AN INTEGRATED FRAMEWORK FOR REPORTING CLINICALLY RELEVANT BIOMARKERS FROM PAIRED TUMOR/NORMAL GENOMIC AND TRANSCRIPTOMIC SEQUENCING DATA IN SUPPORT OF CLINICAL TRIALS IN PERSONALIZED MEDICINE

SARA NASSER[1], AHMET A. KURDOGLU[1], TYLER IZATT[1], JESSICA ALDRICH[1], MEGAN L. RUSSELL[1], ALEXIS CHRISTOFORIDES[1], WIABHAV TEMBE[1], JEFFERY A. KIEFER[2], JASON J. CORNEVEAUX[1], SARA A. BYRON[2], KAREN M. FORMAN[3], CLARICE ZUCCARO[3], JONATHAN J. KEATS[1], PATRICIA M. LORUSSO[4], JOHN D. CARPTEN[2], JEFFREY M. TRENT[2] AND DAVID W. CRAIG[1]*

[1]*Neurogenomics Division, Translational Genomics Research Institute,Phoenix, AZ 85004, USA*
[2]*Integrated Cancer Division, Translational Genomics Research Institute, Phoenix, AZ 85004, USA*
[4]*Yale Cancer Center, Yale School of Medicine,333 Cedar Street, New Haven, CT 06510*
[3]*Barbara Ann Karmanos Cancer Institute, Wayne State University, Detroit, MI*
**E-mail: dcraig@tgen.org*

The ability to rapidly sequence the tumor and germline DNA of an individual holds the eventual promise of revolutionizing our ability to match targeted therapies to tumors harboring the associated genetic biomarkers. Analyzing high throughput genomic data consisting of millions of base pairs and discovering alterations in clinically actionable genes in a structured and real time manner is at the crux of personalized testing. This requires a computational architecture that can monitor and track a system within a regulated environment as terabytes of data are reduced to a small number of therapeutically relevant variants, delivered as a diagnostic laboratory developed test. These high complexity assays require data structures that enable real-time and retrospective ad-hoc analysis, with a capability of updating to keep up with the rapidly changing genomic and therapeutic options, all under a regulated environment that is relevant under both CMS and FDA depending on application. We describe a flexible computational framework that uses a paired tumor/normal sample allowing for complete analysis and reporting in approximately 24 hours, providing identification of single nucleotide changes, small insertions and deletions, chromosomal rearrangements, gene fusions and gene expression with positive predictive values over 90%. In this paper we present the challenges in integrating clinical, genomic and annotation databases to provide interpreted draft reports which we utilize within ongoing clinical research protocols. We demonstrate the need to retire from existing performance measurements of accuracy and specificity and measure metrics that are meaningful to a genomic diagnostic environment. This paper presents a three-tier infrastructure that is currently being used to analyze an individual genome and provide available therapeutic options via a clinical report. Our framework utilizes a non-relational variant-centric database that is scaleable to a large amount of data and addresses the challenges and limitations of a relational database system. Our system is continuously monitored via multiple trackers each catering differently to the diversity of users involved in this process. These trackers designed in analytics web-app framework provide status updates for an individual sample accurate to a few minutes. In this paper, we also present our outcome delivery process that is designed and delivered adhering to the standards defined by various regulation agencies involved in clinical genomic testing.

Keywords: Genomic Testing; Next-Gen Sequencing Analysis; Personalized Medicine

1. Introduction

Cancer onset and progression leads to a variety of genomic events such as chromosomal aberrations and genomic mutations. Cancer develops through a stochastic process that produces cellular heterogeneity and structural complexity of the cellular genome. New advances in the depth and dimensionality of tumor profiling through the use of Next Generation Sequencing (NGS) have emphasized that there are in fact an under-appreciated number of acquired genetic changes and aberrations in cancer genomes that show clonal evidence of being selected. While the function of many of the mutations within a cancer are unknown and may represent passenger events, a subset of these are possibly biologically relevant and therapeutically actionable. Coupled with this knowledge and the ability to sequence a patient's genome within a clinically relevant timeframe, there is an increasing desire to utilize this vast amount data for patient care at an individual level.

National Cancer Institute lists hundred's of FDA approved targeted therapies which is further supplemented as an even larger number of therapies still under clinical investigation. Most notably, the BRAF inhibitors approved by FDA have demonstrated clinical response in patients with BRAF mutations.[1] More recently within our and collaborating research protocols, therapies targeting genomic events like FGFR fusions in Cholangiocarcinoma[2] and EGFR mutations in Lung cancer[3] have shown promising results and in some cases suggested that these targeted therapies instead of chemotherapy may be the best choice of treatment.

