



4-2010

THE SYNTHSEQ APPROACH TO PERSONAL GENOTYPING

Stephanie M. Amato
University of Pennsylvania

Alan S. Futran
University of Pennsylvania

Kevin S. Krebs
University of Pennsylvania

Brian N. Recchione
University of Pennsylvania

Follow this and additional works at: http://repository.upenn.edu/cbe_sdr

Amato, Stephanie M.; Futran, Alan S.; Krebs, Kevin S.; and Recchione, Brian N., "THE SYNTHSEQ APPROACH TO PERSONAL GENOTYPING" (2010). *Senior Design Reports (CBE)*. 20.
http://repository.upenn.edu/cbe_sdr/20

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/cbe_sdr/20
For more information, please contact libraryrepository@pobox.upenn.edu.

THE SYNTHSEQ APPROACH TO PERSONAL GENOTYPING

Abstract

Inspired by the Archon X Prize for Genomics, our research project involves implementing a novel strategy for sequencing the human genome. This prize worth \$10 million will be awarded to the first company to sequence 100 human genomes with 99.999% accuracy in less than 10 days for under \$10,000 each. However, the possibility of winning the X Prize is secondary to the prospect of revolutionizing medical diagnostics. Currently, the genomic state-of-the-art involves identifying SNPs (single nucleotide polymorphisms) that are correlated to certain diseases. Compared to such existing diagnostics, the genome-wide, sequence-based biomarkers that will be made possible by fast and affordable human genome sequencing are staggering.

After six months of thorough investigation and development, we are pleased to present SynthSeq, a cutting-edge, whole-genome sequencing venture based on the novel sequencing-by-synthesis technology. In contrast to high-priced competitors, our inexpensive and comparatively error-free whole genome sequencing solution will prove to be an invaluable diagnostic resource, and it will only become more valuable as advances are made in the field of molecular diagnostics. At our intended retail price of \$5,000, series A investors can expect a worst-case MIRR of 22%, and the ultimate NPV should be no less than \$700 thousand. We are confident that our innovative SynthSeq technology will deliver high-fidelity, low-cost whole genome sequencing to as many as 3,000 customers per year as currently envisioned, with the potential for scale-up to millions.

2010

STEPHANIE M. AMATO

ALAN S. FUTRAN

KEVIN S. KREBS

BRIAN N. RECCHIONE

***THE SYNTHSEQ APPROACH TO
PERSONAL GENOTYPING***

FACULTY ADVISOR: DR. JOHN C. CROCKER

PROFESSOR LEONARD A. FABIANO

APRIL 13, 2010

TABLE OF CONTENTS

<i>ABSTRACT</i>	5
<i>1. INTRODUCTION</i>	7
Purpose for Genetic Screening.....	7
The <i>SynthSeq</i> Approach.....	13
Technology Readiness Assessment.....	15
Customer Requirements.....	18
Conclusions.....	20
<i>2. NEXT GENERATION SEQUENCING TECHNOLOGIES</i>	23
Market Analysis.....	24
Sequencing by Synthesis.....	24
<i>SynthSeq</i> Technology – Asynchronous, Single Molecule Sequencing by Synthesis.....	25
Competitive Analysis.....	28
Illumina.....	28
Roche/454 Sequencing.....	29
Pacific Biosciences.....	31
Knome.....	32
Pricing Analysis.....	34
Conclusions.....	35
<i>3. OPTIMIZING SYNTHSEQ'S THROUGHPUT</i>	37
EMCCD Size of Pixels and Dimensions of Pixel Array.....	38
Viewing Multiple Imaging Areas.....	39
Single Nucleotide Addition and Detection.....	40
Read Length.....	41
Optical Detection Efficiency.....	42
The Coverage Factor.....	46
Throughput of Optimization.....	49

Conclusions.....	51
4. PRE-SEQUENCING PREPARATIONS.....	53
DNA Extraction, Purification, and Preparation.....	54
Layer-by-Layer Assembly and Template Annealing.....	56
Sequencing Reaction Flow Cell	58
Sequencing Station Setup.....	60
Conclusions.....	61
5. CHEMISTRY AND DETECTION.....	63
Initial Imaging of Template Position.....	64
Nucleotide Addition.....	64
DNA synthesis.....	64
Phi 29 Polymerase	66
Probability of Polymerase Binding and Remains on Template.....	67
Reversibly Terminating Tagged Nucleotides	70
Polymerization Rate	72
Probability of Nucleotide Incorporation.....	75
Reaction Conditions.....	76
Optical Detection of Single Molecules.....	77
The Scanner Problem	77
Total Internal Reflection Microscopy.....	78
EMCCD camera.....	81
Dichroic Beam-Splitter for Fluorophore Resolution.....	84
Probability of Properly Identifying Nucleotides.....	86
Overview of Imaging Cycle.....	93
Fluorophore Cleave	95
Terminator Removal and 3'-OH Regeneration.....	95
Conclusions.....	96
6. GENOME ASSEMBLY.....	97
De Novo Sequencing vs. Comparative Genome Assembly.....	98
Data Collection and Processing.....	99
Reassembly of the Reads.....	100
Bioinformatics Hardware Requirements.....	100

Aligner and Computational Demands.....	101
Product Delivery	103
Conclusions.....	104
7. FINANCIAL ANALYSIS	105
Market and Revenue Projection.....	106
Costs, PPE, and Depreciation.....	108
Income Statement.....	115
Working Capital.....	116
Free Cash Flow, Terminal Value.....	119
NPV Valuation.....	121
Equity Shares.....	122
MIRR Analysis.....	124
What-If Scenarios.....	127
Price Sensitivity Analysis.....	129
Conclusions.....	133
8. CONCLUSIONS.....	135
Conclusions.....	135
Acknowledgements.....	138
APPENDIX A: REAGENT SPECIFICATIONS.....	139
APPENDIX B: QIAGEN PROTOCOLS.....	145
APPENDIX C: EQUIPMENT SPECIFICATIONS.....	151
APPENDIX D: EFFICIENCY SIMULATION CODE.....	165
APPENDIX E: FINANCIAL PRO FORMA.....	171
REFERENCES.....	185

ABSTRACT

Inspired by the *Archon X Prize for Genomics*, our research project involves implementing a novel strategy for sequencing the human genome. This prize worth \$10 million will be awarded to the first company to sequence 100 human genomes with 99.999% accuracy in less than 10 days for under \$10,000 each. However, the possibility of winning the X Prize is secondary to the prospect of revolutionizing medical diagnostics. Currently, the genomic state-of-the-art involves identifying SNPs (single nucleotide polymorphisms) that are correlated to certain diseases. Compared to such existing diagnostics, the genome-wide, sequence-based biomarkers that will be made possible by fast and affordable human genome sequencing are staggering.

After six months of thorough investigation and development, we are pleased to present *SynthSeq*, a cutting-edge, whole-genome sequencing venture based on the novel sequencing-by-synthesis technology. In contrast to high-priced competitors, our inexpensive and comparatively error-free whole genome sequencing solution will prove to be an invaluable diagnostic resource, and it will only become more valuable as advances are made in the field of molecular diagnostics. At our intended retail price of \$5,000, series A investors can expect a worst-case MIRR of 22%, and the ultimate NPV should be no less than \$700 thousand. We are confident that our innovative *SynthSeq* technology will deliver high-fidelity, low-cost whole genome sequencing to as many as 3,000 customers per year as currently envisioned, with the potential for scale-up to millions.

1. INTRODUCTION

PURPOSE OF GENETIC SEQUENCING

The first successful sequencing of the human genome began in 1990 with the Human Genome Project. It took until 2003 for the results of the project – the full sequence of the human genome – to be released. The project cost approximately \$3 billion.¹ Since then, Sanger Sequencing, the method used by the Human Genome Project, has been replaced by newer, better sequencing methods. These “next generation” techniques have reduced the sequencing time of a full human genome from the decade-long scale to the week-long scale and have moved closer to

making the “thousand dollar genome” a reality. The prospect of fast, cheap access to any individual’s entire genetic sequence has opened the door for exciting new possibilities in promoting human health.

One of the biggest goals of the Human Genome Project, and of genomic research in the years since the first successful sequencing, has been to link physical conditions, including many diseases, to their genetic roots. This could provide doctors and scientists with the opportunity to diagnose and treat patients for specific ailments that are identified as risks based on that individual’s genetic code. As access to genetic information becomes easier to obtain, and science gains an understanding of the genetic causes of many conditions, a new approach to the diagnosis and treatment of disease is gaining prevalence – personalized medicine. In the near future, access to such information could mean that certain revealing sequences within each patient’s individual genome will be used to authorize the use of specialized medical treatments. While affordable entire genome sequencing and the full implications of personalized medicine are not yet a reality, medicine has already incorporated genetic screening into many facets of patient care.

While the average consumer does not have access to a service that will give them their entire genome at a reasonable cost, there are services that will sequence small parts of the human genome containing small nucleotide polymorphisms (SNPs). SNPs are the most common genetic variations in the human genome. Disease genetics studies aim to link specific SNPs to increased risk for disease, thus allowing services to focus on specific locations in the genome that represent potential markers.² About 3,669 DNA variants associated with diseases and traits have been identified; screening specifically for these variants could provide insight into patient risk for disease without the need for whole-genome sequencing. Companies such as 23andMe, Navigenics and deCode Genetics provide services that will test for genetic variations and give information on susceptibility to some diseases. While sequencing these segments of the genome provides some

insight into the risk factors, this approach is handicapped by our limited knowledge of which genes contribute to disease and their precise mechanisms of action.³

Perhaps a more tangible effect of genome sequencing in current medicine comes in the area of pharmacogenomics, a field that investigates the effects of genetic variations on the body's response to drugs. All people respond to drugs differently, and many drug responses appear to be linked to genetics. Finding the links between genetic polymorphisms and differences in drug metabolizing enzymes and drug transporters allows for the use of genetic tests to predict individuals' responses to different treatments.⁴ Pharmacogenomics could lead to drugs that are prescribed for and administered to each patient depending on his or her genetic makeup. This predictive genomic information could help optimize drug selection and dosing, while avoiding adverse responses by offering advance knowledge of which patients would benefit from a particular drug and how that drug should be administered.⁵

A recent example of the increased prevalence and importance of pharmacogenomics comes in the marketing and administration of the anticoagulant Warfarin. Warfarin is a widely prescribed drug that helps prevent blood clots and heart attacks, but dosing the drug for individual patients is difficult. If a patient is given too high a dose, Warfarin can cause massive bleeding in the brain; in 2007, the drug was the second leading cause of emergency room admissions related to adverse drug response. In 2007, information on genetic testing was incorporated into the drug's product label, and the FDA authorized the marketing of a genetic test associated with the drug. Prior to the approval of additional information on Warfarin's label and the associated genetic test, the only way to attempt to prevent such reactions was trial and error – patients would be given a dose of the drug, which would be adjusted depending on its observed effects. However, data suggesting that patients with certain variations within the CYP2C9 and VKORC1 regions of the genome require lower doses of Warfarin (and are therefore more prone to adverse effects when given the usual dose) led to the use of genetic tests to fine-tune dosage of the drug in high-risk patients.⁶

The increased utility of genetic screening for drug therapies has prompted the FDA to begin thinking about how to approach drugs manufactured for particular genetic groups. Clinical trials are currently designed to test therapies on large and diverse groups of patients to identify an average response that gauges both safety and efficacy. Pharmacogenomics seeks to bring this type of analysis down to a sample size of one. The FDA has begun to add recommendations for tests, as evidenced by the Warfarin case; however, there is still no standard approach for approving a diagnostic test to be used in concert with a given therapy. The FDA has indicated that it is in the process of developing guidelines to tackle this issue. In the meantime, pharmacogenomics will continue to strengthen its foothold in the pharmaceutical industry. There are already a wide range of drugs on the market with associated pharmacogenomic tests, some of which are summarized in Table 1.1 below.⁷

Drug	Purpose	Gene Factor
Atomoxetine HCl (Strattera)	ADHD treatment	Patients with a mutation in the CYP2D6 gene are at risk of suffering serious liver damage.
Clopidogrel (Plavix)	Heart attack prevention	A variation of the CYP2C19 gene interferes with the way the drug is metabolized, rendering it ineffective.
Cetuximab (Erbix) Panitumumab (Vectibix)	Colorectal cancer drug	The drugs indicated only in patients whose tumors express a normal KRAS gene.
Gefitinib (Iressa)	Lung cancer drug	Indicated primarily in the treatment of patients whose tumors have a mutation in the EGFR gene.
Irinotecan (Camptosar)	Colorectal cancer drug	People with a genetic variant suffer side effects because they have fewer liver enzymes to break down the drug.
Tamoxifen (Nolvadex)	Breast cancer drug	Variations in the CYP2D6 gene can make a person metabolize the drug too quickly or not at all.
Warfarin (Coumadin)	Blood thinner	In certain patients, the drug can cause excessive bleeding. Genetic testing can reveal the right dose.

Table 1.1 Drugs on the market with associated pharmacogenomic tests

Genetic screening's most significant impact in the field of medicine has arguably been in cancer treatment. One application of genetic analysis in cancer treatment is evaluating risk of developing certain cancers. Mutations in specific genes have been implicated in many kinds of cancer. For example, mutations in BRCA1 and BRCA2 are positively correlated with a woman's risk of developing breast and ovarian cancer.⁸ Genetic screening – either whole genome sequencing or,

more commonly at this point, SNP analysis to identify specific mutations – could be used to gauge a patient’s risk and take an appropriate action to prevent cancer or catch it in its early stages.

Another application of genetic screening in oncology has been to develop specialized therapies to fight specific tumors. Historically, cancer treatment has been dominated by broadly acting cytotoxins that attack all fast-growing cells, in an approach known as chemotherapy. However, in recent decades the molecular basis of many cancers has been studied and elucidated. Specific sets of genetic defects lead to various types of cancer in many different parts of the body, and two patients with seemingly similar cancers often have dissimilar underlying molecular causes. Thus, many modern therapies are targeted to cancers with very particular genetic mutations. As a result, genetic screening has become increasingly important in oncology. An example of applying genetic screening to discover and apply treatment options is in the tyrosine kinase inhibitor, Imatinib, and its use in fighting Chronic Myeloid Leukemia. The drug is effective against cancer whose specific cause is the fusion of the *BCR* and *ABL* genes, and it represents a powerful tool for patients fighting this cancer.⁹ This novel, individualized approach to treatment requires the discovery and development of tests for biological indicators that help doctors determine how to treat each individual patient. As screening methods become increasingly fast and affordable, genetic screening is sure to assume an even bigger role in cancer treatment.¹⁰

A very recent example of the advances in genetic sequencing with implications for cancer treatment has come in the identification of personalized tumor biomarkers. Using advanced genome sequencing, researchers have reported that they are able to detect mutant DNA in patient plasma isolated from blood samples, and that the genetic rearrangements in this DNA proved to be identical to those found in tumor samples from the same patients. This allows for the detection of chromosomal rearrangement – which has been long recognized as a universal feature of cancer – in a way that was previously impossible. Whereas conventional approaches to genetic sequencing in cancer treatment have looked for single-letter mutations, this “personalized analysis of rearranged

ends” (PARE) approach identifies entire rearrangements in DNA sequences. This has been made possible by the ability to perform massive, parallel, and near-complete sequencing of individual tumor genomes. Identifying the specific mutations behind a patient’s cancer provides valuable information needed to provide appropriate treatment. While the cost of this sequencing is still too high to be widely applied, advancements in cost and throughput of genome sequencing could make it possible to bring this tool into hospitals in the near future.¹¹

While science and medicine have made great strides toward the realization of personalized medicine, challenges persist. Although genetic screening has become increasingly fast and affordable and promises to soon become widely accessible, the great difficulty in analyzing the data it provides has become apparent. Little is known about the links between genes and disease, and even those links that have been found generally account for a small percentage of the overall risk of disease.⁷ A recent study comparing the results of two DNA screening companies, Navigenics and 23andME, which predicted risk of thirteen diseases in five individuals, found that for seven of the diseases, 50% or fewer of the predictions made by the two companies agreed.¹²

There is still work to be done on both fronts of the pursuit of personalized medicine – the accessibility of genetic data and the understanding of the genetic causes of disease. However, it is conceivable that as the price of sequencing the genome decreases and large amounts of genomic data become available, correlational analyses of genetic sequences and diseases will become easier to conduct and will provide more accurate forecasts. One can imagine a day when pharmaceutical companies sequence the entire genome of each patient in phase III trials in order to identify genetic causes of different reactions to a drug. With the average cost per patient hovering around \$26,000 and the cost of sequencing the human genome nearing \$1000, this seems fully plausible¹³. It seems clear that advancements in genetic sequencing are poised to transform medicine in the near future.

THE SYNTHSEQ APPROACH

The *SynthSeq* approach strives to deliver superior, accurate whole genome sequencing to its customers. The main goal of the project is to develop a process that will sequence human genomes at exceptional throughput and low cost. The process has been designed to meet the demand of 3000 genomes per year. The incremental operating costs per genome must be \$10,000 or less apiece, and the start-up capital for the process cannot exceed \$25 million. These parameters were inspired by the *Archon X Prize for Genomics*, which calls for the sequencing of 100 different human genomes, for less than \$10,000 each, in less than 10 days, with an error rate below ten per million

Project Name	The <i>SynthSeq</i> Approach Toward Personal Genotyping
Project Champion	Dr. John Crocker
Project Leaders	Stephanie Amato, Alan Futran, Kevin Krebs, Brian Recchione
Specific Goals	Sequence 3000 human genomes per year for less than \$10,000 per sequenced genome, with a start-up cost of \$25 million or less.
Project Scope	<p>In-Scope:</p> <ul style="list-style-type: none"> • Identify and investigate high throughput screening techniques • Provide in-house whole genome sequencing through application of the most promising technology • Characterize the biological, optical, and computational methods underlying this technology • Select appropriate equipment and staff • Develop a working and profitable business model centered around the aforementioned production level, investment, and genome price <p>Out-of-Scope</p> <ul style="list-style-type: none"> • Fabrication of flow cells • Synthesis of reversible terminator nucleotides • Focused screening of genome (such as SNP screening) • The provision of genetic or medical consultation
Deliverables	<ul style="list-style-type: none"> • Market assessment and competition analysis • Technical feasibility assessment • Full scale manufacturing requirements and protocol • Financial analysis over a 4-year project life cycle
Timeline	<ul style="list-style-type: none"> • Working sequencing prototype within 12 months • Scale-up operations within 2 years • Full-scale production in years 3-4 with simultaneous R&D • Liquidate or sell the company at the conclusion of fifth year

Table 1.2 *SynthSeq's* Project Charter

bases. The first team to meet these criteria will be awarded \$10 million. Outlining the company's specific goals, project scope, deliverables, and a timeline for developing its sequencing process, the *SynthSeq* project charter is presented in Table 1.2.

The scope encompasses the development of the sequencing process in all of its aspects. Sequencing by synthesis principles based on a technology established by Helicos BioSciences lays the foundations of the process, while the use of novel biological, mechanical, and optical innovations makes the *SynthSeq* approach unique. Once the DNA sample is extracted, purified, and sheared, short single strand DNA templates are immobilized on the surface of a flow cell where the sequencing chemistry occurs. The sequencing is dependent on the position of the template strands on the flow cell surface. The soft lithography polydimethylsiloxane (PDMS) flow cells are purchased as one-time use consumables from an outside vendor. They incur low consumables costs and allow the construction of customizable PDMS gaskets that easily adhere to the glass surface of the flow cells.

The chemistry relies heavily on reversible terminator fluorescent nucleotides, which prevent the elongation of the template strand until two small moieties on the nucleotide are chemically altered. This property allows time-controlled asynchronous base pair detection because the polymerization reaction is not occurring in real time. The synthesis of these specific nucleotides is outsourced to a company that specializes in modified nucleotide synthesis. However, an organic chemist will work in research and development to improve the function of these nucleotides since they are an integral part of the sequencing design. Once base addition successfully occurs, intricate optical techniques are used for single molecule detection and subsequent nucleotide identification. The nucleotide fluorescence relates directly to its position on the imaging grid of the flow cell, and the data is recorded in its corresponding fragment's sequence. An open source genome aligner is used to assemble the genome after all the individual DNA templates are sequenced.

The analysis of the final sequenced genome is out of the scope of the project. Currently, customers will not receive any medical consultation or diagnosis of elevated risks of diseases after his or her DNA has been sequenced. The company will refer customers to other companies that specialize in the analysis of genomic sequences. However, recognizing that this analysis can be lucrative, *SynthSeq* may opt to incorporate this service into its business plan in the future.

The goal of the first year is to develop a working prototype of the process. Once the prototype is functioning successfully, the scale-up and commercial sequencing will commence in the second year of operation. The number of sequencing stations will be increased to accommodate the increased production rate of genomic sequences. After scale-up is achieved, the company resources will be allocated to sustaining high throughput commercial sequencing. Also, funds will be directed to research and development for improvements in areas such as sequencing chemistry, optimal reaction conditions, higher throughput design, and better fluorescent detection techniques.

TECHNOLOGY READINESS ASSESSMENT

All next generation sequencing methods, including *SynthSeq*, represent an evolution of DNA sequencing technology. These current approaches are rooted in the motivation to drive down the cost of genome sequencing while increasing throughput and ensuring a high degree of accuracy. *SynthSeq's* ability to meet its process requirements is indebted to the considerable advances made by previous technologies.

Advancements in nanotechnology, nucleotide chemistry, optical devices, microscopy tools, and computational systems make the *SynthSeq* design successful in meeting its cost and throughput goals. The microfluidic flow cells, where the sequencing chemistry occurs, take advantage of micro-fabrication techniques to create a cell on the micro scale. The soft lithography cells reduce the reaction volume significantly compared to more traditional sequencing methods and thereby

conserve the amount of reagents used in the process. Another nano-technological contribution is polyelectrolyte layer by layer assembly, which forms a polymer film that prevents non-specific binding of free dNTPs to the bottom of the flow cell. This reduces background fluorescence and preempts recording the detection of a free nucleotide as a base in the sequence.

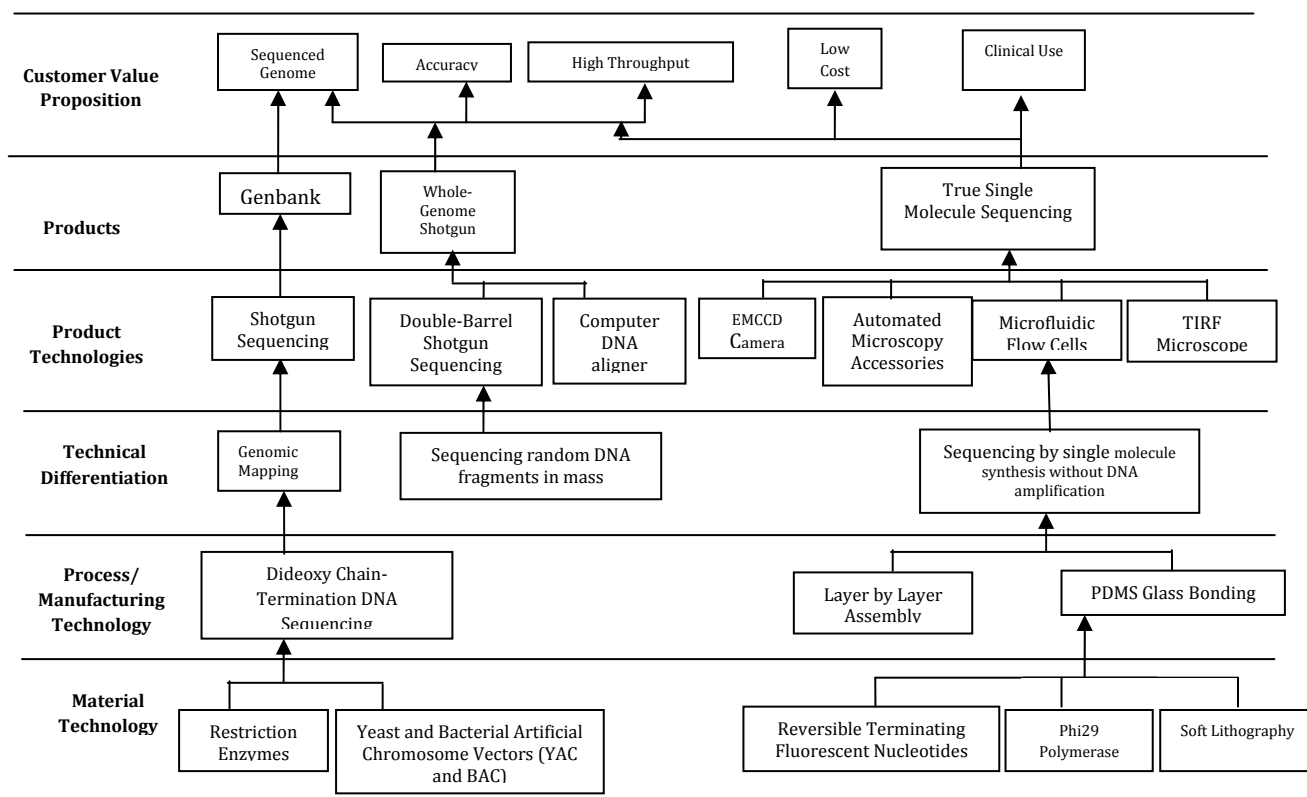


Figure 1.1. Innovation Map for *SynthSeq* Technology

The single molecule detection is made possible by the ability to suspend the DNA polymerization reaction to identify the base added. The development of reversible terminator fluorescent nucleotides allows the designer to control the reaction. The conversion of the 3'-OH of these nucleotides to a small allyl group allows the polymerase to add the base without any steric inhibitions, but prevents the polymerization of additional bases until the hydroxyl group is regenerated. Once the reaction is suspended, the use of a megapixel Electron Multiplying Closed-Coupled Device (EMCCD) camera and automated microscope stage positioners allow the simultaneous detection of single base additions on the order of magnitude of 10^5 DNA templates.

The high pixel resolution and fast frame rate serve as important parameters for high throughput because they allow more template elongations to be observed in a shorter period of time. Further improvements in speed and resolution will increase the throughput capabilities of the process. Also, advancements in genome alignment algorithms will facilitate the assembly of all the collected optical data into the genomic sequence in less time, and using less processing power.

In addition to imaging limitations, the throughput exhibits a strong dependence on the microscope stage accessories. Advancements in stage automation allow the process to image multiple areas in the same flow cell, which dramatically reduces the time necessary to sequence a genome. The velocity and settle time of the stage positioner dictates the speed at which different imaging areas on the flow cell can be analyzed. The piezo nanopositioner provides stage movement on the nanometer scale. This instrument is able to detect the presence of more than one template within a single pixel.

All of the above technologies are integral to the execution of the design goals. Figure 1.2 outlines the *SynthSeq* process. After taking into account capital, operating, and labor costs, as well as market pressures, a *SynthSeq* customer can expect to pay \$5000 for his or her complete genome.

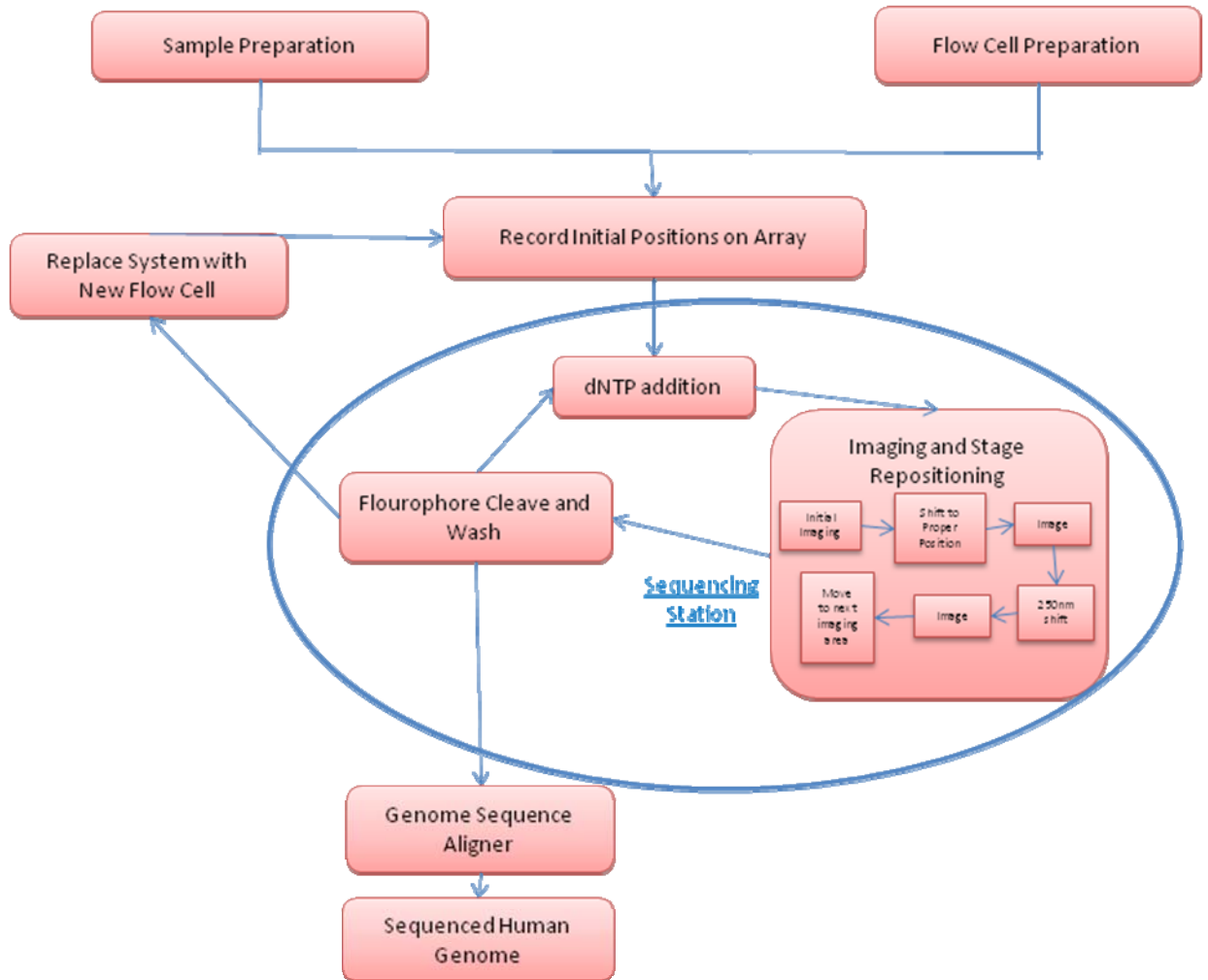


FIGURE 1.2 The SynthSeq process is outlined from the preparation stages to the product delivery.

CUSTOMER REQUIREMENTS

The customer requirements involved in human genome sequencing revolve around a few main issues: accuracy of the sequence, quick results, and a positive customer experience. These items help define the Critical-to Quality (CTQ) variables for the process, which are highlighted in Table 1.3. These variables represent the core drivers of this technology. Without achievements in either accuracy or high throughput, the process will not meet its requirements for sequencing 12 genomes a day with a minimum error rate of 1 per 100,000 bases. The CTQ variables for the

process can be categorized as *new-unique-and difficult* (NUD). NUD requirements represent basic customer needs that must be satisfied before they will purchase our genomic sequencing service. These requirements are not fully established in the genomic sequencing market. The developments made with the CTQ attract customers and are novel to the field.

The accuracy of the genomic sequence is affected by the error rates in each stage of the process. For customer ease, the genomic DNA is collected through a saliva sample, which may introduce some bacteria into the sample, trace amounts of erroneous DNA. During sequencing, errors are introduced through polymerase-facilitated nucleotide addition, single-molecule fluorophore detection by the EMCCD cameras and the elucidation of the identity of the excited nucleotide, and they are manifested in the alignment of the DNA sequence. A barcode system will be instituted to prevent misidentification, with the scanning of each genome taking place prior to sequencing to confirm that the genomic DNA sample corresponds to the correct customer.

The turnaround of the genomic data to the customer should be reasonable. This time frame

Customer Requirements	CTQ Variables	Weight
Accuracy	Sample Prep Error Rate	0.50
	Polymerase Error Rate	
	EMCCD Single Molecule Detection Error Rate	
	Fluorophore Cleavage Error Rate	
	Sequence Aligner Error Rate	
Result Turnaround	EMCCD Resolution	0.30
	EMCCD Frame Rate	
	XY Stage Positioner Velocity	
	Rate of Nucleotide Addition	
	Rate of Fluorophore Cleavage	
Customer Experience	Sample Method	0.20
	Confidentiality	
	Referral to Genome Analysis Service	
	Low Cost	

Table 1.3. Customer requirements satisfied by the *SynthSeq* technology CTQ variables addressed.

is linked directly to the process's throughput of data generation and alignment. Since the *SynthSeq* process is serial in nature, the main limitations to throughput originate in the speed of imaging the

single molecule fluorophores and moving from one imaging area to the next on a flow cell; the frame speed and pixel resolution of the camera and the velocity of the stage controller are key elements. Advancements in these technologies may significantly increase the speed of data generation for our process. *SynthSeq* was designed to successfully sequence twelve genomes in a day using four workstations. Considering both sample preparation and genome assembly, the turnaround is less than four days, providing the customer with a relatively fast result.

Our ability to attract customers to *SynthSeq* is essential for the success of our company. As aforementioned, saliva was selected as the sampling mode over other common methods such as blood and cheek swabs. This collection method was chosen for its non-invasiveness, simplicity and easy purification using a basic kit, while still providing adequate genomic DNA for analysis. *SynthSeq*'s customer price of \$5000 is competitive with the other sequencing technologies (see Chapter 2), so our customers are getting unparalleled value for their money.

CONCLUSIONS

The following chapters will show that an attractive business model can be built around the sequencing of entire human genomes using a novel single molecule Sequencing By Synthesis (SBS) strategy. This process will succeed in sequencing twelve genomes per day for 250 days per year using just four stations composed of reaction, imaging and fluid handling equipment. The genome will be sequenced at a cost of \$747 to the company, while our customers will be charged a competitive price of \$5000 per genome. This price significantly undercuts all current competitors and, compounded with the unprecedented accuracy of our sequencing technology and the positive customer experience provided by the company, should attract a majority of the genome sequencing market to *SynthSeq* while still ensuring comfortable profit margins.

We will describe the basic principles behind our technology and why it is superior to its competitors (Next Generation Sequencing Technologies, Chapter 2), highlight the components that

made possible the unprecedented throughput achieved by the *SynthSeq* method (High Throughput, Chapter 3), explain the steps required to take a human saliva sample and prepare a flow cell containing template DNA on a novel sequencing surface (Preparation Stages, Chapter 4), describe the chemistry behind our unique approach to single molecule SBS, the optical equipment used to image and sequence individual DNA molecules, and the analysis that proves this method can identify the sequence of nucleotides on the molecule (Chemistry and Detection, Chapter 5), review the hardware and software used to assemble a full genome from billions of short sequences (Genome Assembly, Chapter 6) and present a financial analysis proving that *SynthSeq* is a worthwhile investment poised to provide high profits to investors (Financial Analysis; Chapter 7).

2. NEXT GENERATION SEQUENCING TECHNOLOGIES

This chapter will outline the fundamentals of the *SynthSeq* technology, followed by an analysis of the genotyping market and existing companies, and finally a pricing scheme. We will demonstrate that our technology is poised to outperform any existing technology in terms of both cost and throughput, making it a prime target for venture capitalists. Existing genotyping firms that will be assessed include Illumina, 454 Life Sciences, Pacific Biosciences, and Knome. The former three firms offer sequencing platforms in the form of an instrument that can be purchased and used in private labs. Knome has a business model most similar to ours, as they provide a genotyping service to individuals. Further analysis will be provided in the following sections, all of which demonstrates the viability of our technology as a sound investment opportunity for angel investors.

MARKET ANALYSIS

The nature of the genotyping industry is quite volatile due to the constant and swift technological innovations that characterize the field of biotechnology. Consequently, a portion of *SynthSeq's* revenue must be allocated toward staying at the forefront of these technological advances by investing in research and development. Furthermore, the success of our business is highly dependent upon the low cost of the service that we provide. Many competing technologies exist, which would steal some of *SynthSeq's* market share if they were competitively priced. Fortunately, our technology is more accurate and cheaper than that of other existing firms (see the following Competitive Analysis), so *SynthSeq* definitely merits investment.

Another reason that costs need to be kept down is the current state of the market. The fact that it is nascent industry also means that the value of the personal genome is not yet obvious. Compounded by the fact that inferior, cheaper options, such as the SNP analysis, already exist, there is only so much that a customer would be willing to pay for what may turn out to be merely extraneous data on top of the SNP analysis which is already available commercially.¹⁴ Nevertheless, it is *SynthSeq's* firm belief that as more genomes are sequenced and compared, the great value of our product for personalized medicine will be realized.

Potential consumers of *SynthSeq's* genotyping service include several main types. Our target customers run the gamut from healthcare professionals hoping to determine a patient's risk factors to healthy individuals who want a copy of their genome for personal reasons. Scientists in labs across the country will want to sequence genomes cheaply and quickly to build up a comparative database for determining the mutations that lead to various pathologies. As more and more genomes are sequenced, the value of such a database will continually increase, incentivizing still more individuals to reap the benefits of having access to their genomic sequence.

Specifically, individuals with a family history of certain cancers or genetic disorders for which there is no current diagnostic might represent some of our customers. Necessary changes in

medical practice and insurance company policies (not to mention ethical dilemmas) notwithstanding, the technology has the potential to revolutionize the field of medicine, offering novel insights into diseases such as cancer, as well as a plethora of other genetic maladies. The next few years could see *SynthSeq's* genotyping service become a routine medical test, ordered for millions of individuals every year in the United States alone.¹⁵

SEQUENCING BY SYNTHESIS

In the two decades since the Human Genome Project began, sequencing technologies have grown increasingly more time and cost efficient. Perhaps the most significant technological leap in the quest for the thousand-dollar genome has been the advent of non-Sanger sequencing. Sanger sequencing uses modified, tagged nucleotides to terminate newly synthesized DNA fragments at specific bases. The fragments are then size separated. Since the modified terminator nucleotides are tagged, the sequence can be determined from this separation. Since the early 1990s, different implementations of this biochemical method have been the main way of sequencing DNA. However, over the past five years, new “next-generation” sequencing techniques have surfaced, which are exploited in highly parallel DNA sequencing platforms that are reducing the cost of DNA sequencing by orders of magnitude.¹⁶ The specific advancement that has made it possible for *SynthSeq* and other “next-generation” technologies to achieve such high throughput at such a low cost is known as Sequencing-By-Synthesis (SBS).

Sanger sequencing relies on the amplification of target DNA by polymerase chain reaction (PCR) and the ability to terminate this process, followed by the recognition of the identity



Figure 2.1 Sequencing centers producing the Sanger sequence data for mammalian genome projects are factory-like outfits with a large number of personnel.

and relative position of the ddNTP that terminated amplification. By doing this for billions of fragments of the target DNA being sequenced, the overall genome can be reconstructed. This basic idea has been coupled with developments such as capillary electrophoresis, laboratory automation and process parallelization. The first attempts to sequencing the entire genome involved huge sequencing centers run by hundreds of scientists and operators using these methods to construct one genome, taking years at a time and costing billions of dollars (Figure 2.1).¹⁷

SBS is fundamentally different. Instead of running a chemical reaction (PCR) on the target DNA first and then analyzing the products to find a sequence, the methods that use single molecule SBS detect individual DNA molecules and monitor the addition of nucleotides that incorporated during DNA synthesis.¹⁸ Innovations in template preparation, imaging, and genome alignment and assembly have made it possible to directly sequence a strand of DNA as it is synthesized by tracking which nucleotide is being added to each position of the template during the reaction. This method of sequencing DNA has been employed in platforms that cheaply produce large amounts of sequence data in a way that traditional (Sanger) sequencing cannot.¹⁹ There are many platforms that employ SBS, each using it in a unique way.

