PacBio: Long Read Sequencing Capability for Whole Genome Sequencing

PacBio: Long Read Sequencing Capability for Whole Genome Sequencing

Tamara Mason

A Thesis in the Field of Biotechnology

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

November 2021

Abstract

The advancement in cameras and computation ability has allowed the sequencing industry to make revolutionary progress. One of the latest sequencing technologies, called “Third generation sequencing”, relies on the reading of long, continuous fragments of DNA. The long, continuous sequencing of DNA fragments allows for less reliance on post-sequencing analytics and fewer restrictions when identifying difficult-to-assemble regions of the genome. In order to optimize the workflow and make the technology available for research efforts, The Broad Institute’s Genomics Platform added ‘Long Read’ Sequencing to its repertoire of genomic services. Pacific Biosciences released the Sequel II instrument in early 2019 along with a higher output protocol and sequencing chip (8M SMRT Cell) that was reported to produce up to 8 times as much data as the original Sequel I instrument. The new workflow claimed to provide higher output, Q50 accuracy, uniform coverage, and a reduction in GC bias content. This claim was tested by comparing this new instrumentation to the previous Pacific Biosciences version as well as a second-generation sequencing equivalent. Size selection and fragmentation optimizations, as well as other process improvements were undertaken to optimize the lab workflow. Ultimately, it was found that new product is faster and simply more cost-effective than its previous version. With the sinking costs and increased output, it was found to be feasible to move the Pacific Biosciences technology into a more mainstream whole human genome application, which provides an innovative way to identify genomic error, such as variants and mutations, within the population.

Dedication

To my husband J.D. who both brought me perfectly peeled clementines and gave me space as needed during the duration of this thesis.

.

Acknowledgments

This thesis would not have been possible without the work and support of the Long Reads team on the Genomics Platform who created the long reads capability : Maura Costello, Erin LaRoche, and Ally Day. I'd also like to thank those that contributed specifically to the data analysis and work on the pipeline generation end: Brian Granger and Michael DaSilva. Special thank you to Dr. Kiran Garimella, without whom, I do not think the platform's ability to create a long reads product would have occurred. I also thank Pacific Bioscience;s FAS, Michael Weiand.

Finally, I'd like to thank my thesis director Jonna Grimsby, who graciously agreed to work with me on this considerable project.

.

Table of Contents

List of Tables

List of Figures

Chapter I.

Introduction

The field of genomics has undergone much change in the last 25 years. Different types of sequencing technologies have emerged and the advancement has led to new and exciting discoveries. *The Human Genome Project* allowed for the sequencing and analysis of a single full genome in the late 20th century at the cost of $2.7 billion dollars (Lander et al., 2001). Early 21st century advances have led to projects such as All of Us, which aims to sequence a widely diverse group of over 1 million Americans over approximately the same timeline, for approximately half the cost ($1.5 billion dollars) (*All of Us* Research Program Backgrounder, 2019). Constant advancements in genomics allow for projects of escalating size and consequence to be conceived and implemented.

The Discovery of DNA

With crystallographic data produced by Rosalind Franklin and Maurice Wilkins, Watson and Crick are well known throughout the scientific community as being the first to posit the three-dimensional structure of DNA in 1953 (Watson, 1953 and Zallen, 2003). Although this led to research into DNA replication and encoding proteins in nucleic acid; scientists had a more difficult time developing investigation techniques to study nucleic acids as a whole – which have longer, harder to distinguish molecules (Heather, 2016).

Two Dimensional Fractionation

The first nucleic acid sequenced was ribonucleic acid (RNA); shorter than DNA and single stranded. Early techniques measured nucleotide composition, but not necessarily the order of nucleotides within the fragments. In 1965, Robert Holley (et al.) was able to produce the whole nucleic acid sequence of alanine tRNA molecule from yeast using ribonuclease treatments. Also in 1965, Fred Sanger (et al.) applied a radiolabel technique and two-dimensional fractionation to partially digested fragments, which allowed him to sequence many ribosomal and transfer RNAs. Two-dimensional fractionation was then used by Walter Fiers' lab in 1972 to construct a protein-coding gene and then a bacteriophage's genome in 1976 (Heather, 2016). By 1968, Ray Wu and Dale Kaiser published a DNA polymerase method to fill in overhanging 5' ends of DNA with radioactive nucleotides. By adding the nucleotides one at a time, the incorporation of individual bases could be deducted first in the overhang, and then anywhere within the fragment; albeit for only short stretches of DNA (Padmanabhan, 1974).

Polyacrylamide Gel Sequencing Techniques Lead to 'First Generation Sequencing'

Two-dimensional fractionation was replaced in the 1970s by polyacrylamide gel techniques, which sorted fragments by length and gave greater output. Several new workflows utilized this method and 'first-generation' sequencing methodology was developed.

Alan Coulson and Fred Sanger developed a 'plus and minus' system in 1975. This technique first identified a fragment of interest and then created complementary 'primers' to act as starting locations. Eight conditions were tested. Four 'plus' reactions were tested for each nucleotide separately. These were prepared in a polymerase that would cause

degradation in that specific region, leading to the last base in the sequence being the specified nucleotide. Four ‘minus’ reactions were comprised of three of the four nucleotides paired together with a DNA polymerase which allowed standard DNA synthesis until the missing nucleotide was encountered. This allowed one to infer that the next base in the sequence would be the missing nucleotide. Using electrophoresis, the fragment lengths could be compared among the eight conditions and the sequence could be determined. This method was used to sequence the first DNA genome - ϕX174 (also known as PhiX, a well-known positive control genome). (Sanger, 1975)

Allan Maxam and Walter Gilbert's chemical cleavage technique in 1977 also used electrophoresis. Instead of using polymerase to mimic synthesis, Maxam and Gilbert used chemicals to cleave fragments of interest at specific points (Heather, 2016).

Also in 1977, Sanger developed the ‘chain-termination’ technique which takes advantage of adding radiolabeled dideoxynucleotides (which halt synthesis) and normal deoxyribonucleotides (which promote synthesis) to a template DNA fragment in order to create DNA strands of all possible lengths. If one performs four separate reactions with the four different dideoxynucleotides, one can then run the resulting fragments and identify the terminating base based on fragment length. This is known as the dideoxy chain-termination method or Sanger Sequencing. Advancements of this technology included replacing radiolabeled ddNTPs with fluorometric based detection (allowing a single reaction to occur rather than four) and capillary based electrophoresis - leading to more and more automated DNA sequencing instruments. This technology led to the first commercially offered DNA sequencing instrument. (Sanger, 1977)

In order to sequence longer fragments of DNA, a technique called ‘shotgun sequencing’ was developed in the late 1970s that derived nucleotide sequences of overlapping fragments and assembled them post-sequencing (Anderson, 1981). Soon after, many improvements and additions to genomic technologies were made: polymerase chain reactions (PCR), recombinant DNA practices, finding more efficient ddNTPs and more efficient analytic pipeline tools (Heather, 2016). These new discoveries allowed the limits of genomic sequencing to expand rapidly.

Second Generation Sequencing: Massively Parallel Sequencing

The next generation of sequencing (Second Generation Sequencing) relied on massively sequencing fragments in parallel using luminescent methods. Unique to each nucleotide, luminescent pyrophosphates were engineered to fluoresce as their attached nucleotide was synthesized onto an immobilized template fragment of DNA via DNA polymerase. Reading the resulting fluorescent signatures allowed researchers to determine the exact sequence of the fragment via ‘sequencing by synthesis. (Nyrén, 1987) This technology led 454 Life Sciences to produce the first commercial ‘second-generation’ sequencing instrument, GS 20. Sample preparation for this instrument used a bead-based oil emulsion PCR, which produced significantly more output. Subsequent improvements included reaction volume reductions to the micrometer scale and advancements in high resolution imaging (Shendure, 2008).

Many new second-generation sequencing instruments followed the GS 20. Solexa (acquired later by Illumina) produced the predecessor of today’s market share leading instrumentation, the Genome Analyzer. This new instrument utilized similar fluorescent based ‘sequencing-by-synthesis’ techniques, adapter bound fragments that are

immobilized onto complementary lawn oligos rather than bead-based techniques. This allowed fragments to colonize, producing a 'cluster' of identical fragments that could be more easily interpreted by imaging software and also allowed for 'paired-end' sequencing. Paired-end sequencing improved the mapping accuracy of reads to reference sequences, detection of repetitive sequences, and identification of spliced exons and rearranged fragments. (Turcatti, 2008). Although initial instruments had very short read ability (35 basepairs), newer generations of this technology increased read lengths to up to 300 basepairs. In recent years, significant advancements to the flow cell seeding process - replaced with nanowell technology, have increased output many fold, allowing for faster run times and significant reduction in costs (Costello, 2018). Sequencing capabilities have doubled approximately every five months between 2004 and 2010 (Stein, 2010). This increase in capabilities was led by the Illumina sequencing platform, which has a near monopoly of the industry (Greenleaf, 2014).

Second generation sequencing managed to enable many innovations in the field of genomics. In summary, it allowed for a characterization of DNA variation, de novo sequencing of many species, microbiome sequencing, methylation detection, characterization of genetic isoforms, characterization of methylome and transcriptomes, and a better understanding of protein and DNA interactions. (Schadt, 2010). Despite the utility and tremendous contribution of short reads by second generation sequencing, these short reads have made it harder to assemble (de novo) long DNA sequences. Long reads are more powerful in resolving repeat regions (Chin, 2013) and complex genomic regions. As of 2015, more than 160 euchromatic gaps still remained in the human genome assembly, often containing short tandem repeats >1kbp (k basepairs) in GAC

rich regions including inversions, complex insertions, and long tracts of tandem repeats (Chaisson, 2014).

Third Generation Sequencing: Single Molecule Sequencing

Third generation sequencing, which is loosely defined as single-molecule sequencing (SMS) (Schadt, 2010), overcomes many points of contention that have been found with second generation sequencing technologies. The main issue it resolves is the bias introduced by PCR in second-generation sequencing, but it also has the advantage of increased read length and decreased time to prepare libraries and sequence. Overall, potential advantages are higher output, faster turnaround time, longer read length to assist in de novo efforts, longer read length to detect haplotypes and whole chromosome phasing, and higher consensus accuracy for increased sensitivity for rare variants as compared to second generation sequencing data. One major area where this generation of sequencing lags behind second generation sequencing is in terms of cost. Second generation sequencing has been able to scale output to a point where costs are very low. (Schadt, 2010)

Single-molecule sequencing was first performed by Stephen Quake's Lab and commercialized by Helicos BioSciences and was performed very similarly to second generation sequencing, but without PCR amplification. Amplification bias was averted, but the technology was expensive, so the company did not survive (Heather, 2010).

The next version of single molecule real time sequencing (SMRT sequencing) was developed by Pacific Biosciences (PacBio). It utilizes DNA polymerase working in approximately real time along with florescent-labeled nucleotides arranged along a zero-

mode-waveguide to produce a continuous movie of DNA polymerases' activity on prepared long DNA fragments. Zero-mode waveguides (ZMW) are small microfabricated nanostructures (70uM diameter and 100uM in depth) surrounded by aluminum walls and filled with DNA polymerase. Original cells had ~150,000 ZMW per SMRT Cell. The structure of the well allows for the activity of a single DNA fragment's synthesis with fluorescently tagged dNTPs monitored by real time imaging, creating a 'movie' of the polymerase activity which can be analyzed to interpret the sequencing of the DNA. Two to four nucleotides are synthesized per second. Emission patterns of the fluorescent tags both detail the base (distribution) and show subtle, measurable differences that are due to epigenetic modifications. (Bleidorn, 2015)

This initial method developed by Pacific Biosciences allowed for the longest read length available – >1,000 basepairs and often >10,000 basepairs, which then allowed de novo assembly to be less hindered by alignment and consensus necessary for second generation sequencing. This allowed for easier characterization of structural variation, insertions, and deletions. Kinetic information is another output unique to Pacific Biosciences. Enzyme incorporation (kinetics) can be seen in real time during data collection – allowing methylation to be detected easily alongside sequence (Schadt, 2010). Other applied applications included real-time observation of ribosomes as they translated mRNA – which allowed for an up-close glimpse into translational activities at the nucleotide level (Uemura, 2010). Previously, large gaps in genome sequences were filled using high-effort, high-cost capillary-based sequencing (Sanger shotgun sequencing). Pacific Biosciences can be used to resolve structural unknowns and finish genome assemblies in a straightforward manner (Huddleston). One major setback of this

technology is error rate. When introduced, the technology had 20% error rate with random error generation and produced fairly low amounts of data (Bleidorn, 2015). Much effort is aimed to address this error rate in terms of applying backend pipeline optimization and analysis to create a comparable error rate to second generation sequencing alternatives.

Soon after Pacific Biosciences introduced SMRT sequencing, nanopore sequencing methodologies were developed. Nanopore sequencing was developed by Oxford Nanopore Technologies (ONT), first available with the MinION instrument. As opposed to the very large Pacific Biosciences instrumentation, the MinION is the size of a cell phone. The technology is based on a-hemolysin protein's ability to create nanopores within phospholipid layers, which act as a small channel that can be controlled with ionic gradients, which can then filter nucleotides, the bases of which can be identified by the change in ion currents (Feng, 2015). Initial models of the instrument had 512 channels detecting 10 basepairs per second. Double stranded DNA was prepared with a hairpin adapter, which is unzipped and threaded through the pore, allowing both of the DNA strands to be sequenced, and a consensus to be determined. Error rates are as high as 25-40%. Output is 90-490 mbp per run, average length of 6 kbp and max of 150kbp. (Bleidorn, 2015).

Increasingly, data management and processing needs related to computation have become a problem (for all high output sequencing), including data transfer, access control/management, standardization, and output file format standardization (Schadt, 2010). The research presented here will assess how Pacific Biosciences handles these processing needs in order to be an efficacious alternative to ONT technologies.

The Current Status of Pacific Biosciences Long Read Sequencing at the Broad Institute

The Broad Institute's Genomic Platform has worked to first implement and then optimize Pacific Biosciences as a Long Reads product for the past several years. It is a valuable resource to add to the platform's well established second generation sequencing options.

Pipeline

Originally, the long reads analysis pipeline (LRWholeGenomeSingleSample.wdl) was written by Broad Institute Senior Computational Scientist, Dr. Kiran Garimella. He worked in conjunction with the Broad Institute's Data Sciences Platform to analyze Pacific Biosciences data using his novel script in Workflow Description Language (WDL). Soon after, Oxford Nanapore instrumentation was acquired by the Genomics Platform and these scripts were altered to analyze all long read sequencing outputs (third generation sequencing).

Pipeline analysis can be a slow and costly endeavor. Cloud based computing can make analysis quicker, but with the high volume of data output by long read sequencing instruments, the analysis speed would remain hindered. This pipeline was built to utilize an application called "ZMW Sharding". With the patterned SMRT Cell and expected number of ZMWs available, the application splits reads after a certain number of reads have been written to a shard (e.g. ~100,000 basepair section of the read), and only when the next read's ZMW ID does not match the current one. Figure 1 shows the 'ZMW Sharding' application workflow. Essentially, this splits the data-heavy analysis into pieces so that they can be processed much more quickly separately and then merged later, so that less data per node is outputted and can be quickly merged. An application was

built to do variant calling via GATK, DeepVariant, pbsv, Sniffles, etc., meaning that analysis via multiple variant callers is no longer necessary. Efforts are being made to introduce this script to Terra, an openly available cloud space, which will allow the script to live as an open-source analysis tool for researchers to use.

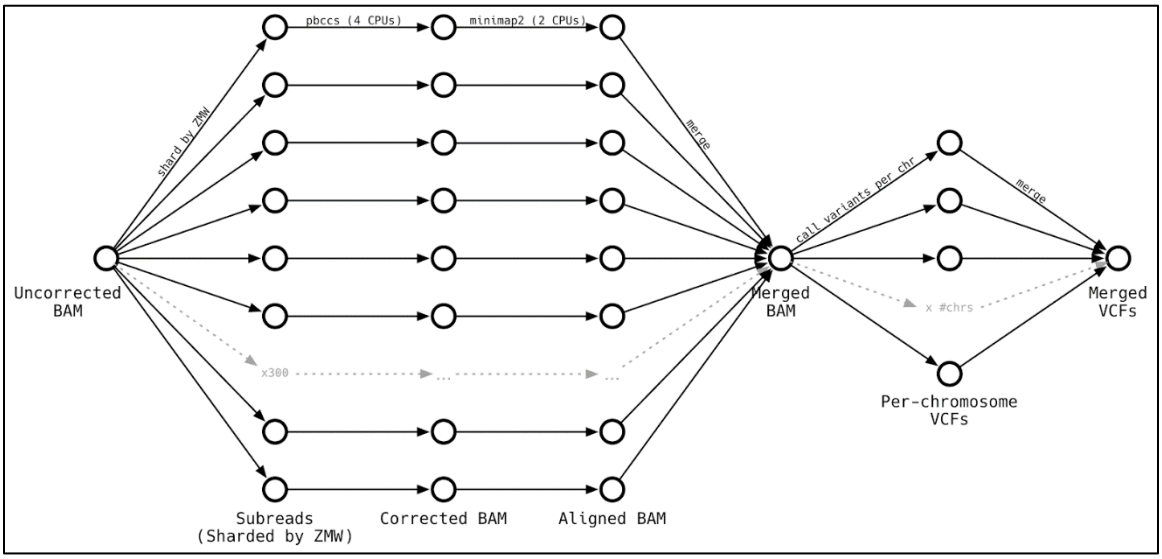


Figure 1: Cloud Based Long Read Pipeline Analysis Strategy

This figure details how the long read WDL endeavors to speed up analysis by fragmenting the data, performing analysis in parallel, and recombining data to have a pipeline that can keep up with the high volume of data production.