Studies to date are largely limited in scope, suffer from small numbers of patients, or therapeutic options, or lack the randomization or prospective design necessary to provide an unbiased assessment. One could critique that inability to access therapeutic drugs recommended by a group of experts (often termed a tumor board), or may indicate investigational agents only available through other trials. Determining the effectiveness of using genomics to inform therapy selection is the subject of numerous research studies, including several studies at the Translational Genomics Research Institute that have driven the development of data analytics pipeline achieving both technical, regulatory, and clinical goals. The pipelines we describe largely originate from two multi-year studies. The first is a study of the feasibility of using molecular-guided therapy for patients with BRAF wild-type metastatic melanoma (BRAFwt MM) as part of the Stand Up To Cancer Melanoma Dream Team. Detailed in other publications in preparation, the inclusion of investigational agents within the pharmacopeia (or compendium of drugs that could be indicated within a report), our pipeline was developed both to satisfy requirements within a laboratory developed test (LDT), regulated under the Center for Medicaid Services(CMS), as well as the Food and Drug Administration (FDA).[4] The second is a study of glioblastoma funded under the Ivy Foundation facilitating the development of the engine of rules identifying therapeutic options within this study.

As these examples show, the framework and mindset in code development for flexible platforms is regulated under multiple agencies. CMS regulates all laboratory testing (except research) performed on humans in the U.S. through the Clinical Laboratory Improvement Amendments (CLIA).[5] Particularly in the case of clinical research trials that impact care of patients, the FDA also provides regulatory oversite. Providing a framework that provides analytical validity for both FDA and CMS, requires understanding that precision, specificity,

sensitivity, reproducibility, repeatability be characterized with datasets that are often generated under a variety of conditions and truly characterizes the limits of detection. Understanding the regulatory environment is changing two involve multiple agencies, requires additional diligence with pipeline reporting, such critically as the ability to provide negative calls. For example traditional uses of the VCF format do not provide negative calls, and the ability to understand both false positives and false negatives is an increasing requirement for a field where the VCF specification could fall short. For example, there should not be ambiguity about where the lack of a variant is a 'no call' or a 'negative for BRAF 600V/E" below 0.01

The challenges of conducting trials and research in personalized genomics is elevated by the fact that next-generation sequencing data analysis requires intensive computational processing.[6] Effectiveness requires proper versioning, rapid turnaround, and reasonable disk requirements. A major aspect is whether the platform allows for implementing improvements and proper versioning within a clinical validated setting. Thus a CLIA certified lab needs to be equipped with high performance computers with additional security layers that can expedite genomic analysis. Additionally, interpretation of the findings requires accessing several genomic databases for annotations, pharmaceutical databases to gather gene-to-drug relations and tie in these databases to provide a clear and concise report.

Annotations and drug-gene matching process needs to be continually updated to include newly discovered variants and drug-gene matches. Further, tumor specific transcript variants exist and may require additional methods for detections and reporting. For instance, the EGFRvIII variant that has been reported in 30-40% of highly aggressive glioblastomas[7] is not available in most annotations and thus cannot be detected via usual methods.

In the following sections we provide an overview of the TGen Personal Genomics System and describe the framework that is being used for genomic testing and data delivery.

2. System Layout for Facilitating Clinical Research Trails in Personalized Genomics

The system developed to support a wide variety of clinical research trails for biopsy to report monitoring, tracking, and clinical reporting is a dynamic and flexible analysis framework which integrates genome and transcriptome sequencing data to inform interpretation of next-generation sequencing data on samples, specifically developed for cancer genomics. Our current platform is being utilized within a CLIA lab and has also been utilized under FDA within larger protocols. An overview is provided in Figure 1, which builds on stages supported by layers.

2.1. *Overview of Stages Going From Consent to Report*

The stages all begin with the patient and the doctor, who enroll and consent patients and provide a biopsy and/or blood draw to our clinical laboratory. This first stage is supported by a patient-centric custom-built clinical data portal. These portals and database described later in section 2.3 collect encoded, predefined non-public health information (PHI) for each patient. The second stage is specimen processing and accessioning whereby the specimens are evaluated for suitability and DNA/RNA analyte is isolated. Each step is managed by a

