



Impacts of Genome Sequencing Technologies

Developing Educational Materials to Create Greater Public Awareness

May 1, 2012

An Interactive Qualifying Project submitted to the faculty of
Worcester Polytechnic Institute in partial fulfillment of the requirements for the
Bachelor of Science degree

Tuhina Bhattacharya

William M. Elmore

Xuyu Qian

Luyang Zhang

Advisor

Professor Zheyang Wu
Department of Mathematical Sciences

Authorship

This report was written as a collaboration of the project team members: Tuhina Bhattacharya, William Elmore, Xuyu Qian, and Luyang Zhang. While some sections were primarily written by one individual, each group member played a role in reading and revising all sections for clarity, as well as ensuring that the views presented in the report were the views of the group as a whole.

Acknowledgements

The team would like to thank the following individuals and departments for their assistance and support throughout the project.

- ❖ Our advisor Professor Zheyang Wu for his continuous guidance and support throughout the project
- ❖ Professors Carolina Ruiz, Elizabeth Ryder, Matthew Ward and John Wilkes for previewing our educational materials and providing their guidance.
- ❖ The undergraduate students and faculty at Worcester Polytechnic Institute who took the time to provide their input on our student and faculty surveys
- ❖ The Worcester Polytechnic Institute Mathematical Sciences Department for their financial assistance in order for the team to afford the necessary resources and equipment
- ❖ The Worcester Polytechnic Institute Bioinformatics and Computational Biology program for their interest and feedback on our educational materials.

Abstract

Educators often use online material to aid in student learning. After reviewing academic and corporate sources, the team determined the primary effects of the advances in genome sequencing technologies. Based on survey responses, it was determined that many people are not familiar with these effects. Consequently, the team developed a series of educational videos and supplemental assignments that provide an introduction to genome sequencing technologies, as well as their worldwide impacts on legislation, economy, forensics, health care, and biological research.

Table of Contents

AUTHORSHIP	2
ACKNOWLEDGEMENTS.....	3
ABSTRACT	4
TABLE OF TABLES	8
TABLE OF FIGURES	9
LIST OF ACRONYMS	10
EXECUTIVE SUMMARY	12
CHAPTER 1: INTRODUCTION	15
CHAPTER 2: BACKGROUND ON GENOME SEQUENCING TECHNOLOGIES	17
2.1 The Sanger Method – 1st Generation Genome Sequencing	17
2.2 Second Generation Sequencing Technologies	19
2.2.1 Roche 454 System	19
2.2.2 Sequencing Systems by Illumina, Inc.	21
2.2.3 Applied Biosystems SOLiD System	25
2.3 Third Generation Sequencing Technologies	28
2.3.1 Helicos BioSciences True Single Molecule Sequencing (tSMS)	28
2.3.2 Pacific Biosciences Single Molecule Real Time Sequencing (SMRT)	31
2.4 Additional Sequencing Methods	33
2.4.1 Direct Sequencing via TEM by ZS Genetics	33
2.4.2 Nanopore Sequencing.....	33
2.4.3 Electronic Scanning Tunneling Microscopy (STM)	34
2.4.4 NanoKnife Edge Probe Method by Revase	34
2.4.5 Sequencing by Synthesis (SBS) by Intelligent Biosystems	36
2.4.6 DNA Nano-ball Method.....	36
2.5 Computational Methods	37
2.5.1 Computational Challenges Posed by New Genome Sequencing Technologies	38
2.5.2 Data Processing and Management	38
2.5.3 Analyzing Sequence Data	39
2.6 Combining Sequencing Technologies	41
2.7 Projections	42

CHAPTER 3: IMPACTS AND APPLICATIONS OF GENOME SEQUENCING.....	44
3.1 Economy	44
3.2 Worldwide Legislation Regarding the Release of Genetic Information.....	48
3.2.1 Implications of Genome Ownership, Privacy, and Disclosure	48
3.2.2 Legislation in the United States	50
3.2.3 Legislation in Europe.....	51
3.2.4 Legislation in Australia	52
3.2.5 Legislation in Japan	53
3.2.6 Legislation in China	53
3.2.7 Legislation in Developing Countries.....	54
3.2.8 Cultural and Religious Barriers.....	54
3.3 Forensics	55
3.4 Medicine	57
3.4.1 Medical Diagnosis, Treatment, and Cancer Research.....	57
3.4.2 Personalized Drug Development	58
3.4.3 Disease Prevention and Vaccination.....	62
3.5 Advances in Genetic Research.....	62
3.5.1 Mutation Discovery and Defining Variability Across Human Genomes	63
3.5.2 Sequencing Clinical Isolates in Strain-to-Reference Comparisons	63
3.5.3 Enabling Metagenomics.....	64
3.5.4 Defining DNA–Protein Interactions and Regulatory Protein Binding.....	64
3.5.5 Exploring Chromatin Packaging and Noncoding RNAs.....	65
3.6 Evolution.....	65
3.7 Agriculture	67
CHAPTER 4: SURVEYING WPI STUDENTS AND FACULTY	70
4.1 Creating and Implementing the Student Survey.....	70
4.2 Results and Discussion of the Student Survey	71
4.2.1 Respondent Demographics	71
4.2.2 Stated Familiarity and Interest Ratings on the Impacts of Genome Sequencing Technologies	72
4.2.3 Student Responses to the Changes Brought By Advances in Sequencing Technologies	75
4.3 Creating and Implementing the Faculty Survey	79
4.4 Results and Analysis of the Faculty Survey	80
CHAPTER 5: DEVELOPING EDUCATIONAL MATERIALS	82
5.1 Creating the Preview Video and Obtaining Feedback.....	82
5.2 Creating and Posting the Final Set of Videos	83

5.3 Creating Supplemental Assignments	84
5.4 Future Usage of the Set of Videos and Supplemental Assignments in the Classroom.....	85
CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS	86
6.1 The Team’s Conclusions on the Impacts of Genome Sequencing Technologies	86
6.2 Future Recommendations	86
REFERENCES.....	87
APPENDIX I: STUDENT SURVEY.....	93
APPENDIX II: FACULTY SURVEY.....	96
APPENDIX III: FEEDBACK FORM FOR VIDEO PREVIEW	98
APPENDIX IV: SUPPLEMENTAL ASSIGNMENTS.....	99

Table of Tables

Table 1: Performance Statistics of the 454 GS FLX+System.....	21
Table 2: A Comparison of the Performance Statistics of the Illumina Genome Analyzer IIX and the HiSeq 2000.....	22
Table 3: A Comparison of the Performance of Recent Illumina Sequencing Systems	25
Table 4: A Comparison of Recent SOLiD Systems Based on Statistical Performance.....	28
Table 5: A Comparison of the Performance Statistics for Helicos tSMS.....	29
Table 6: A Comparison of the Performance Statistics for Pacific Biosciences SMRT	31
Table 7: The Cost of Amazon Elastic Compute Cloud.....	39
Table 8: Specifications of Bench Top Sequencers that are Available as of August, 2011	43
Table 9: Comparison of Employment Estimation (Jobs per Year) by Sector.....	47
Table 10: Genome Sizes and Gene Density in Humans as Compared with Other Organisms.....	59

Table of Figures

Figure 1: A Graphic Comparing a Standard Nucleotide (a) to a Dideoxynucleotide (b)	17
Figure 2: A Graphic Illustrating How the Chain Terminated DNA Fragments Produce a Sequence of Bands	18
Figure 3: A Graphic Illustrating Emulsion-based Clonal Amplification.....	20
Figure 4: A Graphical Representation of Illumina Library Preparation	22
Figure 5: A Graphic Illustrating Illumina Bridge Amplification.....	23
Figure 6: A Graphic Illustrating the Emittance of Fluorescence upon Laser Excitation.....	24
Figure 7: A Graphic Illustrating the Preparation of Clonal Bead Populations via Microreactors	26
Figure 8: A Graphic of a Di-base Probe Emitting Fluorescence as Used in the SOLiD System .	26
Figure 9: A Graphic Showing the Matrix Used to Decode the 2-base Encoding Used in the SOLiD System	27
Figure 10: A Graphic Illustrating the Recording of Single Nucleotide Base Additions in tSMS	30
Figure 11: A Graphic Illustrating DNA Polymerase Adding a Nucleotide and Then Cleaving the Fluorescent Label During SMRT Sequencing.....	32
Figure 12: A Graphic Illustrating the Light Emittance Captured in the ZMW Nano-photonic Chamber During SMRT Sequencing.....	32
Figure 13: A Graphic Showing an Overview of Sequencing Through Scanning Tunneling Microscopy	34
Figure 14: A Graphic Showing the Relationship Between the Probes and the Substrate in OmniMoRA Sequencing.....	35
Figure 15: A Graphic Illustrating the Preparation of DNA Nano-balls.....	37
Figure 16: Structure of Forward and Backward Linkage Associated with the Human Genome Project	45
Figure 17: United States Federal Funding for Genome Related Research between 1988 and 2010	46
Figure 18: Comparison of Employment Estimation by Sector.....	47
Figure 19: Cost per Genome from September 2001 to September 2011	60
Figure 20: An Illustration of an Example of Metagenomic Studies	64
Figure 21: Adoption of Transgenic Crops in USA from 1996 to 2011	68
Figure 22: A Line Chart Comparing the Level of Familiarity on the Impacts of Sequencing Technologies of Biology Related Majors and Non-biology Related Majors	73
Figure 23: A Line Chart Comparing the Percent of Interest on the Impacts of Sequencing Technologies of Biology Related Majors and Non-biology Related Majors	74

List of Acronyms

AHEC	Australian Health Ethics Committee
AIDS	Acquired Immune Deficiency Syndrome
ALRC	Australian Law Reform Commission
ARG	Ancestral Recombination Graph
BLAST	Basic Local Alignment Search Tool
BB	Biology and Biotechnology
BCB	Bioinformatics and Computational Biology
BME	Biomedical Engineering
bp	Base Pair
CLUSTALW	Cluster Alignment Interface with Weights
ddNTP	Dideoxynucleotide-tri-phosphate
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleotide-tri-phosphate
GB	Gigabyte
GINA	Genetic Information Nondiscrimination Act
G8	Group of Eight
HGP	Human Genome Project
HIV	Human Immunodeficiency Virus
IRB	Instructional Review Board
MB	Megabyte
PCR	Polymerase Chain Reaction
RNA	Ribonucleic Acid

SBS	Sequencing by Synthesis
SMRT	Single Molecule Real Time
SNP	Single Nucleotide Polymorphism
ssDNA	Single Stranded Deoxyribonucleic Acid
STM	Scanning Tunneling Microscope
TEM	Transmission Electron Microscopy
tSMS	True Single Molecule Sequencing
UV	Ultraviolet
ZMW	Zero Mode Waveguide

Executive Summary

Genome sequencing technologies are becoming increasingly more advanced with higher throughputs and increased accuracy, and more applications for them are being discovered daily. While having a positive effect on the United States economy, these advances impact a variety of fields including forensics, disease diagnosis and treatment, vaccination, personalized drug development, genetic studies, evolutionary studies, and agriculture. However these positive impacts have also introduced some social issues such as the issues pertaining to ownership, privacy, and legislation related to the control of genetic information. The goal of this Interactive Qualifying Project was to research the effects genome sequencing technologies have had in each of these fields and then communicate them to a global audience.

Online resources can play a critical role in student education as they serve to broaden the boundaries of a course past the information from the instructor and textbook. In addition, the use of multimedia to aid in student comprehension and retention is becoming increasingly common. Thus the team generated online educational material in regards to genome sequencing technologies and their various impacts in other fields.

Furthermore, undergraduates at WPI were surveyed in regards to their interest, opinion, and familiarity with recent genome sequencing technologies and their various applications. WPI faculty were also surveyed, but in regards to their preferences for various aspects of our final deliverables as well as their interest in using our educational materials in their course(s). The information obtained from these two surveys enabled the team to make the educational materials more relevant and useful for the target demographic.

The student survey results indicates that those majoring in biology related fields have a higher amount of familiarity with genome sequencing technologies and their impacts on medicine, compared to students majoring in fields unrelated to biology. In regards to interest, students reported to be more interested in how advances in genome sequencing technologies have impacted cancer research and disease treatment. The impacts of these technologies on legislation and the economy received the lowest amount of interest, but the team determined that this was resultant of the fact that students reported to have the lowest amount of familiarity with the

impacts in these areas. Thus, the team resolved to still include these topics of lower interest in the final series of videos, hoping that the team's introduction to this field will initiate a higher amount of interest.

Even though the level of response from the faculty survey was low, the team was able to obtain a lot of valuable information from conducting this survey. Foremost is the fact that instructors reported to prefer an educational video broken into segments rather than a full length video, due to more flexibility in terms of showing the videos that are most relevant to their specific course. Some also noted that class time was limited, and thus having the materials available online would be more convenient as students could be directed to watch the material and complete the assignments outside of the classroom.

The primary component of the educational materials developed by the team was a series of online videos to educate people about the advances in genome sequencing technologies and the implications of these advances in fields such as law, forensics, medicine, and biological research. Most of the images and text used in these videos were created by the team members, and all of the video animation, script writing, and narration were also done by the team. The pieces of software utilized include Adobe Photoshop CS 5.1, which was used for image creation and adjustment, Adobe Flash CS 5.5, which was used for two-dimensional animation, and Adobe Premiere Pro CS 5.5, which was used to compile the animation sequences, music, narration, and subtitles together. Along with the creation of the visual components, a script was written for each segment of the video and then recorded. Approximately 35 minutes of film was made using this workflow.

Partly in response to the faculty survey and suggestions, the team also developed a short assignment for each segment of the final video. The answers to most of these questions are contained in the series of videos, while others initiate students to conduct further research on a specific area and then participate in a class debate and/or discussion.

The final set of videos was posted on YouTube in order to inform a larger audience. These videos can be accessed through the team's YouTube Channel:

<http://www.youtube.com/user/ImGenTechWPI/videos>. This includes a playlist of the 12 individual segments as well as a playlist with the material organized into three parts: 1st, 2nd, and 3rd Generation Genome Sequencing Technologies, Social Impacts of Genome Sequencing Technologies, and Impacts of Genome Sequencing Technologies Related to Medicine and Biological Research. All of the videos are set to public access, assuring that people all over the world can view the videos as long as their country has no regulations preventing them to view it. This will also enable educators to link to the video(s) and have their students watch it outside of the classroom. It was also requested of the team to send the final set of videos to the WPI Bioinformatics and Computational Biology program for possible future use. The description field underneath each video in the playlist direct instructors to email imgentech12@gmail.com if they wish to access the assignments related to the video.

A future extension of this project would be to assess the effectivity of these educational materials. This could be accomplished through the creation of a focus group of students to assess their understanding of the various impacts of genome sequencing technologies before and after viewing the team's final set of videos and completing some of the supplemental assignments. Another possible continuation of this project would be to create other types of educational materials to communicate the same information, allowing an even larger audience to benefit from the content. These materials could be but are not limited to interactive websites, computer games, and mobile applications.

Chapter 1: Introduction

The advances in genome sequencing technologies in the last few decades have affected a variety of fields including the economy, disease diagnosis and treatment, vaccination, personalized drug development, genetic studies, evolutionary studies, forensic studies, and agriculture. However these positive impacts have also introduced some social issues such as the issues pertaining to ownership, privacy, and legislation related to the control of genetic information.

It can be said that one cannot fully understand the impact a given technology can have on their lives without having a proficient understanding of the technology itself. As a result, in order to create greater public awareness on these issues, the team created a series of educational videos that serve as an introduction to the impact new genome sequencing technologies have had on these fields, while also providing an introduction to the primary genome sequencers playing a role in these industries. The team also conducted a survey in order to determine which topics undergraduate students were more familiar with, which topics they were interested in learning more about, and to also develop a sense of the opinions these students had in regards to the many impacts these technologies have on society. Furthermore, in response to suggestions made by faculty at Worcester Polytechnic Institute, the team also created a set of assignments to accompany these videos. If these components are implemented together, they will allow a student to gain an introduction to the technological, social, and medicinal aspects of rapidly developing sequencing technologies and how it affects their lives, while also being able to demonstrate this understanding to their course instructor through the completion of the assignments. All of the educational materials developed by the team were posted on the internet in order to reach a larger audience and have more of an impact with our project deliverables.

As this field progresses, it is expected that the rapid advances in sequencing technology will enable further development of other fields such as medicine, government, and agriculture. Next generation sequencing technologies have the following technological goals:

- ❖ Achieving a low cost per genome sequenced
- ❖ Achieving faster sequencing
- ❖ Reducing error rates

- ❖ Allowing for larger read lengths to facilitate assembly (for human genome sequencing) but also having the ability to have shorter read lengths (for the accurate sequencing of microbes)
- ❖ Creating a more automated technology in order to reduce the number of lab technicians needed to run the sequencer and interpret the data
- ❖ Creating flexible sequencers and software so that they can be interchanged, allowing for the combination of technologies to overcome their limitations
- ❖ Decreasing the time and complexity of DNA preparation
- ❖ Reducing the number of reagents and/or consumables involved

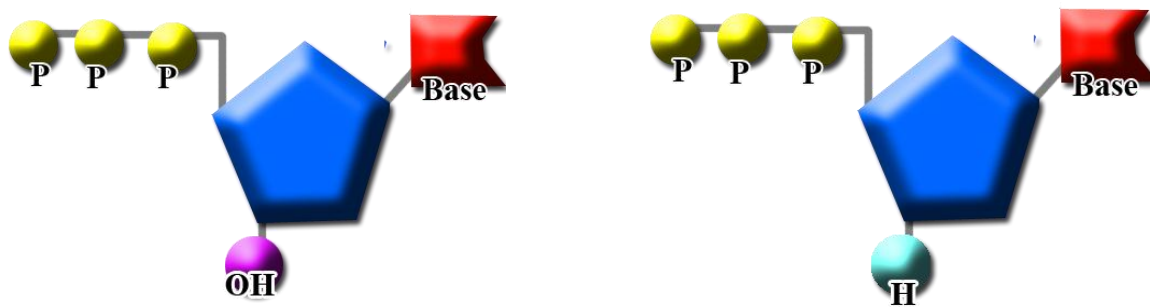
Chapter 2: Background on Genome Sequencing Technologies

This chapter includes information concerning the background of gene sequencing technologies, including the Sanger Method, second and third generation gene sequencers, computational methods, and projections.

2.1 The Sanger Method – 1st Generation Genome Sequencing

The Sanger method, also known as the chain termination method, was developed by Frederick Sanger in 1977. This was the first practical sequencing method and was recognized as the method of choice since its introduction. The Sanger method has been commonly used both in industry and academia. Investigators of the Human Genome Project (HGP) also made use of the Sanger method as the basis of their sequencing technique. Various novel sequencing methods have evolved from the Sanger method. However, compared to the second generation methods and other novel sequencing methods introduced in the 21st century, the Sanger method is low in output and accuracy, and as a result it is no longer commonly used.

In the Sanger method, dideoxynucleotides (ddNTPs), which contain a hydrogen group on the three prime carbon instead of a hydroxyl group, prevent the bonding of addition nucleotides, and thus cause the termination of the DNA chain (Sanger, 1977). Figure 1 below is a graphic from our educational video comparing a normal nucleotide to a dideoxynucleotide.



(a) A Normal Nucleotide with a Hydroxyl Group (b) A Dideoxynucleotide with a Hydrogen Group

Figure 1: A Graphic Comparing a Standard Nucleotide (a) to a Dideoxynucleotide (b)

First, the DNA of interest is put into solution and heated to denature. Primers are prepared by cutting small segments from the complementary strand. A primer is then annealed to one of the

template strands so that the primer's 3' end is located next to the DNA sequence of interest (Sanger, 1977). Either the primer or one of the nucleotides is radioactively or fluorescently labeled to be detected on a gel later. Next the primer-attached DNA solution is divided into four tubes labeled A, T, C, G, each containing all the dNTP, DNA polymerase, and only the corresponding ddNTP, which is at about 1/100 the concentration of the dNTP concentration (1977). As a result, the replicating DNA chains are randomly terminated at the corresponding nucleotide when a ddNTP is bonded to the chain, producing DNA bands of various lengths. Then the DNA is denatured again and run in different lanes on a polyacrylimide gel in order to separate the different sized bands. The DNA fragments are attracted by electrostatic force and the shorter the DNA fragment, the lighter the molecular weight, and thus it travels a longer distance across the gel (1977). Consequently, the gel provides a short-to-long array of DNA fragments of the same nucleotide. The gel is exposed to UV light or X-rays, and a film indicating the DNA sequence is produced (1977). Figure 2 is a set of graphics from our video illustrating the running of the DNA fragments on a gel and the resulting set of bands. In the example shown in Figure 2, the first 5 base sequence is A-G-C-A-T (starting from the bottom of the gel).

