



4-14-2009

High-Throughput, Whole-Genome Sequencing

Gregory J. Bittle
University of Pennsylvania

Boris N. Petkov
University of Pennsylvania

Yonghee Evan Rhee
University of Pennsylvania

Elliot C. Woods
University of Pennsylvania

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

 Part of the [Chemical Engineering Commons](#)

Bittle, Gregory J.; Petkov, Boris N.; Rhee, Yonghee Evan; and Woods, Elliot C., "High-Throughput, Whole-Genome Sequencing" (2009). *Senior Design Reports (CBE)*. 9.
http://repository.upenn.edu/cbe_sdr/9

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/cbe_sdr/9
For more information, please contact libraryrepository@pobox.upenn.edu.

High-Throughput, Whole-Genome Sequencing

Abstract

Since the completion of the Human Genome Project, research focusing on the consequence of known human genetic code has advanced by leaps and bounds. The development of personalized medicine, a field focused on enumerating the effects of individual genetic variations, termed SNPs, has become a reality for those researching the molecular basis of disease. With clinical correlates between genotype and prognosis becoming ever more common, the utility of personal genetic screening has become impossible to ignore. In this report, we present *PennBio*: a whole-genome sequencing company utilizing a novel single-molecule, real time sequencing-by-synthesis technology. Using unique zero-mode waveguides, which have revolutionized single-molecule detection, individual enzymes polymerizing novel phospholinked fluorescence labeled nucleotides can be observed as they sequence genomic template DNA. Modern optical techniques record these fragmented sequences, which are then analyzed by highly efficient alignment algorithms. A personal genomic code will ultimately allow consumers to be aware of their genetic predispositions as the medical community continues to discover them.

Disciplines

Chemical Engineering

Gregory J. Bittle
Boris N. Petkov
Yonghee Evan Rhee
Elliot C. Woods
220 S. 33rd Street
Philadelphia, PA 19104

April 14, 2009

Prof. Leonard A. Fabiano
Dr. John C. Crocker
220 S. 33rd Street
Philadelphia, PA 19104

Prof. Fabiano and Dr. Crocker,

After six months of thorough analysis and careful planning, we are prepared to present *PemBio* – a cutting-edge, whole-genome sequencing operation based upon the innovative SMRT chip technology. Current genetic screening providers focus almost exclusively on mutations already associated with particular diseases, effectively limiting their clinical relevance to the domain of contemporary genomic understanding. A whole-genome sequence, on the other hand, is an invaluable diagnostic and prognostic resource that becomes more effective with advances in the field molecular diagnostics.

An initial Series A investment of \$700,000 would be required to develop a working prototype, followed by \$2.6 million in Series B funding and a year to reach full production. While competitors' prices are well in excess of \$100,000, we have demonstrated a minimum sustainable price of \$2,000 while still offering Series A investors a 16% MIRR. At a more realistic price of \$4,000, Series A investors can expect a worst-case MIRR of 34%, with an ultimate NPV of no less than \$2.5 million.

We are confident that *PemBio* will deliver superior-quality, high-throughput whole-genome sequencing at a price that is expected to bring the promise of personalized medicine to as many as 3,000 customers each year.

Sincere regards,

Gregory J. Bittle

Yonghee Evan Rhee

Boris N. Petkov

Elliot C. Woods

PennBio Genomics

High-Throughput, Whole-Genome Sequencing

Gregory J. Bittle

Boris N. Petkov

Yonghee Evan Rhee

Elliot C. Woods

Project Advisor: Dr. John C. Crocker

Professor Leonard A. Fabiano

Department of Chemical and Biomolecular Engineering

April 14, 2009

Table of Contents

Abstract	1
I. Human Genomics	3
I.1 Current Screening Methods	7
I.2 The <i>PennBio</i> Approach	8
I.3 Technology Readiness Assessment	11
I.4 Customer Requirements	14
I.5 Unprecedented Throughput	15
II. The SMRT Chip Platform	19
II.1 Sequencing-by-Synthesis	20
II.2 The Zero-Mode Waveguide	23

II.3 Phospho-linked Fluorophores	25
III. Competitive Analysis	27
III.1 <i>Illumina</i>	28
III.2 <i>454 Sequencing</i>	29
III.3 <i>VisiGen</i>	30
IV. DNA Polymerization	33
IV.1 The Phi29 Polymerase	37
IV.2 Target DNA Isolation	39
IV.3 Priming and Random Hexamers	40
IV.4 Template Binding	42
IV.5 Immobilization of the Enzyme-Template Complex	42
IV.6 Phospholinked Fluorescent Nucleotides	47
IV.7 Polymerization Rate Comparison	48
IV.8 Dissociation of Synthesized Strands and Re-Complexing with New Primed Templates	49
IV.9 Error Rates and Possible Sources of Error	52
IV.10 Conclusions	54
V. Optical Detection of Single Molecules	55
V.1 Confocal Fluorescence Microscopy	56
V.2 High-Multiplex Confocal Microscopy	57
V.3 Two-Color Wide-Field Microscopy	58
V.4 Fluorescence Detection and Signal to Noise	66
V.5 Conclusions	70

VI. Genome Assembly	71
VI.1 The Human Genome Project	72
VI.2 Whole-Genome Shotgun Sequencing	73
VI.3 The <i>PennBio</i> Strategy	75
VI.4 The Coverage Problem	75
VI.4.1 The Probability that the Polymerase Reached the Base Position	77
VI.2.2 Probability of Misidentifying a Nucleotide	81
VI.2.3 The Complete Negative Binomial Estimate	81
VI.5 Simulation Overview	82
VI.5.1 Genome Generation, Fragmentation, and Polymerization	82
VI.5.2 Reassembly of Random Fragments	83
VI.6 Initial Simulation Results	84
VI.7 Final Error Rate Calculations	86
VI.8 Computing Time and Code Optimization	90
VI.9 Data Collection and Processing	92
VI.10 Computational Demands	93
VI.11 Multiprocessor Speed-Up and Amdahl's Law	95
VI.12 System Selection	96
VI.13 Conclusions	97

VII. Financial Analysis	99
VII.1 Market and Revenue Projection	101
VII.2 Costs, PPE, Depreciation	103
VII.3 Income Statement	111
VII.4 Working Capital	113
VII.5 Free Cash Flow, Terminal Value	117
VII.6 NPV Valuation	119
VII.7 Equity Shares	121
VII.8 MIRR Analysis	123
VII.9 What-If Scenarios	126
VII.10 Price Sensitivity Analysis	128
VII.11 Growth Case	129
VII.11 Conclusions	129
VIII. Conclusions	131
VIII.1 Acknowledgements	133
Appendix A: Reagent Specifications	135
A.1 Sequencing-by-Synthesis Reagents	136
A.2 Deoxyribonucleotide	
Fluoropentaphosphate Reagents	137
A.3 Deoxyribonucleotide Fluoropentaphosphates	141
A.4 Proteins	142
Appendix B: DNA Extraction and Isolation	145

Appendix C: Synthesis of Phospho-linked Nucleotide Pentaphosphates	149
Appendix D: Equipment Specifications	153
D.1 Microscopes and Peripherals	154
D.2 EMCCD Cameras	155
D.3 Nanopositioning Stages	157
D.4 Signal Processing Servers	158
D.4 Reassembly Servers	159
Appendix E: MATLAB Simulation Code	161
E.1 Program Framework	162
E.2 Local Alignment	166
E.3 Vote Counting	167
E.4 Base Assignment	168
E.4 Representative Sensitivity Analysis	169
E.5 Dual-Camera Peak Identification	170
Appendix F: Financial <i>Pro Forma</i>	171
References	185

Abstract

Since the completion of the Human Genome Project, research focusing on the consequence of known human genetic code has advanced by leaps and bounds. The development of personalized medicine, a field focused on enumerating the effects of individual genetic variations, termed SNPs, has become a reality for those researching the molecular basis of disease. With clinical correlates between genotype and prognosis becoming ever more common, the utility of personal genetic screening has become impossible to ignore. In this report, we present *PennBio*: a whole-genome sequencing company utilizing a novel single-molecule, real time sequencing-by-synthesis technology. Using unique zero-mode waveguides, which have revolutionized single-molecule detection, individual enzymes polymerizing novel phospho-linked fluorescence labeled nucleotides can be observed as they sequence genomic template DNA. Modern optical techniques record these fragmented sequences, which are then analyzed by highly efficient alignment algorithms. A personal genomic code will ultimately allow consumers to be aware of their genetic predispositions as the medical community continues to discover them.

I. Human Genomics

Traditionally, the clinical research at the heart of modern evidence-based medicine has been performed across large patient populations. While great efforts are undertaken to ensure that these sample groups are homogeneous, the confounding effects of individual variations are impossible to avoid and particularly difficult to model, imbuing the conclusions of any study with an often-significant degree of uncertainty. Rare conditions that affect only a small portion of the population, or that are not easily identified through clinical observation, are most prone to such mischaracterization – the very conditions that require the most specialized treatment, and lead to the most morbid outcomes. Physicians and their patients are ultimately tasked not only with assessing the likelihoods of these rare events, but also with evaluating the applicability of available evidence to the specific situations at hand. The unfortunate consequences of a chance

misjudgment are commonly seen in adverse drug reactions, treatment inefficacy, and late-stage disease detection.

Over the past few decades, however, the rapid growth of genomics as a medically relevant discipline has been the catalyst for dramatic advances in the practice of *molecular diagnostics*. Through a greater understanding of how DNA is interpreted (transcriptomics) and how its protein products affect cellular processes (proteomics), scientists are, for the first time, beginning to associate abnormal physical conditions with their genetic precursors. This clinical-to-genomic mapping paves the way for a new diagnostic and prognostic paradigm – the idea of *personalized medicine*. Recent translational research into the molecular basis for complex and widespread diseases such as cancer, heart disease, and diabetes has demonstrated the effectiveness of this approach. The physician now makes informed decisions based on the patient's genetic make-up rather than population-average data, allowing for increased confidence and more favorable outcomes, overall.

One of the most visible products of personalized medicine has been the field of *pharmacogenetics*, which seeks to characterize the interplay between individuals' genotypes and their responses to specific medications. Therapeutic parameters including dosage, side effects, and efficacy can be predicted based on genomic data, allowing for the tailoring of an individualized treatment regimen. The field of pharmacogenetics also sees the potential of more rapidly identifying novel drugs for common use in humans.¹ An example of the successful application of pharmacogenetic principles is the use of warfarin. Warfarin is an anti-coagulant

used in treating thrombosis, and is the most widely used anti-coagulant used in North America.² Its utilization, however, is limited by potentially severe adverse reactions to the drug. Genetic variations in the genes *CYP2C9* and *VKORC1* have been shown to be correlated with these dangerous side-effects. With this novel research, new statistical models involving age, weight, gender, and *genotype* are used to gauge dosage for individuals being prescribed warfarin.³

Cancer treatment has long used genetic strategies in the assessment of an individual's stage of tumorigenesis, enabling the selection of more effective treatment regimens. This early form of personalized medicine is growing rapidly as techniques for characterizing the genetic aberrations present in the cancer cells become more easily performed. Cancers, in general, are defined by cells which have lost control of their genetic regulators, largely due to somatic mutations, allowing them to proliferate unchecked. These mutations can vary even within certain types of cancer, and the field of cancer genetics seeks to correlate the genetic mutations to prognostic outcomes. With genetic screening of the cancer being targeted, physicians can start to make prognostic and therapeutic decisions which more accurately suit the specific mutations which have occurred.^{4,5}

An example of this customized treatment based on characterized oncogenesis is the prescription of Gleevec in the treatment of chronic myeloid leukemia (CML). Some 95% of CML cases are genetically characterized by a fusion of the *BCR* and *ABL* genes, forming the *BCR-ABL* fusion protein. Gleevec targets the *ABL* kinase activity, and is an effective treatment

for those with the *BCR-ABL* translocation, making genetic testing an often-used diagnostic test in treating CML.⁶

The application of genetic screening to treat cancer goes beyond the acute stage of the disease. Several types of cancer, called familial cancers, are passed down through generations from parent to child, and account for 5-10% of cancers seen in current oncology wards. Genetic screening can assess an individual's risk, given that the genetic mutations associated with a certain disease have been characterized. With this knowledge at hand, preventative measures can be taken to minimize the chance of the oncogenesis. Mutations in the *BRCA1* and *BRCA2* genes, for example, are associated with increased risk of developing certain breast and ovarian cancers. Discovery of a high-risk mutation in one or both of these genes may prompt an individual to seek prophylactic treatment such as mastectomy or removal of the ovaries.⁷

Oncology is not the only medical field which can benefit from the development of personalized medicine-based preventative treatment. Conditions such as heart disease, diabetes and other complex syndromes such as high blood pressure and high cholesterol have been shown to have significant genetic corollaries. Though both type I and type II diabetes are both suspected of having genetic correlations, type II shows much stronger hereditary influence. Several genes have been shown to be associated with the development of type II diabetes and their genetic variations are being more specifically characterized as research continues.^{8,9}

Personalized medicine is the new frontier in medical research, and in clinical practice. From acute treatment decisions based on the molecular basis of disease to preventative medicine based on pre-onset genetic screening, the potential benefits to society of continued genetic and translational research are immeasurable.

I.1 Current Screening Methods

Contemporary genetic testing services – such as *23andMe* and *Navigenics* – provide interested consumers with information about known biomarkers, that is, physically relevant genetic abnormalities. Numerous traits from eye color to predisposition for certain diseases to body types can be predicted with varying degrees of confidence. These services do not attempt to sequence the individual's complete genome; instead, they identify single-nucleotide polymorphisms (SNPs – pronounced “snips”) in which certain markers are exclusively targeted. *23andMe*, for example, offers 500,000 SNPs associated with 109 traits at \$399 per individual. This gives the consumer price to be 0.08 cents per SNP and \$3.67 per trait.

Databases such as NCBI's *dbSNP* have been established to catalogue the continuously growing number of reported SNPs discovered in current biomedical research. As of April, 2009, the *dbSNP* had over 16,600,000 SNPs identified with many millions added every six months, but few of these have been cited clinically and even fewer have been identified with definite phenotypes. These databases, however, are science intensive, and would intimidate anyone not extraordinarily familiar with molecular genetics, and the consumer is undoubtedly more concerned with their phenotype – disease predispositions, for example – than their actual

genotype – that base 53,457 in their *HOXA9* gene is an adenine. Websites such as *SNPedia.com* seek to connect clinical correlates to known SNP variants. *SNPedia.com* currently offers clinical phenotypic correlates on 5,043 SNPs, covering 307 traits, and is growing steadily.

The limitation of these services lies in the fact that an estimated 10% of the 100 million SNPs predicted to be present in the human genome have been identified, and an even smaller fraction of those have been associated with some known physical manifestation.¹⁰ The SNP-focused genomics products therefore become obsolete as new discoveries are made, as they are restricted in their usefulness to SNPs that were known at the time of the analysis. Whole-genome sequencing operations offer a comprehensive genomic sequence which will be clinically relevant for the rest of the consumer lifetime. By delivering an entire genomic sequence, the customer has been granted access to at least 16 million known SNPs, and over 5,000 correlated SNPs which make them who they are.

I.2 The *PennBio* Approach

PennBio is a newcomer to the biotechnology industry committed to providing exceptionally accurate whole-genome sequencing. The primary goal of the project is to achieve unprecedented genome throughput at a low cost. Specifically, operations have been designed to meet an estimated demand for 3000 human genomes per year at a cost of \$10,000 or less, with a maximum start-up cost not to exceed \$25 million. The rapid production and low price targets were inspired by the Archon X Prize competition, which offers a \$10 million award to the first team to be able to meet these specified goals.

A brief overview of the *PennBio* business plan is presented below in **Table I.1**, providing an outline of the goals, scope, deliverables, and timetable for developing a system for high throughput genomic sequencing.

Project Name	PennBio: <i>High-Throughput Genomic Sequencing</i>
Project Champion	Dr. John Crocker
Project Leaders	Gregory Bittle, Boris Petkov, Evan (Yonghee) Rhee, and Elliot Woods
Specific Goals	Sequence 3000 human genomes per year for no more than \$10,000 per genome sequenced, and with a start-up cost of \$25 million or less.
Project Scope	<p><i>In-Scope:</i></p> <ul style="list-style-type: none"> ◇ Identify and evaluate high-throughput sequencing techniques ◇ Apply the most promising technology to provide in-house whole-genome sequencing ◇ Characterize the biological, optical, and computational methods underlying this technology ◇ Select appropriate equipment and staff ◇ Develop a viable business model, centered around the above production level, investment, and a genome price <p><i>Out-of-Scope:</i></p> <ul style="list-style-type: none"> ◇ Manufacturing of Zero-Mode Waveguides (ZMWs) ◇ Focused screening (such as SNP screening) ◇ The provision of genetic counseling or medical advice
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 5-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 3-5 with concurrent R&D ◇ Liquidate or sell the company at the end of year 5

Table I.1 Project Charter for High Throughput Genomic Sequencing

The scope of the project includes all processing steps from DNA isolation to complete genome reassembly. Once extracted, the DNA will be analyzed by single-molecule real time observation, a highly efficient approach at sequencing that reads code as transcription occurs, as opposed to the original Sanger method, which requires the amplification of the DNA prior to the reading of sequence. The technique adopted for the process closely models that proposed by *Pacific Biosciences*, who uses specially developed zero-mode waveguides (ZMW) and uniquely labeled fluorescent nucleotides. These nucleotides will be produced on-site, as there are no vendors that produce these specific molecules.

The production of the ZMWs, however, will be contracted out to a nanofabrication firm in order to avoid the high capital investment in lithography equipment. Additionally, *PennBio* will not offer any forms of genetic counseling of genome interpretation. This is a rapidly evolving, research-intensive service that is best left to specialists.

The goal of the first year of the product development is the production of a working sequencing prototype. Once the effectiveness of the prototype is verified, scale-up and commercial sequencing will begin in year two. The number of setups will be increased so that the desired throughput for meeting the project goal can be met. Following year two, most of the company resources will be dedicated to full time commercial genome sequencing and research and development for the next innovation in genomic sequencing.

I.3 Technology Readiness Assessment

Following the very first successful genome sequencing, each of the next attempts at human genome sequencing strived toward reducing costs and increasing throughput. With the Archon X Prize setting the desired bar for motivation to achieve the set requirements, the necessary technology has to first develop in order to support the ultimate goals.

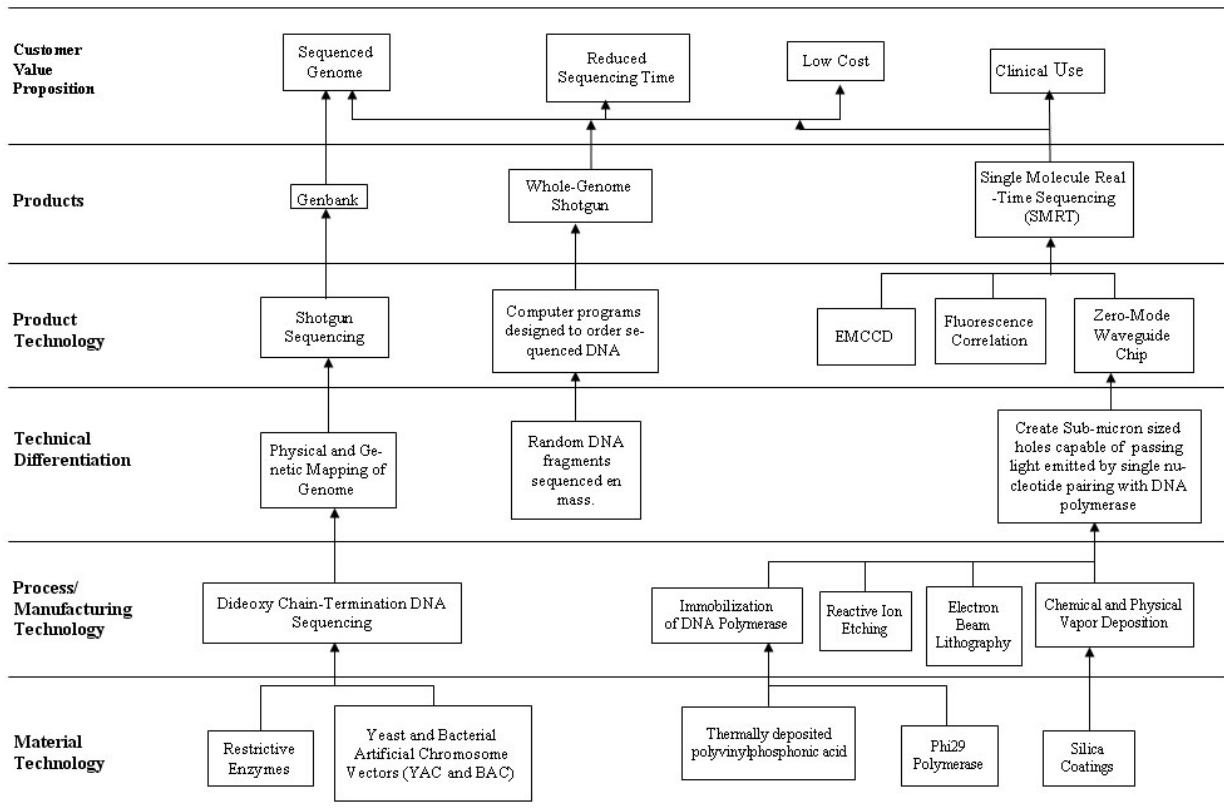


Figure I.1 Innovation map for high-throughput genomic sequencing

With improvements in nanostructure fabrication, novel nucleotide chemistry, genomic assembly algorithms, and optical devices capable of single molecule detection, *PennBio* is able to offer sequencing which meets lofty throughput and price goals never before reached using the SMRT chip platform. The technique requires a relatively low reagent volume, reducing costs significantly when compared to now-obsolete sequencing methods. Additionally, the throughput is maximized as the DNA sequence is read while transcription occurs. In order to carry out SMRT chip-based sequencing, the key supporting technologies of zero-mode waveguides (ZMWs), novel fluorescent molecules, and powerful imaging instrumentation have to be available first (See **Figure I.1**). Due to advances in material sciences, techniques to cheaply and consistently produce ZMW chips make them readily available at the desired specifications. In addition, developments in fluorophore synthesis allows the production of the key sequencing molecules in lab. Possibly the most important development for SMRT sequencing in terms of throughput is the development of Electron Multiplying Closed-Coupled Device cameras or EMCCDs. The cameras allow for high resolution imaging at high frame rates necessary for the real-time reading of massive parallel arrays of microreactors that are the key parts of SMRT. Further improvements to this imaging technology would only serve to increase the throughput of the sequencing techniques.

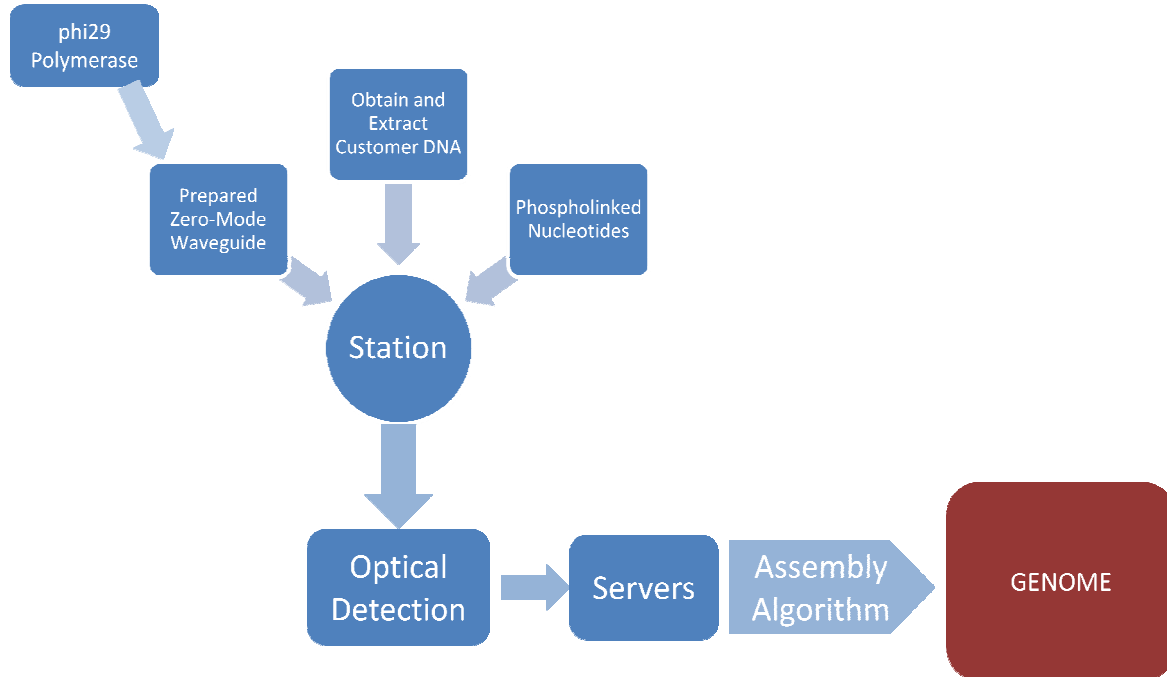


Figure 1.2 Organization of the sequencing process

At \$4,000 per individual, *PennBio* is offering a superior and more comprehensive service than ever before offered: for an estimated 0.02 cents per SNP. More importantly, as new SNPs are discovered and new clinical correlates are published, the *PennBio* customer will have the information necessary to re-examine his or her genome under the guidance on a geneticist. Customers of SNP-limited services, such as *23andMe*, will have to be retested in the future if they hope to maintain an up-to-date genomic perspective on their current health and disease susceptibility.

I.4 Customer Requirements

The two main customer requirements in designing the high throughput genome sequencing technique are high throughput and reading accuracy. In order to meet both of these customer requirements, “New-Unique-Difficult” (NUD) concepts must be implemented. These are addressed by critical to quality variables (CTQ) that define the core of the SMRT technology.

The main limit on this throughput is the resolution capabilities of the EMCCD in use and the frame rate of the EMCCD at the said resolution. As resolution of the EMCCD increases, the frame rate of the camera decreases. SMRT technology relies on recording the necessary chemical reaction in real time. Therefore, the more reactions that can be seen by the cameras, the higher the throughput obtained. The number of reactions that are viewed is also dependent on the waveguide chip itself.

Customer Requirements	CTQ Variables	Weight
<i>Reading Accuracy</i>	Polymerase Error Rate	0.25
	EMCCD Frame Rate	
	Number of Waveguides	
<i>High Throughput</i>	EMCCD Resolution	0.75
	EMCCD Frame Rate	
	Number of Waveguides	
	Rate of Reaction	

Table I.2 Customer Requirement This table describes the customer requirements that are met by SMRT sequencing technology, and the CTQ variables that address the specific customer requirements.

Since the reaction occurs on the waveguide chip, increasing the number of reaction locations on the ZMW increases throughput capabilities. Additionally, running the reaction as fast as possible maximizes the throughput.

In order to successfully view the reaction while maintaining reading accuracy, it is important to maintain an EMCCD frame rate that can match the speed of the reaction. More specifically, the frame rate of the camera must be greater than the rate of the reaction observed in each well. A compromise must be reached between the frame rate of the camera and the resolution that is used in order to maintain the throughput and reading accuracy, while keeping **Table I.2** in mind. Throughput and reading accuracy are in fact very tightly bound. In addition to the compromise between EMCCD speed and resolution, the coverage of the process must be addressed. Increasing the coverage, or amount of times the whole genome is read increases the reading accuracy. However in order to meet this coverage requirement, the throughput must also increase. Therefore, factors such as waveguide number and EMCCD specifications that increase throughput are directly bound to increasing reading accuracy as well.

I.5 Unprecedented Throughput

Single-day turnaround is one of the hallmarks of the SMRT sequencing design, and plays a vital role in setting this technology above the competition. For a given investment in detection equipment, the limit of EMCCD efficiency is achieved by one-to-one waveguide to pixel mapping.

The maximum number of waveguides that can be simultaneously observed ($n_{ZMW,max}$) is therefore defined by the EMCCD camera's field of view:

$$n_{ZMW,max} = 512 \cdot 512 = 262,144 \quad \text{Equation I.1}$$

Some of these pixels will be used to detect a bright pattern to which the nanopositioning stage can align itself, and a margin of error in the alignment will be allowed. Assuming the pixels at the extreme edges of the field of are not involved in sequencing, the actual number of ZMWs under observation (n_{ZMW}) is:

$$n_{ZMW} = 262,144 - (4 \cdot 512) + 4 = 260,100 \quad \text{Equation I.2}$$

Not all of these waveguides will ultimately contain DNA polymerase molecules; rather, the proper immobilization of an enzyme within a ZMW is a Poisson event. The probability of single-occupancy (the only fill state that will produce a meaningful signal) achieves its maximum at 36.8% when $\lambda = 1$. Therefore, the number of polymerase molecules that should be applied to the chip in order to maximize single-occupancy is:

$$I = 1 = \frac{n_{phi29}}{n_{ZMW}} = \frac{n_{phi29}}{260,100}$$

Equations I.3, 4

$$n_{phi29} = 260,100$$

Assuming only 36% single-occupancy is achieved, the number of active ZWMs would be:

$$n_{ZMW,act} = 260,100 \cdot 0.36 = 93,636 \text{ active waveguides} \quad \text{Equation I.5}$$

The polymerization rate of phi29 has been demonstrated to be no less than 4.7 bases per second¹³. At this speed, the number of bases synthesized per second, per SMRT chip is:

$$n_{bp/s} = 93,636 \cdot 4.7 = 440,089 \text{ bp/s} \quad \text{Equation I.6}$$

This process must be sustained until a 9-fold multiplicity is generated, that is, until the total sequenced length is equal to nine times the length of the genome, or 27 gbp. As we will demonstrate in **Chapter V**, such redundancy is necessary if gaps are to be avoided. The time necessary to accomplish the coverage goal is equivalent to the time required to sequence a single human genome on a single SMRT chip, and is calculated to be:

$$\frac{(3 \times 10^9)(9)}{\left(\frac{n_{bp}}{s}\right)} = \frac{2.7 \times 10^{10}}{440,089} = (6135 \text{ s}) \cdot \left(\frac{1}{3600 \text{ s/hr}}\right) = 17.0 \text{ hr} \quad \text{Equation I.7}$$

The SMRT system is the only available genome sequencing platform that offers such low turnaround times, without compromising quality or volume. The waveguide array not only serves to attenuate background noise, but it also provides the geometric precision required for efficient, single-pixel detection. Combined with a purpose-developed polymerase and powerful computing resources, this novel sequencing technology is certainly the most promising among modern whole-genome techniques. In the next chapter, the advantages of SMRT technology will be examined.

II. The SMRT Chip Platform

Pacific Biosciences was founded in 2004 with the goal of developing a low cost, high-throughput genomic sequencing system driven by the observation of single DNA polymerase molecules, working continuously under realistic biological conditions¹¹. By monitoring hundreds of thousands of enzymes in parallel, this technique is distinguished by exceptional sequence data quality and unprecedented throughput.

These capabilities are due in large part to “single-molecule real-time” (SMRT) technology, which is itself a fusion of biochemistry, optical theory, and recent advances in nanofabrication¹². The physical product of SMRT technology is the SMRT chip – a single-use

assay platform that makes the high-fidelity, high-parallel reads possible. Two of the most common challenges to successful fluorescence-based assays—contiguous long sequence reads and negligible fluorescent noise—are simultaneously overcome by this revolutionary design, which for the first time utilizes zero-mode waveguide reaction vessels and phospho-linked fluorescent nucleotides to obtain a high-multiplex, high signal-to-noise output of long DNA sequences.

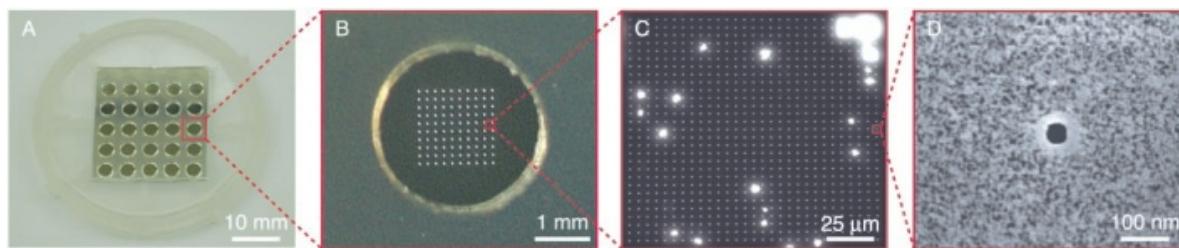


Figure II.1 SMRT chip with array-separation gasket (A) Macroscopic view of a waveguide array (B) Closer view – the bright spots are flaws in the aluminum cladding (C) SEM characterization of a single cylindrical waveguide (D).

II.1 Sequencing by Synthesis

Using zero-mode waveguides (ZMW)—small aluminum wells in the bottom of the reaction vessel which have apertures smaller than the wavelength of the biomolecules’ fluorescent emissions—signal to noise ratios become dominated by the presence or absence of those molecules in the waveguide—see **Figure II.2**. Utilizing this technology, single biomolecules can be observed in action. The concept behind single molecule real-time (SMRT) sequencing relies on the interplay between active DNA polymerases, synthetic nucleotides, and

the zero-mode waveguides. With some preparation, these waveguides can be populated by single DNA polymerase enzymes immobilized in place at the bottom of the wells¹³.

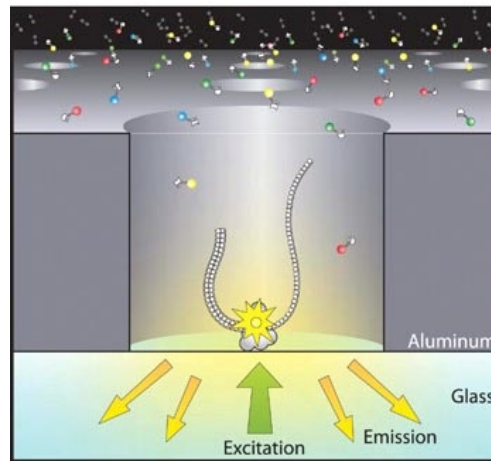


Figure II.2 ZMW with active polymerase. Reproduced from Eid (2009).

DNA polymerases replicate DNA polymers by using single stranded DNA template polymers and synthesizing a complementary strand using free DNA monomers known as nucleotides, matching guanine to cytosine and adenine to tyrosine and vice versa. SMRT sequencing relies on replacing natural nucleotides with synthetically labeled fluorescent nucleotides. These nucleotides are fluorescent until the DNA polymerase incorporates them into the ever-growing complementary strand. Once incorporated, the fluorescent tag is released from the nucleotide, leaving it 'dark'. When immobilized DNA polymerases bound to template DNA incorporate these synthetic nucleotides into a growing complementary strand at the bottom of the zero-mode waveguides, fluorescent emissions from the nucleotides escape the waveguide only as

they are being incorporated into the complementary strand then go dark as their fluorescent moiety is cleaved and diffuses from the waveguide. By labeling the four different bases with four distinguishable fluorescent tags, emissions from the waveguide form a temporal sequence of different wavelengths which directly reflect the sequence of the strand being synthesized and therefore the sequence of the template strand – see **figure II.3**. These emissions can be recorded with high resolution optical devices and the sequences stored for data analysis.

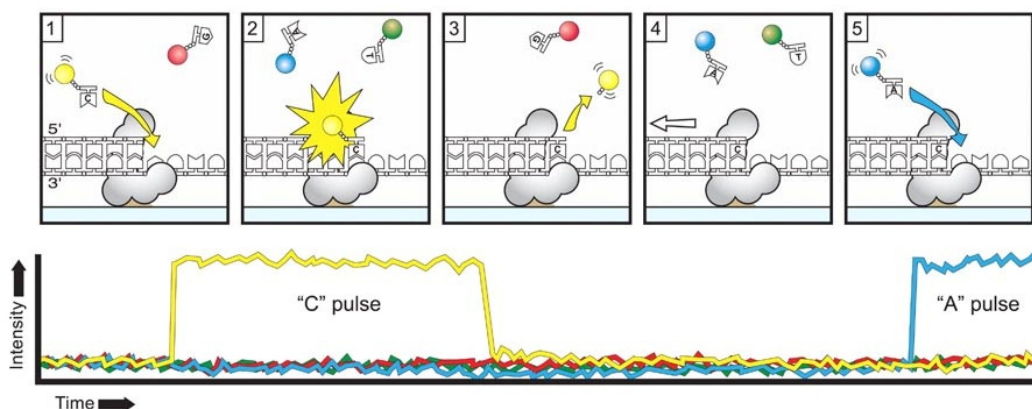


Figure II.3 Sequencing by synthesis process, showing nucleotide addition (1), fluorescent emission and fluorophore cleavage (2, 3), followed by polymerase procession and addition of the next base (4, 5). The intensity time trace is representative of time-series data generated upon signal analysis. Reproduced from Eid (2009).