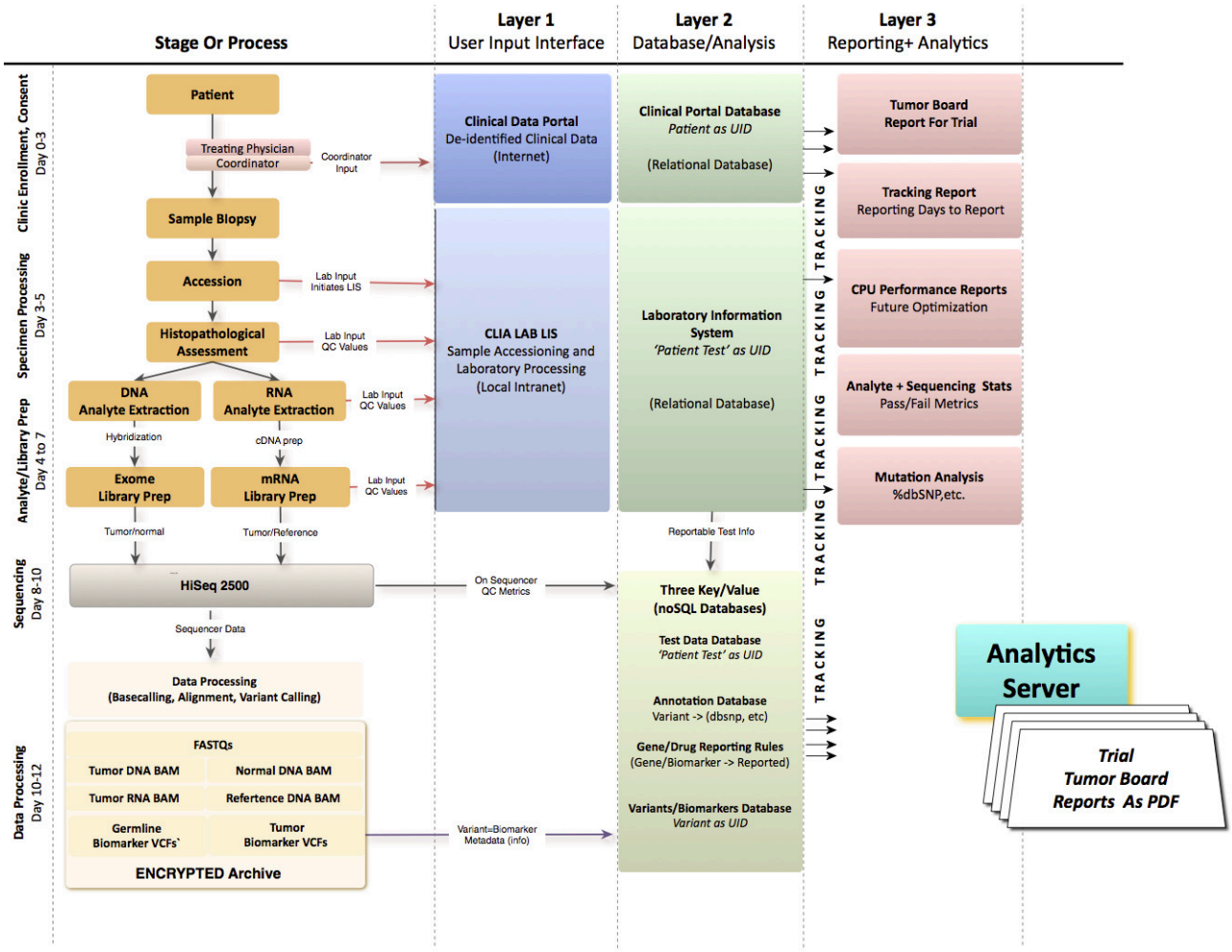

Fig. 1. Data Flow depicting the three layers of our system showcase the flow of samples from the clinic to sequencing and analysis. Finally, an interpretive report is sent back to the clinic

FileMaker Pro12[8] relational database. The third stage is library preparation and sequencing, again supported by a FileMaker Pro database which collects both images and data. As the Illumina 2500 sequencer writes to a shared disk system, data is collected into the relational database. The fourth stage is analysis and reporting, which is a framework built around (1) standardized file formats, (2) multiple quality control checks, (3) automated processing, (4) scheduled re- leases of sequence data, sequencing alignments, and variant calls, and (5) centralized primary data processing. As variants are generated, they are placed into a NoSQL database utilizing a document primary key as a ′biomarker′ .

Our framework is compatible with most external programs by plugging into allowing us to use the best tools and never be obligated to one tool, or feel required to develop an entire pipeline from scratch when components may be added according to their license agreements. Currently, we use external open-source style tools for mapping whereas the rest of the pipeline is largely built on internally developed software. As other groups develop muta-

tion detection tools, our framework will be compatible because of design. Our personalized medicine framework is a modular-based standard driven framework built to allow flexibility of adding/swapping out component analysis programs rapidly. It functions currently on Dell blade systems or appliances though are portable to other Unix-based environments. Fundamentally, many aspects are coded with map-reduce in mind allowing eventual porting to other frameworks. Component programs do not have to be proprietary, and thus other tools may be added if shown to be effective at identifying variants. It is optimized for oncology or cancer, though expandable to other areas such as neurological diseases. Current implementation is within a Torque-style queuing system common to many computing environments whereby jobs are monitored using background processes such that samples flow in an automated fashion to generation of reports.

