 1990, the US Department of Energy and the National Institutes of Health formally founded the Human Genome Project, a $3 billion endeavor projected to be completed in 15 years. This was not only a United States sponsored venture; researchers from the United Kingdom, France, Australia, Japan, and many other countries joined this substantial undertaking. The Human Genome Project employed hierarchical sequencing as it allowed accurate mapping and high quality sequences with less than 1 error per 40,000 bases[9]. This approach also made it possible to share the workload across research centers around the world.

A "rough draft" of the genome was finally published in 2000, made possible by major advances in computing technology. Frustrated at the slow pace of the government-sponsored genome project, in 2000, a private biotechnology firm, Celera Genomics, launched a parallel human genome project and employed the whole-genome shotgun approach. Because of their higher throughput and powerful bioinformatics software, they were able to catch up to the Human Genome Project by 2001 at only a tenth of the cost – $300 million. However, Celera had unrestricted access to the Human Genome Project progress data, which handily served as reference

[8] Mendelsohn, M. L., et al. Department of Energy Office of Energy Research Office of Health and Environmental Research, Subcommittee on Human Genome of the Health and Environmental Research Advisory Committee (1987). *Report on the human genome initiative office of health and environmental research: Report on the human genome initiative office of health and environmental research*. Retrieved from website: http://www.ornl.gov/sci/techresources/Human_Genome/project/herac2.shtml

[9] Mardis ER. Next-Generation DNA Sequencing Methods. Annual Review of Genomics and Human Genetics. 2008. 9: 387-402.

for their shotgun sequencing methods. Nevertheless, Celera's contribution considerably accelerated the process and the essentially complete sequence was announced in 2003, two years ahead of schedule. By 2006, the sequence of the last chromosome was published.

### *2.C.IV. THE ARCHON GENOMICS X PRIZE*

Upon the completion of the Human Genome Project, the X Prize Foundation, based in Playa Vista, California, announced the Archon Genomics X Prize in October of 2006. The X Prize was a joint effort between the X Prize foundation and the J. Craig Venter Science Foundation, and promised a $10-million dollar prize to the first team that could sequence 100 genomes within 30 days or less, with an accuracy of no more than 1 error in every 1,000,000 bases sequenced, and costs no more than $1000 per genome[10].

**Table 1: Archon X Prize Requirements[11]**

| Competition Categories | Minimum Requirements | Best-in-Class Requirements |
|---|---|---|
| **Cost** | $10,000 per genome | $1,000 per genome or less |
| **Speed** | 30 Days | FIRST TO SUBMIT;<br>Must be 30 days or less |
| **Accuracy** | No more than 1 error per 100,000 bases | No more than 1 error per 1,000,000 bases |
| **Completeness** | 95% | 98% |
| **Haplotype Phasing** | 0% | Complete Phasing of Chromosomes (see section 3.8) |

It is no surprise that the requirements of the Archon X Prize, shown above in Table 1 align perfectly with the goals of IonSeq. The Archon X Prize is responsible for spurring the development of so many next generation sequencing technologies. As of completion date of this report, the Archon X Prize remains to be collected.

[10] Express Scripts. (2012). Archon genomics x prize competition guidelines. In New York, New York: Retrieved from http://genomics.xprize.org/sites/genomics.xprize.org/files/docs/AGXP_Competition_Guidelines.pdf
[11] Ibid.

## 2. D. PROJECT GOALS AND REQUIREMENTS

The primary goal of IonSeq is to achieve unprecedented human genome throughput at a low cost and to design operations to deliver accurate sequences for its customers. A brief overview of the IonSeq Project Charter is presented below in Table 2. IonSeq will focus on bringing the best sequencing services to its clients by focusing on areas where optimizations can be made. The $1,000 genome is the XPrize target, but IonSeq will set a per sequenced genome price on the basis of achieving the best returns for the company's investors.

**Table 2: Project Charter for IonSeq DNA Sequencing**

| Project Charter | The IonSeq Approach |
|---|---|
| **Project Name** | DNA Sequencing: $1,000 Genomes using Ion Torrent Technologies |
| **Project Champions** | Dr. John Crocker |
| **Project Leaders** | Kendrick Chow, Tushmit Hasan, Gawain Lau, Joan Liu |
| **Specific Goals** | Design a $1,000 genome sequencing process in the context of a service-based startup company, capable of sequencing 10,000 human genomes per year (equivalent 40 genomes/day for 250 days/year) |
| **Project Scope:** | In-scope:<br>• Optimization of DNA sequencing workflow<br>• Error analysis modeling<br>• Costing and profitability analysis on current technology and optimizations<br>Out-of-scope:<br>• Manufacturing of Ion Torrent technology<br>• Biomedical Analysis of Generated Genome Sequences |
| **Deliverables** | • Business opportunity assessment<br>• Technical feasibility assessment<br>• Full scale service requirements<br>• Financial analysis over four year period |
| **Timeline** | • **Year 1:** Proof-of-concept, Series A funding for 10 genomes/day, 2500 genomes/year<br>• **Year 2:** Start-up service for hospitals/labs/clinics. Series B funding: 40 genomes/day, 10000 genomes/year<br>• **Years 3:** Expand service across the United States.<br>• **Year 4:** Assume acquisition or liquidation event. |

### *2.D.I. IONSEQ REPORT ROADMAP*

From the next section forward, this report will outline the technical and business foundations of IonSeq. Chapter 3 will discuss the background of IonSeq, from a brief history of the base technology to factors and requirements that will be of interest to the company. Chapter 4 will explore the pre-sequencing steps, including DNA extraction/library creation, emulsion PCR, and enrichment. Chapter 5 will go into detail regarding the chip configuration and explore the time scale of the steps within the bead loading and sequencing process. Chapters 6 and 7 will derive models of the kinetics of nucleotide insertions and signal generation from ISFET sensors, which are both important parts of sequencing. Chapter 8 will cover topics involving the reconstruction of the genome, including dephasing, base calling, and alignment. Chapter 9 will explore various optimization options, including ISFET material selection and other sequencing parameters. Chapter 10 will describe the market within which IonSeq intends to operate. Chapter 11 will outline how IonSeq intends to execute this venture, from establishing an execution timeline to detailing the company's operations. Finally, Chapter 12 will present a comprehensive financial analysis of this venture, which will show the strong potential for IonSeq's success.

# 3. COMPANY BACKGROUND: CREATION OF IONSEQ

IonSeq is the newest evolution of Ion Torrent. Building on top of the sequencing technology of the past generation, IonSeq seeks to access a growing market by providing a full genomic sequencing service, thereby simplifying the DNA sequencing process, and increasing the access of Ion Torrent's technology. A major road block in sequencing technology has been the complicated work flows[12]. However, IonSeq will take advantage of the emergence of the desktop sequencing market and the new level of ease it has allowed in genomic sequencing. In this chapter, an overview of Ion Torrent technologies will be covered and customer and technical requirements will be briefly discussed.

## 3.A. ION TORRENT

Ion Torrent was founded by Jonathan Rothberg in 2007, in Guilford, CT. Rothberg, who in 1999 had founded 454 Life Sciences, was no stranger to next generation sequencing. In 2010, Ion Torrent was acquired by San Francisco-based Life Technologies.

[12] Mulhern, J. (2013, February 18). Ion torrent edges illumina in sales battle of benchtop sequencers, says macquarie report. *Bio-IT WOrld*, Retrieved from http://www.bio-itworld.com/news/02/18/13/Ion-Torrent-edges-Illumina-sales-benchtop-sequencers-Macquarie.html

### *3.A.I. OVERVIEW OF ION TORRENT TECHNOLOGY*

All of Ion Torrent technologies are based on semiconductor sequencing. During the polymerization of DNA, hydrogen ions are released, and a change in pH is induced. The Ion Torrent system consists of a series of wells, each containing a bead covered with template DNA strands, located on a small chip, which is placed in Ion Torrent's sequencer machine for sequencing. A single nucleotide is introduced to the wells, and if the dNTP is complementary to the leading template nucleotide, it is incorporated onto the strand. The remaining unreacted nucleotides are washed out of the well, and the next nucleotide flow is loaded. These nucleotide insertions release protons, or $H^+$, which trigger an Ion-Sensitive Field Effect Transistor (ISFET). The ISFET is located upon a complementary metal-oxide-semiconductor (CMOS), which is able to convert the genetic information to digital information.

The sequencing chemistry is a flow-based chemistry originally introduced in the 454 sequencing platform. The DNA fragments on the bead are rendered single stranded and primed, loaded with polymerase, and sequenced. Each well is monitored for insertion events. If there is an insertion, the release of protons from all the strands on the bead creates a positive voltage near the gate region of a transistor, which results in a change in the current flowing through the transistor. This is the fundamental detection process, which may in turn be converted into a voltage signal by collecting the associated current.

Ion Torrent released their first system, the Ion Personal Genome Machine (PGM), in December of 2010. This is the least expensive next generation sequencer on the market, with a list price of approximately $50,000. In addition, runs cost between $300 and $750. However, the PGM is targeted towards smaller genomes and are unable to handle a full human genome.

Ion Torrent has introduced three chips for the PGM—the 314, 316, and 318—each with greater number of wells and output. The latest, PGM 318, is capable of completing a 100-base read

in about eight hours. Ion Torrent also prepares software that streamlines data analysis and preparation kits that accompany the system. The expected output is anticipated to have an accuracy of over 99% after alignment.

In September 2012, Ion Torrent introduced the Ion Proton System which allows for larger chips with higher densities needed for exome and whole genome scale sequencing. The Ion Proton is substantially more expensive at $149,000 but is capable of generating much larger outputs. The first chip, the Proton I is said to be able to give 30x coverage for 2 human exomes, which translates to about 60 million bases.[13] The Ion Proton System's next chip, the Proton II is promoted to be capable of generating 30x coverage for an entire human genome, or 3 billion bases.

### *3.A.II. ION TECHNOLOGY VS. EVERYTHING ELSE*

Ion Torrent technology is unique in that no modified nucleotides or optics are used. This comes into play especially when considering accuracy. Ion Torrent's massively parallel technology also oversamples the DNA sequence, up to 30x coverage, in independent sequencing reactions to allow for high consensus accuracy. In optical DNA sequencing, the basis for other next generation technologies, nucleotides are modified with a fluorescent signature which can be captured under fluorescence illumination. However, the modified nucleotides must have their fluorescent signatures removed before the addition of a new base, which is sometimes performed incorrectly, leading to inaccurate reads. In addition, there are special cases where a series of the same base occurs in a row, such as AAA or TTTT. These are known as homopolymers. With the introduction of the correct nucleotide, the entire homopolymer chain is incorporated at once, instead of one at a time. In optical systems, it is difficult to quantify the increase in intensity of light from homopolymer insertions. In the semiconductor system, the measured pH difference and the resulting voltage difference can be more reliably related to the length of the homopolymer.

[13] Ng, S. B., Turner, E. H., & et al, (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, *461*, 272-276. doi: 10.1038/nature08250

Ion Torrent technology is also superior in terms of speed and scalability. Due to the number of reactions involved in traditional sequencing technologies, there is a considerable delay between bases. However, Ion Torrent's system of flow chemistry is limited only by the kinetics of nucleotide insertions and ISFET sensor behavior.

### *3.A.III. CUSTOMER NEEDS ANALYSIS*

Customers require accurate sequencing – 10 errors per million bases or a Phred rating of Q50. A full definition of the Phred scale and derivation of error rates are covered in Section 8.D. They will not require full analysis of sequenced genomes, but merely the raw data generated from sequencing runs. Clients will not require long base reads, and the company will not be tailoring to customers for de novo sequencing, but for human genome sequences, where there are sequenced human genomes available for comparison and genome reassembly. Furthermore, customers will seek a price per sequenced genome under $5,000, preferring a price tag that approaches $1,000 as outlined by the Archon X Prize.

### *3.A.IV. TECHNICAL REQUIREMENTS*

To achieve these specifications, the important parameters IonSeq plans to address include polymerase selection, nucleotide incorporation rate, well size, well configuration and density, diffusion of protons/nucleotides, and ISFET design. These parameters will be designed to achieve accurate sequencing and greater throughput by reducing run times. The following sections will explore these parameters in detail.

# 4. PRE-SEQUENCING

The IonSeq workflow begins with the collection of DNA samples from its clients and moves through all the functions necessary to yield a full sequenced, reconstructed human genome, in an easy to access file format for the clients' use. Pre-sequencing makes up an important part among the other major elements in the technical workflow, illustrated in Figure 2. This chapter will explore the DNA extraction/library creation, emulsion PCR, and enrichment steps.


**Figure 2: IonSeq workflow.** IonSeq will provide clients with an easy way to retrieve raw human DNA genome sequence from a DNA sample.

## 4.A. PRE-SEQUENCING PREPARATIONS

Before the DNA is sequenced, it must be properly shipped to our laboratory facility and processed. The process for pre-sequencing, which consists of steps crucial to the success of the sequencing run, is thoroughly outlined below.

### *4.A.I. DNA COLLECTION*

Upon requesting IonSeq's services, customers will be mailed a DNA Collection kit that they can use to mail back a DNA sample of the genome to be sequenced. IonSeq is exploring a variety of DNA Self-Collection Kits, many of which are available and compatible for use with Ion Torrent technology. Customers have the option of buying their own DNA extraction kits, or using their own DNA extraction techniques, but IonSeq is no longer able to guarantee a high level of accuracy. Currenlty, IonSeq's suggested kit is Oragene's DNA Self-Collection Kits, which require about 2 mL of saliva from the donor[14]. Other DNA extraction kits that customers may choose to use are Norgen Buccal DNA Collection Kit and Isohelix DNA Buccal Swabs. The DNA sample will be mailed back to IonSeq in a shipping container that was provided along with the DNA collection kit. Upon receiving the sample, the DNA must be purified and incubated overnight at 50°C before extraction[15].

### *4.A.II. DNA EXTRACTION*

DNA extraction will be carried out as an automated process using the Magtration® System 12GC instrument manufactured by PSS Bio Instruments. This process will be completed in IonSeq's laboratory. The device is a bench top unit, and uses paramagnetic-particle technology to purify DNA from the Oragene solution and can purify up to 12 Oragene samples in 30 minutes, using an elution volume of 200 µL[16]. In a test study done by PSS Bio Instruments, which manufactures the Magtration® Systems, the following scattergram of DNA yields was generated, and is shown in Figure 3 below[17]. From the 200 µL sample, the median yield of usable DNA is 3.8 µg, sufficient for the rest of the preparation steps.

[14] Lem, C. S. (2009). Magtration System 12GC: Application data - DNA from Saliva. PSS Bio Instruments technical bulletin (101305), 2
[15] Ibid.
[16] Lem (2009). Magtration
[17] Lem (2009), Magtration

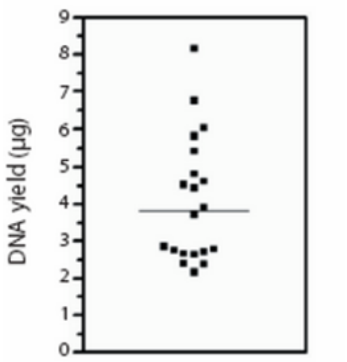



**Figure 3: Scattergram of DNA yields from 200 µL of Oragene/saliva sample.** The horizontal line represents the median yield of 3.8 µg[18].

## *4.A.III. DNA FRAGMENT LIBRARY PREPARATION*

This step is also carried out in the IonSeq laboratory and involves the most "wet-lab" work. First, the DNA is fragmented to appropriately sized, blunt-ended DNA fragments. The fragment DNA is ligated to Ion-compatible adapters, followed by nick repair to complete the linkage between adapters and DNA inserts. For barcoded libraries, Ion Xpress™ Barcode Adapters are available.

### *4.A.III.1) DNA FRAGMENTATION*

The DNA has to be broken down into short fragments that can be easily sequenced. Fragmentation is a random process, hence there is considerable overlap between many of the fragments as shown Figure 4 below.

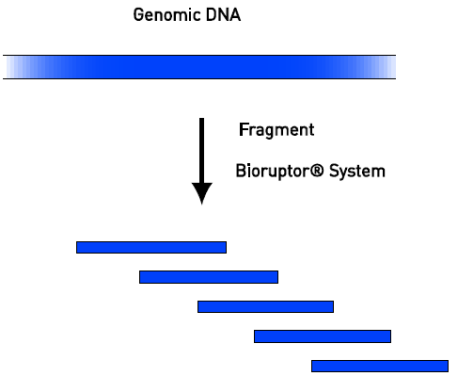



**Figure 4: DNA fragmentation is necessary to allow for easier handling of long genomes. This results in overlap that increases sequencing coverage and improves sequencing accuracy.**

[18] Lem (2009), Magtration

The DNA is fragmented using a BioRuptor® UCD-600 NGS Sonication System. The device accepts input amounts of 100 ng or 1 µg of genomic DNA and fragments the DNA into 100, 200 or 300 base-read fragments. Table 3 shows the library sizes that must be aimed for in order to achieve the desired target read lengths. The recommended read length for the fragment libraries is 200 bp, as will be discussed in detail later in the Dephasing Model section in Chapter 8.A.

**Table 3: Library sizes that must be aimed for in order to achieve the desired target read lengths[19]**

| Target read length* | Median insert size | Median library size |
|---|---|---|
| 300 bases (300base-read library) | ~320 bp | ~390 bp |
| 200 bases (200-base-read library) | ~260 bp | ~330 bp |
| 100 bases (100-base-read library) | ~130 bp | ~200 bp |

* Library sizes are described in this table in terms of the target read length for the library.

The fragmentation profile, the distribution of fragment sizes, can be assessed using a Bioanalyzer® instrument. The samples then must be further prepared using manual procedures and various kits.

### *4.A.iii.2) End Repair*

During fragmentation, the shearing process does not make clean cuts and often the ends of the fragments are damaged. The 5' and 3' ends may contain phosphate or hydroxyl overhangs that will block the ends and interfere with the amplification and sequencing steps downstream. Hence, the ends of the library fragments must be repaired. This is performed hands-on using the end repair buffers and end repair enzymes provided in the Ion Plus Fragment Kit.

### *4.A.iii.3) Adapter Ligation, Nick Repair, and Barcoding*

After the ends are repaired, adapters must be ligated on either of the fragments such that they can be attached to beads for clonal amplification, which is discussed in section 4.A.iv. The adapters are short segments of DNA of known sequences. Two different adapters are used, and one of the adapters is usually modified with a bioton on the 5' end. The biotin attaches to the

[19] Life Technologies. (2012). Ion xpress plus gdna fragment library preparation. In Life Technologies.

streptavidin-coated beads on which the fragments will be clonally amplified. IonSeq will use the Ion Plus Fragment Library Kit, which contains standard A and P1 adapters. It also contains DNA ligase and nick repair polymerases that are necessary to prepare a good adapter-ligated and nick-translated fragment library.

During the adapter ligation step, the option exists to use barcoded adapters instead of the standard A adapter. The Ion Xpress™ Barcode Adapters Kit can be used to create barcoded libraries. The barcoded adapters contain a special sequence of DNA, typically 40 bases long, that will serve to identify the fragment as belonging to a particular genome. This becomes useful if more than one genome is sequenced on a single chip, which can significantly increase throughput.

#### *4.A.III.4) DNA PURIFICATION*

The Agencourt® AMPure® XP Kit can be used to purify the fragment libraries[20]. This step is necessary because not all fragments get attached to adapters during the previous step, and not all fragments get attached to different adapters. Removal of these faulty fragments will help the execution of the next steps of the pre-sequencing process.

#### *4.A.III.5) SIZE SELECTION AND LIBRARY QUANTITATION*

The DNA fragments can now be size-selected using the Pippin Prep™ instrument available from Life Technologies. This gives a tighter size distribution than gel selection and results in a more consistent library size. The library is then quantified using quantitative PCR (qPCR) and the Ion Library Quantitation Kit. It may be necessary to amplify the library in order to ensure that sufficient template preparation reactions can take place on the beads. Appendix A describes how to determine if amplification is required. The final step before template preparation is qualifying and

[20] Life Technologies, Ion Xpress™ Plus

pooling the libraries using qPCR and Bioanalyzer® quantitation[21]. The details of these steps are also listed in Appendix B.

### *4.A.IV. TEMPLATE PREPARATION USING EMULSION PCR (EMPCR)*

The next step is to clone the fragments. In emPCR, DNA fragments are amplified to form a clonal population on beads. The fragments are denatured to form single strands. The strands and beads are mixed in a water-in-oil emulsion such that microreactors are formed in the emulsion each containing one strand and one bead, which then anneal. Reagents required for PCR may now be added and each strand is clonally amplified to form hundreds of thousands of copies on the beads. Figure 5 shows a representation of the emPCR process.

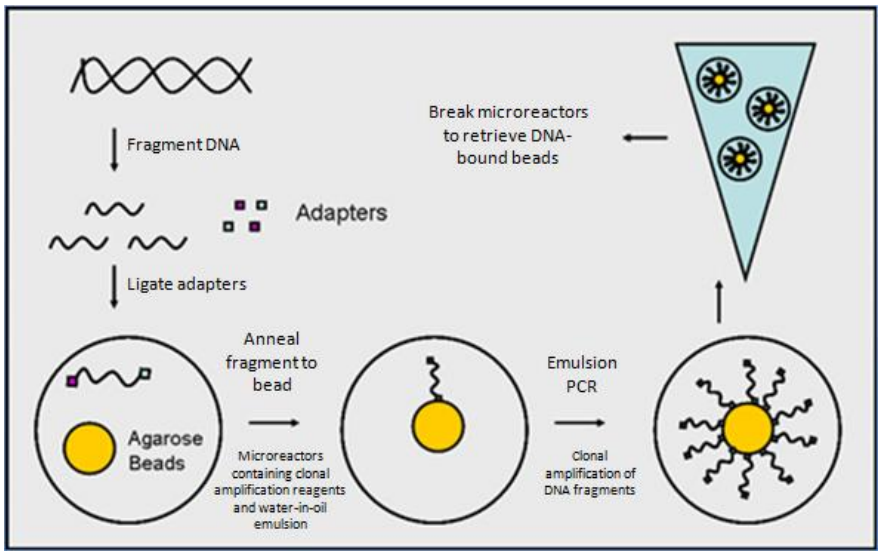



**Figure 5: Clonal Amplification of DNA fragments[22]**

#### *4.A.IV.1) CLONAL AMPLIFICATION AND SAMPLE RECOVERY*

This template preparation step can be carried out using the Ion OneTouch™ 2 System which integrates multiple manual template preparation steps (loading, clonal amplification and sample recovery) into a single system and also enables parallel processing of multiple samples per day

[21] Life Technologies, Ion Xpress™
[22] Life Sequencing. (Producer). (2008). *empcr to ssdna library*. [Print Graphic]. Retrieved from http://www.lifesequencing.com/pages/protocolo-de-secuenciacion?locale=en

through a modular design[23]. Before the sample is loaded into the system, the library has to be diluted using the appropriated template dilution factor (to give a concentration of ~26 pM). The clonal amplification takes place on Ion OneTouch™ 200 Ion Sphere™ Particles. The system recovers and enriches the template positive particles and yields about 10-30% of usable beads for sequencing.[24] Details of calculating the template dilution factor is included in Appendix C.

## 4.B. PRE-SEQUENCING RESULTS

At this juncture, template beads have been produced, with the fragments clonally amplified over the entire surface of the bead, numbering into the hundreds of thousands. Figure 6 is another representation that shows the bead and clonally amplified strands as well as a brief illustration of the complementary sequence to the template strands. This model will be revisited in Chapter 6 when kinetics of nucleotide insertion will be covered.

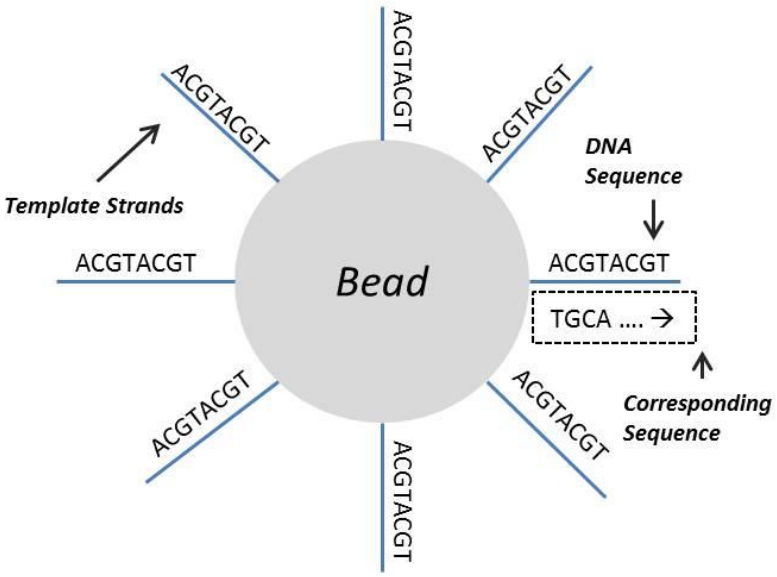



**Figure 6: This bead has template strands clonally amplified on its surface. In reality, over hundreds of thousands of strands would be present.**

[23] "Ion OneTouch™ 2 System." Life Technologies, n.d. Web. <https://products.invitrogen.com/ivgn/product/4474779?ICID=search-product>.
[24] Ion Torrent User Guide. Ion OneTouch™ System. 2012. Publication Part Number 4472430 Rev. E LifeTechnologiesCorp YouTube Channel. Watch Ion OneTouch™ technology in action. http://www.youtube.com/watch?v=fxCY_f0QaZQ

# 5. CHIP CONFIGURATION, WORKFLOW, AND THROUGHPUT

The sequencing chips are the center of the entire IonSeq sequencing process and the foundation of the Ion Torrent technology. On these chips, all of the sequencing of the genome is performed. The DNA template beads are loaded in wells on the chips, nucleotides are flowed through the chips, and the ISFET sensors are fabricated in the chips. Understanding overall chip configuration will help build ideas of the possible ways to increase throughput via structural changes. Furthermore, a key parameter, well diameter, will prove important in proton diffusion and signal generation modeling.

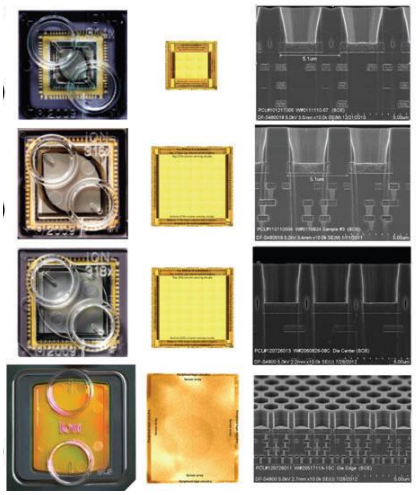
## 5.A. MANUFACTURING NODE, DIE SIZE, AND OTHER CHIP SPECIFICATIONS

Ion Torrent sequencing chips have evolved since the organization's inception, changing chip sizes, semiconductor manufacturing node, well diameters, and well pitch. The chip size is reported

by its length and width, typically in millimeters. The semiconductor node is a standard manufacturing metric; it represents one half of the shortest distance between identical features that can be fabricated. For instance, the 350 nm node indicates that 700 nm is the minimum distance between features on a chip; the 110 nm node dictates a minimum distance of 220 nm. Well diameter is typically in microns and is one of the key parameters of the sequencing chip. Each individual well holds the template bead where the sequencing process occurs. Well pitch is defined as the distance between the centers of neighboring wells. Table 4 outlines the progress of these chips up to the Proton I, which is manufactured using 110 nm node technologies. The previous three, the 314, 316, and 318 models, were manufactured using 350 nm standards.[25]

**Table 4: History and Specification of Past Ion Torrent Sequencing Chips[26][27]**

| Chip | Sensor Count ($10^6$) | Die Size (*mm* x *mm*) | Well Diameter ($\mu m$) | Well Pitch ($\mu m$) |
|---|---|---|---|---|
| 314 | 1.2 | 10.6 x 11.0 | 3.0 | 5.1 |
| 316 | 6.3 | 16.9 x 17.1 | 3.0 | 5.1 |
| 318 | 11.3 | 16.9 x 17.1 | 3.0 | 4.1 |
| Proton I | 165 | 23.7 x 20.0 | 1.25 | 1.68 |



To understand the limits the manufacturing node imposes upon chip fabrication, the distance edge to edge between two wells on a Proton I chip—the difference between the well pitch and well diameter—is 0.43 µm, above the 220 nm limit for the 110 nm node. Theoretically, the Proton I

[25] Merriman, B., Ion Torrent R&D Team, B., & Rothberg, J. (2012). Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis*, *33*, 3397-3417.
[26] Ibid.
[27] Rothberg, Jonathan M, et al. "An Integrated Semiconductor Device Enabling Non-optical Genome Sequencing." *Nature* 475 (2011): 348-52. Print.

would be limited to a well pitch of 1.47 µm with a well diameter of 1.25 µm. Figure 7 illustrates the geometry explained above.


**Figure 7: The dotted line indicates the node while the solid line shows the diameter and pitch. The configuration on the right is at the limit of the 110 nm node.**

## 5.B. Reverse Engineering Proton II Specifications

The Proton II chip has been touted as the first chip to sequence a full human genome; however, it will not be released until Quarter 3 of 2013. Preliminary Ion Torrent specifications state 660 million wells, using the same 110 nm node on the same chip size of 20 mm x 23.7 mm, without providing any information on well size or configuration[28]. Therefore, using this known information, IonSeq can reverse engineer a potential configuration for the Proton II chip.

As the first step, the potential well arrangements are considered below for the Proton I chip; the results of this sizing experiment will help confirm the method for reverse engineering the Proton II chip. Two configurations are considered below in

Figure 8: a square pattern or a hexagonally-packed pattern, as ascertained from the SEM image of the Proton I chip in Table 4.

[28] Rothberg, Jonathan M, et al. "An Integrated Semiconductor Device Enabling Non-optical Genome Sequencing." *Nature* 475 (2011): 348-52. Print.

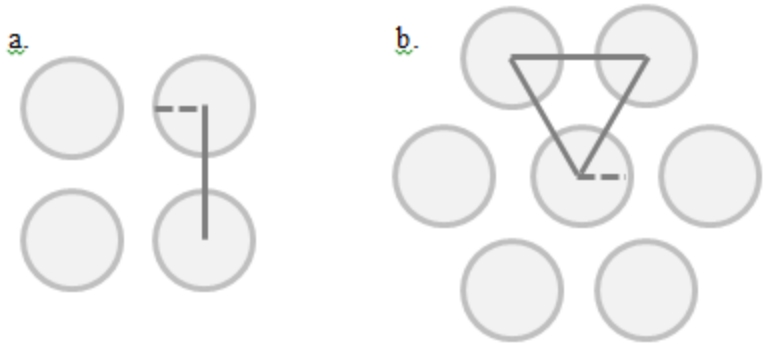



**Figure 8: Two different well arrangements on the chip will yield different well densities and lead to different overall number of wells on a chip.**

The well-to-well pitch is shown by the solid line, and the dotted line indicates the well diameter. Equation 1and Equation 2 are derived well number density expressions for the square packing and the hexagonal packing, respectively.

**Equation 1**

$$Square\ Well\ Number\ Density\ \left[\frac{wells}{\mu m^2}\right] = \frac{1}{(Pitch)^2}$$

**Equation 2**

$$Hexagonal\ Well\ Number\ Density = \frac{3}{6 * \frac{1}{2} * Pitch * \sqrt{Pitch^2 - \left(\frac{Pitch}{2}\right)^2}}$$

For the Proton I, Ion Torrent specifications indicate 165 million wells are fabricated on a chip of 20 mm x 23.7 mm dimensions. To reproduce this value, the number of wells was determined by calculating the density of wells on a square micron basis and multiplying by the area of the chip to yield number of wells.

**Table 5: Well Density and Number Results for 1.25 µm Diameter and 1.68 µm Pitch**

| Proton I | Chip Area ($10^6$ µm$^2$) | Well Number Density (µm$^{-2}$) | Number of Wells ($10^6$) |
|---|---|---|---|
| Square-packed | 474 | 0.354 | 168 |
| Hexagonal-packed | 474 | 0.409 | 194 |

Table 5 outlines the results of from the rest of these calculations. The hexagonal-packed arrangement is too dense to yield the 165 million well specification for the Proton I. However, the square-packed configuration yields 168 million wells, correlating very well to the specification. The

square-packing, in the case of the Proton I, is the most likely configuration. Furthermore, the well density method is validated and can be applied to reverse engineering the Proton II chip.

For the Proton II, the four-fold increase in well number, from 165 to 660 million wells, suggests significant reworking of well geometry. Several possibilities were considered. First, the well diameter was considered to not have changed from 1.25 μm. In this situation, neither the square packed nor the hexagonal-packed configurations are sufficiently dense to permit the 660 million wells as shown in Table 6 as the pitch is limited by the 110 nm manufacturing node.

**Table 6: Well Density and Number Results for Different Diameters and Pitches**

| Proton II | Chip Area ($10^6$ μm$^2$) | Well Number Density (μm$^{-2}$) | Number of Wells ($10^6$) |
|---|---|---|---|
| *1.25 μm diameter, 1.47 μm pitch* | | | |
| Square-packed | 474 | 0.463 | 219 |
| Hexagonal-packed | 474 | 0.564 | 267 |
| *0.70 μm diameter, 0.92 μm pitch* | | | |
| Square-packed | 474 | 1.206 | 572 |
| Hexagonal-packed | 474 | 1.392 | 660 |
| *0.63 μm diameter, 0.85 μm pitch* | | | |
| Square-packed | 474 | 1.392 | 660 |
| Hexagonal-packed | 474 | 1.608 | 762 |

Alternatively, setting the pitch size to its minimum limit, as defined by the 110 nm node, it was found that the well diameter must be 0.70 μm with a well-to-well pitch of 0.92 μm to yield 660 million wells in the hexagonal packed organization. For the square-packed arrangement, the diameter must be 0.63 μm with a pitch of 0.85 μm. It was decided to adopt the hexagonal packing for the Proton II chip due to the proton diffusion considerations with the larger well size as covered in Chapter 6. The kinetics and signal generation models will use 0.70 μm as the well diameter.

To give a sense of the scalability in this node, 1 billion wells requires 0.52 μm diameter wells with 0.74 μm pitch. For 1.5 billion wells, 0.39 μm diameter wells and 0.61 μm pitch are required. It is important to understand, however, that while increasing the number of wells by shrinking well size and pitch are certainly key design parameters, they have important consequences for the emulsion PCR process, sensitive to bead size, and well-to-well crosstalk, an

important signal-to-noise consideration when shrinking feature sizes. Switching to a smaller node would allow larger wells with smaller pitch lengths. An increase in die size, as well as the change in the manufacturing standard, all influence chip cost. For this design, IonSeq will keep the same chip size as the Proton I, in order to maintain compatibility with the Proton machines manufactured by Ion Torrent. However, for potential solutions, it will consider the use of the 110 nm and 32 nm nodes. Pricing, though, will be based off the 110 nm node manufacturing standard.

## 5.C. CHIP FLOW PROCESSES

The overall series of steps on the chip are illustrated in Figure 9. This section will describe each of the steps and introduce more concepts to be covered later in this report.

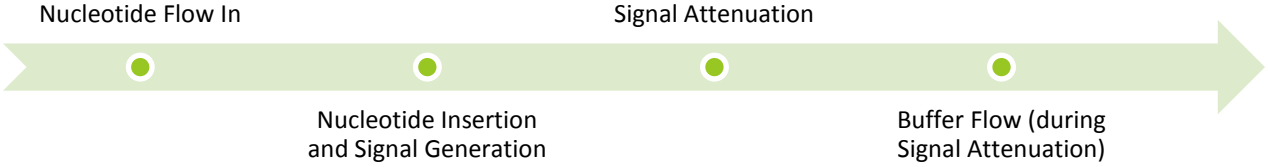



**Figure 9: Workflow on a Proton Chip**

### *5.C.I. NUCLEOTIDE FLOW IN*

Ion Torrent patents state that the nucleotide fluid flow has "a fluid flow Reynolds number of at most 2000, 1000, 200, 100, 50, or 20."[29] To understand the work sequence on a chip, the flows of solutions need to be modeled. Flow of nucleotides onto chip surface is indicated as laminar; a 4 mL/sec flow rate is one of the volumetric flow rates suggested, but this can be taken to be a design parameter[30]. Given the characteristics of the well, and a 1-mm guess for the gap height between the cover of the chip and chip surface, the results for the Reynolds number and amount of time it takes to fill the open volume of the chip are outlined in Table 7. These time scales confirm the claims

[29] Bustillo, J., W. Hinz, K.L. Johnson, J. Leamon, J.M. Rothberg, and J. Schultz. Sequencing nucleic acid comprises disposing template nucleic acids into reaction chambers in contact with or capacitively coupled to chemical-sensitive field effect transistor. Ion Torrent Systems, assignee. Patent GB2561128-A; GB2461128-B. 15 Dec. 2010. Print.
[30] Ibid.

made in the patent of the nucleotide flow nearly instantaneously filling the wells on the chip[31]. However, it would be preferable to minimize the amount of time it takes for this to occur. By taking about 2.5 times the nucleotide injection rate to 10 mL/s, coverage time shrinks by a factor of 2.5, as shown in the tabulated times in the right of Table 7, while maintaining a reasonable, laminar Reynolds number.

**Table 7: Flow times to fill chip volume**

| Volume of Chip (mL) | Volumetric Flow In (mL/sec) | Reynolds Number | Time to Fill (sec) |
|---|---|---|---|
| 0.474 | 4 | 183 | **0.12** |
| 0.474 | 10 | 458 | **0.0474** |

Next, the flow of nucleotides into individual wells needs to be considered. There are two possible means for nucleotides to enter these wells; via diffusion or by a CSTR-like model. Using the diffusivity constant for nucleotides in neutral water, the mean time for nucleotides to diffuse into the well is found by the square of the characteristic length divided by the diffusivity constant. For the CSTR-model, the volumetric flow rate into the well with zero concentration of nucleotides is determined in addition to the residence time, τ. The flow is modeled as a parabolic flow, the velocity is taken at a characteristic height equal to the well diameter, and the volumetric flow rate into an well, given its opening area, is found. The residence time is the open volume of a well divided by the volumetric flow rate into the well. Using a simple exponential expression, $Concentration * (1 - e^{-\frac{time}{\tau}})$, with the results shown in Table 8, it was determined that the time to reach the concentration of nucleotide flow is greater than the diffusion time. Therefore, it is assumed that diffusion dominates, and the nucleotides quickly fill each well within milliseconds.

**Table 8: Times for wells to reach nucleotide flow concentration**

| **Method** | Diffusivity Constant ($m^2/s$) | Characteristic Length (μm) | Volumetric Flow Rate into Well (mL/s) | Residence Time (sec) | Time to Reach Flow Concentration (sec) |
|---|---|---|---|---|---|
| Diffusion | $3.68 \times 10^{-10}$ | 0.70 | --- | --- | **0.0011** |
| CSTR | --- | --- | $5.385 \times 10^{-11}$ | 0.0033 | **0.020** |

[31] Bustillo, J., W.

### *5.C.II. NUCLEOTIDE INSERTION, SIGNAL GENERATION AND ATTENUATION*

The nucleotide insertions are next in the workflow. This is the time it takes for complete reactions of nucleotides at each position in the strands. It will be shown later in the Kinetics section, Chapter 6, that this is on the scale of tens of milliseconds or less (~0.02 seconds). During this reaction time, the signal is being generated, which takes around tenths of a second (~0.1 sec). After all the nucleotide insertions are completed, the ISFET sensors require settling times. Given the parameters stated in the literature and the material used (silicon nitride), as covered in the ISFET section, Chapter 7, of this report, such a settling time can be about 6 seconds.

### *5.C.III. BUFFER FLOW*

Although the signal takes significant time to attenuate, the nucleotides should be washed out with a buffer solution prior to complete signal attenuation. Leaving the nucleotide solution in place for too long may increase the probability of incorrect base insertion; beginning the flow after a period of time sufficient to allow complete reaction will allow full removal of any nucleotides during the remaining time allocated for signal attenuation. This is an important step in the workflow; if any nucleotides are left over from a previous flow cycle, they could insert erroneously and lead to faulty sequences. Following the CSTR model, given a mean residence time of the volume of the chip divided by the buffer flow rate, 10 mL/s, it would take about 2 seconds to for the buffer flow to completely remove all nucleotides. This is much less than the buffer run time of 4.5 seconds required if run at the suggested 4 mL/s rate. Running the buffer during the period of signal attenuation would be sufficient to ensure no leftover nucleotides. Overall, the time scales for each element of a sequencing cycle are outlined in Table 9.

**Table 9: Overview of Time Scales for Elements in the Chip Workflow**

| Step | Time (sec) |
|---|---|
| Flow in | ~0.05 |
| Diffusion into Well | ~0.001 |
| Nucleotide Insertion | ~0.02 |
| Signal Generation | ~0.1 |
| Signal Attenuation | ~6 |
| Removing All Nucleotides from Well | ~2 |
| Overall Sum | ~8.17 |

Therefore, the time for a cycle is about 8 seconds as the signal attenuation part of the workflow dominates this time figure. The number of flows should be greater than the base length of the strands multiplied by the four possible bases. As a sufficient buffer, the number of cycles can be doubled to ensure complete sequencing. Table 10 shows the sizing calculations at 8 seconds per cycle, and includes a barcoding variation, as covered in Section 4.A.iii.3), which would allow multiplexing genomes on one run and increase the overall base length and run times.

**Table 10: Time for Genome Sequencing for various setups**

| Description | Per Cycle (sec) | Base Length | Cycles | Time (hr) |
|---|---|---|---|---|
| **Proton II - No Barcode** | 6 | 200 | 1600 | **3.56** |
| **Proton II - Barcode** | 6 | 240 | 1920 | **4.27** |

The sequencing times correlate with public Ion Torrent data and marketing pitches. The base sequencing in the actual Proton machine takes approximately 4 hours of the 8 hour Proton machine run time[32]. The remaining time outside of sequencing is devoted to base-calling, alignment, and genome reconstruction.