Library Construction

Initial implementation of Long Read Sequencing Applications in the Genomics Platform at The Broad Institute utilized a Circular Consensus Sequencing (CCS) approach. This method assumes that one polymerase enzyme will produce genomic sequencing data for a single template fragment (a single library) multiple times within a

single sequencing run. Without further analysis, the error rates can be as high as 10-15% (Lu, Giordano, & Ning, 2016). However, these errors are random rather than systematic and many can be eliminated through circular consensus sequencing (CCS) analysis applications which can correlate several reads of the same DNA fragment to verify the individual bases.

Library construction includes a process called Solid Phase Reversible Immobilization (SPRI) that utilizes magnetic beads to bind nucleic acid by size and is used for clean up protocols (to remove very large or very small fragments from the reaction solution). Utilizing this method is one way to control size within library construction, but fragmentation and gel-based size selection methodologies are also used ensure consistent DNA size and quality.

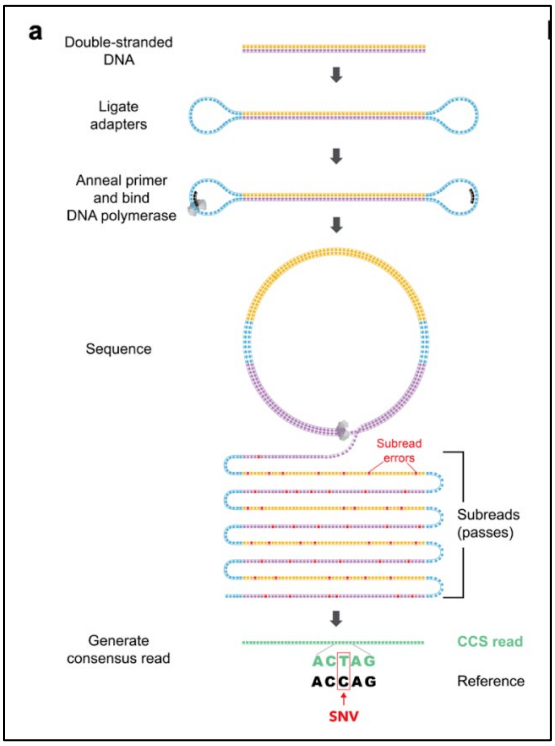


Figure 2: General Pacific Biosciences Workflow for CCS Libraries

Figure 2 shows the general pathway raw DNA will take to produce data output. Figure published in "Highly-accurate long-read sequencing improves variant detection and assembly of a human genome" by Wenger, et al. in Nature Biotechnology, 2019.

After the success of CCS libraries, another type of library, Continuous Long Read (CLR) libraries, were produced. CLR libraries are long strands of DNA library that are not fractionated (or minimally so) that allow for a continuous template to be sequenced. Long reads like these may have higher error rate but can be more impactful for detecting large genomic mutations or rearrangements or for de novo sequencing efforts.

Size Selection Optimization

Size selection is an important step in constructing high quality libraries for sequencing on Pacific Biosciences instrumentation. This size selection allows for accurate sequencing of targeted molecules. Small fragments will take up real estate on a SMRT Cell, but will produce fewer unique bases. Larger fragments can be less stable and can produce fewer high quality reads (depending on library type).

Gel-based size selection of large genomic DNA fragments is completed by a BluePippin from Sage Science for Pacific Biosciences library preparation. This instrument creates a single, low yield DNA fraction, allowing for only enough molecules to sequence on one or two SMRT Cells. Pacific Biosciences has recommended the SageELF Instrument, which takes up to 12 different fractions from each library and allows for collection of a main 10 kb library along with closely-sized fragments to use in cases where additional yield is required; for instance, lab error, instrument failure, or deeper coverage needs.

Fragmentation Optimization

At the 'long read sequencing' product launch in the Broad Institute's Genomics Platform in 2018, the Diagenode's Megaruptor 2 instrument was utilized to fragment DNA into large fragments of a desired size. Fragmentation of genomic DNA is a critical part of library construction. The process must fragment the DNA while also maintaining quality of each fragment. The Megaruptor 2 was built to fragment 1 or 2 libraries between 3-75 kilobases (Megaruptor 2, 2021). The newest version, Megaruptor 3, allows for fragmentation of up to 8 samples to 5-100 kilobases simultaneously (Megaruptor 3, 2021). Initially, 10 kb fragment sizes were desired for CCS libraries. Later, larger fragments of 15kb and 30kb were sought to enhance updated chemistry and to accommodate CLR library construction workflows.

Sequel Operations Over Time

The Genomics Platform obtained a Sequel I instrument to establish the long read sequencing program in 2018. In 2019, in an effort to deepen it's long-read knowledge base and offerings, the platform worked with Pacific Biosciences to validate the Sequel II with updated flow cell chemistry. This included the use of an updated SMRT Cell (8M vs 1M read output); a critical update if the instrument was to compete with Illumina genome output (both in cost and quality).

Chapter II.

Research Methods

This research methods section describes the materials and processes chosen to assess the Genomic Platform's long read sequencing ability on Pacific Biosciences and the manner with which to analyze data and validate new chemistries. In short, this section details how the Long Reads Sequencing team wrote and assessed an analytical pipeline, optimized size selection and fragmentation methodologies, validated new Sequel II instrumentation and chemistry, and carried out various process improvement efforts to enable long read sequencing.

Pipeline Creation

The Genomics Platform's Translational Analysis Group (TAG) led an effort to upload and update (as needed) the long reads analysis pipeline (LRWholeGenomeSingleSample.wdl) analysis script into Terra, an openly available cloud space. This team led the effort to test efficacy, robustness, and equivalency for data sets of varying sizes (starting with tiny data sets and then ensuring that full data sets were able to be run through the pipeline). Comparison data from parallel analysis pipelines was also used to ensure accuracy of output. To stand this analysis platform up, billing projects and project spaces were created, end users were identified, and training of the Long Read's Operations team was scheduled.

Size Selection Optimization

The process by which size selection was performed was optimized over time. Initially, optimization of current library preparation technology was performed and later a new sequencing instrument was acquired. Experiments regarding improving output based on size were also completed in order to optimize output of the long read product.

Optimizing BluePippin Size Selection

In an effort to optimize Sage Science's BluePippin selection to achieve fragment selection of an average of 10 kb (based on the sequencing manufacturer's recommendation), two fragmentation protocols were tested. For both of these protocols, Sage Science's suggested gel and marker configuration of 0.75% gel + S1 marker (optimal for 3-10 kb) Lambda DNA (2ug) at 15 kb was size selected with the BluePippin. 15 kb DNA was prepared by utilizing Covaris g-tubes (mixing 10 kb and 20 kb sheared DNA together as there was no defined setting for 15 kb). The first mode tested was the Range mode – set to 8-14 kb, centered at 11 kb (Pacific Biosciences's suggested input for CCS library construction). The second mode that was tested was tight mode centered at 11 kb (which should include fragments +/- 20%). For each selection condition, DNA recovered from the Blue Pippin cassette within the desired size range was measured based on Caliper LabChip GX fragment analyzer and compared to pre-selection aliquots.

Optimizing Library Construction Output

An experiment was designed to test if the loss of data seen with longer fragments due to template degradation during the recording of a 30 hour sequencing movie could be alleviated by providing slightly shorter, more robust fragments in order to give more

polymerase passes to each library fragment. This could hypothetically improve the CCS process, increase CCS coverage, and reduce error rates. Three initial library insert size conditions were tested: 9-10 kb (the current standard), 8-9 kb (condition 1), and 7-8 kb (condition 2).

To carry out this insert size experiment, libraries were made using DNA from a HapMap control sample for consistency (GWD HG02982, 311 ng/uL input). To fragment the DNA into appropriate sizes, the 9-10 kb condition sample was sheared on the Diagenode Megaruptor 3 with standard settings, speed 46. The 8-9 kb condition sample was sheared on speed 49 based on an initial Megaruptor 3 validation study. The 7-8 kb condition sample utilized a speed of 54 based on specifications in the Megaruptor 3 manual (Megaruptor 3, 2021). For the gel-based size selection portion of the workflow, the BluePippin was utilized to select fractions centered at 9.5 kb (for 9-10 kb), 8.5 kb (for 8-9 kb), and 7.5 kb (for 7-8 kb). All three conditions underwent standard CCS library preparation with SMRTbell Express Template Prep Kit V1.0. For sequencing, Sequel II sequencing settings for CCS run design were applied, with 30 hour movies.

During library preparation, it was found that fragments were sheared slightly too large to fall into the intended size categories. The fragments intended for 7-8 kb were used for 8-9 kb (final size of 8518 basepairs) and fragments intended for 8-9 kb were utilized for 9-10 kb (final size of 9995 bp). Fragment fractions were reselected on BluePippin with targets 750-1000 bp lower than the previous attempt. Resulting fragments were 7870 bases (used for the 7-8 kb condition) and 6714 bases (used for a new 6-7 kb condition). During sequencing, the initial runs failed after the 9-10 kb size condition completed sequencing and during processing of the 8-9 kb size condition. The

8-9 kb size condition was not able to be reworked due to lack of input DNA remaining. The 7-8 kb and 6-7 kb conditions were reworked and sequenced again. There was no attempt to correct the lack of sequencing results for the 8-9 kb condition as it was determined that the three remaining conditions were sufficient for interpreting results.

Validation and Introduction of SageELF Size Selection

To validate the protocol to support Pacific Biosciences CCS Library construction, an experiment was run to identify proper sizing on Sage Science's SageELF. 3 ug of control DNA was sheared on the Diagenode Megaruptor 3. DNA was placed on SageELF at Pacific Biosciences's recommended target of "3400 bp" and 12 elution products were extracted from the gel. Fractionation efficiency was determined using an Agilent BioAnalyzer (12K chip), which uses capillary electrophoresis to for interpretation of size and quality of DNA.

Fragmentation Optimization

The method by which fragmentation was performed was optimized over time. Testing of a new version of the fragmentation instrumentation provided by Diagenode was carried out. Fragmentation experiments for the testing of various fragment input sizes into library construction protocols were performed to optimize sequencing performance.

Megaruptor 2 vs. Megaruptor 3 for DNA Fragmentation

To help determine if the upgrade from Diagenode's Megaruptor 2 to Megaruptor 3 was beneficial, the manufacturer provided a temporary loan of the newer instrument.

Four separate runs of the Megaruptor 3 system at four different speeds (46, 47, 48, and 49) were completed with speeds set between the suggested speeds for CCS input size ranges of 7-10 kb fragments (speed of 49-56) and 10-15kb fragments (speed of 40-46). Five aliquots of HapMap HG00514 DNA at 5 uL each were used (final concentration 16 ng/uL). There was also a separate run on the Megaruptor 2 system using the same control DNA with the standard validated settings for a 10 kb size output. These fragments were run and assessed on the Caliper (Agilent) using the LabChip GX gDNA QC protocol (the same protocol used for standard Pacific Biosciences library construction quality control) to test for equivalency.

For larger fragments intended for CLR library construction, a similar experiment was performed. Three separate runs of the Megaruptor 3 system at 3 different speeds (28, 30, and 32) were carried out, with speed settings set between the suggested speeds for 30 kb fragments. Four aliquots of HapMap HG00514 DNA at 5 uL each were used (final concentration 16 ng/uL). There was also a separate run on the Megaruptor 2 system using the same control DNA with the standard, validated settings for a 30 kb fragmentation. These fragments were also run and assessed on the Caliper (Agilent) using the LabChip GX gDNA QC protocol to analyze fragment size and test for equivalency.

Fragmentation Alterations for Updated CCS HiFi Chemistry

The introduction of Pacific Biosciences's SMRTbell Express Template Prep Kit 2.0 to create High Fidelity (HiFi) CCS libraries necessitated the need to create 15-20 kb fragments for optimal chemistry performance. To determine the best method for this, three separate runs of the Diagenode Megaruptor 3 system were performed using 1 ug NA12878 control gDNA at three different speeds (36, 38, and 40), with speed settings set

between the suggested speeds for 15-20 kb fragments (speed of 36-38) and 20-30 kb fragments (speed of 33-35). The goal was to create fragments at both 15 kb and 20 kb to further evaluate the new library construction chemistry. These results were compared to previous control samples sheared to ~10 kb on the Megaruptor 3. Resulting output material was assessed using the Caliper LabChip GX gDNA QC protocol to test for equivalency and to interpret size distribution and tightness. Resulting fragments underwent library construction with the SMRTbell Express Template Prep Kit V2.0 and were sequenced on the Sequel II to validate chemistry efficiency as well, resulting in primary sequencing data and Caliper LabChip GX data.

Sequel I vs. Sequel II: Sequencing Capability

The Sequel I was a formidable instrument that offered valuable opportunity for researchers to study genomic data, but was found to be difficult to sequence properly due to short reads and the high volume of post-sequencing analysis required with second generation sequencing methods. The introduction of the Sequel II, in combination with the 8M SMRT Cell, allowed the long read sequencing market to begin to compete in output and cost with the previous generation's product.

Sequel I vs. Sequel II, Operational Validation and Pipeline Implementation

The new Sequel II and its 8M SMRT Cell were compared to the Sequel I and the 1M SMRT Cell. The library construction portion of this new workflow was similar to that of Sequel I library preparation (utilizing the SMRTbell Express Template Prep Kit V1). This allowed the focus of the comparison to be on the performance of the instrument and new SMRT cell changes.

We compared 19 previously sequenced tumor-normal pairs (a pair of samples from the same individual, one with tumor attributes) prepared as 15 kb CCS libraries and sequenced on Sequel I with the standard 1M SMRT Cell in 600 minute movies versus the first 11 HapMap 10 kb CCS libraires sequenced on Sequel II with the new 8M SMRT Cell and the newly recommended 1800 minute movies. Pacific Biosciences claimed that the new chip type (which has eight times as many wells as the previous version) enables high quality 30X long read human genomes in just one or two SMRT Cells. This claim, along with others made by Pacific Biosciences, was analyzed using primary output data (data produced directly on instrument without downstream analysis). A full sample list used for the Sequel I and Sequel II comparison can be seen in Appendix I.

A total of 36 SMRT Cells (Table 1) were sequenced on Pacific Biosciences Sequel II during the validation phase (including some from the comparison study of Sequel I vs.Sequel II). Samples were derived from defined Human Genome Structural Variation Consortium trios, clinical samples, and tumor/normal pairs which were constructed into CCS libraries (and one CLR library). Outputs were compared to expectations defined by Pacific Biosciences (refer to Table 2). Operational workflows for the new instrument were defined, and a standard operating procedure was determined.

Table 1: List and origin of Sequel II Validation Samples

*Curated validation samples for ensuring that Sequel II and updated SMRT Cell produced high quality, accurate data. Samples curated included HapMap samples with defined references and several clinical samples that represented a subset submitted by collaborators. All were prepared with the 10 kb "HiFi" Circular Consensus (CCS) protocol except that noted with *, which was prepared with the 30 kb Continuous Long Read (CLR) protocol.*

Sample ID	Sample Type	Relationship to Proband	Sex	Family	Population	Reference BioBank
NA12878 (rep 1)	HapMap	proband	female	CEPH/ UTAH	CEU	NHGRI
NA12878 (rep 2)	HapMap	proband	female	CEPH/ UTAH	CEU	NHGRI
NA12891	HapMap	father	male	CEPH/ UTAH	CEU	NHGRI
NA12892	HapMap	mother	female	CEPH/ UTAH	CEU	NHGRI
HG02984	HapMap	proband	male	GB97	GWD	NHGRI
HG02982*	HapMap	father	male	GB97	GWD	NHGRI
HG02983	HapMap	mother	female	GB97	GWD	NHGRI
HG00514	HapMap	proband	female	SH032	CHS	NHGRI
HG00512	HapMap	father	male	SH032	CHS	NHGRI
HG00513	HapMap	mother	female	SH032	CHS	NHGRI
NA19240	HapMap	proband	female	Y117	YRI	NHGRI
NA19239	HapMap	father	male	Y117	YRI	NHGRI
NA19238	HapMap	mother	female	Y117	YRI	NHGRI
HG00731	HapMap	father	male	PR05	PUR	NHGRI
HG00732	HapMap	mother	female	PR05	PUR	NHGRI
HG00733	HapMap	proband	female	PR05	PUR	NHGRI
RGP_1A	RGP, Familial Trio	mother	female	n/a	n/a	n/a
RGP_1B	RGP, Familial Trio	father	male	n/a	n/a	n/a
RGP_1C	RGP, Familial Trio	proband	female	n/a	n/a	n/a
RGP_2A	RGP, Familial Trio	Mother	female	n/a	n/a	n/a
RGP_2B	RGP, Familial Trio	Father	male	n/a	n/a	n/a
RGP_2C	RGP, Familial Trio	proband	female	n/a	n/a	n/a

Normal	Tumor/Normal Pair	n/a	n/a	n/a	n/a	n/a
Tumor	Tumor/Normal Pair	n/a	n/a	n/a	n/a	n/a
Clinical Sample 1.1	WGS Coverage 1 of 3	n/a	n/a	n/a	n/a	n/a
Clinical Sample 1.2	WGS Coverage 2 of 3	n/a	n/a	n/a	n/a	n/a
Clinical Sample 1.3	WGS Coverage 3 of 3	n/a	n/a	n/a	n/a	n/a
Clinical Sample 2	Clinical	n/a	n/a	n/a	n/a	n/a
Clinical Sample 3	Clinical	n/a	n/a	n/a	n/a	n/a
Clinical Sample 4	Clinical	n/a	n/a	n/a	n/a	n/a
Clinical Sample 5	Clinical	n/a	n/a	n/a	n/a	n/a
Clinical Sample 6	Clinical	n/a	n/a	n/a	n/a	n/a
Clinical Sample 7	Clinical	n/a	n/a	n/a	n/a	n/a
Clinical Sample 8	Clinical	n/a	n/a	n/a	n/a	n/a
Clinical Sample 9	Clinical	n/a	n/a	n/a	n/a	n/a
Clinical Sample 10	Clinical	n/a	n/a	n/a	n/a	n/a

Table 2. Expected output by SMRT Cell type.