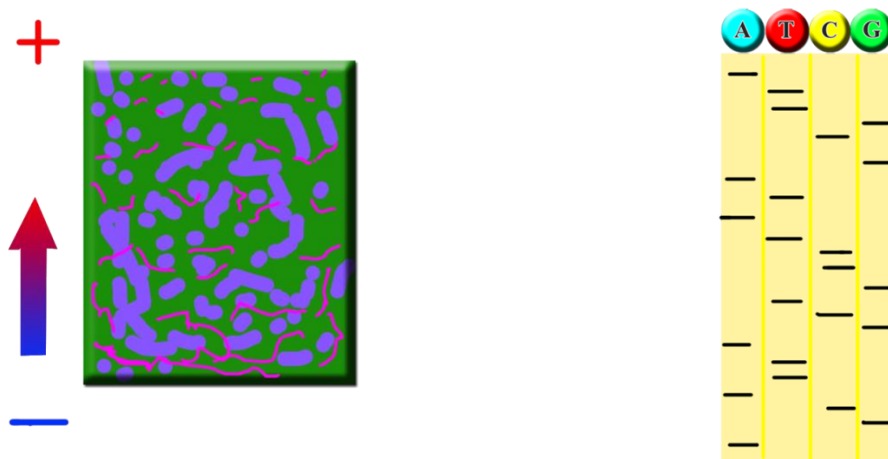


Figure 2: A Graphic Illustrating How the Chain Terminated DNA Fragments Produce a Sequence of Bands

Further improvements to the method include automated sequencing, where the ddNTPs are labeled with four different dyes (Obenrader, 2011). This allows for the use of one single gel lane instead of four separate ones and each dye fluoresces at different wavelengths, which enables automatic reading with lasers (2011)

The limitations of the Sanger method include (Obenrader, 2011):

1. Sequencing accuracy is affected by non-specific binding of the primer to the DNA.
2. The accuracy of the output is affected by DNA secondary structures.
3. The method only allows for a direct reading of 300-1000 bases at a time because the differences in lengths become insignificant for longer DNA fragments.

2.2 Second Generation Sequencing Technologies

Second generation sequencing technologies are further advances upon the Sanger method. The following is a discussion of the sequencing platforms by Roche, Illumina, Inc., and Applied Biosystems.

2.2.1 Roche 454 System

In 2005, the 454 sequencing system was the first of the second generation sequencing platforms to be available as a commercial product (Shendure, 2008). The basis of this device is the detection of pyrophosphate, which was first described in 1985, and shortly afterwards it was incorporated into DNA sequencing technology (2008). The 454 method uses DNA polymerase and primers to create a complementary strand of DNA and then amplify it. These strands have a pyrophosphate group that emits a different light for each possible nitrogenous base. Therefore, the sequence of the light being emitted creates the sequence of nucleotides. The reading length for this method is higher than the Sanger method and does not involve as much time and money in DNA preparation, however, it still requires extensive laboratory work prior to sequencing (2008). Synthesizing all of the information together creates the final assembled sequence. The following is a summary of the steps involved in the 454 method of DNA sequencing:

Generation of a single-stranded template DNA library: First the library can be constructed by any method that gives rise to a mixture of short, adaptor-flanked fragments (Ansorge, 2009). Fortunately this step can be automated and thus proves to be more convenient when compared to the extensive preparation involved in the Sanger method.

Emulsion-based clonal amplification of the library: Then emulsion PCR is used to amplify the DNA library. The amplification here is important because otherwise sufficient light signal intensity for detecting the reaction cannot be obtained during the sequencing step (Ansorge, 2009). Figure 3 is a graphic from our final set of videos, illustrating this step of the process.

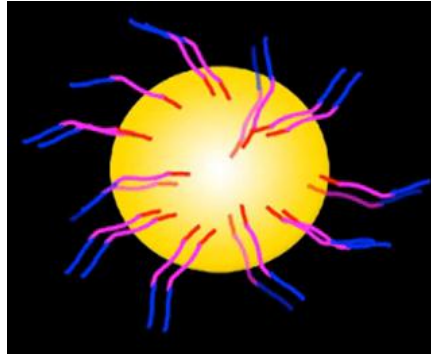


Figure 3: A Graphic Illustrating Emulsion-based Clonal Amplification

Data generation via sequencing-by-synthesis: Each bead with its single amplified fragment is placed at the top end of an etched fiber in an optical fiber chip, created from glass fiber bundles (Ansorge, 2009). In the next step polymerase enzymes and primers are added to the beads, and one unlabeled nucleotide is supplied to the reaction mixture to all beads on the chip, so that synthesis of the complementary strand can begin (2009). The base added to the sequence can be identified based on the light signal emitted.

Data analysis using different bioinformatics tools: Compared to the Sanger method, 454 technology allows for the analysis of 96 samples in parallel in a microtiter plate. Thus it is quite efficient and avoids the large use of gels or polymers that are used in the Sanger method (which limit the number of genomes that can be analyzed in parallel) (Ansorge, 2009). In addition, this method creates read lengths that are more consistent and longer than the lengths generated by the Sanger method, enabling more ease of assembly. It also has a high number of reads per run, and thus is nicknamed the “shotgun” approach (2009). The run time of less than a day and very high accuracy clearly attest to why this technology has been accepted and widely used (2009).

A major limitation of 454 technology is that it cannot deal successfully with homopolymers, for example, CCC or GGG (Ansorge, 2009). The length of the homopolymers can only be inferred from the light signal intensity, which can lead to a greater error rate. Also, currently the per-base

cost of sequencing with the 454 platform is much greater than that of other platforms, such as the SOLiD System and Illumina Solexa. Table 1 below provides performance statistics of the GS FLX+ System, the series of instruments designed by 454 Life Sciences.

Table 1: Performance Statistics of the 454 GS FLX+System
(Data from Roche Diagnostics Corporation, 2011)

GS FLX+ System		
Sequencing Kit	GS FLX Titanium XL+	GS FLX Titanium XLR70
Read Length	Up to 1000 bp	Up to 600 bp
Mode Read Length	700 bp	450 bp
Typical Throughput	700 Mb	450 Mb
Reads per Run	~1 Million shotgun	~1 Million shotgun
Consensus Accuracy	99.997%	99.995%
Run Time	23 hours	10 hours
Sample Input	gDNA or cDNA	gDNA, cDNA, or amplicons (PCR products)

2.2.2 Sequencing Systems by Illumina, Inc.

The Illumina method also uses the principle of chain termination, as described in section 2.1. However, it has substantial improvements in its high output, making commercial DNA sequencing services possible (Cappelletti, 2008). With the help of improved computation, the process is highly automated. By 2008, the Illumina Genome Analyzer Iix was the most advanced platform of the company (2008). A few years later, the company released the HiSeq 2000, which utilizes the same principle of sequencing. Improvements in engineering aspects, including the design of the platform, chemicals reagents being used and computational methods have led to one of the highest outputs currently available on the market (Illumina, 2011). It is important to note that the customizable read length of this system allows for the sequencer to be used in various fields (2011). Table 2 summarizes the performance statistics of these two sequencers.

Table 2: A Comparison of the Performance Statistics of the Illumina Genome Analyzer IIx and the HiSeq 2000

(Based on data from Cappelletti, 2008 and Illumina, 2011)

Performance Statistic	Illumina Genome Analyzer IIx	Illumina HiSeq 2000
Output per Run	20-25 Gb of high quality data	600 Gb of high quality data 6 billion reads per paired-end run
Output per Day	25 Gb	Data not available
Read Length	75 base pairs	Customized: 35, 50, or 100 bp
Accuracy	98.5%	Data not available

The following is a detailed description of the Illumina method for sequencing, with a specific focus on the Illumina Genome Analyzer IIx.

Library preparation: The DNA is first randomly cut into small fragments (~ 75 bases long). Adaptors (specially made from oligonucleotides) are ligated to the fragments. After 6 – 15 cycles of PCR and denaturation, a single stranded DNA library of different fragments is retained (Illumina, 2011). Figure 4, from Ansorge, illustrates these steps.

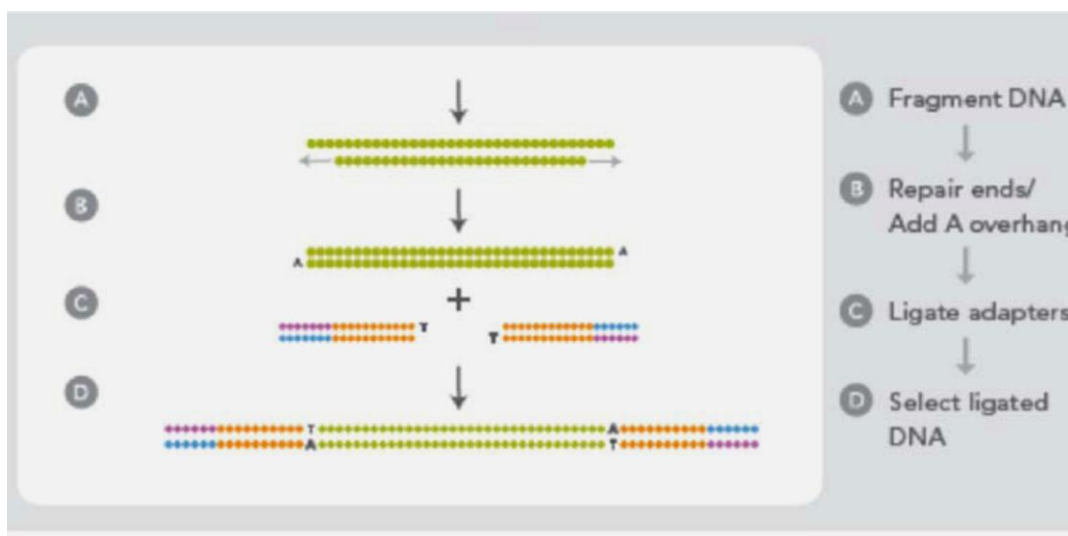


Figure 4: A Graphical Representation of Illumina Library Preparation

(Ansorge, 2009)

Automated cluster generation: The single stranded DNA fragments are washed over the surface of a flow cell, where adaptors are covalently bonded on the surface in a dense forest (Ansorge, 2009). The single stranded fragments (with adaptors on both ends) randomly bind to the adaptors on the inside surface. A process called bridge amplification is conducted and allows for the obtaining of clusters of different fragments distributing randomly on the surface (2009). Bridge amplification allows the generation of immobilized copies of a specific DNA molecule on the oligo-derivatized surface (2009). Figure 5 below is a simplified graphic from our final set of videos illustrating the concept of bridge amplification. For each cluster, there are 1000 copies of the same DNA single strand (forward direction only) densely packed in a circular area of about 1 um in diameter (2009).

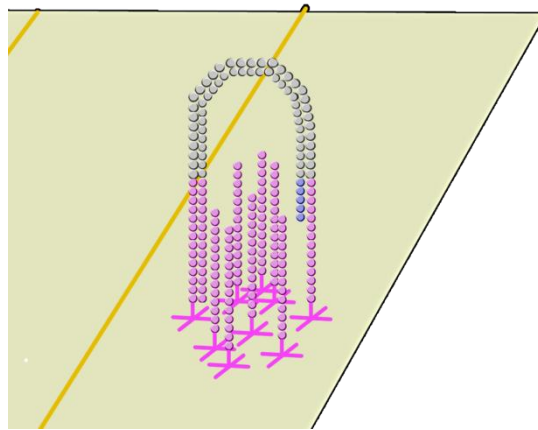


Figure 5: A Graphic Illustrating Illumina Bridge Amplification

Sequencing: Next, sequencing primers (complementary to the adaptor end) and DNA polymerase are added into the flow cell. The four different types of ddNTPs are fluorescently labeled with different dyes and added into the flow cell (Illumina, 2011). Because ddNTPs terminate the chain growth, nucleotides are added one at a time onto the primer. Next, the excessive ddNTPs are washed away. For each cluster, over one thousand of the same fluorescently labeled ddNTPs are gathered together, which results in strong fluorescence under laser excitation (2011). Figure 6 below is a graphic from our final set of videos that illustrates this step. The image is captured by a camera and analyzed by software. Next, the fluorescent dyes are chemically removed, and the ddNTPs are converted to normal dNTPs allowing the next nucleotide to be added (2011). These

steps, when combined create one sequencing cycle. The sequencing cycle is repeated to determine the next nucleotide on the fragments (2011).

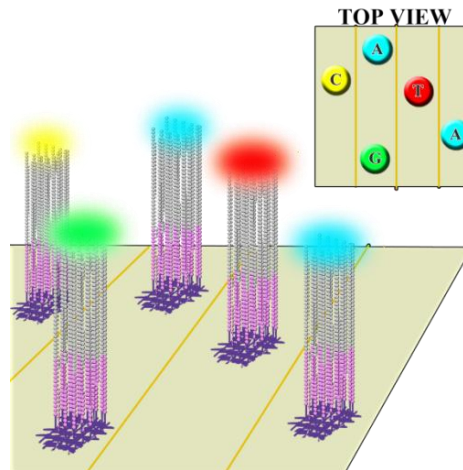







Figure 6: A Graphic Illustrating the Emittance of Fluorescence upon Laser Excitation

Although it seems that adding and sequencing one base at a time is very slow, one flow cell is able to analyze more than 150 million clusters at a time, making the system very efficient (Illumina, 2011). At approximately 20x coverage, software can be used to automatically determine the order of the fragments in order to provide the genome-scale DNA sequence (2011). Table 3 below, courtesy of Illumina, compares the performance of different Illumina sequencing systems developed over the past few years.

Table 3: A Comparison of the Performance of Recent Illumina Sequencing Systems
(Illumina, 2012)

 HiSeq 2000 Redefining the trajectory of sequencing.	 HiSeq 1000 Powerful. Flexible. Scalable.	 HiScanSQ Two proven technologies. One powerful platform.	 Genome Analyzer IIx The most widely cited platform.	 MiSeq My samples. My study. MiSeq.
Output (maximum) 600 Gb	300 Gb	150 Gb	95 Gb	> 1 Gb
Single Reads (maximum) 3 Billion total 187 Million/lane	1.5 Billion total 187 Million/lane	750 Million total 94 Million/lane	320 Million total 40 Million/lane	3.4 Million total 3.4 Million/lane
Paired-end Reads (maximum) 6 Billion 374 Million/lane	3 Billion 374 Million/lane	1.5 Billion 188 Million/lane	640 Million 80 Million/lane	6.8 Million total 6.8 Million/lane
Required input 50 ng with Nextera 100 ng – 1 µg with TruSeq	50 ng with Nextera 100 ng – 1 µg with TruSeq	50 ng with Nextera 100 ng – 1 µg with TruSeq	50 ng with Nextera 100 ng – 1 µg with TruSeq	50 ng with Nextera 100 ng – 1 µg with TruSeq
Read length 2 × 100 bp	2 × 100 bp	2 × 100 bp	2 × 150 bp	2 × 150 bp
Percentage of Bases > Q30 > 85% (2 × 50 bp) > 80% (2 × 100 bp)	> 85% (2 × 50 bp) > 80% (2 × 100 bp)	> 85% (2 × 50 bp) > 80% (2 × 100 bp)	> 85% (2 × 50 bp) > 80% (2 × 100 bp)	> 85% (2 × 50 bp) > 80% (2 × 100 bp)

2.2.3 Applied Biosystems SOLiD System

The SOLiD method of sequencing by Applied Biosystems is different from the sequencers discussed previously because it can be adjusted based on the type of information desired. For example, mate pair libraries can be used if one is trying to find genetic rearrangements (Applied Biosystems, 2011).

Similar to the previous method, the DNA is first amplified via PCR and fragmented, before a primer is hybridized to the fragment. Clonal bead populations are prepared in microreactors with all of the reaction reagents required (Applied Biosystems, 2011). Figure 7 below is a graphic from our video that illustrates this concept. More beads can be put on the slide but it can often result in lower output. Thus, the system is flexible based on the read length, coverage, and speed one desires (2011).

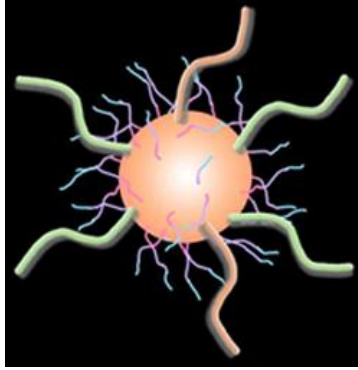


Figure 7: A Graphic Illustrating the Preparation of Clonal Bead Populations via Microreactors

The templates are then denatured, and the beads are separated using an enrichment process that divides the beads with extended templates from the beads that did not react (Life Technologies, 2011). The ligation step utilizes di-base probes to compete for ligation to the primer. Figure 8 below is a simplified graphic from our video illustrating a di-base probe emitting fluorescence. These probes have eight individual bases: three that are degenerate, three that are universal, and two that are the object of interrogation (2011).



Figure 8: A Graphic of a Di-base Probe Emitting Fluorescence as Used in the SOLiD System

A 3 prime modification is performed on the fragments, which allows the beads to bond covalently to the surface of the slide (Applied Biosystems, 2011). These modified beads are deposited on a slide. The extension product that is generated by this process is removed, and the primer goes back to the previous position (N-1) in order to ligate again. Five rounds of ligation are completed for every single sequence tag, and the process ensures that every base is interrogated by two different primers in two separate reactions (2011).

The data is generated based on a multi-base encoding scheme that assigns a particular color to four possible combinations of nitrogenous bases (Life Technologies, 2011). Similar to the Illumina method, a camera passes over and takes pictures of the plates, showing the sequences

present. A special matrix is prepared in order to decode these sequences, mapping out the possible combinations and showing the colors associated with them (2011). This is known as Exact Call Chemistry (2011). Figure 9 below is a graphic from our video illustrating the matrix used to decode the 2-base encoding.

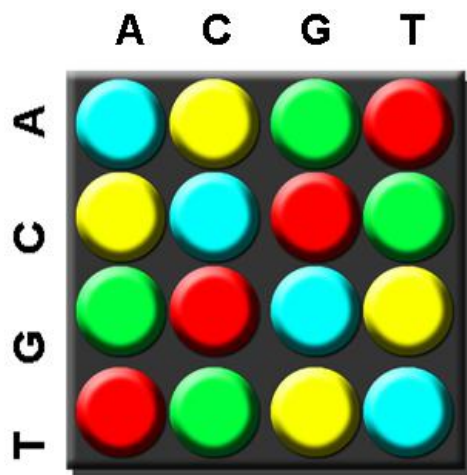


Figure 9: A Graphic Showing the Matrix Used to Decode the 2-base Encoding Used in the SOLiD System

One of the advantages to this system is the ease with which it detects single base insertions and deletions, since the entire color scheme changes completely in the region surrounding the breakpoint of the anomaly (Life Technologies, 2011). The SOLiD system is also useful for detecting small nucleotide polymorphisms (SNPs), which are thought to denote differences in the phenotypes of people (2011). Table 4 below compares the most recent SOLiD systems based on their statistical performance.

Table 4: A Comparison of Recent SOLiD Systems Based on Statistical Performance

(Data from Applied Biosystems, 2011)

System and features	5500 System (1.0 μm microbeads)	5500xl System (1.0 μm microbeads)	5500xl System (0.75 μm nanobeads)
Application-Per-Lane Sequencing	Independent FlowChip lanes allow you to configure read length of chemistry for each lane enabling multiple applications in a single run.		
System Accuracy with Exact Call Chemistry (ECC) Module	Up to 99.99%		
Multiplexing	96 barcodes for both RNA and DNA applications		
Independent Lanes	1–6 (1 FlowChip)	1–12 (2 FlowChips)	1–12 (2 FlowChips)
Throughput	7–9 Gb/day	10–15 Gb/day	>20 Gb/day
Human Genomes/Run	Up to 1 genome (30X average coverage)	Up to 2 genomes (30X average coverage)	Data not available
Maximum Read Lengths	Mate-paired: 2 x 60 bp Paired-end: 75 bp x 35 bp Fragment: 75 bp	Mate-paired: 2 x 60 bp Paired-end: 75 bp x 35 bp Fragment: 75 bp	Fragment: 50 bp
Sequencing Run Type	Yield and run times for 1 lane		
Time for 1 Genome	1 human genome (4–5X average coverage) 7 days		

2.3 Third Generation Sequencing Technologies

Third generation sequencing technologies focus on the concept of adding complementary labeled nucleotides one at a time in order to read the resulting sequence. The following is a discussion of True Single Molecule Sequencing by Helicos BioSciences and Single Molecule Real Time Sequencing by Pacific Biosciences.

2.3.1 Helicos BioSciences True Single Molecule Sequencing (tSMS)

Helicos technology allows for single molecule sequencing without amplification of the DNA via PCR (Xu, et al., 2009). This also enables easier and shorter DNA sample preparation while helping to minimize the chances of error. With 20x coverage, the accuracy of Helicos technology is better than 2nd generation technologies (2009). Table 5 below compares the performance statistics for Helicos True Single Molecule Sequencing in 2009 to the currently available data.

Table 5: A Comparison of the Performance Statistics for Helicos tSMS

(Data from Xu, et al., 2009 and Helicos BioSciences, 2010)

Platform	Read Length	Accuracy at 20X Coverage	Gigabases per Run	Cost per Human Genome	Price of Instrument
Single Molecule tSMS Early 2009	30-35 bp	99.995%	21-28 per 8 day run	70,000	1.35 million
Single Molecule tSMS Early 2011	25-55 bp Average 35 bp	99.995%	21-35 per day (1 gigabase per hour)	70,000	1.35 million

First, the DNA is broken into fragments of 100-200 nucleotides each (Helicos BioSciences, 2010). A poly-A tail is attached to the three prime end of each fragment, in addition to a fluorescently labeled nucleotide. These fragments are hybridized onto the surface of a flow cell, which contains immobilized oligo-T-nucleotides complementary to the poly-A primer (2010). DNA polymerase and fluorescently labeled nucleotides are introduced into the flow cell, where the nucleotides are added one at a time complementary to the template fragment. Similar to Illumina's method of sequencing, a laser is used to illuminate the surface of the flowcell and capture the fluorescent signal emitted. However, Helicos records the addition of each nucleotide on a single DNA fragment as opposed to Illumina's cluster based sequencing system. Figure 10 below is a graphic from our final set of videos that illustrates this concept. This allows for billions of unique fragments to be independently sequenced at the same time (2010). Furthermore, the HeliScope Single Molecule Sequencer also acts as a microscope, enabling the nucleotides to be visually seen through proprietary fluid and optic technology (2010).

