Spring 5-15-2023

Bioinformatic tools to alleviate the annotation bottleneck within precision oncology

Erica Kay Barnell

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Molecular Genetics and Genomics

Dissertation Examination Committee:
Obi L. Griffith, Chair
Malachi Griffith
Meagan A. Jacoby
Timothy J. Ley
David Spencer
Lukas D. Wartman

Bioinformatic Tools to Alleviate the Annotation Bottleneck within Precision Oncology

by
Erica K. Barnell

A dissertation presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2023
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

The body of work presented here would not have been possible without the expertise, guidance, and support from so many individuals. I would first like to thank my undergraduate advisors Greg Storch, David Greenfield, and Mark Manary and my graduate advisors, Philip Needleman, Ira Kodner, and Aadel Chaudhuri for providing me with the courage to pursue a career in academic medicine. I would also like to thank the members of the Griffith Lab, with a special thank you to Katie Campbell, Zach Skidmore, Kelsy Cotto, Kilannin Krysiak, Alex Wagner, Arpad Danos, and Ben Ainscough. These scientists were instrumental to my training and gave me the strength to execute on my thesis research. Additionally, this work would not have been possible without direction from my thesis committee: Obi Griffith, Tim Ley, Malachi Griffith, Meagan Jacoby, Lukas Wartman, and David Spencer. I would also like to thank my parents, Michael and Vicki Barnell, Ken and Jeanne Newcomer, my brothers Andrew Barnell, Christian Newcomer, and Liam Newcomer, and my sisters Katie Rudolf and Annie Williams whose continuous support has helped me become the person I am today. And to Kenneth Newcomer - what an adventure it has been.

Most importantly, I would like to thank the patients and their families for their selfless contribution to the advancement of science. Thank you for allowing me into your lives so that we, together, may help others.

Erica K. Barnell

Washington University in St. Louis

May 2023

Dedicated to Catherine Barnell;

my husband Ken, my son Kenny;

and my sweet Ruby girl.

ABSTRACT OF THE DISSERTATION

Bioinformatic tools to alleviate the annotation
bottleneck within precision oncology

by

Erica K. Barnell

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics and Genomics

Washington University in St. Louis, 2023

Professor Obi L. Griffith, Chair

In the era of advanced ability to perform complex genomic sequencing, precision oncology has been adopted as the ideal paradigm for optimization of outcomes for patients with cancer. However, despite technological advances in all aspects of the massively parallel sequencing pipeline, the application of precision oncology to every clinical workflow has been unattainable. Suboptimal adoption of custom medicine within oncology is attributable to the annotation bottleneck, which currently demands inordinate manual and computational requirements for completion. Alleviation of the annotation bottleneck requires co-development of bioinformatic strategies and analysis knowledgebanks to automate variant identification and variant annotation for clinical utility. The body of work presented here provides validated methods to alleviate the annotation bottleneck within the precision oncology pipeline. The introduction describes the specific aspects of the massively parallel sequencing pipeline that require development. Subsequently, we present three tools (DeepSVR, a Manual Review Standard Operating Procedure, and OpenCAP) that were developed to improve upon existing methods for variant identification and annotation. DeepSVR provides a machine learning approach to improve automated somatic variant calling by reducing false positives associated with sequencing pipelines that are observable by manual reviewers. The Manual Review Standard Operating Procedure provides a systemic and standardized approach for manual review of aligned sequencing reads for sequencing data with paired tumor and normal samples. Finally, the Open-sourced CIViC Annotation Pipeline (OpenCAP) serves as a software to create rationally designed clinical capture panels that are linked to clinical relevance summaries to improve library preparation and clinical annotation. The combined utility of these three tools for alleviation of the analysis bottleneck are demonstrated using a clinical example. Specifically, we developed a targeted clinical capture panel (MyeloSeq) to evaluate recurrent mutations observed in myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML). The MyeloSeq sequencing pipeline incorporated many of the tools described above for variant identification and annotation and provides a succinct output report for physician consumption. When surveying physicians who utilize the MyeloSeq panel, we observed that over 44% of physicians changed their treatment protocol based on the MyeloSeq results. This included 39 new therapeutics prescribes, 4 definitive diagnoses, and 13 changes in treatment plan (stem-cell transplant versus chemotherapy) based on prognostic indicators. This example demonstrates that the developed tools help alleviate the analysis bottleneck within precision oncology and will improve physician’s ability to integrate precision medicine into clinical workflow.

Chapter 1: Introduction

1.1 The massively parallel sequencing pipeline can be used for precision oncology

1.1.1 Introduction to precision oncology and the massively parallel sequencing pipeline

Precision medicine is a broad topic that encompasses all aspects of customizing patient protocols to individuals. Precision oncology is the application of precision medicine to cancer patients. One example of precision oncology is the identification of variants that are characteristic to the individual's tumor and employing custom therapy protocols to improve outcomes. Over the last few decades, technological advancements and changes in the healthcare infrastructure have permitted the integration of precision oncology into clinical practice, which has improved the treatment for many patients with cancer.

Some initial research that highlighted the potential of precision oncology was derived from *The Human Genome Sequencing* project, which was completed in 2000 and required global collaboration from over 20 groups. Using a shotgun sequencing strategy, the project identified 30,000-40,000 protein-coding genes and 1.4M single nucleotide polymorphisms (SNPs).[1] The completion of this project was a revolutionary milestone, however the work required to fully analyze the information contained within the sequence proved to be more extensive than the sequencing itself, resulting in minimal direct clinical benefit.[1] In retrospect, the most valuable outcome from the project was the development of next-generation or massively parallel sequencing (MPS). This method for high-throughput sequencing was implemented near the conclusion of *The Human Genome Sequencing Project* and allowed for early completion of the project.[2] Since the development of high-throughput approaches, the sequencing cost of the human genome has plummeted from $100M in 2001 to $1,000 as of 2015.[3]

Reduction in the cost of sequencing has made it possible to develop sequencing-based approaches for diagnosis and treatment of disease. In theory, and in academia, precision medicine tools developed over the last 15 years have demonstrated success with regards to treating patients based on genetic, environmental, and behavioral factors.[4] Individual cases and small cohorts have provided substantial evidence that clinical practice and patient outcomes are improved when employing precision medicine to treat patients. However, it has been unattainable to scale this process for every individual.[5] This is in part due to the exponential growth of clinically relevant information that impacts patient care. For each patient, millions of base pairs must be sequenced, hundreds of variants must be called, and all must be evaluated for actionability. This process requires extensive computing power for data processing and storage, a high manual burden for variant identification and review, and a cohort of experts for summarization and execution.[6] Successful development of processes that improve these components within precision oncology will facilitate the incorporation of precision medicine into clinical practice.

The following chapter provides an overview of the MPS pipeline and existing obstacles that hinder the current workflow. Specifically, we describe methods for sample procurement, nucleic acid extraction, library preparation / targeted enrichment strategies, various sequencing strategies, somatic variant calling, variant refinement, variant annotation, and variant reporting. We describe the outstanding issues associated with each step in the MPS pipeline with a specific emphasis on the annotation bottleneck. The annotation bottleneck, which is described in detail below, refers to the manual processes of variant identification, variant annotation, and report generation. We ultimately propose solutions to alleviate the annotation bottleneck through development of bioinformatic tools that automate the involved steps. It is our hope that these tools could be used to increase the impact and utility of precision oncology for cancer patients.

1.1.2 Sample procurement

Sample acquisition requires obtaining a relatively pure tumor sample, ideally for comparison against a pure matched-normal sample. Variants identified in the tumor sample are typically compared to variants in the normal sample to label relevant mutations as either somatic or germline. Two broad classes of tumor samples include those from solid tumors and hematologic tumors.

Solid Tumors can be evaluated using a tumor biopsy or surgical resection. Samples can be analyzed fresh, or they can be subsequently prepared as formalin-fixed paraffin-embedded (FFPE) tissue. Although these approaches seem similar, the downstream analysis can be affected dramatically between different protocols. Processing in FFPE samples can potentially alter the DNA, which might change the variants called in the downstream analysis.[7] In addition to processing issues, many samples are a heterogeneous composition of normal tissue, immune cells, and sub-clonal tumor tissue.[8] The dynamic interaction between cancerous cells and adjacent healthy cells results in the blending of DNA and RNA transcripts from the various cell populations, thereby complicating the variant calling pipeline. Finally, underrepresentation of tumor cells in the biopsy can result in false negatives with regards to variant identification. This is especially true when evaluating solid tumors using a liquid biopsy approach (e.g., circulating tumor cells), low-burden tumors (e.g., sarcomas), and tumors that have variants of low variant allele frequency (VAF).

Hematologic tumors can be evaluated using a blood sample or bone marrow biopsy. Unique challenges with hematologic sequencing include reduced tumor tissue purity and lack of normal tissue purity (e.g., blood cell infiltration into the tissues).[9] Additionally, blood samples show differential escape of tumor cells from the bone marrow. Samples from the peripheral blood might differ from stem cells trapped in the bone marrow, which typically harbor primary mutations initiating the malignant transformation.[10] Lack of consistency with tumor sampling can prevent comparison across tissues or between different tumor types and can alter the downstream genetic analysis.