Sequel I data is based on prior working knowledge of sequencing Whole Genome Libraires and Sequel II data is based on described deliverable set forth by Pacific Biosciences.

	Sequel I	Sequel II
Total ZMWs on chip	1 million	8 million
Actual active ZMWs	600,000-800,000	4-6 million
Output in Gb	~10 Gb	200-300 Gb
Mean Read length	~10,000 bp	60,000-70,000 bp
N50 Read Length 50% of reads ≥ than this	~14,000 bp	130,000-150,000 bp
Bases sequenced per min	~24 bp per min	~80 bp per min

Dr. Kiran Garimella performed secondary analysis on this data utilizing the computational workflow he designed. He used this workflow to analyze the ability of the long read sequencers to detect structural variation and compare the output to Illumina-generated sequencing. He also did the analysis to determine if Pacific Biosciences Sequencing could uncover a structural variant that was eluding second generation sequencing attempts.

Process Improvements

Increasing Pacific Biosciences 10 kb CCS Library Yield

Standard operating procedure requires 5-8 ug of gDNA to enter the library production workflow utilizing SMRTbell Express Template Prep Kit V1. Yield from the standard protocol was found to be limited and allowed for very few SMRT Cell sequencing attempts on Sequel instruments (as little as one SMRT Cell, which equates to

~10X human coverage on Sequel II). Generally purposed to clean up small fragments from the elution (referred to as 'clean up'), the SPRI bead concentration (0.5X) may be too stringent, causing loss of yield within the desired size range. Library creation was followed by a gel-based fragment size selection on the BluePippin- the intended step in the workflow for size selection. Increasing SPRI concentration during cleanup prior to gel-based size selection may increase output of the library construction process.

To test this potential solution, four replicate libraries were created. Standard 0.5X SPRI cleanups were performed on two and 1.0X SPRI clean ups were performed on the remaining two aliquots. Each aliquot was quantified via Qubit fluorometric quantitation (Thermo Fisher Scientific) and size selection was evaluated using the Caliper LabChip GX (Agilent) after each SPRI clean-up step during library preparation: after post shearing clean up, after post end-repair clean up, after post adapter ligation clean up, and after size selection clean up.

SMRTbell Express Template Prep Kit V2.0 Validation

SMRTbell Express Template Prep Kit V1 was made obsolete in early 2020 and was no longer available for purchase. The new kit version (V2.0) and a software update had to be verified for performance in order for the CCS Library product offering to continue with the Genomic Platform. SMRTbell Express Template Prep Kit 2.0 claimed to offer shorter library preparation times (12 hours vs. ~3 days). The addition-only nature of the protocol was expected to increase yields without requiring additional input. Resulting libraries were examined to see if they met or exceeded our current in- house averages with respect to library yield, sequencing coverage, and ease of use.

Twelve 15kb CCS libraries were prepared and sequenced utilizing the standard operating procedure of SMRTbell Express Template Prep Kit 2.0, including updates to both reagents/chemistry (reducing preparation time from three days to 12 hours), fragmentation suggestions (10kb to 15kb – a separate analysis was completed for this), and sequencing workflow (in particular the pre-extension time increasing from 2 hours to 8 hours). A total of six reference libraries were chosen, with the majority having two replicates (one lacked a replicate due to sample exhaustion while one had three replicates). The six samples included two from the RM8392 family trio, one control genome from Family GB97, and a trio (child, mother, father) from Family ST116. All were well referenced genomes that would produce well-established results and would be useful for identifying any issues with updated protocol. Results were compared to historical 10 kb libraries in terms of primary output and quality of resulting run(s).

IsoSeq Validation

RNA libraries were created from a K-562 cell line using a standard Pacific Biosciences protocol (PN 101-763-800 Version 02). The total RNA input into reverse transcription and cDNA synthesis was 568 ng. After running the first part of the protocol with two ProNex bead cleanups, 237.7 ng remained. This was above the cutoff concentration recommended in the protocol (5.06ng/uL vs >3.5ng/uL). The "standard" ProNex bead purification (86 uL beads) was chosen for the second bead cleanup to target transcripts ~2 kb. After cDNA synthesis, fragment analysis was run on a BioAnalyzer (Agilent) to ensure the appropriate size was selected for. It peaked at slightly over 2 kb, which was to be expected with the standard size selection SPRI. A second library construction was completed on this cDNA using ProNex beads to size-select for small

(95 uL beads), medium (86 uL beads), and larger (82 uL beads) fragment sizes, the largest of which was ~3 kb and was used for a second validation. Both libraries were sequenced according to the standard Pacific Biosciences IsoSeq protocol and results were compared to expected standards outlined by our Pacific Biosciences Scientific field application scientist.

Chapter III.

Results

Over the course of many experiments and process validations, the Pacific Biosciences Sequencing product within the Broad Institute's Genomics Platform was optimized over time to produce a useful and reliable option for researchers to utilize for long read sequencing in their scientific investigations.

Pipeline Creation

A workspace in Terra to run a test of the pipelines for processing of long read data from Pacific Biosciences or Oxford Nanopore platforms was created and appropriate reference files were uploaded as workspace attributes. The pipelines were written in WDL 1.0 intended for use with Google Cloud Platform via the scientific workflow engine, Cromwell. Processing was designed to be reasonably consistent between both long read platforms and to use platform-specific options or tasks where necessary. The script LRWholeGenomeSingleSample.wdl, ran error correction, alignment, and variant discovery on >= 1 unit of data from the same sample. Initially, a 'tiny set' of data was used to test the efficacy of the pipeline integration into Terra, which was followed by a 'full set' of data (maximum output) to test the robustness of the function. After successful run of appropriately sized data, a test set was run through both the original script and the Terra-based function to prove equivalency.

Once the analysis was deemed acceptable, billing and a dedicated workspace were set up. Appropriate users created Terra accounts (supervisors, project managers, and lab personnel) and were added to the dedicated workspace. A training session was held so that the operations team could run pipeline rather than bioinformaticians. Unfortunately, due to the timing of the pandemic, the training occurred in February 2020 and direct ownership of the analysis pipeline has not yet been transferred at the time.

Size Selection Optimization

Optimizing fragment size output for preferred library types was an important aspect of the general improvement of long read sequencing results. The tighter the size distribution, the better control of sample input going into sequencing, resulting in higher quality data from the instrument.

Optimizing BluePippin Gel-Based Size Selection

The option of utilizing the BluePippin to select a range of fragments rather than a tight selection was ultimately deemed most appropriate for best results. Running the BluePippin in tight mode to collect 10 kb fragments proved to provide too little resulting sample output, where lower output is correlated to lower peaks during fragment analysis (Fig. 3). In order to balance the accuracy of size selection and output yield , an input range of 8-11 kb was deemed best, and the proper settings for this selection were identified to allow for a consistent output.

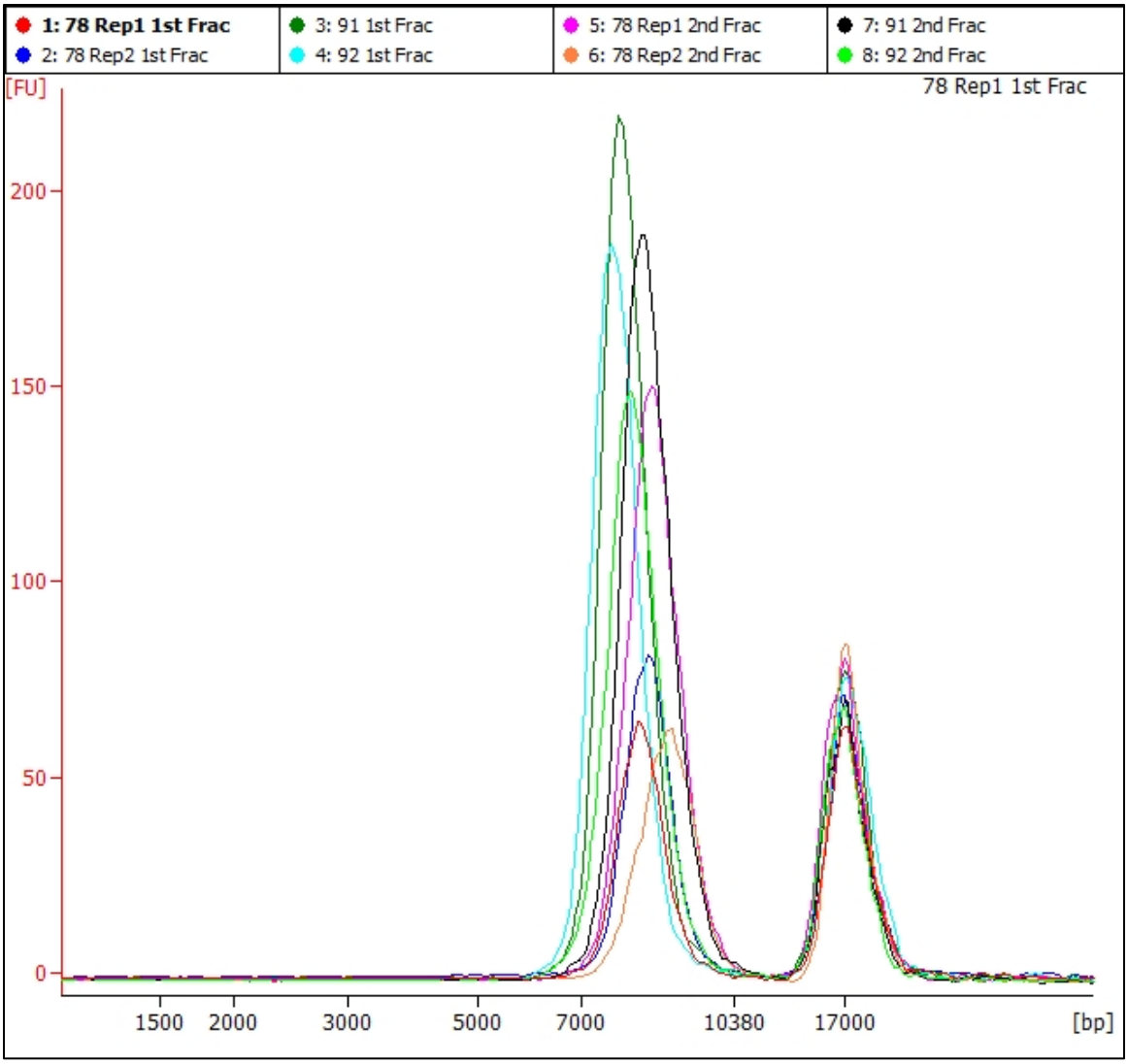


Figure 3: BluePippin Size Selection Caliper Results

Replicates of different fragmentation methods: 1,2, 5, 6 are 10 kb intended fragments, while 3, 4, 7, 8 are 9-11 kb range intended fragments.

Optimizing Output: Balancing Data Generated vs. Potential Quality Increase

The initial sequencing for the fragment size titration initial sequencing run experiment failed after the 9-10 kb library was sequenced and during the 8-9 kb library's sequencing event. The 8-9 kb library was not able to be re-sequenced due to lack of

remaining library. The 7-8 kb and 6-7 kb conditions were re-prepared for sequencing and run again. Resulted data is in Table 3. From this data, one can see that although the polymerase was able to read the insert more times for shorter fragments on average, the longer fragments were able to produce more data. The current condition (9-10 kb) was shown to produce more data in bases and unique molecular yield. With this result, repreparation of the 8 -9 kb condition was not pursued, as we had sufficient evidence supporting maintenance of the current protocol (9-10 kb fragmentation).

Table 3. Primary Sequencing Data Comparing Fragment Size Titration

Table 3 reveals the primary sequencing output of various input sizes into the CCS workflow.

Condition	Target Size	Total Bases (Gb)	Unique Molecular yield (Gb)	Mean polymerase read length	Mean subread length from sequencer (insert size)
Condition 1	9-10kb	176.58	24.32	56,090	8,397
Condition 2	8-9kb	n/a	n/a	n/a	n/a
Condition 3	7-8kb	142.33	14.91	61,684	6,656
Condition 4	6-7kb	107.16	10.36	61,984	5,848

Validation and Introduction of SageELF Size Selection

Twelve SageELF elution products were compared to determine the best 10 kb output condition (Table 4). With a goal of a 10 kb average fragment size, the 6th elution, which targeted 10,233 bases, resulted in a fragment measured at 10,093 bases (as measured on Caliper LabChip GX fragment analyzer). Elution products 7 and 8 also produced fragments within manufacturer recommended sizes that could move through the CCS process as rework or be used for deeper coverage requests.

Table 4. Caliper Output Data for SageELF Size Selection

Quantitative results of SageELF titration experiment. Note that for well 5 size was estimated based on marker size. Well 6 produced optimal results, while wells 7 and 8 produced usable, but less optimal fragmentation results.

Well Number	SageELF Software Target (basepair)	Measured size on 12k BioAnalyzer Run (basepair)
1	20997	too big to resolve
2	18288	too big to resolve
3	16965	too big to resolve
4	14863	too big to resolve
5	12424	Can see slight shoulder, estimated 12-14 kb
6	10233	**10093**
7	8327	8676
8	6878	7722
9	6312	6878
10	5130	6274
11	4083	4883
12	3268	4451

Fragmentation Optimization

The ability to fragment DNA with high consistency and quality was important to maintain high quality inputs and outputs of the long read sequencing workflow.

Megaruptor 2 vs. Megaruptor 3 for Fragmenting DNA

Diagenode Megaruptor 3 validation yielded acceptable results (see Figure 4 below). The Caliper LabChip GX gDNA QC run shows a tighter fragment size using the Megaruptor 3 than the Megaruptor 2 (A01). Additionally, products from the Megaruptor 3 are more centered around 10 kb rather than slightly larger, as observed in product from the Megaruptor 2. Looking at results based on the Megaruptor 3 speed settings, speeds 46 and 47 (B01 and C01 respectively) returned similar results whereas speeds of 48 and 49 (D01 and E01 respectively) yielded slightly shorter fragments. As speed settings of 46

and 47 yielded better results than the others and produced almost identical size distributions in our test, 46 was chosen arbitrarily between the two as the standard speed to fragment high quality DNA to 10 kb.

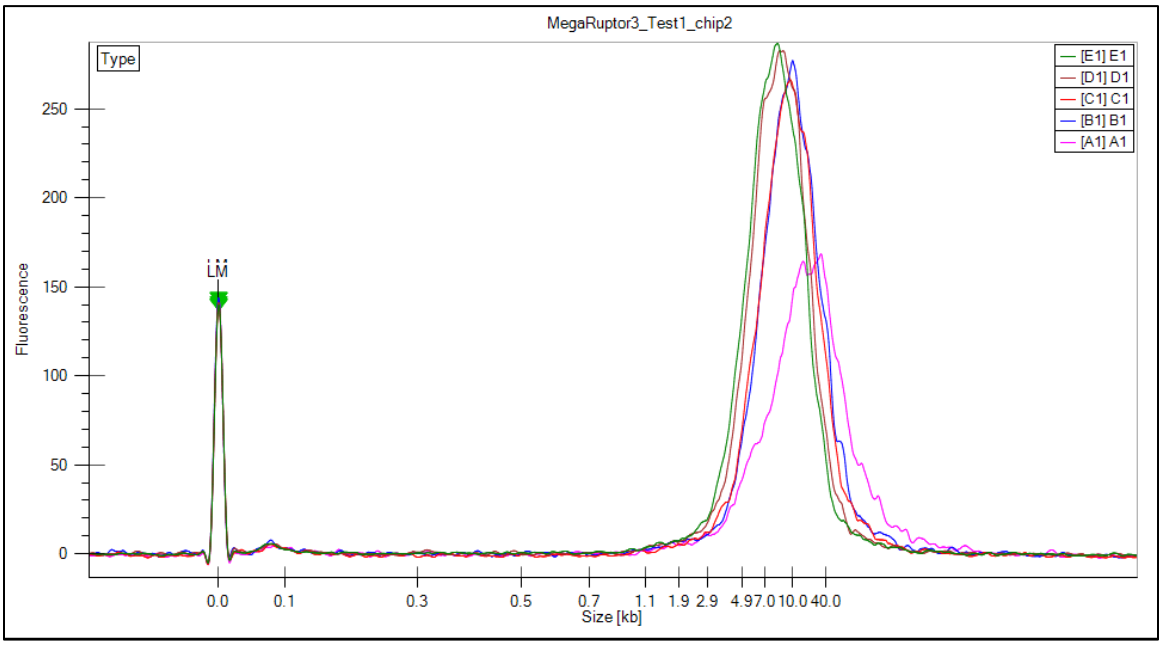


Figure 4. MegaRuptor 3 optimization for 10kb libraries.

Several speeds for fractionation were tested on the MegaRuptor 3 and analyzed using the Caliper.. A01 - DNA sheared on MegaRuptor 2 using the 10-15 kb setting. B01 - DNA sheared on MegaRuptor 3 using the 46 speed setting. C01 - DNA sheared on MegaRuptor 3 using the 47 speed setting. D01 - DNA sheared on MegaRuptor 3 using the 48 speed setting. E01 - DNA sheared on MegaRuptor 3 using the 49 speed setting.

For CLR libraries, results can be seen in Figure 5 below. The Megaruptor 2 control DNA (B01) showed fairly uniform fragmentation but it was slightly smaller than ideal. The Megaruptor 3 28 speed shear product (C01) was more spread out, indicating a less than optimal shearing speed, as a tighter shear slightly larger than 30 kb was

preferred. Product from the 30 speed shear setting (D01) was the closest to our desired results as it had a tighter distribution and had a large portion of DNA larger than 30 kb. Product from the 32 speed shear setting (E01) was sized as intended, but was more widely distributed than desired.