SYNTHSEQ TECHNOLOGY – ASYNCHRONOUS, SINGLE MOLECULE SBS

SynthSeq's technology is based on true single molecule SBS. DNA molecules are immobilized on a solid surface in a flow cell and are individually imaged by an EMCCD based optical detection system as DNA synthesis is carried out. The DNA to be sequenced is denatured and sheared and the billions of fragments of the single stranded DNA are annealed to a glass slide using novel surface chemistry. The molecules are distributed randomly over the surface of the glass slide. This slide is loaded into the flow cell, which is mounted on the microscope stage. Here, the sequence of chemical reaction and imaging steps, which identify the sequence of the individual

fragments, are carried out. DNA amplification is done using unique, modified nucleotides that reversibly terminate the extension reaction. These nucleotides are tagged with a fluorophore according to their identity, G, C, A or T, and have a 3'-allyl group replacing the usual hydroxyl group that temporarily blocks subsequent nucleotide addition. After incorporation of the nucleotides, the reaction is stopped, allowing for the entire slide to be imaged to identify which nucleotide was added to each template fragment on the slide. Then, the modified nucleotides are un-blocked (regeneration of 3'-OH and cleavage of fluorophore), which allows the extension reaction to continue and the identity of the next nucleotide added to be determined.

There are a few aspects of *SynthSeq's* approach to SBS that set it apart from other technologies. Some methods PCR amplify the DNA before sequencing which introduces amplification bias – an over-representation of reads from areas of the genome with between 40% and 65% G+C content. The *SynthSeq* approach directly sequences single molecules taken directly from samples without any artificial amplification. Another defining characteristic of the *SynthSeq* method is that it is asynchronous. DNA synthesis is blocked after each nucleotide addition and imaging steps are carried out in between nucleotide addition steps. The next nucleotide is not added until the reaction has been unblocked. Thus, it is known exactly how many nucleotides should have been added and recorded as part of the sequence. This virtually eliminates insertion and deletion errors. Other technologies use real-time imaging of ongoing reactions. These procedures often suffer from errors in reading homopolymers – stretches of DNA composed of the same, repeated nucleotide. Since our technology pauses following each successive addition, we know exactly how many nucleotides exist in a sequence of repeats.

Figure 2.2 below is a schematic of the steps in the *SynthSeq* process.

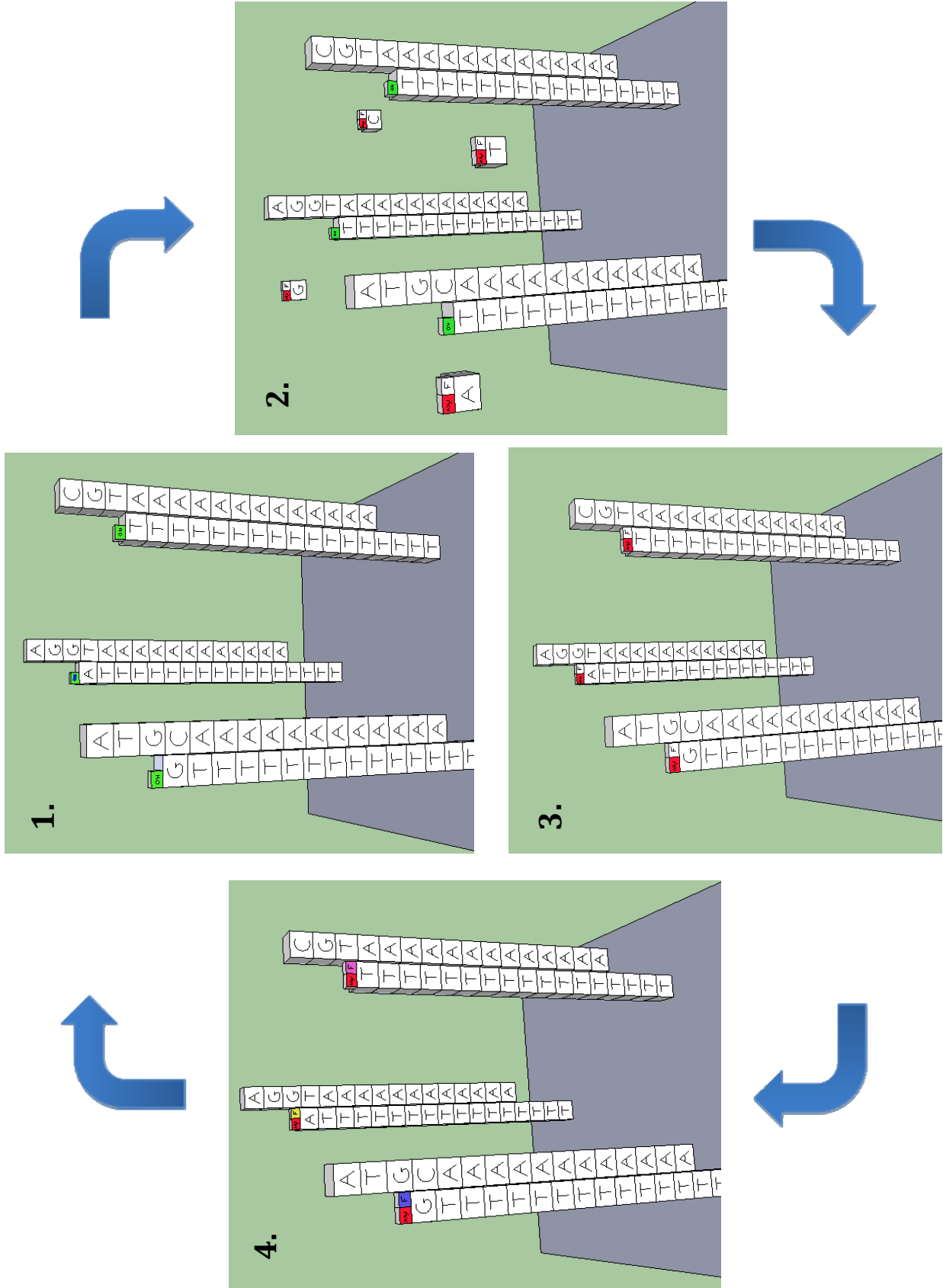


Figure 2.2 *SynthSeq*'s single molecule SBS: 1) Nucleotides modified with 3'-allyl group (shown in red) and fluorophore are introduced to template strands, 2) DNA synthesis occurs and modified nucleotides are added to template strands, 3) Fluorophores are excited (different excited fluorophores shown in blue, yellow and magenta) and molecules are imaged, 4) Fluorophores are cleaved and 3'-OH groups (shown in green) are regenerated.

COMPETITIVE ANALYSIS

Despite the unprecedented throughput of the *SynthSeq* technology that has been demonstrated, it is important to consider possible competitors in the genome sequencing market. Some of the companies that offer parallel services to ours include Illumina, Roche/454 Life Sciences, and Pacific Biosciences. Unfortunately, due to the great amount of information that is proprietary and protected, precise determinations of cost and throughput are often difficult to assess. Furthermore, like all biotechnology related fields, genomic sequencing is characterized by frequent technological advances and increases in efficiency that promote stiff competition. For this reason we will devote 15% of our revenue to research and development, in an effort to stay at the forefront of the industry. In the following sections, we demonstrate the advantages of our technology over the methods currently employed by other firms.

Illumina

Like many other next-generation sequencing firms, Illumina relies upon the parallel sequencing of millions of DNA fragments to determine an individual's genomic sequence. Specifically, they rely on a reversible terminator-based sequencing chemistry known as Solexa. First, randomly fragmented genomic single stranded DNA (ssDNA) strands are attached to a planar, optically transparent surface and then extended and amplified, creating a sequencing flow cell with millions of clusters that contain roughly one thousand copies of a certain template. Next, the clusters are sequenced using four-color DNA sequencing by synthesis technology that employs reversible terminators with cleavable fluorescent dyes, in addition to an undisclosed DNA polymerase. The tagged dNTPs then emit light at different wavelengths following laser excitation. This light is detected using total internal reflection (TIRF) optics. Finally, the reads are aligned against a reference genome using Illumina's own proprietary software.

Some of the advantages of the Solexa technology include the simplicity of their flow cell (random array, like *SynthSeq*), as well as the use of DNA clusters, which intensifies the signal sent to the imager, increasing the signal to noise ratio. Also, after completion of the first read, the templates can be regenerated, enabling a second read from the other end of the clusters, resulting in more data and an internal verification method.

Illumina claims their new HiSeq2000 to be the first commercially available sequencer to enable researchers to obtain ~30x coverage of two human genomes in a single run for under \$10,000 (USD)* per sample.²⁰ However, as demonstrated by Figure 2.3, Illumina’s throughput simply cannot compete with *SynthSeq*’s technology. Although the specific error rates of this instrument could not be found in the Illumina HiSeq2000 brochure or datasheet, it is likely that its increased throughput comes at the cost of accuracy. Researchers have noted that Illumina’s sequencing reads are prone amplification bias. It is likely that this occurs as a result of Illumina’s reliance on PCR to amplify their reads or their cluster amplification technique; *SynthSeq*’s technology demonstrates no such bias.²¹



Figure 2.3 Timeline of Illumina’s technology, which demonstrates their throughput deficit. (Image from Illumina, Inc.)

Roche/454 Sequencing

More competitive than Illumina’s technology, 454 Sequencing relies on longer read lengths to improve throughput. They employ sequencing by synthesis techniques and record light signals using CCD cameras. As shown in Figure 2.4, DNA beads are deposited into PicoTiterPlates, allowing

400,000 parallel reads. Their technology is highly accurate and is characterized by very few substitution errors, allowing *de novo* sequencing to be performed (i.e. no reference genome is necessary). Additionally, the accompanying software is nimble, enabling facile data transfer and fewer IT demands.

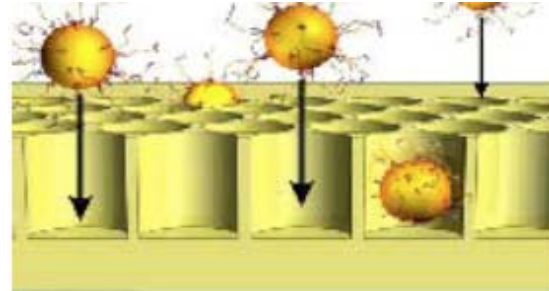


Figure 2.4 454's DNA beads being deposited into a PicoTiterPlater (Image from Illumina, Inc.)

However, as shown in Figure 2.5, 454 (like Illumina's Solexa) depends on PCR, which takes a great deal of time and may result in amplification bias. Additionally, 454 utilizes a pyrosequencing technique. One nucleotide species is introduced during each cycle, which is detrimental to both throughput and reagent costs. *SynthSeq* allows for the introduction all of the nucleotides in concert, while 454's technology necessitates four separate add and wash cycles.²²

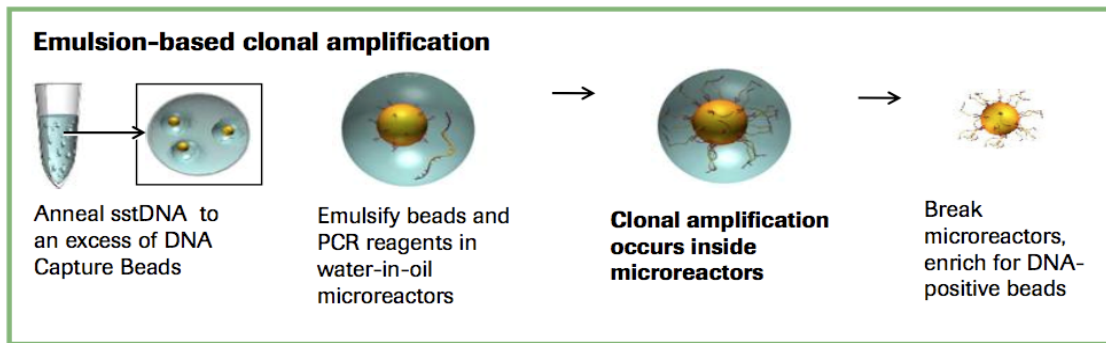


Figure 2.5. Formation of DNA coated microbeads by clonal amplification in water-in-oil microreactors (Image from Roche Diagnostics)



Figure 2.6. Timeline of 454's sequencing technology (Image from Roche Diagnostics)

The resultant timeline for the 454 Sequencing shown in Figure 2.6 indicates far lower throughput than is possible using *SynthSeq* technology. For instance, the quoted run time for one million reads is ten hours using the GS XLR70 sequencing kit, corresponding to just one billion bases per day.²³ Although this may sound like very high throughput, each of *SynthSeq*'s four stations will process

over thirty billion bases per day. *SynthSeq* does not involve PCR, and its alignment process will employ a reference genome rather than *de novo* sequencing, saving time and increasing throughput.

Pacific Biosciences

Pacific Biosciences, founded in 2004, has developed a platform that uses “Single Molecule Real Time” (SMRT™) DNA sequencing technology. This technology is unique in that it employs natural DNA synthesis by a DNA polymerase. Like *SynthSeq* technology, their approach is based on

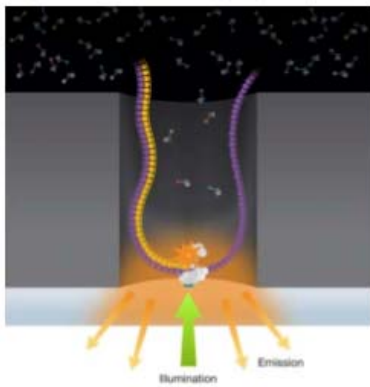


Figure 2.7 PacBio SMRT Sequencing, inside ZMW (Image from Pacific Biosciences)

the imaging of successive fluorescently tagged bases while a single DNA polymerase molecule processes them. However, they employ zero-mode waveguides (ZMWs), tiny aluminum wells that house the growing DNA strands, as opposed to the random array used by the *SynthSeq* technology. In addition to a SMRT cell that ensures a

high signal to noise ratio, PacBio's SMRT technology relies on phospholinked nucleotides that produce fast, accurate reads using DNA synthesis and a detection system that enables the real-time detection of single molecules. The low costs made possible by the PacBio platform (the most similar technology to *SynthSeq*) will surely result in some stiff competition.



Figure 2.8 SMRT sequencing along a DNA single-strand. (Image from Pacific Biosciences)

PacBio asserts that their longer read lengths promote simpler, more accurate assembly because they can fully span repetitive or structurally varied genomic regions that can present problems for shorter read-length platforms. The longer reads also reduce the need for high coverage, significantly lowering operating costs and requiring lower throughput. While these are accurate claims, our technology overcomes the read-length issue with its relatively error-free process (no insertions or deletions, only no-reads) and an aligner designed specifically for short reads. PacBio's SMRT sequencing technology, in contrast, is plagued by insertion and deletion errors that lead to the need for high coverage, severely decreasing throughput.

PacBio also touts the openness of its informatics design, which facilitates "scalable customization and integration with existing infrastructure", as well as the versatility of their SMRT™ Cell, which allows users to easily upgrade their setups. Similarly, they note that as detector technology improves, increases in throughput will be possible without changes to the assay. It is important to realize that while other companies are focused on ensuring the flexibility and user-friendliness of their instruments, *SynthSeq* does not need to concern itself with such issues. Given that we are offering a service instead of a sequencing platform, we have certain inherent advantages over the companies mentioned above. Our sole focus is on ensuring throughput and customer experience, not developing an instrument that requires maintenance and optimization.

Knome

Knome is another firm that may present some competition to *SynthSeq*. Unlike the above companies that peddle expensive genotyping instruments, Knome has a business model that closely parallels our own. Like *SynthSeq*, Knome is focused solely on customer experience and throughput, rather than manufacturing genotyping machines for the scientific community.

Originally situated as a quasi boutique firm selling an expensive product to self-proclaimed “pioneers”, Knome now offers its genotyping service to a wider audience, as a result of downward movement in price over the last couple of years. The identity of the sequencing technology they use is heavily guarded, but it should be assumed that they can compete with any of our other competitors in terms of throughput. Unfortunately, their error rates could also not be found, so gauging their competitiveness is rather difficult. However, with a price tag of \$68,500 for their whole genome sequencing service, *SynthSeq* is in position to significantly undercut them.

In a savvy economic move, Knome has recently developed two separate levels of sequencing to expand their market. In addition to the “KnomeCOMPLETE” (an individual’s entire genome sequence), the company has also started to offer the “KnomeSELECT” option, which includes only the exome, or the protein-coding region of the DNA sequence. The price tag for the limited analysis is \$19,500, more than three times cheaper than the full sequence and, at this stage in personalized medicine, almost equally useful.²⁴ It is important to note that much of Knome’s focus lies in customer experience, including representatives who personally meet with clients to discuss the findings and analysis of a team of geneticists. Customer confidentiality, access to consultation, and providing the latest genomic interpretations are three facets of Knome’s business model (Figure 2.9) that we need to strive for in order to erode their market share, our ability to undercut their price notwithstanding.²⁵



Figure 2.9 The Knome Business Model. (Image from Knome, Inc.)

PRICING ANALYSIS

Knome's cheaper exome sequence option is reminiscent of many other companies offering a similar service at much lower prices. For example, Navigenics and 23andme are two such firms offering a more limited analysis of the genes that have been linked to disease.²⁶ Ostensibly, *SynthSeq* with its advanced chemistry, low error rates, and entire genome capabilities should attract a wide audience if priced lower than our full genome competitors. However, before *SynthSeq* can be absolutely sure of success, the relative utility of the full genome over a "select" analysis consisting exclusively of the protein-coding DNA exome needs to be demonstrated. Until some palpable advantages of full-genome sequencing are demonstrated, a large proportion of the personal genotyping market will continue to purchase the cheaper option. Therefore, our focus needs to be on matching these competitors' prices as well.

One more consideration that *SynthSeq* needs to make in terms of pricing is the downward trend in cost of the human genome over time that is shown in Figure 2.10, below.

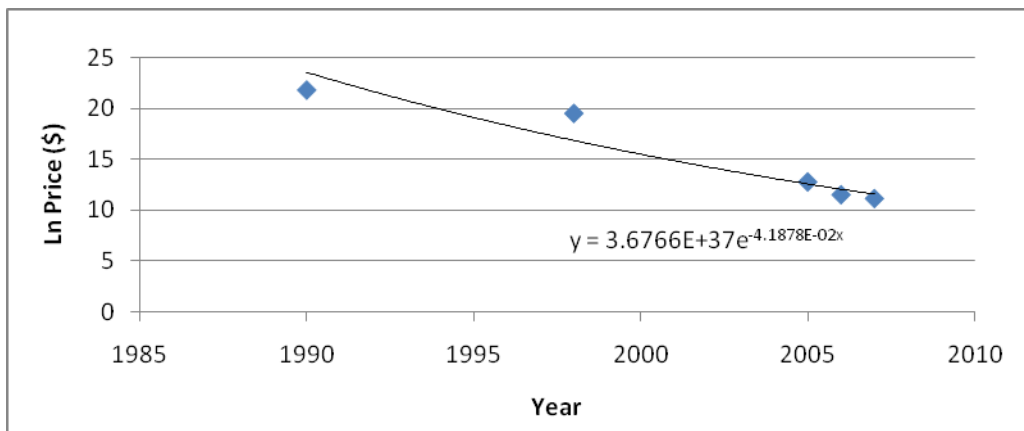


Figure 2.10. Full genome sequencing price trends (data shown for Knome), Year 1 corresponds to 2005 and remaining years follow. Predicted prices are \$43,000, \$26,000 and \$10,500 for years 2011, 2013 and 2015 respectively.

Based on the above trend, the cost of total genome sequencing has been decreasing exponentially over the course of the last decade. *SynthSeq* must consider not only the price at which it will be competitive in the first year, but also how it might need to respond in subsequent years to

the continuing downward trend of genotyping costs. It is important to note that the prices of our competitors are inflated to what the market will tolerate, and do not necessarily reflect the actual cost of producing a genome. As soon as we introduce a cheaper option, other companies may be in a position to match our price, which is another aspect of the market that *SynthSeq* must take into consideration (see Chapter 8 for more detailed financial analysis).

CONCLUSIONS

As demonstrated in the above sections, the *SynthSeq* technology presents a sound investment opportunity. Our superior throughput and very low cost compared to our competitors is a direct result of the relatively error-free nature of our detection strategy (discussed in Chapter 5). Starting out with the best technology is the greatest boon to our future ability to keep costs low. Even when undercutting the closest competitors' prices by more than half, our company will post significant profits. However, remaining in this advantageous position over the next few years will admittedly be a challenge, which is why *SynthSeq* is devoting considerable resources to continued research and development efforts. We cannot over-emphasize the importance of providing our customers with a personalized experience rivaling that of Knome, including but not limited to the latest interpretation of their genomic sequence. For the above reasons, investors should be assured that their money will support the most viable solution to fast and inexpensive personal genotyping.

3. OPTIMIZING *SYNTHSEQ*'S THROUGHPUT

The *SynthSeq* process aims to sequence twelve genomes per day. The technology must generate massive amounts of data in parallel to meet this goal due to the large number of base pairs in a human genome and the nature of sequencing by synthesis. The *SynthSeq* technology remains competitive in the commercial sequencing market because of its high throughput capabilities. Since our process is serial in nature, many unique design decisions are incorporated to transform a typically slow and low throughput, but highly accurate process into a system that meets its desired throughput goal.

Many of the bottlenecks in the *SynthSeq* design originate in the chemistry and detection of single fluorescently tagged nucleotides after they are added to the DNA template strands. Unlike

other technologies that sequence in real time, *SynthSeq* suspends the polymerization reaction to determine which base is incorporated. This places time limitations on the process, but these restrictions are overcome by utilizing mechanical and optical equipment in novel ways.

Many variables affect the overall throughput calculation of the *SynthSeq* process. The following factors were heavily considered during the design stages: pixel size, fields of view, kinetics of chemical reactions, read length, optical efficiency, and coverage. These parameters are ultimately optimized to generate the throughput needed to reach *SynthSeq*'s goal.

EMCCD SIZE OF PIXELS AND DIMENSIONS OF PIXEL ARRAY

The pixel size and dimensions of the pixel grid for the EMCCD camera set the parameters for the optical field of view and the detection volume. Every camera specifies the pixel size inherent to its chip and how many pixels lie on that chip. *SynthSeq* utilizes a megapixel camera with 1024 by 1024 pixels to take advantage of as many pixels in a single imaging area as possible. The pixel is ultimately where a single nucleotide fluorophore is detected and identified. Ideally, a single DNA fragment immobilized on the surface of the reaction chamber will correspond to one pixel. Therefore, the more pixels present on the camera's detection chip, the more DNA templates can be sequenced in one imaging area. The camera chosen by *SynthSeq* has one of the largest pixel array dimensions on the EMCCD market. When new camera innovations arise with more pixels in the imaging chip, that technology can be applied to the *SynthSeq* process to increase throughput even further.

The size of the pixel affects the size of the imaging area, the detection volume, and the kinetics of the nucleotide addition. The camera consists of pixels with standard 13 μ m by 13 μ m dimensions. However, through a 60x objective and a 0.5 magnification changer built into our detection microscope, the pixel size is reduced to 0.5 μ m by 0.5 μ m.

$$\frac{13\mu\text{m}}{60x \times 0.5 \text{ mag changer}} \cong 0.5\mu\text{m pixel length}$$

Equation 3.1

The microscope used in the process offered 100x objectives and other magnification changers, but the resulting pixel size would have been too small for *SynthSeq*'s design. The size of the pixel sets the imaging area to be about 0.25mm²; this value corresponds to the field of view in the microscope.

VIEWING MULTIPLE IMAGING AREAS

A feature unique to *SynthSeq* is its ability to simultaneously sequence templates in multiple imaging areas and then rotate through these fields of view to perform single molecule detection. This practice increases the throughput of this overall process tremendously. If only one imaging area is sequenced at a time, the throughput of the process would fall and sequencing a single genome would take over 1300 hours and consume over 1600 flow cells. This time and use of resources is not feasible for *SynthSeq*.

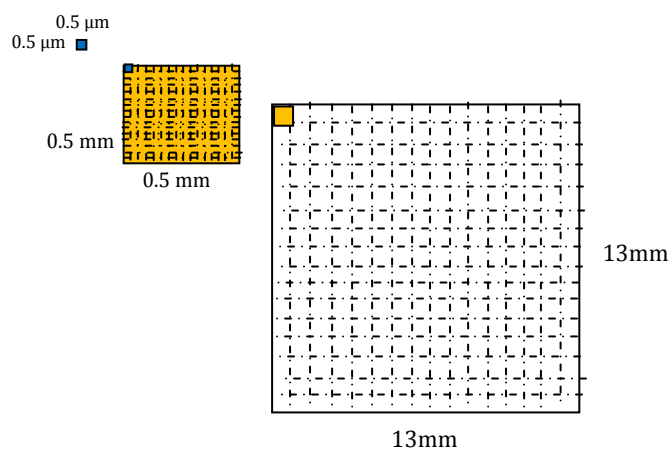


Figure 3.1. The dimensions of the pixel size, imaging area, and overall reaction chamber directly impact the throughput. The white grid represents the reaction chamber in the flow cell that consists of 841 imaging areas. Each imaging area, shown in orange, has 1024 by 1024 pixels, each with dimensions of 0.5μm by 0.5μm (in blue).

To solve this problem, sequencing multiple imaging areas simultaneously is applied. A total of 841 imaging areas fit into the 13mm by 13mm reaction chamber in the flow cell with the 0.25mm² size of one imaging area. This increases the throughput of detection to 2.8×10^8 DNA templates sequenced simultaneously in the flow cell. The time moving from one imaging area to the next is a major factor to the throughput of the process and will be discussed in the next section.

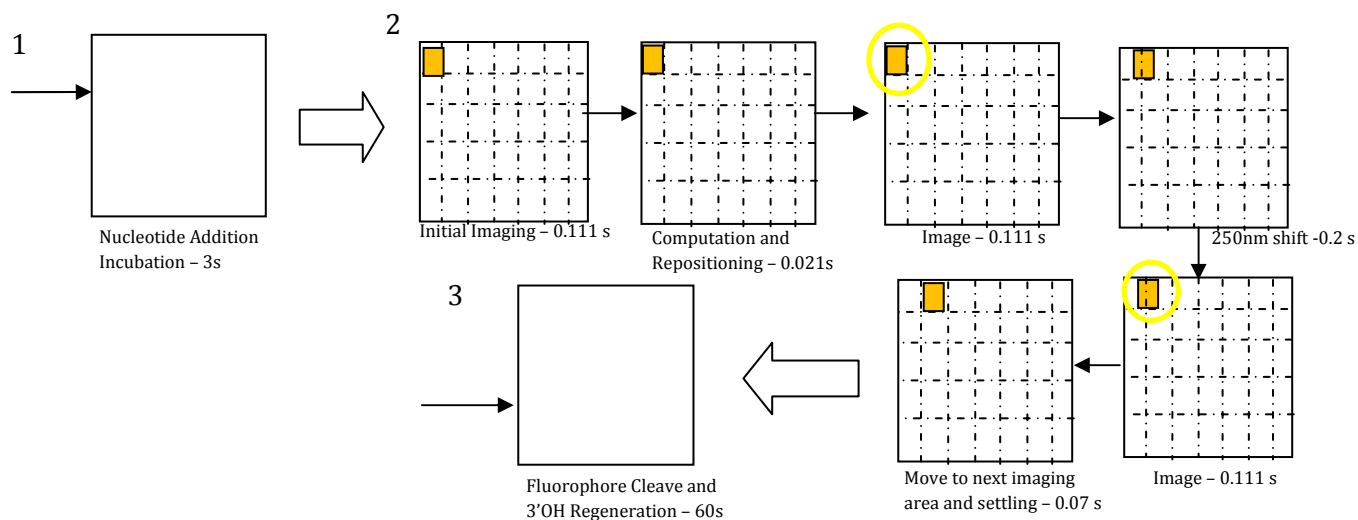


Figure 3.2 The outline of the chemistry and detection cycle for a single nucleotide addition to the DNA templates. (1) The free dNTPs are incubated over the entire reaction chamber. (2) A single imaging area undergoes imaging, shifting, and another imaging before the stage moves on to the next imaging area. (2) The cleave solution is injected into the entire reaction chamber to allow the addition of the next free dNTP round to sequence the next base on the templates.

SINGLE NUCLEOTIDE ADDITION AND DETECTION

The rates of polymerization and fluorophore cleave/3'-OH regeneration reactions as well as the imaging protocol to detect the nucleotides are major time limitations in the design and influence the throughput. Figure 3.2 displays the time involved in every step in the chemistry and detection process. The innovative equipment used at the sequencing station permit the asynchronous sequencing process to work in the proper time frame to achieve our throughput goal.

The incubation of the free dNTP reaction solution in the reaction chamber occurs for 3s. This incubation time does not significantly change the throughput when increased. Therefore, a conservative incubation time was chosen to ensure successful nucleotide addition. The time needed to cleave the fluorophore from the incorporated nucleotide and wash the tags out of the reaction chamber is only 60s. This rate also does not limit the throughput time significantly.

The imaging of the tagged nucleotides provides the most difficult challenge to maintaining the goal of sequencing twelve genomes per day. The imaging time depends on the frame rate of the camera, which is set at nine frames/s. Faster cameras exist on the market, but the EMCCD is chosen for its megapixel imaging grid, not for its frame rate. The two stage automated controllers, X/Y positioner and microscan piezo from ThorLabs move the stage to image multiple imaging areas at once. Since the shift time (0.02s) using the piezo did not affect the throughput, two images per imaging area are incorporated into the design to improve the optical efficiency (discussed further below). The speed of the X/Y positioner (10mm/s) is essential for moving between imaging areas quickly. Overall, each imaging area takes 0.444s and the total time over all imaging areas for the asynchronous addition of one nucleotide to each template is 374s. Achieving the necessary throughput hinges upon this imaging cycle.

READ LENGTH

Another variable considered in the throughput calculation is the read length of each DNA template. The read length is defined as the number of base pairs that are sequenced on a single template strand. The read length sequence is then sent to the aligner to assemble the entire human genome. DNA shearing equipment limits how small the genomic DNA can be fragmented. The specifications of the shearer used in the *SynthSeq* process cut the DNA into a mean length of 100 base pairs. These fragments are immobilized on the flow cell surface. The shear-length exceeds the

read length of the templates to prevent reaching the end of a template strand before all cycles of nucleotide addition have been completed.

SynthSeq's read length is 32 base pairs, which is shorter than many other competing technologies. A shorter read length makes identifying repeating sequences more difficult when aligning the genome. However, *SynthSeq* requires a shorter read length because of the asynchronous nature of its sequencing process. The polymerase that catalyzes nucleotide addition can only remain attached to the DNA template strand for 4.1 hours (see Chapter 5). If the read length is increased, the time required to sequence more bases per template would surpass that time and new polymerases would have to be incubated into the chamber mid-sequencing.

OPTICAL DETECTION EFFICIENCY

Single molecule SBS relies on the detection of individual DNA strands. The pixel count and number of imaging areas limit the number of genomic fragments that can be sequenced asynchronously. The system would be at its ideal level of template detection throughput at 1.05×10^6 templates per imaging area. However, an inherent feature of the *SynthSeq* technology is the random immobilization of DNA templates on the surface of the flow cell's reaction chamber. As a result, the templates are bound randomly on the flow cell surface at a particular area density defined as:

$$n = \frac{\# \text{ templates}}{\# \text{ pixels}} \quad \text{Equation 3.2}$$

The area density of template packing is restricted because the signals from different growing strands contaminate one another when they are too close due to optical diffraction. The intensity of a single fluorescent base appears on the pixel grid as roughly a 300nm diameter circle.

This disc cannot interfere with other template fluorescent circles or pixel boundaries following specific boundary conditions. If one pixel has two template circles, then the pixel becomes inactive because the wavelength emission of a particular base cannot be assigned to the proper template. Also, the intensity ratios from two cameras used to identify bases will not fall in acceptable ranges when overlap occurs. The inability to identify the emitting fluorophore would result in an unidentified base at that position in the sequence.

The random arrangement of templates on the surface causes some pixels to be unusable. The ratio of usable pixels to total pixels becomes the 'efficiency factor' of the template/pixel system and significantly affects the throughput. Higher efficiency allows more genomic fragments to be sequenced simultaneously in a given imaging area, resulting in fewer flow cells used per genome and a reduced total sequencing time. The efficiency is defined by Equation 3.3.

$$\text{efficiency} = \frac{\text{usable pixels}}{\text{total pixels}} \quad \text{Equation 3.3}$$

The probability of one template optical center landing in one pixel follows a statistical distribution. Treating templates as fluorescent points, an optimally loaded system has 36.8% empty pixels, 36.8% pixels containing one template, and 26.4% pixels containing two or more templates. This calculation results in an efficiency factor of 0.368, which corresponds to the maximum yield for the pixel array efficiency when optical diffraction is neglected.²⁷

The optical diffraction of the fluorophores and the 250nm stage shift must be taken into account to obtain an accurate estimate of the system's efficiency. To adapt to this design, a computational simulation is performed using Java code presented in Appendix D. This takes into account the scenarios where a pixel remains usable or becomes inactive due to template-template and template-pixel boundary interactions.

A 1024 by 1024 two-dimensional array models the imaging area of the EMCDD. Each box in the array is broken down further into a 50 by 50 array to evaluate the pixel boundary conditions. A random number generator randomly places numbers within the smaller arrays. The simulation evaluates the number's position relative to both its position in the array and its position in comparison to other random numbers.

The boundary conditions follow certain criteria to determine if a pixel is viable. It is assumed that if 50% or more of the template's area lies within a pixel its intensity ratio can still be

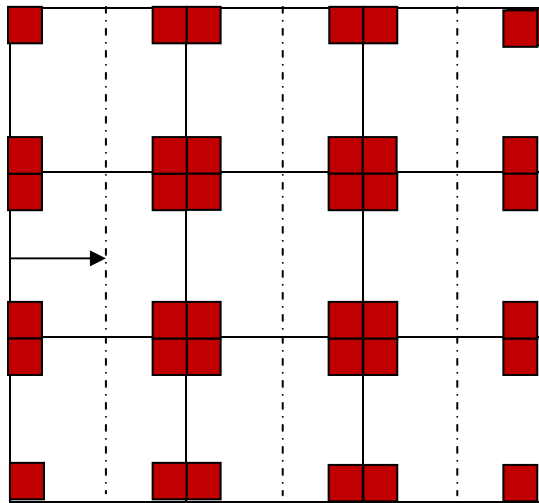


Figure 3.3 The 3 by 3 array represents the larger pixel grid. Since a random coverage is used for template immobilization, computational simulation must be used to determine the statistical efficiency of the pixels. If a template lands on any red area, that data is not usable in the first optical imaging step. The grid is shifted by 250nm to the right shown by the dotted lines. This shift significantly improves the efficiency.

accurately determined, and it is labeled viable.

Therefore, if the simulated template lies anywhere in the pixel other than the four corners, where significant pixel overlap occurs (the red areas of Figure 3.2.), the template is considered viable and the count for that particular pixel is set to 1. If a template lies on one of the four corners, it appears as a diffractive disc straddling 3 or 4 pixels. It is improbable that more than 50% of that disc will lie in one pixel, thus making all four pixels unusable. Only one template can lie in a pixel at once. If the simulated pixel already has a value of 1 (i.e. a template already resides in the pixel), the box becomes unusable and the value of that box in the array is set to -1.

Another constraint on optical detection is contamination from neighboring pixels. As long as less than 10% of the diffractive disc from a neighboring template encroaches upon the target pixel, the target pixel is not contaminated. If any more than 10% of that disc lies in the target pixel,

this contaminating signal causes an intolerable spread of the intensity peaks of the target signal. This infringement makes the target pixels unusable.

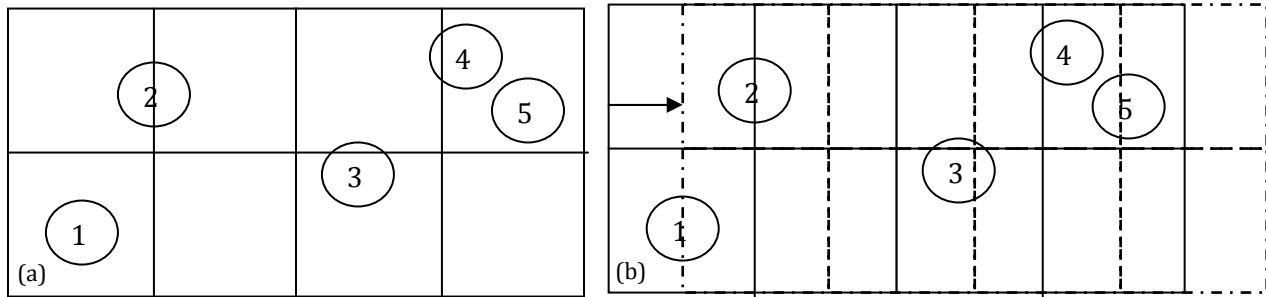


Figure 3.4 (a) The templates are randomly arranged on the pixel grid. Number 1 is a good example of a viable pixel placement. Number 3 is still in a good place because more than 50% of it is in a pixel. Templates 4&5 are not usable because they are both occupying the same pixel. (b) After the shift is applied, template number 2, which was unusable previously is now able to be used in the sequencing.

Using these assumptions, the efficiency of the pixel grid becomes 0.186. This efficiency is about half the efficiency calculated when the templates are treated as points. This drop is expected because simulating the templates as discs imposes additional limitations to what the system can identify and thereby reduces the number of usable pixels.

In order to compensate for this decrease in efficiency, a 250nm stage shift is performed using the piezo positioner. The shift will increase the efficiency by making some templates detectable although they were previously considered in an unfavorable position on the pixel grid (Figure 3.4(b)). The efficiency calculated is also a function of the area density. The optimal area density is determined by trial and error to see which value produces the highest efficiency. With an $n=1.15$, when 1.21×10^6 templates are laid out randomly on the imaging surface of the flow cell, an efficiency of 0.324 is achieved. This value is close to the ideal efficiency obtained when both the optical diffraction and shift were not taken into account. This efficiency value is incorporated into the overall throughput equation.

An ordered array could have been produced on the slide's surface instead of a random coverage of DNA fragments. There exists instrumentation that could produce such an ordered

array at the micron scale, such as the DPN 5000 System from Nanoink, Inc. This system uses Dip Pen Nanolithography (DPN), a technique that uses an atomic force microscope tip to transfer molecules to a surface.²⁸ This could be used to “print” arrays of poly-T tails on the surface of a glass slide that would then produce an ordered array of template DNA fragments. There is one main advantage to this approach; an ordered array makes it easy to track DNA fragments on the surface after each nucleotide addition and imaging step. The templates could be ordered so one template corresponds to exactly one pixel.

There are several reasons why random coverage of single stranded DNA template strands was chosen instead of an ordered array. The first reason is the cost and complexity of machinery. Dip Pen Nanolithography machines can cost up to \$500,000 compared to the machinery required for random coverage, which costs roughly \$17,000. Another reason an ordered array was passed over was that it was unclear whether or not nano-array printing machines could print onto a bio-inert surface. Representatives from nanoink could not guarantee success using the DPN5000 machine in this context. Finally, the array printer would have to produce an ordered array of single molecules; the DPN5000 cannot consistently print only one single stranded DNA oligonucleotide. There would have been many positions on the array that contained two or more template DNA fragments producing unusable data.

THE COVERAGE FACTOR

The random nature of this process requires an over-generation of information in order to assemble a full genomic sequence. The target genome must be multiplied by a coverage factor to ensure no gaps exist in the sequence. This factor increases the total length sequenced and lowers throughput. Therefore, this coverage factor must be optimized to reach the highest throughput while eliminating base pair gaps. At a given base pair position, a success is defined as the existence

of a sequenced fragment containing the base pair in question with the assumption that the aligner will position it properly into the sequence. More critically, the success of a proper base in the sequence depends on the success of binding polymerase to the template strands and maintaining

Step in Sequencing Process	Probability of Success
Polymerase Binding	99%
Polymerase Remaining on Template Strand	99%
Nucleotide Incorporation	99%

high processivity, incorporating a free dNTP into the growing template strand, generating a strong fluorophore signal for detection, and assigning the signal to the proper nucleotide so it can be placed in the genome sequence. Each of these steps in the single molecule sequencing has a probability of failure that contributes to the overall probability of gaps forming in the genome assembly. Therefore, each of these steps received individual probability analysis to estimate the

Table 3.1 Probability of success for each possible point of error in the sequencing process. Because *SynthSeq* is such a low error process every step has a 99% chance of being successful.

depth needed for whole genomic coverage. The probability of success for each of these steps is shown in Table 3.1. These specific analyses are shown in detail in Chapter 5.

