- Blank -

AN ABSTRACT OF THE FINAL REPORT OF

Kelley M. Wanzeck
For the degree of Professional Science Master's in Applied Biotechnology
Presented on November 18, 2011

Title: Uncovering Biological Meaning From Genome-Scale Datasets:
A Challenge and an Opportunity

Internship conducted at:

Intuitive Genomics, Inc.
1005 N. Warson Rd., Suite 223
St. Louis, MO 63132

Supervisor: Dr. Doug Bryant, Chief Technology Officer and Co-Founder

Dates of Internship: 6/20/2011 – 9/30/2011

Abstract approved: _____
Dee R. Denver

As DNA sequencing technologies continue to advance, resulting in increased throughput and decreased costs, both the number of researchers utilizing these technologies and the quantity of data outputted by a single sequencing experiment will, likewise, continue to increase. Currently, DNA sequence data can be generated at a much faster rate than computational tools can be created for the management, storage, and analysis of these large-scale datasets. While large genome-scale datasets have the capacity to fuel the next generation of scientific discovery, leveraging these datasets requires that researchers identify ways to uncover meaning from the data. Both performance of a research project in Dr. Dee Denver's Laboratory and an internship experience with Intuitive Genomics, Inc provided me an opportunity for exploration of methods for uncovering meaning from genome-scale datasets.

Intuitive Genomics, Inc. is a bioinformatics services startup located in St. Louis, Missouri. Serving as the company's Marketing Manager, the ultimate goal of my internship project was significant contribution to the growth and success of the startup. Areas of responsibility included marketing, sales, strategic development of products and services, and administration. Key accomplishments included significant contribution of content to the company's newly launched website, effective market research and competitor analysis, development of marketing materials, successful submission of an application for funding, professional interaction with current and potential customers, and representation of the company at a vendor show. The internship experience exposed me to a dynamic entrepreneurial environment, increased my knowledge of computational science, gave me first-hand experience with a variety of business concepts, enhanced my written and oral communication skills, and increased my confidence in professional interactions with customers and colleagues. These new skills and exposures will prepare me well for a career with a scientific corporation. Intuitive Genomics benefitted from an employee's 100% dedication to the company. I served as an administrative contact, a presence in the office space, conducted a variety of pertinent research initiatives, developed meaningful content for the website, developed marketing materials and served as a point of contact for current and potential customers.

Uncovering Biological Meaning From Genome-Scale Datasets: A Challenge and an Opportunity

By: Kelley M. Wanzeck

A FINAL REPORT

Submitted to Oregon State University

In partial fulfillment of the requirements for the degree of

Professional Science Master's in Applied Biotechnology

Presented at 9 am, November 18, 2011

Commencement June 16, 2012

Professional Science Master's in Applied Biotechnology
Final report of Kelley M. Wanzeck
Presented on November 18, 2011

APPROVED:

Dee Denver, representing Molecular and Cellular Biology
Major Professor

Jeff Chang, Botany and Plant Pathology
Committee Member

Pat Hayes, Crop and Soil Science
Committee Member

I understand that my final report will become part of the permanent collection of the Oregon State University Professional Science Master's Program. My signature below authorizes release of my final report to any reader upon request.

Kelley M. Wanzeck, Author

Thanks to the Intuitive Genomics team for their guidance and support over the course of my internship project.

Thanks to Drs. Dee Denver, Jeff Chang and Pat Hayes for their service as members of my master's committee.

Table of Contents

List of Figures

List of Tables

List of Appendices

1. **Scientific Report**

1.1 Introduction

1.1.1 Identifying the Problem

With the advent of high-throughput DNA sequencing, costs associated with acquiring genome-scale datasets have decreased one hundred fold over the past few years. Simultaneously, advancements in these same DNA sequencing technologies have resulted in tremendous increases in the amount of data that can be generated in a single sequencing run. This increase in throughput and decrease in cost have facilitated explosive growth in biological data generation bringing both challenges and opportunities to the life sciences community. Thus, transitioning from raw data to meaningful results has become an arcane art.

Recent advancements in DNA sequencing technology are responsible for both the decrease in cost and increase in throughput resulting in the current emphasis on DNA sequencing in the scientific research community. The most impactful advancement has been the creation of next-generation sequencing technology, replacing previous sequencing methodologies namely, Sanger Sequencing. Next-generation technologies rely on sequencing-by-synthesis or sequencing-by-ligation chemistries, both permitting massive parallelization of the sequencing reaction. Amplification via bridge amplification or emulsion PCR prior to initiation of the sequencing reaction simultaneously increase the scale at which fragments of DNA can be sequenced. The three most prominent providers of DNA sequencing instruments include Illumina, 454 Life Sciences, and Applied Biosystems (Life Technologies). The HiSeq2000, FLX +, and SOLiD 5500xl, respectively, are changing the genomics landscape through both their availability to researchers and their scale of output. The continued advancement of these new technologies promises to facilitate a new age in scientific discovery.

One of the main challenges brought about by the explosive growth in biological data generation is the availability of computational tools and approaches for the management, storage and analysis of this large-scale genomic data. Currently, biological data can be generated much more quickly than can the tools necessary for the derivation of meaning from these immense datasets.

The massive genome-scale datasets outputted by DNA sequencing instruments hold promise to power the next generation of scientific discovery in biology, however, most scientists are severely ill-equipped to find meaning in this increasingly massive body of data, and often, are even unaware of what questions these data may address. More scientists than ever before have access to DNA sequencing technologies yet few have the expertise themselves or have access to the expertise necessary to interpret the large-scale output. Additionally, it is becoming less practical to hire an employee with bioinformatics expertise to manage a single lab's large-scale data. Not only does hiring an employee for this purpose result in a large monetary investment, but, training this employee to perform the necessary functions for the lab can halt progress toward publication and slow the overall productivity of the lab.

Although one option for data analysis, publicly available bioinformatics tools and pipelines commonly require deep or complete customization prior to use as tools are often poorly documented, exceedingly slow, un-optimized and require expensive computing hardware. Additionally, freeware tools often only narrowly address a biological question, preventing scientists from gaining a deep understanding of the meaning held within their data.

Ultimately, there exists a recognizable gap between the rate at which massive genome-scale datasets can be generated and the rate at which computational resources become available to support the management and analysis of this data. The current overarching challenge for the life sciences community is the generation of mechanisms with the capacity to uncover meaningful biological insights from massive genome-scale datasets. There are a variety of mechanisms through which researchers are attempting to overcome this challenge.

1.1.2 Potential Solutions

Uncovering meaningful insights from high-throughput sequence data has become an arcane art. Additionally, the insights each scientist desires are unique to their individual research projects. Researchers are experimenting with a variety of approaches in an effort to make sense of their large-scale data.

Through both my work with Dr. Dee Denver's Laboratory at Oregon State University and my internship experience with Intuitive Genomics, Inc, I have gained first-hand experience with a couple distinct approaches to the challenge of uncovering meaning form genome-scale datasets.

My work for Dr. Denver's Laboratory involved the annotation of *Caenorhabditis drosophilae (C. drosophilae)* genome contig assemblies. A contig can be defined as a contiguous sequence of nucleotides, representative of a smaller piece of a larger genome. The goal of the project was to gain an understanding of the organization of the genome content through the visualization of predicted protein-coding regions within each genomic contig. Visual images were created through use of WebGBrowse, a publically available, web-based genome browser created by The Center for Genomics and Bioinformatics at Indiana University. A variety of annotation tools, the specifics of which will be outlined below, were utilized to generate the required input for this web server and therefore arrive at this ultimate goal.

The visualization of predicted protein-coding genes within a genomic contig, the deliverable of my research project, is representative of the process of uncovering meaning from genome-scale data. Specifically, the work I completed comprised genome annotation, the final step in a fairly standardized process moving data from raw sequence reads to meaningful output.

This process is most simply depicted as a movement of large-scale data among data forms. Initially, the raw sequence data represents a massive series of called bases (nucleotide bases identified as adenine, guanine, cytosine, or thymine) organized into short sequence reads of a standard length. Prior to DNA sequencing, the genomic material is fragmented into small segments of DNA. Adapters are ligated to the ends of these fragments to facilitate the sequencing reactions. The resulting sequence reads represent the known series of bases comprising these small DNA fragments. Through the process of genome assembly, these sequence reads are assembled into longer consecutive stretches of sequence data dependent on the alignment of the sequence reads to a reference genome or through *de novo* methods. These distinct lengths of contiguous sequence data represent genomic contigs, or pieces of the larger genome. Next, annotation tools can be utilized to characterize the genomic content within each contig. For example, a genome annotation tool may predict sequences within each larger contig that have properties characterizing them as protein-coding. Often, annotation tools used to predict protein-coding regions of a contig output the boundaries of these protein-coding regions. This information allows for the generation of an appropriate input for genome visualization tools. The WebGBrowse tool utilized for my project outputted a visual depiction of the predicted protein-coding regions within every contig assembly. Other approaches may be taken to the characterization of genome content following assembly. Depending on the research project, investigators may be interested in locating transcription factor binding sites, regulatory RNAs, or sequence variations. Ultimately, uncovering meaning from DNA sequence data involves a progression from sequencing, to assembly, to genome annotation, visually depicted in Figure 1 below.

Figure 1: Standard Pipeline for Analysis of DNA Sequence Data

While identifying regions of each contig predicted to code for a protein is an important step in genome annotation, the visual depiction of protein-coding elements was not the overarching goal of my project. Protein-coding regions are only one specific element important for identification when annotating a genome. Therefore, outlining a methodology for the visual depiction of protein-coding regions within a contig was performed for the purpose of establishing a foundation upon which Indiana University's WebGBrowse tool might be utilized for visualization of more complex genome annotations. I aspired to learn the basic functionality of the WebGBrowse genome annotation tool such that the tool could be further exploited for more complex genome annotation by the Denver Lab in the future. Through the process, I explored one simplistic approach to the challenge of uncovering meaning from genome-scale data. The visuals created as a result of this project are much more meaningful than raw sequence reads for the purpose of better understanding the organization and functionality of the *C. drosophilae* genome.

The approach taken by Intuitive Genomics Inc. in an attempt to uncover meaning from large genome-scale datasets will be addressed in the "Business Report" to follow.

1.2 Materials and Methods

The genomic material utilized for this project represented nuclear sequence data from the *C. drosophilae* genome. The nematode genome was previously sequenced by the Denver Lab in a single lane of an Illumina flow cell. The genomic material was used to create a 300 base pair (bp) fragment library and 80 bp paired-end reads were sequenced by an Illumina HiSeq2000. Additionally, the sequence data was assembled by the Denver Lab utilizing the assembly program, Velvet, a new set of algorithms, collectively used to manipulate de Bruin graphs for genomic sequence assembly. Assembly via Velvet yielded 80 nuclear contigs. Lastly, the Denver Lab utilized a tool called blastx to confirm that all contigs were comprised of *C. drosophilae* genomic material as expected. The National Center for Biotechnology Information (NCBI) hosts the blastx tool, used to identify potential protein matches for a sequence of nucleotides through the query of a translated nucleotide sequence against a protein database (NCBI, 2009).

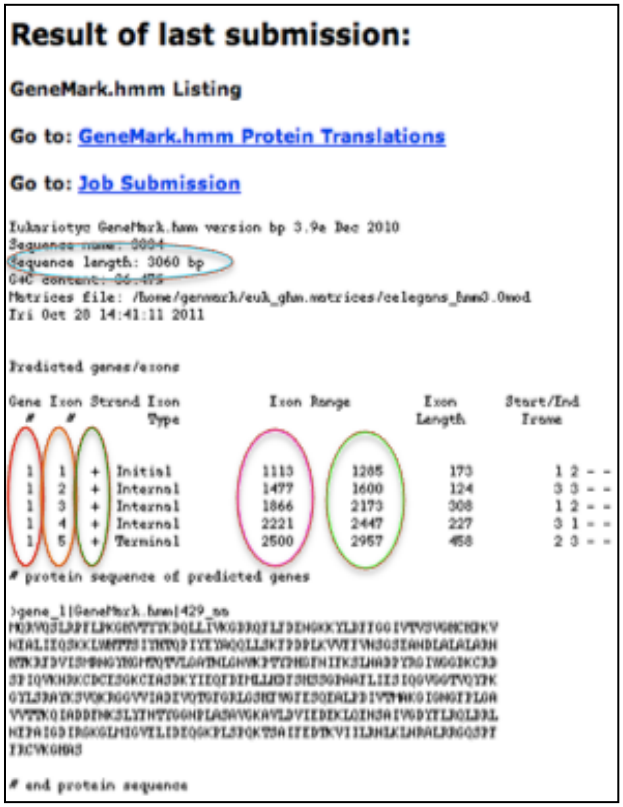
Following sequencing, assembly, and the initial blastx searches, the resulting 80 nuclear contigs were presented to me for continuation with the genome annotation. My goal was to gain an understanding of the genomic content within each contig, particularly protein-coding genes, and to visualize these protein-coding genes through use of WebGBrowse, Each visual would clearly depict the regions within a single contig predicted to be protein-coding. The process of obtaining the appropriate information for the generation of these genomic visuals can be broken into three unique steps seen below. Further, a second NCBI tool was utilized for the purpose of exploring the identity and function of the predicted proteins. This process is explained in a fourth step.

1. Obtaining Exon Boundaries
2. Creating a gff3-Format File
3. File Input and Visualization in WebGBrowse
4. Identifying Protein Name and Function with blastp

1.2.1 Obtaining Exon Boundaries

The first step in the process of generating input for WebGBrowse involved obtaining the necessary information from a protein-prediction software. For my purposes, I utilized the web-based tool, GeneMark™, developed at The Georgia Institute for Technology. Upon visiting the tool at the following site, http://exon.biology.gatech.edu/, the link within the section titled, "Gene Prediction in Eukaryotes" was selected. At this stage in the process, I had also obtained the appropriate Fasta files for the contigs included in the analysis. I copied and pasted the sequence data representing a single contig (beginning with the contig identifier and ending with the last base in the sequence) into the input box marked "Sequence." Alternatively, the Fasta file could have been uploaded. I selected the species most closely related to *C. drosophilae* (*Caenorhabditis elegans*). Selecting a closely related species informed the WebGBrowse algorithm to use the parameters set for the identification of proteins in the closely related species, when

predicting protein-coding regions within the input data. Once the sequence data had been inputted, I began the application by clicking “start GeneMark.hmm”. An example of the resulting output can be visualized in Figure 2. Those data necessary for generation of a gff3 file are highlighted in Figure 2 and identified in the accompanying key. Data should be collected for each predicted protein-coding gene as well as for the predicted exons.



Key:
—Contig Length
—# genes
—# exons
—Strand
—Start position
—End position

Figure 2: Visual Depiction of GeneMark™ Output

GeneMark™ data was collected in an excel file for use in the generation of an appropriate gff3-format file (described below).

1.2.2 Creating a gff3-Format File

While it was not necessary that I be an expert in the generation of files in gff3 format, it was important that I understood the basic structure of this file type such that a simple file could be generated to upload to WebGBrowse. One gff3-formatted file was generated for each contig that was visualized. A typical gff3 file contains nine columns of information that are to be filled with the following content. Each column is separated from the adjacent column by a tab.

1. Sequence ID
2. Source
3. Type
4. Start
5. End
6. Score
7. Strand
8. Phase
9. Attributes

For my purposes, I disregarded columns 6 and 8. Figure 3 will guide my description of the information to be placed in each of the applicable columns.

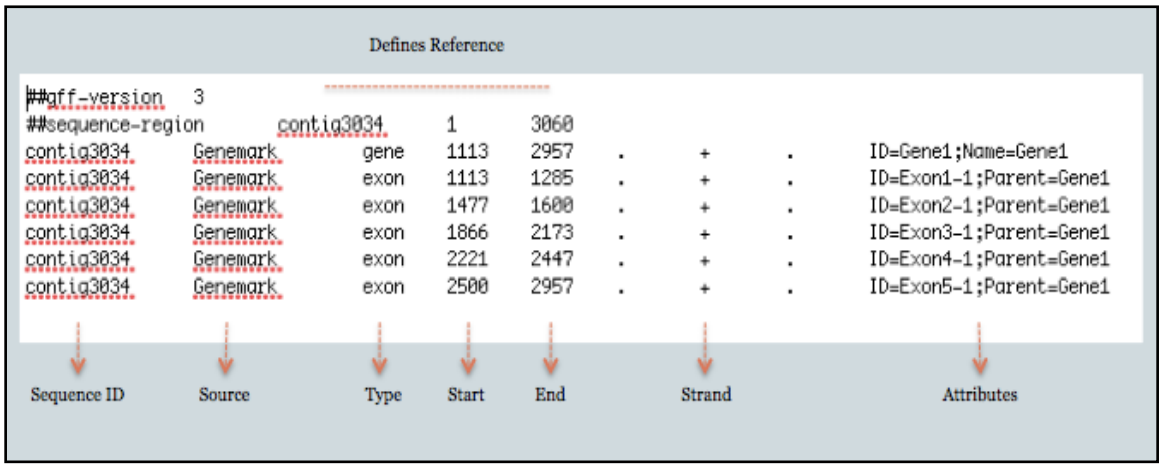


Figure 3: Example of a gff3-Format File

Each file began with a description of the file format being used. This description can be visualized in line number one. Line number two was used to provide a reference onto which the regions of the sequence data described thereafter could be mapped. For my purposes, I used this line to identify the contig I was working with as well as define the boundaries of the contig with the start position being 1 and the end position being the length of the contig. From this point on, each line following must include information collected in the nine columns described above. The sequence ID in column one should be equal to the name of the reference provided in line one. The second column can be any random text identifier. Given that I collected the input data from GeneMark™, I chose to label the "source" column with the text "Genemark". The third column describes the region of the contig that is being defined. Given that GeneMark™ outputted the boundaries for every predicted protein-coding gene including the unique exons within each predicted protein-coding region, each region corresponded to either a gene or an exon. Columns four and five were used to input the start and stop positions for the feature of interest. Because the reference was defined as a sequence beginning at 1 and ending with the length of the contig, the boundaries of each gene or exon should fall within this range. A period can be used as a placeholder for the content in columns six and eight. A "+" or "-" was placed in column seven according to the strandedness of each of the features. Start

position, end position and strandedness should all have been directly collected from the GeneMark™ output. The final column serves a couple distinct purposes. The first component that was included is "ID". The letters "ID" followed by the "=" sign should proceed a name given to the feature being described. I have chosen to label the only gene in the contig identified as 3034, "Gene 1" and the five exons defining this gene, "Exon1-1", "Exon2-1", etc. I followed a similar scheme as I labeled the features in the other gff3 files I generated. In addition to ID, a "Parent" attribute should be included if the features falls within another feature. For example, Exon1-1, Exon2-1, etc are features within Gene 1. Therefore, each of the lines where these features are described should include the attribute "Parent=Gene1." Attributes (ID and Parent) can be separated by a semicolon. The final attribute I chose to include in some of the lines of my gff3 files was "Name". If a feature was given a name, this name appeared in the actual WebGBrowse window as a label of the feature. Given the simplicity of my visualizations, I chose only to label the genes within each of the contigs. The gff3 files were built in a simple text editor and saved as a gff3. To facilitate saving the file in the proper format, the tag ".gff3" was added to the end of the file name.

1.2.3 File Input and Visualization in WebGBrowse

Following generation of each gff3 file, the file was ready to be inputted into WebGBrowse. The web-based tool can be accessed at the following web address: http://webgbrowse.cgb.indiana.edu/cgi-bin/webgbrowse/uploadData. Each file was uploaded via the "choose file" button under the GFF3 file heading. An e-mail address was inputted to facilitate the return to the document for editing at a later date. Following the upload of the file, I was directed to a page comprising two input boxes. The first prompted me to provide a short description of the file. Here, I gave the work submission a name that could be used to identify the uploaded file at a later date (i.e. Contig 3034). Secondly, I was prompted to "add new tracks" based upon the "types" of features built into the gff3 file. For my purposes, the gff3 files were comprised of two feature types, "genes" and "exons." If desired, a track could be added to represent each of these features. A "glyph style" was chosen for each track. "Glyph" describes the physical units used to represent a feature within a WebGBrowse display. Dependent on the desired visual, a number of different styles could be chosen. Each glyph type is displayed and described as a user scrolls though the options, facilitating selection of the most appropriate type for a researcher's needs. For the purpose of the images I generated, it was only necessary for me to add a single track. I added the "gene" track and chose to represent this track using the "line" glyph, "gene" glyph or "generic" glyph. These glyph types automatically displayed the associated exons within the gene boundaries. Upon the addition of each track, I was presented with a window representing various adjustable parameters. For all glyph types there was an option to add an identifier for "Key." The keyword inputted here served as the label for the track in the WebGBrowse display. I labeled the key "gene" when adding the "gene track" for simplicity. For both the "gene" and "generic" glyphs, a "connector-type" must be selected if one desires to visualize the exons as strung together. This parameter exists within the "advanced" parameters of the "generic" glyph settings and in the "transcript" parameters of the "gene" glyph settings. Various parameters within each glyph type additionally facilitate changes in font size, color, outline color, etc., if desired. Additionally, I experimented with the "box" glyph as well as with the addition of a second track displaying only the exons. The visual images for which these amendments were made are labeled accordingly.

Once the "gene" track had been added, including the adjustment of any desired parameters, I choose, "display in GBrowse 2.0" (default). The outputted image displayed both the predicted protein coding gene(s) and the individual exons predicted within this gene(s), both within the context of the contig they were identified in. For simplicity, I minimized the "region" track leaving the "overview" track and the associated "gene" track in the display. To ensure that the only region visualized is the contig itself, the display boundaries were set such that the visual begins with position "0" and ends with the total length of the contig (in bps). For example, the display region for the contig of identifier 3034 was made to read "contig3034:0…3060." Manually setting the display boundaries for each contig ensured that visuals displayed the desired region of the genome to scale. Screenshots were collected such that a visual image of each contig could be archived. Visual displays for 20 genomic contigs are represented in the results section and in the appendix.

1.2.4 Identifying Protein Name and Function with blastp

One additional analysis performed in an effort to better understand the genomic makeup of *C. drosophilae* involved use of NCBI's blastp tool. Blastp searches a protein database utilizing a protein query (NCBI, 2009). When prompted, GeneMark™ outputs the raw protein sequence for those regions of the genome predicted to be protein-coding. This protein sequence, outputted for each predicated gene, was inputted into the blastp tool. The blastp output displayed all protein sequences within a well-kept protein database that matched the queried protein sequence with significant % identity. The blastp output provided me with a prediction for both the identity and function of the predicted protein as well as the identity of other organisms whose genomes comprise a similar protein sequence. To utilize the tool, each protein sequence outputted from GeneMark™ was pasted into the input box beneath the heading "Enter Query Sequence." The protein sequence was queried against the "non-redundant protein sequence" database and run via the algorithm "blastp". The tool can be accessed at the following web address: http://blast.ncbi.nlm.nih.gov/Blast.cgi.

1.3 Results

The goal of this research project was an increased understanding of the genomic content within the *C. drosophilae* genome. For the purpose of this report, visual displays were generated for 20 of 80 total contigs depicting the regions of each contig predicted to be protein-coding. Visuals were generated using the WebGBrowse tool created by Indiana University's Center for Genomics and Bioinformatics. A few different "glyph" styles were chosen in an effort to generate a distinct variety of visual representations. The glyph style chosen for each display is noted in the figure title. A screenshot of the final interface of the WebGBrowse tool as well as a few visual representations of WebGBrowse output can be viewed in Figures 4 – 7. The remainder of the 20 WebGBrowse images can be viewed in Figures 16-32 (appendix pgs. 30-34).

Highlighted in each figure title is the contig identifier and the type of glyph chosen for representation of the genes within each contig. The contig identifier equals the approximate length of each contig. As can be seen from both the below figures and those present in the appendix, a majority of the contigs visualized comprise only a single gene while others comprise multiple. Additionally, there exists extreme variation in both the size and number of exons comprising each predicted gene. Genes comprised of a large number of consecutive exons have greater opportunity for alternative splicing, a process by which exons are put together in unique combinations to yield different proteins from the same gene, and overall, have the potential to contribute more dramatically to the complexity of an organism than do genes with exons that are few and far between. The latter are comprised of a greater percentage of intrinsic material. Visualizing these distinct differences in gene organization can facilitate a deeper understanding of the genome itself.