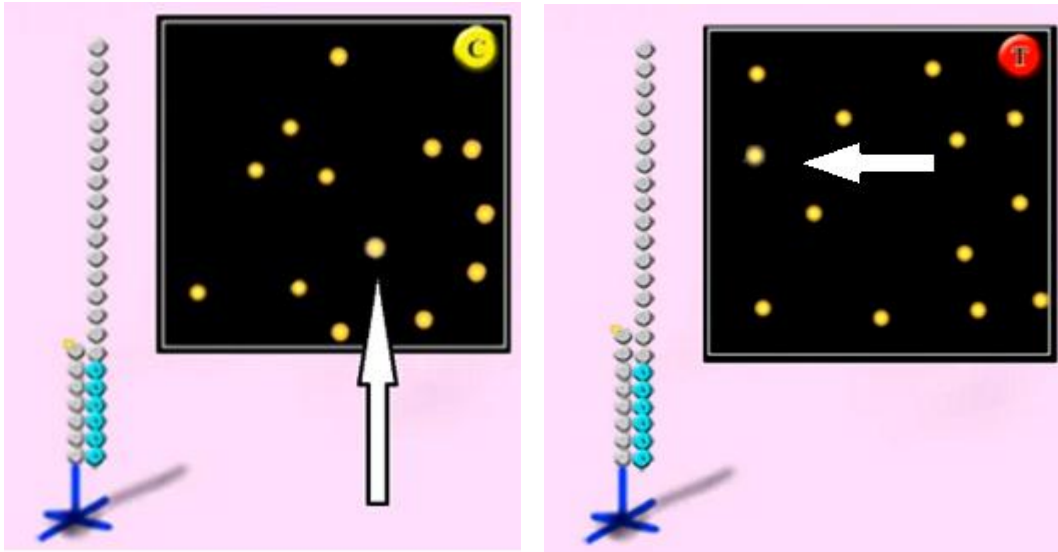


Figure 10: A Graphic Illustrating the Recording of Single Nucleotide Base Additions in tSMS

A disadvantage of the Helicos technology is that it relies on reagents, which are consumable components of the instrument (Helicos BioSciences, 2010). Thus, purchasing the sequencer and software is not sufficient since the reagents need to be bought and supplied in conjunction to how often the sequencer is used. The errors found when testing Helicos technology are of various types. 0.2% were substitution errors (where a noncomplementary nucleotide was added), 1.5% were insertion errors (where an extra nucleotide was added and thus created a frameshift mutation that caused the instructions for the amino acid sequence to be completely off by one nucleotide per codon), and 3% were deletions errors (where one of the nucleotides that should be added is not added, causing a frameshift mutation (Bauman, 2010 and Helicos BioSciences, 2010).

Some interesting aspects of Helicos technology include an easy to use touch screen interface and multiple loading channels, which require little training on the part of the lab technician and versatility of experimental design, as well as the ability to remotely monitor and obtain data from the sequencer and software through a web application (Helicos BioSciences, 2010). This enables technicians to work on other activities and/or have the sequencer run over the weekend, which reduces the cost of labor/man power for the company. Bar code readers enable organization and tracking of DNA samples, which reduces the chances of samples being mislabeled by the lab

technician. It is also important to note that the Helicos tSMS technology can be applied to RNA, making it more versatile (2010).

2.3.2 Pacific Biosciences Single Molecule Real Time Sequencing (SMRT)

Pacific Biosciences' Single Molecule Real Time Sequencing produces much larger read lengths compared to those produced by Helicos. This is resultant of the technology allowing the natural addition of phospholinked nucleotides by DNA polymerase (Xu, et al., 2009). The distinguishing feature between SMRT and tSMS technology is that the data from the tSMS technology is obtained after the 30 minute sequencing period, whereas the SMRT technology enables concurrent visualization of the data during the sequencing time period (Pacific Biosciences, 2010). Table 6 below compares the information on SMRT sequencing in 2009 to the currently available data.

Table 6: A Comparison of the Performance Statistics for Pacific Biosciences SMRT
(Data from Xu, et al., 2009 and Pacific Biosciences, 2011)

Platform	Read Length	Accuracy at 20X Coverage	Preparation vs. Sequencing
Single Molecule SMRT Early 2009	1000-1500 bp	99.3% 15x coverage (80-85% at 1x coverage)	Data not available
Single Molecule SMRT 2011	Greater than 1000, up to 10,000 at times	Data not available	Less than a day to prepare, 30 minutes of sequencing

First, fluorescence is attached to nucleotides, creating phospholinked nucleotides with a different color for each of the four nitrogenous bases. Other single molecule fluorescent approaches have the fluorophore attached directly to the base, but in SMRT technology, the labeling fluorophore is attached to the phosphate chain, which later becomes a permanent part of the growing DNA strand upon cleavage by the enzyme (Pacific Biosciences, 2010). This approach makes it easier to detect the visual signal in real time. Figure 11 below, which is also from our final set of videos, illustrates how DNA polymerase cleaves the fluorescent label off during base addition.

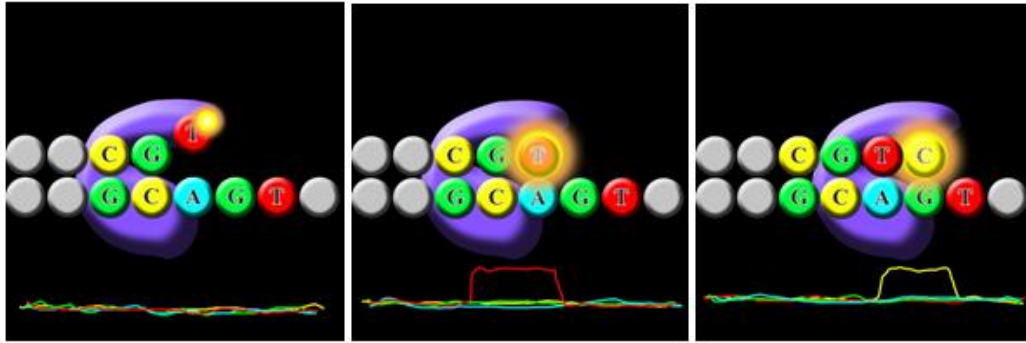


Figure 11: A Graphic Illustrating DNA Polymerase Adding a Nucleotide and Then Cleaving the Fluorescent Label During SMRT Sequencing

Thus, the short read lengths found in Helicos tSMS technology is resultant of having to start and stop the reaction in order to detect the visual signal, and this disruption interferes with the speed of DNA polymerase. The much longer read lengths produced by the SMRT method (compared to tSMS technology) allow for simpler and faster genome assembly (2010).

A key component to SMRT technology is the zero-mode waveguides (ZMWs). ZMWs are simply small pores (on the order of tenths of nanometers) surrounded by metal film and silicon dioxide (Pacific Biosciences, 2010). Each of these acts as a visualization chamber and enables the detection of a single molecule amongst the many surrounding it in the sample. Figure 12 below is graphic from our video that illustrates this concept.

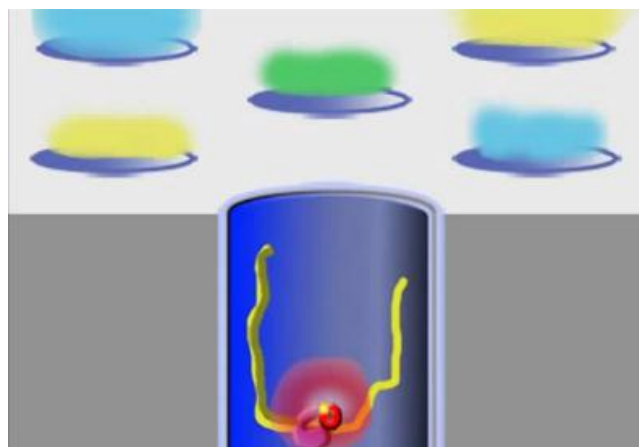


Figure 12: A Graphic Illustrating the Light Emission Captured in the ZMW Nano-photonic Chamber During SMRT Sequencing

The fast detection allows for lower background noise (Xu, et al., 2009). Assembly of the fragments is based on common sequences. The low reagent use and the long reads reduce the cost of SMRT technology.

2.4 Additional Sequencing Methods

There are many additional methods of genome sequencing arising in industry as well. Many of these technologies are less competitive in the market. Competition brought by other concurrent DNA sequencing technologies that are often more accurate and cheaper have led to the cease of published innovation by a few of the companies mentioned in this section. However, these different methods target the sequencing of different lengths of DNA and produce different levels of accuracy, and thus can be tailored according to the specific need. The following is a summary of these methods and their advantages and disadvantages.

2.4.1 Direct Sequencing via TEM by ZS Genetics

Direct sequencing via transmission electron microscopy (TEM) also involves PCR amplification, similar to the 2nd generation sequencing technologies previously discussed. Nucleotides are previously labeled, with four respective heavy elements such as iodine, and their attachment to a substrate is captured through a transmission electron microscope (Xu, et al., 2009). This microscope is based on contrast, and the heavy element is apparent compared to the other elements such as carbon and hydrogen that are part of the DNA (2009). Some issues with this technology include that the speed is limited by the speed of the camera, it involves prior preparation of the DNA and reagent samples, and that it has a high initial capital cost since TEMs can cost about 1 million dollars each (2009).

2.4.2 Nanopore Sequencing

In nanopore sequencing the single stranded DNA (ssDNA) molecule is electrically driven through a nanoscale (on the order of 10^{-9} meters) pore and the change in electrical signals dictates the DNA sequence (different electronic signal from each of the bases) (Xu, et al., 2009). However it was found to have much noise in the electrical signal, while requiring high sampling rates (2009). The final reading is also dependent on the velocity of the DNA driven through the pore. On a higher note, this technology has a relatively low cost (2009).

2.4.3 Electronic Scanning Tunneling Microscopy (STM)

In STM, an electrical vacuum is created, and a sharp metal tip is brought into contact with the surface of the DNA sample. This generates a current on the order of nA (nano-Amperes), which is then translated into an image based on current density (Xu, et al., 2009). However, many limitations come with this technology, including the consistency of the DNA sample preparation (prior to sequencing), the fragility of the DNA (when straightening and stretching the DNA sample on a surface), and also knowing the optimum location for the start of the sequencing because different start locations have given the researchers different results (Schadt, et al., 2010). Figure 13 below, from Schadt and his colleagues, illustrates this method of sequencing.

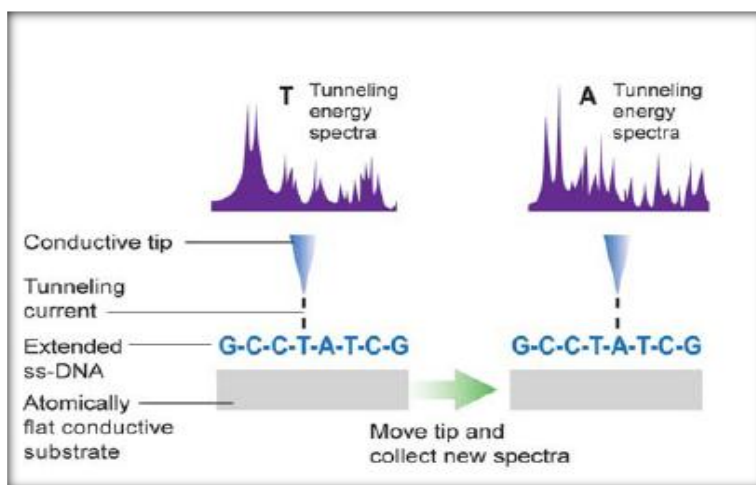


Figure 13: A Graphic Showing an Overview of Sequencing Through Scanning Tunneling Microscopy
(Schadt, et al., 2010)

2.4.4 NanoKnife Edge Probe Method by Revase

The NanoKnife Edge Probe method of sequencing is commercially known as OmniMoRA (Omni Molecular Recognizer Application) by ReVase. The single stranded DNA is first stretched and immobilized in a small channel (on the order of micrometers) (Defense Tech Briefs, 2009). Each electrically conductive probe lines up with a DNA base, excites it electrically, and then recognizes the vibrations produced by the DNA base in response to the electrical signal (2009). Each nitrogenous base has a different vibrational pattern, allowing physical detection of individual bases. When the probe comes in contact with the wrong base in

the sequence, there is no current detected (Xu, et al., 2009). Figure 14 below, courtesy of Blow, shows the relationship between the probes and the substrate.

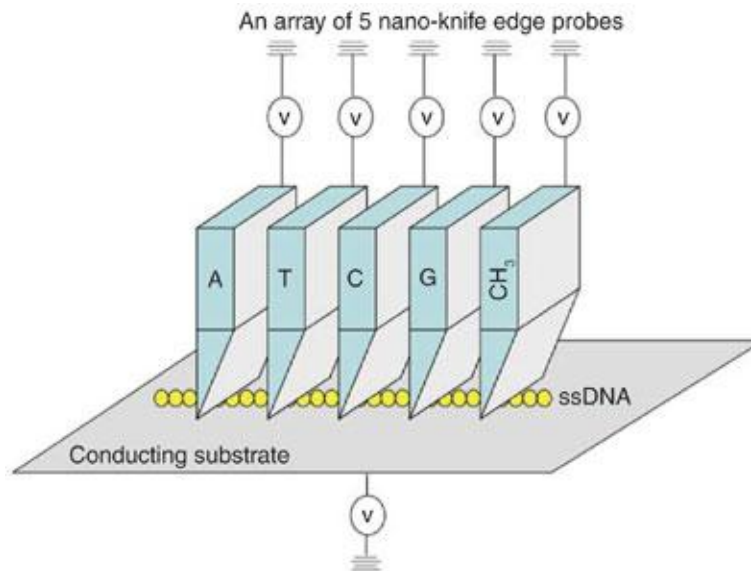


Figure 14: A Graphic Showing the Relationship Between the Probes and the Substrate in OmniMoRA Sequencing (Blow, 2008)

There was also the possibility that this technology would allow the detection of the placement and presence of methyl groups on the genome, which help regulate whether a given gene is expressed (Defense Tech Briefs, 2009). Methyl groups are also thought to be linked to cancer formation, so this technology has impacts beyond genome sequencing (2009).

This technology has its limitations as well. First, much of the research is focused on how to consistently manufacture and use the probe nano-nozzles (Xu, et al., 2009). In addition, they are also looking into incorporating optical, magnetic, and electric stimuli into the probes, thus enhancing the detecting signals (2009). The cost of the nano-knife-edge probes must also be reduced before the technology can be used widely. In addition, to reduce detection errors, it has been proposed that each nucleotide be measured with 64 sets of the four nano-knife edge probes and to use a fifth probe to detect a methyl group (Blow, 2008). This degree of accuracy warrants high cost; thus compared with other methods such as Helicos tSMS, it is not as competitive in the market (Blow, 2008). The company did report to have a high accuracy (99.9%), but this level

of accuracy has been reached by other sequencers as well (Zimmerman, 2008). The current goal of the company is to be able to sequence 100 genomes in less than 10 days, for \$10,000 per genome (2008).

2.4.5 Sequencing by Synthesis (SBS) by Intelligent Biosystems

Intelligent Biosystems has developed the Sequencing by Synthesis (SBS) method which begins with amplification of DNA fragments via PCR. The prepared DNA fragments are contained in flow cells that are loaded into the sequencer (Intelligent BioSystems, 2011). The nitrogenous bases being added to the growing strand are previously prepared to contain an end cap as well as a fluorescent dye that acts as a label (each type of nucleotide has a different removal dye) (Ju, et al., 2006). The fragments are then affixed as an array of spots on a glass chip and treated with reagents including the previously prepared nitrogenous bases, which attach to each of the spots and extend the growing strand of DNA based on the complementary strand. A camera is used to measure the fluorescent dye as it is cleaved (along with the end cap) and the information is recorded to identify the base that was added (2006).

However, the labeling method of SBS introduces issues during the addition of nucleotides to the growing strand by DNA polymerase. This can lead to varying and/or short read lengths, and potential errors (Tettelin and Feldblyum, 2009). Labeling nucleotides also makes the preparation time longer and more costly (2009).

2.4.6 DNA Nano-ball Method

Similar to many of the other sequencing methods, this method begins with creating DNA fragments with restriction enzymes and amplifying them with PCR, producing hundreds of coiled single-stranded DNA. Figure 15 below, courtesy of Mohankumar, shows this step graphically. These DNA Nano-balls are placed in an array and read starting from 10 bases away from the site of anchoring (Drmanac, 2009).

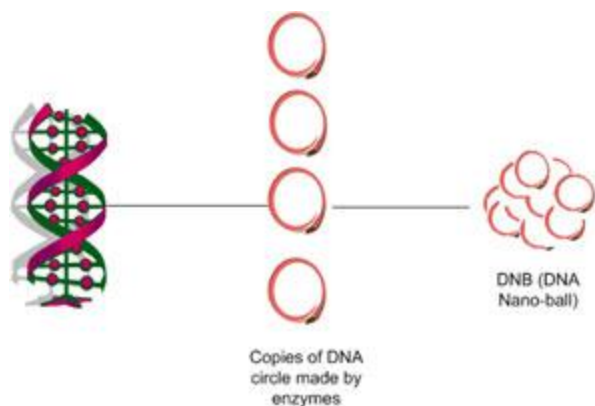


Figure 15: A Graphic Illustrating the Preparation of DNA Nano-balls
(Mohankumar, 2011)

The read length for this method is 62-70 bases (Drmanac, 2009). This method was used to sequence three human genomes, two of which had been previously sequenced using other technologies. An alignment algorithm was used to assemble the sequence reads, and representational biases were assessed in order to check base coverage statistics (2009). Reportedly, calls were made with some confidence, and many of the errors found were discovered to be errors made in the reference genomes. Overall, this process costs approximately 4400 dollars and thus proves to be very affordable (2009).

However, there was a significant issue with errors in the genome sequencing while using this process. Occasionally, aberrant mate-pair gaps appeared which would often lead to length altering structural variants as compared to the reference genomes (Drmanac, 2009). Another source of error in most methods came from the fact that N+1 sequencing methods were used, meaning that the accuracy of one sequence is dependent on the last (2009). This could cause initial errors to cause increasing inaccuracies and carry forward large amounts of errors.

2.5 Computational Methods

Currently two factors are crucial to the development of next-generation DNA sequencing technology: the biological methodologies as well as the computational power. It is important to find reliable and efficient ways to interpret large data and obtain meaningful results. Currently there are several popular areas in which computation power plays an important role. For

example, For example, data mining of DNA databases is necessary to associate certain genes to a disease (Dennis, 2009). With the computation capabilities of computers almost doubling every year, more research is being performed in this area in order to find better algorithms to match the large amount of data acquired by current DNA sequencing technologies (2009).

2.5.1 Computational Challenges Posed by New Genome Sequencing Technologies

Foremost of the challenges posed by the advances in sequencing technologies is the large data storage requirement for the sequencing results. For example, Illumina's Hiseq 2500, which was released in Jan, 2012, has the output of 120 GB of data in 27 hours (Dennis, 2009). In addition, it is estimated that for 20-60 GB results data, a total of 56 TB raw data is involved (2009).

Second is the need for faster processing and interpretation of DNA sequencing results. Next generation sequencers can now produce large amounts of data in the form of short reads every hour, thus requiring heavy computational analysis (Rusk, 2009). More efficient software tools and high-performance hardware are needed to meet the ever-increasing demand.

2.5.2 Data Processing and Management

To address the demand for large storage and hardware computing power, cloud computing is now widely used in industry. Clouds are a large pool of easily usable and accessible virtualized resources (such as hardware, development platform or services) (Dennis, 2009). These resources can be dynamically re-configured to adjust to variable load (2009).

There are several advantages cloud computing can bring to sequencing data processing and management. First is the incomparable computing power it possess, since an entire computing task previously performed by a single super computer can now be processed by a cloud of 100 computers, which often offers more computing power and increased reliability. In addition, cloud computing reduce the cost significantly in terms of the storage hardware and has a relatively low cost for its computing capacity. Table 7 below shows the cost of Amazon Elastic Compute Cloud based on the operating system used.

**Table 7: The Cost of Amazon Elastic Compute Cloud
(Amazon Web Services, 2012)**

Standard On-Demand Instances	Linux/UNIX Usage	Windows Usage
Small (1 core, 1.7GB memory & 160 GB storage)	\$0.080 per hour	\$0.115 per hour
Large (4 cores, 7.5GB memory & 850 GB storage)	\$0.320 per hour	\$0.460 per hour
Extra Large (8 cores, 15GB memory & 1690 GB storage)	\$0.640 per hour	\$0.920 per hour

Furthermore, currently the cost of cloud computing is decreasing by 15% every year (Dennis, 2009). Thus, it is expected that the price will continue to drop over time. This technology also allows users to enjoy the flexibility provided by being able to choose the type of server, operating system and CPU power that best matches the requirements of their task.

2.5.3 Analyzing Sequence Data

Sequencing analysis is the process of using various analytical methods to find the structure, features and functions associated with the particular genomic sequence. This includes sequence alignment, searches against a biological database, and sequence assembly (Durbin, 1998). In order to process the large output of data in genome sequencing technologies, almost all of the algorithms currently being used employ a heuristic method, which provides the largest number of possible results (Dennis, 2009). Thus, there is always a relationship between sensitivity and speed. In most cases, this function relationship can be represented by a curve.

