Addressing the Needs of Research and Clinical Applications for Cell-Free DNA, Exome Selection, and Targeted Panel Selection in a High-Throughput Laboratory Environment

Addressing the Needs of Research and Clinical Applications for Cell-Free DNA, Exome Selection, and Targeted Panel Selection in a High-Throughput Laboratory Environment

Gina Vicente

A Thesis in the Field of Biotechnology

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

November 2021

Abstract

The thesis will address the development of a new targeted enrichment method using a hybrid selection-based approach. While targeted enrichment approaches have been used routinely by many research and clinical labs, the project here addressed several areas in need of improvement in order to increase assay sensitivity, specificity, and provide even coverage across the exome for somatic applications. The goal of this work was to achieve these improvements in a streamlined laboratory workflow optimized for a variety of sample types. It will explore the beginning to end process: from the entirety of the universal hybrid selection product creation, including challenges of early research and development, to implementation of a streamlined hybrid selection method at large-scale production at a clinical level, to data analysis, and the clinical applications for precision medicine. Large-scale genomic sequencing involves both the innovative scientific methods developed in research and overcoming operational challenges to produce repeatable, robust data at scale. The thesis will further discuss how laboratory operations balance custom, complex cancer projects with various sample types, challenging sample inputs such as formalin-fixed paraffin-embedded (FFPE) samples, and low quantity deoxyribose nucleic acid (DNA) samples processed through unique molecular indices (UMI) library preparation, sequencing, and data analysis, while also maintaining quality for the large-scale research and clinical projects that consist of thousands of samples.

Dedication

My thesis is dedicated to my parents. Thank you for the never-ending support and unconditional love.

Acknowledgments

Thank you to the somatic team, R&D team, automation team, Junko, Carrie, Justin, Nicole, Hayley, Sophie, Caroline, Wendy, Defo, Katie, Tim, and Stacey.

Table of Contents

List of Tables

List of Figures

Chapter I.

Introduction

Genomics provides the ability to study the human genome and understand how alterations within the DNA sequence can alter health. The study of genomics is constantly evolving as new technologies and methods are developed to further understand both population genomics and diseases, particularly for cancer. New Next Generation Sequencing (NGS) methods and analysis tools are created and continuously updated to improve sequencing quality, sequencing coverage, workflows, and cost of the assay. The ability to sequence various sample types at scale is necessary for scientific significance in large-scale projects. This thesis will address how laboratory operations implemented a newly developed hybrid selection method suitable for both large-scale projects and custom, complex cancer projects with various sample types and challenging sample inputs such as poor-quality formalin-fixed paraffin-embedded (FFPE) and samples with limited DNA material. The samples are prepared with UMIs and processed through the newly developed targeted enrichment method suitable for exome and targeted selection for various sample types.

Cancer

In the United States, cancer was the second leading cause of death in 2019 (Siegel, 2019). The American Cancer Society projected that cancer caused

approximately 606,880 deaths and another 1,762,450 new cases of cancer in the US alone in 2019 (Siegel, 2019). There are several different types of cancer affecting different tissues and organs, including leading cancers of the lung, breast, and colon. Several cancer genomic databases have been created as open-source information to help further scientific discovery. Some examples of these databases include the National Cancer Institute's The Cancer Genome Atlas (TCGA), Cancer Cell Line Factory (CCLF), and the United Kingdom's The Catalog of Somatic Mutations in Cancer (COSMIC), among many more (Yang *et al*, 2015). Each database has specific objectives and focus, but all contribute to the large amount of publicly available cancer data based on tumor type and varying analysis. For example, TCGA has created data sets using exome sequencing, mRNA sequencing, and SNP arrays (Manier, 2013). As a result, various omics interpretations are used to understand the tumor type and associated genetic variants.

The cancer databases include sequencing and analysis for cancer exomes, genomes, and transcriptomes. In addition to data sources for the cancer genomes, the databases also include analysis tools to find point mutations, structural alterations, and variants. These significant data sources can provide the necessary genomic information to strengthen drug therapies and even personalized genomics (Yang *et al*. 2015).

Challenges of Characterizing the Cancer Genome

These large data sets contain sequencing information from thousands of samples by tumor type to help to correspond the genes associated with those cancers. By characterizing different cancer types, it is possible to find actionable genomic alterations and find targeted drug therapies (Lanman, 2015). By understanding a patient's cancer

sequence, the possibilities of diagnostics, disease monitoring, and clinical trial enrollments have increased significantly (Lennon, 2016).

However, understanding the cancer genome has multiple challenges. Tumor collections can be difficult, particularly for areas that are not easy to access such as cancer in the brain, and sample quality can be poor. Tumor samples prepared for sequencing are often of poor sample quality, such as samples prepared for formalin-fixed paraffin-embedded (FFPE), the standard tumor preservation method used in pathology surgeries (Kokkat, 2013). Limitations of the formalin fixation methods include tumor tissue availability and possible sample modifications due to the formalin fixation (Kokkat, 2013), causing DNA extraction of FFPE samples to result in low DNA quantity for sequencing. Furthermore, cancers can cause multiple genomic alterations that can complicate analysis of what causes cancer in a specific region and what causes cancer to spread (Lennon, 2016). Despite the challenges, there is high potential to associate the mutated genes to understand a patient's cancer and even provide personalized care.

Cell-free DNA (cfDNA)

A particular sample type with immense potential to further cancer discovery is cell-free DNA (cfDNA) within circulating tumor DNA (ctDNA). cfDNA consists of roughly 166 base pair double-stranded DNA (dsDNA) created by apoptosis or release of nuclear DNA into the circulation (Lanman, 2015). The ctDNA within cfDNA can provide the unique ability to sample a patient's blood in real-time, using a simple blood draw. The advantages of using cfDNA as a liquid biopsy compared to a tumor biopsy include the less invasive method of extracting the DNA from the patient while still

having the ability to sequence targeted panels to deep sequencing depth (Lanman, 2015). In addition, while ctDNA has lower concentrations of mutated DNA in comparison to tissue from biopsies, methods have been developed that can quantify the tumor fraction, such as the ichorDNA algorithm (Adalsteinsson, 2017). In addition to a blood draw being less invasive than a traditional tissue biopsy, a blood draw is also much less expensive, at approximately $100-200 (Lennon, 2016). In comparison, tissue biopsies can cost between $1000-4000 (Lennon, 2016). Furthermore, repeated blood draws for ctDNA over shorter time spans, such as days or weeks, are more feasible for patients than repeated tumor biopsies (Weber, 2021).

The importance of tumor heterogeneity, diversity of cells within a tumor, can be better characterized using cfDNA rather than typical tumor biopsies, particularly for gastrointestinal cancer (Parikh, 2019). The single tumor biopsy will provide localized information to the tumor, whereas a cfDNA sampling can provide a more comprehensive analysis of the tumor, particularly for tumor heterogeneity (Parikh, 2019). Because this sample type is advantageous in creating impactful sequencing data, there is a need to establish robust cfDNA sequencing methods that are both repeatable and reproducible at a large scale that also meets patients' needs.

Exome & Targeted Sequencing

Whole-genome sequencing provides the ability to sequence the entirety of an organism's genome. The human genome is approximately 3 billion base pairs long. While whole-genome sequencing costs are decreasing, it may still be more effective to

run whole exome or targeted sequencing depending on the study and sequencing depth required. Exome sequencing targets the exome protein-encoding region, while targeted sequencing allows specific regions to be selected in either non-coding or coding regions and sequenced instead of the entire genome (Gnirke, 2009). The goal to create a 30x coverage human whole genome sequencing that costs approximately ~$1000 is still in progress, as laboratories incorporate the costs of both the library preparation and sequencing (Schwarze, 2019). In contrast, the cost of a human whole exome sequence can be significantly due to the smaller region for coverage and costs as low as $500 with prices constantly dropping (Schwarze, 2018).

Data processing analyses for the raw NGS data output from sequencer are needed to demultiplex the samples, re-align the fragments to the reference human genome, and analyze the sequencing quality of the data produced. Alignment pipelines such as BWA-ALM, BWA-MEM, and the Illumina Dynamic Read Analysis for Genomics (DRAGEN) can be used to analyze the NGS data. The Burrows-Wheeler aligner (BWA-ALN) and the Burrows-Wheeler maximum exact matches (BWA-MEM) are two alignment tools used for genome sequencing analysis (Robinson, 2017). The Picard pipeline from the Broad Institute consists of command-line tools which analyze sequencing data after the alignment to provide quality metrics for whole-genome sequencing, whole-exome sequencing, and RNA sequencing. Picard metrics such as mean target coverage (MTC), % selected, % target bases, and off-target provide quality control (QC) metrics will determine the sequencing quality of the samples discussed in this thesis. MTC is a function of the amount of sequencing devoted to the sample, while the other metrics help provide understanding of the hybrid selection process utilizing the exome or targeted

oligonucleotides that bind to the specific region of interest. In addition, uniformity within the target region is important for sequencing coverage to prevent increased sequencing for regions that may be under-covered (Hasin-Brumshtein *et al*. 2018).

Samples can then be analyzed for further discovery to determine structural variants, indels, single nucleotide polymorphisms, and raw reads to further understand genotypes, variations, or population differences (DePristo, 2011). Exome sequencing for somatic samples provides information regarding somatic copy number alterations, clonal mutations, mutational signatures, and neoantigens (Adalsteinsson, 2017). The information from targeted enrichment sequencing be used to understand populations or applied for patient care.

Development on cfDNA

New and novel methods are necessary to produce high quality data to constantly advance genomics and scientific discoveries. New reagents and sequencing technologies are tested to understand what would be advantageous in large-scale sequencing and genotyping production. New reagents, automation, and sequencing technologies are tested for feasibility and quality. As biotechnology companies create higher quality oligonucleotides for targeted panels paired with the decreasing cost of sequencing, the cost of targeted and exome sequencing has decreased, allowing for larger sample size or deeper coverage of sequencing.

Important factors for choosing new selection methods included the quality of the product, product performance on various sample types, price, ease of workflow, and ability to customize panels with the vendor. The goal was to develop a targeted method

with similar or improved performance compared to the current Illumina exome in production. The universal selection designed particularly for somatic samples includes a high-throughput liquid biopsy workflow. The method can process cfDNA, gDNA, and FFPE samples through extraction, UMI library preparation, hybrid selection, and sequencing. For cfDNA, it begins with extraction from plasma. Other somatic samples can be processed alongside the cfDNA, and different sized panels with different targeted regions can be utilized in this single workflow.