A summary of the data outputted by GeneMark™ including the number of genes predicted within each contig, the total number of exons comprising those genes, and as a result, the percentage of the total contig predicted to comprise exons can be viewed in Table 1. This summary highlights the diversity of protein concentrations among different pieces of the larger genome. In addition to the visuals, this summarized data provides a preliminary means of bettering understanding the organization of the *C. drosophilae* genome.

Use of NCBI's blastp tool facilitated exploration of the identity and function of each predicted protein-coding gene. Input of a number of *C. drosophilae* predicted protein sequences into the blastp tool resulted most commonly in protein matches to *Caenorhabditis elegans, Caenorhabditis remanei,* and *Caenorhabditis briggsae,* species known to be closely related to *C. drosophilae.* In addition, a variety of other species including *Brugia malayi, Loa loa, Trichinella spiralis,* and *Aedes aegypti,* parasitic nematodes and mosquito comprise proteins in the NCBI database that closely match those predicted in the *C. drosophilae* genome. Given that these matches are less expected than matches to other species in the same genus, these results may prove meaningful in understanding the functions of these shared proteins in both *C. drosophilae* and the parasitic species. Use of the blastp tool proved an effecitve means of uncovering

additional meaning from the the information outputted by GeneMark™. A summary of the blastp results can be viewed in Table 2.

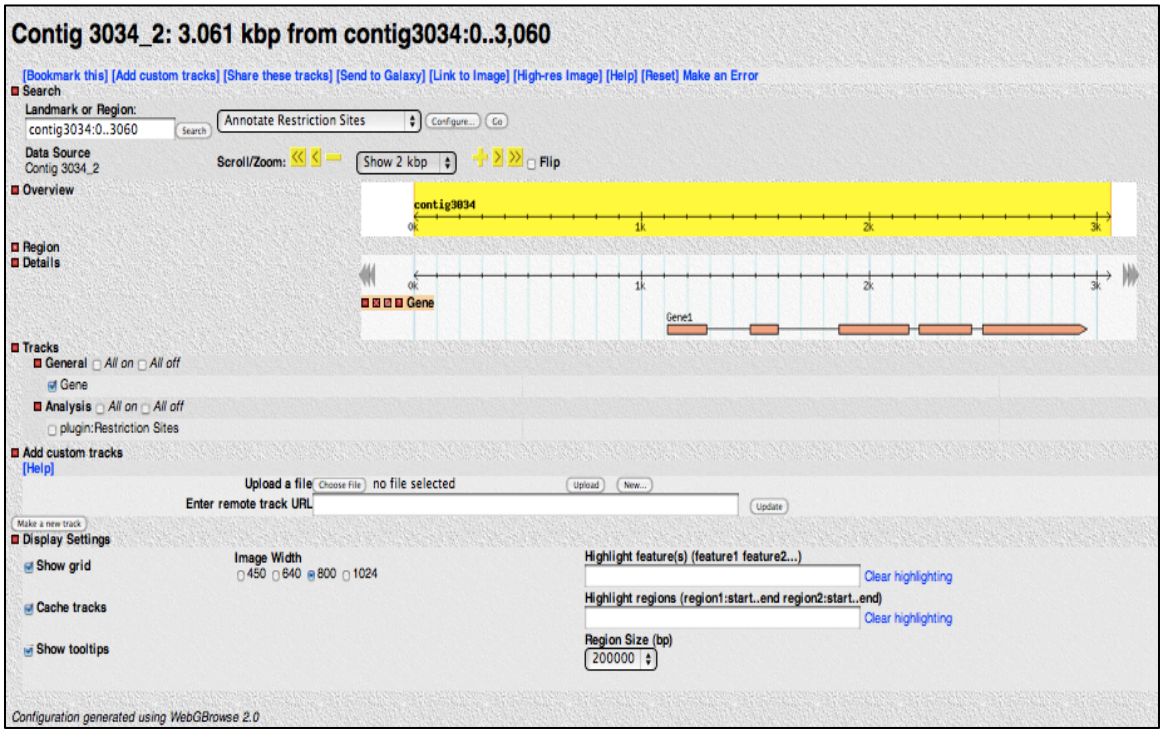


Figure 4: Visual Representation of WebGBrowse Output

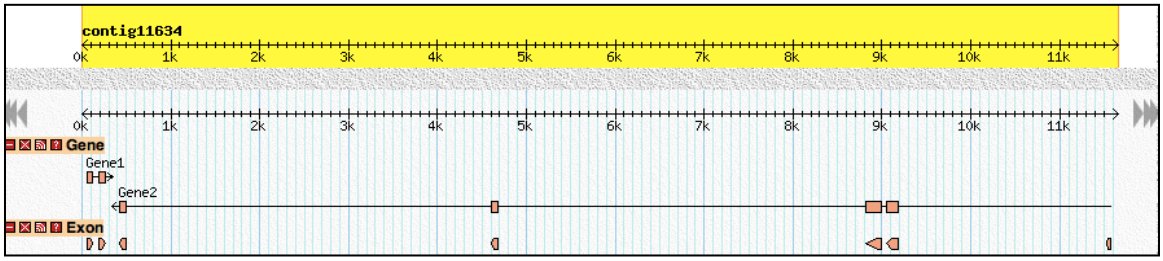


Figure 5: Contig 11634; Gene Track = line glyph; Exon track = generic glyph

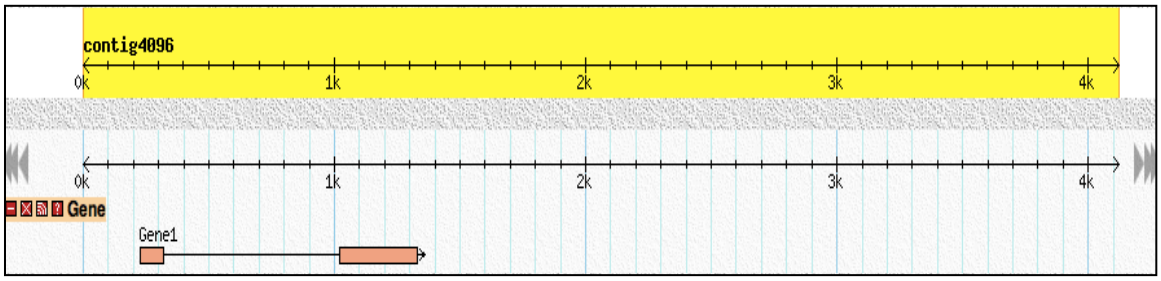


Figure 6: Contig 4096; Gene track = line glyph

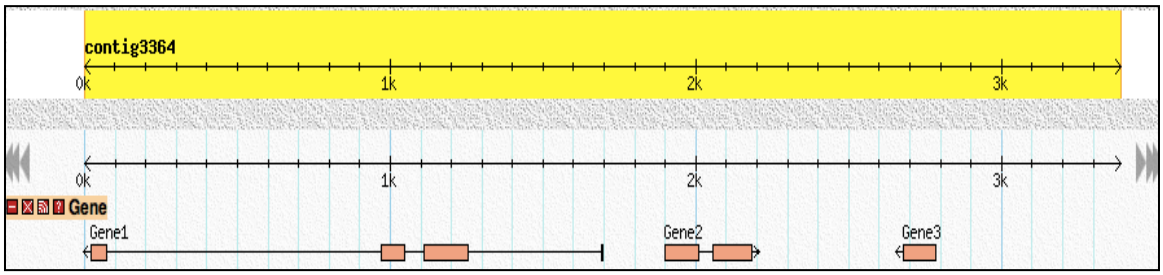


Figure 7: Contig 3364; Gene track = line glyph

Table 1: Summary of GeneMark™ Output for 20 *C. drosophilae* Contigs

Contig Identifier	# Predicted Genes	# Total Exons	% Exon
11634	2	7	5.43%
4096	1	2	9.68%
3364	3	7	18.32%
3309-8	1	3	95.35%
3309-7	1	6	82.25%
3153	1	2	91.44%
3034	1	5	42.16%
2998	1	11	80.13%
2935	1	4	17.73%
2904	1	11	75.73%
2889	2	12	54.00%
2855	1	9	72.30%
2821	1	7	71.13%
2728	1	7	79.27%
2718	1	8	66.14%
2699	1	11	62.57%
2625	1	6	89.89%
2622	1	1	97.70%
2617	1	4	82.97%
2545	1	10	59.28%

Table 2: Summary of Blastp Results for 7 Predicted Protein-Coding Genes

Protein-Coding Gene	11634 1	11634 2	3364 1	3364 2	3309 1	3153 1	3034 1
C. elegans	X	X	X	X	X	X	X
C. remanei	X	X	X	X	X	X	X
C. briggsae	X	X	X	X	X	X	X
B. malayi	X				X	X	
L. loa	X			X	X	X	
T. spiralis	X			X	X		
H. contortus	X						
A. aegypti		X		X	X	X	X
C. quinquefasciatus					X	X	X
G. morsitans morsitans							X

1.4 Discussion

As high-throughout DNA sequencing technologies continue to increase in throughput and decrease in cost, researchers are challenged to uncover biological meaning from these growing genome-scale datasets. Researchers are attempting a variety of approaches to overcome this challenge.

Large-scale genomic data progresses through a fairly standardized pipeline in its movement from raw data to meaningful information. The standard procedure for dealing with high-throughput sequencing data involves acquiring sequence data, assembling the sequence reads and annotating the larger genomic contigs. Unique methods for both assembly and annotation can be chosen in an effort to guide an analysis towards the answer of a specific biological question. Similarly, the information needed from a genome-scale dataset is entirely dependent on the research question being addressed. This basic process of analyzing the content of a DNA sequence comprises a pathway through which researchers can uncover meaning from the seemingly basic series of called bases the investigator has access to following DNA sequencing.

Despite the ability to choose both sequencing instrument and assembly algorithm, the sequencing and assembly pieces tend to be fairly consistent activities in the pipeline. It is in the annotation of assembled sections of the genome where researchers are able to more specifically choose the techniques that will provide them with meaning consistent with their research question.

As was presented in the materials and methods and results sections above, visualization is one way in which meaning can de derived from sequence data. Transforming a series of bases into a visual depiction of the placement of genes within a genomic segment offers researchers a tool to begin to understand the activity of these genes within the larger organism. Various tools including Indiana University's WebGBrowse have been developed for this purpose. Ultimately, the ability to visualize what is not readily apparent in a large genome-scale dataset facilitates a deeper understanding of the genome's organization and supports researchers in their quest to answer biological questions. While the visual depictions outputted by WebGBrowse do not identify the genes within each contig by name or by function, they do show the organization of the gene, identifying the number, size and position of the exons comprising that gene. These visuals provided preliminary information important to an understanding of the overall genomic makeup of *C. drosophilae.*

NCBI's blastp tool provided additional information concerning the identity and function of each predicted protein as well as identified other species whose genomes contain a similar protein sequence. This information further advanced the search for meaning within the initial genomic sequence. While it makes sense that those protein sequences matching most closely with the proteins predicted in *C. drosophilae* would exist in species within the same genus (i.e. *C. elegans, C. remanei,* and *C. briggsae*), protein matches to unexpected species are more informative in terms of understanding the protein's function. As was outlined in the results section and in Table 2, protein matches were made to both parasitic nematode species such as *Brugia malayi* and *Loa loa* and parasitic mosquitos such as *Aedes aegypti.* If the protein in question is important for these species' parasitic activities, presence of a similar protein sequence in *C. drosophilae* may hint at the lifestyle of this organism as well. Further investigation would be necessary to solidify this connection. Ultimately, blastp facilitated the initiation of protein identification and in a simple way aided in the process of uncovering meaning from the raw sequence reads the project began with. In combination, the genome annotation tools GeneMark^{TM}, WebGBrowse and blastp facilitated identification of the regions of 20 *C. drosophilae* contigs predicted to code for protein, facilitated the visualization of all genes and exons predicted within these same contigs and provided preliminary information concerning the identity and function of the predicted proteins.

Considering the larger picture, a broader solution to the challenge of uncovering meaning from genome-scale datasets is the use of computational tools built specifically for the analysis of massive genome-scale data. The field of bioinformatics can be defined as the application of computer science and information

technology to the field of biology and medicine. Applications comprise the creation of algorithms, databases and information systems for the purpose of generating new knowledge in these same fields. Bioinformatics solutions can be built on a per-job basis for individual researchers interested in addressing one specific scientific question.

A number of companies have sprung up in recent years in an attempt to uncover meaning through the development of custom analysis tools and pipelines. Some companies specialize in the manufacture of a single software package for the execution of one narrow function while others offer consulting services and custom-built software solutions.

A summer internship experience with Intuitive Genomics, Inc. exposed me to the goals and new directions of one such bioinformatics services company.

Intuitive Genomics offers cutting edge genomics and bioinformatics services to scientists, companies, and institutes faced with the challenge of analyzing massive genome-scale datasets. Mostly, Intuitive Genomics supports customers interested in the analysis of high-throughput sequencing data but the company is also capable of addressing all types of large-scale biological data. Intuitive Genomics is a service and consulting-based bioinformatics company focused on the generation of custom software and pipelines to specifically address biological questions.

Companies such as Intuitive Genomics offer a solution to the challenge of uncovering meaning from large genome-scale datasets. Approaches taken to address this challenge include the development of software packages, hosting software-as-a-service and offering bioinformatics services and consulting. Software companies create software packages to be run on the customer's own hardware for the purpose of conducting a specific analysis. Companies hosting software-as-a-service manage an online portal supporting a variety of analysis capabilities used by customers and charged on a per-job basis. Bioinformatics consultants and service providers consult with a researcher to learn about the biological question, develop custom software for the analysis of the researcher' s data, analyze the data through use of the custom software, and interpret the results.

The specific goals and approaches of Intuitive Genomics in their effort to uncover meaning from seemingly unmanageable datasets will be outlined more specifically in the business portion of this report.

1.5 Conclusion

The increasing throughout and decreasing costs of DNA sequencing are changing the way researchers approach their biological questions. The rate at which DNA sequence data can be generated far outpaces the rate at which computational tools can be generated to store, manage and analyze this data. This gap creates both a challenge and an opportunity for the life science community. I was granted two unique experiences while pursuing my PSM degree that exposed me to this challenge and opportunity in the research community.

Firstly, a research experience in Dr. Dee Denver's lab provided me with the opportunity to uncover meaning in a series of *C. drosophilae* genomic contigs through visualization of predicted protein-coding regions. This process involved the use of a number of genome annotation tools ultimately yielding a visual representation in WebGBrowse.

Secondly, a 3-month internship with Intuitive Genomics exposed me to an industrial approach to the challenge of large-scale data management and analysis. Through my involvement in discussions related to product development, sales and marketing strategies and business-model creation, I gained a thorough understanding of one company's approach to uncovering meaning in large-scale data.

These two experiences were not mutually exclusive and both contributed to my overall understanding of

current approaches to the analysis of high-throughput sequencing data.

The prevalence of high-throughput DNA sequencing as a research tool and the power the outputted data has to spark scientific discovery will continue to advance with the increasing throughout and decreasing cost of the instrumentation. To stay at the forefront of discovery in the sciences, it will be crucial that researchers are equipped to uncover biological meaning from the massive genome-scale datasets outputted.

Both my research experience in the Denver Lab and internship with Intuitive Genomics have alerted me to mechanisms by which this data can be leveraged, both now and in the future.

2. **Business Report**

2.1 Description of the Business

2.1.1 Overview

Intuitive Genomics is a bootstrapped startup actively delivering expert bioinformatics services and consulting to its customers since incorporation in August 2010. The company's goal is to power the next-generation of scientific discovery through delivery of vital insights and actionable information from these customers' genome-scale datasets.

2.1.2 Founding Team

In August 2010, Intuitive Genomics, Inc, was both founded and incorporated in Corvallis, Oregon.

The founding team was comprised of Dr. Doug Bryant, CTO, and Drs. Todd Mockler and Jim Carrington, scientific advisors. Dr. Doug Bryant has focused his research on applying machine learning to massive biological datasets and is the author of widely used tools for the analysis of high-throughput sequence data, including SuperSplat and Gumby. Dr. Todd Mockler is a faculty member at the Danforth Plant Science Center and a professor in the Center for Genome Research and Biocomputing at Oregon State University. His published work has provided critical tools and approaches for using high-throughput sequence data to understand complex systems. Dr. Jim Carrington is the President of the Donald Danforth Plant Science Center, a member of the National Academy of Sciences, and is internationally recognized for his research on gene silencing. The original founding team comprised a fourth individual, Nathan Williams, MBA MIT-Sloan 2008, no longer with the company as a result of differences in opinion concerning the proposed strategy for the company's future directions.

Doug Bryant and Nathan Williams were childhood friends who began discussions about starting a company upon Nathan Williams' completion of his MBA at MIT-Sloan. Doug Bryant was working on his Ph.D at Oregon State University at the time, performing the majority of his thesis work in Dr. Todd Mockler's laboratory. Upon Doug and Nathan's decision to pursue a bioinformatics services model, Doug Bryant spoke with Drs. Todd Mockler and Jim Carrington, both professors at Oregon State University at the time, about joining together to pursue the idea. Upon agreement from all parties, these four individuals moved forward with the incorporation of Intuitive Genomics, Inc.

At the time of incorporation, each of the four founding members held equal ownership of the S-corporation in the form of vested equity shares.

2.1.3 Administrative Structure

Up until the major organizational change that lead to Nathan Williams' resignation, Nathan served as the company's CEO. Doug Bryant serves as the company's CTO while simultaneously holding a post-doctoral position at the Danforth Plant Science Center. Dr. Jim Carrington, President of the Danforth Plant Science Center and Dr. Todd Mockler, faculty researcher at the same institute, both serve as scientific advisors in the forward movement of the company. Hired initially as a summer intern, I held the position, Marketing

Manager and worked most closely with Nathan Williams on the business side of the corporation.

For the majority of my internship experience, Nathan Williams and I were dedicated to the day-to-day operations of the company, together comprising the business end of the organization. At this time, Drs. Doug Bryant and Todd Mockler were primarily involved with the scientific aspects of the corporation, processing customer data, interpreting the results, and generating reports to be returned to the customer. Even in their shared role comprising the processing of customer jobs, Dr. Bryant, performed the majority of the software development and data processing while Dr. Mockler served in an advisory role. While Drs. Bryant and Mockler were also involved in weekly team meetings and major decisions or milestone events, their involvement was mostly limited to the processing of customer jobs and therefore oscillated with the influx of new clients. This level of interaction was appropriate given both founder's simultaneous commitment to full-time positions at The Danforth Center. Dr. Carrington has served primarily as a scientific advisor and powerful resource, facilitating the introduction of the company to key authorities, resulting in new customers as well as early-stage funding opportunities.

The company' administrative structure (Figure 8) changed abruptly upon Nathan Williams' resignation. At this time, Dr. Bryant stepped forward to aid me in conducting the company's administrative duties. Dr. Bryant will remain the key player in the processing of customer jobs and Drs. Mockler and Carrington will continue to serve as scientific advisors for the company. This change in job responsibilities is highlighted in Figure 9.

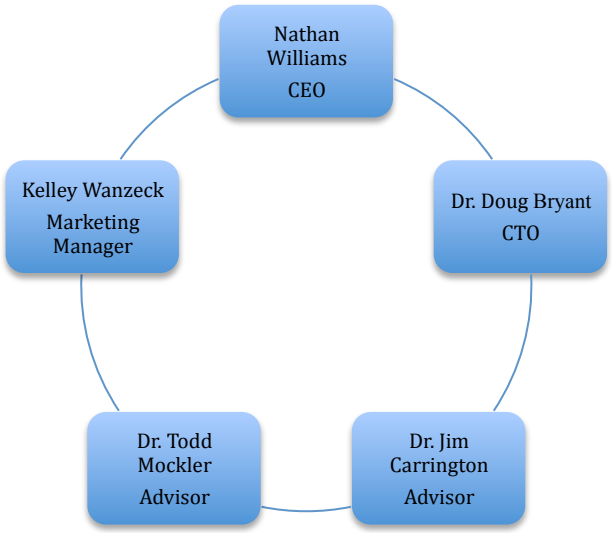


Figure 8: Intuitive Genomics' Basic Administrative Structure

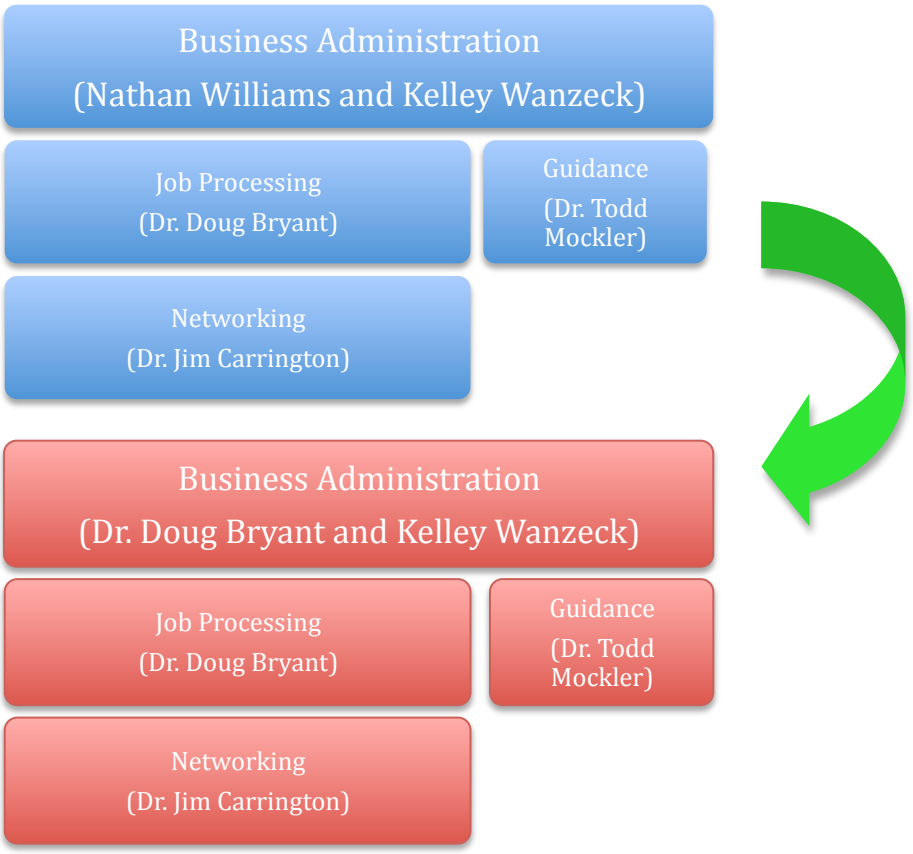


Figure 9: Change in Job Responsibilities as a Result of Change in Management

2.1.4 Location

Given the history of Intuitive Genomics' founding team, the startup company originated in Corvallis, Oregon. At the time of the company's founding, Drs. Jim Carrington and Todd Mockler were both serving as faculty members at Oregon State University and Doug Bryant was finalizing his Ph. D research in Dr. Todd Mockler's Laboratory.

Upon Dr. Jim Carrington's acceptance of his offer to serve as President of the Donald Danforth Plant Science Center (St. Louis, Missouri) and Dr. Todd Mockler's acceptance of a position at the Danforth Center as incoming faculty, the company made an executive decision to move operations to St. Louis, Missouri.

In August 2011, the company made its official move, taking up occupancy at the Bio-Research and Development Growth (BRDG) Park, a bioscience incubator on the Danforth Plant Science Center Campus. Given the company's change in location, the first portion of my internship was conducted in Corvallis, Oregon while the second half was fulfilled in Intuitive Genomics' new office in St. Louis, Missouri. Short biographies of both BRDG Park and The Donald Danforth Plant Science Center can be found in the appendix (pgs. 35 & 36).