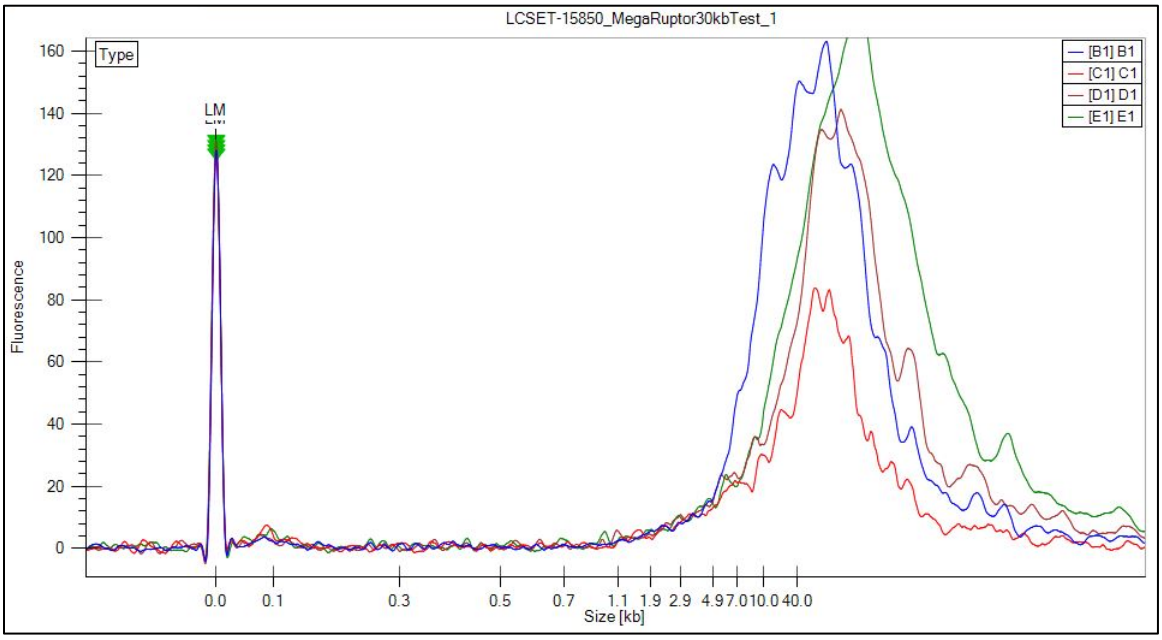


Figure 5. Megaruptor 30kb fragmentation optimization results

B01 - DNA sheared on the MegaRuptor 2 using the 30kb setting (control). C01 - DNA sheared on the MegaRuptor 3 using the 28 speed setting. D01 - DNA sheared on the MegaRuptor 3 using the 30 speed setting. E01 - DNA sheared on the MegaRuptor 3 using the 32 speed setting.

Fragmentation Alterations for Updated CCS HiFi Chemistry

Of the 12 CCS libraries produced (6 libraries, most with replicates), the average longest subread mean was 15,023 bases. The optimal shearing speed on the Diagenode Megaruptor was found to be 36 for 15-20 kb libraries.

Sequel 1 vs. Sequel II

The changes in instrumentation and the new ability to run 8M SMRT Cells allowed the Genomics Platform to pursue the use of Pacific Biosciences for deeper coverage projects, including whole genome sequencing.

Initial Primary Data Comparison

Data below in Table 5 includes a comparison of Sequel 1 and Sequel II validation testing runs: first 11 CCS HapMap trio Sequel II runs (10 kb insert, 1800 minute movies) compared to 19 tumor/normal paired Sequel I runs (15 kb insert, 600 minute movies). The Sequel II results in an immense increase in raw data generation compared to the Sequel I, as well as a significant increase in the speed with which data is generated Table 5). The control charts below display that although the two populations of runs by instrument type are distinct from each other in total based produced (Fig 6), count of high quality reads (Fig. 7), and speed of sequencing (Fig. 8), they are both providing data that was in control for that phase of the experiment. The output of high quality reads between Sequel I and Sequel II are quite different due to the chip types used on each instrument. The Sequel 1 running a 1M well chip has the capacity for 1M reads, and shows an average of 839,163 reads in this data set. The Sequel II runs with an 8M well chip has the capacity for 8M reads, and shows an average of 4,398,745 reads – >4M of which Pacific Biosciences has determined to be within specification. The speed with which the polymerase adds bases to the DNA template was innately important to how long the polymerase can read before losing efficiency. In this case, it can be seen that the Sequel I chemistry averaged 24.5 bases per minute as opposed to the Sequel II chemistry which

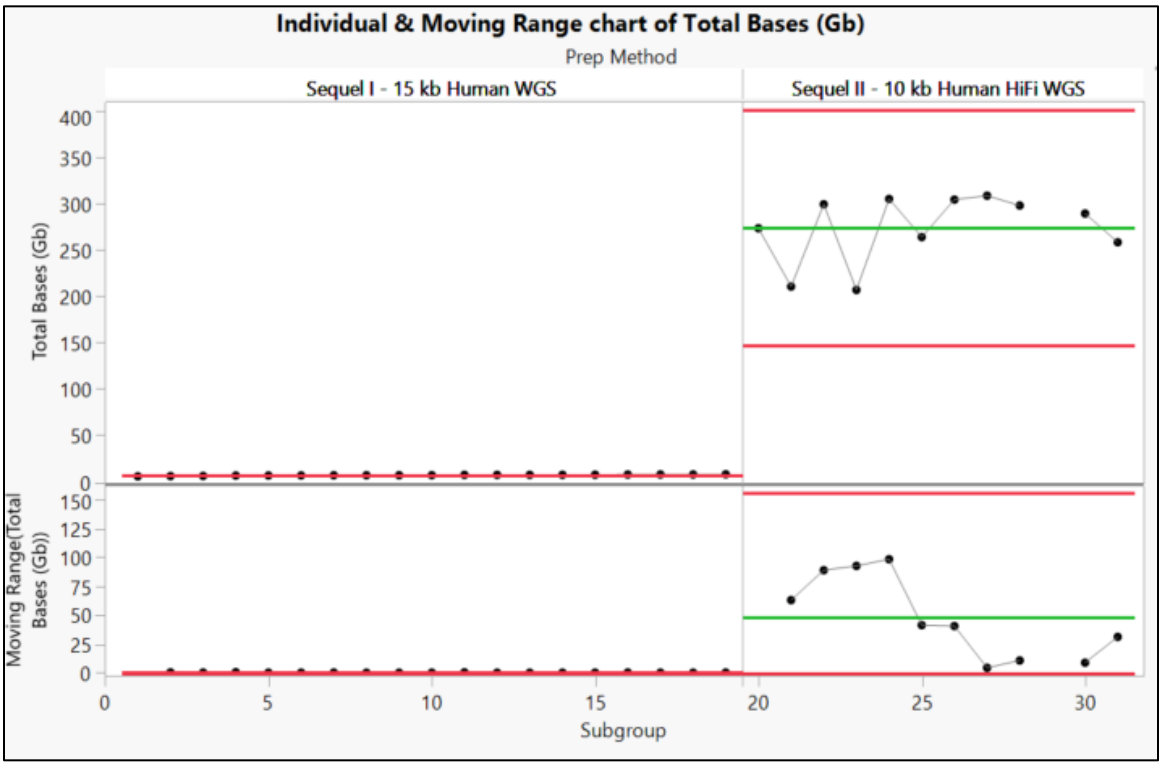
averaged 79.2 bases per minute (Fig 8). This allowed more data to be read over the same period of time.

Table 5. Data comparing Sequel I to Sequel II.

Sequel I instrument runs were from 15 kb Human WGS libraries. Sequel II runs were from 10kb Human HiFi (CCS) WGS libraries.

Sample Name	Prep Method	Total Bases (Gb)	Total "P1" High Quality Reads	Mean Read Length (bp)	N50 Read Length (bp)	N50 Insert Size (bp)	Read length bp/min seq time
K 18 tumor	Sequel I	6.2	870,319	7,133	12,250	11,250	20.42
K 18 tumor	Sequel I	6.4	876,757	7,388	12,250	11,250	20.42
K 18 tumor	Sequel I	6.6	876,980	7,542	12,750	11,250	21.25
K 18 tumor	Sequel I	7.0	855,952	8,237	13,750	12,250	22.92
K 18 tumor	Sequel I	7.1	799,551	8,909	14,750	12,750	24.58
K 13 normal	Sequel I	7.2	870,985	8,270	13,750	12,750	22.92
K 18 tumor	Sequel I	7.3	845,601	8,710	14,250	12,250	23.75
K 18 tumor	Sequel I	7.3	868,602	8,496	14,250	12,250	23.75
K 13 normal	Sequel I	7.4	851,398	8,707	14,750	13,250	24.58
K 18 tumor	Sequel I	7.4	847,360	8,879	14,250	12,750	23.75
K13 normal 10kb+	Sequel I	7.8	848,013	9,230	15,250	13,750	25.42
K 18 tumor	Sequel I	7.8	825,000	9,524	14,750	13,250	24.58
K 18 tumor	Sequel I	7.8	826,556	9,499	15,250	13,250	25.42
K 13 normal	Sequel I	7.8	840,573	9,386	15,750	13,750	26.25
K 18 tumor	Sequel I	7.9	806,951	9,939	15,750	13,250	26.25
K 18 tumor	Sequel I	8.1	798,303	10,269	15,750	13,750	26.25
K 13 normal	Sequel I	8.2	785,521	10,497	16,750	14,750	27.92
K 13 normal	Sequel I	8.2	831,860	10,021	16,750	14,250	27.92
K13 normal	Sequel I	8.4	817,820	10,357	16,750	14,750	27.92
NA12878 (rep 1)	Sequel II	274.0	4,883,086	56,307	140,531	9,893	78.07
NA12878 (rep 2)	Sequel II	211.1	3,733,589	56,884	151,047	9,535	83.92
NA12891	Sequel II	299.9	4,609,504	65,271	144,918	8,815	80.51
NA12892	Sequel II	207.4	4,464,489	46,623	125,810	8,624	69.89
HG02984	Sequel II	305.7	4,699,406	65,071	142,287	8,827	79.05

HG02983	Sequel II	264.5	4,352,652	60,808	141,090	9,394	78.38
HG00514	Sequel II	304.9	4,486,272	67,999	143,797	9,022	79.89
HG00512	Sequel II	309.1	4,584,823	67,454	143,454	8,988	79.70
HG00513	Sequel II	298.6	4,188,587	71,329	149,112	9,369	82.84
NA19239	Sequel II	289.9	4,299,834	67,450	145,164	8,377	80.65
NA19238	Sequel II	258.9	4,083,955	63,421	141,308	8,767	78.50

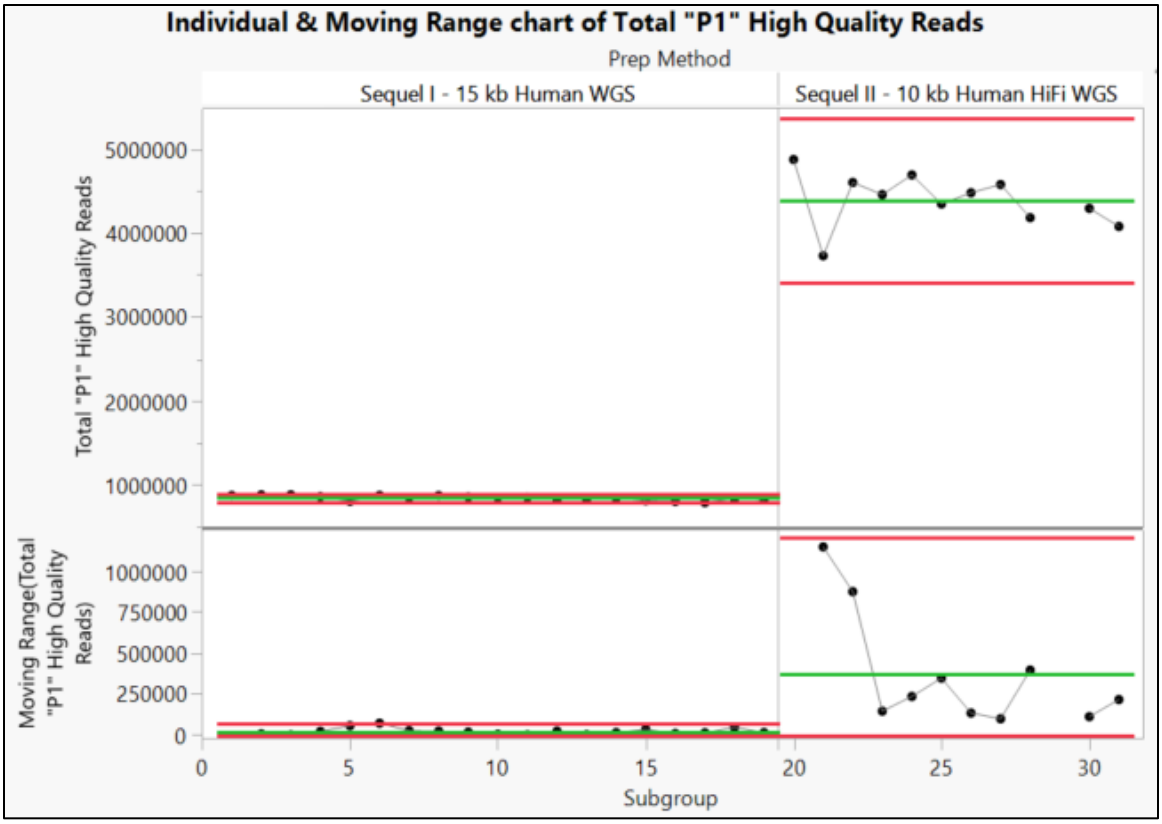


Total Bases (Gb) Limit Summaries

Points plotted	Prep Method	LCL	Avg	UCL	Limits Sigma	Subgroup Size
Individual	Sequel I - 15 kb Human WGS	7.1476	7.47438	7.80116	Moving Range	1
Individual	Sequel II - 10 kb Human HiFi WGS	147.6696	274.8955	402.1213	Moving Range	.
Moving Range	Sequel I - 15 kb Human WGS	0	0.122911	0.401491	Moving Range	1
Moving Range	Sequel II - 10 kb Human HiFi WGS	0	47.853	156.3134	Moving Range	.

Figure 6. Comparison of output (Gb) of Sequel I and Sequel 2 runs.

Sequel I versus Sequel II data in terms of the volume of data generation can be seen in a control chart of the data provided in Table 5.

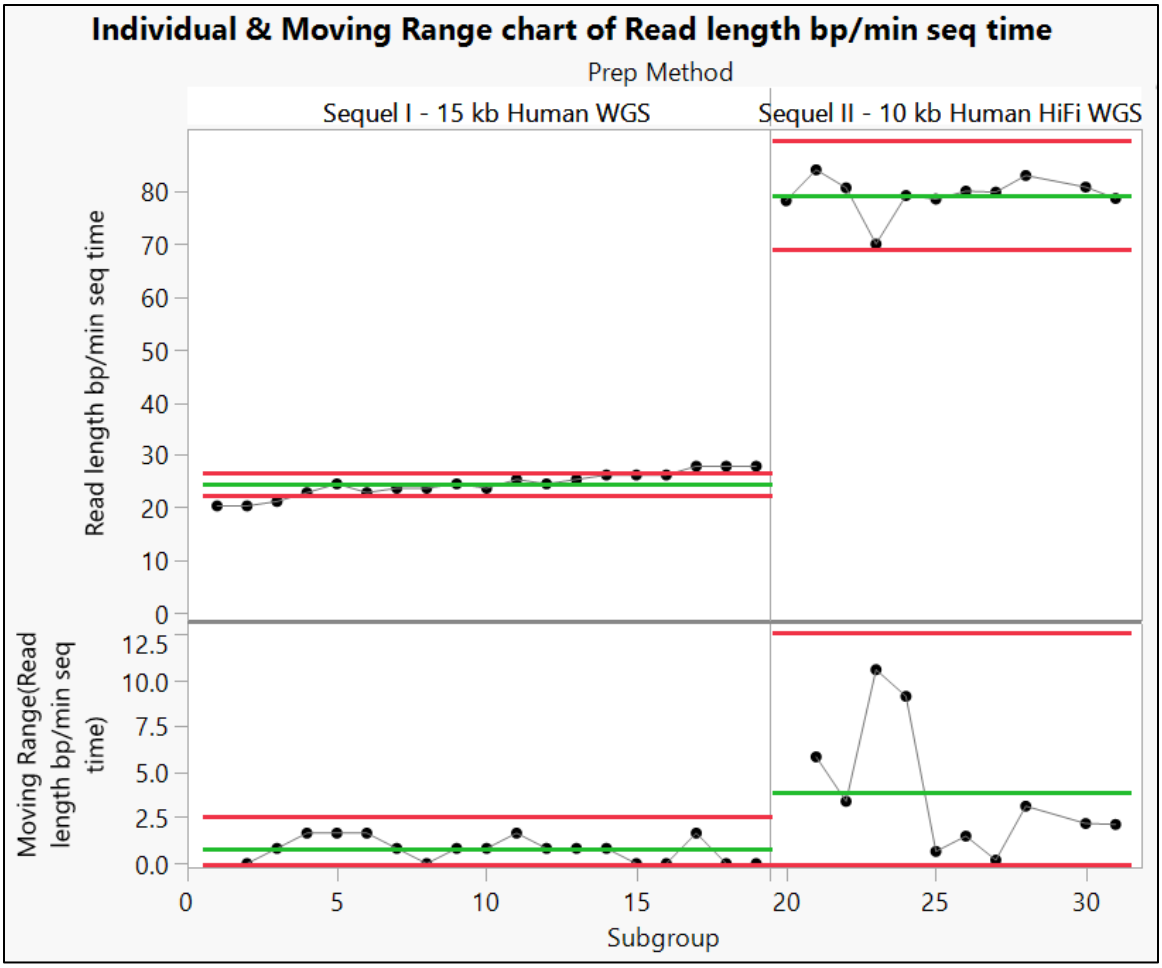


Total "P1" High Quality Reads Limit Summaries

Points plotted	Prep Method	LCL	Avg	UCL	Limits Sigma	Subgroup Size
Individual	Sequel I - 15 kb Human WGS	783062	839163.3	895264.5	Moving Range	1
Individual	Sequel II - 10 kb Human HiFi WGS	3413004	4398745	5384486	Moving Range	.
Moving Range	Sequel I - 15 kb Human WGS	0	21101.17	68927.63	Moving Range	1
Moving Range	Sequel II - 10 kb Human HiFi WGS	0	370763.1	1211110	Moving Range	.