1.1.3 Nucleic acid extraction

Nucleic acid purification requires cell lysis, binding of nucleic acid, washing off non-nucleic acid material, drying of nucleic acid, and elution into a buffer or water. There are many commercially available kits that can perform manual nucleic acid purification and these steps can also be automated using commercially available equipment. Below each step is briefly described:

- **Lyse:** Tissue samples are typically extracted and stored as whole cells. The lysis step is used to disrupt the cellular membrane to expose the nucleic acid (DNA and RNA). Lysis buffer typically comprises a chaotropic agent, which breaks the hydrogen bond network between water molecules and optionally a surfactant to lower surface tension between membrane components and nucleic acid-containing solution. Some chaotropic agents can include: guanidium thiocyanate or magnesium chloride.
- **Bind:** After nucleic acid has been suspended in solution, it can be reversibly bound to a positively charged material for purification. These materials can include magnetic particles, columns, filters, silica beads, or organic solvent-based methods.

- **Wash:** Once the nucleic acid is bound to a positively charged material, remaining substances in the lysate are washed from solution. A washing solution does not disrupt the covalent bond between the nucleic acid and the positively charged material used for purification.
- **Dry:** To ensure proper elution, bound nucleic acid typically needs to be completely devoid of all liquid. This can be accomplished through evaporation or, to avoid degradation, alcohols have been used to expedite the drying step.
- **Elute:** Elution buffers are solvents that displace the nucleic acid from the positively charged material. These buffers are used for purification and concentration of the nucleic acid into a solution.
- **Cleanup:** Elutions can be optionally treated with either RNAse or DNAse to eliminate nucleic acid that is not being used in the pipeline. This further reduces noise associated with the sample and generates a purified solution of the material being analyzed.

After the nucleic acid generation step, assessment of quantity and quality of the final elution is typically performed. This can be accomplished using spectrophotometry and/or electropherograms. Spectrophotometry measures a substance’s ability to absorb a specific wavelength, which in turn is a proxy for concentration and purity. Electropherograms measure the nucleic acid concentration and size using a fluorescent spectrum. Both metrics can be used to ensure sample quality for downstream processing.

1.1.4 Library preparation and target enrichment strategies

Library construction is required to prepare input for the massively parallel sequencing (MPS) pipeline. This process consists of three steps: 1) genomic fragmentation, 2) ligation to custom linkers (e.g., adapters), and 2) polymerase chain reaction (PCR) amplification. Below each step is briefly described:

- **Genome fragmentation:** Fragmentation involves breaking the DNA into smaller pieces. Genomic fragmentation can be accomplished using physical or chemical means. Physical fragmentation methods include sonication, nebulization, or enzymatic reactions. Chemical fragmentation relies on hydroxyl radicals to break DNA into fragments. Relative to physical fragmentation, chemical fragmentation can accommodate more material, but can induce false positives through novel mutations or transversion artifacts.
- **Adaptor ligation:** Adaptors are chemically synthesized double stranded DNA molecules that make sequencing reactions possible. Adaptors are ligated to DNA fragments and may include sequences to allow binding to a flow cell, sequencing primer sites, sample indexes, unique molecular identifier (UMI) sequences, etc. These adaptors are ultimately sequenced and might require removal prior to alignment, depending on the alignment strategy.
- **PCR amplification:** The PCR amplification process creates many copies of a specific DNA (or complementary DNA) segment. PCR requires first denaturing double-stranded DNA (dsDNA) to create single-stranded DNA (ssDNA) using heat. Subsequently, primers bind to targeted ssDNA fragments and DNA polymerase initiates the elongation of ssDNA to create a copied dsDNA. Amplification is typically performed at multiple steps in the sequencing pipeline.

After PCR amplification, target enrichment strategies can be optionally employed to generate a more specific collection of DNA fragments for sequencing. These enrichment strategies are often performed on the constructed sequence library or incorporated into a library construction step. One type of target enrichment strategy includes hybridization capture. This process requires designing specific probes that bind to regions of interest for isolation (e.g., use of strepavidin Beads in combination with biotinylated DNA). Genomic DNA that is not bound to the capture probes will be washed away during elution. The remaining DNA, which is enriched for regions of interest, is amplified using PCR and sequenced. A second type of target enrichment strategy includes amplicon enrichment. Amplicon enrichment entails amplifying regions of interest by PCR using sets of primer sequences designed to target specific genomic loci. This method does not eliminate background noise but rather it preferentially amplifies regions of interest. There are several other types of targeted enrichment strategies that can be used for custom projects.

Library preparation and target enrichment strategies can also employ unique molecular identifiers (UMIs), which are short sequences or molecular tags that can increase acuity in downstream variant calling. Typically, these molecular identifiers are added prior to amplification to tag individual DNA molecules observed in the sample. This allows the individual to assign all amplification products to a single originating DNA molecule after sequencing. Through a process of consensus read formation, individual sequencing-related errors can be discounted, decreasing the effective error-rate of sequencing. UMI-based sequencing can take on many forms, each unique to the individual library preparation.

1.1.5 Massively parallel sequencing approaches

Sequencing is the final step in the data production part of a genomic analysis pipeline. The most commonly used sequencing technique is so-called next-generation (NGS) or high-throughput sequencing, which evaluates millions of sequences in parallel to dramatically reduce time and cost of the analysis. There are at least two platforms, that are approved for clinical utility, that harness the power of next-generation sequencing to efficiently evaluate tumor samples: 1) Illumina sequencing, and 2) ThermoFisher ION Torrent. Illumina sequencing anneals individual reads to a bead or plate using DNA adaptors and the molecule is amplified through PCR. Amplified reads are sequenced by individually adding single fluorescently tagged and blocked-nucleotides to the complementary DNA sequence and exposing the nucleotide to light to produce a characteristic fluorescence. These blocked-nucleotides can then be unblocked to allow for an additional base to bind and the process is repeated until the whole complementary sequence is elucidated. This platform has a high accuracy rate, can evaluate 50-300 base-pairs per read, and has a very high-throughput run (millions to billions of reads per flow cell). Each run takes approximately 2-3 days to complete for as little as $1,000 per 30x whole genome sample. ThermoFisher ION Torrent evaluates hydrogen atoms emitted during polymerization of base pairs, which can be measured as a variation in the solution's pH. This method has a low error rate for substitutions and point mutations, and it is relatively inexpensive with a fast turn-around for data production (2-7 hours per run). However, the platform has higher error rates for insertions and deletions, it cannot read long chains of mononucleotides, and it cannot currently match the throughput of the Illumina sequencing platform.

More recently, a newer class of sequencing technology, called third generation sequencing, has been developed to address several issues that currently exist with next-generation sequencing (NGS). Specifically, third generation sequencing platforms (e.g., PacBio and NanoPore) allow for sequencing of longer reads at a reduced cost relative to NGS-based approaches. PacBio utilizes hairpin adaptors to create a loop of DNA that can be fed through an immobilized polymerase to add complementary base pairs. As each nucleotide is held in the detection volume by the polymerase, a light pulse identifies the base. This platform requires high quality intact DNA with highly controlled fragmentation and can read strands up to 1Mb in length. Oxford NanoPore Sequencing utilizes biological transmembrane proteins that translocalize DNA. Measurement of changes in electrical conductivity as the DNA passes through the pore elucidates sequence reads. This platform can evaluate variable length reads and is inexpensive relative to other technologies. Specifically, the MinION device is completely portable, commercially available, and can evaluate 20-100MB per run. The tradeoff is its low fidelity rate of only ~85%.

For each sequencing platform described above, there are several broad classes of sequencing strategies that can be employed. This includes broad capture of the entire genome (whole genome sequencing), capture of all protein coding exons (whole exome sequencing), or targeted capture of desired loci of the genome (targeted sequencing). Other applications of sequencing include evaluation of transcribed nucleic acid (RNA-sequencing) and evaluation of nucleic acid released into the blood (circulating tumor cells, or cell-free DNA). These methods and applications of sequencing technology are expanding in the face of reduced sequencing costs and increased read accuracy and length.

1.1.6 Alignment and automated variant calling

Following generation of raw sequence read data, alignment to the reference genome is the next step within the sequencing pipeline. The reference genome approximates the complete representation of the genetic sequence for the 4 billion base pairs of human DNA. Using a representative assembly prevents the need to build an assembly each time a genome is sequenced, however, there are trade-offs to this approach. Specifically, due to single nucleotide polymorphisms (SNPs) (and large-scale variants) intrinsic to an individual, the reference genome does not perfectly match any one person. Further, due to repetitive elements (duplications, inverted repeats, tandem repeats), the reference genome is likely incomplete or incorrect in places. Therefore, new genome assemblies are constantly being built to improve our ability to resolve the true human genome sequence. Most recently, GRCh37[11] was published in 2009 and GRCh38[12] was published in 2013. Alignment to the reference genome can be performed using various alignment software and, generally speaking, alignment strategies can either optimize accuracy or processing time. Optimal solutions include either Smith-Waterman[13] or Needleman-Wunsch[14] alignment strategies, which are computationally expensive and process read strands slowly. Alternatively, fast solutions include hash-based algorithms such as Burrows-Wheeler transformation[15], which create shortcuts to reduce alignment time with minimal reduction in accuracy.

The next step in the typical cancer sequencing pipeline is to use paired tumor and normal alignments for germline and somatic variant calling. Germline variant calling consists of identifying single nucleotide polymorphisms (SNPs), insertions/deletions (indels), and structural variants (SVs) that are intrinsic to the normal tissue. An example of a germline variant calling software is the Genome Analysis Tool Kit (GATK)[16], which can be used for all types of variants described above. Somatic variant calling is a similar process, but it requires the variant to be exclusively observed in the tumor tissue and not present in the germline (i.e., normal) tissue. Somatic variant calling may involve looking for single nucleotide variants (SNVs), insertions and deletions (indels), copy number variants (CNVs), structural variants (SVs), and loss of heterozygosity (LOH), depending on the type of sequencing performed. These different types of variants can be identified by using various software (e.g., Strelka[17], MuTect[18], VarScan[19], Manta[20], and CNVKit[21]).