2.1.5 Product and Service Offerings

Intuitive Genomics targets individual researchers, scientific corporations, and research institutions with the goal of helping scientists uncover meaning from their massive genome-scale datasets. Currently, Intuitive Genomics offers bioinformatics consulting and custom software development services in an effort to aide

researchers in this regard. Intuitive Genomics consults with the customer for the purpose of understanding both their bioinformatics challenges and the underlying biological question. The company then develops customized software or a bioinformatics pipeline to specifically address the customer's needs. Although the company specializes in the development of custom software for the analysis of high-throughput sequence data, Intuitive Genomics' expertise in both computational and biological sciences facilitate expert consulting at any point within a project timeline. Intuitive Genomics has offered consulting services to customers anywhere from the design of an experiment to the interpretation of the results following bioinformatics analysis.

Concerning the development and execution of custom software, Intuitive Genomics' service capabilities include automated pipelines in the cloud for common analytics, custom bioinformatics analysis through Intuitive Genomics' pipeline and software customization technology, and outsourced general bioinformatics support.

While the company's current mode of operation is consulting/service-based, the company expects to productize a number of custom software solutions in the near future. Two specific product ideas include a data management and archiving platform for high-throughput sequence data, facilitating the timely retrieval of data by end-users and a software platform automatically connecting high-throughput sequencing users with sequencing service providers. The latter would incorporate an integrated sequencing run management and scheduling solution benefiting both the sequence provider and researcher.

To date, Intuitive Genomics has analyzed terabytes of biological data and delivered valuable hypothesis generating results to a diverse set of customers. Results have covered a wide variety of agriculturally significant species including peach, cherry, strawberry, switchgrass, Brachypodium, as well as fungal and algal species relevant to the biofuels industry. Customers have included government agencies, academic researchers, sequencing service providers, and biotechnology companies.

2.1.6 Long-Term Goals of the Company

Intuitive Genomics recognizes the gap between the rate at which high-throughout sequencing data can be produced and the rate at which publically available computational tools have been developed for the storage, management and analysis of these massive genome-scale datasets.

The company's goal is to power the next-generation of scientific discovery by helping researchers to uncover meaning from large genome-scale datasets. The massive amounts of genomic data that can be outputted by the most recent versions of DNA sequencing instruments have the potential to guide scientists towards cutting edge discoveries if only these researchers had the tools and expertise to parse these massive datasets for the answers to their biological questions. Intuitive Genomics' goal is to aid researchers in this endeavor through the development of custom software and bioinformatics pipelines built specifically to answer an individual scientist's research question. Currently, Intuitive Genomics has focused its efforts on the life sciences and agricultural biology research communities. Focused in these fields, Intuitive Genomics has served a variety of researchers interested in the development of nutrient-rich and sustainable agriculture as well as those focused on the development of effective biofuels.

Ultimately, Intuitive Genomics aspires to impact all industries utilizing high-throughput sequencing as a means of answering biological questions. Naturally, Intuitive Genomics will continue to seek customers in the life sciences and agricultural biology fields but will also support the health science community in its emphasis on personalized medicine, diagnostics and therapeutics.

Long term, Intuitive Genomics hopes to serve as a leader in the bioinformatics community, dedicated to the development of custom solutions for any researchers struggling to uncover meaning from large-scale datasets. The company expects to continue with the consulting/service model while simultaneously introducing a series of software products that will more specifically address individual challenges in this field.

2.1.7 Project Goals

My internship project with Intuitive Genomics, Inc. initiated on June 20th, 2011 and terminated on September 30th, 2011.

Over the course of the internship period, I was responsible for a number of initiatives related to marketing, sales, and strategic positioning of the company in its market niche. As I became aware of from the beginning of my experience, serving in any capacity for a startup company is a dynamic and unpredictable endeavor. In addition to my roles in marketing, sales and strategic development, I also filled administrative roles serving as secretary, account manager and administrative contact for the establishment of the company in its new BRDG Park office. Throughout my various experiences, I acted under the title, Marketing Manager. Major accomplishments and critical experiences will be discussed in further detail in later sections of the report

All tasks focused on during the internship period comprised small initiatives towards achievement of one overarching goal, continued growth and success of Intuitive Genomics. Inc. Overall success of the startup is marked by the ability to meet monthly revenue targets, cover monthly budget items, generate satisfied customers and partners and expand the company's bioinformatics product and service offering. Successful completion of a number of the accomplishments discussed in this report permitted me to significantly impact these markers of company growth. The impacts of my internship project are visually depicted in Figure 10.

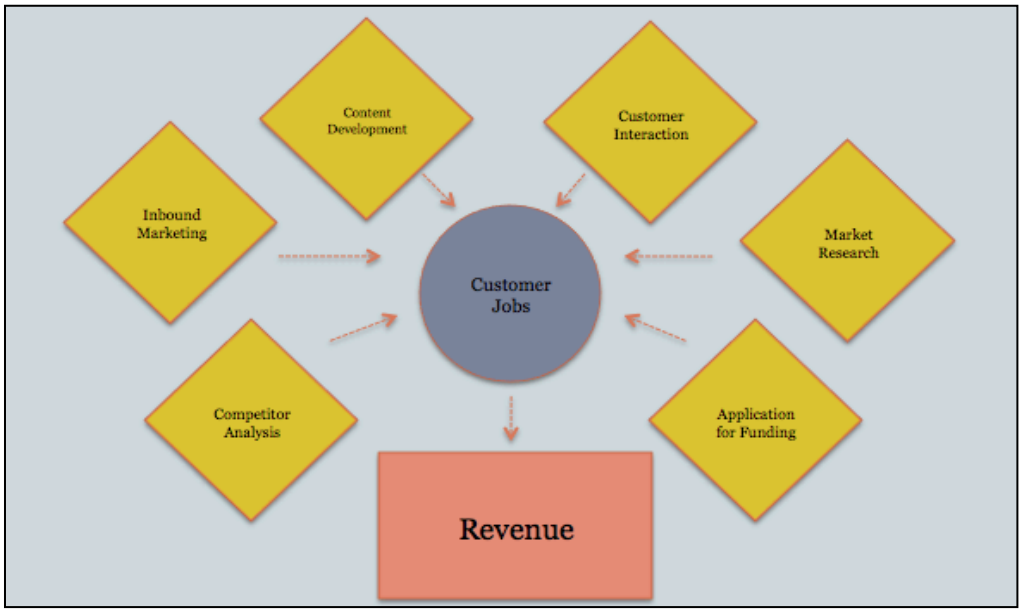


Figure 10: Impacts of My Internship Project

In addition, a number of outcomes of the internship experience marked success on a more personal level (Figure 11). These key experiences enhanced my skill set or exposed me to activities that will make me a more marketable candidate for a position with a scientific corporation.

For one, my exposure to a dynamic entrepreneurial environment, particularly my experience conducting a variety of administrative duties, my exposure to the process of seeking funding opportunities and my witness of a major organizational change early in the life of a startup company, will supplement my education in the sciences and prepare me for a career position with an early stage scientific corporation. Given my scientific background, the internship provided me with experience on the business side of a

corporation, better preparing me to take on a position where I will serve as a liaison between these two distinct subject areas. The internship gave me an opportunity to try out my skills in an environment where dual skill sets in business and science are equally important.

Business-oriented skills acquired over the 15-week internship period include, the execution of project management strategies, skillful interaction and increased confidence in correspondence with potential customers, investors and other related authorities, and professional collaboration with coworkers. Additionally, I fine-tuned my written and oral communication skills and experimented with marketing techniques for the purpose of bringing in new customers.

My summer experience increased my understanding of the genomics market, specifically in the field of plant science and agriculture. Given that my interests lie in the field of personalized genomics and the various emerging applications of DNA sequencing technologies, experience with a company that offers a solution to the challenge of uncovering meaning from large genome-scale datasets alerted me to the use of genomic data for a variety of unique applications. While the majority of Intuitive Genomics' customers are focused on plant research or in the development of biofuels, the same founding principles can be applied to human DNA in applications more focused in my area of interest. Therefore, exposure to the genomics market, even if in a subfield a bit tangent to my interests, provided me with a deep understanding of the immense possibilities in research arising from the ability to generate and analyze genome-scale datasets. Additionally, I gained experience in the translation of highly scientific concepts into common terms and grew immensely in my breadth of bioinformatics knowledge.

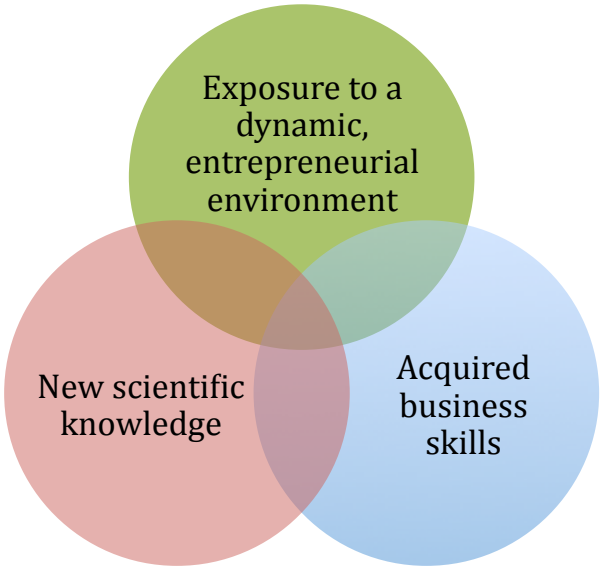


Figure 11: Key Outcomes of the Internship Experience

Success of my internship project was evaluated based upon achievement of both weekly and monthly milestones as well as progress towards a revenue goal set at the beginning of the internship period. Weekly and monthly milestones comprised various deliverables namely development of blog posts and white papers for the newly launched website, development of landing pages for the collection of inbound leads, detailed analyses of various competitor companies, and collection of contact information for potential customer leads. A goal of $45,000 in new customer revenue was established at the start of the internship period. A bonus structure was created around this revenue goal such that a percentage of any new customer business brought in during the internship period was returned in the form of a bonus at the end of the internship.

While weekly pay was stable throughout the course of the internship, a bonus structure indicative of achievement of both weekly/monthly deliverables and progress towards a revenue target proved an effective means of both evaluating and rewarding progress throughout the course of the internship.

2.2 Marketing

2.2.1 Target Market

Intuitive Genomics targets independent researchers, scientific corporations and large research institutes. Initially, the company has focused on life sciences research and agricultural biology. This is in large part because of the nature of the research conducted at both Oregon State University and The Donald Danforth Plant Science Center, two entities that the company has been directly affiliated with since the company's incorporation.

The company's current proximity to the Danforth Plant Science Center provides the company with an initial market from which to bring in new customers and make strong connections in the plant science field. Other key players in the St. Louis Community namely universities such as Washington University and Saint Louis University and large research institutions such as Monsanto provide additional opportunities to forage strong connections and form an initial customer base.

2.2.2 Future Directions

Although the company's initial emphasis has been in agricultural and plant biology, Intuitive Genomics is in no means limited to this narrow market. As the company continues to grow in the next few years, the fields to which Intuitive Genomics' bioinformatics capabilities can be applied will continue to expand. In the future, Intuitive Genomics will expand its target market to include customers seeking the analysis of human genetic material and will exist as a key player in the pursuit of applications of personalized medicine.

Additionally, the company does not plan to pursue the current consulting/service-based model for the remainder of its existence. While this service and consulting model may still be utilized, the company also plans to productize valuable analysis tools for commercial sale to customers. Creation of a product designed to perform a specific analysis function has the potential to attract a large market across a breadth of scientific disciplines and will immediately make the company more appealing to venture capital investors.

In order to properly evaluate the best strategy for the company's future directions, I conducted research on a variety of topics related to strategic positioning over the course of my internship. Goals of this research were strategic positioning of the company within its market niche as well as positioning against competitors in the same field. Research topics included the identification of differentiating factors of Intuitive Genomics' current service and consulting model, and the exploration of other business model approaches to the same scientific problem.

As the future directions of the company migrated towards the productization of specific analysis tools, I participated in team discussions concerning the potential for product development and conducted preliminary market research for these proposed product offerings.

2.2.3 Competitors

In its current capacity as a consulting/service-based bioinformatics company, Intuitive Genomics must remain cognizant of a few key competitors. While few competitors are currently pursuing an identical business model, many are still a threat based upon their end goal of uncovering meaning from massive genome-scale datasets. Competitor strategies include the development of data analysis software packages, hosting software-as-a-service portals, and addition of bioinformatics services to a previously established DNA sequencing service offering. Companies such as Partek (St. Louis, MO), Multicore Ware (St. Louis, MO), and Bioinformatics Solutions, Inc. (Waterloo, Ontario, Canada) offer data analysis software packages

sold to customers and run on these customers' own hardware. DNAnexus (Mountain View, CA), Genome Quest (Westborough, MA) and Knome (Cambridge, MA) offer a cloud-based self-service approach in which customers access a bioinformatics portal created by the company and pay per-job for usage. Appistry (St. Louis, MO) provides the customer with options by making their data analysis solution available for deployment as an appliance or through a remotely hosted pay-per-use service. Companies such as Cofactor Genomics (St. Louis, MO) and Edge Bio (Gaithersburg, MD) offer limited bioinformatics services as an add-on to the DNA sequencing service they already provide. Lastly, most in line with Intuitive Genomics' business model, Data2Bio (Ames, IA), offers outsourced bioinformatics services to customers in the US and abroad.

The above summary of competitors is the result of extensive research on this topic completed over the course of the internship experience. Research in this area comprised both identification of local, national, and international competitor companies as well as classification of these companies based upon their current business strategies. Identifying competitor strategies was key to understanding the potential threat competing companies might play in Intuitive Genomics' selective market.

Intuitive Genomics differentiates itself from its competitors in a few distinct ways.
For one, Intuitive Genomics operates as a service, meaning the company's expert staff both develop the custom software and execute the associated bioinformatics analysis for their customers. In contrast to software companies and companies offering software-as-a-service, customers are not responsible for running software on their own hardware nor paying for access to a bioinformatics portal to facilitate the analysis of their large-scale data.

As discussed in the scientific report, there exists a major gap between the rate at which large scale datasets can be generated and the rate at which computational tools can be developed for the management, storage and analysis of this data. Given this gap, although an increasing number of researchers are utilizing high-throughput sequencing as a means of answering their biological questions, most lack the hardware and expertise to effectively perform an appropriate bioinformatics analysis in their own labs through use of their own resources. Intuitive Genomics addresses this distinct customer need by serving as each customer's on-demand, outsourced bioinformatics division. Intuitive Genomics prevents researchers from a couple unfavorable alternatives: 1) Cobbling together a solution based upon outdated publically available tools and 2) Hiring and training an in-house bioinformatician.

Secondly, Intuitive Genomics is unique from its competitors because the company's team members hold expertise in both biological and computational sciences. It is because of this unique mix of knowledge that the company has the capacity to enter into customer projects at any stage in the process beginning with experimental design and ending with interpretation of results post-analysis. Intuitive Genomics collaborates with customers at any and all stages in project development where Intuitive Genomics' biological and computational expertise may benefit the customer.

Often, customer needs extend beyond the bioinformatics analysis into interpretation of these results in a biological context. Because Intuitive Genomics' team members hold expertise in both biological and computational sciences, the company is not only capable of performing the bioinformatics analysis, but in addition, can consult with a customer concerning the biological meaning behind the results of the analysis. The capacity to offer consulting in experimental design and results interpretation in addition to their primary service, bioinformatics analysis, permits Intuitive Genomics to offer a more complete solution to their customers. Intuitive Genomics has even gone so far as to facilitate the sequencing of customer samples prior to analysis of the data. Intuitive Genomics works with a select few sequencing service providers to facilitate this service. The immense capacity of Intuitive Genomics' service offerings can be visualized in Figure 12. The stars are indicative of those services in which Intuitive Genomics' specializes.

Figure 12: Capacity of Intuitive Genomics' Service Offerings

Lastly, Intuitive Genomics is unique in their handling of each customer project on a per-job basis. Intuitive Genomics expertly and rapidly customizes pipelines and software to meet each customer's unique and exact specifications as determined through close customer collaboration. In contrast to the "one size fits all" software solutions offered by Intuitive Genomics' competitors, the company creates custom solutions that specifically target each researcher's biological question. The result is findings that are remarkably more relevant and significantly higher quality. Customers remain involved throughout the development of their custom software solution ensuring that the solution is built to uncover answers to their specific biological question. Customers are guaranteed the output of meaningful results and only pay for the services they are provided.

2.2.4 Customer Needs and Service Benefits

Considering the current challenges of the life sciences community, Intuitive Genomics' customers have a few distinct needs. For one, customers require bioinformatics expertise in an effort to keep up with the increasing rate at which large genome-scale datasets can be generated. Secondly, customers require expertise in the interpretation of bioinformatics results if they are to gain valuable meaning from the analysis of their large-scale data. Lastly, researchers need custom tools built to perform functions specific to their biological question.

Customers benefit from Intuitive Genomics' services because the company's goals align directly with these customers' needs. As described in the above section, Intuitive Genomics offers bioinformatics expertise, performing custom bioinformatics services *for* its customers. In addition, the company is equipped to enter into collaboration with a customer at any stage in their project timeline. The core team can support customers in the design of their research project, can facilitate sequencing of biological samples, performs the bioinformatics analysis of the outputted data and can aid in the interpretation of analysis results in a biological context. Lastly, Intuitive Genomics' focus on the development of custom software and pipelines ensures that analysis results are succinct with the researcher's biological question. While a number of publically available bioinformatics tools as well as software solutions built by Intuitive Genomics' competitors are capable of performing standard bioinformatics analyses, these solutions do not guarantee researchers output consistent with their needs.

2.2.5 Marketing Strategies

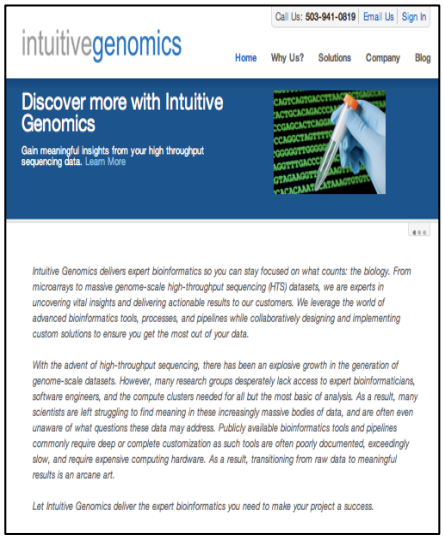
Intuitive Genomics focuses on a couple unique customer channels in an effort to bring business to the company. For one, Intuitive Genomics targets customers directly. Initial customers entering through this channel have done so through word-of-mouth. Personal connections of Intuitive Genomics' founding team resulted in a number of the company's initial customers.

Secondly, the company brings in customers through the formation of strategic partnerships with DNA sequencing providers. Often customers utilizing DNA sequencing providers for their high-throughput sequencing needs are not equipped to perform the downstream analysis of this data. Formation of a

strategic partnership brings to Intuitive Genomics an existing stream of customers while allowing the partnering company to greatly expand its offerings to include custom bioinformatics services. With Intuitive Genomics acting as their virtual bioinformatics division, these partners see increases in both their revenues and competitive positions. This sales channel comprises a more indirect means of bringing in new business. Intuitive Genomics has presently established one such partnership with a sequencing service provider in the St. Louis area and plans to pursue formation of additional partnerships, particularly with providers certified for use of Illumina's sequencing instruments. One of my research projects this summer involved the identification of companies with which potential partnerships might be established.

Over the course of my internship experience, I explored a few different methods for marketing Intuitive Genomics' services. These included direct communication with researchers at the Danforth Plant Science Center and BRDG Park, compilation of potential leads from local university websites to be targeted in a mass e-mail, and the execution of inbound marketing techniques.

Direct communication with researchers as a means of bringing in new business became dramatically easier upon the company's relocation to St. Louis. The company's new office space in BRDG Park on the same campus as the Danfoth Plant Science Center offered the company direct access to potential customers at both facilities. To implement this strategy for bringing in new business, Intuitive Genomics' personnel would communicate with potential customers via e-mail or in person to set up a time for further discussion. In my first few weeks in Intuitive Genomics' new office I had the opportunity to make introductions to a number of other BRDG Park tenants. A couple of these tenants were flagged as potential customers based upon the goals of their companies. Casual conversations and informal social gatherings were established for the purpose of informing the potential customer about the goals and service offerings of Intuitive Genomics. Likewise, Drs. Mockler, Carrington and Bryant kept the company in mind as they interacted with researchers at the Danforth Center. Contact information for any potential customer from the Danforth Center was passed along to Nathan Williams and myself. In order to protect the company's separation from the Danforth Center, Nathan and I were in charge of furthering conversations about Intuitive Genomics with these individuals. Marketing efforts on behalf of BRDG Park and the Danforth Center inform each entity of the happenings at the other. A quarterly newsletter released by BRDG Park in October informed approximately 450 individuals in the St. Louis community of Intuitive Genomics' presence at BRDG Park. In addition to personal introductions, marketing efforts such as this will do well to inform potential customers of the company's existence. Over the course of my internship, a couple key connections were made through my personal communications within the BRDG Park and Danforth Plant Science Center community. Although a key resource for other aspects of company development, this community serves as a relatively limited resource for the purpose of landing new customer jobs.



A second approach to bringing in new customers involved the execution of a variety of inbound marketing strategies. The overall goal of inbound marketing is the creation of an online presence such that interested customers find your company instead of the company having to track down these same customers themselves. Various strategies included the creation of web advertisements via Google Adwords, exploitation of company information via social media networks, posting company-related content to content rating sites like digg and reddit, and optimizing search engines to tie specific keyword searches to company web pages.

The implementation of inbound marketing strategies for Intuitive Genomics first involved the generation of a website. The company's website was launched during my internship period, my contribution to the launch comprising the development of service descriptions for informational pages,

Figure 13: Intuitive Genomics' Newly Launched Website

white papers for landing pages as well as articles for the company's blog. A number of these blog articles and white papers are included in the appendix of this report. The two white papers included comprise user guides for both the selection of an appropriate high-throughput sequencing instrument (appendix pg. 38) and selection of an appropriate Personal Genome Machine (appendix pg. 46). The blog articles represent my opinion on the co-existence of high-throughput sequencing instruments and personal genome machines in both the short (appendix pg. 55) and long term (appendix pg. 57) as well as outline the challenges (appendix pg. 59) and opportunities (appendix pg. 61) resulting from the emerging role of biocomputing in the life sciences.

Another inbound marketing task involved the creation of social media sites for the company. For the minimum purpose of reserving the website, user profiles were created for the company on facebook, twitter, linkedin, and youtube. Intuitve Genomics' business page on facebook as well as twitter profile were used to update followers on the post of new blog articles as well to share other company-related news.

Both Google Adwords and Hubspot software were utilized on a trial basis in an attempt to prove the effectiveness of inbound marketing as a mechanism for bringing in qualified leads. Google Adwords offers users pay-per-click advertising built around specific keywords chosen by the user. The user is only charged when a browser clicks on the ad and can set a daily monetary limit as well as pre-set the length of the advertising period. Hubspot offers users a collection of tracking and analysis tools for the purpose of optimizing their inbound marketing experience. The software can be used to physically build a webpage, landing page or blog article, can be used to monitor the success of competitor companies in their execution of similar marketing techniques, and provides a suite of analysis tools for optimizing the channels through which inbound leads are brought to the company. During the internship period, two Google Adwords trials were implemented and a month free trial of Hubspot was monitored. Overall, these inbound marketing strategies proved to be less effective in bringing in new customer business when compared to direct customer interactions in the company's new work environment in St. Louis.

The final marketing strategy considered in an effort to bring in new customers involved taking advantage of contact lists derived from local universities. Washington University exists as a rich pool of research talent and therefore has the potential to provide a great offering in terms of incoming customers. In an effort to leverage this talent pool we considered creating lists of contacts from local universities based upon receipt of research grants or area of research focus. Upon compilation of faculty lists, this strategy involves sending targeted e-mails to these individuals addressing their potential pain in bioinformatics and introducing the company's expertise. Although considered as a preliminary method for targeting potential customers in the local community, the targeted e-mails have not yet been sent for this purpose.

Besides execution of the above strategies for the purpose of bringing in new customer business, I held a variety of additional responsibilities linked to both sales and marketing. For one, if a potential customer's project involved sequencing of a series of samples prior to analysis of the data, I communicated with partner sequencing facilities to collect information for the customer quote. Based upon quotes relayed to me from the sequencing provider, I was in charge of formulating customer quotes representative of the entire job by combining the costs of sequencing and the bioinformatics analysis required for the project. In addition to obtaining information regarding pricing, I communicated with these same partner facilities to inquire about shipping instructions as well as to obtain updates on the sequencing jobs once they were underway.