Creation of the cfDNA ULP-WGS workflow

Methods were designed in collaboration with Cancer Genomics to sequence cfDNA and the matched normal DNA. An ultra-low pass whole genome sequencing (ULP-WGS) library and sequencing to just 0.1x coverage was able to determine if the cfDNA is suitable for downstream deep coverage exome sequencing by using ichorDNA. ichorDNA is a software that can determine tumor percentages using somatic copy number alterations within the cfDNA (Adalsteinsson, 2017). The ULP-WGS is an initial screen of the patient's cfDNA. The sequencing data from the ULP-WGS screening provides tumor percentage and information about the disease progression or treatment resistance, and whether downstream sequencing can be informative. The goal is to scale-up these developments, allowing for thousands of samples to be screened through the ULP-WGS, ichorDNA analysis, and proceed to exome sequencing if the specification of tumor percentage greater than 10% is met. The ability to process these samples in a clinical laboratory would allow the information obtained from cfDNA sequencing to be used for patient care.

A new selection method was needed that would be able to incorporate the new ULP-WGS workflow along with the somatic workflow including the whole exome sequencing and targeted sequencing for both research and clinical applications. With the cost of sequencing decreasing due to sequencer technology advances, continued work to study specific tumor types with higher coverage sequencing also became more feasible. Cancer studies require targeted enrichment products that can be customized while still maintaining high-quality, high-throughput sequencing. This work resulted in the new method consisting of library preparation and hybrid selection that can use any targeted panel from any vendor, with the option of sequencing to the customer's chosen sequencing coverage depth.

Addressing the Needs of Research and Clinical NGS

The need to sequence cfDNA and capture important cancer target regions at depth will allow researchers to further understand the disease. Because cancer can cause multiple genomic alterations that can complicate analysis, the ability to sequence to various depths provides flexibility for cancer research. Sequencing the exome region can be more cost effective, particularly when approximately 85% of the recognized disease-causing mutations are found in the exome region (Rehm, 2016). Costs are less prohibitive for exome and targeted sequencing. At the Broad Institute Genomics Platform, the cost of one ULP-WGS sample, including fractionation and extraction, is approximately $105. The results of the low coverage 0.1x sequencing provides researchers with information to decide if further sequencing is desired.

Sequencing cfDNA is beneficial at both the research level and clinical level. Clinical samples used for patient treatment are handled in a way that meets Clinical Laboratory Improvement Amendments (CLIA) and College of American Pathologists (CAP) regulations. The work completed here follows guidelines on test development and validation, quality management, and is analyzed further to determine for sequencing depth, sensitivity, and specificity for the assay. The clinical validation will include various sample types, utilize intra-run and inter-run reproducibility testing by trained personnel, and evaluate the test performance to CAP and CLIA's requirements. The creation of a new workflow for both gDNA and cfDNA through exome selection and targeted panel selection has resulted in both meaningful research and clinical applications for the cancer community.

Chapter II.

Materials & Methods

The development methods and production protocols are described in this section. The Somatic team, Research & Development (R&D), automation teams, and laboratory operations production teams created these methods within Broad Institute Genomics Platform laboratories in Biosafety Level 1 (BL1) or Biosafety Level 2 (BL2) settings.

Library Construction Input

Samples used for early development included HapMap cell lines from the International HapMap Project, which mapped the haplotypes of the human genome and resulted in a public reference database. The hybrid selection methods were tested on unique molecular indices (UMI) libraries from HapMap cell lines and validated on various sample inputs. Previous validation work was performed to include UMIs to dual-indexed libraries to improve sequencing accuracy. The 6 nt UMI barcode can be analyzed to help improve accuracy by removing PCR duplicates and other duplicate reads to improve variant sensitivity in downstream analysis. Libraries were created using the KAPA HyperPrep Kits with KAPA Library Amplification Primer Mix (10X) (catalog # KK8504). Customized stubby-Y UMI adapters were ordered from Integrated DNA Technologies (IDT) and titrated to find the optimal library condition. Optimized libraries included a 5 µL addition of UMI adapter added during the adapter ligation step, and 4 µL

of the P7 primer and 4 µL of the P5 primer during PCR. These optimized libraries are part of the complete workflow and are the input into the hybrid selection methods discussed here. These included the creation of libraries from the well-defined HapMap cell line NA12878 and other well-characterized HapMap samples. Samples were extracted using cell pellets or purchased from the Coriell Institute.

Hybrid Selection Developments

IDT's exome panel and xGen® Hybridization and Wash Kit (catalog #1080584) were tested along with Twist Biosciences' custom ordered oligonucleotides. Different experiments were tested in order to achieve the goal of creating a high throughput hybrid selection method suitable to run various sample types through an automated workflow. The UMI libraries were the input into the process, but the input quantity of UMI libraries into the selection process needed to be determined. IDT's protocol recommends 500 ng per sample along with 12-plexed pools. After library construction, NA12878 libraries were quantified using Thermo Fisher's Quant-iT Pico-Green dsDNA kit and normalized to 25 ng/µL on a Hamilton Starlet using Tris-HCl. After normalization, 25 µL of each library was pooled into a single pool of 8 or 12 sample pools for testing and tested alongside a single, non-pooled sample. The different pool size did not create differences in metrics, but larger pool sizes reduced reagent usage and downstream sequencing cost. Therefore, a maximum of 12-plex pools were decided upon and used for future testing with 500 ng of each sample.

DNA Concentration Prior to Capture

The previous selection method, ICE, utilized a SPRI concentration to reduce the large sample volume prior to the hybridization. The SPRI protocol used the SPRI beads from Beckman Coulter, with a 70% ethanol wash, and an elution of 30 µL Tris-HCl. However, the IDT xGen® Hybridization and Wash Kit protocol recommended either a Speed-Vac system or a SPRI concentration (IDT, 2020). In place of a Speed-Vac, a Biotage SPE-DRY lyophilizer dried down sample in a 96-well plate format. The protocol required the DNA 12-plex pools, Human Cot-1, and IDT blocking oligo (catalog #1075476) to dry down. Optimal temperature settings were 50° C for upper temperature and 70° C for lower temperature with input as 60 PSI, upper flow set to 80 L/min, lower flow set to 25 L/min.

Both the lyophilizer and SPRI methods were tested several times using NA12878. Different temperature settings and air flow on this machine were not as sensitive as the SPRI concentration required for ICE. The lyophilizer did not require hands-on processing and provided more consistent quantification results, proving to be the more optimal concentration method.

One concern was the potential for contamination using the Biotage SPE-DRY machine. However, due to the 96 wells on the head, there was little possibility for wells to contaminate each other. Samples were dried down into the bottom of the well using heat and air. In testing, no splashing was observed, and sequencing metrics did not indicate unexpected index reads in pools for evidence of contamination. Furthermore, a water bath sonicator was purchased to clean the head after each use. Good lab practices of weekly cleaning and library contamination metrics are regularly monitored.

Heat and Labware Experiments

Specific heat temperatures are required for DNA denaturation, annealing, and wash temperature optimization. Heating elements during the overnight 16-hour incubation was performed in an Eppendorf twin.tec 96-PCR plate (catalog #95041-438) that consisted of lyophilized pooled samples, Human Cot-1, and IDT Blocking oligo. The wells are then resuspended with the hybridization buffer and oligonucleotides of the targeted panel before going on a 16-hour incubation consisting of 95°C for 30 sec to denature and then a 65°C hold for the remaining overnight time. During the incubation, the dsDNA denatures and the xGen Blocking Oligo bind to non-targeted regions, while the oligonucleotides bind to the target regions. After the overnight incubation, the pooled samples undergo multiple wash steps to remove non-target material. The wash steps are performed at 72° C and at room temperature on an automated Agilent Technologies Bravo liquid handler. Samples then undergo PCR of the targeted material using 2x KAPA HiFi HotStart ReadyMix (catalog #KK2606).

The plate holding samples was changed from an Eppendorf twin.tec 96-PCR plate to a Eppendorf LoBind® twin.tec PCR Plate (catalog # 0030129512) to prevent the targeted DNA material from adhering to the plastic well. 2D barcoded tubes with easy to color-code caps were implemented for oligonucleotide (bait) registration with the Laboratory Information Management System.

Targeted Panel Creation

Twist Biosciences' ability to create customized oligonucleotides in approximately 3 weeks provides the expedited turnaround necessary to create the panel, test them, and analyze the coverages for both small and large customized projects. Twist provided customizable oligonucleotides at a cheaper price along with the flexibility to purchase oligonucleotides separately from the capture reagents, allowing for customization in the selection method. Their technology synthesizes DNA through silicon-based wafers using 9,600 nanowells rather than the traditional 96-wells from traditional DNA synthesis (LeProust, 2016). The smaller volumes in their silicon plates allows for increased efficiency to synthesize thousands of genes per run (LeProust, 2016). Their technology also allows customized ordering along different locations within the genome at both intronic and exonic regions and provides the ability to cover regions that are more difficult to cover during selection and sequencing, while also decreasing cost.

A custom exome panel was designed based on input from Cancer Genomics investigators taking into account the need to study cancer-specific gene regions. An additional 1.8 Mb region, which includes specific genes deemed important to cancer research, was added to Twist's 33 MB Human Core exome. The target territory includes the mitochondrial genome, the ACMG genes, the Online Mendelian Inheritance in Man putative gene sequences, 99% of ClinVar variants, COSMIC genes, the Dana Farber Cancer Institute OncoPanel Genes, and cancer specific regions including TAL1 Enhancer, TERT Promoter, NOTCH 3'UTR, and FOXA1 Promoter (Cibulskis, 2020). The final target territory consists of a 35,086,168-base region with a baited region of

38,886,093 bases, and probe length of approximately 120-mer. The final design was submitted to Twist Biosciences and created using their patented silicon production.

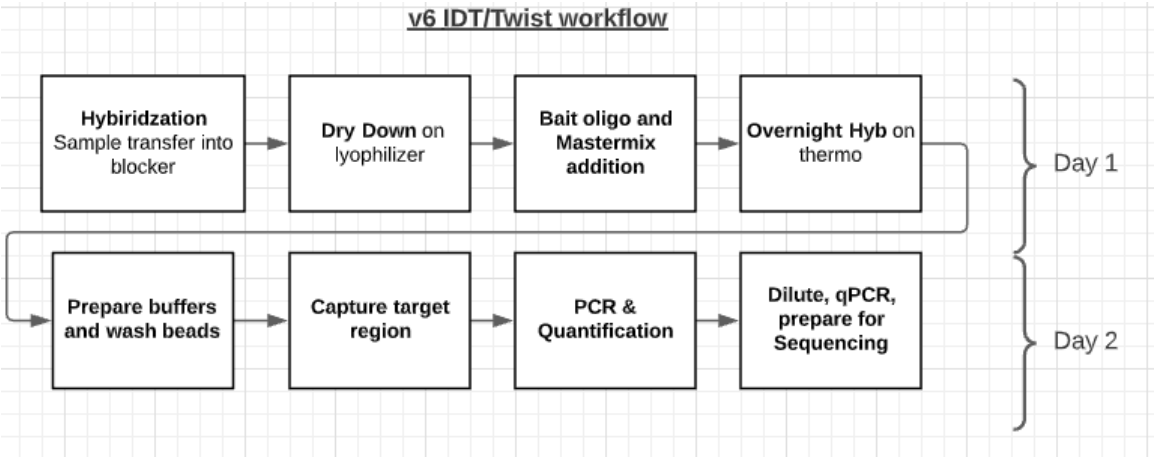

Figure 1. Exome v6.0 IDT exome workflow.