## 5.D. POTENTIAL CHIP THROUGHPUT

IonSeq's bottom line will rely upon the overall throughput of genomes at an appropriate cost. Back of the envelope calculations can give a realistic idea of the throughput for one sequence

[32] Ion Torrent. "The Ion Proton System: Rapid genome-scale benchtop sequencing. Specification Sheet." 2012. <www.lifetechnologies.com/proton>.

run on one Proton machine. The key parameters for determining overall throughput are the read lengths and the percent of active wells. Read length will refer to the length of the DNA template fragments on the spheres, and as a rule of thumb, 90% of total wells on the chip are active. The remaining inactive wells are used baseline readings for signal processing. Throughput is defined in Equation 3 and the number of human genomes is expressed in Equation 4. The human genome is taken to be 3 billion bases in length and coverage is defined as the average number of times a nucleotide in a template has been read. Higher coverage leads to greater accuracy in realignment.

**Equation 3**

$$Throughput\ (GB) = \#\ of\ active\ wells * \frac{Read\ length}{10^9}$$

**Equation 4**

$$\#\ of\ Genomes\ Sequenced = \frac{Throughput\ (GB) * 10^9}{Coverage * Genome\ Size}$$

The Proton I chip is insufficient for sequencing a full human genome, as seen below, even over varying read lengths at 30x coverage. For the Proton II chip, the four-fold increase in the number of wells allows it to handle a human genome. Table 11 lays out the expected throughput for different chip layouts. A hypothetical, 1 billion well, "Proton III" chip is the only arrangement of the three that can handle more than two full human genomes, given the use of barcoding during the sequencing run.

**Table 11: Throughput of Different Proton System Chips**

| | Wells ($10^6$) | % Active | Read length | Throughput (GB) | Coverage | # Human Genomes |
|---|---|---|---|---|---|---|
| **Proton I** | 165 | 0.9 | 100 | 14.85 | 30 | 0.165 |
| | 165 | 0.9 | 200 | 29.7 | 30 | 0.33 |
| | 165 | 0.9 | 300 | 44.55 | 30 | 0.66 |
| **Proton II** | 660 | 0.9 | 100 | 59.4 | 30 | 0.66 |
| | 660 | 0.9 | 200 | 118.8 | 30 | 1.32 |
| | 660 | 0.9 | 300 | 178.2 | 30 | 1.98 |
| **"Proton III"** | 1,000 | 0.9 | 100 | 90 | 30 | 1 |
| | 1,000 | 0.9 | 200 | 180 | 30 | 2 |
| | 1,000 | 0.9 | 300 | 270 | 30 | 3 |

Areas for significant advancement include shifting to a smaller manufacturing node; this would allow for even more wells on the same size chip. For example, at the 32 nm node, 1.5 billion wells,

each of diameter 0.54 μm at 0.60 μm pitch, can be achieved on a single chip; at a read length of 200 bases and 30x coverage, this arrangement can handle 3 full human genomes.

## 5.E. CONCLUDING THOUGHTS ON CHIP ORGANIZATION AND THROUGHPUT

After exploring various chip configurations, evaluating each step of the chip workflow, and generating back of the envelope throughput calculations, several important findings come to light. First, the well diameter of 0.70 μm was determined for the Proton II chip; this value will be used in the kinetics and signal generation sections, Chapters 6 and 7. Second, that the bottleneck in the chip workflow is the signal attenuation. In Chapter 9, optimization options will consider different sensor materials that will work to decrease this attenuation time and decrease overall cycle times. Third, the Proton II is only able to sequence one genome on a chip at a time while a hypothetical Proton III, with one billion wells, may be able to handle 2 human genomes on a single chip.

# 6. KINETICS OF NUCLEOTIDE INSERTIONS

Understanding the behavior of nucleotide (dNTP) incorporation onto the prepared DNA strand template is important to modeling the workflow on the Proton chip and designing different potential configurations for increased throughput. Building the kinetic model for each base incorporation event will prove instrumental in forming the foundation for the signal generation covered in the ISFET section. This chapter compares the model results with Ion Torrent published literature and allows IonSeq to further solidify the validity of these models.

## 6.A. NUCLEOTIDE KINETICS BACKGROUND

The basic mechanism is the nucleophilic attack on the phosphorous of the unbound nucleotide, by a hydroxyl group on the nucleotide in the template strand. A phosphodiester bond joins the two nucleotides, creating a pyrophosphate leaving group and producing a proton as shown in Figure 10. The proton per base incorporated is the measured variable by the semiconductor technology in the Proton chip.

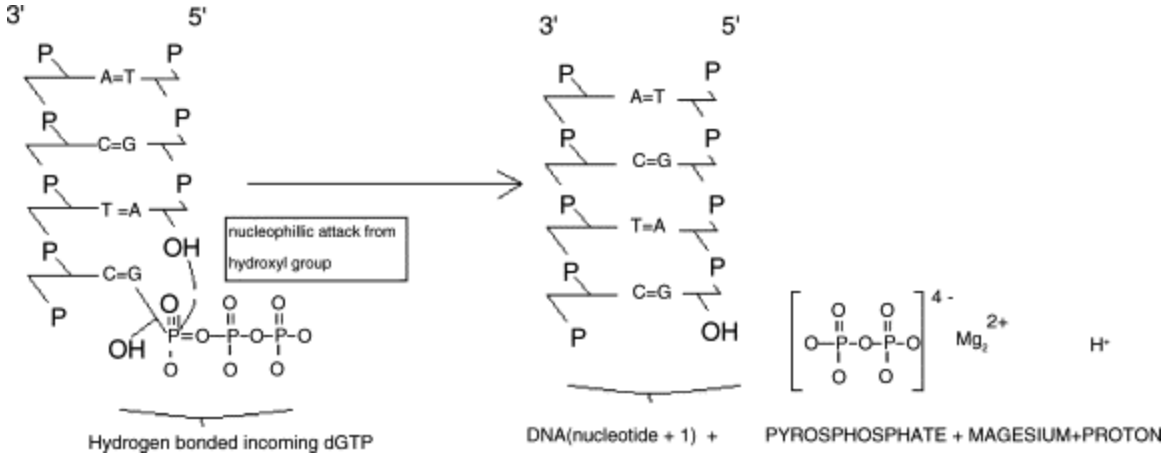



**Figure 10: The mechanism for nucleotide incorporation onto a DNA template strand illustrates the production of protons which is key to Ion Torrent technology.**

Ion Torrent employs proprietary DNA polymerases, tuned for rapid sequencing with low error rates and no proofreading ability[33]. IonSeq has been unable to reverse engineer the rate data for the polymerases due to the numerous factors involved, which included ISFET sensor dynamics. Consequently, relatively fast DNA polymerase data was sought, which provided high fidelity and rapid sequencing times in order to maximize accuracy and throughput. Human mitochondrial DNA polymerase fit the desired characteristics and research has yielded comprehensive kinetic data for matches (italicized values) and mismatches of then nucleotides as shown in Table 12[34]. A potential alternative would be viral DNA polymerase or similar, such as Phi 29, which are typically much more rapid[35]. However, comprehensive kinetic information was lacking. The use of human mitochondrial DNA polymerase will serve as the basis of the design, but IonSeq will contribute R&D resources to developing proprietary polymerase that will give it the competitive advantage.

The following rate information for human mitochondrial DNA polymerase was collected from extensive kinetic work performed at the University of Texas Austin[36]. The rate constant of

[33] Rothberg, Jonathan M, et al. "An Integrated Semiconductor Device Enabling Non-optical Genome Sequencing." *Nature* 475 (2011): 348-52. Print.
[34] Johnson, Allison A., and Kenneth A. Johnson. "Fidelity of Nucleotide Incorporation by Human Mitochondrial DNA Polymerase." *The Journal of Biological Chemistry* 276.41 (2001): 38090-8096.
[35] Esteban, Jose A., Margarita Salas, and Luis Blanco. "Fidelity of Phi29 DNA Polymerase." *The Journal of Biological Chemistry* 268.4 (1993): 2719-726.
[36] Johnson, Allison A., and Kenneth A. Johnson. "Fidelity of Nucleotide Incorporation by Human Mitochondrial DNA Polymerase." *The Journal of Biological Chemistry* 276.41 (2001): 38090-8096.

polymerase is given by $k_{pol}$, and the substrate concentration at half maximum is given by $K_D$. Both of these parameters will be used in the standard Michaelis-Menten mechanism for enzymes.

**Table 12 Human mitochondrial DNA rate data**

| **dNTP : Template Base** | **$K_D$** (μm) | **$k_{pol}$** ($s^{-1}$) |
|---|---|---|
| *A : T* | *0.8* | *45* |
| T : T | 57 | 0.013 |
| C : T | 360 | 0.038 |
| G : T | 70 | 1.16 |
| *C : G* | *0.9* | *43* |
| A : G | 250 | 0.042 |
| T : G | 200 | 0.16 |
| G : G | 150 | 0.066 |
| *T : A* | *0.6* | *25* |
| C : A | 540 | 0.1 |
| G : A | 500 | 0.05 |
| A : A | 25 | 0.0036 |
| *G : C* | *0.8* | *37* |
| A : C | 160 | 0.1 |
| C : C | 140 | 0.003 |
| T : C | 180 | 0.012 |

## 6.B. NUCLEOTIDE DIFFUSION

The protons that are produced from nucleotide incorporation can follow a few different paths the instant after it is produced: diffusing out of the well, remaining in the well, or remaining in the well and interacting with the ISFET. Due to the minuscule volume of each well, which is on the order of picoliters, diffusion becomes the predominant effect. The time scale for a proton to diffuse out of the well is approximately modeled as the square of the length scale divided by the diffusivity of the proton in water as shown in Equation 5. The diffusion of nucleotides is also an important consideration, and this mean residence time is given in Equation 6. Viscosity and pH effects on this diffusivity value are other parameters to consider, but in the context of this design project, they are assumed to not have an effect due to the relative stability of these values throughout the process. For a well size of 0.70 um and the proton diffusivity of water of $9 \times 10^{-9}$ $m^2/s$, the mean residence time for a proton is on the order of $10^{-5}$ seconds.

**Equation 5**

$$Mean\ Proton\ Residence\ Time = \tau_p = \frac{L^2}{D_p}$$

**Equation 6**

$$Mean\ Nucleotide\ Residence\ Time = \tau_n = \frac{L^2}{D_n}$$

Using the production rate of protons and the impulse behavior – a first order decay process – of proton diffusion out of the well, with a characteristic mean residence time outlined above, a convolution of the two functions can be performed to create an overall function of proton count in the well as a function of time. This will be used as part of the modeling of ISFET signal generation. Furthermore, the mean nucleotide residence time can be employed to understand how nucleotides can diffuse into the well. The basic convolution process is outlined in Equation 7, where *x(t)* is the number production of the protons from nucleotide incorporation, *h(t)* is the impulse function of proton diffusion out of the well, and *y(t)* is the number of protons left in the well after diffusion is taken into account. These expressions will be outlined in the Kinetics Model section.

**Equation 7**

$$y(t) = x(t)^* h(t) = \int_0^t x(\lambda)h(t-\lambda)d\lambda$$

## 6.C. KINETICS MODEL

Nucleotide incorporation and generation of protons can be interpreted as a binding model of nucleotides to the template that follows pseudo-first order kinetics, as shown in Equation 8. Following the binding of the polymerase in the pre-sequencing process, nucleotides attach according to Michaelis-Menton kinetics, where the observed rate constant depends upon the concentration of nucleotides in Equation 9. The concentration of nucleotides at a given time is shown in Equation 10 and the corresponding concentration of protons is illustrated in Equation 11. The model will use the rate data for the insertion of nucleotide A to the base T on a template strand.

**Equation 8**

$$\frac{d[dNTP]}{dt} = -k_{obs}[dNTP]$$

**Equation 9**

$$k_{obs} = \frac{k_{pol}[dNTP]}{K_D + [dNTP]}$$

**Equation 10**

$$[dNTP] = [dNTP]_0 e^{-k_{obs}t}$$

**Equation 11**

$$[H^+] = [dNTP]_0 - [dNTP]$$

In these expressions, the concentrations and $K_D$, the binding constant, are reported in μM and the rate constant for polymerization, $k_{pol}$, and overall rate constant, $k_{obs}$, are in units of $s^{-1}$.

Acknowledging that the observed rate constant has nucleotide concentration dependence, the more accurate method would be to integrate as shown in Equation 12 and Equation 13.

**Equation 12**

$$\frac{d[dNTP]}{dt} = \frac{k_{pol}[dNTP]^2}{K_D + [dNTP]}$$

**Equation 13**

$$k_{pol}^{-1}\left(\ln[dNTP] - \frac{K_D}{[dNTP]}\right) + C = t$$

However, it was observed that the observed rate constants do not change significantly within the range of nucleotide concentrations expected. As a result, it is assumed that the observed rate constant at the initial concentration can be used throughout the kinetic model.

Using the kinetics model, IonSeq can also address homopolymers, or stretches of DNA that have the same base code. For example, if the strand has a 3-base homopolymer sequence of 'AAA,' the kinetics of nucleotide insertion over that stretch of bases will differ than that for a single 'A' base or a 2-base homopolymer sequence 'AA.' The 1-base case will be considered first.

## *6.C.I. N = 1 CASE*

For the incorporation of one nucleotide, the n = 1 case, Equation 10 and Equation 11 yield the behavior shown in the top and bottom of Figure 11, respectively.

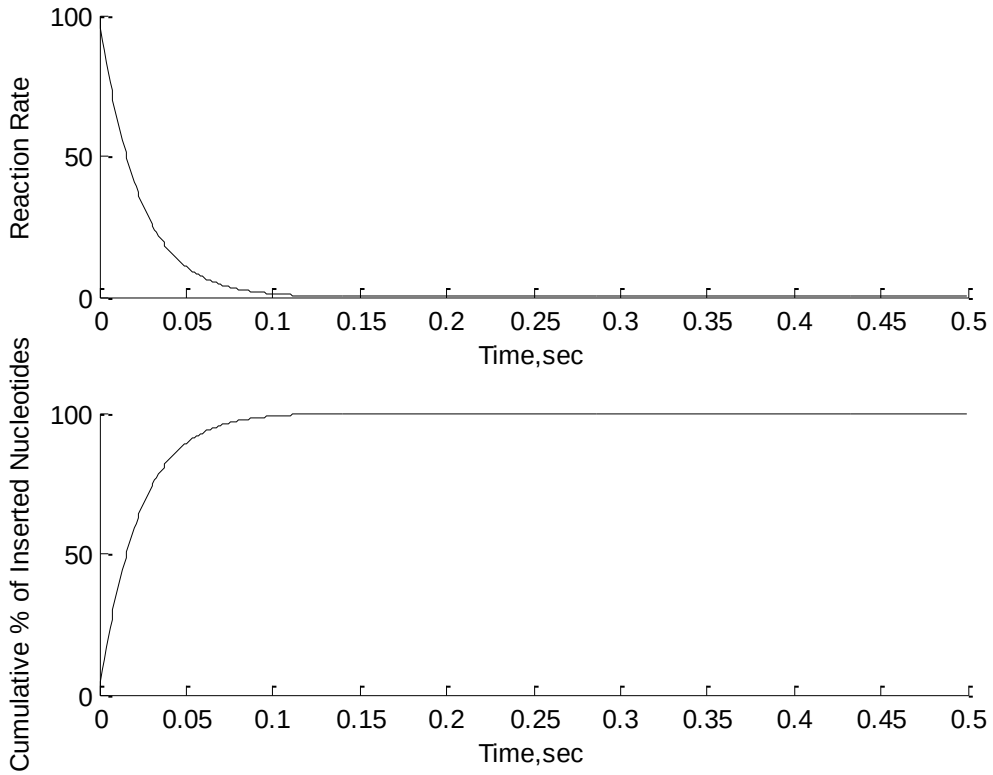



**Figure 11: (top) For one nucleotide incorporation, the kinetics follow a straightforward, first order decay. (bottom) The response of inserted mucleotides, and therefore protons generated, is shown as the percent of possible insertion events.**

Equation 11, which expresses the resulting proton concentration, can be taken to be the cumulative percent of the possible nucleotide insertion events by simply the whole expression by the initial nucleotide concentration, $[dNTP]_0$. For example, if there are 100,000 template strands, by a time of 0.05 seconds, the nucleotides will have inserted into 85% of the total strands. Consequently, 85,000 protons will have been produced as each nucleotide insertion produces one proton. By a time of 0.25, all of the strands will have had a nucleotide inserted, and therefore, 100,000 protons will have been produced. Substituting Equation 11 into Equation 10, and multiplying by the number of strands, yields Equation 14. The $(1-e^{-kt})$ expression is the percent of strands that have had undergone nucleotide insertions. Figure 13 illustrates the behavior of proton generation for this n = 1 case.

**Equation 14**

$$x(t) = \#\ of\ H^{+}Produced = \ Strands * \left(1 - e^{-k_{obs}t}\right)$$

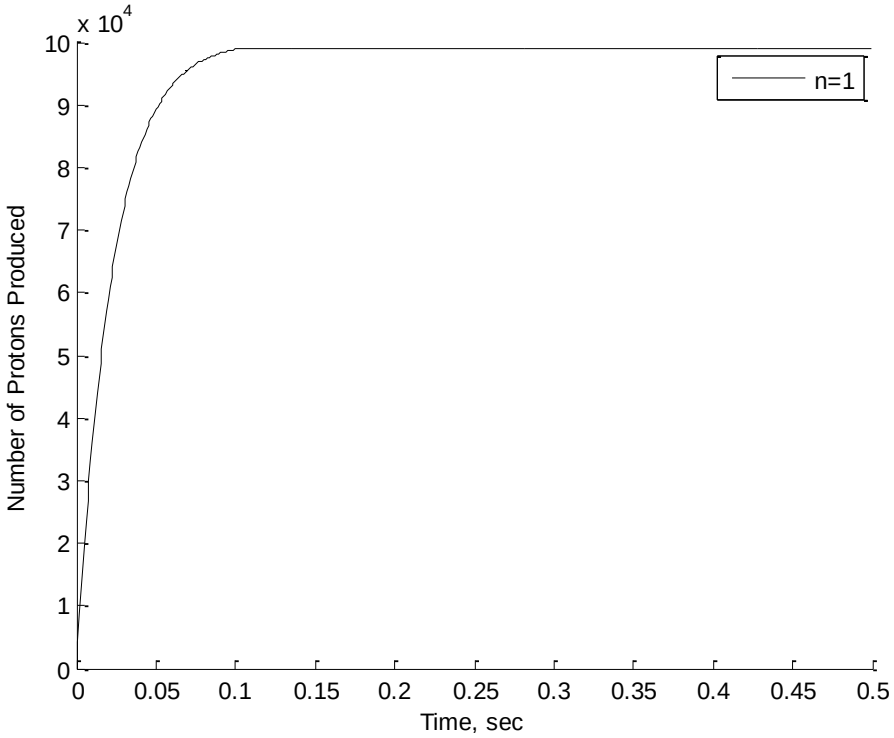



**Figure 12: Number of protons produced over time for the N = 1 case.**

However, this does not yet take into consideration the significant diffusion of protons out of the well. The actual quantity of protons that remain in the well is significantly less than shown in Figure 12. To determine this value, the impulse of proton diffusion out of the well,

Equation 15, must be considered, where $\tau_p$ is the mean residence time for the proton in water as determined in Equation 5.

**Equation 15**

$$h(t) = e^{-\frac{t}{\tau_p}}\,dt$$

Carrying out the convolution between Equation 14 and

Equation 15 yields Equation 16, the overall number of protons left in the well.

**Equation 16**

$$y(t) = Strands * \tau_p \left(1 - \frac{e^{-k_{obs}t}}{1 - k_{obs}\tau_p}\right)$$

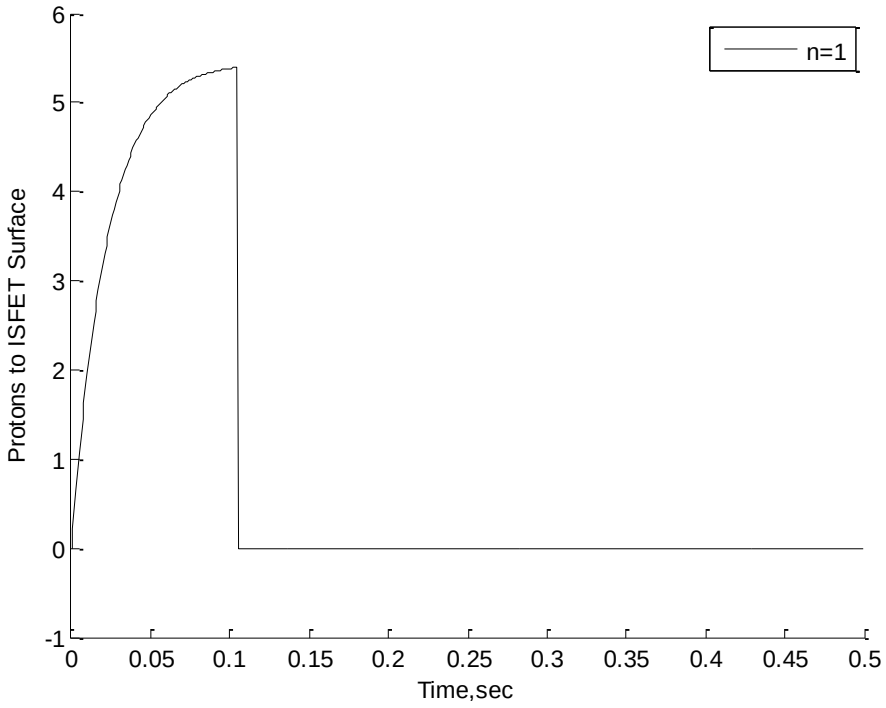



**Figure 13: This proton generation profile over time for n = 1 case shows significantly less protons available in the well due to proton diffusion out of the well.**

Figure 13 shows that, in reality, only 5 or 6 protons remain (out of the possible 100,000 protons generated) in the well after accounting for diffusion effects. The sharp drop-off evident in the same figure is the due to the fact there is no more proton production once all the possible nucleotide insertion events have occurred. However, due to the sensitivity of the ISFET sensors, as explained in Chapter 7, this small number of protons interacting with the sensor surface is sufficient for signal generation.

## *6.C.II. N = 2 CASE*

To account for homopolymers, the sequencing can be modeled along the lines of residence time distribution in series of CSTRs. All the bases in a homopolymer stretch will not react at the same time. This can be shown by taking each base position as a CSTR, and the nucleotide concentration distribution over time will impact the rate at which each nucleotide incorporates in the homopolymer stretch. There is some residence time, the inverse rate constant of base incorporation, which serves the residence time in the RTD model. Concentration, over time *t*, at each base position, *n*, may be described in Equation 17, instead of Equation 10.

**Equation 17**

$$[dNTP]_n = [dNTP]_0 e^{-k_{obs}t} * \frac{(k_{obs}t)^{n-1}}{(n-1)!}$$

For the n = 2 case, a homopolymer sequence of two bases, Equation 17 becomes Equation 18, which is plotted in green in Figure 14. To determine the overall, cumulative expression, Equation 18 needs to be summed with the expression for the n = 1 case, or Equation 10. The summed plot, in black, is shown in the top of Figure 14 and expressed in Equation 19. Equation 20 expresses the concentration of protons produced, simply the difference between the initial nucleotide concentration and cumulative nucleotide expression, Equation 19. By dividing Equation 20 by the initial concentration, the cumulative percent of inserted nucleotides can be derived and shown in the bottom of Figure 14.

**Equation 18**

$$[dNTP]_2 = [dNTP]_0 e^{-k_{obs}t} * (k_{obs}t)$$

**Equation 19**

$$[dNTP] = [dNTP]_0 e^{-k_{obs}t} + [dNTP]_0 e^{-k_{obs}t} * (k_{obs}t)$$

**Equation 20**

$$[H^+] = [dNTP]_0 - [dNTP]_0 e^{-k_{obs}t} - [dNTP]_0 e^{-k_{obs}t} * (k_{obs}t)$$

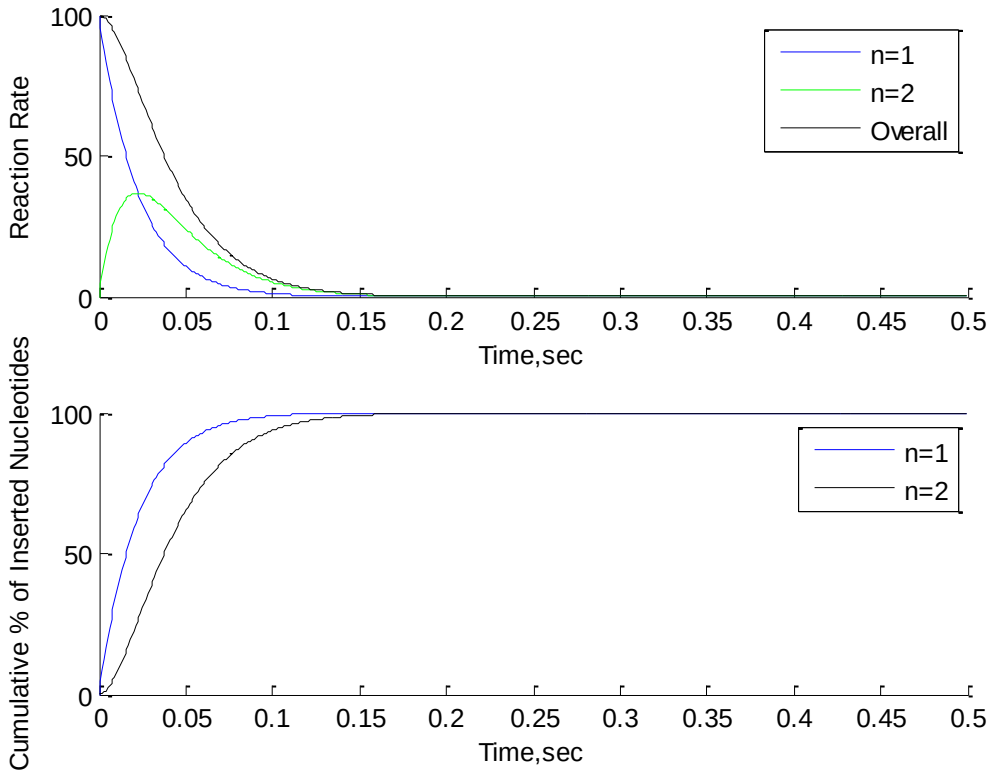



**Figure 14: (top) For the n=2 homopolymer case, the overall reaction rate, in black, is determined by the sum of the blue and green plots. (bottom) The corresponding cumulative percent of inserted nucleotides indicates greater time needed for the n = 2 case to reach the asymptote, the point where all strands have seen nucleotide insertions.**

Following the same logic and procedure of n = 1 case, the expression for total protons produced is shown in Equation 21, instead of Equation 14. The additional exponential term emerges from the CSTR-in-series model and the summation of the concentration expressions as discussed above. Because this is the 2-base homopolymer case, the greatest possible number of protons that can be produced is double the number of strands.

**Equation 21**

$$x(t) = \#\ of\ H^{+}Produced = \ 2 * \ Strands * \left(1 - e^{-k_{obs}t} - e^{-k_{obs}t} * k_{obs}t\right)$$

Figure 15 compares the production of protons between the n = 1 and n = 2 cases. As expected, the ultimate number of protons is doubled for the n = 2 case as there are twice the number of possible

nucleotide insertions in the 2-base homopolymer. Also, observing the time of approach to the asymptotes, the mean reaction time is doubled for n = 2 case.

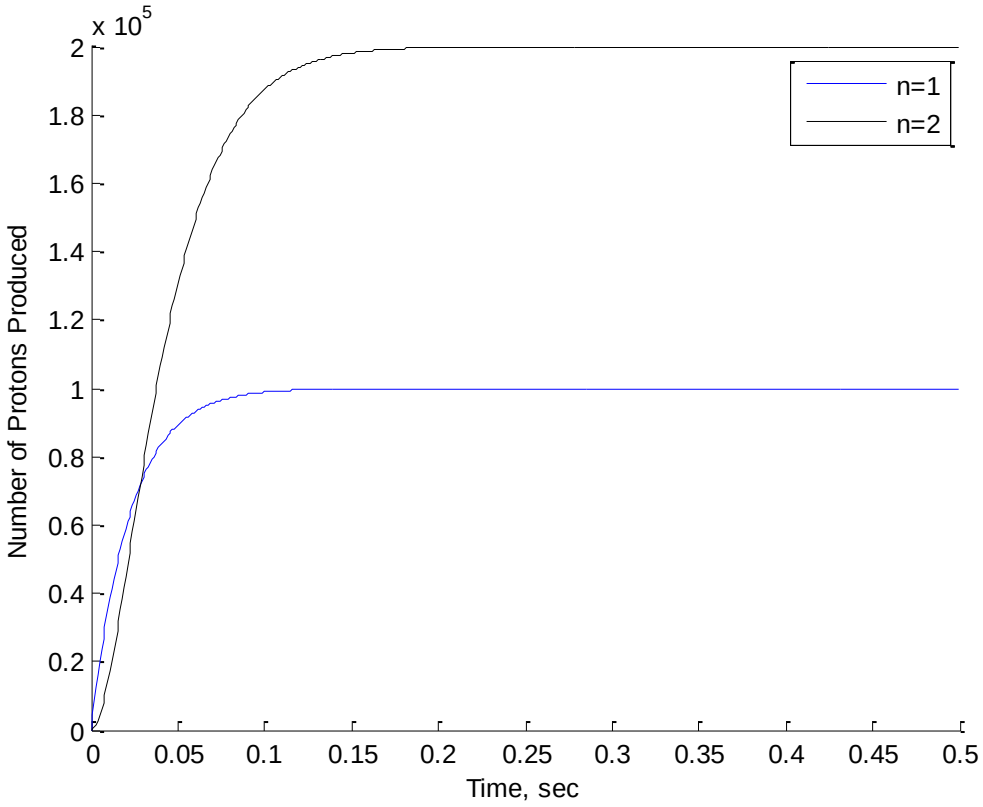



**Figure 15: Proton production for the n = 1 and n = 2 cases show that the total protons is doubled as expected. Furthermore, the mean reaction time for the n = 2 case is twice that of the n = 1 case.**

**To determine the number of protons left in the well after diffusive effects are considered, it is necessary to take the convolution of**
Equation 15, the proton diffusion impulse, with Equation 21, the total proton production expression. This results in Equation 22, and this expression is plotted along with a comparison to the n = 1 case in Figure 16.

**Equation 22**

$$y(t) = 2 * Strands * \left( \tau_p - \frac{\left(\tau_p e^{-k_{obs}t}\right)\left(k_{obs}\left(t - \tau_p(k_{obs}t + 2)\right) + 1\right)}{\left(k_{obs} * \tau_p - 1\right)^2} \right)$$

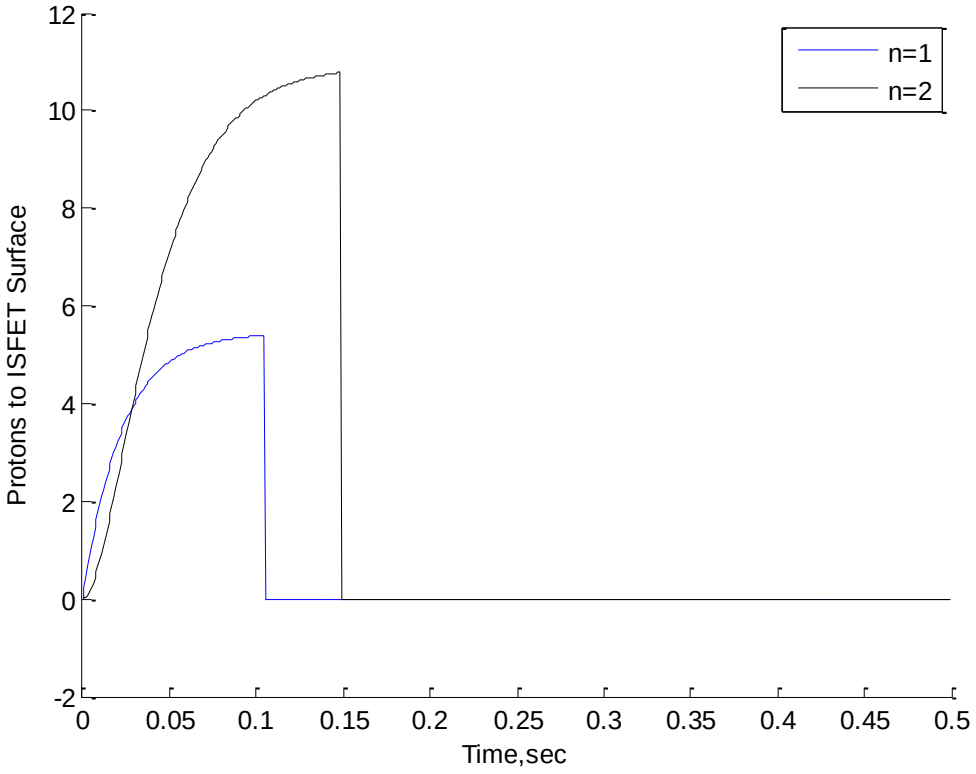



**Figure 16: Protons left in well after diffusion for the n = 1 and n = 2 cases.**

Once again, for the n = 2 case, out of the 200,000 produced protons (100,000 strands for 2-base homopolymer), only about 10 to 11 protons remain in the well after diffusion. The sharp drop off evident in the figure is attributed to the end of proton production and the diffusion of the remaining protons out of the well.

## *6.C.III. N = 3 CASE*

For the n = 3 case, a homopolymer sequence of three bases, Equation 17 becomes Equation 23, which is shown in the red plot in Figure 17. To determine the overall, cumulative expression, Equation 23 needs to be summed with the expression for the n = 1 case, or Equation 10, and the n = 2 case, or Equation 18. The summed plot is shown in the top of Figure 17 in black and in equation form in Equation 24. Equation 25 expresses the concentration of protons produced, simply the

difference between the initial nucleotide concentration and cumulative nucleotide expression, Equation 24. By dividing Equation 25 by the initial concentration, the cumulative percent of inserted nucleotides can be derived and is shown in the bottom of Figure 17 for the n = 3 case.

**Equation 23**

$$[dNTP]_3 = [dNTP]_0 e^{-k_{obs}t} * \frac{(k_{obs}t)^2}{2}$$

**Equation 24**

$$[dNTP] = [dNTP]_0 e^{-k_{obs}t} + [dNTP]_0 e^{-k_{obs}t} * (k_{obs}t) + [dNTP]_0 e^{-k_{obs}t} * \frac{(k_{obs}t)^2}{2}$$

**Equation 25**

$$[H^+] = [dNTP]_0 - [dNTP]_0 e^{-k_{obs}t} - [dNTP]_0 e^{-k_{obs}t} * (k_{obs}t) - [dNTP]_0 e^{-k_{obs}t} * \frac{(k_{obs}t)^2}{2}$$

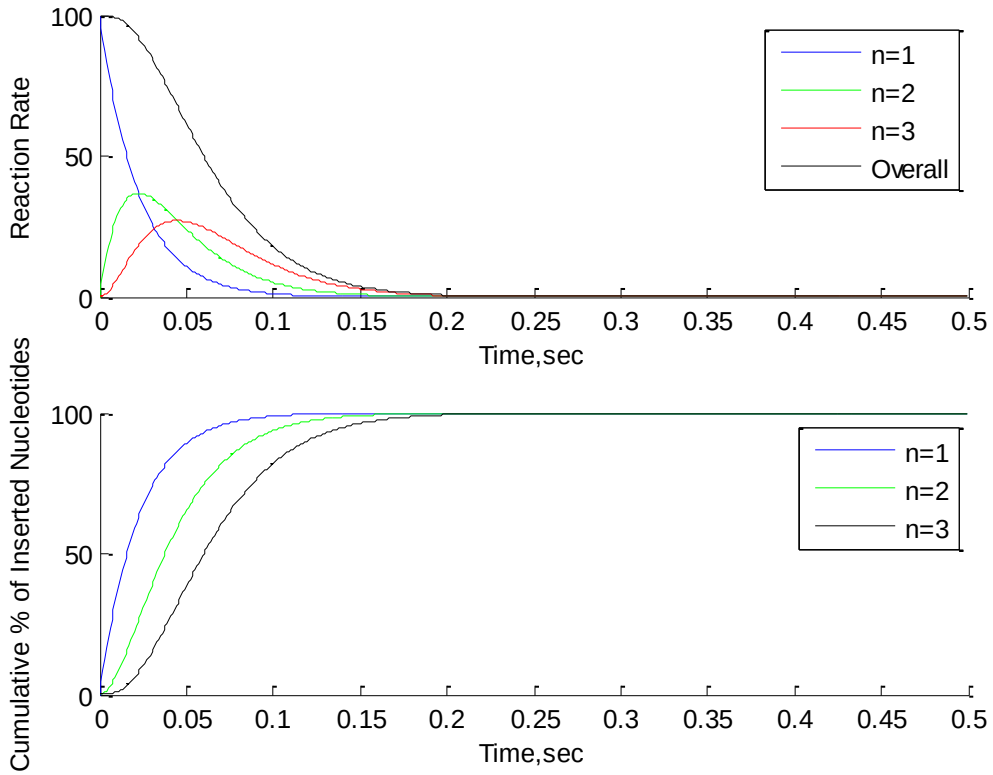



**Figure 17: For the n=3 homopolymer case, the overall reaction rate, in black, is determined by the sum of the blue, green, and red plots, generated from the CSTR-in-series model. (bottom) The corresponding cumulative percent of inserted nucleotides indicates greater time needed for the n = 3 case to reach the asymptote, the point where all strands have seen nucleotide insertions.**

Following the same logic and procedure of n = 1 and n = 2 cases, the expression for total protons produced is shown in Equation 26, instead of Equation 14 and Equation 21. Once again, there is an additional exponential term that emerges from the CSTR-in-series model and the summation of the concentration expressions as discussed above. Because this is the 3-base homopolymer case, the greatest possible number of protons that can be produced is three times the number of strands.

**Equation 26**

$$x(t) = \#\ of\ H^{+}Produced = \ 3 * \ Strands * \left(1 - e^{-k_{obs}t} - e^{-k_{obs}t} * k_{obs}t - e^{-k_{obs}t} * \frac{(k_{obs}t)^2}{2}\right)$$

Figure 18 compares the production of protons among the n = 1, n = 2, and n = 3 cases. As expected, the ultimate number of protons is tripled for the n = 3 case from the n = 1 case as there are three times the number of nucleotides to be inserted for a 3-base homopolymer. Also, observing the time of approach to the asymptotes, the mean reaction time is tripled for n = 3 case from the n = 1 case.

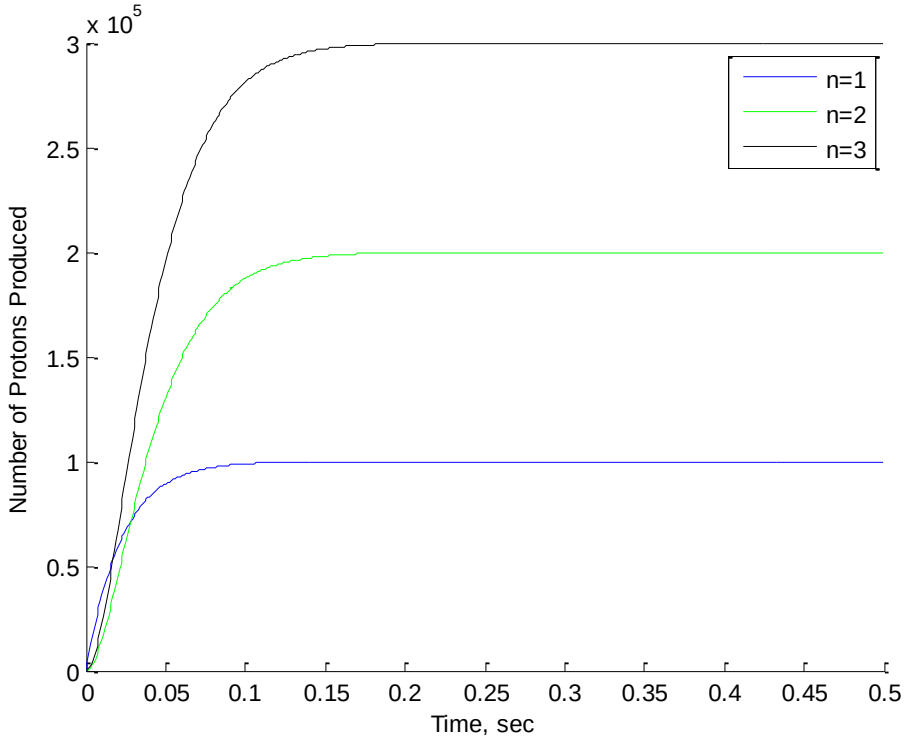



**Figure 18: Proton production for the n = 1, n = 2, and n = 3 cases show that the total protons is tripled as expected for the n = 3 case. Furthermore, the mean reaction time for the n = 3 case is three times that of the n = 1 case.**

**To determine the number of protons left in the well after diffusive effects are considered, it is necessary to take the convolution of**
Equation 15, the proton diffusion impulse, with Equation 26, the total proton production expression. This results in Equation 27, and this expression is plotted along with a comparison to the n = 1 and n = 2 cases in Figure 19.