Particularly during the second portion of the summer following the company's organizational change, I was the lead contact for both current customers and any potential leads. I participated in weekly team meetings, an opportunity for all team members to provide updates on various happenings with the company, as well as attended all customer related social gatherings.

Lastly, I was responsible for the development of marketing materials to be made available at the Annual Danforth Center Fall Symposium. Materials generated for this purpose included a marketing flyer, placed

in all attendees' tote bags as well as available for pick up at the Intuitive Genomics' vendor table, a 1-page advertisement, highlighted in the symposium program, and a powerpoint slide, part of a slideshow live whenever talks were not in session in the auditorium. An 8-foot banner displaying the Intuitive Genomics' logo was also printed for display behind the company's vendor table. As the only Intuitive Genomics' affiliate not attending the research symposium as a Danforth Center employee, I was solely responsible for manning the Intuitive Genomics' table during all sessions of the vendor show. This responsibility required that I speak on behalf of the company when inquires were made as to the company's services, incorporation, and recent move to St. Louis.

2.3 Finances

Upon the incorporation of Intuitive Genomics in August 2010, the four founding team members each owned one-fourth of the company in the form of vested equity shares. As a part of Nathan Williams' resignation, his share in the company was purchased by the remaining three founders resulting in a new financial structure. The financial implications of this change in management are visually represented in Figure 14.

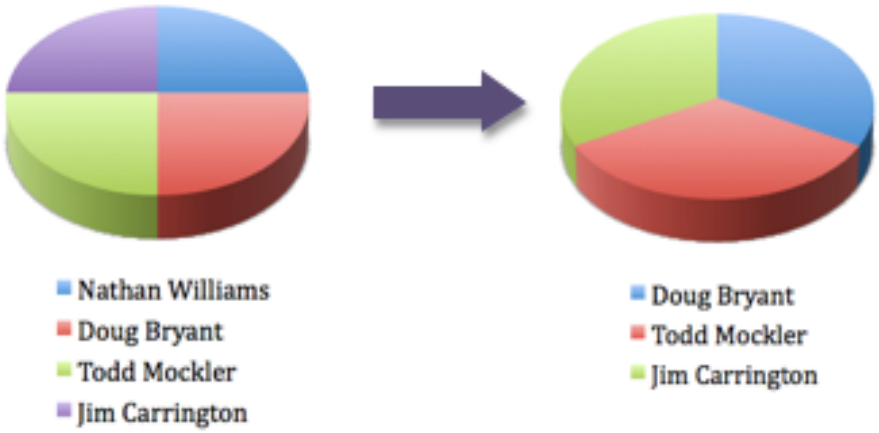


Figure 14: Financial Implications of Change in Management

2.3.1 Expenses

Intuitive Genomics is a privately held bootstrapped startup meaning all operations of the small business up to this point have been funded by initial customer jobs. The company has had customers from day one and has therefore been able to use this initial revenue as a means of supporting the daily operations of the startup. The two major expenses the company faced in the first year of operation included fees associated with the company's incorporation as well as expenses required for use of Amazon's cloud on a per-job basis to run customer bioinformatics jobs. Other minimal expenses were accrued in the generation of a company logo, purchase of business cards for all company affiliates and payment for any social gatherings with current customers.

Upon the company's move to St. Louis, rent for occupancy of the new office space in BRDG Park was introduced as a recurring expense. A 12-month lease was signed starting August 1st resulting in a single payment due at the start of each month from that point forward. This single payment includes rent for occupancy of both a corner office and a single cubicle, all related utility payments (high-speed internet excluded), wifi, local phone service, access to a shared copier/printer/fax machine as well as access to the break room, conference rooms and the mail room on the floor. The company pays an additional monthly fee for high-speed Internet access. Additional non-recurring expenses included the purchase of office

furniture, the purchase of new business cards, the purchase of marketing flyers and a banner for use at the company's first vendor show, and payment for any customer-related social gatherings. Lastly, property insurance, effective one year from the date of issue, was purchased as required by the property management company.

Hired to serve as the company's Marketing Manager for a 3-month period, I was the company's first employee. I was paid an hourly rate based upon a 40-hr work week for a total of 15 weeks. In addition, a bonus structure was put in place at the beginning of the internship period based upon accomplishment of weekly and monthly goals as well as dependent on new customer business brought in during the internship period and calculated relative to a revenue goal of $45,000. My bimonthly wages, including a small fee for shipment of each paycheck, comprised another recurring expense for the company beginning the week of June 20, 2011.

Monthly rent, high-speed Internet, and any use of the Amazon cloud will continue on as the company's main recurring expenses. In addition, I will continue to work for the company on a contract basis generating a small income via an hourly rate that the company will pay out on a monthly basis.

The company will continue to bring in revenue as a result of completed customer jobs. Intuitive Genomics charges customers on a per-job basis for the work they complete. The company's current cost structure comprises both a base cost for the development and execution of custom pipelines and software as well as an incremental charge used to increase the cost of the job based upon the number of samples, number of analyses or complexity of the analysis. Base and incremental costs are both based upon the type of data, size of the dataset, processes required for the analysis as well as extent of customization the job requires. Worked into the total bioinformatics cost is a component representative of the consulting service that inevitably becomes a part of the process.

For customers who only need assistance with the analysis piece of their project, the base and incremental bioinformatics costs are the sole contributors to the quoted price for the job.

Often, customers not only need Intuitive Genomics' services for the analysis of their large genome-scale datasets but they require assistance with the actual sequencing piece. In these cases, Intuitive Genomics must quote the customer a total price representative of both the bioinformatics costs and the cost of sequencing. The sequencing cost is based upon quotes from sequencing service providers with whom the company works closely.

In all cases, the bioinformatics piece is that which brings in direct revenue to the company. Currently, the company has several bioinformatics jobs underway and can therefore expect to receive revenue from these customers based upon the initial quote. Capture of any of the potential customers the company is currently in discussion with would bring in additional revenue.

Given that the overarching goal of all my efforts this summer was to bring in new customer business, the revenue the company will receive as a result of current customer jobs is in part reflective of my efforts. This concept was visually displayed in Figure 10. It is expected that any work I continue to perform for the company will position the company to bring in additional new business.

2.3.2 Fundraising

Aside from revenue, another area of Intuitive Genomics' finances I was directly involved with over the course of the internship was fundraising. As a startup corporation, a good portion of our time upon relocation to St. Louis involved meetings and research in an attempt to identify potential funding opportunities. An initial seed round would financially support validation of the company's go-to-market strategy as well as allow for the expansion of the company's bioinformatics technology to include software products. Specifically, the funds would go towards basic operational expenses, hiring additional technical and customer development staff, and acquiring secure compute infrastructure.

Initially, a number of meetings were scheduled with authorities at the Danforth Plant Science Center and affiliates of investment companies housed in BRDG Park. Goals of these meetings were to provide the company with a list of local contacts for the pursuit of various funding opportunities as well as to provide the company with an informal venue in which to outline a couple distinct product ideas for review by experienced entrepreneurs. The COO of the Danforth Plant Science Center offered a multitude of recommendations in terms of local contacts and two experienced entrepreneurs affiliated with Nidus Partners offered solid advise concerning both the company, itself, and Intuitive Genomics' two most current product ideas.

A meeting with St. Louis County and BioGenerator provided the company with information on two specific local funding opportunities, BioGenerator's i6 Project and St. Louis County's Helix Fund. BioGenerator's i6 Project provides up to $125,000 to companies based upon a specific project proposal. Intuitive Genomics submitted a proposal for the i6 Project for a software product efficient in data compression, storage and retrieval of genome-scale datasets. Intuitive Genomics' initial proposal was chosen for advancement into Phase B of the application process. Movement to Phase B required that the company give a short, 15-minute presentation on the proposed project to the BioGenerator Core Team. If the proposal is selected, Intuitive Genomics will receive up to $125,000 to use towards the creation of the proposed software product. There seems to be a high probability that this project will be funded. Not only would successful funding through the i6 Project provide the company with some initial funds to cover basic operating costs, the hiring of new computational and administrative personnel as well as the purchase of computational infrastructure, but the creation of a product in the form of data compression and storage software would be a great first step in productizing Intuitive Genomics' proprietary software. Existence of an Intuitive Genomics' product in high demand in the genomics field would immediately make the company more appealing to venture capitalists, increasing the company's chances for third party follow-on funding by a venture capitalist or angel investor. St. Louis County's Helix Fund has not been pursued at this point.

The company's next steps concerning the pursuit of funding opportunities will involve submission of proposals for Small Business Innovation Research (SBIR) and/or Small Business Technology Transfer (STTR) Grants through various federal funding agencies. The most immediate proposal submissions will likely be in response to solicitations by the National Science Foundation (NSF) and National Institutes of Health (NIH). I will play a large role in the preliminary research and submission of these proposals and will in this way remain directly involved with Intuitive Genomics' financial situation. A visual depiction of the company's two key revenue channels can be viewed in Figure 15.

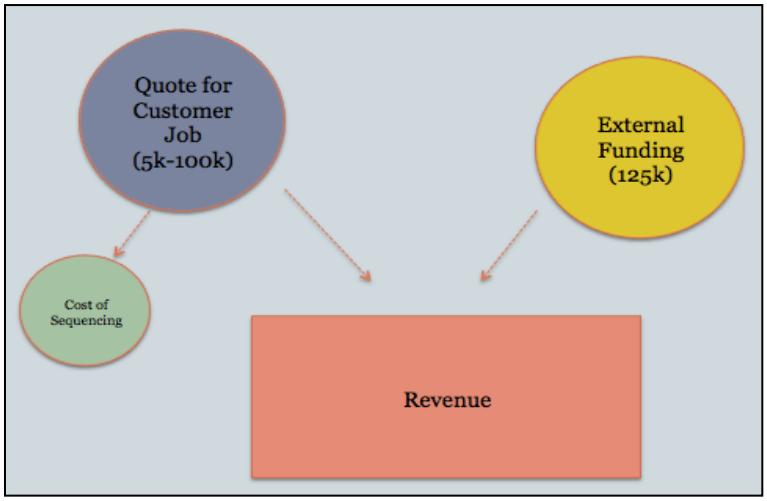


Figure 15: Anticipated Revenue Channels

2.4 Company Management and Human Resources

As described in detail in the section addressing "administrative structure, " I worked most closely with Nathan Williams, CEO, for the first part of the summer. At this time, Nathan Williams and I served as the business end of the company, Dr. Doug Bryant with guidance from Dr. Todd Mockler served as the company's scientific expertise, and Dr. Jim Carrington served in an advisory role. Despite my greater interaction with Nathan Williams, given the small size of the company, I was also very well connected with the other team members.

In a startup environment, all major action items require close collaboration among all team members. In this regard, titles given to team members are reflective of the types of responsibilities these individuals hold but are not necessarily reflective of the experience or qualifications of the individual. I learned quickly the importance of dissolving hierarchies when it comes to establishing leadership within a startup as many times, all team members are experiencing things for the first time. It is important that the management team is poised to learn together when it comes to facing new challenges and making decisions that will affect the downstream success of the company.

Based on the above, my interactions during the internship period were not limited to Nathan Williams because he and I worked most closely. I spent time each week with all members of the management team and felt well exposed to all aspects of the company.

Mentioned briefly in the "administrative structure" section, an organizational change taking place a couple weeks before the end of the internship period resulted in the resignation of Nathan Williams from his management role with the company. Both the decision itself and the transition period that followed marked key learning experiences for me.

The reason for Nathan's resignation was founded in a strategic difference in opinion regarding the future directions of the company. This difference in opinion between Nathan Williams and the other three members of the board of directors ultimately led to his exit from his CEO position with Intuitive Genomics. As a firsthand witness to this major organization change, both the reason for establishment of a board of directors and the importance of defining clear strategic goals were highlighted. Being witness to this change in management also highlighted the dynamic nature of startup life. I learned that it's often imperative to make changes in the early stages of a company's existence that reflect the best interests of the company. The founding team members, also comprising the board of directors in the case of many young startups, are often asked to make sacrifices for the betterment of the company, and must act quickly in making big decisions as these decisions will affect the long term success of the startup.

Doug Bryant, Todd Mockler and Jim Carrington acted on behalf of Intuitive Genomics when they made the decision to offer Nathan Williams a resignation package. While they may have sacrificed the immediate well being of the company by ridding of their CEO, they ultimately made the best decision for the company as opportunity for solid leadership will emerge in the future of the company.

Given my involvement with the company both before and after the organizational change, I was witness to the two very different management styles. The company's focus prior to Nathan's resignation was on the establishment of an online presence with the intent of bringing customers to Intuitive Genomics of their own effort. The focus upon Nathan's exit involved capitalizing on the connections that had been made with potential customers in the local St. Louis area. A few of these potential customers, if obtained, would offer Intuitive Genomics immense and recurring business meaning a steady revenue stream and flexibility to explore various marketing and sales mechanisms. A second focus following the change in management was on the pursuit of funding opportunities. The company is hopeful that success in bringing in some preliminary capital will fund the creation of a software product, increasing Intuitive Genomics' market presence and making the startup more appealing to venture capitalists.

In terms of my involvement with the management team post organizational change, Dr. Doug Bryant and I initiated a much stronger collaboration during the last two weeks of the internship period. Dr. Bryant spent his mornings collaborating with me at BRDG Park concerning customer interactions, funding opportunities and various administrative duties. While Dr. Bryant's increased dedication to Intuitive Genomics on top of his post-doc position has been sufficient during the current period of transition, the continued forward movement of the company is dependent on identification of an individual to serve as the company's CEO. A strong leader's full-time dedication to the company will drive the company rapidly towards success.

While the major job responsibilities of the position I was hired to fill this summer mostly surrounded marketing and sales initiatives, given that a move marked a major company event during the time I was actively with the company, I also took on a variety of administrative roles. In the first few weeks of the company's occupancy of BRDG Park, I was the only Intuitive Genomics affiliate occupying the space. Therefore, I served as the administrative contact for establishment of the company in its new office space. I worked closely with both the Senior Property Manager and Business Development Officer to ensure that all lease documents were properly submitted, last minute maintenance on the new space was completed, and utilities included in our monthly rent were actively running. I was issued key cards for access to the space after hours as well as given keys to the office. I made introductions to neighboring employees, some of whom were targeted as potential customers, and ultimately served as the face of the company for the occupants of BRDG Park during the first few weeks of my time in St. Louis. I relayed all relevant information back to the team members working at the Danforth Center in an effort to ensure that everyone was on board with the happenings in the new office space. A detailed account of all tasks completed over the course of the internship can be viewed in the Internship Journal (appendix pg. 62).

2.5 Conclusion

Overall, my commitment to Intuitive Genomics over the course of a 3-month internship exposed me to a dynamic entrepreneurial environment, educating me in a variety of business arenas. Coupled with my education in the sciences and coursework in a variety of business topic areas, the experiences I acquired working in an industrial setting have prepared me well for a career with a scientific corporation.

Two unique experiences over the course of my pursuit of the Professional Science Master's degree have exposed me to both the challenges of uncovering meaning from genome-scale datasets and the opportunities that have arisen for individual researchers and companies attempting to overcome this challenge. Given my passion for personalized genomics, the use of DNA sequencing data for the purpose of developing individualized diagnostics and therapeutics, exposure to a solution to the challenge of uncovering meaning from large datasets was directly in line with my area of interest.

I am confident that my experiences as a Professional Science Master's Student at Oregon State University have prepared me well to enter into my desired career field.

References

Bio-Research and Development Growth (BRDG) Park at the Danforth Plant Science Center. (2010). http://brdg-park.com/welcome.html

Georgia Institute of Technology. (2010). GeneMark™. http://exon.gatech.edu/

Indiana University. The Center for Genomics and Bioinformatics. (2011). WebGBrowse 2.0, A Web Server for GBrowse. http://webgbrowse.cgb.indiana.edu/cgi-bin/webgbrowse/uploadData

National Center for Biotechnology Information. (2009). BLAST. http://blast.ncbi.nlm.nih.gov/Blast.cgi

The Donald Danforth Plant Science Center. (2011). http://www.danforthcenter.org/default.asp

Appendix

Visual Representations of *C. drosophilae* Genomic Contigs in WebGBrowse

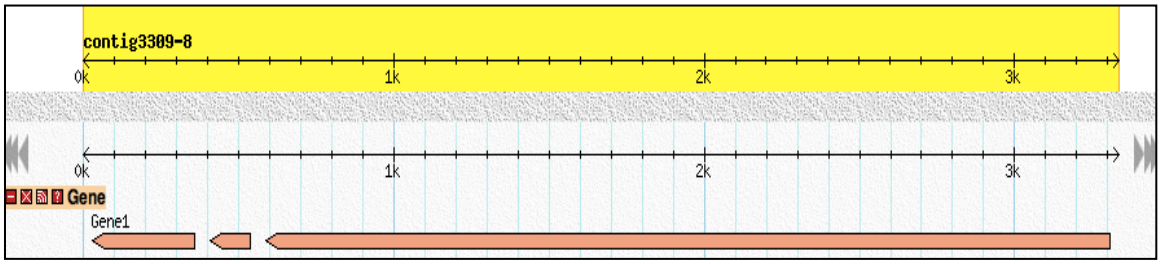


Figure 16: Contig 3309-8; Gene track = generic glyph (Connector = none)

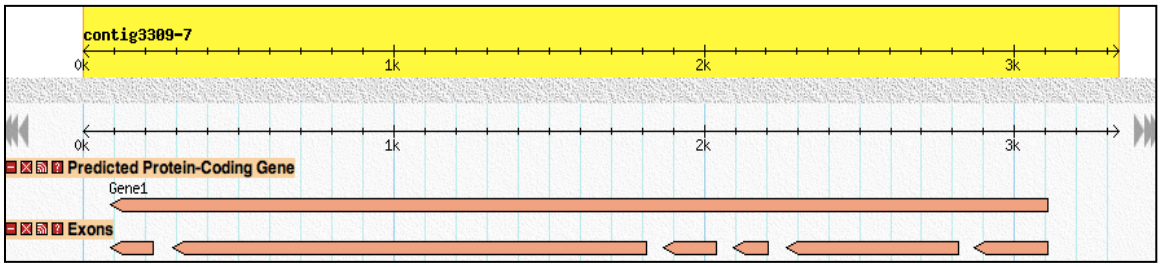


Figure 17: Contig 3309-7; Gene track = box glyph; Exon track = generic glyph

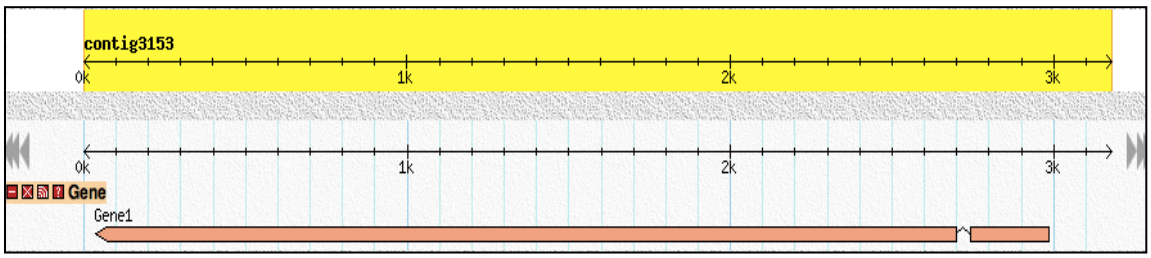


Figure 18: Contig 3153; Gene track = gene glyph

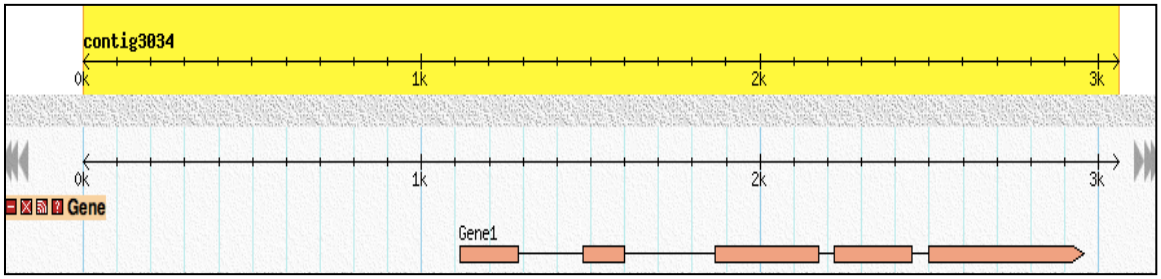


Figure 19: Contig 3034; Gene track = gene glyph (Connector = solid)

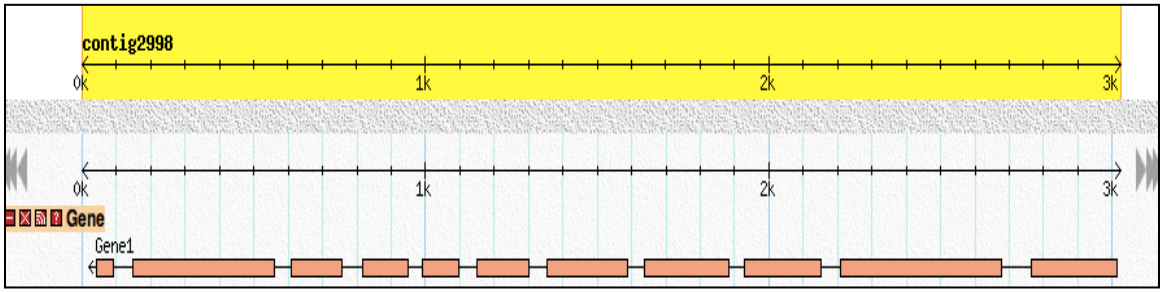


Figure 20: Contig 2998; Gene track = line glyph

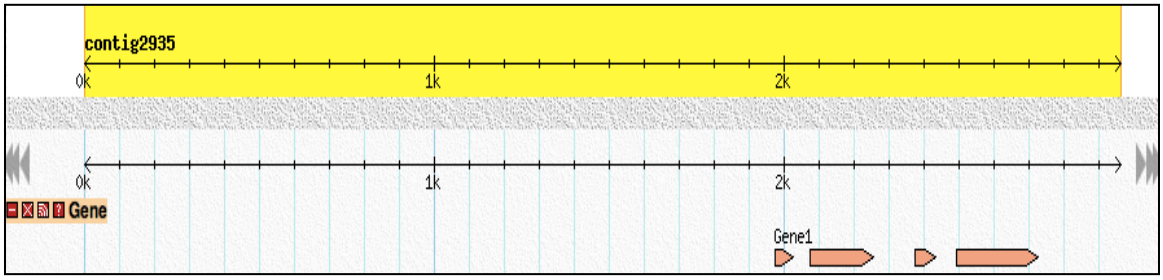


Figure 21: Contig 2935; Gene track = generic glyph (Connector = none)

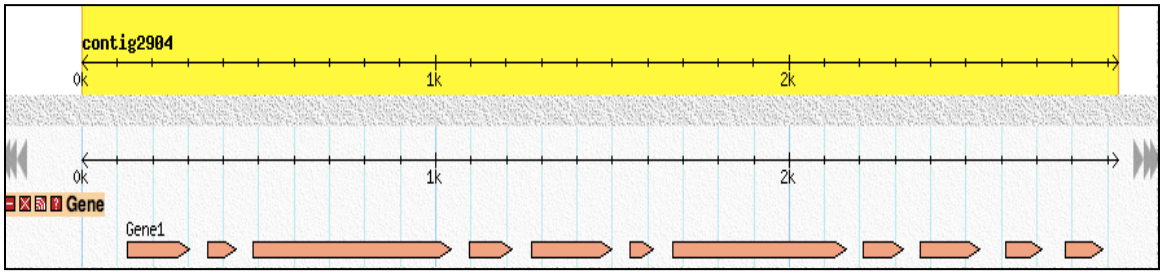


Figure 22: Contig 2904; Gene track = generic glyph (Connector = none)