2.2. *User Input Layers*

User interfaces are developed with the understanding they support trials at early stages which require prototyping by lab staff and clinical teams in an iterative manner before fixing for validation in use in a trial. User interaction happens at two levels: i) Clinical Data Portal and ii) Laboratory Data.

- **Clinical Portal**: The User Layer provides user interfaces to interact with the databases, extract reports and track information. To provide the flexibility for the research trials, pragmatic open source solutions are used, recognizing they are not necessarily appropriate for production clinical environments that are in open networks. Pragmatically, portals for clinical data are designed in WordPress^{TM} , a PHP management system tied in with an additional layer of security. WordPress utilizes extensive plugin framework that allows for easy addition and removal of features. The Clinical Data Portal (fields as shown in Table 1) are paired with their genomic counterpart for tumor board presentation.
- **Laboratory Data Frontend**: Laboratory and Sequencing data is entered in a user friendly FileMaker Pro database[8] described in section 2.3. FileMaker Pro′s flexible and user friendly framework provides ease to handle the massive amount of data that is generated during sequencing.

2.3. *Database/Analysis Layer*

NGS technologies provide a high-resolution and high- throughput approach to identify individual nucleotide bases from DNA samples. The goal of the NGS bioinformatics pipeline is to identify germline and somatic genetic variants events from tumor/normal pairs at the genomic (DNA) level, including coding point mutations and small insertions/deletions, copy number changes, and structural events (intra-chromosomal rearrangements and translocations).

An overview of the Analysis Workflow is provided in Figure 1. Briefly, each flowcell contains up to 4 tumor/normal pairs with an obligate reference control barcoded according to Illumina specifications (the control is described in detail in the Protocol section). Data is written from the HiSeq2500 to the scratch portion of a server in the form of BCL folders within the

Table 1. Clinical Data Portal: An example list of fields collected in the clinical data portal. Columns indicate sections in the portal and rows correspond to the fields within each section.*This is collected for all professions of the disease.

PatientSummary	*PrimaryDx*	*ProgressiveDisease∗*	*CurrentPresentation*
PatientIdentifier	SpecimenSize	SiteOfRecurrence	ComorbidConditions
Date of Consent	SatelliteNodules	DateofRecurrance	MenopausalStatus
PatientAge	StageT ,N,M	Surgery	PreviousCancers
PatientRace	Ulceration	SurgeryResponse	PriorTreatments
PatientEthnicity	Mitoses	SurgeryDate	DrugAllergies
PatientGender	ClarksLevel	SurgeryType	Medications
PatientSummary	MutationBRAF/NRAS/CKIT	RadiationLocation	Physical Exam
	LymphNodeInvolvement	RadiationType	Imaging/Radiology
	IFNType/Cycle	RadiationResponse	WBC/ANC/AlkPhos
	ClinicalTrialVaccine	TreatmentofRecurrance	Proteinuria

IlluminaRunFolders directory. An analysis run is triggered by the Clinical Laboratory Information System (DCLIS); depositing files within the ConversionArea folder that is processed into MergeSheets and SampleSheets. Using a queuing system and write FAIL/COMPLETED system BCL folders/files are converted to FASTQ files (raw sequence) and aligned to the genome using BWA-MEM[9] followed by a standard best-practice cascade of variant calling software tools.