**Equation 27**

$$y(t) = 3 * Strands * \left( \tau_p - \frac{e^{-k_{obs}t}\tau_p \left(-2 - k_{obs}(-1 + k_{obs}\tau_p)\left(-t(2 + k_{obs}t) + \tau_p(6 + k_{obs}t(3 + k_{obs}t))\right)\right)}{2\left(-1_{k_{obs}\tau_p}\right)^3} \right)$$

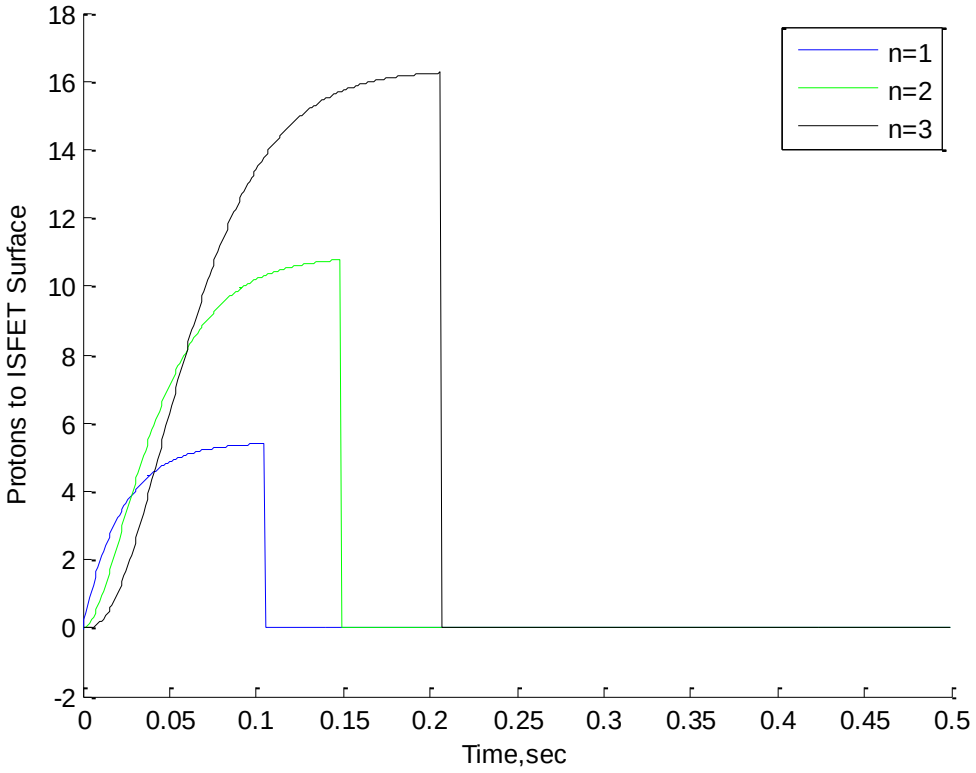



**Figure 19: Protons left in well after diffusion for the n = 1, n = 2, and n =3 cases.**

For the n = 3 case, out of the 300,000 produced protons (100,000 strands for 3-base homopolymer), only about 15 to 16 protons remain in the well after diffusion. The sharp drop off evident in the figure is attributed to the end of proton production and the diffusion of the remaining protons out of the well.

## 6.D. CONCLUDING THOUGHTS ON THE IMPORTANCE OF KINETICS

Overall, the kinetics of nucleotide insertions is significant for several reasons. To help determine the length of time to allow for a nucleotide flow over the chip, the kinetics will show how long the reactions take for various lengths of homopolymers. It is essential not to cut short the nucleotide flow at the risk of not fully sequencing the strands; furthermore, it is important not to keep the nucleotides on the chip longer than necessary as that may increase the probability of faulty insertions. From the derivations explained above, a homopolymer of 3 bases will require a nucleotide flow for about 0.2 seconds. Carrying out an extrapolation, it can be estimated that a homopolymer of five bases will require 0.3 seconds of nucleotide of flow time. Understanding that five percent of the human genome consists of homopolymers of 5 bases or longer, this is an important design parameter to consider[37]. Also, the consideration of diffusion is crucial considering the small length scale of these wells. The convolution calculations performed in this section reveal a magnitude of four decrease from the protons produced and the protons that remain in the well to be recognized by the ISFET (~100,000 protons to ~10 protons). Furthermore, the kinetic behavior of insertions will serve as a key element in the development of the signals generated from the sensors.

[37] Chan, Eugene Y. "Next-Generation Sequencing Methods: Impact of Sequencing Accuracy on SNP Discovery." DNA Medicine Institute. Web. 29 Mar. 2013.

# 7. SIGNAL GENERATION – ISFET TECHNOLOGY

In order to convert nucleotide insertions onto the template strands into interpretable data, sensors are necessary to measure the protons generated in the wells. These sensors convert this information into digital data that can be processed and reconstructed in order to derive the genome sequence. This is the area where semiconductor technology merges with genome sequencing to form the backbone of high throughput semiconductor-based sequencing. This chapter will cover details of the sensors used in IonSeq's method of sequencing and how the signal generated from these sensors can be interpreted as nucleotide insertion events.

## 7.A. ISFET BASICS

The fundamental difference that sets Ion Torrent technology apart from other "next generation" DNA sequencing competitors is its use of ISFETs, ion-sensitive field effect transistors, to measure proton concentrations in each individual well. Upon nucleotide incorporation onto the template strand, the generated protons are sensed by the ISFETs underlying each well. Through

this process, IonSeq can quickly generate sequences and take advantage of the rapid scalability of semiconductor manufacturing.

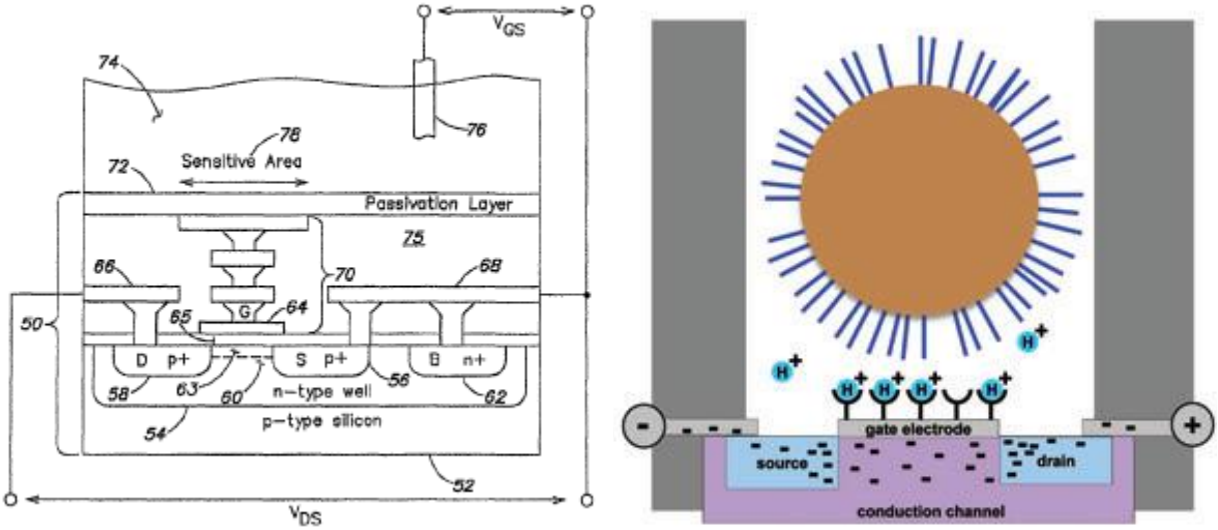



**Figure 20 a. (left) The patent drawing shows the inner construction of the ISFET sensor[38]. b. (right) This cartoon illustrates the template bead and the generated protons on the surface of the ISFET[39].**

Each ISFET is structured similarly to typical metal oxide semiconductor field effect transistors. The sensor consists of p-type regions, which contain sources and drains (56 and 58), and n-type well (54), as shown in Figure 20a. Current flows between the source and drain and is modulated by the activity upon the passivation layer, which is an ion-sensitive membrane, exposed to the analyte solution above it. Material selection in this passivation layer influences the sensor's sensitivity to specific ions. Using silicon nitride, silicon oxynitride, and other aluminum, silicon, or tantalum oxides enable the ISFET to sense protons generated from nucleotide insertions on the strands on the template bead, as illustrated in Figure 20b[40].

[38] Rothberg, Jonathan M., James M. Bustillo, Mark J. Milgrew, Jonathan C. Schultz, David Marran, Todd M. Rearick, and Kim L. Johnson. Methods and Apparatus for Measuring Analytes. Life Technologies Corporation, assignee. Patent 8263336. 11 Sept. 2012. Print.
[39] Merriman, B., Ion Torrent R&D Team, B., & Rothberg, J. (2012). Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis*, *33*, 3397-3417.
[40] Rothberg, Jonathan M., James M. Bustillo, Mark J. Milgrew, Jonathan C. Schultz, David Marran, Todd M. Rearick, and Kim L. Johnson. Methods and Apparatus for Measuring Analytes. Life Technologies Corporation, assignee. Patent 8263336. 11 Sept. 2012. Print.

At the interface between the solution and this layer, an electric potential difference ($\Psi_0$), in units of mV, develops as a direct consequence of the reactions that occur between protons and the surface groups. This potential difference is a function of solution concentration. For silicon nitride based ISFETs, the important surface reactions include a series of protonation and deprotonation reactions as shown by

Equation 28, Equation 29, and Equation 30[41]:

**Equation 28**

$$SiO^- + H^+ \leftrightarrow SiOH \;;\; \frac{[SiO^-][H^+]}{[SiOH]} = \frac{k_{b,1}}{k_{f,1}} = K_1$$

**Equation 29**

$$SiOH + H^+ \leftrightarrow SiOH_2^+ \;;\; \frac{[SiOH][H^+]}{[SiOH_2^+]} = \frac{k_{b,2}}{k_{f,2}} = K_2$$

**Equation 30**

$$SiNH_2 + H^+ \leftrightarrow SiNH_3^+ \;;\; \frac{[SiNH_2][H^+]}{[SiNH_3^+]} = \frac{k_{b,3}}{k_{f,3}} = K_3$$

The kinetics of each of these chemical reactions should be considered for a comprehensive model. However, for the purposes of this design, it is assumed that proton donation reactions dominate the transient response of the signal over the others and that pH changes are small, implying near constant surface potentials and allowing for the linearization of surface reaction equations. Our model shows a pH drop of ~0.4 per base incorporation, which is relatively consistent with Ion Torrent literature stating ~ 0.2 pH drop. [42] This validates this key assumption.

## 7.B. Double Layer Capacitance

A "double-layer capacitance" forms as a result of the physical limitations of the ions approaching the ISFET surface; these particles cannot come any closer than their ionic radius as illustrated in Figure 21. Charge densities, in units of coulombs, on either side of this double layer —

[41] Woias, P., L. Meixner, D. Amandi, and M. Schönberger. "Modelling the Short-time Response of ISFET Sensors." *Sensors and Actuators B: Chemical* 24.1-3 (1995): 211-17. Print.
[42] Merriman, B., Ion Torrent R&D Team, B., & Rothberg, J. (2012). Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis*, *33*, 3397-3417.

in the solution, $\sigma_{dl}$, and on the surface, $\sigma_0$ — are related by the double layer capacitance, $C_{dl}$, in units of farads, and the surface potential, $\Psi_0$. A change in the charge density on the solution side of the double layer is not immediately recognized by the charge density on the surface side; by Equation 31 below, this change forces surface potential to change.

**Equation 31**

$$\sigma_0 = C_{dl}\Psi_0 = -\sigma_{dl}$$

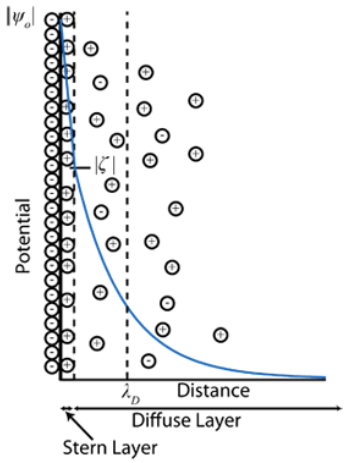



**Figure 21: Due to the ions' shape, they can come no closer than their ionic radius, forming a double layer capacitance, identified as the Stern Layer, above the ISFET surface.[43]**

This double layer capacitance is important in generating the signal needed for recording each base insertion event. Surface potential, $\Psi_0$, is calculated by considering the sensitivity of that potential to pH changes in Equation 32 and Equation 33.

**Equation 32**

$$\Psi_0 = \left(pH_{bulk} - pH_{pzc}\right) * \frac{\Delta\Psi_0}{\Delta pH}$$

**Equation 33**

$$\frac{\Delta\Psi_0}{\Delta pH} = -2.3 * \frac{k_B T}{q}\alpha$$

In these expressions, $\alpha$ is a dimensionless sensitivity parameter—a typical value for silicon nitride is 0.93—$k_B$ is the Boltzmann constant, q is the elementary charge, and T is temperature. $pH_{pzc}$ is the pH at point of zero charge, a material-dependent parameter, which is the pH at which

[43] "Electrokinetics." MIT Laboratory for Energy and Microsystems Innovation, n.d. Web. <http://web.mit.edu/lemi/rsc_electrokinetics.html>.

there is no surface potential on the ISFET surface. The sequencing will be run at pH of 8, which will play a key role in optimizing attenuation time as explained in the Optimization section, Chapter 9. Sample materials and their characteristic values are shown in Table 13. IonSeq will take silicon nitride as the base case material as expressed in Ion Torrent patent[44], but other materials will be explored in the Optimization section.

**Table 13: Collection of parameters for various ISFET materials[45]**

| Metal | Oxide/Nitride | $pH_{pzc}$ | $\Delta V_{TH}$ (mV/pH) | Theoretical $\Psi_0$ (mV) @ pH = 9 |
|---|---|---|---|---|
| Al | $Al_2O_3$ | 9.2 | 54.5 (35° C.) | –11 |
| Zr | $ZrO_2$ | 5.1 | 50 | 150 |
| Ti | $TiO_2$ | 5.5 | 57.4-62.3 (32° C., pH 3-11) | 201 |
| Ta | $Ta_2O_5$ | 2.9, 2.8 | 62.87 (35° C.) | 384 |
| Si | $Si_3N_4$ | 4.6, 6-7 | 56.94 (25° C.) | 251 |
| Si | $SiO_2$ | 2.1 | 43 | 297 |
| Mo | $MoO_3$ | 1.8-2.1 | 48-59 | 396 |
| Hf | $HfO_2$ | 7-4-7.6 | 50-58 | 81.2 |
| W | $WO_2$ | 0.3, 0.43, 0.5 | 50 | 435 |

## 7.C. SIGNAL GENERATION AND ATTENUATION

Signal generation is modeled as the following. $\Psi_0(t)$ is the surface potential function– the actual signal –with time dependence, $\tau$ is the time constant based upon $\tau_0$ – the material-dependent theoretical minimum response time – and the bulk solution pH, and $\Delta\Psi_0$ is the amplitude of the disturbance variable—the generation of protons in the well. Equation 34 encapsulates this relationship among surface potential, amplitude, and material time constant.

**Equation 34**

$$\Psi_0(t) = \Delta\Psi_0 e^{-\frac{t}{\tau}}$$

[44] Rothberg, Jonathan M., James M. Bustillo, Mark J. Milgrew, Jonathan C. Schultz, David Marran, Todd M. Rearick, and Kim L. Johnson. Methods and Apparatus for Measuring Analytes. Life Technologies Corporation, assignee. Patent 8263336. 11 Sept. 2012. Print.
[45] Ibid.

The amplitude is a function of the molar concentration of protons at the passivation surface layer as shown in Equation 35, and the material time constant is a function of pH as seen in Equation 36, where $\tau_0$, again, is the material-dependent theoretical minimum response time.

**Equation 35**

$$\Delta\Psi_0 = \Psi_1 - \Psi_2 = \frac{\sigma_0}{C_{dl,1}} - \frac{\sigma_0}{C_{dl,2}} = \Psi_1\left(1 - \frac{C_{dl,1}}{C_{dl,2}}\right)$$

**Equation 36**

$$\tau = \tau_0 * 10^{\frac{pH}{2}}$$

This results in an involved function for signal, which essentially is a response to the generation of protons in the well that approach the passivation layer.

In these equations, $\sigma_0$ represents the surface charge density, $\sigma_{dl}$ is the charge density on the solution side of the double layer, $C_{dl,1}$ is the initial double layer capacitance, and $C_{dl,2}$ is the double layer capacitance that changes with proton generation. $C_{dl,2}$ is, in turn, a function of the Boltzmann constant, $k_B$, permittivity of free space, $\varepsilon_0$, and the Debye screening length, $\lambda$, as expressed in Equation 37.

**Equation 37**

$$C_{dl} = \frac{k_B \varepsilon_0}{\lambda}$$

The Debye screening length, Equation 38, is a function of the ionic strength of the solution, $I$, which is a function of the charge number of the ionic species, $z_s$, and the molar concentration of those ionic species, $c_s$, that are at the passivation layer, in Equation 39.

**Equation 38**

$$\lambda = \frac{0.3\ nm}{\sqrt{I}}$$

**Equation 39**

$$I = 0.5 \sum_s z_s^2 c_s$$

For this application, $z_s$, is simply +1 for protons and the molar concentrations of protons are based upon the kinetics model in Chapter 6 and the Boltzmann distribution of those protons that reach the passivation layer, as expressed in Equation 40.

**Equation 40**

$$[H^+]_{surface} = [H^+]_{bulk} * e^{-\frac{q}{k_B T}\Psi_0}$$

## 7.D. Results of the ISFET-Signal Model

The signals generated, $\Psi_0(t)$, are included below. These results are for the Proton II chip, with an estimated 0.70 µm well diameter at pitch of 0.92 µm. Figure 22 illustrates the signals generated for up to 3 base long homopolymers. The signal for each case is dependent on the number of protons left in the well as shown in Equation 39. In section 6.C.iii, the number of protons in the well after diffusion was illustrated in Figure 19, and shown again for convenience.

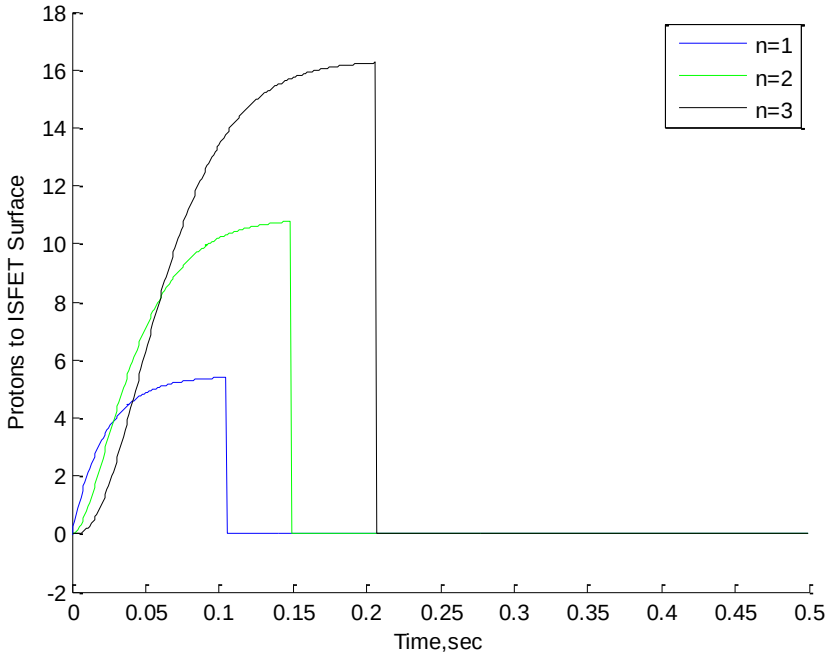



**Figure 19: Protons left in well after diffusion for the n = 1, n = 2, and n =3 cases.**

From the kinetic model results, the signals derived from protons interacting with the ISFET sensors help identify the extent of nucleotide insertions on the template strands. The MATLAB code used to

generate these results are found in Appendix E. The results from Ion Torrent literature, as seen in the right of Figure 22, strongly confirm the validity of the IonSeq model. The signal peak is achieved under half a second, and signal attenuation makes up a significant part of the model. The signal attenuation time can be taken as the time it takes the signal to decrease to just 5% of its peak value; for the example provided, the attenuation time is approximately 6 seconds.

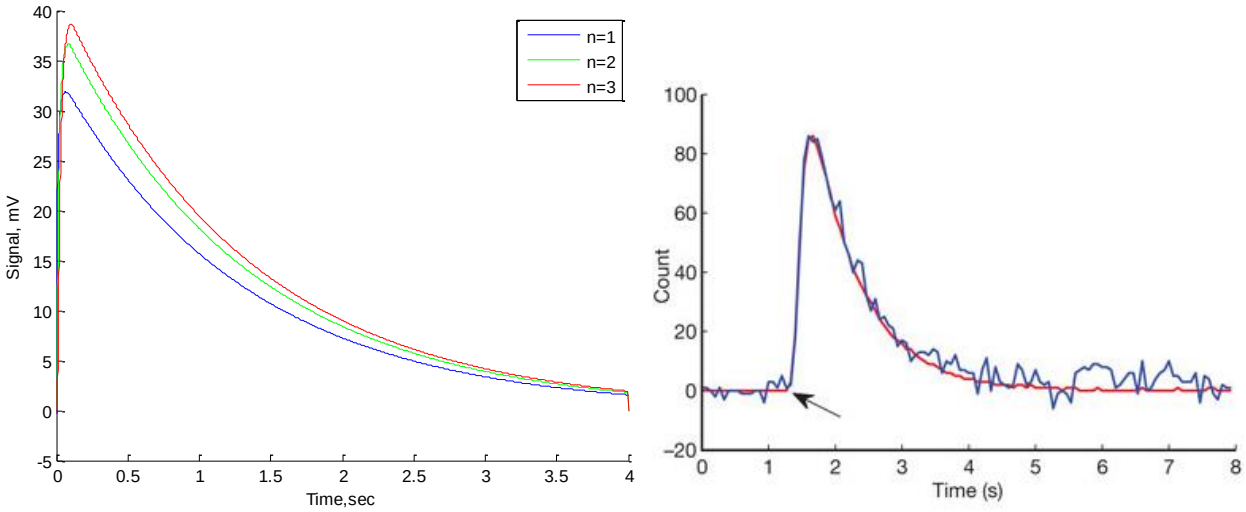



**Figure 22**: **(Left) Signal generation from the change in proton concentration show increasing amplitudes for longer homopolymers. (Right) The experimental and model signal generated by Ion Torrent reaffirms the validity of IonSeq's model shown on the left.[46]**

In reality, because these ISFET sensors are sensitive, the flows of the nucleotide solutions and wash buffers contribute very much to background noise when collecting nucleotide incorporation information. As the bead loading process is probabilistic, typically 10% of the wells are not occupied after the loading process. While this may slightly decrease overall throughput, the non-template bearing wells are important in providing baseline readings for the other template-bearing wells. These signal plots, in practice, are generated by subtracting out the baseline signal plots from confirmed empty wells from the wells with template beads. The model developed here does not include these complexities, but is effectively illustrates the end product, the final signal, used for base calling and generating the genome's sequence.

[46] Rothberg, Jonathan M, et al. "An Integrated Semiconductor Device Enabling Non-optical Genome Sequencing." *Nature* 475 (2011): 348-52. Print.

## 7.E. SHOT NOISE

Due to the nature of the protons interacting with the surface, fluctuations in the current generated in the ISFET create shot noise, which may have an effect on signal fidelity. Shot noise is classified as instrumental noise and is the unavoidable result of the quantum nature of electric charge. The 'packets' of charges in the current, created by protons interacting with the sensor surface, have their own behavior. The number of electrons, which cross the sensor junction in a particular time interval, fluctuates and is not uniform.[47] If the number of electrons that cross the ISFET junction remains constant, there would exist an underlying, base shot noise that could easily be separated from the collected signal. Since this is not the case, noise must be modeled according to the standard deviation of the average number of protons generated, which is just the square root of the number of protons. Because the number of protons that remain in the well, as calculated in the Equation 16, is just an average, and the number of protons in the well, in reality, fluctuates drastically, the total number of protons produced is used in these noise calculations. Therefore, this model dictates that the signal to noise ratio (SNR) is equivalent to the number of protons generated divided by the square root of that value.

**Equation 41**

$$SNR = \frac{Signal}{Noise} = \frac{N}{\sqrt{N}} = \sqrt{N}$$

Over the course of the signal generation model, SNR was calculated from the number of protons generated at that point in time. Figure 23 illustrates the signals for homopolymers up to three bases in length, with the thickness that gives the range of shot noise effect (up to 3 standard deviations) on the accuracy of the signals generated.

[47] Lesurf, Jim. "Sources of Noise." University of St. Andrews, n.d. Web. 7 Apr. 2013. <http://www.st-andrews.ac.uk/~www_pa/Scots_Guide/iandm/part3/page1.html>.

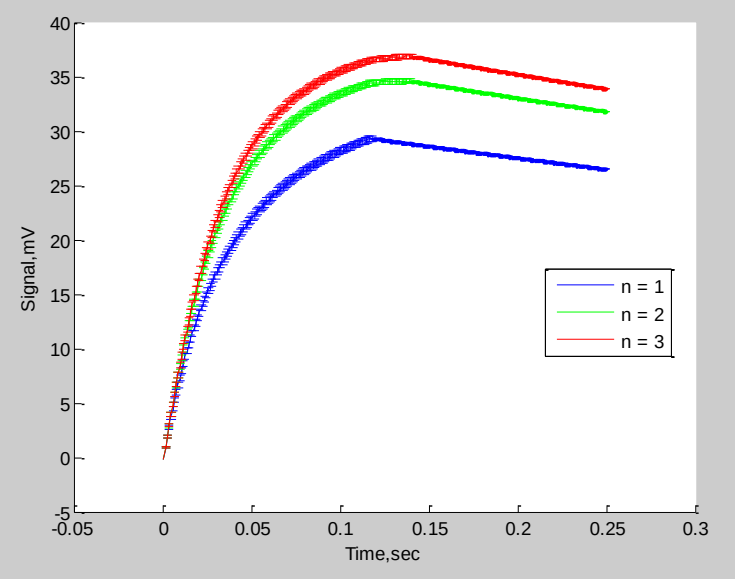



**Figure 23: The signal, for up to three homopolymers shown, is sufficiently strong and clear to allow for accurate base-calling.**

Shot noise, however, is only one of the different types of noise that can be considered. Cross-talk between wells, where protons may be able to diffuse out and into other wells may be an issue that emerges with smaller wells. Thermal noise, or Johnson-Nyquist noise, is another inevitable characteristic of signal collection from electrical conductors and is the result of the thermal agitation of electrons[48]. Further exploration in this area is needed for a more complete model of noise.

## 7.F. ISFET Signal Conclusions

The model outlined in this chapter demonstrates that distinct signals can be generated when protons are produced from nucleotide insertions. Furthermore, it has been shown that the signals for homopolymers up to length three bases are distinguishable by the different peak heights. This is important in reducing base call errors and improving overall accuracy of a sequenced genome. In addition, this model will be crucial to the development of further optimizations of this process, as explored in Chapter 9.

[48] Lesurf, Jim. "Sources of Noise." University of St. Andrews, n.d. Web. 7 Apr. 2013. <http://www.st-andrews.ac.uk/~www_pa/Scots_Guide/iandm/part3/page1.html>.

# 8. DATA ANALYSIS AND GENOME CONSTRUCTION

As the sequencing reactions take place, the signal from the released protons are generated as described in Section 7.C. This signal from each micro-well can now be used to keep track of which bases were inserted during the sequencing reaction and in which order, hence the sequence of the fragment in each well can be deduced. This can be accomplished by through base calling algorithms as discussed in Section 8.B.

However, considerable challenges are posed to signal detection by errors that may occur during the sequencing process, so it is important to be able to optimize the variables such as strand length, flow time, reaction kinetics, etc in order to produce and detect acceptable signals. A model of the sequencing process was created in MATLAB to be able to optimize these variables and the model is discussed further in Section 8.A.

Once the sequence of each fragment is obtained, it is possible to perform realignment to the reference genome, discussed in Section 8.C., thus yielding the final sequence for the entire genome.

## 8.A. DEPHASING MODEL

To model the base calling process, IonSeq created a dephasing model using MATLAB. This model seeks to simulate the sequencing process as it occurs in each well. It models the sequencing process using the Kinetic Monte Carlo Method. The model also quantifies the extent to which dephasing occurs during the sequencing process. It allows the user to specify the number of strands on the bead, the length of each strand, the number of flow cycles, the flow order of the bases, and the flow time for each base. The user may also choose to specify the concentration of the nucleotide flow. The code can be found in Appendix F.

There are hundreds of thousands of identical strands on each bead in each bead, clonally amplified from one fragment. Sequencing takes place simultaneously on each strand. In this way, a large number of protons are generated such that a perceivable signal can be generated. In theory, the sequencing should progress at the same rate on each strand, but if during a particular flow of bases, some of the correct bases fail to incorporate, the corresponding strands get out of phase with the rest of the strands. This can also happen the wrong base is incorporated into the strand. Dephasing has a negative effect on the signal, since it means that there is a lower signal for the correct base incorporation. It also contributed to noise because protons are generated during the incorrect base flow. Hence it is important to be able to predict what percentage of the strands is likely to get dephased and optimize the variables to reduce dephasing.

### *8.A.I. THE KINETIC MONTE CARLO METHOD*

The Kinetic Monte Carlo Method, employed by the model, accounts for the fact that the bases are flowed over the wells for a certain duration of time and that this amount of times affects the probability of incorporation of a match or a mismatch. The flow time of each base can be thought to be comprised of several tiny time segments. During each time segment, there is a

probability of base incorporation at the current position of the polymerase. This probability, as shown in

Equation 42 below, is a function of the observed rate constant, $k_{obs}$, and the length of the time segment, dt, which should be no greater than the inverse of the largest rate constant.

**Equation 42**

$$Probability = k_{obs} * dt$$

The observed rate constant can be calculated from the reaction rate constant, $k_{pol}$, the concentration of the nucleotides, [dNTP], and the dissociation constant, $K_D$, of the polymerase using Equation 43.

**Equation 43**

$$k_{obs} = \frac{k_{pol} * [dNTP]}{K_D + [dNTP]}$$

If, at the current position of the polymerase, the nucleotide that is flowed over the wells is a match according to the template strands, the probability of incorporation is typically higher than if the nucleotide is a mismatch. This is because the reaction rate constants for matches are typically higher than those for mismatches. This means that the mean reaction time, calculated from the inverse of the rate constant as in Equation 44 below, is lower for correct matches and higher for mismatches.

**Equation 44**

$$Mean\ reaction\ time = \frac{1}{k_{pol}}$$

The dephasing model loops through each time segment. The length of the time segments is kept smaller than the smallest mean reaction time. A base incorporation is likely to happen when enough time segments have passed that the mean reaction time is achieved.

## *8.A.II. Quantifying the Extent of Dephasing*

The model keeps track of every time there is a mismatched base incorporated and every time a base is failed to be incorporated. It stores the information regarding the position of the base in question, the strand in which it is present, the time segment and the base flow during which the error occurs. It uses this information to calculate how many strands have become dephased and to what extent. If a mismatched base is incorporated then that strand gets ahead of the others, while if a base is failed to be incorporated, that strand falls behind. Some strands experience both kinds of dephasing and the model generates a distribution of all dephased strands and the extent to which they are dephased.

## *8.A.III. Optimizing Variables*

### *8.A.III.1) Strand Length*

The model was run using a range of different read lengths, and as expected, it was found that the shorter the strand length, the less the dephasing. As the strand length gets longer, there is greater probability for errors to accumulate and hence there is greater extent of dephasing. This can be seen in Figure 24.

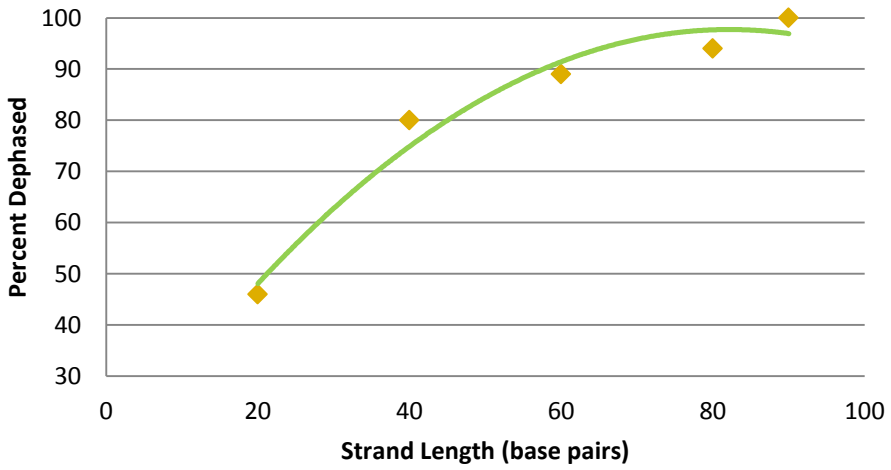



**Figure 24: As the strand length increases, the percentage of strands dephased also increases.**

While shorter read lengths can provide more accurate sequencing results, they are also more difficult to realign to a reference genome. Ion Torrent read lengths are generally 200 bp long[49]. However, as can be seen from Figure 25, given the current variables used in the Dephasing Model, the strands become 99% dephased for read lengths as short as 100 bp.

The charts below show the distribution of dephasing for read lengths of 100 bp and 20 bp. At 100 bp read lengths, not only are 99% of the strands dephased, but they are dephased by several base pairs – as many as 10, although the majority of strands are dephased by 3-5 bp. At 20 bp read lengths, only 46% of the strands are dephased, and the extent of dephasing is also considerably less, with most strands dephased only by 1 bp. The signal generated for 20 bp read lengths will be reliable because 54% of the strands will be providing the correct signal. Although 46% of the strands will be providing incorrect signals, since the strands are dephased by different amounts, the various incorrect signals will not be as strong thus will not interfere greatly with correct signal detection.

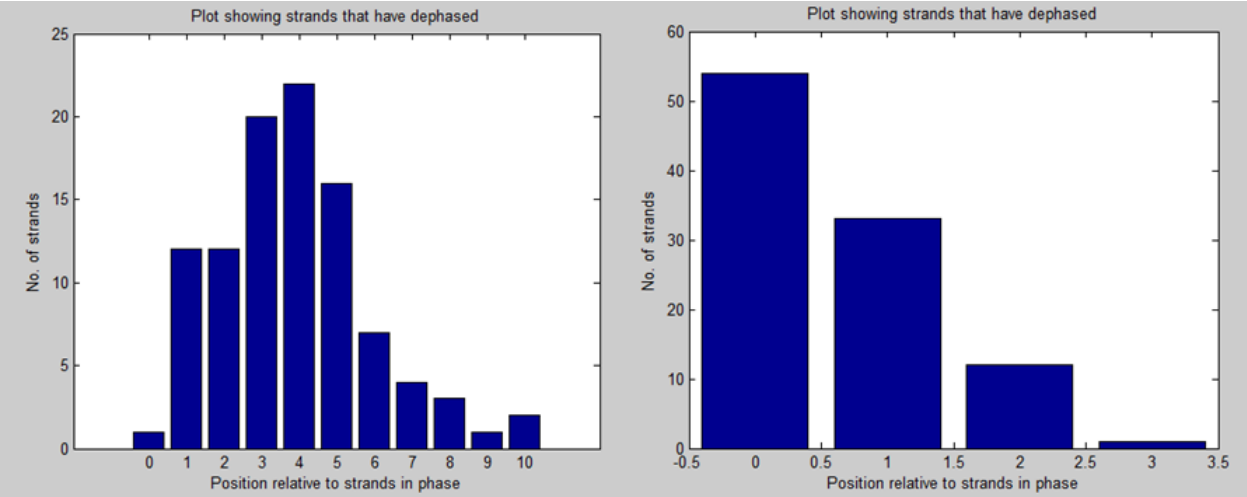



**Figure 25: Distribution of Dephased Strands.** Model run with 100 bp strands (99% dephased) (Left). Model run with 20 bp strands (46% dephased) (Right). At the given rate constants, short strand lengths of 20 bp yield acceptable dephasing. At 100 bp, dephasing becomes problematic.

However, 20 bp is not ideal for the proposed throughput, and it is reasonable to assume that the key to achieving longer read lengths lies with the kinetics of the polymerase used for

[49] Liu Lin (2012), Comparison

sequencing. This will be discussed further in Section 8.A.iii.7. While the dephasing model uses 20 bp read lengths given the polymerase rate constants used, the actual process used by IonSeq will use read lengths that are 200 bp long. With proper adjustments to the kinetics, the error rate at 200 bp can be decreased, as will be further discussed.

### *8.A.iii.2) Number of Strands*

The process of clonal amplification by emulsion PCR can yield as many as 10 million copies of the strands per bead[50]. However, the model was run primarily with only 100 copies of the strand per bead for the sake of time. The model was also run with 1000 copies per bead and it was seen that all other variables remaining unchanged, roughly the same percentage of strands are dephased and the distributions of dephased strands were very similar, as shown in Figure 26 below. Hence, it was considered safe to assume that the model can be easily scaled with respect to number of strands.

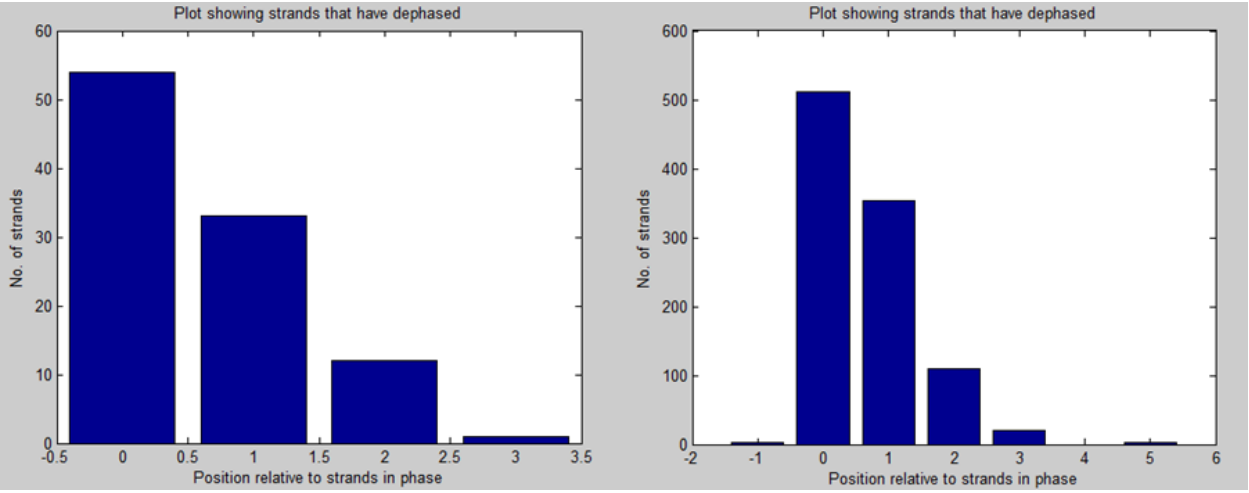



**Figure 26: Number of Dephased Strands.** Model run with 100 strands (46% dephased)(Left). Model run with 1000 strands (48% dephased)(Right). Note that distribution of dephased strands is similar.

[50] Margulies, M., Egholm, M., & Altman, W. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*, 376-280.

### *8.A.III.3) NUMBER OF FLOW CYCLES*

The number of flow cycles determines the number of times bases are flown over the chip. It stands to reason that there may be a need to flow all four bases (A, C, G and T) for each position to ensure the correct incorporation at each of them. This calls for a number of flow cycles that is four times the length of the fragments.

### *8.A.III.4) FLOW ORDER*

The flow order used in this model was A, C, G, T. Changing, reversing or alternately reversing the flow order does not have any effect on the error rate, according to the model. This follows intuition because given that the strand sequences are random, they are not biased towards any particular flow order.

### *8.A.III.5) FLOW TIME*

It is important to select an optimum flow time for the bases, or the length of time during which the bases will flow over the wells. Longer flow times allow for more errors to accumulate and shorter flow times might not allow sufficient time for the signal to build up during the flow. The model returns fewer errors for shorter flow times, but does not account for signal build-up since this aspect was not built into the model. At higher flow times, while the number of strands roughly doubles, it is interesting to note that the vast majority of errors are due to incorporations of the wrong base rather than failure to incorporate a base. This suggests that the longer time provides greater probability of incorporation and while it greatly reduces the chances of a miss, it also increases the chances of a wrong insertion, shown in Figure 27 below. As can be seen in Figure 19 and Figure 22, the time to register the signal peak is roughly 0.25 s[51] and this is the recommended

[51] Merriman, B., Ion Torrent R&D Team, B., & Rothberg, J. (2012). Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis*, *33*, 3397-3417.

value, because it is just high enough to register a signal but low enough to avoid large extents of dephasing. More favorable kinetics can allow for longer flow times as discussed below.