1.1.7 The impact of the massively parallel sequencing pipeline

The development of the massively parallel sequencing pipeline is increasingly sufficient for processing clinical samples in real-time for application to precision oncology. The main factors contributing to this, as outlined above, include reduction in cost of sequencing and full automation of the pipeline. As mentioned, the cost of sequencing continues to drop as technology evolves, which has been inversely correlated with implementation of MPS-based assays within the clinic. This is because successful reimbursement is a meaningful driver of change in healthcare. Demonstrating improvements in outcomes from sequencing approaches are still necessary to increase reimbursement rates for MPS-based tools. Automation of the MPS platform has also increased use of sequencing approaches in healthcare. To employ change in the clinic, physicians require that results are provided in a relatively short time frame (i.e., days). This ensures that the change in therapy induced by the sequencing results provides full benefit to the patient. Recent advancements in sequencing technology and software has permitted the development of pipelines that can generate aligned sequencing data in the requested time frame[5,22], which has increased the use of these pipelines in a clinical setting. These two factors have contributed to advancing diagnostic tools that directly impact patient care and patient outcomes.

1.2 The variant refinement and annotation bottleneck hinders adoption of precision oncology

1.2.1 The somatic variant refinement bottleneck

The output from automated variant calling software is a list of variants that deviate from the reference genome or between tumor and normal samples. Although each program has unique flaws and benefits, in general, one of the biggest problems with automated somatic and germline software is the high burden of false positives in the final report. Therefore, outputs from the automated software require substantial variant refinement prior to variant annotation. Specifically for somatic variants, somatic variant refinement is required to identify a high-quality list of variants associated with an individual’s tumor. This can be accomplished by employing heuristic filters on various sequencing metrics and / or manual review of aligned sequencing reads. Heuristic filtering can include setting minimum thresholds for sequencing metrics such as variant allele frequency (VAF), total coverage, allele read count, or allele read depth. For example, a typical sequencing strategy might recommend that variants require at least 20X coverage in both the tumor and normal sample with a VAF >5%. These numbers can be adjusted based on the experiment and the reagents employed on the samples. Manual review of somatic variants requires direct visualization of aligned sequencing reads using a genomic viewer such as Integrative Genomic Viewer (IGV).[23] This process can be used to manually filter out variants attributable to sequencing and alignment artifacts.

Although somatic variant refinement is an important part of the sequencing pipeline, there are many issues associated with the existing procedures. First, manual review it is incredibly time-consuming and expensive. To illustrate the existing manual review burden using a standard cancer genomics workflow, a previously conducted breast cancer study[24] will be used as an example. In this study, 10,112 variants were identified via automated somatic variant callers, 1,066 variants were filtered using heuristic cutoffs, and 9,046 variants required manual review. Assuming experienced manual reviewers can evaluate 70-100 variant per hour (personal observation), this study would have required >100 hours for manual review. Another issue with manual review is a high level of inter- and intra-lab variability / error due to individual bias and level of experience. This is likely attributable, at least in part, to underreporting of manual review procedures and lack of consensus on proper protocols for the process. Improvement in the variant refinement process in terms of standardization and systematization would prime the MPS pipeline for improved clinical utility.

1.2.2 The variant annotation bottleneck

After identifying a putative list of high-quality tumor-defining variants, variant annotation is required to understand the therapeutic, prognostic, diagnostic, and predisposing implications of the variants as they relate to the patient's tumor. This process requires both defining the clinical relevance of a variant and generating a clinical report that is easily accessible by researchers and physicians. Currently, the process of variant annotation is the largest bottleneck within precision oncology.[25] The annotation process requires a complex and expensive manual analysis to ascertain clinical relevance of genomic findings. Further, this process is not standardized across and between institutions.[26] Components required to improve the annotation bottleneck include co-development of bioinformatic tools and variant knowledgebases that effectively elucidate and annotate clinically actionable variants from sequencing data.

1.3 Existing approaches to alleviate the annotation bottleneck within precision oncology

1.3.1 Existing knowledgebases are insufficient for building precision oncology tools

For successful variant annotation, large knowledgebases must be generated to store the clinical data for consumption. There are currently many approaches for building variant annotation knowledgebases, each with their own benefits and drawbacks. The three main types of knowledgebases, which are described in detail below, include: government-sponsored variant-observation knowledgebases, academic variant interpretation knowledgebases, and industry variant interpretation knowledgebases.

Government-sponsored variant-observation knowledgebases have been developed to assist with open-access variant annotation. Examples of these databases include ClinVar[27]; The Cancer Genome Atlas (TCGA), Genomic Data Commons (GDC)[28]; and the Catalogue Of Somatic Mutations In Cancer (COSMIC)[29]. ClinVar is a crowd-sourced, free, and open knowledgebase that compiles genetic variants in disease. The database captures submissions of variant observations from clinical workflows across all human disease and with highly variable curation of clinical relevance (from none to expert-panel moderated). The GDC, COSMIC, and other similar databases systematically document observed cancer variation and provide a sense of their tumor-specific prevalence. Although some of these variant calls have external validation,[30] most do not provide any curated significance of their clinical relevance. Despite the fact that few variants within the databases have expert-curated interpretations and there is no focus on clinically actionable variants with regard to cancer, government-sponsored databases do have free public access and they are typically funded for regular updates and feature development.

Academic variant interpretation knowledgebases are knowledgebases that are generated and maintained by large academic centers or institutions. Currently, there are at least seven different academic knowledgebases that have been developed with a focus on clinical interpretation of cancer variants. These knowledgebases include: OncoKB[31], Precision Medicine Knowledgebase (PMKB) [32], and the Cancer Genome Interpreter (CGI)[33], among others. These knowledgebases were developed by experts within the field, and many have high quality data with interpretations for variants. However, the majority of these databases use a domain expert curation model, which is ultimately unsustainable as the knowledge within the field continues to grow. Additionally, most of these resources require log-in, have no public application programming interface, limit data use through restricted licenses, and some have no detailed or human-readable clinical interpretation or summary of variants.

Industry variant interpretation knowledgebases are commercially available databases that analyze tumor samples and provide information to clinicians on cancer variants. Foundation Medicine[34] and Guardant Health are two of the most relevant resources used within industry for variant annotation. Typically, industry platforms charge for use and are either reimbursed by the patients or by insurance companies. Although these knowledgebases have provided convincing evidence that they improved patient treatment, all data for these companies is restricted access and methods for variant identification and interpretation are typically unavailable. For example, FoundationOne CDx (produced by Foundation Medicine) does not publish methods for panel development and clinical variant annotation. The FDA filing states that the assay evaluates: 1) genomic locations associated with 15 FDA-approved drugs, 2) 324 genes of interest, and 3) two genomic signatures (microsatellite instability and tumor mutational burden). However, the specific genes, transcript versions, and hotspots used for evaluation are not well described. The lack of transparent molecular details regarding the commercial assays hinders advancements in research and development.

1.3.2 CIViC serves as an optimal knowledgebase for variant curation and annotation

Of all existing knowledgebases, we believe the Clinical Interpretations of Variants in Cancer (CIViC - www.civicdb.org) serves as the optimal knowledgebase for variant curation and annotation.[35] The CIViC database is a fully open, free-to-use knowledgebase, which incorporates clinical evidence associated with a biomedical publication. Evidence supporting specific clinical interpretations is gathered via crowdsourced curation followed by expert review and moderation.

All submissions, revisions, moderations, and comments on CIViC entries are tracked and displayed through the CIViC web interface, providing transparency and clear provenance of all content in the knowledgebase. The CIViC knowledgebase was built to permit both consumption (i.e., searching, browsing, and downloading) of existing entries as well as curation of new content. The knowledgebase has been organized into a four-level hierarchy: Genes, Variants, Evidence Items, and Assertions, whereby each level has its own knowledge model (**Figure 1.1**).

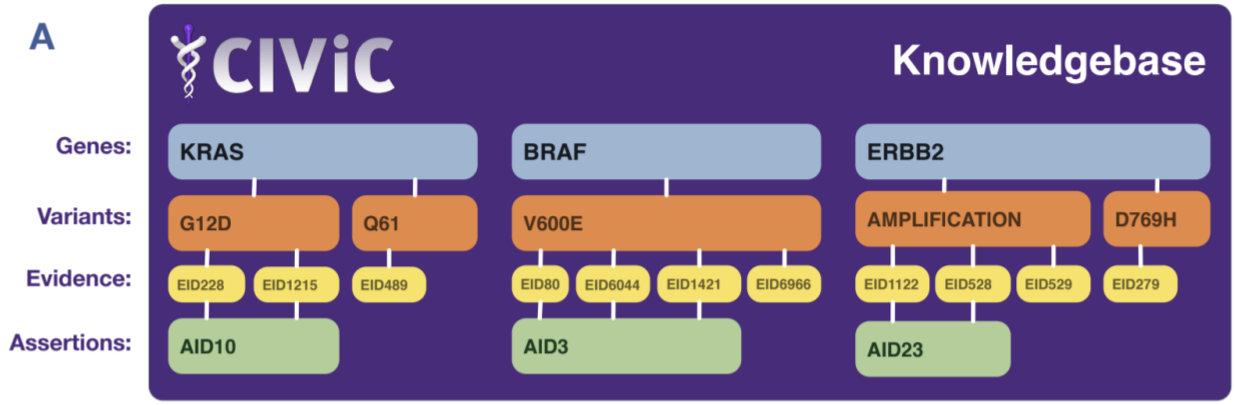


Figure 1.1 Overview of the CIViC knowledge model for the exploration of existing data and content curation. *The CIViC knowledge model consists of four interconnected levels that contribute to the content within CIViC: Genes (blue), Variants (orange), Evidence (yellow), and Assertions (green). Each broadly defined variant is associated with a single gene but can have many lines of evidence linking it to clinical relevance.*

All data created using these knowledge models are available through a web interface (www.civicdb.org) and an application programming interface (API, https://griffithlab.github.io/civic-api-docs). Additionally, all data within CIViC can be moderated via public curation. Adding content involves submitting new Evidence Items or Assertions that subsequently undergo revision and review. Editing content involves adding or revising the clinical summary and / or its associated features.