II.2 The Zero-Mode Waveguide

Existing methods for single-molecule detection, such as fluorescence correlation spectroscopy and near-field confocal microscopy, are characterized by observation volumes on the order of femtoliters (10^{-15} liters)¹⁴. If any specific individual molecule is to be identified with any certainty, then any additional fluorescent species may not be present above pico or nanomolar concentrations. While this background-minimization

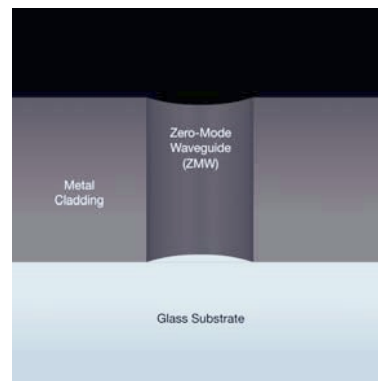


Figure II.4 An illustration of a cylindrical ZMW. Published by Pacific Biosciences, Inc. (2008).

technique is effective, the use of unusually low ligand concentrations in an enzyme-based assay can have undesirable effects on reaction kinetics, or even cause the reaction pathway to deviate from its natural course. DNA polymerase, in particular, has a micromolar Michaelis-Menten constant (K_m), exhibits prohibitively slow polymerization rates at low nucleotide availability. The zero-mode waveguide (ZMW) addresses this limitation by dramatically reducing the observation volume, thereby allowing the use of biologically-relevant conditions.

For light of any given wavelength, an aperture can be constructed through which light cannot propagate¹⁵. The aperture can take on any number of geometries, and its dimensions are functions of the wavelength of the incident radiation. The cylindrical ZMWs on the SMRT chip are designed to be narrower than this cutoff diameter, so that the fluorophore excitation radiation becomes evanescent upon entering, and decays exponentially with distance traveled into the ZMW. In this way, the emission energy profile is sufficient to excite any fluorophores at the very

edge of the waveguide, but too weak to excite the fluorescent species in the solution above. Attainable observation volumes are decreased from femtoliters to zeptoliters (10^{-21} liters), which provides a margin to increase fluorescent species concentration without contributing appreciably to background.

With this knowledge of the incident energy profile, a single DNA polymerase molecule is immobilized at the bottom of each waveguide – precisely within the volume under observation. As this enzyme adds fluorescently labeled nucleotides to a template sequence, it brings them within the observation volume, and an excitation signal is detected. The time required to add a nucleotide to a growing strand is several orders of magnitude greater than the diffusing timescale of these molecules, enabling the polymerization process to be clearly monitored over the constant, but low intensity, noise¹³.

SMRT chip production is very similar to integrated circuit fabrication, and makes use of many of the techniques developed for this industry. With even more recent improvements in the reliability and miniaturization of the manufacturing equipment, it has become widely available, and the process used in zero-mode waveguide creation is now quite routine. Given the submicron

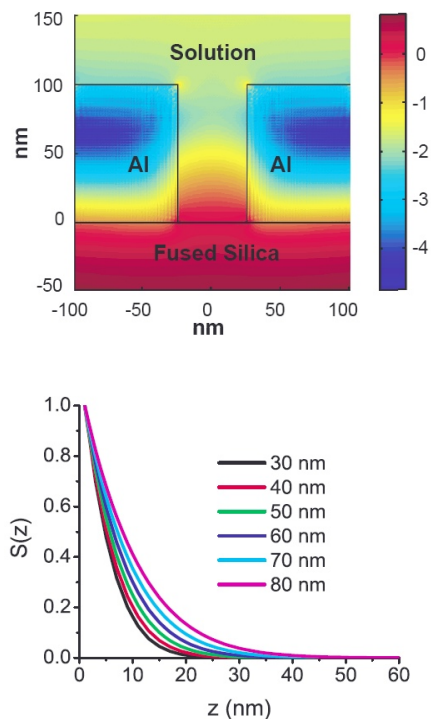


Figure II.5 A heat map of illumination intensity (top frame). Intensity as a function of distance into the ZMW by radius (bottom frame). Reproduced from Levene (2003).

scale of the waveguides, an array density on the order of 10^6 waveguides per mm^2 is easily achieved, enabling extensive parallelization of the sequencing process.

II.3 Phospholinked Fluorophores

Sequencing techniques utilizing fluorescently labeled nucleotides are well documented and have been robust strategies for sequencing; however these approaches have used almost exclusively base-linked nucleotides which present many enzymatic incorporation problems including processivity issues and overall protocol encumbrance as additional bleaching or washing steps are necessary for longer reads.¹⁶ Not only do base-labeled nucleotides form an altered complementary strand, which is sterically disturbed—leading to issues such as increased dissociation events from the enzyme (lowered processivity), these steric issues also lead to active site association issues leading to extremely hampered enzymatic kinetics. In addition, these fluorescent tags also remain on the growing complementary strand leading to increasing background noise levels unless bleaching steps are taken to eliminate previously incorporated bases from the fluorescent read-out. These additional steps interrupt enzymatic activity and these techniques are therefore characterized by relatively short read-lengths as dissociation events become much more frequent. To address this issue, when considering high-throughput entire genomic sequencing, an alternative approach was developed using phospho-linked fluorescently labeled nucleotides which eliminates the issues associated with base-linked fluorophores. The phosphodiester bonds of nucleotides are cleaved when the nucleotides are incorporated into the

growing complementary strand, and phospho-linked fluorophores diffuse away as the cleaved phosphate groups do.

Expected Throughput

In practice process times not equal to 24 hours would lead to undesirable precession in the sequencing time over a period of several days. In order to keep the work schedule constant, 24 hours must be allotted for sequencing.

The throughput goal is 100 genomes per 10 days, which is equivalent to 3000 genomes per year. Given the constraint presented above, and assuming five “sequencing days” per week, the number of stations/chips required is:

$$n_{stations} = \frac{(10 \text{ genomes/day}) \cdot 0.8}{(1 \text{ genomes}/(\text{day} \cdot \text{station}))} = 13 \text{ stations} \quad \text{Equation II.1}$$

Each of these stations will consist of one SMRT chip, simultaneously observed by two EMCCD cameras through an inverted fluorescence microscope.

III. Competitive Analysis

The relatively new market for genomic sequencing is filled with opportunity for growth; however, this opportunity opens up possibilities of competition. In order to move forward with the business model for *PennBio*, it is critical to perform an analysis of the competitive environment and the major companies that offer similar services with slightly different techniques. Besides the SMRT technology used for *PennBio*'s design, three other competing technologies have been analyzed. The companies offering these technologies are *Illumina*, *454 Sequencing*, and *VisiGen*. Since genome sequencing is a relatively new market and most of the products being offered are still in their infancy, not all of the information regarding the products, such as potential costs, is available to the public.

III.1 *Illumina*

Illumina relies on *Solexa* technology in order to sequence their genomes. The basis of the sequencing involves fragmenting the genome into pieces and then attaching adapters to the end of each strand. These adaptors attach to a flow cell surface. Once the strands are attached to the surface, they undergo many cycles of amplification until they form clusters of up to 1000 identical copies. These copies are all single stranded. Following this step, fluorescently tagged dNTPs with reversible termination properties are added along with polymerase. These attach to the end of the strands and are then emit light following laser excitation. *Illumina's* imaging technology then reads out the different wavelengths emitted for each nucleotide. This process is repeated again until the entire strands are read. Then the fragments are assembled into a genome sequence using an unreleased algorithm.^{17,18}

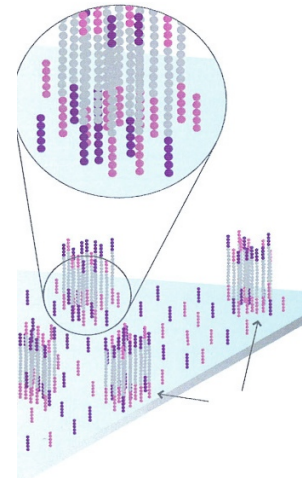


Figure III.1 Illustration of the DNA fragment clusters attached to the flow cell.¹³



Figure III.2 The *Illumina* Genome Analyzer¹²

One of the advantages of the *Illumina* technology is the simplicity of the flow cell design. There are no specific wells or beads that need to be attached to, and the amplification of the clusters allows for a large viewing concentration. The very nature of the clusters themselves guarantee a strong

signal, as they are very concentrated as opposed to the single fluorescent molecule that has to be detected for the *PennBio* design. Also, the optics of the system are much simpler, since the molecules are placed arbitrarily. A major advantage that *Illumina* possesses is an already built sequencing station that mechanizes each sequencing step. These units will be provided to laboratories, so that *Illumina* does not actually sequence the genomes, but the labs that purchase them do.

A major disadvantage to the technology is that it has low throughput compared to *PennBio*'s projected throughput. An entire genome can be sequenced on the order of weeks. This is just the sequencing method. The actual preparation of the flow cell and amplification of the DNA takes one business day. This is significantly longer than *PennBio*'s expected throughput.

III.2 454 Sequencing

454 Sequencing is more competitive than *Illumina* in terms of throughput, though not by much. Like *Illumina*, *454* amplifies a single stranded molecule attached to a surface. However, the surface is a bead that is immersed in an emulsion, creating a microreactor. These beads are then placed in individual wells. The wells are in a PicoTiterPlates that are found in a Genome Sequencer FLX Instrument. Using a fluidic assembly, the sequencer pumps nucleotides over the wells. Upon incorporation with the DNA template, a combination of fluorescent enzymes that react with the template emits different light

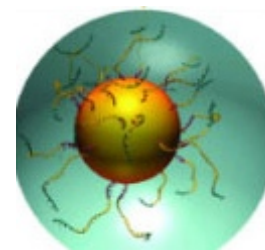


Figure III.3 454's bead microreactor.¹⁴

spectrums based on the nucleotide. Much like the other sequencing techniques discussed, the light emitted is used to come up with a sequence reading. Unlike the rest of the techniques, the 454 technique adds each dNTP one at a time per cycle, and these cycles are repeated until the entire DNA strand is sequenced. With DNA fragments of approximately 400 bp long, the cycling slows down the sequencing time.¹⁹

In addition to this delay, the relatively short fragment length, increased time is required to process and reassemble the genome. As it is discussed later, in the section dealing with reassembly of the genome, shorter fragment lengths make it much more difficult and time consuming to assemble a complete genome from fragments.

III.3 *VisiGen*

VisiGen is the competitor most similar to *PennBio*. *VisiGen* uses SMRT sequencing in order to achieve their high throughput. The major difference between the two competing technologies is *PennBio*'s use of a zero-mode waveguide. *VisiGen* forgoes the use of this plate and randomly immobilizes its polymerase onto the surface of its plate. The biomechanism for both companies use a polymerase that allows DNA to replicate with the addition of fluorescently tagged dNTPs. Using laser excitation, the

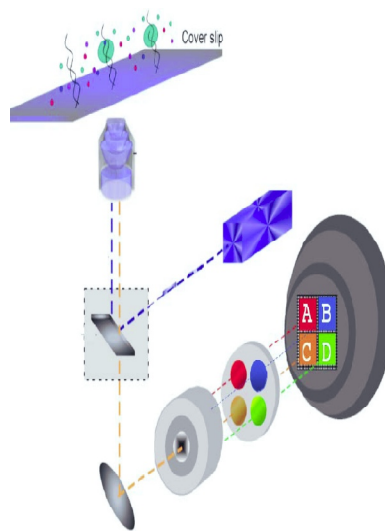


Figure III.4 Basic setup for the *VisiGen* sequencing system.¹⁵

fluorophores emit light that is detected by an EMCCD camera, as can be seen in the figure on the left.²⁰

Using this technique, *VisiGen* avoids the requirement for waveguide use. Though this reduces the cost in purchase of waveguides, it creates several other problems. With the polymerases randomly dispersed with no real separation, interference from other complexes creates potential errors in the sequence read and increase the signal to noise ratio. In addition to this, the field of view used is not maximized, as the polymerases do not necessarily line up with each pixel, as they do in the ZMW for *PennBio*. As seen in **Figure III.4**, *VisiGen* uses a Bayer color filter on its EMCCD. This dedicates at best 4 pixels per polymerase, which is half of the resolution used by *PennBio*, as it is discussed later. This increases the number of microscopes required per station or the number of stations required per genome to be sequenced, increasing the cost well above any savings incurred from not having to purchase waveguides.

IV. DNA Polymerization

The following chapter seeks to describe in detail the synthesis reaction contained by each ZMWs. In order to understand the process, some fundamental biochemistry must be discussed.

To fully conceptualize the reaction taking place, let us begin by considering the reactants, themselves. In sequencing-by-synthesis, single stranded template DNA serves as the foundation of the reaction. It is given the name *template* because it serves as a guide for the synthesis of new strands, and DNA's unique structure makes it ideally replicated by a process so elegant, it could only be found in nature.

DNA's famous double-helix structure is constructed of two anti-parallel strands of nucleic acid polymers made from deoxynucleotides. The strands have a *backbone* made from connecting ribose and phosphate ester linkages labeled from 5' to 3' from the carbons on the ribose ring, like the sides of a ladder, in which the rungs are the nucleobases—guanine, adenine, cytosine, and thymine. The rungs are formed as hydrogen bonding between complementary bases connects the two phosphor-deoxyribose backbones. Adenine interacts with thymine as guanine pairs with cytosine—see **Figure IV.1** below.

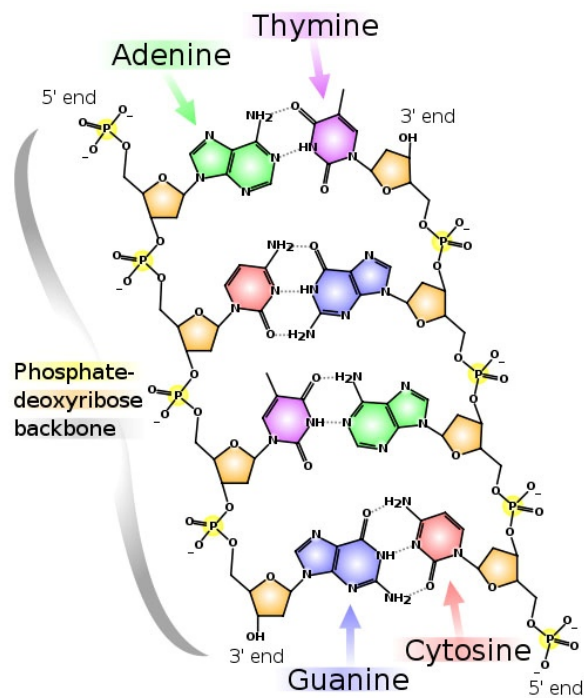


Figure XII.1 DNA's unique ladder-like structure. Detailed are the phosphate-deoxyribose backbone and the hydrogen-bonded nucleobases complementary to one another. Reproduced from Ball (2007).

This unique structure allows DNA to be denatured—when the hydrogen-bonding between complementary strands is disrupted and the two strands are separated from one another—and then new complementary strands synthesized from the two original strands, effectively doubling the number of strands each time. Because the bases in the double-stranded DNA are complementary to each other, each strand holds the entire code, and both strands can be used to make entirely new, complete double-stranded DNA. *In vivo* this process involves a complex interplay between many enzymes which serve distinct functional roles in constructing new complementary strands from the template strands. *In vitro*, this process can be modified to take place with the help of just one enzyme—DNA polymerase. DNA polymerase catalyzes the addition of nucleotide triphosphates to a growing complementary strand by cleaving the phosphate groups from the substrate. Polymerases, however, cannot construct complementary strands *de novo*. They must bind double stranded DNA and build upon the complementary strand in so-called extension reactions. To ready template strands for replication *in vitro*, therefore, template strands are *primed* with short, single stranded DNA polymers complementary to the beginning of the sequence to be replicated—these short nucleotide sequences are called primers. Once primed the polymerase adds the complementary nucleotide, shifts down the DNA and adds the next complementary nucleotide. Once the polymerase reaches the end of the template strand, the polymerase dissociates from the newly synthesized double stranded DNA and can complex with a new, primed template strand. This concentric-cycles scheme for DNA synthesis is detailed in **Figure IV.2**.

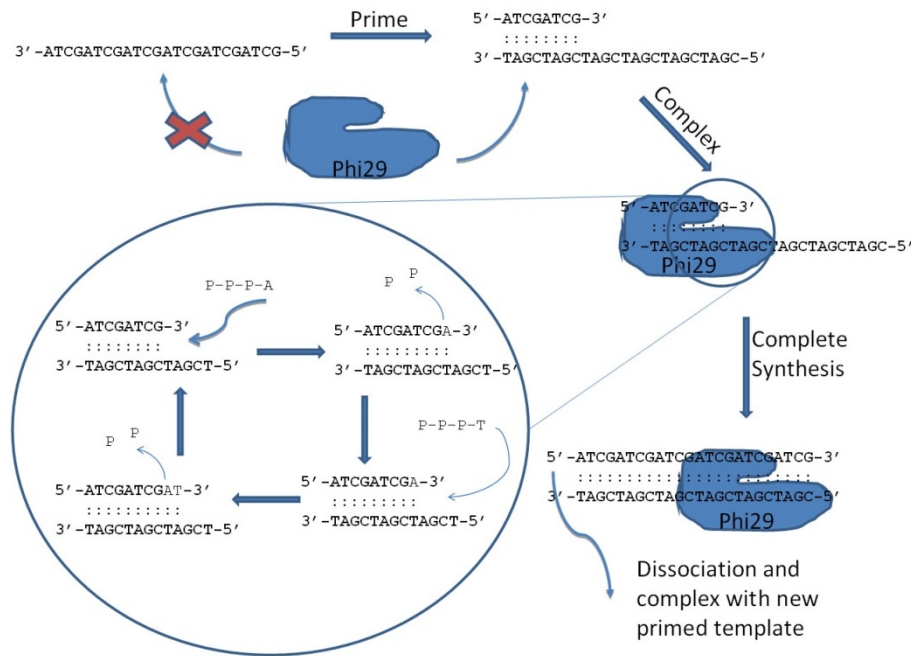


Figure IV.2 *In vitro* DNA replication using primers to allow phi29-polymerase complexing and extension of the complementary strand.

In order to utilize this invaluable enzymatic machinery to sequence template DNA by synthesis, several steps must be taken to prepare the reaction vessel and the starting materials. First, an appropriate DNA polymerase must be chosen which is suited to the task of accurate, speedy DNA replication in the manner necessary for SMRT sequencing. We show that the phi29 polymerase, from the phi29 bacteriophage, is well suited for the job. Next, the template DNA must be prepared. Because an individual's genome is to be the template for sequencing, a tissue sample is needed to obtain a sample of the DNA. Using a whole-genome amplification kit offered by Qiagen, a small amount of genomic DNA, which can be isolated from a non-invasive cheek swab, can be amplified so that plenty of genomic template is available for sequencing. For

the sequencing-by-synthesis reaction, suitable primers are chosen such that no sequence is biased over others, and that the entire genome is sequenced in its entirety. The templates, primers and polymerases are pre-bound and pre-complexed before immobilizing the polymerase into the zero-mode waveguides. This immobilization involves several chemical treatments of the waveguide nanostructure to ensure optimal waveguide occupancy. Lastly, the ingenious phospho-linked fluorescent nucleotides are discussed as their unique properties and behavior as substrates for DNA synthesis are crucial for the success of SMRT sequencing, followed by a discussion of dissociation rates for completed synthesized strands and reassociation of new primed-template strands, polymerization rates and error rates.

IV.1 The Phi29 Polymerase

The polymerase to be used in the SMRT system had to be chosen carefully. Dozens of polymerases are commercially available and most have been extensively documented. Most of these polymerases serve very specific roles: the *Taq* family of polymerases consists of thermally stable polymerases and these are ideal for polymerase chain reactions in which thermo-cycling is utilized to produce extremely high levels of amplification of template DNA; high-fidelity polymerases are available which contain extensive 3'->5' exonuclease activity—essentially backward double-checking of synthesized strand—enabling the proof-reading of the strands being synthesized to give extremely low error rates in synthesized strands; long-template polymerases exist which are known for very high processivity – a sort of ‘polymerase endurance’

– and very low dissociation rates between the polymerase and the template strand. These polymerases are capable of replicating very long template strands.

The bacteriophage, phi29, is a member of a family of phages which mostly infect *Bacillus subtilis*, a ubiquitous bacterium commonly found in soil. This phage carries its genetic code in double-stranded DNA form, and its DNA polymerase has been found to be an exceptional one.²¹ Because of the nature of the phage's minimal biochemical 'luggage' present in its capsule, its polymerase must be capable of replicating the genome of the phage with little of the enzymatic support often present in prokaryotic or eukaryotic systems. The phi29 polymerase has been found to be capable of extremely processive replication in the absence of accessory proteins to aid in the retention of the template strand in the active site of the enzyme. In addition, the polymerase shows strand displacement capabilities while it polymerizes making it able to replicate strands of DNA with complementary strands still partially bound to the template strand as well as overcome secondary structure in single-stranded DNA templates. These properties allow the enzyme to replicate the phage's genome without the use of primases or other accessory proteins commonly found in genomic replication schemes and perform multi-pass replications without dissociation with the template with just one modestly sized (66.7 kDa) monomeric unit, further highlighting the incredible efficiency of this enzyme.²²

In addition to incredible processivity – average replication lengths of over 70 kbp are commonly reported and values as high as several hundred kilobases have been cited in the literature – the synthesis rates and fidelity of the enzyme are also rather impressive. Esteban,

Salas and Blanco reported error rates as low as approximately 10^{-5} and polymerization rates as high as 100 bases per second.^{23,24,25}

In combination, these properties of the phi29 polymerase – extremely high processivity and strand displacement capabilities with very high replication fidelity and synthesis rates – make it an ideal candidate as the enzymatic machinery for the catalysis of isothermal single molecule DNA replication.

IV.2 Target DNA Isolation

Genomic sequencing by synthesis relies on template genomic DNA. To isolate a sample, the REPLI-g whole genome amplification kit by QIAGEN utilizing multiple displacement amplification (MDA) with phi29 polymerase provides a simple and cheap method of DNA isolation and amplification.²⁶ Only a cheek swab from the person whose genome is to be sequenced is needed providing a non-invasive means of collecting their genomic code. The cheek swab collects cells from the inside of the mouth which contain whole genomic DNA. The DNA is then isolated from the cells and amplified to ensure that enough material is present as described in **Appendix B**. As shown in **Figure IV.3**, the REPLI-g Midi kit provides 40 μg of genomic DNA, regardless of the amount of starting material.

The amplified genomic DNA is generally greater than 10 kbp in length and ranges from 2 to 100 kbp. And the REPLI-g Midi kit uses the same phi29 polymerase as the SMRT system, providing the same level of fidelity as is necessary for accurate sequencing. As with the SMRT system, the phi29 polymerase shows sequence displacement competency so that there is no

sequence bias in the amplification.²⁷ These properties make the product of the genomic amplification a perfect template for SMRT sequencing.²⁸

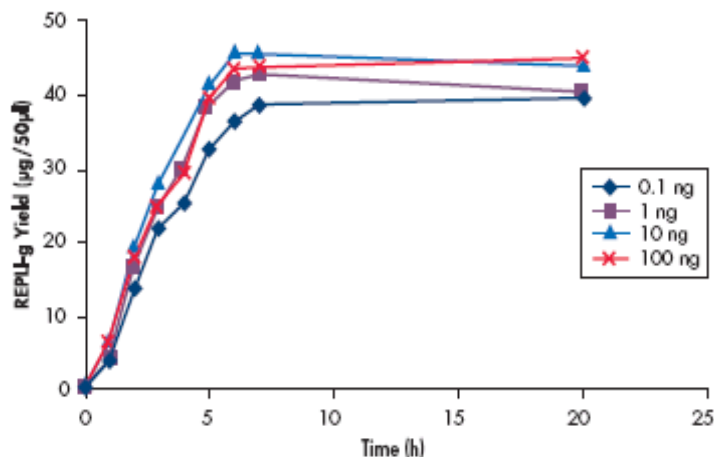


Figure IV.3 Yield of genomic template DNA after treatment with the REPLI-g Midi kit, as given by QIAGEN.

IV.3 Priming and Random Hexamers

With plenty of genomic DNA as a template for SMRT sequencing, the DNA must be primed for sequencing by synthesis. There is no known DNA polymerase which can replicate DNA from a native single strand. The phi29 polymerase requires a portion of the template strand to be double stranded as a starting point for synthesis, as it binds complementary nucleotides to the growing synthesized strand, so a starting synthesized strand must be bound to the template DNA before the reaction can begin. Choosing the right primer is an extremely important task. Because the entire genome must be sequenced, no one sequence can be biased over another. This demands the use of random hexamers.

Random hexamers, a collection of a six nucleotide oligonucleotides with random sequences—designated 5'-NNNNNN-3'—have been used in sequence-independent amplification for years and are well documented, widely available, common reagents for unbiased amplification of DNA.^{29,30} Because every possible combination of nucleotides is present in random hexamer mixes in essentially equimolar amounts, all sequences present in the reaction vessel have the same probability of being amplified. Fidelity Systems has developed a phi29 Random Hexamer mix optimized for sequence-independent phi29 polymerase-mediated DNA amplification, designed by Clyde A. Hutchison and colleagues, which includes random hexamers resistant to 3'->5' exonuclease activity—a feature of some polymerases making them capable of rejecting imperfect priming—providing optimized hexamer-template association resistant to dissociation by phi29 polymerase-mediated exonuclease activity.³¹

Template DNA amplified with the QIAGEN MDA kit is denatured using an alkaline denaturation as opposed to a heat denaturation for several reasons. First, phi29 polymerase is heat sensitive and is inactivated at temperatures above 65°C, while most thermocycling-based polymerization protocols include denaturation steps around 94-98°C. Therefore, annealing—the term given to the association of complementary sequences to form double-strands—of primers their templates is not possible without first inactivating the polymerase. Second, heat denaturation is known to degrade DNA samples and fragmented makes for poor sequencing templates.

The preparation of the template-primer complex is as follows. The DNA isolated with the QIAGEN REPLI-g Midi kit is first sonicated to a mean length of 2kbp to eliminate steric issues associated with entry of the template-polymerase complex into the waveguides. Next, the sonicated template DNA is denatured using a 0.2 M NaOH alkaline solution. After alkaline denaturation, the template DNA is incubated with the primer mix at a concentration of 50 μ M for 3-10min at 30°C.³²

IV.4 Template Binding

Next, the primed template DNA is incubated at 1.5-3 molar excess with biotinylated phi29 polymerases at 4°C for ten minutes in buffer to form the template-polymerase complex.³³ Because no free nucleotides are present, the primed template DNA binds to the active site of the enzyme but the synthesis reaction does not proceed. The template-bound polymerase must be immobilized in the bottom of the waveguides for the synthesis reaction to be accurately observed and recorded.

IV.5 Immobilization of the Enzyme-Template Complex

Immobilization of the polymerase requires extensive preparation of the waveguide nanostructure. Several factors must be taken into consideration when preparing the nanostructure. Though circular ZMWs have been used in many single-molecule detection studies,³⁴ their applications have been highly limited by the inability to selectively immobilize

molecules to the observation volumes immediately above the transparent floor. To address this issue, the dual-material nature of the ZMWs has to be exploited. Selectively reacting one of the two materials in a derivatization reaction enables the manipulation of either of the two different surfaces. Many factors had to be taken into consideration when designing the derivatized surface: stability in aqueous solution, in which the sequencing reactions would take place is a high priority; fluorescent background must be as low as possible to reduce noise in detection during the synthesis; and adsorption of the fluorescently labeled substrates must be low to help keep noise at a minimum. Passivation of mixed material nanostructures is an area of intensive active research, but the most common materials are gold-on-glass based structures.^{35, 36} Aluminum-on-glass structures have many advantages over gold when optical confinement of the ZMWs is considered as it has better reflectivity and a shorter optical skin depth. Aluminum, however, is corrosive in aqueous medium.³⁷ Organophosphorus acids have been shown to react with metal oxides, such as aluminum oxide, while not interacting with silicon dioxide surfaces in aqueous medium, offering a method of protecting the aluminum while leaving the glass of the structure unadulterated.^{38,39}

To selectively passivate the aluminum from protein absorption, polyvinylphosphonic acid (PVPA) is thermally deposited from a 2% aqueous solution of PVPA by incubation at 90°C for 2 minutes and then annealed in a dry oven at 80°C for 10 minutes.⁴⁰ To test the bias of these passivated aluminum surfaces on glass, adsorption of neutravidin as a test protein was conducted on both PVPA treated and untreated aluminum-on-glass nanostructures and the protein fluoresced for visualization with fluorescence microscopy. As shown in **figure IV.4**, the treated

aluminum showed tremendous bias with respect to physisorption of neutravidin. Untreated chips show bright aluminum squares as reflections from the metal intensify fluorescence, while treated chips show dark aluminum squares where little physisorption is found.⁴¹

When treated with phi29 polymerase, similar protein physisorption bias was found. A localization density ratio of over 400:1 on glass over aluminum was conferred, demonstrating the suitable passivation of the aluminum on the ZMW nanostructures for DNA synthesis.⁴²

Preparing the ZMW for immobilization of the polymerase on the glass of the waveguides is further enhanced by the use of an additional biotinylated polyethylene glycol layer. The biotinylated polyethylene glycol (PEG) polymer is deposited on the silicon dioxide by a process known as silanization. Using Biotin-polyethyleneglycol-trimethoxysilane, the glass bottoms of

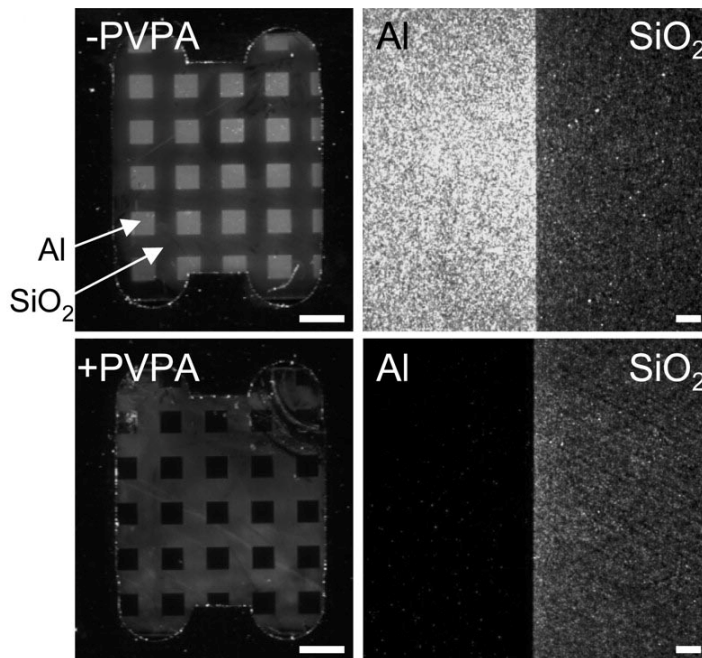


Figure IV.4 PVPA treated (bottom) versus untreated (top) whole aluminum-on-glass chips (left; scale bar, 1 mm) and aluminum glass interfaces (right; scale bar 10 μm) with deposited fluorescently tagged neutravidin. Reproduced from I. Korfach (2007).

the ZMWs are effectively covered in Biotin using the PEG polymers, known commonly as PEGylation, by a reaction between the trimethoxysilane and the silicon dioxide.⁴³ The now biotinylated ZMWs are treated with streptavidin at 22°C for ten minutes at a 2-fold streptavidin to polymerase molar excess.

Biotinylation is a process by which the coenzyme biotin—also known as vitamin B₇ or coenzyme R, molecular formula C₁₀H₁₆N₂O₃S, see **figure IV.5**—is covalently attached to another biomolecule. This technique has been used extensively in laboratory research and biomolecule preparation for decades because biotin and the avidin type proteins bind with an incredible degree of affinity.⁴⁴ The dissociation constant for biotin from avidin is $\sim 10^{-15}$ M making it one of the strongest known non-covalent interactions.⁴⁵

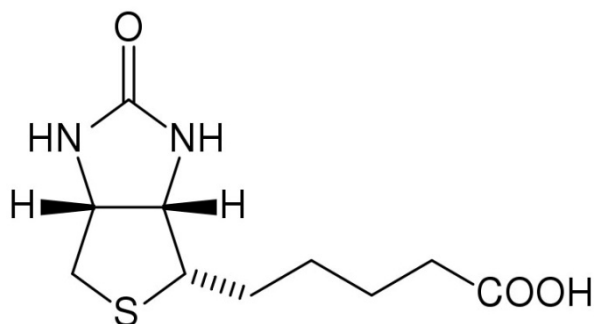


Figure IV.5 Biotin

Because streptavidin is a tetrameric protein which binds biotin stoichiometrically, the protein behaves like a glue between the biotinylated glass and the biotinylated phi29 polymerase. The ZMWs are washed with buffer to remove excess unbound streptavidin, followed by

immobilization of the polymerase-template complex at 4°C for 15 minutes by binding polymerase to streptavidin. Unbound complexes are washed away with reaction buffer.⁴⁶

Immobilization in this manner—using biotinylation—gives orientation consistency to the polymerases present in the ZMWs. This is extremely important for high-throughput sequencing. Randomly distributing the polymerase molecules across the ZMWs leads to a Poisson distribution of occupancy, and optimal loading gives only 36.8% of ZMWs with single molecule occupancy.⁴⁷ Clearly, waveguides with no polymerases will not produce reads, but also, waveguides with two or more polymerases will give reads in which the sequences of the two polymerases cannot be distinguished and throughput is affected. Orientation, however, is not directed by random distribution and misaligned polymerases will not function correctly and throughput could be highly reduced. By utilizing biotinylation binding in the bottom of the ZMWs, Korlach, Turner and colleagues found that 82% of singly occupied ZMWs produced full-length sequence-by-synthesis reads, greatly improving the throughput of the SMRT sequencing system.⁴⁸

IV.6 Phospholinked Fluorescent Nucleotides

By labeling nucleotide bases at the terminal phosphate, see **Figure IV.6** below, several issues of processivity in fluorescence based sequencing are addressed. Because natural polymerase activity cleaves the alpha-beta phosphoryl bond in the phosphonucleotide, the nucleotide incorporated into the growing product strand is a completely unaltered deoxyribose nucleic acid, and the strand grows as normal, as steric hindrance is eliminated. Furthermore, it has been shown that extending the triphosphate moiety to four and five phosphates increased incorporation efficiencies.⁴⁹ In several kinetic studies, Korlach and associates have shown that phi29 polymerase can, when all four dNTPs have been replaced with phospho-linked nucleotides, perform processively over thousands of bases at kinetics reaching levels of those associated with unmodified dNTPs—see **figure IV.7**. In addition, the synthesis of these special nucleotides has been elaborated in the literature—the procedure is detailed in Appendix D.⁵⁰

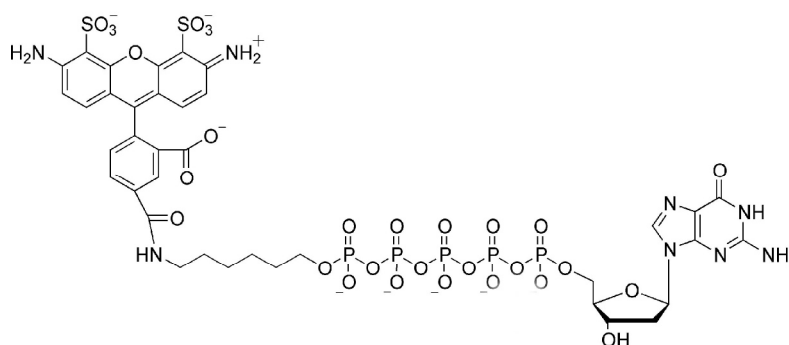


Figure IV.6 Penta-phosphate-linked Alexa Fluor 488 aminohexyl-O-dG5P

IV.7 Polymerization Rate Comparison

With the primed template DNA bound to the phi29 polymerase and this complex bound to the biotinylated glass bottom of the ZMWs, the sequencing by synthesis reaction is ready to proceed. First, an enzymatic oxygen scavenging system, using protocatechuate dioxygenase, is added to the array. Fluorophores are very susceptible to oxidative damage, and it has been shown that dioxygenases added to fluorescence based single-molecule experiments greatly increase the life of the fluorophores.⁵¹ Finally, the four phospho-labeled deoxyribose pentaphosphate nucleic acids are added to the array along with manganese acetate to concentrations of 250 nM (each nucleotide) and 0.5 mM respectively and the polymerization initiates at 30°C for the length of time needed for suitable sequence coverage.