Figure 7. Comparison of high quality reads output for Sequel I and Sequel II.

The output of high quality reads between Sequel I and Sequel II are quite different due to the chip types used on each instrument which can be seen in the number of high quality reads produced.



Read length bp/min seq time Limit Summaries

Points plotted	Prep Method	LCL	Avg	UCL	Limits Sigma	Subgroup Size
Individual	Sequel I - 15 kb Human WGS	22.44699	24.53947	26.63195	Moving Range	1
Individual	Sequel II - 10 kb Human HiFi WGS	68.88778	79.21808	89.54838	Moving Range	1
Moving Range	Sequel I - 15 kb Human WGS	0	0.787037	2.570882	Moving Range	1
Moving Range	Sequel II - 10 kb Human HiFi WGS	0	3.8855	12.69211	Moving Range	1

Figure 8. Comparison of the speed sequencing for Sequel I vs. Sequel II.

It can be seen that the Sequel I chemistry averaged 24.5 bases per minute as opposed to the Sequel II chemistry which averaged 79.2 bases per minute.

Table 6: Expected versus actual output of 8M SMRT Cell run on Sequel II.

***This data was based on first 11 HapMap samples produced and sequenced on 8M SMRT Cells as listed in Table 5 above.*

	Sequel II, SMRT Cell 8M (expected)	**Average for 34 Chip Validation Testing** (actual)**
Total ZMWs on chip	8 million	n/a
Actual active ZMWs	4-6 million	4,517,220
Output in Gb	200-300 Gb	275 Gb
Mean Read length	60,000 - 70,000 bp	62,602 bp
N50 Read Length 50% of reads ≥ than this	130,000 - 150,000 bp	142,593 bp
Bases sequenced per min	~80 bp per min	79.22 bp per min

Validation Testing on Sequel II 8M SMRT Cells

34 CCS sequencing runs on the 8M SMRT Cell on Sequel II were successfully completed (Table 7). From this data set, the average output in total bases was 241.5 Gb, with the maximum output of 394.2 Gb. The average high quality read count was 3,895,345, slightly below the expected 4M output. Also of note, the post-analysis genome coverage output was 8.7 Gb – ranging from 3 Gb on the least successful attempt to 16.8 Gb on the most successful (which correlates to the highest raw output in total bases).

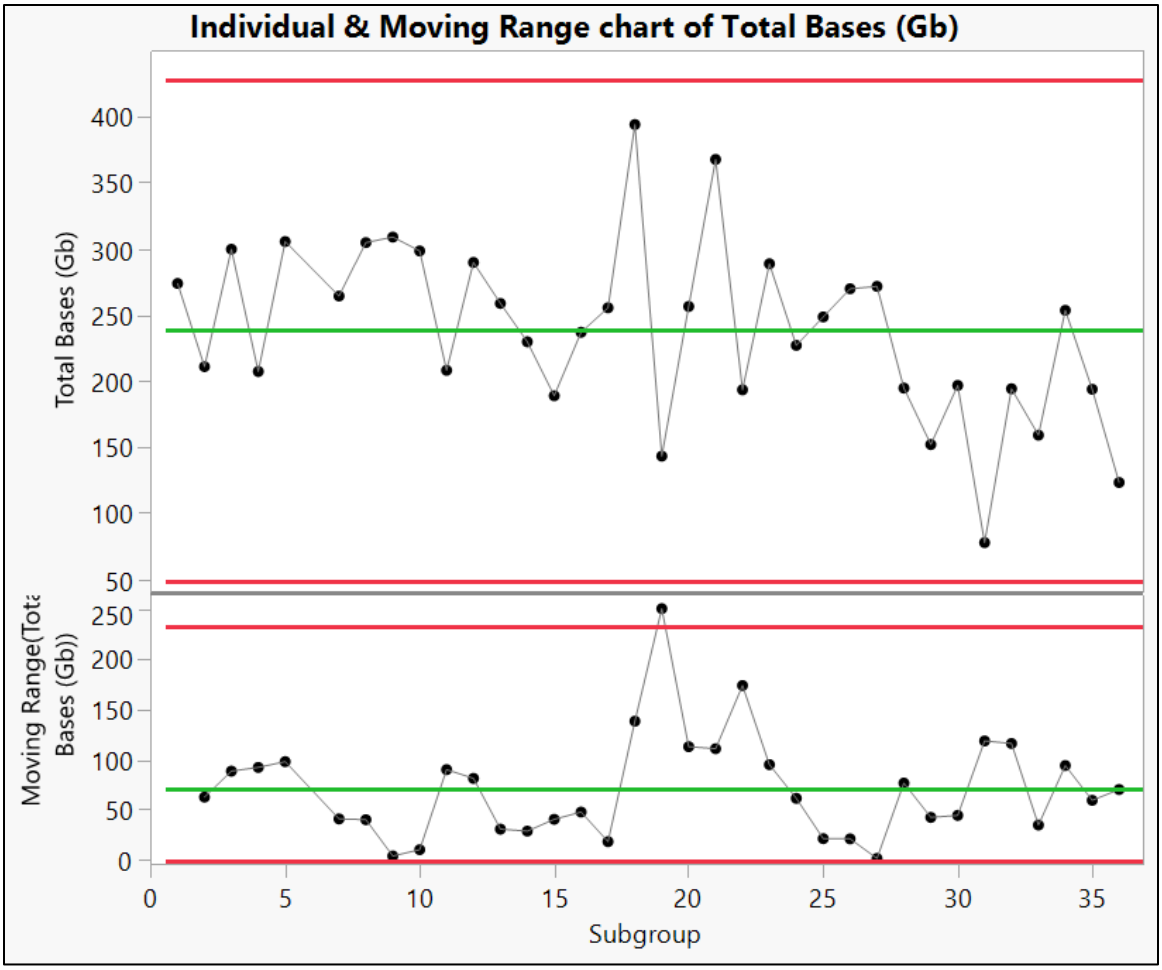
Table 7. Sequel II 8M SMRT Cell sequencing results.

36 SMRT Cells were run on the Sequel II using updated chemistry and the updated instrument itself. Majority were CCS libraries with one CLR. CS label implies Clinical Sample.

Sample ID	Total Bases (Gb)	Total "P1" High Quality Reads	Mean Read Length (bp)	N50 Read Length (bp)	N50 Insert Size (bp)	X Genome Coverage (post CCS)	Unique Molecular Reads (Gb)
NA12878 (rep 1)	274	4,883,086	56,307	140,531	9893	7.48	31.28
NA12878 (rep 2)	211.1	3,733,589	56,884	151,047	9535	5.70	23.88
NA12891	299.9	4,609,504	65,271	144,918	8815	9.19	38.11
NA12892	207.4	4,464,489	46,623	125,810	8624	6.49	n/a
HG02984	305.7	4,699,406	65,071	142,287	8827	10.75	42.33
HG02982	69.1	4,658,259	14,856	26,261	24809	n/a	n/a
HG02983	264.5	4,352,652	60,808	141,090	9394	8.65	34.63
HG00514	304.9	4,486,272	67,999	143,797	9022	10.20	41.14
HG00512	309.1	4,584,823	67,454	143,454	8988	11.97	41.18
HG00513	298.6	4,188,587	71,329	149,112	9369	10.65	42.79
NA19240	208.3	3,246,114	64,222	146,023	8715	6.53	24.78
NA19239	289.9	4,299,834	67,450	145,164	8377	10.11	38.26
NA19238	258.9	4,083,955	63,421	141,308	8767	8.73	35.02
HG00731	229.9	3,732,766	61,640	139,353	8867	7.82	28.22
HG00732	189.1	3,306,626	57,238	139,957	8235	5.65	21.99
HG00733	237.1	3,903,335	60,792	124,324	8406	7.98	29.53
RGP_1A	255.6	4,166,087	61,425	120,302	10269	11.27	37.85
RGP_1B	394.23	5,973,313	66053	132466	10101	16.81	19.98
RGP_1C	143.4	3,932,239	36474	75630	10047	n/a	19.86
RGP_2A	256.6	3715886	69130	146593	12417	11.42	n/a
RGP_2B	367.8	4709522	78143	155576	11083	15.91	n/a
RGP_2C	193.6	3020164	64154	140538	10363	7.12	26.50
Normal	288.9	4270005	67676	151598	10484	5.65	n/a
Tumor	227.2	3684861	61665	134216	9519	8.28	30.00
CS 1.1	248.7	3868287	64336	148993	11809	10.00	n/a
CS 1.2	270	4206340	64249	144902	11835	10.94	n/a
CS 1.3	271.8	4297813	63291	144953	11773	11.02	n/a
CS 2	194.9	3301538	59056	136263	10000	7.59	27.87
CS 3	152.2	2671407	56979	135352	10892	6.01	23.56

CS 4	196.9	2995383	65820	150075	9816	7.76	24.66
CS 5	78	1215518	64410	148881	9562	3.01	9.60
CS 6	194.4	3430517	56716	139200	10260	7.47	28.80
CS 7	159.3	2713199	58841	141499	10153	6.09	22.11
CS 8	253.7	3956225	64176	151808	9564	9.52	31.18
CS 9	194	4583566	42376	124053	10460	7.28	35.30
CS 10	123.5	3087073	40066	127613	10424	4.19	22.76

In the validation data set (Fig. 9 and Fig. 10), it can be seen that one flowcell mid-validation had a lower than average output and this aligns to a lower than average N50 read length. This corresponds to sample RGP_1C, which through the N50 Read Length metric, appears to have underperformed. Hypothetically, this could be due to DNA degradation prior to library construction which can impact the efficiency of the polymerase during sequencing. Figure 9 and 10 also shows that the total data produced during the validation sequencing events, outside of this particular sample, were all within statistical control.

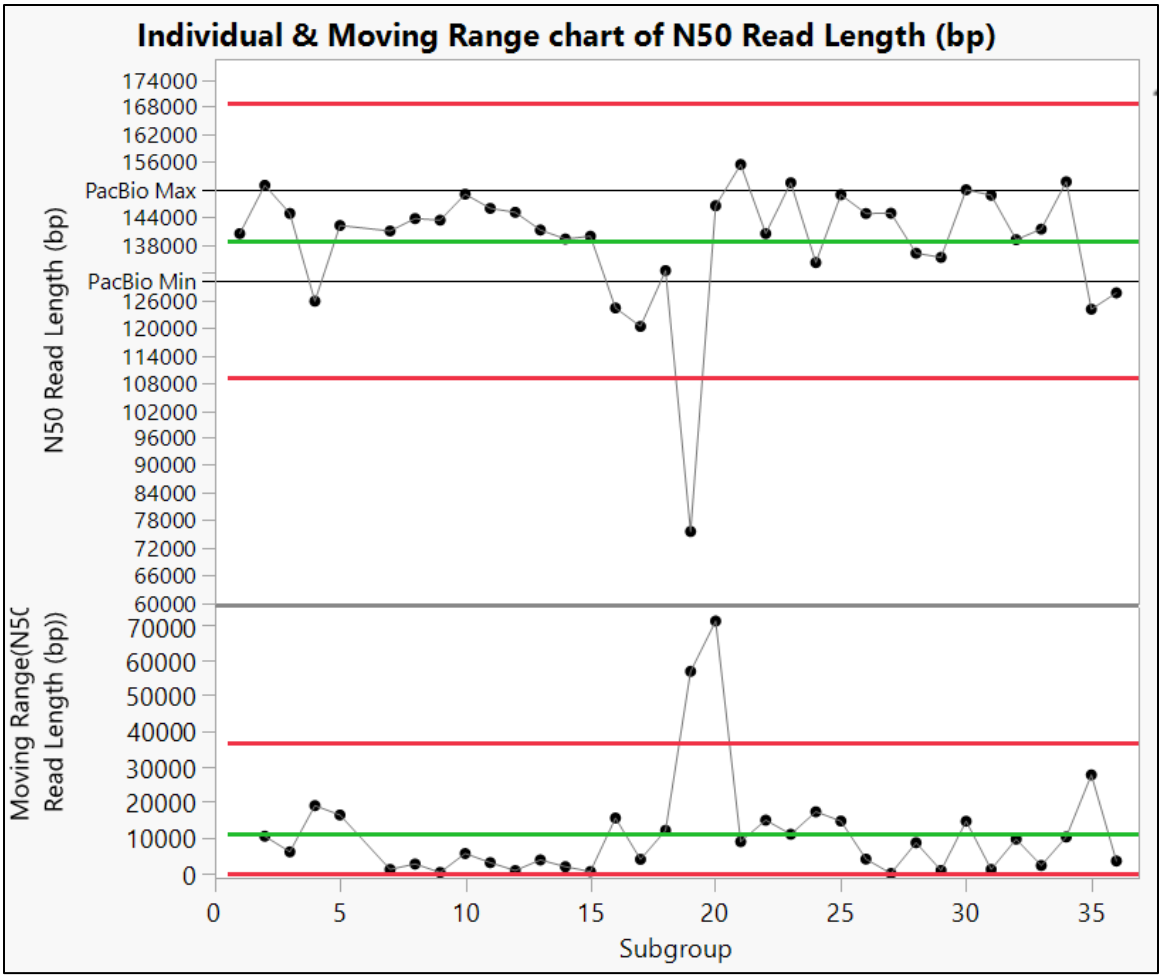


Total Bases (Gb) Limit Summaries

Points plotted	LCL	Avg	UCL	Limits Sigma	Subgroup Size
Individual	48.8811	238.6609	428.4406	Moving Range	1
Moving Range	0	71.38118	233.1689	Moving Range	1

Figure 9. Total bases produced (Gb) in the Sequel II validation experiment.

Total bases produced was a raw data metric that shows the potential for coverage from any given SMRT Cell.



N50 Read Length (bp) Limit Summaries

Points plotted	LCL	Avg	UCL	Limits Sigma	Subgroup Size
Individual	109185.8	139105.3	169024.9	Moving Range	1
Moving Range	0	11253.53	36760.01	Moving Range	1

Figure 10: N50 read length results from Sequel II validation experiment.

N50 Read Length shows the number of sequenced fragments that had this number of bases sequenced by a single polymerase at this length or greater.

Secondary Analysis Performed by Dr. Kiran Garimella

Dr. Kiran Garimella presented his secondary analysis of Sequel II output data on June 5, 2019. His data set included 32 SMRT Cells from Table 7 (16 HGSV available samples and 16 clinical samples – including trios and tumor/normal pairs).

In figure 11, the first row of data shows NA12878 uncorrected, which has an average of 10% error. The second row is NA12878 again, but now corrected via CCS analysis. This correction shows an average of 1% error, which is a similar error rate to second generation sequencing counterparts. Third and fourth rows in Fig, 11 are NA12891 corrected and NA12892 corrected (the parents). 94% of errors remaining in corrected data sets after CCS were shown to be 1 basepair (bp) indels. This data shows that uncorrected coverage per 8M SMRT Cell was ~142.5X raw coverage, while corrected coverage was 10X (as expected). When CCS analysis and error correction were performed, the abundance of resequencing of fragments allowed for removal of single subread errors when the majority of subreads are at consensus. This demonstrates that Pacific Biosciences has become much more comparable in terms of error rate to Illumina second generation sequencing applications.

Figure 11. Chromosome 6 of CEPH/UTAH Trio Visualized in IGV.

Figure provided by Dr. Kiran Garimella of data viewed in Integrative Genomics Viewer (IGV). Samples include (from top to bottom): NA12878 uncorrected, NA12878 corrected, NA12891 corrected, and NA12892 corrected.