In addition, computational methods are necessary for the final assembly of the genome, once each fragment is sequenced. The heuristic algorithm is commonly used along with a scoring system to evaluate each set of bases obtained by the sequencer (Dennis, 2009). The number of bases evaluated at a time as well as the possibility of errors is also related to the sensitivity of the algorithm. Higher sensitivity also can result in lower speed, thus there is a trade-off between having a faster method of sequencing a genome and having an accurate assembled sequence

2.5.3.1 Sequence Alignment

A large part of DNA sequencing analysis lies in the alignment of these sequences. Sequence alignment refers to a way of arranging sequences of DNA, RNA, or proteins in order to identify similarities between them (Mount, 2004). These similarities can be resultant of several factors, including the function, structure, and evolutionary history between the sequences (2004). Thus sequence alignment is the key to finding important regions and determines the function of each part. Global sequence alignment refers to comparing both sequences up to their complete length, while local sequence alignment only does the searches in conserved regions (2004). The following is an example of alignment, where there are 6 matches, 4 gaps, and 1 mismatch (Batzoglou, 2006).

```
A--TGG--ACGT--  
ACTG--CCCGTA
```

According to a scoring function, every match, gap, and mismatch is associated with a score to determine the proficiency of the alignment, as shown in the example equation below (Batzoglou, 2006).

$$\text{Score } F = (\# \text{ matches}) \times \text{assumed match} - (\# \text{ mismatches}) \times \text{assumed mismatch} - (\# \text{gaps}) \times \text{assumed gap}$$

Often an aligner is needed during DNA sequencing analysis. In other words, upon acquiring the genome sequence of a patient with a given disease, an aligner is required to find what particular part of the genome is actually causing the disease. Two of the most widely used algorithms to do sequence alignment are BLAST and CLUSTALW.

2.5.2.2 BLAST Algorithm

The abbreviation, BLAST, stands for Basic Local Alignment Search Tool. This algorithm is mainly used for searching and comparing sequences. BLAST is one of the fastest comparing algorithms since it emphasizes speed rather than sensitivity (Gish, 1990). Due to how new genome sequencing technologies have a higher throughput, an algorithm with high speed but lower accuracy is often preferred (Gustav, 2007). The key to its high speed lies in the fact that

BLAST uses a heuristic method, which means it make a guess about the possibility for there being a match, thus improving the searching speed greatly (2007).

BLAST and many other algorithms primarily use a similarity search because it allows for the identification of conserved functions (Eddy, 2004). The output of a BLAST program consists of many parts, including a header, graphic overview, description, alignment, and statistics (2004). Also one advantage of BLAST is that parameters can be changed to improve the speed, but of course lower sensitivity comes with the higher speed (2004).

Based on the advances being made in the field of genome sequencing technology, it is believed that people will either continue to modify BLAST, or simply develop a new algorithm with higher speed and relatively satisfactory sensitivity. If routine sequencing of individual human genomes becomes a reality in the future, a faster algorithm is required, as well as a biological method with higher throughput.

2.5.2.3 CLUSTALW Algorithm

CLUSTALW, which stands for Cluster Alignment Interface with Weights, is another way to find the alignment between short read lengths. This algorithm is widely used for multiple sequence alignment (Thompson, et al., 1994). Sequence alignment refers to the arranging of DNA sequences to identify the similarity of a given region, which is often the functional region (1994). In this algorithm, individual weights are assigned to each sequence in partial alignment in order to distinguish the difference between similar and divergent sequences (1994).

2.6 Combining Sequencing Technologies

There has also been research into the possibility of combining different sequencing technologies in order to overcome the limits of each technology when used on its own. For example, when Roche 454 technology was combined with the Sanger method it enabled the amplification of DNA fragments that cannot propagate in an *E. coli* host, since 454 does not rely on the use of this host (Tettelin and Feldblyum, 2009). However the limitation of 454 with sequencing large homopolymers (large sequences of a repeated single base) can be overcome with the Sanger

method that produces more accurate results when it comes to homopolymers (2009). As a result, combining sequencing technologies allows each to compensate for the limitations of the other.

Unfortunately, combining these technologies introduces many issues as well. For one thing the data from these machines are quite large and are accompanied by the issues of data storage, in addition to each sequencer often having differing software applications that can process and analyze the output of each sequencer. In addition, the differences in accuracy, read length, and DNA preparation can make it difficult for the results of a combined method to be reliable and consistent.

2.7 Projections

In order to meet the goals for the next generation of sequencers, as stated in Chapter 1, engineers are now researching methods of modifying the current equipment to allow for more convenience for the user while creating higher amounts of accurate and fast output. Some of these investigators are considering adding a “bar code like” feature to DNA sequencing machines so that they can differentiate between matching target DNA from different people (Wright, et al., 2011). Currently, many sequencers enable the simultaneous sequencing of multiple samples, often through having several channels, but it is difficult to differentiate between matching samples of target DNA. As a result, adding DNA bar code tags (labeled molecules) to the ends of DNA fragments during the library preparation would enable the tracking of individual patients (2011).

Many of the DNA sequencers currently being manufactured and sold are ideal for large sequencing centers instead of small research and clinical laboratories. These large systems are also too expensive of an investment (in terms of capital) for small laboratories and companies as well. As a result, some companies are now developing smaller bench top genome sequencers that are more affordable and take up less space (Wright, et al., 2011). Table 8 provides information on the currently available bench top sequencers. These, however, are more geared towards small genome sequencing (bacterial), and PCR product sequencing (fragments of larger genomes). The goal is to design bench top sequencers that are able to sequence whole human genomes (2011).

Table 8: Specifications of Bench Top Sequencers that are Available as of August, 2011
(Data from Wright, 2011)

Platform	Real Length	Run time	Output (per run)
Mi-Seq (Illumina)	1x35 bp	4 hours	120 MB
	2x100 bp	19 hours	680 MB
	2x150 bp	27 hours	1 GB
GS Junior (Roche)	400 bp	10 hours	35 MB
PGM (Ion Torrent)	100 bp	1-2 hours	1 GB

It is thought that the future of algorithms used in genome sequencers may have modifications in the following fields in order to meet the demands of the improvements in new genome sequencing technologies. In most heuristic algorithms there is a scoring mechanism used as the criteria to evaluate each word, which is directly related to the sensitivity of the algorithm (Dennis, 2009). Thus, improving the scoring mechanism can directly result in improved sensitivity. Also the statistics model can be improved to better reflect the needs of the user. Improving the statistics model may allow users to be able to adjust the sequencer to produce the sensitivity desired.

Chapter 3: Impacts and Applications of Genome Sequencing

This chapter includes research on the social impacts and medical applications of next generation gene sequencing technologies. Advances in this field exercise significant influence in the economy, legislation, forensics, health care, biological studies, and agriculture.

3.1 Economy

Research in genome technologies has brought tremendous impacts to economy and industry. These economic impacts consist of three parts: direct impacts (regarding specific expenditures), indirect impacts (regarding suppliers), and induced impacts (regarding the additional economic impacts from the spending of suppliers and employees in the overall economy, as well as the additional industries enabled by advances in genome sequencing technologies) (Battelle Technology Partnership Practice, 2011).

Figure 16 below, courtesy of Battelle Technology, is a diagram showing the structure of forward and backward linkage associated with the Human Genome Project (HGP), and highlights the many fields that have grown due to the advances in genome sequencing technologies. An assessment report on the economic impacts of genome research from Battelle was made available in 2011, which draws most data from government statistics. According to the findings in this report, between 1988 and 2010 the human genome sequencing projects in the United States (not constrained to just the Human Genome Project) generated an economic output of \$795 billion (Battelle, 2011). These projects created 3.8 million job-years of employment and \$244 billion of personal income (2011). The indirect and induced outputs are far beyond this direct output.

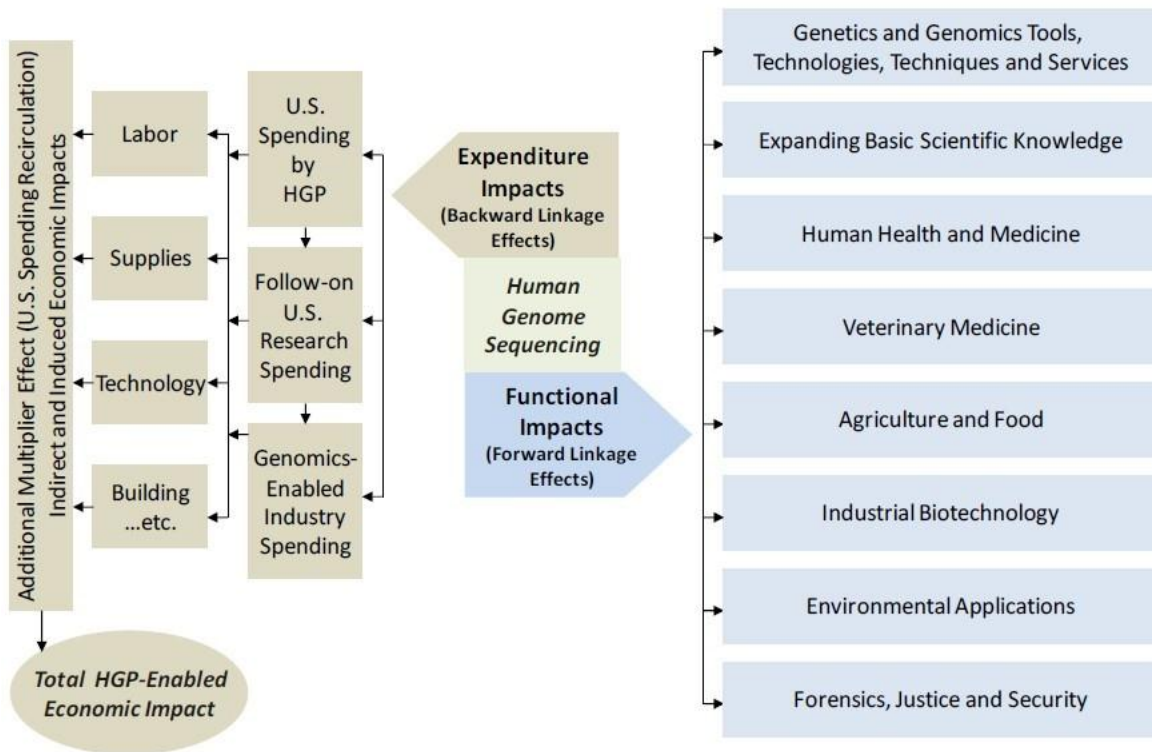


Figure 16: Structure of Forward and Backward Linkage Associated with the Human Genome Project
(Battelle, 2011)

The reported federal investment from the U.S. government during the period of the Human Genome Project was \$5.6 billion (Battelle, 2011). Further investment on genome technology research after the HGP between the years of 2004 and 2010 was 28% more than the original investment (2011). This investment was the foundation to generating the economic output of \$795 billion above, and thus shows a Return on Investment to the U.S. economy of 141 to 1. In other words, every \$1 of federal investment contributed to the generation of \$141 in the nation's economy (2011). In 2010, the total genomics-driven output is \$265 billion (2011). Figure 17 below is a graph illustrating the increase in U.S. federal funding towards genome related research between the years of 1988 and 2010.

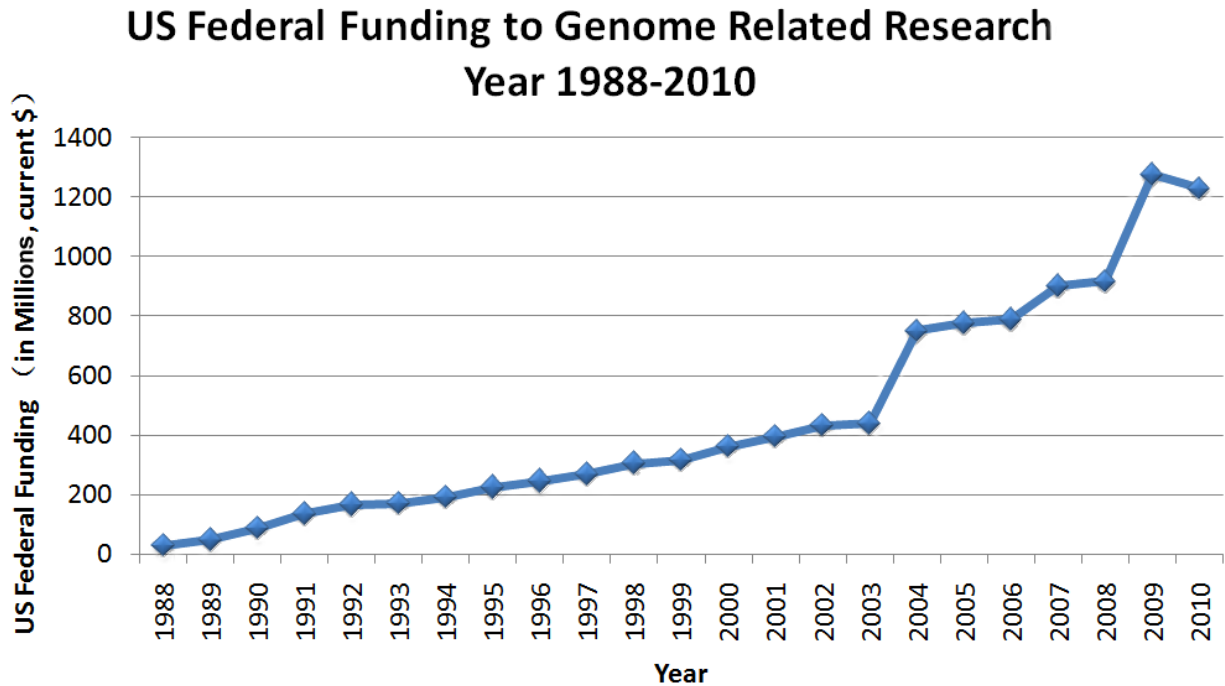


Figure 17: United States Federal Funding for Genome Related Research between 1988 and 2010
(Data from Battelle, 2011)

There are six major sectors within genomics related research: Research and Development (R&D), Instruments and Equipment, Related Biologics and Diagnostic Substances, Related Testing, Related Pharmaceuticals, and Related Bioinformatics (Battelle, 2011). A comparison table of employment estimation of the sectors is shown below in Table 9. Figure 18 illustrates this information in a graph.

Table 9: Comparison of Employment Estimation (Jobs per Year) by Sector

(Data from Battelle, 2011)

Genomic Sector	2010	2003	2000	1992
Genomics R&D and Related Biotechnology	13,323	13,140	8,275	2,378
Genomic Instruments and Equipment	11,704	15,727	10,957	9,917
Genomics-Related Biologics and Diagnostic Substances	7,234	9,427	7,145	2,243
Genomic Related Testing	5,142	1,644	1,301	542
Genomics-Related Bioinformatics	792	1,430	667	174

Pharmaceutical sector R&D	2010	2003	2000
Genomics-Related Pharmaceutical R&D	9,109	2,337	743

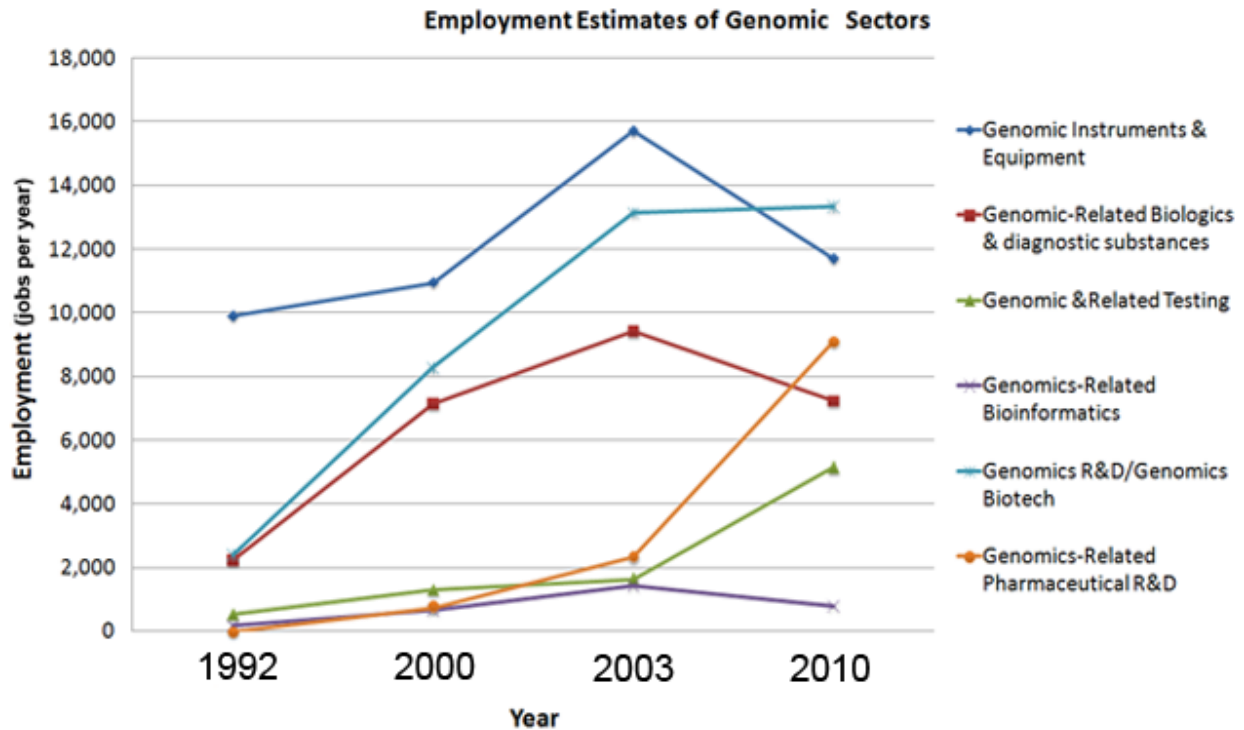


Figure 18: Comparison of Employment Estimation by Sector

(Data from Battelle, 2011)

As illustrated by the data, genomic instruments and equipment is the primary influencing factor among all of the sectors, followed by genomics related research and development and genomics related pharmaceutical research. The drop in employment seen in Figure 19 is resultant of the completion of the HGP (Battelle, 2011). The recent economic recession has also dampened the employment in the field of genomics. On the other hand, the investment and economy output continued to increase after the completion of the HGP. In addition, the genomics-related share of the pharmaceutical industry has continued to significantly grow in both employment and output since 2003.

3.2 Worldwide Legislation Regarding the Release of Genetic Information

The advent of faster and more available human genome sequencing has led many countries around the world to consider creating legislation to either protect the privacy of individuals who are sequenced, and/or to assure that information from one's genome is not used in way that creates detrimental consequences to the sequenced individual. Other issues such as sequenced minors and the genetic information of the deceased also create controversy.

3.2.1 Implications of Genome Ownership, Privacy, and Disclosure

A majority of the genome sequencing performed today involves informed consent. One issue with informed consent is that individuals giving consent for their genetic information to be reviewed by others cannot foresee all of the consequences of this action and assess them prior to making the appropriate decision (Wright, et al., 2011). As a result there is a limitation of the word *informed* in the phrase *informed consent*. For example, a person may be asked by an employer to disclose their genetic information to see if they have a predisposal for a given disease and whether they require a vaccine prior to working in the laboratory. The patient may consent, knowing that they do not have a history of that disease, but may face consequences if the employer finds that the employee has a predisposal to cancer and thus does not want to hire him/her anymore on the assumption that the employee may have to leave work for treatment in the future.

Another issue is not always being able to obtain complete family consent (Wright, et al., 2011). For example, a patient may wish to have a gene therapist create a pedigree and inform them on

the chances of the patient's child being born with a particular condition. In this case, the patient may be okay with releasing his/her genetic information, but the other family members who are also involved in the development of that pedigree may not all wish to give consent. It is also difficult when the family member in question is deceased because one cannot assume that releasing their genetic makeup would be according to the wishes of the deceased person (2011). Similarly, it is difficult to prevent one from drawing conclusions based on seeing non-genetic data that pertains to rare disorders, which can be linked to diseases that can arise in the future. It can also be difficult to prevent someone from drawing conclusions from a donor profile (in regards to organ and blood donation) (2011).

The process of obtaining consent for the release of genetic information can be further complicated if the patient involved is a minor. For example, the parents may agree to the release because it may enable other clinicians and analyzers to help the child in regards to a treatment of a disease, but the child may grow up to regret that release of genetic information if it prevents him/her from gaining proper health insurance even after they are completely healthy (Wright, et al., 2011).

Some patients may also not wish to know if they are predisposed to certain diseases. The "right not to know" must also be respected. In one case, James Watson, one of the Nobel Prize winners for the discovery of the structure of DNA, agreed to have his genome analyzed and available to the public except for a particular protein that links to Alzheimer's disease, because he personally did not want to know whether he was predisposed to it or not (Wright, et al., 2011).

Furthermore, medical records are increasingly becoming electronic in an effort to reduce the use of paper, and to increase productivity by allowing for faster searching. However, with medical data (including data from whole or partial genome sequencing) being electronic, an additional level of security is needed to surround these files in order to prevent unlawful access to patient information. As a result, advances and availability of genome sequencing causes growth in the electronic security industry but also induces fear in some patients who may be concerned that their information will not remain confidential despite laws such as GINA (which is discussed in section 3.2.2) due to the records being electronic. One proposed solution for this is to have patients give consent for disclosure prior to genetic testing, since it is very difficult to keep that

information private once it is obtained and analyzed (Wright, et al., 2011). In this way, all tested patients would have their genetic information available in a database and they would be fully aware of this prior to agreeing to have their genome sequenced. Another solution to the issue of securing electronic medical records that contain genome data and analysis results is to make that part of the record available only at certain times and to certain professionals to minimize the leaking of information (2011).

In any case, the line needs to be drawn between releasing genomic data (even if the patient gives consent) for research purposes and releasing it for clinical purposes (treating the patient for instance). In some cases, the clinician is conducting his/her own research, so it can be extremely difficult to differentiate between whether the patient is releasing the information for personal purposes or to further the purposes of others.