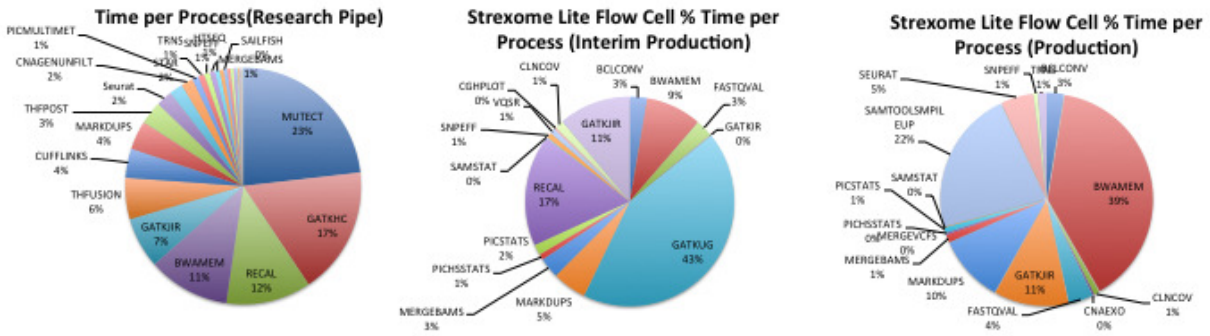

Fig. 2. Percentage of CPU Times utilization listed for three different pipelines highlight the importance of optimization. In the research pipeline, "Mutect" which was not optimized took the largest chunk of CPU hours. The middle chart display the CPU time distribution for the Interim Pipeline: germline variant caller GATKUG caller was taking a large chunk of CPU hours, replacing it with Samtools significantly improved turn-around time.

2.3.1. *Databases*

At the backbone of our infrastructure sits a database layer, which consists of annotation databases, drug rule matching databases and clinical databases. Although, each database is

built in a different environment to cater to its users, we aggregate the relative information so it can be used for reporting and tracking.

- **Sequencing Database** Sequencing data is built in a user friendly FileMaker Pro database.[8] This relational database captures relations between "Patients", "Samples", "Orders" and "Sequencing Statistics". FileMakerPro′s audit trail allows realtime tracking for document changes which are critical for a CLIA lab.
- **Annotations** Annotation is handled by committing sources of annotation into the database (typically by first exporting a text copy of the public database), and then performing an annotate action on the variant collection. The annotate action appends additional fields to all variants that are in annotated regions. We chose to handle annotation this way for two reasons: It allows for the natural retention of a "snapshot" of the annotation version that was used for the analysis. Many public variation databases are frequently modified, and some do not follow a strict versioning scheme that uniquely identifies a copy of the data. This can help with maintaining the repeatability of an analysis. Databases of many forms can be represented easily. Inserting an annotation as a field on a variant record allows us to create queries that use it. Solving the data representation and access problem of modern sequencing helps sustains the progress of genomics by a) using expert and analysis time more efficiently and b) by allowing even small labs to perform complicated knowledge extraction from the abundance of genotypic information that is available.
- **Biomarker Database** This database will integrate the disparate gene state annotation data and compile all available genomic information to facilitate the efficient and effective access for knowledge mining of the various dimensions of gene states. We will also include genes that become evident only from integrated analysis of genome and transcriptome analysis, such as a gene mapping to a large hemizygous deletion region and contains an obviously inactivating frameshift or nonsense mutation in the retained allele. In cases, where RNA data is present we integrate RNA Allele specific expression for the variants.
- **Statistics Database** This is a project centric noSQL database that holds a sample's sequencing information and collects all statistics that are used in determining the quality of the run. Statistics such as total bases aligned, mean target coverage help decide the accuracy of the alignment. In addition, performance metrics are also collected for each run, thus allowing analytics report generation (discussed in section /citerep).
- **Drug Rule Matching** We developed a conceptual framework for annotating the relationships between genomic alterations and drug response. This drug-rule matching algorithm identifies drug candidates for individual genomic alterations, including somatic mutations, indels, gene fusions, DNA copy number changes, and RNA expression changes, based on literature-curated evidence within a structured framework. The primary annotation source is PubMed publications, with other information sources captured when appropriate. New drug-gene associations from the literature will be added to the drug rule matching database through versioning.

2.3.2. *Genetic Variant Calling*

Our framework uses tumor and constitutional sample for somatic variants. Several tools have been developed to identify somatic events including Mutect,[10] Strelka[11] to identify somatic mutations, GATK,[12] samtools[13] callers to identify germline variants. Figure 2.3 compares our research pipeline, the interim production pipeline and production pipeline. Some software that become bottlenecks were identified in the research and interim-production settings and were optimized and/or replaced for realtime production.

2.4. *Reporting and Analytics Layer*

This layer allow end-users to interact with the framework at any given time point and generate interim reports. This layer is also responsible of sample tracking, delivery and maintenance.

- **Test Tracking:** A Clinical Genomic test consists of multiple processes that start with receiving the sample, followed by sample isolation, library preparation, sequencing, analysis and report generation. Users/Clients often are interested in tracking the status of their samples to estimate progress or get a priori information. We provide multiple trackers which are designed in JasperSoft^{TM}, one such tracker depicted in Figure 3 can only be visualized by authorized personnel authenticated using JasperSoft's authentication. *JasperReports Server* uses the Spring sub-project, Acegi Security, for authentication and authorization.[14]