The laboratory workflow for the Exome v6.0 with the Twist Broad custom exome or custom panel hybridization.

Quality Controls for Custom Panel Testing

In order to ensure that each panel and synthesis thereafter were able to capture the expected target regions, a simplified quality control workflow was developed. A well-defined 16-gene panel was ordered in excess from Twist Biosciences to be used for the process control panel, and it consisted of a 55,857 bp panel size with 587 probes. The panel consistently performed with a %selected metric between 91-95% through multiple hybridizations performed manually by multiple users and through automated testing. The workflow consisted of triplicate HapMap NA12878 hybridized on the same PCR plate using the process control alongside triplicates of HapMap NA12878 hybridized with any new panel for testing and following the workflow depicted in Figure 1. By processing

the two panels on the same PCR plate, the lab can analyze the in-process PCR quantification and downstream sequencing metrics to eliminate the possibility of a poor hybridization event.

All the samples are then pooled together and sequenced on an Illumina MiSeq sequencer for testing. After data is demultiplexed, aligned, and processed through the Picard Pipeline using the hg19 reference, the samples receive Picard NGS metrics. Specific metrics such as sequencing depth, %selected, and %target bases at 20x as listed in Table 1 are analyzed to see if target regions were hybridized appropriately. Further analysis to check the regions of interest by performing a coverage analysis of the target regions within the panel can also be run.

Table 1. Quality control metrics for testing panels.

Hs Bait set	Product Order Sample	Fold 80 Penalty	Mean Target Coverage	Selected bases %	Target Bases 20x %	Target Territory	Zero Coverage Targets
BroadPanCancer2019	SM-I3Q34	1.223071	196.914423	90.1891	99.2236	1668012	0.3985
BroadPanCancer2019	SM-I3Q3S	1.237062	228.856455	87.971	99.5106	1668012	0.3871
BroadPanCancer2019	SM-I3Q3G	1.239244	211.910774	89.3624	99.4839	1668012	0.3871
twist_proc_contr_v1_1	SM-I3Q3H	4.096497	819.299318	94.7204	100	56161	0
twist_proc_contr_v1_1	SM-I3Q35	4.421179	884.235893	95.1991	100	56161	0
twist_proc_contr_v1_1	SM-I3Q3T	4.250575	850.115009	91.9259	100	56161	0

Here, a specific 396 gene panel, BroadPanCancer2019, is tested alongside the Process control bait, twist_proc_control_v1_1. The process control bait is a small territory that will receive %selected bases in the low-mid 90s. Lower % selected metrics may indicate processing issues.

Research Validation

The laboratory research validation of the new method was completed in October 2019. Due to the customer need to create a research-based product, the research products were tested and launched before clinical testing. Samples were tested with the new method and compared to the current Illumina exome capture. The research validation included varying sample types as listed in Table 2, Table 3, and Table 4. It included 12 FFPE samples with 150ng input, 12 FFPE samples with 300 ng input, 41 gDNA samples with 100ng input, and 20 cfDNA samples of 10ng input with their 20 gDNA normal pairs. The differing sample types of cfDNA, gDNA, and FFPE samples were sourced from previously processed NCI ALCH lung cancer or prostate samples or cell lines. They provided comparable representation of the types of samples run through the somatic workflow to test expected performance. Different batches, named Library Construction sets (LCSETs) were processed through the lab in 96-well format. Sequencing in batches of approximately 24 samples were placed onto the Illumina NovaSeq 6000 sequencer.

Ongoing pipeline analysis continues with pipeline improvements. The two aligners compared were the Burrows-Wheeler aligner (BWA-ALN) and the Burrows-Wheeler maximum exact matches (BWA-MEM), which are alignment tools used for genome sequencing analysis (Robinson, 2017). The Exome v2.0 utilized BWA-ALN, while the Exome v6.0 was tested with BWA-MEM. BWA-MEM is an improved version for alignment (Robinson, 2017).

Table 3. Research validation of 41 whole blood samples with 100ng input.

Sample ID	External ID	LCSET	Input	Quality	PF Bases	% selected	%target bases at 20x	% target bases at 100x	Zero Coverage Targets
SM-GSXY6	ALCH-B006-NB1-A-1-0-D-A673-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	16,734,858,136	93.1409	97.7456	47.7454	1.4112
SM-GSXY6	Deleted per Client Request	LCSET-15216	100 ng	Whole Blood:Whole Blood	20,568,176,028	92.8431	97.9941	77.635	1.3055
SM-GYUN3	ALCH-B029-NB1-A-1-0-D-A679-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	34,900,117,984	92.6655	98.1209	96.2198	1.2566
SM-HVG32	ALCH-B1TB-NB1-A-1-0-D-A76B-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	32,537,750,092	92.6715	98.0564	93.5874	1.3109
SM-H4EQN	ALCH-B0DF-NB1-A-1-0-D-A695-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	36,559,125,036	92.6305	98.2463	96.129	1.1186
SM-GYUK9	ALCH-AF3Q-NB1-A-1-0-D-A678-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	39,020,071,252	92.773	98.1689	96.9696	1.214
SM-GYUPY	ALCH-B063-NB1-A-1-0-D-A680-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	40,306,249,200	92.623	98.1689	97.1869	1.2282
SM-HZ7VV	ALCH-B1YT-NB1-A-1-0-D-A76I-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	38,716,726,176	92.7165	98.2136	96.7104	1.1436
SM-GBOA9	SM-EBKCH	LCSET-15216	100 ng	Whole Blood:Whole Blood	31,494,741,276	92.6461	98.3651	94.62	0.9943
SM-GBO9H	SM-EBKCL	LCSET-15216	100 ng	Whole Blood:Whole Blood	37,645,858,224	92.7603	98.3499	96.6626	1.0325
SM-GBO9T	SM-EBKCM	LCSET-15216	100 ng	Whole Blood:Whole Blood	36,879,152,656	92.823	98.3008	96.261	1.057
SM-GBOA5	SM-EBKCF	LCSET-15216	100 ng	Whole Blood:Whole Blood	39,659,513,292	92.7616	98.3798	97.1093	1.0335
SM-GBO9O	SM-EBKCS	LCSET-15216	100 ng	Whole Blood:Whole Blood	38,961,355,308	92.7499	98.3278	96.7211	1.0736
SM-GBO9I	SM-EBKCT	LCSET-15216	100 ng	Whole Blood:Whole Blood	40,122,381,764	92.7479	98.4027	97.0368	0.9772
SM-GBO9X	SM-EBKCR	LCSET-15216	100 ng	Whole Blood:Whole Blood	34,349,154,572	92.7424	98.3371	95.5839	1.008
SM-GBO9S	SM-EBKCE	LCSET-15216	100 ng	Whole Blood:Whole Blood	46,802,332,320	92.6463	98.4293	97.589	0.9958
SM-GBO9N	SM-EBKCK	LCSET-15216	100 ng	Whole Blood:Whole Blood	44,578,468,204	92.8301	98.3602	97.3592	1.0242
SM-HVOK6	NA12878_Aug2018_18	LCSET-15216	100 ng	DNA:DNA Genomic	48,810,936,080	92.9957	98.2086	97.5688	1.2189
SM-HVOJZ	NA12878_Aug2018_12	LCSET-15216	100 ng	DNA:DNA Genomic	41,513,657,228	93.1877	98.185	97.1391	1.2282
SM-HVOKC	NA12878_Aug2018_24	LCSET-15216	100 ng	DNA:DNA Genomic	29,696,880,340	93.2206	98.0942	93.1571	1.2708
SM-HVOKO	NA12878_Aug2018_36	LCSET-15216	100 ng	DNA:DNA Genomic	52,533,813,440	93.1367	98.2643	97.5416	1.1744
SM-HVOJT	NA12878_Aug2018_6	LCSET-15216	100 ng	DNA:DNA Genomic	38,796,690,376	93.1337	98.1471	96.588	1.2473
SM-HVOKI	NA12878_Aug2018_30	LCSET-15216	100 ng	DNA:DNA Genomic	44,734,206,696	93.2098	98.1919	97.3642	1.2395
SM-GYUO4	ALCH-B04H-NB1-A-1-0-D-A679-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	34,236,631,204	93.0394	98.2245	94.9408	1.1397
SM-GEAZG	ALCH-AEKP-NB1-A-1-0-D-A628-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	36,103,203,828	92.6293	98.2562	95.9562	1.1524
SM-GSXXT	ALCH-AFDR-NB1-A-1-0-D-A673-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	53,375,869,608	92.5986	98.2191	97.6727	1.2155
SM-GYUO6	ALCH-B04I-NB1-A-1-0-D-A679-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	47,725,901,456	92.6216	98.2832	97.4027	1.1494
SM-B4MKQ	ALCH-AC6Z-NB1-A-1-0-D-A49J-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	48,056,892,800	92.3992	98.2095	97.5393	1.216
SM-B4MKB	ALCH-AC5E-NB1-A-1-0-D-A49J-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	45,734,406,852	92.508	98.2407	97.5517	1.1773
SM-GSXYI	ALCH-B024-NB1-A-1-0-D-A673-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	28,121,698,020	92.6993	98.1734	90.426	1.1734
SM-H4EQ2	ALCH-B0CK-NB1-A-1-0-D-A695-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	29,566,650,968	92.3579	98.1185	93.0413	1.2483
SM-B4MIL	ALCH-AC00-NB1-A-1-0-D-A49J-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	38,025,386,100	92.8631	98.1437	96.9304	1.24
SM-GSXXH	ALCH-AF1T-NB1-A-1-0-D-A673-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	39,803,013,480	92.8468	98.1352	96.3936	1.2551
SM-C1RVE	ALCH-ACGN-NB1-A-1-0-D-A49Q-36	LCSET-15216	100 ng	DNA:DNA Genomic	41,616,491,744	92.7571	98.182	97.313	1.2224
SM-GYUJY	ALCH-AEZR-NB1-A-1-0-D-A678-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	34,217,782,020	92.7927	98.1452	95.9994	1.2277
SM-GYUKL	ALCH-AFB1-NB1-A-1-0-D-A678-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	40,373,671,416	92.4534	98.2063	97.1775	1.2042
SM-GEB16	ALCH-AENU-NB1-A-1-0-D-A628-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	47,814,851,956	92.7031	98.2576	97.431	1.1465
SM-GYUP6	ALCH-B05C-NB1-A-1-0-D-A680-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	36,927,015,836	92.5294	98.297	96.5711	1.1005
SM-HZ7VL	ALCH-B1YB-NB1-A-1-0-D-A76I-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	41,417,526,156	92.939	98.2092	97.2936	1.2038
SM-B4MJY	ALCH-AC3R-NB1-A-1-0-D-A49J-36	LCSET-15216	100 ng	Whole Blood:Whole Blood	40,168,977,664	92.8448	98.1604	97.2291	1.2615

Table 4. Research validation of 20 cfDNA and their corresponding 22 gDNA normal samples.