3.2.2 Legislation in the United States

Advances in new genome sequencing technologies also introduce questions of intellectual property. With genome sequencing, one can now patent the sequencing technologies, specific sequences of DNA and regulatory RNA, diagnostic test kits, chemical reagents, and other components of genome sequencing. However, different countries provide different amounts of protection for intellectual property, and whole genome sequencing can sometimes cross international borders (Wright, et al., 2011). The information obtained from whole genome sequencing may also interfere with existing patents in regards to diagnoses, treatments, and prevention procedures in medicine.

Fortunately, the United States Genetic Information Nondiscrimination Act (GINA) prohibits health insurance and employers from discriminating on the basis of one's genetic makeup (U.S. Department of Health and Human Services, 2009). However, it does allow employers to keep genetic information in individual personnel files as long as the employees have allowed for the disclosure of that information and it is clear that they are not obligated to disclose their genetic information (National Library of Medicine, 2012). This type of information includes prenatal test results, ancestry and predisposal to certain diseases, and paternity test results (2012). It also prevents employers from requiring potential employees to have genetic tests done without

justification (2012). The initial act was introduced in 2008, but many revisions and provisions have been added since.

In addition to GINA, which is a federal act, many states also have related laws that vary in how protective they are. One limitation of GINA is that it does not refer to insurances other than health insurance, such as disability insurance and life insurance (Department of Health and Human Services, 2009). It also does not always cover companies with fewer than fifteen employees (2009).

3.2.3 Legislation in Europe

In addition to the United States, Europe is also taking part in developing a relationship between genome sequencing technologies and their application to the health care system. However, countries in Europe differ in their attitude towards genetic testing. Currently, there is no uniformed legislation against genetic discrimination by the European Union (Borry, 2008).

The countries of the United Kingdom and Belgium allow direct-to-customer genetic tests to be provided by private companies. Previously in the UK, genetic tests were confined to only specialized clinics and the information could not be used by employers or insurance companies (GeneWatch UK, 2006). Therefore there was no legislation to prevent discrimination according to genetic make-up. But in recent years, direct-to-customer testing has come into the market, and many genetic discrimination cases, especially those in the U.S., have raised the debate in the UK as to whether genetic discrimination should be prohibited by the law (BBC, 2011). As a result, the 2010 Equality Act restricts the use of genetic tests by employers and insurance companies. It can be seen that the UK is moving in the same direction as the United States in regards to this issue.

On the other hand, some European countries ban direct-to-customer genetic tests completely. This group of countries includes Germany, France, Portugal and Switzerland (Borry, 2008). Legislation has been made so that genetic tests can only be carried out through a medical doctor with the consent of the patient. In addition, the doctor must inform the patient of how meaningful the test is and what the possible consequences will be. In a statement from the Council of Europe

in 2003, it was pointed out that many ethical, social, and legal issues have been raised by directly marketing genetic tests to customers (2008).

Overall, most European countries show concern regarding the usage of genetic information. Though their aims to avoiding issues regarding genetic discrimination are similar, countries differ in how they specify the manner and degree of the legislation. Some countries elect to form general regulations and leave the details for special committees. Whether centralized legislation at the European Union level will form remains to be seen.

3.2.4 Legislation in Australia

Genetic research is also developing quickly in Australia. In 2007, more than 220 genetic tests were available for the use in medical diagnosis and risk (Sandra, 2007). As a result, cases of genetic discrimination started occurring rapidly. Reports show that by the year 2000 there were 48 anonymously reported cases of genetic discrimination in Australia, including cases involving life insurance and employment (Stewart, 2010). This issue has since then drawn public attention and the government took action by creating the Australian Law Reform Commission (ALRC) and the Australian Health Ethics Committee (AHEC) to conduct a comprehensive inquiry in terms of protecting genetic information (2010).

It is important to note that Australia has a national health care system that is not risk-rated. Also unlike the US, there is no direct relationship between employers and insurance agencies, which causes the government to use a different approach to solving this issue (Stewart, 2010).

In terms of health insurance, the ALRC and the AHEC state that little change needs to be made in the current legislation (Stewart, 2010). It is held that genetic information obtained from personal and family genetic testing is no different from any other information for insurance purposes. This opinion legalizes private insurers to take genetic information into consideration when offering their insurance products.

However, in regards to employment, the ALRC and the AHEC recommend that some changes should be made to the current legislation (Stewart, 2010). Even though there is little information showing that employers in Australia are using genetic screening, it is inevitable that employers

will someday try to make use of the new technologies as the cost decreases and the results become more reliable. These committees suggest for the prohibition of employers using genetic testing except for rare cases where the health of employees are closely related to the working environment (2010).

Thus, even though the Australian government has not yet incorporated genetic discrimination into legislation, it does accept the recommendations made by the ALRC and AHEC. As a result, it can be expected that changes will eventually be made to the federal legislation.

3.2.5 Legislation in Japan

Unlike the countries previously mentioned, Japan currently lacks a regulatory framework to control how genetic information is used (Porter, 2010). While the idea of prohibiting genetic discrimination has been mentioned in some guidelines; there is no restriction on insurance companies and employers regarding how they can make use of results from genetic testing (2010). However, this situation has not yet caused many reported incidences genetic discrimination.

3.2.6 Legislation in China

Furthermore, China has not created any legislation in terms of preventing genetic discrimination and it seems that this situation will not change in the next few years. However, there are several reasons behind this. Foremost is the fact that there is a low amount of public awareness regarding the application of genome sequencing technologies in the medical field (Mao, 1998). Though China built the Beijing Genomics Institute in 1999, which is one of the premier genome sequencing centers in the world, this institute is mostly focused on research projects such as the Human Genome Project, SARS Project, and the sequencing of plants and animals (Beijing Genomics Institute, 2012). In addition, due to the lack of funding and expertise, genetic services in China are still underdeveloped (Mao, 1998).

Moreover, cases of genetic discrimination usually occur in fields like insurance and employment. Unlike the United States, China has a different policy regarding insurance and pensions, where employers pay directly the country's government instead of private insurance companies, and the

government then distributes this to employees where applicable (Mao, 1998). As a result, few cases of genetic discrimination by insurance companies have been reported.

But rare occurrences of genetic discrimination in employment do exist. It has been reported that China's first claim regarding genetic discrimination in 2010 was unsuccessful (Isaacs, 2010). In this case, three civil servant candidates sued the Foshan Human Resources and Social Security Bureau in Guangdong Province for allegedly discriminating against them during the recruitment process on the basis that they have a gene predisposing them to thalassaemia (a blood disease) (2010). The local court eventually ruled in favor of the employer based on the fact that thalassaemia can be a severe disease. However, this ruling seems contradictory to the fact that China protects carriers with Hepatitis B from discrimination through the legislation, due to the large number of people in China that are afflicted with Hepatitis B (Isaacs, 2010). It remains to be seen how legislation will develop as more cases of genetic discrimination in employment occur in the future.

3.2.7 Legislation in Developing Countries

There is much less debate over genetic discrimination in developing countries. It is reported that there were approximately 5500 genetic specialists in 1998 across the world, and 3300 of them worked in developed countries (Mao, 1998). This situation has not changed significantly over the years since this statement due to the fact that genetic research is a privilege of developed countries. Unfortunately, developing countries have other more pressing issues to address such as poverty, poor health care, and illiteracy before they can address issues relating to gene sequencing.

3.2.8 Cultural and Religious Barriers

Another ethical issue relating to the advances in genome sequencing technologies is the cultural/religious barriers. Since genome sequencing technologies are turning something that was perceived as a rare and expensive occurrence into a more affordable, versatile, faster, and useful method in medicine, the future of whole genome sequencing may encounter interference from various religions and cultures. For example, if based on genome sequencing a doctor determines that the best way to treat an infant's genetic disease is to develop a personalized drug, the

parent(s) may disagree stating that doing such an act interferes with “God’s Will” (Wright, et al., 2011). As a result, even though whole genome sequencing technologies seems to be growing at a logarithmic rate, it could be easily halted by religious and cultural issues.

3.3 Forensics

The DNA sequencing of loci has been useful in both solving cases and identifying the remains of victims. Specifically, the advent of second and third generation sequencing technologies have had an immediate impact on the field of forensics. In fact, Dr. Mitchell Holland and his colleagues used the 454 Genome Sequencer FLX to analyze the ability to get reliable sequence data from STR and Y STR loci as well as look at mitochondrial DNA sequences all in a single reaction (Holland, et al., 2011). This allowed for the identification of certain loci using the binding of primers and for the using of sequence based barcodes to identify individuals (2011). The significant aspect about this is that this team claimed to be able to analyze many loci and hundreds of individuals in a single run of the genome sequencer (2011). They also claim that fragment length is not a factor, making more sensitive systems possible. The pros of this system are an increased discrimination potential and a low cost solution.

Moreover, Casework Genetics is using Illumina technology in forensic applications as well. Specifically, Casework Genetics used the Illumina Infinium HumanOmni 1-Quad BeadChip in their genotyping operations, and were able to genotype over a million SNP’s per sample (Illumina, 2011). Typically, short tandem repeat polymorphisms are used for DNA fingerprinting, which involves bases that are repeated multiple times, and this region of bases being searched out using PCR amplification and gel electrophoresis (2011). However, this method is noted as being slow and does not solve the issue of mixed samples. Casework Genetics views searching out single nucleotide polymorphisms, also known as SNP’s, as the future of forensics, as SNP’s are highly conserved within populations, and as such are fairly unique (2011). It was also found that using the Illumina Infinium HumanOmni 1-Quad BeadChip in conjunction with Illumina’s iScan system allowed for more automation and organization, as the team was able to scan beadchips in just a few minutes (2011). The sequencing was found to be as accurate as other methods and the Illumina system extended the limit of detectability within mixtures of samples (2011). In summation, the Illumina system is ideal for dealing with complex

mixtures, and gives hope for the tracking of fugitives, determining the presence of an individual in a public place, and even understanding criminal groups.

However, there are also some issues with using gene sequencing as part of forensics. For one thing, legislation has kept up with the increased gathering and analysis for the UK's National DNA Database (ESRC, 2008). It is proposed that developments of technology must be met with respective developments in law, and this is exemplified in the reclassification of mouth swabs as non-intimate DNA collection (2008). Occurrences such as this verify the belief of some that the accuracy of DNA evidence makes exceptional legal procedures necessary. As such, G8 countries have developed a protocol for the proper exchange of DNA evidence between nations, which includes weighing the rights of victims versus the rights of the accused, and considering that most innovations become practices through their application to specific cases, not merely by theorizing (2008). It is thought that ethical considerations may be overlooked in the future as the use of gene sequencing in forensics becomes more routine (2008).

It is also difficult to build the trust of citizens about the retention of genetic evidence. In fact, some technologies could restrict access to genetic databases, with the main impacts being on privacy, autonomy, and social justice (ESRC, 2008). Any accessible privacy enhancing technologies would increase the autonomy of the system, but would not administer proper social justice because it would put those without access to technology or who at a disadvantage. However, non-accessible technologies reverse the issues described above. An additional issue is obtaining the trust of citizens in regards to the storage of genetic information in a database with genetic information from criminals (2008). This practice is done to simplify searches during criminal cases, but it can cause citizens to feel uncomfortable.

Furthermore, the sequencing of children in cases of missing or kidnapped children also creates some dilemmas. These issues include gaining consent for minors, as described previously, and determining which parts of the genome should be disclosed to children and/or parents (ESRC, 2008). In addition, as it becomes a common practice to obtain DNA samples from family members in order to track down criminals or gain insight into their genetic predispositions to criminal activity, issues such as people not wanting to be associated with criminal investigations

arise (2008). It also introduces another problem when a familial link is found that proves that criminality can run in a family, since the pool of possible suspects in a given case can then increase (2008).

3.4 Medicine

Advances in human genome sequencing have contributed to many aspects of the medical field including disease treatment, cancer research, personalized drug development, and vaccination. This has led to a new personalized approach in medicine that is specific to the patient and the disease in question.

3.4.1 Medical Diagnosis, Treatment, and Cancer Research

Generally speaking, personalized medicine involves the early prediction and detection of certain diseases, treatment and the development of drugs specific to the patient. Personalized medicine covers a wide range in medical process, from making predictions about an individual's susceptibility to a certain disease, to the actual treatment (Duke Medicine, 2011).

Prenatal testing is another area where the introduction of DNA sequencing has the potential to create large improvements for the patient and the clinician (Wright, et al., 2011). Currently, prenatal testing is done using chorionic villus sampling (CVS) and amniocentesis, which are both invasive ways to acquire a sample of the fetal DNA from inside the mother, and can involve risks such as miscarriage and tissue irritation. Research is being done on cell-free fetal DNA (cffDNA) that circulates in maternal blood plasma during pregnancy (2011). CffDNA is from the placental cells and thus could allow for prenatal testing that only requires a blood sample from the mother, an assay to separate the cffDNA from the maternal cells, and an appropriate gene sequencer. Unfortunately, much more research is required in this field before this future application can be materialized (2011).

In addition, there are many mutants of the human immunodeficiency virus (HIV) and many have become resistant to commonly used drugs. It has become a practice to sequence the specific mutant of the virus in order to effectively treat the patient (Intelligent BioSystems, 2011). However, improvements in sequencing technologies are desired in order to be able to measure low population mutants (Intelligent BioSystems, 2011).

Furthermore, the advances in genome sequencing technologies, especially those regarding frequency of error, speed, and cost per genome have allowed gene sequencing to become a part of treating many cancer patients. Mutations that lead to cancer vary so highly across patients that some researchers believe that each patient's genome needs to be sequenced in order to provide optimum cancer treatment (Intelligent BioSystems, 2011). The Cancer Genome Anatomy project has been launched by the National Cancer Institute to understand the varying mutations that lead to cancer (2011).

3.4.2 Personalized Drug Development

Personalized medicine can be defined as a special kind of medicine that allows for the customization of healthcare by using the genetic or related information of the patients in question (Duke Medicine, 2011). Pharmacogenomics is a field regarding the study of the way genetic variations affect how people respond to drugs (2011). As a result, pharmacogenomics and advances in DNA sequencing have built the foundation to the personalized medicine industry.

To make personal drug development a realizable idea, several factors must be taken into account:

- ❖ Genetic tests to prove the relationship between an individual's genes and a given disease are needed.
- ❖ It is necessary to successfully locate the genetic variant that causes the disease.
- ❖ A drug needs to be developed that targets either the deficient phenotype of that gene, or provides direct gene therapy.
- ❖ The response of an individual to the personalized drug must be analyzed for safety and effectivity.

3.4.2.1 Current Development and Data

Since the completion of the Human Genome Project by the U.S. Department of Energy and the National Institute of Health, the idea of developing medicine tailored to specific patients according to their genome has become feasible. One of the initial goals of the HGP was to associate particular genes to specific diseases (Gillham, 2011). Table 9 below compares genomes between humans and other organisms.

**Table 10: Genome Sizes and Gene Density in Humans as Compared with Other Organisms
(Data from Human Genome Project, 2003)**

Organism	Estimated size (base pairs)	Estimated gene number	Average gene density	Chromosome number
<i>Homo sapiens</i> (human)	3.2 billion	~25,000	1 gene per 100,000 bases	46
<i>Mus musculus</i> (mouse)	2.6 billion	~25,000	1 gene per 100,000 bases	40
<i>Drosophila melanogaster</i> (fruit fly)	137 million	13,000	1 gene per 9,000 bases	8
<i>Arabidopsis thaliana</i> (plant)	100 million	25,000	1 gene per 4,000 bases	10
<i>Caenorhabditis elegans</i> (roundworm)	97 million	19,000	1 gene per 5,000 bases	12
<i>Saccharomyces cerevisiae</i> (yeast)	12.1 million	6,000	1 gene per 2,000 bases	32

Among the genomes researchers have sequenced, more than half of them have been found to be made up of repeating sequences. A large amount of research shows that most genetic diseases are caused by mutation in these regions (Gillham, 2011). In fact, in as early as 1993 scientists were able to successfully associate two copies of the APOE4 variant mutation with Alzheimer's disease (Angier, 1993). This study showed that people with this certain gene variation are eight times more likely to get Alzheimer's disease, and triggered the development of personalized drugs for the treatment of Alzheimer's (1993).

Currently scientists have developed several therapies, many of which have been used in practice. In 2006, 13 prominent examples of personalized drugs, treatments and diagnostic products were available (Personalized Medicine Coalition, 2011). By the end of 2011, this number increased to 72, of which the average increasing rate is about 40.8% (2011). In addition, as of 2011, 33 pharmacogenomic biomarkers are included on FDA-approved drug labels (Personalized Medicine Coalition, 2011). Most personalized drugs target a specific biomarker and the group of people who have the given biomarker can possibly benefit from using the drug (2011). Furthermore, the Personalized Medicine Coalition stated in their annual report that 30% of all personalized treatments in late clinical development, 50% of all treatments in early clinical development, and 60% of all treatments in preclinical development now rely on biomarker data

(2011). In the future, creating a database of biomarkers and their corresponding personalized drugs could significantly improve disease diagnosis and treatment in terms of speed and efficiency.

Currently there are several incentives that encourage the development of personalized drugs . The first incentive is the decreasing cost of genome sequencing. Figure 19 is a graph indicating the decreasing sequencing cost per genome from Sept, 2001 to Sept, 2011, with data courtesy of the National Human Genome Research Institute.

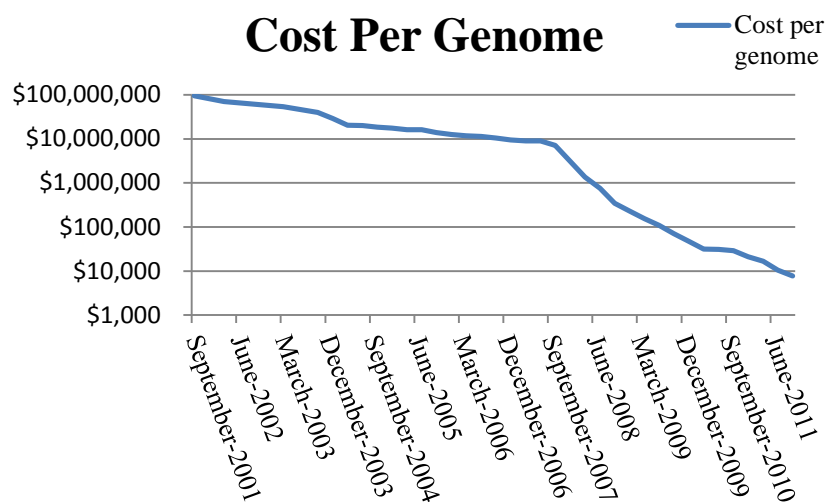


Figure 19: Cost per Genome from September 2001 to September 2011
(Data from National Human Genome Research Institute, 2011)

It is clear that the cost per genome is decreasing every year since 2001. However, the next generation sequencing technologies, which commercialized around 2007, have caused increased savings in recent years.

The second incentive encouraging personalized drug development is that there is an increased amount of investment in this field. It has been reported that Siemens signed its first personalized medicine diagnostics deal with two drug production corporations, both of whose main goal is to develop genetic tests that will determine which patients will benefit from which therapies (Bloomberg, 2012). Furthermore, according to the Personalized Medicine Coalition, there has

been a 75% increase in the investment for personalized drugs over the last five years (Personalized Medicine Coalition, 2011). Most of the investment is towards genetic testing in order to associate certain mutant variants with diseases. Moreover, according to TriMark Publications in 2011, the market for molecular diagnostics, which includes tests related to that of personalized drugs, will more than double in the U.S. to \$5.5 billion in 2016 (from its \$2.5 billion in 2010) (Bloomberg, 2012).

3.4.2.3 Future Difficulties

Unfortunately, there are many limitations to personalized drugs, including technical and social barriers. Currently, most of the personalized medicine available is not 100% guaranteed to work on the general population. For example, research has shown that the drug Xalkori, developed by Pfizer Inc., might help patients who have non-small cell lung cancer (NSCLC) (Personalized Medicine Coalition, 2011). However, only about five percent of NSCLC patients can be effectively treated with Xalkori. These patients have tumors resultant of a mutated version of the anaplastic lymphoma kinase (ALK) gene (2011). Due to this low percentage of effectiveness, this drug is often the last choice in clinical settings (2011).

Moreover, the progress in identifying important genetic variants is a bit slow due to the fact that there is usually little evidence showing that a common genetic variant will increase the risk of a given disease (Garber, 2009). An additional issue is that there has been little study to address the clinical utility of personalized medicine (2009). For example, it may be difficult for a doctor to convince his or her patients to spend the additional money for a genetic test in the hopes of receiving personalized medicine, especially if the doctor does not know how well the drug will work.

Finally, many consumers are concerned that this area of pharmacology has not matured enough to be deemed completely safe and effective. In addition, personalized drugs require not only the patient's sequencing information, but it also relies on the development of many other areas such as understanding the functions of certain regions of DNA and RNA.

3.4.3 Disease Prevention and Vaccination

Risk assessment is the most widely-recognized application of current technologies. The tools that enable risk assessment and predication include health risk assessment, family history genome information and clinical decision support (Bauman, 2010). Some of the applications of genomic and personalized medicine include disease susceptibility, diagnosis and prognosis, pharmacogenomics and monitoring (2010). To put these applications into practice, genetic testing is sometimes integrated into clinical practice.