The total coverage is established using a negative binomial distribution. The purpose of finding this distribution is to determine how many copies of the genome must be sequenced to reach the maximum overall error rate of 10^{-5} or 1 in every 100,000 bases sequenced. This function models the probability that there will be a certain number of failures before a predetermined number of successes can occur. One of the variables in this distribution is the probability a success will occur. The success rates of each individual source of error in the process contribute to the overall probability of success. The total probability is the product of the probability from each component (Table 3.1), which equals 0.95. Equation 3.4 is the formula for the negative binomial distribution.

$$nb(x;r,p) = \binom{x+r-1}{r-1} p^r (1-p)^x$$

Equation 3.4

Where x is the number of failures, r is the number of successes, and p is the probability of success. In this system, the number of failures equals mean coverage - 1, and the number of successes equals 1. Every base position in the genomic sequence needs at least one a success or identified base to add to the sequence. The coverage is how many copies are required of a particular base position. The failures are represented by X's, or unidentified bases, in that position when processed by the

Data	Description
5	Number of failures
1	Threshold number of successes
0.95	Probability of a success
Result	Description
2.96875E-07	Negative binomial distribution

Table 3.2 The coverage is determined using the negative binomial distribution. With a probability of success at 0.95, the error rate is well below the desired maximum error rate. The process uses a 6-fold coverage.

aligner. The failure number is varied until the probability of error is under 10^{-5} . After this evaluation and application of Equation 3.4, the coverage factor is calculated as 6 (Table 3.2). Therefore, six copies of the genome must be sequenced in the process to accurately reconstruct a full human genome. This coverage provides an overall error rate of 2.97×10^{-7} , which is lower than the 10^{-5} target.

Many commercial competitors require more than 20-fold coverage for their sequencing technology. This is a major advantage for *SynthSeq* because it allows a higher throughput. This lower coverage is made possible by the asynchronous nature of the process. Since the total number of nucleotide additions is known, no deletion or insertion errors exist. If a base is not properly identified in a particular position, then an X is placed in its spot and the sequencing continues in the proper order. Also, *SynthSeq* does not need to worry about homopolymers – repetitive bases in a single template – because reversible terminator nucleotides allow for the stepwise addition of nucleotides through those sequences. Finally, since the *SynthSeq* process does not rely on amplification of the extracted consumer DNA, it avoids amplification bias inherent to PCR. The lower number of sequencing errors gives some flexibility in the need for additional coverage.

THROUGHPUT OPTIMIZATION

A critical feature of this technology is that it accomplishes exceptional throughput capabilities without sacrificing highly accurate sequencing. Many design variables are taken into account when developing a cost-effective process with the required twelve genomes per day capability. These factors are optimized to achieve our throughput goal with four sequencing stations. The variables of the process that were optimized, and the design decisions made for each variable, are displayed in Table 3.3.

First, the pixel size is dependent on the objectives and magnification changers available for the microscope system. The 60x objective with the 0.5 magnification changer sets the pixels in the imaging area to roughly $0.25\mu\text{m}^2$ and the total imaging area to 0.25mm^2 . The number of imaging areas in the

Throughput Variable	Optimized Value/Decision
Pixel Size	0.5 μm
Reaction Chamber Dimension	13mm by 13mm
Read Length	32 base pairs
Optical Efficiency	0.324
Coverage Factor	6-fold

Table 3.3 Many factors were optimized to determine the exact throughput of the *SynthSeq* process.

13mm by 13mm flow cell reaction chamber is 841. These areas set the parameters for how many DNA templates can be sequenced simultaneously in one flow cell. Equation 3.5 shows the ideal number of templates that can be sequenced simultaneously in one flow cell:

$$N_{\text{templates ideal}} = 1024 * 1024 * 841 \text{ imaging areas} = 8.8 * 10^8 \text{ DNA templates} \quad \text{Equation 3.5}$$

Where one imaging area contains a 1024 by 1024 pixel grid. Incorporating the read length determines how many base pairs are being sequenced in a single flow cell, which is calculated in Equation 3.6.

$$N_{\text{bp ideal}} = 1024 * 1024 * 841 \text{ imaging areas} * 32 \text{ base pairs} = 2.8 * 10^{10} \text{ base pairs} \quad \text{Equation 3.6}$$

If **100% efficiency** could be achieved, one flow cell would be more than adequate to sequence a human genome. However, the other factors must be accounted for to get a more accurate calculation of *SynthSeq's* throughput.

The major limitations to the throughput of this process are the efficiency and coverage factors and the cycle time associated with the identification of each nucleotide. A higher efficiency correlates to more templates being sequenced at once and less data being discarded because of optical detection ambiguity. A small increase or decrease in the efficiency significantly alters the throughput results. The real number of base pairs sequencing in one flow cell includes the efficiency factor.

$$n_{bp} = 1024 * 1024 * 841 \text{ imaging areas} * 32 \text{ base pairs} * 0.824 = 9.1 * 10^8 \text{ base pairs} \quad \text{Equation 3.7}$$

The coverage factor dictates how many total base pairs must be sequenced during the process. The 6-fold coverage factor forces the overall sequencing time to increase as well as the amount of equipment needed to maintain the 12 genomes/day goal. This coverage requires the sequencing of $18 * 10^9$ base pairs. Equation 3.8 confirms that two flow cells are required to ensure that the adequate number of bases are sequenced to properly align the genome.

$$\frac{18 * 10^9 \text{ bases in 6 fold coverage}}{9.1 * 10^8 \text{ bases sequenced per flow cell}} = 1.9 \approx 2 \text{ flow cells per genome}$$

Equation 3.8

The completion of the chemistry and detection cycle for one nucleotide addition sets the time for the whole process occurring at the sequencing station. These times are limited by the polymerase and cleave kinetics, frame rate of the EMCCD, velocity of the piezo controller, and the velocity and settle time of the X/Y stage positioner. The time required to sequence one base across all templates in the reaction chamber is about 7.3 minutes. This includes polymerization, imaging, and cleaving steps. This chemistry and detection cycle occurs 32 times per flow cell to sequence each template to its read length. Consequently, each flow cell runs for 4 hours. In between flow cell

sequencing, a technician must dispose of the finished cell, reapply oil to the microscope stage, attach the next flow cell properly, check for errors in the automation process, and begin the sequencing cycles. This procedure takes an estimated time of fifteen minutes.

After summing all these process times, the sequencing of one genome requires about 8.25 hours. Each sequencing station can sequence up to three genomes per day assuming operation occurs for 24 hours a day. One night operator will be working to ensure sequencing progresses as planned and to change the flow cells when necessary. To achieve the goal of twelve genomes per day, *SynthSeq* requires four sequencing stations.

CONCLUSIONS

SynthSeq maintains a highly accurate sequencing process with unprecedented throughput capabilities. The process design is optimized to sequence twelve genomes per day with an error rate on the order of 10^{-7} , and 6-fold coverage. As demonstrated, random coverage is not detrimental to the throughput of the process; the 32.4% efficiency achieved is comparable to the predetermined efficiency value stated in the literature.

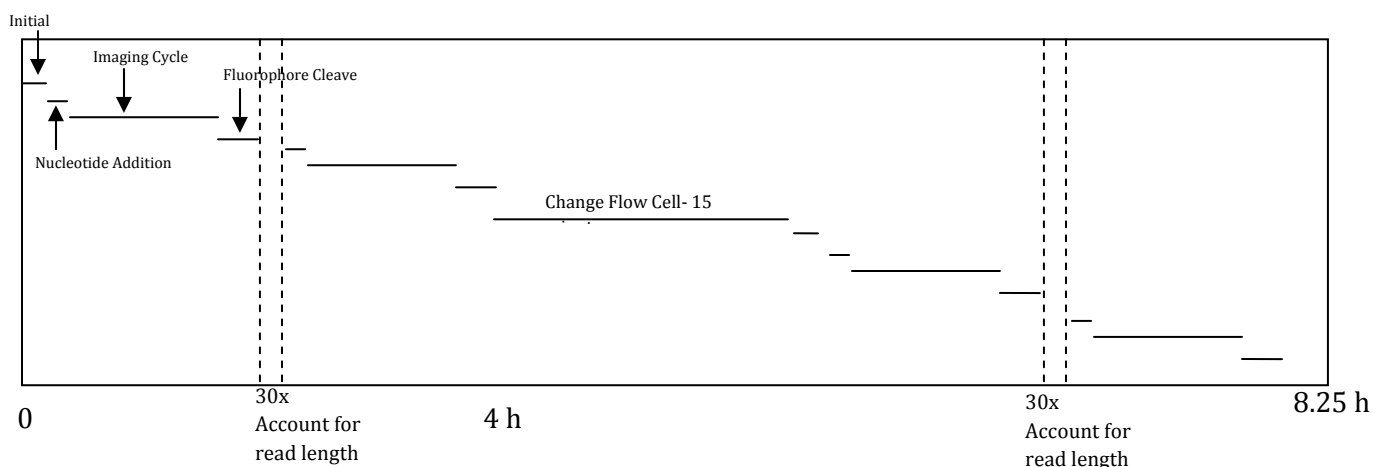


Figure 3.5 The time needed to sequence a single human genome in the *SynthSeq* process is outlined in this Gantt chart.

The Gantt chart in Figure 3.5 displays the time needed for each step in the sequencing and the total time found using the throughput calculation. Now that proof of the throughput has been established, the following chapters will discuss the details of the sample and flow cell preparation, the chemistry and detection cycle for adding a nucleotide, and the assembly of the genome.

4. PRE-SEQUENCING PREPARATIONS

Before the actual sequencing of the human genome can commence, the DNA must be purified, the reaction chamber flow cell must be assembled, and the sequencing station must be assembled. This chapter will outline the necessary preparation stages that need to be performed before the actual chemistry and nucleotide detection occurs. The DNA sample is received as a saliva sample for customer ease. The DNA in this sample is extracted, purified, sheared, and modified to fit the requirements of the *SynthSeq* process. The genomic DNA fragments are immobilized randomly on the flow cell reaction chamber surface. The bottom glass cover slip of the flow cell undergoes layer by layer assembly to prevent the non-specific binding of fluorescent

nucleotides to the flow cell surface. Once the flow cell is assembled with the immobilized DNA templates, it is attached to the sequencing station via the microscope stage.

DNA EXTRACTION, PURIFICATION, AND PREPARATION

Covaris s2 Shearing Conditions	
Base Pair Size	100bp
Duty Cycle	20%
Intensity	5
Cycles per Burst	200
Frequency Sweep	yes
Z Height	6mm
Temperature	6 °C
Seconds (time)	480

End Repair Mixture Components	
Reagent	Volume Required (μL)
DNA sample	29
Nuclease-free water	46
t4 DNA ligase buffer with 10mM ATP	10
dNTP mix	4
T4 DNA polymerase	5
Klenow enzyme	1
T4 PNK	5

Total Volume	100
---------------------	-----

Genomic sequencing begins with the

collection of the customer’s DNA. Since saliva is not the most pure source of genomic information, the need for high purity sample collection is high. The DNA Genotek Oragene® kit solves this problem with a solution that mixes with saliva samples for preservation. A kit is mailed to the customer; he or she spits into the prepared sample and mails their DNA sample back to *SynthSeq*. The sample is then loaded into the Magtration 12GC, manufactured by PSS Bio Instruments, which uses paramagnetic-particle technology to purify the DNA from the

Oragene® solution. The elution volume is 200µL with a median yield of **Table 4.1** Shearing conditions for DNA samples

3.8µg/200µL of DNA and a median $A_{260/280}$ ratio of 1.95.²⁹ The DNA is now isolated, but its length is too long for the purpose of the design. To rectify this problem, the DNA sample is inserted into the Covaris s2 Shearer, where Adaptive Focused Acoustics energy, more commonly used in the process of breaking up kidney stones, shears the double stranded DNA. By using the specifications found in Table 4.1, we can obtain a target peak for a length of 100 base pairs.³⁰

Adding poly-A tail Mixture

Table 4.2 Components used to repair ends of sheared DNA.

Reagent	Volume Required (µL)
DNA sample	32
Klenow buffer	5
dATP	10
Klenow exo (3' to 5' exo minus)	3
Total Volume	50

Table 4.3 Components used to add poly-A tail to DNA.

The DNA strands are then purified using the QIAquick Purification Kit, and the ends are repaired by mixing the DNA with the reaction mixture found in Table 4.2 and incubated.³¹ The DNA once again is purified using the QIAquick Purification Kit. The poly-A tail with fluorescent tag is added to the DNA fragments using the reaction mixture found in Table 4.3 and incubated.³² The sample is then run through a Qiagen MinElute Purification Column Kit.³³ Both Qiagen kit protocols can be found in Appendix B.

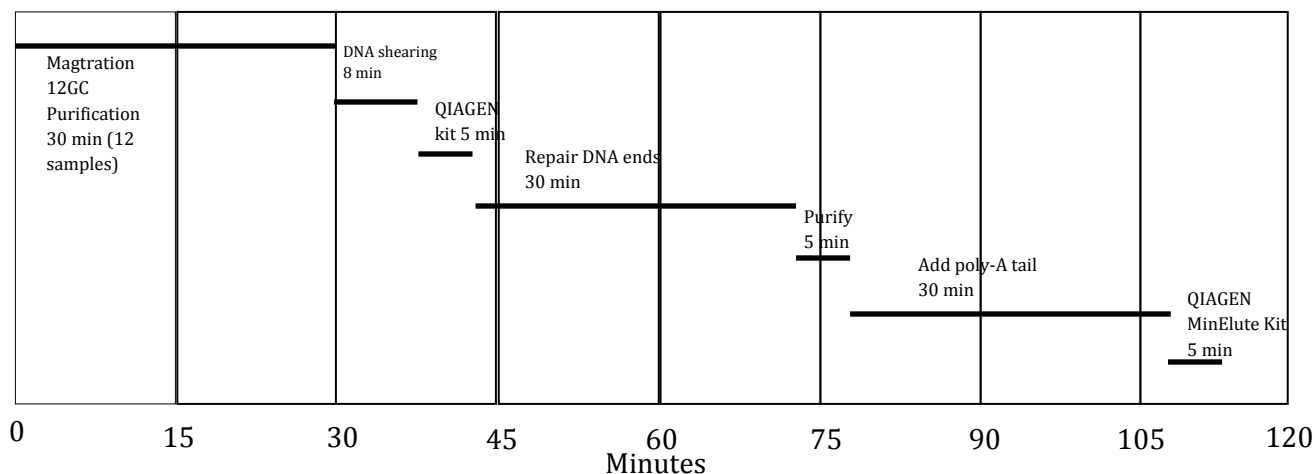


Figure 4.1 The Gantt chart is for the DNA sample preparation from the customer’s saliva to the 100 base pair DNA template fragments ready to be annealed to the flow cell surface. The total time of preparation takes about 2 hours.

The Gantt chart (Figure 4.1) shows the time each step in the sample preparation takes to complete. The Magtration purification can purify all twelve genomic DNA samples needed for daily operation at once. All these procedures are performed to prepare the samples sent to the sequencing station the following day. The times depicted in the Gantt chart affect the customer turnaround because a day is required for DNA preparation before it is sequenced. However, the DNA extraction, purification, and shearing steps do not impact the overall throughput, as they do not constitute the rate-limiting stage of the entire process, which is actually the *SynthSeq* chemistry and nucleotide detection (refer to Chapter 5).

LAYER-BY-LAYER ASSEMBLY AND TEMPLATE ANNEALING

In order to carry out the single molecule SBS process, the fragments of the DNA being sequenced must be bound to the glass slide. The surface must be able to bind to these fragments but prevent non-specific binding of free nucleotides and other molecules during the sequencing process to prevent imaging errors. To accomplish this, it is necessary to coat the surface of the slide with a bio-inert surface that can still bind to the DNA fragments. The solution to this problem requires layer by layer deposition of polymer onto the glass followed by immobilization of single stranded DNA strands onto that surface that serve to capture the template DNA.

The formation of a bio-inert surface on the glass slide is accomplished using layer-by-layer (LbL) deposition. LbL deposition is a method that assembles ultrathin films on a substrate by submerging the substrate, in this case a glass slide, back and forth between two dilute baths of oppositely charged polyelectrolytes. There are many advantages to LbL deposition. It is a relatively inexpensive procedure that allows for control over the structure, composition and thickness of the deposited film.

For this process, the two polymers being used to coat the glass slides for use in *SynthSeq*'s single molecule SBS approach are poly(acrylic acid) (PAA) and poly(acrylamide) (PAAm). The LbL process is done using an automated dipping machine, DS-50 SLIDE STAINER 115V. The glass slide is alternately dipped in dilute solutions of the PAA and PAAm with three rinses (2 min, 1 min, 1 min) in between each polymer application. Dipping the substrate into each of the polymer solutions once completes one bilayer on its surface. A polymer trilayer is used in this process, so three cycles of polymer addition are completed in order to create the bio-inert surface that will prevent free nucleotide absorption. The multilayer films are thermally cross-linked at 90°C under vacuum for eight hours.³⁴

After completing the bio-inert surface, the slide must be bio-functionalized such that the template DNA fragments to be sequenced can be immobilized on its surface. The sample preparation steps of the process include DNA purification, denaturation, fragmentation and

functionalization steps; the functionalization step is particularly important for the immobilization of these DNA strands onto a glass slide. Single stranded poly-A tails are added to the beginning of each of the sequences, allowing the fragments to anneal to a complementary poly-T oligonucleotide. These oligonucleotides are used to bio-functionalize the polymer surface of the glass slide.

Single stranded DNA oligonucleotides can be covalently bound to the carboxyl groups of a PAA surface by carbodiimide activation. The aminated ssDNA oligonucleotide is diluted in acid buffer (10 mM 2-(N-morpholino)ethansulfonic acid buffer (MES buffer, pH 6; Sigma-Aldrich) containing 10 mM MgCl₂ (Sigma-Aldrich), 5 mM 1-ethyl-3(3-dimethylaminopropyl)carbodiimide (EDC; Sigma- Aldrich), and 0.33 mM N-hydroxysulfosuccinimide (NHSS; Sigma- Aldrich)) and the glass substrate is submerged in the DNA solution. The concentration of the DNA in the solution and the incubation time (approximately 1 hour) is optimized to reach the desired coverage on the surface of the polymer-coated slide.³⁵

The modified template single stranded DNA fragments are annealed to the poly-T tails on the bio-functionalized, bio-inert surface of the glass slide using standard oligonucleotide annealing procedure. The poly-T oligonucleotide coated slide is immersed in a solution of the template DNA and heated to 90-95°C for 3-5 minutes and allowed to cool to room temperature. After slides are coated, functionalized and annealed to the template DNA they are stored at 4°C until ready to use.³⁶

SEQUENCING REACTION FLOW CELL

The microfluidic flow cell serves as the sequencing reaction vessel. Because its dimensions are on the micro scale, the reaction volume needed for the polymerization and cleaving are significantly reduced. Technological developments in soft lithography have provided a facile route

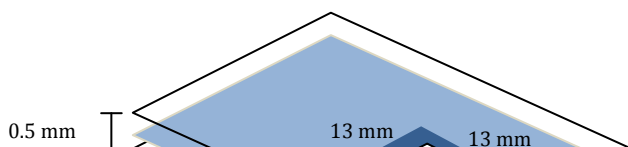


Figure 4.2 This drawing depicts the dimensions of the PDMS flow cell used in the sequencing. The top and bottom of the flow cell are composed of glass. The area of the reaction chamber is as shown. Not pictured are channels etched into the PDMS, two inlet channels and one outlet, connected to the reaction chamber.

to microfabrication of microfluidic channels in a flow cell. Soft lithography includes a family of techniques involving a soft polymeric replica (i.e. PDMS) cured to a hard master mold to create a soft stamp.³⁷ The polymer is compatible with the chemicals used in the process so the flow cell will not corrode or degrade while sequencing. The flow cells will be externally manufactured by CiDRA® Precision Services, LLC, using their SlipStream™ technology. They develop consumable, custom flow cells made as per design.³⁸ In this sequencing process, a standard microscope coverslips serve as the top and bottom of the device. The PDMS is cut according to the design specifications of a reaction chamber 13mm by 13mm by 0.5mm. Figure 4.2 displays a rough representation of the flow cell dimensions.

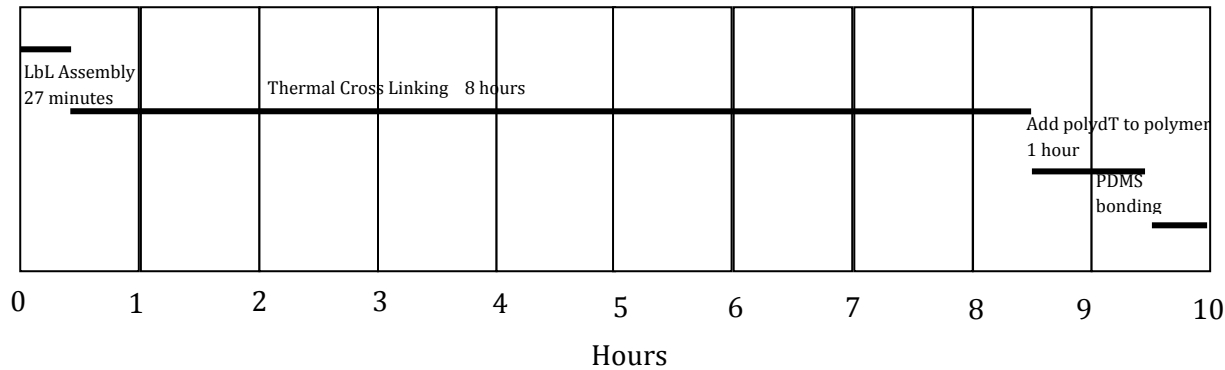


Figure 4.3 The time required to complete the flow cell preparation is shown in this Gantt chart.

Layer-by-layer assembly is performed on the glass slides, and the poly-T tails are immobilized on the surface. The sample’s DNA is annealed to the poly-T tails. Then, the PDMS stamp is bonded to the surface of the slides. The bonding is done using partial curing techniques, and requires 35 minutes. This procedure results in high average bond strength and gives the most flexibility in time, temperature, and cleanliness considerations.³⁹ Two inlet channels are fabricated into the cell, one for the nucleotide reaction solution and one for the cleave wash. The cell includes an outlet channel to allow continuously flushing of solutions out of the system. An outline of the entire flow cell preparation is depicted in Figure 4.3, and the steps occur over a 10 hour time period. Like the sample preparation, the flow cell preparation is performed a day prior to the actual genome sequencing and does not directly affect the process’s throughput, but does affect the turnaround.

SEQUENCING STATION SETUP

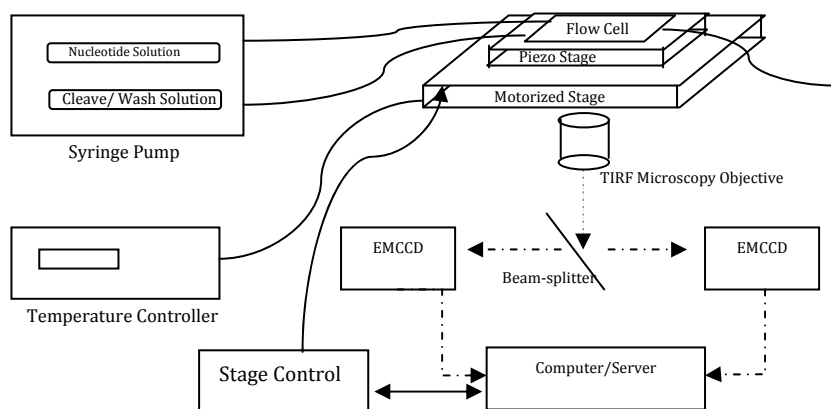


Figure 4.4 The station setup is a combination of instruments all integral to the success and throughput of *SynthSeq*'s process.

The instruments involved in the sequencing station make the throughput of the design possible. The station setup is depicted in Figure 4.4. An infusion only syringe pump with automated controls is

used to inject reaction solutions into the flow cell at specified flow rates and time increments. The reaction chamber is maintained at 40°C for all chemical reactions using a temperature controller connected directly to the stage. The flow cell is positioned directly on a microscan piezo controller with is compatible with the motorized X/Y microscopy stage controller. The stage accessories are placed on the stage of a TIRF microscope. For the design of the sequencing station, the ThorLabs MAX201 Motorized X/Y Stage along with the ThorLabs Microscan Piezo X/Y/Z Controller was chosen over an automated stage position with higher reproducibility. The MAX201 moves the stage to the next imaging area at 1µm repeatability. The microscan piezo will then shift the stage to the proper position after an initial imaging to align the templates to their original data points. The two picture adjustment system saves capital because the stage controller with high reproducibility of 0.1µm is estimated to cost about \$50,000, which is significantly higher than the roughly \$17,000 costs in the chosen system.

When the lasers excite the fluorophores, the emission wavelengths are reflected back through the objective and into the dichroic beam-splitter, which separates the light at 576nm. The

two ranges of wavelengths are then sent to their corresponding EMCCD cameras to detect the spatially dependent signals of each growing template strand. The data collected is then sent to the computer and servers for genome assembly.

CONCLUSIONS

Although these preparation stages are not major factors in achieving the throughput goal of the system, they are essential to the proper functioning of the design. The DNA extraction, purification, shearing, and poly-A addition allows the genomic DNA to anneal to the prepared flow cell reaction chamber surface. This surface is pre-treated with layer by layer assembly to prevent free fluorescent nucleotides from binding to the glass surface and distorting the emission of the incorporated nucleotides. The sequencing stations consist of various pieces of equipment necessary for the successful execution of the chemistry and detection sequencing cycle.

The preparation stages do affect the product turnaround. Both the sample and flow cell preparation must be performed the day prior to sequencing a particular customer's genome. One day is added to the turnaround as a result of these steps. After the completion of these preparations, the actual sequencing of the DNA fragments takes place. This sequencing protocol directly affects the throughput of the process.

5. CHEMISTRY AND DETECTION

The bottleneck in the design's throughput occurs during the sequencing chemistry and detection steps. The serial nature of base addition and optical detection to determine the next nucleotide on the DNA template creates the most time constraints in collecting genome data. For every nucleotide base addition, a rotation is performed until the 32 base read length is reached for the template strands immobilized on the flow cell surface. An outline of the cycle, Figure 5.1, displays the three main modules: nucleotide

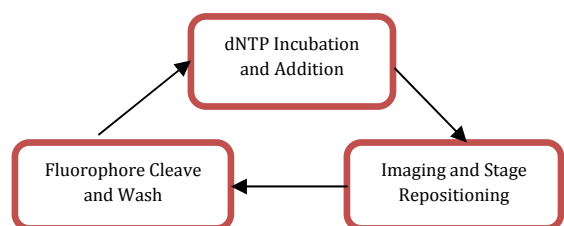


Figure 5.1 The chemistry and detection cycle starts with nucleotide addition, then continues with optical detection, and ends with removing the blocking group on the nucleotide.

addition, stage movements and imaging for base identification, and fluorophore-blocking group chemical cleave.

INITIAL IMAGING OF TEMPLATE POSITION

Before the nucleotide reaction solution is injected in the flow cell, the initial positions of the templates must be recorded. A fluorophore attached to the poly-A tail of the single stranded DNA sample is excited. Their emission is detected on the pixel grid of the EMCCD camera, and the signal is sent to the computer to record its position. The fluorophore is not cleaved off from the template, but instead allowed to photobleach. It is assumed that the fluorophore will completely photobleach before the first nucleotide is added to the system. The location of each template is important for accurately recording the sequence generated. The position of each detected fluorescent nucleotide will correspond spatially to a particular growing strand to generate its sequence. The initial positions will act as fiducials (i.e. landmarks) to facilitate moving the stage to the proper place when switching between imaging areas. Cross correlation functions are applied to adjust the imaging areas to their original positions based on the initial imaging of the flow cell. Once these positions are recorded, the aforementioned cycle can begin.

NUCLEOTIDE ADDITION

DNA STRUCTURE AND SYNTHESIS CATALYZED BY DNA POLYMERASES

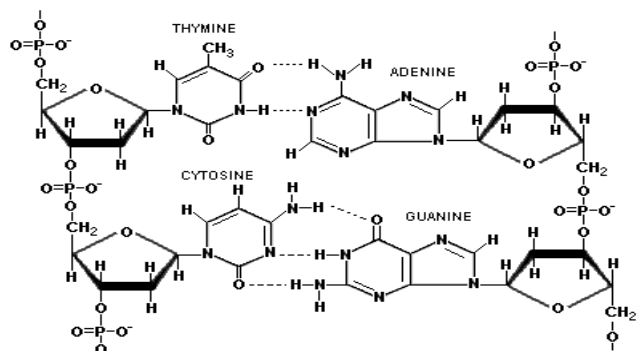


Figure 5.2 The structure of DNA has 4 nucleotide bases. The cytosine forms 3 hydrogen bonds with guanine, and thymine forms 2 hydrogen bonds with adenine. The negatively charged backbone consists of alternating phosphate and deoxyribose sugars.

Deoxyribonucleic acid (DNA) consists of two long polynucleotide chains composed of four different types of nucleotide subunits. One chain is referred to as a single-stranded

DNA, and when two chains interact through hydrogen bonding, they assume the double helix conformation. Nucleotides are composed of a five-carbon deoxyribose sugar with single phosphate group on the 5' carbon and a nitrogen-containing base. The four bases found in DNA are adenine (A), cytosine (C), guanine (G), and thymine (T). The two ring base purines are paired with the single ring pyrimidines; A always pairs with T and G always pairs with C.⁴⁰ The negatively charged backbone of DNA is comprised of alternating sugar and phosphate linkages called phosphodiester bonds. These bonds occur at the 3'-OH group on the deoxyribose and the phosphate group attached to the 5' carbon. The formation of these phosphodiester bonds is at the crux of nucleotide addition.

Enzymatic proteins called DNA polymerases catalyze the formation of the phosphodiester bond during *in vitro* DNA synthesis. They facilitate the formation of hydrogen bonds between the free deoxyribonucleoside triphosphates and their complementary bases on the template DNA strand. This synthesis proceeds in the 5' to 3' direction because each subsequent nucleotide requires a free 3'-OH group to form the phosphodiester bond. In order for the synthesis to commence, a short region of double stranded DNA with an exposed 3'-OH group must exist to act as a primer for the addition. Polymerases need a primer because they attach to a double stranded DNA before beginning base addition. The correct nucleotide has a greater affinity for the polymerase. Correct complementary base binding to the template offers the most energetically favorable pairing. With the proper

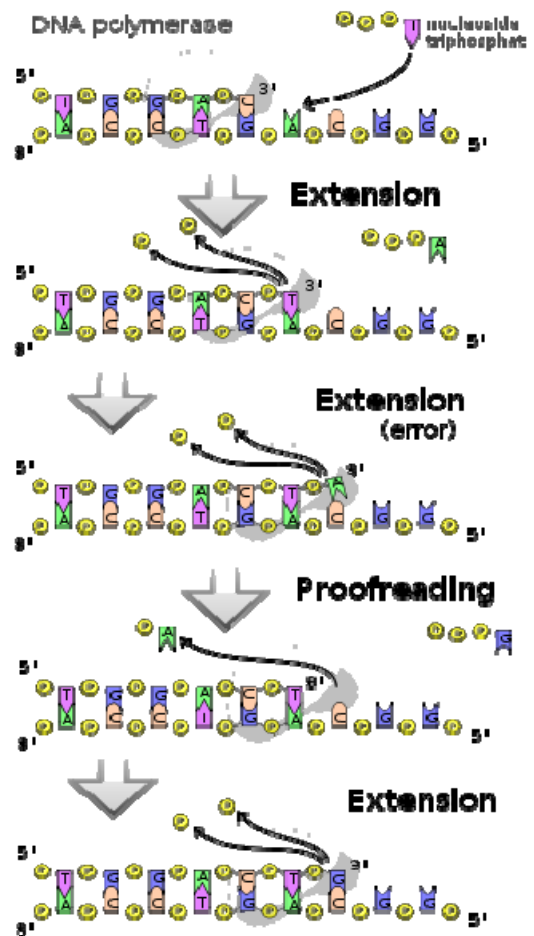


Figure 5.3 DNA polymerases follow specific steps when catalyzing DNA synthesis polymerization. Phi29 polymerase operates in this depicted fashion when incorporating the nucleotide analogs in this technology.

base pair, the enzyme tightens around the active site and undergoes a conformational change immediately before addition.

Some DNA polymerases exhibit high processivity, polymerizing many nucleotides to the 3' end of the chain before falling off the DNA template. Other DNA polymerases are distributive and incorporate just one nucleotide and then fall off the DNA template. When a polymerase incorrectly adds a base pair, its proofreading mechanism is utilized. The removal of the wrong base occurs from 3' to 5' exonuclease properties of certain polymerases that remove nucleotides from the 3' end. This reduces errors introduced in DNA synthesis and genome replication.

PHI29 POLYMERASE

Multiple DNA polymerases are found in human and bacterial cells indicating specialized roles for the different enzymes in various aspects of DNA replication. However, they have highly conserved structure among the species and very little variation in their catalytic mechanisms. Polymerases are grouped into families dependent on their similarities in protein structure and function. Many high fidelity, eukaryotic and bacterial based polymerases are found in Family B. These polymerases, in addition to accurate nucleotide addition, intrinsically have exonuclease proofreading activity.⁴¹ Phi29 DNA polymerase, a member of the B-type family, was chosen for the process because of its high fidelity, incorporation speed, and good processivity.

Bacteriophage phi29 DNA polymerase is a small (ca. 68kDa), replicative polymerase from the *Bacillus subtilis* phage phi29. It assists in DNA polymerization reactions and terminal protein deoxynucleotidylatation. Phi29 polymerase also has degradative activities, pyrophosphorolysis and 3'-5' exonuclease activity. This polymerase exhibits processivity of greater than 70 kilobases with uninterrupted synthesis and strand displacement ability. Significant amino acid conservation spans across many of the eukaryotic polymerases including phi29 to form a polymerization active site in

several groups of nucleic acid synthesizing enzymes. The synthesis originates from the C-terminus domain of the enzyme. The binding of phi29 to DNA primer-template structures is enhanced by the presence of metal ions that are known to activate DNA polymerization.⁴² The polymerase adequately recognizes chain-terminating agents, a property essential to the sequencing process.

The polymerase exhibits a low error rate of 10^{-5} , which increases the probability of a nucleotide being successfully incorporated into the template strand. This error rate corresponds to natural nucleotides with no additional groups attached.⁴¹ For certain dNTP analogs, they do not incorporate due to steric or electrostatic interactions that interfere with the polymerase active site. In the *SynthSeq* technology, dye-dye interactions do not come into play because the dye is cleaved off before the next addition is made. With these analogs, a threshold number of base incorporations exist and the threshold increases with increasing length of the linker between the base and the dye. Incorporations up to 40 bases have been observed.⁴³ This restriction is one of the factors contributing to a shorter read length in the *SynthSeq* technology.

Despite these polymerization concerns when utilizing tagged dNTPs, the position of the fluorophore on the nitrogenous base instead of the 3'-OH and the cleavable nature of the dye makes these incorporation issues negligible. After fluorescence detection and cleavage, the blocker group is not involved in the subsequent reaction. The nucleotide is reverted back to its natural state, and then base addition may continue without affecting the activity of the polymerase.⁴⁴

PROBABILITY POLYMERASE BINDS AND REMAINS ON TEMPLATE

If a DNA polymerase is not bound to a template strand, that fragment of DNA cannot be sequenced. The design, however, assumes that every template strand has a polymerase bound initially at the polyT/polyA tail at its base. The binding coefficients typical to DNA polymerases like phi29 are on the nanomolar scale. A K_d (Equation 5.1) on this small magnitude signifies a very tight

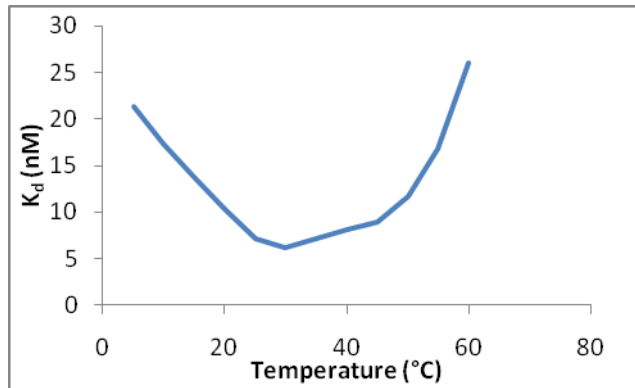


Figure 5.4 The plot of the binding coefficient as a function of temperature for Taq polymerase. At 40°C, the K_d value is 8.1nM.

bond between the polymerase and DNA because the rate constant of binding is significantly larger than that of dissociation.

$$K_d = \frac{k_{off}}{k_{on}} = \frac{[template][phi29]}{[template-phi29]} \quad \text{Equation 5.1}$$

The probability of finding the double stranded DNA template occupied by a polymerase is represented by:

$$P_{bound} = \frac{[template-phi29]}{[template] + [template-phi29]} \quad \text{Equation 5.2}^{45}$$

From algebraic manipulation of Equation 5.1, the ratio of free polymerase to bound polymerase:

$$\frac{[phi29]}{[template-phi29]} = \frac{K_d}{[template]}$$

Equation 5.3

The binding constant used for the phi29 polymerase is estimated to be comparable to that of Taq polymerase at 40°C (Figure 5.4), which is 8.1nM.⁴⁶ The DNA concentration is determined by divided the optimized area density (1210000) by the volume of one pixel observation volume (125μm³); the concentration was calculated to be 9680 templates/ μm³. Using these values, the ratio of free polymerases to bound polymerases is 8.37*10⁻⁷. Since the ratio is very small, the

majority of the polymerases are bound to the double stranded DNA templates, and the amount in free solution is negligible.

Equations 5.2 and 5.3 were combined to establish the concentration of phi29 polymerase needed to attain a 99% probability that the polymerase binds to the template. The phi29 polymerase must be incubated into the system at 0.78µM to achieve this probability goal. The polymerase at this concentration is incubated for 15 minutes with the immobilized template strands. These actions are done to ensure polymerase attaches to all the templates.

Once the polymerase is bound to the template, a probability of detachment exists. As the read length of the DNA fragments increases, the polymerase's ability to remain on the DNA is reduced. One reason the phi29 polymerase was chosen is because of its extremely high processivity, travelling up to 70,000 bases before dissociating from the strand.⁴⁷ However, this processivity was determined without interruption to the polymerization process. Since *SynthSeq* uses reversible terminator nucleotides and the reaction is continually suspended, the amount of time the polymerase remains on the strand must be evaluated opposed to the number of base pairs. Since the polymerase travels at a rate of 4.7 bases/s (see Polymerization Rate) and remains attached to the DNA for 70,000 bases, the amount of time it is expected to remain on the strand is 4.14 hours. The amount of time each flow cell undergoes the chemistry and detection cycle is 4 hours. The constant suspension of the nucleotide addition does not affect the enzyme's high processivity in the context of *SynthSeq* sequencing.

The occurrence of the release event can be modeled by the exponential cumulative distribution function because it describes the times between events in a process in which these events occur continuously such as nucleotide addition using the polymerase. Since the number of events occurring is greater than 0, the probability that the polymerase is unattached from the strand is represented by the following equation:

$$P(X \leq x) = F(x, \lambda) = 1 - e^{-\lambda x} \quad \text{Equation 5.4}^{48}$$

Where X is the variable denoting number of bases traveled, x is the read length, and k is the rate parameter. The read length in this process is 32 base pairs, and the rate parameter is $1/70,000$ base pairs⁻¹. With the parameters are considered, the calculated error rate is $4.57 \cdot 10^{-4}$, and the probability of the polymerase remaining on the strand after the read length in the process is 99.95%.