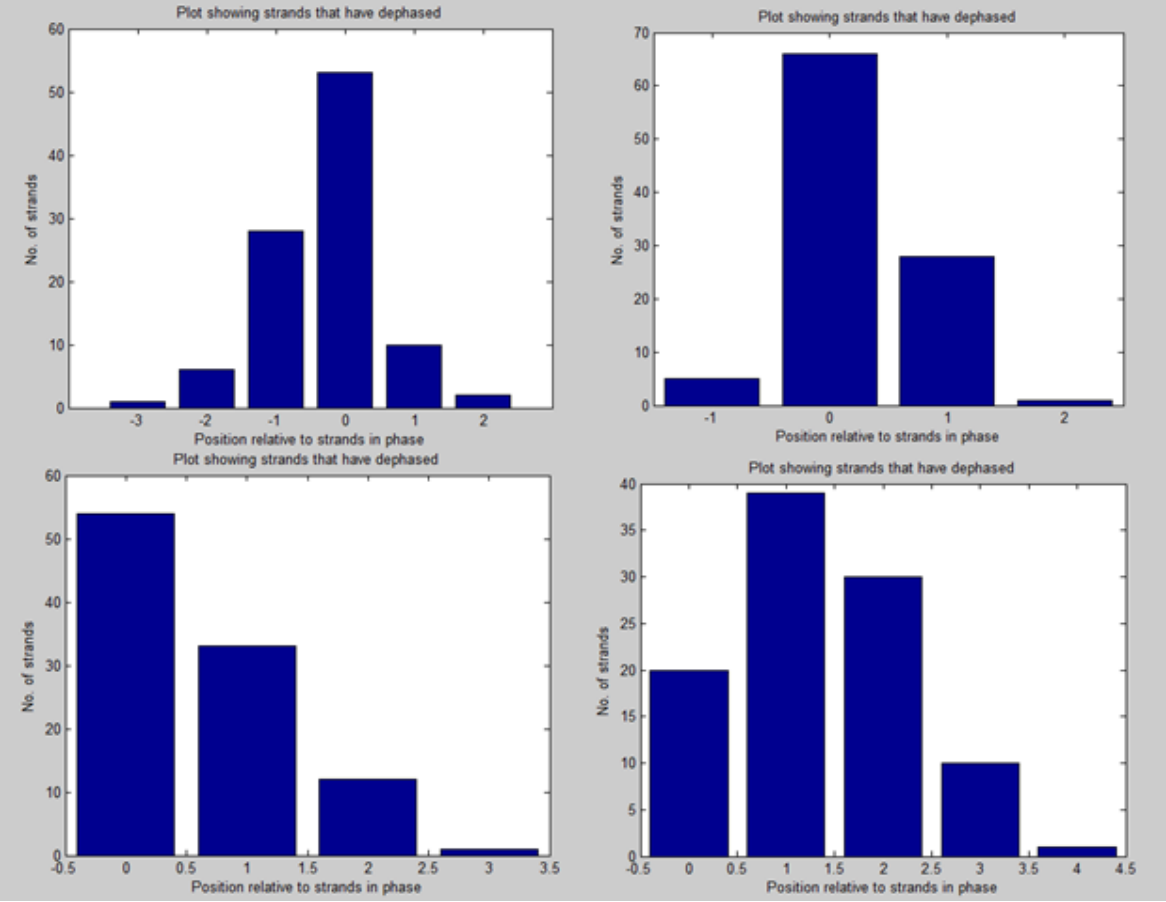



**Figure 27: Dephasing Model Distributions.** Model run with flow time 0.05 s (38% dephased)(Top Left). Model run with flow time 0.1 s (50% dephased) (Top Right.) Model run with flow time 0.25 s (46% dephased) (Bottom Left). Model run with flow time 0.5 s (80% dephased)(Bottom Right). At shorter flow times there are more missed incorporations, but less dephasing. At longer flow times there are more mismatches and greater dephasing.

### *8.A.iii.6) Nucleotide Concentration*

It is recommended that nucleotide concentrations be below 500 µM.[52] The model shows that at lower concentrations, there is less dephasing, as seen in Figure 28 below. It is also important not to make the concentrations too low, lest there not be enough nucleotides. In addition, as can be

[52] Bustillo, J., W. Hinz, K.L. Johnson, J. Leamon, J.M. Rothberg, and J. Schultz. Sequencing nucleic acid comprises disposing template nucleic acids into reaction chambers in contact with or capacitively coupled to chemical-sensitive field effect transistor. Ion Torrent Systems, assignee. Patent GB2561128-A; GB2461128-B. 15 Dec. 2010. Print.

seen from Equation 43 for very low nucleotide concentrations, the observed polymerase rate constants will become too small, and the probability of incorporation will also become too small for proper incorporation.

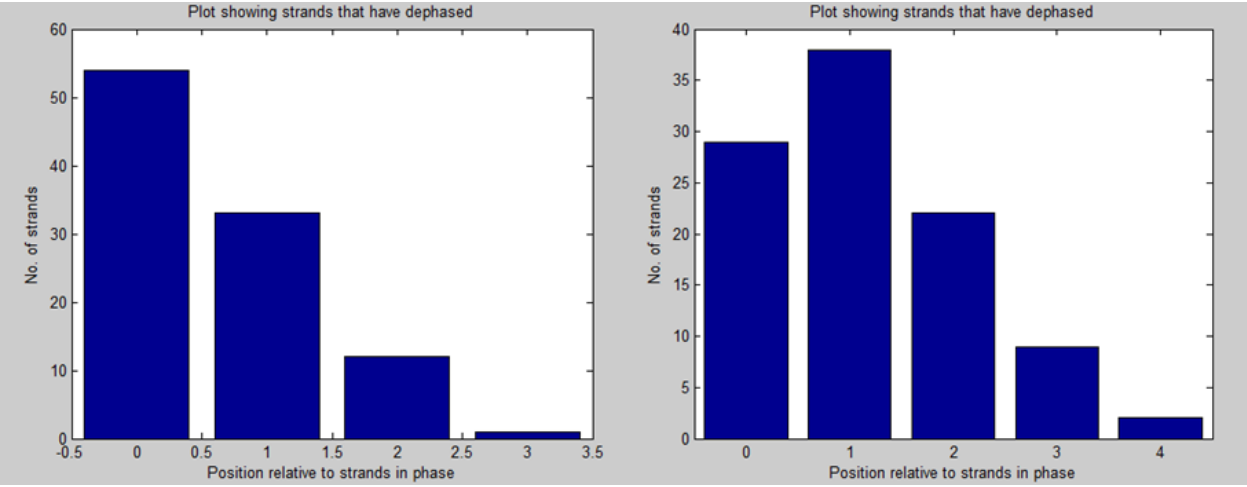



**Figure 28: Distribution of dephased strands.** Model run with nucleotide concentration 100 μM (46% dephased)(Left). Model run with nucleotide concentration 400 μM (72% dephased)(Right). Note that the distributions are only somewhat similar, but the major difference lies in the percent dephased.

To determine the minimum nucleotide concentration required, Equation 45 can be used below.

**Equation 45**

$$Nucleotide\ concentration = \frac{No.of\ potential\ nucleotide\ insertions}{N_{Avogadro} * Reservoir\ volume}$$

$$No.of\ bases = Strand\ length * No.of\ strands * No.of\ wells$$

The concentration of nucleotides flowed in should be at minimum sufficient to account for all potential nucleotide insertions. The concentration of bases can be calculated as shown from Equation 45 using the reservoir volume, which is the volume over the wells through which the nucleotides will be flowed. The actual volume of the wells is negligible compared to the reservoir volume as shown in Figure 29 below, and hence does not need to be accounted for.

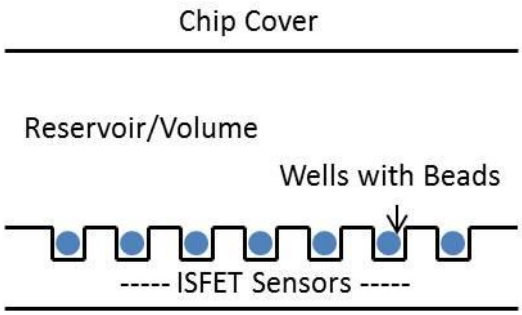



**Figure 29: This side view of the sequencing chip shows the reservoir volume to be much greater than the volume of the wells. (Not to scale)**

As mentioned in Chapter 5 on chip configuration, the dimensions of the chip can be taken to be the following in Table 14.

**Table 14: Sequencing Chip Dimensions**

| | |
|---|---|
| Die Area | 20 mm x 23.7 mm |
| Gap Height | 1 mm |
| No. of wells | 660 million |

If the Dephasing Model is run using 20 bp strand lengths and 100 strands per bead, then the minimum concentration required, calculated using the above equations and dimensions, is 4.2 nM. This concentration is too small to generate high enough probabilities to ensure base incorporations, even correct ones. This is to be expected because the beads are designed to contain hundreds of thousands of strands. Hence it was decided to calculate the minimum nucleotide concentration using 100,000 strands, the number of strands proposed to be an a template bead. A strand length of 200 bp was used in the calculation because that is the strand length that will be actually used in the process. This yields a minimum nucleotide concentration of 42 μM and to be safe, 100 μM was the concentration used throughout all calculations.

### *8.A.III.7) POLYMERASE RATE CONSTANTS*

The model used rate constants of human mitochondrial DNA polymerase, as explained in Kinetics section, Chapter 6. The table of the constants is reproduced here for convenience.

**Table 12: Human mitochondrial DNA rate data**

| dNTP : Template Base | $K_D$ (μm) | $k_{pol}$ ($s^{-1}$) |
|---|---|---|
| A : T | 0.8 | 45 |
| T : T | 57 | 0.013 |
| C : T | 360 | 0.038 |
| G : T | 70 | 1.16 |
| C : G | 0.9 | 43 |
| A : G | 250 | 0.042 |
| T : G | 200 | 0.16 |
| G : G | 150 | 0.066 |
| T : A | 0.6 | 25 |
| C : A | 540 | 0.1 |
| G : A | 500 | 0.05 |
| A : A | 25 | 0.0036 |
| G : C | 0.8 | 37 |
| A : C | 160 | 0.1 |
| C : C | 140 | 0.003 |
| T : C | 180 | 0.012 |

However, these rate constants are not particularly suitable for the sequencing reactions in high throughput sequencing. The $k_{pol}$ values of some of the mismatch incorporation are rather high. For example, the incorporation of G on T has $k_{pol}$ 1.16 $s^{-1}$ compared to other rate constants on the order of 0.001 $s^{-1}$. The $k_{pol}$ value of the correct incorporation of T on A is also rather low (25 $s^{-1}$) as compared to the other correct incorporation values, which are all closer to 40 $s^{-1}$.

Using the $k_{pol}$ values in Table 12, the error rates are incredibly high and more than 95% of the strands become dephased. Thus for successful sequencing runs, it is extremely important to use better polymerases. Ion Torrent has not released any information regarding the polymerase they use, given that it is their trade secret, but it is very likely that the Torrent R&D teams have designed their own polymerases by introducing mutations to these polymerases that can have more favorable rate constants.

Having run the model using different values of the rate constants, it is recommended that the $k_{pol}$ values for all mismatches be lower than approximately 0.01 and the $k_{pol}$ values for all matches be around the same ballpark as 40. This creates a much better distribution of dephased strands and with read lengths of 200 bp only about 40% of the strands become dephased, as in Figure 30 below. The better the rate constants, which in our case indicates higher $k_{pol}$ values for matches and lower $k_{pol}$ values for mismatches, the fewer the errors will be and also the longer the base flow time can be.

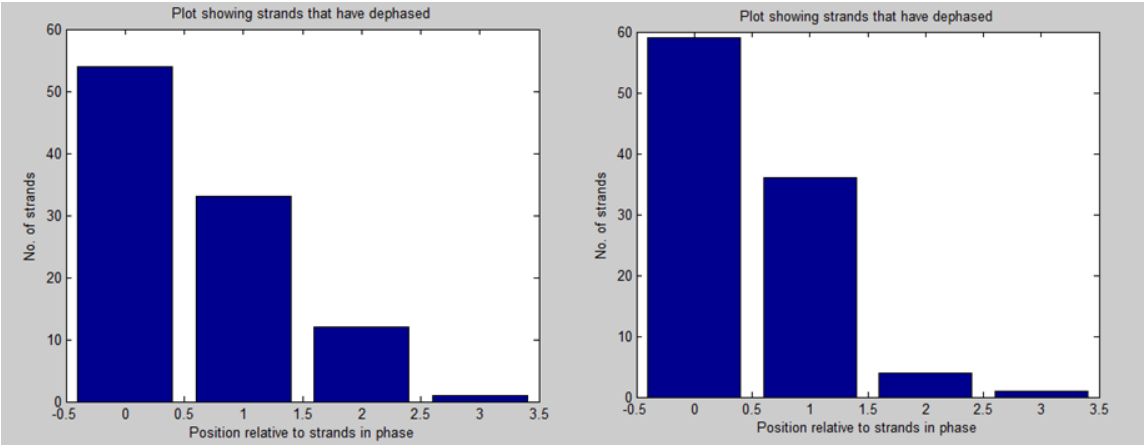



**Figure 30: Distribution of dephased strands.** Model run with human mitochondrial DNA rate constants with 20 bp read length (46% dephased) (Left). Model run with recommended rate constants with 200 bp read length (41% dephased) Note that distribution of dephased strands similar.

With the recommended rate constants, the 200 bp long fragments to be used in IonSeq's process will provide acceptable error rates. In fact, it may even be possible to achieve longer read lengths as shown in the Figure 31 below given that even with 300 bp read lengths, less than 50% of the strands are dephased. This is a recommended area to be explored in the future to achieve better realignment.

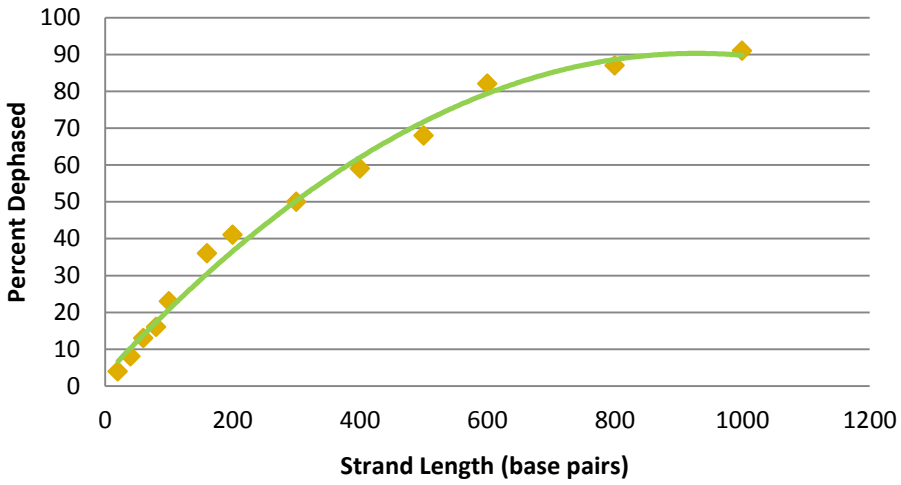



**Figure 31: Percent of strands dephased over different strand lengths shows that a strand length of 200 bp is acceptable.**

### *8.A.III.8) SUMMARY*

Using the dephasing model to predict how changing the variables can affect dephasing, the following values are recommended for the process in Table 15 below.

**Table 15: Summary of Recommended Values based off of Dephasing Model**

| | |
|---|---|
| Target strand length of library | 200 bp |
| Number of flow cycles | 800 |
| Flow order | any |
| Flow time | 0.25 s |
| Nucleotide concentration | 100 µM |

It is also recommended that efforts be made to design polymerases with favorable kinetics. Polymerization rate constants for all mismatches should be at least two orders of magnitude less than 1 and these constants for all correct matches should be around the same ballpark as 40 $s^{-1}$.

## 8.B. BASE CALLING

When raw data is generated from the sequencing runs, the servers need to convert that information into an actual sequence for distribution to the client. This process includes accurately

recognizing homopolymers, distinguishing between single base differences. There exists a wealth of proprietary and open-source base calling algorithms, and there continues to be new algorithms developed with advancing next generation sequencing technologies.

## *8.B.i. CallSim Base Calling*

One example is the software application CallSim, which uses a base calling algorithm applicable to data from the Ion Proton System. The algorithm processes a single read using a Monte Carlo approach and is not dependent upon information from any other read in the data set. It accounts for the random nature of the polymerase on the DNA molecules associated with a single sequencing well. The pseudocode is outlined in Figure 32.

```
signal:= 0 ;  # initialize the signal value
for (all polymerase/molecules that are not stalled)
{
  if (base == 'N') { skip over this base }
  else if (Rand[0,1] < P_stop)  { polymerase status:= stalled }
  else while((base==flow base) & (!stalled) & (no failure to add base))
  {
    if ( Rand[0,1] > P_skip )      # add base
    {
      position++;                  # polymerase moves to next base
      signal++;                    # signal produced
      if (position == last base) { polymerase status:= stalled }
    }
  }
} next polymerase/molecule
```

**Figure 32: The CallSim algorithm is one of the several available base-calling methods for use with Ion Torrent technology[53].**

## *8.B.ii. IonSeq Base Calling*

IonSeq's base calling model is rudimentary, but works well with high fidelity polymerases. Via the kinetic Monte Carlo method, there are a certain number of time segments within which base insertions may occur. The probability of base insertion is based upon the rate constants, and the size of these time segments. During one nucleotide flow, the number of insertion events is summed

[53] Morrow, J., & Higgs, B. (2012). Callsim: Evaluation of base calls using sequencing simulation. . *ISRN Bioinformatics*, *2012*, 10 pages. doi: 10.5402/2012/371718

across all the time segments. This yields total insertion events, which is then divided by the number of strands in the model. This resulting number gives the average number of bases inserted per strand. However, this number may prove to be unclear. For example, a value of 2.7 most likely corresponds to a value of 3 bases, but a value of 2.45 leaves ambiguity between 2 or 3 base insertions. IonSeq's base calling algorithm is to round this value to the nearest integer. Table 16 outlines the steps of the base calling process described above for 50 strands. The "# of Flow Base Inserted" column correlates exactly to the "Strand Base" column shown at the right of the table.

**Table 16: The Base Calling Process for 50 Strands**

| Flow # | Flow Base | *Time 1* | *Time 2* | *Time 3* | *Time 4* | *Time 5* | Sum | Sum/ Strands | # of Flow Base Inserted | Strand Base |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | C |
| 2 | G | 76 | 51 | 15 | 5 | 3 | 150 | 3 | 3 (G) | C |
| 3 | C | 81 | 17 | 2 | 0 | 0 | 100 | 2 | 2 (C) | C |
| 4 | A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | G |
| 5 | T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | G |
| 6 | G | 35 | 9 | 6 | 0 | 1 | 51 | 1.02 | 1 (G) | C |
| 7 | C | 0 | 1 | 0 | 0 | 0 | 1 | 0.02 | 0 | T |
| 8 | A | 110 | 34 | 0 | 0 | 0 | 144 | 2.88 | 3 (A) | T |
| 9 | T | 20 | 14 | 8 | 5 | 1 | 48 | 0.96 | 1 (T) | T |
| 10 | G | 1 | 0 | 0 | 0 | 0 | 1 | 0.02 | 0 | A |
| 11 | C | 84 | 10 | 2 | 0 | 0 | 96 | 1.92 | 2 (C) | G |
| 12 | A | 3 | 0 | 0 | 0 | 0 | 3 | 0.06 | 0 | G |
| 13 | T | 2 | 0 | 0 | 0 | 0 | 2 | 0.04 | 0 | C |
| 14 | G | 36 | 12 | 1 | 1 | 1 | 51 | 1.02 | 1 (G) | T |
| 15 | C | 2 | 2 | 0 | 0 | 0 | 4 | 0.08 | 0 | T |
| 16 | A | 119 | 20 | 2 | 0 | 0 | 141 | 2.82 | 3 (A) | T |

Figure 33 provides a more visual representation of the base calling results. It shows the raw insertion values in the top graph while the base-calling results are shown in the bottom graph. Although the graphs look very similar, there are differences in fidelity. The graphs to the left are for a lower fidelity polymerase while the graphs to the right are for a higher fidelity polymerase. The lower fidelity polymerase yields a noisier collection of raw data, which can lead to incorrect base calls. The MATLAB code for this algorithm can be found in Appendix G.

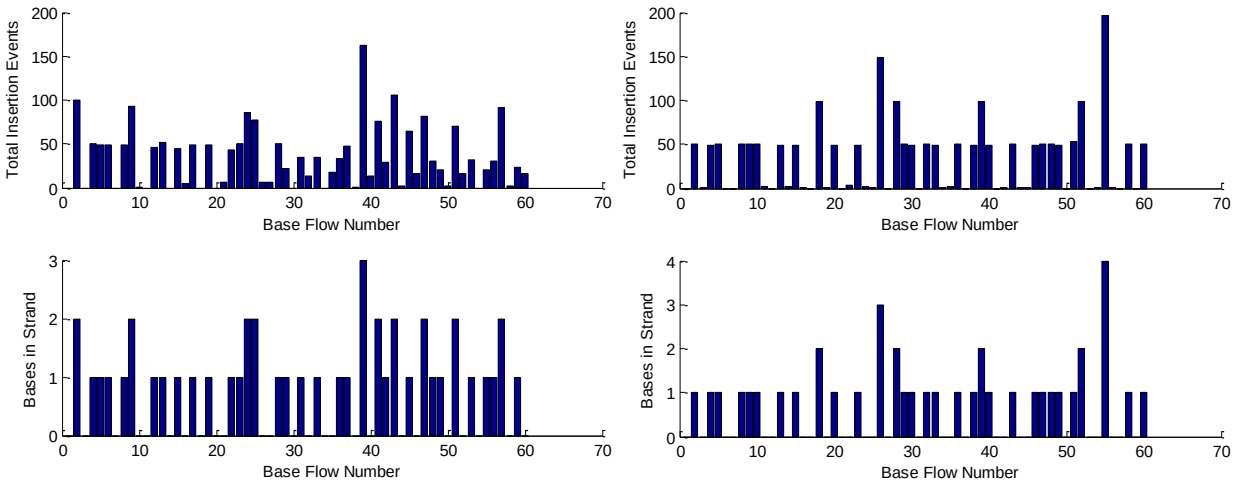



**Figure 33a: Base calling for slower polymerase rates shows the values of some total insertion events to be between integer base values and may lead to incorrect base calls. b: Faster polymerase rates reduce this ambiguity in base calling.**

Here, there exists significant ambiguity at some base flows, which lead to erroneous base calls, and may impact the accuracy of the base calls of later flows. This is the result of relatively slow rate constants for correct insertions. If the rate constants for slower correct insertions increased to the speed of the other insertions (~40 $s^{-1}$), Figure 33 illustrates much more accurate base-calling.

Table 17 lists the results the derived sequence for a sample genome, with 200 base strands. There were 14 mismatches in this example predominately at the very end of the strand as bolded. However, further observation reveals that this onset of errors was the cause of a missing nucleotide insertion at the double starred position. This missing insertion caused the rest of the sequence result, which is correct, to be moved up one position. Current aligner algorithms would recognize this single error and make the appropriate corrections. IonSeq's base calling algorithm works well with the developed MATLAB model but is not tested for real applications where it must interpret the ISFET signals as shown in Figure 22 in Chapter 7. IonSeq will adopt a more reliable base-calling algorithm proven to be highly accurate.

**Table 17: Comparison between Sample DNA Strand and Derived Sequence**

| Genome Sequence | Sequence Result |
|---|---|
| T | T |
| A | A |
| T | T |
| A | A |
| A | A |
| G | G |
| T | T |
| T | T |
| C | C |
| C | C |
| C | C |
| C | C |
| A | A |
| C | C |
| C | C |
| A | A |
| T | T |
| G | G |
| A | A |
| T | T |
| C | C |
| C | C |
| T | T |
| T | T |
| G | G |
| C | C |
| A | A |
| T | T |
| C | C |
| A | A |
| C | C |
| A | A |
| C | C |
| T | T |
| G | G |
| A | A |
| T | T |
| A | A |
| G | G |
| T | T |
| T | T |
| T | T |
| A | A |
| T | T |
| T | T |
| T | T |
| C | C |
| C | C |
| G | G |
| T | T |
| G | G |
| G | G |
| C | C |
| A | A |
| G | G |

| Genome Sequence | Sequence Result |
|---|---|
| C | C |
| G | G |
| G | G |
| G | G |
| C | C |
| A | A |
| A | A |
| A | A |
| A | A |
| G | G |
| T | T |
| T | T |
| C | C |
| A | A |
| T | T |
| G | G |
| A | A |
| A | A |
| C | C |
| G | G |
| G | G |
| T | T |
| C | C |
| C | C |
| A | A |
| T | T |
| C | C |
| A | A |
| T | T |
| G | G |
| T | T |
| C | C |
| A | A |
| G | G |
| A | A |
| C | C |
| C | C |
| C | C |
| G | G |
| A | A |
| G | G |
| T | T |
| G | G |
| G | G |
| C | C |
| A | A |
| G | G |

| Genome Sequence | Sequence Result |
|---|---|
| C | C |
| G | G |
| G | G |
| G | G |
| T | T |
| G | G |
| G | G |
| G | G |
| C | C |
| G | G |
| T | T |
| C | C |
| A | A |
| A | A |
| T | T |
| T | T |
| C | C |
| T | T |
| A | A |
| C | C |
| T | T |
| G | G |
| C | C |
| G | G |
| T | T |
| C | C |
| C | C |
| T | T |
| C | C |
| G | G |
| C | C |
| C | C |
| T | T |
| C | C |
| C | C |
| G | G |
| A | A |
| T | T |
| T | T |
| G | G |
| T | T |
| A | A |
| A | A |
| C | C |
| C | C |

| Genome Sequence | Sequence Result |
|---|---|
| A | A |
| G | G |
| A | A |
| G | G |
| T | T |
| C | C |
| T | T |
| G | G |
| C | C |
| G | G |
| A | A |
| T | T |
| A | A |
| C | C |
| G | G |
| T | T |
| G | G |
| G | G |
| C | C |
| A | A |
| T | T |
| G | G |
| C | C |
| C | C |
| C | C |
| C | C |
| C | C |
| C | C |
| C | C |
| **C** | **T**** |
| **T** | **C** |
| C | C |
| **C** | **G** |
| G | G |
| **G** | **T** |
| **T** | **G** |
| **G** | **A** |
| A | A |
| **A** | **T** |
| **T** | **A** |
| **A** | **C** |
| **C** | **A** |
| **A** | **T** |
| **T** | **G** |
| **G** | **A** |
| A | A |
| A | A |
| **A** | **T** |

## 8.C. REALIGNMENT

The Ion Reporter™ Software can be used to reassemble the reads and create a report file containing the full genome sequence including variant lists that can be delivered to the customer.

The Ion Reporter™ Software can utilize cloud computing technology; thus the realignment can be carried out on external servers. The reads can be securely transferred to a centralized server hosted by Ion Torrent using secure https protocol and then stored using 256 bit encryption technology[54].

The realignment can also be carried out using algorithms such as Novoalign, NextGEne and Partek, all of which are compatible with reads from Ion Proton™ Sequencers. The realignment is done by mapping the reads to a reference genome, available in the public domain, as opposed to de novo alignment. De novo alignment requires more complex alignment algorithms and long read lengths; hence, alignment to a reference genome is the best strategy.

## 8.D. Error Rates

Measuring the error rates that emerge from sequencing runs is crucial for understanding the extent of a sequence's accuracy and for evaluating the necessary rigor for mapping and alignment processes after the sequencing runs. An easy to interpret metric is important for quick interpretation throughout the sequencing industry.

### *8.D.i. Phred Quality Rating*

The Phred quality rating, Q, is common metric to measure the accuracy of base calls.[55] Equation 46 shows that the Phred quality rating is based upon the probability of an error or incorrect base call. Table 18 outlines the standard Phred ratings and their respective probabilities for error.

**Equation 46**

$$Q = -10 * \log_{10} P$$

[54] IonTorrent. (Producer). (2012). *Learn more about ion torrent software*. [Web Video]. Retrieved from http://www.youtube.com/watch?v=g0ze9Dp9qu0

[55] Richterich, Peter. "Estimation of Errors in "Raw" DNA Sequences: A Validation Study." *Genome Research* 8.3 (1998): 251-259. Web. 31 Mar. 2013.

**Table 18: Phred Quality Scores and Corresponding Probabilities**

| Phred Score (Q) | Probability of Error (P) | Base Accuracy (%) |
|---|---|---|
| 10 | 1 out of 10 bases | 90 |
| 20 | 1 out of 100 bases | 99 |
| 30 | 1 out of 1,000 bases | 99.9 |
| 40 | 1 out of 10,000 bases | 99.99 |
| 50 | 1 out of 100,000 bases | 99.999 |

## *8.D.II PHRED RATINGS FROM IONSEQ BASE CALLING ALGORITHMS*

Using this metric, IonSeq can easily measure the accuracy of their sequences and compare to common standards in the industry. Furthermore, these numbers are important in meeting customer requirements. As outlined in section 3.A.iii., IonSeq's customers require a Phred score of 50, or 99.999% accuracy. Error rates will be heavily dependent on polymerase kinetics; higher polymerization rate constants for correct nucleotide matches and lower rate constants for mismatches will lead to greater accuracy. Three different cases, consisting of different rate constants, are evaluated; these values are tabulated in Table 19, based off the values found in Table 12. The rate constants that are changed from the given values in Table 12 are shown in bold.

**Table 19: Rate Constants for Three Cases**

| dNTP : Template Base | $K_D$ (μm) [Case 1/2/3] | $k_{pol}$ ($s^{-1}$) [Case 1/2/3] |
|---|---|---|
| A : T | [0.8, 0.8, 0.8] | [45, 45, 45] |
| T : T | [57, 57, 57] | [0.013, 0.013, **0.0013**] |
| C : T | [360, 360, 360] | [0.038, 0.038, **0.0038**] |
| G : T | [70, 70, 70] | [1.16, **0.016, 0.0016]** |
| C : G | [0.9, 0.9, 0.9] | [43, 43, 43] |
| A : G | [250, 250, 250] | [0.042, 0.042, **0.0042**] |
| T : G | [200, 200, 200] | [0.16, **0.016**, **0.0016**] |
| G : G | [150, 150, 150] | [0.066, 0.066, **0.0066**] |
| T : A | [0.6, 0.6, 0.6] | [25, **40**, 40] |
| C : A | [540, 540, 540] | [0.1, **0.01**, **0.001**] |
| G : A | [500, 500, 500] | [0.05, 0.05, **0.005**] |
| A : A | [25, 25, 25] | [0.0036, 0.0036, 0.0036] |
| G : C | [0.8, 0.8, 0.8] | [37, **40**, 40] |
| A : C | [160, 160, 160] | [0.1, **0.01**, **0.001**] |
| C : C | [140, 140, 140] | [0.003, 0.003, 0.003] |
| T : C | [180, 180, 180] | [0.012, 0.012, **0.0012**] |

Using these values, the error rates from the base calling algorithm, discussed in section 8.B.ii and found in Appendix G, can be found. Working with 100 strands each with a length of 200 bases, the algorithm is run and the counts of mismatched nucleotides are tabulated. These values are averaged across all the strands, converted into overall percent error, and then Equation 46 is used to derive the Phred score. The results are outlined in Table 20; if compared with the manipulated rate constants in Table 19, the Phred score increases by about 10 when the rate constants are favorably changed by factor of 10. This is a result of the logarithmic definition of the Phred score. To achieve the customer requirement of 50 Phred score, IonSeq will need to develop its own proprietary polymerases to significantly reduce mismatches; as the base calling model results demonstrate, this can be theoretically be achieved.

**Table 20: Phred Scores for the Cases of Different Rate Constants**

| Case | Percent Error | Phred Score |
|---|---|---|
| Case 1 | 0.5% (1 error out of 200 bases) | 23 |
| Case 2 | 0.09% (0.18 errors out of 200 bases) | 30.5 |
| Case 3 | 0.005% (0.01 errors out of 200 bases) | 43 |

## 8.E. Future Potential of Data Analysis

The IonSeq process generates large amounts of data from each microwell and powerful computational and bioinformatics tools, like those discussed in Sections 8.B. and 8.C. are required to take the raw reads and convert them into the final genome sequence that will be delivered to the customer. Through progress and development in both sequencing technology and computational technology, there are opportunities of achieving even higher throughput in the future. As more complex base calling and realignment algorithms are created, it may be possible to sequence genomes with even lower coverage with greater accuracy. There are several opportunities for improvements in sequencing technologies as discussed in Section 8.A. By improving polymerase kinetics, increasing read lengths, and decreasing signal detection times, it will be possible to reduce

time required for the sequencing and sequencing costs. With reductions to the required coverage, significant increases to throughput can also be made. Hence, it is important for IonSeq to keep up with the rapid advancements being made in these technologies, such that it is possible to take full advantage of these advancements and maximize throughput.

# 9. OPTIMIZATION OPTIONS

Now that the kinetics and ISFET sensor models have been thoroughly explored and developed in Chapters 6 and 7, optimization of the signal strength can be performed. IonSeq seeks to decrease attenuation time to improve throughput as well as increase the signal differentiation between homopolymers for more accurate base calling. In this chapter, two options will be considered and the third choice will be developed by combining aspects of the first two options.

## 9.A. OPTION A: GATE INSULATOR MATERIAL – SIGNAL STRENGTH OR ATTENUATION TIME AND GENOMIC OUTPUT

Ion Torrent's Proton II chip is to be composed of Silicon Dioxide insulator layer on a Silicon Nitride gate insulator layer. Signal strength is a function of the difference in surface potential and electrolyte solution potential. Surface potential of an ISFET is based off the Nernst equation for a two proton layers and follows the site-binding model, as previously shown in Equation 33[56]. The

main variables behind the ISFET surface potential are $\kappa$, the selectivity constant, and the difference between the bulk fluid *pH* and $pH_{pzc}$, the pH at zero potential. For Silicon Nitride at 298K, the selectivity constant is 0.93 and the $pH_{pzc}$ is 6.8. To achieve a positive surface potential, the bulk fluid pH must be greater than the $pH_{pzc}$. The Proton II chip is run with a bulk pH of 8.0. By substituting the gate insulator layer material, the $pH_{pzc}$ and selectivity constant will change. By keeping the bulk fluid pH at 8.0, the pH difference of the bulk fluid and $pH_{pzc}$ can be maximized for a $pH_{pzc}$ lower than that of Silicon Nitride. For a gate insulator material, $pH_{pzc}$ can be calculated by Equation 47[57].

**Equation 47**

$$pH_{pzc} = -\log_{10}(K_a K_b)^{0.5} = 0.5\,(pK_a + pK_b)$$

Selectivity constants are a function of a material's sensitivity value. Sensitivity is defined derivative of the site-binding model with respect to pH; Equation 33 can be rearranged to see this dependence[58]. At 298K, Silicon Nitride has a sensitivity value between 52-58 mV/pH for a pH range of 1-13 and a selectivity of 0.93. By obtaining a material's sensitivity, the selectivity constant can be obtained; the results are listed in Table 21.

[56] Woias, P., L. Meixner, D. Amandi, and M. Schönberger. "Modelling the Short-time Response of ISFET Sensors." *Sensors and Actuators B: Chemical* 24.1-3 (1995): 211-17. Print.
[57] Chiang, Jung-Lung. *Study on the pH-Sensing Characteristics of ISFET with Aluminum Nitride Membrane*. Diss. 2002.
[58] R.E.G. van Hal, J.C.T. Eijkel, P. Bergveld, A novel description of ISFET sensitivity with the buffer capacity and double-layer capacitance as key parameters, Sensors and Actuators B: Chemical, Volume 24, Issues 1–3, (1995): 201-205.

**Table 21: Determination of Selectivity Constants**

| | $pH_{pzc}$ | Sensitivity (mV/pH) | Obtained κ |
|---|---|---|---|
| $Si_3N_4$ | 6.8[59] | 52-58[60] | 0.93[61] |
| $Ta_2O_5$ | 2.8[62] | 57.1-58.3[63] | 0.98 |
| $TiO_2$ | 6.1[64] | 56.2[65] | 0.95 |
| $SnO_2$ | 6.0[66] | 58.9[67] | 1.00 |
| $PbTiO_3$ | 1.8[68] | 56-59[69] | 0.97 |

Optimization by materials offers two options: to increase signal strength or shorten attenuation time. Signal strength increase is reliant on the difference in pH of the bulk fluid and the $pH_{pzc}$ and the selectivity constant. In the case of signal strength increase, attenuation time remains the same, but signal-to-noise ratio increases. Shortened attenuation time is reliant on the potential of the bulk solution, which influence both the potential of the surface of the gate insulator and the potential of the solution after a reaction. Theoretically, shortened attenuation times decreases the amount of base pair mismatches, increases the potential gap between homopolymer reads, and decreases the time for nucleotide turnover. The decrease in base pair mismatches and increase in potential gap between homopolymer reads offers a significant increase in accuracy. The decrease in time for nucleotide turnover lower the amount of time required for the sequencing of a single strand, thus increases overall output of the chip.

[59] Dutta, J. C. "Modeling Ion Sensitive Field Effect Transistors for Biosensor Applications." *International Journal of Advanced Research in Engineering and Technology*. (2010): 38-57.
[60] Chiang J, Chou J, Chen Y; Sensitivity and hysteresis properties of a-wo3,ta2o5, and a-si:h gate ion-sensitive field-effect transistors. Opt. Eng. 0001;41(8):2032-2038.
[61] Ibid. 53
[62] Natishan, P. M., E. McCafferty, and G. K. Hubler. "Surface Charge Considerations in the Pitting of Ion-Implanted Aluminum." *Journal of The Electrochemical Society* 135 (1988): 321.
[63] Chiang, Jung-Lung, Jung-Chuan Chou, and Ying-Chung Chen. "Sensitivity and hysteresis properties of a-WO3, Ta2O5, and a-Si: H gate ion-sensitive field-effect transistors." *Optical Engineering* 41.8 (2002): 2032-2038.
[64] Preočanin, T., & Kallay, N. (2006). Point of zero charge and surface charge density of TiO2 in aqueous electrolyte solution as obtained by potentiometric mass titration. *Croatica chemica acta*, *79*(1), 95-106.
[65] Jung-Chuan Chou, Lan Pin Liao, Study on pH at the point of zero charge of TiO2 pH ion-sensitive field effect transistor made by the sputtering method, *Thin Solid Films*, 476:1. (2005): 157-161.
[66] Liao, Hung-Kwei, et al. "Study on $pH_{pzc}$ and surface potential of tin oxide gate ISFET." *Materials chemistry and physics* 59.1 (1999): 6-11.
[67] Ibid.
[68] Jan, Shiun-Sheng, et al. "Preparation and properties of lead titanate gate ion-sensitive field-effect transistors by the sol-gel method." *Japanese journal of applied physics* 41.2A (2002): 942-948.
[69] Ibid.

For the scenario of a single base incorporation, where the bulk solution pH is 8, Silicon Nitride peaks at 28.95mV, requires 3.20 seconds to degrade the signal to 90% of the maximum value, and the signal ratio, defined as the ratio of the difference in signal between the second and third nucleotide incorporation to the difference in signal between the first and second nucleotide incorporation, of 0.4433. The signal for this base case is shown in Figure 34. This value must be less than one, where a ratio of one indicates the most amount of differentiation between homopolymers and smaller ratio indicates a more quickly diminishing homopolymer response signal leading to the inability to distinguish between homopolymer at shorter homopolymer chains.

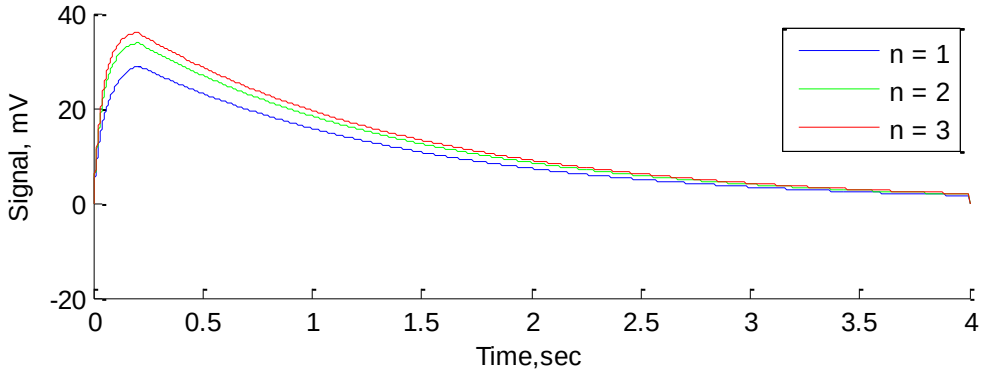
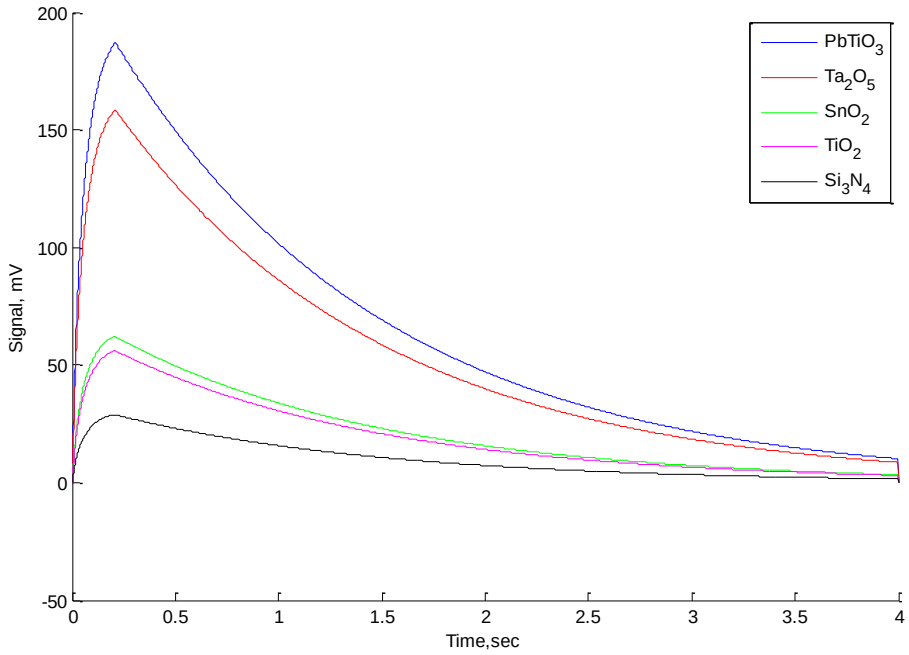



**Figure 34: Signal Base Case – Silicon nitride**

Keeping the bulk solution pH constant, surface potentials for the materials increase approximately two-fold and range from 56.18 mV to 187.18mV; the largest surface potential occurs for Lead Titanate and the smallest surface potential occurs for Titanium Oxide, shown in Figure 35.




| Material | Maximum Signal |
|---|---|
| $PbTiO_3$ | 187.18 mV |
| $Ta_2O_5$ | 158.61 mV |
| $SnO_2$ | 62.25 mV |
| $TiO_2$ | 56.18 mV |
| $Si_3N_4$ | 28.95 mV |

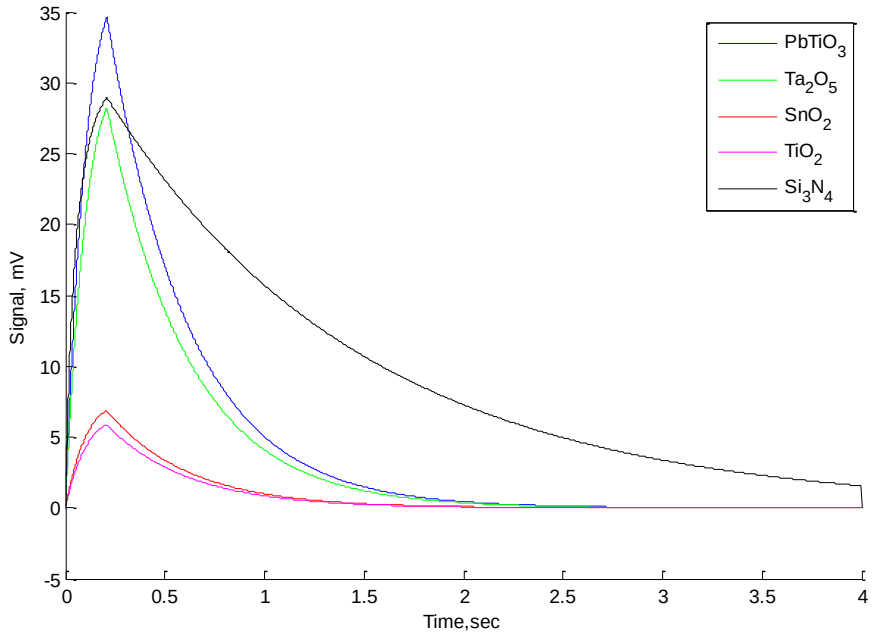
**Figure 35: Signal generation for a single nucleotide incorporation of various materials at bulk pH of 8 by decreasing order of maximum signal output**

Using Silicon Nitride, several factors limits the bulk fluid pH, which affects both the signal strength and the attenuation time. An optimal bulk solution pH lies at a number which maximizes the difference between itself and the $pH_{pzc}$ of Silicon Nitride and is at a suitable pH which would not denature the DNA polymerase. The nature of the high $pH_{pzc}$ of Silicon Nitride limits its bulk fluid pH, which limits both, its signal strength and its attenuation time.