Within CIViC, a set of structured knowledge models have been developed to formally represent cancer variant interpretations. At the Gene-level, the CIViC interface displays the Gene Name, Gene Summary, an external link to The Drug Gene Interaction Database[36], useful citations on the overall clinical relevance of the gene, and link-out details from MyGene.info[37]. At the Variant-level, the CIViC interface shows the Variant Summary, Variant Type, HGVS expressions, ClinVar IDs, the Variant Evidence Score, representative Variant Coordinates / Transcript, associated Assertions, and external data from MyVariant.info[37]. These two CIViC features provide high-level summaries of the Gene and Variant Records that currently exist within the database.

The foundational unit of the CIViC knowledge model is the CIViC Evidence Item (EID). Evidence Items follow a structured model with 12 required fields (Gene name, Variant Name, Source Type, Variant Origin, Disease, Evidence Statement, Evidence Type, Evidence Level, Evidence Direction, Clinical Significance, and Trust Rating) and several additional optional fields (e.g., Associated Phenotypes). Based on the Evidence Type, additional required or optional fields become available (e.g., Predictive Evidence Types require a Drug Name). Evidence Items serve as an easily readable clinical statement derived from a publication to support a variant's implication in clinical relevance.

Within CIViC, clinical interpretation occurs at the Assertion-level. CIViC Assertions summarize the clinical relevance of a variant in a specific disease context using a collection of Evidence Items. Assertions are further categorized by type, whereby four different Assertion Types are supported in the CIViC framework: Predictive, Diagnostic, Prognostic, and Predisposing. Each CIViC Assertion include a Gene, Variant, Variant Origin, Disease, Assertion

Type, Assertion Direction, Clinical Significance, a short clinical Summary, and a longer Description. Fields unique to Assertions include annotation with clinical guidelines such as Association for Molecular Pathology (AMP) Tier and Level, American College of Medical Genetics and Genomics (ACMG) codes, National Comprehensive Cancer Network (NCCN) guideline/version, Food and Drug Administration (FDA) approvals/diagnostics, Associated Phenotypes, Drug names, and Drug Interaction Types.

At all levels of the CIViC knowledgebase, curation practices and clinical interpretations align with existing guidelines approved by cancer variant interpretation consortiums. For example, Predictive, Prognostic or Diagnostic Assertions utilize the somatic variant interpretation guidelines, providing an Association for Molecular Pathology (AMP) Tier (I-IV) and Level (A-D).[38] Similarly Predisposing Assertions utilize the American College of Medical Genetics and Genomics (ACMG) guideline classifications (Pathogenic, Likely Pathogenic, Likely Benign, Benign and Variant of Unknown Significance), their predicate ACMG evidence codes (i.e., PVS1, PP2, etc.), and rules for combining criteria.[39]

CIViC's unique framework could serve as a platform to impact the precision oncology pipeline. The structured data models are linked to a public API that can be accessed remotely to pull data required for variant annotation. Additionally, all submissions and revisions are public with links to the publication supporting clinical claims, which permits transparency. The database allows for external curation to alleviate the annotation bottleneck but requires expert moderation, which provides credibility to variant annotations. In summary, given the unique features of the CIViC knowledgebase, data within CIViC could be effectively used to automate variant annotation.

1.3.3 Genomic reporting and genomic literacy in healthcare

Once a list of annotated somatic variants has been identified for cancer patients, the information must be effectively communicated to oncologists. The ability to relay relevant findings to oncologists in a timely manner has been an incredible challenge within oncology.[26] Misinterpretation and overinterpretation of genomic findings is prevalent in the field for both direct-to-consumer genomic tests and those under laboratory developed test (LDT) regulation.[40,41] Additionally, interpretation of reports by physicians is highly variable, which leaves room for error in correctly applying clinical outcomes to treatment protocols.[42] Finally, there is minimal training regarding the communication of complex genomic information to patients. As a result, patients can become confused by genomic terminology and are unable to make informed decisions about their treatment and disease.[43] There is a need to generate standardized reports to facility oncologist and patient understanding of variant interpretation to improve the quality of care.

1.3.4 Other outstanding issues within the annotation bottleneck

With recent expansion of pan-cancer sequencing efforts, in both research and clinical settings, there has been a rapid increase in the number of variants that require clinical annotation.[44–48] Given the existing computational and manual requirements for variant identification and interpretation described above, there is a severe need to automate and normalize clinical classification of somatic variants.[49,50] This can be accomplished through building automated tools that incorporate automated approaches to improve annotation practice. Specifically, software can be built to eliminate false positive variants identified by traditional automated callers. Additionally, for the remaining variants that require manual review, a standardized guideline can be developed and validated to ensure that variants are being properly identified. Subsequently, existing databases can be used to build automated annotation software that generated variant annotation reports for physician use. For example, variant annotations contained by the CIViC knowledgebase can be used to build highly accessible variant annotation reports for integration into patient care. Utility of these reports can be validated by surveying oncologists who use genomic data for altering treatment protocols for their patients. This dissertation will attempt to address these issues through building and validating the aforementioned bioinformatic tools for alleviation of the annotation bottleneck within precision oncology.

Chapter 2: A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data

2.1 Preamble

The following chapter has been published:

Ainscough, B.J.*, **Barnell E.K.*,** Griffith, M., Rohan, T.E., Govindan, R., Mardis, E.R., Swamidass. J.S., Griffith O.L. Deep Learning Approach to Automating Somatic Variant Refinement. A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nature Genetics*, November 5, 2018.

As an author of the published manuscript, and in compliance with the editorial policies at Nature Genetics, the cited publication is included in full in the following chapter. The Author Contributions are included within the publication (**Chapter 2.9**); however, I would like to distinguish my role from the other first-author in this study. Both co-first authors contributed figures, tables, code, and assisted with packaging and implementing the ultimate model used for automated refinement. Ben Ainscough was responsible for collecting data and performing preliminary model development. I was responsible for understanding inter-reviewer agreement, validating the models using independent testing sets, and developing a tutorial/documentation for DeepSVR.

2.2 Summary

Cancer genomic analysis requires accurate identification of somatic variants in sequencing data. Manual review to refine somatic variant calls is required as a final step after automated processing. However, manual variant refinement is time-consuming, costly, poorly standardized, and non-reproducible. Here, we systematized and standardized somatic variant refinement using a machine

learning approach. The final model incorporates 41,000 variants from 440 sequencing cases. This model accurately recapitulated manual refinement labels for three independent testing sets (13,579 variants) and accurately predicted somatic variants confirmed by orthogonal validation sequencing data (212,158 variants). The model improves on manual somatic refinement by reducing bias on calls otherwise subject to high inter-reviewer variability.

2.3 Introduction

Somatic variant callers are commonly used to identify somatic variants from aligned sequence reads in cancer genomics studies and in clinical cancer assays.[5] These callers attempt to statistically model sample purity, sequencing errors, zygosity, ploidy, and other factors. Post-processing of called variants is an approach we term 'somatic variant refinement' and is an important, and distinct, next step from variant calling. Somatic variant refinement eliminates false positives from a candidate somatic variant list through heuristic filtering and manual review. Heuristic filtering includes setting project-specific thresholds for sequencing features such as read coverage depth, variant allele fraction (VAF), base quality metrics, and others. Manual review requires direct examination of aligned reads using a genomic viewer such as Integrative Genomic Viewer (IGV)[23,51] to identify false positives that are consistently missed by automated somatic variant callers.

Somatic variant refinement remains indispensable for accurate analysis of cancer data, especially as cancer genomics is brought into the clinic, where variants are used to guide therapy.[38,52] Manual reviewers look for patterns that are neglected or unavailable to standard variant callers to alter confidence in a variant call. For example, confidence is reduced if: all supporting reads are oriented in the same read direction; a variant is supported exclusively by overlapping reads from short DNA fragments; a variant is located in or near homopolymer stretches, short repeats, or other low-complexity sequences; supporting reads indicate multiple mismatches relative to the reference genome; variant support is identified in the normal data track; variant support occurs exclusively at the ends of sequencing reads; in addition to other factors. If the number of problematic variant reads at a locus is high, a reviewer may label a variant identified by a somatic variant caller as a false positive.

In our experience, somatic variant refinement can dramatically improve the quality of final variant calls by eliminating large percentages of false positives from automated callers. However, despite extensive use of somatic variant refinement in clinical and translational genomics, filtering and refinement protocols are usually unstated or only briefly mentioned. Some illustrative examples of this reporting are, "mutations … were called with MuTect and filtered with oxidation and panel of normal samples filters to remove artefacts,"[53] or, 'all indels were manually reviewed in IGV.'[54] These excerpts exemplify a prevalent history of under-reporting variant refinement details from our institute and others.[55–57]

Discrepancies in manual review procedures may result in significant inter- and intra-lab variability and error. To address the issue of reproducibility, our group generated a standard operating procedure for somatic variant refinement through the use of manual review.[58] However, even with complete conformity of manual review standard operating procedures, the process is time-consuming and expensive. Automated somatic variant callers can identify thousands of variants per cohort, which corresponds to hundreds of hours of manual review by a highly trained staff.[24] Machine learning could automate somatic variant refinement and essentially eliminate this bottleneck, reducing the required time and expense associated with variant identification.