REVERSIBLY TERMINATING TAGGED NUCLEOTIDES

Many SBS techniques struggle with the problem of counting homopolymers – DNA sequences that repeat the same base one or more times. One way to overcome this issue is to use nucleotides that are incorporated onto a template DNA strand but block addition of any other nucleotides, so that they could progress through these homopolymeric regions one base pair at a time.⁴⁹ However, in order to continue the process of SBS, the ability to add nucleotides after blocking the extension would have to be restored. One of the most unique features of *SynthSeq*'s method of single molecule SBS is that each imaging step is carried out in between nucleotide addition steps; this is accomplished using modified, tagged nucleotides that reversibly terminate the extension reaction of DNA synthesis.

The ideal nucleotide for this process would succeed in both of the functional modifications – reversible termination and tagged for identity of the base – with only one physical modification. Little information on the structure and synthesis of such a modified nucleotide exists in the public domain; however, private companies have reported data suggesting that they have synthesized one. Helicos Biotechnologies possesses proprietary reversible terminator nucleotides that contain only one modification. These maintain a hydroxyl group at the 3' position and have a base modified with a propargylamine linked via a cleavable linker to a fluorescent dye tethered to an inhibitor. Helicos reports efficient incorporation, blocking and regeneration of extension using these nucleotides⁵⁰.

The nucleotides used by *SynthSeq* have two major modifications that make it possible for this process to be carried out. First, the nucleotides contain a tag on the base on the 1' position of the five-carbon sugar via a cleavable linker. The identity of the tag depends on the identity of the base; each has an associated fluorophore that emits light at a specified wavelength. Second, the nucleotides have a reversible terminator group blocking the 3'-OH, which serves to terminate DNA synthesis.⁵¹

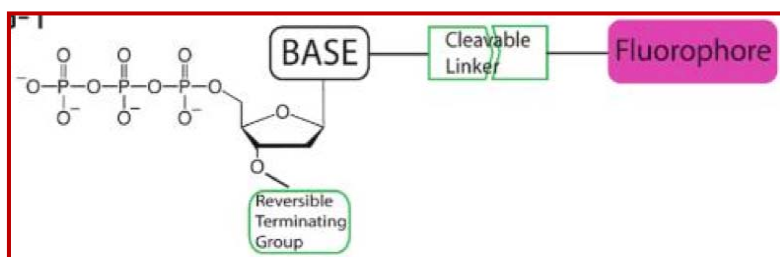


Figure 5.5 *SynthSeq*'s modified reversible terminator nucleotides. Modifications label the nucleotide according to which base it contains and block the 3'-OH group to terminate DNA extension.⁴⁴

The nucleotides are tagged on the base of the nucleotide by a fluorophore that corresponds to that base. The four bases and their corresponding fluorophores are: Guanine/bodipy-650, Cytosine/bodipy-FL-510, Adenine/ROX, and Tyrosine/R6G. The fluorophores are attached via a cleavable allyl linker that enables the release of the fluorophore after imaging. They are attached at the 5-position of the pyrimidines (T and C) and the 7- positions of the purines (G and A).⁵²

The 3'-OH group of these nucleotides have been capped with a small chemically reversible moiety. It is of crucial importance that this modification not be too large or bulky, so that these nucleotides will still be suitable substrates for DNA polymerase, and that it can be

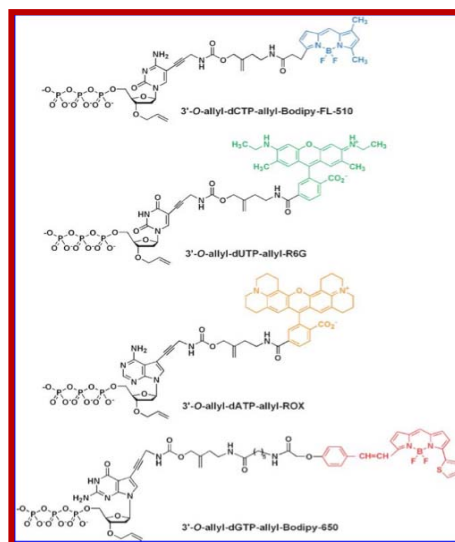


Figure 5.6 *SynthSeq*'s four modified nucleotides: 32-O-Allyl-dCTP-allyl-bodipy-FL-510, 32-O-allyl-dUTP-allyl-R6G, 32-O-allyl-dATP-allyl-, and 32-O-allyl-dGTP-allyl-bodipy-650⁴⁴

converted to a hydroxyl group in order to enable further extension reactions after imaging. Replacing the hydroxyl group with a small allyl group serves to block further nucleotide addition while preserving DNA polymerase's ability to incorporate the nucleotide onto a DNA template.

After nucleotide addition and imaging steps, two things must be done in order to enable the cycle to be repeated: the fluorescent label must be cleaved from the base and the 3' allyl blocking group must be converted to a hydroxyl group. These two actions are both accomplished in one single deallylation step using a Palladium catalyzed reaction with Na_2PdCl_4 and $\text{P}(\text{PhSo}_3\text{Na})_3$ in Thermopol reaction buffer.

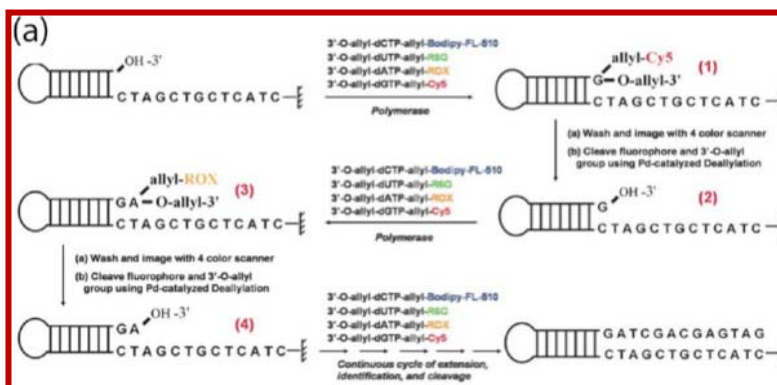


Figure 5.7 Reversible terminator nucleotides are added to the template strand and imaged. Then, the fluorophore is cleaved and the 3'-OH group regenerated allowing for subsequent base addition and imaging steps.⁴⁴

POLYMERIZATION RATE

The phi29 polymerase will serve in this technology as the enzyme catalyzing nucleotide addition to the single stranded DNA templates immobilized in the flow cell. The polymerase requires 0.1 to 10 μM of dNTP concentration to maintain fast, accurate, and processive synthesis. It also shows no preference for labeled or unlabeled nucleotides and is not hindered by the additional tag moieties of the reversible terminator labeled nucleotides used in this design. Once the fluorophore is cleaved, it diffuses quickly away from the DNA strand and gives the next

complementary base the ability to bind and interact with the polymerase. The tag will typically diffuse away from the template in about 2 to 10 μs after cleavage.⁴⁵ The cleave process provides enough time to wash through the cell to ensure the tag is removed and will not interfere with the next nucleotide addition.

Enzymes like phi29 polymerase are among the most selective and powerful catalysts known. An understanding of its rate of chemical reaction and how these rates change with varying conditions such as temperature and concentration of substrate is essential to develop an optimal nucleotide incubation time for addition. This polymerase with the help of a manganese metal ion activator has only one substrate, the free nucleotide analog. The polymerase has an energetic affinity to the proper base to complement the template strand, and the polymerase and nucleotide bind to form an enzyme-substrate complex. Then, the chemical reaction occurs and binds the free nucleotide to its complementary base. A common equation to model this enzymatic reaction is the Michaelis-Menten equation:

$$V = \frac{k_{cat}[E_0][S]}{K_M + [S]} \quad \text{Equation}$$

5.5

Where V is the velocity or rate of the reaction; k_{cat} is the enzyme complex dissociation rate constant; $[E_0]$ is the initial enzyme concentration; $[S]$ is the substrate concentration; K_M is the Michaelis-Menten constant.³⁹

This chemical system deviates from the traditional enzyme kinetics. The polymerase is assumed to be bound to every DNA fragment due to high concentrations of enzyme incubated with the templates for 15 minutes. The enzyme concentration does not come into play in the rate equation because it's always available to bind to the free nucleotide substrates. Its concentration remains constant. Also, the chemistry supposes the formation of the enzyme-substrate is rapid. The limiting reaction is the dissociation of the complex into the polymerase and bound nucleotide.

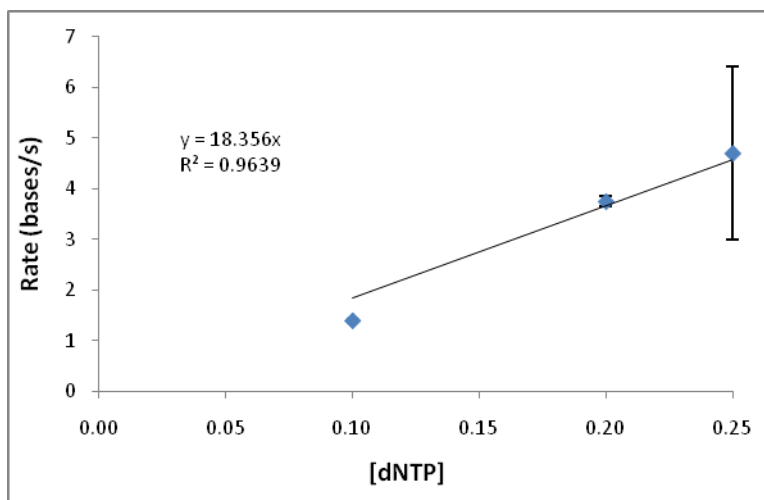
Therefore, a first order rate law can be proposed for the nucleotide addition reaction using phi29 polymerase.

$$r = kC_{\text{dNTP}} \quad \text{Equation 5.6}$$

Where r is the rate of reaction; k is the rate constant; C_{dNTP} is the concentration of free nucleotides in the reaction solution.

This rate law is supported by kinetic data collected by Eid et al. The experimentally determined phi29 DNA synthesis rate for a single polymerase is 4.7 ± 1.7 bases/s at 30°C. The high standard deviation is due to the existence of two kinetic states independent of dNTP, template, or other such experimental variations. The rate can be 2base/s or 4base/s depending on its state, which explains possible rate fluctuations at this higher concentration. At higher temperatures, more bases can be incorporated in a finite period of time. The temperature of the cell cannot exceed 65°C in the flow cell because the polymerase will denature and cease to function properly.⁵³ This process runs all of the sequencing chemistry at 40°C, so the data used is a conservative estimate of the rate for this system. Experiments determined the rate of single molecule DNA synthesis increased as the concentration of nucleotide reagents increased (Figure 5.8)⁵⁴.

Figure 5.8 The phi29 polymerase follows a first order rate law with a rate constant of 18.356s^{-1} . Along with the data point mentioned in the text, at 100nM , the rate was 1.4 ± 0.03 base/s and at 200nM , the rate was 3.75 ± 0.10 base/s. The determination of this rate equation allows the rate to be calculated from any free dNTP concentration.



The linear relationship between the rate and various nucleotide concentrations confirms the first order rate law of the reaction. For *SynthSeq*, the addition reaction is run at an individual dNTP concentration of $0.074\mu\text{M}$. The $0.074\mu\text{M}$ corresponds to the maximum concentration where an average of one fluorophore is in the detection volume of $0.0075\mu\text{m}^3$ at one period of time.

$$7.5 \times 10^{-21} \text{ m}^3 \times 4 \text{ molecules} = 7.4 \times 10^{-8} \frac{\text{mol}}{\text{L}} \times 1000 \frac{\text{L}}{\text{m}^3} = 6.022 \times 10^{-18} \frac{\text{molecules}}{\text{mol}} \approx 1 \text{ molecule}$$

Equation 5.7

The detection volume describes the volume above the flow cell surface where the fluorophores can be excited and their emissions' can be detected. The detection volume will be more clearly defined in the following section, Optical Detection of Single Molecules. The purpose of this concentration is to reduce the background fluorescence during single molecule detection. Any lower concentrations will further reduce background, but the phi29 polymerase rate will be negatively affected. This concentration of 0.074μM correlates to a rate of 0.738bases/s or approximately 1s/base. To maximize the success of the base addition, the reaction solution is incubated for 3s.

PROBABILITY OF NUCLEOTIDE INCORPORATION

Nucleotide addition to the fragmented genomic DNA template is integral to the sequencing process. The phi29 polymerase's fidelity rate is high, and its error rate is low on the order of 10⁻⁵. Consequently, the probability of the nucleotide successfully being incorporated into the template is dependent on the initial nucleotide concentration, the kinetics of the polymerization reaction, and the set incubation time before imaging takes place.

The polymerization follows a first order rate law (Equation 5.8). The differential equation that represents the change in dNTP concentration over time is as shown below:

$$\frac{dC_{dNTP}}{dt} = -kC_{dNTP} \quad \text{Equation 5.8}$$

Where k is the rate constant of the reaction in s⁻¹. When this ordinary differential equation is solved by separation of variables over time, t, the concentration profile equation in μM is formed:

$$C_{dNTP} = C_{dNTP_0} e^{-kt} \quad \text{Equation 5.9}$$

Where C_{dNTP_0} is the initial nucleotide concentration in μM. For this reaction, the initial dNTP concentration used is 0.074μM and the rate constant determined in Figure 5.8 is 18.356s⁻¹, which makes the concentration specific to this system:

$$C_{dNTP} = 0.074 e^{-18.356t} \quad \text{Equation 5.10}$$

Typically, a reaction is adequately complete in the time when about $1/e$ of its initial concentration is reacted. In this system, this occurs at a concentration of $0.027\mu\text{M}$ and a time of $1/k$ or 0.054s . To reduce the error rate of the nucleotide addition, the dNTP reaction solution is incubated in the flow cell for 3s . This incubation time is 60-fold more than the expected incubation time to add a base. Therefore, the error rate is represented by e^{-60} or 8.76×10^{-27} , and the probability of the nucleotide being successfully polymerized is 99.99%.

REACTION CONDITIONS

The reaction conditions of the nucleotide addition are specified for the process. The reaction is run at 40°C throughout the entire sequencing cycle. This temperature was chosen to increase enzymatic kinetics as well as run the deallylation chemical cleave reaction for the reversible terminator nucleotides at an adequate temperature. The temperature of the system cannot exceed 65°C because the enzyme will denature. The temperature of the stage and flow cell is maintained by a temperature controller with a set point of 40°C .

Because of the small scale dimensions of the reaction chamber in the flow cell, convective fluid flow of the injected solutions is negligible compared to the diffusion. The free dNTPs diffuse into the detection volume to bind to the polymerase active site and undergo addition.

OPTICAL DETECTION OF SINGLE MOLECULES

SCANNER PROBLEM

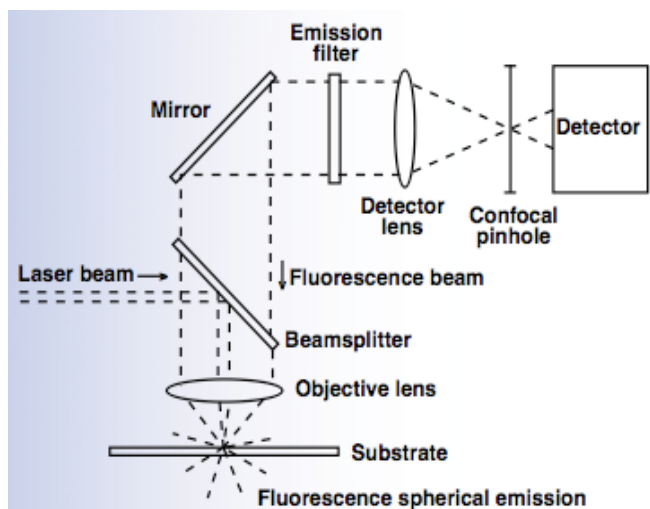


Figure 5.9 The detection mechanism of the ScanArray ExpressHT. (Image from PerkinElmer Life Sciences, Inc.)

Our initial plan was to simplify the imaging process by making use of a scanning instrument such as PerkinElmer's ScanArray ExpressHT. Such instruments are commonly used in similar applications requiring extensive spatial multiplexing. The flow cell would be inserted into the confocal laser scanner, and the DNA strands would be scanned sequentially after each addition step.

Despite meeting the 5µm pixel resolution requirement of the system and supporting the simultaneous detection of four wavelengths, the scanner was simply unable to meet the throughput demands associated with full genome sequencing. PerkinElmer reports the scan speed of its instrument to be less than 2.5 minutes for a 20mm x 30mm area at 10 micron resolution.⁵⁵ This is very fast. However, it does not turn out to be fast enough to sequence the roughly half a billion DNA strands that are necessary to construct an entire human genome at six-fold coverage in a reasonable time frame. The following equation demonstrates the failure of the scanner to meet throughput demands.

$$\begin{aligned}
 & (2.5 \text{ minute scan} / \text{quoted area-cycle}) * (0.28 \text{ quoted areas} / \text{flow cell area}) * (400 \text{ } 0.5\mu\text{m} \\
 & \text{resolution pixels} / 10\mu\text{m resolution pixel}) * (2 \text{ flow cells} / \text{genome}) * (32 \text{ cycles} / \text{flow cell}) * \text{Equation 5.11} \\
 & (1\text{hr} / 60\text{min}) / (0.32 \text{ pixel efficiency}) = \underline{\underline{1000 \text{ hours} / \text{genome}}}
 \end{aligned}$$

Note that this staggering figure represents just the required scanning time for each genome. As demonstrated by the above equation, the confocal laser scanner is not a viable option for addressing *SynthSeq*'s considerable throughput requirements.

TOTAL INTERNAL REFLECTION MICROSCOPY

Total Internal Reflection Fluorescence Microscopy (TIRFM) is an optical technique used to observe single molecule fluorescence. The method reduces the background when detecting the addition of a single base pair. This optical phenomenon exploits events occurring at surfaces. Light strikes an interface between two optical media of different refractive indices.

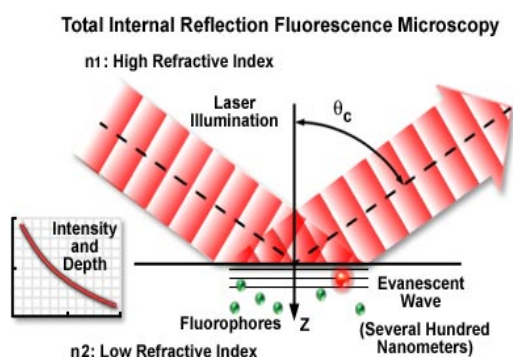


Figure 5.10 For proper conditions, light illuminated on a surface will totally reflect from the surface at a critical angle. An evanescent wave propagates into the second medium at exponentially decaying strength. This wave can excite fluorophores.⁵⁸

Distance (Nanometers)	Relative Intensity
1	0.99
10	0.92
100	0.43
1000	0.0002

Table 5.11 The data shows the intensity of the evanescent wave decreases as the wave travels further away from the surface. This intensity decays exponentially. By a depth of 100nm, more than half the strength of the wave is lost. At the 30nm depth chosen for this technology, the evanescent wave maintains about 75% of its intensity.⁵⁸

A refractive index is the factor by which the velocity of propagation in a medium is decreased relative to the velocity of light in vacuum.⁵⁶ If the light is incident at an angle greater than the critical angle, then it undergoes total reflection. Beyond the angle of total reflection, the electromagnetic field of incoming light continues to extend into the second medium. The strength of

this evanescent wave decreases exponentially from the surface, and the effects of the wave only extend about 100nm.

In the *SynthSeq* process, the observational depth for fluorophore detection is assumed to be 30nm because the wave maintains about half of its strength at that point. This estimation may provide more background fluorescence than expected since fluorophores, although less likely, can be excited and fluoresce at further distances away from the surface than 30nm. In order to confirm the DNA template strand's first 38 nucleotides do not exceed the detection depth, the root mean square distance of the ssDNA from the surface is calculated using Equation 5.12.

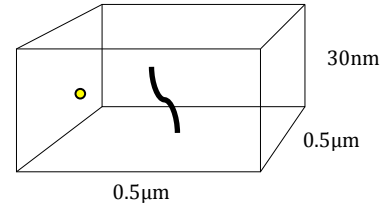


Figure 5.11 The detection volume for each incorporated nucleotide in the template strand is dependent on the evanescent wave intensity depth.

$$\sqrt{R^2} = l_n^{1/2} + l_{\text{contour}}^{1/2}$$

Equation 5.12

Where l_n is the Kuhn length of ssDNA (4nm)⁵⁷ and l_{contour} is the contour length of the strand defined as $n_{\text{bases}} / 1.5 \text{ nm}$. For a ssDNA of 38 base pairs, the root mean square distance of the end of the strand to the surface is only 10nm, which is well within the observation depth of TIRFM.

The conditions for total reflection and the critical angle follow Snell's Law when the angle of refraction is 90°.

$$\theta(c) = \sin^{-1}\left(\frac{n_2}{n_1}\right)$$

Equation 5.13

Where $\theta(c)$ is the critical angle, n_2 is the refractive index of the specimen analyzed, and n_1 is the refractive index of the objective. The refractive index of the first medium must be larger than the second medium's index.

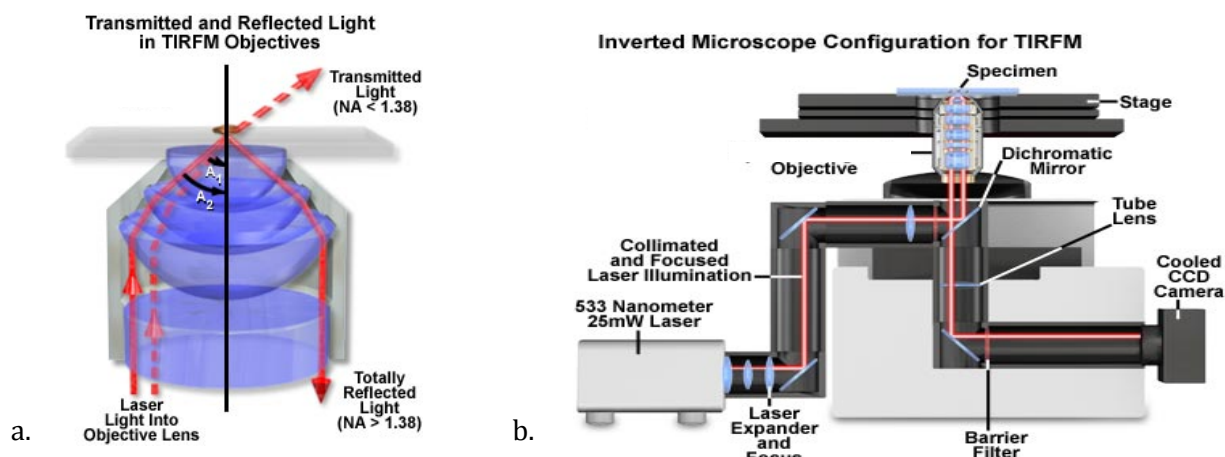


Figure 5.12 (a) The laser enters the objective at an angle where TIRFM occurs, and the light is totally reflected back into the objective lens. After it goes back into the lens, the light is sent to a dichromatic beam-splitter and then EMCCD cameras. (b) The configuration of a typical inverted TIRM is depicted.⁵⁸

This technique is utilized in this design for several important reasons. TIRFM provides a limited depth for fluorophore excitation near the surface of the flow cell. Only the fluorophores attached to the incorporated nucleotide base remain close enough to the surface to be excited by the evanescent wave. Therefore, the background fluorescence is reduced. Less tagged nucleotides in the reaction solution will emit light. TIRFM offers high contrast images with a good signal to noise ratio to provide a sharper detection of single molecules.⁵⁸

SynthSeq uses an Olympus XI81 inverted microscope with TIRFM capabilities. The microscope is equipped with 4 line TIRF lasers capable of exciting all four reversible terminator nucleotides simultaneously.⁵⁹ The instrument uses an APO 60x objective with a 1.49 numerical aperture (NA). The NA improves TIRF capabilities. Barrier filters are also put in place to prevent excitation light from interfering with the detection of the specific wavelength emissions of the four nucleotides. A 0.5 magnification changer is also put in place to adjust, along with the objective, the pixel length of the EMCCD from its standard 13 μm to 0.5 μm . The pixel size change reduces the imaging area and detection volume to help the throughput of the process. If the pixel size remained larger, the concentration of dNTP needed to maintain an average of one tagged nucleotide in the detection volume at any point in time would be too low for the polymerase to function properly.

The X/Y stage positioner and microscan piezo controller are compatible with the stage of this microscope model.

EMCCD CAMERA

Electron Multiplying Charged Coupled Device cameras provide the high resolution and single molecule detection capabilities required to execute high throughput genomic sequencing. An EMCCD is a quantitative digital camera capable of single photon events while maintaining high quantum efficiency because of its unique electron multiplying structure built into its sensor. An electron multiplying structure is built into the chip to allow weak signals (e.g. emission from a single fluorophore) to be multiplied before any read out noise is incorporated into the output. This amplification makes the read noise negligible. Since the chip does not require an image intensifier, the approximately the full quantum efficiency of the silicon sensor can be utilized. The EMCCD sensors use a shift and gain register. The gain can be increased and tuned so extremely weak signals can be detected above the read noise of the camera at any readout speed. This property allows the technology to surpass tradition CCD cameras that experience a larger read out noise detection limits with higher pixel readout speeds. ⁶⁰

This process uses Andor's megapixel iXon^{EM}+888 back-illuminated EMCCD to detect single molecule events in the sequencing by synthesis process. For the station set up and the execution of fluorophore identification, two cameras are used at each sequencing station. This specific model is suggested because it has a fast frame rate, on the order of 10^6 pixels, and high quantum efficiency.

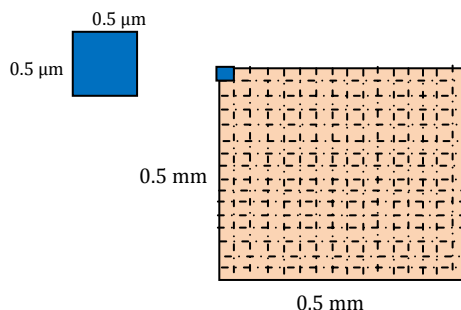


Figure 5.13 A 1024 by 1024 pixel imaging area contains pixels with dimensions of $0.5\mu\text{m}$ by $0.5\mu\text{m}$.

The frame rate of the camera influences throughput. This camera operates at a full frame rate of 8.9 frames/s (i.e. 0.111 s/frame). This camera parameter determines the time needed for the camera to detect the excited fluorophores' emitted photons and transfer them into a signal. The speed determines time required on each imaging area before the stage can move to the next one. Since the reaction is not detected in real time, the frame rate does not have to keep up with the kinetics of the reaction and does not have to be significantly fast. The camera images the fluorophores while the synthesis is suspended. Therefore, the 9 frames/s is sufficient for this process and does not present any major time limitations in the chemistry and detection cycle.

Although the frame speed of the camera impacts the throughput, the imaging area and number of pixels are more important for the *SynthSeq* process. The megapixel chip is essential to the throughput. The camera has a 1024x1024 active pixel array where each pixel is 13 μ m by 13 μ m in size. As mentioned previously, a 60x objective and 0.5 magnification changer in the microscope changes the pixel size to 0.5 μ m by 0.5 μ m. The imaging area due to the size of the pixel array makes one imaging area approximately 0.5mm by 0.5mm. The template strand sequencing is spatially dependent to where the DNA is immobilized on the flow cell surface. The pixel can detect the addition of bases to the DNA template residing solely in that pixel. Therefore, increasing the number of pixels in each imaging area allows more templates to be sequenced at once. However, the efficiency factor determined through the simulation described in Chapter 3 greatly limits the throughput of a single imaging area despite the capacity for 10⁶ templates to be sequenced simultaneously.

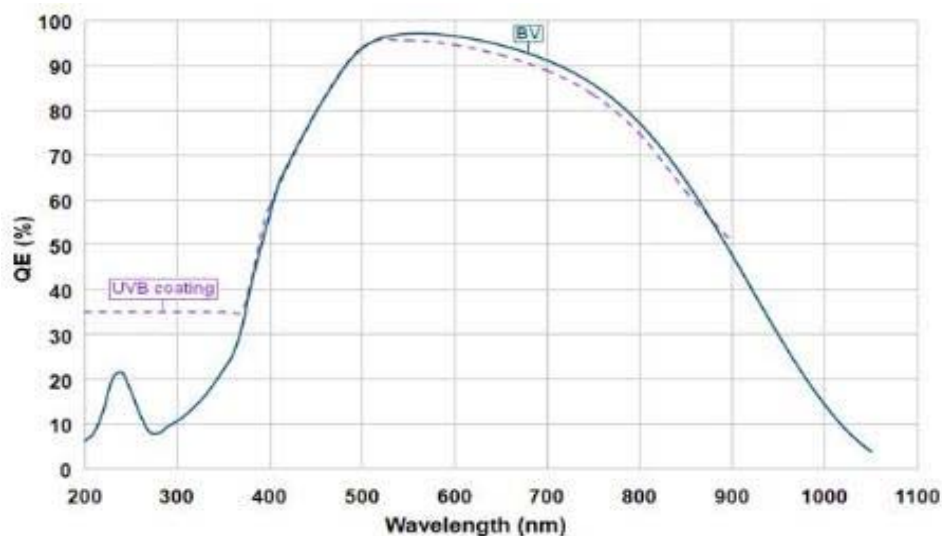


Figure 5.14 The quantum efficiency for an EMCCD camera remains over 90% for the 500nm to 700nm wavelength range, which is applicable to the *SynthSeq* process.⁶¹

The detection of a fluorescent molecule relies on a high quantum efficiency (QE) and favorable signal to noise ratio. The QE is defined as the percentage of photons that are actually detected and then transmitted as electrons by the camera. Higher QE generally produces a better quality reading of the fluorophores because it has a higher signal. The quantum efficiency is a function of the emission wavelength of the light being detected. The iXon^{EM} +888 has quantum efficiencies greater than 95% for the emission wavelengths used in the process (Figure 5.14). Back-illuminated detectors have a 4-fold greater QE than front-illuminated cameras of the same pixel size. A high quantum efficiency promotes a high signal to noise ratio (SNR). SNR is a measure of how much signal is ruined by noise. The higher this ratio becomes the less prominent the noise is in an EMCCD. The overall noise is a combination of read noise, shot noise, background noise, and dark current. The electron multiplying gain in an EMCCD are sufficient to effectively eliminate read noise. Since the cameras are thermoelectrically cooled to -95°C, the dark current is also eliminated from the total noise.⁶¹ Therefore, when using an EMCCD, only the shot noise and background noise needs to be taken into account. The evaluation of signal to noise and how it pertains to the fluorescence detection in the *SynthSeq* process is discussed in the section discussing the probability of misidentifying a nucleotide.

DICHROIC BEAM-SPLITTER FOR FLUOROPHORE RESOLUTION

As a result of the scanner fiasco, *SynthSeq* needed to find an alternate imaging solution for its technology. A dichroic beam-splitter is another necessary component of our imaging setup. We plan to resolve the four different wavelengths emitted by the fluorophores using two CCD cameras and the dichroic beam-splitter, which will allow us to determine the identity of the captured light based on differences in intensity ratios between the two cameras. The four fluorophores transmit at different wavelengths, so a certain wavelength is chosen that will divide them into pairs. The emission wavelengths of the fluorophores are 510nm, 550nm, 602nm, and 650nm.⁶² Consequently, an intermediate cut-off wavelength

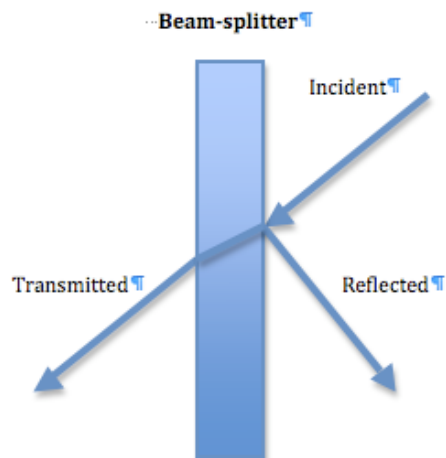


Figure 5.15 The dichroic beam-splitter, showing the incident, reflected and transmitted light pathways

(the incident wavelength at which 50% of light is transmitted, and 50% is reflected) of 576nm has been chosen for the dichroic beam-splitter. Basically, the beam-splitter will reflect 100% of the 510nm light into the “green” camera, and only 70% of the 550nm light. At the opposite end of the spectrum, 100% of the 650nm light and proportionally less of the 602nm light (again, perhaps 70%, enough to differentiate the captured light based on intensity) will be sent to the “red” CCD camera. When split optimally, the aggregate intensity of each wavelength of light across both camera should be comparable. The different fluorophore signals are then resolved based on their respective “green/red” ratios, which is based on the relative intensity of the light hitting each camera.

Thus, each fluorophore will be resolved based on its ratio of green light to red light, as determined by the pair of CCD cameras (see Figure 5.16).

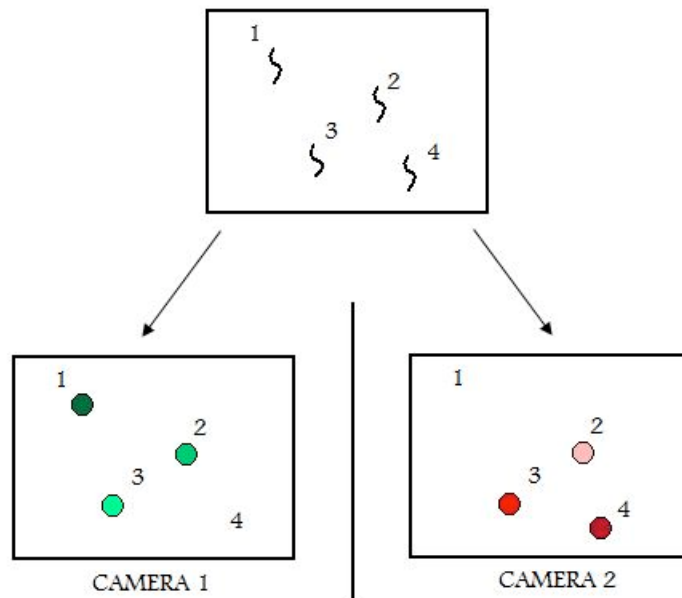


Figure 5.16 Resolution of the four fluorophores using two EMCCD cameras, one for green light and the other for detection of red light

In the next section we will provide an error analysis of our optics system, which demonstrates the feasibility of the technique described above.

Unfortunately, finding an appropriate beam-splitter that would allow our desired resolution of the fluorophores was not as easy as we were expecting. Many optics companies are striving to achieve higher resolution rather than a more diffuse splitter like the one our design requires. When we investigated the R/T curves for available splitters, we found that the reflectance would allow for a satisfactory split in intensity between the first pair of fluorophores, but then the next two wavelengths could not be differentiated because the transmittance curve had already peaked, resulting in the full transmission of both wavelengths of light. Before moving on to a different approach, we asked the technical support staff at Chroma, a company that specializes in optical technology, if they might be able to offer us some alternative. Fortunately, they informed us that their engineers would be able to fashion a beam-splitter to our specifications at a cost of just \$250 dollars, available in 4-6 weeks.⁶³ With a satisfactory solution having been reached, we were in a

position to move forward with our intended optical setup (the R/T curve of the resulting splitter is shown in Figure 5.17). The next step is to confirm its viability by subjecting it to error analysis.

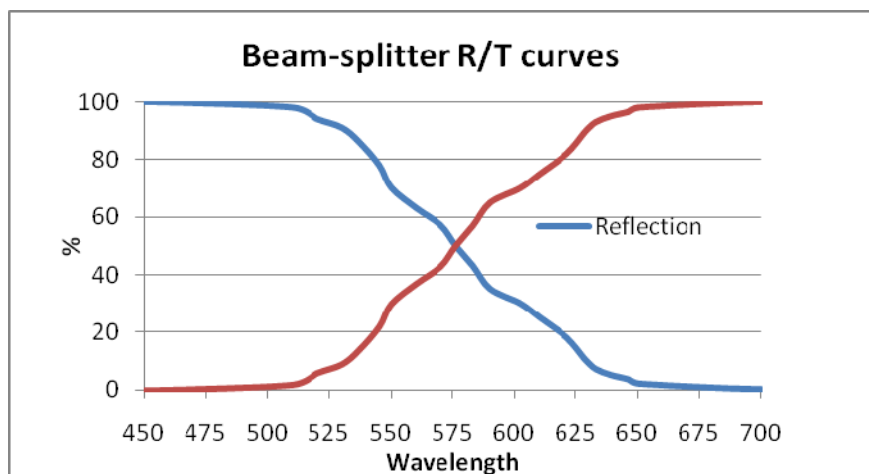


Figure 5.17 The beam-splitter curves allow us to predict the relative reflection and transmittance percentages for each of the fluorophores based on their emission wavelength.

PROBABILITY OF PROPERLY IDENTIFYING NUCLEOTIDES

After each nucleotide addition step in the sequencing process, there is an imaging step that seeks to find the identity of the nucleotide added to each template. The fluorophore attached to the recently added nucleotide is excited, and the light it emits is directed towards a beam-splitter. Based on the wavelength of the emitted light, different fractions of the light's total intensity are transmitted to two different cameras, one that collects red light and the other that collects green light. A quantity that compares the intensity of light recorded by each camera is used to identify the nucleotide.

The quantity used to compare the intensity of light in each camera will be called the dimensionless relative intensity (I). In order to correct for the differences in total intensity of light emitted that occurs in each event due to differences in orientation of the template molecule, the

quantity used is non-dimensionalized by dividing by the total intensity light captured by both of the cameras combined. I is defined by Equation 5.14.

$$I = \frac{G - R}{R + G} \quad \text{Equation 5.14}$$

Where G is the intensity of light captured in the green camera and R is the intensity of light captured in the red camera.

For the following discussion on nucleotide identification, the following assumptions will be used. Four nucleotides – to be called nucleotides 1, 2, 3 and 4 in order of increasing emission light wavelength – emit at evenly spaced wavelengths and with the same intensity. A theoretical beam-splitter is used that separates light at a wavelength exactly in between nucleotides 2 and 3. The splitter sends 100% of the light emitted by nucleotide 1 to the green camera, 70% of the light emitted by nucleotide 2 to the green camera and 30% of the light emitted by nucleotide 2 to the red camera, 30% of the light emitted by nucleotide 3 to the green camera and 70% of the light emitted by nucleotide 3 to the red camera, and 100% of the light emitted by nucleotide 4 to the red camera. Thus, using Equation 5.14 and the assumptions for the behavior of the four fluorophores and the nature of the beam-splitter, the four I values for the nucleotides when performing as expected are:

$$I_1 = 1, \quad I_2 = 0.4, \quad I_3 = -0.4, \quad I_4 = -1$$

It is of crucial importance to *SynthSeq's* process that nucleotides not be misidentified. Each of the four peaks in the graph below represent the distribution of I values for each of the four types of nucleotides in the process. If everything works correctly, each time an imaging step occurs, the data from the fluorescence of the fluorophore on the most recently added nucleotide of a template fragment will fall close to the average I value of the associated with that nucleotide. The case of a misidentified nucleotide occurs when the I value for a reading falls in the range of I values for a nucleotide other than the one that was just added. The following discussion seeks to show the probability of this kind of an error occurring.

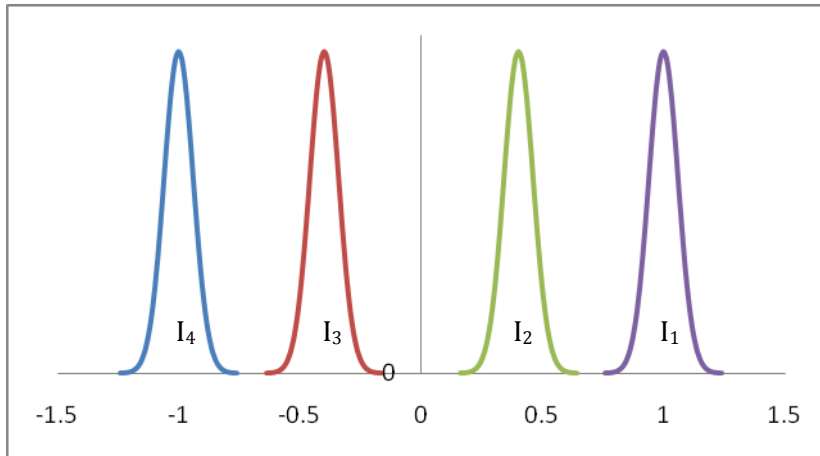


Figure 5.18 Schematic of the distribution of I values for the four fluorophores. $I_4 = -1.0$, $I_3 = -0.4$, $I_2 = 0.4$, $I_1 = 0$

The nature of the signal recorded by the cameras depends on a number of factors. The number of photoelectrons (N_{PE}) measured by the camera represents the signal strength. This number depends on a number of factors. The specific absorbance (SA) is a measure of how much energy (in this case, the number of excitation photons, N_{EX}) is absorbed by the fluorophore. The optical efficiency (OE) describes how many of the emitted photons are captured by the objective. The quantum efficiency (QE) is a measure of how many of the photons captured by the objective are converted into photoelectrons in the detector. The number of photoelectrons resulting from the excitation of one fluorophore is described by Equation 5.15:

$$N_{PE} = N_{EX} (SA)(OE)(QE) t_{exp} \quad \square$$

Equation 5.15

Where t_{exp} is the exposure time.