Rates of polymerization of the phi29 polymerase utilizing phospho-linked nucleotides exhibit classic Michaelis-Menten saturation kinetics. Consistent with Michaelis-Menten kinetics, maximum saturation velocities, V_{max} , and substrate concentration at half-maximum velocity, $K_{1/2}$, values can be calculated for any fluorophore/nucleotide combination and a kinetic fingerprint equation developed for predicting polymerization velocities, V_{el} , as a function of nucleotide concentration, C .⁵²

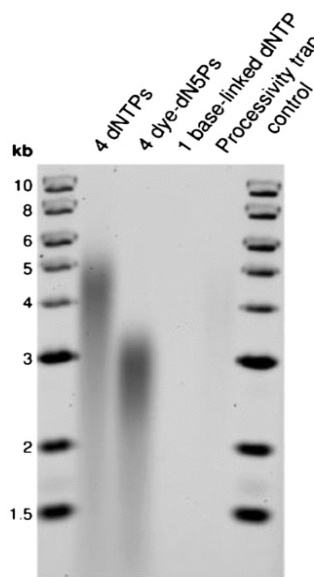


Figure IV.7 DNA products after 5 minute extension reactions. Native dNTPs are compared to phospho-linked dN5Ps and conditions with just one base-linked dNTP alongside a negative control.

The elongation velocity is given by:

$$V_{el} = \frac{V_{max} \cdot C}{K_{\frac{1}{2}} + C} \quad \text{Equation IV.1}$$

To optimize the read accuracy, nucleotide concentrations have to be carefully tuned to optical read capabilities. Background fluorescent noise is somewhat concentration dependent, however, the use of the ZMWs provides a wide window of concentrations which provide acceptable signal to noise ratios. The small detection volume surrounded by reflective aluminum limits penetrative light to no further than a few nanometers into the waveguide and additionally limits diffusion-mediated fluorescent occupancies to the order of 2-10 μ s versus nucleotide incorporation events on the order of milliseconds, providing easily distinguishable pulses. Base discrimination and read confidence are based on pulse and inter-pulse durations (see fluorescence discussion below). Nucleotide concentrations of 250nM each provide an average polymerization rate of 4.7 ± 1.7 bases/second with an acceptable error rate and signal to noise ratio.⁵³

IV.8 Dissociation of Synthesized Strands and Re-Complexing with New Primed Template Strands

After a polymerase has completed synthesis of a complementary strand from a primed template strand, the newly synthesized double strand dissociates from the enzyme and the enzyme is left free to re-associate with another primed template and continue sequencing. The

re-complexing of a new primed template is a diffusion-mediated event. Because the binding of primed DNA templates to polymerases is quite an exergonic reaction,⁵⁴ we assumed that any primed-template DNA which made its way to the bottom of a waveguide containing an empty polymerase would bind and proceed with synthesis.

To calculate the rate at which primed-template DNA enter the waveguides, Fick's law of diffusion was applied to free DNA molecules in solution. Fick's law of diffusion has the form:

$$J = -D \frac{\partial \phi}{\partial x} \quad \text{Equation IV.2}$$

Where, J is the flux (in $\frac{\text{moles}}{\text{m}^2 \cdot \text{second}}$); D is the diffusion coefficient of the DNA (in $\frac{\text{m}^2}{\text{sec}}$); and $\frac{\partial \phi}{\partial x}$ is the special derivative of the concentration gradient.

Robertson *et al.* have developed a correlation between DNA length and its diffusion coefficient of the form: $D = 2.3453 \cdot (\text{length in } \mu\text{m})^{-0.567}$.⁵⁵ This correlation gives our average two kilobase DNA fragments diffusion coefficients of $2.488 \frac{\mu\text{m}^2}{\text{sec}}$. This value is fixed for our fragment length at reaction temperature.

The concentration gradient, however, is manipulatable, and it must be. In order for our fragments to be distinguished from one another, there must be some distinguishable signal to the recording computers that an old fragment is done being sequenced and a new one has started, or reassembly will be unnecessarily more complicated. In this regard, we calculated the rates of diffusion into the ZMWs for various concentration gradients. At a low limit of 0.1 ng per 50 μl

(a lower limit based on a hypothetical failed genomic amplification using minimal genomic template, which correlates to a molarity of 1.62 picomolar) the flux is shown to be:

$$J = 2.488 \frac{\mu m^2}{sec} \cdot \frac{1.62 \times 10^{-12} moles}{.1 \mu m \cdot liter} = \frac{4.03 \times 10^{-26} moles}{\mu m^2 \cdot sec} \quad \text{Equation IV.3}$$

At the high limit of 40µg per 50µl (based on undiluted amplified genomic product, corresponding to a molarity of 0.647 micromolar) the flux is found to be:

$$J = 2.488 \frac{\mu m^2}{sec} \cdot \frac{6.47 \times 10^{-7} moles}{.1 \mu m \cdot liter} = \frac{1.61 \times 10^{-20} moles}{\mu m^2 \cdot sec} \quad \text{Equation IV.4}$$

When these fluxes are multiplied by the area of the waveguides (100nm in diameter yields well areas of 7854 nm²) and Avogadro’s number (6.02 x 10²³ strands of DNA per mol) the diffusion rate into the waveguides of the DNA strands is found:

	Lower Limit	Upper Limit
<i>Diffusion Rate (strands/s)</i>	0.00019	76.11
<i>Dissociation/Recomplex Lag</i>	5,256 s	0.013 s

Table IV.1 Diffusion rates calculated from correlated DNA diffusion coefficients and the consequential lag times between dissociation of completed complementary synthesized strands and reassociation with new primed-template DNA strands, at reasonable limits of operation.

Clearly, at the lower limit, throughput would be highly compromised, as on average, hours would be spent waiting for free polymerases to bind new template strands. The higher limit presents another problem, however, as lag time between fragments would be on the order of inter-pulse widths, meaning that no distinction between the last base of a fragment completing

synthesis and the first base of a fragment starting synthesis could be made, and the fragments could not be separated, causing severe reassembly problems. In order to optimize distinction between fragments yet reduce impact on throughput, DNA template concentration is chosen where the lag time between fragments is just longer than 75% of inter-pulse durations, such that the majority of fragments are distinguished, yet throughput is not significantly affected. Using data collected from Eid *et al.*, a cumulative exponential distribution was evaluated to find that three-quarters of inter-pulse durations would be less than one second long. To then find the DNA concentration at which mean lag time between dissociation and re-complexing events was longer than one second, the flux has to be calculated and then the DNA concentration backed-out using Fick's Law:

$$\frac{1 \text{ strand}}{1 \text{ second}} \cdot \frac{1 \text{ mol}}{6.02 \times 10^{23} \text{ strands}} \cdot \frac{1}{0.007854 \mu\text{m}^2} = \text{flux of } \frac{2.115 \times 10^{-22} \text{ moles}}{\mu\text{m}^2 \cdot \text{sec}} \quad \text{Equation IV.5}$$

$$\frac{2.115 \times 10^{-22} \text{ moles}}{\mu\text{m}^2 \cdot \text{sec}} \cdot 0.1 \mu\text{m} \cdot \frac{\text{sec}}{2.488 \mu\text{m}^2} \cdot \frac{10^{15} \mu\text{m}^3}{1 \text{ liter}} = \text{Concentration of } 8.5 \text{ nM} \quad \text{Equation IV.6}$$

With an optimized DNA concentration of 8.5 nanomolar calculated, the final reactants can be added and the reaction begun.

IV.9 Error Rates and Possible Sources of Error

Errors are currently on the order of 0.214%. Of these errors, the majority (44%) are deletions. Deletions occur from incorporation events which are too short, or when inter-pulse

durations are too short to be confidently detected. Nucleotides with no fluorescent label—so called ‘dark nucleotides—can be sources of deletion errors however, HPLC analysis show that the phospho-linked nucleotide composition is over 99.5% pure,⁵⁶ and kinetic studies have shown that phi29 polymerase shows no discrimination between phospho-linked and native nucleotides.⁵⁷ Additionally, statistical models predicting pulse-width distributions and projected probabilities of pulse detection show excellent consensus with the deletion rates observed. Future research and development efforts will focus on reducing these errors by further modifying the enzyme to reduce the fraction of short incorporation events as well as increasing camera frame-rate to strengthen the resolution of inter-pulse widths.⁵⁸

Insertion errors are caused mostly by dissociation of cognate nucleotides from the enzyme before the formation of the phosphodiester bond to the growing complementary strand resulting in duplications. As with deletions, these errors can be addressed by modifying the enzyme to reduce the free-energy of the nucleotide substrate in the active site thereby reducing the dissociation constant for cognate nucleotides. Mismatches were accountable to spectral misassignments between fluorophores with close emission wavelengths. These errors can addressed in future experiments by using dyes with more separation between emission wavelengths, as well as increasing the camera sensitivity.⁵⁹

IV.10 Conclusions

With a waveguide of 512 by 512 ZMWs with 1 micron square pitch spacing, an area of 0.318 mm^2 makes up the reaction vessel interface. With a liberal 1mm tall aqueous reaction mixture added, our reaction volume comes to 0.318 mm^3 or $0.318 \text{ }\mu\text{l}$. Once our template-complexed polymerase has been immobilized on our nanostructure the reaction mixture is added which is comprised of the following, in molar amounts shown, and the sequencing reaction proceeds at 30°C :

Compound	Concentration	Comments
<i>DNA</i>	8.5 nM	Template fragments of which are replicated forming sequencing output
<i>Primers</i>	50 μM	Random hexamers prime template strands for sequencing by synthesis
<i>dN5Ps</i>	250 nM (each)	Substrate for nucleotide addition to growing complementary strands
<i>ACES</i>	50 mM	pH buffer for biochemical reactions in the range 6.1-7.5
<i>Potassium Acetate</i>	70 mM	Salt buffer
<i>Dithiothreitol</i>	5 mM	Reducing reagent which deprotects thiolated DNA for efficient processivity of polymerases
<i>Manganese Acetate</i>	0.5 mM	Metal cation necessary for function of many DNA polymerases

Table IV.2 Combined at 30°C , these components initiate the reaction and sequences are read at the rates described above.

V. Optical Detection of Single Molecules

Several techniques exist for the focused analysis of single molecules. While varying slightly in execution, these all derive their resolving power from the minimization of one of two system parameters: background fluorescence, and observation volume. Either approach has a similar effect, namely, to reduce the probability of detecting a signal from more than one fluorescently labeled molecule at any given moment, enabling confident distinction between true signal and background noise.

The following chapter takes a closer look at the optical system required to meet the high throughput required for success. An analysis of the microscopic systems and the EMCCD

cameras required is performed to make sure that the setup is feasible. A key design element of the optical system, the use of two cameras to view the sample, is thoroughly examined, as it is critical in increasing throughput. The remainder of the chapter looks at the capabilities of those cameras used to determine whether or not they meet the detection requirements for successful observation of the sample volume.

V.1 Confocal Fluorescence Microscopy

One of the most widespread methods in practice today, due to its effectiveness and relative simplicity, is confocal microscopy¹⁴. As illustrated by **Figure V.1**, a laser beam is brought to its diffraction limited focus within a particular probe volume using a high-aperture objective lens. A pinhole (typically 50-100 μm in diameter) can also be placed at the interface with the sample, rejecting any light that remains out of focus. In conjunction with diffractive manipulation of the beam, this allows for approximately cylindrical observation volumes of 0.5-1.0 fl ($\sim 0.5 \mu\text{m}$ in diameter and $\sim 1.0 \mu\text{m}$ in height).

As labeled molecules enter the detection volume, the red-shifted photons emitted by the excited fluorophores are focused through the same pinhole and objective lens before being reflected by a dichroic mirror into the detection apparatus. Here the beam is divided equally between two

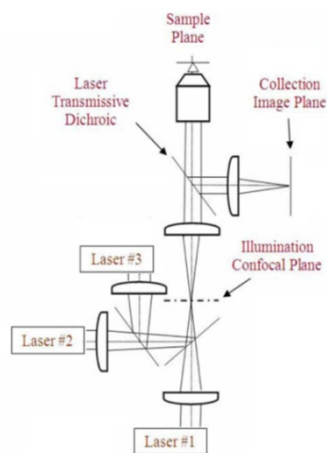


Figure V.1 A typical confocal microscopy setup, adapted from Lundquist (2008).

avalanche photodiodes by a beam splitter. If multiple emissions wavelengths are present, dichroic mirrors can be utilize for color isolation. Alternatively, individual wavelengths within the beam can be separated by prisms and focused onto characteristic regions of a charge-coupled device (CCD) detector. In this setup, both the intensity and position of the light would be recorded, allowing simple differentiation among many distinct fluorophores.

V.2 High-Multiplex Confocal Microscopy

For applications like SMRT sequencing that require spatial multiplexing, scanning confocal microscopy has been the standard measurement technique. This method is proven and well understood, but the frame rate is limited by the period of the scanning mechanism and by design, it cannot provide continuous observation of any single site. Our throughput requirements, however, demand the simultaneous monitoring of over 250,000 individual waveguides, completely ruling out scanning microscopy as a viable detection method. Rather, we will make use of recent advances in high-multiplex confocal microscopy to facilitate the collection of real-time, high-sensitivity fluorescence data⁶⁰.

The greatest single innovation in this field is the use of holographic phase masks (HPMs) to split a single excitation laser beam into an array of sub-beams at the same wavelength. These HPMs can be customized to generate almost any pattern of excitation radiation and almost any wavelength, and are thus readily adaptable to any number of sample volume configurations.

In consideration of the multiple excitation wavelengths required by the four fluorophores in use, multiple arrays from different sources could be combined using relay lenses and dichroic mirrors into a common illumination plane before being directed by the objective lens onto the sample. The emitted light is then collected through the same objective, deflected 90 degrees, split, and focused onto two single photon sensitive EMCCD chips. Previous applications of this CCD-based

technique have utilized prism assemblies to disperse the emitted light over several CCD pixels, providing continuous color separation for the spatial identification of wavelength.

In order to obtain high-resolution spectra, however, 15 pixels would be required for each ZMW, severely limiting the observational capacity of each EMCCD. Moreover, the entire purpose of a ZMW is to attenuate background noise by reducing observation volume. This renders confocal techniques, even those which support simultaneous spatial multiplexing, unnecessarily complex and expensive.

V.3 Two-Color Wide Field Microscopy

Our detection process will utilize some of the simplest illumination and observation methods, relying almost entirely on the properties of the ZMW and the back-illuminated

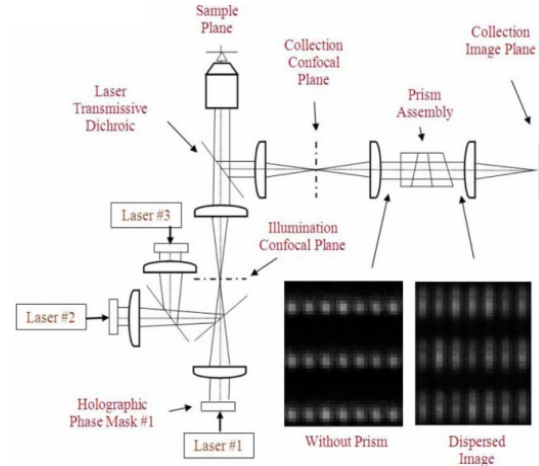


Figure V.2 High-multiplex confocal microscopy setup. The images in the bottom right illustrate the spatial wavelength identification method. Reproduced from Lundquist (2008).

EMCCD to facilitate high-quality signal acquisition. Excitation energy will not be provided by laser, but by a mercury arc lamp whose output will be split, wavelength filtered, then recombined onto the ZMW array. This requires no complicated optics, such as phase masks, while still providing uniform illumination. A high numerical aperture, low-distortion objective will enable the observation of the entire ZMW array simultaneously with the precision necessary to map each waveguide to a single EMCCD pixel. The EMCCDs themselves are mounted in a manual precision dual-port camera adapter for reliable CCD-to-chip alignment.

The fluorophore detection technique has also been designed to be optically simple and equipment efficient. Rather than devoting 15 pixels to each ZMW for spatial identification, we devote only two. Proper identification is performed by recording separate wavelength spectra on each pixel, and taking the ratio of intensities between the two. Physically, this is accomplished by using a dichroic mirror in the beam splitter at the core of the precision alignment adapter. Wavelengths higher than the cutoff are reflected to one camera, while the rest pass through to the second. Since both CCDs are aligned to the same waveguide array with 1-to-1 mapping, each pair of pixels will correspond to a single ZMW, and record the progress of a single polymerase molecule. The ratio of emissions intensities detected on each CCD will determine which fluorophore was excited, thereby identifying the nucleotide just added to the sequence. Despite doubling the number of cameras required, this method yields a 15-fold reduction in the number of pixels needed to identify a fluorophores, thereby reducing the overall detection equipment requirements 7.5-fold.

In order to guarantee that the individual peaks corresponding to the presence of a specific fluorophore, and thus a specific base, could be detected using the dual camera system, a read error analysis was performed using the Monte Carlo statistical method on the specific fluorophore peaks. Monte Carlo measures the independent fluctuation of multiple variables that affect a function. This is a perfect statistical method for analyzing the variation of light intensity read by both cameras.

As described earlier, the method used to determine the identity of the fluorophore emitting the light involves comparing the intensities hitting each camera. Since each fluorescent molecule has an individual light intensity distribution over a range of wavelengths, creating a cutoff wavelength for each camera divides the intensity readings between two cameras. With the individuality of each intensity distribution in mind, then **the ratio of intensities for each peak**

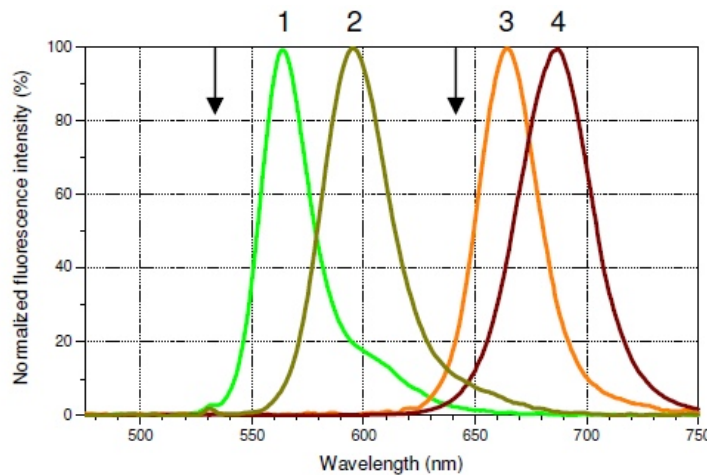


Figure V.3 Intensity Distribution of Four Fluorophores. The figure shows the intensity distributions over a range of wavelengths for fluorophores coupled with (1) dATP, (2) dTTP, (3) dGTP, and (4) dCTP. The camera wavelength ranges split the peaks at 638 nm, with camera 1 capturing the majority of dATP and dTTP, and camera 2 capturing the majority of the dCTP and dGTP intensities.

should be molecule specific. That is R_{peak} for each fluorescently tagged dNTP should be specific so that identification of each base is possible. R_{peak} is given by:

$$R_{peak} = I_{c2}/I_{c1} \quad \text{Equation V.1}$$

where I_{c2} is the intensity reading for camera 2 and I_{c1} is the intensity reading for camera 1. Using the intensity distributions in **Figure V.3**, the areas under each peak can be found. The wavelength cutoff between camera 1 and camera 2 was chosen to be at 638nm wavelength, at the intersection of the intensity distributions for peaks 2 and 3. This allows for relatively balanced intensity readings for each camera. Using ImageJ software, the area under the intensity curve for each peak is measured over the wavelength ranges for each camera, giving I_{c1} and I_{c2} readings.

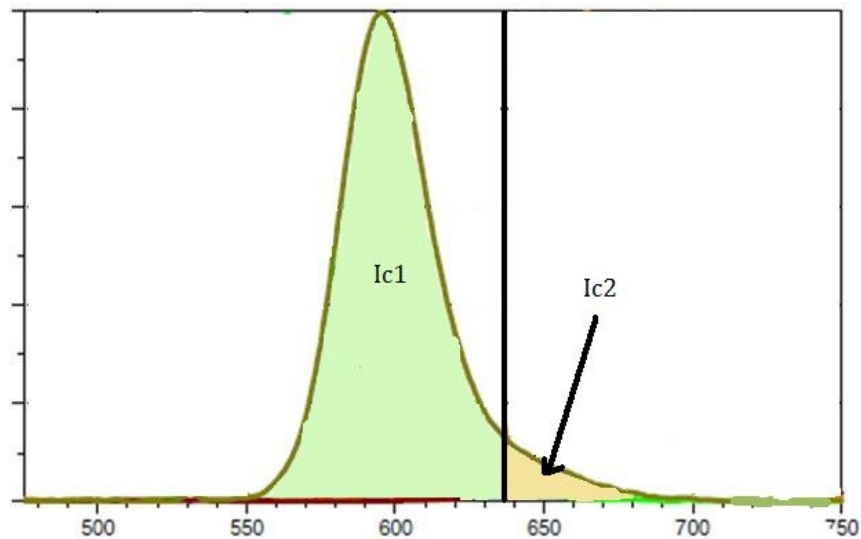


Figure V.4 Intensity Division of dTTP Fluorophore for Cameras 1 and 2. The figure above shows the normalized peak for the dTTP attached fluorophore. The line down the middle represents the wavelength cutoff for cameras 1 and 2, with camera 1 on the left and camera 2 on the right. The area under the peak is measured for the wavelength coverage of each camera using ImageJ software. The same process was repeated for the other three fluorophores.

Using these values, the ratio of intensity readings of camera 1 vs. camera 2 can be calculated. For example, I_{c1} and I_{c2} for dTTP were found to be 28178 and 1801, respectively, giving an intensity ratio of $R_{peak}=0.063915$.

$$R_{peak} = \mathbf{1801/28178} = \mathbf{0.0639145} \quad \text{Equation V.2}$$

In order to estimate the deviation from this value, a Monte Carlo statistical analysis was performed using MATLAB. The Monte Carlo method requires knowledge of the standard deviation of the peak intensities captured by each camera. This is a difficult proposition, since *Invitrogen* does not provide data on the variation of the intensity vs. wavelength variation for their Alexa-Fluor dyes. An alternative method for finding the standard deviation is possible using brightness deviation values (See Table V.1)²⁰. Since brightness correlates with light intensity, we assume that the standard deviation for the brightness also directly correlates with the standard deviation for each peak area of the normalized intensity curves. Therefore it is possible to find the standard deviation percentage for each fluorophore peak.

$$\text{Standard Deviation \%} = \text{Brightness} / \text{Brightness Std. Deviation} \quad \text{Equation V.3}$$

$$\text{Standard Deviation \%}_{dTTP} = \mathbf{39/2781} \times \mathbf{100} = \mathbf{1.4024\%} \quad \text{Equation V.4}$$

As shown above, the standard deviation percentage for the dTTP intensity is 1.4024%. By multiplying the standard deviation percentage by the I_{ci} values, a standard deviation for the intensity readings of each fluorophore for each camera can be estimated.

$$\text{Camera 1 Std.Dev}_{dTTP} = \text{Std.Dev.}\% \times I_{c1} = 395.1607 \quad \text{Equation V.5}$$

The standard deviation readings for each peak and for each camera are shown in **Table V.1** below. These standard deviations can then be used to carry out a Monte Carlo simulation using MATLAB. The simulation examines the distribution of R_{peak} using **Equation V.1** and the standard deviations found for I_{c1} and I_{c2} . The normal distributions for I_{c1} and I_{c2} are first generated in MATLAB using the following code:

```
ca_cb= (randn(n,1)*std)+mean;
```

where std is the standard deviation of the intensities found earlier, n is the number of iterations, and mean is the average value of the intensity per camera, or in this case I_{ci} . This process is performed for the intensity distribution of each peak for every camera. The Monte Carlo simulation is carried out using the distribution generated for each camera for a specific peak.

```
peak2 = c2_2./c1_2;
```

where the above code is analogous to **Equation V.1** for the dTTP intensity distribution. The process is repeated for dATP, dGTP, and dCTP. Once the distribution of each peak ratio is generated, the next step is to calculate the medians and the standard distributions found for each fluorophore. These are listed in **Table V.1** on the previous page. By plotting the histograms of each fluorophore camera intensity ratio, it is evident that there is no overlap between the ratios

for each fluorophore peak. It is clear that by using the dual camera peak identification setup would not produce base mismatches due to overlap of R_{peak} values. The complete MATLAB code for the simulation is found in **Appendix E.6**.

	dATP	dTTP	dGTP	dCTP
<i>Peak</i>	1	2	3	4
<i>Brightness</i>	6,446	2,781	4,865	2,691
<i>Brightness Std. Dev.</i>	109	39	92	41
<i>Brightness Rel Std Dev (%)</i>	0.01691	0.014024	0.018911	0.015236
<i>I_{c1}</i>	24,198	18,178	548	110
<i>Cam 1 Std. Dev.</i>	409.181	385.160	10.363	1.675
<i>I_{c2}</i>	226	1,801	25,175	31,303
<i>Cam 2 Std. Dev.</i>	3.822	25.257	476.074	476.932
<i>R_{peak}</i>	0.0093	0.0639	45.9398	284.5727
<i>Ratio Std. Dev.</i>	6.36×10^{-5}	3.65×10^{-4}	0.352	1.76

Table V.1 Data for Dual Camera Error Analysis. The following table shows data required for performing a Monte Carlo simulation demonstrating that the reliability of using a two camera system in order to detect 4 different fluorescent molecules. I_{c1} , and I_{c2} , are values obtained using ImageJ for the areas underneath the intensity distribution curves from Fig. IV.3. The areas are measured for the intensities for the wavelength range of each camera, set at 638 nm. The Camera standard deviation columns are obtained by multiplying the Standard deviation % column with the I_{c1} and I_{c2} columns. The Ratio Standard Deviation column is the standard deviation calculated using a Monte Carlo simulation in MATLAB.

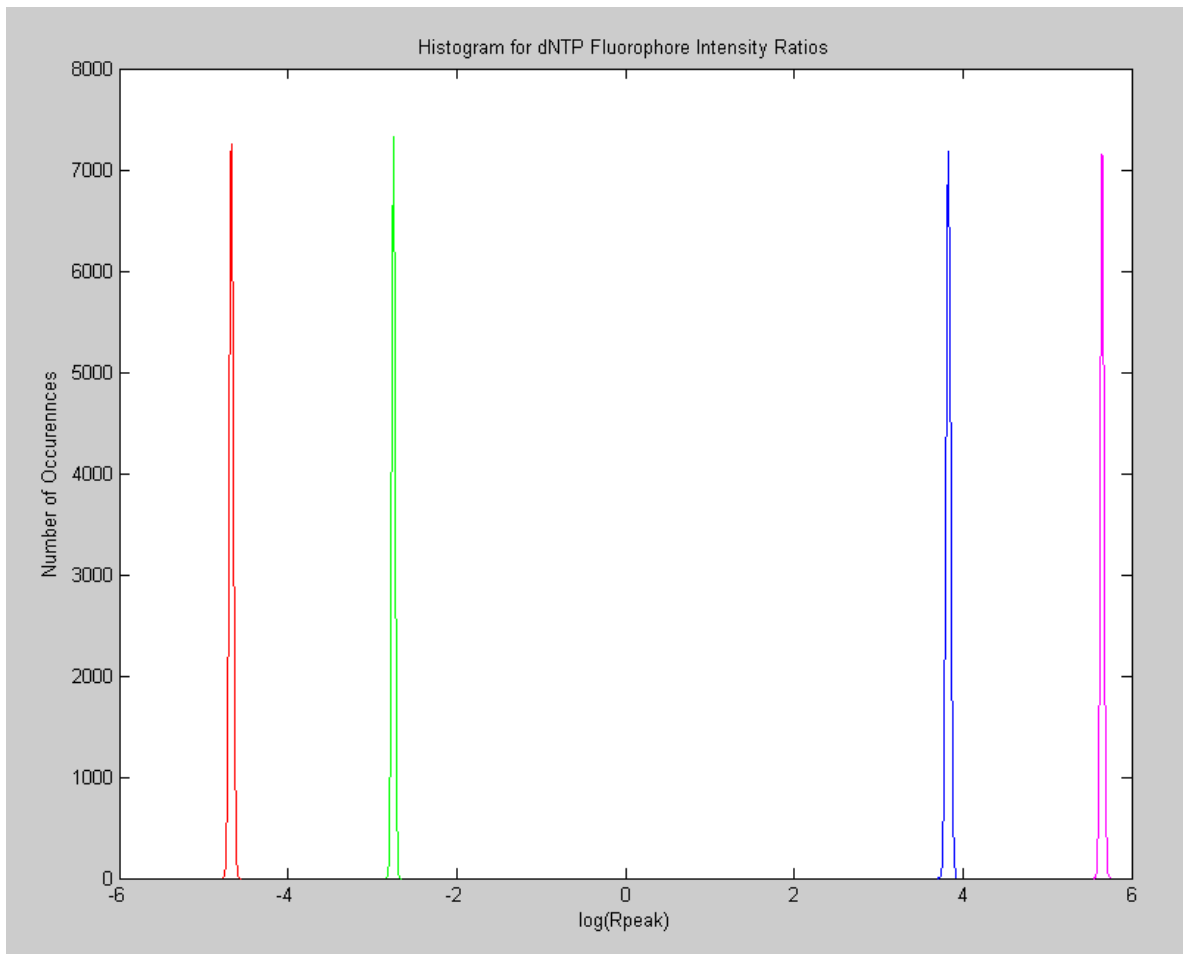


Figure V.5 The above image shows histograms produced for the distribution of $\log(R_{peak})$ values for dATP, dTTP, dGTP, and dCTP using the Monte Carlo method in MATLAB. The red, green, blue, and magenta curves are the histograms for dATP, dTTP, dGTP, and dCTP, respectively. It is clear that the lack of overlap between each R_{peak} value demonstrates that the two camera system should not result in base mismatches due to R_{peak} value overlap.

V.4 Fluorescence Detection and Signal to Noise

Electron Multiplying Charge Coupled Devices (EMCCDs) present a critical component in the design for high throughput genome sequencing. EMCCDs are capable of capturing single photon events at high read-out speeds. The chip uses the principles of impact ionization in order to register and multiply the presence of an electron. In order to deal with issues of background noise, EMCCDs employ an Electron Multiplying (EM) solid state register that amplify the signal from the electrons before passing through the output amplifier. The extra register at the end of the first register allows for the amplification of the signal without requiring an image intensifier, which would add noise to the image, lowering the camera performance. This elimination of background noise is critical when dealing with light signals from an individual molecule. This design is perfect for use in conjunction with the waveguide containing the biochemical reaction.⁶¹

The light emitted from each waveguide is lined up with the individual EMCCD pixels using a mechanical stage. The EMCCD desired for the setup is the iXon+ DU-897E from Andor. The camera provides a 512x512 resolution chip with individual pixels that are 16 μ m in size. This dedicates a single pixel to manage the light from an individual waveguide.⁶²



Figure V.6 Andor iXon+ DU-897 EMCCD Camera (Andor Tech. PLC, Belfast, Northern Ireland)

An important aspect to photodetection is the quantum efficiency (QE) of the CCD. The QE is defined as the percentage of photons that are actually detected and then transmitted as electrons by the photodetector. The QE of a camera is important in determining the signal to noise when detecting the light used. In general the higher that the QE of the camera being used

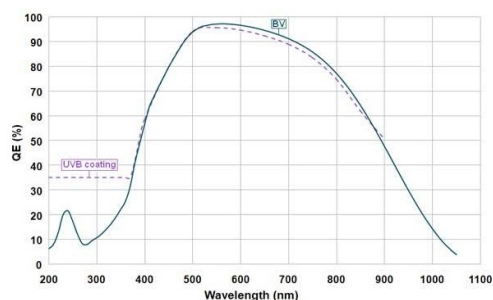


Figure V.7 Wavelength dependency of QE of the Andor iXon+ Du-897 EMCCD camera.²⁴

is, the better the better the quality of the reading will be.

One of the most important determinants of the QE of an EMCCD is the wavelength of the light that is detected.

For the purpose of the sequencing process, wavelengths

between the range of 500nm and 700nm need to be detected by the EMCCD. According to the QE vs

wavelength curves generated by Andor (see **Figure V.7**), the iXon+ 897 ranges in QE from 0.95 to 0.98 within the desired wavelength range. This is as close to an ideal QE as possible with the current technology allows.²⁴

Another important quality of the EMCCD is its readout speed. The camera should be able to detect the light produced from the reaction faster than the actual reaction occurs. That is to say the frame rate of the camera should be significantly faster than the excitation period of an individual fluorophore. This is important, since the fluorophore is excited while it is being attached to the DNA template. Once the next base is added, the fluorophore is no longer detected and the next signal has to be dealt with. Therefore in order to prevent base readings from blurring together, the frame rate of the camera should be faster than the nucleotide addition

rate for the phi 29 polymerase. With a frame rate of 31 frames per second (See Appendix D.3) the iXon+ 897 easily surpasses the maximum dNTP addition rate of 4 bps.

In order to successfully identify the presence of a fluorescent molecule, it is important to have a favorable signal to noise ratio. Signal to noise ratio simply describes the relevant signal from the fluorophore as compared to the extraneous information. This extraneous information can be due to several different factors, including background noise and shot noise. Background noise includes any signal coming from ambient sources. Shot noise occurs due to statistical fluctuations of finite number of particles detected as a result of random arrival time. The number of photons collected by the detector can be described as a Poisson distribution and the noise can be described as the standard deviation of that or the square root of the average number of photons collected. Therefore, a simplified version of the signal to noise ratio can be modeled by:

$$R_{SN} = \frac{n_p}{\sqrt{n_p}} = \sqrt{n_p} \quad \text{Equation V.1}$$

where n_p is the average number of photons detected.²⁰ From this equation, it is easy to see that as the number of photons detected increase, the signal to noise ratio increases. When accounting for noise from the signal transmission in the light detector and for the baseline noise read or the background read, the relationship becomes:

$$R_{SN} = \frac{n_p}{\sqrt{2(n_p - bkg)}} \quad \text{Equation V.2}$$

where bg is the background noise, and $\sqrt{2}$ is a correction for the noise from the EMCCD camera.²⁰ Considering all of these conditions, a signal to noise ratio near 100 would be favorable for a reliable read. Using these parameters, Eid and colleagues tested a waveguide containing phi29 polymerase and fluorophore labeled dNTPs. The lowest signal to noise ratio found was 356 for the dTTP fluorophore. This number is well above the desired ratio of 100, meaning that using the same fluorophores, the light signal can be reliably detected. The experiment also proved that a large number of photons are emitted by the fluorophores, on the order of 250000, making the background noise inconsequential.²⁰

In order to account for the potential difference in EMCCD cameras used in the experiment, one can consider the quantum efficiency (QE) of the camera.

$$R_{SN} = \frac{QE \cdot n_p}{\sqrt{QE \cdot n_p + n_n^2}} \quad \text{Equation V.3}$$

where n_n refers to the sensor noise.⁶³ Since, n_p is so high, the sensor noise is not as important, it can be ignored. Using the QE of around 0.98 for the camera in our experiment

$$R_{SN} = 0.99 \frac{n_p}{\sqrt{n_p}} \quad \text{Equation V.4}$$

which for a large number of photons emitted by the fluorophores gives a favorable signal to noise ratio estimation.²⁴

V.5 Conclusions

By using two iXon+897 EMCCD cameras along with a high-multiplex confocal microscopy setup per station, the high throughput imaging necessary to view the throughput of the biochemical reaction on the chip is possible. After analysis of the dual camera setup, it is clear that it is possible to view and differentiate the four fluorophores in the ZMW by dedicating a total of two pixels per waveguide, maximizing the throughput of the setup. Upon considering the signal to noise conditions from the setup, the use of iXon+897 EMCCD in conjunction with the fluorophores used to tag the DNA, the high signal to noise ratio proves to be favorable for quality performance viewing of the reaction taking place, and a subsequently lower error rate in reading the DNA sequence.

VI. Genome Assembly

The field of *bioinformatics* encompasses the development of databases, algorithms, and other computational techniques for the indexing and analysis of biological information, including DNA sequences.⁶⁴ This unique discipline lies at the intersection of computer science, biology, mathematics, and medicine, providing the very tools required to compile and analyze genomic data. The functionality provided is similarly diverse, ranging from protein modeling to gene mapping, and even to determining the evolutionary history of a particular organism. Indeed, without these advanced capabilities, a full-genome sequence would be of little practical value, as its relevant information content could not be decoded.