Current software used for automated somatic variant calling includes VarScan,[59] SAMtools,[60] Pindel,[61] Sniper,[62] Strelka,[17] and MuTect,[18] among others. To improve on these algorithms, researchers have incorporated machine learning models to reduce the false positive rate intrinsic to automated somatic variant callers.[63,64] These initial attempts show promise for using machine learning approaches for somatic variant refinement; however, use of small training datasets (fewer than 3,000 variants) and limited number of cancer types prevents extrapolation of existing models onto a wide variety of sequencing data.[65]

Here we present a robust model that automates somatic variant refinement. We show that use of this model could substantially reduce a major bottleneck in cancer genomic analysis while improving reproducibility and inter-lab comparability in genomic studies and in clinical settings. This model is built on a training dataset of 41,000 variants from 21 studies, with 440 cases derived from nine cancer subtypes. All cases include paired tumor and normal samples that have been sequenced, evaluated for somatic variants using automated callers, and manually reviewed by individuals (an estimated 585 hours of manual effort). For each variant, we assembled 71 features to train the model including cancer type, sample type, tumor / normal read depth, tumor / normal VAF, base quality, mapping quality, etc. To our knowledge, this is the largest dataset assembled to develop a machine learning approach for somatic variant detection. This dataset includes both solid and hematological malignancies, covers a broad range of average mutation burden, and includes data from multiple different sequencing pipelines (**Table 2.1**). This broad representation supports the generalizability of machine learning for somatic variant refinement.

Table 2.1 Machine learning model development included a variety of different approaches. The cancer sequence data used to develop machine learning models included a variety of different tumor subtypes, sequencing approaches and manual review calls.

Malignancy	**Training set**	**Hold out test set**	**Total**
Leukemia (n = 243)	5,815	2,877	8,692
Lymphoma (n = 23)	1,263	628	1,891
Breast (n = 135)	8,986	4,320	13,306
Small-cell lung (n = 18)	9,177	4,601	13,778
Glioblastoma (n = 17)	844	412	1,256
Melanoma (n = 1)	185	100	285
Colorectal (n = 1)	842	419	1,261
Gastrointestinal stromal (n = 1)	70	31	101
Malignant peripheral nerve sheath (n = 1)	288	142	430
Total	27,470	13,530	41,000

Sequencing methods	**Training set**	**Hold out test set**	**Total**
Capture sequencing	9,479	4,755	14,234
Exome sequencing	9,367	4,677	14,044
Genome sequencing	8,624	4,098	12,722

Variant calls	Training set	Hold out test set	Total
Somatic	12,266	6,115	18,381
Ambiguous	7,189	3,454	10,643
Fail	5,909	2,945	8,854
Germline	2,106	1,016	3,122

2.4 Methods and experimental procedures

2.4.1 Training data

We assembled manual variant refinement data from 21 different cancer genomic studies conducted at the McDonnell Genome Institute (MGI), including 11 genomic discovery cohorts, 1 clinical trial, and 9 case studies.[22,55,66–75] Samples present in multiple studies were eliminated by removing all sample pairs with more than 70% co-occurrence of genomic mutations. In total, 440 sample pairs were evaluated, with 266 samples derived from hematologic malignancies and 174 samples derived from solid tumors (**Table 2.1**). Samples were only included if paired tumor/normal sequencing data and manual somatic variant refinement calls were available. Sequencing data from this cohort were analyzed using standard cancer genome pipelines at the McDonnell Genome Institute over a period of several years.[5] Briefly, sequencing data were produced using genome, exome, or custom capture sequencing. Reads were aligned to reference genome hg19/GRCh37 using Burrows–Wheeler aligner (BWA)[76] or BWA-MEM,[77] duplicates were marked by Picard,[78] and variants were called with SAMtools[60] or (predominantly) the union of SAMtools and VarScan.[19] Variants identified by automated callers were annotated and subjected to false positive filtering strategies such as removal of variants with low VAF (for example, < 5%) or low coverage (for example, < 20×). Much of the raw sequencing data from these 21 cancer genomic studies is publicly available, and all variant calls, manual review data, and associated features required for model development are provided in a publicly available GitHub repository (see **2.9 Data Availability**).

Manual variant refinement for all projects was performed by individuals at the MGI, who recently described a standard operating procedure for this process.[58] In this operating procedure, reviewers manually refine variants using four distinct classes: 'somatic'—a variant that has sufficient sequence read data support in the tumor in the absence of obvious sequencing artifacts; 'ambiguous' —a variant with insufficient sequence read data support to definitively classify the variant; 'germline'—a variant that has sufficient support in the normal sample beyond what might be considered attributable to tumor contamination of the normal; and 'fail'—a variant with low variant sequence read data support and/or reads that indicate sequencing artifacts, yet has acceptable variant coverage. In accordance with the standard operating procedure, as reviewers call variants, they often provide additional notes or tags describing the reason for each call.

Germline and fail calls represent two distinct types of failure for somatic variant calling. However, since germline and fail calls rarely invoke different downstream analysis procedures, they were merged into one class called 'failed'. Therefore, the machine learning model was developed for 'somatic', 'ambiguous', and 'fail' classes. All manual variant refinement results were standardized to a one-based coordinate system using the convert_zero_one_based Python tool. Relevant metrics were extracted from the bam files using bam-readcount. Bam file metrics were merged with cancer type and reviewer information. All continuous features were normalized to fall between 0 and 1 using Scikit-learn's MinMaxScaler.[79] All categorical variables were onehot boolean indexed to split any feature with n categories into an n column boolean array. Following processing, the training dataset included 71 features.

2.4.2 Model development and analysis

Logistic regression, random forest, and deep learning were tested as alternative models for somatic variant refinement. A logistic regression model was implemented using the keras library. Scikit-learn was used to implement the random forest model.[79] The random forest was trained using the parameters n_estimators = 1,000 and trees max_features = 8. The deep learning model was implemented using the keras library as a feed-forward neural network with the input layer equaling the number of features, four hidden layers with 20-node hidden layers, and an output layer equaling the three outputs (somatic, ambiguous, fail). The input and hidden layers used a hyperbolic tangent (tanh) activation function, the output layer used a softmax activation function. Categorical cross-entropy was used as a loss function and the Adam optimizer was used over 700 epochs with a batch size of 2,000. L2 regularization was used with a weight of 0.001.

To compare model performance, one-versus-all receiver operator characteristic curves were generated, and area under the curve metrics (AUC) were quantified using scikit-learn. We used multiple out-of-sample model validation strategies on the 41,000-variant dataset. We randomly selected two thirds of the data to serve as a training set and the remaining one-third served as the hold out test set. On the training set, we performed tenfold cross-validation for model selection and hyper-parameter tuning. When models and hyper parameters were selected, a model was trained on the training set and evaluated against the hold out test set to understand model performance. Model performance was decomposed by performing a cross-tabulation analysis on data features including reviewer, disease type, and sequencing depth (**Appendix 1, Table S1**).

Reliability diagrams were used to determine if model outputs could be interpreted as the probability of a manual variant refinement call. Model output, which was a continuous value, was plotted for 10 equally distributed bins that were separated by whether the model's output matched or did not match the manual variant refinement call. For each bin, we calculated the ratio between the number of sites where the model agreed with the call and the total number of sites in the bin. It is expected that if the model output estimates a well-scaled probability, then the calculated ratio will be correlated to an identity line (x = y). Pearson correlation coefficient was used to test for a well-scaled probability using the scipy.stats.pearsonr function.[80]

Feature importance for the deep learning model was calculated by using the cross-validation dataset. Each of the 71 features was independently shuffled and change in average AUC was determined by comparing baseline performance to shuffled performance. The random forest feature importance metric was obtained from scikit-learn's built in feature_importances_ parameter on a trained random forest model.

2.4.3 Validation of model performance by independent sequencing data with orthogonal validation

To assess model performance on orthogonal sequencing data, we evaluated variant calls from a single AML case, AML31, that had extensive orthogonal validation sequencing. Genome sequencing data (average coverage = 312×) were previously produced for AML31 and evaluated using seven different variant callers. Orthogonal custom capture validation sequencing (average coverage = 1,000×) was used to validate the 192,241 variants identified by any of the seven variant callers (MuTect, Seurat, Shimmer, Sniper, Strelka, VarScan, Bassovac).[22] Variants identified as somatic by orthogonal sequencing (the 'Platinum SNV List') were considered true positives (n = 1,343). Variants that were identified by only one of the seven callers, but not validated by orthogonal custom capture sequencing, were considered false positives (n = 190,898). Features were obtained from genome sequencing bam files for every site that was called by at least one of seven variant callers in the original study and had been selected for targeted re-sequencing (n = 192,241). The random forest and deep learning models were used to predict calls for each of the sites in the AML31 dataset and ROC figures were used to illustrate model performance.