In this process, two sources contribute to noise. The first source of noise is from statistical fluctuations in the number of photons detected due to the random nature of their emission and capture. The number of photoelectrons that result from a signal can be described as a Binomial distribution. Thus, this contribution to noise can be described by the standard deviation of that distribution which is the square root of the average number of photoelectrons: $\sigma_{N_{PE}} = \sqrt{N_{PE}}$. The

second source of noise is from background. Background is the number of photoelectrons that are not a direct result of a signal (N_{BG}), where signal is defined as the excitation of a fluorophore that is attached to the last nucleotide added to a template DNA strand. The main contribution to background comes from fluorophores on non-incorporated, free nucleotides. A design of the diffusion of free nucleotides found that an average of one free nucleotide will be in one pixel of the image during the exposure time (see Reaction Conditions section). Thus, number of photoelectrons from background equal number of photoelectrons from signal: $N_{DE} = N_{BE}$. The total noise is calculated by combining these two sources of noise, resulting in Equation 5.16.

$$Noise = \sqrt{N_{DE} + N_{BE}} = \sqrt{2N_{DE}} \quad \text{Equation 5.16}$$

And the signal to noise ratio (SNR) is described by Equation 5.17.

$$SNR = \frac{Signal}{Noise} = \frac{N_{DE}}{\sqrt{2N_{DE}}} = \sqrt{\frac{N_{DE}}{2}} \quad \text{Equation 5.17}$$

In order to calculate the probability that statistical fluctuations and background noise will lead to the misidentification of a nucleotide (i.e. that the I value falls in the range of the wrong nucleotide), it is necessary to calculate the standard deviation of I for each nucleotide ($\sigma_{I,i}$). If the standard deviation is large, then there will be a large number of misidentified nucleotides, which could cause difficulties aligning the genome sequence and, ultimately, could lead to an inaccurate sequence. In order to calculate $\sigma_{I,i}$, the deviations in G and R must be propagated (see Equation 5.14). Doing so yields the following equation.

$$\left(\frac{\sigma_{I_i}}{I_i}\right)^2 = \frac{\sigma_{G_i}^2 + \sigma_{R_i}^2}{(G_i - R_i)^2} + \frac{\sigma_{G_i}^2 + \sigma_{R_i}^2}{(G_i + R_i)^2} \quad \text{Equation 5.18}$$

Where $I = \frac{G - R}{R + G}$, $G_i = (T_i)N_{DE}$, $R_i = (1 - T_i)N_{DE}$, $\sigma_{G_i} = \sqrt{N_{DE}(T_i) + (0.5)N_{BG}}$,

$\sigma_{R_i} = \sqrt{N_{DE}(1 - T_i) + (0.5)N_{BG}}$, and T stands for transmittance, which equals 1 for T_1 , 0.7 for T_2 ,

0.3 for T_3 , and 0 for T_4 . Plugging each of these values in and solving for σ_{I_i} yields Equation 5.19.

$$\sigma_{I_i} = (2T_i - 1) \sqrt{\frac{2}{(2T_i - 1)^2 N_{PE} + N_{PE}}}$$

Equation 5.19

Using Equation 5.15, Equation 5.19 and the equation relating the energy of a photon to its wavelength ($E = \frac{hc}{\lambda}$) it is possible to relate the power of the excitation laser to the standard deviation of I. Thus, it is possible to calculate a minimum power necessary to achieve a given level of certainty in the measurements. It was decided that if the distance halfway between the mean values for I₁ and I₂ (equal to the distance between I₃ and I₄ and also the shortest distance between any two mean I values) was greater than five standard deviations of I, then the number of errors due to noise (1 in 1,744,278) would be sufficiently low. Thus, the laser used in this process must be sufficiently powerful to provide enough photoelectrons to keep σ_{I_i} less than .06. Using the condition $\sigma_{I_i} < .06$, it is found that the excitation light must be strong enough to provide 1125 photoelectrons.

Next, it is necessary to determine if the laser being used to excite fluorophores is powerful enough to satisfy the condition. Equation 5.15 is used to find out how many photoelectrons the laser can produce to strike one fluorophore. The power of the laser being used goes up to 5W. First, this power must be converted to a number of excitation photons (N_{ex}) using the equation $E = \frac{hc}{\lambda}$. With an excitation light wavelength of 550nm, it is found that 2.76x10¹⁸ photons/W are produced by the laser. Since the ratio of the area of one pixel to the area of the entire imaging area is 1/1000000, the number of photons striking that area can be estimated as 2.76x10¹²/W.

$$\frac{A_{pixel}}{A_{image}} N_{ex} = N_{px}$$

Equation 5.20

Where N_{ex} is the excitation photons and N_{px} is the photons hitting the pixel.

Next, it is necessary to find the specific absorbance and quantum yield of the fluorophores. For this, a widely used fluorophore, Rhodamine 6G, will be used to represent the four fluorophores

in this system. The specific absorbance is determined by first finding the absorption cross section with Equation 5.21 and the known molar extinction coefficient of Rhodamine 6G ($\epsilon = 1.16 \cdot 10^5$ L/(mol)(cm))⁶⁴, and then dividing that value by the area of one pixel ($2.25 \cdot 10^{-9}$ cm).

$$\sigma = 1000 \ln(10) \frac{\epsilon}{N_A} = 3.82 \cdot 10^{-21} \text{ cm}^2 \quad \text{Equation 5.21}$$

The specific absorbance is found to be approximately $1.8 \cdot 10^{-7}$. The quantum yield of Rhodamine 6G is 0.95.

The only remaining terms in Equation 5.11 are QE, OY and t_{exp} . The quantum efficiency for the Andor L888SS EMCCD at the wavelengths being used is approximately 0.95.⁶⁵ The optical efficiency can be estimated to be 0.25. The exposure time in this process is 1/9 seconds. Putting all of these terms into Equation 5.15 gives a value of 12500 photoelectrons/W or 62500 photoelectrons at the full power of 5W. This value is well above the 1125 photoelectrons (which corresponds to a power of only .09 W) necessary to ensure that error due to noise is negligible.

Another difficulty in determining the identity of a nucleotide involves photobleaching – the photochemical destruction of a fluorophore. If a fluorophore photobleaches before a strong enough signal is produced, the result is a “dark read” and it is not possible to determine the identity of the nucleotide. The probability of photobleaching is a function of excitation intensity and time, concentration of oxygen, and temperature.⁶⁶ Decreasing the power of the excitation can reduce photobleaching, however this would also result in more noise and less certain measurements. Thus, it is necessary to optimize the power of the laser to accomplish low noise and low rates of photobleaching.

For common organic fluorophores, the probability of photobleaching each time the fluorophore is hit by a photon ranges from 10^{-6} - 10^{-5} . Photobleaching can be significantly reduced using several methods, including enzymatic deoxygenation using glucose oxidase/catalase. Using various methods, photobleaching probabilities as low as 10^{-9} have been reported.⁶⁷ For this

analysis, it is assumed that photobleaching occurs at a rate of 1 in 5000000 events and that a dark read occurs when a fluorophore is photobleached before half of the total imaging time has expired. In Figure 5.19, the probability of a dark read and the standard deviation of I for fluorophores on nucleotide 1/4 and 2/3 (nucleotides 1 and 4 have a slightly higher standard deviation of I than nucleotides 2 and 3, see Equation 5.19) are plotted against laser power. Since the standard deviation of I is proportional to the square root of the power, it decreases rapidly and then reaches a relatively stable lower limit, while the probability increases linearly as the power increases.

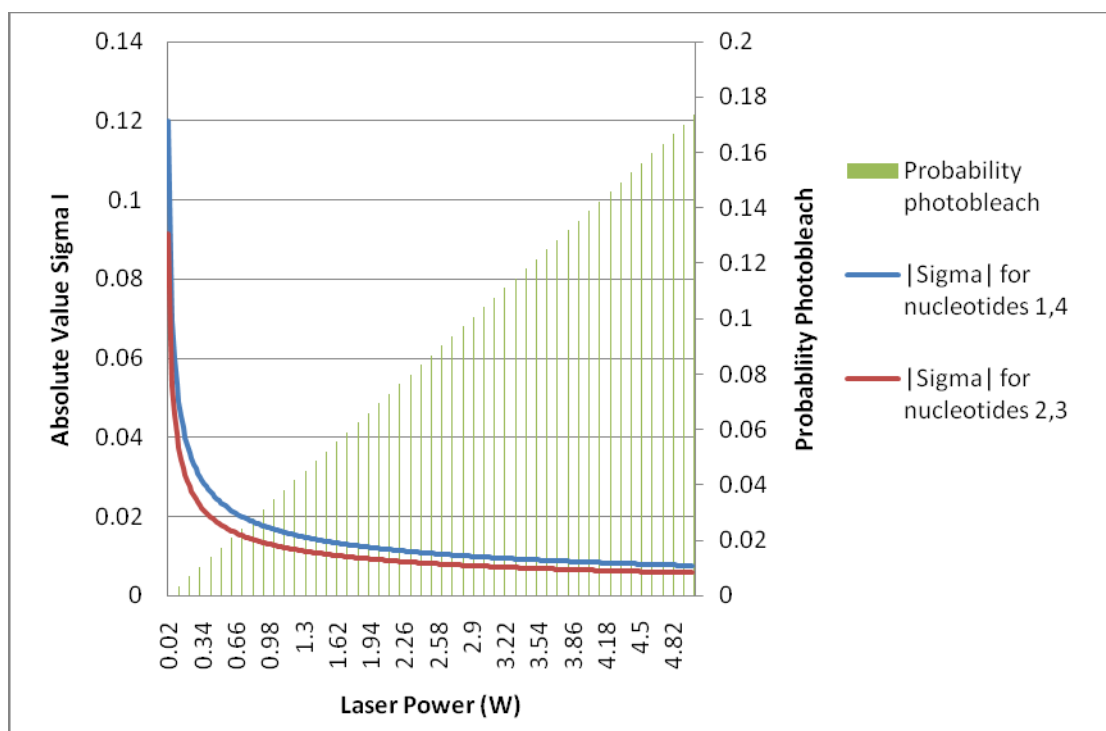


Figure 5.19 The standard deviation of I and the probability of a dark read vs. laser power

From relationships between the probability of a dark read and standard deviation of I vs. laser power, a laser power that minimizes dark reads and the standard deviation of I can be chosen. At a laser power of 280mW, the less than 1% of reads will be dark reads and the standard deviation of I is .032 for nucleotides 1 and 4 and .024 for nucleotides 2 and 3. These values are small enough

that noise can still be neglected as a significant source of error when attempting to identify nucleotides.

While it is encouraging that noise does not contribute to misidentification of nucleotides to any significant extent, this is not the only possible way that this error can occur. For example, if two template strands lie too close to one another on the surface of the slide, it is possible that emission light from these two reactions could contaminate one another. This contaminating light could lead to a value of I that does not correspond to the actual nucleotide on the template strand. There needs to be a way to correct for such errors.

This problem is solved by assigning a value to each template strand that quantifies the accuracy of its readings. This value will be termed the Z-score. It is defined as the average distance of I from the nearest mean I value. This quantity is defined by Equation 5.22.

$$\frac{\sum_i |I - \text{nearest } I|}{32}$$

Equation 5.22

The Z-score allows the identification of template strands whose fluorescence measurements often do not fall onto one of the four values that represents one of the four nucleotides. The data associated with any template strand whose Z-score is greater than three standard deviations of I can be discarded. As a result, even if one reading from an unusable template strand gives a value of I that corresponds to a nucleotide that is not the actual nucleotide present, this error will not be counted since the Z-score for that template will be high, and all of the readings associated with it will be discarded.

OVERVIEW OF THE IMAGING PROCESS

After the addition of each nucleotide to the growing DNA template, a specific imaging sequence is performed to detect the fluorescence from the base, as depicted in Figure 5.20. This sequence path is automated before the process begins to minimize the time and labor involved.

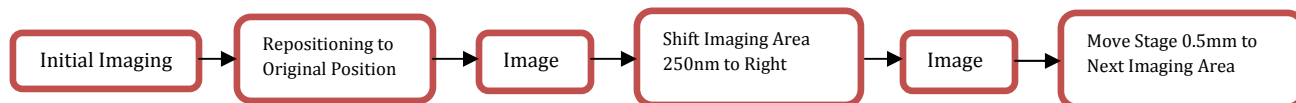


Figure 5.20 The procedure for gathering nucleotide identity data from a newly added base follows this sequence.

Initially, the area being analyzed is imaged to detect where templates are in relation to their initially recorded positions. The initial positions act as fiducials to determine how much the imaging area spatially deviated from its original position. The imaging area is readjusted using the piezo to position the imaging area in its proper place. This step is important because the position of each base addition detected during fluorescence corresponds to a particular DNA fragment template. If the position references were not accurate, the correct sequences of the templates would not be generated.

Once the stage is positioned, the nucleotide fluorophores in the imaging area are excited using the TIRF microscopy lasers. The emission is reflected back into the objective and sent to the dichromatic beam-splitter. Once the splitter separates the light at the determined wavelength, the two new beams are sent to the cameras for detection and then to the servers for genome data assembly.

In order to increase the optical efficiency of the templates on the pixel grid, the stage is shifted by 250nm. By performing this shift, the issues associated with multiple templates on one pixel and templates overlapping multiple pixels are reduced. After the shift occurs, the area is imaged and the data is computed so the intensities detected still correspond spatially to their original template positions. The stage must then move to the next programmed imaging area. This movement is accomplished using the XY stage positioner on the microscope stage, which moves at a

velocity of 10mm/s. Once the stage has settled and the flow cell is stationary, the imaging sequence is reinitialized for the new imaging area.

FLUOROPHORE CLEAVE

TERMINATOR REMOVAL AND 3'-OH REGENERATION

After nucleotide addition and imaging steps, the termination of the DNA synthesis reaction must be reversed. This involves the cleaving of the fluorophore attached to the nucleotide's base and the regeneration of the 3' hydroxyl group. Both of these steps are accomplished through one Pd catalyzed deallylation reaction.

The deallylation mixture that is introduced through the flow cell consists of Thermopol I reaction buffer, Na_2PdCl_4 and $\text{P}(\text{PhSO}_3\text{Na})_3$. This reaction is optimally run at 60°C , however at this temperature there is a risk of interfering with the binding affinity and performance of the phi29 polymerase, thus this reaction (and the entire process) is instead run at 40°C . At this temperature, the polymerase performance is not compromised. Since this reaction is being performed below its optimal temperature, the deallylation mixture is left in the flow cell for one minute to ensure full removal of the fluorophore and complete conversion of the 3'-O-allyl group to a 3'-OH group. To complete the deallylation/cleave step, Tris HCl buffer is introduced through the flow cell to remove the Pd complex. After this procedure, the system is ready for the next nucleotide addition and imaging step.⁶⁸

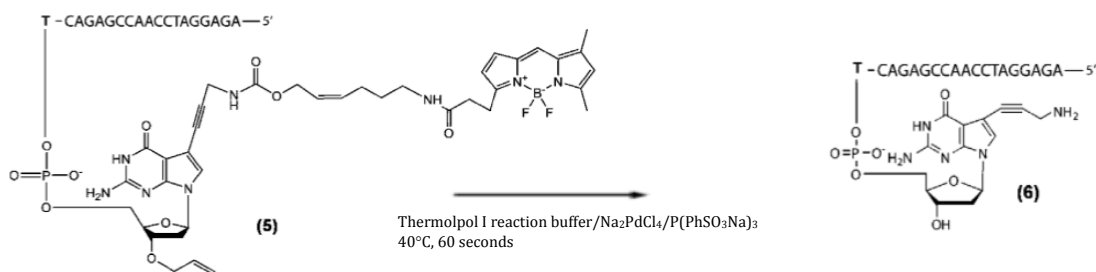


Figure 5.21 The fluorophore is cleaved and the 3'-OH is regenerated using Pd-catalyzed deallylation chemistry

CONCLUSIONS

In this chapter, we described the processes that lie at the heart of *SynthSeq's* technology – the chemical addition of reversibly terminating, tagged nucleotides, their imaging, and their identification. We outlined the initial imaging of all of the DNA strands using fluorophores attached to the poly-A tail of the fragments; this allowed for the mapping of the surface and establishment of fiducials to track the position and identity of each DNA molecule. We described the chemistry involved in our process including the polymerase used, the characteristics of the modified nucleotides used, the rate and fidelity of nucleotide incorporation, and the regeneration of nucleotide function. We also described the hardware and analysis used to accurately detect and identify single molecules including a TIRF microscope, specialized stage with nano-positioner, EMCCD camera, and beam-splitter. We showed that each step could be carried out with high confidence in its success. In short, we demonstrated that *SynthSeq* is able to accurately carry out direct, asynchronous, single molecule sequencing by synthesis.

The sequence of steps outlined in this chapter provided the most difficult technical challenges in *SynthSeq's* business and define the throughput capabilities of our process. Ultimately, the chemistry and detection portions of this process are what allow for the affordable, high throughput sequencing of the full human genome that sets *SynthSeq* apart from past technologies and current competitors.

6. GENOME ASSEMBLY

This chapter focuses on the alignment of the genome, which represents a very resource-intensive part of *SynthSeq*'s technology in terms of money and time. Just like many of the other

issues that *SynthSeq* has faced in developing its sequencing technology, the alignment of the genome presents a fundamental trade-off between cost and throughput. This trade-off is manifested in the amount of processing power dedicated to sequencing the genomes. If an insufficient amount of processing capacity is available, the process will bottleneck as detected reads continually pile up. Additionally, the method of parallelization across server cores will have a profound effect on turnaround. Assuming enough processing power is allocated to avoid data back-up, the primary issue becomes turnaround. Due to the fact that the process can be made highly parallel, the sky is the limit when it comes to the theoretical alignment speed. However, a bank of multi-core processors can quickly become prohibitively expensive. We have developed our alignment strategy with the intention of providing enough processing capacity such that the chemistry and detection steps will be throughput-limiting. Although this approach resulted in higher costs for the alignment side of the operation, we believe that the throughput it allows is worth the cost. Note that throughput is not only reliant upon the chemistry and detection steps, but also the ability to align bases at a rate similar to that of their detection. Ample processing power must be allocated to the genomic alignment; otherwise a bottleneck in the system will arise.

This processing allowance is separate from turnaround, which is based on how we parallelize the sequencing itself. Each genome is sequenced by exactly one server, parallelized across its 32 cores. This method results in a 30 hour alignment time per genome, based on our conservative estimates (see Equation 6.1). The following sections delineate the alignment requirements and demonstrate its feasibility. It is important to note that the throughput estimates in this chapter are conservative. During the start-up year, we may find that the optimization of our alignment strategy significantly reduces our expected processing demands, lowering operating costs considerably.

DE NOVO SEQUENCING VS. COMPARATIVE GENOME ASSEMBLY

Within the field of genotyping, there exist two different approaches to genome alignment, and both are frequently used. *De novo* sequencing is characterized by the independent alignment of multiple reads without the use of a reference genomic template, whereas *comparative genome assembly* relies on an existing genome to aid in alignment.⁶⁹ Generally, longer read technologies are better suited to *de novo* sequencing, while shorter read-length systems rely heavily on *comparative genome assembly*. Acknowledging that *SynthSeq* constitutes a short read length technology (32 base pairs), our reassembly strategy utilizes the public-domain human genome sequence in concert with comparative alignment software. Aligner algorithms, such as *indexDPgenomic* and BFAST, which will be discussed later in the chapter, are designed specifically for comparative genome assembly of short reads. They can therefore carry out successful alignment of reads in a fraction of the time and with far less complexity than would be required if employing a *de novo* assembly strategy.⁷⁰

DATA COLLECTION AND PROCESSING

Each sequencing station direct connects to a signal acquisition server in an adjacent room. This unit scores the frames from the two EMCCD cameras, identifies the fluorophores in each viable pixel, and then converts the data into a form that the aligner can process. No-reads are converted into Xs, and any sequence that fails the Z-score test is thrown out. The data for each genome is translated into a nucleotide sequence using standard signal processing techniques and sent directly to a dedicated server for alignment.

Alignment speed is primarily based on memory access rates and multithreading capability. Optimized reassembly is achieved only by minimizing the time required by each individual task while maximizing the number of tasks that can be performed simultaneously. In the same way that the parallelization of steps is fundamental to the overall process design, individual genome

fragments can be distributed to multiple cores to speed up processing. Given the complexity of modern computing systems, the result of this parallelization cannot be perfectly predicted, which underscores the importance of optimizing the alignment process during the start-up period. Such an optimization might allow us to dramatically reduce the equipment costs in series B.

REASSEMBLY OF THE READS

Once the frame data is converted to nucleotide sequences, our chosen algorithm aligns this sequence data (i.e. the reads) to the reference genome. The aligner produces a catalog of nucleotide sequences, positioning them in the appropriate part of the genome. After all fragments are organized in this way, a consensus is reached for every position, and a value of A, C, T, G, or X is assigned to each position accordingly, generating the final genome. As previously mentioned, the primary errors associated with *SynthSeq* are no-reads, which are represented by an “X” signifying that the identity is inconclusive (see Chapter 5). Our ability to image after the addition of each nucleotide for a known number of cycles largely precludes both insertion and deletion errors that plague many other technologies. Note that the aligner itself does not make errors, but it will faithfully relay any errors emanating from the detection side of the technology.

BIOINFORMATICS HARDWARE REQUIREMENTS

A typical bioinformatics setup for a genotyping application includes the following elements:

- A minimum two Dual Core Intel Xeon Processor (or a single Quad Core) with a speed of 2.6GHz (or higher) and at least 4 GB of RAM per-core

- At least 250GB of hard disk space with each core, or high-bandwidth access to a central store as an alternative.
- At least 10 TB of centralized storage
- High-performance Port Switch offering high-bandwidth networking across all servers⁷¹

The *SynthSeq* solution to each of these hardware requirements is detailed in the following sections.

Processor – IBM Power 750 Express Server

The IBM Power 750 Express Server was chosen for its immense processing power and energy efficiency. With 32 cores, each possessing 4GB of RAM, and a 3.0 GHz processor, this server meets all bioinformatics requirements and is more economical than competing servers possessing fewer cores.

Central Storage – Netdisk 9010N Storage Server

In addition to highly powerful servers that are able to align the sequencing data to the reference genome, we also need a server that can store all of the data that results. For this purpose, we have chosen the Netdisk 9010N, a storage server with 10TB of hard drive space, meeting the recommended minimum amount of storage capacity. When combined with the 128GB hard-drives on each of our processors, however, the aggregate amount of disk space far exceeds the minimum.

Server Networking Solution – Dell PowerConnect 6224 24 Port Switch

The final piece of the hardware puzzle is a port switch that allows all of the various components to communicate with each other, promoting parallel processing and high-bandwidth

data transfer. We have chosen the Dell PowerConnect 6224 Port Switch for this purpose, due to its low cost, impressive specifications, and use in similar bioinformatics setups.

ALIGNER SOFTWARE AND COMPUTATIONAL DEMANDS

The above components represent the hardware requirements, but there also exists the issue of which aligner software to use. Fortunately, Helicos Biosciences offers a genome aligner that is specifically designed for short read-lengths, called *indexDPgenomic*. This aligner supports fast, mismatch tolerant indexed alignment of short reads. It is an improved version of indexDP that is aimed at aligning reads to large reference sets, making it ideal for *SynthSeq* sequencing methods. Also, because the software is open source, we do not need to worry about any costs associated with the alignment program itself.

For the *indexDPgenomic* aligner, we have decided to break up the alignment of each genome into 15-million-read chunks that will each be processed by one 4GB core. These jobs are completed in parallel across the 32 cores of an individual server, reducing the alignment time significantly. A bank of these servers is connected via broadband connection, and our 10TB storage server serves the total processing requirements of twelve genomes per day. Using benchmarks determined for the alignment of human genome reads with the BFAST aligner (also open-source), we are able to estimate our expected processing demands.

$$(12 \text{ genomes / day}) * (18 \text{ Gigabases / genome}) * (480\text{GB RAM} * 1 \text{ day} / 55 \text{ Gigabases}) * (1 \text{ server} / 128\text{GB RAM}) = 14.7 \text{ servers required} \rightarrow \underline{15 \text{ servers}}$$

Equation 6.1

Studies employing the BFAST aligner report the daily alignment of 55 billion human bases to a reference genome using processing power equivalent to four of our servers. Fortunately, our low coverage requirement dictates that we need to align just a fraction of that number of bases per

genome. Based on Equation 6.1 above, a bank of 15 32-core servers with 4GB RAM per core is sufficient to achieve our desired throughput of twelve genomes per day, and this setup may even allow for significant throughput increases on the chemistry and detection side of *SynthSeq*.⁷² The calculated processing requirement translates to approximately 30 hours per genome using one of the 32-core servers with 128 GB RAM to align each genome.

Depending on how the processing is allocated, the actual time for each genome to be aligned will vary, but this amount of processing power should be sufficient based on these available conservative benchmarks. Furthermore, the researchers note that the alignment was neither multi-threaded nor parallel, two options that would greatly improve calculated rates and are supported by BFAST and *indexDPgenomic*.⁷³ Starting out with just one or two servers during the series A period will allow us to optimize the alignment protocols and gain a better understanding of the exact processing capacity that will be required for the scale-up in subsequent years.

PRODUCT DELIVERY

The final piece of the puzzle is to determine how we will deliver the sequences to our customers and at what cost. From the sources on this subject, the data file of the human genome would be about 750MB uncompressed (Equation 6.2). Including the reference human genome along with the customer's genome, in addition to our analysis and recommendations or at the very least the names of some referral companies might be desirable. Allowing for the inclusion of all of these things, one 2GB flash drive per customer should be sufficient (smaller if we zip the sequences). However, the low cost of such drives limits the utility of shrinking the files. If we can compress the data files to a great enough extent, we may be able to e-mail them in the future. This

practice might call customer confidentiality into question, however, so we may instead upload them to a secure server from which customers can download their sequence, the reference human genome sequence, a FAQ sheet, and perhaps a list of referral companies that can further analyze their genome. At least for the start-up phase of *SynthSeq*'s technology, we plan to send genome sequences to our customers using flash drives with the latest encryption technology. The equations used to determine the size of the raw data file are below:

$$3,080,400,000 \frac{\text{bases}}{\text{human genome}} * 2 \frac{\text{bits}}{\text{base}} * \frac{1 \text{ byte}}{8 \text{ bits}} * \frac{1 \text{ MB}}{1,048,576 \text{ bytes}} = 734.4 \text{ MB per uncompressed human genome}$$

Equation 6.2

CONCLUSIONS

As mentioned previously, one of the primary advantages of *SynthSeq* technology is its lack of insertion and deletion errors. The fact that no-reads are the only source of error alleviates our need to achieve the high coverage requirements characteristic of other technologies. Researchers have noted that *indexDPgenomic* is designed specifically to accommodate insertion and deletion errors, as opposed to other aligners created under the assumption that substitutions are more prominent.⁷⁴ *IndexDPgenomic* was chosen due to its use in short-read systems such as *SynthSeq*, but due to the program's penchant for insertion and deletion errors, we will need to evaluate its speed in relation to other aligners, including BFAST. During our start-up period, we will have the opportunity to optimize our alignment procedures. Fortunately, both *indexDPgenomic* and BFAST are open-source, allowing us the freedom to test them without any additional cost. Using conservative benchmarks and parallel processing across a single server per genome, we have determined that we can prevent data backup and achieve 30 hour genome alignment, using a total of fifteen servers, ensuring both the throughput of our process and a respectable turnaround.

7. FINANCIAL ANALYSIS

All previous chapters have dealt with the biochemical processes and technological hurdles of this project, yet the financial analysis of the technology is equally important to the success of the company. This analysis will reveal whether or not the project is financially feasible; without an existing market and the ability to generate a profit, no investors will agree to fund it. The technology would fail from a business standpoint. This financial analysis will prove that the project is indeed viable. Both NPV and MIRR analyses will be used to gauge the project's profitability. All necessary calculations and figures are included herein. It is important to note that all the calculated values in this analysis are contingent upon projected earnings that are based on stated assumptions.

Analysis begins with the revenue projections based on genomic throughput and price. A separate sensitivity analysis is done on the genome price because of the importance of revenue in the financial model. Total costs and depreciation explanations follow the revenue projections and allow us to build an income statement for the project. The income statement shows earnings, but we want to see free cash. Therefore, we need to adjust earning figures into cash ones by examining working capital and other cash-affecting items.

The company's value can then be determined by combining the terminal value analysis and discounted cash flow analysis. This is followed by the rate of return analysis for the two series of investors involved in this project. This section is more difficult than conventional analysis, since there are both series A and series B investors, who participate in two separate rounds of investment. In order to simplify these added complications, an equity stake analysis is conducted. With all of these explanations completed, the calculations and final results will be presented on a single page spreadsheet for convenience.

Multiple what-if scenarios will be discussed to see how they would impact our bottom line. The scenario analysis will include a genome price sensitivity analysis to study the project's

limitations on profit. Once a sales price is established from the sensitivity analysis, two final scenarios are performed to pass final judgment on the potential success of the *SynthSeq* platform.

MARKET AND REVENUE PROJECTION

The genomic sequencing market is a relatively new market that has so far only been accessible to wealthy individuals. However, the efforts of various companies are working to make the market accessible to the general public. The *Archon X Prize* has contributed to the technological improvements seen in recent years. The direct result of these technological improvements is a price drop that should drastically expand the market for personal genomic sequencing. This expansion is uncertain due to the paucity of available data for current sales volume. Revenue projections are therefore difficult to accurately determine for this young and unpredictable market.

The following sections will assume our genome price to be \$10,000, since the *Archon X Prize for Genomics* is given to the first team to successfully manufacture and sell a genome for this price.⁷⁵ A sensitivity analysis will be conducted in a later section to determine the actual sale price of each genome. This determination is crucial for a proper financial analysis to be conducted on the project. We must also assume that at 100% manufacturing capacity, 3000 genomes will be sold each year. Thus, the company will gross \$30 million using these initial assumptions.

There are four stages of growth and development: the research stage, the scale up stage, the sales stage, and the termination stage. The research period encompasses the development of a working prototype and all necessary capital is funded by series A angel investors. There is no revenue during this research phase, and only limited staff is employed. After a working prototype is successfully developed, the scale up stage begins with series B investors providing the rest of the necessary capital. New staff is added, and the step up increase allows for the first sales to begin. We will assume that the company will be at 50% manufacturing capacity during the scale up phase. We reach 100% design capacity during the sales stage, and the company is fully functional. The

terminal stage marks the end of the sales and a terminal value is calculated based on prior free cash flow projections. At this time, the company will hopefully have made a profit and will either be sold to a larger firm or liquidated for revenue.

The following time scale will be used in the following sections: the research stage will take one year, the scale up stage will last another year, and the sales stage will continue for three additional years. The terminal stage ends the life of the company and a terminal value can be calculated. The five year lifetime of the company is reasonable based on the rapid technology improvements in the genome sequencing market. The revenue projections for these four phases are shown in Table 7.1.

Revenue Projections						
Year	<i>2011</i>	<i>2012</i>	<i>2013</i>	<i>2014</i>	<i>2015</i>	<i>2016</i>
Stage Name	Research	Scale Up	Sales	Sales	Sales	Terminal
Design						
Capacity	0%	50%	100%	100%	100%	100%
Revenue	\$0.00	\$15,000.00	\$30,000.00	\$30,000.00	\$30,000.00	Terminal Value

Table 7.1 Revenue Projections for the following 5 years (\$ in thousands).

It is important to note that inflation will be ignored for the entirety of the financial analysis. Inflation calculations over such a small timescale are unnecessary. Moreover, the 1986 Financial Accounting Standards Board determined that inflation accounting is unnecessary for financial statements, which lie at the heart of this analysis.⁷⁶

COSTS, PPE, AND DEPRECIATION

A number of different costs are associated with running our company. These costs fall into four distinct categories: equipment costs for the research phase and the sales phase, annual research costs, and annual sales costs. Equipment purchased for the research stage includes all laboratory machines necessary for developing the working prototype. This includes sample preparation machines, TIRF microscopes, EMCCD cameras, and processors. The annual research cost is essentially employee salaries and lab space rent during this stage. Equipment purchased during the sales stage is comprised of all additional laboratory equipment needed to scale up production and reach 100% design capacity. The annual sales cost is the yearly burn rate of the company once it is able to generate revenue. This burn rate is substantially greater than the research stage burn rate. Inventory, research and development, and sales costs are added to salary and rent during this stage. These costs are summarized in Table 7.2.

series A angel investors fund the cost of the research stage equipment purchases and the cost of research for the first year. These investors are wealthy individuals who are willing to participate in the risky research stage for developing new technology. For this project, the series A investment comes to about \$1.7 million.

series B investors join after a working prototype is developed and the research stage is completed, thereby avoiding the high risk shared by the series A investors. The series B investment is used to fund the purchase of all remaining equipment necessary for scaling up, as well as three months worth of salary, rent, and inventory. The remaining salary, rent, and inventory costs can be funded by generated revenue, assuming the company reaches 50% design capacity by the end of that fiscal year. This will ensure the solvency of our company during this time. The total necessary investment during series B is \$2.85 million. It will be funded one year after the series A investment, upon completion of the research stage.

As previously stated, the labor costs during the research stage are different from labor costs in the following stages. Workers during the research stage consist of one secretary, two interns,

and four senior scientists, one of which will also serve as the company's Chief Technical Officer (CTO). The four senior scientists will not receive their full salary during prototype development. Instead, they will receive two-thirds of their ordinary salary in exchange for a 10% equity stake in the company. This decision is the result of a compromise between the series A investors and the four scientists. The investors want to limit the risk of their capital as much as possible, so they try to cut salary costs by agreeing to give the employees a stake in the company. Under normal conditions, the investors receive 85% of the company and give the employees 15%, especially when the workers lack entrepreneurial experience. However, the added financial risk associated with biotechnology companies may warrant series A investors to demand 90% ownership in the company instead.

During the scale up stage and sales stages, the labor costs will be significantly higher for two reasons. First, the four senior scientists will now receive their full salaries and begin working in research and development to keep the company competitive in a fast-paced genomic market. Secondly, additional staff will be hired to meet the throughput demands of the company. New staff members include a CEO, a CFO, two junior scientists, an organic chemist, three operators, two salespeople for marketing, an HR representative, and an IT specialist. The junior scientists will work on sample preparations, and the organic chemist will work to develop the reagents in-house.

To meet our genome quota, we need to hire operators to work night shifts in order to keep production going at all times. The sequencing of 3,000 genomes can be realized if production operates around the clock for 250 days each year, at a rate of twelve genomes per day. The company must operate for 24 hours a day for five days a week to achieve this level of production. No operator can work more than 40 hours in a week, and one must be on site at all times to remove and load new flow cells to keep the sequencing running throughout the night. Since there are 120 working hours per week, three operators will be necessary to meet our throughput demands.

Rental costs are also different during the research stage and in subsequent stages. Space is minimized during prototype development, but expansion is necessary to meet company demands. Based on available rental costs and room, *SynthSeq* will be located in Boston, MA.⁷⁷

Inventory costs for this project include flow cells, flash drives to store DNA read outs for each customer, sample preparation kits, and reagents. The total cost of these items depends on the exact stage in question. The research stage is only involved in building a prototype; there are no associated inventory costs. At 100% design capacity during the sales stage, the total inventory cost is \$691,000. During the scale up stage, we operate at 50% capacity and the total cost is halved.

Lastly, the operating costs include R&D and sales costs. These costs also only exist once the research phase is complete. Our project will use 15% of revenue on research and development costs and 3% of revenue on sales related costs. These are typical numbers for a biotechnology startup company.

Cost Estimates

Research Stage (Series A)

Equipment Costs for Research Stage

Item	Unit Cost	Quantity	Total Cost
Magratriation 12GC	\$36,000.00	1	\$36,000.00
Covaris s2 Shearer	\$55,000.00	1	\$55,000.00
Syringe Pump	\$3,006.00	2	\$6,012.00
LBL DS-50 SLIDE STAINER 115V	\$16,734.00	1	\$16,734.00
EMCCD	\$44,574.00	4	\$178,296.00
Olympus TIRF Microscope	\$200,000.00	2	\$400,000.00
MAX201 stage controller	\$6,460.00	2	\$12,920.00
SCXY2100B positioner	\$9,990.00	2	\$19,980.00
Teklynx Label Matrix Software	\$338.50	1	\$338.50
Datanmax-O Neil E4203	\$218.75	3	\$656.25
Beam Splitter	\$535.00	1	\$535.00
IBM Power 560 Express	\$133,415.00	3	\$400,245.00
IBM Power 750 Express Servers	\$500.00	2	\$1,000.00
Dell PowerConnect 6224 (Port Swi	\$101,952.00	1	\$101,952.00
NetDisk 9010N 10TB Storage Towr	\$1,015.00	3	\$3,045.00
Temperature Controller	\$1,999.00	1	\$1,999.00
Fluorescence Barrier Filter	\$810.81	2	\$1,621.62
	\$250.00	2	\$500.00
Total:			\$1,236,834.37

Labor Costs

Personnel	Salary	Quantity	Total Cost
CTO	\$80,000.00	1	\$80,000.00
Senior Scientist	\$80,000.00	3	\$240,000.00
Secretary	\$30,000.00	1	\$30,000.00
Intern	\$3,000.00	2	\$6,000.00
Total:			\$356,000.00

Inventory Costs

Item	Unit Cost	Quantity	Total Cost
Flow Cells	\$75.00	6000	\$450,000.00
Oragene DNA Kits	\$14.75	3000	\$44,250.00
QIAquick Purification Kit	\$1.85	6000	\$11,100.00
MinElute Purification Column	\$1.98	3000	\$5,940.00
Reagents	\$50.00	3000	\$150,000.00
SanDisk Cruzer USB flash drive	\$10.00	3000	\$30,000.00
Total:			\$691,290.00

Cost Estimates

Post-Research Phase (Series B)

Equipment Costs after Research Stage

Item	Unit Cost	Quantity	Total Cost
Syringe Pump	\$3,006.00	3	\$9,018.00
EMCCD	\$44,574.00	6	\$267,444.00
Olympus TIRF Microscope	\$140,000.00	3	\$420,000.00
MAX201 stage controller	\$6,460.00	3	\$19,380.00
SCXY2100B positioner	\$9,990.00	3	\$29,970.00
Opticon OPL-6735 Barcode Scanne	\$218.75	3	\$656.25
IBM Power 560 Express	\$133,415.00	3	\$400,245.00
Beam Splitter	\$500.00	3	\$1,500.00
IBM Power 750 Express Servers	\$71,366.40	14	\$999,129.60
Temperature Controller	\$810.81	3	\$2,432.43
Dell PowerConnect 6224 (Port Swi	\$1,015.00	3	\$3,045.00
Fluorescence Barrier Filter	\$250.00	3	\$750.00
Total:			\$2,153,570.28

Labor Costs

Personnel	Salary	Quantity	Total Cost
CEO	\$180,000.00	1	\$180,000.00
CFO	\$150,000.00	1	\$150,000.00
CTO	\$120,000.00	1	\$120,000.00
Senior Scientist	\$120,000.00	3	\$360,000.00
Junior Scientist	\$60,000.00	2	\$120,000.00
Organic Chemist	\$90,000.00	1	\$90,000.00
Secretary	\$30,000.00	1	\$30,000.00
Salesperson	\$60,000.00	2	\$120,000.00
HR Representative	\$60,000.00	1	\$60,000.00
IT Specialist	\$50,000.00	1	\$50,000.00
Operator	\$50,000.00	3	\$150,000.00
Total:			\$1,410,000.00

Inventory Costs

Item	Unit Cost	Quantity	Total Cost
Flow Cells	\$75.00	6000	\$450,000.00
Oragene DNA Kits	\$14.75	3000	\$44,250.00
QIAquick Purification Kit	\$1.85	6000	\$11,100.00
MinElute Purification Column	\$1.98	3000	\$5,940.00
Reagents	\$50.00	3000	\$150,000.00
SanDisk Cruzer USB flash drive	\$10.00	3000	\$30,000.00
Total:			\$691,290.00

Rental Costs				Rental Costs			
Item	Cost/sqft	Sqft	Total Cost	Item	Cost/sqft	Sqft	Total Cost
Biochemical Laboratory	\$33.00	400	\$13,200.00	Biochemical Laboratory	\$33.00	800	\$26,400.00
Sequencing Laboratory	\$33.00	300	\$9,900.00	Sequencing Laboratory	\$33.00	600	\$19,800.00
Office Space	\$20.00	1000	\$20,000.00	Office Space	\$20.00	1000	\$20,000.00
Total:			\$43,100.00	Total:			\$66,200.00

Rental Costs				Rental Costs			
Item	Cost/month	Months	Total Cost	Item	Cost/month	Months	Total Cost
Utilities	\$5,000.00	12	\$60,000.00	Utilities	\$5,000.00	12	\$60,000.00
Maintenance	\$1,000.00	12	\$12,000.00	Maintenance	\$1,000.00	12	\$12,000.00
Total:			\$72,000.00	Total:			\$72,000.00

Operating Costs				Operating Costs			
Item	Sales	% of Sales	Total Cost	Item	Sales	% of Sales	Total Cost
Research	\$30,000,000.00	15%	\$4,500,000.00	Research	\$30,000,000.00	15%	\$4,500,000.00
Sales	\$30,000,000.00	3%	\$900,000.00	Sales	\$30,000,000.00	3%	\$900,000.00
Total:			\$5,400,000.00	Total:			\$5,400,000.00

Total Annual Costs:		Total Annual Costs:	
Series A Total	Series B Total	Series A Total	Series B Total
\$1,707,934.37	\$2,825,417.28	\$471,100.00	\$7,639,490.00

Table 7.2 Cost Estimates for series A and B.