Sample ID	External Sample	LCSET	Input	Root Material Type	PF Bases	Selected Bases %	Target Bases 20x %	Target Bases 100x %	Zero Coverage Targets %
SM-HTPB6	PCProject_0159_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	2.5838E+10	92.8776	98.0305	77.8111	1.0574
SM-HTPD9	PCProject_0133_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	4.7629E+10	93.0887	97.8393	91.3372	1.3672
SM-HREW4	PCProject_0119_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	3.3014E+10	92.0904	97.8687	90.5504	1.4029
SM-HTPCK	PCProject_0068_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	5.5766E+10	92.9741	98.0525	95.3846	1.305
SM-HTPB7	PCProject_0093_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	5.6093E+10	91.5888	98.4104	96.93	0.9929
SM-HTPBN	PCProject_0118_BLOOD_2_P	LCSET-15304	10 ng	Plasma:Plasma	3.7083E+10	92.2086	97.8416	93.0225	1.493
SM-HREWI	PCProject_0189_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	4.2593E+10	91.8743	98.2537	95.8797	1.0976
SM-HTPCE	PCProject_0134_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	4.4887E+10	93.0553	97.8829	95.4039	1.42
SM-HRQZB	PCProject_0103_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	5.0875E+10	93.0681	97.9839	95.9508	1.4078
SM-HREXF	PCProject_0088_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	5.3732E+10	91.9579	98.1259	96.4427	1.2664
SM-HTPCC	PCProject_0050_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	3.3062E+10	92.1394	97.865	89.9512	1.4401
SM-HTPB5	PCProject_0172_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	4.6735E+10	92.0317	97.9279	95.7156	1.4137
SM-HRDOH	PCProject_0203_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	4.6668E+10	92.8559	98.148	95.7018	1.2351
SM-HTPCI	PCProject_0109_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	3.9065E+10	92.1735	98.081	90.9108	1.2351
SM-HRDOL	PCProject_0341_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	4.3859E+10	92.8974	98.2961	95.9423	1.0638
SM-HRQZI	PCProject_0437_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	4.3848E+10	91.864	98.1962	94.9699	1.17
SM-HTPCW	PCProject_0030_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	6.3171E+10	92.6635	98.4271	96.9092	0.962
SM-HRQZH	PCProject_0294_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	5.1064E+10	91.6686	97.9572	96.5895	1.4391
SM-HREWE	PCProject_0064_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	3.692E+10	93.0581	97.9286	94.0857	1.3843
SM-HREVW	PCProject_0083_BLOOD_P	LCSET-15304	10 ng	Plasma:Plasma	4.4787E+10	92.9226	97.9332	95.495	1.4802
SM-I74N2	PCProject_0088_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	3.5626E+10	92.3836	98.1498	95.1694	1.2126
SM-I74NN	PCProject_0203_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	4.026E+10	92.7267	98.1501	95.9823	1.2199
SM-I74O1	PCProject_0093_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	3.4473E+10	92.1893	98.1623	94.4308	1.1597
SM-I74NP	PCProject_0159_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	4.0633E+10	92.6316	98.2019	95.6983	1.1656
SM-I74OB	PCProject_0068_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	3.9611E+10	92.0798	98.2857	96.719	1.0981
SM-I74MP	PCProject_0050_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	3.3785E+10	92.3269	98.1728	95.2725	1.1964
SM-I74NR	PCProject_0030_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	3.4558E+10	92.0982	98.1704	92.9919	1.146
SM-I74OM	PCProject_0106_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	3.652E+10	92.1997	98.203	95.3416	1.1969
SM-I74MO	PCProject_0134_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	3.7161E+10	92.6581	98.1722	95.9539	1.1597
SM-I74NO	PCProject_0118_BLOOD_2_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	3.3561E+10	92.3191	98.1667	94.929	1.1685
SM-I74P2	PCProject_0064_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	3.3671E+10	92.0073	98.0418	88.5531	1.2209
SM-I74NC	PCProject_0109_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	4.0831E+10	92.2297	98.192	96.2504	1.1749
SM-I74NM	PCProject_0341_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	3.1014E+10	92.8084	98.1141	92.2627	1.1871
SM-I74OS	PCProject_0133_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	3.5008E+10	92.7297	98.1557	94.7874	1.1617
SM-I74OY	PCProject_0187_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	3.2305E+10	92.8458	98.1285	93.8804	1.1905
SM-I74ML	PCProject_0172_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	3.6092E+10	92.7981	98.1952	95.5203	1.1475
SM-I74NB	PCProject_0103_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	2.6952E+10	92.3995	98.0865	88.8849	1.2204
SM-I74MM	PCProject_0083_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	4.0812E+10	92.1152	98.2709	96.2836	1.1235
SM-I74N3	PCProject_0294_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	4.2478E+10	92.6026	98.2277	96.7008	1.1499
SM-I74OA	PCProject_0437_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	3.8064E+10	92.678	98.1523	95.7052	1.216
SM-I74NE	PCProject_0189_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	4.8102E+10	92.4269	98.1395	94.7421	1.2243
SM-I7IJI	PCProject_0119_BLOOD_BC	LCSET-15305	150ng	Whole Blood:Buffy Coat	2.533E+10	92.7729	97.9954	85.1186	1.2943

Clinical Validation

Various sample inputs and sample types were tested in the clinical validation. Validations require testing for repeatability and precision using sample replicates within the processing run and throughout multiple processing runs (Chesher, 2008). Samples included in the validation include the well-characterized HapMap sample NA12878 in replicate, various HapMap cell lines to create the Panel of Normals, tumor-normal pairs of various tumor and normal material types, and HapMap 5-plex, 10-plex- and 20-plex pools (Table 5). The cell line pools are listed in Table 6 and were run in replicate to test for sensitivity and compared against previous truth data. The data pipelines included Illumina's Dynamic Read Analysis for Genomics (DRAGEN) somatic pipeline in comparison to the current Picard pipeline utilizing the BWA-MEM pipeline.

Furthermore, in process quality control methods such as contamination and sample identity checks were run. A sample identity check using a genotyping tool using 98 SNPs was run on the validation samples and compared to sequencing data to find the Log of Odds (LOD) scores. Positive LOD scores indicate the sample identities match, while negative LOD scores indicate different sample identities. Samples were analyzed through both BWA-MEM and DRAGEN with hard-clipping, to compare the differences between the analysis pipelines.

Future testing for the clinical validation will include tumor-normal pairs including FFPE tumor with blood normal, FFPE tumor with saliva normal, fresh frozen tissue

tumor with blood normal, fresh frozen tissue with saliva normal, and FICOLL blood (PBMCs) tumor with saliva normal to find the various tumor inputs and sensitivity.

Table 5. Batch 1 of samples for the clinical validation of the Exome v6.0.

Use	PDO Sample	Somatic New Patient ID	Total PF Bases	% Selected Bases	% Target Bases at 100X	Zero Coverage Targets %
HapMap Pool	SM-KZIBE	10plex_100ng_1	42,541,250,488.00	91.566	93.734	1.203
HapMap Pool	SM-KZIBQ	10plex_100ng_2	33,622,798,732.00	91.608	96.043	1.214
HapMap Pool	SM-KZIC3	10plex_100ng_3	33,110,338,732.00	91.671	94.845	1.21
HapMap Pool	SM-KZICF	20plex_100ng_1	32,213,670,388.00	91.654	87.29	1.195
HapMap Pool	SM-KZICR	20plex_100ng_2	33,103,889,328.00	91.576	95.076	1.163
HapMap Pool	SM-KZID4	20plex_100ng_3	34,095,453,292.00	91.405	96.259	1.119
HapMap Pool	SM-KZIDG	5plex_100ng_1	41,993,353,660.00	91.162	97.317	1.124
HapMap Pool	SM-KZIDS	5plex_100ng_2	50,279,667,172.00	91.195	97.378	1.092
HapMap Pool	SM-KZIBF	5plex_100ng_3	45,822,013,276.00	91.774	80.234	1.268
NA12878 titration	SM-KZIDU	NA12878_100ng_1	31,993,566,628.00	91.68	91.721	1.368
NA12878 titration	SM-KZIBH	NA12878_100ng_2	40,429,660,372.00	91.625	96.67	1.325
NA12878 titration	SM-KZID7	NA12878_10ng_1	41,550,497,408.00	91.868	94.126	1.381
NA12878 titration	SM-KZIDJ	NA12878_10ng_2	46,032,351,096.00	91.87	95.451	1.36
NA12878 titration	SM-KZIDV	NA12878_300ng_1	61,889,401,460.00	91.242	97.566	1.29
NA12878 titration	SM-KZIBI	NA12878_300ng_2	45,950,901,492.00	91.289	97.122	1.334
NA12878 titration	SM-KJTBA	NA12878_50ng_2	42,139,705,228.00	91.606	96.478	1.294
NA12878 titration	SM-KZIBU	NA12878_50ng_2	40,378,647,388.00	91.696	95.241	1.327
PON creation	SM-KZIC4	SM-JLV7I_NA10865_Oct2019	37,437,407,600.00	91.676	89.313	1.266
PON creation	SM-KZICG	SM-JLV7J_NA12817_Oct2019	29,809,972,524.00	91.517	93.579	1.288
PON creation	SM-KZIBG	SM-JLV7O_NA10864_Oct2019	35,477,573,388.00	91.569	95.439	1.35
PON creation	SM-KZIBS	SM-JLV7P_HG03919_Oct2019	47,710,314,496.00	91.412	97.27	1.339
PON creation	SM-KZIC5	SM-JLV7Q_HG01791_Oct2019	38,403,837,664.00	91.138	95.094	1.078
PON creation	SM-KZICH	SM-JLV7R_NA18864_Oct2019	33,973,765,212.00	91.473	96.529	1.305
PON creation	SM-KZIDI	SM-JLV7U_NA18526_Oct2019	46,799,610,588.00	91.552	95.16	1.353

Clinical Somatic Exome sample list utilizing the Broad custom exome panel.

Varying Process Steps per Sample Type

The different samples require slightly different processing. For example, differences in cfDNA processing do not require shearing because the DNA is already fragmented to approximately 160bp. In contrast, gDNA does require shearing to approximately 150 bp size. There are best practices for handling certain sample types and input, and the laboratory requires systems to recognize the product type to determine the correct processing steps. Currently, there are approximately seven different products they are running: Exome for Cell-Free Liquid Biopsy from cfDNA ULP Libraries v6, Exome Express for Cell-Free Liquid Biopsy from non-cfDNA ULP Libraries v6, Exome

for Cell-Free Liquid Biopsy from non-cfDNA ULP Libraries v6, Express Somatic Human WES v6 with a 28-day turnaround time, Exome v6.0 Somatic Human WES - Research, Exome v6.0 Somatic Human WES - Clinical, Exome v2.0 - CLIA Somatic Exome using Illumina Content Exome (ICE) and the various custom panel products.