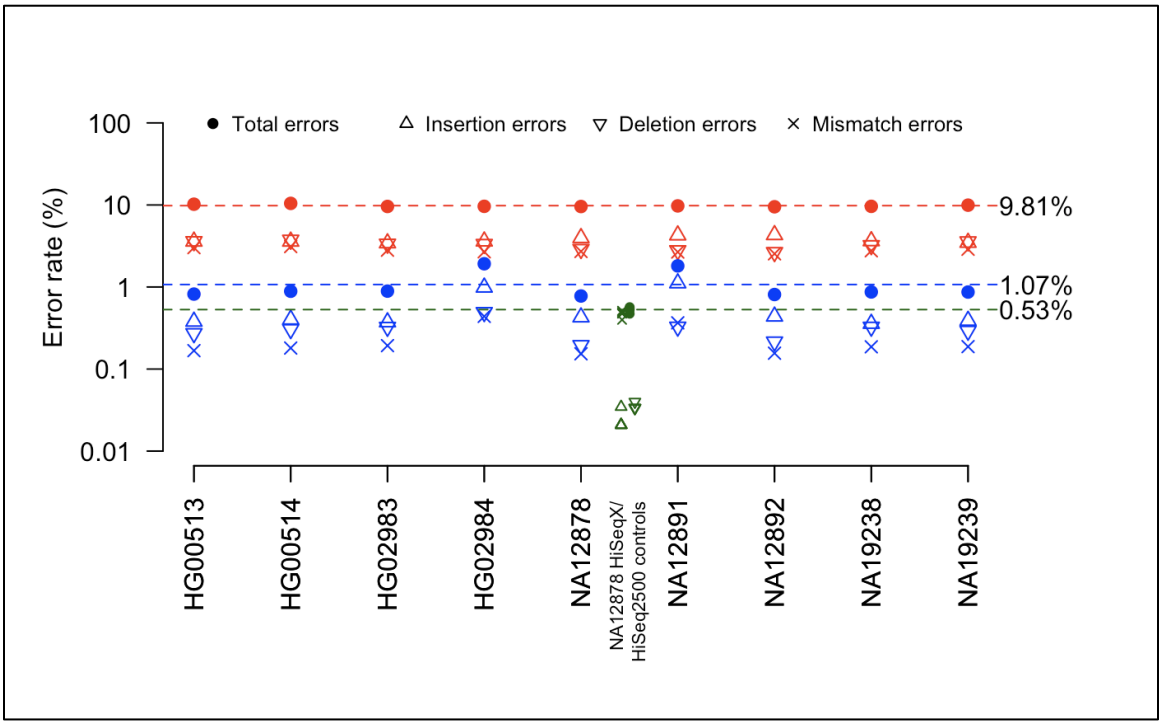


Figure 12. Comparison of Error Rates: CCS, CLR, and Second Generation Sequencing

Figure provided by Dr. Kiran Garimella. Blue corresponds to Pacific Biosciences generated CCS Library. Red corresponds to Pacific Biosciences generated CLR library. Green corresponds to second generation sequencing.

Comparison of Pacific Biosciences Sequel II and Illumina HiSeq X

Illumina Hiseq X and Pacific Biosciences Sequel II secondary data was directly compared in IGV (Integrative Genomics Viewer). Analysis of variation in PacBio data revealed a deletion that was much less apparent in the Illumina data. In Figure 13, the variant phase (blue/red verticals) in the HiSeq X data (A) appears unclear, while quite evident (red) in B, C, and E (PacificBiosciences data). The short reads produced by second generation sequencing led to a coverage drop, MQ0 reads, and mis-mapped reads. Using PacBio sequencing data, there is clear and consistent evidence for a deletion in

NA12878 replicates (B/C) and in the mother (E), while absent in the father (D). This deletion was not easily discernable in NA12878 sequenced on HiSeq X (A). Pacific Biosciences long reads allowed this deletion event to be detected when it was not previously observed with Illumina data, while also confirming Mendelian concordance.

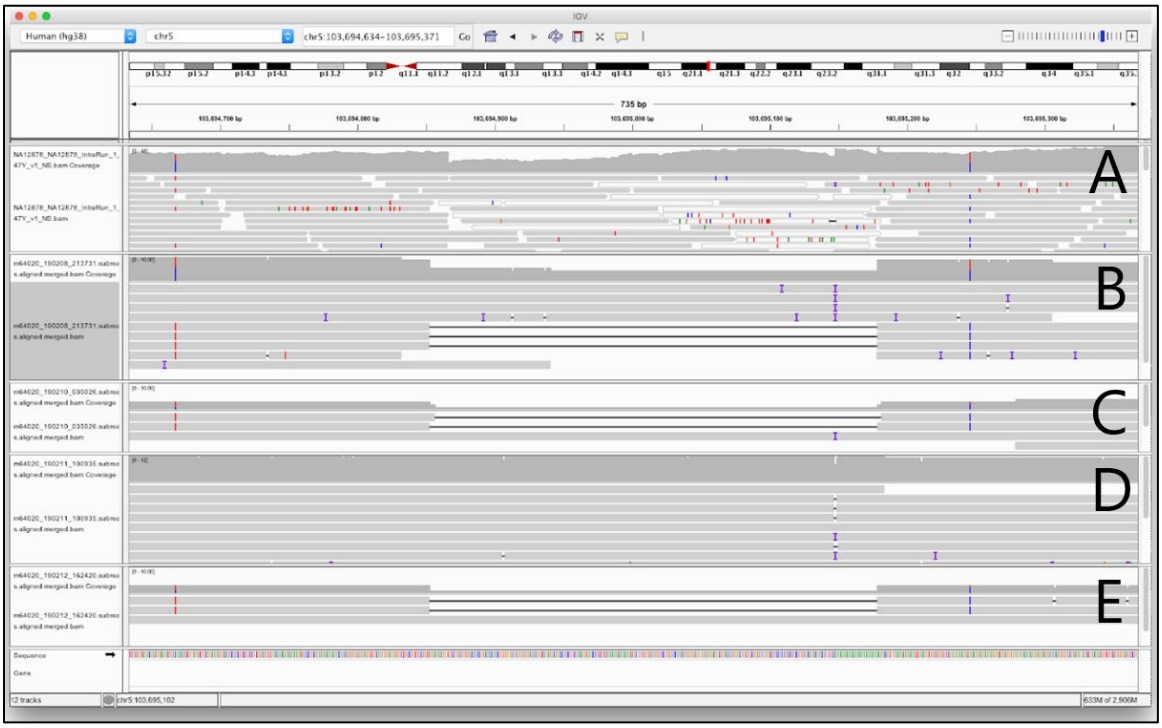


Figure 13 Pacific Biosciences vs. Illumina Structural Variant Detection on Chr5

Dr. Kiran Garimella provided an Integrative Genomics Viewer (IGV) comparison of chromosome 5 showing the CEPH/UTAH family trio as sequenced on the (A) Illumina HiSeq X and the Pacific Biosciences Sequel II (B, C, D, E). A, B, C show proband NA12878.D is NA12891, the father and E is NA12892, the mother.

Another case where long read sequencing demonstrated superior detection of variants was in this same proband, NA12878 on chromosome 1 (Fig. 14). Dr. Kiran

Garimella again provided IGV data from the NA12878 Trio, from Chromosome 1. In this figure, A, B and C represent NA12878 (the proband). A represents Illumina HiSeq X data while B and C represent Sequel II data. D represents NA12891 (the father) and E represents NA12892 (the mother), both from Sequel II data. There was no clear evidence of an inversion in the short read second generation sequencing data (A), but the inversion was clearly evident in the proband (B/C) and the mother (E) using Sequel II long reads that spanned this inversion event.

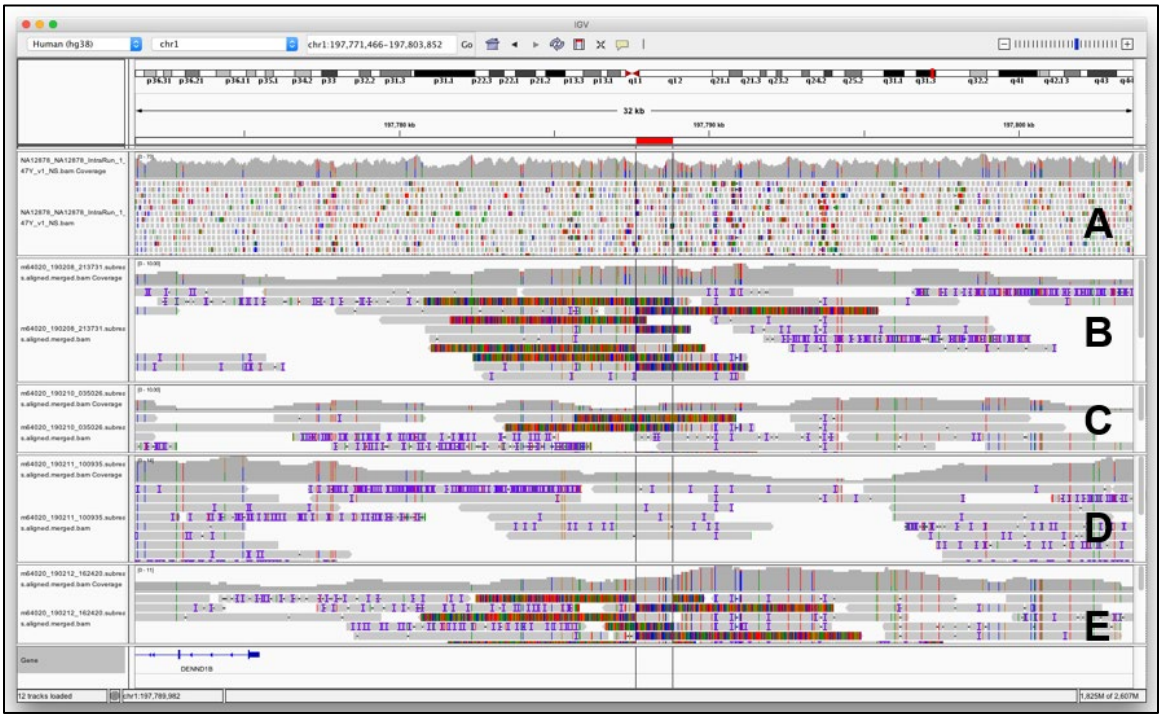


Figure 14. Pacific Biosciences vs. Illumina Structural Variant Detection on Chr1

Dr. Kiran Garimella provided IGV data of NA12878 Trio's Chromosome 1. A, B, C: NA12878 (proband) – A on Illumina HiSeq X and B/C on Sequel II. D: NA12891 (father) and E: NA12892 (mother) on Sequel II.

On Chromosome 6, at loci characterized by high genomic diversity (Fig. 15), sequencing of NA12878 on HiSeq X (top row, Fig. 15) showed inconsistent sequencing coverage while sequencing of NA12878 on Pacific Biosciences Sequel II (bottom row, Fig. 15) exhibited uniform coverage across this diverse region.

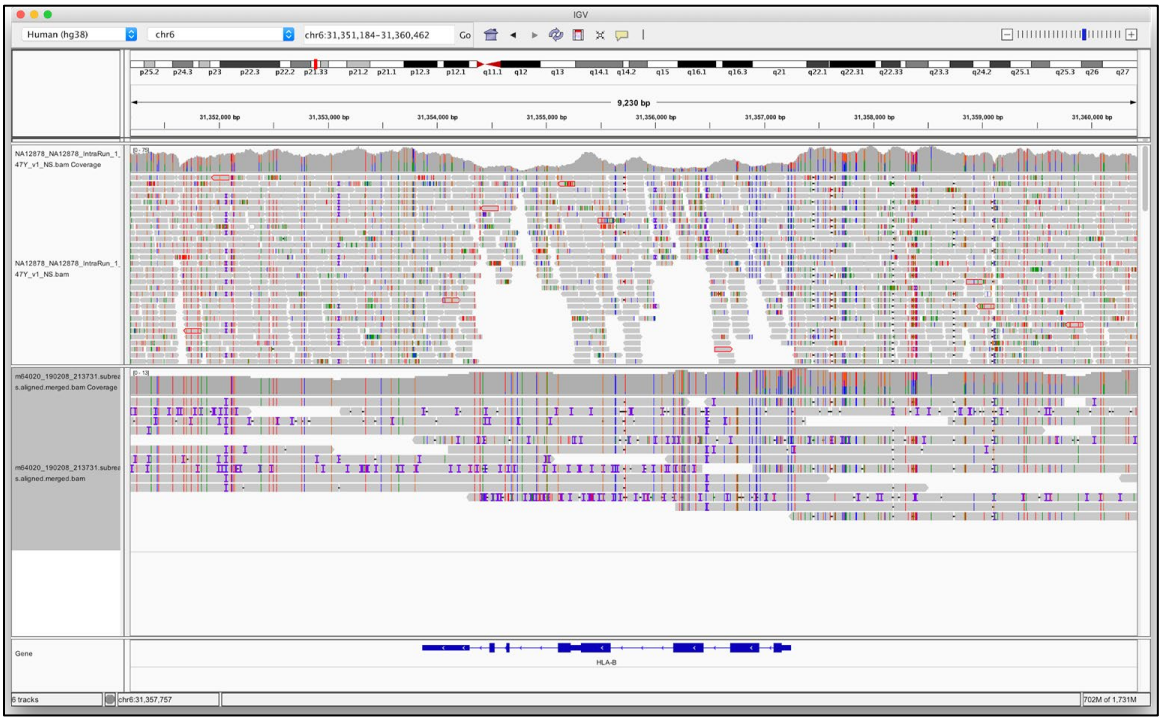


Figure 15. IGV coverage comparison for NA12878 HiSeq X and Sequel II data.

Figured provided by Dr. Kiran Garimella. The evenness of coverage is displayed for NA12878 across a portion of the genome with high genomic diversity (located on Chromosome 6).

Process Improvements

Ongoing process improvements are critical for continuously developing the highest quality and most efficiently prepared product offered by the Genomics Platform to our collaborators.

Increasing Pacific Biosciences 10 kb CCS Library Yield

It was hypothesized that library yields post-ligation may show a significant increase in yield (~5%) when 1.0X SPRI was used instead of 0.5X SPRI. Looking at Figures 16A and 16B, there was no significant difference in yield observed when comparing 0.5X and 1.0X SPRI cleanups in post-shear or post-end repair clean ups (end-repair occurs after fragmentation and before ligation). The accompanying Paired Student's t tests also showed no significant difference in yield. Figure 16C, however, shows that a statistical difference in yield was observed when comparing 0.5X and 1.0X SPRI for a post-ligation clean up. Comparing actual output, approximately 5% more library material was recovered with a 1X SPRI compared to the 0.5X SPRI recommended in the standard protocol. Furthermore, insert sizes did not appear to be affected by using the 1X SPRI, as an excess of low molecular weight material was not observed on BioAnalyzer (Agilent, 12k chip) when compared to a 0.5X SPRI cleanup (Fig. 17).

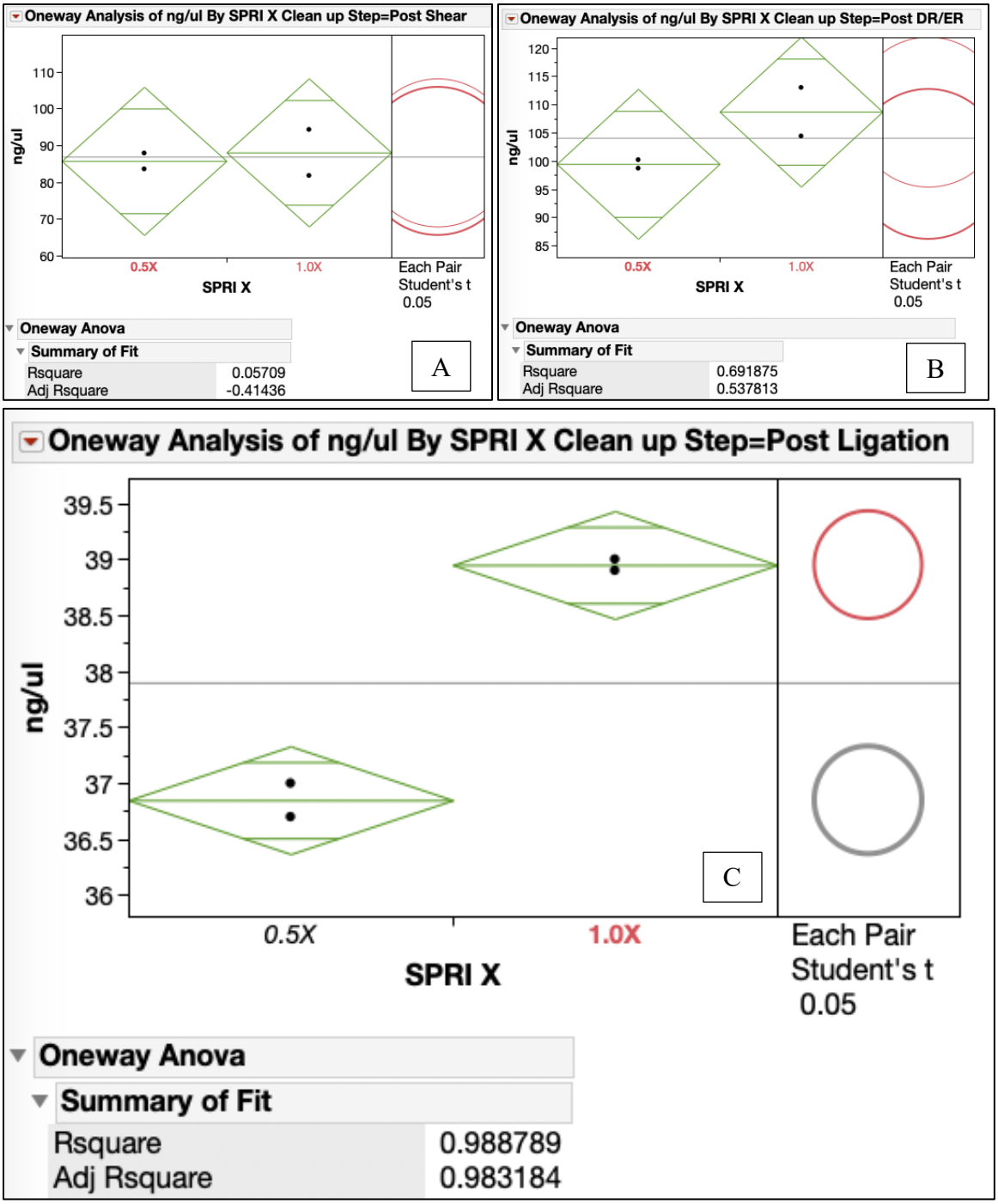


Figure 16. Comparison of yields (ng/uL) recovered from 0.5X (standard protocol) versus 1.0X SPRI cleanups at three different steps of library construction.

(A) shows comparison post-shearing. (B) shows comparison post-end repair (mid construction). (C) shows the comparison post-ligation.