Our company will use the 5-year MACRS depreciation schedule in order to maximize the amount of tax savings from the accelerated tax schedule. Depreciation is a non-cash expense that decreases the pre-tax income from which taxes are deducted. An accelerated depreciation schedule is ideal for short lived projects such as this because of the significant impact the savings has on the NPV and MIRR analyses. The 5-year MACRS depreciation percentages in order are 20%, 31%, 19.2%, 11.52%, 11.52%, and 5.76%.

Depreciation Schedule					
MACR Tax Schedule	20.00%	32.00%	19.20%	11.52%	11.52%
Year	<i>2011</i>	<i>2012</i>	<i>2013</i>	<i>2014</i>	<i>2015</i>
series A Equipment	\$1,236.83				
Depreciation		(\$247.37)	(\$395.79)	(\$237.47)	(\$142.48)
series B Equipment		\$2,153.57			
Depreciation			(\$430.71)	(\$689.14)	(\$413.49)
Initial Net PPE	\$0.00	\$1,236.83	\$3,143.04	\$2,316.54	\$1,389.92
PPE Purchased/(Sold)	\$1,236.83	\$2,153.57	\$0.00	\$0.00	\$0.00
Less: Total Depreciation	\$0.00	(\$247.37)	(\$826.50)	(\$926.61)	(\$555.97)
Final Net PPE	\$1,236.83	\$3,143.04	\$2,316.54	\$1,389.92	\$833.95

Table 7.3 Depreciation Schedule using the 5-year MACRS schedule (\$ in thousands).

The series A equipment is labeled as the research equipment purchase, while series B equipment is labeled as the sales equipment purchase. The ending net PPE figures in Table 7.3 represent the balance sheet amounts for how much property and equipment the company owns. Total depreciation is found on the income statement (Table 7.4) and decreases pre-tax income, thereby lowering taxes. During the what-if scenarios, we will explore the effect of the research phase taking two years instead of just one. Under these circumstances, the series B equipment and

its depreciations begin one year later on the income statement. As a result, the total depreciations for each year will change.

INCOME STATEMENT

Income Statement					
Year	<i>2011</i>	<i>2012</i>	<i>2013</i>	<i>2014</i>	<i>2015</i>
Revenue	\$0.00	\$15,000.00	\$30,000.00	\$30,000.00	\$30,000.00
Cost of Sales	(\$115.10)	(\$483.85)	(\$829.49)	(\$829.49)	(\$829.49)
Gross Profit	(\$115.10)	\$14,516.16	\$29,170.51	\$29,170.51	\$29,170.51
Operating, SG&A Expenses	(\$356.00)	(\$4,110.00)	(\$6,810.00)	(\$6,810.00)	(\$6,810.00)
Depreciation	\$0.00	(\$247.37)	(\$826.50)	(\$926.61)	(\$555.97)
Pre-Tax Income	(\$471.10)	\$10,158.79	\$21,534.01	\$21,433.90	\$21,804.54
Tax @ 40%	\$188.44	(\$4,063.52)	(\$8,613.60)	(\$8,573.56)	(\$8,721.82)
Net Income	(\$282.66)	\$6,095.27	\$12,920.41	\$12,860.34	\$13,082.72
Design Capacity	0%	50%	100%	100%	100%
Margins					
Gross Margin	0.00%	96.77%	97.24%	97.24%	97.24%
Profit Margin	0.00%	40.64%	43.07%	42.87%	43.61%

Table 7.4 The income statement showing gross and profit margins (\$ in thousands).

The total costs appearing in Table 7.4 are broken down into the cost of sales and operations, or the SG&A expenses. The cost of sales includes all costs that are directly involved in the making of the product, both fixed and variable costs. Fixed costs include rent and utility costs because these do not vary with the number of goods produced and sold. Variable costs are the cost of inventory because this cost is dependent on the number of genomes sequenced each year. The gross profit is calculated by simply subtracting our company's revenue by the total cost of sales. The gross margin

is the percentage of how much money is left from the revenue after paying the cost of sales. The operating, SG&A expenses deal with selling, general, and administrative fees associated with the company's operation. These expenses include the fixed cost of salary, as well as the variable costs from research and development and sales. These variable costs are based on the amount of genome sequenced, as well as the price charged for the sequencing. The pre-tax income can now be calculated by subtracting the operating, SG&A expenses and depreciation from gross profit.

Federal taxes are currently set around 35%, but total taxes increase to near 40% with the addition of state tax. Note that taxes in the first year are actually positive because the company has only incurred losses; we receive money from the government in what is known as a tax shield. This shield does not always apply, but for the purposes of our analysis we will make this assumption.

WORKING CAPITAL

The generated income statement allowed us to calculate net income, but this is not equivalent to cash. We need cash figures in order to conduct an NPV and MIRR analysis, so we must adjust net income for cash items. One of these cash items that must be adjusted for is working capital.

Working capital is the amount of capital needed by a company for it to operate normally. Essentially, working capital is a portion of profits allocated for daily operations; it can be described as current assets minus current liabilities. Current assets are anything that can quickly be converted to cash while current liabilities are all bills and debts that need to be paid back. Thus, working capital is the remaining money left over after the company pays off its debts.

This analysis will cover the four main working capital items: accounts receivable, inventory, accounts payable, and cash reserve. Accounts receivable are earnings that have not been paid for in

cash yet. We will assume that the customer will have 30 days to pay for the product. With this being the case:

$$\text{Accounts Receivable} = \frac{\text{Revenue (\$)}}{\text{Year}} * \frac{1 \text{ Year}}{365 \text{ Days}} * 30 \text{ Days} \quad \text{Equation 7.1}$$

Inventory comprises the items needed to sequence the genomes. This includes reagents, QIAGEN kits, flow cells, etc. and can be found in Table 7.2. Since new inventory will be purchased each month:

$$\text{Inventory} = \frac{\text{Inventory Cost(\$)}}{\text{Year}} * \frac{1 \text{ Year}}{365 \text{ Days}} * 30 \text{ Days} \quad \text{Equation 7.2}$$

Accounts payable are the company's bills that have yet to be paid for in cash. These bills will be paid every 30 days; in this way accounts payable are the opposite of accounts receivable. Since operating costs and rent are the company's main bills:

$$\text{Accounts Payable} = \frac{(\text{Rent} + \text{Operating Costs(\$)})}{\text{Year}} * \frac{1 \text{ Year}}{365 \text{ Days}} * 30 \text{ Days} \quad \text{Equation 7.3}$$

Cash reserves are the cash kept on hand to pay for future salaries. Since three months salary are held on reserve:

$$\text{Cash Reserve} = \frac{\text{Salary(\$)}}{\text{Year}} * \frac{1 \text{ Year}}{12 \text{ Months}} * 3 \text{ Months} \quad \text{Equation 7.4}$$

The change in these four working capital items allows net income to be changed into cash. These changes are detailed in Table 7.5.

Working Capital					
Year	<i>2011</i>	<i>2012</i>	<i>2013</i>	<i>2014</i>	<i>2015</i>
Working Capital Item Estimates					
Accounts Receivable	\$0.00	\$1,232.88	\$2,465.75	\$2,465.75	\$2,465.75
Inventory	\$0.00	\$28.41	\$56.82	\$56.82	\$56.82
Accounts Payable	\$9.46	\$233.28	\$455.19	\$455.19	\$455.19
Cash Reserve	\$89.00	\$352.50	\$352.50	\$352.50	\$352.50
Changes in Working Capital					
(Increase)/Decrease in A/R	\$0.00	(\$1,232.88)	(\$1,232.88)	\$0.00	\$0.00
(Increase)/Decrease in Inv	\$0.00	(\$28.41)	(\$28.41)	\$0.00	\$0.00
(Increase)/Decrease in A/P	\$9.46	\$223.82	\$221.92	\$0.00	\$0.00
(Increase)/Decrease in C/R	(\$89.00)	(\$263.50)	\$0.00	\$0.00	\$0.00
Total Change in Working Capital	(\$79.54)	(\$1,300.97)	(\$1,039.37)	\$0.00	\$0.00

Table 7.5 Working Capital and Changes in Working Capital (\$ in thousands).

Any increase in assets decreases cash because the company needs to spend money to buy necessary items. Any decrease in assets increases cash because selling equipment generates cash for the company. Any increase in liability increases cash because borrowing from a bank means there is more cash to spend, at least temporarily. Any decrease in liabilities decreases cash since it indicates that loans and debts were paid back in cash.

An increase in accounts receivable should decrease cash because net income sees a change that can't be taken into account until the company receives the cash. The following fiscal year will see that money paid, which results in a decrease in accounts receivable, assuming no new changes occur. An increase in inventory decreases cash because the company needs to spend more cash to buy enough supplies. An increase in accounts payable increases cash because the company has acquired cash in taking out loans and incurring debt. An increase in cash reserves decreases cash because it ties up cash into holding that will be used to pay for salary in the future.

Purchasing PPE, or plant, property, and equipment, immediately decreases cash. This is not shown on the income statement because the income statement reflects a company's operational efficiency, not one-time cash expenses. Such purchases are instead slowly amortized. Selling of equipment follows the same rules, in that it immediately generates cash but is not part of revenue on the income statement. Purchasing and selling equipment is not part of the company's everyday operations.

Issuing common stock suffers the same problem as PPE. Issuing stock for series A and B investors immediately generates cash, but it is an isolated event, divorced from revenue. Repurchasing stock from the public market decreases cash, but again is not seen on the income statement. PPE and stock exchanges are accounted for on the free cash flow statement.

FREE CASH FLOW, TERMINAL VALUE

Now that working capital changes have been calculated and other cash items have been accounted for, net income can be fully converted into free cash flow. Free cash flow is thus equal to net income plus changes in working capital plus PPE changes plus stock issuances (Table 7.6). The free cash flows can now be used to perform an NPV and MIRR analysis because they represent the actual cash received by the owners.

Free Cash Flow					
Year	<i>2011</i>	<i>2012</i>	<i>2013</i>	<i>2014</i>	<i>2015</i>
Net Income	(\$282.66)	\$6,095.27	\$12,920.41	\$12,860.34	\$13,082.72
<u>Cash Flow Statement</u>					
Cash From Operating Activities					
Plus: Depreciation	\$0.00	\$247.37	\$826.50	\$926.61	\$555.97
<i>Changes in Working Capital</i>					
(Increase)/Decrease in A/R	\$0.00	(\$1,232.88)	(\$1,232.88)	\$0.00	\$0.00
(Increase)/Decrease in Inv	\$0.00	(\$28.41)	(\$28.41)	\$0.00	\$0.00
(Increase)/Decrease in A/P	\$9.46	\$223.82	\$221.92	\$0.00	\$0.00
(Increase)/Decrease in C/R	(\$89.00)	(\$263.50)	\$0.00	\$0.00	\$0.00
<i>Total Change in Working Capital</i>	(\$79.54)	(\$1,300.97)	(\$1,039.37)	\$0.00	\$0.00
Cash From Investing Activities					
(Purchase)/Selling of Equipment	(\$1,236.83)	(\$2,153.57)	\$0.00	\$0.00	\$0.00
Cash From Financing Activities					
Issuance of Common Stock	\$1,707.93	\$2,825.42	\$0.00	\$0.00	\$0.00
Free Cash Flow	\$108.90	\$5,713.52	\$12,707.54	\$13,786.95	\$13,638.69

Table 7.6 Free Cash Flow from Net Income (\$ in thousands).

Before an NPV and MIRR analysis can be conducted, the terminal value of the company must be determined. This is done using the perpetuity growth model:

$$\text{Terminal Value} = \text{Cash Flow} * \frac{(1 + g)}{(r - g)} \quad \text{Equation 7.5}$$

In this equation, the cash flow used is the final free cash flow listed in Table 7.6. The variable g is the growth rate of the cash flow and the company. The variable r is the discount rate. The terminal value calculated with this formula represents the present value of all continuing

future cash flows. This calculation is theoretical because it assumes that the free cash flow is predictable.

NPV VALUATION

The net present value is used to predict the present worth of an investment assuming a certain level of future profitability. It is the sum of every cash flow, which for these purposes is the free cash flow plus the terminal value. This theoretical model is heavily dependent on the chosen discount rate. The discount rate varies by industry, based on the risk associated with that industry. Since the biotechnology is one of the riskier and more unpredictable industries, the discount rate used will be 25% and 30%. Both of these rates will be used to create a broad range of NPVs. The sales stage is significantly less risky than the research stage. Therefore, the discount rate during research is 50%. Table 7.7 details the calculations of present value and NPV using a discount rate of 25%.

NPV Valuation @25%							
Year	<i>2010</i>	<i>2011</i>	<i>2012</i>	<i>2013</i>	<i>2014</i>	<i>2015</i>	<i>Terminal Value</i>
T	<i>0</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>
Free Cash Flow	\$0.00	\$108.90	\$5,713.52	\$12,707.54	\$13,786.95	\$13,638.69	\$54,554.77
Discount Rate	0%	50%	25%	25%	25%	25%	25%
Present Value	\$0.00	\$72.60	\$3,656.65	\$6,506.26	\$5,647.14	\$4,469.13	\$14,301.21
Investments	(\$1,707.93)	(\$2,825.42)					
Discount Rate	0%	25%					
Present Value	(\$1,707.93)	(\$2,260.33)					
Net Present Value	\$30,684.71						

Table 7.7 NPV Analysis at 25% Discount Rate for series B (50% Rate for series A) (\$ in thousands).

To calculate NPV, all projected discounted free cash flows and the terminal value are converted into present values and then summed together. The present values of the initial investments are also calculated and subtracted from the original sum. It is important to take care when dealing with the two rounds of investment because they are not treated equally in these calculations. series A investment is equal to its present value because it is not a future cash flow, but an outflow that occurs in the present. The series B investment, however, occurs one year from the series A investment and is therefore discounted to find its present value. While the first free cash flow is discounted 50% due to the risk of series A, the series B investment is discounted by 25% because the risk involved in developing a prototype has been eliminated.

The terminal value of the company is also discounted to the present because it is a future projected value. The growth rate for these calculations was assumed to be 0%. The sum of all present values is the net present value, adding all positive future cash flows with negative investments. With our previously stated assumptions, our company's investors will make over \$30 million. To provide a complete financial picture, NPV analyses will be shown for myriad scenarios.

EQUITY SHARES

One way to measure profitability is through an IRR analysis, but this is difficult to perform due to the two sets of investors that participate in this project. These investments, done at two different times, force us to carefully track how much of the free cash flows go to series A investors and how much go to series B investors. First, we need to compare how much the two groups have invested in the project. series A investors put in about \$1.7 million at the project's outset and series B investors put in about \$2.85 million one year later. To account for the time difference, series A investment must be discounted forward by 50% to properly compare the two numbers.

Percentage of Investments			
	Investment	FV Investment	Percentage
series A Investors	\$1,707.93	\$2,561.90	47.6%
series B Investors	\$2,825.42	\$2,825.42	52.4%
Total		\$5,387.32	100%

Table 7.8 Compares series' Percentages of Investments (\$ in thousands).

Looking at Table 7.8, we can see that series B investors should keep 52.4% of the company once they enter, while series A investors keep the remaining 47.6%. series B investors will get 52.4% of the company, series A investors will keep 90% of the remaining 47.6%, and the scientists will keep the last 10% of that 47.6%. The scientists have a share in the company because of the initial trade they made in salary for company equity with the series A investors. The breakdown of company ownership is detailed in Table 7.9.

If NPV values are calculated for each year, we can see how each owner's share value increases over time. The calculation of equity percentage is essential for seeing how much of the free cash flows go to each group. This information can now be used to properly conduct a rate of return analysis.

Equity Percentage						
Year	2011	2012	2013	2014	2015	2016
Scientists	10.0%	4.8%	4.8%	4.8%	4.8%	4.8%
series A Investors	90.0%	42.8%	42.8%	42.8%	42.8%	42.8%
series B Investors	0.0%	52.4%	52.4%	52.4%	52.4%	52.4%
Total	100%	100%	100%	100%	100%	100%
NPV @25%	(\$1,635.33)	(\$239.02)	\$6,267.24	\$11,914.38	\$16,383.51	\$30,684.71
Share Values vs. Time						
Scientists	\$0.00	\$0.00	\$298.03	\$566.58	\$779.11	\$1,459.19
series A Investors	\$0.00	\$0.00	\$2,682.31	\$5,099.22	\$7,011.95	\$13,132.71
series B Investors	\$0.00	\$0.00	\$3,286.90	\$6,248.58	\$8,592.44	\$16,092.81
Total	\$0.00	\$0.00	\$6,267.24	\$11,914.38	\$16,383.51	\$30,684.71

Table 7.9 Equity Percentage and Share Values of All Company Owners (\$ in thousands).

MIRR ANALYSIS

A traditional rate of return analysis uses the internal rate of revenue (IRR) figure, but in this case it is a flawed model. An IRR analysis on our project would produce highly overstated yields since it entails large positive cash flows. The IRR analysis assumes that free cash flows are reinvested at the IRR being calculated. This does not describe our company, which instead has only two rounds of investment with no free cash flow being reinvested.

The alternative modified internal rate of return (MIRR) is used instead. For this calculation, we need to know the finance and reinvestment rates. The finance rate is annual percentage rate (APR) that our company would have to pay back lenders with if negative cash flows are generated. We can assume the APR to be the prime loan interest rate, which is predicted to stay at its current value of 3.25%.⁷⁸ The reinvestment rate is the interest rate owners would receive on any positive cash flows. We can assume this rate to be 0.5%, which is the current yield on a 6 month US

Treasury bill.⁷⁹ This is seen as the risk-free rate; although the biotechnology industry could assume a higher reinvestment rate because of its inherent risk, we will remain conservative in our assumptions. In this way, a more moderate and realistic profitability analysis is conducted.

MIRR calculations can be seen in Table 7.10. Investments are first defined; all free cash flows are then divided by the equity corresponding equity percentages. MIRR is then determined based on those numbers. Note that series B investors have one less cash flow term, since they entered the project one year after the series A investors.

MIRR Calculations							
Year		<i>2011</i>	<i>2012</i>	<i>2013</i>	<i>2014</i>	<i>2015</i>	<i>2016</i>
Free Cash Flows		\$108.90	\$5,713.52	\$12,707.54	\$13,786.95	\$13,638.69	\$54,554.77
Equity Percentage							
series A		90.0%	42.8%	42.8%	42.8%	42.8%	42.8%
series B		0.0%	52.4%	52.4%	52.4%	52.4%	52.4%
	Investment	Divided Free Cash Flows					
Cash Flows							
series A	(\$1,707.93)	\$98.01	\$2,445.32	\$5,438.68	\$5,900.66	\$5,837.21	\$23,348.82
series B	(\$2,825.42)	\$2,996.49	\$6,664.56	\$7,230.66	\$7,152.91	\$28,611.64	
series A MIRR		71%					
series B MIRR		80%					

Table 7.10 MIRR Calculation. The finance rate is 3.25% and the reinvestment rate is .5% (\$ in thousands).

A pro forma income statement, which summarizes the first part of the financial analysis, has been prepared for the *SynthSeq* process (Table 7.11). The statement shows free cash flows, NPV, and MIRR. There are two terminal values present on the sheet, one on top for a discount rate of 30% and one on the bottom for a discount rate of 25%.

Year	Case 1					2016
	2011	2012	2013	2014	2015	
Income Statement						
Revenue	\$0.00	\$15,000.00	\$30,000.00	\$30,000.00	\$30,000.00	\$30,000.00
Cost of Sales	(\$115.10)	(\$483.85)	(\$829.49)	(\$829.49)	(\$829.49)	(\$829.49)
Operating, SG&A Expenses	(\$356.00)	(\$4,110.00)	(\$6,810.00)	(\$6,810.00)	(\$6,810.00)	(\$6,810.00)
Depreciation	\$0.00	(\$247.37)	(\$826.50)	(\$926.61)	(\$555.97)	(\$555.97)
Pre-Tax Income	(\$471.10)	\$10,158.79	\$21,534.01	\$21,433.90	\$21,804.54	\$21,804.54
Tax @ 40%	\$188.44	(\$4,063.52)	(\$8,613.60)	(\$8,573.56)	(\$8,721.82)	(\$8,721.82)
Net Income	(\$282.66)	\$6,095.27	\$12,920.41	\$12,860.34	\$13,082.72	\$13,082.72
Cash Flow Statement						
Cash From Operating Activities						
Plus: Depreciation	\$0.00	\$247.37	\$826.50	\$926.61	\$555.97	\$555.97
Changes in Working Capital						
(Increase)/Decrease in A/R	\$0.00	(\$1,232.88)	(\$1,232.88)	\$0.00	\$0.00	\$0.00
(Increase)/Decrease in Inv	\$0.00	(\$28.41)	(\$28.41)	\$0.00	\$0.00	\$0.00
(Increase)/Decrease in A/P	\$9.46	\$223.82	\$221.92	\$0.00	\$0.00	\$0.00
(Increase)/Decrease in C/R	(\$89.00)	(\$263.50)	\$0.00	\$0.00	\$0.00	\$0.00
Total Change in Working Capital	(\$79.54)	(\$1,300.97)	(\$1,039.37)	\$0.00	\$0.00	\$0.00
Cash From Investing Activities						
(Purchase)/Selling of Equipment	(\$1,236.83)	(\$2,153.57)	\$0.00	\$0.00	\$0.00	\$0.00
Cash From Financing Activities						
Issuance of Common Stock	\$1,707.93	\$2,825.42	\$0.00	\$0.00	\$0.00	\$0.00
Free Cash Flow	\$108.90	\$5,713.52	\$12,707.54	\$13,786.95	\$13,638.69	\$13,638.69
NPV @ 30%						\$45,462.31
NPV @ 25%						\$54,554.77
TV @ NPV 30%						\$23,348.82
TV @ NPV 25%						\$28,611.64
% of Design Capacity						
Investment	0%	50%	100%	100%	100%	100%
Series A @ 42.8% of Equity	(\$1,707.93)	\$46.61	\$2,445.32	\$5,438.68	\$5,900.66	\$5,837.21
Series B @ 52.4% of Equity	(\$2,825.42)	\$2,996.49	\$6,664.56	\$7,230.66	\$7,152.91	\$7,152.91
NPV @ 30%	\$23,275.29					
NPV @ 25%	\$30,684.71					
Series A MIRR						71%
Series B MIRR						80%

Table 7.11 Pro Forma for a genome price of \$10,000

WHAT-IF SCENARIOS

Up until now, the financial analysis has been conducted using certain assumptions about the research stage, the scale up and sales stages, and the terminal stage. If these assumptions fail to hold, certain deviations will greatly affect the NPV and MIRR analyses. To study how profitability could change, different case scenarios were defined for these three different stages. Table 7.12 explains these scenarios.

What-If Scenarios		
Research Stage		
Best Case	S1	The startup stage takes one year, as planned.
Worst Case	S2	The startup stage takes two years.
Scale Up and Sales Stage		
Best Case	D1	The first year of sales has 50% capacity, with remaining three years at 100%.
Worst Case	D2	The first year of sales has 30% capacity, the next year's capacity is 70%, and the final two years have full capacity.
Terminal Stage		
Best Case	T1	The company stays profitable. Revenue stays constant.
Middle Case	T2	The company starts to decline due to rising competition. Price reduces by half after one year in sales.
Worst Case	T3	Improved rival technology cuts customer base immediately. Equipment is sold at 50% of face value.

Table 7.12 What-If Scenarios

Informing scenario 2 of the terminal stage, Figure 2.10 shows the recent trend in prices for a fully sequence genome, beginning with the Human Genome Project and ending with Knome.^{80, 81} This trend shows an exponential decrease in the price of genome sequencing. This trend is likely to continue, so a price decrease in future years is an important case to consider. However, no company has been able to produce a fully sequenced genome with acceptable accuracy rates for less than \$5,000; this price has not changed in many years. A new technology may help lower this

price in the next 5 years, but no company will undersell their own cost of sequencing the genome. As a result, we will use the price of \$5,000 as the lower limit on price reduction in this scenario.

Table 7.13 shows the complete NPV and MIRR summary for the twelve different scenarios.

Scenario Summary							
	Tree		Case	NPV @ 30%	NPV @ 25%	series A MIRR	series B MIRR
S1	D1	T1	1	\$23,275.29	\$30,684.71	71%	80%
		T2	2	\$10,708.19	\$14,244.15	52%	56%
		T3	3	\$10,904.06	\$12,808.41	44%	46%
	D2	T1	4	\$19,741.07	\$26,762.17	69%	77%
		T2	5	\$8,149.85	\$11,424.12	50%	53%
		T3	6	\$7,369.84	\$8,885.87	39%	39%
S2	D1	T1	7	\$17,412.60	\$24,095.13	58%	80%
		T2	8	\$8,102.63	\$11,346.04	43%	57%
		T3	9	\$7,896.27	\$9,794.09	36%	46%
	D2	T1	10	\$14,693.96	\$20,957.10	56%	77%
		T2	11	\$5,991.86	\$8,928.67	41%	53%
		T3	12	\$5,177.63	\$6,656.06	32%	39%

NPV and MIRR Range			
	High	Medium	Low
	(Average of Top 3)	(Median)	(Average of Bottom 3)
NPV	\$27,180.67	\$11,125.05	\$5,941.85
series A	66%	47%	35%
series B	79%	55%	42%

Table 7.13 Complete NPV and MIRR Analysis of the Twelve What-If Scenarios (\$ in thousands).

Ideally, we would like to provide an MIRR of 50% for series A investors and an MIRR of 30% for series B investors, or better. Except under worst case scenarios, these desired rates are achieved. A copy of all twelve scenarios' pro forma is provided in Appendix E. It is important to remember that all of these scenarios assumed the price of the genome to be \$10,000.

PRICE SENSITIVITY ANALYSIS

The same analysis can now be conducted for various genome prices to determine at what price the project is no longer profitable. Table 7.14 contains these results, with genome prices ranging from \$10,000 to \$1,000.

Price Sensitivity Analysis									
Genome Price	NPV			MIRR					
	High	Medium	Low	High A	Medium A	Low A	High B	Medium B	Low B
\$10,000.00	\$27,180.67	\$11,147.38	\$5,967.79	66%	47%	36%	79%	55%	42%
\$9,000.00	\$23,828.92	\$9,494.89	\$4,920.38	63%	45%	33%	75%	51%	39%
\$8,000.00	\$20,477.09	\$7,803.22	\$3,872.84	60%	41%	30%	72%	47%	35%
\$7,000.00	\$17,125.26	\$6,086.44	\$2,825.31	56%	38%	28%	65%	42%	31%
\$6,000.00	\$13,773.43	\$4,320.18	\$1,848.37	52%	34%	24%	59%	37%	26%
\$5,000.00	\$10,421.60	\$2,719.42	\$697.63	47%	31%	22%	53%	31%	22%
\$4,000.00	\$7,069.76	\$1,144.92	(\$453.12)	41%	23%	16%	45%	22%	16%
\$3,000.00	\$3,717.93	(\$511.99)	(\$1,565.57)	32%	14%	11%	34%	12%	7%
\$2,000.00	\$200.31	(\$1,919.85)	(\$3,025.96)	21%	4%	-5%	19%	-1%	-12%
\$1,000.00	(\$2,922.62)	(\$3,278.05)	(\$4,704.31)	-6%	-10%	-47%	-14%	-18%	-100%
Breakeven Price	\$1,935.86	\$3,363.66	\$4,393.76						

Table 7.14 Genome Price Sensitivity Analysis for NPV and MIRR (\$ in thousands).

Under best case scenarios, a price just over \$1,900 will allow us to break even. However, the worst case scenario must be considered to determine a reasonable price to charge for the genome. The break-even price is just under \$4,400. If we charge around \$5,000 per genome, we will still make sizable returns, especially since the chance of worst scenarios is low. This is a comfortable price, since our closest competitor, Knome, charges \$68,000 for the same service. It is important to note that these prices are not representative of the actual cost of genome sequencing. Since machinery is a one-time cost and operational costs depend on revenue, this per genome cost is calculated using labor, inventory, and rental costs alone. The resultant drop-dead price for genome assembly is accordingly determined to be \$746.50.

The MIRR yields we hope to achieve for series A and B investors are also important. Ideally, series A investors, who undertook large financial risks to fund this project, would receive 50% returns on their investment. The series B investors undertook less risk, so a 30% return on their investments is reasonable. Looking at Table 7.14, series A investors would see an MIRR of 52% if we sold each sequenced genome for \$6,000 under best conditions; series B investors would see an MIRR of 53% at this sales price. Although achieving an MIRR of 50% for series A is not possible except under best case scenarios, it is important to remember that the assumptions that went into the best case are not implausible. It is also important to remember that the actual MIRR of series A investment is greater than 52%, since the senior scientists who took salary hits in their first year with *SynthSeq* own a share of the total success. Bearing that in mind, *SynthSeq* will charge \$5,000 per sequenced genome.

With \$5,000 having been selected as the price of the *SynthSeq* genome, three additional case studies can be conducted to better gauge the financial success of the platform. The first of these cases studies the minimum throughput the company can absorb while still breaking even. The second considers how long the company must operate in order to break even, assuming expected throughput is achieved. The final case considers the potential for Big Pharma to utilize our sequencing process as part of phase III clinical trials. All of these cases will be performed under optimal conditions which assume one year for research and development, one year for scale-up, and maintaining original sales price for the entire operation period.

The first case considered involves a situation where unexpectedly low throughput in series A necessitates high capital costs in series B to purchase enough machinery and rental room to still meet minimum genome requirements. The throughput is divided by a value α , the same number that machinery, reagents, flow cells, and rental costs are multiplied by to recover overall throughput. Table 7.15 presents how NPV and series A and B MIRR vary with alpha.

Reduced Throughput Factor			
α	NPV @25%	series A	series B
1	\$11,947.25	52%	56%
2	\$9,676.93	42%	43%
3	\$7,175.38	35%	34%
4	\$4,673.82	29%	27%
5	\$2,172.27	24%	22%
6	(\$329.29)	20%	17%
5.86	\$0.00	21%	18%

Table 7.15 Reduced Throughput Alpha Factor's Effect on NPV (\$ in thousands).

Based on Table 7.15, *SynthSeq* will not lose money unless a minimum throughput is 5.86 times less than is currently anticipated with the technology. Considering our expected throughput, this corresponds to performing 35 times coverage on the whole genome just to get enough data to sequence the genome. This proves that even if some unexpected failure in throughput is discovered during series A, *SynthSeq* will most likely remain financially safe.

The second case considered here involves a situation where a competitor enters the market and forces early termination of the company in order to avoid failure from revenue loss. The reality of another start-up biotechnology company entering the market and undercutting *SynthSeq* is a possibility, considering the rapid innovations in this field each year. Table 7.15 shows how NPV and series A and B MIRR vary with operating years for the company.

Early Termination Effects			
Years Operation	NPV @25%	series A	series B
5	\$11,947.25	51%	54%
4	\$3,896.44	27%	26%
3	\$1,473.57	22%	19%
2	(\$1,615.74)	2%	-8%
2.44	\$0.00	14%	9%

Table 7.15 Early Termination's Effect on NPV (\$ in thousands).

Based on Table 7.15, *SynthSeq* will only lose money if the company is in existence for fewer than 2.5 years. Since the first year includes the research and design period, only part of the first year needs to be spent operating at 100% capacity to bring the company into a positive NPV. Therefore, *SynthSeq* is generally well protected against threats from competitors that would hinder profits before enough sales are made.

The final case in this financial analysis highlights a conceivable, near-future application of our technology that would contribute to personalized medicine. This scenario studies the financial feasibility of entering into an agreement with Big Pharma to provide genomic sequencing for patients in phase III clinical trials. Most phase III trials include between 500-3,000 patients and last for at least three years. At this point in the process, toxicity and dosage tests have been performed for a new drug on a small test group; phase III trials study exactly how effective a drug is for a larger population and whether or not it is better than current treatments.⁸² Genetic screening in phase III trials could serve to find a genetic link between effectiveness of a drug and/or study the effects of genetics on pharmacokinetics. It is possible to envision large pharmaceutical companies paying an additional amount for each patient in the clinical trial to conduct more personalized tests to see if certain genetic material affects drug response in patients. Since the average cost per patient of phase III clinical trials is \$26,000, the relative cost of sequencing each patient's genome might become economically justifiable at the low prices being proposed by *SynthSeq*, especially if a lower cost were negotiated for such a deal.

This financial analysis has already proven that the bottom line price for sequencing a genome is \$746.50, including inventory, rental, and labor fees. The purchased equipment from series A and B can handle the demands of sequencing 3,000 genomes per year. If a large pharmaceutical company agreed to pay for continual research costs as well as \$1,000 per sequenced genome, it would be highly beneficial to enter into contract with such a company.

SynthSeq would provide exclusive genome sequencing for a low price to the pharmaceutical company, while getting guaranteed business and protection from outside competitors for some negotiated time. With market security and large savings in research and development costs, such an agreement for phase III clinical trials would be highly profitable.

CONCLUSIONS

This financial analysis has provided strong evidence that our company is a profitable investment. Even when considering various less-than ideal situations, we see that investors will still receive acceptable returns on their investment. The sensitivity analysis shows that an ample price margin exists for our investors. Additional case studies further prove that profitability margins greatly outweigh the risk of failure of the *SynthSeq* business plan, as well as provide the potential for genome sequencing to enter the arena of personalized medicine.

It is important to remember that these financial models are hypothetical. They provide a guide for investors and reduce as much risk as possible, but they cannot accurately produce the future of the genome sequencing market. Financial risk exists in this project, but no more than it would for any reasonable fiscal endeavor. This company is a biotechnology firm that lacks many competitors due to the technology advances made in this project. Until the genomic market tempers with time, profit margins will remain high.

8. CONCLUSIONS

The goal of this undertaking was to design a process that would enable the sequencing of twelve entire genomes per day for 250 days per year with less than ten errors per million bases and incremental operating cost per genome of less than \$10,000 using less than \$25 million in start-up capital. The designed process satisfied all of these criteria, and we have shown that it does so with a lower error rate than competitors and at a lower cost. The technology that allows *SynthSeq* to provide such a highly accurate sequence with such high throughput and at such low prices is its unique approach to single molecule sequencing by synthesis.

The advent of non-Sanger sequencing approaches has opened up the door to sequencing technologies that are far superior to those originally used in the Human Genome Project. *SynthSeq's* method uses those advantages, and takes it one step further. Since our method is

asynchronous (i.e. nucleotide identification is done in between DNA synthesis steps), it makes it possible to easily identify and eliminate errors. Since the number of nucleotides that should have been added to the template strand is known exactly, insertion and deletion errors are eradicated. The stepwise addition of nucleotides also avoids errors that stem from homopolymers. Furthermore, since PCR is not employed, amplification bias is avoided. By quantifying the average deviation from expected values with the Z-score, data from template strands with systematic errors is discarded. These are fundamental advantages that *SynthSeq*'s sequencing procedure has over its competitors.

The cost for *SynthSeq* to sequence one genome – which includes labor, rental and inventory costs – is \$747. The company will charge customers \$5000; this price is well below competitors' prices. The only other company currently selling the service to sequence an entire human genome is Knome. It costs Knome \$5,000 to sequence one genome, and they charge customers \$68,500 per genome. Illumina claims it is able to sequence the genome for \$10,000, and the price to a customer, when it is released, is sure to be well above that. Thus, *SynthSeq* will provide an accurate genome sequence to customers at a fraction of the cost of its competitors and provide significant returns to investors. Charging \$5000 per genome, *SynthSeq* expects to have a positive NPV over four years. In the worst-case scenario, the NPV of the company is \$700 thousand and in the best-case scenario, the NPV of the company is \$10.5 million. This corresponds to a ROI for a series A investor of 22% in the worst-case scenario and 47% in the best-case scenario.

SynthSeq has applied cutting edge technologies in a novel way to create a process that performs better than any other on the market. We developed a genome sequencing process that provides an affordable, reliable product in a short period of time. *SynthSeq* has presented an attractive business plan that is poised to provide considerable returns on investment for its financial backers. Further, *SynthSeq* has developed a technology and a business plan that could help

revolutionize medical care by facilitating the average consumer's access to this crucial information. We firmly believe that the technology we have presented herein will significantly further medicine's ability to predict, prevent and treat disease.

ACKNOWLEDGEMENTS

We would like to take this opportunity to thank all of the people who contributed to our senior design project. Thanks especially to Dr. John C. Crocker for his sage advice and patience over the course of this endeavor. His extensive knowledge about DNA sequencing technology was vital to our understanding of the project details and its successful completion. We greatly appreciate Dr. Daeyeon Lee showing us his layer-by-layer laboratory, which helped us nail down the fundamentals of flow cell preparation. Thanks also to Parijat Sarkar, who was instrumental in helping us develop a simulation to model the efficiency of our random array. Finally, we would like to thank Professors Leonard A. Fabiano and Warren D. Seider for their involvement in the senior design process, as well as the industrial consultants who devoted their time to attend our weekly team meetings.

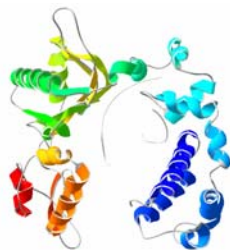
APPENDIX A: REAGENT SPECIFICATIONS

A.1. SAMPLE PREP REAGENTS

Species	Formula	Manufacturer	Price
PB buffer (concentrated)	N/A	Qiagen	100mL for \$68.00
PE buffer (concentrated)	N/A	Qiagen	100mL for \$68.00
EB buffer	N/A	Qiagen	250mL for \$27.00
T4 DNA Ligase buffer with 10mM ATP	N/A	New England BioLabs	100,000 units* for \$252.00
dNTP mix	N/A	Qiagen	800μL for \$153.00
T4 DNA Polymerase	Protein Structure**	New England BioLabs	750 units*** for \$244.00
Klenow Enzyme	Protein Structure+	New England BioLabs	1000 units++ for \$224.00
T4 Polynucleotide Kinase	Protein Structure+++	New England BioLabs	2500 units++++ for \$212.00
Klenow exo(3' to 5' exo minus)	N/A	Fisher Scientific	100 units for \$53.27

* One unit is defined as the amount of enzyme required to give 50% ligation of HindIII fragments of λ DNA (5' DNA termini concentration of 0.12 μ M, 300- μ g/ml) in a total reaction volume of 20 μ l in 30 minutes at 16°C in 1X T4 DNA Ligase Reaction Buffer.