This chapter, however, will focus exclusively on the aspect of bioinformatics that is most relevant to *PennBio*'s business plan – *genome assembly*. With a sample of some 27 million fragments of random length and sequence being generated during the observation of the SMRT chip, this final step in genome “production” is perhaps the most difficult and resource-intensive. Many different assembly techniques exist, all varying in experimental fragment length, knowledge of the target sequence, and specific alignment algorithms.

The following discussion of the Human Genome Project and Celera Genomics illustrates quite clearly the rapid, almost quantum advances that have been made in genome reassembly in the last two decades, and provide important background and precedent for the *PennBio* method. Further detail will then be given regarding the specific considerations and calculations that were involved in the development of a reliable, efficient assembly procedure. Finally, the results of an original Monte Carlo assembly simulation will be discussed with respect to its usefulness as a model validation tool, and as a means of predicting the computational resources necessary to meet *PennBio*'s throughput goals.

VI.1 The Human Genome Project

The early whole-genome sequencing ventures, including the Human Genome Project (HGP, 1990-2003), took what is now considered a brute force approach to sequencing determination⁶⁵. The complete human genome was fragmented into long strings, some 150 kbp in length, and distributed to laboratories around the world for analysis. These genome fragments were sequenced in a processive manner, starting at one end and proceeding to the other, one base

position at a time, before the completed fragments were mapped to chromosomes and reassembled into a single consensus genome. This method is referred to as the “hierarchical shotgun” approach, and was selected by the HGP principally for its ability to accurately map repeat-rich sequences, and also because it allowed the project workload to be shared across several analysis sites. In the *Nature* article announcing the endeavor’s completion, the authors insist that “the advantages of this more conservative approach outweighed the additional cost, if any.”⁶⁶

During its thirteen year timeframe, the HGP was the focus of over 20 public molecular biology laboratories and approximately \$3 billion in public funding. Its contributions to genetic science, particularly as an ardent proponent of the human genome as public-domain information, are indisputable, but many have questioned the efficiency of their sequencing method. In fact, just eight years into the Project, a private biotechnology firm – Celera Genomics – launched a parallel human genome sequencing venture based on the whole-genome shotgun (WGS) technique.

VI.2 Whole-Genome Shotgun Sequencing

Whole-genome shotgun sequencing is distinguished from the hierarchical method by its short read lengths. In contrast to the 150 kbp fragments used by the publicly-funded effort, Celera generated fragments a mere 550 bp in length⁶⁷. The entire fragment library was then aligned simultaneously to create a full consensus genome, rather than first producing long intermediate sequences. By 2001, Celera had caught up to the public HGP, and the two

competitors published their draft genomes within two months of each other. The price of Celera’s finalized sequence, however, was reported to be \$300 million – a tenth of the public funding required. As striking as this cost differential is, it must be considered in context. Celera’s project had, on its first day, free and unrestricted access to the public project’s progress, which was updated *daily*. The extant sequence information made the mapping of short fragments much less uncertain than the initial *de novo* assembly case, and permitted the use of less rigorous assembly techniques⁶⁸. It is difficult to say whether the approach would have been as successful without a template, incomplete though it was. Indeed, the scientists on the public side may have been right to choose the more painstakingly accurate method.

But regardless of what could have been, Celera’s success confirmed that WGS sequencing was a viable and time-efficient alternative to the older techniques. Modern sequencing operations are nearly unanimous in their acceptance of this approach, and it has benefitted considerably from advances in computing power and assembly algorithms.

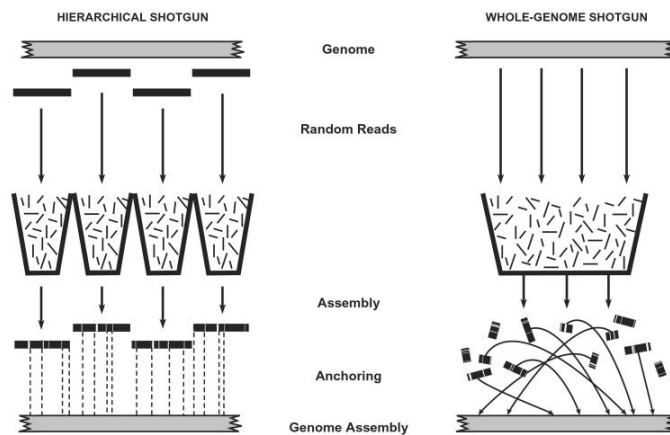


Figure VI.1 A graphical comparison of the hierarchical (left) and whole-genome shotgun (right) sequencing methods. Reproduced from Mihai (2004).

VI.3 The *PennBio* Strategy

PennBio's reassembly strategy combines the sensitivity of WGS with the public-domain human genome sequence in a process often called *comparative genome assembly*⁶⁵. As in the classic technique, customer genomes are randomly fragmented into 2000 bp lengths, but rather than being aligned to each other to produce a consensus sequence, each fragment is mapped to the available genome. Given that individual sequences differ by one base in everything thousand, on average, and the polymerase-related error rate is extraordinarily low modern alignment algorithms can execute the mapping with a success rate of nearly 100%, and in a fraction of the time and computational complexity required by *de novo* assembly⁴⁶.

In this way, *PennBio* provides its customers with the most sophisticated of modern analysis technique, resting on the shoulders of nearly two decades of human genomic study. Exceptional accuracy, not only in identifying known SNPs, but at every base position is what makes each sequence an essential tool in the most advanced molecular diagnosis both today, and in the future as our medical understanding of the genome continues to grow. At the same time, low cost and unprecedented throughput for bring this indispensable resource to the average customer for the first time.

VI.4 The Coverage Problem

The most significant shortcoming of the shotgun method is its dependence on an overgeneration of information, *i.e.*, the requirement that the total length sequenced is several

times longer than the target genome, itself⁶⁹. With no such redundancy, the random nature of the synthesis reaction would certainly yield gaps in the final genome. But as the total length sequenced increases, the probability of missing any particular base position decreases. The relationship follows a binomial distribution:

$$\Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{Equation VI.1}$$

where K is a random variable for the number of successes, k is a realization of this variable, n is the number of trials performed, and p is the probability of success for each trial. Since the number of successes is typically specified by the requirements of the reassembly algorithm, an alternative formulation of this distribution – the negative binomial – is more useful. This expresses the probability, $f(n)$, that k successes will be achieved in n trials:

$$f(n) = \binom{n-1}{n-k} p^k (1-p)^{n-k} \quad \text{Equation VI.2}$$

Given a particular base position, a success is achieved if a sequenced fragment exists containing that base, under the assumption that it can be properly positioned by a reassembly routine. Defined more rigorously, a success is achieved if the polymerase under observation has at any point added a nucleotide at that particular position, generating a measureable signal, which was then translated into a character in a random fragment which can be accurately positioned in the final genome. Each of these steps in the single-position sequencing operation has a likelihood of failure, and contributes to the determination of p . Therefore, each of these processes underwent

individual statistical modeling in order to generate a final estimate for the redundancy required for complete coverage. The following analysis is performed at the single base position level.

VI.4.1 Probability that the Polymerase Reached the Base Position

If a certain base is to be sequenced the polymerase must, at very least, physically arrive at that particular position. Since polymerization is unidirectional, the probability that this requirement is satisfied can be divided into two sub-probabilities: that the polymerase attaches to the template strand before a given position, and that it does not release the template before reaching it.

Phi29, and all DNA polymerases, will only bind to double stranded DNA. Since our templates are necessarily single-stranded, short lengths of primer ssDNA are annealed to it, providing the enzyme with attachment points. Template priming in this design is accomplished by means of random-sequence six-base fragments of ssDNA (hexamers). These are demonstrated to bind to the target ssDNA genome fragments randomly, and since polymerization can only begin where a primer is bound, the actual sequenced strings will, themselves, have a random distribution that is both a function of fragment length and the relative primer and target compositions.

Random primer binding is modeled by the Poisson distribution, with a rate parameter, λ , and a total nucleotide distance, x (often called *exposure*)⁷⁰:

$$\Pr(R = r) = \frac{(\lambda x)^r e^{-(\lambda x)}}{r!} \quad \text{Equation VI.3}$$

This equation expresses the probability of observing r primer binding events over a length of x bases of ssDNA. In this case, λ is defined as the molar ratio primer to template, which is equivalent to the ratio of the number of primer hexamers, $n_{primers}$, template fragments, n_{frag} :

$$\lambda = \frac{n_{primers}}{n_{frag}} = \frac{n_{primers}}{l_{genome} / \mu_{frag}} \quad \text{Equation VI.4}$$

where l_{genome} is the length of the human genome (3 gbp) and μ_{frag} is the mean length of a fragment (2 kbp). Taking the genome and mean fragment lengths as constants, this parameter can be optimized by varying the number of primer molecules present for a given coverage multiplicity.

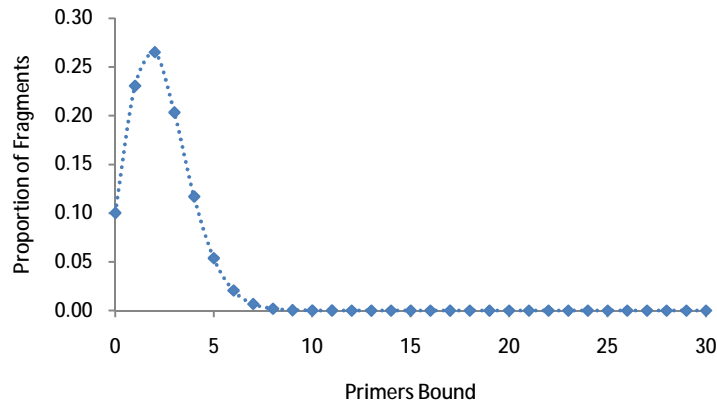


Figure VI.2 The optimal priming strategy. A less aggressive approach reduces primer-excision lags and conserves the relatively expensive reagent (blue).

Realistically, there will be a large excess of template DNA in the reaction mixture. Therefore, low concentrations of unprimed template are not expected to have a significant effect on the overall sequencing rate. It becomes more important to consider the number of primers per fragment, since the polymerase must pause to remove these if they are encountered during polymerization, causing a noticeable delay. In this case, the acceptable proportion of unprimed fragments was set at 10^{-1} , yielding a rate parameter of 2.3 (equal to the mean), which represents a 6-fold decrease in the amount of random primer required for each sequencing reaction without sacrificing performance.

Assuming priming is optimal and the enzyme has begun to sequence its template, there is a probability of it detaching before polymerization is complete. This probability increases with total distance traveled along the template – a characteristic distance that varies greatly from enzyme to enzyme. Indeed, phi29 was selected in part due to its capacity for very long read lengths, with a reported mean value of 70 kbp. Like primer binding, the occurrence of a release is

a Poisson event. Therefore, the distance before the first release can be modeled by the cumulative gamma distribution, with a mean of 70,000 and shape parameter, k , equal to 1. In this special case, the gamma distribution is equivalent to the exponential distribution with the rate parameter, λ , equal to $1/70,000$:

$$\Pr(X \leq x) = 1 - e^{-x/\lambda} \quad \text{Equation VI.5}$$

where X is a random variable denoting the number of bases traveled, x is a realization of X , and $\Pr(X \leq x)$ is the probability of a release occurring at or before x . For a mean fragment length of 2000 bp, the probability of release at or before the terminal base was 2.8% – fairly low, even assuming the polymerase covered the full length of the fragment.

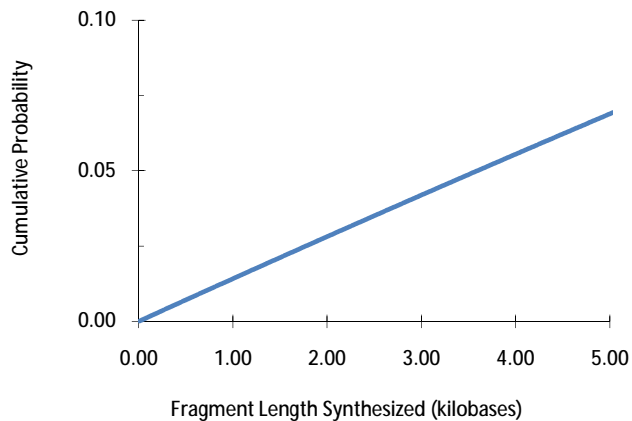


Figure VI.3 The probability of incomplete polymerization with respect to fragment length.

VI.4.2 Probability of Misidentifying a Nucleotide

While phi29 is capable of remarkably high fidelity, the system's optical limitations result in an error rate of 1 in every 500 nucleotides. Like primer binding, the total number of errors per fragment, R , over a distance, x , is Poisson distributed with an expected rate of λ^x . Therefore, for a sequenced string of 1000 bp, the mean number of errors is 9, with a 99% confidence interval of [0,22].

VI.4.3 The Complete Negative Binomial Estimate

Given the parameters derived above, the probability of success at any particular base position is 0.963. Unambiguous nucleotide assignment requires three successful trials per position ($k = 3$), and an incomplete coverage rate of 1 per 100,000 bases sequenced.. Evaluation of the distribution with these values yields a final minimum estimate of 7-fold redundancy.

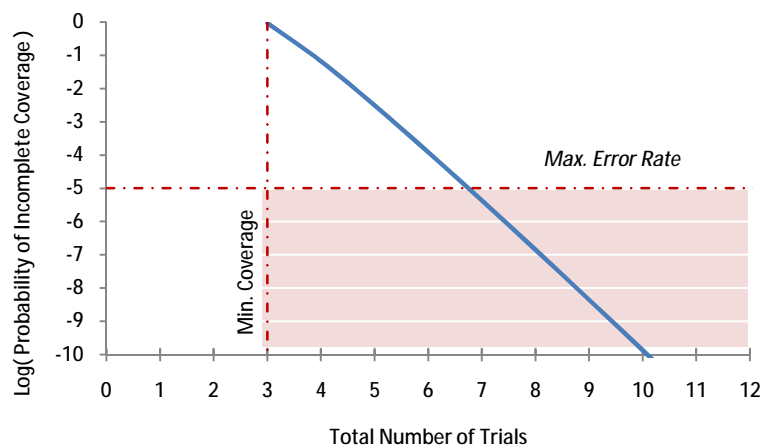


Figure VI.4 Negative binomial estimate of required coverage multiplicity. Points within the shaded area satisfy by the coverage and error rate specifications, and are considered acceptable.

This multiplicity value is critical to the calculation of throughput, which in turn is central to PennBio's business model. With so many random events involved in DNA sequencing, further validation of these estimates is certainly warranted. A Monte Carlo MATLAB program was developed to simulate the phi29-mediated polymerization process – allowing the manipulation of release rate, various error rates, and target genome length – followed by an algorithm designed to reassemble the sequenced fragments and assess any errors made. While the realistic genome generation routine is purely for purposes of model validation, the latter reassembly portion bears remarkable similarity to alignment software in use today, and could be used to reassemble personal genomes during operation.

The purpose of this section is not to delve too deeply into the programming behind this simulation, but rather to illustrate the rigor *PennBio's* process evaluation, and to provide repeatable, statistically sound evidence of the project's practical viability.

VI.5 Simulation Overview

VI.5.1 Genome Generation, Fragmentation, and Polymerization

A random genome of specified length is generated by the *randseq* function and stored as the “consensus” genome – the simulation-level equivalent of the HGP sequence. Single-nucleotide polymorphisms (SNPs) are then inserted at exponentially distributed intervals to create the “actual” genome sequence of interest – the customer genome. This “actual” genome is sonicated into fragments of normally distributed lengths about a mean of 1000 bp, which are fed

to a phi29 polymerase. As it polymerizes each fragment, the enzyme introduces errors at exponentially distributed intervals. Deletions are also inserted to reflect possible detection faults, and to evaluate the robustness of the alignment algorithm to phase shifts.

This polymerization protocol is repeated a specified number of times, randomly storing or deleting fragments as they are created according to the output of a uniform random number generator, with a probability of fragment storage of 1%. Fragment generation ceases when the desired coverage multiplicity is achieved, simulating the randomness with which the polymerase sequences the excess of DNA fragments present in the reaction mixture and allowing evaluation of the previous redundancy estimate.

VI.5.2 Reassembly of Random Fragments

These fragments are then locally aligned to the “consensus” genome by the built-in Smith-Waterman algorithm (*swalign* function), recording the starting point of the alignment and fragment length. These data are used to keep a running tally of nucleotide “votes” at each position in the genome. After all fragments are aligned in this way, the votes are counted, and a value of A, C, T, G, or X is assigned to each position, generating the “final” genome, with X signifying ‘blank’ or ‘inconclusive.’

A short script at the end of the program then compares the “final” genome to the “actual” starting genome and generates an error rate. The application of various sensitivity analysis scripts allows the examination of the effects of various computational and biological parameters on the error rate, which must be less than 1 in 100,000 bases.

VI.6 Initial Simulation Results

Due to memory restrictions, the initial simulations were carried out on genomes with lengths of less than 20.000 kbp, with 10.000 kbp being the most common length used. The final error rates of these first simulations were on the order of 10^{-3} – well above the specified threshold to 10^{-5} . Many of these resulted from a failure of the alignment algorithm to correctly recognize deletions within the fragments, so the *swalign* subroutine was then modified by varying the penalties for opening gaps in the sequence, and the penalties for extending these gaps. This improved the mean error rate to 10^{-4} , which seemed to be independent of any further adjustments to program parameters.

Subsequent examination of the particular error locations revealed that they were caused by the concurrence of insufficient coverage (only two bases per position) and ambiguous base definition (a tie vote – the two bases were not identical). This was inconsistent with our statistical coverage predictions. The program responded to such situations by designating that base ‘X’ as instructed, eventually resulting in an error in the final alignment. Two options existed in correcting this unanticipated deficiency – either establishing a protocol by which the HGP genome would resolve the ambiguity, or increasing the multiplicity.

Initially, the first course of action appeared more attractive, as it required little adjustment to the physical and temporal requirements of the sequencing system. Moreover, since the sequencing technique is based on knowledge of the human consensus sequences, an extremely accurate tie breaker was readily available. Deferring to the consensus sequence in the cases of

ambiguous base assignment was expected to fail if and only if the base in question was a SNP. In this scenario, completely random assignment would actually be more reliable, as by definition of “SNP,” it would not agree with the known sequence. Therefore, the expected error rate was a product of the SNP frequency (10^{-3}) and the frequency of insufficient, ambiguous coverage (10^{-4}), which was on the order of 10^{-7} errors per base sequenced.

While this seems acceptably low, and is certainly below the threshold of 10^{-5} , it is nonetheless unsatisfying. First, the probability must be put into perspective, remembering that a human genome is approximately 3 gbp long. We would therefore expect to misidentify an average of 30 SNPs on each genome sequenced. Second, and perhaps even more importantly, SNPs are considerably more significant in the molecular diagnosis of diseases. This is why most sequencing technologies, until now, have relied on SNP screening – providing only the most effective data at a reasonable price. To allow even 30 SNPs per genome to be improperly analyzed would dramatically decrease our product’s diagnostic power, and could make it difficult to displace the SNP screening assays already in common use.

Instead, the decision was made to increase coverage multiplicity until the average error rate was below 1.00×10^{-5} errors per genome. As the simulation is constructed, this was a very simple operation, merely requiring the manipulation of a single input constant (*mult*) within the sensitivity analysis loop. Values between 8.00 and 11.00 were examined with a step size of 0.25 before concluding that a multiplicity of 9.00 proved the optimal balance between error rate and physical equipment limitations.

VI.7 Final Error Rate Calculations

Satisfied with the program's performance, and with the biological accuracy of the input parameters, the final objective of this simulation was to determine the expected error rate in each sequenced genome. To this end, a large sample was generated by simulation iteration. The inputs were fixed, and are given in **Table VI.1.A**. The iteration routine collected final error rates from $n = 660$ independent trials using 20,000 kbp target genomes. These were then consolidated into a single table of count, given in **Table VI.1.B**, which shows the number of trials that were completed with a certain number of errors.

A. Inputs		B. Results	
Parameter	Value	Errors Per Trial	Count
<i>Genome Length</i>	20 kbp	0	593
<i>Fragment Length</i>	1 kbp	1	48
<i>Multiplicity</i>	9-fold	2	14
<i>SNP Rate</i>	1/1000	3	5
<i>Error Rate</i>	1/500	4	0
<i>Deletion Rate</i>	1/500	5	0

Table VI.1 Input parameters in the determination of error rate (A). The simulation results, as count data (B).

Like many of the molecular-level processes described in previous sections, instances of disagreement between the final genome and the target genome were assumed to be rare and Poisson distributed. In order to test this hypothesis, the index of dispersion was computed for this data set using the formula:

$$D = \frac{s^2}{m} = \frac{0.207}{0.138} = 1.502 \quad \text{Equation VI.6}$$

where σ^2 is the sample variance, and μ is mean number of errors per 20.0 kbp trial. In the case of the Poisson distribution, the mean and variance are equal, and $D = 1$. When the index of dispersion is greater than one, the events are said to be *overdispersed*, and are better characterized by a negative binomial distribution. This adds a dispersion parameter, α , to the calculation, and better accounts for unobserved heterogeneity in sample. In fact, the Poisson distribution is simply a special case of the negative binomial with $\alpha = 0$. If this can be shown to be the case with the sample data, a Poisson approximation would be valid.

This possibility was evaluated by performing a negative binomial regression *nbreg* on the sample (Stata 10.2), which estimates, among other things, the dispersion parameter. For this sample, the reported value was $\alpha = 0.099$ with a standard error of 0.067 . A z-test of the hypothesis that $\alpha = 0$, however, returned a p-value of 0.138 . We therefore fail to reject the null hypothesis, and are justified in our use of the Poisson distribution in fitting the data.

The maximum-likelihood estimator of the error rate is given by the formula:

$$\hat{I}_{MLE} = \frac{\sum_{i=1}^n R_i}{n} \quad \text{Equation VI.7}$$

This is equivalent to the arithmetic mean of the number of errors, R , over n independent trials.

The error rate for the actual sample is therefore given by:

$$\hat{I}_{MLE,avg} = \frac{91 \text{ errors}}{(660 \text{ trials}) \cdot (20,000 \text{ bp/trial})} = 6.89 \times 10^{-6} \text{ errors/bp} \quad \text{Equation VI.8}$$

Upper (UL) and lower limits (LL) of the confidence interval can then be constructed about this value using the chi-square distribution⁷²:

$$LL = \frac{C_{2(R+1), (1-\alpha/2)}^2}{2 \cdot (20,000 \cdot n_{avg})}$$

$$UL = \frac{C_{2R, \alpha/2}^2}{2 \cdot (20,000 \cdot n_{avg})}$$

Equations VI.9, 10

where R is the total number of errors observed, α is the significance level, and $(20,000 \cdot n)$ represents the sample in total bases. Several possible confidence intervals are given in **Table VI.2**, and show that that only at $\alpha = 0.0001$ (99.99% CI) does the upper limit cross the 10^{-5} threshold. This provides quantitative evidence that the probability of producing a genome with an error rate higher than 10^{-5} is extremely, and acceptably, low under the specified biological conditions.

	95% CI	99% CI	99.9% CI	99.99% CI
<i>Lower Limit</i>	5.55×10^{-6}	5.17×10^{-6}	4.76×10^{-6}	4.43×10^{-6}
<i>Upper Limit</i>	8.46×10^{-6}	8.98×10^{-6}	9.61×10^{-6}	1.02×10^{-5}

Table VI.2 Confidence intervals for various values of α , centered about a mean of $\lambda_{avg} = 6.89 \times 10^{-6}$.

As mentioned previously, reliable SNP identification is essential if current SNP-screening services are to be displaced by whole-genome sequencing. Indeed, if the error rate in the most important segments of the genome were found to exceed the 10^{-5} threshold, the value of average error rate would be meaningless. The code was therefore run for an additional $n = 660$

iterations, this time with a new subroutine that tracked the SNP-specific error rate. The Poisson rate parameter was calculated, using the formula state above, to be:

$$\hat{I}_{MLE,SNP} = \frac{2 \text{ errors}}{26453 \text{ SNPs}} = 7.56 \times 10^{-5} \text{ errors/SNP} \quad \text{Equation VI.11}$$

Likewise, the upper and lower limits of the confidence intervals about this mean are:

$$LL = \frac{c_{2(R+1), (1-\alpha/2)}^2}{2 \cdot (n_{SNPs})} \quad \text{Equations VI.12, 13}$$

$$UL = \frac{c_{2R, \alpha/2}^2}{2 \cdot (n_{SNPs})}$$

Table VI.3 shows the confidence intervals for the SNP-specific error rate, which were calculated using **Equations VI.12** and **VI.13**. Unlike the average error rate, this parameter does not, at first, appear to be below the error threshold. The discrepancy, however, arises from the much smaller sample size. While the average error rate was calculated from 1.32×10^7 bp over 660 trials, only 26,453 SNPs were recorded in twice as many (1,320) trials. Given the confidence intervals, however, it is clear that even for $\alpha = 0.05$, we cannot reject the hypothesis that $\lambda_{SNP} = 10^{-5}$. Judging from experience in calculating λ_{avg} , the mean will continue to fall as the sample size increases.

	95% CI	99% CI	99.9% CI	99.99% CI
<i>Lower Limit</i>	9.16×10^{-6}	3.91×10^{-6}	1.21×10^{-6}	3.79×10^{-7}
<i>Upper Limit</i>	2.73×10^{-4}	3.51×10^{-4}	4.56×10^{-4}	5.56×10^{-4}

Table VI.3 Means and 99.9% confidence intervals for the average and SNP-specific error rates.

VI.8 Computing Time and Code Optimization

With the error rate under control, the practical issue of computing time required to sequence a complete genome was addressed. As previously mentioned, memory was a limiting factor in the choice of representative genome length. In addition to this restriction, the time required to generate, fragment, and reassemble the 10.0 kbp genome was approximately 15 minutes. In a best case scenario where computing time scales linearly with target genome length – a relationship that is, in fact quadratic – this translated into a 3 gbp sequencing time of 2.70×10^8 seconds, or about 8.65 years.

The physical computing specifications were inflexible, so code optimization became the primary focus of attention. Preliminary alignment script analysis was carried out using the MATLAB Profiler tool, which records and graphically displays the number of calls to each function, and the total amount of time these calls represent. This revealed the alignment algorithm itself (*swalign*) to be responsible for an overwhelming proportion of computing time and memory usage. The Smith-Waterman algorithm, while very robust to any number of input errors, is known to be very resource demanding⁷³. Both the alignment time and memory requirements scale by the *product* of the two sequence lengths⁷⁴. Given this non-linear relationship, decreasing sequence length was expected to produce significant decreases in processing time. These lengths were, however, fixed either by nature or by other practical considerations. Instead, a hypothetical “short fragment” was created, allowing the alignment

process to be divided into a fast position determination step, followed by a much more restricted full-fragment alignment.

First, the beginning 10 bp (1%) of the fragment of interest were aligned to the full consensus genome. This enabled the same positioning accuracy and repeatability as full-fragment alignment, but took a fraction of the time, as it reduced the memory requirements 100-fold. The alignment starting point was then fed to a second *swalign* call which aligned the full fragment to that particular section of the genome, once again dramatically reducing computing time, despite adding extra steps. With this two-phase method in place, computing time for a 10.0 kbp genome decreased from 15 minutes to 15 seconds – a 60-fold reduction.

The code was further optimized according to the MATLAB standards outlined in “Techniques for Improving Performance,” by replacing all non-essential cell arrays with matrices, as MATLAB cannot accelerate *for* loops involving cell arrays⁷⁵. Vectorization of all remaining loops, where possible, also increased performance. Finally, the full script and certain complex operations within it were converted to functions, allowing MATLAB to more efficiently load them into memory before execution. Computing time was further reduced from 15 seconds to less than 5 seconds for a 10.0 kbp target genome at 9.00-fold multiplicity.

Target genomes up to 1.0 megabases in length were evaluated using the final program. Computing time was approximately 3.1 hours for full assembly, with an average error rate of 8.5 errors per genome ($n = 2$). This is a significant improvement over the original 10.0 kbp targets, and primarily serves to demonstrate the program’s much-improved memory utilization.

(The complete simulation script, including subroutines, can be found in **Appendix E**.)

VI.9 Data Collection and Processing

Each sequencing station will be directly connected to a signal acquisition server in the adjacent room. This unit will compare the frames of the two EMCCD cameras, identify the fluorophores in each pixel, and convert the frame-based data into time-series data. This is then translated into a nucleotide sequence by standard signal processing techniques and sent directly to the genome reassembly server, rather than being stored. By not storing the fragments until a complete set is generated, both sets of computers are utilized continuously, and no precious processing time is wasted.

The genome reassembly servers are extraordinarily sophisticated, and were selected for their exceptional memory efficiency and processing bandwidth. The sheer volume of data passing through them in a single day demands such high-performance features. Local sequence alignment speed, in particular, is particularly responsive to memory access rates and multithreading capability. Indeed, successful reassembly is achieved only by minimizing the time required by each individual task while maximizing the number of tasks that can be performed simultaneously. Just as parallel operation is central to the overall process design, the individual genome fragments are distributed to the many subunits of a multiple-core processor. But with the complexity of modern computing systems, the result of this parallelization is not always obvious, and deserved a more rigorous treatment.

A preliminary assessment has been based on the properties of the reassembly simulation, providing an estimate of the total computation required. Standardized performance benchmarks were then used to more precisely define acceptable system specifications.

VI.10 Computational Demands

Using the finalized version of the assembly program described in **Section VI.5**, the functional relationship between target genome length and alignment time was determined. Rather than calculating total, single CPU processing time, however, the process was examined at the single-fragment alignment level. That is, the time required to align a single 1.0 kbp fragment was evaluated as a function of reference genome length. As shown in **Figure VI.5**, this effect is linear, and has the form:

$$time = 0.00021 \cdot genlength \quad R^2 = 0.9999 \quad \text{Equation VI.14}$$

where *time* is in seconds, and *genlength* is the reference genome length in kilobases. In this case, the intercept was set to zero, as the time required to create and manipulate empty matrices is negligibly small with respect to any non-trivial alignment operations.

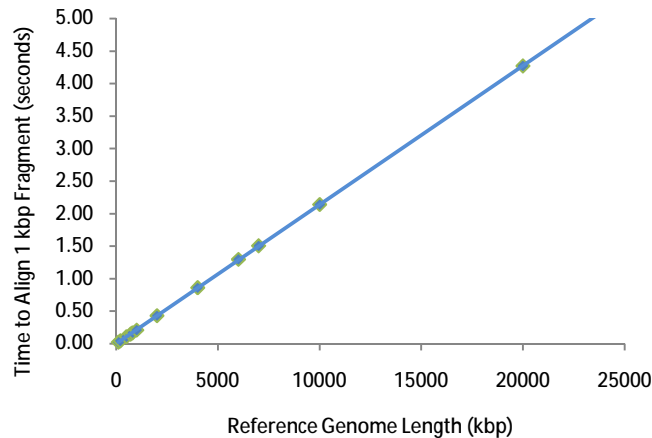


Figure VI.5 The time required to align a single, 1 kbp fragment to a reference genome of variable length. The line is a least-squares regression fit, with $R^2 = 0.9999$.

These alignments were performed on an ordinary desktop computer with a 2.4 GHz Intel Core 2 Duo 6600 processor and 2.0 GB RAM. Given the relatively low computational power of this system, the total time required to align a 1.0 kbp fragment to a 3.0 gbp genome (3,000,000 kbp) was determined to be 630 seconds. Assuming a 3-fold increase in calculation speed could be accomplished by translating the code into a more efficient language (such as C/C++), 210 seconds could, in practice, be realized. Properly scaled by the number of fragments, this translates into a total genome reassembly time of 9.72×10^{10} seconds, or approximately 3,000 years at 9-fold coverage multiplicity (see Appendix A for detailed calculations). This value is obviously and obstacle with a successful business model, and can only be overcome by way of extraordinary computational resources, especially those that allow high-speed, massively-parallel task execution.

VI.11 Multi-Processor Speed-Up and Amdahl's Law

The decrease in processing time afforded by the use of multiple processors (or multiple processor cores) is fundamental to data-intensive computing strategies. In order to best understand the benefit of a multiple-CPU approach, each computational step in the overall sequencing process has been analyzed and separated into its “serial” and “parallelizable” parts. Those designated “serial” must be performed in order, as each operation requires input from the one before. Any tasks that can be accomplished independently are said to be “parallelizable,” and to the extent that expenses allow, are simultaneously addressed by multiple processing units.

The multiplicative factor by which performance is increased when additional processors are added is referred to as “speedup,” and can be calculated using Amdahl's Law. This formula expresses speedup, S , as a function of the number of processing units available, N , and the fraction of the task that can be parallelized, P :

$$S = \frac{1}{(1 - P) + \frac{P}{N}} \quad \text{Equation VI.15}$$

If 75% of an operation can be parallelized, for example, and it is divided among 4 processors, the calculation speed is expected to increase 2.3 times. This is equivalent to saying that the time required to complete the operation would decrease 2.3-fold.

The alignment routine itself is responsible for an overwhelming majority of the total computing time. In fact, as the genome length becomes very large, the fraction of the computing time required by this segment of the program goes to approaches unity. Fortunately, this process

is almost 100% parallelizable, *i.e.*, $P \approx 1$. Each single-fragment alignment is completely independent of the others, with the only interaction occurring in the final vote counting procedure. Therefore, each computation-intensive alignment task could be delegated to a separate processing unit. In this unusually clear-cut case, each additional processor is expected to increase the overall processing speed by the same amount, and none of the diminishing returns modeled by Amdahl's law will be encountered.

VI.12 System Selection

In order to reduce the overall data processing time to a single day, a bank of IBM p560 Express servers will be utilized. These units are distinguished by remarkably high processor bandwidths, large L1 and L2 caches, and most importantly, several independent processing cores. They are also designed for practicality – being extraordinarily energy efficient and compact. **As shown in the previous section, the time required to align a 1.0 kbp fragment to a full-length target genome on a typical office computer is:**

$$time_{pc} = 0.00021 \cdot (3 \times 10^6 \text{ kbp}) = 630 \text{ s} \qquad \text{Equation VI.16}$$

Assuming the use of a lower-level programming language increases performance 3-fold, the time required would be 210 seconds.

A single-core p560 server is at least 10,000 time faster than the computer used in the above calculation, as measured by IBM's commercial processor workload (CPW) index. (The scale for this metric is normalized such that the processing speed of a midrange IBM

System i server is defined to be equal to 1.00). The server model selected also has eight independent CPU cores. Since the alignment routine is almost 100% parallelizable, processing speed is increase by the factor:

$$S = \lim_{P \rightarrow 1} \left(\frac{1}{(1-P) + \frac{P}{8}} \right) = 8 \quad \text{Equation VI.17}$$

The maximum final genome assembly time at 9.00-fold multiplicity is therefore:

$$t_{server} = \frac{9.00 \cdot (210 \text{ s/frag}) \cdot (3,000,000 \text{ frag})}{10,000 \cdot (8 \text{ cores})} = (70,875 \text{ s}) \cdot \left(\frac{1}{3600 \text{ s/hr}} \right) = 19.7 \text{ hr} \quad \text{Equation VI.18}$$

VI.13 Conclusions

While the assembly of a full human genome is, indeed, a complex undertaking, it has been demonstrated not only to be feasible, but adherent to *PennBio*'s quality standards, and consistent with single-day sequencing operations. The alignment program has been proven effective and efficient, and has itself validated the probabilistic sequencing models developed. Given sufficient computing resources – which are well within reason and budget – this program can be run from start to finish in approximately 19.7 hours, as shown above, leaving a generous margin for technical difficulties or delays. As a whole, the quantitative evidence provided in this chapter has been fundamental to the practical evaluation of the *PennBio* sequencing approach, and ultimately, the viability of its business model.

VII. Financial Analysis

While the previous sections discussed the biochemical and technical aspects of this project, it is imperative to analyze the financial aspect of the technology to determine whether the project is financially feasible. If there is no existing market and no cash can be generated, no investors would fund this project. Consequently, the technology would not be exposed to the public. The financial analysis will show that the project is profitable. It will explain how that decision was determined using NPV and MIRR analysis, and how those figures were calculated. It is important to note, however, that these valuations are based on projected earnings, which in turn, are based on several assumptions.

The analysis begins with revenue projections based on the genomic throughput and price. Because the entire model depends heavily on the revenue, a separate sensitivity analysis is done on the genome price. Next, the total costs and depreciation are explained. Knowing these two elements leads us to build an income statement. But because the income statement shows earnings, and we want free cash, we want to adjust those earning figures into cash figures. To do so, we then examine working capital and other cash affecting items.