Bacterial genome sequencing has enabled the development of effective vaccines. Two examples are vaccines for bacterial meningitis and for the many diseases pertaining to the Streptococcus strains (Tettelin and Feldblyum, 2009). Essentially, the knowledge of the DNA sequences allowed Tettelin and his colleague to find the resulting proteins, proteins functions, and the antigens of these bacteria that can be identified by the human immune system. The antigenicity (ability of structural and chemical structures on the bacterium to create an adaptive immune response), and pathogenicity (ability to cause disease) allows researchers to develop vaccines that can safely allow the creation of memory cells, which have the potential to reduce the symptoms and effects upon a second infection with the same strain due to the faster proliferation of antibodies by B lymphocytes (Bauman, 2010). Specifically, the vaccines developed for meningitis and streptococcus related diseases are made as “cocktails” that are a combination of denatured, attenuated, and/or modified antigens of these strains (2009).

3.5 Advances in Genetic Research

The advances made in second and third generation sequencing technologies paired with existing information on the human, bacterial, and viral genomes have enabled the identification of genotypic and phenotypic variances resultant of nucleotide deletions, insertions, inversions, frameshifts, and splicing (Tettelin and Feldblyum, 2009). Finding these differences and the relationships between genotype and phenotype will allow for further development of personalized medicine, better diagnosis, epidemiology, and disease prevention through more effective vaccines. For example, the 454 method was used to find and map more than 1000 structural differences between two humans that were larger than 3000 bases, while using a standard human genome as the reference (2009). The 454 method was also used in combination

with Illumina sequencing technology to detect regulatory RNAs that play a role in gene expression (2009).

Thus, molecular biology has been significantly impacted with the advances in genome sequencing technology because it is now easier to manipulate genes due to the knowledge of their sequences. Manipulating genes allows the discovery of gene function, regulation, and interaction, in addition to protein function, product, and interaction (2009). The following is an overview of various applications of second generation sequencing technologies in genetic researches.

3.5.1 Mutation Discovery and Defining Variability Across Human Genomes

Next generation sequencing technologies facilitate the discovery of mutations that determine phenotypes (Mardis, 2008). Previously, discovery of mutations were usually conducted with a direct focus on select regions of a genome of interest. However, the high speed of new sequencing technologies has made whole-genome search for mutations possible. Unfortunately, the short read length from the next generation methods, such as that of Helicos BioSciences, imposes certain limitations to this approach (2008).

These advances in gene sequencing have also led to variability studies across human genomes. These studies are based on population levels. Currently, large numbers of human genomes can be sequenced to provide data for variability studies (Mardis, 2008). Rare alleles can be detected at the sequence level, while allowing for the capturing of large insertions and deletions at the genome scale (2008).

3.5.2 Sequencing Clinical Isolates in Strain-to-Reference Comparisons

For the study of pathogens, such as bacteria and viruses whose genomes mutate constantly, researchers would like to obtain rapid data about antibiotic susceptibility and resistance (Mardis, 2008). The advances in sequencing technologies assist in this by providing complete genome sequences of different strains (2008). Since all of these sequencing methods obtain the fragment library from one single genome, variants in the clinical strains can be quickly identified.

3.5.3 Enabling Metagenomics

Metagenomics refers to the study of DNA fragments directly taken from environmental microbial and viral populations, instead of a cultured lab setting (Mardis, 2008). Samples are taken directly from the environment, so that the sequence reads are in rough proportion to the population frequency of each microbe, from which relative abundance can be estimated (2008). This concept is illustrated in Figure 20, which is also a graphic from the team's final set of videos. A particular example is the study of billions of microbial species that humans live in symbiosis with, called the human microbiome. The third generation sequencing methods in particular demonstrate extremely high efficiency with short bacterial genomes because of the short read lengths produced by these sequencers, ideal for facilitating metagenomics studies.

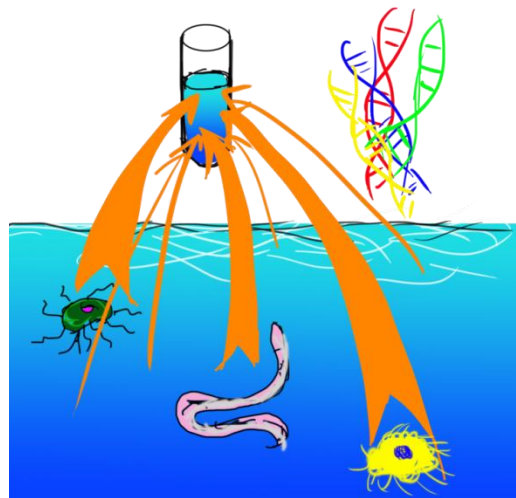


Figure 20: An Illustration of an Example of Metagenomic Studies

3.5.4 Defining DNA–Protein Interactions and Regulatory Protein Binding

DNA-protein interactions refer to the nuclear interactions between DNA and proteins that control DNA packaging into histones and/or the transcription of DNA into mRNA (Mardis, 2008). Previously these studies were only conducted on simple organisms with small genomes. However, with the assistance of second and third generation sequencing technologies researchers are able to understand the interaction processes in complex organisms with large genomes (2008).

Furthermore, the transcription of genes is regulated by protein binding sites on DNA. These binding sites are specific sequences recognized by regulatory proteins (Bauman, 2010). Using hybridization of fluorescently labeled DNA fragment enables genome-wide evaluation of binding sites (Mardis, 2008). The high throughput of next generation sequencers can provide enough data for all sites in the human genome within a single instrument run (2008).

3.5.5 Exploring Chromatin Packaging and Noncoding RNAs

Chromatin packaging is the packaging of genomic DNA into histones, which largely determines the transcription of genes (Bauman, 2010). Using next generation sequencing platforms, researchers are now able to better understand the histone binding locations genome-wide (Mardis, 2008). These advances have also allowed for the discovery of noncoding RNAs and their locations, which include regulatory RNAs (2008). These types of RNA serve important functions in the cell cycle, in fact, mutations in noncoding RNAs are closely related to cancer and other human diseases (2008).

3.6 Evolution

Advances made in genome sequencing technologies have also affected the progress in the field of evolutionary biology. The practice of comparative mapping is becoming more commonplace, and promises to have a huge impact in both healthcare and academic applications. By comparing the complete genomes of livestock, it has been discovered that the genetic mutations responsible for muscular dystrophy and acute stress syndrome are the same ones that have an effect on malignant hypothermia in humans (O'Brien, et al., 1999). Similar mapping of the X and Y chromosomes in marsupials have shed light on the origin and divergence of mammalian sex chromosomes (1999). It is also interesting to note that as few as 13 translocation mutations separate the human genome from that of the cat family (1999). It is expected that further comparative mapping will reveal whether the linkage associations that are depicted are accidental or whether they have been selected and retained due to dependence on those genes. Comparative mapping also creates the opportunity to highlight important genomic events such as epidemics, reproductive isolation, and survival in terms of their effect on forming new species (1999).

Second and third generation sequencing technologies that specifically identify SNP's are also having a great effect in determining positive selection within ecosystems. In humans, this data is being used to identify loci where particular alleles are being favored (Voight, et al., 2006). These alleles have not yet reached fixation, meaning that this information can be used to show which environmental and regional factors push alleles into fixation (2006). One example is the study of alleles controlling chemosensory perception, gametogenesis, spermatogenesis, and fertilization. These have been identified as targets of selection in human-chimpanzee genetic divergence, and data suggests that these are relatively recent selections (2006).

The field of phylogenetics, the study of the evolutionary development of organisms, relies heavily on a single assumption: the species observed today can be symbolized as leaf nodes in a developmental tree in which ancestral species are internal nodes (Siepel, 2009). As one travels along the branches of this tree, a genome changes through mutations until it reaches the genome of a modern-day species. Factors such as topology and length of the branch can determine the ancestral sequence of a particular species (2009).

The purpose of the statistical modeling and population considerations that are taken into account during the analysis of these trees is to attempt to reconstruct the ARG, or Ancestral Recombination Graph, for a species (2009). To clarify, this graph is not a tree. It begins with a single chromosome, and a tree can be extracted from it by taking a particular fork at any recombination event in the ARG (2009). These ARGs take into account population sizes, geographical barriers, and other causes of recombination that are possible at any evolutionary fork. The main caveat in this method is that if a certain number of genealogies (at least 4) are present within the same species, the wrong "topology" may be favored, resulting in the wrong phylogeny for a species (2009). This has been shown to happen rarely though, but more research is being done regarding the development of a more consistent estimator.

In the end, this method shows promise for the prediction of evolutionary similarities between humans and different species of primates. However, substitutions may be over-counted, resulting in a bias towards the prediction of genes and negative/positive selection (Siepel, 2009).

Unfortunately, the ARGs that link humans to primates on an evolutionary scale are difficult to

build based on current genomes alone (2009). Developing a method to construct this to a high accuracy would open up a large realm of possibilities, but before this can be done, computational intractability and difficulties in modeling biological phenomena (such as the difference between species in recombination hot-spots) must be overcome (2009). The possibilities for genomic impact on evolutionary biology are huge, and it could be possible to further trace the phylogeny of a variety of organisms if these hurdles are overcome. The implications of this research hold a particular impact on religion, science, and the way we view the environment.

3.7 Agriculture

Since the 1990s, sequencing technologies have been widely used in the agriculture industry. Using genetic information, scientists can identify genes that are responsible for specific phenotypic traits. Any beneficial genes (from plants or animals) discovered from these studies can be incorporated into our food sources if deemed safe (World of Genetics, 2006). Thus far scientists have sequenced the large majority of agricultural crops and have discovered a variety of important genes including those related to bacterial resistance, growth, and taste, which has led to the large amount of transgenic products currently available (2006).

One of the aims of transgenic products is to increase resistance to pests, such as bacteria and insects. By 1998, there were approximately 175 registered bio-pesticide active ingredients and 700 products (World of Genetics, 2006). In the case of cotton, the use of *Bacillus thuringiensis* in crops reduced the amount of chemical pesticides used by 450,000 kg in 1998 (2006). It can be seen that transgenic products featuring pest resistance have benefited people greatly by reducing costs and protecting the environment, thus making it extremely attractive to manufacturers and environmental protection agencies (2006).

Another related that receives great interest is the development of transgenic products to provide additional nutrition value. It is thought that the problem of hunger and malnutrition in underdeveloped countries can be addressed by developing transgenic rice with additional nutritional content (Baggott, 2006). Fortunately, most of the technical problems to develop such crops have been solved. The idea and details of “golden rice” was first published in Science in 2000 (Ye, 2000). This variant of rice is designed with a precursor of pro-vitamin A, and it is hoped that this will help prevent blindness and death due to a deficiency of vitamin A, which is

very common in developing countries (2000). Even though more varieties of transgenic rice are being developed and improved upon, “golden rice” raises some social concerns. In some cases, people have concerns regarding the safety of “golden rice” (Ingo, 2010).

In 2006, the group of countries growing 97% of global transgenic crops was comprised of the United States (53% of crops), Argentina (17%), Brazil (11%), Canada (6%), India (4%), China (3%), Paraguay (2%) and South Africa (1%) (Cornejo, 2011). Figure 21 below displays a graph from the USDA (United States Department of Agriculture) indicating the adoption of transgenic crops in the USA from 1996 to 2011. This graph illustrates how transgenic crops now account for more than two-thirds of all crops grown in the United States, which is resultant of the fact that people in the U.S. generally accept the adoption of genetically engineered crops (2011).

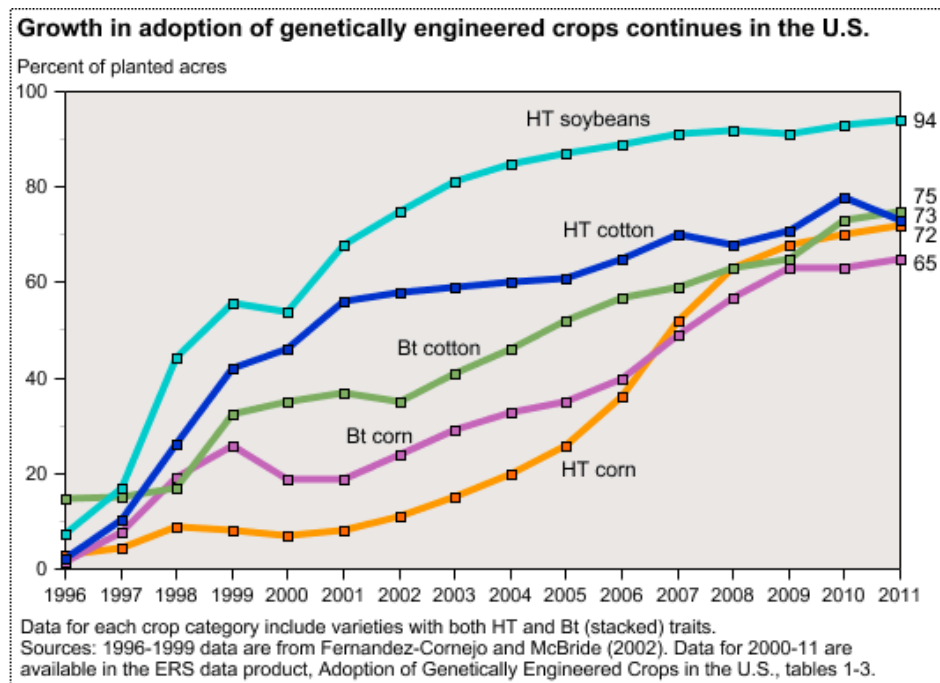


Figure 21: Adoption of Transgenic Crops in USA from 1996 to 2011

(Cornejo, 2011)

Transgenic foods can have many benefits including having additional nutritional value, and in some cases having better taste, which are both characteristics that have been found to be important to most consumers (2011).

However, there are also a lot of concerns regarding transgenic foods, including allergy and safety risks, possible harm to other organisms in the environment, and ethical issues such as tampering with nature through genetic engineering across species (U.S Department of Energy, 2008). In fact, a poll by Eurobarometer showed that the many European consumers find transgenic food dangerous (Baggott, 2006). Unfortunately, product safety is difficult to prove without long-term testing, so it is hoped that data in the years to come will enable consumers to make a more informed decision regarding the safety of ingesting transgenic crops. In addition, it seems that many countries are divided into two parties regarding the legislation of transgenic food (U.S. Department of Energy, 2008). While transgenic crops are adopted widely in countries such as the U.S, Argentina, Brazil and Canada, a lot of countries have banned transgenic food, such as France, Ireland and Russia (2008).

Chapter 4: Surveying WPI Students and Faculty

Undergraduate students at Worcester Polytechnic Institute were surveyed in regards to their interest, opinion, and familiarity with recent genome sequencing technologies and their various impacts. Members of WPI faculty were surveyed in regards to their preferences for various aspects of our final deliverables as well as their interest in using our educational materials in their course(s). The information obtained from these two surveys enabled the team to make the educational materials more relevant and useful for the target demographic. Both of these studies were approved to be exempt from further review by the WPI Instructional Review Board (IRB).

4.1 Creating and Implementing the Student Survey

This study primarily involved a nomothetic approach to using survey methods in order to answer the following questions:

1. Are students familiar with recent genome sequencing technologies and their impacts on government, medicine, research, and society?
2. Which impacts of genome sequencing technologies are students most interested in learning more about?

The team developed a list of questions to address these topics, revised them for clarity, and then proceeded to obtain an IRB exemption from further review. After obtaining this approval, the questions were inserted into an online form using SurveyMonkey, and distributed through email to the WPI moderated standing list for undergraduate students. All of these students had the option to participate and were told that their responses would be kept anonymous and confidential. A copy of the student survey can be found in Appendix I.

Basic demographic information was obtained through the survey for the purposes of categorization and data analysis. The final draft of the student survey is included in Appendix I. No contact information was obtained from the student survey.

The responders were grouped into categories based on their major (biology related majors vs. majors unrelated to biology) and the overall opinion, familiarity, and interest of each group was to be compared and then incorporated into creating a finished product that reflects the needs and

interests of the students most likely to encounter this video in their future courses. The survey is primarily quantitative and option-based in order to prevent the survey from being too time consuming and involved for participants. The quantitative statistics will allow for generalizations to be created in order to answer the two primary research questions.

4.2 Results and Discussion of the Student Survey

The team received 243 responses from WPI students in the first 24 hours since sending out the survey invitation. This response rate was higher than the team's expectations.

4.2.1 Respondent Demographics

Generally speaking, the respondents to the student survey have a uniform class distribution, as was desired by the team. However, the mode of the data is the junior class. This may be resultant of the fact that all four team members are juniors as well, and thus their names on the survey email may have been recognized by their fellow 3rd year classmates, moreso than students from other years. Yet, statistically speaking, there was no significant difference between the four groups (Freshmen, Sophomores, Juniors, and Seniors).

Respondents were also asked of their major to see whether having a biology background impacted their responses. It was found that one-third of the students were biology related majors, such as Biology and Biotechnology (BB), and Biomedical Engineering (BME). The team was specifically interested in seeing whether students having a biology related educational background would exhibit more confidence in their understanding of DNA sequencing technologies, as well as whether they would show a higher interest in related topics such as the impacts on disease diagnosis, treatment, and prevention.

It is also important to note that the respondents are restricted to WPI undergraduate students. As a result, this factor needs to be taken into account when drawing conclusions from this component of the study.

4.2.2 Stated Familiarity and Interest Ratings on the Impacts of Genome Sequencing Technologies

In addition to the overall interest shown regarding the various social and medical impacts of the advances in genome sequencing technologies, the team also wished to compare the amount of interest shown by different groups of undergraduate students. Generally, most respondents (62.3%) have at least some interest in learning more about the technical components of modern DNA sequencing methods. This ratio is greater than what was expected due to the amount of prerequisite knowledge required in order to fully understand current sequencing methods. This response shows that these WPI students are aware of that the technology behind genome sequencing is advancing at a fast rate and are thus genuinely interested in learning more about it.

To check the specific interests shown in each area, the team made use of the crosstab feature of SurveyMonkey with one group made up of the biology related majors discussed previously, and the other group made up of those who have less of a biology related background. Overall, the team expected the biology related group to be more familiar with how new genome sequencing technologies work. However, the team was also interested in learning how each of these groups view the various impacts of these technological advances.

The respondents who were biology related majors, were found to have a higher confidence in their familiarity with the impacts relating to the medical field, such as personalized drug development and vaccination. Correspondingly, there were twice as many people expressing interests in these areas as opposed to the number of individuals expressing an interest in the impacts on the economy and legislation. It is worth noting that, though the economy is greatly affected by the advances in genome sequencing technologies, it received the lowest interest by both groups being analyzed. Some of these respondents discussed this with members of the team, and it was found that their low interest was resultant of their lack of awareness regarding how advances in genome sequencing affect the economy. This feedback indicates that one's familiarity on a topic can affect their interest level in regards to learning more about the topic. As a result, the team decided that it was important to include in our educational material topics that received low familiarity and interest rates rather than purely having topics with high interest, with the hope that our introduction to the topic will stimulate further interest.

For the respondents who are not biology related majors, the median familiarity rate is only 1 (out of 5) for the impacts on ownership and privacy, legislation, economy and personalized medicine. This result indicates that they are not aware of the impacts of new DNA sequencing technologies, especially in areas other than the medical field. This group reported to have the highest familiarity on the impacts on genetics and evolutionary studies, which received an average familiarity rate of 2.38 (5 being the highest).

In terms of the respondents' interest, impacts on ownership and privacy, cancer research and genetic studies were the areas that received the highest amount of interest: 36%, 40% and 38% of all respondents respectively. In contrast, only 25 respondents reported to be interested in the impacts on the economy.

In addition, two line charts were produced to compare the respondents' confidence in their familiarity with the various impacts on genome sequencing technologies (Figure 22 below), and the percent of interest shown by biology related majors versus non-biology related majors (Figure 23 below).