Validation data were also obtained from TCGA exome sequencing data that had orthogonal validation.[30] Using the minor allele frequency (MAF) file (mc3.v0.2.9.CONTROLLED_lt3_b.maf) described by Ellrot et al.,[30] we identified a cohort of 19,917 variants from 106 tumor/normal pairs for model validation (**Appendix 1, Table S2**). This cohort was identified by removing un-powered validation, non-exonic variants, and potential germline calls from the original MAF file. Additionally, eligible variants required original identification via exome sequencing and orthogonal validation via targeted capture (target_status ≠ 'NaN'). Variants were labeled as true positives if they passed the Broad Institute's initial quality check (that is, 'FILTER' = 'PASS') and were statistically powered (that is, 'target_status' = '_powered'). Variants were labeled as false positives using the following tools: DetOxoG, strand bias, The Broad Institute's Panel of Normals, and ExAC47. Any TCGA sample that had at least 20 false positives and 20 false negatives validated on TCGA exome data via targeted capture was eligible for classifier validation. To test model performance on these data, we trained a deep learning model on the entire training dataset and made predictions for all variants in the independent test samples. We assessed the model performance using ROC curves as outlined above. To overcome batch effects associated with new data, we re-trained the model 15 times using incremental amounts of the test data (0%–75% with 5% increments) and employed the new model on the remaining variants.

2.4.4 Validation of model performance by independent sequencing data with manual review

To assess model robustness when employed on external data, an independent test dataset was assembled from 37 additional paired tumor/normal cases (13,579 variants) that were not included in the training set (**Appendix 1, Table S3**). Model development, variant predictions, and accuracy metrics were employed as described in the orthogonal validation analysis.

2.4.5 Annotations of clinical relevance

All variants identified as somatic by either manual somatic refinement or by the deep learning classifier (n = 21,100) were evaluated for clinical significance. MR-specific calls were defined as variants identified as somatic by the manual review pipeline but labeled as ambiguous or fail by the classifier. Classifier-specific calls were defined as variants identified as ambiguous or fail during manual review but identified as somatic by the classifier. Variants were annotated using the CIViC database.[35] To evaluate overlap with the CIViC database, coordinates were queried from the CIViC interface using the public API. Given that not all variants within CIViC can be analyzed using whole genome or whole exome sequencing, we used the provided Sequence Ontology IDs to filter out variants that cannot be analyzed using DNA-sequencing, such as ‘increased expression’ or ‘methylation’. Using coordinates queried from the CIViC interface, we determined overlap between discrepant calls and CIViC annotations.

2.4.6 Re-review of conflicting calls

A subset of variants whereby the original manual review call disagreed with the classifier call were re-reviewed using IGV. Using a standard operating procedure for manual review setup and execution,[58] we created IGV snapshots for 40 clinically relevant MR-specific calls, 53 clinically relevant classifier-specific calls, 43 non-clinically relevant MR-specific calls, and 43 non-clinically relevant classifier-specific calls. These 179 variants were manually re-reviewed by seven individuals who were blinded from original manual review calls. To analyze the 179 discordant variants, a consensus call was established as the ‘true label’ by aggregating the seven calls provided by blinded individuals. To be considered a consensus, the most common choice had to exceed any other choice by at least two votes. Any other distribution of votes resulted in that variant being classified as ‘no consensus’.

2.4.7 Statistical tests used

All plots were produced using the MatPlotlib library.[81] ROC curves were generated, and AUC metrics were calculated using scikit-learn. To assess reviewer agreement, we used Fleiss’ Kappa statistic, which is a statistic that lies between -1 and 1 where a Kappa statistic at or below 0 indicates poor agreement and above 0 indicates good agreement. For reliability diagrams, the binomial proportion confidence intervals were calculated for each bin. Pearson correlation coefficient comparing colored points to the diagonal line was calculated to assess the output of the respective model. Pearson correlation coefficient was used to test for a well-scaled probability using the scipy.stats.pearsonr function.

2.5 Results

2.5.1 Data assembly and standardization

The 41,000 called and reviewed variants used to train the model were derived from 440 individual tumors, which represent nine cancer types. Sequencing methods were evenly split between capture sequencing (14,234 variants), exome sequencing (14,044 variants), and genome sequencing (12,722 variants). Among all manually reviewed variant calls, 18,381 were confirmed as somatic, 10,643 were assessed as ambiguous, 8,854 as failed, and 3,122 as germline. The training data include both hematopoietic (10,583 variants) and solid tumors (30,417 variants), which often have distinct characteristics during manual variant refinement (**Table 2.1**).

2.5.2 Model development

Three models were developed (logistic regression, random forest, and deep learning) using the 41,000-variant dataset. To guard against overfitting, we randomly selected one-third of the dataset as a hold out test set and used the remaining two-thirds as a training set. Using a tenfold cross-validation strategy, all three models (logistic regression, random forest, and deep learning) achieved better than random performance on variant classification. The logistic regression model demonstrated the worst performance (average area under the curve (AUC) = 0.89) and limited ability to classify ambiguous calls (AUC = 0.79). Both random forest and deep learning models performed well across all classes attaining an average AUC of 0.98 and 0.96, respectively (**Figure 2.1a**). Performance of the hold out test set mirrored the tenfold cross validation (example of deep learning output in **Appendix 2, Figure S1a**). For the hold out test set, decomposition of model performance based on disease, reviewer, and sequencing depth showed no change in model performance for the deep learning and random forest models (example of deep learning cross-tabulation analysis in **Appendix 1, Table S1**).

Reliability diagrams were used to determine whether model outputs could be interpreted as a well-scaled probability. Comparing the reliability diagrams for each model indicated that the random forest model and the deep learning model produced outputs that are most closely scaled to a probability. The random forest model and the deep learning model yielded Pearson correlation coefficients (r) of 0.99 and 1.00, respectively (**Figure 2.1b**). The logistic regression model output was most divergent from a well-scaled probability with r = 0.29. When reliability diagrams were plotted independently for each class (somatic, ambiguous, and fail) for the deep learning and random forest models, all classes produced well-scaled outputs (example of deep learning output in **Appendix 2, Figure S2**).

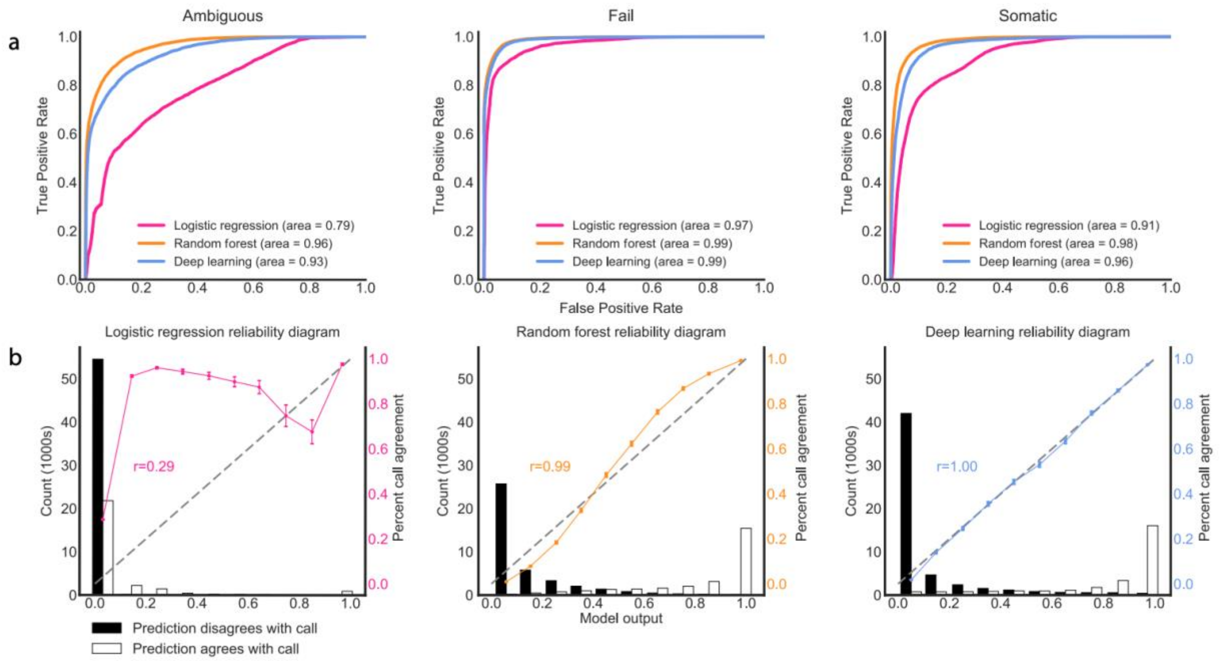


Figure 2.1 Deep learning and random forest models during 10-fold cross validation. ***a*** *Comparison of performance of three machine learning models via receiver operating characteristic (ROC) area under the curve (AUC). Performance was parsed by the three classification classes (ambiguous, fail, and somatic) for cross validation data (n=27,470 variants).* ***b*** *Graphs depict how model outputs scaled to a probability (between 0 and 1) using cross validation data (n=27,470 variants). Bar graphs show 10 equally distributed bins of model output. The bar graphs plot the number of model calls that agree and disagree with the manual review call. The diagonal line indicates a perfectly scaled probabilistic prediction. The colored points display the ratio of predictions that agree with the call to the total number of predictions for a given bin. Binomial proportion confidence intervals were calculated for each bin. Pearson's correlation coefficient comparing colored points to the diagonal line was calculated to assess the output of the respective model.*

2.5.3 Feature importance

The feature importance analysis determined which features were important for making model predictions. For the deep learning model, feature importance was ranked using the average change in the AUC after randomly shuffling individual features. For the random forest model, the built-in feature importance metric was used. To assess how manual reviewers rank feature importance, seven experienced manual reviewers at our institute ranked the top 15 (of 71) features that were most important in their manual review decision-making process. Feature ranks were normalized, and average importances across the seven reviewers were used to determine feature importance for manual reviewers. All three lists were rank normalized for comparison. Comparison shows that the models rely on many features that expert manual reviewers also use to make classification decisions (**Figure 2.2**). The random forest feature importance was moderately correlated to the deep learning and manual reviewer feature importance (Pearson r = 0.47 and 0.50, respectively). The deep learning importance was only weakly correlated with manual reviewer survey results (Pearson r = 0.17). Of note, both the random forest model and the deep learning model ranked reviewer identity higher than reviewers themselves ranked this feature. Similarly, cancer type was ranked as an important feature for both models but was not ranked highly by manual reviewers.