Of the materials selected, the $pH_{pzc}$ is significantly lower than that of Silicon Nitride, allowing for a wider range of allowable bulk fluid pH Because the optimal pH for the DNA polymerase lies between 5 and 8[70], a bulk fluid pH of 7 would be lower than that of the base case but offers a conservative pH range for the influx of protons and for the maintaining of integrity of

[70] Lopes, D. O., et al. "Analysis of DNA polymerase activity in vitro using non-radioactive primer extension assay in an automated DNA sequencer." *Genet Mol Res* 6 (2007): 250-255.

the DNA polymerase. Application of the new bulk fluid pH causes an attenuation time of 1.156 seconds and a signal ratio of 0.6058. The results are illustrated in Figure 36.




| Material | Maximum Signal |
|---|---|
| $PbTiO_3$ | 34.63 mV |
| $Ta_2O_5$ | 28.26 mV |
| $SnO_2$ | 6.87 mV |
| $TiO_2$ | 5.87 mV |

**Figure 36: Signal generation for a single nucleotide incorporation of various materials at bulk pH of 7 by decreasing order of maximum signal output compared to Si3N4 at bulk pH of 8**

Optimization by material shortens attenuation time 3.5-fold, but of the materials selected for optimization, Lead Titanate offers a higher maximum signal at $pH_{bulk}$ of 7 than Silicon Nitride at $pH_{bulk}$ of 8. A comparison between Lead Titanate at $pH_{bulk}$ of 7 and Silicon Nitride at $pH_{bulk}$of 8 can be found below in Table 22. The key takeaway from this optimization is the increase in the signal ratio, which would imply more accurate base calling when encountering homopolymers.

**Table 22: Comparison of Silicon Nitride at bulk pH of 8 and Lead Titanate at bulk pH of 7**

| | Silicon Nitride | Lead Titanate | Fold Change |
|---|---|---|---|
| **Maximum Signal** | 28.95 mV | 34.63 mV | 1.20 |
| **Attenuation Time (99%)** | 6 seconds | 2.10 seconds | -2.76 |
| **Signal Ratio** | 0.4433 | 0.6058 | 1.37 |

Throughput from the use of Lead Titanate requires the calculation of a cycle time. Nucleotides are washed following the rise time for signal generation. Because Silicon Nitride's rise time is insignificant (<0.1 seconds), the total cycle time is equal to the attenuation time plus the buffer flow time of 2 seconds. For the case of Lead Titanate, the rise time for the signal is 0.21 seconds, 10% of the total attenuation time, and is extremely significant. The cycle time for a Lead Titanate insulator gate chip would be the sum of the rise time, the attenuation time (2.31 seconds) and the buffer time, which would yield a cycle time of 4.31 seconds. A comparison of throughput from Proton II to a theoretical Proton III, based off the Proton II, made of Lead Titanate, and at bulk pH of 7, can be seen below in Table 23. These changes reduce throughput by a factor of 1.85.

**Table 23: Comparison of Silicon Nitride at bulk pH of 8 and Lead Titanate at bulk pH of 7**

| Strand Length (base pairs) | Cycle Time (sec) | Cycles | Throughput (hr/genome) |
|---|---|---|---|
| **Current Proton II** | 8 | 1600 | 3.56 |
| **Proposed Proton III** | 4.31 | 1600 | 1.92 |

## 9.B. Option B: Well-Size and Organization – Attenuation Time and Genomic Output

At present, Ion Torrent's Proton II chip contains 660 million wells at a well diameter 0.70 μm on a nodal length of 110 microns. With those methods, it takes about almost 3 hours for a single genome. However, current technology has driven node length to a commonly available 32nm, with a 22nm node length in development[71]. By decreasing node length, more wells could be fitted within a single chip, allowing two optimization options for throughput: decreased strand size for faster turnover or the ability to apply genetic tags for the sequencing of one more genomes.

[71] "3-D, 22nm: New Technology Delivers An Unprecedented Combination of Performance and Power Efficiency." Intel, Web. <http://www.intel.com/content/www/us/en/silicon-innovations/intel-22nm-technology.html>.

**Table 24: Comparison of well changes between Proton II and Proton III**

| | Well Diameter (μm) | Attenuation Time | Potential Ratio | Maximum Signal |
|---|---|---|---|---|
| **Current Proton II – 110nm** | | | | |
| 660M | 0.70 | 6 seconds | 0.4433 | 28.95mV |
| **Proposed "Proton III" – 32nm** | | | | |
| 660M | 0.85 | 4 seconds | 0.4356 | 29.33mV |
| 1.0B | 0.70 | 4 seconds | 0.4433 | 28.95mV |
| 1.5B | 0.54 | 4 seconds | 0.4958 | 25.96mV |

Applying a 32nm node length to a chip allows for three different scenarios: 660 million wells at 0.85μm per well, 1 billion wells at 0.70μm per well, or 1.5 billion wells at 0.54μm per well, outlined in Table 24. By decreasing well size, the significance in the difference between homopolymers increased, indicated by the potential ratio. This decrease in well size also decreased the maximum signal output, where a decrease in well size decreased the maximum signal output, which lowers the signal-to-noise ratio.

By implementing a genetic tag at approximately 40 base pairs in length, the amount of cycles required increases but allows for the possibility of increasing throughput by increasing total genome output. For a genetic tag attached to 200 base pair strands, it would require approximately 750 million wells for a single genome. Utilizing the 32nm case, two genomes would be completed in 2.13 hours. Table 25 gives a potential workflow.

**Table 25: Theoretical Workflow**

| Description | Well Diameter (μm) | Cycle Time (sec) | Base Length | Cycles | Time (hr) | Throughput (hr/genome) |
|---|---|---|---|---|---|---|
| **Single Genome** | | | | | | |
| Proton II – 660M @ 110nm | 0.70 | 8 | 200 | 1600 | 3.56 | 3.56 |
| "Proton III" – 1.5B @ 32nm | 0.54 | 6 | 100 | 800 | 1.33 | 1.33 |
| **Barcoded – Two+ Genomes** | | | | | | |
| Proton II – 1B @ 110nm | 0.51 | 8 | 240 | 1920 | 4.27 | 2.13 |
| "Proton III" – 1.5B @ 32nm | 0.54 | 6 | 240 | 1920 | 3.20 | 1.60 |

Implementation of a shorter sequence in more wells allows for less nucleotide cycles for faster genomic output. For a 32nm chip at a 100 base pair strand, it would require 1.5 billion wells but would halve the amount of nucleotide cycles to 800. The single strand output would have the same attenuation time, 4 seconds, and a total output time of 0.89 hours. Utilizing a lower base pair strand limits the possibility of longer homopolymer sequences, allowing for increased accuracy in homopolymer readings in addition to the increase in potential ratio.

## 9.C. Option A+B: Application of 32nm Technology to Lead Titanate Gate Insulator Layer

With Ion Torrent's technology, the Proton II, being at 660 million wells utilizing 110nm nodal length, it requires approximately 3.5 hours to generate the full raw data from a genome. Applying both optimization cases leads to a Lead Titanate gate insulator chip with 1.5 billion wells at a 32nm node. This resulting theoretical signal response is shown in Figure 37.

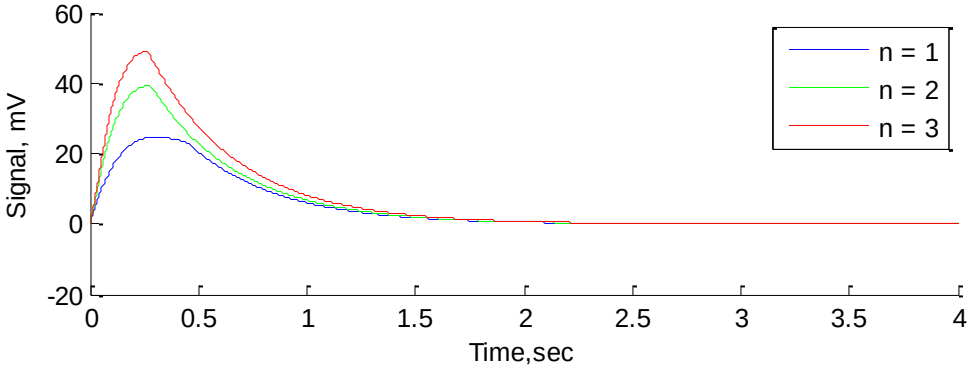



**Figure 37: Proposed signal generated from Theoretical Proton III Chip**

Using the 32nm technology with the Lead Titanate layer effectively shortens the attenuation time to 2.23 seconds, offers a maximum signal of 24.67, and offers an extremely high potential ratio of 0.632. However, unlike the case for the Proton II, where the rise time determines when dNTPs are washed out of the well and is insignificant (~0.1 seconds), the time required for the maximal signal is 0.31 seconds, approximately 14% of the amount of time required for 99% signal

attenuation. Cycle time, thus, is now the sum of the 99% signal attenuation time and the time required to reach the maximum signal. Two options remain: one genome per chip with 100 base pair strands or two genomes per chip with 240 base pair strands.

**Table 26: Proposed Proton III Technology**

| | Attenuation Time (Seconds) | Potential Ratio | Maximum Signal |
|---|---|---|---|
| **Current Proton II – 110nm node, 660 Million Wells** | 6 | 0.4433 | 28.95mV |
| **Proton III – 32nm node, 1.5 Billion Wells** | 2.23 | 0.6632 | 24.67mV |

| Strand Length (base pairs) | Cycle Time (sec) | Cycles | Time (hr) | Throughput (hr/genome) |
|---|---|---|---|---|
| **Current Proton II** | | | | |
| 200 | 8 | 1600 | 3.56 | 3.56 |
| **Proposed Proton III** | | | | |
| 100 | 4.63 | 800 | 1.03 | 1.03 |
| 240 | 4.63 | 1920 | 2.47 | 1.23 |

## 9.D. Optimization Conclusions

The potential ratio of the proposed Proton III, shown in Table 26, would have a 1.5 fold increase over the Proton II, effectively showing a significant accuracy increase by maximizing difference between homopolymers sequences. Accuracy with homopolymers can further be increased by utilizing a shorter base pair length. This allows for less dephasing and shortened possible homopolymer lengths. Sequencing throughput of the proposed Proton III using a 32nm node and Lead Titanate chip leads to a greater than a 3-fold increase in the sequencing throughput required compared to the Proton II. However, described in Chapter 11, the current, stock Proton sequencer has a run time of 8 hours, four of which is devoted to actual base flows and the remaining four is devoted to base-calling and alignment. The 3-fold increase in sequencing throughput described in this section only influences one half of the machine's run time. Therefore, throughput for the machine overall is only increased by 1.5 times. Implications on the IonSeq's overall throughput is further discussed in Chapter 11.

# 10. MARKET ANALYSIS

Next Generation Sequencing (NGS) has the potential to become a disruptive technology that will allow scientists to access DNA data like never before. It is the fastest growing and "most attractive" segments of a potentially $7.1 Billion Genomics Space[72]. The NGS market itself was worth just over one billion dollars in 2011, and is expected to double by 2016. The market is currently in a highly volatile growth stage, where new biotech companies sprout up constantly. However, IonSeq is confident in our ability to provide a novel service and occupy a niche in the biotechnology sector. This chapter will paint the market landscape and describe some of the competitors with which IonSeq must contend in order to achieve success.

## 10.A. MARKET OUTLOOK

Approximately 25 companies compete in the U.S. next-generation sequencing services market, with new competitors entering the market every year. Given the high expected growth rate, several

[72] Decisive Bio-Insights. (2013). Next generation sequencing: Market size, segmentation, growth and trends by provider. (2nd ed.). Culver City, CA: DeciBio, LLC.

competitors are expected to enter the market through 2016[73]. Advancements in technology are reducing the cost of sequencing systems, and making them more appealing to those who could not previously afford the equipment[74]. In addition, manufacturers are expanding their product line to include smaller, "personal sized" machines. Most importantly, the throughput processes themselves have improved, reducing cost and time per run'. There are also great advancements in the data processing software and data processing capabilities.

However, there does appear to be potential barriers in the NGS market that is of interest to IonSeq. NGS Service Providers are starting to appear, and we must work to take our market share. NGS Service Providers prove attractive to smaller laboratories as there is a lower cost, and quicker response time. We will not be competing with the four key players in the US: Illumina, Roche, Life Technologies, and Qiagen[75], but rather use them as vendors if necessary. We feel this is a relationship that will benefit all parties.

As previously mentioned, IonSeq's target market is a Direct-to-Consumer Sequencing Services market, providing exome and genome sequencing to individual customers and physicians. The NGS-DTC market is expected to form and mature over the next five years[76]. The average market price per genome is expected to decrease over the next few years, from the current price of approximately $4,000[77], to reach the X-Prize goal of $1,000 per genome. Services offered by companies targeting this market include whole-genome sequencing, exome sequencing, de novo sequencing. However, the scope of IonSeq's operations will be limited to genome and small chain

[73] Bird, C. (2012, May 1). Next-gen sequencing services: An expanding role in clinical applications opens new markets. *Genetic Engineering & Biotechnology News*, *32*(9), Retrieved from http://www.genengnews.com/gen-articles/next-gen-sequencing-services/4088/
[74] Companies and Markets. (2011). *Strategic analysis of the u.s. next generation sequencing markets*. Frost and SUllivan. Retrieved from http://www.companiesandmarkets.com/Market/Healthcare-and-Medical/Market-Research/Strategic-Analysis-of-the-U-S-Next-Generation-Sequencing-Markets/RPT915231
[75] Companies and Markets (2011). *Strategic*
[76] Bird, C. (2012) Next-gen
[77] Bird, C. (2012) Next-gen

sequences, with further expansion likely in the future. This NGS-DTC market is expected to reach $550 million by 2015, nearly half of the entire NGS market[78].

Due to the above circumstances, IonSeq is confident that its ability to offer customers a unique personal sequencing service will lead it to successes in the market. IonSeq's innovation spans across five categories critical to a successful business: Customer Value, Products and Services, Technical Differentiation, Process Technology, and Material Technology. The relationship between each of IonSeq's unique offerings is diagramed below, in Figure 38. As stated previously, IonSeq's value proposition lies in sequencing speed and throughput, which emerge as a result of the CMOS chips with ISFET sensors. IonSeq's services also allow customers to avoid burdening high capital costs for sequencing equipment and maintenance and labor costs involved in upkeep of that equipment.

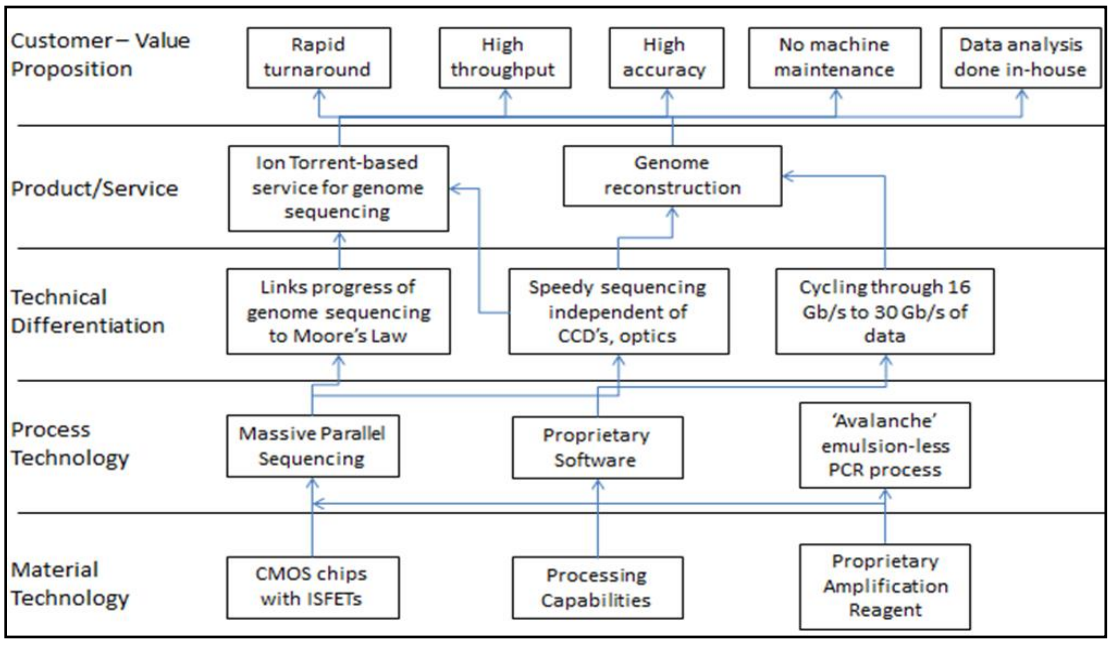



**Figure 38: Innovation Map for IonSeq**

[78] Bird, C. (2012) Next-gen

## 10.B. COMPETITOR ANALYSIS

While IonSeq is offering a unique service, it is important to be aware of the other genome sequencing technologies and platforms available in the market. Genome sequencing technology sees frequent technological advances and increases in efficiency and competitiveness. It will be crucial for IonSeq to keep up to date with breakthroughs in throughput achieved by any company and IonSeq's success in the long-run is dependent upon Ion Torrent R&D being able to remain competitive in the market. This section examines some of the competing technologies in the market.

### *10.B.I. COMPETING GENOME SEQUENCING PLATFORMS*

Genome Sequencing Platforms include current NGS market occupiers, who currently act as vendors for laboratories. These companies primarily sell equipment and engage in research to further sequencing technologies.

#### *10.B.I.1) ILLUMINA GA / HISEQ SYSTEM*

Illumina sequencing platforms are currently the most widely used platforms in the market. In 2006, the Genome Analyzer (GA) was released by Solexa and in 2007 Solexa was purchased by Illumina. The GA, like Ion Torrent, also uses the sequencing-by-synthesis approach, however it differs vastly in that reads are conducted through optical detection. The single DNA fragments are grafted onto the flowcell and are made to form clusters by bridge amplification. All four kinds of nucleotides, each attached to a different cleavable fluorescent dye and a removable blocking group, are flowed over the flowcell at the same time. Each nucleotide incorporation results in chain termination, and the fluorescence from each cluster is detected using a CCD camera. Reagents then need to be flown in to remove the fluorescent moiety and unblock the chain so that next round of nucleotides can be flowed[79].

[79] Liu (2012). Comparison

The greatest strength of the Illumina technology lies in the fact that it can generate an immense amount of data. The latest Illumina GA can yield outputs of 85 GB/run. In 2010, Illumina launched HiSeq 2000, which can yield outputs of 600 GB/run. The reported accuracy of this technology is very high, above PHRED scores of Q30. The dominant error type is substitutions rather than insertions or deletions[80]. Yet there are limitations to advancements in the technology. It is unlikely that the technology can yield read lengths greater than 200 bp because of signal decay and dephasing due to incomplete cleavage of fluorescent labels or terminating moieties. Limitations in CCD technology also pose obstacles to increases in throughput[81].

### *10.B.I.2) ROCHE 454 SYSTEM*

Roche 454 was the first commercially successful next generation system. It makes use of pyrosequencing, which makes its method of detection of incorporation different from Ion Torrents, but otherwise the technology is almost identical. Clonal amplification is achieved by emulsion PCR, which creates copies of each fragment on individual beads and the sequence on each bead is determined by pyrosequencing, which detects nucleotide incorporation using a flash of light, which is emitted when diphosphate, the bi-product of nucleotide incorporation, reacts with the enzymes sulfurylase and luciferase. The nucleotides are flown over the wells one at a time and the platform keeps track of which cells emit the flash of light for a particular nucleotide.

In 2005 Roche 454 was able to achieve read lengths of 100-150 bp and 20 Mb of data per run. In 2008 the 454 GS FLX Titanium System was launched which could attain 700 bp read lengths with 99.9% accuracy within 24 hours[82].

However, this technology has had trouble competing in the market with Illumina, given that it has not been able to generate nearly as much data, has greater per-base cost of sequencing and is greatly limited by homopolymers. Since there is no terminating moiety preventing multiple

[80] Quail, M., & et al, (2012). A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers.*BMC Genomics*, *13*, 341.
[81]Shendure, J., & Ji, H. (2008). Next generation dna sequencing. *Nature Biotechnology*, *26*(10), 1135-1145.
[82] Liu (2012). Comparison

consecutive incorporations in a given cycle, the length of all homopolymers has to be inferred from signal intensity. At longer homopolymer lengths this becomes difficult. Hence, indel errors are more dominant than substitution errors[83].

While Ion Torrent faces the same problem of homopolymers, it has greater potential of being able to read longer homopolymer lengths because it relies on rapidly evolving transistor technology and is not limited by advancements in CCD technology. Given the similarities in the two technologies, Ion Torrent also has the potential to achieve longer read lengths in the future.

#### *10.B.I.3) ABI SOLiD System*

The SOLiD platform was purchased by Applied Biosystems in 2006. It makes use of emulsion PCR to amplify the DNA fragments, but unlike the other platforms, which use sequencing by polymerization, the SOLiD platform uses sequencing by ligation. All possible 8-mer oligonucleotides are hybridized to the template simultaneously, but only those with a specific two-base-pair combination can bind strongly enough to be ligated onto the growing strand. The identity of the fifth base can be decoded by the color of the fluorescent dye at the unligated end of the 8-mer. After several rounds of ligation, the newly synthesized strand is melted off, and a second round of ligation is initiated with the new primer offset from the first by one base. In this way five rounds are ligation are carried out, which allows the complete sequence to be inferred.

Between 2007 and 2010 several SOLiD platforms were released, but none of these have managed to outcompete the yield of Illumina or the read lengths of 454. Hence its applications have become restricted to targeted resequencing and transcriptome research[84].

#### *10.B.I.4) Third Generation Sequencer Technology*

The third generation of genome sequencing is already under development, the most prominent technologies being those employed by Pacific Biosciences (PacBio) and Oxford

[83] Shendure (2008). Next-Generation
[84] Liu (2012), Comparison

Nanopore. These technologies eliminate the need for PCR, which shortens DNA preparation time, and also reduces bias and error caused by PCR. PacBio has introduced the single-molecule real-time (SMRT) sequencing, which makes use of direct observation of the enzymatic reaction in real time. While this technology is able to achieve the longest read lengths, 1300 bp, its throughput is much lower than those of second generation sequencers, making this technology unsuitable for service companies such as IonSeq. The Nanopore technology uses the concept of putting a thread of single-stranded DNA across α-haemolysin pore, which can cause different levels of disruptions in a continuous ionic flow based on which base is passed through it. While this technology is also very promising, it has not been developed sufficiently to determine its market competitiveness[85].

## *10.B.II. COMPETING GENOME SEQUENCING SERVICES COMPANIES*

As the cost of genome sequencing is rapidly declining, the genomics service industry is also becoming increasingly lucrative. As such IonSeq is likely to face competition from several sources.

### *10.B.II.1) COMPLETE GENOMICS*

Complete Genomics (CG) provides genome sequencing and analysis services. Much like IonSeq, CG is also dedicated solely to human DNA sequencing. CG provides two primary services – the Standard Sequencing Service and the Cancer Sequencing Service. The services largely differ from those of IonSeq's primarily because with each service from CG, the customers receive reports on summary statistics, variants including SNPs, indels, etc. These services are aimed at directly allowing customers to efficiently characterize the full spectrum of genetic variants that exist in the population and conducting large-scale genome-wide association studies.

There are two primary components of the CG sequencing technology: DNA nanoball (DNB) arrays and combinatorial probe-anchor ligation (cPAL) reads. The DNA fragments are packed onto a silicon chip and amplified such that all copies are connected in a head-to-tail configuration,

[85] Liu (2012), Comparison

forming long single molecules which then ball up into DNBs, which are approximately 200nm in diameter. A ligase enzyme attaches a different fluorescent molecule to each type of nucleotides in every DNB. The sequence is determined by imaging the fluorescence[86].

Complete Genomics has one of the most competitive sequencing platforms in the market, and has been successfully competing with the Illumina HiSeq 2000, especially in applications that seek to identify single nucleotide variants in human populations. CG is very likely to be IonSeq's biggest competitor.

### *10.B.II.2) GENE BY GENE DNA DTC*

DNA DTC was launched at the end of 2012, as a division of Gene by Gene, a company that provides on DNA testing focusing on ancestry, health, research and paternity. The company's newest division, DNA DTC, aims to utilize next generation sequencing of exomes and whole genomes for genome-wide association studies, human mitochondrial tests, and offer whole genome sequencing services. DNA DTC plans to use the Illumina HiSeq platform and is currently offering a price of $695 for exome sequencing[87].

DNA DTC is likely to provide competition for IonSeq going forward, but at present enough information is not available to determine at what level. DNA DTC's services also appear to be targeted more towards genomics research centers rather than clinical institutions, so it is likely that its market will not have a major overlap with that of IonSeq's.

### *10.B.II.3) EDGEBIO*

EdgeBio is a bioinformatics company specializing in next generation sequencing technologies and applications. EdgeBio has provided genomics services since 2009. They serve as a sequencing and bioinformatics provider for many companies and research institutions worldwide.

---

[86] Complete Genomics. (2011). introduction to complete genomics' sequencing technology. In *Complete Genomics Media*. Moutain View, CA: Complete Genomics. Retrieved from http://media.completegenomics.com/documents/Technology White Paper.pdf

[87] Croft, K. (2012, November 29). Gene by gene launches dna dtc: Offers highly reliable, cost-effective dna testing to institutional clients worldwide. *Market Watch: The Wall Street Journal*. Retrieved from http://www.marketwatch.com/story/gene-by-gene-launches-dna-dtc-2012-11-29

EdgeBio uses a broad range of high throughput sequencing platforms including Illumina HiSeq2000 and MiSeq, as well as Ion Torrent PGM and Proton. Their clients are able to create an online account with them through which they can monitor the progress of their projects[88].

EdgeBio's genomics services are likely to be in direct competition with those provided by IonSeq. There are even instances of technology overlap, so customers wishing to use Ion Torrent technology are going to have choices among service companies. However, since EdgeBio simply purchases equipment from Illumina, Ion Torrent and other companies, and has no affiliation whatsoever with them, they are not able to take advantage of the rapid advancements of the sequencing technologies. This is where IonSeq has a major advantage.

#### *10.B.II.4) 23ANDME*

23andMe is a personal genomics company operating since 2007. 23andMe focuses more on providing DNA testing services rather than DNA sequencing ones. The company genotypes the DNA using microarray technology (specifically, the Illumina OmniExpress Plus) to identify which genetic variant the individual possesses. This allows the customers to assess their inherited traits, genealogy and congenital risk factors. Genotyping services offered for $99[89].

23andMe will not be a direct competitor for IonSeq since it does not provide genome sequencing services. Genome sequencing has an inherent advantage over genotyping because it can provide the entire sequence as opposed to simply the information regarding which known variant an individual possesses. Genotyping by means of microarrays is also unable to identify previously unknown variants.

[88] Edge Bio. (n.d.). *Next generation dna sequencing services*. Retrieved from http://www.edgebio.com/sequencing
[89] 23andMe. (n.d.). *How it works*. Retrieved from https://www.23andme.com/howitworks/

# 11. STRATEGY AND IMPLEMENTATION

IonSeq's success will not only rest on its technical soundness and strong market position, but it will also require impeccable execution. This chapter will review the timeline for meeting Series A and Series B goals, delineate the structure of each work day, outline the general labor and material requirements, and review other business requirements.

## 11.A. MEETING SERIES A AND SERIES B GOALS

Series A will serve as the prototype stage, a proof-of-concept, of the Ion Torrent technology. IonSeq will purchase the capital equipment, support labor costs, and procure the materials necessary to sustain the sequencing of 10 genomes/day. For the first quarter of the first year, IonSeq will purchase one Proton Sequencer/Server and perform practice sequence runs using the DNA samples from the founding members. The CEO, CTO, CFO, secretary, and the engineers will be employed at this stage. IonSeq will be in frequent contact with research laboratories that employ

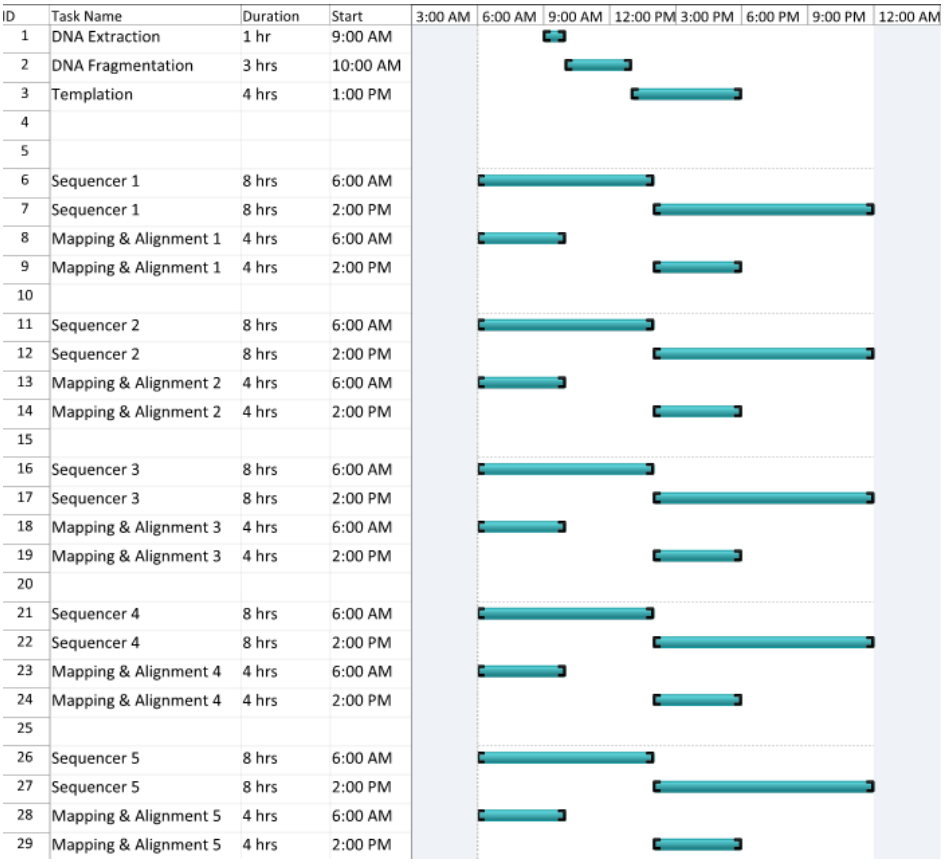
this technology for collaboration purposes. By the end of the first quarter, IonSeq intends to prove a throughput of 1 genome per day. For the second quarter, IonSeq will purchase an additional Sequencer/Server package, and this period will be devoted to proving the work flow that consists of running each sequencer twice in a 16 hour working day. By the end of the second quarter, a throughput of 4 genomes per day should be achieved. Once this is proven, a marketing manager will begin efforts in promoting IonSeq to potential clients. The third quarter will include the additional purchase and incorporation of the remaining three sequencers. Technicians will be hired and trained during this period as well. At the end of the third quarter, expected throughput is 8 genomes/day. The fourth quarter will involve perfecting the work flow, and achieving consistent 10 genomes/day throughput. A sales manager will be hired at this time and trained.

Series B investments will be expected at the end of the first year in preparation for commercial launch at the beginning of the second year. At the beginning of this phase, 15 additional sequencers/servers will be purchased. The engineers and technicians will train new, incoming technicians. The sales/marketing manager will be working in tandem to generate sales. Over the first three quarters during this year, IonSeq will have put online 5 sequencers in each quarter. During the fourth quarter, IonSeq will be able to prove 40 genomes/day output – though actual sales may prove to be less than throughput.

## 11.B. WORK DAY

The Gantt chart in Figure 39 illustrates Series A start-up phase of IonSeq, which supports a throughput of 10 genomes per day or 2,500 genomes over a year, using five Proton II machines (at specifications) with any multiplexing. Pre-sequencing steps can be executed in parallel during the day in preparation for sequencing in the next day. Each sequencer runs for 8 hours, out of which 2-4 hours is committed to collecting raw data from the sequencing chip and the remaining time is committed to base calling, alignment, and genome reconstruction. Mapping and alignment is

completed on an accompanying server, off of the sequencer machine; this process will take an additional 4 hours. The Gantt chart is built upon these specified times.



| ID | Task Name | Duration | Start |
|---|---|---|---|
| 1 | DNA Extraction | 1 hr | 9:00 AM |
| 2 | DNA Fragmentation | 3 hrs | 10:00 AM |
| 3 | Templation | 4 hrs | 1:00 PM |
| 4 | | | |
| 5 | | | |
| 6 | Sequencer 1 | 8 hrs | 6:00 AM |
| 7 | Sequencer 1 | 8 hrs | 2:00 PM |
| 8 | Mapping & Alignment 1 | 4 hrs | 6:00 AM |
| 9 | Mapping & Alignment 1 | 4 hrs | 2:00 PM |
| 10 | | | |
| 11 | Sequencer 2 | 8 hrs | 6:00 AM |
| 12 | Sequencer 2 | 8 hrs | 2:00 PM |
| 13 | Mapping & Alignment 2 | 4 hrs | 6:00 AM |
| 14 | Mapping & Alignment 2 | 4 hrs | 2:00 PM |
| 15 | | | |
| 16 | Sequencer 3 | 8 hrs | 6:00 AM |
| 17 | Sequencer 3 | 8 hrs | 2:00 PM |
| 18 | Mapping & Alignment 3 | 4 hrs | 6:00 AM |
| 19 | Mapping & Alignment 3 | 4 hrs | 2:00 PM |
| 20 | | | |
| 21 | Sequencer 4 | 8 hrs | 6:00 AM |
| 22 | Sequencer 4 | 8 hrs | 2:00 PM |
| 23 | Mapping & Alignment 4 | 4 hrs | 6:00 AM |
| 24 | Mapping & Alignment 4 | 4 hrs | 2:00 PM |
| 25 | | | |
| 26 | Sequencer 5 | 8 hrs | 6:00 AM |
| 27 | Sequencer 5 | 8 hrs | 2:00 PM |
| 28 | Mapping & Alignment 5 | 4 hrs | 6:00 AM |
| 29 | Mapping & Alignment 5 | 4 hrs | 2:00 PM |



**Figure 39: Gantt chart for work day flow for Series A startup period.**

For Series B, at a throughput of 40 genomes per day or 10,000 genomes over a year, IonSeq will employ 20 Proton II machines, organized over a similar work flow as shown for 5 Proton II machines. However, another option is to employ barcoding of the genomes, which would allow the sequencing of two genomes on one run of the Proton Sequencer. Figure 40 shows a revised Gantt chart, illustrating the need for only 3 Proton Sequencers in order to sequence at least 10 genomes in a day. The additional change is the continuous running of the servers to finish the mapping and alignments of the increased number of genomes processed per sequencer.

| ID | Task Name | Duration | Start |
|---|---|---|---|
| 1 | DNA Extraction | 1 hr | 9:00 AM |
| 2 | DNA Fragmentation | 3 hrs | 10:00 AM |
| 3 | Templation | 4 hrs | 1:00 PM |
| 4 | | | |
| 5 | | | |
| 6 | | | |
| 7 | Sequencer 1 | 8 hrs | 6:00 AM |
| 8 | Sequencer 1 | 8 hrs | 2:00 PM |
| 9 | Mapping & Alignment 1 | 4 hrs | 6:00 AM |
| 10 | Mapping & Alignment 1 | 4 hrs | 10:00 AM |
| 11 | Mapping & Alignment 1 | 4 hrs | 2:00 PM |
| 12 | Mapping & Alignment 1 | 4 hrs | 6:00 PM |
| 13 | Sequencer 2 | 8 hrs | 6:00 AM |
| 14 | Sequencer 2 | 8 hrs | 2:00 PM |
| 15 | Mapping & Alignment 2 | 4 hrs | 6:00 AM |
| 16 | Mapping & Alignment 2 | 4 hrs | 10:00 AM |
| 17 | Mapping & Alignment 2 | 4 hrs | 2:00 PM |
| 18 | Mapping & Alignment 2 | 4 hrs | 6:00 PM |
| 19 | Sequencer 3 | 8 hrs | 6:00 AM |
| 20 | Sequencer 3 | 8 hrs | 2:00 PM |
| 21 | Mapping & Alignment 3 | 4 hrs | 6:00 AM |
| 22 | Mapping & Alignment 3 | 4 hrs | 10:00 AM |
| 23 | Mapping & Alignment 3 | 4 hrs | 2:00 PM |
| 24 | Mapping & Alignment 3 | 4 hrs | 6:00 PM |

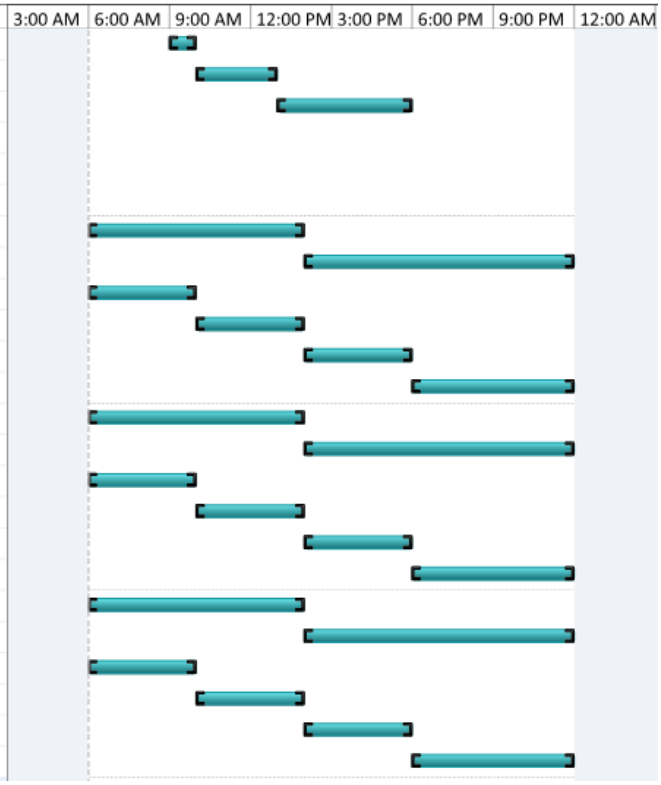



**Figure 40: Gantt Chart for Series A with barcoding option shows a need for fewer Proton II machines but the mapping & alignment servers will be run more often.**

Taking into consideration the optimizations covered in section 9, the sequencer run time can be reduced to about 6 hours. It was shown that by choosing lead titanate as the sensor surface material, the attenuation time can be cut down almost by a factor of 4. However, this will only affect the actual sequencing run and not the data analysis time (4-6 hours on the sequencer). Therefore, the run time for the sequencer is only reduced from 8 to 6 hours even despite the significant drop in signal attenuation time. As Figure 41 shows, this results in the mapping & alignment time to become the bottleneck.

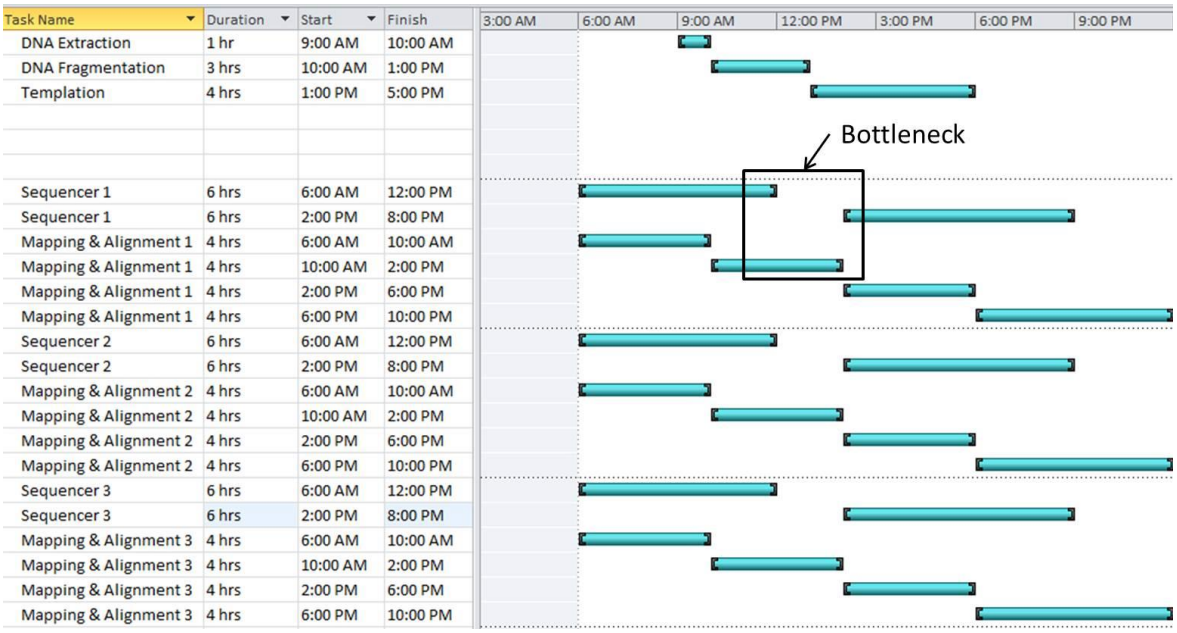



**Figure 41: This Gantt chart for the optimized sequencing chip shows that the mapping & alignment processes become the bottlenecks.**

The additional down time for the sequencers will prove wise as the extra time can be used for upkeep and maintenance. However, it is obvious that cutting down on the machine and server run times will also depend on optimizing the data analysis steps, not just the sequencing cycles on the chip.