Patient Name	Active Date	Test ordered	Samples Received	Extraction Started	Extraction Complete	Library Prep Started	Library Prep Complete	On Sequencer (Flowcell Created)	Analysis Started	Analysis Completed
SU2C Melanoma - C017-0002 C017_0002_20140610_T2_A1STX	2014-06-13 +110	Strexome	A01652-Tumor 2014-06-13	-	-	-	-	✓ 2014-07-02	✓	-
			A01601-Tumor 2014-06-11	A01632-Tumor 2014-06-12	✓	A02003-Tumor 2014-06-30	✓	✓ 2014-07-02	✓	-
			A01601-Normal 2014-06-11	A01632-Normal 2014-06-12	✓	A02003-Normal 2014-06-30	✓	✓ 2014-07-02	✓	-
UCSF PED GLIOMA - PNCO003-1 PNCO003-1 C021_0001_20140916_T3_TSMRU	2014-09-17 +14	Strexome	A03942-Normal 2014-09-17	A04022-Normal 2014-09-19	✓	A04100-Normal 2014-09-24	✓	✓ 2014-09-26	✓	-
			A03942-Tumor 2014-09-17	A04022-Tumor 2014-09-19	✓	A04100-Tumor 2014-09-24	✓	✓ 2014-09-26	✓	-

Fig. 3. An Internal Tracker provides information from sample receipt to final report generation. Each step provides an internal ID and the number of days utilized. Total Active Time is color coded providing quick information on the number of days since the sample receipt.

- **Archiving:** Every sample analyzed in the CLIA lab is encrypted and archived. Encryption is a slow process, thus archiving is run as a maintenance task. We use a two-level *gpg* encryption using asymmetric keys, the top level encrypts the entire package and a second layer of encryption is provided for documents that may contain patient-specific data.

3. Validation

FDA and CMS require extensive testing for repeatability and reproducibility. Thus each flowcell contains tumor/normal pairs with a barcoded reference control COLO829.[15] This reference control is used to validate a flowcell by generating performance measures for a run. Reporting

metrics is a challenge as we need to carefully consider the fact that traditional performance metrics might not apply in a marker-positive framework.[16] In Table 2, pre-production performance metrics are reported on the full range. Genomic tests conducted on approximately 3 billion base pairs of the human genome return only a small number of variants. Thus for these rare-events, the number of true negatives in the test will always be much larger than the true positives or false negatives. Taking into consideration only the reportable range addresses the issues in production. Table 2 also bring into light that Accuracy and Sensitivity (which are traditionally reported) might not be the best indicators of performance of a test in a genomic framework. This is illustrated by two examples where specificity is $> 99\%$ and sensitivity is at 50% in cases when the false positives are greater than true positives, whereas Positive Predictive Values(PPV) reports more reliable numbers.

Table 2. Performance Metrics on preproduction COLO829, hypothetical examples and production COLO829. Hypothetical examples indicate that sensitivity and specificity are not the best indicators of performance.

Sample	FN	FP	TN	TP	FDR	$Sensitivity$	$Specificity$	PPV
PreCOLO829_1	14	119	53535119	116	50.64%	89.23%	99.99%	49.36%
PreCOLO829_2	27	101	53535129	101	50.00%	78.91%	100.00%	50.00%
PreCOLO829_3	2	17	12716974	14	54.84%	87.50%	100.00%	45.16%
PreCOLO829_4	2	18	12716973	14	56.25%	87.50%	100.00%	43.75%
PreCOLO829_5	2	19	12716972	14	57.58%	87.50%	100.00%	42.42%
PreCOLO829_6	2	16	12716975	14	53.33%	87.50%	100.00%	46.67%
Hypothetical_1	15	17	12716975	1	94.44%	6.25	99.99	5.56
Hypothetical_2	1	20	12716975	1	95.24%	50.00	99.99	54.31
ProdCOLO829_1	23	25	1000000	258	8.83%	91.81%	99.97%	91.17%
ProdCOLO829_2	33	33	1000000	248	11.74%	88.26%	99.98%	88.26%
ProdCOLO829_3	26	11	1000000	255	4.14%	90.75%	99.97%	95.86%
ProdCOLO829_4	28	16	1000000	253	5.95%	90.04%	99.97%	94.05%
ProdCOLO829_5	26	20	1000000	255	7.27%	90.75%	99.97%	92.73%
ProdCOLO829_6	25	22	1000000	256	7.91%	91.10%	99.97%	92.09%

3.1. *Targeted Variant Detection*

Personalized Genome Testing is a growing field with emergent need for tools that are more focussed on actionable events. In this section, we illustrate with an example the need for development of such new tools. Analyses indicated that $92 - 94\%$ of human genes undergo alternative splicing, 86% with a minor isoform frequency of 15% or more.[17] EGFRvIII is a functional and permanently activated mutation of the epidermal growth factor receptor EGFR, a protein that contributes to cell growth and has been well validated as a target for cancer therapy.[18,19] Existing tools to detect isoforms presence using differential expression of a test and control sample,[20] these tools are quite promising but require much difficult to acquire RNA-Seq control. Additionally, Cuflinks[21] quantifies the transcript by read alignment and