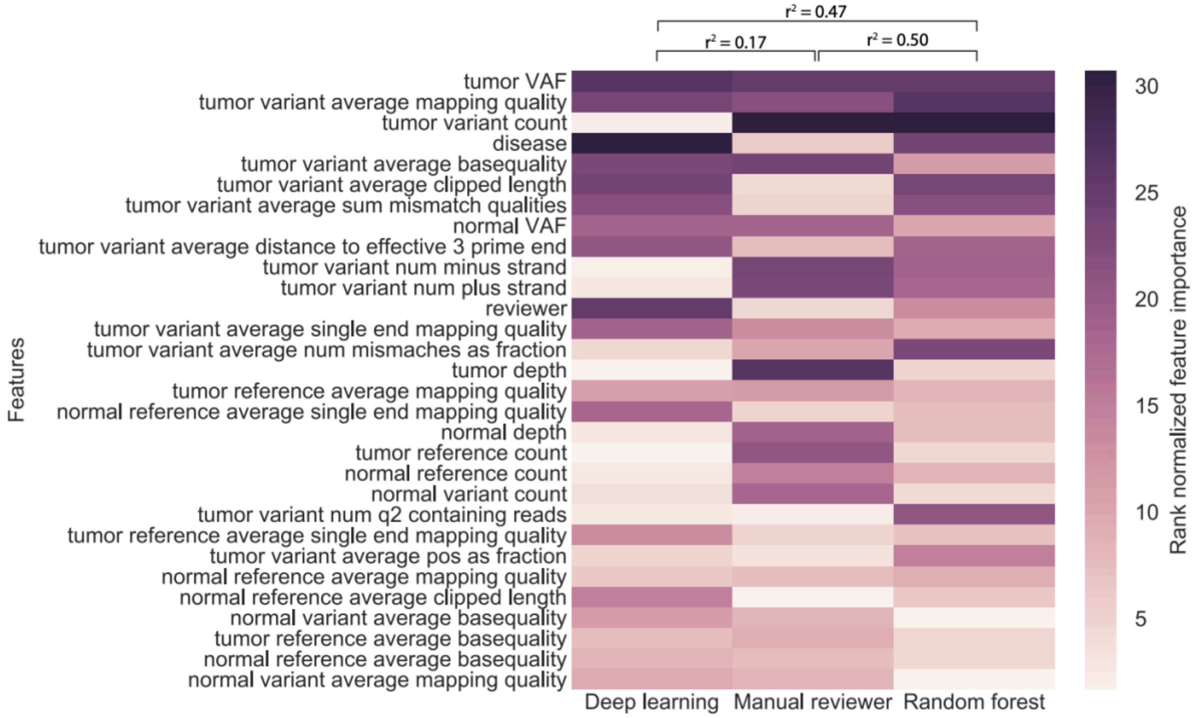


Figure 2.2 Machine learning models and manual reviewers feature use when making classifications. ***a*** *Features ranked as important by random forest and deep learning models were also ranked highly by experienced manual reviewers (n=71 features). Human manual reviewer feature importance was determined by asking 7 individuals to rank feature importance. Single feature impact for the deep learning model was obtained by training a model on the entire training dataset (n=27,470 variants) then shuffling each feature individually and calculating the mean ROC AUC for all three variant classes. The change in mean ROC AUC for all classes was sorted and plotted. Random forest feature importance was obtained via scikit-learn's feature importance parameter. All feature importance metrics were ranked normalized. The random forest feature importance is moderately correlated to the deep learning and manual reviewer feature importance (Pearson's r= 0.47 and 0.50 respectively). The deep learning importance was weakly correlated with manual reviewer survey results (Pearson's r=0.17). The top 30 (of 71) most important features are shown.*

We hypothesized that the cancer type feature was mediated by differences between liquid and solid tumors. Specifically, the concentration of leukemia cells in normal tissue for patients with high circulating counts is higher than in solid tissue malignancies with circulating tumor cells.[82] This contamination ultimately increases the risk that a somatic variant will be mis-called. To test this hypothesis, we collapsed the cancer type features to a single solid/liquid boolean. Using the deep learning model as an example, the tenfold cross-validation performance for models trained with individual tumor types was similar to models trained with a simplified tumor type (solid/liquid boolean) (**Appendix 2, Figure S1b**).

2.5.4 Inter-reviewer variability

Reviewer identity was highly ranked by both the deep learning and random forest models, indicating reviewer-specific patterns in manual review. To quantify the variability between manual reviewers, we had three independent reviewers call a random subset of 176 sites from the training dataset. This resulted in three independent review calls for each of the 176 variants. Reviewers achieved fair agreement with a Fleiss' Kappa statistic of 0.37.[83] When evaluating all calls in the inter-reviewer variability analysis, 77.3% showed good or acceptable agreement (that is, all three reviewers agreed on the call or reviewers only disagree between ambiguous and somatic or ambiguous and fail calls) (**Figure 2.3a**). Model performance was correlated with reviewer agreement such that when all three reviewers called a variant as somatic, the model produced a high somatic probability (average output > 0.8). Conversely, when all reviewers agreed that a call was fail, the model produced a low somatic probability (average output < 0.2). As expected, in situations where there was inter-reviewer variability, the model produced a wider distribution of somatic probabilities (**Figure 2.3b-c**). Together, these results indicate that there is as much as 22.7% disagreement among reviewers, especially on ambiguous calls.

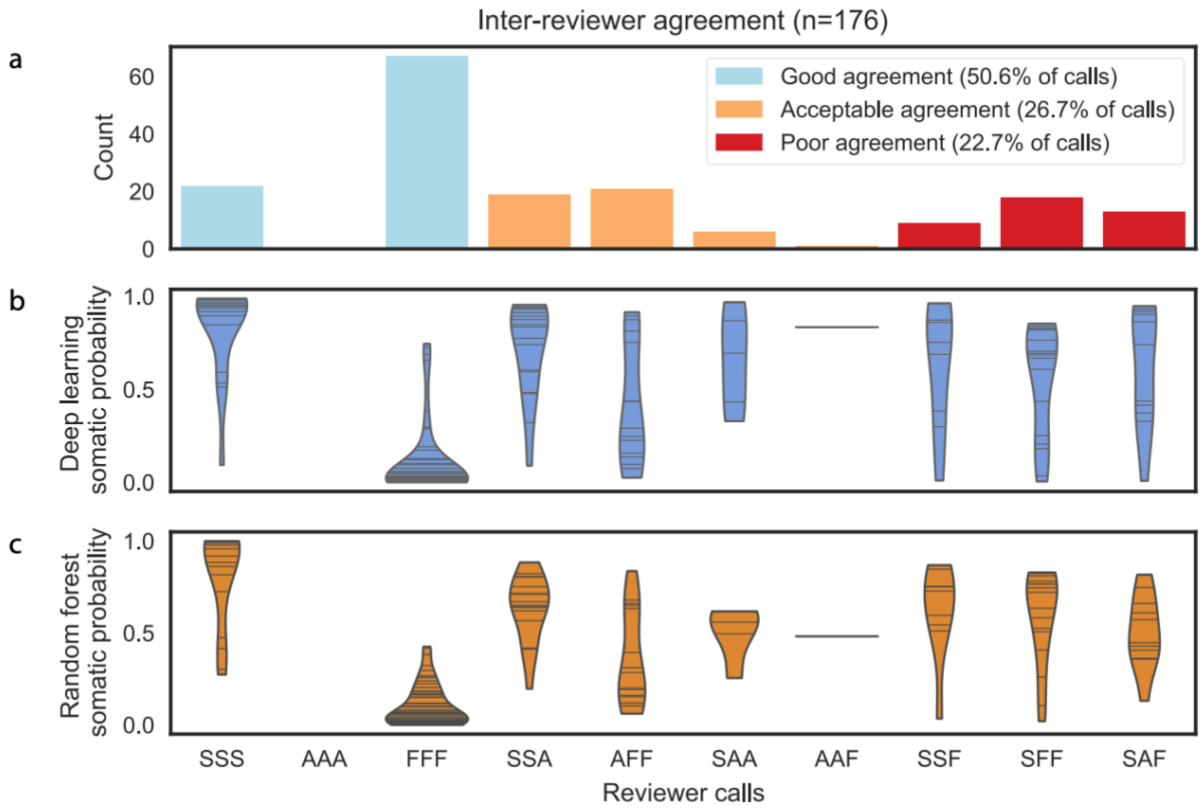


Figure 2.3 Model confidence closely parallels reviewer confidence. *When reviewers exhibit strong agreement on a variant call, the model outputs confident probabilities (>0.8 or <0.2), whereas when reviewers exhibit inter-reviewer variability for a variant call, the model outputs inconclusive probabilities (0.2 > and < 0.8).* ***a*** *Bar graphs show binned agreement of 3 reviewers for 176 variants. The x-axis outlines all possible permutations of agreement among three reviewers. The y-axis outlines the frequency of each permutation. 'S' denotes a somatic call, 'A' denotes an ambiguous call, and 'F' denotes a fail call. 'SSS' is the case where all three reviewers call the same variant somatic and the other permutations follow a similar pattern (e.g., 'SAF'= somatic, ambiguous, fail). It is considered (1) good agreement when all three reviewers agree, (2) acceptable agreement when reviewers only disagree between ambiguous and somatic or ambiguous and fail calls, (3) and poor agreement when one reviewer calls a variant somatic while another calls a variant fail.* ***b*** *Violin plots of deep learning somatic probability whereby the horizontal lines indicate the occurrence of a probability, and the width indicates the distribution of probabilities (n = 528 variants [176 variants for each of the 3 reviewers]).* ***c*** *Violin plots of random forest somatic probability whereby the horizontal lines indicate the occurrence of a probability, and the width indicates the distribution of probabilities (n = 528 variants [176 variants for each of the 3 reviewers]).*

Model outputs that do not depend on reviewer identity are most desirable to reduce the impact of idiosyncratic criteria on ultimate calls. Therefore, new models were developed after removing the reviewer feature from the training data to assess performance in situations when the reviewer is unknown. Using the deep learning model as an example, tenfold cross-validation with all 71 features resulted in an average AUC of 0.960, whereas tenfold cross-validation without the reviewer feature resulted in an average AUC of 0.956. This experiment illustrates expected performance on de novo data that does not include a reviewer feature (**Appendix 2, Figure S1c**).