Once we have the projected free cash flows, we can then value the company by combining terminal value analysis and discounted cash flow analysis. Next, we do a rate of return analysis for the investors. Because our project involves two rounds of investments from two separate investors (series A and series B), this section becomes a little trickier than the conventional analysis. To simplify the complications created by the multiple investments and investors, we conduct an equity stake analysis. This will then complete our return of rate analysis for the two groups of investors. Having completed all these explanations, we will put everything together onto a single page spreadsheet.

Finally, we discuss multiple what-if scenarios and how that impacts the bottom line. The scenario analysis will also involve a genome price sensitivity analysis to determine at which point the entire project would lose money.

VII.1 Market and Revenue Projection

The genome sequencing market is a relatively new and volatile market, originally tailored to high net worth individuals. Currently, one existing company, Knome®, is charging \$100,000 per client. Furthermore, due to technological improvement incentives like the Archon X Prize competition, technology can be expected to improve. If that were the happen, costs would significantly drop, and the price charged to clients would also fall dramatically. Little information is known about sales volume, but even if it were known, the current economic recession may render that figure irrelevant. Consequently, revenue projections are difficult to nail down because genome sequencing is not a mature and predictable market.

For purposes of illustration, we will assume that our genome price is \$10,000 for the next several sections. Later on, we will conduct a sensitivity analysis on this price because the entire financial analysis is heavily dependent on this figure. We also assume that 100% of our design capacity is to sell the throughput of 3000 genomes per year: That is, our design capacity would allow the company to gross \$30MM per year.

The growth and development of the company fall into four stages: the research stage, the scale up stage, the sales stage, and the terminal stage. During the research stage, scientists develop a working prototype. There is no revenue and all the needed capital is provided by the series A investor. Next, the scale up stage is where a working prototype has been developed, at which point series B investors fund the rest of the necessary capital. New staff is added and there is a step up increase after the company starts to make sales at a percentage of the design capacity.

Financial Analysis

Ideally, this figure will be 50% of nominal capacity. After that, the sales stage is when the company is fully functional and makes 100% of design capacity. Grown sufficiently, the company makes the most money in this stage. Finally, the terminal stage is how the company will end. A terminal value of the company will be calculated based on the prior free cash flow projection. However, because there are multiple scenarios in the terminal stage, this will be discussed in more detail in a later section.

Ideally, the research stage will take one year, the scale up stage another one year, the sales stage three years, and the terminal stage will have one terminal value associated with it. Because the genome sequencing market is a volatile, technology-related market, the company will have a short lifetime.

The following table summarizes revenue projections.

Revenue Projections						
Year	<u>2010</u>	<u>2011</u>	<u>2012</u>	<u>2013</u>	<u>2014</u>	<u>2015</u>
Stage Name	Research	Scale Up	Sales	Sales	Sales	Terminal
Design Capacity	0%	50%	100%	100%	100%	100%
Revenue	\$ -	\$ 15,000.0	\$ 30,000.0	\$ 30,000.0	\$ 30,000.0	Terminal Value

Table VII.1 Revenue projections for the following 5 years. (\$ in thousands).

It should be noted that inflation is ignored from all financial analysis. The inflation calculations are unnecessary and are only relevant in settings experiencing hyperinflation. In fact, the Financial Accounting Standards Board determined in 1986 that inflation accounting is unnecessary for financial statements.

VII.2 Costs, PPE, Depreciation

There are multiple, different costs associated with running *PennBio*. The costs can be divided into four categories: research equipment purchase, research annual cost, sales equipment purchase, and sales annual cost. The research equipment purchase includes all necessary laboratory equipment for developing a working prototype. This includes microscopes, computers, stations, EMCCD, etc. The research annual cost is how much the company spends during the research stage. This mainly includes salary and rent. The sales equipment purchase is the rest of the laboratory equipment purchased once a working prototype has been developed. This equipment is used to scale up production. The sales annual cost is the annual burn rate of the company when the company is able to generate revenue. This burn rate is significantly higher than the previous burn rate: In addition to salary and rent, inventory costs, research and development costs, and sales costs have been added. These cost figures are summarized in table X.2.

The series A investors are paying for the research equipment purchase and one year's value of research annual cost as an initial fund. Because the research stage is the riskiest stage (developing new technology), the series A investors are usually wealthy angel investors. In this project, the series A investment comes to a total of \$1.2MM.

The series B investors come to invest once the risk of research has been reduced: only when a working prototype has been developed, do the series B investors give funds to scale up production. Their investment will be used for the sales equipment purchase, and 3 months worth of salary, rent, and inventory, assuming the company will reach 50% of design capacity. The rest of the salary, rent, and inventory will be funded by the revenue generated, but 3 months worth of capital should be enough for the company to stay liquid. The series B investment comes to a total of \$3.5MM, which will be funded ideally one year after the series A investment, or whenever the research stage is finished.

The labor costs for the research stage and the non-research stages differ. During the research stage, the company must try to minimize all forms of cost, and salary is no exception. Workers consist of a single secretary and four senior scientists, one of whom will also function as a chief technical officer (CTO). In addition, because salary costs are large, and the series A investor is trying to minimize his capital's exposure to unnecessary risk, the scientists and series A investor have come up with an agreement: the scientists will receive 70% of their ordinary salary until a prototype is developed, in exchange for 10% of the company. Normally, in any venture, the investor receives about 85% of the company to justify for risk, especially with people inexperienced with entrepreneurship. But because this is a biotechnology company, and biotech companies carry significantly higher risk, the series A investor feels he must receive 90% instead. Once a prototype has been developed, the scientists will receive their full salary. The labor costs after the research stage are significantly higher: a CEO, junior scientists, salespeople, and an IT specialist have been added.

The rental costs for the research stage is lower than the non-research stages. Space is minimized during prototype development but once it's been developed, the space required for sequencing and the molecular biology lab are expanded.

The inventory costs include only the SMRT chips and reagents. These costs depend on the design capacity: at 100% of design capacity, the total inventory cost will be \$690,000. At 50% design capacity, the chip costs will be half that. Since the research phase only involves prototype development, and not revenue generation, there are no inventory costs.

Finally, the operating costs include research and development, and sales costs. These costs only apply during the stage beyond the research stage. They have been estimated as a percentage of revenue.

The Gantt chart in **Figure VII.1** on the facing page outlines a potential two day operational period scenario in order to determine the number of technicians required for optimal sequencing throughput. **In consideration of the number of lab technicians required in order to manage the throughput of the system, it is important to consider the amount of time required for the preparation of the key parts solutions required for sequencing. The most important steps are the genome extraction and amplification, and the preparation of the waveguides for sequencing. Since the genome isolation requires a 16 hour incubation period for maximum extraction, it is unnecessary to wait for incubation during the work day. Therefore, it is best to prepare the DNA solution at the end of the work day, and then allow for incubation overnight. Upon returning to work the next morning, the amplified genomic DNA is ready for use. It is important to note that a**

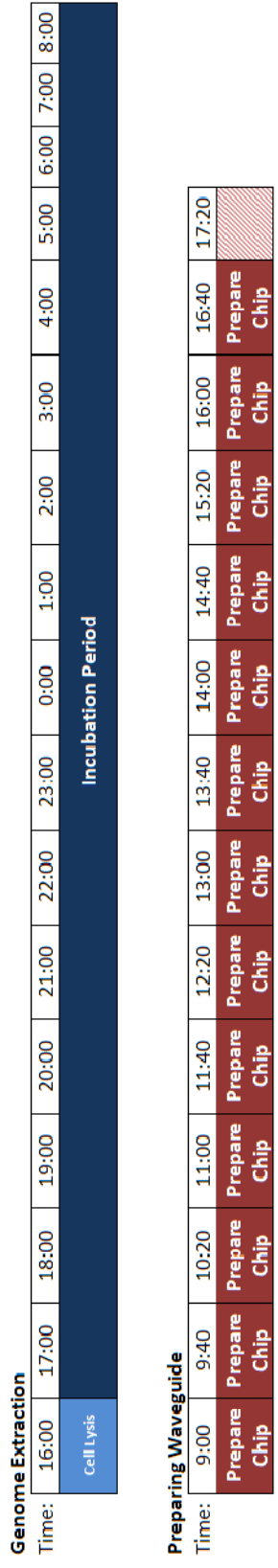


Figure VII.1 Gantt chart representation of technician work schedule.

large number of samples can all be prepared in one batch, so that this time scale assumes the preparation of approximately 2 days worth of DNA for sequencing for one day of DNA extraction.

The preparation of a single chip, including immobilization of the DNA polymerase and the loading of the DNA takes approximately 40 minutes. In order to safely meet throughput requirements, 13 chips must be prepared in one day. As Figure VII.1 demonstrates, it is possible for a single technician to prepare all 13 chips in a single day, not accounting for necessary breaks. Once the chips are prepared, they can be set in the sequencing apparatus and allowed to run overnight. Assuming a very efficient lab technician that does not eat and works 20 minutes of overtime, it is possible that in this scenario, one lab technician can handle the throughput required to meet the sequencing goals. However, in order to account for potential complications and problems, another lab technician should be available so that the work load can be divided, easily meeting the throughput required. Also, this scenario assumes that several chips cannot be prepared at once. Considering the small size and standardization of the chips, it is reasonable to assume that several, if not all, of the chips used in one day cannot be prepared at the same time, up to the point of the addition of the DNA, which is actually the final step in the preparation of the chips. Assuming this, two lab technicians should be sufficient to handle the throughput requirements.

Cost Estimates

Equipment Costs for Prototype Development				Rest of Equipment After Prototype Development			
Item	Unit Cost	Quantity	Total Cost	Item	Unit Cost	Quantity	Total Cost
EMCCD	\$ 32,500.00	4	\$ 130,000.00	EMCCD	\$ 32,500.00	20	\$ 650,000.00
Nanopositioner	72,000.00	2	144,000.00	Nanopositioner	72,000.00	10	720,000.00
Microscope	30,000.00	2	60,000.00	Microscope	30,000.00	10	300,000.00
Station Setup	10,000.00	2	20,000.00	Station Setup	10,000.00	10	100,000.00
Personal Comp	1,000.00	5	5,000.00	Personal Comp	1,000.00	7	7,000.00
Data Processor	4,000.00	2	8,000.00	Data Processor	4,000.00	13	52,000.00
Server	70,650.00	2	141,300.00	Server	70,650.00	12	847,800.00
Bio Lab Equip	200,000.00	1	200,000.00				
Total:			<u>\$ 708,300.00</u>				<u>\$ 2,676,800.00</u>

Research Phase (Series A)

Personnel	Salary	Quantity	Total Cost
CTO	\$ 84,000.00	1	\$ 84,000.00
Senior Scientist	84,000.00	3	252,000.00
Secretary	40,000.00	1	40,000.00
<i>*During research phase, personnel receive 70% of original salary, because they are also paid 10% of equity stake.</i>			
Total:			<u>\$ 292,000.00</u>

Sales Phase (Series B)

Personnel	Salary	Quantity	Total Cost
CEO	\$ 250,000.00	1	\$ 250,000.00
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
Salesperson	50,000.00	2	100,000.00
Secretary	40,000.00	1	40,000.00
IT person	40,000.00	1	40,000.00
Total:			<u>\$ 1,030,000.00</u>

Inventory Costs			
<i>Item</i>	<i>Unit Cost</i>	<i>Quantity</i>	<i>Total Cost</i>
SMRT Chip	\$ 200.00	3000	\$ 600,000.00
Reagents	30.00	3000	90,000.00
Total:			<u>\$ 690,000.00</u>
Rental Costs			
<i>Item</i>	<i>Cost per sqft</i>	<i>Sqft</i>	<i>Total Cost</i>
Molecular Bio Lab	\$ 126.00	400	\$ 50,400.00
Sequencing Space	126.00	200	25,200.00
Office Space	25.00	1000	25,000.00
Total:			<u>\$ -</u>
Operating Costs			
<i>Item</i>	<i>Cost per month</i>	<i>Month</i>	<i>Total Cost</i>
Utilities	5,000.00	12 mo	60,000.00
Maintenance	1,000.00	12 mo	12,000.00
Total			<u>\$ 172,600.00</u>
Operating Costs			
<i>Item</i>	<i>Sales</i>	<i>% of Sales</i>	<i>Total Cost</i>
Research	\$ 30,000,000.00	5%	\$ 1,800,000.00
Sales	30,000,000.00	3%	900,000.00
Total			<u>\$ 2,700,000.00</u>
Total Annual Costs			<u>\$ 4,718,600.00</u>

Table VII.3 Cost Estimates

The company will use a 5 year MACRS depreciation schedule because the accelerated tax schedule will provide the company with tax savings. Depreciation is a non-cash expense but it still affects the pre-tax income, from which taxes are deducted. If the pre-tax income decreases, taxes will also decrease. And tax, unlike depreciation, is a cash expense. An accelerated depreciation schedule for short lived projects like this will have a significant impact on the NPV and MIRR analysis. The depreciation percentages for MACRS in order are 20%, 32%, 19.2%, 11.52%, 11.52%, and 5.76%.

Depreciation Schedule					
<u>MACR Tax Schedule:</u>	<u>20.00%</u>	<u>32.00%</u>	<u>19.20%</u>	<u>11.52%</u>	<u>11.52%</u>
Year	<u>2010</u>	<u>2011</u>	<u>2012</u>	<u>2013</u>	<u>2014</u>
	\$				
Series A Equipment	708.3				
Depreciation		\$ (141.7)	\$ (226.7)	\$ (136.0)	\$ (81.6)
Series B Equipment		\$ 2,676.8			
Depreciation			\$ (535.4)	\$ (856.6)	\$ (513.9)
	\$	\$			
Beginning Net PPE	-	708.3	\$ 3,243.4	\$ 2,481.4	\$ 1,488.9
PPE Purchased/(Sold)	708.3	2,676.8	-	-	-
Less: Total Depreciation	-	<u>(141.7)</u>	<u>(762.0)</u>	<u>(992.6)</u>	<u>(595.5)</u>
	\$				
Ending Net PPE	708.3	\$ 3,243.4	\$ 2,481.4	\$ 1,488.9	\$ 893.3

Table VII.4 Depreciation Schedule using the 5 year MACRS schedule. (\$ in thousands).

The series A equipment is the research equipment purchase, and the series B equipment is the sales equipment purchase. The ending net PPE figures are balance sheet items and represent how much property and equipment the company owns. The total depreciation is what will appear on the income statement and will decrease the pre-tax income. In a later section, we will explore a what-if scenario of when the research stage takes two years instead of one. In that case, the series B equipment and its depreciations would simply be shifted to the right one year and the total depreciations would be changed as well.

VII.3 Income Statement

Income Statement					
Year	<u>2010</u>	<u>2011</u>	<u>2012</u>	<u>2013</u>	<u>2014</u>
Revenue	\$ -	\$ 15,000.0	\$ 30,000.0	\$ 30,000.0	\$ 30,000.0
Cost of Sales	<u>(172.6)</u>	<u>(643.6)</u>	<u>(988.6)</u>	<u>(988.6)</u>	<u>(988.6)</u>
Gross Profit	(172.6)	14,356.4	29,011.4	29,011.4	29,011.4
Operating, SG&A Expenses	(292.0)	(2,380.0)	(3,730.0)	(3,730.0)	(3,730.0)
Depreciation	-	(141.7)	(762.0)	(992.6)	(595.5)
Pre-Tax Income	(464.6)	11,834.7	24,519.4	24,288.8	24,685.9
Tax @ 40%	<u>185.8</u>	<u>(4,733.9)</u>	<u>(9,807.8)</u>	<u>(9,715.5)</u>	<u>(9,874.3)</u>
Net Income	\$ (278.8)	\$ 7,100.8	\$ 14,711.6	\$ 14,573.3	\$ 14,811.5
Design Capacity	0%	50%	100%	100%	100%
Margins					
Gross Margin	0.0%	95.7%	96.7%	96.7%	96.7%
Profit Margin	0.0%	47.3%	49.0%	48.6%	49.4%

Table VII.5 The income Statement showing the gross and profit margins. (\$ in thousands).

The total costs are divided into cost of sales and operating, SG&A expenses. The cost of sales (also known as costs of goods sold, COGS) includes costs that are directly involved in the making of the goods. It is a sum of fixed costs and variable costs. The fixed costs include rent and overhead because rent and overhead costs do not vary with the number of goods sold. The variable costs are the inventory costs because the inventory cost is a function of how many genomes are sequenced. Subtracting the cost of sales from the revenue is the gross profit. The gross margin is a percentage showing how much money is left from the revenue after paying for the cost of sales. The operating and SG&A (Selling, general, and administrative) expenses are costs that are associated with managing the business. These costs are mainly salary, but also include the research and development cost and the sales cost. Subtracting these new expenses and the depreciation gives the pre-tax income.

Federal taxes are set around 35% but when state tax is added, the tax can be rounded up to 40%. In the first year, because the company has actually lost money, the tax figure is positive and the company actually receives money from the government. This is called negative taxes, or tax shield. Sometimes this does not apply, but we will assume that this holds with our company.

VII.4 Working Capital

The income statement led us to net income. But net income is not equivalent to cash. In order to get to cash figures, and subsequent NPV and IRR analysis, we need to adjust net income for cash items. There are multiple cash items to adjust for. Change in working capital is one of them.

Working capital is how much capital a company needs to operate normally. If a company makes \$1000, it can't give all of it to the owner: a portion of it has to be allocated for the company to run its day-to-day operations. This allocated cash is the working capital and it can be described as current assets minus current liabilities. Current assets are things the company has that can be converted to cash quickly. Current liabilities are bills and debts the company has to pay quickly. Working capital is current assets minus current liabilities because it is the money left over after having paid all its imminent bills.

There are four main working capital items we will work with: accounts receivables, inventory, accounts payable, and cash reserve. Accounts receivables are earnings that haven't received cash payment yet. For example, after making a sale, the company records its earnings even when it hasn't received cash payment yet. The client usually has about 30~60 days to pay. We will choose 30 days. Because all account receivables convert into revenue:

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

Inventory is what the company needs to produce the goods it sells. In our case, they are the SMRT chips and reagents. Our company will buy new inventory every month.

$$Inventory = \frac{Inventory\ Cost(\$)}{Year} * \frac{1Year}{365Days} * 30\ Days \quad \text{Equation VII.2}$$

Accounts payable is the opposite of accounts receivable: they are bills the company records (and subtracts from revenue) but haven't handed in the cash yet. We will pay bills every 30 days. Normally, in the following equation, cost of sales is used instead of rent and operating cost, but because we have more detailed information, we can modify the equation. Cost of sales usually has rent, operating costs, and inventory costs all buried inside so many people use cost of sales as a proxy. Because all account payables turn into rent and operating costs:

$$Accounts\ Payable = \frac{Rent + Operating\ Costs(\$)}{Year} * \frac{1Year}{365Days} * 30\ Days \quad \text{Equation VII.3}$$

Cash reserve is cash on hand needed to pay for future salary. We will reserve 3 months worth of salary.

$$Cash\ Reserve = \frac{Salary(\$)}{Year} * \frac{1Year}{12Months} * 3\ Months \quad \text{Equation VII.4}$$

Change in working capital items change net income into cash.

Working Capital					
<i>Year</i>	<u>2010</u>	<u>2011</u>	<u>2012</u>	<u>2013</u>	<u>2014</u>
Working Capital Item Estimates					
	\$				
Accounts Receivable	-	\$ 1,232.9	\$ 2,465.8	\$ 2,465.8	\$ 2,465.8
Inventory	-	28.4	56.7	56.7	56.7
Accounts Payable	24.5	135.5	246.5	246.5	246.5
Cash Reserve	73.0	257.5	257.5	257.5	257.5
Changes in Working Capital					
	\$				
(Increase)/Decrease in A/R	-	\$ (1,232.9)	\$ (1,232.9)	\$ -	\$ -
(Increase)/Decrease in Inv	-	(28.4)	(28.4)	-	-
Increase/(Decrease) in A/P	24.5	111.0	111.0	-	-
(Increase)/Decrease in C/R	<u>(73.0)</u>	<u>(184.5)</u>	<u>-</u>	-	-
Total Change in Working Capital	\$ (48.5)	\$ (1,334.8)	\$ (1,150.3)	\$ -	\$ -

Table VII.6 Working Capital and Change in Working Capital. (\$ in thousands).

Any increase in assets will decrease cash. It makes sense because the company needs to spend money to buy things. Any decrease in assets will increase cash because selling equipment or any other asset results in cash. An increase in liability also increases cash. If a company borrowed money from a bank, it has more money to spend, before having to spend it back. A decrease in liability results in decrease of cash because the company has to spend cash to pay back debt.

An increase in account receivable since last fiscal year should decrease cash because net income has taken this account but the company hasn't actually received the money yet. By next fiscal year, however, the company will have been paid, which is reflected by a decrease in accounts receivables (assuming no additional A/R accrues). An increase in inventory decreases cash because the company needs to spend cash to buy inventory. An increase in accounts payable increases cash because net income is revenue minus accounts payable, but the company hasn't paid the bill yet. An increase in cash reserve decreases cash because cash has to be held back to pay future salary and that cash can't be used to pay the owners or else the company will cease to operate normally.

Two other cash items need mentioning. Purchasing PPE (plant, property, and equipment) decreases cash immediately. However, a PPE purchase isn't shown on the income statement. Instead, it is slowly amortized. This is because the income statement reflects the company's operational efficiency, which doesn't necessarily involve one-time cash expenses. Selling of equipment is the same: it immediately generates cash, but it isn't part of revenue because revenue only reflects the company's natural operations. Selling unwanted equipment isn't part of operations: it's a one-time thing.

Issuing common stock is the same. Issuing common stock to investors (series A and series B investors) immediately generates cash but isn't part of revenue. Repurchasing existing shares from the public market would decrease cash but isn't reflected on the income statement.

All cash items are found in the cash flow statement.

VII.5 Free Cash Flow, Terminal Value

After getting change in working capital and other cash items from the cash flow statement, we can convert net income into free cash flow.

Free Cash Flow					
Year	<u>2010</u>	<u>2011</u>	<u>2012</u>	<u>2013</u>	<u>2014</u>
Net Income	\$ (278.8)	\$ 7,100.8	\$ 14,711.6	\$ 14,573.3	\$ 14,811.5
<u>Cash Flow Statement</u>					
Cash From Operating Activities					
Plus: Depreciation	-	\$ 141.7	\$ 762.0	\$ 992.6	\$ 595.5
Changes in Working Capital					
(Increase)/Decrease in A/R	-	(1,232.9)	(1,232.9)	-	-
(Increase)/Decrease in Inventory	-	(28.4)	(28.4)	-	-
Increase/(Decrease) in A/P	24.5	111.0	111.0	-	-
(Increase)/Decrease in C/R	<u>(73.0)</u>	<u>(184.5)</u>	-	-	-
Total Change in Working Capital	(48.5)	(1,334.8)	(1,150.3)	-	-
Cash From Investing Activities					
(Purchase)/Selling of Equipment	(708.3)	(2,676.8)	-	-	-
Cash From Financing Activities					
Issuance of Common Stock	<u>1,200.0</u>	<u>3,500.0</u>	-	-	-
Free Cash Flow	164.5	\$ 6,730.9	\$ 14,323.4	15,565.9	\$ 15,407.1

Table VII.7 Free Cash Flow from Net Income. (\$ in thousands).

The free cash flows are used for NPV and IRR analysis because they represent real cash received by the owners.

Before conducting NPV and IRR analysis, however, we need to determine the terminal value of the company. We accomplish this by using the perpetuity growth model:

$$Terminal\ Value = Cash\ Flow * \frac{(1 + g)}{r - g} \qquad \text{Equation VII.5}$$

The cash flow is the last free cash flow. The parameter g is the growth rate of the cash flow and the company. The parameter r is the discount rate. This terminal value can be described as the present value of all the continuing future cash flows. This concept is highly theoretical because it assumes that future cash flows are predictable. Note: when g = 0, the equation simplifies into CF / r, which is the basic perpetuity model.

Terminal Value	
Last Free Cash Flow	\$ 15,407.1
Discount Rate	25%
Growth Rate	Terminal Value
	\$
15%	154,070.6
	\$
0%	61,628.2
	\$
-15%	38,517.6

Table VII.8 Terminal Value Examples

VII.6 NPV Valuation

The net present value is a mathematical model used to describe how much richer one would become today if he were to undertake the investment. It is the sum of all the present values of every cash flow, which in this case, is the free cash flow and the terminal value. It is a theoretical model that is heavily dependent on the discount rate.

NPV Calculation @ 25%							
Year	<u>2009</u>	<u>2010</u>	<u>2011</u>	<u>2012</u>	<u>2013</u>	<u>2014</u>	<u>Terminal Value</u>
T	<u>0</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>
Free Cash Flow		\$ 164.5	\$ 6,730.9	\$ 14,323.4	\$ 15,565.9	\$ 15,407.1	\$ 61,628.2
Discount Rate		<u>50%</u>	<u>25%</u>	<u>25%</u>	<u>25%</u>	<u>25%</u>	<u>25%</u>
Present Value		\$ 109.7	\$ 4,307.8	\$ 7,333.6	\$ 6,375.8	\$ 5,048.6	\$ 16,155.5
Investments	\$ (1,200.0)	\$ (3,500.0)					
Discount Rate	<u>0%</u>	<u>25%</u>					
Present Value	\$ (1,200.0)	\$ (2,800.0)					Sum of All Present Values (NPV)
							<u>\$ 35,330.9</u>

Table VII.9 Net Present Value Calculation Using Discount Rate of 25%. (\$ in thousands).

The discount rate used depends on the industry. The riskier the industry, the higher the discount rate is. Because the biotechnology industry is risky and unpredictable, the discount rate used will be 25% and 30%. Both will be used to create a range of multiple NPVs. The discount rate during the research stage, however, will be 50% because developing new technology is much riskier. Table 8 summarizes the calculations using a discount rate of 25%.

In order to calculate the net present values, all the projected free cash flows and terminal value are discounted back into present values. Those present values are then summed. Next, we have to subtract out the present values of the cost of the project – the initial investments. Our calculation becomes trickier because we have two rounds of investments at two different times from two different investors. Some qualitative reasoning is required. The present value of the first investment, the series A investment of \$1.2MM, is equal to itself because it is not a future cash flow: it is a cash outflow that occurs in the present. The second investment, the series B investment of \$3.5MM, occurs at the end of 2010, when a working prototype has been developed. Unlike the first cash inflow that is discounted at a rate of 50%, the series B investment is actually discounted at a rate of 25%. This is because the purpose of the 50% discount rate was to take into account the extraordinary risk of developing new technology. Because series B is funded only when that extra risk is taken away, it is discounted at 25%, not 50%.

Also note that the terminal value of the company is also discounted to the present. This is because the terminal value is a future projected value. We have assumed that the growth rate will be 0%. The sum of all the present values is the net present value: it is called net present value because it adds all the positive future cash flows, and the present negative investments to get a net sum. This number tells us that undertaking the *PennBio* project will make the investors \$35MM richer today. Of course, keep in mind that this is only a model and it depends heavily on several assumptions. It is good practice to do multiple analyses under different scenarios and assumptions.

VII.7 Equity Shares

There is another profitability measure called the IRR, but because there are two different investors, each investing at two different times, we need to create an internal bookkeeping of how much equity each group of investors own. To clarify, how much of the free cash flows do the series A investors get to keep, and how much do the series B investors keep?

To get an initial picture, we compare how much the two investors have put into the project. The series A investors put in \$1.2MM at the beginning and the series B investors put in \$3.5MM one year later. Because of the time difference, the series A investment needs to be

Percentage of Investments			
	<u>Investment</u>	<u>FV Investment</u>	<u>Percentage</u>
Series A Investors	\$ 1,200.0	\$ 1,800.0	34%
Series B Investors	3,500.0	<u>3,500.0</u>	<u>66%</u>
Total		\$ 5,300.0	100%

Table VII.10 Comparing Percentages of Investments. (\$ in thousands).

discounted forward by 50% to compare the two numbers correctly.

From this result, it can be argued that the series B investors should keep 66% of the company once they come in. On the other hand, because the series A investors have put in more sweat equity and feel that they undertook more risk that the discount rate doesn't quite cover, both could agree that the series A investors will receive more equity. However, for the sake of simplicity and because the issue of sweat equity is a very subjective one, we will avoid that here. The series B investors will receive 66% of equity, the series A investors will keep 90% of the remaining 34%, and the scientists will keep 10% of the remaining 34%. In the case where the

research stage takes two years, the series A investment will be discounted forward twice by 50%: series B investors would receive 54%. **Table VII.11** summarizes company ownership.

Equity Percentage						
<i>Year</i>	<u>2010</u>	<u>2011</u>	<u>2012</u>	<u>2013</u>	<u>2014</u>	<u>2015</u>
Scientists Series A Investors Series B Investors	10%	3%	3%	3%	3%	3%
Investors Series B Investors	90%	31%	31%	31%	31%	31%
Investors	<u>0%</u>	<u>66%</u>	<u>66%</u>	<u>66%</u>	<u>66%</u>	<u>66%</u>
Total	100%	100%	100%	100%	100%	100%
NPV @ 25%	\$ (1,090.3)	\$ 417.5	\$ 7,751.0	\$ 14,126.8	\$ 19,175.4	\$ 35,330.9

Shares Values vs. Time

Scientists Series A Investors Series B Investors		\$				
Scientists Series A Investors Series B Investors	\$ -	14.2	\$ 263.2	\$ 479.8	\$ 651.2	\$ 1,199.9
Investors Series B Investors	-	127.6	2,369.2	4,318.0	5,861.2	10,799.2
Investors	<u>-</u>	<u>275.7</u>	<u>5,118.6</u>	<u>9,329.0</u>	<u>12,663.0</u>	<u>23,331.7</u>
Total	\$ -	\$ 417.5	\$ 7,751.0	\$ 14,126.8	\$ 19,175.4	\$ 35,330.9

Table VII.11 Equity Percentage and Share Values of Owners. (\$ in thousands).

By calculating the NPV values at each year, we can also calculate each owner's share value as a function of time.

Defining equity percentage is an important bookkeeping task because percentage of ownership determines how much of the free cash flows generated each group will get to keep. Using that information, we can conduct a rate of return analysis.

VII.8 MIRR Analysis

Conventional rate of return analysis uses the IRR (internal rate of return) figure. However, there are many flaws in that model. Consequently, using IRR for our analysis would yield highly overstated and inaccurate results, especially in projects with large positive cash flows. In using the IRR, the calculation assumes that the free cash flows will be reinvested at the rate being calculated. This is a critical mistake because in the *PennBio* investment, there are only two rounds of investments and none of the free cash flows are being reinvested. In fact, McKinsey consultants advise avoiding using the IRR.

The alternative rate of return metric will be the MIRR (Modified internal rate of return). In this metric, we must specify the finance rate and the reinvest rate. The finance rate is the APR (Annual percentage rate) the company would have to pay to debt lenders if there are any negative cash flows. The reinvest rate is the rate the owners would receive on the positive cash flows. For the finance rate, we will assume a standard 4.4% on a bank term loan. For the reinvest rate, we will assume a 4.9% return. It is the current yield on a 3 month and 6 month US Treasury bill, which can be seen as the risk free rate. Alternatively, we could use a higher reinvest rate since the investors can be assumed to experienced enough to earn more than the risk free rate, especially if they are investing in the risky biotech industry. However, we chose the risk free rate to be conservative in our profitability analysis.

The MIRR calculations are simple: First, the investment is defined. Then all the free cash flows the investor would receive are divided by the equity percentage. An MIRR can be determined from those figures. The following table summarizes the calculations.

MIRR Calculations

Year	<u>2010</u>	<u>2011</u>	<u>2012</u>	<u>2013</u>	<u>2014</u>	<u>2015</u>	
Free Cash Flows	\$ 164.5	\$ 6,730.9	\$ 14,323.4	\$ 15,565.9	\$ 15,407.1	\$ 61,628.2	
Equity Percentage							
Series A	90%	31%	31%	31%	31%	31%	
Series B	0%	66%	66%	66%	66%	66%	
Cash Flows	<u>Investment</u>	<u>Divided Free Cash Flows</u>					
Series A	\$ (1,200.0)	\$ 148.0	\$ 2,665.4	\$ 5,672.1	\$ 6,164.1	\$ 6,101.2	\$ 24,404.8
Series B	\$ (3,500.0)	\$ 3,769.3	\$ 8,021.1	\$ 8,716.9	\$ 8,628.0	\$ 34,511.8	
Series A MIRR	<u>77%</u>						
Series B MIRR	<u>87%</u>						

Table VII.12 MIRR Calculation. Finance rate at Term loan APR of 4.4%, reinvest rate at 6-month T-bill of 4.9%. (\$ in thousands).

Note that an IRR calculation would yield around 150% returns, a grossly overstated figure. Also, the series B cash flow has one less term. This is because the series B investors came in one year after the series A investors did.

To summarize the first part of our financial analysis, we created a pro forma income statement, adjusted to get free cash flows, and then used those to arrive at an NPV and MIRR analysis. The following page summarizes everything we have done so far. There are two terminal values: the one on the top uses a discount rate of 30% and the one on the bottom uses 25%.

Year	2010	2011	2012	2013	2014	2015
<u>Income Statement</u>						
Revenue	\$ -	\$ 15,000.0	\$ 30,000.0	\$ 30,000.0	\$ 30,000.0	
Cost of Sales	(172.6)	(643.6)	(988.6)	(988.6)	(988.6)	
Operating, SG&A Expenses	(292.0)	(2,380.0)	(3,730.0)	(3,730.0)	(3,730.0)	
Depreciation	-	(141.7)	(762.0)	(992.6)	(595.5)	
Pre-Tax Income	(464.6)	11,834.7	24,519.4	24,288.8	24,685.9	
Tax @ 40%	185.8	(4,733.9)	(9,807.8)	(9,715.5)	(9,874.3)	
Net Income	\$ (278.8)	\$ 7,100.8	\$ 14,711.6	\$ 14,573.3	\$ 14,811.5	
<u>Cash Flow Statement</u>						
Cash From Operating Activities	\$ -	\$ 141.7	\$ 762.0	\$ 992.6	\$ 595.5	
Plus: Depreciation						
Changes in Working Capital						
(Increase)/Decrease in A/R	\$ -	\$ (1,232.9)	\$ (1,232.9)	\$ -	\$ -	
(Increase)/Decrease in Inventory	-	(28.4)	(28.4)	-	-	
Increase/(Decrease) in A/P	24.5	111.0	111.0	-	-	
(Increase)/Decrease in C/R	(73.0)	(184.5)	-	-	-	
Total Change in Working Capital	\$ (48.5)	\$ (1,334.8)	\$ (1,150.3)	\$ -	\$ -	
Cash From Investing Activities	(708.3)	(2,676.8)	-	-	-	
(Purchase)/Selling of Equipment						
Cash From Financing Activities	1,200.0	3,500.0	-	-	-	
Issuance of Common Stock	\$ 164.5	\$ 6,730.9	\$ 14,323.4	\$ 15,565.9	\$ 15,407.1	\$ 51,356.9 (Terminal Value)
Free Cash Flow	0%	50%	100%	100%	100%	100%
% of Design Capacity						
<u>Investment Divided Free Cash Flows</u>						
Series A @ 31% of Equity	\$ (1,200.0)	\$ 148.0	\$ 2,057.4	\$ 4,378.1	\$ 4,757.9	\$ 4,709.3
Series B @ 66% of Equity	(3,500.0)	4,445.0	9,458.8	10,279.3	10,174.5	40,697.9
NPV @ 30%	\$ 26,959		Series A MIRR			77%
NPV @ 25%	\$ 35,331		Series B MIRR			87%

VII.9 What-If Scenarios

So far, the previous financial analysis holds under certain assumptions. But what if those assumptions aren't true? What if the project undergoes certain deviations? Then what will happen to our financial analysis?

To observe, how the project's profitability would change, different case scenarios were defined for the research stage, scale up and sales stage, and the terminal stage. The following table summarizes the different case scenarios and their results.

What-If Scenarios	
Research Stage	
Best Case	S1) The start up stage takes one year, as planned.
Worst Case	S2) The start up stage unexpectedly takes two years.
Scale Up and Sales Stage	
Best Case	D1) The first year of sales has 50% of design capacity, as planned. The remaining three years run at full 100% capacity.
Worst Case	D2) The first year of sales has 30% capacity. The next year's capacity is 70%. The remaining two years of sales have full 100% capacity.
Terminal Stage	
Best Case	T1) The company stays profitable. The revenue stays constant perpetually.
Middle Case	T2) The company starts to decline due to rising competitors. Earnings decrease 15% each year.
Worst Case	T3) Due to an improved rival technology, customers stop showing up immediately. Equipments are sold at 50% of face value, a total of \$1.69MM.