** T4 Polymerase Structure:



*** One unit is defined as the amount of enzyme that will incorporate 10 nmol of dNTP into acid-precipitable material in a total reaction volume of 50 μ l in 30 minutes at 37°C (5) in 1X T4 DNA Polymerase Reaction Buffer with 33 μ M dNTPs including [3 H]-dTTP, 70 μ g/ml denatured herring sperm DNA and 50 μ g/ml BSA.

+ Klenow Enzyme Structure:



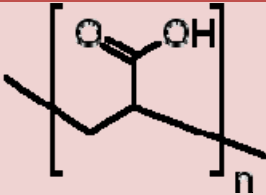
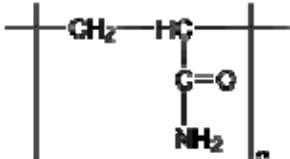
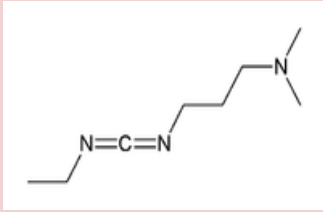
++ One unit is defined as the amount of enzyme required to convert 10 nmol of dNTPs to an acid-insoluble form in 30 minutes at 37°C.

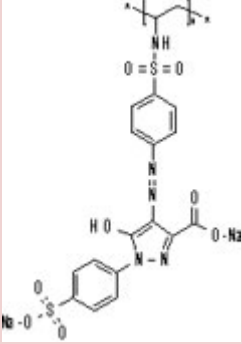
+++T4 Polynucleotide Kinase Structure:



**** One Richardson unit is defined as the amount of enzyme catalyzing the incorporation of 1 nmol of acid-insoluble [32 P] in a total reaction volume of 50 μ l in 30 minutes at 37°C in 1X T4 Polynucleotide Kinase Reaction Buffer with 66 μ M [γ - 32 P] ATP (5×10^6 cpm/ μ mol) and 0.26 mM 5'-hydroxyl-terminated salmon sperm DNA (1).

A.2. FLOW CELL PREP REAGENTS

Species	Formula /Structure	Manufacturer	Price
Poly(acrylic acid)		Sigma-Aldrich	10 g for \$219
Poly(acrylamide)		Sigma-Aldrich	N/A
MES buffer, pH 6	$C_6H_{13}NO_4S$	Fisher Scientific	125mL for \$45.00
Magnesium Chloride	$MgCl_2$	Sigma- Aldrich	1 mL for \$37.70
1-ethyl-3(3-dimethylaminopropyl)carbodiimide		Sigma- Aldrich	N/A
N-hydroxysulfosuccinimide	$C_4H_4NNaO_6S$	Sigma-Aldrich	1 g for \$323.00

polydT tails		Acros Organics	5 g for \$102.20
--------------	--	-------------------	---------------------

A.3. NUCLEOTIDE ADDITION REAGENTS

Species	Formula	Manufacturer	Price
dNTPs	Different for each tagged nucleotide	TriLink	\$20,000 custom synthesis
ACES buffer	$C_4H_{10}N_2O_4S$	Fisher Scientific	600g for \$643.81
Potassium Acetate	CH_3COOK	Fisher Scientific	2.5kg for \$633.23
Dithiothreitol	$HSCH_2CH(OH)CH(OH)CH_2SH$	Fisher Scientific	5 g for \$107.20
Manganese Acetate	$C_4H_6MnO_4$	Fisher Scientific	100g for \$191.21
Phi29 Polymerase	Protein structure*	New England BioLabs	1250 units** at 10,000units/mL for \$244.00

* Phi29 Polymerase Structure:



** One unit is defined as the amount of enzyme that will incorporate 0.5 pmol of dNTP into acid insoluble material in 10 minutes at 30°C.

A.4. FLUOROPHORE CLEAVE REAGENTS

Species	Formula	Manufacturer	Price
Thermopol I buffer	N/A	New England BioLabs	6mL for \$14.00
Sodium tetrachloropalladate (II)	Na_2PdCl_4	Sigma-Aldrich	1 g for \$102.50
$\text{P}(\text{PhSO}_3\text{Na})_3$	N/A	N/A	N/A
Tris HCl buffer	$\text{HSCH}_2\text{CH}(\text{OH})\text{CH}(\text{OH})\text{CH}_2\text{SH}$	Fisher Scientific	1100mL for \$308.40

APPENDIX B: QIAGEN PROTOCOLS

QIAquick PCR Purification Kit Protocol

using a microcentrifuge

This protocol is designed to purify single- or double-stranded DNA fragments from PCR and other enzymatic reactions (see page 8). For cleanup of other enzymatic reactions, follow the protocol as described for PCR samples or use the MinElute Reaction Cleanup Kit. Fragments ranging from 100 bp to 10 kb are purified from primers, nucleotides, polymerases, and salts using QIAquick spin columns in a microcentrifuge.

Important points before starting

- Add ethanol (96–100%) to Buffer PE before use (see bottle label for volume).
- All centrifugation steps are carried out at 17,900 \times g (13,000 rpm) in a conventional tabletop microcentrifuge at room temperature.
- Add 1:250 volume pH indicator I to Buffer PB (i.e., add 120 μ l pH indicator I to 30 ml Buffer PB or add 600 μ l pH indicator I to 150 ml Buffer PB). The yellow color of Buffer PB with pH indicator I indicates a pH of ≤ 7.5 .
- Add pH indicator I to entire buffer contents. Do not add pH indicator I to buffer aliquots.
- If the purified PCR product is to be used in sensitive microarray applications, it may be beneficial to use Buffer PB without the addition of pH indicator I.

Procedure

1. **Add 5 volumes of Buffer PB to 1 volume of the PCR sample and mix. It is not necessary to remove mineral oil or kerosene.**
For example, add 500 μ l of Buffer PB to 100 μ l PCR sample (not including oil).
2. **If pH indicator I has been added to Buffer PB, check that the color of the mixture is yellow.**
If the color of the mixture is orange or violet, add 10 μ l of 3 M sodium acetate, pH 5.0, and mix. The color of the mixture will turn to yellow.
3. **Place a QIAquick spin column in a provided 2 ml collection tube.**
4. **To bind DNA, apply the sample to the QIAquick column and centrifuge for 30–60 s.**
5. **Discard flow-through. Place the QIAquick column back into the same tube.**
Collection tubes are re-used to reduce plastic waste.
6. **To wash, add 0.75 ml Buffer PE to the QIAquick column and centrifuge for 30–60 s.**
7. **Discard flow-through and place the QIAquick column back in the same tube. Centrifuge the column for an additional 1 min.**

IMPORTANT: Residual ethanol from Buffer PE will not be completely removed unless the flow-through is discarded before this additional centrifugation.

8. Place QIAquick column in a clean 1.5 ml microcentrifuge tube.
9. To elute DNA, add 50 μ l Buffer EB (10 mM Tris-Cl, pH 8.5) or water (pH 7.0–8.5) to the center of the QIAquick membrane and centrifuge the column for 1 min. Alternatively, for increased DNA concentration, add 30 μ l elution buffer to the center of the QIAquick membrane, let the column stand for 1 min, and then centrifuge.

IMPORTANT: Ensure that the elution buffer is dispensed directly onto the QIAquick membrane for complete elution of bound DNA. The average eluate volume is 48 μ l from 50 μ l elution buffer volume, and 28 μ l from 30 μ l elution buffer.

Elution efficiency is dependent on pH. The maximum elution efficiency is achieved between pH 7.0 and 8.5. When using water, make sure that the pH value is within this range, and store DNA at -20°C as DNA may degrade in the absence of a buffering agent. The purified DNA can also be eluted in TE buffer (10 mM Tris-Cl, 1 mM EDTA, pH 8.0), but the EDTA may inhibit subsequent enzymatic reactions.

10. **If the purified DNA is to be analyzed on a gel, add 1 volume of Loading Dye to 5 volumes of purified DNA. Mix the solution by pipetting up and down before loading the gel.**

Loading dye contains 3 marker dyes (bromophenol blue, xylene cyanol, and orange G) that facilitate estimation of DNA migration distance and optimization of agarose gel run time. Refer to Table 2 (page 15) to identify the dyes according to migration distance and agarose gel percentage and type.

MinElute PCR Purification Kit Protocol

using a microcentrifuge

This protocol is designed to purify double-stranded DNA fragments from PCR reactions resulting in high end-concentrations of DNA (see page 12). Fragments ranging from 70 bp to 4 kb are purified from primers, nucleotides, polymerases, and salts using MinElute spin columns in a microcentrifuge.

Important points before starting

- Add ethanol (96–100%) to Buffer PE before use (see bottle label for volume).
- All centrifugation steps are carried out at 17,900 x g (13,000 rpm) in a conventional tabletop microcentrifuge at room temperature.
- Add 1:250 volume pH indicator I to Buffer PB (i.e., add 120 μ l pH indicator I to 30 ml Buffer PB or add 600 μ l pH indicator I to 150 ml Buffer PB). The yellow color of Buffer PB with pH indicator I indicates a pH of ≤ 7.5 .
- Add pH indicator I to entire buffer contents. Do not add pH indicator I to buffer aliquots.
- If the purified PCR product is to be used in sensitive microarray applications, it may be beneficial to use Buffer PB without the addition of pH indicator I.

Procedure

1. **Add 5 volumes of Buffer PB to 1 volume of the PCR reaction and mix. It is not necessary to remove mineral oil or kerosene.**
For example, add 250 μ l of Buffer PB to 50 μ l PCR reaction (not including oil).
2. **If pH indicator I has been added to Buffer PB, check that the color of the mixture is yellow.**
If the color of the mixture is orange or violet, add 10 μ l of 3 M sodium acetate, pH 5.0, and mix. The color of the mixture will turn to yellow.
3. **Place a MinElute column in a provided 2 ml collection tube in a suitable rack.**
4. **To bind DNA, apply the sample to the MinElute column and centrifuge for 1 min.**
For maximum recovery, transfer all traces of sample to the column.
5. **Discard flow-through. Place the MinElute column back into the same tube.**
6. **To wash, add 750 μ l Buffer PE to the MinElute column and centrifuge for 1 min.**
7. **Discard flow-through and place the MinElute column back in the same tube. Centrifuge the column for an additional 1 min at maximum speed.**
IMPORTANT: Residual ethanol from Buffer PE will not be completely removed unless the flow-through is discarded before this additional centrifugation.
8. **Place the MinElute column in a clean 1.5 ml microcentrifuge tube.**

9. To elute DNA, add 10 μ l Buffer EB (10 mM Tris-Cl, pH 8.5) or water to the center of the membrane, let the column stand for 1 min, and then centrifuge for 1 min.

IMPORTANT: Ensure that the elution buffer is dispensed directly onto the center of the membrane for complete elution of bound DNA. The average eluate volume is 9 μ l from 10 μ l elution buffer volume.

Elution efficiency is dependent on pH. The maximum elution efficiency is achieved between pH 7.0 and 8.5. When using water, make sure that the pH value is within this range, and store DNA at -20°C as DNA may degrade in the absence of a buffering agent. The purified DNA can also be eluted in TE buffer (10 mM Tris-Cl, 1 mM EDTA, pH 8.0), but the EDTA may inhibit subsequent enzymatic reactions.

10. If the purified DNA is to be analyzed on a gel, add 1 volume of Loading Dye to 5 volumes of purified DNA. Mix the solution by pipetting up and down before loading the gel.

Loading dye contains 3 marker dyes (bromophenol blue, xylene cyanol, and orange G) that facilitate estimation of DNA migration distance and optimization of agarose gel run time. Refer to Table 3 (page 15) to identify the dyes according to migration distance and agarose gel percentage and type.

APPENDIX C: EQUIPMENT SPECIFICATIONS

SAMPLE PREPARATION EQUIPMENT

PSS BioInstruments Magtration 12GC



Size	500(W) x 500(D) x 570(H) mm
Samples on board	1 to 12
Sample Volumes	100 or 200 μ L for DNA
Elution Volumes	50, 100 or 200 μ l
Purification time (1 · 12 samples)	30 minutes for DNA
Unit Price	\$40,000

Covaris s2 Shearer



Operating Environment	<ul style="list-style-type: none"> • Temperature: 15°C–25°C • Maximum humidity: 80% at 31°C; 50% at 40°C
System Components	<p>When purchased through Applied Biosystems, the Covaris S2 System consists of the following components:</p> <ul style="list-style-type: none"> • Covaris S2 Machine • Covaris 2-series machine holder for (one) 1.5 mL microfuge tube • Covaris 2-series machine holder for (one) 0.65 mL microfuge tube • Covaris 2-series machine holder for (one) tube(13 x 65 mm) • Dell™ Latitude laptop computer • VWR® Compact Chiller, Model 117-612
Power Requirements for Covaris S2 Machine	<ul style="list-style-type: none"> • US: 120 V (+/- 10%), 60 Hz (+/- 10%) • Japan: 100 V, 50–60 Hz • International: 220–240 V, 50–60 Hz • Power: ~300 W
Unit Price	\$55,000

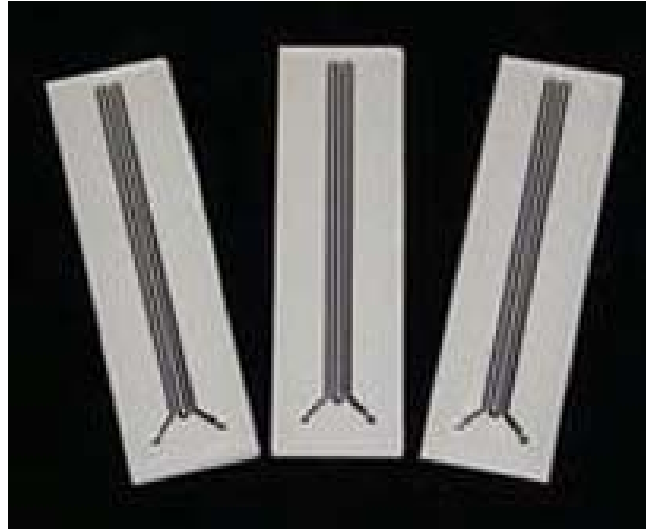
FLOW CELL PREPARATION EQUIPMENT

DS-50 SLIDE STAINER



Input Line Pressure	10-100 psia
Flow-Through Rate	Factory set via regulator Approx. 1L/min.
Recirculation Rate	Approximately 3.5L/min.
Wash Tank Capacity	Approximately 1.750 mL
Electrical Power	115VAC nominal, 50/60 Hz 220-240 VAC nominal
Power Consumption	3.6A peak
Unit Price	\$16,734

CiDRA® Precision Services SlipStream



Optical Flatness (all surfaces)	Less than 500nm/mm
Bottom & Top Plate Thickness	100um to 1cm
Top & Bottom Plate Thickness Tolerance	Less than +/-20um
Top & Bottom Plate Parallelism Tolerance	Less than +/-10um
Channel Thickness	25um to 700um
Channel Width	50um to >1cm
Channel Dimensional Accuracy	+/-30um
Surface Roughness (Ra) for All Optical Surfaces	<2nm
Maximum Operational Temperature	200°C
Maximum Pressure	20psi
Maximum Size	2,000cm ²
Unit Price	\$75.00

STATION ASSEMBLY EQUIPMENT

Harvard Apparatus Pump 22 Infusion Only Syringe Pump



# of Syringes	2
Accuracy	0.35 %
Average Linear Force	47 lbs
Communications	RS-232, TTL
Communications Level	2
Depth English	5.5 in
Depth Metric	14 cm
Display	LED, 3-1/2 digit, numeric
Flow Rate Maximum	55.1 ml/min
Flow Rate Minimum	0.002 μ l/hr
Height English	11 in
Height Metric	28 cm
Input Power	0.5 A, 30 W
Motor	1/4 microstepping, 0.9° micro step angle motor
Motor Stepper 1 Revolution of Lead Screw	3,200 at 1/4 stepping
Net Weight English	10 lb
Net Weight Metric	4.5 kg
Non Volatile Memory	Storage of all settings
Pressure	Standard
Pump Configuration	Standard
Pump Function	Infusion Only
Unit Price	\$3,006

ThorLabs MAX201 Motorized X/Y Stage



Travel	75 mm x 50 mm (3" x 2")
Maximum Speed	10 mm/s
Accuracy	1 μ m
Repeatability (Unidirectional)	1 μ m
Resolution (micro step size)	40 nm
Load Capacity (max)	5 kg
Cable Included	3 m Long
Compatible w/ Olympus IX71	✓
Compatible w/ Zeiss Axiovert 200	
Compatible w/ Nikon TE2000	✓
Unit Price	\$6,460

ThorLabs Microscan Piezo XYZ Controller SCXYZ100B



X- and Y-Axis Travel	100 μm
Z-Axis Travel	80 μm
Positioning Resolution	25 nm
Feedback	Strain Gauges
Load Capacity	100 g on Top Surface
Stiffness	0.4 N/ μm in X or Y
Resonant Frequency	>70 Hz
Max Voltage	75 V
Unit Price	\$9,990

Olympus IX81 Motorized Inverted Microscope with TIRFM Capabilities



Objective Lens	APO 60x Oil Immersion NA = 1.49 Magnification Changer = 0.5
TIRF	4 line lasers corresponding to four nucleotide analog excitation wavelengths
Unit Price	\$200,000

Andor iXon^{EM} +888 Back-Illuminated EMCCD



Active Pixels	1024 x 1024
Pixel Size (W x H; μm)	13 x 13
Image Area (mm)	13.3 x 13.3
Active Area Pixel Well Depth (e-, typical)	80,000
Gain Register pixel well depth (e-, typical)	730,000 \pm 2
Max Readout Rate (MHz)	10
Frame Rates (frames per sec)	8.9 (full frame)
Read Noise (e-)	< 1 to 47 @ 10 MHz
Unit Price	\$44,574



425 Sullivan Avenue, Suite #3, South Windsor, CT 06074 USA
 Tel: 1 (860) 290-9211, Fax: 1 (860) 290-9566

www.andor.com

Quotation Ref: O3ZEOA20000P

Issue Date: Friday, 05 February 2010

Valid Until: 30 Days beyond issue date

Ms. Stephanie Amato
 University of Pennsylvania
 Philadelphia, PA 19104
 USA
 Email : steph.m.amato@gmail.com

Technical Inquiries: Charles Fanghella
 Tel: (860) 290-9211
 Order Inquiries: Aida Dubois
 Tel: 860-290-9211 X 201

Item	Description	Part #	Qty	Unit Price	Discount%	SubTotal
1	1024x1024,13um,EMCCD,BV,10MHz,-100C	DU-888E-C00-#BV	1	48,000.00	12.00	42,240.00
2	IXON PCI Controller Card	CCI-23	1	1,800.00	12.00	1,584.00
3	Imaging Software	SOLIS (I)	1	1,500.00	50.00	750.00
4	Software Develop Kit - CCD PCI System	ANDOR-SDK-CCD	1	500.00	100.00	0.00

TOTAL :	44,574.00
----------------	------------------

All prices in USD

Please contact us if you have any further questions or would like to place an order.

All Prices in U.S. dollars, F.O.B South Windsor, CT

- Pricing is valid for 30 days from issue date.
- Andor's standard Terms and Conditions apply
- Estimated shipping date: 4-6 Weeks ARO
- Payment Terms : NET 30 days upon approval.
- Shipping method: Air Freight
- Warranty: 1 year parts and labor from date of shipment

Authorized by:

Charles Fanghella

THIS PRICING INFORMATION IS CONFIDENTIAL



IBM Power 750 Express Server



Power 750 Express rack-mount server

POWER7 processor modules - one per processor card	6-core 3.3 GHz or 8-core 3.0 GHz or 8-core 3.3 GHz or 8-core 3.55 GHz ¹
Sockets	1 to 4
Level 2 (L2) cache	256 KB per core
Level 3 (L3) cache	4 MB per core
Memory	8 GB ² to 512 GB of RDIMM DDR3 Active Memory Expansion
Solid State Drives (SSD)	Up to eight SFF drives
Disk drives	Up to eight SFF SAS drives
Disk capacity	Up to 2.4 TB
Media bays	Slimline for DVD-RAM Half height for tape drive or removable disk
PCI Adapter slots	Two PCI-X 2.0; Three PCI Express 8x
Unit Price	\$101,953

Netdisk 9010N Storage Server



Processor	Dual Core Intel processor running at 2GHz +
Main Memory	DDR2 1GB (Optional 2GB to 4GB)
Flash	2GB with NAS and OS (Optional for MS HomeServer)
LAN	Two Gigabit LAN ports
Network Service	DHCP client/server (default is DHCP client/Static) Network Protocol: CIFS/SMB, NFS, FTP iSCSI (optional)
Storage/RAID	Twelve 3.5" SATA/SATA-II Hot-Swap Drive 500GB, 1TB, 1.5TB or 2TB RAID 5, 6, 10, 1 and 0 with Hot Spare supported
USB	2x USB 2.0 connectors
Dimension	202mm x 424mm x 490mm (W)8.27" x (H)17.32" x (D)20"
System Power	400W PSU (Redundant P/S is optional)
Software	Operating System: Linux/FreeBSD, (Optional Microsoft Home Server) Backup and Recovery Client Support: Supports MS Windows 2000, Windows XP and Windows 2003, Linux and Mac
Unit Price	\$1,999

Dell PowerConnect 6224 24 Port Switch



Port Attributes	24 10/100/1000BASE-T auto-sensing Gigabit Ethernet switching ports 4 SFP combo ports for fiber media support 10 Gigabit Ethernet uplink modules (optional) 48Gbps Stacking module (optional) Auto-negotiation for speed, duplex mode and flow control Auto MDI/MDIX Port mirroring Flow-based port mirroring Broadcast storm control
Performance	Switch Fabric Capacity 136 Gb/s Forwarding Rate 95 Mpps Up to 8,000 MAC Addresses 256MB of CPU SDRAM 32MB of Flash Memory
Availability	Spanning Tree (IEEE 802.1D) and Rapid Spanning Tree (IEEE 802.1w) with Fast Link Support Multiple spanning trees (IEEE 802.1s) Supports Virtual Redundant Routing Protocol (VRRP) External redundant power support with PowerConnect EPS-470 (sold separately) Cable diagnostics Optical transceiver diagnostics
Unit Price	\$1,400

APPENDIX D: EFFICIENCY SIMULATION CODE

JAVA CODE

```
import java.util.Random;

public class RandomInteger1{

public static void main(String [] args){

// a x b = dimensions of the 2D arrays

int a = 1024;

int b = 1024;

// d is the number of templates

int d = 1100*1100;

//c = # of little boxes per big box

int c = 50;

//r = radius

int r = 15;

int [][] x = new int [a][b];

//y is the second 2D array, 0.25 um shifted over

int [][] y = new int [a][b+25];

double count = 0;

Random randomGenerator = new Random();

for (int i = 1; i <= d; i++)

{ int rint1 = randomGenerator.nextInt(a*c);

int rint2 = randomGenerator.nextInt(a*c);

if(!((rint1%c<r-2&&rint2%c<r-2)&&!(rint1%c<c-r-2&&rint2%c<r-2)&&!(rint1%c<r-2&&rint2%c<c-r-2)&&!(rint1%c<c-r-2&&rint2%c<c-r-2)&&x[rint1/c][rint2/c]==0)

{ x[rint1/c][rint2/c] = 1;

if(rint1%c<r-2&&rint1/c-1>=0)
```

```

{ x[rint1/c-1][rint2/c] = -1;}
if(rint1%c>c-r-2&&rint1/c+1<a)
{ x[rint1/c+1][rint2/c] = -1;}
if(rint2%c<r-2&&rint2/c-1>=0)
{ x[rint1/c][rint2/c-1] = -1;}
if(rint2%c>c-r-2&&rint2/c+1<a)
{ x[rint1/c][rint2/c+1] = -1;}
}
else{
if(!(rint1%c<r-2&&rint2%c<r-2)&&!(rint1%c<c-r-2&&rint2%c<r-2)&&!(rint1%c<r-
2&&rint2%c<c-r-2)&&!(rint1%c<c-r-2&&rint2%c<c-r-2)&&x[rint1/c][rint2/c]==1)
{ x[rint1/c][rint2/c] = -1;}
else{
if(rint1%c<(r/1.412)&&rint2%c<(r/1.412)&&rint1/c-1>=0&&rint2/c-1>=0)
{x[rint1/c-1][rint2/c-1] = -1;}
if(rint1%c<(r/1.412)&&rint2%c>(c-r/1.412)&&rint1/c-1>=0&&rint2/c+1<a)
{x[rint1/c-1][rint2/c+1] = -1;}
if(rint1%c>(c-r/1.412)&&rint2%c<(r/1.412)&&rint1/c+1<a&&rint2/c-1>=0)
{x[rint1/c+1][rint2/c-1] = -1;}
if(rint1%c>(c-r/1.412)&&rint2%c>(c-r/1.412)&&rint1/c+1<a&&rint2/c+1<a)
{x[rint1/c+1][rint2/c+1] = -1;}
}}
rint2 += 25;
if(!(rint1%c<r-2&&rint2%c<r-2)&&!(rint1%c<c-r-2&&rint2%c<r-2)&&!(rint1%c<r-
2&&rint2%c<c-r-2)&&!(rint1%c<c-r-2&&rint2%c<c-r-2)&&y[rint1/c][rint2/c]==0)
{ if(rint2/c>=25)
{y[rint1/c][rint2/c] = 1;}

```

```

if(rint1%c<r-2&&rint1/c-1>=0)
{ y[rint1/c-1][rint2/c] = -1;}
if(rint1%c>c-r-2&&rint1/c+1<a)
{ y[rint1/c+1][rint2/c] = -1;}
if(rint2%c<r-2&&rint2/c-1>=0)
{ y[rint1/c][rint2/c-1] = -1;}
if(rint2%c>c-r-2&&rint2/c+1<a)
{ y[rint1/c][rint2/c+1] = -1;}
}
else{
if(!(rint1%c<r-2&&rint2%c<r-2)&&!(rint1%c<c-r-2&&rint2%c<r-2)&&!(rint1%c<r-
2&&rint2%c<c-r-2)&&!(rint1%c<c-r-2&&rint2%c<c-r-2)&&y[rint1/c][rint2/c]==1)
{ y[rint1/c][rint2/c] = -1;}
else{
if(rint1%c<(r/1.412)&&rint2%c<(r/1.412)&&rint1/c-1>=0&&rint2/c-1>=0)
{y[rint1/c-1][rint2/c-1] = -1;}
if(rint1%c<(r/1.412)&&rint2%c>(c-r/1.412)&&rint1/c-1>=0&&rint2/c+1<a)
{y[rint1/c-1][rint2/c+1] = -1;}
if(rint1%c>(c-r/1.412)&&rint2%c<(r/1.412)&&rint1/c+1<a&&rint2/c-1>=0)
{y[rint1/c+1][rint2/c-1] = -1;}
if(rint1%c>(c-r/1.412)&&rint2%c>(c-r/1.412)&&rint1/c+1<a&&rint2/c+1<a)
{y[rint1/c+1][rint2/c+1] = -1;}
}}}
for(int j=0; j<=a-1;j++)
{ for(int k=0;k<=b-1; k++)
{ if(x[j][k]==1)

```

```
count++;}
}
for(int j=0; j<=a-1;j++)
{ for(int k=25;k<=b+24; k++)
{ if(y[j][k]==1&&x[j][k-1]!=1)
count++;}
}
System.out.println(count);
System.out.println(count/(a*b));
}}
```


APPENDIX E: FINANCIAL PRO FORMA

Year	Case 1					2016
	2011	2012	2013	2014	2015	
Income Statement						
Revenue	\$0.00	\$15,000.00	\$30,000.00	\$30,000.00	\$30,000.00	
Cost of Sales	(\$115.10)	(\$483.85)	(\$829.49)	(\$829.49)	(\$829.49)	
Operating, SG&A Expenses	(\$356.00)	(\$4,110.00)	(\$6,810.00)	(\$6,810.00)	(\$6,810.00)	
Depreciation	\$0.00	(\$247.37)	(\$826.50)	(\$926.61)	(\$555.97)	
Pre-Tax Income	(\$471.10)	\$10,158.79	\$21,534.01	\$21,433.90	\$21,804.54	
Tax @ 40%	\$188.44	(\$4,063.52)	(\$8,613.60)	(\$8,573.56)	(\$8,721.82)	
Net Income	(\$282.66)	\$6,095.27	\$12,920.41	\$12,860.34	\$13,082.72	
Cash Flow Statement						
Cash From Operating Activities	\$0.00	\$247.37	\$826.50	\$926.61	\$555.97	
Plus: Depreciation						
Changes in Working Capital						
(Increase)/Decrease in A/R	\$0.00	(\$1,232.88)	(\$1,232.88)	\$0.00	\$0.00	
(Increase)/Decrease in Inv	\$0.00	(\$28.41)	(\$28.41)	\$0.00	\$0.00	
(Increase)/Decrease in A/P	\$9.46	\$223.82	\$221.92	\$0.00	\$0.00	
(Increase)/Decrease in C/R	(\$89.00)	(\$263.50)	\$0.00	\$0.00	\$0.00	
Total Change in Working Capital	(\$79.54)	(\$1,300.97)	(\$1,039.37)	\$0.00	\$0.00	
Cash From Investing Activities						
(Purchase)/Selling of Equipment	(\$1,236.83)	(\$2,153.57)	\$0.00	\$0.00	\$0.00	
Cash From Financing Activities						
Issuance of Common Stock	\$1,707.93	\$2,825.42	\$0.00	\$0.00	\$0.00	
Free Cash Flow	\$108.90	\$5,713.52	\$12,707.54	\$13,786.95	\$13,638.69	\$45,462.31 \$54,554.77
TV @ NPV 30%						
TV @ NPV 25%						
% of Design Capacity	0%	50%	100%	100%	100%	
Investment						
Series A @ 42.8% of Equity	(\$1,707.93)	\$46.61	\$2,445.32	\$5,438.68	\$5,900.66	\$5,837.21
Series B @ 52.4% of Equity	(\$2,825.42)	\$2,996.49	\$6,664.56	\$7,230.66	\$7,152.91	\$28,611.64
NPV @ 30%	\$23,275.29					
NPV @ 25%	\$30,684.71					
				Series A MIRR	71%	
				Series B MIRR	80%	

Year	Case 2					2016
	2011	2012	2013	2014	2015	
Income Statement						
Revenue	\$0.00	\$15,000.00	\$15,000.00	\$15,000.00	\$15,000.00	
Cost of Sales	(\$115.10)	(\$483.85)	(\$829.49)	(\$829.49)	(\$829.49)	
Operating, SG&A Expenses	(\$356.00)	(\$4,110.00)	(\$4,110.00)	(\$4,110.00)	(\$4,110.00)	
Depreciation	\$0.00	(\$247.37)	(\$826.50)	(\$926.61)	(\$555.97)	
Pre-Tax Income	(\$471.10)	\$10,158.79	\$9,234.01	\$9,133.90	\$9,504.54	
Tax @ 40%	\$188.44	(\$4,063.52)	(\$3,693.60)	(\$3,653.56)	(\$3,801.82)	
Net Income	(\$282.66)	\$6,095.27	\$5,540.41	\$5,480.34	\$5,702.72	
Cash Flow Statement						
Cash From Operating Activities						
Plus: Depreciation	\$0.00	\$247.37	\$826.50	\$926.61	\$555.97	
Changes in Working Capital						
(Increase)/Decrease in A/R	\$0.00	(\$1,232.88)	\$0.00	\$0.00	\$0.00	
(Increase)/Decrease in Inv	\$0.00	(\$28.41)	(\$28.41)	\$0.00	\$0.00	
(Increase)/Decrease in A/P	\$9.46	\$223.82	\$0.00	\$0.00	\$0.00	
(Increase)/Decrease in C/R	(\$89.00)	(\$263.50)	\$0.00	\$0.00	\$0.00	
Total Change in Working Capital	(\$79.54)	(\$1,300.97)	(\$28.41)	\$0.00	\$0.00	
Cash From Investing Activities						
(Purchase)/Selling of Equipment	(\$1,236.83)	(\$2,153.57)	\$0.00	\$0.00	\$0.00	
Cash From Financing Activities						
Issuance of Common Stock	\$1,707.93	\$2,825.42	\$0.00	\$0.00	\$0.00	
Free Cash Flow	\$108.90	\$5,713.52	\$6,338.50	\$6,406.95	\$6,258.69	\$20,862.31 \$25,034.77
TV @ NPV 30%						
TV @ NPV 25%						
% of Design Capacity	0%	50%	100%	100%	100%	
Investment						
Series A @ 42.8% of Equity	(\$1,707.93)	\$46.61	\$2,445.32	\$2,712.80	\$2,742.10	\$2,678.65
Series B @ 52.4% of Equity	(\$2,825.42)	\$2,996.49	\$3,324.27	\$3,360.17	\$3,282.42	\$13,129.66
NPV @ 30%	\$10,708.19					52%
NPV @ 25%	\$14,244.15					56%
						Series A MIRR
						Series B MIRR

Year	Case 3					2016
	2011	2012	2013	2014	2015	
Income Statement						
Revenue	\$0.00	\$15,000.00	\$30,000.00	\$30,000.00	\$0.00	
Cost of Sales	(\$115.10)	(\$483.85)	(\$829.49)	(\$829.49)	(\$138.20)	
Operating, SG&A Expenses	(\$356.00)	(\$4,110.00)	(\$6,810.00)	(\$6,810.00)	(\$1,410.00)	
Depreciation	\$0.00	(\$247.37)	(\$826.50)	(\$926.61)	(\$555.97)	
Pre-Tax Income	(\$471.10)	\$10,158.79	\$21,433.01	\$21,433.90	(\$2,104.17)	
Tax @ 40%	\$188.44	(\$4,063.52)	(\$8,613.60)	(\$8,573.56)	\$841.67	
Net Income	(\$282.66)	\$6,095.27	\$12,920.41	\$12,860.34	(\$1,262.50)	
Cash Flow Statement						
Cash From Operating Activities						
Plus: Depreciation	\$0.00	\$247.37	\$826.50	\$926.61	\$555.97	
Changes in Working Capital						
(Increase)/Decrease in A/R	\$0.00	(\$1,232.88)	(\$1,232.88)	\$0.00	\$2,465.75	
(Increase)/Decrease in Inv	\$0.00	(\$28.41)	(\$28.41)	\$0.00	\$56.82	
(Increase)/Decrease in A/P	\$9.46	\$223.82	\$221.92	\$0.00	(\$443.84)	
(Increase)/Decrease in C/R	(\$89.00)	(\$263.50)	\$0.00	\$0.00	\$0.00	
Total Change in Working Capital	(\$79.54)	(\$1,300.97)	(\$1,039.37)	\$0.00	\$2,078.74	
Cash From Investing Activities						
(Purchase)/Selling of Equipment	(\$1,236.83)	(\$2,153.57)	\$0.00	\$0.00	\$0.00	
Cash From Financing Activities						
Issuance of Common Stock	\$1,707.93	\$2,825.42	\$0.00	\$0.00	\$0.00	
Free Cash Flow	\$108.90	\$5,713.52	\$12,707.54	\$13,786.95	\$1,372.20	\$1,695.20 TV @ NPV 30%
% of Design Capacity	0%	50%	100%	100%	100%	\$1,695.20 TV @ NPV 25%
Investment						
Series A @ 42.8% of Equity	(\$1,707.93)	\$46.61	\$2,445.32	\$5,438.68	\$5,900.66	\$587.29
Series B @ 52.4% of Equity	(\$2,825.42)	\$2,996.49	\$6,664.56	\$7,230.66	\$719.66	\$889.06
NPV @ 30%	\$10,904.06			Series A MIRR	44%	
NPV @ 25%	\$12,808.41			Series B MIRR	46%	

Year	Case 4					2016
	2011	2012	2013	2014	2015	
Income Statement						
Revenue	\$0.00	\$9,000.00	\$21,000.00	\$30,000.00	\$30,000.00	
Cost of Sales	(\$115.10)	(\$345.59)	(\$622.10)	(\$829.49)	(\$829.49)	
Operating, SG&A Expenses	(\$356.00)	(\$3,030.00)	(\$5,190.00)	(\$6,810.00)	(\$6,810.00)	
Depreciation	\$0.00	(\$247.37)	(\$826.50)	(\$926.61)	(\$555.97)	
Pre-Tax Income	(\$471.10)	\$5,377.05	\$14,361.40	\$21,433.90	\$21,804.54	
Tax @ 40%	\$188.44	(\$2,150.82)	(\$5,744.56)	(\$8,573.56)	(\$8,721.82)	
Net Income	(\$282.66)	\$3,226.23	\$8,616.84	\$12,860.34	\$13,082.72	
Cash Flow Statement						
Cash From Operating Activities						
Plus: Depreciation	\$0.00	\$247.37	\$826.50	\$926.61	\$555.97	
Changes in Working Capital						
(Increase)/Decrease in A/R	\$0.00	(\$739.73)	(\$986.30)	(\$739.73)	\$0.00	
(Increase)/Decrease in Inv	\$0.00	(\$17.05)	(\$22.73)	(\$17.05)	\$0.00	
(Increase)/Decrease in A/P	\$9.46	\$135.05	\$177.53	\$133.15	\$0.00	
(Increase)/Decrease in C/R	(\$89.00)	(\$263.50)	\$0.00	\$0.00	\$0.00	
Total Change in Working Capital	(\$79.54)	(\$885.22)	(\$831.49)	(\$623.62)	\$0.00	
Cash From Investing Activities						
(Purchase)/Selling of Equipment	(\$1,236.83)	(\$2,153.57)	\$0.00	\$0.00	\$0.00	
Cash From Financing Activities						
Issuance of Common Stock	\$1,707.93	\$2,825.42	\$0.00	\$0.00	\$0.00	
Free Cash Flow	\$108.90	\$3,260.22	\$8,611.84	\$13,163.33	\$13,638.69	\$45,462.31 TV @ NPV 30%
% of Design Capacity	0%	30%	70%	100%	100%	\$54,554.77 TV @ NPV 25%
Investment						
Series A @ 42.8% of Equity	(\$1,707.93)	\$46.61	\$1,395.34	\$3,685.77	\$5,633.76	\$5,837.21
Series B @ 52.4% of Equity	(\$2,825.42)	\$1,709.84	\$4,516.54	\$6,903.60	\$7,152.91	\$28,611.64
NPV @ 30%	\$19,741.07			Series A MIRR		69%
NPV @ 25%	\$26,762.17			Series B MIRR		77%