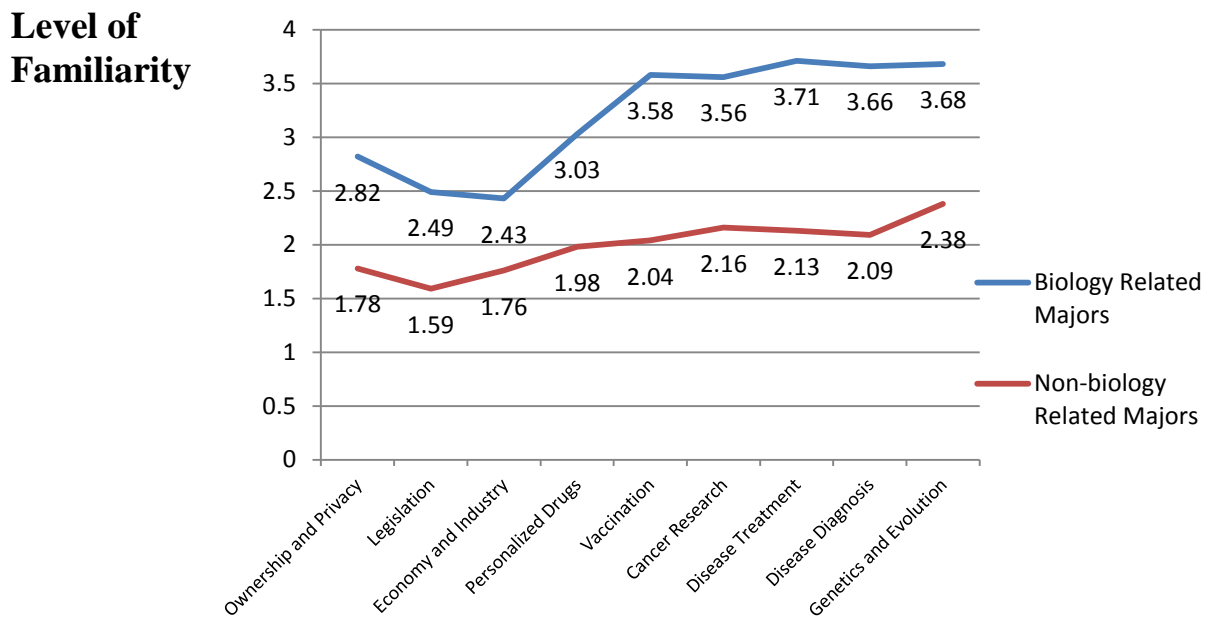


Figure 22: A Line Chart Comparing the Level of Familiarity on the Impacts of Sequencing Technologies of Biology Related Majors and Non-biology Related Majors

Percent of Interest

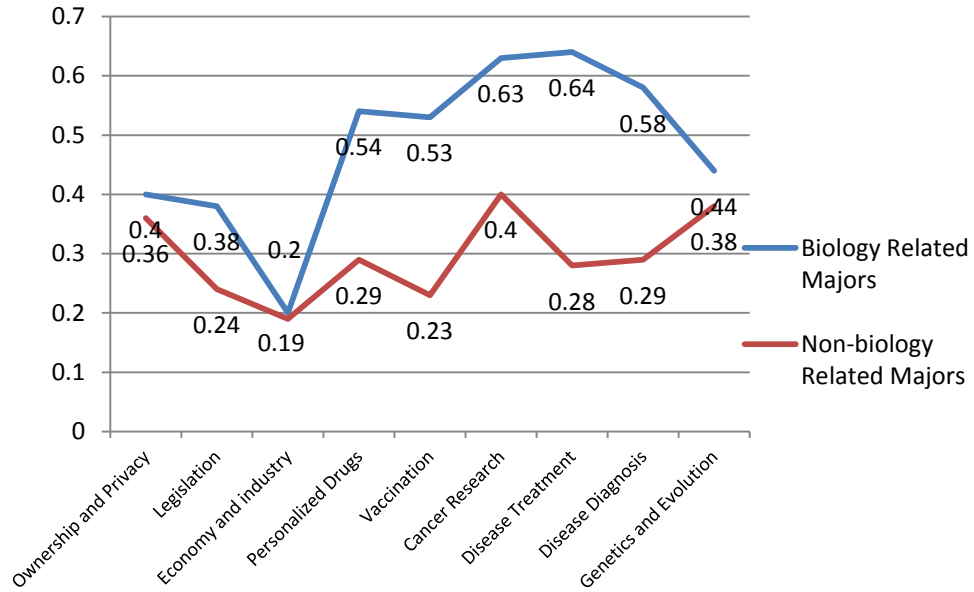


Figure 23: A Line Chart Comparing the Percent of Interest on the Impacts of Sequencing Technologies of Biology Related Majors and Non-biology Related Majors

A T-Test was performed to see whether the differences between the responses of biology related majors versus non-biology related majors were statistically significant, and indeed there was a statistically great difference between the responses of these two groups. Respondents with a background in biology tended to state that they were more familiar and interested in the topics compared to those who did not have as large of a background in biology. The team's goal was to depict the data using straight line segments and compare the shapes, so that an idea of whether this trend applies to a general population could be obtained.

For the stated familiarity rate, the line segments are very similar in the shape. This allows the team to conclude that the average college student is at least somewhat familiar with how advances of sequencing technology has impacted the field medicine, while other areas such as legislation, and economy, are not as widely known. However, this similarity cannot be found with regards to the interest percentage except for the fact that both groups have the least interest in the impacts on the economy. Because of their background knowledge, and possibly due to their career goals, biology related majors displayed a higher level of interest in the areas related to medicine and biological research.

4.2.3 Student Responses to the Changes Brought By Advances in Sequencing Technologies

The second part of the student survey was primarily focused on the reactions and opinions of the respondents in regards to the potential changes that could be brought upon by new sequencing technologies. Various situations were covered, such as the possibility of having individuals sequenced on a regular basis, and how people think about the potential social and legal issues that may arise, such as genetic discrimination. These survey questions can be found in Appendix I, and are noted as questions 7 through 15.

Statistical analysis was completed for all of these questions but only a selection of these findings is included in the sections below. Specifically, chi-square tests were used to indicate whether there was an association between the responders' personal views and the respondents' background (biology related majors versus non-biology majors). The null hypothesis in each case assumes that these two variables are independent. Fisher exact tests were also used when the sample size was small. As the name indicates, the test calculates the exact probability instead of a simple approximation produced by chi-square test. As a result, it is useful when there is not a large enough sample to make accurate approximations from. Fisher exact tests were performed using MATLAB code written by Giuseppe Cardillo.

4.2.3.1 Issues Pertaining to Legislation, Privacy, and Disclosure

Question 7 of the survey asked: *A person's genetic sequence can indicate whether they are predisposed to certain diseases. Do you think the government should prohibit health insurance companies from discriminating on the basis of one's genetic makeup?*

The great majority (73.4%) of the student respondents stated that there should be legislation prohibiting health insurance companies from discriminating on the basis of genetic. In addition, 88.9% of all respondents believe that a person's DNA sequence should remain private. These responses show that students are concerned about the proper use and disclosure of genetic information. As discussed in section 3.2, the United States already has such legislation in place, but a worldwide consensus is still far away. About one-fourth of survey respondents do not support having such legislation.

When comparing the responses from biology related majors versus non-biology related majors, the chi-square value χ_s^2 was found to be 1.31. The corresponding P value is 0.2524, which is greater than 0.2. As a result, the null hypothesis, which assumes an individual's major does not affect his or her choice, cannot be rejected. Based on this result, the team concludes that the majors of the respondents did not have a great influence on how they view genetic discrimination by insurance companies. Generally three-fourths of all respondents support the idea of prohibiting genetic discrimination by insurance companies.

The responses from different class years were also statistically compared. The chi-square value χ_s^2 was calculated to be 5.86. The corresponding P value is 0.1186, which is less than 0.2. As a result, the team rejected the null hypothesis that an individual's class does not affect his or her choice. However, based on these results, it can be said that there exists an association between the class year of respondents and their view on genetic discrimination. For example, more 3rd year students than any other class responded that they support the idea of prohibiting insurance companies from discriminating based on one's genetic makeup. This response was followed by 4th year students. It is interesting that upperclassmen seemed to be more concerned about genetic discrimination than first and second year students.

4.2.3.2 Issues Pertaining to the Economy

Question 9 of the survey asked: *The Human Genome Project was the second most expensive scientific program in human history. Do you think further government investment into this field will have a positive effect on United States economy?*

It was found that the respondents were divided in regards to their opinion about whether advances in genome sequencing will boost the economy. Approximately half of them (46.6%) believe it can improve the economy, while the other half (45.1%) hold the opinion that the investment can have positive effect but is not worth the investment. 8.3% stated that they were not sure about their opinion. Since these respondents reported to have a low level of familiarity with how DNA sequencing technologies affect the economy (lowest amount of reported familiarity and lowest percent of interest, as seen in Figures 22 and 23 respectively), it would be very interesting to conduct a survey after students watch the video in order to gather data on how

they review the relationship between advances in sequencing technologies and the economy after learning more about the topic.

It is important to note that more than half of all respondents stated that they were not sure or that they do not see the investment having a positive effect on the U.S. economy. The responses from biology related majors versus non-biology related majors were compared, and the chi-square value χ_s^2 was calculated to be 2.82. The corresponding P value is 0.2411, which is greater than 0.2. As a result, the null hypothesis, which assumes a person's major does not affect his or her choice, cannot be rejected. Thus, the team concludes that the majors of the respondents did not have a significant amount of influence on how they view this topic. However, the survey results show that respondents with more of a background in biology hold a more positive view of genetic technology.

Responses from different class years were statistically compared as well even though there was a smaller sample size. As a result, a Fisher exact test was chosen to be done. Using the Monte Carlo Method, the P value was calculated to be 0.5888, which is quite large. This indicates that there is almost no association between the respondents' class year and their view on how sequencing technology affects the economy. About one-tenth of all respondents believe that there is no positive outcome on the economy compared to the amount invested. Based on these results and the research the team conducted on the impacts on the economy, it can be deduced that many people are not aware of the positive impacts sequencing technologies have on our economy.

4.2.3.3 Opinions Regarding the Impacts on Medicine and Biologic Research

62.5% of respondents stated that they believed that new sequencing technologies will make it possible to prevent and cure cancer. At the same time, about 28.8% of the respondents said they were not sure if the technologies would be able to enable this. 8.7% believed that these technologies would not lead to the possibility of preventing and curing cancer. Currently, cancer is difficult to both prevent and cure, so these answers are not outliers at all.

In addition, 71.5% of respondents stated that further research into sequencing technologies will change the diagnosis and treatment of disease dramatically. However, there were still respondents who were not sure (19.8%) or thought that the diagnosis and treatment of disease would not change dramatically (8.7%).

Furthermore, 83.7% of respondents stated that they believe advances in sequencing technologies would bring further insights on the study of genetics and evolution. Compared to the previous two questions in this series, respondents seemed to have a more positive outlook regarding the potential of genetic and evolutionary studies. Given the many of the current genome sequencing technologies is being used to further the study of genetics and evolution, this feedback from the respondents was reasonable.

Question 10 of the survey asked: *Personalized drug development is made possible by recent genome technologies. These drugs are made on the basis of one's genetic makeup. It is potentially more effective, but it can cost more and has certain safety concerns. If you were given the option of using a personalized drug for a severe disease, would you use it?*

Approximately two-thirds of the respondents expressed that they would be willing to use a personalized drug for a severe disease, while one-third of the respondents were not sure about using this treatment. It is understandable that the high cost and safety concerns may make them people rethink the prospect of using personalized drugs.

The responses from biology related majors and non-biology related majors were compared once again. For this result there was a smaller sample size, and thus a Fisher exact test was used. The Monte Carlo Method was used and the P value was calculated to be 0.5888, which is quite large. This indicates there is almost no association between respondents' background and their willingness to use the personalized drug. This may be resultant of the fact that personalized medicine is still a relatively new area that people still need more information on before they can make an informed decision.

Again, responses from different class years were statistically compared using the Monte Carlo Method. Here the expected frequency for people answering no is so small that fisher exact test has to be used again. The P value was calculated to be 0.0717, which also indicates that there is no significant association between respondents' class year and their willingness to use personalized drugs. Generally there are very few respondents who directly state that they would refuse personalized medicine, but one-third of them are still not sure for due to considerations regarding cost, safety, and effectiveness.

4.2.3.4 Opinions Regarding Volunteer Participation in Genome Related Research

Question 15 of the survey asked: *Imagine that a large genome research project is being carried out on a population basis around the globe, and it is recruiting volunteers to have their genome sequenced as data for the research. The volunteers are anonymous and only demographical information is recorded. Would you volunteer to participate in the study?*

Nearly three quarters of the respondents were willing to participate in a worldwide genome research project by having their genome sequenced with only demographical information recorded. For this question there was no option regarding unsure responders, and thus, the other quarter of respondents reported that they would not be willing to participate in such a study. Based on this response, in theory, if such a research project is carried out in the future, there would be a sufficient number of volunteers in order to collect sound data.

4.3 Creating and Implementing the Faculty Survey

The second component of this study aimed to find whether faculty members teaching courses in biology, biomedical engineering, bioethics, government, and biostatistics would be interested in using our educational materials in their courses. In addition, the team hoped to use the results from the survey to envision how our potential educational materials would be presented to students and whether the video should be full length or broken into shorter segments based on topic. The overall opinion and interest of the participants were analyzed and then incorporated into the creation of the educational materials. Similar to the student survey, the faculty survey was made primarily option based in order to prevent the survey from being too time consuming and involved.

Similar to the methods involved in creating the student survey, the team developed a list of questions, revised them for clarity, and then proceeded to obtain an IRB exemption from further review. After obtaining this review, the questions were inserted into an online form using SurveyMonkey, and distributed through email to the moderated standing list for WPI faculty. All of these faculty members had the option to participate and were told that their responses would be kept anonymous and confidential.

Demographic information regarding the courses they teach was obtained through the survey for the purposes of categorization and data analysis. The final draft of the faculty survey is included in Appendix II. Unlike the student survey, faculty participants had the option to provide their email address if they were interested in previewing our educational materials for potential use in their courses, and/or if they were interested in providing suggestions to the team. However the email field was kept separate from the other response fields and as a result did not allow for identification of the participants based on their answers to the other questions.

4.4 Results and Analysis of the Faculty Survey

The number of responses obtained from the faculty survey was much lower than the student survey. Only 8 responses were collected from WPI faculty, and these respondents comprised the following departments: Biology and Biotechnology, Mathematics, Humanities and Arts, Computer Science, and Social Science and Policy Studies.

All but one participant showed at least some interest in showing the team's educational video to their class. In addition, all but one participant preferred the video to be broken up into segments based on topic rather than being a full length 40 minute video. This preference is probably resultant of the fact that class time is very limited, and as a result, having the video available online where students can be directed to watch it as a homework or extra credit assignment is much more convenient and time efficient. The team also determined that breaking the video up into segments allowed the video to be applicable to a larger range of courses, as course instructors could select the segments that are most relevant to their course for their students to view. For example, a professor of the social sciences department may select the segments pertaining to ownership and genome privacy, legislation, and economics for his/her course, while

a professor of biotechnology courses may select the segments pertaining to the genome sequencing technologies themselves as well as their various impacts on the medical field.

Furthermore, the team was interested to find that approximately half of the responders stated that they would show the video and/or specific video segments in class, while the other half stated that they preferred the videos be available online for students to view on their own. The team would have like to have more respondents in this case in order to have a statistically preferred viewing medium. However, the team addressed this issue by posting the final set of videos on YouTube so that course instructors could decide for themselves whether they wanted to show the video and/or segments in class or direct students to view it on their own. Creating YouTube links also enables instructors to link to it through the course site (myWPI in the case of this institution). Likewise to the question pertaining to the preferred viewing medium, approximately half of the respondents preferred to have supplemental written assignments while the other half did not. The team addressed this issue by creating a short assignment for each video segment and making it available to those who are interested in using it.

The impacts made by sequencing technologies on disease treatment as well as genetic and evolutionary studies received the highest amount of interest from the faculty. It is also important to note that the faculty members were interested in the impacts sequencing technologies have on the economy, since this received the lowest interest from the student respondents.

As a follow-up to the survey, one faculty member suggested that the team look into researching the social and government implications of genome sequencing technologies in other countries as well. It was further suggested that this information could initiate a classroom debate where students research the views of various countries and participate in a model UN type of discussion. The team responded to this suggestion and this information was incorporated into both the educational video through the Impacts on Worldwide Legislation segment, as well as the supplemental assignment for that segment, which includes debate topics.

Chapter 5: Developing Educational Materials

The following sections describe the methodology of developing the final set of online videos and their supplemental assignments. It is hoped that using these materials in conjunction will allow people to learn about this subject as well as demonstrate their understanding to an instructor.

5.1 Creating the Preview Video and Obtaining Feedback

The primary component of the educational materials developed was the set of online videos. Most of the images and text used in these videos were created by the team members, and all of the video animation, script writing, and narration was also done by the team. Adobe Photoshop CS 5.1 was used for image creation and adjustment, Adobe Flash CS 5.5 was used for 2-dimensional animation, and Adobe Premiere Pro CS 5.5 was used to compile the animation sequences, music, narration, and subtitles together. Autodesk Maya 2012 was used to create a model that was later used as an image in many segments of the video. In addition, Camtasia Studio 7.1.1 was used to record the narration as well as to create the tutorial in the Algorithms to Align Sequences video segment. Along with the creation of the visual components, a script was written for each segment of the video and then recorded. All together, approximately 35 minutes of film was made using this workflow. The sources for any borrowed images or music clips were credited at the end of each video.

In order to obtain suggestions for improvement during the video creation process, the team compiled a preview video that contained the following segments: First Generation Sequencing Technology, Second Generation Sequencing Technologies, Third Generation Sequencing Technologies, Impacts on the Economy, and Impacts on Personalized Medicine. These were in a preliminary form and thus did not yet have an introduction and credit sequence.

WPI professors who expressed an interest in previewing our material were provided with these video segments along with a feedback form. This form asked for their rating on the informational content, animations, use of subtitles, and the music, while asking how willing they would be to show videos similar to their students in the future. A copy of this feedback form can be found in Appendix III.

Overall, the response was very high, as the professors stated that the preview segments were well done, and that the videos provided a brief but engaging overview on each of the topics. In fact, the team was invited to a luncheon by the Bioinformatics and Computational Biology (BCB) program to share some of the segments from the preview video to students in the WPI BCB program, who also found our work thought-provoking.

5.2 Creating and Posting the Final Set of Videos

The group continued to create videos of individual segments, such as Impacts on Disease Prevention and Treatment, and Impacts on Forensics. A credit sequence was created and added to each video as well. These 12 segments were then combined into three broad sections: 1st, 2nd, and 3rd Generation Genome Sequencing Technologies, Social Impacts of Genome Sequencing Technologies, and Impacts of Genome Sequencing Technologies Related to Medicine and Biological Research. A general introduction segment was created as well.

At this point, the videos were uploaded onto YouTube as the 12 individual segments as well as the three compiled segments, and a playlist was created for each form (the 12 segments, and the three part video). The team chose YouTube as the medium of choice because it allows for maximum exposure and thus will enable a very large audience to benefit from the information. It was also requested of the team to send the final set of videos to the WPI Bioinformatics and Computational Biology program for possible future use.

In both cases, viewers are able to contact the team via email for a copy of the related assignments, in the case they are interested in using it in their classroom. Viewers are also encouraged to leave constructive comments for improvement (once approved). The description field of each video contains a message pertaining to this.

All of our videos can be accessed through the team's YouTube Channel:

<http://www.youtube.com/user/ImGenTechWPI/videos>. The final list of videos and their run time is as follows:

3 Part Video

- ❖ 1st, 2nd, and 3rd Generation Genome Sequencing Technologies (14:40)
- ❖ Social Impacts of Genome Sequencing Technologies (11:16)
- ❖ Genome Sequencing Technologies: Impacts on Medicine and Biological Research (7:47)

Individual Segments

- ❖ Impacts of Genome Sequencing Technologies: Introduction (1:38)
- ❖ 1st Generation Genome Sequencing: The Sanger Method (2:33)
- ❖ 2nd Generation Sequencing Technologies: Roche 454, Illumina, and Applied Biosystems (4:40)
- ❖ 3rd Generation Sequencing Technologies: Helicos tSMS and Pacific SMRT Technology (3:24)
- ❖ Using Algorithms to Align Sequences (4:20)
- ❖ Genome Sequencing Technologies: Impacts on Ownership and Privacy (2:10)
- ❖ Genome Sequencing Technologies: Impacts on Worldwide Legislation (4:24)
- ❖ Genome Sequencing Technologies: Impacts on the Economy (2:27)
- ❖ Genome Sequencing Technologies: Impacts on Forensics (2:32)
- ❖ Genome Sequencing Technologies: Impacts on Disease Prevention and Treatment (1:43)
- ❖ Genome Sequencing Technologies: Impacts on Personalized Medicine (2:27)

5.3 Creating Supplemental Assignments

Partly in response to the faculty survey and suggestions, the team also developed a short assignment for each segment of the final video. The final set of assignments is included in Appendix IV. Since users are able to view the videos by individual segments in addition to the three part video, questions were developed for each segment of the video. The answers to most of these questions are contained in our video, while others initiate students to conduct further research on a specific area. These assignments will be made available to instructors upon request. The description field underneath each video in the playlist will direct instructors to email imgentech12@gmail.com if they wish to access the assignments related to the video.

Most of the questions are either identification or open response type questions to test a student's understanding of the material presented in the video. A few questions ask the student to go

beyond the material in the video by conducting additional research. It is hoped that these research based questions will instigate students to learn more about this field, as the advances in these technologies may very well impact their lives at some point. In addition, due to the suggestion made in regards to creating debate topics for the classroom, the assignment for the Impacts on Worldwide Legislation segment includes three debate topics where students can conduct further research and represent different countries around the world in a discussion similar to the set up of a model United Nations.

5.4 Future Usage of the Set of Videos and Supplemental Assignments in the Classroom

Furthermore, some of the faculty who provided input to the team during the creation and development process have expressed tentative interest in using our educational materials in their related courses in the future. This includes faculty members from the Bioinformatics and Computational Biology Department (BCB), Biology and Biotechnology Department (BB), and Social Science and Policy Studies Department (SSPS).

Fortunately, since all four members of the team will remain at WPI for at least one more year, the team members will be able to assist in the transition and implementation process if applicable, and hopefully view the positive results from adding our materials to WPI courses. After graduating from WPI, the group plans on having at least one member of the team periodically monitor the team email address and the YouTube account in order to respond to any requests made by educators for more information.

Chapter 6: Conclusions and Recommendations

This chapter begins with a collection of the conclusions from the whole team. In addition, recommendations for future modifications and extensions of the project itself are also included.

6.1 The Team's Conclusions on the Impacts of Genome Sequencing Technologies

Through the process of researching and creating educational material for all the topics the team covered, many different aspects of the impacts gene sequencing has on society were uncovered. As such, it is the team's conclusion that each successive generation of gene sequencers has a significant impact on society. Technologies are becoming more streamlined and efficient, and more applications for them are being discovered daily. The genome industry has a positive effect on the United States' economy and has the potential to improve our quality of life through a variety of avenues such as personalized medicine, forensic applications, and genetic research. However, more legislation from nations around the world is necessary in order to preserve human rights and ensure that the information obtained from gene sequencing is used for good rather than for personal gain. The team concludes that the world is close to a time when gene sequencing will be fast, affordable, and widely used for the improvement of our lives.