Table VII.13 What If Scenarios

Scenario Summary

Tree		Case	NPV @ 30%	NPV @ 25%	Series A MIRR	Series B MIRR	
S1	D1	T1	1	\$ 26,959	\$ 35,331	77%	87%
		T2	2	22,349	27,758	69%	76%
		T3	3	16,670	19,619	58%	62%
	D2	T1	4	23,025	30,964	75%	84%
		T2	5	18,414	23,392	66%	73%
		T3	6	12,736	15,253	53%	57%
S2	D1	T1	7	26,733	35,103	58%	81%
		T2	8	22,122	27,530	58%	71%
		T3	9	16,444	19,391	58%	57%
	D2	T1	10	22,799	30,737	54%	78%
		T2	11	18,188	23,164	54%	67%
		T3	12	12,510	15,025	54%	52%

NPV and MIRR Range

	High <i>(Average of top 3)</i>	Medium <i>(Median)</i>	Low <i>(Average of bottom 3)</i>
NPV	<u>\$ 33,799.4</u>	<u>\$ 22,573.7</u>	<u>\$ 13,423.3</u>
Series A	<u>73%</u>	<u>58%</u>	<u>54%</u>
Series B	<u>84%</u>	<u>72%</u>	<u>55%</u>

Table VII.14 Complete NPV and MIRR Summary of 12 What-If Scenarios. (\$ in thousands).

Even under these various scenarios, the *PennBio* project looks very comfortable. A copy of all 12 pro forma is provided in **Appendix F**. One crucial assumption to note here is that all these scenarios have the genome price selling at \$10,000. As mentioned in the introduction, the revenue generation is the most important assumption because all our financial analysis depends on it.

VII.10 Price Sensitivity Analysis

We will repeat the same analysis under various genome prices to see under which prices the project would lose money. The following table summarizes the results. The genome prices range from \$10,000 to \$500.

Genome Price Sensitivity Analysis									
Genome Price	NPV			MIRR					
	High	Medium	Low	High A	Medium	Low	High B	Medium	Low
\$ 10,000	\$ 33,799	\$ 22,574	\$ 13,423	73%	58%	54%	84%	72%	55%
9,000	29,806	19,780	11,628	71%	56%	51%	80%	69%	52%
8,000	25,797	16,969	9,815	67%	53%	48%	76%	64%	49%
7,000	21,788	14,157	8,002	63%	49%	45%	71%	60%	45%
6,000	17,794	11,364	6,206	59%	45%	41%	65%	55%	40%
5,000	13,785	8,553	4,393	54%	40%	36%	59%	49%	35%
4,000	9,776	5,737	2,580	48%	35%	31%	51%	41%	29%
3,000	5,783	2,874	784	40%	27%	23%	41%	32%	21%
2,000	1,773	83	(1,029)	28%	17%	13%	27%	18%	10%
1,000	(2,211)	(2,609)	(2,971)	5%	-2%	-8%	-2%	-7%	-10%
500	(3,530.7)	(4,045.7)	(4,402.1)	-16%	-30%	-31%	-22%	-88%	-100%
Breakeven Price	\$ 1,600.0	\$ 2,000.0	\$ 2,600.0						

Table VII.15 Genome Price Sensitivity Analysis for NPV and MIRR. (\$ in thousands).

As mentioned before, the price of the genome is difficult to nail down. But we see here that charging at least \$3000 per genome can still make returns. This is a very comfortable price margin, especially considering that one of the existing competitors, Knome®, is charging \$100,000.

VII.11 Growth Case

We look at one final scenario. When the company's margins are healthy, it would be wise to expand its operations. This would only happen when the research stage, scale up stage, and the development stage occur as planned – that is, a case 1 scenario. This would be the best case scenario. In our financial model, we will make the design capacity grow exponentially at 50% during the sales stage: the company will operate at a capacity of 100%, 150%, and 225%. Here, the average of the 25% and 30% NPV is \$54MM, the series A MIRR is 95%, and the series B MIRR is 110%. At the end of **Appendix F**, is the pro forma for this case.

VII.12 Conclusions

The entire financial analysis has provided ample evidence to show that *PennBio* would be a profitable investment. Next, the what-if scenarios have shown that the investors would receive high returns, even under worst case scenarios. Finally, the sensitivity analysis has revealed that an ample price margin exists before the investors lose their capital.

Of course, financial models are insightful, but they are still models. They help to guide the investors and to reduce as much risk as possible. Models do not determine the future. And

because no one can predict the future, there will always be a degree of financial risk. But our financial analysis has covered as much as possible. One reason why the venture looks so profitable is chiefly due to the nature of the market: It is a biotech firm in an early, immature market. Not many competitors exist with our technology. And this is typically what one would expect in such settings. Until the market matures, profit margins will be remarkably high.

VIII. Conclusions

The primary objective of this venture was to characterize and develop a novel whole-genome sequencing technique with a particular focus on accuracy and extraordinary throughput. This was achieved through the application of cutting-edge single-molecule, real-time detection technology. Fundamental to this approach are several breakthroughs which pushed production to new heights while maintaining sequence fidelity. Zero-mode waveguides, which simultaneously provide exceptionally high signal-to-noise ratios and the geometric consistency that permits one-to-one waveguide to pixel mapping, allow for unprecedented clarity in single-molecule observation. Novel phospho-linked nucleotides, which emit signal pulses while being incorporated into sequenced DNA then go dark as their fluorescent moiety is cleaved from the unadulterated growing complementary strand, facilitate unparalleled sequence-reads and

Conclusions

polymerization rates. Two precision-aligned EMCCD cameras collect this high-quality signal and relay it to a powerful server bank, where it is converted into a complete, personal genome sequence using a unique two-color ratio approach, and custom-optimized sequence alignment algorithms.

At full capacity, *PennBio* can supply 1 genome per SMRT chip station per day. With a mere 13 stations, production is expected to reach 3,000 full genomes per year. This level of throughput well exceeds the industry *status quo*, and in conjunction with the minimum of capital investment per unit output, translates into affordable and readily available genomic data for our customers. And at a retail price of \$4,000 per genome, we stand to gross up to \$12.0 million per year in revenues. We offer our investors a unique opportunity to profit financially from the relentless advances made every day in understanding and practical use of the human genome.

Moreover, this financial situation appears to be optimal in the current technological environment. Using 1-to-1 polymerase to pixel ratios, and relatively simple optics, we have minimized our fixed costs, while the utilization of high-performance enzymes and streamlined data analysis maximizes revenue. We therefore do not believe that another whole-genome sequencing firm could enter the market and be more profitable than *PennBio* – not without substantial advances in optical detection, computing, and biotechnology.

VIII.1 Acknowledgements

We are especially indebted to Dr. John C. Crocker for his inspiration and much-appreciated criticism throughout the course of this project. His steady guidance, from the initial development of a design concept to the final analyses and write-up, was fundamental to our success. We appreciate the tour of the EMCCD camera setup in his lab, which was very important in visualizing our physical product concept.

We would also like to thank Professor Leonard A. Fabiano and Professor Warren D. Seider for their contributions and assistance each week, without which the completion of this project would not have been possible. And of course, we wish to extend our gratitude to the consultants for their insight during design team meetings.

Conclusions

Appendix A:

Reagent Specifications

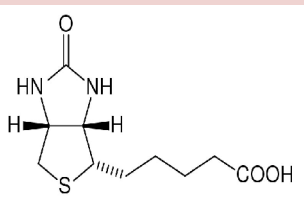
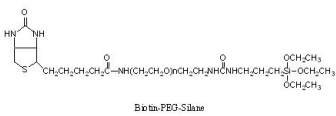
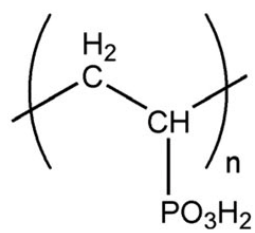
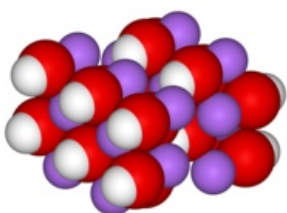
A.1 Sequencing-by-Synthesis Reagents

A.2 Deoxyribonucleotide Fluorophosphate Reagents

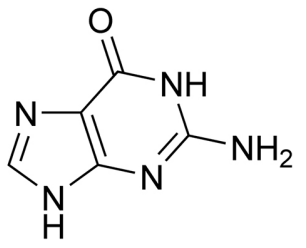
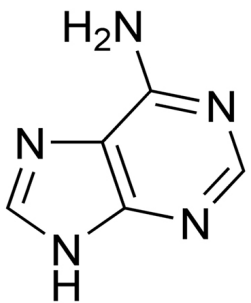
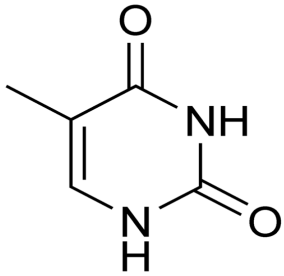
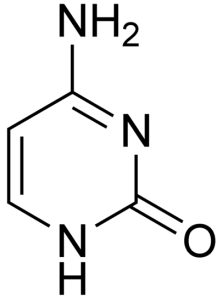
A.3 Deoxyribonucleotide Fluorophosphates

A.4 Proteins


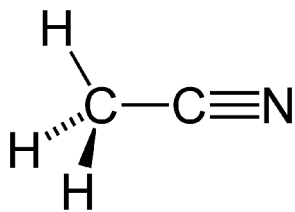
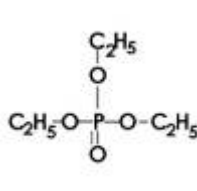

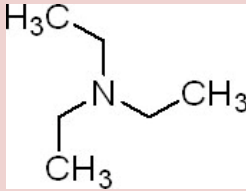
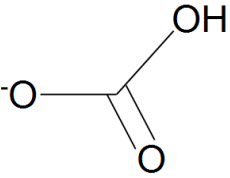
Sequencing-by-Synthesis Reagents

Species	Structure	Formula	Manufacturer	Price (\$)
Biotin	 The structure shows the biotin ring system, which is a bicyclic urea derivative fused to a five-membered thiophene ring. A long aliphatic chain is attached to the thiophene ring, ending in a carboxylic acid group (-COOH).	$C_{10}H_{16}N_2O_3S$	(Various)	--
Biotin-polyethyleneglycol-trimethylsilane	 The structure shows a biotin molecule linked via a carbonyl group to a polyethylene glycol (PEG) chain. The PEG chain is terminated with a trimethylsilyl group (-Si(CH3)3). The label 'Biotin-PEG-Silane' is present below the structure.	--	Laysan Bio	480 per 500 mg
Polyvinylphosphonic acid	 The structure shows a repeating unit of a polymer chain. The backbone consists of carbon atoms, with two hydrogen atoms attached to one carbon and a hydrogen atom and a phosphonic acid group (-CH2PO3H2) attached to the adjacent carbon. The unit is enclosed in large parentheses with a subscript 'n'.	$(C_2H_5PO_3)_n$	Sigma-Aldrich	215 per gram
Random Hexamers	See individual Nucleotides on page B.2	5'-NNNNNN-3'	Fidelity Systems	22 per 4 µg
Reaction Buffer	(Various)	50mM ACES: $C_4H_{10}N_2O_4S$ 75mM KC_2O_2 5mM dithiothreitol: $C_4H_{10}O_2S_2$	(Various)	--
Sodium Hydroxide	 A ball-and-stick model of a sodium hydroxide (NaOH) molecule. It consists of one red oxygen atom, one white hydrogen atom, and one purple sodium atom.	NaOH	The Science Company	11.5 per 500 g

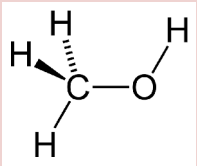
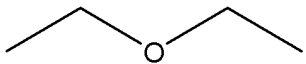
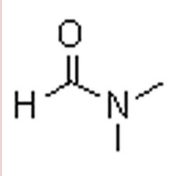
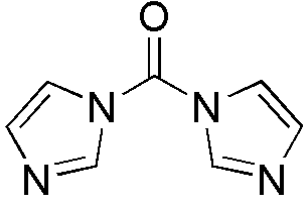
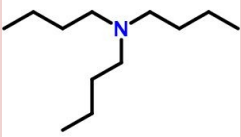
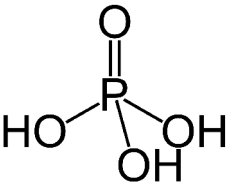
Deoxyribonucleotide Fluoropentaphosphate Reagents

Species	Structure	Formula	Manufacturer	Price (\$)
Guanine		$C_5H_5N_5O$	Cole-Parmer	25.30 per 25 g
Adenine		$C_5H_5N_5$	Cole-Parmer	96.30 per 25 g
Thymine		$C_5H_6N_2O_2$	Cole-Parmer	47.10 per 25 g
Cytosine		$C_4H_5N_3O$	Cole-Parmer	56.60 per 5 g

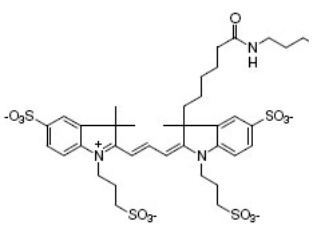
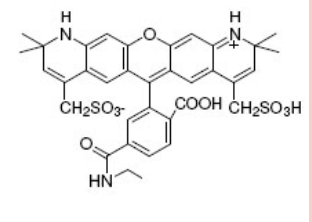
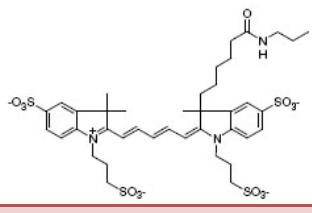
Appendix A: Reagent Specifications

Species	Structure	Formula	Manufacturer	Price (\$)
Fmoc-6-aminohexylphosphate		$C_{21}H_{25}NO_3$	AnaSpec	120 per gram
Anhydrous acetonitrile		C_2H_3N	Sigma-Aldrich	761 per 18 liters
Anhydrous triethylphosphate		$C_6H_{15}O_4P$	Sigma-Aldrich	156.50 per 4 liters
Phosphorus oxychloride		$POCl_3$	Sigma-Aldrich	105 per 100 mL
Triethylamine		$C_6H_{15}N$	Sigma-Aldrich	102 per 2 liters
Bicarbonate		HCO_3^-	Sigma-Aldrich	47.30 per 500 grams (NH ₄ ⁺ salt)

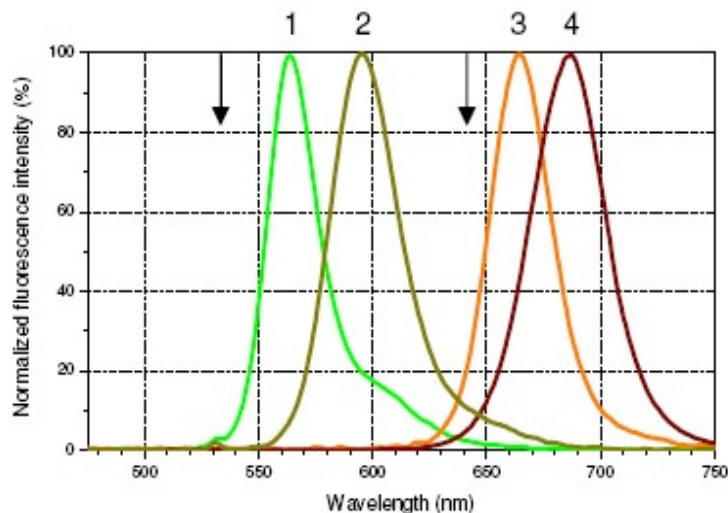
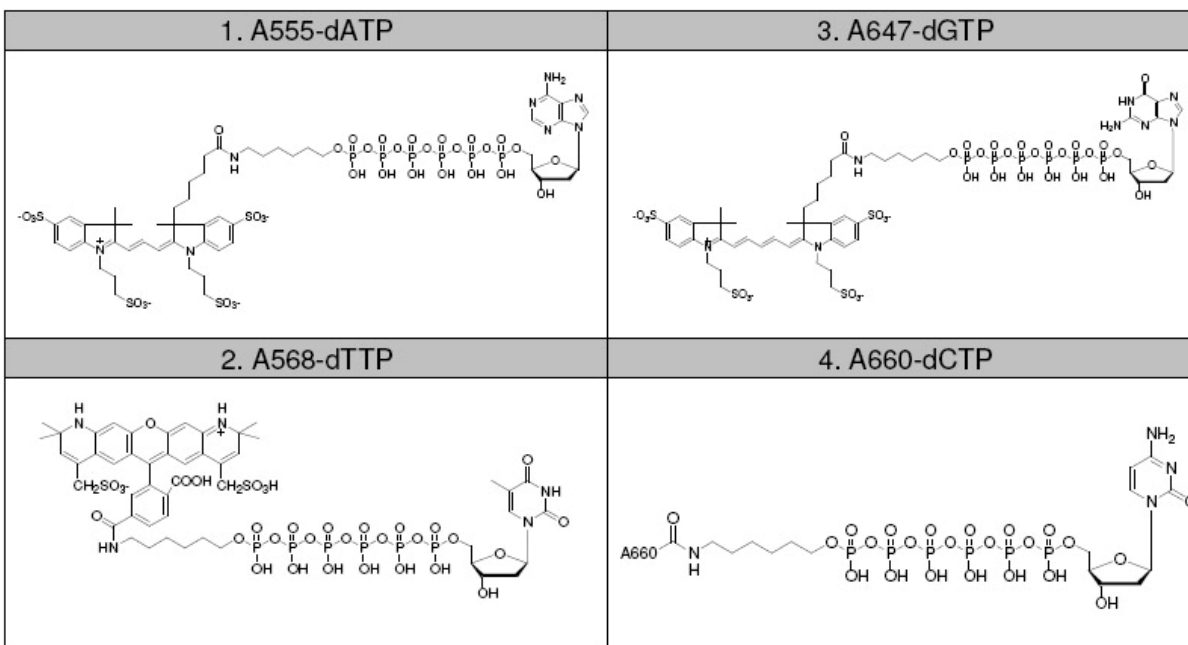
Appendix A: Reagent Specifications

Species	Structure	Formula	Manufacturer	Price (\$)
Methanol		CH ₃ OH	Sigma-Aldrich	40.90 per liter
Diethyl Ether		C ₄ H ₁₀ O	Sigma-Aldrich	199.50 per 4 liters
Dimethylformamide		C ₃ H ₇ NO	Sigma-Aldrich	24 per 250 mL
1,1'-Carbonyldiimidazole		C ₇ H ₆ N ₄ O	Sigma-Aldrich	36.70 per 10 g
tributylamine		[CH ₃ (CH ₂) ₃] ₃ N	Cole-Parmer	35.90
Phosphoric Acid		H ₃ PO ₄	Cole-Parmer	47.74 per 500mL (85% sol'n)

Appendix A: Reagent Specifications

Species	Structure	Formula	Manufacturer	Price (\$)
Magnesium Chloride	Cl – Mg – Cl	MgCl ₂	Gallade Chemical	104.96 per 5 g
Alexa Fluor A555			Invitrogen	240 per 1 mg
Alexa Fluor A568			Invitrogen	240 per 1 mg
Alexa Fluor A647			Invitrogen	240 per 1 mg
Alexa Fluor A660	Not Available		Invitrogen	240 per 1 mg

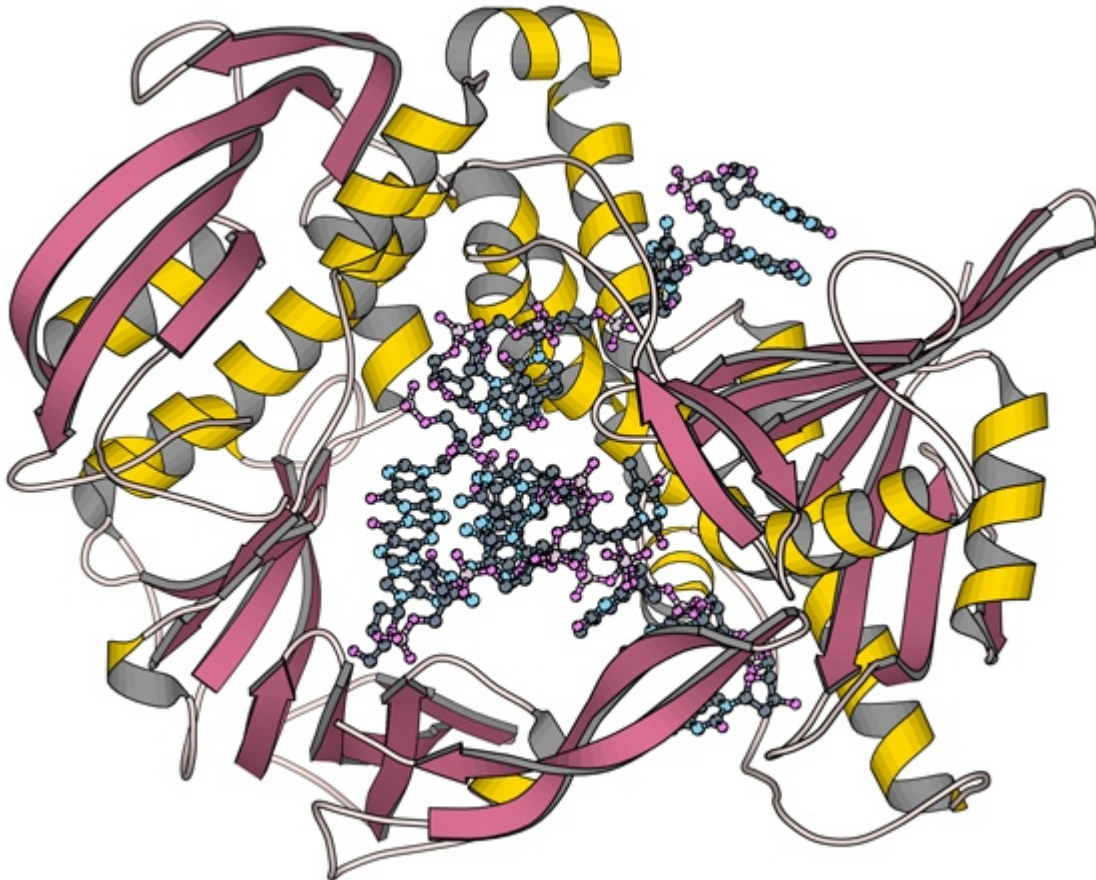
Deoxyribonucleotide Fluoropentaphosphates



Molecular structures of the phospho-linked deoxyribonucleotide pentaphosphates, and their normalized fluorescence emission spectra of their fluorophores. The two excitation wavelengths used are indicated by arrows (532 and 643nm). Reproduced from Eid, *et al.* 2009.

Proteins:

ϕ 29 Polymerase



Reproduced from Berman, *et al.* 2007.

Theoretical molecular weight: 66,714 Daltons.

Quaternary Structure: Monomeric.

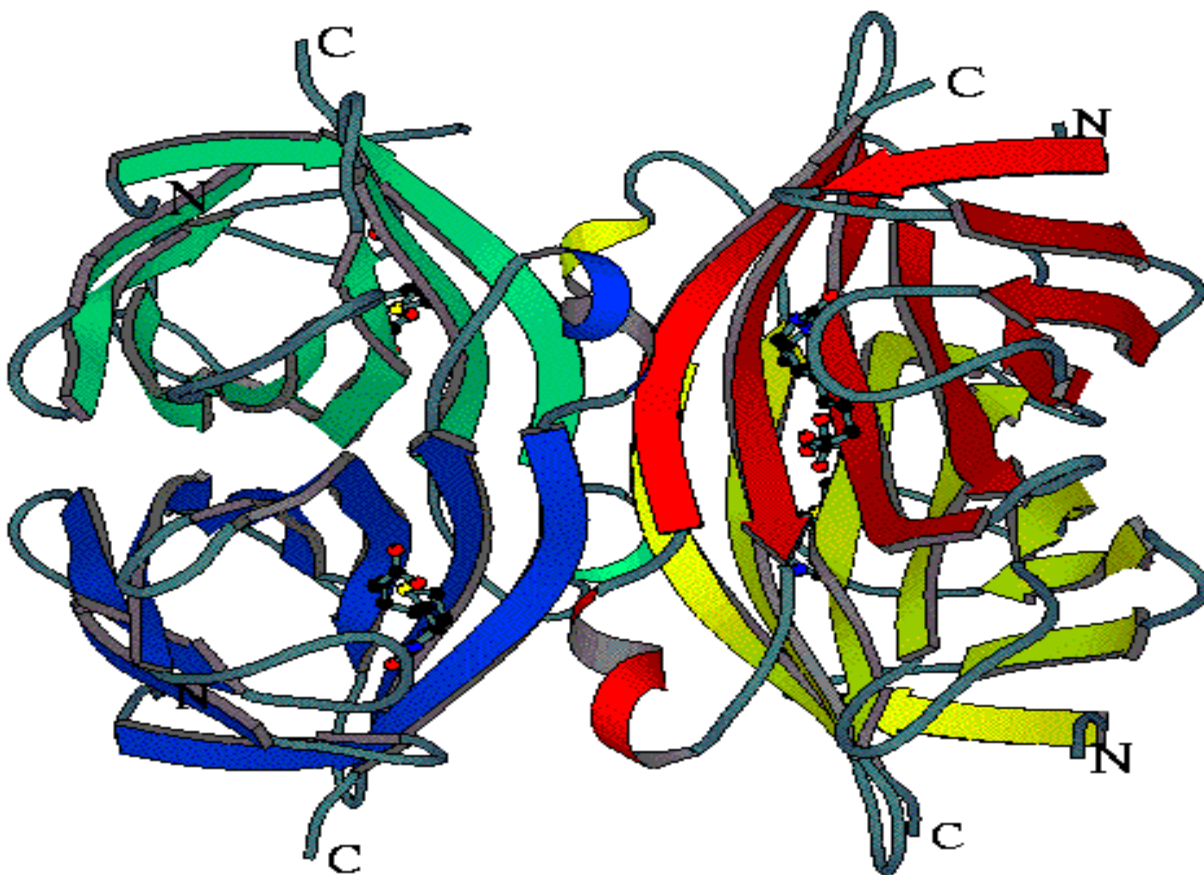
Manufacturer: New England Biolabs

Source: An *E. coli* strain that carries the phi29 DNA Polymerase gene from bacteriophage phi29

Cost: 1,250 units at 10,000 units/ml for \$244.00

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.

Streptavidin



Reproduced from Berman HM, *et al.* 2000.

Theoretical molecular weight: 52,800 Daltons.

Quaternary structure: Tetrameric.

Manufacturer: New England Biolabs

Source: *Streptomyces avidinii*

Cost: 1 mg as 1 mg/ml for \$56.00

Appendix B:

DNA Extraction and Isolation

Protocol provided by QIAGEN:

Whole genome amplification from buccal cells using the REPLI-g[®] Midi Kit

This procedure has been adapted by customers and is for whole genome amplification from buccal cells using the REPLI-g Midi Kit. The procedure is optimized for air-dried buccal swabs with cotton or Dacron[®] tips, and brushes or swabs with an ejectable head (e.g., Whatman[®] Omni Swab). Other swab types may also be used. **The procedure has not been thoroughly tested and optimized by QIAGEN.**

Note: This protocol may be adapted for use with the REPLI-g Mini Kit, using the same reaction setup. In rare cases, potential inhibitors present in the starting material may have inhibitory effects on amplification when using the REPLI-g Mini Kit. In these cases, we recommend using the REPLI-g Midi Kit. Alternatively, upstream genomic DNA purification can be performed (e.g., using a QIAamp Kit) with subsequent whole genome amplification of the purified DNA following the standard protocol in the *REPLI-g Mini/Midi Handbook*.

IMPORTANT: Please consult the “Safety Information” and “Important Notes” sections in the *REPLI-g Mini/Midi Handbook* before beginning this procedure. For safety information on the additional chemicals mentioned in this protocol, please consult the appropriate material safety data sheets (MSDSs) available from the product supplier.

Equipment and reagents to be supplied by user

- Microcentrifuge tubes
- Microcentrifuge
- Water bath or heating block
- Vortexer
- Pipets and pipet tips
- Ice
- Nuclease-free water
- TE buffer (10 mM Tris-Cl; 1 mM EDTA, pH 8.0)
- Swabs, such as sterile Omni Swabs (available from Whatman), or Puritan[®] applicators with plastic shafts and cotton or Dacron tips (available from Hardwood Products)*

* This is not a complete list of suppliers and does not include many important vendors of biological supplies.

Important points before starting

- To collect a sample, scrape a fresh swab firmly against the inside of each cheek 6 times. Ensure that the person providing the sample has not consumed any food or drink in the 30 minutes prior to sample collection. Start the DNA amplification procedure within 2 hours of collection.
- For best results, the template DNA should be >2 kb in length with some fragments >10 kb.
- REPLI-g Midi DNA Polymerase should be thawed on ice (see step 7). All other components can be thawed at room temperature.
- A DNA control reaction can be set up using 10 ng (1 μ l) control genomic DNA (e.g., REPLI-g Human Control Kit, cat. no. 150090).

Things to do before starting

- Prepare Buffer DLB by adding 500 μ l nuclease-free water to the tube; mix thoroughly and centrifuge briefly.
- **Note:** Reconstituted Buffer DLB can be stored for 6 months at -20°C . Buffer DLB is pH-labile. Avoid neutralization with CO_2 .
- Set a water bath or heating block to 30°C .
- All buffers and reagents should be vortexed before use to ensure thorough mixing.

Procedure

1. **Place the swab in a 1.5 ml microcentrifuge tube. Add 1 ml TE buffer and vortex for 10 s.**
If using an Omni Swab, eject the swab head by pressing the end of the inner shaft towards the swab head.
If using a cotton or Dacron swab, separate the swab head from its shaft by hand or by using scissors.
2. **Remove the swab from the microcentrifuge tube using forceps. Squeeze as much liquid as possible out of the swab by pushing the swab against the side of the microcentrifuge tube.**
IMPORTANT: The swab must be removed from the microcentrifuge tube prior to cell lysis (step 5).
3. **Centrifuge the microcentrifuge tube containing buccal cells at maximum speed for 10 s. Discard the supernatant and wash the buccal cells by resuspending the pellet in 1 ml TE and vortexing for 1 min.**
4. **Centrifuge the microcentrifuge tube containing buccal cells at maximum speed for 10 s. Discard the supernatant and resuspend the buccal cell pellet in 30 μ l TE.**
5. **Add 35 μ l reconstituted Buffer DLB to the resuspended buccal cells and mix by pipetting up and down 3 times. Place the microcentrifuge tube on ice for 10 min.**

- 6. Add 35 µl Stop Solution to the lysed buccal cells and mix by pipetting up and down 3 times.**

Note: 10 µl lysed and neutralized buccal cells are used in a 50 µl REPLI-g reaction.

- 7. Thaw REPLI-g Midi DNA Polymerase on ice. Thaw all other components at room temperature, vortex, and centrifuge briefly.**

The REPLI-g Midi Reaction Buffer may form a precipitate after thawing. The precipitate will dissolve by vortexing for 10 s.

- 8. Prepare a master mix on ice according to Table 1. Mix and centrifuge briefly.**

IMPORTANT: Add the master mix components in the order listed in Table 1. After addition of water and REPLI-g Midi Reaction Buffer, briefly vortex and spin down the mixture before addition of REPLI-g Midi DNA Polymerase. The master mix should be kept on ice and used immediately upon addition of the REPLI-g Midi DNA Polymerase.

Table 1. Preparation of Master Mix

Component	Volume/reaction
Nuclease-free water	10 µl
REPLI-g Midi Reaction Buffer	29 µl
REPLI-g Midi DNA Polymerase	1 µl
Total volume	40 µl

- 9. Add 40 µl master mix to 10 µl lysed and neutralized buccal cells (step 6).**

- 10. Incubate at 30°C for 8–16 h.**

Maximum DNA yield is achieved using an incubation time of 16 h. After incubation at 30°C, heat the water bath or heating block up to 65°C if the same water bath or heating block will be used in step 11.

- 11. Inactivate REPLI-g Midi DNA Polymerase by heating the sample at 65°C for 3 min.**

- 12. Store amplified DNA at 4°C for short-term storage or –20°C for long-term storage.**

DNA amplified using the REPLI-g Midi kit should be treated as genomic DNA with minimal freeze-thaw cycles. Storage of nucleic acids at low concentration over a long period of time may result in acid hydrolysis. We therefore recommend storage of nucleic acids at a concentration of at least 100 ng/µl.

QIAGEN REPLI-g Kits are for use only as licensed by Amersham Biosciences Corp (part of GE Healthcare Bio-Sciences) and QIAGEN GmbH. The Phi 29 DNA polymerase may not be re-sold or used except in conjunction with the other components of this kit. See U.S. Patent Nos. 5,854,033, 6,124,120, 6,143,495, 5,001,050, 5,198,543, 5,576,204, and related U.S. and foreign patents. The REPLI-g Kit is developed, designed, and sold for research purpose only.

QIAGEN kit handbooks can be requested from QIAGEN Technical Service or your local QIAGEN distributor.

Selected kit handbooks can be downloaded from <http://www.qiagen.com/literature/default.aspx>.

Material safety data sheets (MSDS) for any QIAGEN product can be downloaded from www.qiagen.com/ts/msds.asp.

Trademarks: QIAGEN®, REPLI-g® (QIAGEN Group); Dacron® (E. I. du Pont de Nemours and Company); Eppendorf® (Eppendorf-Netheler-Hinz GmbH); Puritan® (Hardwood Products Company); Whatman® (Whatman International Ltd.).

Appendix C:

Synthesis of Phospho-Linked Nucleotide Pentaphosphates

Reproduced from J. Korlach *et al.* 2008

The synthesis is described using Alexa Fluor 488-aminoethyl-dG5P (A488-dG5P) as an example.

Fmoc-6-aminoethylphosphate: Fmoc-6-aminohexanol (1 g, 2.94 mMoles) was co-evaporated with anhydrous acetonitrile (2 ×20ml) then suspended in 10 ml anhydrous triethylphosphate. Phosphorus oxychloride (550 μ l, 5.88 mMoles, 2 eq.) was added to the stirring suspension. After 2 hours, HPLC showed disappearance of the Fmoc-aminohexanol. The reaction was quenched by the addition of 100 ml 0.1M triethylamine bicarbonate (pH 6.8) and stirred for 30 minutes. The compound was purified by reverse phase HPLC on a Waters Xterra C18 RP 30×100 column using an acetonitrile gradient in 0.1M triethylamine bicarbonate. The fractions containing product were evaporated, followed by co-evaporation with methanol (2 ×). The residue was triturated twice with 100 ml diethylether and dried under vacuum to give a white powder. Yield: 1.24 g, 68% as bis-triethylamine salt. HPLC 98%.

Fmoc-6-aminoethylphosphate: Fmoc-6-aminoethylphosphate (200 mg, 320 μ Moles) was co-evaporated twice with anhydrous acetonitrile, then taken up in 2 ml anhydrous DMF. 1,1'-Carbonyldiimidazole (CDI, 207 mg, 1280 μ Moles, 4 Eq.) was added and stirred at ambient temperature for 4 hours. Methanol (77 μ l, 1920 μ Moles) was added and stirred 30 minutes. Tributylamine-H₂PO₄ (3200 μ Moles, 10 Eq.), prepared by mixing equimolar amounts of tributylamine and 85% phosphoric acid followed by co-evaporation 3 times with anhydrous acetonitrile, was dissolved in 4 ml anhydrous DMF and added to the reaction. The reaction mixture was stirred 16 hours. HPLC showed 3% Fmoc-aminoethylphosphate remaining. The reaction mixture was diluted to 50 ml with 0.1M TEAB, and was purified by RP HPLC on a Waters Xterra C18 RP 30×100 column using an acetonitrile gradient in 0.1M triethylamine bicarbonate. The fractions containing product were evaporated, followed by co-evaporation with methanol (2 ×). The residue was co-evaporated with anhydrous acetonitrile. Yield: 186 mg, 73% as tris-TEA salt. HPLC 96%.