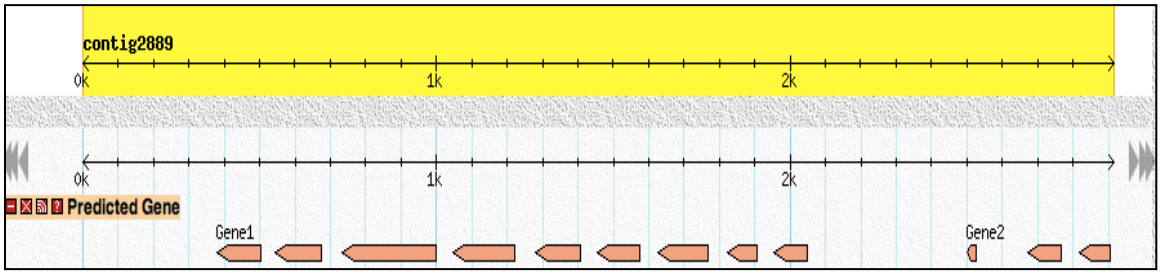


Figure 23: Contig 2889; Gene track = generic glyph (Connector = none)

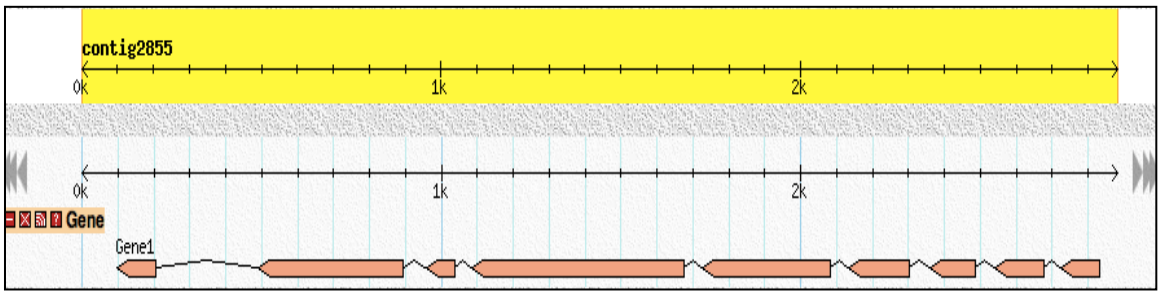


Figure 24: Contig 2855; Gene track = generic glyph (Connector = hat)

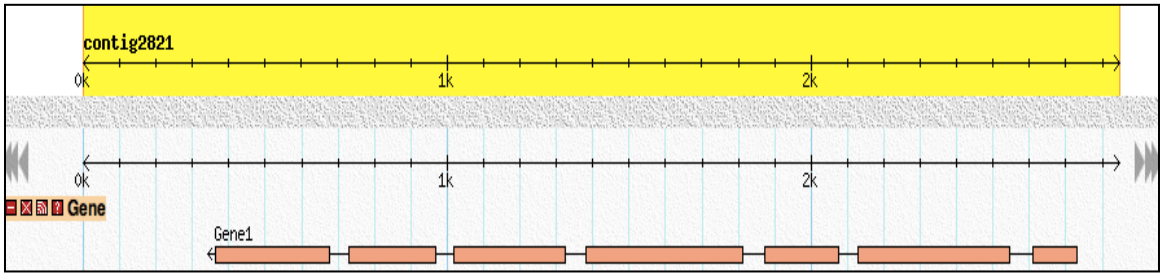


Figure 25: Contig 2821; Gene track = line glyph

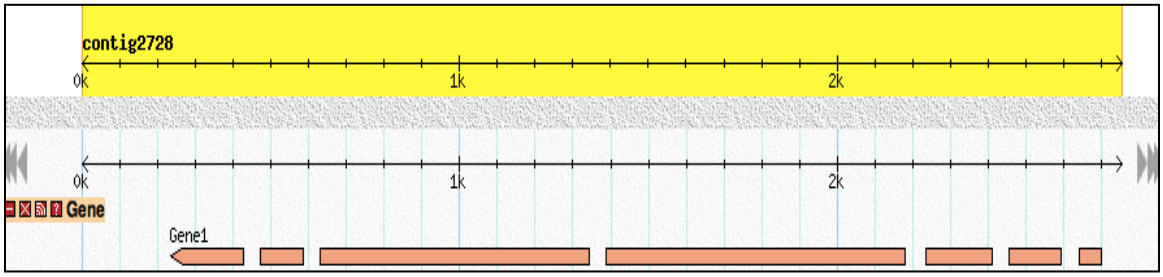


Figure 26: Contig 2728; Gene track = gene glyph (Connector = none)

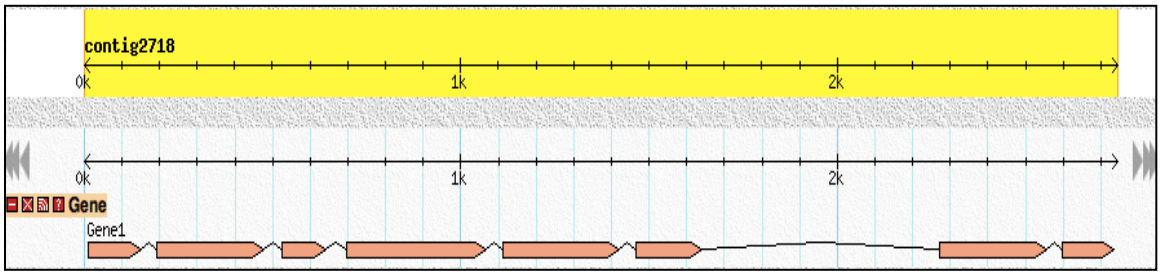


Figure 27: Contig 2718; Gene track = generic glyph (Connector = hat)

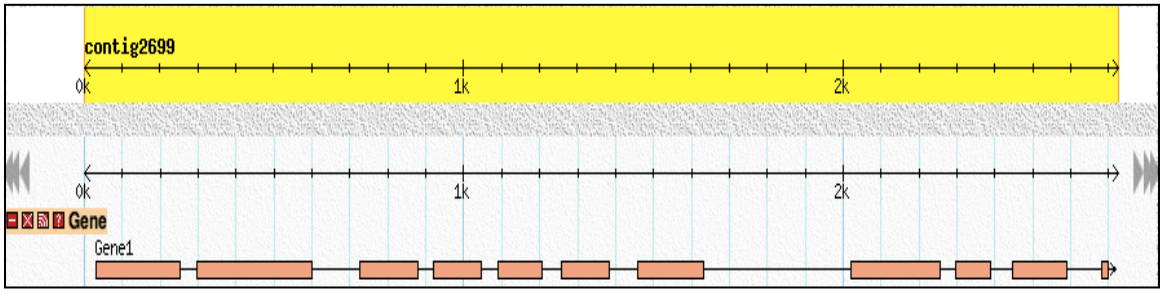


Figure 28: Contig 2699; Gene track = line glyph

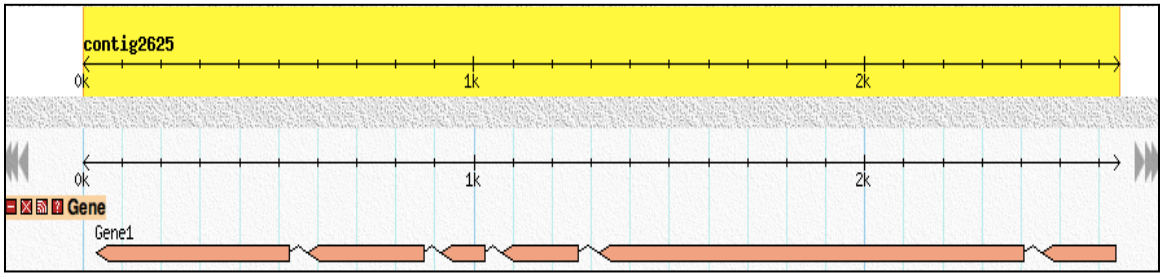


Figure 29: Contig 2625; Gene track = generic glyph (Connector = hat)

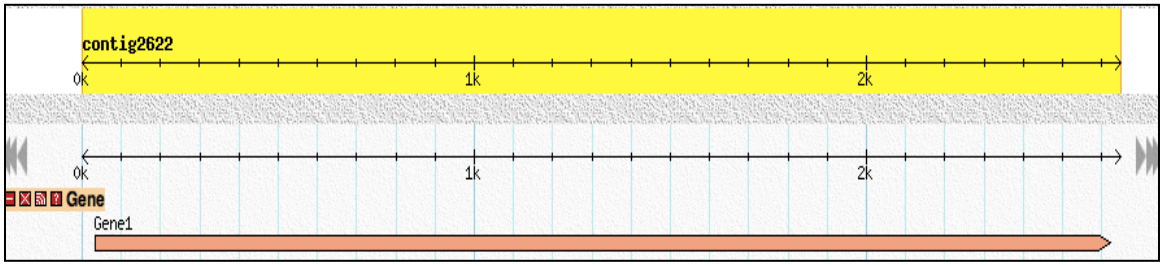


Figure 30: Contig 2622; Gene track = generic glyph (Connector = solid)

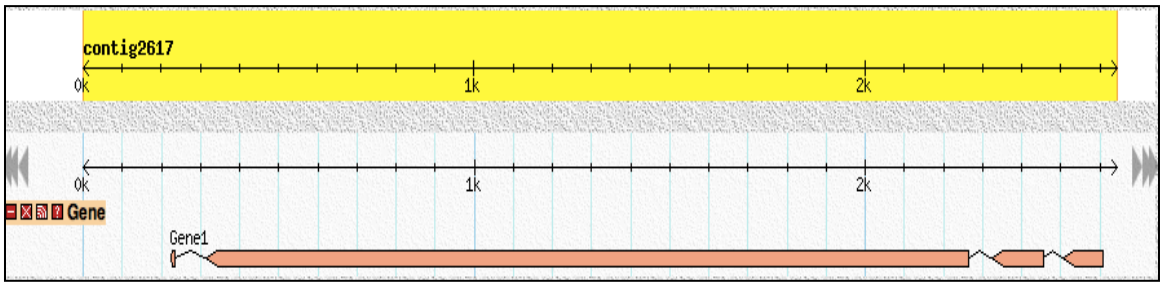


Figure 31: Contig 2617; Gene track = generic glyph (Connector = hat)

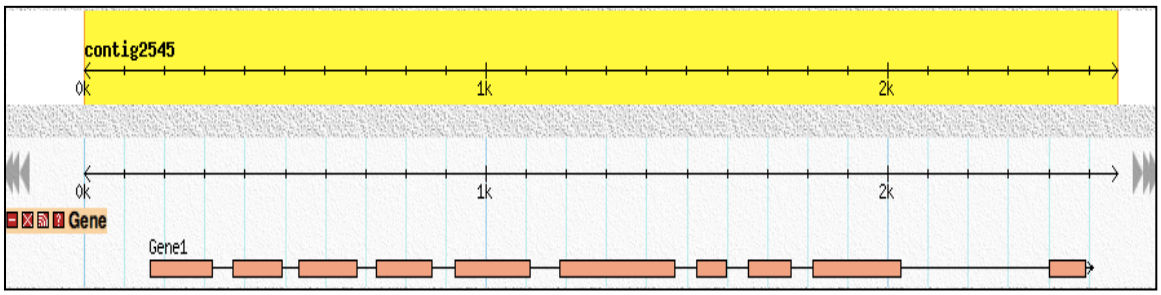


Figure 32: Contig 2545; Gene track = line glyph

Bio-Research and Development Growth (BRDG) Park
at The Danforth Plant Science Center
St. Louis, Missouri

BRDG Park aspires to "…help life sciences and clean-tech companies bridge research, resources and relationships to achieve commercial success" (BRDG Park, 2010). Tenants of BRDG Park benefit from world-class wet labs, office space and an on-site workforce-training program. BRDG Park's presence on the Danforth Plant Science Center Campus offers emerging scientific enterprises an ideal combination of interactions between top scientists and access to state-of-the-art core facilities namely, technologically advanced greenhouses, growth chambers, microscopy, proteomics, and tissue transformation facilities (BRDG Park, 2010).

Located in suburban St. Louis County, Missouri, the research park is being developed by Wexford Science + Technology, a privately held real estate development and investment company that has developed six major research parks nationwide. The single building currently representing BRDG Park will soon expand to a multi-building campus further impacting the bioscience community in St. Louis.

Figure 33: BRDG Park Building 1

The current building is home to an expanding number of promising plant and life sciences companies at a variety of commercialization stages. Current tenants include Divergence (Monsanto), leader in the development of products for the control of parasites in agriculture, St. Louis Community College, offering an on-site biotech workforce development and training program, Phycal LLC, an algae biotechnology laboratory, SyMyco, developers of a family of biological fertilizers, etc. Intuitive Genomics is one of the most recent tenants to occupy an office space.

BRDG Park has a commitment to:

- Provide world-class scientific research and development facilities
- Help attract investment capital and government funding sources
- Support workforce development through an on-site college program
- Maintain a progressive environment for commercial growth and success

(BRDG Park, 2010)

The Donald Danforth Plant Science Center
St. Louis, Missouri

Located in suburban St. Louis, Missouri, The Donald Danforth Plant Science Center is a not-for-profit research institute with a mission of improving the human condition through plant science. Scientists at the Center are engaged in research that strives to enhance the nutritional content of plants, increase agricultural production to create a sustainable food supply, reduce the use of pesticides and fertilizer, develop new and sustainable biofuels, and generate scientific ideas and technologies that will contribute to the economic growth of the St. Louis region (Danforth Plant Science Center, 2011).

Specific areas of research focus at the Center include biofuels, biofortification, disease resistance, drought tolerance, pesticide and fertilizer reduction, and biosafety and regulation (Danforth Plant Science Center, 2011).

Events such as the Annual Danforth Fall Symposium and Ag Showcase facilitate the congregation of leading plant scientists on-site creating a tremendous opportunity for networking and idea sharing.

Ultimately, The Danforth Plant Science Center provides a tremendous resource for the St. Louis community and further supports the fast growing bioscience community in the region.

Figure 34: Danforth Center Welcome Sign

Figure 35: Front Entrance of Danforth Center

Images From Internship Experience

Figure 36: View of Danforth Center from BRDG Park Office

Figure 37: Intuitive Genomics' BRDG Park Office

Figure 38: Intuitive Genomics' Vendor Table at Danforth Symposium

A Comprehensive Guide to High-Throughput Sequencing Platforms

Introduction

Researchers today have choices when it comes to their high-throughput sequencing (HTS) needs. Given the variety of available HTS platforms, it is important that researchers are informed of the key differences among platforms such that they can effectively select a technology most appropriate for their project goals. This resource illustrates key features of a few of the most popular HTS platforms including Illumina's HiSeq 2000, Roche 454's FLX Titanium XL+, and Applied Biosystems' SOLiD™ 5500xl. An overview of the basic characteristics and performance parameters of each platform can be viewed in Table 3. The accompanying sections further explore the offerings of each platform and highlight those features that differentiate the various platforms from one another.

Table 3: Key Differences Among the Leading High-Throughput Sequencing Platforms

Company	Illumina	Roche 454	Applied Biosystems (Life Technologies)
Platform Name	HiSeq 2000	FLX Titanium XL +	SOLiD 5500xl
Chemistry	Reversible Terminator (SBS)	Pyrosequencing (SBS)	Exact Call Chemistry (ligation-based)
Prep Surface	Flow cell (single and dual)	Beads in a PicoTiterPlate	Individual beads (1.0 um) attached to 1-2 flow chips
Amplification	Bridge Amplification	Emulsion PCR	Emulsion PCR
Cost of instrument	$690,000	~$500,000	$595,000
Run Time	2 -11 days (dual flow cell)	23 hrs	2 – 7 days
Throughput	Up to 55 Gb / day (2 x 101 bp)	700 Mb / day	10 – 15 Gb/day
Read Length	2 x 101 bp	Up to 1,000 bp Mode - 700 bp	MP: 2 x 60 bp PE: 75 bp x 35 bp Fragment: 75 bp
Read Quality	> 80% > Q30 (99.9%) (2 x 101 bp)	99.997% consensus accuracy (15x coverage)	99.99% on the highest percent of bases ≥ Q40
Associated Accessories / Software	cBOT cluster generation IlluminaCompute Data processing, storage, analysis	REM e System automated emPCR GS Data Analysis Software Package	SOLiD™ EZ Bead™ System automated emPCR LifeScope™ Genomic Analysis Solutions
Cost of Accessories	cBOT - $55,000	REM e System - $18,900 GS Software included in cost of instrument	SOLiD™ EZ Bead™ System- $60,000
Reagent costs / run	$5,750	$3,495	$3,200

*SBS = Sequencing-by-synthesis *bp = base pair *MP = Mate pair *PE = Paired end *emPCR = Emulsion PCR

Illumina HiSeq 2000

Introduction

Illumina broadly released the HiSeq 2000 sequencing platform in March of 2009, boasting the platform's ability to sequence two human genomes (30x coverage) in a single instrument run for under $10,000 per genome[1]. Other highlights included the instrument's output, user experience, and cost-effective operation.

Chemistry

The HiSeq 2000 leverages Illumina's proven and widely adopted reversible terminator-based sequencing-by-synthesis (SBS) chemistry. This chemistry, coined TruSeq, supports the detection of single bases as they are incorporated into growing DNA strands. Each dNTP is coupled to a unique fluorescently labeled terminator. The respective terminator is imaged as each dNTP is incorporated, followed by cleavage to allow for the incorporation of the next base. All four reversible-terminator-bound dNTPs are present during each sequencing cycle facilitating natural competition and minimizing incorporation bias. Individual bases are called based upon fluorescent signal intensity measurements made during each cycle[2].

Cost and Throughput

At ~$690,000 list price, the HiSeq 2000 marks the most expensive of the popular HTS platforms. Yet, the Illumina Hieq 2000 is also the instrument with the highest throughput. The instrument outputs a maximum of 600 Gigabases (Gb) per sequencing run, that is ~ three billion pairs of 101 base pair (bp) reads. Dependent on the number of cycles, a sequencing run can take as few as 2 and as many as 11 days generating ~55 Gb of sequence data/per day for a 2 x 101 bp run[2]. Max read lengths settle around 101 bp, short in comparison to those generated by Roche 454's platform. Reason for the increased throughput offered by the HiSeq 2000 can be found in Illumina's innovative dual-surface imaging method in addition to the available option to run one or two flow cells per sequencing run[1].

Multiplexing is possible on the HiSeq 2000 facilitating the sequencing of up to 12 samples per sequencing lane (96 samples per flow cell) through use of Illumina-provided sample preparation kits[3]. In addition, NuGen recently launched a sample preparation kit enabling multiplexing of up to 384 samples per sequencing lane (3072 samples per flow cell)[4]. The presence of eight independently configurable lanes on each flow cell supports sequencing of up to eight different sample types in a single run[3]. In addition, the choice to use one or two flow cells in any given run increases both the number and type of samples that can be sequenced simultaneously (up to 192 samples (6144 samples with NuGen's kit) of 16 different types). The choice to run one or two flow cells also supports the simultaneous run of applications requiring different read lengths[1]. The platform supports the preparation of samples for paired-end runs as well as mate-pair library preparation, both facilitated through use of specialized Illumina reagent kits and supportive of greater efficiency in sequence assembly.

Software and Accessories

In conjunction with their HiSeq 2000 platform, Illumina offers IlluminaCompute, a computing architecture developed for the processing and analysis of the platform's resulting genomic data[5]. The system comprises hardware (blade servers from Dell and modular storage from Isilon), software, and support services. IlluminaCompute does not rely on a pre-existing computer infrastructure and can be expanded or reconfigured to address researchers' changing sequencing needs[5].

Additionally, Illumina offers the cBot, an automated system for the generation of clonal clusters from single molecule DNA templates (via bridge amplification)[6]. Hands-on-time for this accessory is less than ten minutes compared to the more than six hours required for manual sample preparation.

Distinguishable Benefits

Overall, advantages of the HiSeq 2000 as compared to the other HTS platforms include:

- Greatest throughput per run
- Simultaneous sequencing of samples requiring different read lengths

Best Suited Applications

Based upon the above advantages, the HiSeq 2000 is best equipped for sequencing projects requiring large volumes of throughput such as those concerned with sequencing whole genomes; the sequencing of entire human genomes has been successfully completed on Illumina's platform. Additionally, the system offers significant benefits for gene expression and epigenetic profiling as compared to those offered by microarrays. Illumina's platform allows for the generation of richer transcript profiles while maintaining the cost and throughput expected from microarray technologies.

The HiSeq 2000's immense throughput makes the platform a good candidate for RNA-seq studies of alternative splicing, a recommendation that does not come without its drawbacks. Given the HiSeq 2000's generation of short reads, a significant number of reads are required to facilitate sufficient alignment and to generate knowledge of the transcript structure and therefore alternative splicing events. In addition, the short nature of the reads increases the probability of spurious alignments, particularly when the target genome is large or highly repetitive, and can therefore lead to significantly more false discoveries. Although short, the tremendous number of reads generated by the HiSeq 2000 platform makes RNA-seq a feasible application for this technology.

Finally, the HiSeq 2000's flexibility in terms of the choice to run one or two flow cells facilitates the simultaneous run of samples requiring different read lengths. For projects requiring this capability, researchers may benefit from the ability to run all of their samples at once as opposed to paying for two separate runs on a different instrument.

Roche 454 FLX Titanium XL+ (FLX +)

Introduction

454 Life Sciences announced the launch and immediate availability of the new GS FLX+ System in June 2011[7]. The instrument's ability to generate sequencing reads up to 1,000 bp in length presents a major milestone in the life science's industry, making it the first high-throughput sequencing technology to deliver millions of bases from reads with accuracy and lengths that are comparable to Sanger-based methods[7]. Early access projects have revealed the critical importance of the system's extended read lengths for a variety of applications including de novo sequencing and assembly of whole genomes, comprehensive transcriptome profiling, and metagenomic characterization of environmental samples[7].

The new GS FLX instrument is available as a new instrument or as an on-site upgrade to the existing instrument featuring a redesigned reagent compartment to accommodate the larger reagent volume of the Titanium Sequencing Kit[7].

Chemistry

The FLX + utilizes a variety of SBS chemistry, pyrosequencing, which relies on the detection of pyrophosphate release upon nucleotide incorporation. Solutions of adenine (A), cytosine (C), guanine (G), and thymine (T) dNTPs are sequentially added and removed from the reaction[8]. Incorporation of the correct, complementary dNTP by DNA polymerase results in the stoichiometric release of pyrophosphate (PPi) followed by conversion of PPi to ATP. The ATP acts as fuel to the luciferase-mediated conversion of luciferin to oxyluciferin, generating visible light in an amount proportional to the quantity of ATP[8]. Quantification of the intensity of visible light resulting upon addition of each nucleotide solution allows for the appropriate base calls to be made.

Cost and Throughput

At a cost of ~$500,000, the FLX + is the least expensive of the popular HTS platforms. Sacrificed in light of the lower cost is throughput, averaging only ~700 Mb per run. However, a single run on the FLX + takes a mere 23 hours as compared to a single run on either the HiSeq 2000 or SOLiD 5500xl which may take anywhere from two days to just under two weeks[9]. Read length for the FLX + averages around 700 bp and peaks at 1000 bp[9]. These values are significantly higher than those offered by either the HiSeq 2000 or SOLiD 5500xl and indeed represent a distinguishing feature of the FLX+.

The FLX+ has 132 multiplex identifiers at its disposal, manufactured for use with the technology. In addition, gaskets are available for the purpose of segregating the associated Pico Titer Plate into 2, 4, 8 or 16 sections supporting the simultaneous sequencing of up to 16 different sample types[9]. Similar to the HiSeq 2000 and SOLiD 5500xl, reagent kits manufactured by 454 support the preparation of samples for paired-end runs as well as facilitate mate-pair library preparation.

Software and Accessories

The GS Data Analysis Software package provided with Roche 454's FLX System comes at no additional cost[9]. This lies in contrast to the data analysis packages offered by Roche 454's competitors. The provided software package includes tools which may used to investigate complex genomic variation in samples including de novo assembly (GS de novo assembler), reference guided alignment and variant calling (GS read mapper), and low abundance variant identification and quantification (GS amplicon variant analyzer). Additionally, a monitor and desktop, integrated into the instrument, facilitate real time data processing. GS FLX+ computing stations are also available for purchase from 454 Life Sciences[10].