6.2 Future Recommendations

Based on the team's experiences during the course of the project, we propose the following recommendations for those who wish to either conduct a project very similar to this and/or wish to continue this project:

- ❖ Encourage faculty members to respond to surveys and/or shorten the email and survey so that they are more likely to respond.
- ❖ Clarify between when survey respondents should answer with only one response or multiple.
- ❖ Create storyboards for each animation sequence as a team rather than solely by the animator in order to facilitate the process of combining images, voice, music, and animation all created by separate individuals.
- ❖ Use a software program such as Autodesk Maya to create three dimensional models and animations in order to add a more professional look to the final videos.

- ❖ Create a focus group of students to assess the understanding of various impacts of genome sequencing technologies before and after having individuals watch the final set of videos. Then analyze this data to determine whether the video proves to be effective in its goal to educate others about this subject.
- ❖ Consider the creation of other types of educational materials to communicate the same information, such as an interactive website, a computer game, or a mobile application.

References

- Abrahams, E., & Silver, M. (2009). *Personalized medicine for diabetes: the case for personalized medicine*. *Personalized Medicine Coalition*, 3(4), 680-684.
- Ahn, C. (2007). *Pharmacogenomics in drug discovery and development*. *Genomics and Informatics*, 5(2), 41-45.
- Angier, N. (1993). *Scientists propose novel explanation for alzheimer's*. *New York Times*, Retrieved March 2012 from <http://www.nytimes.com/1993/11/09/news/scientists-propose-novel-explanation-for-alzheimer-s.html>
- Ansorge, W. (2009). *Next-generation dna sequencing techniques*. *New Biotechnology*, 25(4), 195-203.
- Applied Biosystems. (2011). *Solid system mate paired libraries detect and define large genetic rearrangements*. Retrieved December 2011 from http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_057555.pdf
- Baggott, E. (2006). *A wealth deferred*. *Harvard International Review*, 28. Retrieved December 2011 from http://findarticles.com/p/articles/mi_hb137/is_3_28/ai_n29315531/
- Batzoglou, S. (2006). *Sequence alignment and dynamic programming*. *Stanford AI Lab*. Retrieved April 2012, from http://ai.stanford.edu/~serafim/CS262_2006/Slides/CS262_2006_Lecture2.ppt
- Bauman, R. (2010). *Microbiology with diseases by taxonomy*. 3rd. San Francisco: Benjamin Cummings.
- Battelle Technology Partnership Practice (2011). *Economic impact of the human genome project*. *Battelle.org*. Retrieved January 2012, from <http://www.battelle.org/publications/humangenomeproject.pdf>
- BBC. (2011). *Parents 'want child gene tests'*. *BBC News*. Retrieved March 2012 from <http://www.bbc.co.uk/news/health-13099090>

- Bloomberg.(2012). *Siemens signs first personalized medicine deals*, *Bloomberg Business Week*
- Blow, N. (2008). *Dna sequencing: generation next-next*. *Nature Methods*, 5, 267-74.
- Borry, P. (2008). *Europe to ban direct-to-consumer genetic tests?* *Nature Biotechnology*, 26, 736-737.
- Cappelletti, M. (2008). *Innovations in genomic analysis: downstream analysis of illumina sequencing data*. *Illumina Inc*. Retrieved December 2011, from http://mi.caspur.it/workshop_NGS09/docs/Cappelletti_NGS09.pdf
- Cornejo, J. (2011). *Adoption of genetically engineered crops in the u.s.. United States Department of Agriculture*. Retrieved April 2012, from <http://www.ers.usda.gov/Data/BiotechCrops/>
- Defense Tech Briefs. (2009). *Dna sequencing technique can produce genome in less than a minute*. *Defense Tech Briefs*, Retrieved March 2012 from <http://www.defensetechbriefs.com/component/content/article/4967>
- Dennis, W. (2009). *Next-generation sequencing data analysis and computation*. *Laboratory for Personalized Medicine Harvard*, Retrieved April 2012 from Next-Generation Sequencing Data Analysis and Computation
- Drmanac, R, et al. (2009). *Human genome sequencing using unchained base reads on self assembling dna nanoarrays*. *Science* 327(5961): 78–81.
- Duke Medicine. (2011). *Personalized medicine*. *U.S News*, Retrieved March 2012 from <http://health.usnews.com/health-conditions/cancer/personalized-medicine>
- Durbin, R. M., Eddy, S. R., & Krogh, A. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. (1st ed.). Cambridge: Cambridge University Press.
- Eddy, S. R. (2004). *Where did the blosum62 alignment score matrix come from?* *Nature Biotechnology*, 22, 1035-1036. Retrieved April 2012 from <http://www.nature.com/nbt/journal/v22/n8/pdf/nbt0804-1035.pdf>
- ESRC. (2008). In S. Sturdy (Chair). *Genetic suspects: emerging forensic uses of genomic technology*. Paper presented at ESRC Genomics Policy and Research Forum, Edinburgh, Scotland.
- Gaber, A. M., & Tunis, S. R. (2009). *Does comparative-effectiveness research threaten personalized medicine?* *New England Journal of Medicine*, 360, 1925-1927. Retrieved March 2012 from <http://www.nejm.org/doi/full/10.1056/NEJMp0901355>

- Garrity, G. M., et al., (2009). *Studies on monitoring and tracking genetic resources*. Informally published manuscript, Retrieved March 2012 from <http://www.cbd.int/doc/programmes/abs/studies/study-regime-05-en.pdf>
- GeneWatch UK (2006). *Genetic discrimination by insurers and employers: still looming on the horizon*. *GeneWatch*, Retrieved January 2012 from <http://www.genewatch.org/uploads/f03c6d66a9b354535738483c1c3d49e4/GeneticTestingUpdate2006.pdf>
- Geneway Research. (2009). *Next-generation genomic sequencing*. Retrieved January 2012 from http://www.genewayresearch.com/genomic_seq.php.
- Gillham, N. W. (2011). *Genes, chromosomes, and disease: from simple traits, to complex traits, to personalized medicine*. (1st ed.). FT Press.
- Gustav, W. (2007). *Bioinformatics explained: blast versus smith-waterman*. *CLC Biology*, Retrieved April 2012 from <http://www.clcbio.com/index.php?id=1098>
- Helicos BioSciences Corporation. (2010). *True direct dna measurement*. Retrieved November 2011 from [http://www.helicosbio.com/Portals/0/Documents/Helicos tSMS Technology Primer.pdf](http://www.helicosbio.com/Portals/0/Documents/Helicos_tSMS_Technology_Primer.pdf)
- Helicos BioSciences Corporation. (2010). *Helicos genetic analysis system*. Retrieved November 2011 from http://www.helicosbio.com/Portals/0/Documents/Helicos_SalesSpec.pdf
- Holland, M., McQuillan, M. & Boese, B. (2011). *Next generation sequencing of forensic dna loci using 454 life sciences technology*. Retrieved November 2011 from http://www.promega.com/~media/files/resources/conference_proceedings/ishi_20/oral_presentations/holland.ashx?la=en
- Illumina. (2011). *Casework genetics uses illumina technologies to decipher forensic dna mixture evidence*. Retrieved November 2011 from http://www.illumina.com/Documents/products/appnotes/appnote_forensics_dna_mixture.pdf
- Illumina (2012). *Hiseq sequencing systems*. *Illumina Inc.*. Retrieved January 2012, from http://www.illumina.com/Documents/systems/hiseq/datasheet_hiseq_systems.pdf
- Illumina (2011). *Hiseq systems*. *Illumina Inc.*. Retrieved December 2011, from http://www.illumina.com/systems/hiseq_systems.ilmn
- Intelligent BioSystems. (2011). *Ibs sequencing by synthesis technology*. Retrieved December 2011 from <http://www.intelligentbiosystems.com/index-1%20mod%201.html>
- Intelligent BioSystems. (2011). *Dna resequencing and clinical requirements*. Retrieved

- December 2011 from <http://www.intelligentbiosystems.com/index-2%20mod%201.html>
- Isaacs, J. M., Deng, J. W., & Lauffs, A. W. (2011). *China employment law update*. Baker & McKenzie, Retrieved March 2012 from <http://www.bakermckenzie.com/NLChinaEmploymentLawUpdateApr11/>
- Ju, J., et al. (2006). *Four-color dna sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators*. *Proceedings of the National Academy of Sciences*, 103(52), 19635-40.
- Life Technologies. (2011). *Discover the 5500 series solid sequencers*. Retrieved November 2011 from http://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_088661.pdf
- Mak, C. (2011). *Next-generation sequence analysis*. *Nature*, 29, 45-46.
- Mao, X. (1998). *Chinese geneticists' views of ethical issues in genetic testing and screening: evidence for eugenics in china*. *The American Journal of Human Genetics*, 63(3), 688-695.
- Mardis, E. (2008). *The impact of next-generation sequencing technology on genetics*. *Trends in Genetics*, 24(3), 133-141.
- Mohankumar, D. (2011). *Nanoball sequencing*. Retrieved December 2011 from <http://dmohankumar.wordpress.com/2011/06/05/nanoball-sequencing/>
- Mount, D. (2004). *Bioinformatics: sequence and genome analysis*. (2nd ed.). New York: Cold Spring Harbor Laboratory Press.
- Obenrader, S. (2011). *The sanger method*. Davidson.edu. Retrieved November 2011, from http://www.bio.davidson.edu/Courses/Molbio/MolStudents/spring2003/Obenrader/sanger_method_page.htm
- O'Brien, S. J., et al. (1999). *The promise of comparative genomics in mammals*. *Science*, 286(5439), 458-81.
- Pacific Biosciences. (2010). *Beyond next generation sequencing: single molecule real time (smrt) technology*. Retrieved November 2011 from http://microarrays.ucsd.edu/pacbio/PacBio_SMRT_Sequencing_Whitepaper.pdf
- Personalized Medicine Coalition (2011). *The Case for Personalized Medicine*. Retrieved February 2012 from http://www.personalizedmedicinecoalition.org/sites/default/files/files/Case_for_PM_3rd_edition.pdf

- Porter, G. (2010). *Genetic tests and insurance in japan: the case for a regulatory framework of choice. Predictive and Genetic Testing in Asia*. 145-165.
- Richter, B. and Sexton, D. (2009). *Managing and analyzing next generation sequence data. PLoS Computational Biology*, 5(6), doi:10.1371/journal.pcbi.1000369
- Roche Diagnostics Corporation. (2011). *Gs flx system*. Retrieved December 2011 from <http://454.com/products/gs-flx-system/index.asp>
- Rusk, N. (2009). *Focus on next-generation sequencing data analysis. Nature Methods*, 6. Retrieved February 2012 from <http://www.nature.com/nmeth/journal/v6/n11s/full/nmeth.f.271.html>
- Sandra, T. (2007). *Investigating genetic discrimination in australia. Australia Journal of Emerging Technology and Society*, 5(2), 63-83.
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). *DNA sequencing with chain-terminating inhibitors. Biochemistry*, 74(12), 5463-5467.
- Schadt, E. E., Turner, S., & Kasarskis, A. (2010). *A window into third-generation sequencing. Human Molecular Genetics*, 19(R2), 227-240.
- Shendure, J., and Ji, H. (2008). *Next-generation DNA sequencing. Nature Biotechnology*, 26, 1135-1145.
- Siepel, A. (2009). *Phylogenomics of primates and their ancestral populations. Genome Research*, 19, 1929-41.
- Stewart, K. B. (2010). *Genetic discrimination: australian experiences and policies. GeneWatch*, Retrieved April 2012 <http://www.councilforresponsiblegenetics.org/genewatch/GeneWatchPage.aspx?pageId=189>
- Tettelin, H., and Feldblyum, T. (2009). *Bacterial genome sequencing*, 551, 231-247.
- Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). *Clustalw: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Research*, 22(22), 4673-4680.
- U.S Department of Energy Genome Programs. (2008). *Genetically modified foods and organisms*. Retrieved February 2012 from <http://genomics.energy.gov>
- U. S Department of Health and Human Services, (2009). *"Gina" the genetic non-discrimination act of 2008: information for researchers and health care professionals*. Retrieved February 2012 from <http://www.genome.gov/Pages/PolicyEthics/GeneticDiscrimination/GINAInfoDoc.pdf>

- U. S National Library of Medicine. (2012). *The genetic information nondiscrimination act (gina)*. Retrieved from <http://ghr.nlm.nih.gov/spotlight=thegeneticinformationnondiscriminationactgina>
- Voight, B. F., Kudravalli, S., Wen, X., & Pritchard, J. K. (2006). *A map of recent positive selection in the human genome*. *PLoS Biology*, 4(3), 446-58.
- World of Genetics. (2006). *Agricultural genetics*. *World of Genetics*, Retrieved February 2012 from <http://www.bookrags.com/research/agricultural-genetics-wog/>
- Wright, C., et al. (2011). *Next steps in the sequence: implications of whole genome sequencing for health in the uk*. *PHG Foundation*, Retrieved December 2011 from <http://www.phgfoundation.org/reports/10364/>
- X Ye, S Al-Babili, A Klöti, J Zhang, P Lucca, P Beyer, I Potrykus. (2000). *Engineering the provitamin a (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm*. 287(5451), 303-305.
- Xu, M., Fujita, D., & Hanagata, N. (2009). *Perspectives and challenges of emerging single-molecule dna sequencing technologies*. *Small*, 5(23), 2638-49.
- Zimmerman, J. (2008). *Building a genome in a minute*. Retrieved March 2012 from <http://www.mdatechnology.net/update.aspx?id=a5362>

Appendix I: Student Survey

IQP: Impacts of Genome Sequencing Technologies

1. What is your class?
 Freshman Sophomore Junior Senior Graduate student
2. What is your major?
 BME BB Other Biology related majors
 Statistics Other Engineering Other Sciences Other majors
3. How many courses in Biology, BME, Biostatistics and/or Bioethics have you taken?
 0-2 3-5 5-10 10-15 15 or more
4. Do you have any background knowledge on the technological aspects of modern DNA sequencing methods? (eg. How a strand of DNA is sequenced)
 Yes No
5. Are you interested in learning the technological aspects of the modern DNA sequencing methods?
 Highly interested Some interest Little interest
 Not interested

6. Are you familiar with the following aspects in which potential impact brought by the new genome technologies would occur? In addition, are you interested in learning more about these aspects?

(Gradient: 1 for not familiar at all, 5 for very familiar, Check the aspects you are interested in learning about)

Impacts	1	2	3	4	5	Yes, I am Interested
Social impacts	-	-	-	-	-	-
Ownership, Privacy and Disclosure of personal genomes						
Legislation (Genetic Information Nondiscrimination Act, GINA)						
Economy and Industry						
Human Medicine	-	-	-	-	-	-
Personalized Drug Development						
Vaccination						
Cancer Research						
Disease Treatment						
Disease Diagnosis						
Scientific Research	-	-	-	-	-	-
Genetics and Evolution						
Agriculture						

Questions 7-15 are based on the aspects of the impacts mentioned above.

7. A person's genetic sequence can indicate whether they are predisposed to certain diseases. Do you think the government should prohibit health insurance companies from discriminating on the basis of one's genetic makeup?

Yes No

8. Do you think that a person's DNA sequence should remain private?

Yes No

9. The Human Genome Project was the second most expensive scientific program in human history. Do you think further government investment into this field will have a positive effect on United State's economy?
- Yes No
- It could have positive effect, but the investment would be more useful if it was allotted towards other research.
10. Personalized drug development is made possible by recent genome technologies. These drugs are made on the basis of one's genetic makeup. It is potentially more effective, but it can cost more and has certain safety concerns.
- If you were given the option of using a personalized drug for a severe disease, would you use it?
- Yes No Not sure
11. The occurrence of certain types of cancer is closely related with mutations in a person's genome. Do you believe that new genome technologies will enable researchers to find ways to prevent and cure cancer in the next decade?
- Yes No Not sure
12. Do you believe that further genome research will dramatically change diagnosis and treatment of disease in the next decade?
- Yes No Not sure
13. Do you think new genome sequencing technologies would bring insights on studies of genetics and evolution?
- Yes No Not sure
14. If you could get your entire genome sequenced for under \$100 someday, would you choose to be sequenced?
- Yes No
15. Imagine that a large genome research project is being carried out on a population basis around the globe, and it is recruiting volunteers to have their genome sequenced as data for the research. The volunteers are anonymous and only demographical information is recorded. Would you volunteer to participate in the study?
- Yes No

Appendix II: Faculty Survey

IQP: Impacts of Genome Sequencing Technologies

1. What department are you in?

2. What is your concentration/specialization?

3. What courses do you teach at WPI?
Undergraduate _____
Graduate _____
4. Would you be interested in showing a video to your class about recent genome sequencing technologies and their impacts on government, industry, research, medicine, and society?
 Highly interested Some interest little interest
 Not interested
5. If you are interested in showing such a video, would you prefer a full-length (40-50 min) video or videos broken into segments?
 Full length Segments
6. Would you prefer to show students the video in class, post it on myWPI or give students a YouTube link?
 In class myWPI YouTube link
 Other forms, please explain _____
7. Would you prefer to have questions relating to the contents of the video and answer keys as assignments/extra credit for your students?
 Yes No

8. Are you familiar with the following aspects in which potential impact brought by the new genome sequencing technologies would occur? In addition, are you interested in learning and/or teaching more about these aspects?

(Gradient: 1 for not familiar at all, 5 for very familiar, Check the aspects you are interested in learning about)

Impacts	1	2	3	4	5	Yes, I am Interested
Social impacts	-	-	-	-	-	-
Ownership, Privacy and Disclosure of personal genomes						
Legislation (Genetic Information Nondiscrimination Act, GINA)						
Economy and Industry						
Human Medicine	-	-	-	-	-	-
Personalized Drug Development						
Vaccination						
Cancer Research						
Disease Treatment						
Disease Diagnosis						
Scientific Research	-	-	-	-	-	-
Genetics and Evolution						
Agriculture						

9. **(Optional)** If you are interested in the video and would like to see a preview of it, please leave your email address so that we may contact you. _____

Appendix III: Feedback Form for Video Preview

Name:

1) How thorough did you find the information covered in the video?

Not informative 1 2 3 4 5 Very informative

2) Please rate the quality of the animations

Not good at all 1 2 3 4 5 Very good

3) Did you find the subtitles helpful?

Not helpful 1 2 3 4 5 Very helpful

4) Did you find the music distracting?

Very distracting 1 2 3 4 5 Not at all distracting

5) At this stage, how willing would you be to show/post a video similar to this to your class?

Not at all 1 2 3 4 5 Extremely willing

Additional Comments:

Please email the group with the feedback in the body of the email, or attached to the email.

Alternatively, you may print this out and give it to Professor Zheyang Wu. We appreciate your time and thank you for your feedback!

Appendix IV: Supplemental Assignments

Based on faculty interest, the team has developed supplemental assignments to support the content of our videos. The answers to most of these questions are contained in our video, while others initiate students to conduct further research on a specific area. These questions allow students to display their understanding of the content to the professor using it in their course.

Introduction

1. Who developed the chain termination sequencing method?
2. Name some of the companies who have developed second and third generation sequencing technologies.
3. How do you think these advanced genome sequencers will affect our society?

1st, 2nd, and 3rd Generation Sequencing Technologies

1. Based on the video segment about the Sanger Method, summarize how the chain termination method of sequencing works.
2. How did each of the second generation technologies mentioned in the video improve upon the Sanger Method?
3. What are the differences between the technologies developed by Roche, Illumina, and Applied Biosystems?
4. What are the similarities and differences between Helicos BioSciences sequencing technology and Pacific Biosciences sequencing technology?

Using Algorithms to Align Sequences

1. Why is developing algorithms important in modern biological data processing?
2. Name two algorithms widely used in computational biology.
3. There are many variants of BLAST; find information about each of them and explain how they are different from each other.

Social Impacts Regarding Genome Related Ownership, Privacy, and Disclosure

1. What are some of the issues with the transfer of genetic information between research institutions and between countries?
2. Imagine that you wish to have a gene therapist develop a pedigree and provide the chances that one of your children will be born with muscular dystrophy. What are some of the issues that will have to be addressed regarding the sequencing of the members of your family? How would you address the question of sequencing a deceased family member and their rights?

Impacts on Worldwide Legislation

1. Which country was first to create legislation against genetic discrimination? What is the full name of this act?

2. What are some of the factors that affect how countries deal with genetic discrimination??
3. Debate Topics

It is suggested that students research the views of different countries and represent them in a model UN type of debate to address the following questions.

- ❖ “Should we use human genome sequencing to enable improved disease diagnosis, treatment, and prevention?”
- ❖ “Should we ban or limit the use of genetic testing in order to prevent legal cases of genetic discrimination?”
- ❖ “Should there be a government agency or committee to oversee occurrences of genetic discrimination?”

Impacts on the Economy

1. Which genomic related area(s) currently offers the most job opportunities?
2. The Human Genome Project is one of the most invested projects by the U.S. government. What year did this project start? What year did the HGP end?
3. What are the results and implications for the funding of genetic research?

Impacts on Forensics

1. What are STRs and why are they involved in forensic studies?
2. What are some of the social and legal issues involved in sequencing individuals for forensic evidence?

Impacts on Disease Diagnosis, Treatment, and Prevention

1. For the most part, the sequencing of what organism(s) led to the development of vaccines?
2. Explain what a pedigree is, and how it allows someone to know the probability that a patient will develop a given disease?
3. What is the specific characteristic of HIV that makes it difficult to treat, and how have the advances in genome sequencing technology improved the treatment of AIDS patients?

Impacts on Personalized Medicine

1. Find information about an FDA approved personalized drug and explain how it works. (Please include what this drug targets and how it helps treat a given disease.)
2. Do you support the investing of large amounts of money into genetic testing to develop personalized medicine? State reasons for your argument.

Impacts on Genetics and Evolutionary Studies

1. Describe metagenomics.
2. Describe comparative genomics and provide two examples of applications of this.