Aminoethyl-dG5P: Fmoc-6-aminoethylphosphate (186 mg, 233 μ Moles) was co-evaporated twice with anhydrous acetonitrile, then taken up in 3ml anhydrous DMF. CDI (150 mg, 930 μ Moles, 4 Eq.) was added and stirred at ambient temperature for 4 hours. Methanol (56 μ l, 1400 μ Moles) was added and stirred 30 minutes. dGTP (TEA salt, 350 μ Moles, 1.5 Eq.) was co-evaporated 3 times with anhydrous acetonitrile and suspended in 2 ml anhydrous DMF. The Fmoc-aminoethylphosphoimidazolite reaction was added to the dGTP solution followed by anhydrous MgCl₂ (3500 μ Moles, 333 mg, 10 Eq.). The reaction was stirred 18 hours. HPLC showed 28% of the Fmoc-aminoethylphosphate converted to

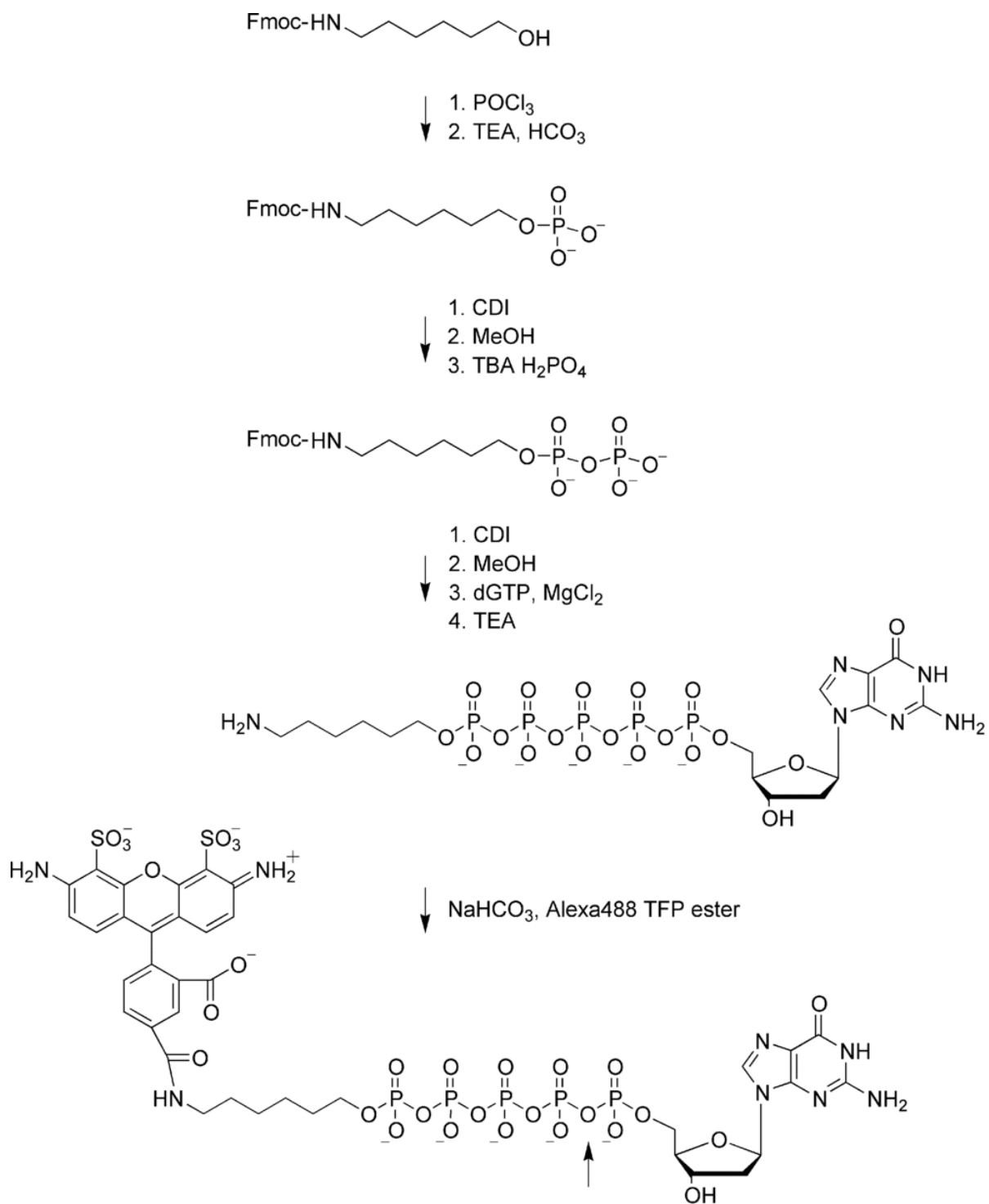
Fmoc-aminohexyl-dG5P. The reaction mixture was diluted to 125 ml with 0.1M TEAB, and was purified by RP HPLC on a Waters Xterra C18 RP 30 ×100 column using an acetonitrile gradient in 0.1M triethylamine bicarbonate. The fractions containing product were evaporated, followed by co-evaporation with methanol (2 ×). The residue was taken up in 20 ml 10% TEA/water and stirred 16 hours to remove the Fmoc protecting group from the amine on the linker. Triethylamine was evaporated, water was added to 25 ml and the solution was extracted 3 times with 25 ml diethyl ether. The product was purified from the aqueous layer by anion exchange chromatography on Q sepharose FF using a TEAB gradient. Yield 42 μMoles, 18%, HPLC 98%.

Alexa Fluor 488-aminohexyl-dG5P (A488-dG5P): Aminohexyl-dG5P (1 μMole) was dissolved in 0.05 M NaHCO₃ pH 8.7 (200 μl) and was added to 1 mg Alexa Fluor 488-TFP ester (Invitrogen, Carlsbad, CA, USA). The mixture was briefly sonicated. After 4 hours, HPLC showed no active ester remaining. The product was identified by characteristic PDA scan. The compound was purified by IEX on Q sepharose FF with a TEAB gradient. The product was further purified by RP HPLC on a Waters Xterra RP C18 19 ×100 column using an acetonitrile gradient in 0.1M TEAB. The fractions containing pure product were evaporated, followed by co-evaporation with methanol (2 ×). The residue was dissolved in water and was quantitated by UV-Vis spectrophotometry. Yield 370 nMoles 37%, HPLC 99%.

The other dNTPs were derivatized with Alexa Fluor 633 NHS ester (aminohexyl-dA5P), Alexa Fluor 680 NHS ester (aminohexyl-dC5P), and Alexa Fluor 568 NHS ester (aminohexyl-dT5P).

Appendix C: Nucleotide Synthesis

Synthesis Scheme:



Appendix D:

Equipment Specifications

D.1 Microscopes and Peripherals

D.2 EMCCD Cameras

D.3 Nanopositioning stages

D.4 Signal Processing Servers

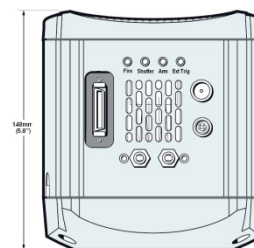
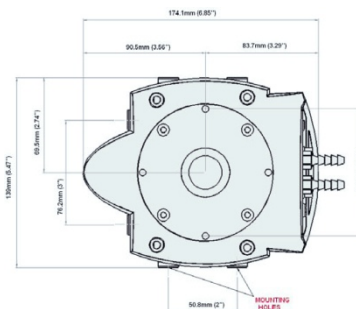
D.5 Reassembly Servers

Olympus IX71 Inverted Research Microscope



Objective Lens	UPLSAPO 60XW
	<ul style="list-style-type: none"> ◇ Plan apochromat ◇ Water immersion ◇ NA = 1.20 ◇ WD = 0.28 mm
Illumination Source	U-LH100HGAP0 Mercury Lamp
	<ul style="list-style-type: none"> ◇ Aspherical optics ◇ Apochromatic lens ◇ Average lamp life: 300 hours
Mirror/Filter Unit	U-MWIB2
Unit Price	\$30,000

Andor iXon EM+ DU-897 Back-Illuminated EMCCD



Active Pixels	512 x 512
Pixel Size	16 μm
Image area	8.2 x 8.2 mm
Pixel Well Depth	160,000 (max 220,000)
Gain Register PWD	800,000
Readout Rate	10 MHz (max)
Frame Rate	31 fps
QE at 600 nm	95%
Unit Price	\$32,500

Appendix D: Equipment Specifications



426 Sullivan Avenue, Suite #3, South Windsor, CT 06074 USA
 Tel: 1 (888) 290-9211, Fax: 1 (888) 290-8688
 www.andor.com

Quotation Ref: 03ZEQA1000FR

Issue Date: Friday, 30 January 2009

Valid Until: 30 Days beyond Issue date

Mr. Boris Petkov
 University of Pennsylvania
 Dept of Engineering and Applied Science
 CBE
 Philadelphia, PA 19104
 USA
 Email : borisp@seas.upenn.edu

Technical Inquiries: Charles Fanghella
 Tel: (860) 290-9211
 Order Inquiries: Aida Dubois
 Tel: 8602909211 X.201

Item	Description	Part.#	Qty	Unit Price	Discount%	SubTotal
1	512x512,16um,EMOCD,BV,10MHz,-100C	DU-897E- CS0-#8V	1	34,500.00	12.00	30,300.00
2	IXON PCI Controller Card	CCI-23	1	1,800.00	15.00	1,500.00
3	Imaging Software	SOLIS (I)	1	1,500.00	50.00	700.00
4	Software Develop Kit - CCD PCI System	ANDOR- SDK-CCD	1	500.00	100.00	0.00

TOTAL : 32,500.00

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



PI P-587 6-Axis Piezo Stage with E-710.6CD Digital Controller



Active Axes	X, Y, Z, θ_x, θ_y, θ_z
Max Travel (X, Y, Z)	800, 800, 200 μm
Max Angle (θ_x, θ_y, θ_z)	± 0.5, ± 0.5, ± 0.5 mrad
Open/Closed Loop Resolution (X, Y)	0.9/2.2 nm
Open/Closed Loop Resolution (Z)	0.4/0.7 nm
Open/Closed Loop Resolution (θ_x, θ_y)	0.05/0.1 mrad
Open/Closed Loop Resolution (θ_z)	0.1/0.3 mrad
Automation	Auto-Alignment with CCD Feedback
Unit Price	\$72,000

IBM System x3450 79483CX (Signal Processing Server)



Processor (CPU)	Quad-Core Intel Xeon E5462
Processor Speed	2.80 GHz
CPUs	2
Front Side Bus	1 GHz
Internal L2 Cache	12 MB
RAM	8 GB (800 MHz DDR2)
Hard Disk	250 GB, 7200 RPM SATA II
Communications	Integrated Dual Gigabit Ethernet
Form Factor	1U Rack
Unit Price	\$4,000

IBM System p560 Express 8234-EMA2 (Reassembly Server)



Processor (CPU)	8-Core IBM POWER6
Processor Speed	3.6 GHz
CPUs	1
Front Side Bus	1 GHz
Int. L2/L3 Cache	8 MB / 32 MB
RAM	8 GB (800 MHz DDR2)
Hard Disk	2 x 146 GB, 15,000 RPM SAS
Communications	Integrated Dual Gigabit Ethernet
Form Factor	4U Rack
Unit Price	\$70,650

Appendix E:

MATLAB Simulation Code

- E.1 Program Framework*
- E.2 Local Alignment*
- E.3 Vote Counting*
- E.4 Base Assignment*
- E.5 Representative Sensitivity Analysis*
- E.6 Dual-Camera Peak Identification*

Program Framework with Genome Generation Routine

```

1  function [finalerrorrate snpgenome finalgenome] = genome(genlength,delrate,errorrate,mult)
2
3  clear snpgenome finalgenome votefrac maxm index errorloc errorpos error half snppos ...
4  fragcell alignment score start counter finalerrorrate
5
6  % -- Set sequence constants -- %
7  fragmean = 1000;
8  fragstdev = 250;
9  snprate = 1/1000;
10
11 % -- Set alignment parameters -- %
12 gapo = 4;
13 gape = 21;
14
15 % -- Set counter variables -- %
16 fragind = 1;
17 dels = 0;
18 errs = 0;
19 snps = 0;
20
21 % -- Estimate size of fragment library cell array -- %
22 fragcell = cell(1,round(length(tmv)*mult/(1.2*fragmean)));
23
24 % -- Preallocate vote counting matrices -- %
25 Avotecount = zeros(1,genlength);
26 Cvotecount = zeros(1,genlength);
27 Tvotecount = zeros(1,genlength);
28 Gvotecount = zeros(1,genlength);
29 Xvotecount = zeros(1,genlength);
30 sumcount = zeros(1,genlength);
31 votefrac = zeros(5,genlength);
32
33 % -- Preallocate error analysis cell arrays -- %
34 half = cell(0,0);
35 error = cell(0,0);
36 errorpos = cell(1,genlength);
37 errorloc = cell(0,0);
38
39 % -- Generate reference (consensus) genome -- %
40 tmv = randseq(genlength);
41 tmv = nt2int(tmv);
42
43
44 %%%%%%%%%%%
45 %SNP Insertion%
46 %%%%%%%%%%%
47
48 snplambda = genlength*snprate;
49 m = -round((5*snplambda));
50 snppos = cell(0,genlength);
51 snpgenome = tmv;
52
53 while m <= genlength
54     clear vvv
55     vvv = round(exprnd(1/snprate));

```



```

56         if (vvv + m) <= 0
57             m = m + vvv;
58         elseif ((vvv + m) > 0)
59             if (vvv+m) < genlength
60                 uuu = unidrnd(4);
61                 snpgenome(m+vvv) = uuu;
62                 snppos{1,m+vvv} = 1;
63                 m = m + vvv;
64                 snps = snps + 1;
65             else
66                 m = m + vvv;
67             end
68         end
69     end
70
71     genlength = length(snpgenome);
72
73
74     %%%%%%%%%%%
75     %Fragmentation%
76     %%%%%%%%%%%
77
78     length1 = 0;
79
80     while length1 <= (mult*genlength)
81         lastcut = 1;
82         while (lastcut < genlength)
83             ttt = unidrnd(50);
84             fraglength = round(normrnd(fragmean,fragstdev));
85             cutsite = fraglength + lastcut;
86             if ttt == 1
87                 if (cutsite >= genlength)
88                     fragcell{1,fragind} = snpgenome(1,lastcut:genlength);
89                     fragind = fragind+1;
90                 else
91                     fragcell{1,fragind} = snpgenome(1,lastcut:cutsite);
92                     fragind = fragind+1;
93                 end
94             else
95                 lastcut = cutsite;
96             end
97         end
98         length2 = zeros(1,mult*1.5*genlength/fragmean);
99
100        for k = 1:length(fragcell)
101            length2(1,k) = length(fragcell{1,k});
102        end
103
104        length1 = sum(length2(1,:));
105    end
106
107    fragkeep = length(fragcell);
108
109
110    %%%%%%%%%%%
111    %Error Insertion%
112    %%%%%%%%%%%
113
114    for n = 1:fragind-1
115
116        errorlambda = length(fragcell{1,n})*errorrate;
117        m = -round((5*errorlambda));

```

Appendix E: MATLAB Simulation Code

```
118         errorfrag = fragcell{1,n};
119
120         while m <= length(fragcell{1,n})
121             clear vv
122             vv = round(exprnd(1/errorrate));
123             if (vv + m) <= 0
124                 m = m + vv;
125             elseif ((vv + m) > 0)
126                 if (vv+m) < length(fragcell{1,n})
127                     uuu = unidrnd(4);
128                     errorfrag(1,(m+vv)) = uuu;
129                     m = m + vv;
130                     errs = errs + 1;
131                 else
132                     m = m + vv;
133                 end
134             end
135         end
136     fragcell{1,n} = errorfrag(1,:);
137 end
138
139 %%%%%%%%%%%
140 %Deletion Insertion%
141 %%%%%%%%%%%
142
143 for n = 1:fragind-1
144
145     k = 1;
146     uuu = poissrnd(delrate*length(fragcell{1,n}));
147     delnum = zeros(1,uuu);
148     fragchar1 = fragcell{1,n};
149
150     while k <= uuu+1
151         gam = round(gamrnd(k,1/delrate));
152         if gam <= length(fragchar1)
153             delnum(1,k) = gam;
154             k = k+1;
155         else
156             k = k+1;
157         end
158     end
159
160     delnum = sort(delnum(1,:));
161     delfrag = [];
162
163     i = 1;
164     k = 1;
165     m = 1;
166
167     while i <= length(fragchar1)
168         if k < length(delnum)
169             if delnum(1,k) == i
170                 k = k + 1;
171                 i = i + 1;
172                 dels = dels + 1;
173             else
174                 delfrag(1,m) = fragchar1(1,i);
175                 m = m + 1;
176                 i = i + 1;
177             end
178         else
179
```

Appendix E: MATLAB Simulation Code

```

180             delfrag(1,m) = fragchar1(1,i);
181             m = m + 1;
182             i = i + 1;
183         end
184     end
185     fragcell{1,n} = delfrag;
186 end
187
188 %%%%%%%%%%%%%%
189 %Local Alignment%
190 %%%%%%%%%%%%%%
191
192 [Avotecount,Cvotecount,Gvotecount,Tvotecount,Xvotecount,sumcount] = ...
193     localalign1(Avotecount,Cvotecount,Gvotecount,Tvotecount,Xvotecount,sumcount,...
194     fragkeep,fragcell,tmv,gapo,gape,genlength);
195
196 %%%%%%%%%%%%%%
197 %Base Assignment%
198 %%%%%%%%%%%%%%
199
200 [finalgenome snpgenome] = baseassign1(snpgenome,genlength,Avotecount,...
201     Cvotecount,Gvotecount,Tvotecount,Xvotecount,sumcount,votefrac);
202
203 %%%%%%%%%%%%%%
204 %Error Evaluation%
205 %%%%%%%%%%%%%%
206
207 % -- Count sites with less than 3-fold coverage -- %
208 for n = 1:genlength
209     if sumcount(1,n) == 2
210         half{1,n} = n;
211     else
212     end
213 end
214
215 % -- Compare final genome to true sequence -- %
216 % Determine number and position of errors
217 for n = 1:genlength
218     errorpos{1,n} = snpgenome(n) - finalgenome(n);
219 end
220
221 k = 1;
222
223 for n = 1:genlength
224     if errorpos{1,n} ~= 0
225         error{1,k} = 1;
226         k = k+1;
227     end
228 end
229
230 % -- Calculate actual coverage multiplicity -- %
231 multact = length1/genlength;
232
233 % -- Calculate error rate in ppm -- %
234 finalerrrate = 1000000*(length(error)/genlength);
235 end

```

Local Alignment Subroutine

```

1  function [Avotecount,Cvotecount,Gvotecount,Tvotecount,Xvotecount,sumcount] = ...
2      localalign (Avotecount,Cvotecount,Gvotecount,Tvotecount,Xvotecount,sumcount, ...
3          fragkeep,fragcell,tmv,gapo,gape,genlength)
4
5  for n = 1:fragkeep-1
6
7      if length(fragcell{1,n}) > 1
8
9          % -- A single fragment is selected from the fragment library -- %
10         frag = fragcell{1,n};
11         length3 = length(frag);
12
13         % -- A shorter fragment is created from the first 10 bases -- %
14         if length3 > 10
15             short = frag(1:10);
16         else
17             short = frag;
18         end
19
20         % -- This short fragment is aligned with the reference genome, recording starting point-- %
21         [scorevalue1 alignvalue1 startvalue] = swalign(short, tmv,...
22             'alphabet','nt','gapopen',gapo,'extendgap',gape);
23
24         startvalue = startvalue(2,1);
25
26         if startvalue<=1
27             incr1 = 0;
28             startvalue = 1;
29         else
30             incr1 = 1;
31         end
32
33         if genlength-startvalue<=(20+length3)
34             incr2 = genlength-startvalue-length3;
35         elseif genlength-startvalue>(20+length3)
36             incr2 = 20;
37         end
38
39         % -- The full fragment is aligned to the stretch of reference genome immediately after the starting point -- %
40         [scorevalue alignvalue startvalue1] = swalign(frag,tmv((startvalue-incr1):startvalue+length3+incr2),...
41             'alphabet','nt','gapopen',gapo,'extendgap',gape);
42
43         alignseq = nt2int(alignvalue(1,:));
44
45
46         % -- The base position votes are counted before moving to the next fragment -- %
47         [Avotecount,Cvotecount,Gvotecount,Tvotecount,Xvotecount,sumcount] = votecount(genlength, ...
48             startvalue,alignseq,Avotecount,Cvotecount,Gvotecount,Tvotecount,Xvotecount,sumcount);
49
50     end
51 end

```

Vote-Counting Subroutine

```
1 function [Avotecount,Cvotecount,Gvotecount,Tvotecount,Xvotecount,sumcount] = ...
2     votecount(genlength,startvalue,alignseq,Avotecount, Cvotecount, ...
3         Gvotecount,Tvotecount,Xvotecount,sumcount)
4
5 % -- The aligned sequences are received from the alignment routine -- %
6 for k = 1:length(alignseq)
7
8     % -- Each base is identified, indexed, and counted -- %
9     if startvalue+k-1<=genlength
10
11         if alignseq(1,k) == 1
12             Avotecount(1,startvalue+k-1) = Avotecount(1,startvalue+k-1) + 1;
13
14         elseif alignseq(1,k) == 2
15             Cvotecount(1,startvalue+k-1) = Cvotecount(1,startvalue+k-1) + 1;
16
17         elseif alignseq(1,k) == 4
18             Tvotecount(1,startvalue+k-1) = Tvotecount(1,startvalue+k-1) + 1;
19
20         elseif alignseq(1,k) == 3
21             Gvotecount(1,startvalue+k-1) = Gvotecount(1,startvalue+k-1) + 1;
22
23         else
24             Xvotecount(1,startvalue+k-1) = Xvotecount(1,startvalue+k-1) + 1;
25         end
26     end
27 end
28
29 % -- Total votes at each position are computed for the assignment subroutine -- %
30 sumcount(1,:) = Avotecount(1,:) + Cvotecount(1,:) + Tvotecount(1,:) + Gvotecount(1,:);
31
32
33 end
```

Base Assignment Subroutine

```
1 function [finalgenome snpgenome] = baseassign1(snpgenome,genlength,...
2         Avotecount,Cvotecount,Gvotecount,Tvotecount,Xvotecount,...
3         sumcount,votefrac)
4
5 % -- Counts at each position are converted to proportions -- %
6
7 votefrac(1,1:genlength) = Avotecount(1,1:genlength)./sumcount(1,1:genlength);
8
9 votefrac(2,1:genlength) = Cvotecount(1,1:genlength)./sumcount(1,1:genlength);
10
11 votefrac(4,1:genlength) = Tvotecount(1,1:genlength)./sumcount(1,1:genlength);
12
13 votefrac(3,1:genlength) = Gvotecount(1,1:genlength)./sumcount(1,1:genlength);
14
15 votefrac(5,1:genlength) = Xvotecount(1,1:genlength)./sumcount(1,1:genlength);
16
17 % -- The the index of the highest vote proportion determines the base assignment -- %
18
19 [maxm index] = max(votefrac,[],1);
20
21 % -- Finalgenome (computed) and snpgenome (true) are converted to nucleotides -- %
22
23 finalgenome = int2nt(index);
24 snpgenome = int2nt(snpgenome);
25
26 end
```

Representative Sensitivity Analysis Script (for Deletion Rate)

```

1      % -- Set sequence constants -- %
2      genlength = 20000;
3      errorrate = .001;
4      mult = 9;
5
6      % -- Create matrix of deletion frequencies to be tested -- %
7      deltest(1,1)= 0.001;
8      deltest(1,2)= 0.005;
9      deltest(1,3)= 0.01;
10     deltest(1,4)= 0.05;
11     deltest(1,5)= 0.1;
12     deltest(1,6)= 0.3;
13
14     % -- Preallocate arrays -- %
15     errorc = cell(0,0);
16     analysis = zeros(length(deltest),7);
17     l = 1;
18
19     trials = 120;
20     delrate = deltest(1,1);
21     % -- Test 'm' levels of deletion frequencies with a sample size of 'trials' at each level -- %
22     for m = 1:length(deltest)
23         totalerrorrate = zeros(trials,6);
24         for n = 1:trials
25
26             [t1 t2 finalerrorrate snpgenome finalgenome] = genome(genlength,delrate,errorrate,mult);
27
28             % -- Record computing time and overall error rate for each trial -- %
29             totalerrorrate(n,2) = t1;
30             totalerrorrate(n,3) = t2;
31             totalerrorrate(n,4) = finalerrorrate;
32
33             % -- Record the value and position of each error -- %
34             for q = 1:length(snpgenome)
35                 if snpgenome(q) - finalgenome(q) ~= 0
36                     errorc{l,1} = q;
37                     errorc{l,2} = snpgenome(q);
38                     errorc{l,3} = finalgenome(q);
39                     l = l+1;
40                 end
41             end
42         end
43     end
44
45     % -- Calculate means and standard deviations for each level of deletion frequency -- %
46     analysis(m+1,1) = 1/delrate;
47     analysis(m+1,2) = sum(totalerrorrate(1:trials,2))/trials;
48     analysis(m+1,3) = std(totalerrorrate(1:trials,2));
49     analysis(m+1,4) = sum(totalerrorrate(1:trials,3))/trials;
50     analysis(m+1,5) = std(totalerrorrate(1:trials,3));
51     analysis(m+1,6) = sum(totalerrorrate(1:trials,4))/trials;
52     analysis(m+1,7) = std(totalerrorrate(1:trials,4));
53     delrate = deltest(1,m);

```

Monte Carlo Simulation for Dual Camera Peak Detection

```

1 % Monte Carlo Simulation for Dual Camera Peak Detection
2 % Basic Equation C2/C1- denotes Intensity reading in camera 2 vs. Intensity
3 % reading for camera 1.
4 %
5 %-----
6 % Generating n samples for a normal distribution for the wavelength
7 % intensities picked up by each camera:
8 %         ca_cb= (randn(n,1)*std)+mean;
9 %         a is the camera number
10 %        b is the peak corresponding to the wavelength
11 %        std is the standard deviation
12 %        mean is the mean
13     n = 100000;%-----number of iterations
14     c1_1 = ( randn(n,1)*409.1812)+24198;
15     c2_1 = ( randn(n,1)*3.821595)+226;
16     c1_2 = ( randn(n,1)*395.1607)+28178;
17     c2_2 = ( randn(n,1)*25.25674)+1801;
18     c1_3 = ( randn(n,1)*10.363)+548;
19     c2_3 = ( randn(n,1)*476.074)+25175;
20     c1_4 = ( randn(n,1)*1.675957)+110;
21     c2_4 = ( randn(n,1)*476.9316)+31303;
22 %-----
23 % Simulation run for each peak
24     peak1 = c2_1./c1_1;
25     peak2 = c2_2./c1_2;
26     peak3 = c2_3./c1_3;
27     peak4 = c2_4./c1_4;
28 %-----
29 % Plotting the Histograms
30     [n1, xout1] = hist(log(peak1), 50);
31     [n2, xout2] = hist(log(peak2), 50);
32     [n3, xout3] = hist(log(peak3), 50);
33     [n4, xout4] = hist(log(peak4), 50);
34
35     plot(xout1, n1, 'r', xout2, n2, 'g', xout3, n3, 'b', xout4, n4, 'm');
36
37 %-----
38 % Output for each peak
39     peak1_mean = mean(peak1)
40     peak1_std = std(peak1)
41     peak1_err = peak1_std/n^(0.5)% Finds standard error.
42
43     peak2_mean = mean(peak2)
44     peak2_std = std(peak2)
45     peak2_err = peak2_std/n^(0.5)
46
47     peak3_mean = mean(peak3)
48     peak3_std = std(peak3)
49     peak3_err = peak3_std/n^(0.5)
50
51     peak4_mean = mean(peak4)
52     peak4_std = std(peak4)
53     peak4_err = peak4_std/n^(0.5)

```


Appendix F:

Financial Pro Forma

F.1 Case 1

F.2 Case 2

F.3 Case 3

F.4 Case 4

F.5 Case 5

F.6 Case 6

F.7 Case 7

F.8 Case 8

F.9 Case 9

F.10 Case 10

F.11 Case 11

F.12 Case 12

F.13 Growth Case

Appendix F: Financial Pro Forma

Case 1 Year	2010	2011	2012	2013	2014	2015
<u>Income Statement</u>						
Revenue	\$ -	\$ 15,000.0	\$ 30,000.0	\$ 30,000.0	\$ 30,000.0	
Cost of Sales	(172.6)	(643.6)	(988.6)	(988.6)	(988.6)	
Operating, SG&A Expenses	(292.0)	(2,380.0)	(3,730.0)	(3,730.0)	(3,730.0)	
Depreciation	-	(141.7)	(762.0)	(992.6)	(595.5)	
Pre-Tax Income	(464.6)	11,834.7	24,519.4	24,288.8	24,685.9	
Tax @ 40%	185.8	(4,733.9)	(9,807.8)	(9,715.5)	(9,874.3)	
Net Income	\$ (278.8)	\$ 7,100.8	\$ 14,711.6	\$ 14,573.3	\$ 14,811.5	
<u>Cash Flow Statement</u>						
Cash From Operating Activities	\$ -	\$ 141.7	\$ 762.0	\$ 992.6	\$ 595.5	
Plus: Depreciation						
Changes in Working Capital						
(Increase)/Decrease in A/R	\$ -	\$ (1,232.9)	\$ (1,232.9)	\$ -	\$ -	
(Increase)/Decrease in Inventory	-	(28.4)	(28.4)	-	-	
Increase/(Decrease) in A/P	24.5	111.0	111.0	-	-	
(Increase)/Decrease in C/R	(73.0)	(184.5)	-	-	-	
Total Change in Working Capital	\$ (48.5)	\$ (1,334.8)	\$ (1,150.3)	\$ -	\$ -	
Cash From Investing Activities	(708.3)	(2,676.8)	-	-	-	
(Purchase)/Selling of Equipment						
Cash From Financing Activities	1,200.0	3,500.0	-	-	-	
Issuance of Common Stock	\$ 164.5	\$ 6,730.9	\$ 14,323.4	\$ 15,565.9	\$ 15,407.1	\$ 51,356.9 (Terminal Value)
Free Cash Flow	0%	50%	100%	100%	100%	100%
% of Design Capacity						
<u>Investment Divided Free Cash Flows</u>						
Series A @ 31% of Equity	\$ (1,200.0)	\$ 148.0	\$ 2,057.4	\$ 4,378.1	\$ 4,757.9	\$ 4,709.3
Series B @ 66% of Equity	(3,500.0)	4,445.0	9,458.8	10,279.3	10,174.5	40,697.9
NPV @ 30%	\$ 26.959		Series A MIRR			77%
NPV @ 25%	\$ 35.331		Series B MIRR			87%

Case 2 Year	2010	2011	2012	2013	2014	2015
Income Statement						
Revenue	\$ -	\$ 15,000.0	\$ 30,000.0	\$ 30,000.0	\$ 30,000.0	\$ 30,000.0
Cost of Sales	(172.6)	(643.6)	(988.6)	(988.6)	(988.6)	(988.6)
Operating, SG&A Expenses	(292.0)	(2,380.0)	(3,730.0)	(3,730.0)	(3,730.0)	(3,730.0)
Depreciation	-	(141.7)	(762.0)	(992.6)	(595.5)	(595.5)
Pre-Tax Income	(464.6)	11,834.7	24,519.4	24,288.8	24,685.9	24,685.9
Tax @ 40%	185.8	(4,733.9)	(9,807.8)	(9,715.5)	(9,874.3)	(9,874.3)
Net Income	\$ (278.8)	\$ 7,100.8	\$ 14,711.6	\$ 14,573.3	\$ 14,811.5	\$ 14,811.5
Cash Flow Statement						
Cash From Operating Activities	\$ -	\$ 141.7	\$ 762.0	\$ 992.6	\$ 595.5	\$ 595.5
Plus: Depreciation						
Changes in Working Capital						
(Increase)/Decrease in A/R	\$ -	\$ (1,232.9)	\$ (1,232.9)	\$ -	\$ -	\$ -
(Increase)/Decrease in Inventory	-	(28.4)	(28.4)	-	-	-
Increase/(Decrease) in A/P	24.5	111.0	111.0	-	-	-
(Increase)/Decrease in C/R	(73.0)	(184.5)	-	-	-	-
Total Change in Working Capital	\$ (48.5)	\$ (1,334.8)	\$ (1,150.3)	\$ -	\$ -	\$ -
Cash From Investing Activities	(708.3)	(2,676.8)	-	-	-	-
(Purchase)/Selling of Equipment						
Cash From Financing Activities	1,200.0	3,500.0	-	-	-	-
Issuance of Common Stock	\$ 164.5	\$ 6,730.9	\$ 14,323.4	\$ 15,565.9	\$ 15,407.1	\$ 29,102.2 (Terminal Value)
Free Cash Flow						\$ 32,740.0
% of Design Capacity	0%	50%	100%	100%	100%	100%
Investment Divided Free Cash Flows						
Series A @ 31% of Equity	\$ (1,200.0)	\$ 148.0	\$ 2,057.4	\$ 4,378.1	\$ 4,757.9	\$ 4,709.3
Series B @ 66% of Equity	(3,500.0)	4,445.0	9,458.8	10,279.3	10,174.5	21,620.8
NPV @ 30%	\$ 22,349	Series A MIRR				69%
NPV @ 25%	\$ 27,758	Series B MIRR				76%

Appendix F: Financial Pro Forma

Case 3 Year	2010	2011	2012	2013	2014	2015
<u>Income Statement</u>						
Revenue	\$ -	\$ 15,000.0	\$ 30,000.0	\$ 30,000.0	\$ 30,000.0	
Cost of Sales	(172.6)	(643.6)	(988.6)	(988.6)	(988.6)	
Operating, SG&A Expenses	(292.0)	(2,380.0)	(3,730.0)	(3,730.0)	(3,730.0)	
Depreciation	-	(141.7)	(762.0)	(992.6)	(595.5)	
Pre-Tax Income	(464.6)	11,834.7	24,519.4	24,288.8	24,685.9	
Tax @ 40%	185.8	(4,733.9)	(9,807.8)	(9,715.5)	(9,874.3)	
Net Income	\$ (278.8)	\$ 7,100.8	\$ 14,711.6	\$ 14,573.3	\$ 14,811.5	
<u>Cash Flow Statement</u>						
Cash From Operating Activities	\$ -	\$ 141.7	\$ 762.0	\$ 992.6	\$ 595.5	
Plus: Depreciation						
Changes in Working Capital						
(Increase)/Decrease in A/R	\$ -	\$ (1,232.9)	\$ (1,232.9)	\$ -	\$ -	
(Increase)/Decrease in Inventory	-	(28.4)	(28.4)	-	-	
Increase/(Decrease) in A/P	24.5	111.0	111.0	-	-	
(Increase)/Decrease in C/R	(73.0)	(184.5)	-	-	-	
Total Change in Working Capital	\$ (48.5)	\$ (1,334.8)	\$ (1,150.3)	\$ -	\$ -	
Cash From Investing Activities	(708.3)	(2,676.8)	-	-	-	
Cash From Financing Activities	1,200.0	3,500.0	-	-	-	
Issuance of Common Stock	\$ 164.5	\$ 6,730.9	\$ 14,323.4	\$ 15,565.9	\$ 15,407.1	\$ 1,692.6 (Terminal Value)
Free Cash Flow						
% of Design Capacity	0%	50%	100%	100%	100%	100%
<u>Investment Divided Free Cash Flows</u>						
Series A @ 31% of Equity	\$ (1,200.0)	\$ 148.0	\$ 2,057.4	\$ 4,378.1	\$ 4,757.9	\$ 4,709.3
Series B @ 66% of Equity	(3,500.0)	4,445.0	9,458.8	10,279.3	10,174.5	1,117.7
NPV @ 30%	\$ 16,670		Series A MIRR			58%
NPV @ 25%	\$ 19,619		Series B MIRR			62%

<u>Case 4</u> Year	2010	2011	2012	2013	2014	2015
<u>Income Statement</u>						
Revenue	\$ -	\$ 9,000.0	\$ 21,000.0	\$ 30,000.0	\$ 30,000.0	
Cost of Sales	(172.6)	(505.6)	(781.6)	(988.6)	(988.6)	
Operating, SG&A Expenses	(292.0)	(1,840.0)	(2,920.0)	(3,730.0)	(3,730.0)	
Depreciation	-	(141.7)	(762.0)	(992.6)	(595.5)	
Pre-Tax Income	(464.6)	6,512.7	16,536.4	24,288.8	24,685.9	
Tax @ 40%	185.8	(2,605.1)	(6,614.6)	(9,715.5)	(9,874.3)	
Net Income	\$ (278.8)	\$ 3,907.6	\$ 9,921.8	\$ 14,573.3	\$ 14,811.5	
<u>Cash Flow Statement</u>						
Cash From Operating Activities	\$ -	\$ 141.7	\$ 762.0	\$ 992.6	\$ 595.5	
Plus: Depreciation						
Changes in Working Capital						
(Increase)/Decrease in A/R	\$ -	\$ (739.7)	\$ (986.3)	\$ (739.7)	\$ -	
(Increase)/Decrease in Inventory	-	(17.0)	(22.7)	(17.0)	-	
Increase/(Decrease) in A/P	24.5	66.6	88.8	66.6	-	
(Increase)/Decrease in C/R	(73.0)	(184.5)	-	-	-	
Total Change in Working Capital	\$ (48.5)	\$ (874.7)	\$ (920.2)	\$ (690.2)	\$ -	
Cash From Investing Activities	(708.3)	(2,676.8)	-	-	-	
(Purchase)/Selling of Equipment						
Cash From Financing Activities						
Issuance of Common Stock	1,200.0	3,500.0	-	-	-	\$ 51,356.9 (Terminal Value)
Free Cash Flow	\$ 164.5	\$ 3,997.8	\$ 9,763.6	\$ 14,875.7	\$ 15,407.1	\$ 61,628.2
% of Design Capacity	0%	30%	70%	100%	100%	100%
<u>Investment Divided Free Cash Flows</u>						
Series A @ 31% of Equity	\$ (1,200.0)	\$ 148.0	\$ 1,222.0	\$ 2,984.4	\$ 4,546.9	\$ 4,709.3
Series B @ 66% of Equity	(3,500.0)	2,640.1	6,447.7	9,823.6	10,174.5	40,697.9
NPV @ 30%	\$ 23,025	Series A MIRR				75%
NPV @ 25%	\$ 30,964	Series B MIRR				84%