Sailfish[22] provide expression levels for an isoform using an alignment-free approach. Table 3) contains FPKM/RPKM values reported by Cuflinks and Sailfish. Intuitively, a high R/FPKM value for this variant would imply its presence. In Table 3, GBM13 an experimentally validated sample reported an FPKM 165.354, whereas GBM 7 with no evidence of EGFRvIII reported an FPKM value of 820.109. Since these tools rely on the presence of reads in individual exons, they lack to provide evidence of a contiguous segment that defines this particular isoform. Thus R/FPKM values indicate the presence of reads across the whole region, as this region is a subset of the wild-type EGFR, the R/FPKM values might be misleading. Following the objective to detect the presence of variants (such as fusions, isoforms) we use a targeted de Novo assembly approach focusing only on certain region of the genome. In a clinical setting where only certain genes have actionable drugs, an approach to assemble and detect clinically actionable variants seems suitable. We have used this approach in past to correctly detect FGFR fusion.[2] A Denovo assembler Trinity[23] is used to assemble a region (a region is acquired by initial alignment), which are subsequently aligned to the reference genome for validation check using BWA-MEM.[9] Guided Assembly for this variant provides a clear response (Yes/No) which is required in clinical testing.

Table 3. EGFRvIII detection using Cufflinks, Sailfish and Guided assembly approach. GBM 6,7,12 all have a high FPKM value but upon examination evidence of EGFRvIII variant was not found for these samples. *EGFRvIII was experimentally validated for this sample.

	Cufflinks (FPKM)		Sailfish (RPKM)		Guided Assembly
Samples	EGFRvIII variant	Wild-type	EGFRvIII	Wild-type	EGFRvIII
GBM3	148.281	6.44284	71.1637	3.38737	Yes
GBM4	171.925	152.204	101.339	77.7003	No
GBM5	485.282	284.416	120.435	232.007	No
GBM6	550.976	432.228	238.849	175.429	Yes
GBM7	820.109	9.90112	382.24	6.24513	No
GBM8	150.059	43.8713	83.3531	21.5956	No
GBM10	266.228	31.7377	136.81	15.5846	Yes
GBM11	18.3666	0.321978	3.91242	0	No
GBM12	653.088	0.000121561	136.032	20.8175	No
GBM13*	165.354	0.000114743	12.5984	0.210588	Yes

4. Discussion

We have presented a three layer system, whereby inputs at layer 1 are through established prototyping solutions are supported by a second layer of automated interpretive engine utilizing intense knowledge mining genes that are known to be directly altered cancer, or who play critical roles in molecular mechanisms that are the targets of pharmaceutical agents. This organization allows for rapid prototyping, implementing, and analytically validating data analysis pipelines in support of open protocols and clinical research trials of personalized medicine built around a three layer design supporting data collection, analysis, and report delivery from

consent to reporting. The first layer leveraging of prototyping environments whereby experiment and clinical collaborators can design interfaces, fitting in with a structured relational and non-relational databases provide a capability of data collection and tracking to create a development cycle that is both agile and rationale.

In a second layer, we describe a framework where both relational and non-relational databases are used to collect all information linking to a patient as it moves through various stages to the identification of biomarkers. Use of non-relational key-value document stores through non-relational databases provides design flexibility of the first layer. Overall, utilization of both relation and non-relational document store databases is based on integration around two these two key concepts 'patients' and 'biomarkers', for which all other concepts depend. Moving across multiple studies with this 'variant' or 'patient' centric conceptually works with with most bioinformatics pipelines that are focused on identifying tumor specific (somatic) mutations or biomarkers that compare or germline variants utilizing a variety of tools. A key is that two samples such as tumor and normal tracking to a single parent object. Overall, the mindset that our processes identify biomarkers' from 'patients' in a multi-step linear workflow, define joining as always building around 'patients' and 'biomarkers'. Utilization standardized reporting frameworks such as Jaspersoft at the third level provide environments that provide consistent and flexible reporting consistent with industry standards, fitting over the second database layer. Reports may be of the type that represent 'tumor board reports' or 'sensitivity/specificity' reports for supporting regulatory agencies. Reports may be tracking of where a specimen is in a system, or may be CPU utilization at a particular point in time.