## 11.C. Business Requirements

### *11.C.i. Supply Chain Requirements*

IonSeq's main suppliers will include Ion Torrent/Life Technologies for the sequencing chips and sequencing kits, and distributors such as Sigma Aldrich and Fischer Scientific for raw materials and wet laboratory supplies.

### *11.C.ii. Facility Requirements – Space*

There will be the need for sufficient dry and wet lab space to hold the 5 and 20 sequencers/servers required in Series A and Series B phases including the pre-sequencing

equipment. In addition, space will be needed for administrative purposes. For Series A, IonSeq will use a 1,400 sq. ft. lab space. For Series B, IonSeq will expand into a 3,700 sq. ft. facility. The sequencers can be stacked in pairs using a special rack, taking up approximately 5 sq. ft. of space. Each server will be placed next to the sequencers, and each server takes up about 2 sq. ft. Other important pieces of equipment include a storage refrigerator, the Ion Touch 2 systems, and the DNA extraction materials along with space reserved for administrative purposes.

## *11.C.III. EQUIPMENT REQUIREMENTS*

The Ion Proton System itself consists of the Sequencer and the Torrent Server. In-depth specifications are listed in Appendix D. The Sequencer not only performs the raw sequencing, but also carries out preliminary base-calling actions. The rest of the genome's base calling and alignment is performed off the Sequencer and on the Torrent Server. The Sequencer, weighing in at 130 lbs., consists of dual 8-core Intel® processors with 128 MB of memory, 11 TB of storage, and an NVIDIA® GPU processor all run by open-source Ubuntu® operating system. The Server is similarly specified, but includes 27 TB of storage, and two NVIDIA® GPU processors, weighing about 120 lbs. The entire Ion Proton™ system is connected to a cloud server with a large enough disk space to store the data generated from the Sequencer at a throughput of 40 genomes/day.

The size of a haploid human genome is roughly 3 giga-base pairs (Gb). The hardware space required to store 1 Gb is 250 megabytes (MB)[90]. Thus the space required to store data on the server for a 40 genomes/day throughput with 30X coverage is shown in the calculations in Equation 48.

**Equation 48**

$$40\ genomes * \frac{3\ Gb}{genome} * 30X\ coverage * \frac{250MB}{Gb} = 900\ GB$$

This is approximately 1 TB generated per day. After realignment, discussed below, the data from the raw reads will be discarded and the final report will be stored on the cloud computing server

[90] Discussion with Dr. Brian Gregory, Assistant Professor of Biology, University of Pennsylvania

for a period of one month. Within this time the customers will be able to securely access the server and download the report. Over the course of the month, 20 TB of storage on the cloud server will be needed. Delivery of the product will be carried out over this cloud service to eliminate the need for physical shipping. IBM SmartCloud Enterprise services will be considered and priced to handle the large genome files that IonSeq will be producing.

### *11.C.IV. INFORMATION TECHNOLOGY PLAN*

Apart from the equipment needed for sequencing, capable business computers will be needed for administrative purposes. Four machines will be purchased for Series A for use by the CEO, CTO, CFO, and secretary. Furthermore, IonSeq will invest into security systems for its IT infrastructure to protect its equipment and its products from outside interference. Considering the ambiguous HIPAA privacy regulations around DNA sequences and the potential for future rules, IonSeq will ensure that the data collected in the sequencing runs are safely stored and are transferred to the client over secure servers[91].

### *11.C.IV. LABOR*

The Chief Executive Officer will be in charge of the major strategy decisions for IonSeq and will have significant experience in field of next-generation DNA sequencing. The Chief Technology Officer will be the chief engineer in charge of coordinating the other engineers and R&D direction of the company. The Chief Financial Officer will be responsible for ensuring the financials of the company are in order and consistent positive cash flow is maintained. Furthermore, they will be in contact with investors. The sales manager will be responsible for reaching out to pharmaceutical companies and other potential clients. The marketing manager will be responsible for promoting IonSeq among the DNA sequencing community, both in industry and in academia. The secretary will

[91] "HIPAA, the Privacy Rule, and Its Application to Health Research." NCBI, Web. <http://www.ncbi.nlm.nih.gov/books/NBK9573/>.

be organizing the administrative side of the office, receiving incoming samples, and aiding the CFO in accounting needs. The engineers will be the managers of the technicians, contribute to R&D efforts, troubleshoot, and aid the workflow. The technicians will be responsible for pre-sequencing, running the sequencers, carrying out alignment on the servers, and troubleshooting. Details regarding the breakdown on compensation, equity, and working hours can be found in the following Financials section.

# 12. FINANCIAL ANALYSIS

As prefaced, IonSeq will be the service arm of Ion Torrent, engaging in a rapidly developing market. As a high-risk, high-reward biotechnology firm, IonSeq must be projected to perform well to satisfy its investors. The following financial analysis will address the profitability of this venture, exploring the investor's rate of return and net present value, after evaluating various income statements and cash flows.

Assuming that the base Ion Torrent technology is thus far market proven by research labs, the financial landscape for IonSeq will be addressed in the two phases: the scale-up stage and the revenue-generating stage. The first phase will generate no revenue, and will be funded with a Series A investment, covering the necessary equipment, labor, materials, and administrative costs to achieve a throughput of 10 genomes/day over 250 days/year, or 2,500 genomes/year. After the first year, with the assistance of a significant Series B funding, IonSeq will ramp up its facilities and labor requirements in order to achieve a four-fold increase in throughput – 10,000 genomes per year – through the remaining three years in this financial forecast. At that point, the rapid advancement of the field may pose new challenges for IonSeq, and new technologies will influence

the business model beyond this evaluation. IonSeq will seek acquisition by a major pharmaceutical corporation, aiming to create a significant footprint in personalized medicine.

## 12.A. REVENUE PROJECTIONS

Projecting sales is a crucial part for a profitability analysis. IonSeq will not intend to meet the $1,000 sequencing cost per genome that has been an established goal of the Archon XPrize; instead, understanding the value-add of a genome sequencing service, will price each genome at a premium. Sensitivity analyses performed later in this section as well as the history of cost/genome in Figure 42b will show a $2,000/genome price tag to be appropriate for healthy investors' rates of return.

Observing the rate of increase of human genomes sequenced over the past few years in Figure 42a, there is an exponential increase in the number of genomes, limited primarily by the sequencing technology. While it is difficult to forecast IonSeq sales, it is not difficult to see the growing demand for human genome sequencing. This growth can give some direction in revenue projections for IonSeq. Over the Series B, three year period, IonSeq predicts sequencing sales of 5,000, 7500, and 10,000 genomes as outlined in Table 27.

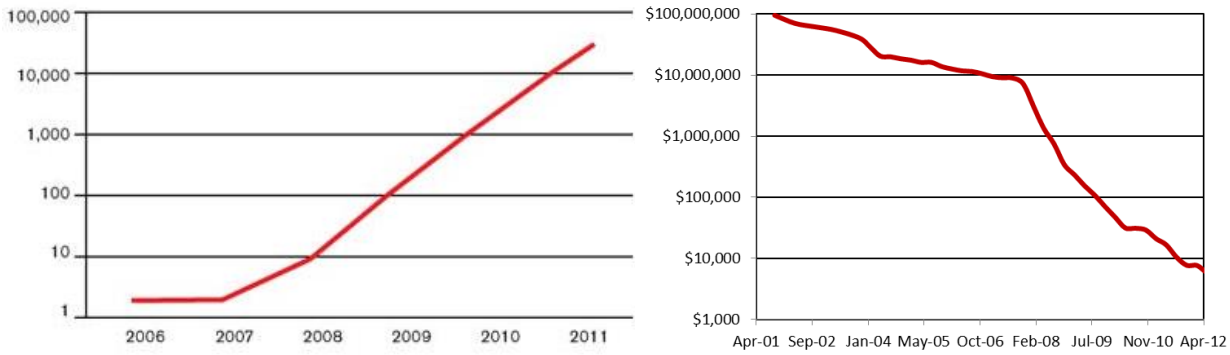



**Figure 42a: The number of genomes sequenced over the past six years has exponentially grown[92]. b: This rapid increase has primarily been the result of decreasing costs per genome[93].**

[92] Duncan, D. (2011, September 23). A dna tower of babel: As more and more people's genomes are decoded, we need better ways to share and understand the data. *MIT Technology Review*, Retrieved from http://www.technologyreview.com/news/425521/a-dna-tower-of-babel/

[93] Wetterstrand, K. (2013, Feburary 11). *dna sequencing costs: Data from the nhgri genome sequencing program (gsp)* . Retrieved from http://www.genome.gov/sequencingcosts/

**Table 27: Projection of Revenue for Years 1-4**

| | Year 1 | Year 2 | Year 3 | Year 4 |
|---|---|---|---|---|
| **Units Sold** | 2,500 | 5,000 | 7,500 | 10,000 |
| **Price per Unit** | $0.00 | $2,000.00 | $2,000.00 | $2,000.00 |
| **Net Revenue** | $0 | $10,000,000 | $15,000,000 | $20,000,000 |

## 12.B. VARIABLE AND FIXED COSTS

To help inform IonSeq's price tag per genome, costs must be evaluated, including variable costs (per genome), fixed equipment costs, overhead costs, and budgets for R&D and sales/marketing. These costs are extensively catalogued in Table 28and Appendix H.

**Table 28: Variable Costs of a Sequencing Run**

| Equipment | Price |
|---|---|
| - PII™ Chip | $350.00 |
| - Ion Plus Fragment Library Kit | $25.00 |
| - Ion PI™ Template OT2 200 Kit | $83.00 |
| - Ion PI™ Sequencing Kit | $67.50 |
| - Ion Proton Controls Kit | $100.00 |
| - 10 M NaOH [$71.40 for 100 mL] | $1.40 |
| - Isopropanol (99.7%) [$265 for 20 kg] | $0.34 |
| - Nuclease-free Water [$96.70 for 5 L] | $0.19 |
| - 10 L of 1N HCl [$91.20 for 10 L] | $0.09 |
| - Ethanol (200 proof) [$315 for 6-500 mL bottles] | $1.05 |
| - 500 mL TE Buffer, 1X Solution pH 8.0, low EDTA | $0.63 |
| - Pipette Tips | |
| P2 | $0.27 |
| P20 | $0.27 |
| P200 | $0.27 |
| P1000 | $0.32 |
| - Thin Wall PCR Tubes, Flat Cap | $5.00 |
| - 1.5 mL microcentrifuge tubes | $0.20 |
| - Agencourt AMPure XP - PCR Purification | $2.11 |
| - Agilent® High Sensitivity DNA Kit | $4.61 |
| - MagaZorb DNA Common Kit-200 | $2.00 |
| - Ion Xpress™ Barcode Adapters 1-96 Kit (partial purchase) | $1.56 |
| - Bioruptor® NGS 0.65 ml Microtubes for DNA Shearing (500 tubes) | $0.34 |
| - Pippin Prep™ Kit 2010 | $4.50 |

The significant assumption is the cost of the Proton chips. There is a lack of information regarding the actual production cost of a single chip. Based off Ion Torrent retail price, IonSeq decided to base the variable cost of the chip to be half of the quoted sales price. Instead of the quoted $700 price, IonSeq used a variable cost of $350 to evaluate profitability. In total, including reagents and other materials, the cost of sequencing a genome is $645.

The major pieces of capital equipment are listed in Table 29 and Appendix I. The most significant items are the Proton Sequencers and Servers. The bundle is quoted at $249,000. IonSeq will take this price as being an accurate value, and it will dominate total capital investment.

**Table 29: Components of Capital Equipment**

| Equipment | Unit Price |
|---|---|
| - Ion Proton II, including Ion Server | $224,000.00 |
| - Maxwell Research System | $30,000.00 |
| - Ion OneTouch 2 System | $19,000.00 |
| - Nitrogen (grade 4.8, 99.998% or better) | $70.00 |
| - Water Purification System (Elga Purelab Flex 3) | $5,000.00 |
| - Multistage gas regulator (VWR, 55850-422) | $375.00 |
| - Lab Freezer | $1,000.00 |
| - Uninterruptable Power Supply (UPS) | $200.00 |
| - Microcentrifuge | $1,995.00 |
| - Galaxy Mini Centrifuge | $401.25 |
| - Pipettes | |
| P2 | $335.00 |
| P20 | $297.00 |
| P200 | $297.00 |
| P1000 | $297.00 |
| - 1 L Glass Bottles | $9.40 |
| - Vortex Mixer | $800.00 |
| - Thermal Cycler | $8,000.00 |
| - Tygon Tubing | $2.00 |
| - Magnetic Stirrer | $230.00 |
| - Magnetic Stir Bars | 10 |
| - Vacuum filtration system (pore size 0.45 um) | $83.40 |
| - Orion 3-Star Plus Benchtop Meter Kit with probes | $752 |
| - Squirt bottles | $5.00 |
| - 50 mL Syringe | $1.85 |
| - DynaMag™-2 Magnet | $531 |
| - Agilent® 2100 Bioanalyzer® instrument | $19580 |
| - Heat Block/Water Bath | $160 |
| - Incubator | $183 |
| - BioRuptor® NGS Sonication System | $13000 |
| - Pippin Prep™ System | $15000 |

## 12.C. LABOR

During phase A, labor requirements are less than for Series B when throughput is quadrupled. In Phase A, a CEO, CFO, CTO, secretary, 2 engineers, 4 technicians, and a marketing manager will be employed. The engineers and technicians will oversee the pre-sequencing process as well as the five Ion Torrent Proton sequencing machines. The work day will be 16 hours, from 6 AM to 10 PM. Each technician and engineer will work 8 hour shifts. Two technicians will be responsible for all the pre-sequencing processes, over an 8 hour period. Two other technicians will oversee the sequencing machines in 8 hour shifts. An engineer will be on hand during each 8 hour shift as well.

In Phase B, a CEO, a CFO, a CTO, a secretary, 6 engineers, 10 technicians, a marketing manager, and a sales manager will be employed. The engineers and technicians will oversee the pre-sequencing process, which is still highly paralleled. However, they will be overseeing 20 Ion Torrent sequencing machines. The work day will be 16 hours, from 6 AM to 10 PM. Each technician and engineer will work 8 hour shifts. Two technicians will be responsible for all the pre-sequencing processes, over an 8 hour period. 8 other technicians will oversee the sequencing machines in 8 hour shifts. There will be 4 technicians overseeing 10 machines over their 8 hour shift. Three engineers will be on hand during each 8 hour shift.

## 12.D. LOCATION

IonSeq will be located in the suburbs of Boston, Massachusetts, due to the abundance of research centers around the Boston area (Massachusetts Institute of Technology, Massachusetts General Hospital, Harvard Medical School) and pharmaceutical companies located in the vicinity, including Merck, Teva Pharmaceuticals, Perkin Elmer Life Sciences, Celgene, and Novartis. The rent

for office space and lab space will approximately $30 per square foot for 1,400 sq. ft. for the Series A phase. Expanding for Series B will require 3,700 sq. ft. at $23 per square foot[94].

## 12.E. Other General and Administrative

Delivery of the product will be carried out over a cloud service to eliminate the need for physical shipping. IBM SmartCloud Enterprise services were considered and priced to handle the large genome files that IonSeq will be producing. To allow clients to access their sequences at any time over the period of a month will require $23,717 per month for the IBM SmartCloud service. At the expected throughput of 40 genomes/day or 800 genomes/month, that comes out to a $29 premium per genome, respectable considering the conveniences of using this cloud service.

Utilities, legal/accounting fees, telephone service, lab insurance, and other supplies and postage will make up the other general and administrative expenses.

## 12.F. Depreciation Schedule

IonSeq will employ the 5-year MACRS depreciation schedule to assist in increasing tax savings from capital expenditures. The depreciation discounts are 20%, 32%, 19.20%, 11.52%, 11.52%, and 5.76%. However, since this report is evaluating IonSeq on the four year time scale, only the first three percentages will be employed. Year 0 depreciation will not be tallied. Depreciation will count under the operating expenses, distributing the burden of the capital expenditures across the life time of the company.

## 12.G. Working Capital

Working capital is an important element in the financial health of the company, working to support the company's obligations until accounts receivable are on hand. Essentially, it covers the difference between the company's current assets and current liabilities, measuring the firm's

[94] Cummings Properties. (n.d.). *Affordable lab space for lease*. Retrieved from http://www.cummingsproperties.com/lab_space.htm

liquidity. This value includes cash reserve, inventory, accounts payable, and accounts receivables. It will be factored into the necessary capital investments in the first production year. At the beginning of the second production year, with sufficient cash on hand from revenue generated from the production year, working capital will not be significant factor. The value of the working capital accounted for in the first production year will be added back into the cash flow statement at acquisition.

Inventory for 7 days will be considered, expressed in Equation 49. These will be the sequenced genomes, in raw data form, ready to be shipped or transmitted to the customer.

**Equation 49**

$$Inventory\ (I) = \frac{Revenue}{365\ days} * 7\ days$$

Accounts receivables will be based upon 30 days, assuming that customers will have 30 days to pay, and this value is shown in Equation 50. This is based off the revenue generated.

**Equation 50**

$$Accounts\ Receivables\ (AR) = \frac{Revenue}{365\ days} * 30\ days$$

Accounts payable by the company will be based upon 30 days, which includes the costs of goods sold, as shown in Equation 51.

**Equation 51**

$$Accounts\ Payable\ (AP) = \frac{Cost\ of\ sales}{365\ days} * 30\ days$$

Cash reserve will cover 30 days of operation expenses, salaries, and other general expenses, and is calculated in Equation 52.

**Equation 52**

$$Cash\ Reserve\ (CR) = \frac{Operating\ Expenses}{365\ days} * 30\ days$$

In sum, the working capital is the difference between the company's current assets and current liabilities and can be expressed by Equation 53, taking all the previous elements above into account.

**Equation 53**

$$Working\ Capital = CR + I + AR - AP$$

## 12.H. KEY RATIOS

From the income statement, three key ratios can be calculated and each provides key information about the health of the company. The gross margin is calculated as the gross profit divided by revenue. This paints a picture of the company's profitability just based upon revenues and costs of goods sold. As an advanced biotechnology company, IonSeq seeks incoming revenue to be substantially greater than the costs to sequence. In Appendix J, Pro Forma 1 shows a gross margin of a healthy 67%. The operating margin is found by dividing pre-tax income by revenue. This illustrates the burdens of income taxes levied on the company, and IonSeq still has a healthy margin about 40%. The profit margin is final profitability ratio, the net income after tax divided by revenue. IonSeq maintains a profit margin of 25%-32% at the price point of $2,000.

Furthermore, from the cash flow statement, current and quick ratios are important metrics to ensure liquidity[95]. Ratios greater than one indicate that the company will be able to cover its liabilities with incoming cash. The current ratio is simply current assets divided by current liabilities. The quick ratio is the difference of current assets and inventory divided by current liabilities. In Pro Forma 1, IonSeq maintains healthy current and quick ratios around 3.60 to 4.74.

## 12.I. INVESTMENTS/EQUITY DISTRIBUTION

The Series A investment must cover total working capital, capital equipment, and all costs of materials needs to meet cited throughput, in the first year. Series B investment must cover the additional capital costs for upgrading our facilities for higher throughput as well as working capital

[95] Berman, Karen, and Joe Knight. *Financial Intelligence for Entrepreneurs*. Boston: Harvard Business, 2008. Print.

costs. Series A investors will receive a greater return on their investment due to the greater risk inherited in the startup phase. Series B will invest one year later in the first revenue-generating year. In the first year, Series A investors will invest and take 90% equity in the company, with the founding members receiving a 10% equity stake. Series B investors will make their investment at the beginning of year 2, and they will take a 32% share of the company, reducing Series A investors' equity stake to 61% and the founders to 7%. This breakdown is shown in Appendix J.

## 12.J. Determining Rate of Return

In order to determine the return on investments for both Series A and Series B investors, the ultimate value of IonSeq needs to be derived. Valuation is highly subjective and is an area of dispute among the founders, investors, and potential acquirers. For this analysis, IonSeq will use the Perpetuity Growth Model[96], which yields a prospective terminal value as calculated in Equation 54:

**Equation 54**

$$Terminal\ Value = Cash\ Flow * \frac{1 + growth\ rate}{discount\ rate - growth\ rate}$$

The discount rate will for the terminal value calculation will be taken to the discount rate attributed to Series B investors, or 25%. Series A investors will take a 50% discount rate due to the risk in their investment. At a sample growth rate of 5%, IonSeq's terminal value is $94,996,923, as opposed to $70,311,677 assuming a growth rate of 0%.

The net present value (NPV) of the company is the sum of present values of each yearly cash flow, and the present values can be calculated in Equation 55.

**Equation 55**

$$Present\ Value\ of\ Cash\ Flow = \frac{Discounted\ Cash\ Flow}{(1 + Discount\ Rate)^{Year}}$$

[96] Damodaran, Aswath. "Closure in Valuation." NYU Stern, n.d. Web. 31 Mar. 2013.

Discounted cash flow is the net earnings minus depreciation. Cash flows are delineated in Appendix J. The NPV of IonSeq is $39,322,347 at the end of year 4 at a proposed growth rate of 5%.

Modified Internal Rate of Return (MIRR), expressed in Equation 56, is a more desirable metric to measure the investors' rates of return[97]. Straightforward IRR makes a significant assumption that positive cash flows are reinvested into the company.

**Equation 56**

$$MIRR = \sqrt[n]{\frac{(Future\ Value\ of\ Positive\ Cash\ Flow\ at\ Reinvestment\ Rate)}{(Present\ Value\ of\ Initial\ Outlays\ at\ Finance\ Rate)}} - 1$$

Reinvestment rate is the rate of return the company can expect to earn from investing their capital in other low-risk financial vehicles—3 year Treasury Yield, 0.38% as of March 21st, 2013, and the finance rate is the annual percentage rate paid to lenders—this value can be estimated from the U.S. Treasury's Long Term Rate Data, 2.76% as of March 21st, 201398. In one instance, as seen in Pro Forma 1, Series A investors receive a MIRR of 102.98% and Series B investors have a MIRR of 93.43% after the end of year 4. These are appropriate returns on investments for a high-risk biotechnology startup.

## 12.K. SENSITIVITY ANALYSES

To better understand the impact of price/genome on returns on investment, a sensitivity analysis was carried out, where the price tag was changed from $1,000 to $2,000. The results of the MIRRs and IRRs are shown in Figure 43. With a price per genome set at $2,000, IonSeq's investors can expect to see very respectable returns on their investments. Furthermore, the use of MIRR over IRR is justified as IRR overestimates investors' true returns.

[97] "Modified Internal Rate Of Return - MIRR." Investopedia, Web. <http://www.investopedia.com/terms/m/mirr.asp>.
[98] U.S. Department of the Treasury, (2013). *Daily treasury long term rate data*. Retrieved from website: http://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/TextView.aspx?data=longtermrate

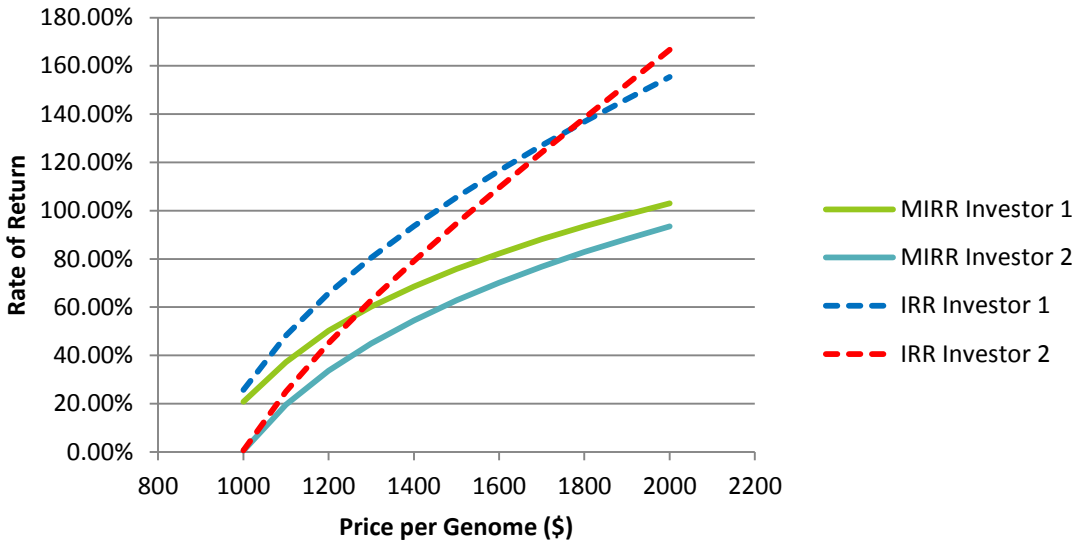



**Figure 43: Sensitivity analysis shows the rate of increase in IRR/MIRR falling off as price grows.**

Another important sensitivity analysis includes varying the prospective sales numbers. In Figure 44, healthy rate of returns exist even if the sale projections are overestimated by 25%. However, there is a steep drop if IonSeq severely underperforms. If the sales projections are overestimated by 75%, investors will have suffered losses on their investments. On the opposite side, the rates of return are limited by IonSeq's maximum throughput of 10,000 genomes per year, resulting in the leveling out as seen in the same figure.

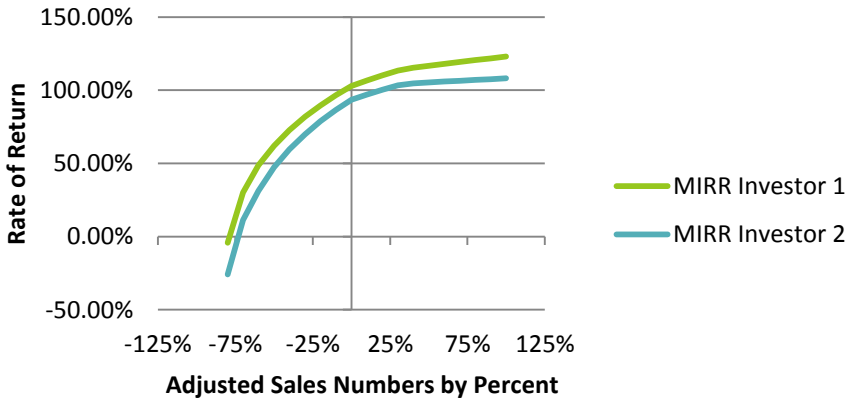



**Figure 44: An overestimation of sales by 75% would lead negative returns for investors.**

## 12.L. BARCODING SCHEME – BRIEF FINANCIAL PICTURE

If barcoding is employed with the 'Proton III' chip with 30x coverage and 1 billion wells, that slashes the capital requirements in half. The labor staff will be maintained as the same number of genomes must be processed. At the same price point, $2,000/genome, the MIRR for Series A investors is 127.65% and for Series B investors, 102.4%, with a NPV of $47,781,308 and a terminal value of $108,422,128 at a proposed 5% growth rate. The overall pro forma is found in Appendix K.

## 12.M. FINANCIAL TAKEAWAYS

From the complete financial analysis performed in this chapter, IonSeq has the potential to handsomely reward its investors with returns on investments as expected for this type of high-risk, high-reward biotechnology venture. Sensitivity analyses help to support IonSeq's claims for their investors' returns; even if sales are overestimated, both investors in Series A and Series B phases will still receive healthy returns of 103% and 93%, given investments of $3,682,886 in Year 1 and $4,510,491 in Year 2. Furthermore, IonSeq has structured its finances so to maintain high current and quick ratios to ensure quick liquidity when the acquisition or liquidity event arrives. At the end of the four year window, IonSeq expects an NPV of $39,322,347 and an overall terminal value of $94,996,923 at a conservative growth rate of 5%. This financial picture of IonSeq demonstrates a remarkably strong candidate for investment.

# 13. CONCLUSIONS

Taking advantage of the scalability of semiconductor technology in the context of genomics and competitively entering the marketplace with a unique service business model, IonSeq is projected to be a significant player in the rapidly expanding genomics industry. The current Proton II technology can be employed to deliver rapid, accurate genome sequences—at a throughput of 10,000 genomes per year—with healthy returns on investment at a cost per genome of only $2,000.

As the industry is rapidly evolving, innovations will most certainly challenge IonSeq to remain ahead of the curve. The team has a firm understanding of the technology and is fully aware of the technical bottlenecks, from nucleotide kinetics to dephasing and error rates, which limit genomic throughput. The potential solutions presented here for a proposed Proton III chip, including the transition to different sensor materials and more advanced manufacturing standards, indicate IonSeq's commitment to finding new ways to deliver sequenced genomes faster to its clients. The era of personalized medicine is here, and IonSeq is here to deliver.

# 14. ACKNOWLEDGEMENTS

We would like to express our greatest gratitude for the guidance and advice provided by Dr. John Crocker and Dr. Warren Seider during every single weekly meeting. It was always a pleasure to engage in challenging discussions regarding the intricacies of this design project.

We would also like to thank Dr. Brian Gregory for his insight into next generation DNA sequencing and his willingness to share his valuable experience with us.

Furthermore, we deeply appreciate the industrial consultants who stuck with us through a topic that was not in their areas of expertise, and who found ways, regardless, to contribute their knowledge ranging from kinetics modeling to financial analyses that proved to be very helpful. We would particularly like to thank Mr. John Wismer and Dr. Richard Bockrath.

# 15. APPENDICES

## APPENDIX A

### Determining if library amplification is required

The unamplified library can be quantified by quantitative PCR (qPCR) with the Ion Library Quantitation Kit. This kit directly determines the library dilution that gives a suitable concentration for template preparation (~26 pM).

1. Determine the Template Dilution Factor (TDF) for the unamplified library with the Ion Library Quantitation Kit.

2. Dilute the **unamplified** library for the qPCR as follows.
   - 100 ng-input: 1:1000 dilution
   - 1 µg-input: 1:2000 dilution

3. Calculate the number of template preparation reactions that can be performed with the unamplified library as follows:

$$\text{No. of reactions} = \frac{\text{library volume in µL} \times \text{TDF}}{\text{volume per template preparation reaction in µL}}$$

The volume per template preparation reaction is:
- 20 µL if using the Ion OneTouch™ 200 Template Kit v2 or the Ion OneTouch™ 200 Template

If the estimated number of template preparation reactions is sufficient for the experimental requirements, no amplification is necessary.

Source: Ion Torrent User Guide. Ion Xpress™ Plus gDNA Fragment Library Preparation. 2012. Publication Part Number 4471989 Rev. E downloaded from Life technologies website.
http://products.invitrogen.com/ivgn/product/4471269

## APPENDIX B

# Qualifying and pooling barcoded libraries

1. **Assess the size distribution of individual barcoded libraries**
   Analyze an aliquot of each barcoded library with an Agilent® High Sensitivity DNA Kit, as indicated in the following table.

| Library type | Unamplified | | Amplified | |
|---|---|---|---|---|
| Input amount | 100 ng | 1 µg | 50–100 ng | 1 µg |
| Library aliquot | 1 µL | 1 µL, 1:5 | 1 µL, 1:10 | 1 µL, 1:10 |

2. **Pool barcoded libraries using qPCR (unamplified libraries or amplified libraries)**

   1. Use the Ion Library Quantitation Kit to directly determine the library dilution by quantitative real-time PCR (qPCR) for each individual barcoded library

   2. Dilute each barcoded library according to its dilution factor. This will result in a library concentration of ~26 pM.

   3. Prepare at least 20 µL of a barcoded library pool by mixing equal volumes of the diluted barcoded libraries. The library pool will be at the correct concentration for template preparation using the appropriate template kit. No further dilution of the library pool is necessary.

3. **Pool barcoded libraries using Bioanalyzer® quantitation (amplified libraries only)**

   1. From the Bioanalyzer® analysis used to assess the individual barcoded library size distribution, determine the molar concentration in pmol/L of each barcoded library using the Bioanalyzer® software.

   2. Prepare an equimolar pool of barcoded libraries at the highest possible concentration.

   3. Determine the molar concentration of the library pool.

   4. Determine the dilution factor that gives a concentration of ~26 pM. This concentration is suitable for template preparation using either Ion Xpress™ Template Kits or Ion OneTouch™ Template Kits. Use the following formula:

$$\textbf{Dilution factor} = \frac{\textbf{Library pool concentration in pM}}{\textbf{26 pM}}$$

   **Example**
   The library pool concentration is 10,000 pM.
   Dilution factor = 10,000 pM/26 pM = 385
   Thus, 1 µL of library pool mixed with 385 µL of Low TE (1:385 dilution) yields approximately 26 pM. Use this library dilution for template preparation.

Source: Ion Torrent User Guide. Ion Xpress™ Plus gDNA Fragment Library Preparation. 2012. Publication Part Number 4471989 Rev. E downloaded from Life technologies website.
http://products.invitrogen.com/ivgn/product/4471269

## APPENDIX C

# Determining Template Dilution Factor for emPCR

For 10–30% of positive Ion Sphere™ Particles, a Template Dilution Factor is required that gives 70 × 106 molecules per 5 µL

Use a conversion factor of 8.3 nM = 5 × 109 molecules/µL for the region of interest, excluding peaks outside of the desired range; and use the following formula:

$$\text{Template Dilution Factor} = \text{Library concentration in nM} * \frac{5 * 109 \frac{\text{molecules}}{\mu\text{L}}}{8.3\ \text{nM}} * \frac{5\mu\text{L}}{70 * 106\ \text{molecules}}$$

**Example**

If the library concentration is 10 nM,

$$\text{Template Dilution Factor} = 10\ \text{nM} * \frac{5 * 109 \frac{\text{molecules}}{\mu\text{L}}}{8.3\ \text{nM}} * \frac{5\mu\text{L}}{70 * 106\ \text{molecules}} = 430$$

Thus, 1 µL of library mixed with 429 µL of Low TE (1:430 dilution) yields approximately 70 × 106 molecules per 5 µL.

| Dilution | Dilution factor | Target number of molecules per reaction† | Library volume | Water |
|---|---|---|---|---|
| 1 | ½ × Template Dilution Factor | $140 \times 10^6$ | 1 µL | [(½ × Template Dilution Factor)–1] µL |
| 2 | Template Dilution Factor | $70 \times 10^6$ | 50 µL Dilution 1 | 50 µL |
| 3 | 2 × Template Dilution Factor | $35 \times 10^6$ | 50 µL Dilution 2 | 50 µL |

† Molecules/5 µL input in the amplification solution.

Source: Ion Torrent User Guide. Ion Xpress™ Plus gDNA Fragment Library Preparation. 2012. Publication Part Number 4471989 Rev. E downloaded from Life technologies website.
http://products.invitrogen.com/ivgn/product/4471269

## APPENDIX D

| **Specifications for the Ion Proton™ Sequencer** | |
|---|---|
| Working environment (for indoor use only) | Temperature: 68-77° F (20-25° C)<br>Humidity: 40-60%, noncondensing<br>Altitude: <6,500 ft (2,000 m)<br>Clearances:<br>12 in (30.5 cm) in rear<br>4 in (10 cm) on left side<br>4 in (10cm) on right side<br>4 in (10 cm) from front edge of bench to sequencer bezel<br>36 in (90 cm) aisle in front of bench for operator access<br><br>Optional rack mounting with two Ion Proton™ Sequencers per rack† |
| Gas Supply | Connection: 0.25 in push-to-connect fitting<br>Pressure: 30 psi<br>Composition: nitrogen (grade 4.8, 99.998% or better) |
| Other connections | Ethernet: 1 GigE<br>USB: 2x USB 2.0 |
| Power | Voltage: 100 V (min) to 240 V (max)<br>Current: 14 A (max)<br>Frequencing: 50/60 Hz<br>Power Draw: 1,350 W |
| Dimensions | Width: 21.3 in/54.2 cm<br>Depth: 30.5 in/77.5 cm<br>Height: 18.7in/47.4 cm |
| Weight | Crated for shipment: 200 lb/90.7 kg<br>Free-standing: 130 lb/59 kg |
| Instrument compute hardware | Processor: Dual 8-core Intel® Xeon® Sandy Bridge<br>Memory: 128 GB RAM<br>FPGA: Dual Altera® Stratix® V<br>GPU processor: 1 x NVIDIA® Tesla® C2075<br>Storage: 11 TB (SSD and HDD)<br>Operating system: Ubuntu® 11:10 |
| **Specifications for the Proton™ Torrent Server*** | |
| Product configuration | A single free standing tower computer appliance, included with the purchase of the Ion Proton™ System. Includes Torrent Suite Software with all necessary software components to deliver signal processing, base calling, read alignment, and variant calling. |
| Processor | Dual 8-core 2.9 GHz CPUs |
| Memory | 128 GB RAM |

| | |
|---|---|
| GPU processor | 2x NVIDIA® Tesla® GPUs |
| Storage (approx.) | 27 TB (sufficient for storage of >50 Ion PI™ Chip runs) |
| Operating system | Ubuntu 10.04 |
| Dimensions (approx.) | Width: 8.5 in/21.8 cm<br>Depth: 28 in/71.4 cm<br>Height: 17 in/43 cm |
| Weight (approx.) | 120 lb/55 kg |
| Power | Voltage: 100 V (min) to 240 V (max)<br>Frequency: 50/60 Hz<br>Current: 12 A (max)<br>Power Draw: 1,100 W |
| **Specifications for the Ion OneTouch™ System** | |
| System | Ion OneTouch™ System (Cat No. 4470001) includes:<br>• Ion OneTouch™ Instrument<br>• Ion OneTouch™ ES |
| Dimensions and weight | • Ion OneTouch™ Instrument:<br>(12 in x 16 in x 14 in, 23 lb; 30 cm x 41 cm x 36 cm, 10.4 kg)<br>• Ion OneTouch™ ES:<br>(9.5 in x 12.5 in x 11 in, 12 lb; 24 cm x 32 cm x 28 cm, 5.4 kg) |
| System run time | 4 hours total time, minutes of hands-on time |
| Throughput | Supports template preparation for Ion 314™ chips,<br>Ion 316™ chips, and Ion 318™ chips* |
| Library types | Supports template preparation with a broad range of libraries used for various applications:<br>• Genomic DNA (fragment and mate-paired)<br>• Amplicon<br>• RNA (cDNA) |
| Operating environment | Temperature: 15-25°C; humidity: 20–80%, noncondensing |
| Consumables | Ion OneTouch™ System Template Kit (Cat No. 4468660) |
| Power requirements | 110/220 V (US/International) |
| Multiplexing | Up to 384 barcoded libraries for DNA- or RNA-based applications |

Source: "The Ion Proton™ System: Rapid Genome-scale Benchtop Sequencing." Life Technologies. <http://tools.invitrogen.com/content/sfs/brochures/CO111809_Specification%20Sheet_Ion%20Proton%20System_0712.pdf>.