Appendix F: Financial Pro Forma

Case 5 Year	2010	2011	2012	2013	2014	2015
<u>Income Statement</u>						
Revenue	\$ -	\$ 9,000.0	\$ 21,000.0	\$ 30,000.0	\$ 30,000.0	\$ 30,000.0
Cost of Sales	(172.6)	(505.6)	(781.6)	(988.6)	(988.6)	(988.6)
Operating, SG&A Expenses	(292.0)	(1,840.0)	(2,920.0)	(3,730.0)	(3,730.0)	(3,730.0)
Depreciation	-	(141.7)	(762.0)	(992.6)	(595.5)	(595.5)
Pre-Tax Income	(464.6)	6,512.7	16,536.4	24,288.8	24,685.9	24,685.9
Tax @ 40%	185.8	(2,605.1)	(6,614.6)	(9,715.5)	(9,874.3)	(9,874.3)
Net Income	\$ (278.8)	\$ 3,907.6	\$ 9,921.8	\$ 14,573.3	\$ 14,811.5	\$ 14,811.5
<u>Cash Flow Statement</u>						
Cash From Operating Activities	\$ -	\$ 141.7	\$ 762.0	\$ 992.6	\$ 595.5	\$ 595.5
Plus: Depreciation	\$ -	\$ (739.7)	\$ (986.3)	\$ (739.7)	\$ -	\$ -
Changes in Working Capital	\$ -	(17.0)	(22.7)	(17.0)	-	-
(Increase)/Decrease in A/R	-	66.6	88.8	66.6	-	-
(Increase)/Decrease in Inventory	24.5	-	-	-	-	-
Increase/(Decrease) in A/P	(73.0)	(184.5)	-	-	-	-
(Increase)/Decrease in C/R	\$ (48.5)	\$ (874.7)	\$ (920.2)	\$ (690.2)	\$ -	\$ -
Total Change in Working Capital	(708.3)	(2,676.8)	-	-	-	-
Cash From Investing Activities	1,200.0	3,500.0	-	-	-	-
(Purchase)/Selling of Equipment	\$ 164.5	\$ 3,997.8	\$ 9,763.6	\$ 14,875.7	\$ 15,407.1	\$ 29,102.2 (Terminal Value)
Cash From Financing Activities	0%	30%	70%	100%	100%	100%
Issuance of Common Stock	\$ 164.5	\$ 3,997.8	\$ 9,763.6	\$ 14,875.7	\$ 15,407.1	\$ 32,740.0
Free Cash Flow	\$ 164.5	\$ 3,997.8	\$ 9,763.6	\$ 14,875.7	\$ 15,407.1	\$ 32,740.0
% of Design Capacity	0%	30%	70%	100%	100%	100%
<u>Investment Divided Free Cash Flows</u>						
Series A @ 31% of Equity	\$ (1,200.0)	\$ 148.0	\$ 1,222.0	\$ 2,984.4	\$ 4,546.9	\$ 4,709.3
Series B @ 66% of Equity	(3,500.0)	2,640.1	6,447.7	9,823.6	10,174.5	21,620.8
NPV @ 30%	\$ 18,414	Series A MIRR		66%		
NPV @ 25%	\$ 23,392	Series B MIRR		73%		

Case 6 Year	2010	2011	2012	2013	2014	2015
<u>Income Statement</u>						
Revenue	\$ -	\$ 9,000.0	\$ 21,000.0	\$ 30,000.0	\$ 30,000.0	
Cost of Sales	(172.6)	(505.6)	(781.6)	(988.6)	(988.6)	
Operating, SG&A Expenses	(292.0)	(1,840.0)	(2,920.0)	(3,730.0)	(3,730.0)	
Depreciation	-	(141.7)	(762.0)	(992.6)	(595.5)	
Pre-Tax Income	(464.6)	6,512.7	16,536.4	24,288.8	24,685.9	
Tax @ 40%	185.8	(2,605.1)	(6,614.6)	(9,715.5)	(9,874.3)	
Net Income	\$ (278.8)	\$ 3,907.6	\$ 9,921.8	\$ 14,573.3	\$ 14,811.5	
<u>Cash Flow Statement</u>						
Cash From Operating Activities	\$ -	\$ 141.7	\$ 762.0	\$ 992.6	\$ 595.5	
Plus: Depreciation						
Changes in Working Capital						
(Increase)/Decrease in A/R	\$ -	\$ (739.7)	\$ (986.3)	\$ (739.7)	\$ -	
(Increase)/Decrease in Inventory	-	(17.0)	(22.7)	(17.0)	-	
Increase/(Decrease) in A/P	24.5	66.6	88.8	66.6	-	
(Increase)/Decrease in C/R	(73.0)	(184.5)	-	-	-	
Total Change in Working Capital	\$ (48.5)	\$ (874.7)	\$ (920.2)	\$ (690.2)	\$ -	
Cash From Investing Activities	(708.3)	(2,676.8)	-	-	-	
(Purchase)/Selling of Equipment						
Cash From Financing Activities	1,200.0	3,500.0	-	-	-	\$ 1,692.6 (Terminal Value)
Issuance of Common Stock						
Free Cash Flow	\$ 164.5	\$ 3,997.8	\$ 9,763.6	\$ 14,875.7	\$ 15,407.1	\$ 1,692.6 Value
% of Design Capacity	0%	30%	70%	100%	100%	100%
<u>Investment Divided Free Cash Flows</u>						
Series A @ 31% of Equity	\$ (1,200.0)	\$ 148.0	\$ 1,222.0	\$ 2,984.4	\$ 4,546.9	\$ 4,709.3
Series B @ 66% of Equity	(3,500.0)	2,640.1	6,447.7	9,823.6	10,174.5	1,117.7
NPV @ 30%	\$ 12,736					
NPV @ 25%	\$ 15,253					
			Series A MIRR			53%
			Series B MIRR			57%

Appendix F: Financial Pro Forma

Case Z Year	2010	2011	2012	2013	2014	2015	2016
<u>Income Statement</u>							
Revenue	\$ -	\$ -	\$ 15,000.0	\$ 30,000.0	\$ 30,000.0	\$ 30,000.0	
Cost of Sales	(172.6)	(172.6)	(643.6)	(988.6)	(988.6)	(988.6)	
Operating, SG&A Expenses	(292.0)	(292.0)	(2,380.0)	(3,730.0)	(3,730.0)	(3,730.0)	
Depreciation	-	(141.7)	(226.7)	(671.4)	(938.2)	(595.5)	
Pre-Tax Income	(464.6)	(606.3)	11,749.7	24,610.0	24,343.2	24,685.9	
Tax @ 40%	185.8	242.5	(4,699.9)	(9,844.0)	(9,737.3)	(9,874.3)	
Net Income	\$ (278.8)	\$ (363.8)	\$ 7,049.8	\$ 14,766.0	\$ 14,605.9	\$ 14,811.5	
<u>Cash Flow Statement</u>							
Cash From Operating Activities	\$ -	\$ 141.7	\$ 226.7	\$ 671.4	\$ 938.2	\$ 595.5	
Plus: Depreciation	\$ -	\$ -	\$ (1,232.9)	\$ (1,232.9)	\$ -	\$ -	
Changes in Working Capital	-	-	(28.4)	(28.4)	-	-	
(Increase)/Decrease in A/R	24.5	-	111.0	111.0	-	-	
(Increase)/Decrease in Inventory	(73.0)	-	(184.5)	-	-	-	
(Increase)/Decrease in A/P	\$ (48.5)	\$ -	\$ (1,334.8)	\$ (1,150.3)	\$ -	\$ -	
(Increase)/Decrease in C/R	(708.3)	-	(2,676.8)	-	-	-	
Total Change in Working Capital	1,200.0	-	3,500.0	-	-	-	
Cash From Investing Activities	\$ 164.5	\$ (222.1)	\$ 6,764.9	\$ 14,287.1	\$ 15,544.1	\$ 15,407.1	\$ 51,356.9 (Terminal Value)
(Purchase)/Selling of Equipment							
Cash From Financing Activities							
Issuance of Common Stock							
Free Cash Flow	0%	0%	50%	100%	100%	100%	100%
% of Design Capacity							
<u>Investment Divided Free Cash Flows</u>							
Series A @ 39% of Equity	\$ (1,200.0)	\$ 148.0	\$ (199.9)	\$ 2,651.4	\$ 5,599.6	\$ 6,092.3	\$ 6,038.6
Series B @ 56% of Equity	(3,500.0)	3,818.9	8,065.3	8,774.9	8,697.5	34,790.1	\$ 24,154.3
NPV @ 30%	\$ 26.733						58%
NPV @ 25%	\$ 35.103						81%

Case 8 Year	2010	2011	2012	2013	2014	2015	2016
<u>Income Statement</u>							
Revenue	\$ -	\$ -	\$ 15,000.0	\$ 30,000.0	\$ 30,000.0	\$ 30,000.0	\$ 30,000.0
Cost of Sales	(172.6)	(172.6)	(643.6)	(988.6)	(988.6)	(988.6)	(988.6)
Operating, SG&A Expenses	(292.0)	(292.0)	(2,380.0)	(3,730.0)	(3,730.0)	(3,730.0)	(3,730.0)
Depreciation	-	(141.7)	(226.7)	(671.4)	(938.2)	(595.5)	(595.5)
Pre-Tax Income	(464.6)	(606.3)	11,749.7	24,610.0	24,343.2	24,685.9	24,685.9
Tax @ 40%	185.8	242.5	(4,699.9)	(9,844.0)	(9,737.3)	(9,874.3)	(9,874.3)
Net Income	\$ (278.8)	\$ (363.8)	\$ 7,049.8	\$ 14,766.0	\$ 14,605.9	\$ 14,811.5	\$ 14,811.5
<u>Cash Flow Statement</u>							
Cash From Operating Activities	\$ -	\$ 141.7	\$ 226.7	\$ 671.4	\$ 938.2	\$ 595.5	\$ 595.5
Plus: Depreciation							
Changes in Working Capital							
(Increase)/Decrease in A/R	\$ -	\$ -	\$ (1,232.9)	\$ (1,232.9)	\$ -	\$ -	\$ -
(Increase)/Decrease in Inventory	-	-	(28.4)	(28.4)	-	-	-
Increase/(Decrease) in A/P	24.5	-	111.0	111.0	-	-	-
(Increase)/Decrease in C/R	(73.0)	-	(184.5)	-	-	-	-
Total Change in Working Capital	\$ (48.5)	\$ -	\$ (1,334.8)	\$ (1,150.3)	\$ -	\$ -	\$ -
Cash From Investing Activities	(708.3)	-	(2,676.8)	-	-	-	-
(Purchase)/Selling of Equipment							
Cash From Financing Activities	1,200.0	-	3,500.0	-	-	-	-
Issuance of Common Stock	\$ 164.5	\$ (222.1)	\$ 6,764.9	\$ 14,287.1	\$ 15,544.1	\$ 15,407.1	\$ 29,102.2 (Terminal Value)
Free Cash Flow	\$ 164.5	\$ (222.1)	\$ 6,764.9	\$ 14,287.1	\$ 15,544.1	\$ 15,407.1	\$ 32,740.0
% of Design Capacity	0%	0%	50%	100%	100%	100%	100%
<u>Investment Divided Free Cash Flows</u>							
Series A @ 39% of Equity	\$ (1,200.0)	\$ 148.0	\$ (199.9)	\$ 2,651.4	\$ 5,599.6	\$ 6,092.3	\$ 12,832.0
Series B @ 56% of Equity	(3,500.0)	3,818.9	8,065.3	8,774.9	8,697.5	18,482.3	
NPV @ 30%	\$ 22,122		Series A MIRR		58%		
NPV @ 25%	\$ 27,530		Series B MIRR		71%		

Appendix F: Financial Pro Forma

Case 9 Year	2010	2011	2012	2013	2014	2015	2016
Income Statement							
Revenue	\$ -	\$ -	\$ 15,000.0	\$ 30,000.0	\$ 30,000.0	\$ 30,000.0	\$ 30,000.0
Cost of Sales	(172.6)	(172.6)	(643.6)	(988.6)	(988.6)	(988.6)	(988.6)
Operating, SG&A Expenses	(292.0)	(292.0)	(2,380.0)	(3,730.0)	(3,730.0)	(3,730.0)	(3,730.0)
Depreciation	-	(141.7)	(226.7)	(671.4)	(938.2)	(595.5)	(595.5)
Pre-Tax Income	(464.6)	(606.3)	11,749.7	24,610.0	24,343.2	24,685.9	24,685.9
Tax @ 40%	185.8	242.5	(4,699.9)	(9,844.0)	(9,737.3)	(9,874.3)	(9,874.3)
Net Income	\$ (278.8)	\$ (363.8)	\$ 7,049.8	\$ 14,766.0	\$ 14,605.9	\$ 14,811.5	\$ 14,811.5
Cash Flow Statement							
Cash From Operating Activities	\$ -	\$ 141.7	\$ 226.7	\$ 671.4	\$ 938.2	\$ 595.5	\$ 595.5
Plus: Depreciation	\$ -	\$ -	\$ (1,232.9)	\$ (1,232.9)	\$ -	\$ -	\$ -
Changes in Working Capital	\$ -	\$ -	(28.4)	(28.4)	-	-	-
(Increase)/Decrease in A/R	-	-	111.0	111.0	-	-	-
(Increase)/Decrease in Inventory	24.5	-	-	-	-	-	-
Increase/(Decrease) in A/P	(73.0)	-	(184.5)	-	-	-	-
(Increase)/Decrease in C/R	\$ (48.5)	\$ -	\$ (1,334.8)	\$ (1,150.3)	\$ -	\$ -	\$ -
Total Change in Working Capital	(708.3)	-	(2,676.8)	-	-	-	-
Cash From Investing Activities	1,200.0	-	3,500.0	-	-	-	-
(Purchase)/Selling of Equipment	\$ 164.5	\$ (222.1)	\$ 6,764.9	\$ 14,287.1	\$ 15,544.1	\$ 15,407.1	\$ 15,407.1
Cash From Financing Activities	0%	0%	50%	100%	100%	100%	100%
Issuance of Common Stock	-	-	-	-	-	-	-
Free Cash Flow	\$ 164.5	\$ (222.1)	\$ 6,764.9	\$ 14,287.1	\$ 15,544.1	\$ 15,407.1	\$ 15,407.1
% of Design Capacity	0%	0%	50%	100%	100%	100%	100%
Investment Divided Free Cash Flows							
Series A @ 39% of Equity	\$ (1,200.0)	\$ 148.0	\$ (199.9)	\$ 2,651.4	\$ 5,599.6	\$ 6,092.3	\$ 6,038.6
Series B @ 56% of Equity	(3,500.0)	3,818.9	8,065.3	8,774.9	8,697.5	955.5	\$ 663.4
NPV @ 30%	\$ 16,444		Series A MIRR				58%
NPV @ 25%	\$ 19,391		Series B MIRR				57%
							\$ 1,692.6 (Terminal Value)
							\$ 1,692.6 Value

Appendix F: Financial Pro Forma

Case 11	2010	2011	2012	2013	2014	2015	2016
Year							
<u>Income Statement</u>							
Revenue	\$ -	\$ -	\$ 9,000.0	\$ 21,000.0	\$ 30,000.0	\$ 30,000.0	
Cost of Sales	(172.6)	(172.6)	(505.6)	(781.6)	(988.6)	(988.6)	
Operating, SG&A Expenses	(292.0)	(292.0)	(1,840.0)	(2,920.0)	(3,730.0)	(3,730.0)	
Depreciation	-	(141.7)	(226.7)	(671.4)	(938.2)	(595.5)	
Pre-Tax Income	(464.6)	(606.3)	6,427.7	16,627.0	24,343.2	24,685.9	
Tax @ 40%	185.8	242.5	(2,571.1)	(6,650.8)	(9,737.3)	(9,874.3)	
Net Income	\$ (278.8)	\$ (363.8)	\$ 3,856.6	\$ 9,976.2	\$ 14,605.9	\$ 14,811.5	
<u>Cash Flow Statement</u>							
Cash From Operating Activities	\$ -	\$ 141.7	\$ 226.7	\$ 671.4	\$ 938.2	\$ 595.5	
Plus: Depreciation							
Changes in Working Capital							
(Increase)/Decrease in A/R	\$ -	\$ -	\$ (739.7)	\$ (986.3)	\$ (739.7)	\$ -	
(Increase)/Decrease in Inventory	-	-	(17.0)	(22.7)	(17.0)	-	
Increase/(Decrease) in A/P	24.5	-	66.6	88.8	66.6	-	
(Increase)/Decrease in C/R	(73.0)	-	(184.5)	-	-	-	
Total Change in Working Capital	\$ (48.5)	\$ -	\$ (874.7)	\$ (920.2)	\$ (690.2)	\$ -	
Cash From Investing Activities	(708.3)	-	(2,676.8)	-	-	-	
(Purchase)/Selling of Equipment							
Cash From Financing Activities	1,200.0	-	3,500.0	-	-	-	
Issuance of Common Stock	\$ 164.5	\$ (222.1)	\$ 4,031.8	\$ 9,727.4	\$ 14,853.9	\$ 15,407.1	
Free Cash Flow	0%	0%	30%	70%	100%	100%	
% of Design Capacity							
							\$ 29,102.2 (Terminal Value)
							\$ 32,740.0
<u>Investment Divided Free Cash Flows</u>							
Series A @ 39% of Equity	\$ (1,200.0)	\$ 148.0	\$ (199.9)	\$ 1,580.2	\$ 3,812.5	\$ 5,821.8	\$ 12,832.0
Series B @ 56% of Equity	(3,500.0)	2,276.0	5,491.3	8,385.3	8,697.5	18,482.3	
NPV @ 30%	\$ 18,188						54%
NPV @ 25%	\$ 23,164						67%
							Series A MIRR
							Series B MIRR

Case 12 Year	2010	2011	2012	2013	2014	2015	2016
Income Statement							
Revenue	\$ -	\$ -	\$ 9,000.0	\$ 21,000.0	\$ 30,000.0	\$ 30,000.0	
Cost of Sales	(172.6)	(172.6)	(505.6)	(781.6)	(988.6)	(988.6)	
Operating, SG&A Expenses	(292.0)	(292.0)	(1,840.0)	(2,920.0)	(3,730.0)	(3,730.0)	
Depreciation	-	(141.7)	(226.7)	(671.4)	(938.2)	(595.5)	
Pre-Tax Income	(464.6)	(606.3)	6,427.7	16,627.0	24,343.2	24,685.9	
Tax @ 40%	185.8	242.5	(2,571.1)	(6,650.8)	(9,737.3)	(9,874.3)	
Net Income	\$ (278.8)	\$ (363.8)	\$ 3,856.6	\$ 9,976.2	\$ 14,605.9	\$ 14,811.5	
Cash Flow Statement							
Cash From Operating Activities	\$ -	\$ 141.7	\$ 226.7	\$ 671.4	\$ 938.2	\$ 595.5	
Plus: Depreciation	\$ -	\$ -	\$ (739.7)	\$ (986.3)	\$ (739.7)	\$ -	
Changes in Working Capital	\$ -	\$ -	(17.0)	(22.7)	(17.0)	\$ -	
(Increase)/Decrease in A/R	24.5	-	66.6	88.8	66.6	-	
(Increase)/Decrease in A/P	(73.0)	-	(184.5)	-	-	-	
Total Change in Working Capital	\$ (48.5)	\$ -	\$ (874.7)	\$ (920.2)	\$ (690.2)	\$ -	
Cash From Investing Activities	(708.3)	-	(2,676.8)	-	-	-	
(Purchase)/Selling of Equipment	1,200.0	-	3,500.0	-	-	-	
Cash From Financing Activities	\$ 164.5	\$ (222.1)	\$ 4,031.8	\$ 9,727.4	\$ 14,853.9	\$ 15,407.1	
Issuance of Common Stock	0%	0%	30%	70%	100%	100%	\$ 1,692.6 (Terminal Value)
Free Cash Flow							\$ 1,692.6 Value
% of Design Capacity							100%
Investment Divided Free Cash Flows							
Series A @ 39% of Equity	\$ (1,200.0)	\$ 148.0	\$ (199.9)	\$ 1,580.2	\$ 3,812.5	\$ 5,821.8	\$ 6,038.6
Series B @ 56% of Equity	(3,500.0)	2,276.0	5,491.3	8,385.3	8,697.5	955.5	
NPV @ 30%	\$ 12,510		Series A MIRR		54%		
NPV @ 25%	\$ 15,025		Series B MIRR		52%		

Appendix F: Financial Pro Forma

<u>Growth Case</u> Year	2010	2011	2012	2013	2014	2015
<u>Income Statement</u>						
Revenue	\$ -	\$ 15,000.0	\$ 30,000.0	\$ 45,000.0	\$ 67,500.0	
Cost of Sales	(172.6)	(643.6)	(988.6)	(1,333.6)	(1,851.1)	
Operating, SG&A Expenses	(292.0)	(2,380.0)	(3,730.0)	(5,595.0)	(8,392.5)	
Depreciation	-	(141.7)	(762.0)	(992.6)	(595.5)	
Pre-Tax Income	(464.6)	11,834.7	24,519.4	37,078.8	56,660.9	
Tax @ 40%	185.8	(4,733.9)	(9,807.8)	(14,831.5)	(22,664.3)	
Net Income	\$ (278.8)	\$ 7,100.8	\$ 14,711.6	\$ 22,247.3	\$ 33,996.5	
<u>Cash Flow Statement</u>						
Cash From Operating Activities	\$ -	\$ 141.7	\$ 762.0	\$ 992.6	\$ 595.5	
Plus: Depreciation						
Changes in Working Capital						
(Increase)/Decrease in A/R	\$ -	\$ (1,232.9)	\$ (1,232.9)	\$ (1,232.9)	\$ (1,849.3)	
(Increase)/Decrease in Inventory	-	(28.4)	(28.4)	(28.4)	(42.5)	
Increase/(Decrease) in A/P	24.5	111.0	111.0	111.0	166.4	
(Increase)/Decrease in C/R	(73.0)	(184.5)	-	-	-	
Total Change in Working Capital	\$ (48.5)	\$ (1,334.8)	\$ (1,150.3)	\$ (1,150.3)	\$ (1,725.4)	
Cash From Investing Activities	(708.3)	(2,676.8)	-	-	-	
(Purchase)/Selling of Equipment						
Cash From Financing Activities	1,200.0	3,500.0	-	-	-	
Issuance of Common Stock	\$ 164.5	\$ 6,730.9	\$ 14,323.4	\$ 22,089.6	\$ 32,866.6	\$ 109,555.5 (Terminal Value)
Free Cash Flow	\$ 164.5	\$ 6,730.9	\$ 14,323.4	\$ 22,089.6	\$ 32,866.6	\$ 131,466.6
% of Design Capacity	0%	50%	100%	150%	225%	225%
<u>Investment Divided Free Cash Flows</u>						
Series A @ 31% of Equity	\$ (1,200.0)	\$ 148.0	\$ 2,057.4	\$ 4,378.1	\$ 6,751.9	\$ 10,046.0
Series B @ 66% of Equity	(3,500.0)	4,445.0	9,458.8	14,587.5	21,704.4	86,817.6
NPV @ 30%	\$ 46,003	Series A MIRR				95%
NPV @ 25%	\$ 62,032	Series B MIRR				110%

References

- 1 Shastry, B. S. (2006). **Pharmacogenetics and the concept of individualized medicine.** *Pharmacogenomics J.* 6 (1): 16–21.
- 2 Holbrook, A. M., Pereira, J. A., Labiris, R., *et al.* (2005). **Systematic overview of warfarin and its drug and food interactions.** *Arch. Intern. Med.* 165(10): 1095–106.
- 3 Schwarz, U. I. (November 2003). **Clinical relevance of genetic polymorphisms in the human CYP2C9 gene.** *Eur. J. Clin. Invest.* 33(Suppl 2): 23–30.
- 4 Mansour, J. C., Schwarz, R. E. (August 2008). **Molecular mechanisms for individualized cancer care.** *J. Am. Coll. Surg.* 207(2): 250–8.
- 5 van't Veer, L. J., Bernards, R. (April 2008). **Enabling personalized cancer medicine through analysis of gene-expression patterns.** *Nature* 452(7187): 564–70.

References

- 6 Saglio, G., Morotti, A., Mattioli, G., *et al.* (December 2004). **Rational approaches to the design of therapeutics targeting molecular markers: the case of chronic myelogenous leukemia.** *Ann. N. Y. Acad. Sci.* 1028: 423–31.
- 7 National Cancer Institute (2002). **Genetic Testing for BRCA1 and BRCA2: It's Your Choice.** <http://www.cancer.gov/cancertopics/factsheet/risk/brca>
- 8 Hanis, C. L., *et al.* (1996). **A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2.** *Nature Genet.* 13: 161-166.
- 9 Lyssenko *et al.* (2008). **Clinical risk factors, DNA variants, and the development of type 2 diabetes.** *N Engl J Med* 359 (21): 2220–2232.
- 10 National Center for Biotechnology Information (2009). **SNP summary.** http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi
- 11 Pacific Biosciences, Inc. (2008). **Company Background.** 31 March 2009. http://www.pacificbiosciences.com/assets/files/pacbio_background.pdf
- 12 Pacific Biosciences, Inc. (2008). **Technology Background.** 31 March 2009. http://www.pacificbiosciences.com/assets/files/pacbio_technology_background.pdf
- 13 Eid, J., Turner, S., *et al.* (2009). **Real-Time DNA Sequencing from Single Polymerase Molecules.** *Science* 323: 133-138.
- 14 Nie and Zare (1997). **Optical detection of single molecules.** *Annu. Rev. Biophys. Biomol. Struct.*, 25: 567-596.
- 15 Levene M. J., *et al.* (2003). **Zero-mode waveguides for single-molecule analysis at high concentrations.** *Science* 299: 682–686.
- 16 Augustin, M. A., *et al.* (2001). **Progress towards single-molecule sequencing: enzymatic synthesis of nucleotide-specifically labeled DNA.** *J. Biotechnol.* 86: 289-301.
- 17 Illumina, Inc. (2008). **Genome Analyzer Specification Sheet.** 5 April 2009. http://www.illumina.com/downloads/GenomeAnalyzer_SpecSheet.pdf
- 18 Center for Genome Research and Biocomputing (2008). **DNA Sequencing with Solexa® Technology.** 5 April 2009. http://corelabs.cgrb.oregonstate.edu/files/illumina/SS_DNAsequencing.pdf
- 19 454 Sequencing (2009). **Products & Solutions.** 5 April 2009. <http://www.454.com/products-solutions/how-it-works/index.asp>

-
- 20 VisiGen Biotechnologies, Inc. (2008). **Real-Time DNA Sequencing**. 3 April 2009.
<http://www.abrf.org/Other/ABRFMeetings/ABRF2005/Hardin.pdf>
- 21 Blanco, L. and Salas, M. (1984). **Characterization and purification of a phage phi 29-encoded DNA polymerase required for the initiation of replication**. *PNAS USA*, 81: 5325-5329.
- 22 Blanco, L., *et al.* (1989). **Highly Efficient DNA Synthesis by the phage phi29 DNA Polymerase. Symmetrical mode of DNA replication**. *J. Biol. Chem.*, 264: 8935-8940.
- 23 Esteban, J. A., *et al.* (1993). **Fidelity of phi29 DNA polymerase**. *J. Biol. Chem.*, 268(4): 2719-2726.
- 24 Blanco, L., *et al.* (1996). **Relating structure to function in phi29 DNA polymerase**. *J. Biol. Chem.* 271: 8509-8512.
- 25 Soengas, M. S., Gutierrez, M. S. and Salas, M. (1995). **Helix-destabilizing Activity of ϕ 29 Single-stranded DNA Binding Protein: Effect on the Elongation Rate During Strand Displacement DNA Replication**. *J. Mol. Biol.* 253(4): 517-529.
- 26 QIAGEN (2009). **REPLI-g Mini & Midi Kits**.
<http://www1.qiagen.com/Products/GenomicDnaStabilizationPurification/replig/RepliGMiniMidiKits.aspx>
- 27 Dean, F. B., *et al.* (2002). **Comprehensive human genome amplification using multiple displacement amplification**. *PNAS. USA* 99: 5261.
- 28 <http://www1.qiagen.com/Products/GenomicDnaStabilizationPurification/replig/RepliGMiniMidiKits.aspx#Tabs=t1>
- 29 Fleischmann, R. D., *et al.* (1995). Whole-genome random sequencing and assembly of Haemophilus influenza Rd. *Science*.
- 30 Reyes, G. R., *et al.* (1991). **SISPA of complex DNA populations**. *Molecular and Cellular Probes*.
- 31 Hutchison, C. *et al.* (2005). **Cell-free cloning using ϕ 29 DNA polymerase**. *PNAS*.
- 32 Dean, F., *et al.* (2008). **Comprehensive human genome amplification using multiple displacement amplification**. *PNAS*.
- 33 Eid, J., Turner, S., *et al.* (2009). **Real-Time DNA Sequencing from Single Polymerase Molecules**. *Science* 323: 133-138.
- 34 Levene MJ, *et al.* (2003). **Zero-mode waveguides for single-molecule analysis at high concentrations**. *Science* 299: 682–686.
-

References

- 35 Liu Y, et al. (2004). **Biosensing based upon molecular confinement in metallic nanocavity arrays.** *Nanotechnology* 15: 1368–1374.
- 36 Xia YN, et al. (1996). **Shadowed sputtering of gold on V-shaped microtrenches etched in silicon and applications in microfabrication.** *Adv Mat* 8: 765–768.
- 37 Vargel, C. (2004). **Corrosion of Aluminium.** (Elsevier, Amsterdam).
- 38 Michel R., et al. (2002). **A novel approach to produce biologically relevant chemical patterns at the nanometer scale: Selective molecular assembly patterning combined with colloidal lithography.** *Langmuir* 18: 8580–8586.
- 39 Mutin P. H., et al. (2004). **Selective surface modification of SiO₂-TiO₂ supports with phosphonic acids.** *Chem Mater* 16: 5670–5675.
- 40 Ramsier, R. D., Henriksen, P. N., Gent, A. N. (1988). **Adsorption of phosphorus-acids on alumina.** *Surf Sci* 203: 72–88.
- 41 Korlach, J, Turner, S, et al. (2008). **Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures.** *PNAS* 105(4) 1176.
- 40 Korlach, J, Turner, S, et al. (2008). **Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures.** *PNAS* 105(4) 1176.
- 43 Laysan Bio (2009). **PEG Reagents.**
http://laysanbio.com/index.php?submenu=Products&src=gendocs&link=Products_new&category=Main
- 44 Hofmann, K., & Kiso, Y. (1976). **An approach to the targeted attachment of peptides and proteins to solid supports.** *PNAS USA.* 73: 3516-3518.
- 45 Green, N. (1963). **The Use of [14C] Biotin for Kinetic Studies and for Assay.** *Biochem J*, 89: 585-591
- 46 Eid, J., Turner, S., et al. (2009). **Real-Time DNA Sequencing from Single Polymerase Molecules.** *Science* 323: 133-138.
- 47 Korlach, J, Turner, S, et al. (2008). **Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures.** *PNAS* 105(4) 1176.
- 48 Eid, J., Turner, S., et al. (2009). **Real-Time DNA Sequencing from Single Polymerase Molecules.** *Science* 323: 133-138.
- 49 Kumar, S., et al. (2005). **Terminal phosphate labeled nucleotides : synthesis, applications and linker effect on incorporation by DNA polymerases.** *Nucleosides, Nucleotides, and Nucleic Acids* 24: 401-408.

-
- 50 Korlach, J. *et al.* (2008). **Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides.** *Nucleosides, Nucleotides and Nucleic Acids.* 27: 1072
- 51 Aitken, C.E., Marshall, R.A., Puglisi, J.D. (2008). **An oxygen scavenging system for improvement of dye stability in single-molecule fluorescence experiments.** *Biophys J* 94: 1826-1835.
- 52 Korlach, J. *et al.* (2008). **Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides.** *Nucleosides, Nucleotides and Nucleic Acids.* 27: 1072
- 53 Eid, J., Turner, S., *et al.* (2009). **Real-Time DNA Sequencing from Single Polymerase Molecules.** *Science* 323: 133-138.
- 54 Datta, K; LiCata, V; (2003). **Thermodynamics of the binding of Thermus aquaticus DNA polymerase to primed-template DNA.** *Nucleic Acid Research.* Vol. 31, No. 19 **5590-5597**
- 55 Robertson, *et al.* (2006). **Diffusion of isolated DNA molecules : dependent on length and topology.** *PNAS.*
- 56 Eid, J., Turner, S., *et al.* (2009). **Real-Time DNA Sequencing from Single Polymerase Molecules.** *Science* 323: 133-138.
- 57 Korlach, J. *et al.* (2008). **Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides.** *Nucleosides, Nucleotides and Nucleic Acids.* 27: 1072
- 58 Eid, J., Turner, S., *et al.* (2009). **Real-Time DNA Sequencing from Single Polymerase Molecules.** *Science* 323: 133-138.
- 59 Eid, J., Turner, S., *et al.* (2009). **Real-Time DNA Sequencing from Single Polymerase Molecules.** *Science* 323: 133-138.
- 60 Lundquist, P., *et al.* (2008). **Parallel confocal detection of single molecules in real time.** *Optics Letters* 33(9): 1026-1028.
- 61 **What is EMCCD Technology.** 29 March 2009. http://www.emccd.com/what_is_emccd/
- 62 Andor (2009). **iXonEM+897.** 15 February 2009. http://www.andor.com/scientific_cameras/ixon/models/default.aspx?iProductCodeID=3
- 63 Pawley, J. B., (1995). **Handbook of Biological Confocal Microscopy.** Springer. 363-364.
- 64 National Center for Biotechnology Information (2004). **Bioinformatics.** 13 April 2009. <http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html>
- 65 Pop, Mihai, *et al.* (2004) **Comparative genome assembly.** *Briefings in Bioinformatics,* 5(3): 237-248.

References

- 66 The International Human Genome Sequencing Consortium (2001). **Initial sequencing and analysis of the human genome.** *Nature* 409: 860-921.
- 67 Venter, J. C., *et al.* (2001). **The sequence of the human genome.** *Science* 291: 1304-1351.
- 68 Waterson, R. H., *et al.* (2002). **On the sequencing of the genome.** *PNAS* 99(6): 3712-3716.
- 69 Roach, J. C. (1995). **Random subcloning.** *Genome Reseach* 5: 464-473.
- 70 Studier, W. F. (1989). **A strategy for high-volume sequencing of cosmid DNAs: Random and directed priming with a library of oligonucleotides.** *PNAS* 86: 6917-3921.
- 71 Wang, D., *et al.* (2000). **Estimation of the mutation rate during error-prone polymerase chain reaction.** *J. Comp. Biol.* 7(1,2): 143-158.
- 72 Sahai and Khurshid (1992). **Confidence intervals for the mean of a Poisson distribution: A review.** *Biometrical Journal* 35(7): 857-867.
- 73 Peltola, H., *et al.* (1984). **SEQAID: A DNA sequence assembling program based on a mathematical model.** *Nucleic Acid Res.*, 12(1) 307-231.
- 74 Farrar, M. (2007). **Striped Smith-Waterman speeds database searches six times over other SIMD implementations.** *Bioinformatics* 23(2): 156-161.
- 75 The MathWorks, Inc. (2009). **Techniques for Improving Performance.** 24 March 2009.
http://www.mathworks.com/access/helpdesk/help/techdoc/index.html?access/helpdesk/help/techdoc/MATLAB_prog/f8-784135.html