Our assay captures coding mutations and structural events within cancer genes. The final output, an interpretive Cancer Panel Report, provides the physician with a list of agents that are associated with tumor specific DNA mutations. Importantly, mutations can have positive or negative correlations to drugs and our system highlights both. This unbiased report includes all relevant patient related information along with both basic and detailed information related to the tumor′s mutational spectrum, and candidate relationships with known therapies.

The framework and data structures we use as part of trials in personalized medicine are conceptually fitting into either three layers supporting a multi-step linear process moving from patient to biomarker sets the mechanism for how data is integrated and analyzed in support of patient of care. The modular-based standard driven framework allows flexibility of adding/swapping out component analysis programs rapidly, thus is not constraint by the tools used. Staying with goals of Bench to Bedside, our future direction is in developing and improving tools that focus towards clinical applications and integrating state-of-the-art software within the system.

5. Acknowledgements

This research was supported by Stand Up To Cancer: Melanoma Research Alliance Melanoma Dream Team Translational Cancer Research Grant (SU2C-AACR-DT0612). Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. Research reported in this publication was also supported by generous philanthropic contributions from the Dorrance Family Foundation and Ben and

Catherine Ivy Foundation.

References

1. P. B. Chapman, A. Hauschild, C. Robert, J. B. Haanen, P. Ascierto, J. Larkin, R. Dummer, C. Garbe, A. Testori, M. Maio, D. Hogg, P. Lorigan, C. Lebbe, T. Jouary, D. Schadendorf, A. Ribas, S. J. O'Day, J. A. Sosman, J. M. Kirkwood, A. M. M. Eggermont, B. Dreno, K. Nolop, J. Li, B. Nelson, J. Hou, R. J. Lee, K. T. Flaherty, G. A. McArthur and BRIM-3 Study Group, *N Engl J Med* **364**, 2507 (Jun 2011).
2. M. J. Borad and et al, *PLoS Genet* **10**, p. e1004135 (Feb 2014).
3. G. J. Weiss, W. S. Liang, M. J. Demeure, J. A. Kiefer, G. Hostetter, T. Izatt, S. Sinari, A. Christoforides, J. Aldrich, A. Kurdoglu, L. Phillips, H. Benson, R. Reiman, A. Baker, V. Marsh, D. D. Von Hoff, J. D. Carpten and D. W. Craig, *PLoS One* **8**, p. e76438 (2013).
4. U. Food and D. Administration, Paving the way for personalized medicine: Fda's role in a new era of medical product development:fda's role in a new era of medical product development.
5. C. for Medicare and M. Services, Clinical laboratory improvement amendments.
6. M.-P. Schapranow, *Analyze Genomes*, tech. rep., Hasso Plattner Institute.
7. C. A. Del Vecchio, C. P. Giacomini, H. Vogel, K. C. Jensen, T. Florio, A. Merlo, J. R. Pollack and A. J. Wong, *Oncogene* **32**, 2670 (May 2013).
8. Filemaker pro (2013).
9. H. Li, *arXiv* **1303**, p. 3997 (2013).
10. K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander and G. Getz, *Nat Biotechnol* **31**, 213 (Mar 2013).
11. C. T. Saunders, W. S. W. Wong, S. Swamy, J. Becq, L. J. Murray and R. K. Cheetham, *Bioinformatics* **28**, 1811 (Jul 2012).
12. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly and M. A. DePristo, *Genome Res* **20**, 1297 (Sep 2010).
13. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and 1000 Genome Project Data Processing Subgroup, *Bioinformatics* **25**, 2078 (Aug 2009).
14. L. T. Ben Alex Ben Alex, Spring security (July 2014).
15. E. D. Pleasance and et al, *Nature* **463**, 191 (Jan 2010).
16. L. Tang and X.-H. Zhou, *Stat Med* **32**, 620 (Feb 2013).
17. E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth and C. B. Burge, *Nature* **456**, 470 (Nov 2008).
18. M. W. Pedersen, M. Meltorn, L. Damstrup and H. S. Poulsen, *Ann Oncol* **12**, 745 (Jun 2001).
19. J. L. Munoz, V. Rodriguez-Cruz, S. J. Greco, V. Nagula, K. W. Scotto and P. Rameshwar, *Mol Cancer Ther* (Jul 2014).
20. Y. Katz, E. T. Wang, E. M. Airoldi and C. B. Burge, *Nat Methods* **7**, 1009 (Dec 2010).
21. C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold and L. Pachter, *Nat Biotechnol* **28**, 511 (May 2010).
22. R. Patro, S. M. Mount and C. Kingsford, *Nat Biotechnol* **32**, 462 (May 2014).
23. B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. Leduc, N. Friedman and A. Regev, *Nat Protoc* **8**, 1494 (Aug 2013).