Each product has different deliverable targets and turnaround times listed in Table 7. The deliverables accommodate the different sample types and depth of sequencing each product may need. Turnaround times for sample intake to data delivery are determined based on the requirement and priority of the sample. Clinical samples that will impact patient treatment require a shorter turnaround time in contrast to a large-scale research project consisting of thousands of samples that can require a longer turnaround time.

Methods for Production Implementation

The hybrid selection protocol was then altered to fit the needs of the large-scale production lab with automation and sample tracking. The somatic exome team runs approximately six hundred or more samples of differing types per week, and the hybrid selection protocol was optimized to run at a customizable but high-throughput scale. In order to run samples at high scale, protocols require samples to run in 96-well format, on an Eppendorf twin.tec® PCR plate or Thermo Fisher Matrix™ 0.5mL ScrewTop Tubes in Barcoded Latch Racks. Samples are then processed on automated liquid handlers for reproducibility and repeatability.

The implementation of the finalized product from research to production also required sample tracking through LIMS and significant work design changes. The full capture methods were automated on the Hamilton Starlet liquid handler and Agilent Bravo liquid handler. Each liquid handler was fitted to mimic the manual pipetting, including 2D flatbed scanners installed onto both liquid handlers to scan the 2D barcodes associated with each sample or reagent. Automated scripts were created to mimic the manual pipetting and includes labware such as Eppendorf twin.tec® PCR plate or Thermo Fisher Matrix™ 0.5mL ScrewTop Tubes in Barcoded Latch Racks. Heated reactions occur in a full 96-well Eppendorf Thermocycler at programmed cycles.

The hybridization set-up adds the reagent blocker and Human Cot-1and samples are then dried down together in the Biotage SPE-DRY 96-head lyophilizer. The lyophilized samples are then resuspended using the oligonucleotide panel (exome or custom targeted panel) and hybridization buffer before going onto an overnight hybridization overnight for approximately 16 hours. Custom panels are registered into

the LIMS system and tracked during sample addition, to ensure downstream analysis is properly associated with the panel.

The capture protocol was designed for higher throughput as well. In place of using heat blocks during the manual protocols for each wash, an automated protocol on the Agilent Bravo with heat blocks at two locations (both approximately programmed to 72°C) mimicked mixing steps performed at heated temperatures manually. PCR steps, library quantification, and qPCR quantification were also performed on the Agilent Bravo. Sample transfers are documented in the LIMS system using 2D barcoded tubes and tracked before sample loading on the Illumina NovaSeq 6000 S4 flowcell, approximately 24 samples per lane to reach approximately 27 Gb of PF data per sample.

Chapter III.

Results

Targeted Panel Quality Control Results

The analysis of the laboratory's quality control steps ensures the designed custom targeted panels can capture the regions of interest. Specific gene regions may be important for different applications such as large research projects or precision medicine. Table 1 lists the sequencing quality control metrics reviewed to analyze that in-laboratory processing is within quality specifications and consistent with previous batches. The amount of sequencing, % selected, and target bases covered are reviewed. If the process control bait is not consistent with previous batches with %selected approximately 90-94%, there is a possibility of an issue within the lab, and the batch should be reworked. If the process control bait looks consistent and within specification limits, the targeted panel is evaluated for performance using the Picard sequencing metrics. A coverage analysis of the regions of interest can determine the performance of the specific targets the panel intended to cover. Table 8 provides the coverage analysis output from the target PanCancer Panel for three specific target regions. Out of the 18,323 probes, only 18 probes appeared to be under-covered. Table 8 lists three of the 18 probes with less than 20x mean target coverage. If the design does not capture the regions of interest with the anticipated coverage levels, the collaborator may re-design the probe panel and re-run the quality control test.

Table 8. Example of the coverage analysis output.

Target	**Chromosome**	**effLength**	**MeanCov**	**LCS.SampleMean**
chr7_101459311_101459373	7	63	18.3651	18.3065
chr12_69205237_69205371	12	135	17.2667	17.2667
chr22_22221612_22221730	22	119	16.563	16.563

The mean target coverage metric, column "Mean Cov" in the table, provides information on the sequencing coverage provided per target location within the genome (Table courtesy of Junko Tsuji, 2019).

Differences Between Capture Methods

The Exome v2.0 using Illumina Capture Exome (ICE) capture method included a UDI (unique dual index) library with a DNA of libraries input going into hybrid selection of 625 ng. This previous capture method utilized the Illumina Rapid Exome kit and required a solid phase reversible immobilization (SPRI) based concentration in order to create a pool volume small enough to start the first hybridization. The new selection method includes a duplex UMI-enabled (unique molecular indices) library with an input of 500 ng into hybrid selection. The capture kit is the IDT xGen Wash kit paired with the Broad customize exome from Twist Biosciences.

A major difference between the two workflows is the process timing (Figure 2). The ICE method required two hybridizations and workflow consisting of two eight-hour days of automated liquid handling, while the Twist requires a single overnight hybridization reaction to reduce the processing time to approximately six hours split between two days. Both protocols accommodate for the pooling of samples of similar

quality and quantity prior to hybridization to improve sample performance. Samples of similar input are pooled together, for example gDNA samples are pooled together while FFPE samples are pooled together in a separate pool. In addition, samples with similar PCR quantification values are pooled together to prevent uneven read coverage in downstream sequencing. Due to improvements in reagents and sequencing technology, the clinical somatic Twist IDT exome costs $600 versus the clinical ICE exome priced at $1200.

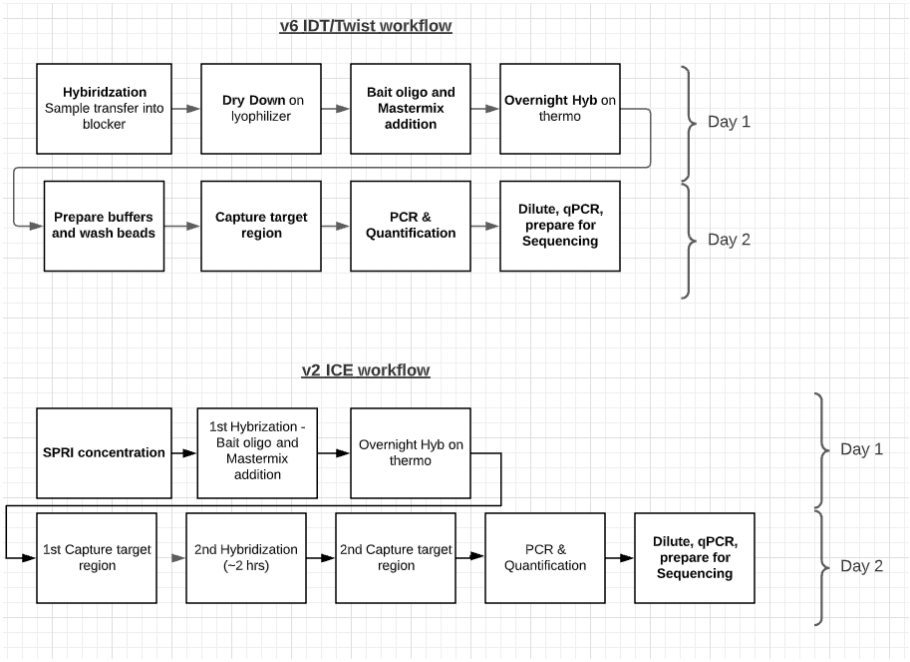

Figure 2. Differences between the Exome v2.0 and Exome v6.0 workflow.

The significant differences include the number of hybridizations and processing time.

Results of Research Validation: FFPE Input Titration

The main deliverable for the sequencing product of the somatic exome is 85% target bases at 100x. This metric indicates that the NGS output of target bases has 85% of the target region within the panel covered at 100x. The depth of coverage at 100x determines how much sequencing and number of reads the sample received.

The research validation tested the varying inputs for FFPE tumors and normals, gDNA samples, and cfDNA samples. For the FFPE derived data with varying inputs of 100ng vs 300ng, samples were booked over 1 lane of NovaSeq 6000 S4 (Figure 3). 19 of the 24 samples received over 85% target bases at 100 x with approximately 15Gb of bases. The 5 samples that did not meet coverage had lower coverage in the sample pool. The %selected for all the samples was greater than 93%, indicating that 93% of the bases aligned were either near or on the bait region. This higher percentage of reaching the baited region of interest allows for less sequencing to reach coverage. The FFPE samples had varying quality, and different inputs of 150 ng or 300 ng do not have a statistical significance on the data output.

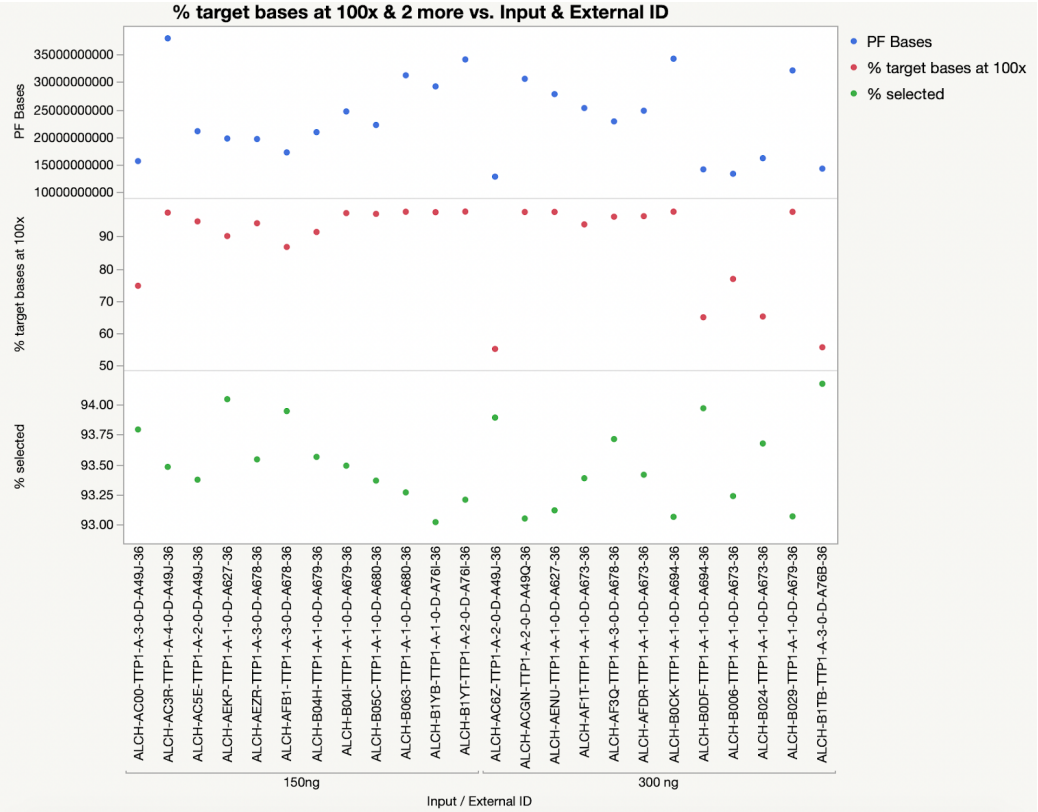

Figure 3. Results of the input titration for the Research FFPE samples.