Roche 454 offers the REM e System as a platform accessory[11]. This liquid handler is designed to fully automate the emulsion PCR enrichment and sequence primer annealing steps in the 454 sequencing workflow. Use of the automated system can reduce up to five hours of hands-on work to 15 minutes of liquid handler setup, improves consistency by enhancing PCR enrichment accuracy and supports a wide variety of library types and all GS FLX and GS Junior Titanium Series emulsion formats. Although the REM e System appears to be significantly less expensive ($18,900) in comparison to the automated systems offered by Roche 454's competitors ($55,000 and $60,000), this is not necessarily the case, as the REM e System requires that the module be integrated into an acquired liquid handling platform, resulting in an additional investment in capital equipment[11].

Distinguishable Benefits

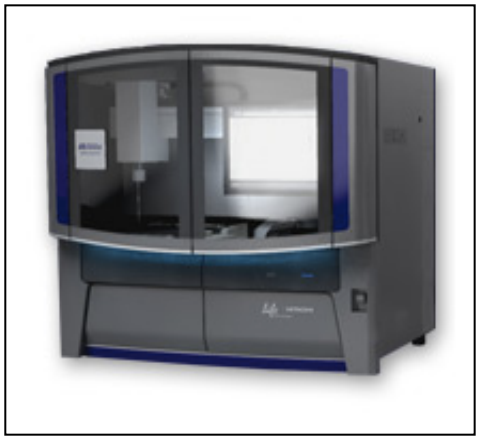
Overall, advantages of the FLX+ as compared to the other available HTS platforms include:

- Longest read length averaging 700 bp and peaking at 1000 bp
- Shortest run time at 23 hours
- Least expensive instrument cost at $500,000
- 132 multiplex identifiers
- Integrated monitor and desktop facilitating real time data processing and eliminating the need for a compute cluster
- GS Data Analysis Software included in cost of sequencer

Best Suited Applications

Given the key attributes of Roche 454's technology, the FLX+ is best suited for applications that do not require large-scale throughput and that benefit from accurate and relatively simplistic alignment of sequence data. For applications such as targeted re-sequencing, amplicon sequencing, and sequencing of bacterial and viral genomes that do not require tremendous throughput, Roche 454's platform offers a means of obtaining sequence reads as long as 1 Kilobase for a lesser overall capital investment in instrument cost and in a shorter amount of time. For applications such as metagenomic analysis, the 132 multiplex identifiers available for use allow for the simultaneous sequencing of 132 unique samples on a single Pico Titer Plate.

Applied Biosystems (ABI) SOLiD™ 5500xl



Introduction

ABI (Life Technologies) announced the launch of the SOLiD 5500xl genome sequencer in November 2010. The new instrument was designed to deliver the industry's fastest and most accurate genomic data for cancer biology and genetic disease research[12].

Chemistry

The SOLiD™ platform leverages Exact Call Chemistry (ligation-based sequencing), facilitated by a set of four fluorescently labeled di-base probes that compete for ligation to the sequencing primer[13]. Specificity of the di-base probe is achieved by interrogating every 1st and 2nd base in each ligation reaction. Multiple cycles of ligation, detection and cleavage are performed with the number of cycles determining the eventual read length[13]. Following a series of ligation cycles, the extension product is removed and the template is reset with a primer complementary to the n-1 position for a second round of ligation cycles. Five rounds of primer reset are performed prior to the run's completion[13].

Cost and Throughput

The SOLiD™ 5500xl finds itself in between the offerings of the Illumina HiSeq 2000 and Roche/454 FLX+ in terms of cost, throughput, and run time. Sold for $595,000, the SOLiD™ 5500xl generates 10 -15 Gb per day and supports a run time of between two days and one week[14]. With the introduction of high-density nano-bead technology, planned for the second half of 2011, the system is predicted to deliver 30 to 45 Gb/day, improving the platform's performance to a level more comparable to Illumina's HiSeq 2000 in terms of throughput per day[14]. Maximum read lengths vary depending on the library preparation, 60 bp for mate pair reads, 75 bp x 35 bp for paired end reads, and 75 bp for basic fragment reads, all significantly shorter than those generated by the FLX+ as well as shorter than those generated by the HiSeq 2000[14].

In addition to 96 barcodes for use in both RNA and DNA applications[14], the SOLiD™ 5500xl system utilizes two flow chips each comprising six independently addressable and configurable lanes[14]. This feature provides researchers with a choice of the number of samples to run each time the instrument is used as well as supports simultaneous sequencing of samples for up to 12 unique applications. Unique to this platform, reagent consumption is engineered independently for each lane meaning users only pay for reagent consumables in the active lanes when performing a partial run[15]. Paired-end runs and mate pair library preparation are supported by this platform.

Software and Accessories

Owners of 5500 Series Systems have access to the optimized algorithms and analysis pipelines of LifeScope™ Genomic Analysis Solutions[16]. Researchers have multiple options for the incorporation of LifeScope™ into their laboratory pipelines. They can install LifeScope™ Server Software on their own hardware, use a cloud computing option, or use preinstalled, preconfigured LifeScope™ hardware in conjunction with their own pipelines[16]. LifeScope™ Software facilitates sequence mapping and variant detection for a variety of workflows including whole genome sequencing, targeted re-sequencing and SNP detection.

In addition, the optional SOLiD™ EZ Bead™ System automates the SOLiD™ System workflow from emulsion PCR to templated bead deposition requiring less than 1-hour of hands-on time[17]. This lies in stark contrast to the ~ six hours that would be required for this same process if performed manually.

Distinguishable Benefits

Overall, advantages of the SOLiD™ 5500xl as compared to the other available HTS platforms include:

1. Pay-per-lane consumables eliminating reagent waste when performing a partial run
2. Superior accuracy

Best Suited Applications

Based upon the features listed above, ABI's SOLiD™ 5500xl is best suited for projects involving the detection of minor variants in heterogeneous samples as well as projects requiring rapid turn-around times. The high accuracy claimed by Life Technologies at > 99.99% in combination with the platform's relatively high throughput make this platform the ideal choice for experiments requiring detection of slight variants as proven accuracy greatly diminishes the opportunities for false discoveries.

"Pay-per-lane consumables" is unique to the SOLiD™ series platforms and offer researchers tremendous flexibility in the execution of their experimental timeline. The extremely low volumes of sequencing reagents utilized per lane coupled with each flow chip's construction of individually addressable and configurable lanes allow researchers to perform partial runs without fear of reagent waste. Although the typical run time for this instrument ranges between two and seven days, a single lane can be sequenced in one day to facilitate a quicker turn-around.

Conclusion

HTS platforms are very competitive with each other in many regards; however, there are unique sets of distinguishing features that should be taken into account when selecting a platform for a given research application.

The immense throughput offered by Illumina's HiSeq 2000 supports the completion of whole genome sequencing projects such as those involving human genomes. The HiSeq 2000 is currently the best choice of platform for projects involving the sequencing of samples requiring different read lengths as their simultaneous sequencing is supported through use of two flow cells.

Based upon the remarkable read length (up to 1 Kilobase) leveraged by this platform, Roche 454's FLX+ is most applicable for projects requiring accurate alignment of sequence data to a reference including targeted re-sequencing and metagenomic analysis. For applications where throughput is not as important, long read lengths facilitate more efficient and simplistic data alignment. Applications involving a large number of samples may benefit from the 132 multiplex identifiers made available by the FLX+ platform. Lastly, for projects where time is more important than throughput, such as in amplicon sequencing or the sequencing of bacterial or viral genomes, the FLX+ platform offers results in the shortest amount of time.

ABI's SOLiD™ 5500xl is most applicable to projects involving the detection of minor variations in sequence data as well as for projects that require a quick turn-around time. The platform's superior accuracy supports the detection of minor variants and individually configurable lanes on the platform's flow chip facilitate partial sequencing runs without the waste of reagents.

Lastly, while all three platforms offer an instrument for automation of sample preparation, Illumina's cBOT and Roche 454's REM e System reduce the ~ five hours that would be required for manual preparation to ~ 10 minutes and ~15 minutes, respectively while ABI's SOLiD™ EZ Bead™ System only reduces library preparation time to just under an hour. The cBOT and EZ Bead™ System are comparable in price at $55,000 and $60,000, respectively. The REM e System appears much lower in price at $18,900 but requires the additional purchase of a liquid handling platform.

The distinguishing features highlighted by each platform collectively support a wide range of life science applications. Knowing which platform to use for each application will facilitate achievement of the desired results and offer overall project success.

References

1. Illumina Press Release. "Illumina Announces HiSeq™ 2000 Sequencing System" 12 Jan 2010. http://investor.illumina.com/phoenix.zhtml?c=121127&p=irol-newsArticle&ID=1374339

2. Illumina (2010). Specification Sheet: Illumina® Sequencing. HiSeq™ 2000. http://www.illumina.com/documents/products/datasheets/datasheet_hiseq2000.pdf

3. Illumina (2011). Multiplexing Sample Preparation Oligonucleotide Kit. http://www.illumina.com/products/multiplexing_sample_preparation_oligonucleotide_kit.ilmn

4. NuGen (2010). Encore™ 384 Multiplex System. http://www.nugeninc.com/nugen/index.cfm/products/next-gen-sequencing/encore-384-multiplex-system/

5. Illumina (2011). Data Sheet: Systems and Software. IlluminaCompute. http://www.illumina.com/Documents/products/datasheets/datasheet_illumin acompute.pdf

6. Illumina (2011). cBOT Cluster Generation System http://www.illumina.com/systems/cbot.ilmn

7. 454 Sequencing. News. "Roche Launches GS FLX+ Sys.tem Offering High-Quality, Sanger-like Reads with the Power of Next-Generation Throughput". 28 June 2011. http://my454.com/resources-support/news.asp?display=detail&id=163

8. 454 Sequencing (2011). The Technology. http://my454.com/products/technology.asp

9. 454 Sequencing (2011). GS FLX+ System. http://my454.com/products/gs-flx-system/index.asp

10. 454 Sequencing (2011). Analysis Software. http://my454.com/products/analysis-software/index.asp

11. 454 Sequencing (2011). REM e System. http://my454.com/products/automation/index.asp

12. Life Technologies Press Release. "Life Technologies Launches New SOLiD Sequencer to Drive Advances in Cancer Biology and Genetic Disease Research." 1 Nov. 2010. http://www.lifetechnologies.com/news-gallery/press-releases/2010/life-techologies-lauches-ew-solid-sequecer-to-drive-advaces-i--c.html

13. Applied Biosystems by Life Technologies (2011). Overview of SOLiD™ Sequencing Chemistry. http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems/solid-sequencing-chemistry.html

14. Applied Biosystems by Life Technologies (2011). 5500 Series Genetic Analysis Systems. http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html

15. Applied Biosystems by Life Technologies (2011). Specification Sheet. 5500 Series Genetic Analysis Systems. http://media.invitrogen.com.edgesuite.net/solid/pdf/CO18235-5500-Series-Spec-Sheet-F2.pdf

16. Applied Biosystems by Life Technologies (2011). Spec Sheet. LifeScope™ Genomic Analysis Solutions. http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_095821.pdf

17. Applied Biosystems by Life Technologies (2011). SOLiD™ EZ Bead™ System. https://products.appliedbiosystems.com/ab/en/US/partnerMkt/ab?cmd=catNavigate2&catID=607361&tab=DetailInfo

A Comprehensive Guide to Personal Genome Machines

Introduction

Researchers today have choices when it comes to their genome sequencing needs.
Recent advancements in the field of next-generation sequencing have resulted in the advent of personal genome machines (PGMs), smaller-scale, bench-top genome sequencers marketed by Illumina (MiSeq), Roche 454 (GS Junior), and Life Technologies (PGM™ Sequencer). This recent emergence promises to bring DNA sequencing directly into individual laboratories.

Given the variety of available sequencing instruments, it's important that researchers are informed of key differences among platforms such that they can select a technology most appropriate for their project goals. This resource illustrates key features of the MiSeq, GS Junior, and PGM™ Sequencer, the three most prominent bench-top platforms. An overview of the basic characteristics and performance parameters of each platform can be viewed in Table 4. The accompanying sections further explore the offerings of each platform and highlight those features that differentiate the various platforms from one another.

Table 4: Key Differences Among the Leading Personal Genome Machines

Company	Illumina	Roche 454	Life Technologies
Platform Name	MiSeq	GS Junior	Ion Torrent PGM™ Sequencer
Chemistry	Reversible Terminator - SBS	Pyrosequencing - SBS	Semiconductor sequencing
Prep Surface	Flow cell	Beads in wells of a PicoTiterPlate	Ion Sphere™ particles in wells of a semiconductor chip
Amplification	Bridge Amplification	Emulsion PCR	Emulsion PCR
Cost	$125,000 $400-$750 / run	~$110,000 $1000 / run	< $50,000 < $500 / run
Instrument dimensions	27.0" × 22.2" × 20.6"	15.8" x 23.6" x 15.8"	24" x 20" x 21"
Throughout	> 1 Gb (2 x 150 bp)	35 Mb / run	314 > 10 Mb 316 > 100 Mb 318 > 1 Gb
Run Time	4 – 27 hrs (1 x 35 bp – 2 x 150 bp)	10 hrs (sequencing) 2 hrs (data processing)	2 hrs
Read Length	150 bp	400 bp	200 bp (2011) 400 bp (2012)
Read Quality	> 75% > Q30 (2 x 150 bp)	Q20 for a 400 bp rea	> Q20 at end of ~ 100 bp read
Associated Software	Cluster generation, sequencing and data analysis included in MiSeq instrument	REM e System – automated emPCR ($18,900) GS Data Analysis Software Package + accompanying PC (Included)	The Ion OneTouch™ System – automated emPCR (< $5,000) Torrent Server and Torrent Suite Software (Included)

Illumina MiSeq

Introduction

Illumina introduced the MiSeq on January 11, 2011 emphasizing the next-generation sequencer's capacity for integration of amplification, sequencing and data analysis in a single instrument. Integration of cluster generation within the 2 sq. ft. instrument eliminates the need for auxiliary hardware, saving valuable bench space. Initial orders for the MiSeq were expected starting in April 2011 with shipment of the first commercial units expected in summer 2011[1].

Chemistry

The MiSeq platform leverages Illumina's proven and widely adopted reversible terminator-based sequencing-by-synthesis (SBS) chemistry. This chemistry, coined TruSeq, supports the detection of single bases as they are incorporated into growing DNA strands. Each dNTP is bound to a unique fluorescently-labeled terminator. The respective terminator is imaged as each dNTP is incorporated, followed by cleavage to allow for the incorporation of the next base. All four reversible-terminator-bound dNTPs are present during each sequencing cycle facilitating natural competition and minimizing incorporation bias. Individual bases are called based upon fluorescent signal intensity measurements made during each cycle[2].

Illumina's previous use of their SBS chemistry in a variety of earlier platforms marks an advantage of the MiSeq. Justin Johnson of EdgeBio comments on the benefits of using new technology as opposed to one that has been previously proven[3]. He concludes that MiSeq's use of the same chemistry as the HiSeq could be appealing to the research community for a variety of reasons. Most appealing is the fact that the sequencing technology has already been proven. Researchers familiar with the HiSeq, in essence, already know what they are purchasing and can utilize the same protocols and reagents as those used for the HiSeq[3]. While facilitating a seamless transition, use of the same chemistry also has the potential to limit the MiSeq platform, preventing the instrument from evolving to anything more than simply a "mini-HiSeq."

Cost, Throughput, and Read Quality

Expected to be priced under $125,000 with individual run prices ranging from $400-$750/run, the MiSeq requires the largest initial capital investment, but is competitive in terms of subsequent costs/sequencing run[1]. > 1 Gigabase (Gb) of sequence data can be expected from a single 2 x 150 base pair (bp) run[4] and a single run on the MiSeq has yielded as much as 1.7 Gb of data[3].

Dependent on the number of cycles, a sequencing run can take as few as 4 hours (1 x 35 bp) and as many as 27 hours (2 x 150 bp)[5]. Max read lengths settle around 150 bp, shorter than those generated by both the GS Junior and PGM™ Sequencer.

Illumina's MiSeq takes the lead considering read quality. In a recent presentation, Illumina claimed an average Q-score of 31 (Q31) for its internally generated data based upon a read length of ~ 100 bp[3]. On average, ~75% of reads in a 2 x 150 bp run have quality scores higher than Q30[5].

Multiplexing is possible on the MiSeq platform facilitating multiplexed PCR amplicon sequencing and other small-scale projects[5]. Like all other Illumina sequencing platforms, MiSeq supports paired-end runs, critical for a broad range of applications including amplicon sequencing, sequencing of complex genome regions, and efficient mapping[1]. All data is generated in the equivalent of one lane of an Illumina flow cell[4]. The decreased size of the flow cell size results in increased fluidics and therefore increased efficiency.

Software and Accessories

The compact, all-in-one MiSeq platform incorporates cluster generation, paired-end fluidics, and complete data analysis, eliminating the need for auxiliary hardware and saving valuable laboratory bench space[5]. While data analysis software for both the GS Junior and Ion Torrent are also included in the price of the instrument, the MiSeq platform is the only next-generation sequencer that integrates amplification, sequencing, and data analysis in a single instrument with a footprint of less than 2 sq ft[5]. Data analysis includes on instrument base calling, alignment, and variant calling.

The MiSeq platform's incorporation of cluster generation technology eliminates the need for the cBot, an automated system for the generation of clonal clusters required for the preparation of samples to be sequenced on most other TruSeq Illumina platforms.

Distinguishable Benefits

Overall, advantages of the MiSeq as compared to the other PGM platforms include:

- Greatest throughput per run
- Superior read quality
- Compact, all-in-one platform incorporating cluster generation, sequencing, and data analysis
- Competitive cost / sequencing run

Best Suited Applications

Based upon the above advantages, the MiSeq platform is best equipped for small-scale sequencing projects requiring increased throughput as well as superior read quality. The Miseq facilitates performance of standard experiments such as amplicon sequencing, clone checking and small genome sequencing offered as an alternative to capillary electrophoresis (CE) sequencing. In addition, the platform facilitates the execution of powerful next-generation sequencing applications including multiplexed PCR amplicon sequencing, targeted re-sequencing, ChIP-Seq and small RNA sequencing[5].

The MiSeq platform is a good choice for researchers with limited laboratory space. The integration of cluster generation, sequencing and data analysis in an easy-to-use 2 sq. ft instrument brings a powerful system to individual research labs while simultaneously preserving valuable bench space.

Roche 454 GS Junior

Introduction

In 2010, 454 Sequencing launched the GS Junior, a bench-top variety of their previous sequencing platforms utilizing Titanium sequencing technology[6].

Chemistry

The GS Junior utilizes a variety of SBS chemistry, coined pyrosequencing, which relies on the detection of pyrophosphate release upon nucleotide incorporation. Solutions of A, C, G, and T dNTPs are sequentially added and removed from the reaction[7]. Incorporation of the correct, complementary dNTP by DNA polymerase results in the stoichiometric release of pyrophosphate (PPi) followed by conversion of PPi to ATP. The ATP acts as fuel to the luciferase-mediated conversion of luciferin to oxyluciferin, generating visible light in an amount proportional to the quantity of ATP[7]. Quantification of the intensity of visible light resulting upon addition of each nucleotide solution allows for the appropriate base calls to be made.

Cost, Throughput, and Read Quality

At a cost of ~$110,000, the GS Junior is less expensive that the MiSeq yet more expensive that the PGM™ Sequencer[6]. Cost/sequencing run is expected to be much higher than the other PGM platforms at ~$1000/ run. The platform is currently capable of outputting 35 Megabases (Mb) of data/ run and produces read lengths of 400 bp with a quality score of Q20[8]. Read lengths are much longer than those offered by the MiSeq or PGM™ Sequencer, comprising a distinguishing factor of the GS Junior. Sequencing takes 10 hours and data processing/analysis an additional 2 hours[8].

While the GS Junior has 132 multiplex identifiers at its disposal, manufactured for use with the technology, the GS Junior Pico Titer Plate is comprised of a single gasket only supporting sequencing of a single sample type[8]. Similar to the MiSeq, reagent kits manufactured by 454 support the preparation of samples for paired-end runs.

Software and Accessories

The GS Data Analysis Software Package is provided at no additional cost with the purchase of a GS Junior sequencing instrument. The software package includes tools to investigate complex genomic variation in samples including de novo assembly (GS de novo assembler), reference guided alignment and variant calling (GS read mapper), and low abundance variant identification and quantification (GS amplicon variant analyzer)[8]. In addition to the complete suite of point-and-click data analysis software, the GS Junior System includes a high-performance desktop computing station. In contrast to the Miseq platform, the included desktop computing station requires additional space in a laboratory setting. Disregarding associated data analysis infrastructures, the MiSeq, GS Junior and PGM™ Sequencer are all fairly similar in size.

Additionally, Roche 454 offers the REM e System as a platform accessory[9]. This liquid handler is designed to fully automate the emulsion PCR enrichment and sequence primer annealing steps in the 454 Sequencing workflow. Use of the automated system can reduce up to five hours of hands-on work to 15 minutes of liquid handler setup, improves consistency by enhancing PCR enrichment accuracy and supports a wide variety of library types and all GS FLX and GS Junior Titanium Series emulsion formats. In order to enjoy automation offered by the REM e System, the system must be purchased for $18,900 and requires an additional investment in a liquid handler platform[9].

Distinguishable Benefits

Overall, advantages of the GS Junior as compared to the other available PGMs include:

- Longest read length averaging 400 bp
- Attendant PC for run processing and data analysis included

Best Suited Applications

Given the key attributes of Roche 454's technology, the GS Junior is best suited for applications that do not require large-scale throughput and that benefit from accurate and simplistic alignment of sequence data. For applications such as amplicon sequencing, sequence capture, whole genome sequencing of microbial genomes, metagenomics and transcriptome sequencing that do not require tremendous volumes of throughput, the GS Junior creates an opportunity for longer read lengths.

In addition, the accompanying PC may lead researchers to choose the GS Junior over other sequencing platforms. The included PC is pre-installed with GUI-based GS Data Analysis Software facilitating *de novo* assembly, reference mapping, and amplicon variant analysis[8]. For researchers with lab space large enough to support the GS Junior instrument as well as accompanying PC, separation of library prep and sequencing from data analysis may be more appropriate for the physical layout of a laboratory space.

Life Technologies Personal Genome Machine (PGMTM) Sequencer

Introduction

Life Technologies announced the launch of the PGM™ Sequencer on December 14, 2010[10]. Ion Torrent, a busines unit of Life Technologies is credited with the invention of the semiconductor device on which the PGM is based[10].

Chemistry

Ion Torrent pairs semiconductor technology with simple sequenicng chemistry to faclitate base calling via their PGM™ Sequencer[11]. Natually, incorporation of a nucleotide into a growing DNA strand by DNA polymerase results in the release of a hydrgon ion. Ion Torrent uses a high-density array of micro-machined wells to perform this biochemical process in a massively parallel way. Each well holds a different DNA template. Beneath the wells reside an ion-sensitive layer and an ion sensor. The charge from an ion released upon incorporation of a nucleotide changes the pH of the solution; this pH change is detected by the ion sensor. The sequencer calls the base by converting the chemical information collected into digital information. The PGM™ sequencer sequentially floods the chip with one nucleotide after another, measuring the change in voltage upon nucleotide incorporation. Because the technology permits direct detection (no scanning, no cameras, no light), each nucleotide incorporation is recorded in seconds and the time required for sequencing is diminished tremendously[11].