## APPENDIX E

### MATLAB Code for Kinetics and Signal Model

```
%% Design Parameters, changeable
WellDiameter = 0.70; %um
BeadSize = WellDiameter/1.25; %um, should be a large fraction of well
diameter
dNTPconc = 100; %uM
Strands = 100000; %number of templates on a bead

Temp = 310; %K, must be aware of temp sensitivity for polymerases
pHbulk = 8;

select_coeff = 0.93; %material dependent
pHpzc = 7; %material dependent, "pH at point of zero charge"

tau_0 = 130; %microseconds, material dependent, ranges from 60-200, smaller
means signal attenuates faster

%Set time
tf = 4;
dt_inc = 0.001;
t_inc = 0:dt_inc:tf;

%Flow
VolFlowRate = 4; %mL/s

%%do not modify

WellDepth = WellDiameter; %um
WellVolume = 3.14*((WellDiameter*10^-4)/2)^2*(WellDepth*10^-4); %mL

BeadVolume = 4*3.14/3*(BeadSize/2*10^-4)^3; %mL
WellVolumeAvail = WellVolume - BeadVolume; %mL

kB = 1.38*10^-23;
q = 1.60*10^-19; %coulombs, elementary charge

ProtonDiff = 9*10^-9; %m^2/s
tau_p = WellDepth^2/(ProtonDiff*(10^6)^2); %mean diffusion time for protons
out of well

%Signal Generation Values
deltaPSI_deltaPH = ((2.3*kB*Temp)/q)*select_coeff*1000;
PSI_not = (pHbulk-pHpzc)*deltaPSI_deltaPH;

dielectric = 80.1;
permittivity = 8.854*10^-12; %Farads/m
Zs = 1; %charge number of ionic species
Cs_initial = (10^-pHbulk)*exp(-q/(kB*Temp)*PSI_not/1000); %molar
concentration of species, M
```

```
IonicStrength_initial = 0.5*Zs^2*Cs_initial;
DebyeScreeningLength_initial = 0.3/(IonicStrength_initial^0.5)/10^9;
C_dl_initial = dielectric*permittivity/DebyeScreeningLength_initial;

%Signal Attenuation
tau_s = tau_0*(10^(pHbulk/2))/10^6; %seconds

%%Kinetic Data
%k_pol, s^-1
AtoA_kpol = 0.0036;
AtoC_kpol = 0.1;
AtoG_kpol = 0.042;
AtoT_kpol = 45;

CtoA_kpol = 0.1;
CtoC_kpol = 0.003;
CtoG_kpol = 43;
CtoT_kpol = 0.038;

GtoA_kpol = 0.05;
GtoC_kpol = 37;
GtoG_kpol = 0.066;
GtoT_kpol = 1.16;

TtoA_kpol = 25;
TtoC_kpol = 0.012;
TtoG_kpol = 0.16;
TtoT_kpol = 0.013;

%KD, uM
AtoA_KD = 25;
AtoC_KD = 160;
AtoG_KD = 250;
AtoT_KD = 0.8;

CtoA_KD = 540;
CtoC_KD = 140;
CtoG_KD = 0.9;
CtoT_KD = 360;

GtoA_KD = 500;
GtoC_KD = 0.8;
GtoG_KD = 150;
GtoT_KD = 70;

TtoA_KD = 0.6;
TtoC_KD = 180;
TtoG_KD = 200;
TtoT_KD = 57;

% Observed rate constant
AonA = ((AtoA_kpol*dNTPconc)/(AtoA_KD+dNTPconc));
AonC = ((AtoC_kpol*dNTPconc)/(AtoC_KD+dNTPconc));
AonG = ((AtoG_kpol*dNTPconc)/(AtoG_KD+dNTPconc));
```

```
AonT = ((AtoT_kpol*dNTPconc)/(AtoT_KD+dNTPconc));
ConA = ((CtoA_kpol*dNTPconc)/(CtoA_KD+dNTPconc));
ConC = ((CtoC_kpol*dNTPconc)/(CtoC_KD+dNTPconc));
ConG = ((CtoG_kpol*dNTPconc)/(CtoG_KD+dNTPconc));
ConT = ((CtoT_kpol*dNTPconc)/(CtoT_KD+dNTPconc));
GonA = ((GtoA_kpol*dNTPconc)/(GtoA_KD+dNTPconc));
GonC = ((GtoC_kpol*dNTPconc)/(GtoC_KD+dNTPconc));
GonG = ((GtoG_kpol*dNTPconc)/(GtoG_KD+dNTPconc));
GonT = ((GtoT_kpol*dNTPconc)/(GtoT_KD+dNTPconc));
TonA = ((TtoA_kpol*dNTPconc)/(TtoA_KD+dNTPconc));
TonC = ((TtoC_kpol*dNTPconc)/(TtoC_KD+dNTPconc));
TonG = ((TtoG_kpol*dNTPconc)/(TtoG_KD+dNTPconc));
TonT = ((TtoT_kpol*dNTPconc)/(TtoT_KD+dNTPconc));

%%%%%%%
n = 3; %homopolymer length
NumberofProtonsProduced_AtoT = zeros(n,length(t_inc));
NumberofProtonsAfterDiffusion_AtoT = zeros(n,length(t_inc));
NumberofProtonsAfterDiffusion = zeros(n,length(t_inc));
pH_transient_AtoT = zeros(n,length(t_inc));
TotalProtonsProduced = zeros(n,length(t_inc));
TotalProtonsAfterDiffusion = zeros(n,length(t_inc));
TimeCutoff = zeros(1,n);

dNTPconc_transient = zeros(n,length(t_inc));
dNTPconc_transient(1,:) = dNTPconc;

SNR = zeros(n,length(t_inc));
Error = zeros(n,length(t_inc));

for y = 1:length(t_inc)
    dNTPconc_transient(1,y) = dNTPconc*(1-exp(-t_inc(y)*AonT));
    dNTPconc_transient(2,y) = dNTPconc*(1-exp(-t_inc(y)*AonT))*(1-exp(-
t_inc(y)*AonT/2));
    dNTPconc_transient(3,y) = dNTPconc*(1-exp(-t_inc(y)*AonT))*(1-exp(-
t_inc(y)*AonT/3));
    dNTPconc_transient(4,y) = dNTPconc*(1-exp(-t_inc(y)*AonT))*(1-exp(-
t_inc(y)*AonT/4));
end

%% Incorporation
%for base A to strand T
d = 0; e = 0; f = 0;

%1 Base
for i=1:length(t_inc)
    NumberofProtonsProduced_AtoT(1,i) =
AonT*dNTPconc*t_inc(i)*WellVolumeAvail*6.022*10^14;
    if d == 0
        TimeCutoff(1) = t_inc(i);
    end
    if NumberofProtonsProduced_AtoT(1,i)>Strands
        NumberofProtonsProduced_AtoT(1,i)=Strands;
        NumberofProtonsAfterDiffusion_AtoT(1,i) =
NumberofProtonsAfterDiffusion_AtoT(1,t1);
```

```
        NumberofProtonsAfterDiffusion(1,i) =
NumberofProtonsAfterDiffusion_AtoT(1,t1)*exp(-t_inc(i)-
(TimeCutoff(1))/tau_p);
        d = 1;
        pH_transient_AtoT(1,i) = -
log10((NumberofProtonsAfterDiffusion_AtoT(1,i)/(6.022*10^23)/WellVolumeAvail*
1000+10^-8));
    else
        NumberofProtonsAfterDiffusion_AtoT(1,i) =
AonT*dNTPconc*WellVolumeAvail*6.022*10^14*(tau_p*t_inc(i)-tau_p^2*(1+exp(-
t_inc(i)/tau_p)));
        NumberofProtonsAfterDiffusion(1,i) =
AonT*dNTPconc*WellVolumeAvail*6.022*10^14*(tau_p*t_inc(i)-tau_p^2*(1+exp(-
t_inc(i)/tau_p)));
        t1 = i;
        pH_transient_AtoT(1,i) = -
log10((NumberofProtonsAfterDiffusion_AtoT(1,i)/(6.022*10^23)/WellVolumeAvail*
1000+10^-8));
    end
    SNR(1,i) = (NumberofProtonsProduced_AtoT(1,i))^0.5;
end

TotalProtonsProduced(1,:) = NumberofProtonsProduced_AtoT(1,:);
TotalProtonsAfterDiffusion(1,:) = NumberofProtonsAfterDiffusion_AtoT(1,:);
TotalProtonsPostDiffusion(1,:) = NumberofProtonsAfterDiffusion(1,:);

%2 Bases
for m=1:length(t_inc)
    NumberofProtonsProduced_AtoT(2,m) = (AonT*dNTPconc*t_inc(m) +
dNTPconc*(exp(-AonT*t_inc(m))-1)) * 6.022*10^14 * WellVolumeAvail;
    if e == 0
        TimeCutoff(2) = t_inc(m);
    end
    if NumberofProtonsProduced_AtoT(2,m)>Strands
        NumberofProtonsProduced_AtoT(2,m) = Strands;
        NumberofProtonsAfterDiffusion_AtoT(2,m) =
NumberofProtonsAfterDiffusion_AtoT(2,t2);
        NumberofProtonsAfterDiffusion(2,m) = Strands*exp((-t_inc(m)-
TimeCutoff(2))/tau_p); %find time at which number of protons produced =
Strands
        e = 1;
    else
        NumberofProtonsAfterDiffusion_AtoT(2,m) =
6.022*10^14*WellVolumeAvail*(dNTPconc*(tau_p*exp(-AonT*t_inc(m))/(1-
AonT*tau_p)-tau_p)+AonT*dNTPconc*tau_p*(t_inc(m)-tau_p));
        NumberofProtonsAfterDiffusion(2,m) =
6.022*10^14*WellVolumeAvail*(dNTPconc*(tau_p*exp(-AonT*t_inc(m))/(1-
AonT*tau_p)-tau_p)+AonT*dNTPconc*tau_p*(t_inc(m)-tau_p));
        t2 = m;
    end
end

%Total for 2 bases
TotalProtonsProduced(2,:) = NumberofProtonsProduced_AtoT(2,:) +
NumberofProtonsProduced_AtoT(1,:);
```

```
TotalProtonsAfterDiffusion(2,:) = NumberofProtonsAfterDiffusion_AtoT(2,:) +
NumberofProtonsAfterDiffusion_AtoT(1,:);
TotalProtonsPostDiffusion(2,:) = NumberofProtonsAfterDiffusion(2,:) +
NumberofProtonsAfterDiffusion(1,:);
SNR(2,:) = (TotalProtonsProduced(2,:)).^0.5;

%3 Bases
for o=1:length(t_inc)
    NumberofProtonsProduced_AtoT(3,o) = (1/3)*dNTPconc*(3*AonT*t_inc(o)-
2*exp(-3*AonT*t_inc(o)/2)+3*exp(-AonT*t_inc(o))+6*exp(-AonT*t_inc(o)/2)-7) *
6.022*10^14 * WellVolumeAvail;
    if f == 0
        TimeCutoff(3) = t_inc(o);
    end
    if NumberofProtonsProduced_AtoT(3,o)>Strands
        NumberofProtonsProduced_AtoT(3,o) = Strands;
        NumberofProtonsAfterDiffusion_AtoT(3,o) =
NumberofProtonsAfterDiffusion_AtoT(3,t3);
        NumberofProtonsAfterDiffusion(3,o) = Strands*exp((-t_inc(o)-
TimeCutoff(3))/tau_p); %find time at which number of protons produced =
Strands
        f = 1;
    else
        NumberofProtonsAfterDiffusion_AtoT(3,o) =
6.022*10^14*WellVolumeAvail*( (1/3)*dNTPconc*(3*AonT*tau_p*(t_inc(o)-tau_p) -
4*tau_p*exp(-(3*AonT*t_inc(o)/2))/(2-3*AonT*tau_p) + 3*tau_p*exp(-
AonT*t_inc(o))/(1-AonT*tau_p) + 12*tau_p*exp(-AonT*t_inc(o)/2)/(2-AonT*tau_p)
- 7*tau_p) );
        NumberofProtonsAfterDiffusion(3,o) = 6.022*10^14*WellVolumeAvail*(
(1/3)*dNTPconc*(3*AonT*tau_p*(t_inc(o)-tau_p) - 4*tau_p*exp(-
(3*AonT*t_inc(o)/2))/(2-3*AonT*tau_p) + 3*tau_p*exp(-AonT*t_inc(o))/(1-
AonT*tau_p) + 12*tau_p*exp(-AonT*t_inc(o)/2)/(2-AonT*tau_p) - 7*tau_p) );
        t3 = o;
    end
end

%Total for 3 bases
TotalProtonsProduced(3,:) = NumberofProtonsProduced_AtoT(3,:) +
TotalProtonsProduced(2,:);
TotalProtonsAfterDiffusion(3,:) = NumberofProtonsAfterDiffusion_AtoT(3,:) +
TotalProtonsAfterDiffusion(2,:);
TotalProtonsPostDiffusion(3,:) = NumberofProtonsAfterDiffusion(3,:) +
TotalProtonsPostDiffusion(2,:);
SNR(3,:) = (TotalProtonsProduced(3,:)).^0.5;

%% Signal
Signal = zeros(n,length(t_inc));
SignalMaintain = zeros(1,n);

Sigmas = 3; %standard deviations

%Signal for 1 base
for j=1:length(t_inc)
    if (AonT*dNTPconc*t_inc(j)*WellVolumeAvail*6.022*10^14) < Strands %if the
number of protons produed is less than the number of possible incorporations
```

```
        Signal(1,j) = PSI_not*(1-
C_dl_initial/(dielectric*permittivity/(0.3/10^9)*0.5^0.5*Zs*(AonT*dNTPconc*We
llVolumeAvail*6.022*10^14*(tau_p*t_inc(j)-tau_p^2*(1+exp(-
t_inc(j)/tau_p)))/(6.022*10^23*WellVolumeAvail)*1000*exp(-
q/(kB*Temp)*PSI_not/1000)+Cs_initial)^0.5))*exp(-t_inc(j)/tau_s);
        SignalMaintain(1,1) = Signal(1,j);
        Error(1,j) = Sigmas*Signal(1,j)/SNR(1,j);
    else
        break
    end
end

for k = 1:(length(t_inc)-j)
    Signal(1,k+j-1) = SignalMaintain(1,1)*exp(-t_inc(k+1)/tau_s);
end

%Signal for 2 bases
for l = 1:length(t_inc)
    if TotalProtonsProduced(2,l) < 2*Strands
        Signal(2,l) = PSI_not*(1-
C_dl_initial/(dielectric*permittivity/(0.3/10^9)*0.5^0.5*Zs*(TotalProtonsAfte
rDiffusion(2,l)/(6.022*10^23*WellVolumeAvail)*1000*exp(-
q/(kB*Temp)*PSI_not/1000)+Cs_initial)^0.5))*exp(-t_inc(l)/tau_s);
        SignalMaintain(1,2) = max(Signal(2,:));
        Error(2,l) = Sigmas*Signal(2,l)/SNR(2,l);
        if l>1
            if Signal(2,l)<Signal(2,l-1)
                break
            end
        end
    else
        break
    end
end

for p = 1:(length(t_inc)-l)
    Signal(2,p+l-1) = SignalMaintain(1,2)*exp(-t_inc(p+1)/tau_s);
end

%Signal for 3 bases
for v = 1:length(t_inc)
    if TotalProtonsProduced(3,v) < 3*Strands
        Signal(3,v) = PSI_not*(1-
C_dl_initial/(dielectric*permittivity/(0.3/10^9)*0.5^0.5*Zs*(TotalProtonsAfte
rDiffusion(3,v)/(6.022*10^23*WellVolumeAvail)*1000*exp(-
q/(kB*Temp)*PSI_not/1000)+Cs_initial)^0.5))*exp(-t_inc(v)/tau_s);
        SignalMaintain(1,3) = max(Signal(3,:));
        Error(3,v) = Sigmas*Signal(3,v)/SNR(3,v);
        if v>1
            if Signal(3,v)<Signal(3,v-1)
                break
            end
        end
    else
        break
    end
```

```
end

for w = 1:(length(t_inc)-v)
    Signal(3,w+v-1) = SignalMaintain(1,3)*exp(-t_inc(w+1)/tau_s);
end

%% Plotting
figure
subplot(2,1,1);
tInt = 500;
hold on
plot(t_inc(1:tInt),TotalProtonsPostDiffusion(1,1:tInt),'b');
plot(t_inc(1:tInt),TotalProtonsPostDiffusion(2,1:tInt),'g');
plot(t_inc(1:tInt),TotalProtonsPostDiffusion(3,1:tInt),'r');
xlabel('Time,sec');ylabel('Total Protons in One Well after
Diffusion');legend('n = 1','n = 2', 'n = 3');
hold off
subplot(2,1,2);
hold on
plot(t_inc,Signal(1,:),'b');
plot(t_inc,Signal(2,:),'g');
plot(t_inc,Signal(3,:),'r');
xlabel('Time,sec');ylabel('Signal, mV');legend('n = 1','n = 2', 'n = 3');
hold off

figure
hold on
X_Axis = 250;
errorbar(t_inc(1:X_Axis),Signal(1,1:X_Axis),Error(1,1:X_Axis),'b');
errorbar(t_inc(1:X_Axis),Signal(2,1:X_Axis),Error(2,1:X_Axis),'g');
errorbar(t_inc(1:X_Axis),Signal(3,1:X_Axis),Error(3,1:X_Axis),'r');
xlabel('Time,sec');ylabel('Signal,mV');legend('n = 1', 'n = 2', 'n = 3');
hold off

%% Flows
WellPitch = WellDiameter + 0.22; %um
Gap = 1; %mm
DieWidth = 20; %mm
DieLength = 23.7; %mm
DieArea = DieWidth*DieLength; %mm
DieCrossSect = Gap*DieWidth; %mm^2
VolumeofChip = DieArea*Gap; %mm^3

NumberofWells = 660000000;

HydraulicDiameter = 4*Gap*DieWidth/(DieWidth*2+DieLength*2);

Viscosity = 0.001; %kg/(m-s)
Density = 1000; %kg/m^3
KVisc = Viscosity/Density; %m^2/s
FlowVelocity = (VolFlowRate/60)*1000/DieCrossSect; %mm/s
Re = FlowVelocity*HydraulicDiameter/(KVisc*1000^2);

TimeforNucleotideCover = VolumeofChip/VolFlowRate; %s
```

## APPENDIX F

## MATLAB Code for Dephasing

```
%% Set length of strand, no. of strands,and no. of cycles of nucleotide flow
% Also pick the bases you want included in the flow cycle
% 1 = A; 2 = C; 3 = G, T = 4; so each cycle = ACGT

Length = 200;
Strands = 100;
FlowCycles = 80;
F = [1 2 3 4];

% Set the time for the flow of one base and divide it into segments of time
dt
dt = 0.02; % should be less than 0.023
flowtime = 0.25; %s
timesegments = round(flowtime/dt);

%% Generate Random Sequence
% Create empty matrix that will contain the same sequence in each strand
% Column = strand; Row = Base position
% Fill up the matrix with each strand (bearing the same sequence)

Sequence = randseq(Length);
SequenceMatN = zeros(Length,Strands);

for a = 1:Strands
    SequenceMatN(:,a) = Sequence;
    SequenceMat = char(SequenceMatN);
end

%% Set sequence of nucleotides that will be flowed in

FlowSequenceN = repmat(F,1,FlowCycles);
FlowSequence = transpose(int2nt(FlowSequenceN));

%% Confirm base incorporation
ConfirmedBase = zeros(Length,Strands);

%%Nucleotide Concentration
dNTPconc = 100; %uM

%%Kinetic Data

%k_pol, s^-1

AtoA_kpol = 0.0036;
AtoC_kpol = 0.01; %0.1
AtoG_kpol = 0.042;
AtoT_kpol = 45;
```

```
CtoA_kpol = 0.01; %0.1
CtoC_kpol = 0.003;
CtoG_kpol = 43;
CtoT_kpol = 0.038;

GtoA_kpol = 0.05;
GtoC_kpol = 40; %31
GtoG_kpol = 0.066;
GtoT_kpol = 0.0116; %1.16

TtoA_kpol = 40; %25
TtoC_kpol = 0.012;
TtoG_kpol = 0.016; %1.6
TtoT_kpol = 0.013;

%KD, uM

AtoA_KD = 25;
AtoC_KD = 160;
AtoG_KD = 250;
AtoT_KD = 0.8;

CtoA_KD = 540;
CtoC_KD = 140;
CtoG_KD = 0.9;
CtoT_KD = 360;

GtoA_KD = 500;
GtoC_KD = 0.8;
GtoG_KD = 150;
GtoT_KD = 70;

TtoA_KD = 0.6;
TtoC_KD = 180;
TtoG_KD = 200;
TtoT_KD = 57;


%% Probabilities of incorporating
AonAProb = ((AtoA_kpol*dNTPconc)/(AtoA_KD+dNTPconc))*dt;
AonCProb = ((AtoC_kpol*dNTPconc)/(AtoC_KD+dNTPconc))*dt;
AonGProb = ((AtoG_kpol*dNTPconc)/(AtoG_KD+dNTPconc))*dt;
AonTProb = ((AtoT_kpol*dNTPconc)/(AtoT_KD+dNTPconc))*dt;
ConAProb = ((CtoA_kpol*dNTPconc)/(CtoA_KD+dNTPconc))*dt;
ConCProb = ((CtoC_kpol*dNTPconc)/(CtoC_KD+dNTPconc))*dt;
ConGProb = ((CtoG_kpol*dNTPconc)/(CtoG_KD+dNTPconc))*dt;
ConTProb = ((CtoT_kpol*dNTPconc)/(CtoT_KD+dNTPconc))*dt;
GonAProb = ((GtoA_kpol*dNTPconc)/(GtoA_KD+dNTPconc))*dt;
GonCProb = ((GtoC_kpol*dNTPconc)/(GtoC_KD+dNTPconc))*dt;
GonGProb = ((GtoG_kpol*dNTPconc)/(GtoG_KD+dNTPconc))*dt;
GonTProb = ((GtoT_kpol*dNTPconc)/(GtoT_KD+dNTPconc))*dt;
TonAProb = ((TtoA_kpol*dNTPconc)/(TtoA_KD+dNTPconc))*dt;
TonCProb = ((TtoC_kpol*dNTPconc)/(TtoC_KD+dNTPconc))*dt;
```

```
TonGProb = ((TtoG_kpol*dNTPconc)/(TtoG_KD+dNTPconc))*dt;
TonTProb = ((TtoT_kpol*dNTPconc)/(TtoT_KD+dNTPconc))*dt;

%% Dephase Counter
DephaseCounter = zeros(Length*Strands,6);
drow=1;
% col 1 = base flow no.
% col 2 = time segment
% col 3 = column no.
% col 4 = position no.
% col 5 = miss?
% col 6 = mismatch?

%% Sequencing
% For every match, place a 1 in the corresponding cell
% For every mismatch, place a -1 in the corresponding cell

for b=1:length(FlowSequence) % loop through every base flowed in
    for t=1:timesegments % loop through every time segment
        for c=1:Strands % loop through every column
            for d=1:Length % loop through every position (until finding the
first unfilled position)
                if ConfirmedBase(d,c) == 0;
                    r=d;
                    if     FlowSequence(b,1) == 'A'
                        if     SequenceMat(r,c) == 'A'
                               RN=rand;
                               if RN<=AonAProb;
                                  ConfirmedBase(r,c)=-1;
                                  DephaseCounter(drow,1)=b;
                                  DephaseCounter(drow,2)=t;
                                  DephaseCounter(drow,3)=c;
                                  DephaseCounter(drow,4)=d;
                                  DephaseCounter(drow,6)=1;
                                  drow=drow+1;
                               end
                        elseif SequenceMat(r,c) == 'C'
                               RN=rand;
                               if RN<=AonCProb;
                                  ConfirmedBase(r,c)=-1;
                                  DephaseCounter(drow,1)=b;
                                  DephaseCounter(drow,2)=t;
                                  DephaseCounter(drow,3)=c;
                                  DephaseCounter(drow,4)=d;
                                  DephaseCounter(drow,6)=1;
                                  drow=drow+1;
                               end
                        elseif SequenceMat(r,c) == 'G'
                               RN=rand;
                               if RN<=AonGProb;
                                  ConfirmedBase(r,c)=-1;
                                  DephaseCounter(drow,1)=b;
                                  DephaseCounter(drow,2)=t;
                                  DephaseCounter(drow,3)=c;
                                  DephaseCounter(drow,4)=d;
                                  DephaseCounter(drow,6)=1;
```

```
                drow=drow+1;
            end
        elseif SequenceMat(r,c) == 'T'
                RN=rand;
            if RN<=AonTProb;
                ConfirmedBase(r,c)=1;
            else
                DephaseCounter(drow,1)=b;
                DephaseCounter(drow,2)=t;
                DephaseCounter(drow,3)=c;
                DephaseCounter(drow,4)=d;
                DephaseCounter(drow,5)=1;
                drow=drow+1;
            end
        end
    elseif FlowSequence(b,1) == 'C'
        if      SequenceMat(r,c) == 'A'
                RN=rand;
            if RN<=ConAProb;
                ConfirmedBase(r,c)=-1;
                DephaseCounter(drow,1)=b;
                DephaseCounter(drow,2)=t;
                DephaseCounter(drow,3)=c;
                DephaseCounter(drow,4)=d;
                DephaseCounter(drow,6)=1;
                drow=drow+1;
            end
        elseif SequenceMat(r,c) == 'C'
                RN=rand;
            if RN<=ConCProb;
                ConfirmedBase(r,c)=-1;
                DephaseCounter(drow,1)=b;
                DephaseCounter(drow,2)=t;
                DephaseCounter(drow,3)=c;
                DephaseCounter(drow,4)=d;
                DephaseCounter(drow,6)=1;
                drow=drow+1;
            end
        elseif SequenceMat(r,c) == 'G'
                RN=rand;
            if RN<=ConGProb;
                ConfirmedBase(r,c)=1;
            else
                DephaseCounter(drow,1)=b;
                DephaseCounter(drow,2)=t;
                DephaseCounter(drow,3)=c;
                DephaseCounter(drow,4)=d;
                DephaseCounter(drow,5)=1;
                drow=drow+1;
            end
        elseif SequenceMat(r,c) == 'T'
                RN=rand;
            if RN<=ConTProb;
                ConfirmedBase(r,c)=-1;
                DephaseCounter(drow,1)=b;
                DephaseCounter(drow,2)=t;
```

```
                    DephaseCounter(drow,3)=c;
                    DephaseCounter(drow,4)=d;
                    DephaseCounter(drow,6)=1;
                    drow=drow+1;
                end
            end
        elseif FlowSequence(b,1) == 'G'
            if      SequenceMat(r,c) == 'A'
                    RN=rand;
                if RN<=GonAProb;
                    ConfirmedBase(r,c)=-1;
                    DephaseCounter(drow,1)=b;
                    DephaseCounter(drow,2)=t;
                    DephaseCounter(drow,3)=c;
                    DephaseCounter(drow,4)=d;
                    DephaseCounter(drow,6)=1;
                    drow=drow+1;
                end
            elseif SequenceMat(r,c) == 'C'
                    RN=rand;
                if RN<=GonCProb;
                    ConfirmedBase(r,c)=1;
                else
                    DephaseCounter(drow,1)=b;
                    DephaseCounter(drow,2)=t;
                    DephaseCounter(drow,3)=c;
                    DephaseCounter(drow,4)=d;
                    DephaseCounter(drow,5)=1;
                    drow=drow+1;
                end
            elseif SequenceMat(r,c) == 'G'
                    RN=rand;
                if RN<=GonGProb;
                    ConfirmedBase(r,c)=-1;
                    DephaseCounter(drow,1)=b;
                    DephaseCounter(drow,2)=t;
                    DephaseCounter(drow,3)=c;
                    DephaseCounter(drow,4)=d;
                    DephaseCounter(drow,6)=1;
                    drow=drow+1;
                end
            elseif SequenceMat(r,c) == 'T'
                    RN=rand;
                if RN<=GonTProb;
                    ConfirmedBase(r,c)=-1;
                    DephaseCounter(drow,1)=b;
                    DephaseCounter(drow,2)=t;
                    DephaseCounter(drow,3)=c;
                    DephaseCounter(drow,4)=d;
                    DephaseCounter(drow,6)=1;
                    drow=drow+1;
                end
            end
        elseif FlowSequence(b,1) == 'T'
            if      SequenceMat(r,c) == 'A'
                    RN=rand;
```

```
                                if RN<=TonAProb;
                                    ConfirmedBase(r,c)=1;
                                else
                                    DephaseCounter(drow,1)=b;
                                    DephaseCounter(drow,2)=t;
                                    DephaseCounter(drow,3)=c;
                                    DephaseCounter(drow,4)=d;
                                    DephaseCounter(drow,5)=1;
                                    drow=drow+1;
                                end
                            elseif SequenceMat(r,c) == 'C'
                                    RN=rand;
                                if RN<=TonCProb;
                                    ConfirmedBase(r,c)=-1;
                                    DephaseCounter(drow,1)=b;
                                    DephaseCounter(drow,2)=t;
                                    DephaseCounter(drow,3)=c;
                                    DephaseCounter(drow,4)=d;
                                    DephaseCounter(drow,6)=1;
                                    drow=drow+1;
                                end
                            elseif SequenceMat(r,c) == 'G'
                                    RN=rand;
                                if RN<=TonGProb;
                                    ConfirmedBase(r,c)=-1;
                                    DephaseCounter(drow,1)=b;
                                    DephaseCounter(drow,2)=t;
                                    DephaseCounter(drow,3)=c;
                                    DephaseCounter(drow,4)=d;
                                    DephaseCounter(drow,6)=1;
                                    drow=drow+1;
                                end
                            elseif SequenceMat(r,c) == 'T'
                                    RN=rand;
                                if RN<=TonTProb;
                                    ConfirmedBase(r,c)=-1;
                                    DephaseCounter(drow,1)=b;
                                    DephaseCounter(drow,2)=t;
                                    DephaseCounter(drow,3)=c;
                                    DephaseCounter(drow,4)=d;
                                    DephaseCounter(drow,6)=1;
                                    drow=drow+1;
                                end
                            end
                        end
                        if ConfirmedBase(d,c) == 0; % if no match/mismatch has
been made, move on, otherwise check the next position
                            break
                        end
                    end
                end
            end
        end
    end
end
% Details:
% loops through every base that is flowed in
```

```
% for every time segment, does the following
% loops through every column
% searches for the first open position
% determines whether there is a match
% accounts for homopolymers (i.e. checks to see if the position after a
% match is also a match)


%% Counting Mismatches

% MismatchCounter = list of all columns with mismatches extracted from
% DephaseCounter
% MismatchColandPos = same list as above (col#1) but also showing the
position of
% each mismatch (col#2)
LengthDC=length(DephaseCounter);
MismatchCounter = zeros(LengthDC,1);
mrow=1;
for mis=1:LengthDC
    if DephaseCounter(mis,6)==1
        MismatchCounter(mrow)=DephaseCounter(mis,3);
        mrow=mrow+1;
    end
end

MismatchCounter(MismatchCounter == 0) = []; % gets rid of all unnecessary
zeros in MismatchCounter

% StrandsAhead = all unique columns that appear in MismatchCounter
StrandsAhead = unique(MismatchCounter);

% StrandsAheadCounter = the number of times each element in StrandsAhead
% appears in MismatchCounter (i.e. by how many positions that column is
ahead)

% StrandsAheadTable = table showing all columns with mismatches (col#1) and
how many
% mismatches in each column (col#2)

% AheadNumbers = list of the unique numbers by which strands are ahead
% (i.e. some strands are ahead by 1, some strands are ahead by 2 and so on)

% AheadNumbersHist = how many times each of the elements in AheadNumbers
% appears (see next entry)

StrandsAheadTable(:,1)=StrandsAhead; % all col #s that are behind
if length(StrandsAhead) == 1 % if 1 or more mismatches in 1 column
   StrandsAheadTable(:,2) = length(MismatchCounter); % no. of mismatches =
length of MissesCounter
   AheadNumbers = length(MismatchCounter);
   AheadNumbersHist = 1;
elseif isempty(MismatchCounter) == 1 % if no mismatches
    StrandsAhadTable = zeros(1,2); % column with zeros
    StrandsAheadCounter = zeros(1,1);
```

```
    AheadNumbers = 0;
    AheadNumbersHist = hist(StrandsAheadCounter,AheadNumbers);
else
    StrandsAheadCounter = hist(MismatchCounter,StrandsAhead);
    StrandsAheadTable(:,2)=transpose(StrandsAheadCounter);
    AheadNumbers = unique(StrandsAheadCounter);
    AheadNumbersHist = hist(StrandsAheadCounter,AheadNumbers);
end

% AheadTable = table showing how many strands (col#2) are ahead by how much
% (col#1)
if isempty(AheadNumbersHist)==1
    AheadTable = zeros(1,2);
else
    AheadTable(:,1)=transpose(AheadNumbers);
    AheadTable(:,2)=transpose(AheadNumbersHist);
end

figure (1)
bar(AheadTable(:,1),AheadTable(:,2))
title('Plot showing strands that have dephased ahead')
xlabel('Position ahead of undephased strands')
ylabel('No. of strands')

%% Counting Misses

% MissesCounter = list of all columns with misses extracted from
% DephaseCounter (only those from the last time segment)
% MissesColandPos = same list as above (col#1) but also showing the position
of
% each miss (col#2)
LengthDC=length(DephaseCounter);
MissesCounter = zeros(LengthDC,1);
msrow=1;
for miss=1:LengthDC
    if DephaseCounter(miss,2)==timesegments
    if DephaseCounter(miss,5)==1
        MissesCounter(msrow)=DephaseCounter(miss,3);
        msrow=msrow+1;
    end
    end
end

MissesCounter(MissesCounter == 0) = []; % gets rid of all unnecessary zeros
in MismatchCounter

% StrandsBehind = all unique columns that appear in MissesCounter
StrandsBehind = unique(MissesCounter);

% StrandsBehindCounter = the number of times each element in StrandsBehind
% appears in MissesCounter (i.e. by how many positions that column is behind)

% StrandsBehindTable = table showing all columns with misses (col#1) and how
many
% misses in each column (col#2)
```

```
% BehindNumbers = list of the unique numbers by which strands are behind
% (i.e. some strands are behind by 1, some strands are behind by 2 and so on)

% BehindNumbersHist = how many times each of the elements in BehindNumbers
% appears (see next entry) (how many strands are behind by e.g. 1)

StrandsBehindTable(:,1)=StrandsBehind; % all col #s that are behind
if length(StrandsBehind) == 1 % if 1 or more misses in 1 column
   StrandsBehindTable(:,2) = length(MissesCounter); % no. of misses = length
of MissesCounter
   BehindNumbers = length(MissesCounter);
   BehindNumbersHist = 1;
elseif isempty(MissesCounter) == 1 % if no misses
    StrandsBehindTable = zeros(1,2); % column with zeros
    StrandsBehindCounter = zeros(1,1);
    BehindNumbers = 0;
    BehindNumbersHist = hist(StrandsBehindCounter,BehindNumbers);
else
    StrandsBehindCounter = hist(MissesCounter,StrandsBehind);
    StrandsBehindTable(:,2)=transpose(StrandsBehindCounter);
    BehindNumbers = unique(StrandsBehindCounter);
    BehindNumbersHist = hist(StrandsBehindCounter,BehindNumbers);
end

% BehindTable = table showing how many strands (col#2) are behind by how much
% (col#1)
if isempty(BehindNumbersHist)==1
    BehindTable = zeros(1,2);
else
    BehindTable(:,1)=transpose(BehindNumbers);
    BehindTable(:,2)=transpose(BehindNumbersHist);
end

figure (2)
bar(BehindTable(:,1),BehindTable(:,2))
title('Plot showing strands that have dephased behind')
xlabel('Position behind of undephased strands')
ylabel('No. of strands')

%% Counting All Dephases

% DephasedStrands = how many are ahead (#1) and behind (#2)
DephasedStrands(1,1) = length(StrandsAhead);
DephasedStrands(1,2) = length(StrandsBehind);

% CommonLength = longest column of DephasedStrands - giving the future
% columns enough space
CommonLength = max(DephasedStrands);

% CommonStrands = list of strands that are both ahead and behind
% NotCommonStrandsA = list of strands that are only ahead
% NotCommonStrandsB = list of strands that are only behind
% col 1 = strand no.
% col 2 = how much ahead
```

```
% col 3 = how much behind
% col 4 = overall dephase (+ve = ahead; -ve = behind)

CommonStrands = zeros(CommonLength,4);
csrow=1;
NotCommonStrandsA = zeros(CommonLength,4);
ncsrow1=1;
NotCommonStrandsB = zeros(CommonLength,4);
ncsrow2=1;

for x = 1:length(StrandsAhead)
    if ismember(StrandsAhead(x),StrandsBehind)==1
        CommonStrands(csrow,1)=StrandsAhead(x);
        csrow=csrow+1;
    else
        NotCommonStrandsA(ncsrow1,1)=StrandsAhead(x);
        ncsrow1=ncsrow1+1;
    end
end

for y = 1:length(StrandsBehind)
    if ismember(StrandsBehind(y),StrandsAhead)==0
        NotCommonStrandsB(ncsrow2,1)=StrandsBehind(y);
        ncsrow2=ncsrow2+1;
    end
end

for xx = 1:CommonLength
    for yy = 1:size(StrandsAheadTable,1)
        if CommonStrands(xx,1) == StrandsAheadTable(yy,1)
            CommonStrands(xx,2) = StrandsAheadTable(yy,2);
        end
        if NotCommonStrandsA(xx,1) == StrandsAheadTable(yy,1)
            NotCommonStrandsA(xx,2) = StrandsAheadTable(yy,2);
        end
    end
    for zz = 1:size(StrandsBehindTable,1)
        if CommonStrands(xx,1) == StrandsBehindTable(zz,1)
            CommonStrands(xx,3) = StrandsBehindTable(zz,2);
        end
        if NotCommonStrandsB(xx,1) == StrandsBehindTable(zz,1)
            NotCommonStrandsB(xx,3) = StrandsBehindTable(zz,2);
        end
    end
    CommonStrands(xx,4) = CommonStrands(xx,2)-CommonStrands(xx,3);
    NotCommonStrandsA(xx,4) = NotCommonStrandsA(xx,2)-
NotCommonStrandsA(xx,3);
    NotCommonStrandsB(xx,4) = NotCommonStrandsB(xx,2)-
NotCommonStrandsB(xx,3);
end

% CommonStrandsTable = (#1) strand no. (#2) how much ahead (#3) how much
% behind (#4) how much dephased - only for strands both ahead and behind
CommonStrandsLength = nnz(CommonStrands(:,1));
CommonStrandsTable = CommonStrands(1:CommonStrandsLength,:);
```

```
% NotCommonStrandsATable = (#1) strand no. (#2) how much ahead (#3) how much
% behind (#4) how much dephased - only for strands ahead
NotCommonStrandsALength = nnz(NotCommonStrandsA(:,1));
NotCommonStrandsATable = NotCommonStrandsA(1:NotCommonStrandsALength,:);

% NotCommonStrandsBTable = (#1) strand no. (#2) how much ahead (#3) how much
% behind (#4) how much dephased only for strands behind
NotCommonStrandsBLength = nnz(NotCommonStrandsB(:,1));
NotCommonStrandsBTable = NotCommonStrandsB(1:NotCommonStrandsBLength,:);

% DephasedStrandsBTable = (#1) strand no. (#2) how much ahead (#3) how much
% behind (#4) how much dephased - for all strands
DephasedStrandsTable =
[CommonStrandsTable;NotCommonStrandsATable;NotCommonStrandsBTable];

% DephasedTable = (#1) how much dephased (#2) how many strands

DephasedList = sort(DephasedStrandsTable(:,4));
DephasedListUnique = unique(DephasedList);
DephasedListHist = hist(DephasedList,DephasedListUnique);

DephasedTable(:,1) = DephasedListUnique;
DephasedTable(:,2) = transpose(DephasedListHist);

TotalStrandsAhead = nnz(StrandsAhead)
TotalStrandsBehind = nnz(StrandsBehind)
TotalStrandsDephased = length(DephasedStrandsTable)
TotalStrandsPerfect = Strands-TotalStrandsDephased

for xxx=1:size(DephasedTable,1)
    if DephasedTable(xxx,1) == 0
        DephasedTable(xxx,2) = DephasedTable(xxx,2) + TotalStrandsPerfect;
    end
end

if ismember(0,DephasedTable) == 0
        DephasedTable = [DephasedTable;[0 TotalStrandsPerfect]];
end

figure (3)
bar(DephasedTable(:,1),DephasedTable(:,2))
title('Plot showing strands that have dephased')
xlabel('Position relative to strands in phase')
ylabel('No. of strands')
```