Results of Research Validation: gDNA Results

Figure 4 shows the sequencing results from the gDNA normal samples with 100 ng input. 42 of the 44 samples met coverage of 85% target bases at 100x. Of the two samples that were low, one sample dropped out significantly at 47.7% target bases at 100x and another at 77.6% target bases at 100x. The smaller amount of data it received is due to the smaller sample representation in the pool, indicating that the samples may have had a smaller volume aliquoted into the pool in comparison to the other samples. However, both receive high %selected and slightly higher zero coverage targets. Overall, despite the uneven pooling, the gDNA normal samples performed consistently due to their high-quality input.

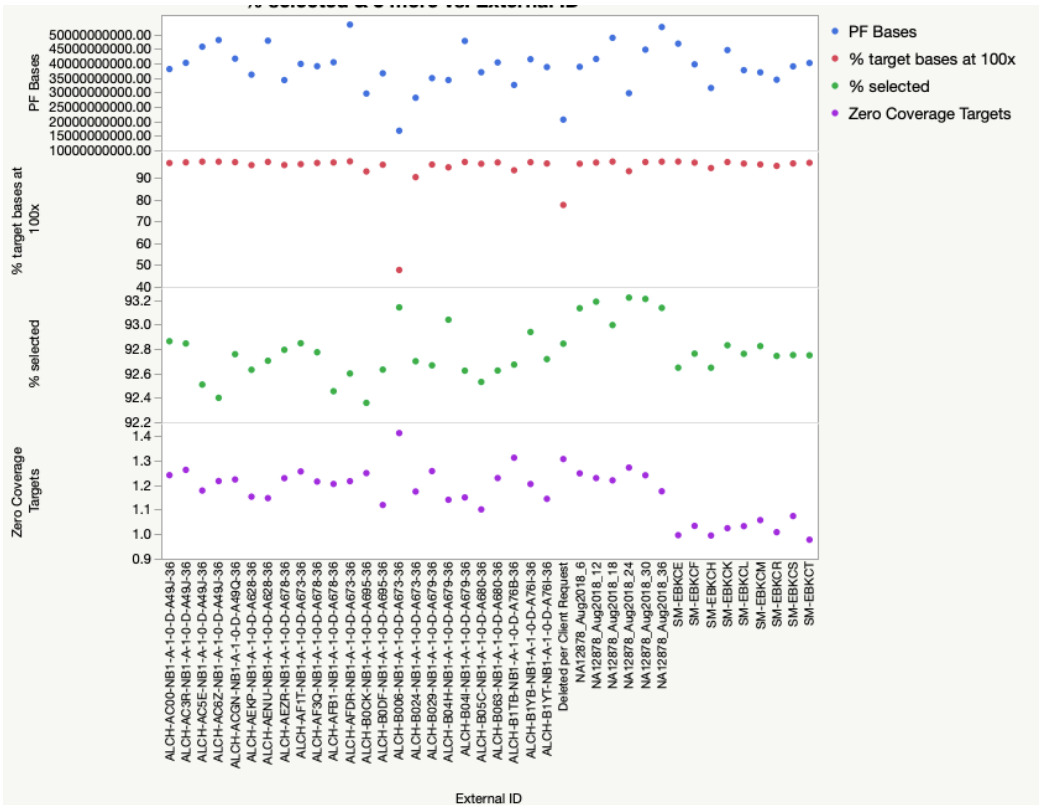

Figure 4. Results of gDNA normal samples with 100ng input.

Results of Research Validation: cfDNA and gDNA pair results

Figure 5 represents the data from the cfDNA and gDNA tumor-normal pairs. The cfDNA had an input of 10ng, while the gDNA normal samples had 150ng. The samples were pooled in a group of 20 cfDNA and 22 gDNA, with each pool receiving one lane on an Illumina NovaSeq 6000 S4 flowcell. The samples received various amounts of data, ranging from 25Gb to 60Gb, indicating the pooling volume for these samples were uneven. However, the percent selected bases for both cfDNA and gDNA were consistently averaging 92% and nearly every sample met 85% target coverage bases at

100x except one sample. The cfDNA and gDNA performance indicates that the new selection method can consistently capture regions of interest within the exome.

For all the research validations, the analysis pipeline used the BWA-MEM aligner with hg19, and samples were analyzed through the Broad Institute's Picard pipeline. The BWA-MEM aligner used for the validation has improvements such as alignment of fragments 100-1000 bp long, chimeric alignment, ability to handle long reference genomes, and is much faster than other aligners (Robinson, 2017).

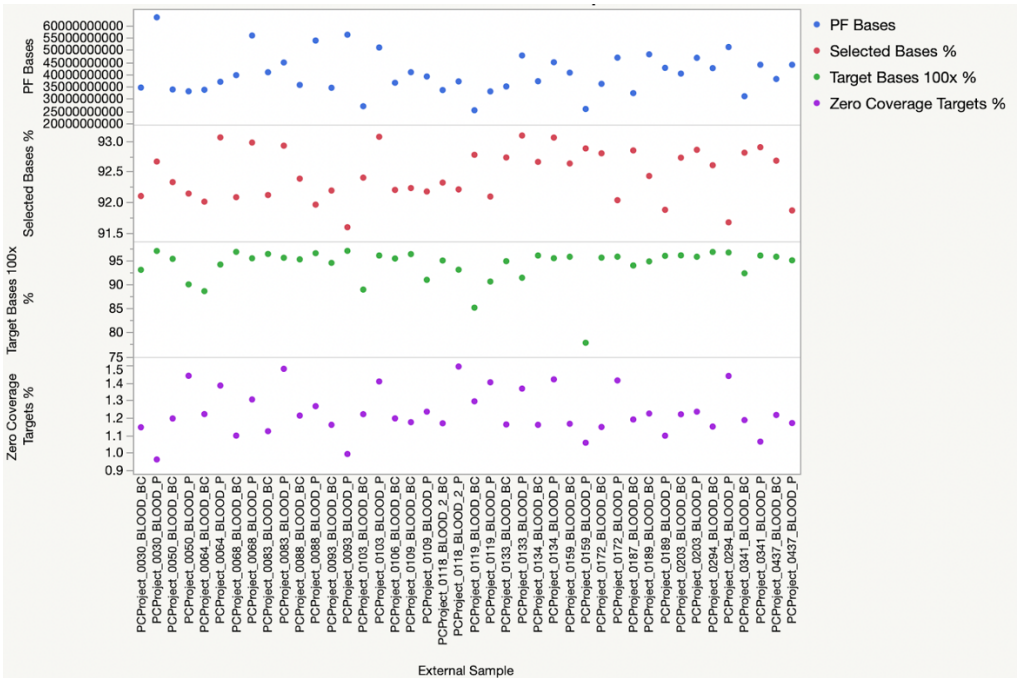

Figure 5. Results of cfDNA & gDNA pairs from Table 3.

Results of Clinical Validation

A collection of different sample types and biological specimens will be tested in the clinical validation. Multiple batches of samples will be analyzed with intra-run and

inter-run repeatability to control for batch effects and to prove repeatability and reproducibility of the clinical assay. The first batch of samples listed in Table 5 included the well-characterized NA12878 with an input titration varying from 10 ng, 50 ng, and 300 ng input run triplicate and HapMap pools consisting of multiple samples in the pooled stock sample: 5-plex, 10-plex, and 20-plex run in triplicate. Seven extra HapMaps were added to the pool to create a full 24-plex to mimic production sequencing runs and included for PoN creation.

The samples listed in Table 5 received 29 to 60 Gb through sequencing, reaching the minimum amount of approximately 27 Gb. All samples received the 85% target bases at 100x except one replicate of the 5-plex pool that received 80% target bases at 100x. The only noticeable difference was that this sample had a smaller insert size than the other samples. All samples had consistent % selected metrics at approximately 91%. The extra seven HapMaps will be added to the Panel of Normals (PoN). The PoN is a set of normal samples from healthy individuals used as a comparison in somatic variant calling analysis and can be used for accuracy and noise elimination in analysis.

The data analysis included a comparison of Illumina's Dynamic Read Analysis for Genomics (DRAGEN) pipeline to the current Picard pipeline utilizing the BWA-MEM to compare sensitivity and specificity. For sensitivity of the assay, the HapMap 5, 10, and 20-plex pools were run in replicates to compare against previous truth data. DRAGEN had improved performance for SNV and INDEL detection. To test for specificity of the assay, the NA12878 replicates were analyzed using the Broad BWA-MEM with Picard pipeline against DRAGEN with hard-clipping. The DRAGEN analysis uses the PoN to help remove false positives. Overall, DRAGEN with hard-

clipping had improvements in sensitivity and specificity. Further analysis analyzing batch 1 samples showed that the uniformity across the exome for both FFPE and cfDNA were more even for the Exome v6.0 with the Broad custom exome panel created by Twist at a 50x coverage, as seen in Figure 6.

Other tumor-normal pairs will be processed in the reproducibility and repeatability testing batches including FFPE tumors with blood normal samples, FFPE tumors with saliva normal samples, fresh frozen tissue with blood normal samples, fresh frozen with saliva normal samples, and FICOLL Blood (PBMCs) with saliva normal samples. NA12878 replicates will be analyzed to prove repeatability through the different processing batches, and ongoing analysis pipeline improvements and the creation of a larger PoN will continue to improve analysis pipelines.

Increased Throughput

The newly designed workflow allows for increased throughput through the lab. Multi-plexing the indexed libraries before hybrid selection increases the sample reactions per well, in addition to a reduction in oligonucleotide reagent cost. To accommodate increased scale per 96-well plate, sample tracking and index barcode association were enabled on automated liquid handlers. The automated protocols can pool indexed libraries into the same well, while also checking and preventing libraries with the same barcodes from getting pooled together. Furthermore, the use of 2D barcode scanning for samples and targeted panels to ensure the correct oligonucleotides are added to the correct sample. These features allow samples of differing panels on the

same hybridization plate, increasing the number of samples to be processed on automation.

Chapter IV.