Cost, Throughput, and Read Quality

At $50,000, the PGM™ Sequencer is the cheapest personal genome machine. Costs per run of less than $500 compete with the Illumina MiSeq for cheapest cost per run and is much less than the cost per run for Roche 454's GS Junior. The PGM™ Sequencer generates varying volumes of throughput depending on the semiconductor chip used. While the platform is currently only capable of outputting slightly more than 10 Mb on its 314 chip[12], exponential increases in throughput are expected with the advancement of the platform's associated silicon chip. Upon availability of the 316 and 318 chips in 2011, >100 Mb/run and >1 Gb/run are expected, respectively[12]. The 316 chip is already generating ~7x more data than the 314 and was due out in July 2011[3]. The Ion 314, 316 and 318 chips are on track to demonstrate a 100x scalability path in 2011 moving from 10 Mb of output to 1 Gb. The throughput predicted for the 318 chip will bring the PGM™ Sequencer into direct competition with Illumina's MiSeq for highest throughput. While throughput of the PGM™ Sequencer is expected to continue to increase with the advancement of the silicon chip, based upon the MiSeq and GS Juniors' exploitation of already proven technologies, these platforms may have reached their peak throughput at their current capacities.

Although seemingly a disadvantage, the lower throughout offered by a single run on the PGM™ Sequencer is key to the flexibility offered in terms of experimental design[3]. In addition to its ability to sequence multiple libraries simultaneously (via multiplexing), the PGM™ Sequencer supports the simultaneous run of different types of input for different experiments. An attempt at the same using Illumina's Miseq or 454's GS Junior would require the use of barcodes, typically causing bias and affecting the quality of results[3].

Setting it apart from the rest, the PGM™ Sequencer has a run time of a mere 2 hours[12]. Read lengths are expected to reach 200 bp sometime in 2011 and increase to 400 bp in 2012. Predicted advancements in throughput and read length will render the PGM™ Sequencer competitive with the MiSeq (throughput) and the GS Junior (read length). The PGM™ Sequencer supports multiplexing of samples as well as paired-end sequencing.

Considering a recently released data set, the 314 chip yielded a quality score of Q17 at base 100[3]. Although not seemingly significant in comparison to those values generated by the MiSeq, this marks an improvement over the chip's performance in January when the quality score at base 100 hovered around Q10[3]. Read quality is expected to see continued improved with the advancement of the technology.

Software and Accessories

Included with the purchase of a PGM™ Sequencer are the Torrent Server and Torrent Suite Software. Upon completion of a run on the PGM™ Sequencer, data is automatically transferred to the Torrent Server configured to run the Torrent Suite Software. Raw ion signals are converted to base calls and stored in industry-standard SFF or FASTQ files[13]. These data can then be processed by a variety of commercially available software packages for applications such as variant detection, RNA-Seq, ChIP-Seq or genome assembly[13]. A potential disadvantage of the PGM™ Sequencer in comparison to the MiSeq and GS Junior is that fact that the included software does not perform the data analysis itself. In contrast to its competitors, the Torrent Suite Software simply converts the raw signals coming off the sequencing machine into a usable format (SFF or FASTQ) such that other software packages can be used to analyze the data[13].

Additionally, Life Technologies offers the Ion OneTouch™ System, an automated sample preparation system[14]. The instrument reduces hands-on time to five minutes combining sample loading, clonal amplification and sample recovery into a single, automated process. The system is scalable, supporting the Ion 314, 316 and 318 chip and was introduced in April of 2011 for an introductory price of under $5,000[14].

Distinguishable Benefits

Overall, advantages of the PGM™ Sequencer compared to the other available PGMs include:

3. Uniform coverage
4. Chemistry close to native molecular processes
5. Use of semiconductor sequencing technology
6. Choice of throughput within a single platform
7. Flexibility in experimental design
8. Least expensive instrument cost at $50,000

Best Suited Applications

Based upon the features listed above, Life Technologies PGM™ Sequencer is best suited for projects requiring the examination of hard to access portions of genomes. The simplicity of both synthesis and detection translate into exceptionally uniform coverage providing access to regions of the genome that other technologies would have difficulty sequencing[15]. Uniformity of coverage also reduces the amount of sequence necessary to have confidence in the data.

Semiconductor sequencing leverages the simple and natural biochemistry of DNA synthesis. The technology keeps the process as close to that which occurs naturally resulting in unprecedented quality of sequence data[15].

Other reasons to choose the PGM™ Sequencer include its flexibility in experimental design offered as a result of its lower throughput and the opportunity to choose the throughput generated for any given application through choice of a semiconductor sequencing chip (314, 316 or 318).

Additionally, the PGM™ Sequencer is the least expensive personal genome machine and leverages semiconductor sequencing technology, on track to reveal a 100x improvement in base yield in a single year[15]. This trend lays a path for semiconductor sequencing to deliver scalable, simple and rapid DNA sequencing to the research and clinical communties with minimal effort required to upgrade to the latest technology[15].

Conclusion

Although all three PGMs discussed above are capable of executing a number of similar applications, each platform comprises a unique set of distinguishing features that should be taken into account when choosing between the three.

Common applications of all three PGMs include amplicon sequencing, small genome sequencing, sequencing of barcoded libraries as well as sequencing of paired-end reads.

The high throughput offered by the Illumina MiSeq makes this technology a good candidate for whole transcriptome studies and other RNA-seq applications. Although currently the leader in terms of throughput at 1 Gb per sequencing run, the PGM™ Sequencer looks to challenge this claim with the coming introduction of the 318 chip, also promising throughput as high as 1 Gb. These high-throughput platforms are also ideal for highly multiplexed PCR amplicon sequencing, ChIP-Seq, and small RNA sequencing[5]. Applications requiring a small number of reads, such as the sequencing of small RNAs, low complexity transcriptomes (e.g. viruses, bacteria) or targeted gene expression should be left to the PGM™ Sequencer's 314 / 316 chips or performed on Roche 454's GS Junior.

If cost is a key factor in a researcher's platform decision, the PGM™ Sequencer is the best choice for a couple of reasons. Not only is the initial cost of the instrument the lowest of the three at $50,000 but, this same platform competes with the MiSeq for lowest cost of reagents. Additionally, the semiconductor technology on which the PGM™ Sequencer is built provides upgrades to the instrument in the form of a simple silicon chip. In comparison to the efforts that would be required to upgrade to a new Illumina or 454 sequencing instrument, the PGM™ Sequencer offers the most cost effective means of continually acquiring the latest technology.

Another factor to consider when deciding between platforms is the uneven playing field on which the PGM™ Sequencer and remaining two personal genome sequencers currently reside in terms of sequencing chemistry. While Illumina's MiSeq and Roche 454's GS Junior rely on an identical chemistry as that utilized for these companies' larger-scale HTS platforms, the Ion Torrent's semiconductor sequencing technology is new to the field. Investors in new sequencing instruments must make a decision between proven and new technology is choosing which platform is most appropriate for their research needs. While the MiSeq and GS Junior leverage proven technology and allow researchers familiar with the reagents and protocols utilized for these platform's parent instruments the ability to simply re-use that which they already know, the PGM™ Sequencer does not offer this advantage in terms of familiarity. What the new semiconductor sequencing technology does provide is a means of "rapid validation" of other sequencing studies based upon its reliance on an entirely different chemistry[3].

As discussed in the corresponding sections above, the PGM™ Sequencer has the potential to continue to improve in both read quality and throughput with the advancement of the associated technology. Given that the platform is still in its infancy, researchers investing in this platform take a chance on the anticipated increase in both of these arenas.

Ultimately, as concluded by Justin Johnson of EdgeBio, it's hard to make a solid comparison between technologies while one platform is still in its infancy. While the MiSeq is itself a new platform, the supporting chemistry is not new. In stark contrast, the Ion Torrent and its associated semiconductor technology have only been available for a few months. As all previous platforms experienced in their infancy, the Ion Torrent is expected to engage in a period of "working out the kinks[3]." The benefits of each platform in terms of associated chemistry will hold more weight once every technologies' platform has been established and chemistry has been proven.

Additionally, the availability of associated software may sway individuals in their decisions to purchase one platform over another. While all three platforms advertise the inclusion of data analysis software in the

platform package, each has something a bit different to offer. Unique to the MiSeq is the incorporation of cluster generation, sequencing and data analysis in a single instrument. Data analysis for this platform includes base calling, alignment and variant analysis.

Both the GS Junior and the PGM™ Sequencer offer analysis software but in the form of external equipment. Included with the purchase of a GS Junior is the GS Data Analysis Software Package as well as a high-performance desktop computing station. The software package includes tools to investigate complex genomic variation in samples including de novo assembly (GS de novo assembler), reference guided alignment and variant calling (GS read mapper), and low abundance variant identification and quantification (GS amplicon variant analyzer)[8]. While purchase of a PGM™ Sequencer includes both the Torrent Server and Torrent Suite Software for no additional cost, the included software does not perform the data analysis itself. In contrast to the other two platforms, the Torrent Suite Software simply converts the raw signals coming off the sequencing machine into a usable format (SFF or FASTQ) such that other software packages can be used to analyze the data. Based upon the above information, researchers interested in the most extensive data analysis software should invest in the GS Junior while those concerned about the excess space required of a server or desktop may be inclined to purchase the MiSeq.

Lastly, options for automation of library prep may influence a researcher's platform decision. The MiSeq holds a tremendous advantage in this regard due to the inclusion of cluster generation within the instrument itself. For no additional cost and requiring no additional space the MiSeq reduces the hands-on time required for cluster generation to a mere 10 minutes of instrument setup. Although not a part of the instrument nor included in the cost of the instrument, the REM e System for the GS Junior offers automation of library prep for an additional investment in capital equipment ($18,900 + cost of liquid handler platform). Similarly, Life Technologies offers an automated sample preparation instrument, the Ion OneTouch™ System, for an additional investment of ~$5,000. Researchers interested in dramatically decreasing sample preparation time can become equipped to do so with the MiSeq, GS Junior and PGM™ Sequencer. Even with the additional investment in the Ion OneTouch™ System, the total cost of Ion Torrent's platform is the cheapest option. For those concerned about laboratory space, Illumina's integrated platform only requires purchase of a single instrument yet offers the benefits of automated library preparation.

The distinguishing features highlighted by each platform collectively support a wide range of genomic applications. Knowledge of which platform to use for any given application supports achievement of project goals and offers overall project success.

References

1) Illumina Investor Relations. “Illumina Announces MiSeq(TM) Personal Sequencing System.” 11 Jan. 2011. http://investor.illumina.com/phoenix.zhtml?c=121127&p=irol-newsArticle&ID=1515239&highlight=

2) Illumina (2011). Sequencing Technology. http://www.illumina.com/technology/sequencing_technology.ilmn

3) Genome Web. “Ion Torrent vs. MiSeq: E. coli Sequencing Kicks off Desktop Sequencer Comparison.” 5 July 2011. http://www.genomeweb.com//node/973043?hq_e=el&hq_m=1045046&hq_l=1&hq_v=b7e135008b

4) Illumina (2011). Sequencing Portfolio. http://www.illumina.com/systems/sequencing.ilmn

5) Illumina (2011). MiSeq Personal Sequencing System. http://www.illumina.com/systems/miseq.ilmn

6) Genomeweb. In Sequence. Roche to Launch Scaled-Down 454 Sequencer in 2010; JGI First Test Site for Long GS FLX Reads. 1 Dec. 2009. http://www.genomeweb.com/sequencing/roche-launch-scaled-down-454-sequencer-2010-jgi-first-test-site-long-gs-flx-read

7) Qiagen (2011). Principle of Pyrosequencing Technology. http://www.pyrosequencing.com/DynPage.aspx?id=7454

8) 454 Sequencing (2011). GS Junior. http://454.com/products/gs-junior-system/index.asp

9) 454 Sequencing (2011). REM e System. http://454.com/products/automation/index.asp

10) Life Technologies Press Release. Life Technologies Launches Ion PGM Sequencer. 14 Dec. 2010. http://www.lifetechnologies.com/news-gallery/press-releases/2010/life-techologies-lauches-io-pgm-sequecer.html

11) Life Technologies (2011). Ion Torrent. Technology: How Does it Work? http://www.iontorrent.com/technology-how-does-it-work-more/

12) Life Technologies (2011). Ion Torrent. Technology: How Does It Perform? http://www.iontorrent.com/technology-how-does-it-perform?/

13) Life Technologies (2011). Ion Torrent Specification Sheet. http://www.iontorrent.com/lib/images/PDFs/ion_prod_a.pdf

14) Life Technologies Press Release. Life Technologies Announces an Automated Sample Preparation System for its Ion PGM™ Sequencer. 13 April 2011. http://www.lifetechnologies.com/news-gallery/press-releases/2011/life-techologies-aouces-a-automated-sample-preparatio-system-for.html

15) Life Technologies (2011). Ion Torrent. Application Note - Performance Spring 2011. http://www.iontorrent.com/lib/images/PDFs/performance_overview_application_note_041211.pdf

High-Throughput Sequencing Platforms and Personal Genome Machines: Friend or Foe?
Part 1: Five Reasons Why These Technologies Will Successfully Coexist (for now)

Recent advancements in the field of next-generation sequencing have resulted in the advent of personal genome machines (PGMs), smaller-scale, bench-top genome sequencers marketed by Illumina (MiSeq), Life Technologies (Ion Torrent), and Roche 454 (GS Junior). This recent emergence promises to bring DNA sequencing directly into individual laboratories, inevitably affecting the current high-throughput sequencing (HTS) market in the process.

Below are five reasons why HTS platforms and PGMs will successfully coexist, at least in the short term:

1. PGMs are currently only useful for unique and limited applications.

Similar to the old argument that the introduction of HTS platforms would soon render microarray technology obsolete, some fear that the introduction of the PGM will eliminate the need for larger-scale HTS instruments. While the rapid advancement of PGM technology does bring with it this threat, the unique purposes for which high-throughput sequencers and PGMs are intended make the introduction of PGMs no immediate threat to the HTS market. While HTS platforms are commonly utilized for the purpose of whole genome, exome, transcriptome, and ChIP analyses, PGMs find their niche in amplicon sequencing, clone checking and small (bacterial/viral) genome sequencing.

2. PGMs make possible the segregation of projects based upon throughput needs.

HTS platforms and PGMs are each intended for unique purposes. While the capacity of a HTS platform is required to generate the throughput necessary for the analysis of a large genome, much lower throughout is required for the sequencing of amplicons or constructs. Prior to the recent advent of PGMs, HTS platforms were utilized for a wide variety of DNA sequencing projects, independent of the throughput required. Often, the throughput generated exceeded the project's needs.

PGMs will make possible the segregation of projects based upon throughput needs. HTS platforms will no longer be bogged down with smaller projects for which a PGM is sufficient. Given the increase in cost and time associated with the use of HTS platforms, the existence of PGMs will allow researchers to save these resources for projects that require the greater throughput offered by these platforms. For all other small-scale projects, a PGM will generate far less costly results in a shorter period of time. The segregation of projects based upon throughput needs will drastically increase the efficiency with which researchers are able to achieve their genomics goals.

3. HTS Platforms and PGMs are created by the same companies.

Three of the most popular HTS platforms include Illumina's HiSeq 2000, Roche 454's FLX+ and Applied Biosystems' (Life Technologies) SOLiD 5500xl. Each of these three companies also currently markets a PGM. Illumina's MiSeq and Roche 454's GS Junior act through use of the same chemistry as these companies' larger-scale HTS platforms yet are physically smaller in size and offer lower throughout. In contrast, Life Technologies' Ion Torrent offers smaller instrument size and lower throughput through use of semiconductor sequencing technology, new to the industry. Overall, the fact that companies are simultaneously marketing HTS platforms and PGMs speaks of these companies' hopes for their coexistence. In the absence of a disruptive entrant these companies will certainly work to keep their PGM and HTS lines both healthy and differentiated such that both types of sequencing technology will be useful and have a significant place in research.

4. The capabilities of these two technologies will complement each other.

While HTS platforms and PGMs were created and are used for unique purposes, the output of one can be used to enhance that of the other.

It may be assumed that labs utilizing a HTS platform for a project have no need for the capabilities of a PGM and vice versa. In reality it's expected that research projects will commonly benefit from the use of both. A HTS platform might be used for the purpose of sequencing a whole genome. A PGM might then be utilized for the targeted re-sequencing of an area of interest. In this way not only will the genomics community find the coexistence of these two technologies useful, but individual labs may also find reason to utilize both for a single project. PGMs will not eliminate the need for HTS platforms, but will be used simultaneously as a means of gathering further support for a scientific claim.

5. PGMs will increase the appetite for HTS.

Further, far from eliminating focus on the previously existing HTS platforms, introduction of the PGM will lead to continued education of the scientific community on the applications of HTS and render this genomic technology within reach of individual research labs. The significantly lower cost of PGMs in comparison to HTS platforms will allow labs on a tighter budget to experiment with genomic technologies, significantly expanding the market. PGMs will expose individual labs to high-throughput sequencing technology, increasing the chances that these same labs will invest in a HTS platform down the road.

Conclusion

In contrast to the expected long-term effects, the recent advent of the PGM is predicted to complement the already thriving HTS market. For five main reasons listed above, HTS platforms and PGMs are expected to thrive simultaneously in the short term. Created by the same companies, PGMs will increase the market's appetite for HTS, increase the efficiency with which genome-scale projects can be completed, and allow for the segregation of projects based upon throughput needs. Coming soon, I'll explore five reasons why PGMs are predicted to eventually largely replace HTS platforms.

High-Throughput Sequencing Platforms and Personal Genome Machines: Friend or Foe?
Part 2: Five Reasons Why PGMs will Replace HTS Platforms In the Long Term

Recent advancements in the field of next-generation sequencing have resulted in the advent of personal genome machines (PGMs), smaller-scale, bench-top genome sequencers marketed by Illumina (MiSeq), Life Technologies (Ion Torrent), and Roche 454 (GS Junior). This recent emergence promises to bring DNA sequencing directly into individual laboratories, inevitably affecting the current high-throughput sequencing (HTS) market in the process.

As discussed in Part I, PGMs are expected to coexist with HTS platforms in the short term due to their ability to complement one another and the unique applications intended for each technology. In direct contrast, I predict PGMs will largely disrupt the HTS market in the long term. Below are five reasons why:

1. PGMs have immense intangible benefits.

PGMs are specialized platforms intended to direct sequencing back into individual research labs. They are particularly useful for applications that don't require the scale of throughput generated by previously established HTS platforms.

Although currently used for smaller-scale genome projects, scientists will soon discover the benefits of shorter turn-around times, decreased cost, decreased unit size, and total control of project timeline offered by individual ownership of a PGM. These benefits will ultimately lead to market dominance by the PGM.

2. PGMs will lower the barrier to entry to the HTS market.

As discussed in part 1, introduction of the PGM will further educate the scientific community on the applications of HTS and render this genomic technology within reach of individual research labs. The significantly lower cost of PGMs in comparison to HTS platforms will allow labs on a tighter budget to experiment with genomic technologies, significantly expanding the market. In addition the accessibility of PGMs will encourage labs that have never considered sequencing as a means of addressing their scientific questions, the opportunity to do so. Ultimately, this exposure to next-generation sequencing will render DNA sequencing a more standard practice for a wide variety of research labs and PGMs will gain prevalence as standard laboratory equipment.

3. PGM technology is expected to advance exponentially with time.

Evidence for the long-term success of the PGM can be seen through the example of Life Technologies' Ion Torrent. In their development of the Ion Torrent, Life Technologies has utilized the entire semiconductor supply chain infrastructure, a collective $1 trillion dollar investment made over the course of the last 40 years. Because of the direct use of semiconductor-based technology, the Ion Torrent is highly likely to, at a minimum, follow the trajectory of Moore's Law. We can already see this playing out as an upgrade to the original Ion Torrent, released during the first half of 2010, acts ten times as fast as the original technology. By 2012, the PGM is expected to decode, in a mere two hours, all 20,000 human protein-coding genes. Ion Torrent Founder, Dr. Jonathan M. Rothberg, has boldly declared that "there isn't a technology that the Ion Torrent will not pass in a very short period of time, no matter how far ahead they are."

Furthermore, Rob Carlson's DNA synthesis and sequencing cost curves support the exponential rate of advancement in the DNA sequencing industry. Carlson's curves show the dramatic decrease in the cost per base of DNA sequencing with time, a direct result of the continued technological advancement within the industry. Given the rapid nature of this progression, I expect PGMs to render HTS platforms obsolete in the long term.

4. PGM capabilities will expand to encompass the vast majority of scientists' needs.

While the technical offerings of HTS platforms and PGMs are currently unique, justifying their coexistence, this will not be the case forever. Soon enough, the technical capabilities of the PGM, including throughput, will match those of the HTS platforms of today, eliminating the need for these much larger units for a growing number of applications. PGMs are the way of the future, and will gradually take control of the genomics market as their capabilities grow to encompass more of the common tasks required by life scientists.

5. New technologies will reduce the large capital investment required to stay at the forefront of technological advancement.

PGM and HTS manufacturers make money by employing the razorblade model; they make most of their money on the sale of consumables as opposed to the sale of the machine hardware itself. Therefore, it makes sense for each manufacturer to "lock-in" customers to their particular platform by selling sequencing machines at a lower cost than their competitors.

New PGM technologies offering simple upgradable machines will increase customer "lock in" by decreasing the chance that current customers invest in an entirely new platform to acquire the latest technology. Life Technologies' Ion Torrent serves as a prime example. The company advertises, "The chip is the machine" emphasizing their new instrument's unique use of semiconductor technology, allowing for upgrades through the simple swap of a silicon chip. Technological advancements such as that exemplified by the Ion Torrent promise to render PGMs lesser upkeep and more accessible to the most recent technology as compared to larger-scale HTS platforms, often requiring replacement and re-calibration of an entire instrument in an effort to upgrade. Easily upgradable technology such as that employed by Life Technologies is sure to be the way of the future.

Conclusion

Just like the personal computer revolution largely moved everyday computing from the mainframes to the desktop (and now to our pockets), DNA sequencing will follow a similar path.

I think we all can agree that the development of computing was the most significant technical advancement of the 20th century. PGMs will likely be a top contender for the most significant advancement of the 21st.

The Emerging Role of Biocomputing in the Life Sciences
Part 1: The Challenges

With the advancement of high-throughput DNA sequencing technology, costs associated with acquiring genome-scale datasets have decreased one hundred fold over the last three years, facilitating explosive growth in biological data generation. This increase in the volume of available sequence data has necessitated a push for the advent of computationally driven analytics. The decreasing costs and increasing throughput of DNA sequencing technologies have brought with them both challenges and opportunities for the life sciences community.

Four challenges arising from the advancement of high-throughput sequencing technologies are outlined below:

1) Data Generation Currently Outpaces Technical Capabilities to Manage the Data

Increased sequencing throughput means a tremendous increase in the sheer amount of data generated in a single sequencing run and introduces technical challenges in managing this data. For example, Illumina's Hiseq 2000 is capable of outputting as much as 600 Gb of data in a single run, introducing a number of technical problems in data management and downstream analysis. These problems may include acquiring the computational power and parallel algorithms to efficiently analyze the data, dedicating the necessary time to transfer these large quantities of data, and blocking out time to collaborate with other researchers regarding the data.

2) Lack of Tool Maturity

Few publically available computational tools are effective for the complete range of analytics necessary for high-throughput sequencing data. While there exist a variety of publically available tools and pipelines for the purpose of analyzing large-scale data, these freeware applications often require deep or complete customization as tools are commonly poorly documented, exceedingly slow, un-optimized, and require expensive computing hardware. As a result, transitioning from raw data to meaningful results is an arcane art. While the current status quo for researchers facing new bioinformatics challenges is to cobble together a custom one-off solution through use of publically available software, this approach often only narrowly addresses the issue at hand.

3) Acquisition of Talent with the Rare Combination of Skills Necessary for Successful Analysis

The development of effective analytics tools requires an individual with a unique set of talents. Identifying proper talent to enable the development of effective computational tools can be a challenge within itself. Not only is it necessary that the individual be gifted in the field of computer science (including knowledge of algorithms designed for optimization and massively parallel systems) facilitating the software and pipeline development aspect of such a task, but the individual must also have a strong background in the life sciences. The ideal candidate for the generation of computational tools for the analysis of large genome-scale datasets is not only capable of building a powerful tool, but also has a deep understanding of the biological question the tool is intended to address. Identifying individuals who excel in both regards can be a challenging feat.

4) Minimal Awareness of Capabilities

Scientists are often unaware of the multitude of potential DNA sequencing applications. Although the use of genomic techniques has increased in popularity as the costs of DNA sequencing have decreased with time, it's still common for scientists to be unaware of what questions these large-scale datasets may address and therefore incapable of mining the outputted data for valuable insights.