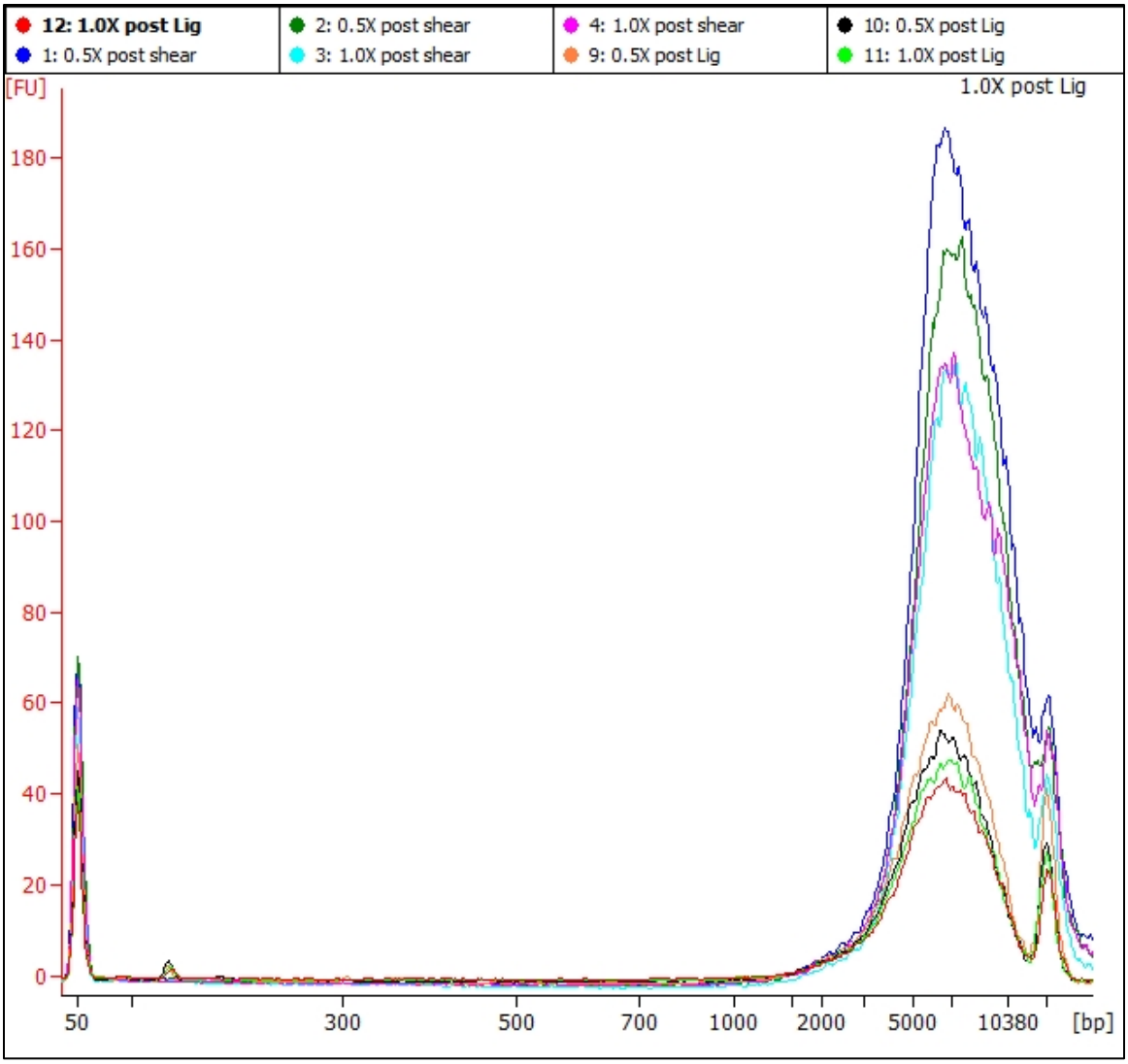


Figure 17: Agilent Bioanalyzer results of SPRI comparison.

1.0X and 0.5X SPRI product sizes were measured on Agilent BioAnalyzer. Post-shear and post-ligation products for 1.0X SPRI are in wells 3/4 and 11/12, respectively. Post-shear and post-ligation products for 0.5X SPRI are in wells in 1/2 and 9/10 respectively.

SMRTbell Express Template Prep Kit 2.0 Validation

When testing the efficacy of version 2 SMRTbell Express Template Prep kits, samples were run using both V1 and V2 kits. The averages of the V1 subset (n=14) and V2 subset (n=12) were calculated to compare yields and quality of runs.

Table 8: Comparison of Current (V1) vs. New (V2) Express Template Kit Outputs
This table details the average primary sequencing data output for V1 and V2 Libraries.

	V1 (n = 14)	V2 (n=12)
Insert Size (bp)	10000	15000
Movie Time (hours)	30	30
Immobilization time (hours)	default	2
Pre-Extension Time (hours)	12 or 2	8 or 3
Total Bases (Gb)	186.26	358.22
Polymerase RL (bp)	53561	88608
Polymerase N50 (bp)	129270	192527
Longest Subread (bp)	8625	15023
Longest Subread N50 (bp)	9426	15379
P0 %	54	47.5
P1 %	44.5	50.4
P2 %	1.51	2.17
Control Total Reads	4898.36	6466.75
Control Poly RL Mean (bp)	48107.00	46677.75
Control Concordance Mean	0.844	0.848
Control Concordance Mode	0.88	0.87
Local Base Rate	2.11	2.23

IsoSeq Validation

Coverage and quality of IsoSeq data met expectations based on experimental output compared to expected as set by the manufacturer. The successful test of the protocol (as seen in Table 9) confirmed our ability to meet industry equivalents in-house (within the Broad Institute Genomics Platform). Figure 18 shows that the 3300 (A) insert

performed more efficiently than that of the 2200 (B) based on the consistency of insert read length observed, regardless of HQ read length. Further work should be done to standardize a lab workflow, as there are different size-selection protocols that can be performed, depending on the quality and size specifications requested by collaborators for their specific research requirements. For now, this can be handled on a case-by-case basis.

Table 9: IsoSeq Library Preparation and Sequencing Output

After the initial Standard SPRI preparation was prepared, three more libraries were prepared with intent to gather smaller, medium or larger fragments based on SPRI concentration to identify the optimal SPRI concentration for IsoSeq library preparation. Two of these were chosen for sequencing.

Library	Beginning Material (ng)	Incoming LC ng/ul	Post SeqPrep Qubit (ng/ul)	Frag Size	Sequencing Yield (Gb)
Standard SPRI (86uL)	568	187	16.2 (8.6%)	2200	471.37
Small SPRI (95uL)	568	187	41.4 (22%)	2500	n/a
Medium SPRI (86uL)	568	187	26.2 (14%)	2800	n/a
Large SPRI (82uL)	568	187	16 (9%)	3300	403.96

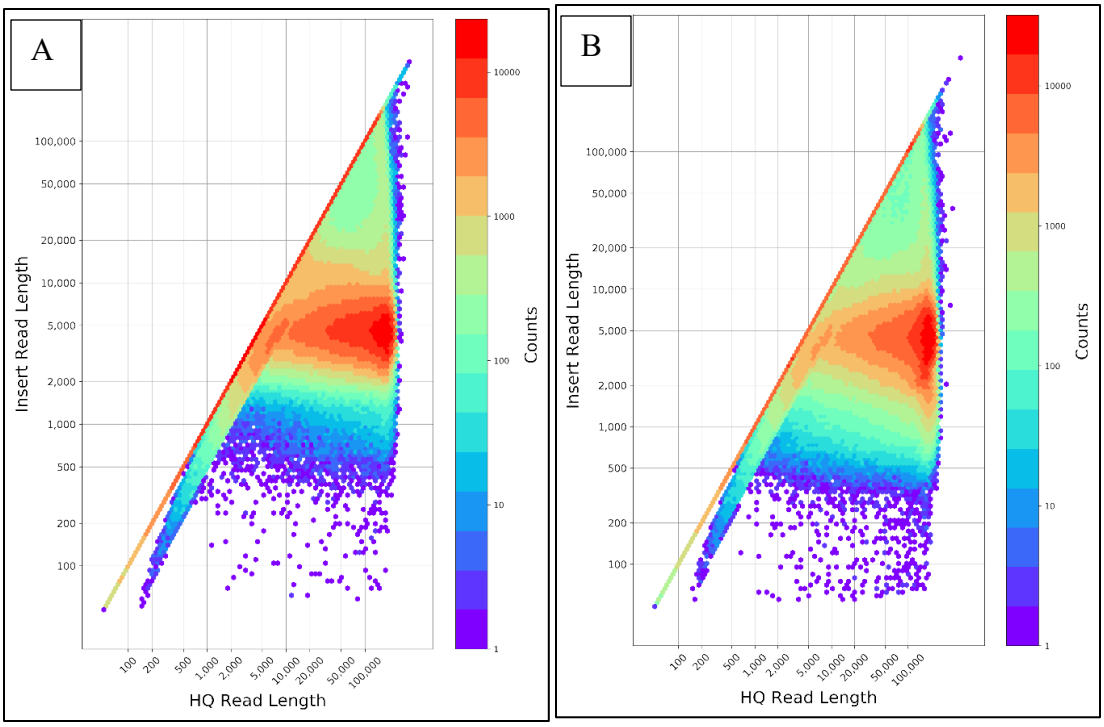


Figure 18. Insert quality (x-axis, HQ read length) plotted against insert read length (y-axis) for 3300 bp (A) and 2200 bp (B) IsoSeq libraries.

Chapter IV.

Discussion

A considerable amount of work was done to optimize the Pacific Biosciences workflow within the Broad Institute Genomics Platform. Pipeline updates were made to make the analysis more "user-friendly". Size selection and fragmentation workflows were optimized with new instruments. The sequencing instrument itself (the Sequel) was updated, as was the SMRT Cell capability. These, along with other process improvements drastically improved the output of the Pacific Biosciences sequencing capability within the platform, making it a more accessible option to all.

Pipeline Creation

A written WDL pipeline from a general cloud platform was successfully transitioned into Terra. Lab Operations and TAG team worked together to define and optimize the workflow including inputs and outputs of the analysis system and methods to deliver results to the end-user.

Although the end goal of delegating full analysis duties to the Operations team has not yet been completed, much work has been done to aid in this eventual effort. The pipeline was the first of its kind to successfully transition into Terra (the open-source, cloud based analysis platform) from the operations standpoint, which means that if so chosen, analysis function can be delegated to operations teams, which reduces the number of handoffs and transitions during operational workflow. This would potentially

reduce turnaround times and this model of using Terra may also prove beneficial to other lab processes on the Genomics Platform that could benefit from a more efficient analysis process within Terra.

Size Selection Optimization

Since reintroducing Pacific Biosciences Sequencing as a long read sequencing product into the Broad Institute Genomics Platform in 2018, the Sage Science BluePippin was used for size selection prior to constructing a sequence-able library. This was not the method recommended by Pacific Biosciences, but the demand for long read sequencing was moderate, so the protocol remained somewhat inconsistent during product introduction. As demand grew and the process moved from Sequel I to Sequel II, there was greater demand to make the process more precise. Initially, experiments were performed to identify the Blue Pippin conditions to precisely create fragments as closely sized to 10-11 kb as possible. As demand grew, fragment length optimization tests were carried out in order to enable sequencing error correction and increase coverage, without sacrificing structural variation. Finally, the Sage Science SageELF was acquired in order to acquiesce to manufacturer recommendations and provide a better product to customers by producing fragment selection in a faster and efficient manner. Each of these steps required continued testing and validation of the workflow to maintain quality and only introduce protocol changes that proved equivalent or superior to the current method.

Of the two methods tested to produce a tight size selection centered around 11 kb, the first mode, set to a range of 8-14 kb and centered at 11 kb, performed well to produce a tight selection centered around 11kb, with substantial output (ng) recovered after selection. The second mode tested was tight mode, centered at 11 kb (which should

include fragments +/- 20%). Tight mode performed adequately, but produced a wider range of size selection products and had lower overall yield recovered after the selection. The range method was advantageous and was chosen as the standard operating procedure moving forward..

To determine whether introducing shorter library fragments into sequencing would result in higher sequencing quality and coverage, three size conditions were tested. Of the three conditions tested, 9-10 kb, 7-8 kb, and 6-7 kb, the 9-10 kb far outperformed the others in terms of sheer total data produced (Gb). If these conditions had resulted in similar raw output, analysis of total coverage of the human genome, error rate after CCS analysis, SNV (single nucleotide variant) and structural variant calling metrics would have been assessed by a bioinformatician. As the sequencing output from the standard size range of 9-10 kb was so statistically distinct, further downstream analysis was abandoned and the initial hypothesis that shorter fragments may increase coverage was rejected. No change to the process was made based on this experiment.

Introduction of the Sage Science SageELF was intended to provide more accuracy in size-selection in addition to time savings. The BluePippin created just a single cut while the SageELF created multiple cuts, allowing for multiple selection products as options to move forward with into downstream library construction. Finally, it was determined that the SageELF allowed for a more consistent and controlled input into the library construction workflow, and, ultimately into sequencing. This allowed us to set realistic expectations regarding long read sequencing data output for our collaborators.

Fragmentation Optimization

The ability to fragment DNA with high consistency and accuracy is important to maintain high quality inputs and outputs for long read sequencing.. Implementation of the Diagenode Megaruptor 3 for both CCS and CLR libraries was important to maintain quality not only of the fragmentation, but also the overall quality of the sequencing data that is delivered to collaborators. From these experiments, it was determined that for CCS libraries, samples were best suited for fragmentation at a speed of 46 to obtain 10 kb DNA fragments and that for CLR libraries, samples were best suited for fragmentation at a speed of 30 to obtain 30 kb DNA fragments. Updating the CCS library protocol with the SMRTbell Express Template Prep Kit 2.0 to a 15 kb input into library construction was fairly straightforward at that point, and a speed of 36 was determined to be optimal for this size. The validation of the Megaruptor 3 allowed for a higher throughput system to be implemented, increasing lab capacity and efficiency.

Sequel I vs. Sequel II

A comparison of Sequel I and Sequel II revealed the many benefits and opportunities that came with the advent of Sequel II and its greatly increased output.

Comparison of Sequel I and Sequel II

Comparison of the primary data output of Sequel I and Sequel II revealed that the Sequel II is opening a lot of new-use cases for long read sequencing research. The simple ability to increase movie length from 600 minutes to 1800 minutes drastically increased sequencing output. This, in combination with the Sequel II's ability to utilize the 8M SMRT Cell, allows The Genomics Platform to be in a place to offer a truly important

technology to the genomic sequencing community as the turnaround time and cost become more equivalent to that of sequencing alternatives on the market.

Comparison of Sequel I vs II showed that although both instrument outputs were in statistical control for primary data output, the output of the Sequel II was considerably increased. The average output (Fig. 6) of the Sequel I (1M SMRT Cell, 600m movie) was 7.47 Gb of raw data. In comparison, the average output of the Sequel II (8M SMRT Cell, 1800 movie) was 274.89 Gb, amounting to 36 times more raw data). This pattern follows when comparing high quality reads (839 k vs. 4,398 k, Fig. 7). This aligns with the SMRT Cell size expectations (assuming it is optimal for >50% of available ZMW to load with single fragment as P1). One of the most important figures that highlights an important feature of Sequel II in enabling such increased output is Figure 8, showing sequencing speed, i.e. read length bp/min. The ability for the polymerase to bind and detect at such increased speeds allows for the Sequel II to outpace the Sequel I in yet another category.

Through this testing, SMRT Cell primary output was compared to that of the expectations set by the manufacturer, Pacific Biosciences. Testing aligned with all expectations – active ZMW, total output, mean read length, N50 read length, and bases sequenced per minute. Looking at the control charts for these metrics, all appeared to perform consistently and well within control limits. The sheer increase in output (both raw coverage and post-analysis whole genome coverage) allows this product to produce output that makes it comparable to alternative genomic sequencing methods.

Secondary Analysis and Illumina Comparison

Dr. Kiran Garimella performed secondary analysis of Sequel II output data which included 32 cells from Table 7 (16 HGSV available samples and 16 clinical samples – including trios and tumor/normal).

In analysis of the CEPH/UTAH trio, Dr. Garimella showed that the addition of CCS consensus analysis reduces error rate by an order of magnitude (~10% to ~1%) with remaining errors consisting mostly of single base pair indels. The error rate in this CCS data aligns with that of second generation sequencing performed on Illumina HiSeq X. Directly comparing data from HiSeq X and Sequel II, it was noted that short reads often lead to a coverage drop, MQ0 reads, and mis-mapped reads. In the case of NA12878, there was clear evidence of a deletion on Chromosome 5 in IGV (Integrative Genomics Viewer) when viewing Pacific Biosciences data. This deletion was unconfirmed in the HiSeq X data. Mendelian evidence of this deletion was also observed when comparing proband to mother. In the case of inversion, the advantage of Sequel II was seen in the CEPH trio on Chromosome 1, as an inversion was not at all evident in HiSeq X data but was evident in Sequel II data. Comparing regions that are traditionally difficult for short fragment sequencing to cover due to mis-mapping (high GC and repeat regions, for example), it was seen in the genomic loci covered in Chr1 that the Pacific Biosciences Sequencing produced an excellent evenness of coverage while the region was problematic for HiSeq X (Fig. 15), demonstrating the importance of long read sequencing for improved sequencing coverage of these difficult regions..

Although Illumina sequencing remains at the forefront of genomic sequencing efforts due to cost and ease of use, the Pacific Biosciences Sequel II and 8M SMRT Cell

are becoming less and less costly and are providing valuable solutions for second generation sequencing problems, possibly making Pacific Biosciences a true market contender in the near future.