## APPENDIX G

## MATLAB Code for Base-Calling

```
%% Set length of strand, # of strands,and # of cycles of nucleotide flow
% Also pick the bases you want included in the flow cycle
% 1 = A; 2 = C; 3 = G, T = 4; so each cycle = ACGT

Length = 200;
Strands = 50;
FlowCycles = 75;
F = [4 3 2 1];

% Set the time for the flow of one base, divide it into segments of time dt
dt = 0.02; % should be less than 0.023
flowtime = 0.1;
timesegments = round(flowtime/dt);

%% Generate Random Sequence
% Create empty matrix that will contain the same sequence in each strand
% Column = strand; Row = Base position
% Fill up the matrix with each strand (bearing the same sequence)

Sequence = randseq(Length);
SequenceMatN = zeros(Length,Strands);

for a = 1:Strands
    SequenceMatN(:,a) = Sequence;
    SequenceMat = char(SequenceMatN);
end

%% Set sequence of nucleotides that will be flowed in

FlowSequenceN = repmat(F,1,FlowCycles);
FlowSequence = transpose(int2nt(FlowSequenceN));

%% Confirm base incorporation
ConfirmedBase = zeros(Length,Strands);

%%Nucleotide Concentration
dNTPconc = 100; %uM

%%Kinetic Data
%k_pol, s^-1

AtoA_kpol = 0.0036;
AtoC_kpol = 0.01; %0.1
AtoG_kpol = 0.042;
AtoT_kpol = 45; %45

CtoA_kpol = 0.01; %0.1
```

```
CtoC_kpol = 0.003;
CtoG_kpol = 43; %43
CtoT_kpol = 0.038;

GtoA_kpol = 0.05;
GtoC_kpol = 40; %37
GtoG_kpol = 0.066;
GtoT_kpol = 0.016; %1.16

TtoA_kpol = 40; %25
TtoC_kpol = 0.012;
TtoG_kpol = 0.016; %1.6
TtoT_kpol = 0.013;

%KD, uM

AtoA_KD = 25;
AtoC_KD = 160;
AtoG_KD = 250;
AtoT_KD = 0.8;

CtoA_KD = 540;
CtoC_KD = 140;
CtoG_KD = 0.9;
CtoT_KD = 360;

GtoA_KD = 500;
GtoC_KD = 0.8;
GtoG_KD = 150;
GtoT_KD = 70;

TtoA_KD = 0.6;
TtoC_KD = 180;
TtoG_KD = 200;
TtoT_KD = 57;


%% Probabilities of incorporating
AonAProb = ((AtoA_kpol*dNTPconc)/(AtoA_KD+dNTPconc))*dt;
AonCProb = ((AtoC_kpol*dNTPconc)/(AtoC_KD+dNTPconc))*dt;
AonGProb = ((AtoG_kpol*dNTPconc)/(AtoG_KD+dNTPconc))*dt;
AonTProb = ((AtoT_kpol*dNTPconc)/(AtoT_KD+dNTPconc))*dt;
ConAProb = ((CtoA_kpol*dNTPconc)/(CtoA_KD+dNTPconc))*dt;
ConCProb = ((CtoC_kpol*dNTPconc)/(CtoC_KD+dNTPconc))*dt;
ConGProb = ((CtoG_kpol*dNTPconc)/(CtoG_KD+dNTPconc))*dt;
ConTProb = ((CtoT_kpol*dNTPconc)/(CtoT_KD+dNTPconc))*dt;
GonAProb = ((GtoA_kpol*dNTPconc)/(GtoA_KD+dNTPconc))*dt;
GonCProb = ((GtoC_kpol*dNTPconc)/(GtoC_KD+dNTPconc))*dt;
GonGProb = ((GtoG_kpol*dNTPconc)/(GtoG_KD+dNTPconc))*dt;
GonTProb = ((GtoT_kpol*dNTPconc)/(GtoT_KD+dNTPconc))*dt;
TonAProb = ((TtoA_kpol*dNTPconc)/(TtoA_KD+dNTPconc))*dt;
TonCProb = ((TtoC_kpol*dNTPconc)/(TtoC_KD+dNTPconc))*dt;
TonGProb = ((TtoG_kpol*dNTPconc)/(TtoG_KD+dNTPconc))*dt;
TonTProb = ((TtoT_kpol*dNTPconc)/(TtoT_KD+dNTPconc))*dt;
```

```
%% Sequencing
% For every match, place a 1 in the corresponding cell
% For every mismatch, place a -1 in the corresponding cell

Counter = zeros(length(FlowSequence),timesegments);

% Details:
% loops through every base that is flowed in
% for every time segment, does the following
% loops through every column
% searches for the first open position
% determines whether there is a match
% accounts for homopolymers (i.e. checks to see if the position after a
% match is also a match)

for b=1:length(FlowSequence) % loop through every base flowed in
   time=0;
    for t=1:timesegments % loop through every time segment
        for c=1:Strands % loop through every column
            for d=1:Length % loop through positions (until first unfilled)
                if ConfirmedBase(d,c) == 0;
                    r=d;
                    if     FlowSequence(b,1) == 'A'
                        if     SequenceMat(r,c) == 'A'
                                  RN=rand;
                               if RN<=AonAProb;
                                  ConfirmedBase(r,c)=-1;
                                  Counter(b,t) = Counter(b,t) + 1;
                               end
                        elseif SequenceMat(r,c) == 'C'
                                  RN=rand;
                               if RN<=AonCProb;
                                  ConfirmedBase(r,c)=-1;
                                  Counter(b,t) = Counter(b,t) + 1;
                               end
                        elseif SequenceMat(r,c) == 'G'
                                  RN=rand;
                               if RN<=AonGProb;
                                  ConfirmedBase(r,c)=-1;
                                  Counter(b,t) = Counter(b,t) + 1;
                               end
                        elseif SequenceMat(r,c) == 'T'
                                  RN=rand;
                               if RN<=AonTProb;
                                  ConfirmedBase(r,c)=1;
                                  Counter(b,t) = Counter(b,t) + 1;
                               end
                        end
                    elseif FlowSequence(b,1) == 'C'
                        if     SequenceMat(r,c) == 'A'
                                  RN=rand;
                               if RN<=ConAProb;
                                  ConfirmedBase(r,c)=-1;
                                  Counter(b,t) = Counter(b,t) + 1;
                               end
```

```
        elseif SequenceMat(r,c) == 'C'
                   RN=rand;
               if RN<=ConCProb;
                   ConfirmedBase(r,c)=-1;
                   Counter(b,t) = Counter(b,t) + 1;
               end
        elseif SequenceMat(r,c) == 'G'
                   RN=rand;
               if RN<=ConGProb;
                   ConfirmedBase(r,c)=1;
                   Counter(b,t) = Counter(b,t) + 1;
               end
        elseif SequenceMat(r,c) == 'T'
                   RN=rand;
               if RN<=ConTProb;
                   ConfirmedBase(r,c)=-1;
                   Counter(b,t) = Counter(b,t) + 1;
               end
        end
    elseif FlowSequence(b,1) == 'G'
        if     SequenceMat(r,c) == 'A'
                   RN=rand;
               if RN<=GonAProb;
                   ConfirmedBase(r,c)=-1;
                   Counter(b,t) = Counter(b,t) + 1;
               end
        elseif SequenceMat(r,c) == 'C'
                   RN=rand;
               if RN<=GonCProb;
                   ConfirmedBase(r,c)=1;
                   Counter(b,t) = Counter(b,t) + 1;
               end
        elseif SequenceMat(r,c) == 'G'
                   RN=rand;
               if RN<=GonGProb;
                   ConfirmedBase(r,c)=-1;
                   Counter(b,t) = Counter(b,t) + 1;
               end
        elseif SequenceMat(r,c) == 'T'
                   RN=rand;
               if RN<=GonTProb;
                   ConfirmedBase(r,c)=-1;
                   Counter(b,t) = Counter(b,t) + 1;
               end
        end
    elseif FlowSequence(b,1) == 'T'
        if     SequenceMat(r,c) == 'A'
                   RN=rand;
               if RN<=TonAProb;
                   ConfirmedBase(r,c)=1;
                   Counter(b,t) = Counter(b,t) + 1;
               end
        elseif SequenceMat(r,c) == 'C'
                   RN=rand;
               if RN<=TonCProb;
                   ConfirmedBase(r,c)=-1;
```

```
                                Counter(b,t) = Counter(b,t) + 1;
                            end
                        elseif SequenceMat(r,c) == 'G'
                                RN=rand;
                            if RN<=TonGProb;
                                ConfirmedBase(r,c)=-1;
                                Counter(b,t) = Counter(b,t) + 1;
                            end
                        elseif SequenceMat(r,c) == 'T'
                                RN=rand;
                            if RN<=TonTProb;
                                ConfirmedBase(r,c)=-1;
                                Counter(b,t) = Counter(b,t) + 1;
                            end
                        end
                    end
                    if ConfirmedBase(d,c) == 0;
                    % if no match/mismatch has been made, move on,
                    % otherwise check the next position
                        break
                    end
                end
            end
        end

    end

end

CounterSum_PreRounding_Insertions = sum(Counter,2);
CounterSum = round(round(4*sum(Counter,2)./Strands)/4);
%rounds to nearest 0.25, then rounds to nearest integer
Standard = sum(CounterSum);
%The proper length of the strands if no dephasing

NumberofDephased = 0;
NumberofDephased_Behind = 0;
NumberofDephased_Ahead = 0;

for y = 1:Strands
    if nnz(ConfirmedBase(:,y)) ~= Standard
        NumberofDephased = NumberofDephased + 1;
    end
    if nnz(ConfirmedBase(:,y)) < Standard
        NumberofDephased_Behind = NumberofDephased_Behind + 1;
    end
    if nnz(ConfirmedBase(:,y)) > Standard
        NumberofDephased_Ahead = NumberofDephased_Ahead + 1;
    end
end
PercentDephased = NumberofDephased/Strands;
PercentDephased_Behind = NumberofDephased_Behind/Strands;
PercentDephased_Ahead = NumberofDephased_Ahead/Strands;

subplot(2,1,1)
```

```
cc = 60;
hold on
bar(CounterSum_PreRounding_Insertions(1:cc));
xlabel('Base Flow Number');ylabel('Total Insertion Events');
hold off
subplot(2,1,2)
hold on
bar(CounterSum(1:cc));xlabel('Base Flow Number');ylabel('Bases in Strand');
hold off

%Error rates
ErrorRates = zeros(1,Strands); %Gives percent error, dephased not counted
for s=1:Strands
    Errors = tabulate(ConfirmedBase(:,s));
    if Errors(1,1) == -1
        ErrorRates(1,s) = Errors(1,3);
    end
end
AverageError = mean(ErrorRates);

%Derive Sequence from Base Calling Algorithm
GeneratedSequence = zeros(Length,1);
gsrow = 1;

for zz = 1:length(CounterSum)
    GSRow = CounterSum (zz,1);
    GeneratedSequence(gsrow:(gsrow+GSRow),1) = char(FlowSequence(zz));
    gsrow = gsrow + GSRow;
end

for yy = 1:length(GeneratedSequence)
    if GeneratedSequence(yy) == 'A'
        GeneratedSequence(yy) = 'T';
    elseif GeneratedSequence(yy) == 'T'
        GeneratedSequence(yy) = 'A';
    elseif GeneratedSequence(yy) == 'G'
        GeneratedSequence(yy) = 'C';
    elseif GeneratedSequence(yy) == 'C'
        GeneratedSequence(yy) = 'G';
    end
end

DerivedSequence = char(GeneratedSequence);

%Any errors from the given sequence and derived sequence
MatchError = zeros(length(DerivedSequence),1);
for xx = 1:length(DerivedSequence)
    if DerivedSequence(xx) ~= SequenceMat(xx)
        MatchError(xx,1) = 1;
    end
end
TotalMatchError = sum(MatchError);
```

## APPENDIX H

# Components of Costs of Goods Sold

| Equipment | Price |
|---|---|
| - PII™ Chip | $350.00 |
| - Ion Plus Fragment Library Kit | $25.00 |
| - Ion PI™ Template OT2 200 Kit | $83.00 |
| - Ion PI™ Sequencing Kit | $67.50 |
| - Ion Proton Controls Kit | $100.00 |
| - 10 M NaOH [$71.40 for 100 mL] | $1.40 |
| - Isopropanol (99.7%) [$265 for 20 kg] | $0.34 |
| - Nuclease-free Water [$96.70 for 5 L] | $0.19 |
| - 10 L of 1N HCl [$91.20 for 10 L] | $0.09 |
| - Ethanol (200 proof) [$315 for 6-500 mL bottles] | $1.05 |
| - 500 mL TE Buffer, 1X Solution pH 8.0, low EDTA | $0.63 |
| - Pipette Tips | |
| P2 | $0.27 |
| P20 | $0.27 |
| P200 | $0.27 |
| P1000 | $0.32 |
| - Thin Wall PCR Tubes, Flat Cap | $5.00 |
| - 1.5 mL microcentrifuge tubes | $0.20 |
| - Agencourt AMPure XP - PCR Purification | $2.11 |
| - Agilent® High Sensitivity DNA Kit | $4.61 |
| - MagaZorb DNA Common Kit-200 | $2.00 |
| - Ion Xpress™ Barcode Adapters 1-96 Kit (partial purchase) | $1.56 |
| - Bioruptor® NGS 0.65 ml Microtubes for DNA Shearing (500 tubes) | $0.34 |
| - Pippin Prep™ Kit 2010 | $4.50 |

## APPENDIX I

# Components of Capital Equipment

| Equipment | Unit Price |
|---|---|
| - Ion Proton II, including Ion Server | $224,000.00 |
| - Maxwell Research System | $30,000.00 |
| - Ion OneTouch 2 System | $19,000.00 |
| - Nitrogen (grade 4.8, 99.998% or better) | $70.00 |
| - Water Purification System (Elga Purelab Flex 3) | $5,000.00 |
| - Multistage gas regulator (VWR, 55850-422) | $375.00 |
| - Lab Freezer | $1,000.00 |
| - Uninterruptable Power Supply (UPS) | $200.00 |
| - Microcentrifuge | $1,995.00 |
| - Galaxy Mini Centrifuge | $401.25 |
| - Pipettes | |
| P2 | $335.00 |
| P20 | $297.00 |
| P200 | $297.00 |
| P1000 | $297.00 |
| - 1 L Glass Bottles | $9.40 |
| - Vortex Mixer | $800.00 |
| - Thermal Cycler | $8,000.00 |
| - Tygon Tubing | $2.00 |
| - Magnetic Stirrer | $230.00 |
| - Magnetic Stir Bars | 10 |
| - Vacuum filtration system (pore size 0.45 um) | $83.40 |
| - Orion 3-Star Plus Benchtop Meter Kit with probes | $752 |
| - Squirt bottles | $5.00 |
| - 50 mL Syringe | $1.85 |
| - DynaMag™-2 Magnet | $531 |
| - Agilent® 2100 Bioanalyzer® instrument | $19580 |
| - Heat Block/Water Bath | $160 |
| - Incubator | $183 |
| - BioRuptor® NGS Sonication System | $13000 |
| - Pippin Prep™ System | $15000 |

## APPENDIX J

Pro Forma Case 1

| Year | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|
| **Income Statement** | | | | |
| *Revenue* | *$0* | *$10,000,000* | *$15,000,000* | *$20,000,000* |
| *Cost of Sales* | *($1,612,500)* | *($3,225,000)* | *($4,837,500)* | *($6,450,000)* |
| *Gross Profit* | *($1,612,500)* | *$6,775,000* | *$10,162,500* | *$13,550,000* |
| Operating, SG&A Expenses | | | | |
| Sales & Marketing | ($85,000) | ($150,000) | ($150,000) | ($150,000) |
| Research & Development | ($370,000) | ($1,180,000) | ($1,330,000) | ($1,480,000) |
| General and Administration | ($249,600) | ($674,204) | ($674,204) | ($674,204) |
| Depreciation of Operating Assets | $0 | ($956,854) | ($1,530,967) | ($918,580) |
| *Total Operating Expenses* | *($704,600)* | *($2,961,058)* | *($3,685,171)* | *($3,222,784)* |
| | | | | |
| **Pre-tax Income** | ($2,317,100) | $3,813,942 | $6,477,329 | $10,327,216 |
| Tax @ 40% (negative income carried over) | | ($598,737) | ($2,590,932) | ($4,130,886) |
| | | | | |
| **Net Income** | ($2,317,100) | $3,215,205 | $3,886,397 | $6,196,329 |
| **Gross Margin** | 0.00% | 67.75% | 67.75% | 67.75% |
| **Operating Margin** | 0.00% | 38.14% | 43.18% | 51.64% |
| **Net Profit Margin** | 0.00% | 32.15% | 25.91% | 30.98% |
| | | | | |
| **Cash Flow Statement** | | | | |
| **Operating Activities** | | | | |
| Net Earnings | ($2,317,100) | $3,215,205 | $3,886,397 | $6,196,329 |
| Depreciation | $0 | $956,854 | $1,530,967 | $918,580 |
| | | | | |
| **Working Capital Estimates** | | | | |
| Accounts Receivable | $0 | $821,918 | $1,232,877 | $1,643,836 |
| Inventory | $0 | $191,781 | $287,671 | $383,562 |
| Accounts Payable | $0 | $265,068 | $397,603 | $530,137 |
| Cash Reserve | $0 | $243,375 | $302,891 | $264,886 |
| Working Capital Changes | | | | |
| Accounts Receivable | 0 | ($821,918) | ($410,959) | ($410,959) |
| Inventories | 0 | ($191,781) | ($95,890) | ($95,890) |
| Accounts Payable | $0 | $265,068 | $132,534 | $132,534 |
| Cash Reserve | $0 | $243,375 | $59,516 | ($38,004) |
| | | | | |
| **Total Working Capital** (only include 2015) | $0 | $992,005 | $1,425,836 | $1,762,147 |
| | | | | |
| **Investing Activities** | | | | |
| Property, Plant & Equipment, COGS/Ops. | ($1,265,786) | ($3,518,486) | $0 | $0 |
| **Financing Activities** | | | | |
| Issuance of Common Stock | $3,682,886 | $4,510,491 | $0 | $0 |
| | | | | |
| **Free Cash Flow** | $100,000 | $4,307,210 | $9,619,443 | $17,577,919 |
| **Current Ratio** | N/A | 4.74 | 4.59 | 4.32 |
| **Quick Ratio** | N/A | 4.02 | 3.86 | 3.60 |

| | | | |
|---|---|---|---|
| **Terminal Value** | $70,311,677 | | |
| Final Year CF | $17,577,919 | Growth Rate | 0% |
| Discount Rate (A) | 50% | | |
| Discount Rate (B) | 25% | | |

| | *2014* | *2015* | *2016* | *2017* | *Terminal Value* |
|---|---|---|---|---|---|
| *Year* | *1* | *2* | *3* | *4* | *5* |
| Free Cash Flow | $100,000 | $4,307,210 | $9,619,443 | $17,577,919 | $70,311,677 |
| Present Value | $66,667 | $2,756,614 | $4,925,155 | $7,199,916 | $23,039,730 |
| | | | | | |
| Investments | $3,682,886 | $4,510,491 | | | |
| Present Value | $3,682,886 | $3,608,393 | | | |
| | | | | **NPV** | **$30,696,803** |
| | Investment | Disc. Invest. | % Share | | |
| **Investor 1** | $3,682,886 | $5,524,329 | 55% | | |
| **Investor 2** | $4,510,491 | $4,510,491 | 45% | | |
| | Sum: | $10,034,820 | | | |

**Equity in Company**

| NPV | 1 | 2 | 3 | 4 | Terminal Value 5 |
|---|---|---|---|---|---|
| | ($3,616,219) | ($4,467,998) | $457,157 | $7,657,073 | $30,696,803 |
| | | | | | |
| Founders | 10% | 7% | 7% | 7% | 7% |
| Investor 1 | 90% | 61% | 61% | 61% | 61% |
| Investor 2 | 0% | 32% | 32% | 32% | 32% |
| | | | | | |
| Founders | $0 | $0 | $31,087 | $520,681 | $2,087,383 |
| Investor 1 | $0 | $0 | $279,780 | $4,686,129 | $18,786,444 |
| Investor 2 | $0 | $0 | $146,290 | $2,450,263 | $9,822,977 |

**MIRR**

| | |
|---|---|
| Reinvestment Rate | 0.38% |
| Finance Rate | 2.76% |

| | | | | | Terminal Value | |
|---|---|---|---|---|---|---|
| Year | 1 | 2 | 3 | 4 | 5 | **NPV** |
| Cash Flow | $100,000 | $4,307,210 | $9,619,443 | $17,577,919 | $70,311,677 | |
| Investor 1 | $90,000.00 | $2,636,012 | $5,887,099 | $10,757,687 | $43,030,746 | -$384,700.08 |
| Investor 2 | | $1,378,307 | $3,078,222 | $5,624,934 | $22,499,737 | $27,584,544.83 |

| **Present Value CF** | | |
|---|---|---|
| Investor 1 | $3,682,886 | |
| Investor 2 | | $4,510,491 |

| **Future Value CF** | | | | | | Sum |
|---|---|---|---|---|---|---|
| Investor 1 | $91,376 | $2,666,177 | $5,931,926 | $10,798,566 | $43,030,746 | $62,518,791 |
| Investor 2 | | $1,394,080 | $3,101,661 | $5,646,309 | $22,499,737 | $32,641,786 |

| | |
|---|---|
| **MIRR Investor 1** | **102.98%** |
| **MIRR Investor 2** | **93.43%** |

| | |
|---|---|
| IRR Investor 1 | 167.58% |
| IRR Investor 2 | 0.60% |

## APPENDIX K

Pro Forma Case 2

| Year | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|
| **Income Statement** | | | | |
| *Revenue* | ***$0*** | ***$10,000,000*** | ***$20,000,000*** | ***$20,000,000*** |
| *Cost of Sales* | ***($1,612,500)*** | ***($3,225,000)*** | ***($6,450,000)*** | ***($6,450,000)*** |
| *Gross Profit* | ***($1,612,500)*** | ***$6,775,000*** | ***$13,550,000*** | ***$13,550,000*** |
| Operating, SG&A Expenses | | | | |
| Sales & Marketing | ($85,000) | ($150,000) | ($150,000) | ($150,000) |
| Research & Development | ($370,000) | ($1,180,000) | ($1,480,000) | ($1,480,000) |
| General and Administration | ($249,600) | ($674,204) | ($674,204) | ($674,204) |
| Depreciation of Operating Assets | $0 | ($552,834) | ($884,534) | ($530,720) |
| *Total Operating Expenses* | ***($704,600)*** | ***($2,557,038)*** | ***($3,188,738)*** | ***($2,834,924)*** |
| **Pre-tax Income** | ($2,317,100) | $4,217,962 | $10,361,262 | $10,715,076 |
| Tax @ 40% (negative income carried over) | | ($760,345) | ($4,144,505) | ($4,286,030) |
| | | | | |
| **Net Income** | ($2,317,100) | $3,457,617 | $6,216,757 | $6,429,045 |
| **Gross Margin** | 0.00% | 67.75% | 67.75% | 67.75% |
| **Operating Margin** | 0.00% | 42.18% | 51.81% | 53.58% |
| **Net Profit Margin** | 0.00% | 34.58% | 31.08% | 32.15% |
| | | | | |
| **Cash Flow Statement** | | | | |
| **Operating Activities** | | | | |
| Net Earnings | ($2,317,100) | $3,457,617 | $6,216,757 | $6,429,045 |
| Depreciation | $0 | $552,834 | $884,534 | $530,720 |
| | | | | |
| **Working Capital Estimates** | | | | |
| Accounts Receivable | $0 | $821,918 | $1,643,836 | $1,643,836 |
| Inventory | $0 | $191,781 | $383,562 | $383,562 |
| Accounts Payable | $0 | $265,068 | $530,137 | $530,137 |
| Cash Reserve | $0 | $210,167 | $262,088 | $233,007 |
| Working Capital Changes | | | | |
| (Increase)/Decrease Accounts Receivable | 0 | ($821,918) | ($821,918) | $0 |
| (Increase)/Decrease Inventories | 0 | ($191,781) | ($191,781) | $0 |
| Increase/(Decrease) Accounts Payable | $0 | $265,068 | $265,068 | $0 |
| Increase/(Decrease) Cash Reserve | $0 | $210,167 | $51,921 | ($29,081) |
| | | | | |
| **Total Working Capital** (only include 2015) | $0 | **$958,798** | $1,759,348 | $1,730,268 |
| | | | | |
| **Investing Activities** | | | | |
| Property, Plant & Equipment, COGS/Ops | ($816,584) | ($1,947,584) | $0 | $0 |
| **Financing Activities** | | | | |
| Issuance of Common Stock | $3,233,684 | $2,906,382 | | |
| | | | | |
| **Free Cash Flow** | $100,000 | $4,516,415 | $12,492,521 | $20,651,834 |
| **Current Ratio** | N/A | 4.62 | 4.32 | 4.26 |
| **Quick Ratio** | N/A | 3.89 | 3.60 | 3.54 |

| | | | |
|---|---|---|---|
| **Terminal Value** | $82,607,336 | | |
| Final Year CF | $20,651,834 | Growth Rate | 0% |
| Discount Rate (A) | 50% | | |
| Discount Rate (B) | 25% | | |

| | 2014 | 2015 | 2016 | 2017 | Terminal Value |
|---|---|---|---|---|---|
| Year | 1 | 2 | 3 | 4 | 5 |
| Free Cash Flow | $100,000 | $4,516,415 | $12,492,521 | $20,651,834 | $82,607,336 |
| Present Value | $66,667 | $2,890,506 | $6,396,171 | $8,458,991 | $27,068,772 |
| | | | | | |
| Investments | $3,233,684 | $2,906,382 | | | |
| Present Value | $3,233,684 | $2,325,105 | | | |
| | | | | **NPV** | **$39,322,317** |
| | Investment | Disc. Invest. | % Share | | |
| **Investor 1** | $3,233,684 | $4,850,526 | 63% | | |
| **Investor 2** | $2,906,382 | $2,906,382 | 37% | | |
| | Sum: | $7,756,908 | | | |

| **Equity in Company** | | | | | Terminal Value |
|---|---|---|---|---|---|
| NPV | 1 | 2 | 3 | 4 | 5 |
| | ($3,167,017) | ($2,601,617) | $3,794,554 | $12,253,545 | $39,322,317 |
| | | | | | |
| Founders | 10% | 8% | 8% | 8% | 8% |
| Investor 1 | 90% | 72% | 72% | 72% | 72% |
| Investor 2 | 0% | 20% | 20% | 20% | 20% |
| | | | | | |
| Founders | $0 | $0 | $303,564 | $980,284 | $3,145,785 |
| Investor 1 | $0 | $0 | $2,732,079 | $8,822,552 | $28,312,068 |
| Investor 2 | $0 | $0 | $758,911 | $2,450,709 | $7,864,463 |

| **MIRR** | |
|---|---|
| Reinvestment Rate | 0.38% |
| Finance Rate | 2.76% |

| | | | | | Terminal Value | |
|---|---|---|---|---|---|---|
| Year | 1 | 2 | 3 | 4 | 5 | NPV |
| Cash Flow | $100,000 | $4,516,415 | $12,492,521 | $20,651,834 | $82,607,336 | |
| Investor 1 | $90,000.00 | $3,251,819 | $8,994,615 | $14,869,320 | $59,477,282 | $0.00 |
| Investor 2 | | $903,283 | $2,498,504 | $4,130,367 | $16,521,467 | $0.00 |

| **PV of (-) CF** | | |
|---|---|---|
| Investor 1 | $3,233,684 | |
| Investor 2 | | $2,906,382 |

| **FV of (+) CF** | | | | | | Sum |
|---|---|---|---|---|---|---|
| Investor 1 | $91,376 | $3,289,031 | $9,063,104 | $14,925,824 | $59,477,282 | $86,846,616 |
| Investor 2 | | $913,620 | $2,517,529 | $4,146,062 | $16,521,467 | $24,098,678 |

| | |
|---|---|
| **MIRR Investor 1** | **127.65%** |
| **MIRR Investor 2** | **102.40%** |

| | |
|---|---|
| IRR Investor 1 | 209.31% |
| IRR Investor 2 | 192.04% |

## APPENDIX L

# Original Problem Statement

**11. A Moore's Law for DNA Sequencing: $1,000 Genomes using Ion Torrent Technology (recommended by John C. Crocker, U. Penn)**

The first human genome was published in 2003, and was the result of over $3 billion of public funding for the Human Genome Project (HGP). Around the same time, a privately funded company, Celera Genomics, using superior technology, published its own genome for just one tenth the cost: $300 million. The content of a single human genome has immense utility as a research tool for understanding the molecular origin of disease. Currently, many researchers are focused on even a greater opportunity and technical challenge – personalized medicine. If the specific genome of an *individual* is known, then it can be used to predict their future predilection for different diseases, or to tailor more effective life-saving therapies for them; e.g., for cancer.

One impediment to personalized medicine is the current high cost of genotyping: you can have your complete genome sequenced today commercially [1], but it comes with a price-tag of >$100,000. To stimulate further progress, in 2006 the X Prize Foundation announced the *Archon X Prize for Genomics* [2], which will award $10 million to the first team to sequence 100 different human genomes, for less than $10,000 apiece, in less than 10 days, with an error rate below ten per million bases. Several firms have been working toward the challenge for several years, including *454* Life Sciences, Pacific Biosciences, Helicos Biosciences [3] and GnuBio [4]. Despite dueling press releases and considerable startup funding, the prize remains unclaimed. In the summer of 2011, a new player, Ion Torrent [5], entered the scene with a CMOS chip that could sequence millions of short fragments of DNA, and it is now preparing a claim to the prize [6].

While these whole genome technologies are exciting, a larger impediment looms to threaten the idea of personalized medicine—*a complete lack of clinical data*. In total, only a few hundred genetic polymorphisms that interact with clinical treatments have been discovered to date, and sequencing just those polymorphisms costs a negligible amount. The most basic assumption of personal medicine is that if we had a large enough data set of whole patient genomes, we could discover millions of interactions between polymorphisms and clinical outcomes for different treatments using simple correlational data-mining. The central question is who will construct that dataset? The most plausible candidate is the large pharmaceutical companies themselves. Today, dosage decisions and safety contraindications for pharmaceuticals are based on large data sets formed during clinical trials, at a cost of ~$10,000 per participant. If the cost of producing a whole genome were reduced to ~$1,000, then complete patient genotyping could become a standard practice during all clinical trials; essentially piggybacking the construction of a personal medicine database on the existing infrastructure.

This project is to design a '$1,000 genome' process within the context of a small startup company using the Ion Torrent Technology [7]. The team's business model will be a service company, receiving large numbers of patient samples (e.g., cheek swabs) from a pharma company client, and the electronic delivery of the corresponding whole genome data to the client within 30 days of sample receipt. Product sequences should have an error rate of less than ten parts per million for non-repeating intronic sequences.

The core (Ion Torrent) technology consists of a CMOS chip covered in millions of microscopic wells, each with $H^+$ ion sensing transistors at its base, covered with a simple fluid handling layer. Patient DNA is fragmented and individual fragments are clonally amplified onto microspheres, which are loaded onto the chip, with one sphere per well, see Figure 1. During the sequencing process, DNA polymerase enzymes are loaded onto the DNA strands to be sequenced (roughly 1 million copies per bead/well)), but are stalled due to the lack of available nucleotide dNTPs. When a solution of a single dNTP is added, all the polymerases waiting for that nucleotide (i.e., stalled at a location having a complementary base on their template) will incorporate the base and advance, releasing a single $H^+$ per enzyme. The synchronous release of these ions causes a transient change in the pH in the well, which is detected by the transistor as a voltage pulse, and transduced to a host computer. After the bases have all been added, the $H^+$ and unreacted nucleotides are washed away (termed a 'flow' cycle) and one of the other 4 nucleotides is washed over the chip. After 400 such flow/wash cycles, strands of up to 100 bases can be sequenced. These fragments of the genome are then reassembled using statistical techniques on a central server.

An emphasis will be placed on process design to optimize sequence throughput as well as sample preparation and clonal amplification, relevant biochemical reactions/kinetics, microfluidic fluid mechanics and chip layout and detector signal to noise, as all limit and determine ultimate process throughput and cost. All sequencing approaches will need to be validated by bioinformatic reconstruction of mock data, most likely using the bioinformatics toolkit in Matlab. Creative solutions by the team to improve the base technology are welcome, but the as published technology will be costed for comparison purposes. Some of the raw data will contain errors due to finite signal to noise in the ion measurements, homopolymer insertions, etc. The team will need to verify that these errors can be weeded out robustly to allow a final error rate of less then 10 parts per million in the final product sequence.

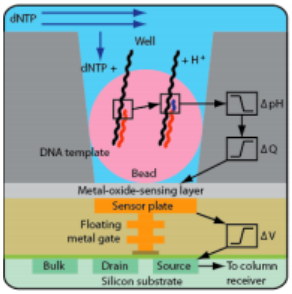



**Figure 1:** Shows a single microbead covered with multiple DNA template strands, in a pH detecting microwell on a CMOS

Since winning an X prize is not an acceptable business model, the team will evaluate the viability of a small venture capital funded biotech start-up, serving large pharma companies as their clients. The team will assume a Series A funded period in which the technology is demonstrated at a scale comparable to the X prize rate – 10 genomes/day * 250 days/year = 2,500 genomes/yr. The team will determine their market price per genome based upon their own incremental and capital costs, as well as those estimated for their two closest competitors, but this is expected to be in the $100-300 range. After Series B funding, the assumed throughput with be quadrupled to 10,000 genomes per year. The financial analysis should seek a significant, positive NPV over a total four year time horizon with an appropriate IRR for a biotech startup with VC funding. Acceleration of genotyping volume can be considered if market analysis suggests adequate demand. If the overall financial analysis looks favorable, the team should also estimate the capital requirements to expand their operational throughput to a plausible ultimate demand (after the widespread adoption of personalized medicine) of $10^6$ genomes/year.

References

1. http://www.knome.com/
2. http://genomics.xprize.org
3. http://www.helicosbio.com/
4. http://gnubio.com/
5. http://www.iontorrent.com/
6. http://bits.blogs.nytimes.com/2012/07/23/cheaper-computer-power-leading-to-sequencing-genome/
7. Rothberg, J. M., et al., *Nature*, **475**, 348-352, 2011.

# 16. REFERENCES

"3-D, 22nm: New Technology Delivers An Unprecedented Combination of Performance and Power Efficiency." Intel, Web. <http://www.intel.com/content/www/us/en/silicon-innovations/intel-22nm-technology.html>.

Barbaro, Massimo, et al. "Fully electronic DNA hybridization detection by a standard CMOS chip." *Sensors and Actuators B* 118 (2006): 41-46.

Berman, Karen, and Joe Knight. *Financial Intelligence for Entrepreneurs*. Boston: Harvard Business, 2008. Print.

Bird, C. (2012, May 1). Next-gen sequencing services: An expanding role in clinical applications opens new markets. *Genetic Engineering & Biotechnology News*, *32*(9), Retrieved from http://www.genengnews.com/gen-articles/next-gen-sequencing-services/4088/

Bustillo, J., W. Hinz, K.L. Johnson, J. Leamon, J.M. Rothberg, and J. Schultz. Sequencing nucleic acid comprises disposing template nucleic acids into reaction chambers in contact with or capacitively coupled to chemical-sensitive field effect transistor. Ion Torrent Systems, assignee. Patent GB2561128-A; GB2461128-B. 15 Dec. 2010. Print.

Chan, Eugene Y. "Next-Generation Sequencing Methods: Impact of Sequencing Accuracy on SNP Discovery." DNA Medicine Institute. Web. 29 Mar. 2013.

Chiang, Jung-Lung. *Study on the pH-Sensing Characteristics of ISFET with Aluminum Nitride Membrane*. Diss. 2002.)

Chiang J, Chou J, Chen Y; Sensitivity and hysteresis properties of a-wo3,ta2o5, and a-si:h gate ion-sensitive field-effect transistors. Opt. Eng. 0001;41(8):2032-2038.

Chiang, Jung-Lung, Jung-Chuan Chou, and Ying-Chung Chen. "Sensitivity and hysteresis properties of a-WO3, Ta2O5, and a-Si: H gate ion-sensitive field-effect transistors." *Optical Engineering* 41.8 (2002): 2032-2038.

Companies and Markets. (2011). *Strategic analysis of the u.s. next generation sequencing markets*. Frost and SUllivan. Retrieved from http://www.companiesandmarkets.com/Market/Healthcare-and-Medical/Market-Research/Strategic-Analysis-of-the-U-S-Next-Generation-Sequencing-Markets/RPT915231

Complete Genomics. (2011). introduction to complete genomics' sequencing technology. In *Complete Genomics Media*. Moutain View, CA: Complete Genomics. Retrieved from http://media.completegenomics.com/documents/Technology White Paper.pdf

Croft, K. (2012, November 29). Gene by gene launches dna dtc: Offers highly reliable, cost-effective dna testing to institutional clients worldwide. *Market Watch: The Wall Street Journal*. Retrieved from http://www.marketwatch.com/story/gene-by-gene-launches-dna-dtc-2012-11-29

Damodaran, Aswath. "Closure in Valuation." NYU Stern, n.d. Web. 31 Mar. 2013.

Decisive Bio-Insights. (2013). Next generation sequencing: Market size, segmentation, growth and trends by provider. (2nd ed.). Culver City, CA: DeciBio, LLC.

Duncan, D. (2011, September 23). A dna tower of babel: As more and more people's genomes are decoded, we need better ways to share and understand the data. *MIT Technology Review*, Retrieved from http://www.technologyreview.com/news/425521/a-dna-tower-of-babel/

Dutta, J. C. "Modeling Ion Sensitive Field Effect Transistors for Biosensor Applications." *International Journal of Advanced Research in Engineering and Technology*. (2010): 38-57.

Esteban, Jose A., Margarita Salas, and Luis Blanco. "Fidelity of Phi29 DNA Polymerase."*The Journal of Biological Chemistry* 268.4 (1993): 2719-726.

Esfandyarpour, Hesaam, Bo Zheng, R. Fabian W. Pease, and Ronald W. Davis. "Structural Optimization for Heat Detection of DNA Thermosequencing Platform Using Finite Element Analysis." *Biomicrofluidics* 2.2 (2008): 024102.

Express Scripts. (2012). Archon genomics x prize competition guidelines. In New York, New York: Retrieved from http://genomics.xprize.org/sites/genomics.xprize.org/files/docs/AGXP_Competition_Guidelines.pdf

Gibson, G., & Muse, S. (2009). *A primer of genome scienc*. (3 ed.). Sunderland, Massachusetts, USA: Sinauer Associates, Inc. Publishers.

Glenn, Travis C. "Field Guide to Next-generation DNA Sequencers." *Molecular Ecology Resources* (2011).

"HIPAA, the Privacy Rule, and Its Application to Health Research." NCBI, Web. <http://www.ncbi.nlm.nih.gov/books/NBK9573/>.

Ion Torrent. "The Ion Proton System: Rapid genome-scale benchtop sequencing. Specification Sheet." 2012. <www.lifetechnologies.com/proton>.

Lek, Monkol. "Challenges in Improving Ion Torrent Raw Accuracy." *BioLektures*. 30 Aug. 2011. Web. 16 Mar. 2013. <http://biolektures.wordpress.com/2011/08/22/challenges-in-improving-ion-torrent-raw-accuracy-part-3/>.

Lem, C. S. (2009). Magtration System 12GC: Application data - DNA from Saliva. PSS Bio Instruments technical bulletin (101305), 2

Life Technologies. (2012). Ion xpress plus gdna fragment library preparation. In Life Technologies.

Liu, L Yinhu Li, Siliang Li, et al. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, *2012*, Retrieved from http://www.hindawi.com/journals/bmri/2012/251364/

Jan, Shiun-Sheng, et al. "Preparation and properties of lead titanate gate ion-sensitive field-effect transistors by the sol-gel method." *Japanese journal of applied physics* 41.2A (2002): 942-948.

Johnson, Allison A., and Kenneth A. Johnson. "Fidelity of Nucleotide Incorporation by Human Mitochondrial DNA Polymerase." *The Journal of Biological Chemistry* 276.41 (2001): 38090-8096.

Jung-Chuan Chou, Lan Pin Liao. *Study on pH at the point of zero charge of TiO2 pH ion-sensitive field effect transistor made by the sputtering method*. Thin Solid Films 476.1. (2005): 157-161.

Liao, Hung-Kwei, et al. "Study on $pH_{pzc}$ and surface potential of tin oxide gate ISFET." *Materials chemistry and physics* 59.1 (1999): 6-11.

Mardis ER. Next-Generation DNA Sequencing Methods. Annual Review of Genomics and Human Genetics. 2008. 9: 387-402.

Margulies, M., Egholm, M., & Altman, W. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*, 376-280.

Mendelsohn, M. L., et al. Department of Energy Office of Energy Research Office of Health and Environmental Research, Subcommittee on Human Genome of the Health and Environmental Research Advisory Committee (1987). *Report on the human genome initiative office of health and environmental research: Report on the human genome initiative office of health and environmental research*. Retrieved from website: http://www.ornl.gov/sci/techresources/Human_Genome/project/herac2.shtml

Merriman, B., Ion Torrent R&D Team, B., & Rothberg, J. (2012). Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis*, *33*, 3397-3417.

"Modified Internal Rate Of Return - MIRR." Investopedia, Web. <http://www.investopedia.com/terms/m/mirr.asp>.

Morrow, J., & Higgs, B. (2012). Callsim: Evaluation of base calls using sequencing simulation. . *ISRN Bioinformatics*, *2012*, 10 pages. doi: 10.5402/2012/371718

Mulhern, J. (2013, February 18). Ion torrent edges illumina in sales battle of benchtop sequencers, says macquarie report. *Bio-IT WOrld*, Retrieved from http://www.bio-itworld.com/news/02/18/13/Ion-Torrent-edges-Illumina-sales-benchtop-sequencers-Macquarie.html

National Academy of Sciences. Division on Earth & Life Studies, Board on Life Sciences. (2013). *Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease* . Retrieved from National Academy of Sciences website: http://dels.nas.edu/Report/Toward-Precision-Medicine-Building-Knowledge/13284

Natishan, P. M., E. McCafferty, and G. K. Hubler. "Surface Charge Considerations in the Pitting of Ion-Implanted Aluminum." *Journal of The Electrochemical Society* 135 (1988): 321

Niedringhaus TP, Milanova D, Kerby MB, et al. Landscape of Next-Generation Sequencing Technologies. Analytical Chemistry. 2011, 83, 4327–4341

Ng, S. B., Turner, E. H., & et al, (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, *461*, 272-276. doi: 10.1038/nature08250

Preočanin, T., & Kallay, N. (2006). Point of zero charge and surface charge density of TiO2 in aqueous electrolyte solution as obtained by potentiometric mass titration. *Croatica chemica acta*, *79*(1), 95-106.

Purushothaman, Sunil, Chris Toumazou, and Chung-Pei Ou. "Protons and Single Nucleotide Polymorphism Detection: A Simple Use for the Ion Sensitive Field Effect Transistor."*Sensors and Actuators B: Chemical* 114 (2006): 964-68.

Quail MA, Smith M, Coupland P, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 2012, 13:341

R.E.G. van Hal, J.C.T. Eijkel, P. Bergveld, A novel description of ISFET sensitivity with the buffer capacity and double-layer capacitance as key parameters, Sensors and Actuators B: Chemical, Volume 24, Issues 1–3, (1995): 201-205.

Rothberg, Jonathan M, et al. "An Integrated Semiconductor Device Enabling Non-optical Genome Sequencing." *Nature* 475 (2011): 348-52. Print.

Rothberg, Jonathan M., James M. Bustillo, Mark J. Milgrew, Jonathan C. Schultz, David Marran, Todd M. Rearick, and Kim L. Johnson. Methods and Apparatus for Measuring Analytes. Life Technologies Corporation, assignee. Patent 8263336. 11 Sept. 2012. Print.

Shendure, J., & Ji, H. (2008). Next generation dna sequencing. *Nature Biotechnology*, *26*(10), 1135-1145.

Son, Anna. "IBISWorld Industry Report 62151: Diagnostic & Medical Laboratories in the US." *IBISWorld*. Dec. 2012. Web. <www.ibisworld.com>.

*The pharmacogenomics knowledgebase*. (n.d.). Retrieved from http://www.pharmgkb.org/.

Tinoco, Ignacio, Leonard S. Lerman, George Cahill, et al. Health and Environmental Research Advisory Committee (HERAC). "Report on the Human Genome Initiative Office of Health and Environmental Research:Prepared for Dr. Alvin W. Trivelpiece Director, Office of Energy Research." . http://www.ornl.gov/sci/techresources/Human_Genome/project/herac2.shtml (accessed April 1, 2013).

U.S. Department of the Treasury, (2013). *Daily treasury long term rate data*. Retrieved from website: http://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/TextView.aspx?data=longtermrate

Wetterstrand, K. (2013, Feburary 11). *dna sequencing costs: Data from the nhgri genome sequencing program (gsp)* . Retrieved from http://www.genome.gov/sequencingcosts/

Williams R, et al. 2006. Amplification of complex gene libraries by emulsion PCR. *Nature Methods | Protocol*. 3(7):545-550

Woias, P., L. Meixner, D. Amandi, and M. Schönberger. "Modelling the Short-time Response of ISFET Sensors." *Sensors and Actuators B: Chemical* 24.1-3 (1995): 211-17. Print.