	Case 5					
Year	2011	2012	2013	2014	2015	2016
Income Statement						
Revenue	\$0.00	\$9,000.00	\$10,500.00	\$15,000.00	\$15,000.00	
Cost of Sales	(\$115.10)	(\$345.59)	(\$622.10)	(\$829.49)	(\$829.49)	
Operating, SG&A Expenses	(\$356.00)	(\$3,030.00)	(\$3,300.00)	(\$4,110.00)	(\$4,110.00)	
Depreciation	\$0.00	(\$247.37)	(\$826.50)	(\$926.61)	(\$555.97)	
Pre-Tax Income	(\$471.10)	\$5,377.05	\$5,751.40	\$9,133.90	\$9,504.54	
Tax @ 40%	\$188.44	(\$2,150.82)	(\$2,300.56)	(\$3,653.56)	(\$3,801.82)	
Net Income	(\$282.66)	\$3,226.23	\$3,450.84	\$5,480.34	\$5,702.72	
Cash Flow Statement						
Cash From Operating Activities						
Plus: Depreciation	\$0.00	\$247.37	\$826.50	\$926.61	\$555.97	
Changes in Working Capital						
(Increase)/Decrease in A/R	\$0.00	(\$739.73)	(\$123.29)	(\$369.86)	\$0.00	
(Increase)/Decrease in Inv	\$0.00	(\$17.05)	(\$22.73)	(\$17.05)	\$0.00	
(Increase)/Decrease in A/P	\$9.46	\$135.05	\$22.19	\$66.58	\$0.00	
(Increase)/Decrease in C/R	(\$89.00)	(\$263.50)	\$0.00	\$0.00	\$0.00	
Total Change in Working Capital	(\$79.54)	(\$885.22)	(\$123.82)	(\$320.33)	\$0.00	
Cash From Investing Activities						
(Purchase)/Selling of Equipment	(\$1,236.83)	(\$2,153.57)	\$0.00	\$0.00	\$0.00	
Cash From Financing Activities						
Issuance of Common Stock	\$1,707.93	\$2,825.42	\$0.00	\$0.00	\$0.00	
Free Cash Flow	\$108.90	\$3,260.22	\$4,153.52	\$6,086.62	\$6,258.69	\$20,862.31 TV @ NPV 30%
						\$25,034.77 TV @ NPV 25%
% of Design Capacity	0%	30%	70%	100%	100%	
	Investment				Divided Free Cash Flows	
Series A @ 42.8% of Equity	(\$1,707.93)	\$46.61	\$1,395.34	\$1,777.66	\$2,605.00	\$2,678.65
Series B @ 52.4% of Equity	(\$2,825.42)	\$1,709.84	\$2,178.34	\$3,192.17	\$3,282.42	\$13,129.66
NPV @ 30%	\$8,149.85					
NPV @ 25%	\$11,424.12					
				Series A MIRR		50%
				Series B MIRR		53%

Year	Case 6					2016
	2011	2012	2013	2014	2015	
Income Statement						
Revenue	\$0.00	\$9,000.00	\$21,000.00	\$30,000.00	\$0.00	
Cost of Sales	(\$115.10)	(\$345.59)	(\$622.10)	(\$829.49)	(\$138.20)	
Operating, SG&A Expenses	(\$356.00)	(\$3,030.00)	(\$5,190.00)	(\$6,810.00)	(\$1,410.00)	
Depreciation	\$0.00	(\$247.37)	(\$826.50)	(\$926.61)	(\$555.97)	
Pre-Tax Income	(\$471.10)	\$5,377.05	\$14,361.40	\$21,433.90	(\$2,104.17)	
Tax @ 40%	\$188.44	(\$2,150.82)	(\$5,744.56)	(\$8,573.56)	\$841.67	
Net Income	(\$282.66)	\$3,226.23	\$8,616.84	\$12,860.34	(\$1,262.50)	
Cash Flow Statement						
Cash From Operating Activities						
Plus: Depreciation	\$0.00	\$247.37	\$826.50	\$926.61	\$555.97	
Changes in Working Capital						
(Increase)/Decrease in A/R	\$0.00	(\$739.73)	(\$986.30)	(\$739.73)	\$2,465.75	
(Increase)/Decrease in Inv	\$0.00	(\$17.05)	(\$22.73)	(\$17.05)	\$56.82	
(Increase)/Decrease in A/P	\$9.46	\$135.05	\$177.53	\$133.15	(\$443.84)	
(Increase)/Decrease in C/R	(\$89.00)	(\$263.50)	\$0.00	\$0.00	\$0.00	
Total Change in Working Capital	(\$79.54)	(\$885.22)	(\$831.49)	(\$623.62)	\$2,078.74	
Cash From Investing Activities						
(Purchase)/Selling of Equipment	(\$1,236.83)	(\$2,153.57)	\$0.00	\$0.00	\$0.00	
Cash From Financing Activities						
Issuance of Common Stock	\$1,707.93	\$2,825.42	\$0.00	\$0.00	\$0.00	
Free Cash Flow	\$108.90	\$3,260.22	\$8,611.84	\$13,163.33	\$1,372.20	\$1,695.20 TV @ NPV 30%
						\$1,695.20 TV @ NPV 25%
% of Design Capacity						
	0%	30%	70%	100%	100%	
Investment						
Series A @ 42.8% of Equity	(\$1,707.93)	\$46.61	\$1,395.34	\$3,685.77	\$5,633.76	\$587.29
Series B @ 52.4% of Equity	(\$2,825.42)	\$1,709.84	\$4,516.54	\$6,903.60	\$719.66	\$889.06
NPV @ 30%	\$7,369.84			Series A MIRR		39%
NPV @ 25%	\$8,885.87			Series B MIRR		39%

Year	Case 7						
	2011	2012	2013	2014	2015	2016	2017
Income Statement							
Revenue	\$0.00	\$0.00	\$15,000.00	\$30,000.00	\$30,000.00	\$30,000.00	
Cost of Sales	(\$115.10)	(\$115.10)	(\$483.85)	(\$829.49)	(\$829.49)	(\$829.49)	
Operating, SG&A Expenses	(\$356.00)	(\$356.00)	(\$4,110.00)	(\$6,810.00)	(\$6,810.00)	(\$6,810.00)	
Depreciation	\$0.00	(\$247.37)	(\$395.79)	(\$668.19)	(\$831.63)	(\$555.97)	
Pre-Tax Income	(\$471.10)	(\$718.47)	\$10,010.37	\$21,692.32	\$21,528.88	\$21,804.54	
Tax @ 40%	\$188.44	\$287.39	(\$4,004.15)	(\$8,676.93)	(\$8,611.55)	(\$8,721.82)	
Net Income	(\$282.66)	(\$431.08)	\$6,006.22	\$13,015.39	\$12,917.33	\$13,082.72	
Cash Flow Statement							
Cash From Operating Activities							
Plus: Depreciation	\$0.00	\$247.37	\$395.79	\$668.19	\$831.63	\$555.97	
Changes in Working Capital							
(Increase)/Decrease in A/R	\$0.00	\$0.00	(\$1,232.88)	(\$1,232.88)	\$0.00	\$0.00	
(Increase)/Decrease in Inv	\$0.00	\$0.00	(\$28.41)	(\$28.41)	\$0.00	\$0.00	
(Increase)/Decrease in A/P	\$9.46	\$0.00	\$223.82	\$221.92	\$0.00	\$0.00	
(Increase)/Decrease in C/R	(\$89.00)	\$0.00	(\$263.50)	\$0.00	\$0.00	\$0.00	
Total Change in Working Capital	(\$79.54)	\$0.00	(\$1,300.97)	(\$1,039.37)	\$0.00	\$0.00	
Cash From Investing Activities							
(Purchase)/Selling of Equipment	(\$1,236.83)	\$0.00	(\$2,153.57)	\$0.00	\$0.00	\$0.00	
Cash From Financing Activities							
Issuance of Common Stock	\$1,707.93	\$0.00	\$2,825.42	\$0.00	\$0.00	\$0.00	
Free Cash Flow	\$108.90	(\$183.71)	\$5,772.89	\$12,644.21	\$13,748.96	\$13,638.69	\$45,462.31 TV @ NPV 30%
% of Design Capacity	0%	0%	50%	100%	100%	100%	\$54,554.77 TV @ NPV 25%
Investment							
Series A @ 42.8% of Equity	(\$1,707.93)	\$46.61	(\$78.63)	\$2,470.73	\$5,411.58	\$5,884.40	\$5,837.21
Series B @ 52.4% of Equity	(\$2,825.42)	\$3,027.63	\$6,631.35	\$7,210.74	\$7,152.91	\$28,611.64	\$23,348.82
NPV @ 30%	\$17,412.60			Series A MIRR	58%		
NPV @ 25%	\$24,095.13			Series B MIRR	80%		

Year	Case 8						2016	2017
	2011	2012	2013	2014	2015	2016		
Income Statement								
Revenue	\$0.00	\$0.00	\$15,000.00	\$15,000.00	\$15,000.00	\$15,000.00	\$15,000.00	
Cost of Sales	(\$115.10)	(\$115.10)	(\$483.85)	(\$829.49)	(\$829.49)	(\$829.49)	(\$829.49)	
Operating, SG&A Expenses	(\$356.00)	(\$356.00)	(\$2,760.00)	(\$4,110.00)	(\$4,110.00)	(\$4,110.00)	(\$4,110.00)	
Depreciation	\$0.00	(\$247.37)	(\$395.79)	(\$668.19)	(\$831.63)	(\$555.97)		
Pre-Tax Income	(\$471.10)	(\$718.47)	\$11,360.37	\$9,392.32	\$9,228.88	\$9,504.54		
Tax @ 40%	\$188.44	\$287.39	(\$4,544.15)	(\$3,756.93)	(\$3,691.55)	(\$3,801.82)		
Net Income	(\$282.66)	(\$431.08)	\$6,816.22	\$5,635.39	\$5,537.33	\$5,702.72		
Cash Flow Statement								
Cash From Operating Activities								
Plus: Depreciation	\$0.00	\$247.37	\$395.79	\$668.19	\$831.63	\$555.97		
Changes in Working Capital								
(Increase)/Decrease in A/R	\$0.00	\$0.00	(\$1,232.88)	\$0.00	\$0.00	\$0.00	\$0.00	
(Increase)/Decrease in Inv	\$0.00	\$0.00	(\$28.41)	(\$28.41)	\$0.00	\$0.00	\$0.00	
(Increase)/Decrease in A/P	\$9.46	\$0.00	\$112.86	\$110.96	\$0.00	\$0.00	\$0.00	
(Increase)/Decrease in C/R	(\$89.00)	\$0.00	(\$263.50)	\$0.00	\$0.00	\$0.00	\$0.00	
Total Change in Working Capital	(\$79.54)	\$0.00	(\$1,411.93)	\$82.55	\$0.00	\$0.00	\$0.00	
Cash From Investing Activities								
(Purchase)/Selling of Equipment	(\$1,236.83)	\$0.00	(\$2,153.57)	\$0.00	\$0.00	\$0.00	\$0.00	
Cash From Financing Activities								
Issuance of Common Stock	\$1,707.93	\$0.00	\$2,825.42	\$0.00	\$0.00	\$0.00	\$0.00	
Free Cash Flow	\$108.90	(\$183.71)	\$6,471.93	\$6,386.13	\$6,368.96	\$6,258.69	\$20,862.31	TV @ NPV 30%
% of Design Capacity	0%	0%	30%	70%	100%	100%	\$25,034.77	TV @ NPV 25%
Investment								
Series A @ 42.8% of Equity	(\$1,707.93)	\$46.61	(\$78.63)	\$2,769.91	\$2,733.19	\$2,725.84	\$2,678.65	\$10,714.60
Series B @ 52.4% of Equity	(\$2,825.42)	\$3,394.25	\$3,349.25	\$3,340.24	\$3,282.42	\$13,129.66		
NPV @ 30%	\$8,102.63			Series A MIRR	43%			
NPV @ 25%	\$11,346.04			Series B MIRR	57%			

Year	Case 9						
	2011	2012	2013	2014	2015	2016	2017
Income Statement							
Revenue	\$0.00	\$0.00	\$15,000.00	\$30,000.00	\$30,000.00	\$0.00	
Cost of Sales	(\$115.10)	(\$115.10)	(\$483.85)	(\$829.49)	(\$829.49)	(\$138.20)	
Operating, SG&A Expenses	(\$356.00)	(\$356.00)	(\$4,110.00)	(\$6,810.00)	(\$6,810.00)	(\$1,410.00)	
Depreciation	\$0.00	(\$247.37)	(\$395.79)	(\$668.19)	(\$831.63)	(\$555.97)	
Pre-Tax Income	(\$471.10)	(\$718.47)	\$10,010.37	\$21,692.32	\$21,528.88	(\$2,104.17)	
Tax @ 40%	\$188.44	\$287.39	(\$4,004.15)	(\$8,676.93)	(\$8,611.55)	\$841.67	
Net Income	(\$282.66)	(\$431.08)	\$6,006.22	\$13,015.39	\$12,917.33	(\$1,262.50)	

Cash Flow Statement

Cash From Operating Activities

Plus: Depreciation	\$0.00	\$247.37	\$395.79	\$668.19	\$831.63	\$555.97	
Changes in Working Capital							
(Increase)/Decrease in A/R	\$0.00	\$0.00	(\$1,232.88)	(\$1,232.88)	\$0.00	\$2,465.75	
(Increase)/Decrease in Inv	\$0.00	\$0.00	(\$28.41)	(\$28.41)	\$0.00	\$56.82	
(Increase)/Decrease in A/P	\$9.46	\$0.00	\$223.82	\$221.92	\$0.00	(\$443.84)	
(Increase)/Decrease in C/R	(\$89.00)	\$0.00	(\$263.50)	\$0.00	\$0.00	\$0.00	
Total Change in Working Capital	(\$79.54)	\$0.00	(\$1,300.97)	(\$1,039.37)	\$0.00	\$2,078.74	

Cash From Investing Activities

(Purchase)/Selling of Equipment	(\$1,236.83)	\$0.00	(\$2,153.57)	\$0.00	\$0.00	\$0.00	
Cash From Financing Activities							
Issuance of Common Stock	\$1,707.93	\$0.00	\$2,825.42	\$0.00	\$0.00	\$0.00	

Free Cash Flow	\$108.90	(\$183.71)	\$5,772.89	\$12,644.21	\$13,748.96	\$1,372.20	\$1,695.20
							TV @ NPV 30%
							\$1,695.20
							TV @ NPV 25%
							\$587.29

% of Design Capacity

	Investment	0%	50%	100%	100%	100%
Series A @ 42.8% of Equity	(\$1,707.93)	\$46.61	(\$78.63)	\$2,470.73	\$5,411.58	\$5,884.40
Series B @ 52.4% of Equity	(\$2,825.42)	\$3,027.63	\$6,631.35	\$7,210.74	\$719.66	\$889.06

NPV @ 30%

NPV @ 25%

Series A MIRR

Series B MIRR

\$7,896.27

\$9,794.09

36%

46%

Year	Case 10						2017
	2011	2012	2013	2014	2015	2016	
Income Statement							
Revenue	\$0.00	\$0.00	\$9,000.00	\$21,000.00	\$30,000.00	\$30,000.00	
Cost of Sales	(\$115.10)	(\$115.10)	(\$345.59)	(\$622.10)	(\$829.49)	(\$829.49)	
Operating, SG&A Expenses	(\$356.00)	(\$356.00)	(\$3,030.00)	(\$5,190.00)	(\$6,810.00)	(\$6,810.00)	
Depreciation	\$0.00	(\$247.37)	(\$395.79)	(\$668.19)	(\$831.63)	(\$555.97)	
Pre-Tax Income	(\$471.10)	(\$718.47)	\$5,228.63	\$14,519.71	\$21,528.88	\$21,804.54	
Tax @ 40%	\$188.44	\$287.39	(\$2,091.45)	(\$5,807.88)	(\$8,611.55)	(\$8,721.82)	
Net Income	(\$282.66)	(\$431.08)	\$3,137.18	\$8,711.83	\$12,917.33	\$13,082.72	
Cash Flow Statement							
Cash From Operating Activities							
Plus: Depreciation	\$0.00	\$247.37	\$395.79	\$668.19	\$831.63	\$555.97	
Changes in Working Capital							
(Increase)/Decrease in A/R	\$0.00	\$0.00	(\$739.73)	(\$986.30)	(\$739.73)	\$0.00	
(Increase)/Decrease in Inv	\$0.00	\$0.00	(\$17.05)	(\$22.73)	(\$17.05)	\$0.00	
(Increase)/Decrease in A/P	\$9.46	\$0.00	\$135.05	\$177.53	\$133.15	\$0.00	
(Increase)/Decrease in C/R	(\$89.00)	\$0.00	(\$263.50)	\$0.00	\$0.00	\$0.00	
Total Change in Working Capital	(\$79.54)	\$0.00	(\$885.22)	(\$831.49)	(\$623.62)	\$0.00	
Cash From Investing Activities							
(Purchase)/Selling of Equipment	(\$1,236.83)	\$0.00	(\$2,153.57)	\$0.00	\$0.00	\$0.00	
Cash From Financing Activities							
Issuance of Common Stock	\$1,707.93	\$0.00	\$2,825.42	\$0.00	\$0.00	\$0.00	
Free Cash Flow	\$108.90	(\$183.71)	\$3,319.59	\$8,548.52	\$13,125.34	\$13,638.69	\$45,462.31 TV @ NPV 30%
% of Design Capacity	0%	0%	30%	70%	100%	100%	\$54,554.77 TV @ NPV 25%
Investment							
Series A @ 42.8% of Equity	(\$1,707.93)	\$46.61	(\$78.63)	\$1,420.75	\$3,658.67	\$5,617.49	\$5,837.21
Series B @ 52.4% of Equity	(\$2,825.42)	\$1,740.98	\$4,483.33	\$6,883.67	\$7,152.91	\$28,611.64	\$23,348.82
NPV @ 30%	\$14,693.96			Series A MIRR			56%
NPV @ 25%	\$20,957.10			Series B MIRR			77%

Year	Case 11							
	2011	2012	2013	2014	2015	2016	2017	
Income Statement								
Revenue	\$0.00	\$0.00	\$9,000.00	\$10,500.00	\$15,000.00	\$15,000.00		
Cost of Sales	(\$115.10)	(\$115.10)	(\$345.59)	(\$622.10)	(\$829.49)	(\$829.49)		
Operating, SG&A Expenses	(\$356.00)	(\$356.00)	(\$2,220.00)	(\$3,300.00)	(\$4,110.00)	(\$4,110.00)		
Depreciation	\$0.00	(\$247.37)	(\$395.79)	(\$668.19)	(\$831.63)	(\$555.97)		
Pre-Tax Income	(\$471.10)	(\$718.47)	\$6,038.63	\$5,909.71	\$9,228.88	\$9,504.54		
Tax @ 40%	\$188.44	\$287.39	(\$2,415.45)	(\$2,363.88)	(\$3,691.55)	(\$3,801.82)		
Net Income	(\$282.66)	(\$431.08)	\$3,623.18	\$3,545.83	\$5,537.33	\$5,702.72		
Cash Flow Statement								
Cash From Operating Activities								
Plus: Depreciation	\$0.00	\$247.37	\$395.79	\$668.19	\$831.63	\$555.97		
Changes in Working Capital								
(Increase)/Decrease in A/R	\$0.00	\$0.00	(\$739.73)	(\$123.29)	(\$369.86)	\$0.00		
(Increase)/Decrease in Inv	\$0.00	\$0.00	(\$17.05)	(\$22.73)	(\$17.05)	\$0.00		
(Increase)/Decrease in A/P	\$9.46	\$0.00	\$68.47	\$88.77	\$66.58	\$0.00		
(Increase)/Decrease in C/R	(\$89.00)	\$0.00	(\$263.50)	\$0.00	\$0.00	\$0.00		
Total Change in Working Capital	(\$79.54)	\$0.00	(\$951.80)	(\$57.25)	(\$320.33)	\$0.00		
Cash From Investing Activities								
(Purchase)/Selling of Equipment	(\$1,236.83)	\$0.00	(\$2,153.57)	\$0.00	\$0.00	\$0.00		
Cash From Financing Activities								
Issuance of Common Stock	\$1,707.93	\$0.00	\$2,825.42	\$0.00	\$0.00	\$0.00		
Free Cash Flow	\$108.90	(\$183.71)	\$3,739.01	\$4,156.76	\$6,048.62	\$6,258.69	\$20,862.31	
							TV @ NPV 30%	
							TV @ NPV 25%	
% of Design Capacity	0%	0%	30%	70%	100%	100%		
	Investment						Divided Free Cash Flows	
Series A @ 42.8% of Equity	(\$1,707.93)	\$46.61	(\$78.63)	\$1,600.25	\$1,779.05	\$2,588.74	\$10,714.60	
Series B @ 52.4% of Equity	(\$2,825.42)	\$1,960.95	\$2,180.04	\$3,172.24	\$3,282.42	\$13,129.66		
NPV @ 30%	\$5,991.86						Series A MIRR 41%	
NPV @ 25%	\$8,928.67						Series B MIRR 53%	

Year	Case 12							2017
	2011	2012	2013	2014	2015	2016	2017	
Income Statement								
Revenue	\$0.00	\$0.00	\$9,000.00	\$21,000.00	\$30,000.00	\$0.00		
Cost of Sales	(\$115.10)	(\$115.10)	(\$345.59)	(\$622.10)	(\$829.49)	(\$138.20)		
Operating, SG&A Expenses	(\$356.00)	(\$356.00)	(\$3,030.00)	(\$5,190.00)	(\$6,810.00)	(\$1,410.00)		
Depreciation	\$0.00	(\$247.37)	(\$395.79)	(\$668.19)	(\$831.63)	(\$555.97)		
Pre-Tax Income	(\$471.10)	(\$718.47)	\$5,228.63	\$14,519.71	\$21,528.88	(\$2,104.17)		
Tax @ 40%	\$188.44	\$287.39	(\$2,091.45)	(\$5,807.88)	(\$8,611.55)	\$841.67		
Net Income	(\$282.66)	(\$431.08)	\$3,137.18	\$8,711.83	\$12,917.33	(\$1,262.50)		
Cash Flow Statement								
Cash From Operating Activities								
Plus: Depreciation	\$0.00	\$247.37	\$395.79	\$668.19	\$831.63	\$555.97		
Changes in Working Capital								
(Increase)/Decrease in A/R	\$0.00	\$0.00	(\$739.73)	(\$986.30)	(\$739.73)	\$2,465.75		
(Increase)/Decrease in Inv	\$0.00	\$0.00	(\$17.05)	(\$22.73)	(\$17.05)	\$56.82		
(Increase)/Decrease in A/P	\$9.46	\$0.00	\$135.05	\$177.53	\$133.15	(\$443.84)		
(Increase)/Decrease in C/R	(\$89.00)	\$0.00	(\$263.50)	\$0.00	\$0.00	\$0.00		
Total Change in Working Capital	(\$79.54)	\$0.00	(\$885.22)	(\$831.49)	(\$623.62)	\$2,078.74		
Cash From Investing Activities								
(Purchase)/Selling of Equipment	(\$1,236.83)	\$0.00	(\$2,153.57)	\$0.00	\$0.00	\$0.00		
Cash From Financing Activities								
Issuance of Common Stock	\$1,707.93	\$0.00	\$2,825.42	\$0.00	\$0.00	\$0.00		
Free Cash Flow	\$108.90	(\$183.71)	\$3,319.59	\$8,548.52	\$13,125.34	\$1,372.20	\$1,695.20	TV @ NPV 30%
							\$1,695.20	TV @ NPV 25%
% of Design Capacity								
Investment	0%	0%	30%	70%	100%	100%		
Series A @ 42.8% of Equity	(\$1,707.93)	\$46.61	(\$78.63)	\$1,420.75	\$3,658.67	\$5,617.49	\$587.29	\$725.53
Series B @ 52.4% of Equity	(\$2,825.42)	\$1,740.98	\$4,483.33	\$6,883.67	\$719.66	\$889.06		
NPV @ 30%	\$5,177.63			Series A MIRR	32%			
NPV @ 25%	\$6,656.06			Series B MIRR	39%			

REFERENCES

-
- ¹ Nyren, P., "The History of Pyrosequencing." *Methods Mol Biol.*, 2006; 373:1-14
- ² Shendure, Jay, and Hanlee Ji. "Next-generation DNA Sequencing." *Nature Biotechnology* 26.10 (2008): 1135-145.
- ³ Cass, Stephen. "Technology Review: Cheap DNA Sequencing Will Drive a Revolution in Health Care." *Technology Review: The Authority on the Future of Technology*. MIT, Mar.-Apr. 2010. Web. 21 Mar. 2010. <<http://www.technologyreview.com/biomedicine/24587/>>.
- ⁴ Shastry, BS. "Pharmacogenetics and the Concept of Individualized Medicine." *The Pharmacogenomics Journal* 16.21 (2006): 16-21.
- ⁵ "Pharmacogenomics: Medicine and the New Genetics." *Oak Ridge National Laboratory*. Web. 13 Apr. 2010. <http://www.ornl.gov/sci/techresources/Human_Genome/medicine/pharma.shtml>.
- ⁶ Ndegwa S. "Pharmacogenomics and warfarin therapy." *Issues Emerg Health Technol.* 2007 Oct;(104):1-8.
- ⁷ Cass, Stephen. "Technology Review: Cheap DNA Sequencing Will Drive a Revolution in Health Care." *Technology Review: The Authority on the Future of Technology*. MIT, Mar.-Apr. 2010. Web. 21 Mar. 2010. <<http://www.technologyreview.com/biomedicine/24587/>>.
- ⁸ BRCA1 and BRCA2: Cancer Risk and Genetic Testi - National Cancer Institute." *NCI Fact Sheet*. National Cancer Institute. Web. 01 Apr. 2010. <<http://www.cancer.gov/cancertopics/factsheet/Risk/BRCA>>.
- ⁹ Saglio G, Morotti A, Mattioli G, Messa E, Giugliano E, Volpe G, Rege-Cambrin G, Cilloni D. "Rational approaches to the design of therapeutics targeting molecular markers: the case of chronic myelogenous leukemia." *Ann N Y Acad Sci.* 2004 Dec;1028:423-31.
- ¹⁰ Van't Veer, Laura J., and Rene Bernards. "Enabling Personalized Cancer Medicine through Analysis of Gene-expression Patterns." *Nature* 452 (2008): 564-70.
- ¹¹ Leary, Rebecca J. et al."Development of Personalized Tumor Biomarkers Using Massively Parallel Sequencing." *Science Translational Medicine* 2.20 (2010): 1-7.

¹² Ng, Pauline C., Sarah S. Murray, Samuel Levy, and J. Craig Venter. "An Agenda for Personalized Medicine." *Nature* 461 (2009): 724-26.

¹³ "Phase 3 Clinical Trial Costs Exceed \$26,000 per Patient." *Life Sciences World*. 13 Oct. 2006. Web. 11 Apr. 2010. <<http://www.lifesciencesworld.com/life-science-news/view/11080?page=16>>.

¹⁴ Wolinsky, Howard. "The Thousand-dollar Genome: Genetic Brinkmanship or Personalized Medicine?". *Science and Society*. 8: 10, 900-903. (2007).

¹⁵ Mardis, Elaine R. "Anticipating the \$1,000 Genome". *Genome Biology*. (2006). 7:112 , p. 1-4.

¹⁶ Shendure, Jay, and Hanlee Ji. "Next-generation DNA Sequencing." *Nature Biotechnology* 26.10 (2008): 1135-145

¹⁷ Schuster, Stephan C. "Next-generation Sequencing Transforms Today's Biology." *Nature Methods* 5.1 (2008): 16-18

¹⁸ Gupta, Pushpendra K. "Single-molecule DNA Sequencing Technologies for Future Genomics Research." *Cell Press* 26.11 (2008): 602-11

¹⁹ Michael, Metzker L. "Sequencing Technologies - the next Generation." *Nature Reviews* 11 (2010): 31-46

²⁰ "Introducing HiSeq2000". *Illumina, Inc.* 2010.

²¹ Goren, Alon et al. "Chromatin Profiling by Directly Sequencing Small Quantities of Immunoprecipitated DNA". *Nature Methods*. 2009. DOI:10.1038, 1-5.

²² Shendure, Jay and Hanlee Ji. "Next-generation DNA Sequencing". *Nature Biotechnology*. 26:10, 1135-1145 (2008).

²³ "Genome Sequencer FLX System: More Applications, More Publications". *Roche Diagnostics / 454 Life Sciences*. www.roche-applied-science.com.

-
- ²⁴ Palmer, Roxanne. "Knome Offers Thriftier Gene Sequencing". 5/18/2009.
<http://www.xconomy.com/boston/2009/05/18/knome-offers-thriftier-gene-sequencing/>
- ²⁵ Knome Website. 4/2/2010. <http://www.knome.com/home/service/process.html>
- ²⁶ McBride, Ryan. "Knome Challenged to Keep in Step with Falling Genetic Sequencing Prices". 1/20/2010.
<http://www.xconomy.com/boston/2010/01/20/knome-challenged-to-keep-in-step-with-falling-genetic-sequencing-prices/2/>
- ²⁷ Ju, Jingyue et al. "Four-color DNA Sequencing by Synthesis Using Cleavable Fluorescent Nucleotide Reversible Terminators." *Proceedings of the National Academy of Science* 103.52 (2006): 19635-9640.
- ²⁸ Nanoink Inc. *DPN 5000 Desktop NanoFabrication System*. Nanoink. 28 Feb. 2010.
<<http://www.nanoink.net/d/DPN5000Brochure.pdf>>
- ²⁹ "High-Throughput DNA Purification with the Magtration 12GC." *Oragene DNA*. DNA Genotek. Web. 2 Feb. 2010. <http://www.dnagenotek.com/pdf_files/MKAN007_Magtration.pdf>.
- ³⁰ "Covaris DNA Shearing." K Biosciences. Web. 3 Feb. 2010.
<http://www.kbioscience.co.uk/instrumentation/acoustics/acoustics-DNA_shear.htm>.
- ³¹ "QIAquick Spin Handbook." QIAGEN, Mar. 2008. Web. 12 Feb. 2010.
<<http://www1.qiagen.com/>>.
- ³² "SureSelect Target Enrichment System Protocol." Agilent Technologies. Web. 5 Feb. 2010.
<http://www.opengonomics.com/Upload/file/PDF/Product_Literature/G336090010_SureSelect_Protocol_v1_2.pdf>.
- ³³ "MinElute Handbook." QIAGEN, Mar. 2008. Web. 12 Feb. 2010. <<http://www1.qiagen.com/>>.

-
- ³⁴ Yang, Sun Y., Jonas D. Mendelsohn, and Michael F. Rubner. "New Class of Ultrathin, Highly Cell-Adhesion-Resistant Polyelectrolyte Multilayers with Micropatterning Capabilities." *Biomacromolecules* 4.4 (2003)
- ³⁵ Laib, Stephan, and Brian D. MacCraith. "Immobilization of Biomolecules on Cycloolefin Polymer Supports." *Analytical Chemistry* 79.16 (2007): 6264-270
- ³⁶ "Protocol for Annealing Oligonucleotides." *Sigma Aldrich*. Web. 01 Apr. 2010. <<http://www.sigmaaldrich.com/life-science/custom-oligos/custom-dna/learning-center/annealing-oligos.html>>.
- ³⁷ Kim, Pilnam, and Keon Woo Kwon. "Soft Lithography for Microfluidics: a Review." *BIOCHIP JOURNAL* 2.1 (2008): 1-11. Web.
- ³⁸ "CiDRA Precision Services | SlipStream Flow Cell." *Precision Services | Custom Machining Services & Flow Cell Manufacturing*. Web. 06 Apr. 2010. <<http://www.cidraprecisionservices.com/flow-cell/technologies-capabilities/slipstream.html>>.
- ³⁹ Eddings, Mark A., and Michael A. Johnson. "Determining the Optimal PDMS–PDMS Bonding Technique for Microfluidic Devices." *J. Micromech. Microeng.* 18 (2008). Web.
- ⁴⁰ Alberts, Bruce. *Molecular Biology of the Cell*. New York: Garland Science, 2008. Print.
- ⁴¹ "Pfam: Family: DNA_pol_B (PF00136)." *Pfam: Home Page*. Web. 10 Apr. 2010. <<http://pfam.sanger.ac.uk/family/PF00136>>.
- ⁴² Blanco, Luis, and Margarita Salas. "Relating Structure to Function." *THE JOURNAL OF BIOLOGICAL CHEMISTRY* 271.15 (1996): 8509-512. Web.
- ⁴³ Tasara, T., B. Angerer, M. Damond, and H. Winter. "Incorporation of Reporter Molecule-labeled Nucleotides by DNA Polymerases. II. High-density Labeling of Natural DNA." *Nucleic Acids Res.* 10.31 (2003): 2636-646. Web.
- ⁴⁴ Guo, Jia, Lin Yu, and Jingyue Ju. "An Integrated System for DNA Sequencing by Synthesis Using Novel Nucleotide Analogues." *Acc. Chem. Res.* 10 (2009).

⁴⁵ Phillips, Rob, Jane Kondev, and Julie Theriot. *Physical Biology of the Cell*. New York: Garland Science, 2009. Print.

⁴⁶ Nucl. Acids Res. -- Datta and LiCata 31 (19): 5590 Figure KG774TB1." *Oxford Journals | Life Sciences | Nucleic Acids Research*. Web. 06 Apr. 2010. <<http://nar.oxfordjournals.org/cgi/content-nw/full/31/19/5590/GKG774TB1>>.

⁴⁷ "Highly Efficient DNA Synthesis by the Phage Phi 29 DNA Polymerase. Symmetrical Mode of DNA Replication. — JBC." *The Journal of Biological Chemistry*. Web. 06 Apr. 2010. <<http://www.jbc.org/content/264/15/8935.abstract>>.

⁴⁸ "Exponential Distribution -- from Wolfram MathWorld." *Wolfram MathWorld: The Web's Most Extensive Mathematics Resource*. Web. 06 Apr. 2010. <<http://mathworld.wolfram.com/ExponentialDistribution.html>>.

⁴⁹ Bowers, Jayson et al. "Virtual Terminator Nucleotides for Next-generation DNA Sequencing." *Nature Methods* 6.8 (2009): 593-95.

⁵⁰ Bowers, Jayson et al. "Virtual Terminator Nucleotides for Next-generation DNA Sequencing." *Nature Methods* 6.8 (2009): 593-95.

⁵¹ Lanrong, Bi, Dae Hyun Kim, and Jingyue Ju. "Design and Synthesis of a Chemically Cleavable Fluorescent Nucleotide, 3'-O-Allyl-dGTP-allyl-Bodipy-FL-510, as a Reversible Terminator for DNA Sequencing by Synthesis." *Journal of the American Chemical Society* 128.8 (2006): 2542-543.

⁵² Ju, Jingyue et al. "Four-color DNA Sequencing by Synthesis Using Cleavable Fluorescent Nucleotide Reversible Terminators." *Proceedings of the National Academy of Science* 103.52 (2006): 19635-9640.

⁵³ "Phi29 DNA Polymerase(M0269), Mesophilic DNA Polymerases, NEB." *New England Biolabs Homepage*. Web. 11 Apr. 2010. <<http://www.neb.com/nebecomm/products/productM0269.asp>>.

⁵⁴ Eid, J., and Et. Al. "Real-time DNA Sequencing from Single Polymerase Molecules." *Science* 5910.323 (2009). Web.

⁵⁵ "ScanArray Express Specifications Sheet". PerkinElmer Life Sciences.

www.perkinelmer.com/lifesciences.

⁵⁶ Encyclopedia of Laser Physics and Technology - Refractive Index, Index of Refraction." *RP Photonics Consulting - Laser and Amplifier Design, Nonlinear Optics, Fiber Optics, Fiber Lasers and Amplifiers, Ultrashort Pulses, Optoelectronics, Consultant, Training*. Web. 06 Apr. 2010. <http://www.rp-photonics.com/refractive_index.html>.

⁵⁷ Tothova, J., B. Brutovsky, and V. Lisy. "Monomer Motion in Single- and Double-stranded DNA Coils." Web.

⁵⁸ "Specialized Microscopy Techniques - Total Internal Reflection Fluorescence (TIRF)." *Olympus Microscopy Resource Center*. Web. 06 Apr. 2010.

⁵⁹ Snedeker, Joseph. "TIRFM Estimate." Message to the author. E-mail

⁶⁰ "What Is EMCCD Technology And How Is It Used." *EMCCD - Electron Multiplying Charge Coupled Device*. Web. 06 Apr. 2010. <http://www.emccd.com/what_is_emccd/>.

⁶¹ "IXon 888 EMCCD Camera - 1024 X 1024 Frame Transfer CCD Sensor." *Andor Technology - EMCCD SCMOS ICCD CCD Scientific Cameras, Spectrographs and Microscopy Systems*. Web. 13 Apr. 2010. <http://www.andor.com/scientific_cameras/ixon/models/default.aspx?iProductCodeID=2>.

⁶² Ju, Jingyue et al. "Four-color DNA Sequencing by Synthesis Using Cleable Fluorescent Nucleotide Reversible Terminators". *PNAS*. (2006). 103:52, 19635-19640.

⁶³ Reichman, Jay. *Handbook of Optical Filters for Fluorescence Microscopy*. Chroma Technology Corp, 2001.

⁶⁴ "Rhodamine 6G." *Laser Photomedicine and Biomedical Optics at the Oregon Medical Laser Center*. Web. 06 Apr. 2010. <<http://omlc.ogi.edu/spectra/PhotochemCAD/html/rhodamine6G.html>>.

⁶⁵ "Unravelling Sensitivity - Signal-To-Noise." *EMCCD - Electron Multiplying Charge Coupled Device*. Web. 06 Apr. 2010. <http://www.emccd.com/what_is_emccd/unraveling_sensitivity/Signal_to_Noise_in_CCDs/>.

⁶⁶ Zondervan, Rob, Kulzer Florian, Sergei B. Orlinski, and Michel Orrit. "Photoblinking of Rhodamine 6G in Poly(vinyl Alcohol): Radical Dark State Formed through the Triplet." *Journal of Physical Chemistry A* 107.35 (2003): 6770-776

⁶⁷ Pawley, James B. "Photobleaching." *Handbook of Biological Confocal Microscopy*. 3rd ed. New York, NY: Springer, 2006. 690-702

⁶⁸ Ju, Jingyue et al. "Four-color DNA Sequencing by Synthesis Using Cleavable Fluorescent Nucleotide Reversible Terminators." *Proceedings of the National Academy of Science* 103.52 (2006): 19635-9640.

⁶⁹ Pop, Mihai, et. al. (2004) Comparative genome assembly. Briefings in *Bioinformatics*, 5(3):237-248

⁷⁰ Eid, J., Turner, S., et al. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* 323: 133-138.

⁷¹ Avak Kahvejian, PhD. Director of Business Development. Helicos BioSciences Corporation. Personal Communication. 3/26/2010

⁷² Homer, Nils et al. "BFAST: An Alignment Tool for Large Scale Genome Resequencing". *PLoS ONE*. 2009. 4:11, e7767.

⁷³ "SourceForge.net: Bfast." *SourceForge.net: Find and Develop Open Source Software*. Web. 06 Apr. 2010. <http://sourceforge.net/apps/mediawiki/bfast/index.php?title=Main_Page>.

⁷⁴ Pushkarev, Dmitry et al. "Single-molecule Sequencing of an Individual Human Genome". *Nature Biotechnology*. 2009. 27:9, 847-850.

⁷⁵ "Archon Genomics X Prize." X Prize Foundation. Web. 16 Mar. 2010. <<http://genomics.xprize.org/>>.

⁷⁶ Loth, Richard. "What You Should Know About Inflation." *Investopedia.org*. Investopedia. Web. 2 Apr. 2010. <<http://www.investopedia.com/articles/01/021401.asp>>.

⁷⁷ Cummings Properties. Advertisement. *Cummings Properties*. Web. 30 Mar. 2010.

<<http://www.cummings.com/directory.html#sitedir>>.

⁷⁸ "Prime Loan Interest Rate Forecast." The Financial Forecast Center, 29 Mar. 2010. Web. 30 Mar.

2010. <<http://www.forecasts.org/prime.htm>>.

⁷⁹ "Forecast of 6 Month U.S. Treasury Bill Yield." The Financial Forecast Center, 7 Feb. 2010. Web.

30 Mar. 2010. <<http://www.forecasts.org/6mT.htm>>.

⁸⁰ "Knome Lowers Price of Full Genome From \$350,000 to \$99,000." *The Genetic Geneologist*. Nov.

2007. Web. Feb.-Mar. 2010. <<http://www.thegeneticgenealogist.com/2009/04/11/knome-lowers-price-of-full-genomefrom-350000-to-99000/>>.

⁸¹ Knome Challenged to Keep in Step with Falling Genetic Sequencing Prices." Xconomy. Web. 2 Apr.

2010. <<http://www.xconomy.com/boston/2010/01/20/knome-challenged-to-keep-in-step-with-falling-geneticsequencing-prices/>>.

⁸² Diamond, Scott L. "Drug Discovery and Development." University of Pennsylvania, Philadelphia. 8

Dec. 2009. Lecture.