Process Improvements

Each small process improvement leads to the achievement of the overarching goal of this project; to optimize the workflow of human genome sequencing applications on Pacific Biosciences Sequel II and to make the workflow comparable in efficiency and cost to that of the industry standard.

In the experiment to test an increase yield by altering SPRI bead concentration, a 5% increase in final yield (post ligation) was seen when doubling the SPRI bead ratios from 0.5X to 1.0X. Output was 39 ng/L (1287 ng yield) with 1X vs. 37 ng/L (1221 ng yield) with 0.5X, or a total of 66 extra ng in yield. However, this increase is quite modest and the cost of SPRI reagents are high, putting into question whether this change is worthwhile. This experiment had a low # of samples, but the small difference did appear consistent between replicates and was statistically significant. A final verification of this result would be quite costly as it would require sequencing multiple samples to ensure that there wasn't a skewing of insert sizes or introduction of excess adapter (thus reducing overall sequencing coverage). In the end, the proposal of changing from a 0.5X to 1.0X SPRI was chosen to not be implemented in order to balance cost with potential gains in yield.

Comparing the SMRTbell Express Template Prep workflow version's direct sequencing outputs, it was seen that the new chemistry increased total bases sequenced largely, in part, due to the polymerase's capacity for longer reads, allowing for larger

insert sizes (15 kb vs. 10 kb), which, theoretically could allow for a higher number of unique molecules sequenced. Additionally, this new chemistry dramatically decreased the time to prepare libraries and allowed the team to reduce turnaround times for the library construction portions of the process from 3 days (including several overnight incubations) to 12 hours. Overall, the SMRTbell Express Template Prep Workflow version upgrade resulted in an increase in overall yield and general sequencing quality.

The IsoSeq validation experiment yielded adequate data to conclude that quality and data metrics met initial expectations. Although the output of sequencing preparation was less than the intended 30% yield suggested by Pacific Biosciences, the output of this protocol was sufficient to run test-sequencing across two types of libraries – one size selected for 2200 bp and one selected for 3300 bp. The insert length versus read length plots (provided by the Pacific Biosciences SMRT Link run data, Fig. 18) showed that the 3500 bp fragment may have been slightly better selected than the 2200 bp fragment. Both libraries appeared to have properly applied adapters and normal sequencing polymerase activity in terms of early termination events. In conclusion, coverage and quality of data met industry standard expectations for IsoSeq libraries. Further work may be done to optimize the workflow in the lab as there are varying size selections one can perform based on incoming quality requirements and collaborator interests. Overall, it was confirmed that this in-house IsoSeq preparation meets industry equivalents.

Future Impact of Long Read Sequencing

Due to limitations in previous versions of the long-read sequencing capability, it has mostly been used for microbial and highly de novo applications. The introduction of the Sequel along with the 8M SMRT cell created an opportunity to research the best ways

to run and optimize the instrument, as well as to discover new improvements to genome sequencing research that long reads can add. One area of interest for the application of continuous long read sequencing is single patient studies for personalized medicine. Long read sequencing could be used in defining whether there are any variants unique to a disease or to individuals. Long read sequencing is particularly valuable as it can uncover previously undetectable variants in historically difficult-to-characterize regions of the genome. Without the limitations of second generation sequencing, hard to sequence areas (for example GC-rich areas or tandem repeats) are now available to sequence for comparison across populations.

This work has demonstrated that a Pacific Bioscience long-read derived CCS human genome compares in quality, accuracy, and cost to that produced by the industry's current standard methodology – Illumina Sequencing-by-Synthesis. There are many areas of study that have been hindered by the limitations of second generation sequencing but have been made possible by third generation sequencing options including performing de novo sequencing efforts, understanding secondary structure implications, and exploring the importance of repetitive regions within the genome. Long read sequencing will allow for many new critical studies to be performed and for the broadening of existing data sets.

Appendix 1.

Sequel I vs. Sequel II Comparison Sample List

Table 10: Sequel I/II Comparison Sample List

Sample Name	Sequencing Instrument	Preparation Method	Movie Time
K 18 tumor	Sequel I	CCS 15kb Library	600m
K 18 tumor	Sequel I	CCS 15kb Library	600m
K 18 tumor	Sequel I	CCS 15kb Library	600m
K 18 tumor	Sequel I	CCS 15kb Library	600m
K 18 tumor	Sequel I	CCS 15kb Library	600m
K 13 Normal	Sequel I	CCS 15kb Library	600m
K 18 tumor	Sequel I	CCS 15kb Library	600m
K 18 tumor	Sequel I	CCS 15kb Library	600m
K 13 Normal	Sequel I	CCS 15kb Library	600m
K 18 tumor	Sequel I	CCS 15kb Library	600m
K13 Normal 10kb+	Sequel I	CCS 15kb Library	600m
K 18 tumor	Sequel I	CCS 15kb Library	600m
K 18 tumor	Sequel I	CCS 15kb Library	600m
K 13 Normal	Sequel I	CCS 15kb Library	600m
K 18 Tumor	Sequel I	CCS 15kb Library	600m
K 18 tumor	Sequel I	CCS 15kb Library	600m
K 13 Normal	Sequel I	CCS 15kb Library	600m
K 13 Normal	Sequel I	CCS 15kb Library	600m
K13 Normal	Sequel I	CCS 15kb Library	600m
NA12878 (rep 1)	Sequel II	CCS 10kb Library	1800m
NA12878 (rep 2)	Sequel II	CCS 10kb Library	1800m
NA12891	Sequel II	CCS 10kb Library	1800m
NA12892	Sequel II	CCS 10kb Library	1800m
HG02984	Sequel II	CCS 10kb Library	1800m
HG02983	Sequel II	CCS 10kb Library	1800m
HG00514	Sequel II	CCS 10kb Library	1800m
HG00512	Sequel II	CCS 10kb Library	1800m
HG00513	Sequel II	CCS 10kb Library	1800m
NA19239	Sequel II	CCS 10kb Library	1800m
NA19238	Sequel II	CCS 10kb Library	1800m

Appendix 2.

SMRTbell Express Template Preparation Kits: V1 and V2

Table 11: Primary Data by SMRT Cell: SMRTbell Express Template Prep kits: V1

Sample Name	**A**	**B**	**C**	**D**	**E**	**F**	**G**
Pre-Extension Time (h)	12	12	12	12	2	2	2
Total Bases (Gb)	237.08	255.58	161.44	143.35	142.33	107.16	172.31
Polymerase RL (kb)	60.79	61.43	33.07	36.47	61.68	61.98	44.17
Polymerase N50 (kb)	129.32	120.30	76.10	75.63	158.60	158.32	131.28
Longest Subread (bp)	8,387	10,041	8,850	8,897	7,256	7,004	6,660
Longest Subread N50 (bp)	8,406	10,269	10,637	10,047	7,773	6,806	7,559
P0 %	49.40	46.60	34.90	48.00	70.70	78.10	49.40
P1 %	48.70	52.00	60.90	49.10	28.80	21.60	48.80
P2 %	1.91	1.44	4.20	2.90	0.49	0.32	1.79
Control Total Reads	3,488	5,250	1,992	1,913	2,806	2,416	13,189
Control Poly RL Mean (kb)	47.66	47.34	30.99	31.21	56.55	55.61	50.63
Control Concordance Mean	0.85	0.85	0.85	0.85	0.84	0.83	0.84
Control Concordance Mode	0.89	0.89	0.89	0.87	0.87	0.87	0.87
Local Base Rate	2.13	2.15	2.03	2.16	2.27	2.21	2.05

Sample Name	H	I	J	K	L	M	N
Pre-Extension Time (h)	2	2	2	12	2	2	2
Total Bases (Gb)	240.26	116.89	223.92	256.65	159.30	194.40	196.92
Polymerase RL (kb)	49.42	39.03	51.31	69.13	58.84	56.72	65.82
Polymerase N50 (kb)	127.95	122.47	132.45	146.59	141.50	139.20	150.08
Longest Subread (bp)	8,525	7,717	8,278	11,786	8,943	9,303	9,097
Longest Subread N50 (bp)	9,402	9,204	9,216	12,417	10,153	10,260	9,816
P0 %	37.10	61.50	43.60	52.40	65.70	56.40	62.10
P1 %	60.70	37.50	54.60	46.40	33.90	42.80	37.40
P2 %	2.19	1.00	1.82	1.29	0.47	0.81	0.49
Control Total Reads	1,686	10,039	10,115	3,372	5,945	2,838	3,528
Control Poly RL Mean (kb)	57.19	45.31	52.34	48.59	47.30	50.41	52.36
Control Concordance Mean	0.84	0.84	0.84	0.85	0.84	0.84	0.84
Control Concordance Mode	0.87	0.87	0.87	0.89	0.87	0.87	0.87
Local Base Rate	2.07	2.00	2.09	1.99	2.15	2.08	2.18

Primary output data used to calculate V1 kit averages in Table 8.

Table 12: Primary Data by SMRT Cell: SMRTbell Express Template Prep kits: V2

Sample Name	**A**	**B**	**C**	**D**	**E**	**F**
Pre-Extension Time (h)	8	8	8	8	3	8
Total Bases (Gb)	352.41	366.97	338.51	363.42	366.04	323.06
Polymerase RL (kb)	86.91	84.31	89.52	90.58	87.01	86.11
Polymerase N50 (kb)	195.94	187.64	197.00	197.89	182.51	194.41
Longest Subread (bp)	14,282	15,214	15,308	15,573	14,446	17,092
Longest Subread N50 (bp)	14,673	15,387	15,253	15,458	14,889	17,928
P0 %	47.40	43.20	51.10	47.90	44.80	51.30
P1 %	50.60	54.40	47.30	50.10	52.60	46.90
P2 %	2.00	2.48	1.68	1.96	2.61	1.85
Control Total Reads	3,975	4,942	5,117	5,324	10,912	4,189
Control Poly RL Mean (kb)	37.26	44.38	43.16	42.25	56.20	41.49
Control Concordance Mean	0.85	0.85	0.85	0.85	0.84	0.85
Control Concordance Mode	0.87	0.87	0.87	0.87	0.87	0.87
Local Base Rate	2.24	2.19	2.24	2.27	2.26	2.17

Sample Name	**G**	**H**	**I**	**J**	**K**	**L**
Pre-Extension Time (h)	8	8	3	3	3	3
Total Bases (Gb)	241.60	301.63	366.04	452.11	387.79	439.07
Polymerase RL (kb)	86.57	82.58	87.01	93.54	93.61	95.54
Polymerase N50 (kb)	196.06	184.47	182.51	196.99	199.67	195.23
Longest Subread (bp)	16,532	16,105	14,446	14,188	13,430	13,657
Longest Subread N50 (bp)	17,460	16,439	14,889	14,613	13,795	13,768
P0 %	64.10	52.90	44.80	36.90	45.30	39.80
P1 %	34.80	45.60	52.60	60.40	51.80	57.50
P2 %	1.12	1.53	2.61	2.66	2.85	2.72
Control Total Reads	1,038	2,121	10,912	8,000	9,943	11,128
Control Poly RL Mean (kb)	35.14	39.69	56.20	52.63	53.26	58.48
Control Concordance Mean	0.86	0.85	0.84	0.84	0.84	0.85
Control Concordance Mode	0.89	0.89	0.87	0.87	0.87	0.87
Local Base Rate	2.12	2.03	2.26	2.31	2.36	2.30

Primary output data used to calculate V1 kit averages in Table 8.

References

(2019). All of Us Research Program Backgrounder. Retrieved from National Institutes of Health: All of Us Research Program: https://allofus.nih.gov/news-events-media/media-toolkit/all-us-research-program-backgrounder

(2019). Human Whole Genome Sequencing. Retrieved from Pacific Biosciences: https://www.pacb.com/applications/whole-genome-sequencing/human/

(2019). Products + Services: Latest System Release. Retrieved from Pacific Biosciences: https://www.pacb.com/products-and-services/sequel-system/latest-system-release/

Anderson, S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Research, 9(13), 3015-3027.

Au KF, Underwood JG, Lee L, Wong WH (2012) Improving PacBio Long Read Accuracy by Short Read Alignment. PLOS ONE 7(10): e46679. https://doi.org/10.1371/journal.pone.0046679

Bleidorn, C. (2015). Third generation sequencing: Technology and its potential impact on evolutionary biodiversity research. Systematics and Biodiversity. 14. 1-8. 10.1080/14772000.2015.1099575.

Chaisson, M.J.P., John Huddleston, Megan Y. Dennis, Peter H. Sudmant, Maika Malig, Fereydoun Hormozdiari, . . . Evan E. Eichler. (2014). Resolving the complexity of the human genome using single-molecule sequencing. Nature, 517(7536), 608-611G.

Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, . . . Jonas Korlach. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods, 10(6), 563-9.

Competition and Markets Authority. (2019, April 17). Illumina, Inc. / Pacific Biosciences of California, Inc. merger inquiry. Retrieved December 12, 2019 from https://www.gov.uk/cma-cases/illumina-inc-pacific-biosciences-of-california-inc-merger-inquiry

Costello, M., Fleharty, M., Abreu, J., Farjoun, Y., Ferriera, S., Holmes, L., … Gabriel, S. (2018). Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC genomics, 19(1), 332. doi:10.1186/s12864-018-4703-0

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. (2012) Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS ONE 7(11): e47768. https://doi.org/10.1371/journal.pone.0047768

Federal Trade Commission. (2019, December 17). FTC Challenges Illumina's Proposed Acquisition of PacBio. Retrieved December 31, 2019, from https://www.ftc.gov/news-events/press-releases/2019/12/ftc-challenges-illuminas-proposed-acquisition-pacbio

Feng, Y., Zhang, Y., Ying, C., Wang, D., & Du, C. (2015). Nanopore-based Fourth-generation DNA Sequencing Technology. Genomics, Proteomics & Bioinformatics, 13(1), 4-16.

Greenleaf, William J, & Sidow, Arend. (2014). The future of sequencing: Convergence of intelligent design and market Darwinism. Genome Biology (Online Edition), 15(3), 303.

Haque, F., Li, J., Wu, H., Liang, X., & Guo, P. (2013). Solid-state and biological nanopore for real-time sensing of single chemical and sequencing of DNA. Nano Today, 8(1), 56-74.

Heather, J., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 1-8.

Holley, R., Apgar, J., Everett, G., Madison, J., Marquisee, M., Merrill, S., . . . Zamir, A. (1965). Structure of a Ribonucleic Acid. Science, 147(3664), 1462-1465.

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., . . . Eichler, E. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. Genome Research, 24(4), 688-696.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., ... Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860-921. https://doi.org/10.1038/35057062

Levene, M., Korlach, J., Turner, S., Foquet, M., Craighead, H., & Webb, W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. Science (New York, N.Y.), 299(5607), 682-686.

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics, Proteomics & Bioinformatics 14(5), p. 265 - 279. https://doi.org/10.1016/j.gpb.2016.05.004

Megaruptor 2. (n.d.). Retrieved March 05, 2021, from https://www.diagenode.com/en/p/megaruptor2-1-unit

Megaruptor 3. (n.d.). Retrieved March 05, 2021, from https://www.diagenode.com/en/p/megaruptor-3

Nyrén, P. (1987). Enzymatic method for continuous monitoring of DNA polymerase activity. Analytical Biochemistry, 167(2), 235-238.

Pacific Biosciences. (2018, November 1). Illumina to Acquire Pacific Biosciences for Approximately $1.2 Billion, Broadening Access to Long-Read Sequencing and Accelerating Scientific Discovery. Retrieved December 12, 2020, from https://www.pacb.com/press_releases/illumina-to-acquire-pacific-biosciences-for-approximately-1-2-billion-broadening-access-to-long-read-sequencing-and-accelerating-scientific-discovery/

Padmanabhan, R., Ernest Jay, & Ray Wu. (1974). Chemical Synthesis of a Primer and Its Use in the Sequence Analysis of the Lysozyme Gene of Bacteriophage T4. Proceedings of the National Academy of Sciences of the United States of America, 71(6), 2510-2514.

Sanger, F., Brownlee, G., & Barrell, B. (1965). A two-dimensional fractionation procedure for radioactive nucleotides. Journal of Molecular Biology, 13(2), 373,IN1-398,IN4.

Sanger, F., & Coulson, A. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology, 94(3), 441,IN19,447-446,IN20,448.

Sanger, F., S. Nicklen, & A. R. Coulson. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America, 74(12), 5463-5467.

Schadt, E.E., Michael D. Linderman, Jon Sorenson, Lawrence Lee, & Garry P. Nolan. (2010). Computational solutions to large-scale data management and analysis. Nature Reviews Genetics, 11(9), 647-657.

Schadt, E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. Human Molecular Genetics, 19(R2), R227-R240.

Shendure, J., Ji, H., (2008). Next-generation DNA sequencing. Nature Biotechnology, 26(10), 1135-1145.

Stein, L. (2010). The case for cloud computing in genome informatics. Genome Biology, 11(5), 207-207.

Turcatti, Gerardo, Romieu, Anthony, Fedurco, Milan, & Tairi, Ana-Paula. (2008). New class of cleavable fluorescent nucleotides: Synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucleic Acids Research, 36(4), E25.

Uemura, S., Colin Echeverría Aitken, Jonas Korlach, Benjamin A. Flusberg, Stephen W. Turner, & Joseph D. Puglisi. (2010). Real-time tRNA transit on single translating ribosomes at codon resolution. Nature, 464(7291), 1012-7.

Watson, J. D. & F. H. C. Crick. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature, 171(4356), 737-738.

Wenger, Aaron M, Peluso, Paul, Rowell, William J, Chang, Pi-Chuan, Hall, Richard J, Concepcion, Gregory T, . . . Hunkapiller, Michael W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology, 37(10), 1155-1162.

Zallen, D.T., (2003). Despite Franklin's work, Wilkins earned his Nobel. Nature, 425(6953), 15.