Discussion

Ultimately, the decision to develop and implement the new hybrid selection method was made upon several factors including a phase-out of the Illumina Rapid Exome kit used in Exome v2.0 but most importantly, the improved data quality and workflow improvements. These included flexibility within the protocol to utilize a targeted oligonucleotide from any vendor, with no limitations on the size of the targeted panel, flexibility of the input of library (gDNA, cfDNA, FFPE), and the ability to choose the sequencing coverage output. While each product has a set coverage deliverable, the flexibility of the workflow allows customers to choose an increased coverage of a library or even choose to send their library through multiple assays (ULP, targeted selection, and exome selection) for a comprehensive analysis of their sample.

The research validation proved the Exome v6.0 version had better coverage over the Exome v2.0 due to the better coverage of Exome v6.0's custom exome panel. The NovaSeq 6000 sequencer provided increased data output at a lower cost due to the updated sequencer technology, allowing approximately 27 Gb of data per sample to reach coverage of 85% at 100x.

The clinical validation data provided further confirmation of the sample performance using the HapMap pools (5-plex, 10-plex, 20-plex) replicates and the well-defined cell line NA12878 replicates through DRAGEN analysis. The DRAGEN pipeline with hard-clipping showed improvement for both assay specificity and sensitivity. Future validation batches will be performed to further analyze sample types and data analysis pipelines are constantly improved upon for sample interpretation.

Table 9. Differences between Exome v2.0 and Exome v6.0.

	Exome v2.0	**Exome v6.0**
Library input	UDI-enable 150 bp fragments	UDI & UMI-enabled 150 bp fragments
Hybrid Selection panel and wash kit	Illumina Rapid Exome kit	Twist Biosciences Exome & Broad custom content. xGen® Hybridization and Wash Kit
Hybridization events	2 (1 of 2 overnight)	1 (overnight)
Sequencer & Cycle	Illumina HiSeq 2500 2x76 cycle	Illumina NovaSeq 6000 2x151bp
Processing time (Including incubations)	48 hours	36 hours

Importance of Cancer Samples

The improved Exome v6.0 product has a hybrid selection method that can handle not only the newly designed custom exome panel, but multiple different targeted panels with various DNA quality and DNA quantity input. With only slight differences in upfront processing, different DNA types such as gDNA and cfDNA can be processed on the same hybridization plate, preventing potential batch effects for tumor-normal pairs. Furthermore, the Broad custom exome panel from Twist has improved target regions specific for cancer regions. There are specific probe updates in regions such as TAL1 Enhancer, TERT Promoter, NOTCH 3'UTR, FOXA1 Promoter among others (Cibulskis, 2020). The increased coverage in these specific cancer regions will be beneficial in cancer research to determine mutations, copy number variations, gene expression, and methylation for different cancer tumor tissues (Weinstein *et al*. 2013).

Furthermore, somatic samples can have low DNA material after extraction, due to the cancer type or the tumor sampling method. These samples may have only enough DNA input for one attempt to create a library and be sequenced. These samples often represent critical time points in a patient's care, and it is vital to receive data for the patient or the research team's cohort study. Workflow design and error prevention steps are important for these samples. In a laboratory setting, unexpected errors and challenges can occur. Personnel are properly trained on laboratory protocols, how to operate the automated robotics, and learn to troubleshoot issues. The workflow designed is intended to minimize errors and risk to samples. The pooling and targeted panel checks within the protocol discussed in the Methods and Materials are examples of the error prevention features that can be implemented on automated liquid handlers. However, other workflow designs such as color coding 96-well PCR plates to identify a specific material or reagent are simple visual cues to the lab user. The different colored caps for the 0.5uL Matrix tubes are associated with certain custom panels, for example, the Broad custom exome panel has been standardized with a blue colored cap. These visual cues are simple ways to ensure efficient processing of samples and reduces user error, with particular care for somatic samples. The pull boxes discussed in the Methods and Materials section create the same physical locations with the fridge or freezer to visualize the amount of work queued in the system. These pull boxes create organization within the laboratory and visual queues for both the testing personnel and management operations.

Advantages of the New Capture Method for the Laboratory Workflow

The new hybrid selection method for targeted panels and exomes has several benefits including improved exome coverage, quicker process turnaround time, and decreased price. The turnaround time for oligonucleotide production from Twist Biosciences can be as little as 3-4 weeks, allowing the panel quality control test to be performed quickly as well. A coverage analysis of the region of interest can be analyzed, and collaborators can receive data quickly to understand if certain areas require higher coverage and alter the design. Another improvement is the versatility of the new method, which allows for various sample types and inputs to be run through a standardized capture method. The overall hybridization consists of one overnight hybridization and full processing time of approximately 36 hours as listed in Table 9. Different bait designs with different sequencing coverages can be processed on the same 96-well plate for increased throughput. Different cancer studies often require customization, high throughput processing, or a combination of both, all of which the Exome v6.0 method provides.

Sequencing data improvements for the customer for the selection method include more even sequencing coverage over the exome region seen in Figure 6. Utilizing the NovaSeq S4 flowcells that produce more data output, less sequencing is required to cover 85% of the exome at 100x coverage. For targeted panels that may require deep sequencing, the UMI-enabled libraries are able to help eliminate potential barcode swapping of the i5 and i7 index, known as cross-talk (MacConaill, L.E. 2018). The 6 nt UMI in the libraries can be analyzed to remove all cross-talk and help understand PCR

duplicates. UMIs are particularly important for somatic analysis such as somatic copy number alterations in cfDNA applications, where it is necessary to have obtained deep sequencing (MacConaill, L.E., 2018). Improved analysis pipelines, such as the alignment and variant calling pipeline DRAGEN, were tested and improved upon for implementation of the clinical Exome v6.0 method.

Due to more even coverage across the exome region and improved sequencing technology, the cost of sequencing has decreased for samples processed with the new Exome v6.0 capture method. The clinical cost of Exome v2.0 was approximately $1200 while the cost of Exome v6.0 is approximately $600. These developments have resulted in significant price decreases to support cancer researchers in the mission to understand cancer genomes and uncover more discoveries.

One application of the new selection method described here involves a metastatic triple negative breast cancer study (mTNBC) that collected ctDNA from patients throughout their cabozantinib monotherapy study (Weber, 2021). Blood draws for plasma were collected from 42 ctDNA samples from 35 patients and the ichorDNA algorithm was run to determine tumor fraction (Weber, 2021). For samples with >10% tumor, exome sequencing was run and for samples with <10% tumor, targeted panel of 396 genes PanCancer custom panel. The analysis shows single nucleotide variants (SNVs) were found in both the targeted panel sequencing and exome sequencing and is a proven method to show that ctDNA monitoring can be useful with potential for clinical applications (Weber, 2021).

As diseases such as cancer become increasingly complex to understand, there is a need for NGS technologies to continuously improve and develop new methods. Genomic

sequencing of gDNA and cfDNA can be utilized and paired with exome or targeted panels to create impactful sequencing data. cfDNA can be advantageous due to the less invasive blood collection method and its unique ability to sample tumor fractions at different time points. The targeted enrichment method using hybrid selection developed here allows for customization while also allowing for high-throughput sequencing, providing flexibility for cancer projects that vary in cohort size. Standardization and development of cfDNA liquid biopsies analysis is needed to allow future clinical applications (Wu, 2020). The analytical improvements for sensitivity and specificity and continuous ongoing analytical pipeline developments in the method discussed here will continue to aid in the cancer community's advancement toward therapies for patient care.

Appendix.

Additional Figures

Table 2. Research validation of FFPE samples with varying input of 150 ng or 300 ng DNA.

Sample ID	**Input**	**Quality**	**PF Bases**	**% Selected**	**%Target bases at 100x**	**Zero Coverage Targets**
SM-B4MKK	300 ng	FFPE	12,766,499,880	93.8914	55.2225	1.6089
SM-C1RV8	300 ng	FFPE	30,595,974,328	93.048	97.5244	1.3628
SM-GEAWD	300 ng	FFPE	27,805,047,672	93.1162	97.5458	1.3197
SM-GSXXI	300 ng	FFPE	25,271,007,140	93.3847	93.6926	1.4274
SM-GYUKA	300 ng	FFPE	22,836,301,212	93.712	96.0656	1.4039
SM-GSXXU	300 ng	FFPE	24,778,104,132	93.4138	96.2357	1.4465
SM-H4EML	300 ng	FFPE	34,257,229,176	93.0617	97.6242	1.3506
SM-H4EN7	300 ng	FFPE	14,091,857,804	93.9695	65.0003	1.5238
SM-HVG33	300 ng	FFPE	14,226,816,408	94.1739	55.7355	1.6216

SM-GSXY7	300 ng	FFPE	13,290,996,120	93.2355	76.8447	1.5168
SM-GSXYJ	300 ng	FFPE	16,128,156,304	93.6749	65.2417	1.443
SM-GYUN4	300 ng	FFPE	32,119,759,976	93.0657	97.5766	1.3755
SM-B4MJS	150ng	FFPE	37,971,186,228	93.4795	97.332	1.3178
SM-B4MK5	150ng	FFPE	21,047,013,980	93.3724	94.6292	1.4328
SM-B4MIF	150ng	FFPE	15,601,469,188	93.7921	74.7352	1.493
SM-GEAVN	150ng	FFPE	19,718,871,836	94.0453	90.0928	1.3887
SM-GYUJZ	150ng	FFPE	19,620,193,064	93.5423	94.0527	1.4191
SM-GYUKM	150ng	FFPE	17,188,709,940	93.9459	86.7262	1.4651
SM-HZ7VM	150ng	FFPE	29,213,027,288	93.0179	97.4657	1.3373
SM-HZ7VW	150ng	FFPE	34,125,802,312	93.2052	97.6567	1.352
SM-GYUO5	150ng	FFPE	20,878,248,200	93.5629	91.3581	1.3608
SM-GYUO7	150ng	FFPE	24,649,643,112	93.4894	97.1899	1.3006

Table 6. List of cell lines within the 5, 10, and 20-plex pools.

HapMap Pool / Plex	**5-plex**	**10-plex**	**20-plex**	
Cell Lines within the Pool / Plex	HG02922	HG01112	HG00096	NA18939
	NA19625	NA20845	HG00268	NA19017
	HG01583	NA19648	HG00419	NA19625
	HG00096	HG01595	HG00759	NA19648
	HG01500	HG00759	HG01112	NA20845
		HG01565	HG01595	NA20502
		NA19017	HG01500	HG01051
		NA18939	HG01565	HG01879
		NA20502	HG01583	HG03742
		HG00268	HG02922	NA18525

Each pool was created with approximately 500 ng of each cell line to create the pool.