Conclusion

Recent advancements in the field of DNA sequencing have resulted in deceased costs of sequencing and increased throughout of sequence data. These rapid changes have necessitated the development of computational tools to facilitate the analysis of these large datasets. The changing environment of DNA sequencing has brought about both challenges and opportunities for the life sciences community. Four of the biggest challenges faced by the community were introduced above. A second blog, coming soon, will discuss the opportunities the changing DNA sequencing landscape has enabled.

The Emerging Role of Biocomputing in the Life Sciences
Part 2: The Opportunities

Despite the recent challenges that have arisen as a result of the decreasing costs and increasing throughput of DNA sequencing technologies (explored in part 1), these trends have simultaneously created a variety of opportunities for the life sciences community.

The three biggest opportunities these advancements have presented are outlined below:

1) The Rapidly Decreasing Cost of Sequencing is Making the Technology More Accessible

The decreasing cost of DNA sequencing has made the technology more readily available to research laboratories and scientific corporations. Genomic technologies are becoming increasingly utilized in a variety of research settings and DNA sequencing instruments are finding their way into individual research labs with the generation of the personal genome machine. As the cost of DNA sequencing continues to decrease, the number of research arenas for which DNA sequencing is standard protocol will continue to expand.

2) Leveraging the Next Generation of Discoveries Requires Investment in New Computational Resources

The increasing performance of DNA sequencing technologies is rendering computational biology a much more important and prevalent aspect of life sciences research and movement in this direction is becoming more common at leading research institutions. Just as individual researchers have realized the benefits of the decreasing costs of DNA sequencing, larger research institutions must support the changing needs and directions of the scientists comprising them and are therefore looking to acquire both the sequencing instrumentation and computational resources necessary to support this changing approach to research.

During his recent talk at the Donald Danforth Plant Science Center, President (and Intuitive Genomics co-founder), Dr. Jim Carrington highlighted computational technologies as key to the new directions of the Danforth Center. Dr. Carrington discussed his efforts to recruit to the Center leading life scientists proficient in computational and high-throughput technologies and his recent launch of a new biocomputing core facility. These changes will enhance the Center's capacity to leverage large-scale datasets and uncover the next generation of key discoveries from these massive amounts of data.

3) New Companies are Emerging to Help Tackle Recent Challenges

The recent advancement of high-throughput sequencing technologies has created an opportunity for the emergence of companies focused on the customization of computational tools for the analysis of large genome-scale datasets (Intuitive Genomics is one example). These companies provide a variety of tools, turnkey pipelines, and/or services to help researchers navigate the complexity of large genome-scale datasets and target the powerful discoveries that lie within.

Conclusion

The decreasing costs and increasing throughput associated with DNA-sequencing are changing the genomics landscape. Various challenges (part 1) as well as opportunities (outlined above) have resulted from these recent advancements. As evidenced by the new directions of the Danforth Center and the goals of emerging companies like Intuitive Genomics, computational tools and resources are becoming increasingly important in the life sciences as advancements in high-throughput sequencing technologies render increasingly massive bodies of data.

Internship Journal

WEEK 1

Monday June 20th, 2011
- Met on campus with Nathan, Doug and Todd
- Signed NDA and Compensation Contract
- Discussed IG business model
- Introduced to IG competitors, current partners, prospects
- Determined metrics for direct, partner and inbound marketing components of business model
- Introduced to GeneRocket
- Discussed current state of website
- Provided materials for continued education in marketing and bioinformatics industry
- Discussed internship timeline
- Began brainstorming components of cookbook for CBI
- Communicated with UGA regarding Green Pacific quote (HiSeq price confirmation)

Tuesday June 21st, 2011
- Prepared library prep and HiSeq portion of Green Pacific quote
- Obtained new e-mail account
- Wrote up short summaries of bioinformatics solutions
- Began reading *Inbound Marketing* and Next-Gen Sequencing Resource

Wednesday June 22nd, 2011
- Skype conversation with Nathan
- GP quote
- Generated personal bio
- Revised Bioinformatic Solutions text + added bullet points

Thursday June 23rd, 2011
- Researched various social media networks (Digg, Reddit, LinkedIn)
- Developed a LinkedIn profile
- Finalized bioinformatic solutions
- Finished reading Next-Gen DNA Sequencing article
- Began brainstorming ideas for images for bioinformatic solutions

Friday June 24th, 2011
- Skype conversation with Nathan
- Finished brainstorming bioinformatics solution image ideas (e-mailed TCM/updated google doc)
- Competitor evaluation for Ambry Genetics
- Read *Inbound Marketing*
- Began LinkedIn profile summary

WEEK 2

Monday June 27th, 2011

- Reviewed Ambry Genetics' competitor evaluation to prepare for tomorrow's meeting/ uploaded my notes to google doc
- Updated Ambry Genetics' info in google doc
- Researched next gen sequencing technologies
- Read *Inbound Marketing*
- Went to campus to check out the Mockler lab posters -> image ideas?

Tuesday June 28th, 2011

- Communicated with Myriam Belanger regarding sample prep pricing
- Generated draft GP quote using past quote as template
- Transcribed meeting notes (GP Quote, Ambry Genetics' Competitive Evaluation, bioinformatics solutions text and images)
- Transcribed NGS notes
- Meeting with Nathan to discuss and prepare weekly report
- Edited content on front page of website

Wednesday June 29th, 2011

- Researched presence of industry blogs/ added blogs to google reader
- Began writing PGM opinion blog (Top 5 Reasons HTS and PGMs will coexist in the short term)
- Team Meeting to discuss progress and next steps
- Submitted draft GP quote to Nathan based upon 10% increase in sequencing costs

Thursday June 30th, 2011

- Confirm accuracy of GP quote
- Skype meeting with Nathan to touch base on GP Quote, Blog posts and to develop July Cookbook
- Continue writing PGM opinion blog (review e-mail from Todd and "blog topics" Google doc)
- Research current opinions on PGMs
- Visit campus to check out images on Todd's research posters
- Begin seeking Google keywords for landing pages

Friday July 1st, 2011

- Called Myriam to confirm UGA pricing
- Discussion with Nathan to finalize GP quote
- Typed up image ideas for solution pages
- Finalized opinion blog – uploaded to Google docs
- Searched Google keywords for landing pages
- Read *Inbound Marketing*

WEEK 3

Tuesday July 5th, 2011

- 9 am catch up with Nathan (Blog opinion posts, upcoming projects)
- Finalized and uploaded "Solution Image Ideas" and "Google Keywords Results" to Google docs
- Revised 2 opinion blog posts and re-submitted to team via e-mail
- Filled out and sent Nathan my w-4
- Began reading pdf on landing pages
- 5 pm meeting with team

Wednesday July 6th, 2011

- Additional edits to bioinformatics solutions
- Read *Inbound Marketing* – SEO
- Follow up on Google Keyword selection
- Began reading EdgeBio article on MiSeq vs. Ion Torrent
- Began making charts for whitepaper

Thursday July 7th, 2011

- Read EdgeBio, Genome Web and Illumina articles / presentations on MiSeq – Ion Torrent comparison
- Whitepaper research / writing
- Began list of companies for competitive evaluations

Friday July 8th, 2011

- Researched potential competitors/partners
- Competitive Evaluation – Cofactor Genomics
- Sent Allison e-mail inquiring about quote for customer videos
- Read pdf on landing pages
- Communicated with Ryan Creason concerning customer videos

WEEK 4

Monday July 11th, 2011

- Sent message to Jessica/brainstorm related to upcoming move
- Edited latest version of IG executive summary
- Finished reading pdf on landing pages
- Sync with Nathan
- Researched Wash U's GTAC / created spreadsheet to compare pricing

Tuesday July 12th, 2011

- Edited latest version of IG executive summary
- Prepared meeting notes/ reviewed Cofactor Genomics Competitive Eval
- Prepared notes for meeting with Ryan Creason
- Meeting with Nathan to prepare meeting notes
- Filled in parts of complete/collab ecosystem spreadsheet
- Whitepaper writing – PGM platforms
- 5 pm team meeting

Wednesday July 13th, 2011

- Edited website content (bioinformatics solutions text, bulleted text, highlights text, text on homepage, text in sitemap)
- 10 am – meeting with Ryan of CreasonCreations
- 1 pm – meeting with Nathan to discuss website content
- Filled in content on complete/collab ecosystem spreadsheet
- Sent out e-mail to team highlighting changes to website content
- Whitepaper writing

Thursday July 14th, 2011

- Whitepaper writing
- Posted whitepaper draft to Google docs for initial feedback
- Read IM chapter on Social Networking
- Set up a StumbleUpon account
- Filled in content on complete/collab ecosystem spreadsheet

Friday July 15th, 2011

- Google Keyword optimization
- Re-worked numbers for GPB quote (based upon incremental project stages)
- Meeting with Nathan/Doug to discuss GPB quote (x2)
- Created a FB page for IG
- Filled in content on complete/collab ecosystem spreadsheet

WEEK 5

Monday July 18th, 2011

- Took call from Apollonia regarding payment terms => expect call from Myriam tomorrow
- Sent e-mail to team regarding UGA quote (Library prep and HiSeq)
- E-mailed Jessica regarding specifics of St. Louis arrival/ Rent
- Read through and edited 2 Opinion Blog Posts
- Read through and edited HTS Whitepaper
- Filled in relevant content in tomorrow's meeting notes
- Keyword optimization

Tuesday July 19th, 2011

- Keyword optimization => Google search results
- Updated complete/collab ecosystem spreadsheet according to above findings
- Met with Nathan to prepare weekly meeting notes
- Team Meeting
- Read *Inbound Marketing*

Wednesday July 20th, 2011

- Reviewed and revised blogs and whitepaper according to team feedback
- Sent out latest versions of blog and whitepaper to team
- Finished *Inbound Marketing*
- (Website Launched)

Thursday July 21st, 2011

- Linked website to FB fan page and twitter page
- Exchanged dialogue with team concerning IG social media presence
- Research/Action based upon Inbound Marketing suggestions
- Sent e-mail to team concerning UGA quote for GPB job (awaiting confirmation from UGA)
- Filled in content in complete/collab ecosystem spreadsheet

Friday July 22nd, 2011

- Finalized complete/collab ecosystem spreadsheet
- Met with Nathan to discuss keyword optimization and landing page content
- Sent follow-up e-mail to Myriam regarding GPB job
- Sent e-mails to team regarding editing my blog posts/landing page and results of keyword optimization

WEEK 6

Monday July 25th, 2011

- Edited blog posts according to Nathan's feedback
- Briefly investigated Google Analytics => no data yet
- Made a list of tools that may be useful for website analytics
- Generated a spreadsheet to begin keeping track of IG and competitor web grades
- Added a short Bio to IG Twitter account
- Edited blog posts according to Doug's feedback
- Re-evaluated Google keywords for optimization
- Created a new sheet in complete/collab ecosystem spreadsheet for competitors
- Began reading Todd/Doug's RNA-Seq book Chapter

Tuesday July 26th, 2011

- Filled in content in Google keywords optimization spreadsheet
- Edited blog posts according to Todd's feedback and send out final version
- Completed 5 am Solutions competitive evaluation
- Updated formatting of all competitive evaluations in Google docs
- Sync with Nathan to fill in today's meeting notes
- Google keyword optimization according to [exact] match
- Updated formatting of competitive evaluations to SWOT format
- Team meeting

Wednesday July 27th, 2011

- Meeting with Nathan to Sync/ decide on keywords for trial Adwords campaign
- Began editing white paper according to Nathan's feedback
- Meeting with Nathan to discuss progress of GPB job
- Called Myriam to inquire about pricing for new sequencing strategy
- Finalized white paper edits according to Nathan's feedback
- Continued reading RNA-Seq chapter

Thursday July 28th, 2011

- Finished reading RNA-seq chapter
- Read through white paper edits and e-mailed to Nathan for second round of feedback
- Developed content for landing page (Topic: Bioinformatics Tools)
- Read through updated IG financial plan (to be discussed next Tuesday)
- Research for HTS white paper (reagent/software costs)
- Second round of HTS white paper edits => send to Doug and Todd for feedback

WEEK 7

Monday August 1st, 2011
- Completed major PGM white paper edits

Tuesday August 2nd, 2011
- Edited HTS white paper -> addition of automated sample prep capabilities
- Finalized first draft content for PGM white paper
- Meeting with Nathan to discuss August Cookbook and meeting notes
- Organized written notes
- Began drafting RNA-Seq blog
- Team meeting
- Created a Google docs collection to house info on sequencing providers as well as a collection to keep track of customer/provider interactions

Wednesday August 3rd, 2011
- Updated interaction log with Myriam Belanger
- Organized written notes
- Researched competitor strategies utilizing complete/collab spreadsheet as a starting point

Thursday August 4th, 2011
- Rogue Rafting Trip

Friday August 5th, 2011
- Rogue Rafting Trip

WEEK 8

Monday August 8th, 2011
- Finished packing for St. Louis

Tuesday August 9th, 2011
- Drove to Portland
-

Wednesday August 10th, 2011
- Travel day
- Caught up on e-mail

Thursday August 11th, 2011
- Toured Danforth Center with Jessica
- Gained access to BRDG Park Office Space
- Organized e-mail and oriented myself on new leads
- Researched public transit in St. Louis
- Obtained BRDG Park Access cards for Nathan, Doug and I
- Shopped with Doug for office furniture (Office Max, Target, STL Office Supply)
- E-mailed Mark Gorski to alert of package arriving tomorrow
- Filled out and sent Doug cover sheet for GPB samples to be sent to GGF

Friday August 12th, 2011
- Organized Gmail and updated iGoogle, Google Reader
- 10 am Sync with Nathan
- Read through briefly and sent out latest versions of HTS and PGM white papers
- Discussed Denver lab informatics with Doug
- Obtained GPB package
- GoToMeeting with Hubspot
- Contacted GPB regarding arrival of samples (expected 6, received 4)
- Communicated with Doug and Nathan concerning strategy for sending GPB samples to GGF
- Communicated with GPB concerning # of samples expected vs. received
- Communicated with Myriam concerning expected sample arrival date and change in sequencing strategy

WEEK 9

Monday August 15th, 2011
- Organized IG Gmail
- Researched Illumina Genome Network
- Researched Illumina Genome Network Partners
- Organized Competitor Strategy spreadsheet
- Began Competitor Strategy powerpoint
- Troubleshooted with Nathan – Intuitive Genomics e-mail
- Caught up on weekend and morning e-mail

Tuesday August 16th, 2011
- Researched Macbook hotkeys
- Continued crafting Competitor Strategy powerpoint
- Updated Competitor Strategy spreadsheet
- Competitive Evaluation of DNAnexus
- Met briefly with STG regarding internet connection
- Team Meeting at BRDG Park => 5 pm
- Crafted Illumina Genome Network powerpoint

Wednesday August 17th, 2011
- Completed Competitor Strategy powerpoint
- Completed preliminary competitor evaluations for Bio::Neos and Duke Institute
- Read about potential funding options – i6 and MTC
- Met with St. Louis County to discuss criteria and application process for the Helix Fund
- Developed Bio for BRDG Park Website => Send to team, then on to Mark
- Identified potential partners that do not offer NGS technologies

Thursday August 18th, 2011
- Met with Mark to discuss BRDG Park signage and i6 Project Co-applicant Meeting
- Downloaded relevant application for various funding options and uploaded to Google docs
- Generated document compiling application process and deadlines for various funding options; uploaded to Google docs
- Left message with Myriam Belanger concerning RNA-seq quote for Jennifer Normanly
- Re-issued IG Bio to Mark
- Updated Google calendar: i6 info meeting and Illumina User Group Meeting
- Confirmed new phone number
- Continued drafting RNA-seq blog

Friday August 19th, 2011
- Registered for Agrigenomics User Group Meeting (August 25th)
- Recorded previous project quotes from GGF
- Mocked up quote for Jennifer Normanly's RNA-seq project
- RNA-seq blog – split into 2
- Attempted to call/ e-mailed Myriam regarding Jennifer Normanly quote

WEEK 10

Monday August 22nd, 2011

- Posted link to BRDG Park Tenant Page to FB, Twitter
- Spoke with Myriam Belanger/Travis Glenn regarding Jennifer Normanly quote
- Drafted Jennifer Normanly quote
- RSVP'd to Biogenerator i6 info session (August 29th)
- Spoke with GFI (Stephanie) regarding use of printer/copier/scanner
- WebGBrowse Project
- Edited RNA-seq blogs -> sent out to team

Tuesday August 23rd, 2011

- Researched and ordered business cards for Todd, Doug and I
- Registered DB for Agrigenomics Meeting Thursday
- Filled in weekly meeting agenda
- Uploaded competitor evaluations to Google docs
- Inquired and obtained long distance access code for office phone
- Danforth Tea Time to collect Cardinals ticket (3 pm)
- Weekly Team Meeting (5 pm)

Wednesday August 24th, 2011

- Phone system logistics (long distance access code/voicemail)
- Picked up business cards
- Researched Danforth Faculty
- Researched types of startup funding
- Met with Kristen Hinzman (Tour of common areas and delivery of tenant handbook)
- Reviewed competitive evaluations

Thursday August 25th, 2011

- Illumina Agrigenomics User Meeting

Friday August 26th, 2011

- Introductions to BRDG Park Tenants (plan for 2nd floor social)
- Laptop connected to shared printer
- Researched investment terminology
- Researched applicable research techniques
- Added header to HTS and PGM white papers
- Researched Illumina Meeting sponsors
- Reviewed MTC application – due Sept. 1st

WEEK 11

Monday August 29th, 2011
- Welcome lunch
- Printed MTC application
- Furniture Shopping
- CET BED Program and BioGenerator i6 Project Info Session
- Finalized data collection for WebGBrowse

Tuesday August 30th, 2011
- MTC Application
- Add Google doc describing Hudson-Alpha pricing
- CET, Coalition for Plant and Life Sciences, CORTEX Research
- "Conversations with Jim Carrington" => Collect content for next blog article

Wednesday August 31st, 2011
- Mocked up quote for sample prep and pooling of 8–12 DNA samples with GGF
- Drafted blog based upon "Conversations" series; sent out to team for review
- E-mailed/Left a message with Myriam regarding quote for sample prep and pooling
- Introductions with Todd Michael
- Champagne Toast to celebrate IG's first year

Thursday September 1st, 2011
- Added hyperlinks to "Conversations" blog
- Finalized quote with Myriam for DDPSC/OSU library prep
- Received new HiSeq pricing from GGF -> added to pricing google doc
- Attended Mockler Lab meeting to hear Doug's practice thesis defense
- Toured Hubspot Trial with Nathan
- Identified competitors for Hubspot/Twitter following
- Updated Hubspot keyword grader
- Skimmed "Business Model Generation"

Friday September 2nd, 2011 (Tulsa)
- Keyword Grader -> classified relevance of keywords
- Created a landing page using Hubspot template
- I6 application
- Draft ad and flyer for Symposium

WEEK 12

Monday September 5th, 2011 (Labor Day)

Tuesday September 6th, 2011
- Monitored Hubspot trial (Cofactor Genomics, Appistry)
- Researched SOLiD Sequencing chemistry
- Tuesday meeting agenda
- Set meetings for discussion of symposium materials, competitor strategy presentation
- Determined best date for PSM Final Presentation/ Reserved room
- Began reworking content for "Conversations" blog
- Read Danforth Newsletter
- Tuesday team meeting – 5 pm
- Linked new blog article to social media

Wednesday September 7th, 2011
- Revised IG Executive Summary for Atlas
- Met with Nathan to discuss Symposium materials
- Met with team to discuss lead compilation strategy
- Added blog post link to dig and reddit
- Communicated with GGF concerning DDPSC/OSU project details (shipping and billing addresses)
- Re-wrote 'Conversations' blog
- Began working on Symposium powerpoint slide

Thursday September 8th, 2011
- Picture and text to Nathan for personalized author section on blog
- Re-structured Conversations blog => challenges and opportunties surrounding biocomputing
- Revised competitor strategy presentation
- Began compiling potential leads from Wash U Med School

Friday September 9th, 2011
- Revised and sent "Role of Biocomputing" blog to JC
- Furniture delivered
- Monitored Hubspot and added companies to complete/collab spreadsheet
- Brainstorm/discussion for i6 Project application
- Began developing landing page
- Lead compilation
- Research competition for i6 project
- Happy Hour with Bioglow

WEEK 13

Monday September 12, 2011
- Assembled office furniture
- Finalized content for "Illumina Sequencing" landing page
- Finalized content for Symposium ½ page ad and powerpoint slide
- Sat in on call to Kate Sydney (Knome)
- Reviewed i6 application
- Researched "data compression" space
- Compiled list of SLU Med faculty

Tuesday September 13, 2011
- Read "Business Model Generation"
- Brainstormed tagline and description for Google Adwords campaign
- Met with Nathan to discuss competitor strategy presentation
- Posted PGM Opinion Blog Part 2 to social media sites
- Compiled list of SLU Med faculty
- Wednesday Meeting agenda
- Began updating competitor strategy presentation

Wednesday September 14th, 2011
- Finalized Symposium materials (Printed flyers and sent remainder to K. Mackey)
- Lead Compilation (incorporated TCM's contacts and researched applicable departments)
- i6 application

Thursday September 15th, 2011
- Finalized and submitted i6 application
- Provided UGA quote for Monsanto job

Friday September 16th, 2011
- Meeting with Sam Fiorello
- Typed up notes based upon meeting with Sam
- Reviewed and sent out latest version of HTS whitepaper
- Updated Competitor Strategy Presentation
- Weekly team meeting

WEEK 14

Monday September 19th, 2011

- Made final revisions to HTS White paper
- Reviewed "Role of Biocomputing" blog and sent out to team
- Generated content for landing page to house HTS white paper
- Registered for Research Commercialization webinars

Tuesday September 20th, 2011

- Reviewed and provided edits to landing page content developed for HTS Guide
- Meeting with team to discuss change in leadership
- Brainstorm for decision to continue working for IG
- Reviewed compensation agreement and payment up through today

Wednesday September 21st, 2011

- Created an e-mail template to "thank individuals for stopping by our booth" at Symposium
- Began research on SBIR opportunities
- Meeting with Doug to prioritize tasks/discuss change of leadership and new job offer
- Data Compression Software Research
- Organization for report on Nathan's plate
- Began listening to Commercialization Lecture #1

Thursday September 22nd, 2011

- Sent e-mail to team regarding job offer (Meeting Friday to discuss terms)
- Listened to remainder of commercialization lecture # 1
- Updates on status of CRM e-mailed to team
- Reminders for Danforth Symposium e-mailed to Doug
- Data Compression Research
- Began crafting 1 pager for Data Compression Project

Friday September 23rd, 2011

- Sent introduction e-mails to all customers directing all future communications to me
- Collected Danforth Symposium Flyers from Kathleen
- Created google doc to keep track of customer interactions

WEEK 15

Monday September 26th, 2011

- Finalized revamp of Symposium flyers and returned to Kathleen
- Responded to Richard Yu's e-mail concerning GPB job
- Jim's Surprise Birthday Party
- Inpromptu meeting with Shaukat to discuss joint offerings
- Phone Conversation with Maxim from the BALSA group

Tuesday September 27th, 2011

- Researched Paychex to determine expectations for next pay period
- Made copies of Paychex statements for my own records
- Conference call with Atlas Client (Art Krieg) to discuss large sequencing and bioinformatics project
- Discussion with Todd and Doug concerning job offer
- Preparation for Symposium
- NSF iCorps Research

Wednesday September 28th, 2011

- Danforth Symposium
- Processed Payroll for 9/16 – 9/30
- Communicated with Nathan to re-establish IG domain (web and e-mail)

Thursday September 29th, 2011

- Danforth Symposium
- Set up call forwarding
- Set up Google voice number
- Sent out summary info of NSF iCorps Program

Friday September 30th, 2011

- Danforth Symposium
- E-mailed committee members to alert them to my return this weekend
- Cleaned up both Google accounts, internship documents and Google docs
- Tested BRDG Park internet and Google voice e-mail
- Finalized employee contract with Doug and Todd
- Outlined 2 additional NSF programs
- Began brainstorming components for PSM Final Report

- Blank -