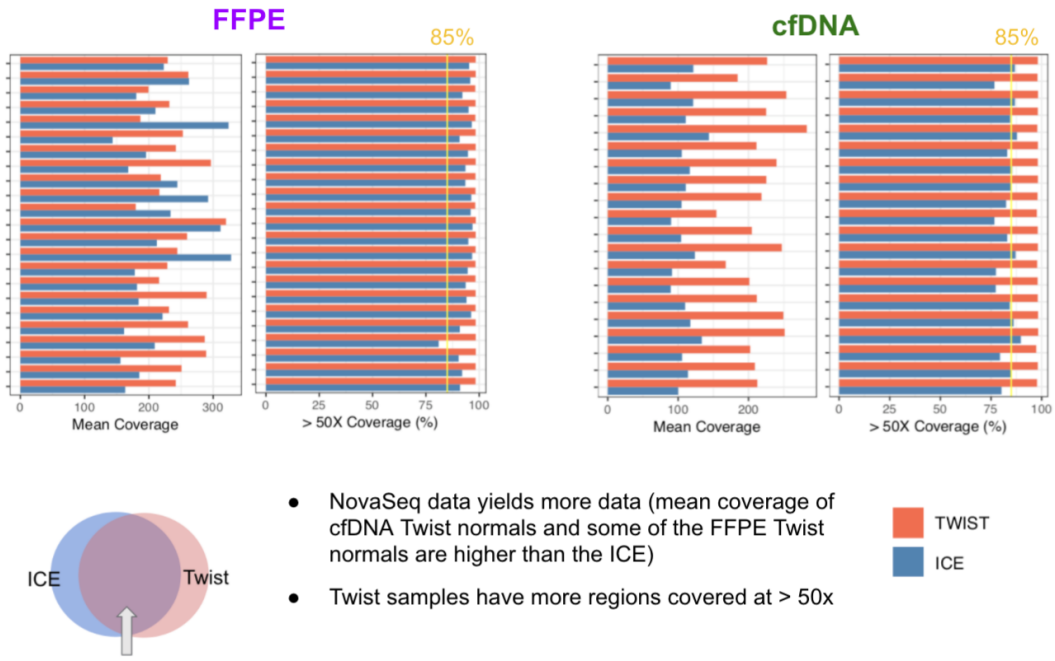

Figure 6. Coverage differences between Exome v2.0 and Exome v6.0 in FFPE and cfDNA.

The Exome v6.0 samples, labeled TWIST here, were found to have more uniform coverage at 50x across the selected region than Exome v2.0 samples, labeled ICE (Figure courtesy of Junko Tsuji, 2020).

Table 7. List of available exome and custom panel products.

Product Name	**Deliverable**	**Turn Around Time (TAT)**
Exome for Cell-Free Liquid Biopsy from cfDNA ULP Libraries v6	85% target bases at 100x	6-8 weeks
Express Exome for Cell-Free Liquid Biopsy from non-cfDNA ULP Libraries v6	85% target bases at 100x	28 days
Exome for Cell-Free Liquid Biopsy from non-cfDNA ULP Libraries v6	85% target bases at 100x	6-8 weeks
Express Somatic Human WES v6	85% target bases at 100x	28 days
Exome v6.0 Somatic Human WES - Research	85% target bases at 100x	6-8 weeks
Exome v6.0 Somatic Human WES - Clinical	85% target bases at 100x	21 days *In development
Exome v2.0 - CLIA Somatic Exome using Illumina Content Exome (ICE)	150x MTC	21 days
Custom panel products	Variable: 250x MTC, 500x MTC, 10,000x Raw coverage, 25,000x Raw coverage	6-8 weeks

Each product requires different processing due to sample type, workflow, TAT, and deliverable.

Bibliography

Adalsteinsson, V., Ha, G., Freeman, S., Choudhury, A.D., Stover, D., Parsons., Gydush, G, Reed, S.,…. Getz, G, Love, C., Meyerson, M. (2017) Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nature Communications*. 8:1324.

Bettegowda, C., Sausen, M., Leary, R.J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, BR., Wang, H., Luber, B., Alani, R.M., Antonarakis, E.S., Azad, N.S., Bardelli, A., Brem, H., Cameron, J.L., Lee, C.C., Fecher, L.A., Gallia, G.L., Gibbs, P., Le, D., Giuntoli, R.L., Goggins, M., Hogarty, M.D., Holdhoff, M., Hong, S.M., Jiao, Y., Juhl, H.H., Kim, J.J., Siravegna, G., Laheru, D.A, Lauricella C, Lim M, Lipson EJ, Marie SK, Netto GJ, Oliner KS, Olivi A, Olsson L,…. Papadopoulos N, Diaz LA Jr. (2014) Detection of circulating tumor DNA in early- and late-stage human malignancies. *Science Translational Medicine*. 6(224):224ra24.

Chen, Q., Zhang, Z. Wang, S., Lang, J. (2019) Circulating Cell-Free or Circulating Tumor DNA in the Management of Ovarian and Endometrial Cancer. O*nco Targets Ther*; 12: 11517–11530.

Chen, Z., Yuan, Y., Chen, X., Chen, J., Lin, S., Li, X., Du, H. (2020) Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Scientific Reports* 10, 3501.

Cibulskis, C. (2020) "Implementation and Performance Assessment of Somatic Exome v6.0". Broad Institute Internal.

Dash, S., Kinney, N.A., Varghese, R.T., Garner, H.R, Wu-chun, F, Anandakrishnan, R. (2019) Differentiating between cancer and normal tissue samples using multi-hit combinations of genetic mutations. *Sci Rep* 9, 1005.

DePristo, M., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas., M.A., Hanna, M., McKenna, A., Fennel, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43, 491–498.

Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T., Young, G., Fennell, T.J., Allen, A., Ambrogio, L., Berlin, A.M., Bluemenstiel, B., Cibulskis, K., Friedrich, D., Johnson, R., Juhn, F., Reilly, B., Shammas, R., Stalker, J., Sykes, S., Thompson, J., Walsh, J., Zimmer, A., Zwirko, Z., Gabriel, S., Nicol, R., Nusbaum, C. (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome biology*, 12(1), R1.

Garraway, L. A. and Lander, E. S. (2013) Lessons from the cancer genome. *Cell*, 153 (1), 17-37.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe., D., Lander, E., Nusbaum, C. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature biotechnology*, 27(2), 182–189.

Goldratt, Eliyahu M. The Goal: a Process of Ongoing Improvement. Great Barrington, MA. *North River Press*, 2014.

Hasin-Brumshtein, Y., Ramirez, M.C.M., Arbiza, L., Zeitoun, R. (2018) "The Importance of Coverage Uniformity Over On-Target Rate for Efficient Targeted NGS." Twist Biosciences White Paper.

Kokkat, Theresa J *et al*. (2013) Archived formalin-fixed paraffin-embedded (FFPE) blocks: A valuable underexploited resource for extraction of DNA, RNA, and protein. *Biopreservation and biobanking* vol. 11,2: 101-6.

Kulkarni, P. and Frommolt, P. (2017) Challenges in the Setup of Large-scale Next-Generation Sequencing Analysis Workflows. *Computational and structural biotechnology Journal* vol. 15 471-477.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

Lanman, R.B., Mortimer, S.A., Zill, O.A., Sebisanovic, D., Lopez, R., Blau, S., Collisson, E.A., Divers, S.G., Hoon, D.S., Kopetz, E.S., Lee, J., Nikolinakos, P.G., Baca, A.M., Kermani, B.G., Eltoukhy, H., Talasaz, A. (2015) Analytical and Clinical Validation of a Digital Sequencing Panel for Quantitative, Highly Accurate Evaluation of Cell-Free Circulating Tumor DNA. *PLoS One*. 10(10).

Lennon, N., Adalsteinsson, V., Gabriel, S. (2016) Technological considerations for genome-guided diagnosis and management of cancer. *Genome Medicine*, 8:112, 1-10.

LeProust, Emily. Rewriting DNA Synthesis. (2016) *Chemical Engineering Progress*, 30-36.

MacConaill, L.E., Burns, R.T., Nag, A., Coleman, H.A., Slevin, M.K., Giorda, K., Light, M., Lai, K., Jarosz, M., McNeill, M.S., Ducar, M.D., Meyerson, M., Thorner, A.R. (2018) Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics* 19, 30.

Manier, S., Park, J., Capelletti, M. et al (2018) Whole-exome sequencing of cell-free DNA and circulating tumor cells in multiple myeloma. *Nat Commun.* 2018; 9: 1691.

Parikh, A. R., Leshchiner, I., Elagina, L., Goyal, L., Levovitz, C., Siravegna, G., Livitz, D., Rhrissorrakrai, K., Martin, E. E., Van Seventer, E. E., Hanna, M., Slowik, K., Utro, F., Pinto, C. J., Wong, A., Danysh, B. P., de la Cruz, F. F., Fetter, I. J., Nadres, B., Shahzade, H. A., … Corcoran, R. B. (2019). Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers. *Nature medicine*, 25(9), 1415–1421.

Rehm, H., Bale, S. Bayrak-Toydemir, P, Berg, J., Brown, K., Deignan, J., Friez, M., Funke, B., Hegde, M., Lyon, E. (2013). ACMG clinical laboratory standards for next-generation sequencing. *Genetics in Medicine,* 15, 733-747.

Robinson, K. M., Hawkins, A. S., Santana-Cruz, I., Adkins, R. S., Shetty, A. C., Nagaraj, S., Sadzewicz, L., Tallon, L. J., Rasko, D. A., Fraser, C. M., Mahurkar, A., Silva, J. C., & Dunning Hotopp, J. C. (2017). Aligner optimization increases accuracy and decreases compute times in multi-species sequence data. *Microbial genomics*, 3(9), e000122.

Schwarze, K., Buchanan, J., Fermont, J.M., Taylor, J.C., Wordsworth, S. (2020) The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genet Med* 22, 85–94.

Schwarze, K., Buchanan, J., Taylor, J.C.,Wordsworth, S. (2018) Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet Med* 20, 1122–1130.

Weber, Z., Collier, K., Tallman, D., Forman, J., Shukla, S., Asad, S., Rhoades, J., Freeman, S., Parsons, H., Williams, N., Barroso-Sousa, R., Stover, E., Mahdi, H., Cibulskis, C., Lennon, N., Ha, G., Adalsteinsson, V., Tolaney, S., Stover, D. (2021) Modeling clonal structure over narrow time frames via circulating tumor DNA in metastatic breast cancer. *Genome Medicine.* 13.

Weinstein, JN et al. (2013) The Cancer Genome Atlas Research Network., Genome Characterization Center., The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 45, 1113–1120.

Wu, J., Hu, S., Zhang, L., Xin, J., Sun, C., Wang, L., Ding, K., & Wang, B. (2020). Tumor circulome in the liquid biopsies for cancer diagnosis and prognosis. *Theranostics*, 10(10), 4544–4556.

Yadong, Y., Dong, X., Xie, B., Ding, N., Chen, J., Li, Y., Zhang, Q., Qu, H., Fang, X. (2015) Databases and Web Tools for Cancer Genomics Study. *Genomics, Proteomics & Bioinformatics,* Volume 13, Issue 1, 46-50.