3.5.5 Independent sequencing data with orthogonal validation

To validate model performance on unfiltered 'raw' variant calls, deep learning and random forest models were used to predict manual review labels for 192,241 putative somatic variants in the acute myeloid leukemia case (AML31) described by Griffith et al.24 This case study had deep (312×) genome sequencing data as well as ultra-deep (1,000×) orthogonal custom capture validation for all 192,241 predicted variant sites. Variants validated by the custom capture data were considered true positives and those that failed validation were considered false positives (**Appendix 1, Table S2**). When comparing somatic model predictions to validation sequencing results, the deep learning model and the random forest model achieved receiver operating characteristic (ROC) AUCs of 0.95 and 0.96, respectively (**Figure 2.4a**).

Additional sequencing data were obtained from The Cancer Genome Atlas (TCGA) dataset. Specifically, we obtained a cohort of 106 TCGA tumor-normal pairs that had original exome sequencing and subsequent targeted orthogonal validation.[30] This cohort comprised eight cancer types and 19,917 total variants, whereby 17,109 were true positives and 2,808 were false positives (**Appendix 1, Table S2**). When employing the deep learning model on this dataset, average ROC AUC for each cancer type ranged from 0.724–0.878, and average ROC AUC for all variants was 0.78 (**Figure 2.4b-c**). To overcome batch effect, orthogonal validation calls were randomly selected in increments of 5% (from 0%–75%) to include in re-training the model. The newly trained model was used to predict calls for remaining variants in the TCGA dataset. When re-training the model using incremental amounts of the testing set, the total AUC improved. After incorporating 20% of the TCGA data, the model attained a ROC AUC of 0.90 and when incorporating 75% of the TCGA data, the model attained a ROC AUC of 0.93 (**Figure 2.4d**).

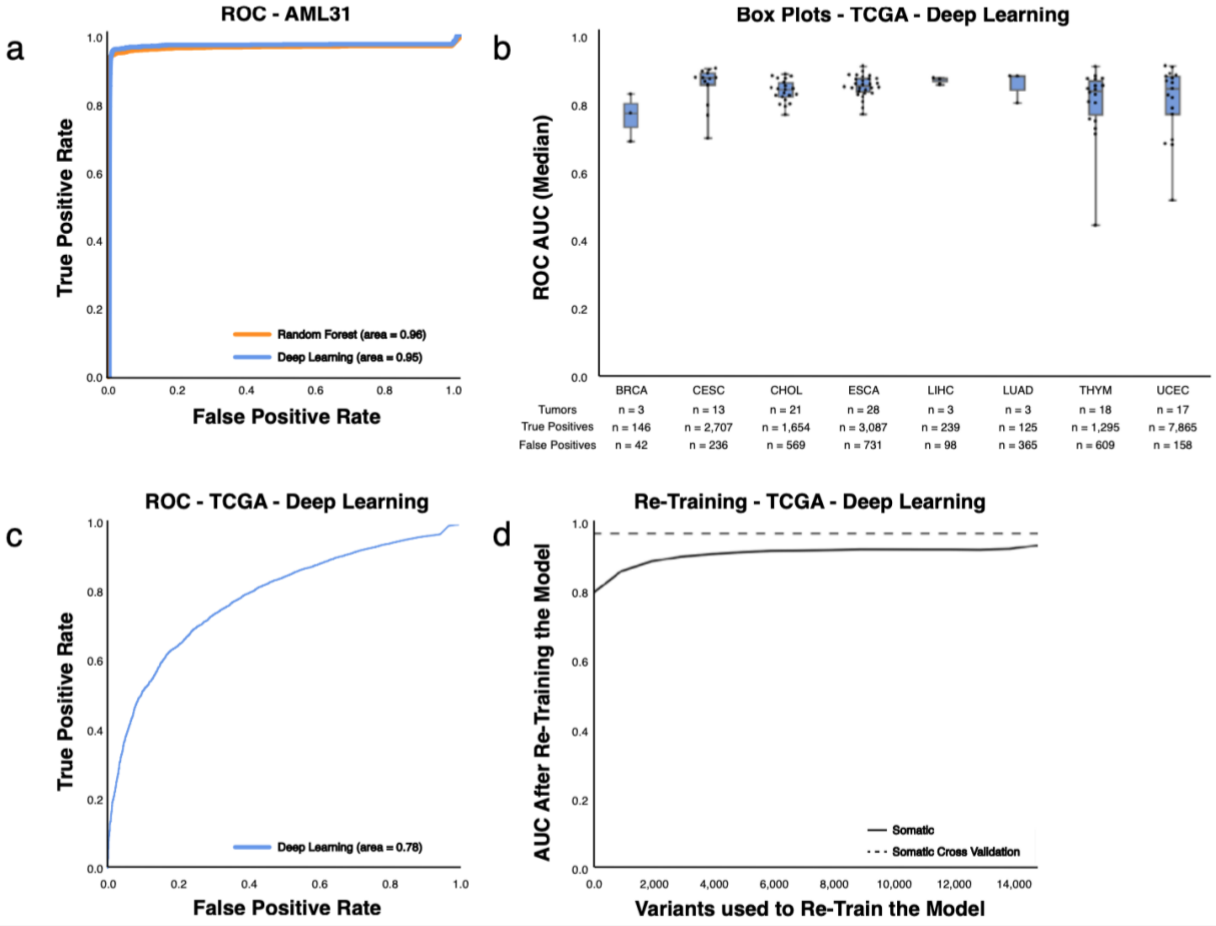


Figure 2.4 Machine learning models accurately predict orthogonal validation sequencing results. ***a*** *A single AML case with 312X genome sequencing had 7 automated somatic variant callers identify 192,241 putative somatic variants. Orthogonal sequencing at ~1,000X was performed for all 192,241 variants to identify true positives and false positives. The random forest and deep learning models predicted labels for all variants using the 312X genome sequencing data as input. Model accuracy was determined by comparing model predictions to orthogonal*

sequencing labels. ***b*** *Box plots describe the median ROC AUC for each of 8 TCGA cancer types (n=106 tumor/normal pairs; n=19,917 variants). Each dot represents a single TCGA tumor/normal pair, the centre represents the 50th percentile, the lower and upper limits of the box represent 25th and 75th percentiles, respectively, and whiskers represent data minimum and maximum. The table below the boxplots shows information on the total number of samples assayed and the distribution of true positive and false positive calls for each cancer type.* ***c*** *ROC AUC for all TCGA data (n = 19,917 variants) using the deep learning classifier trained on the 41,000 variants described in Table 1. d. Change in ROC AUC after retraining the deep learning model with increments of the TCGA data. TCGA data was partitioned in random stratified increments of 5% (from 0-75%) and used to train a new model (increments = 1,327 variants). The x-axis outlines the number of test variants included in re-training. The y-axis plots the resulting model's ROC AUC.*

2.5.6 Independent sequencing data with manual review validation

To test model performance on external manual review data, three independent datasets were obtained whose characteristics differed from the training set. These datasets included 4 small-cell lung cancer (SCLC) cases with 2,686 variants, 14 follicular lymphoma (FL) cases with 1,723 variants, and 19 head and neck squamous cell carcinoma (HNSCC) cases with 9,170 variants (**Appendix 1, Table S3**). The SCLC cases were sequenced independently from the training set SCLC cases, utilized different methods for automated somatic variant calling, and were reviewed by new manual reviewers. The FL cases had a unique distribution of call classes (50.2% somatic, 49.8% fail, and 0% ambiguous) when compared to the training set (44.8% somatic, 29.2% fail, and 26% ambiguous). The HNSCC cases represented a new tumor type and were aligned to a different version of the human reference genome (GRCh38).

For the deep learning model, ROC AUC for independent test sets (n = 37 cases) ranged from 0.78–0.92 for somatic variants, 0.74–0.92 for failed variants, and 0.43–0.47 for ambiguous variants (**Figure 2.5**). When re-training the model using incremental amounts of the testing set, as described above, model performance improved. For the deep learning model, inclusion of approximately 250 manual review calls restored performance to levels observed in cross-validation

(**Figure 2.5**). Initial model performance for the deep learning model outperformed the random forest model, especially for somatic and fail variants (**Appendix 2, Figure S3**).

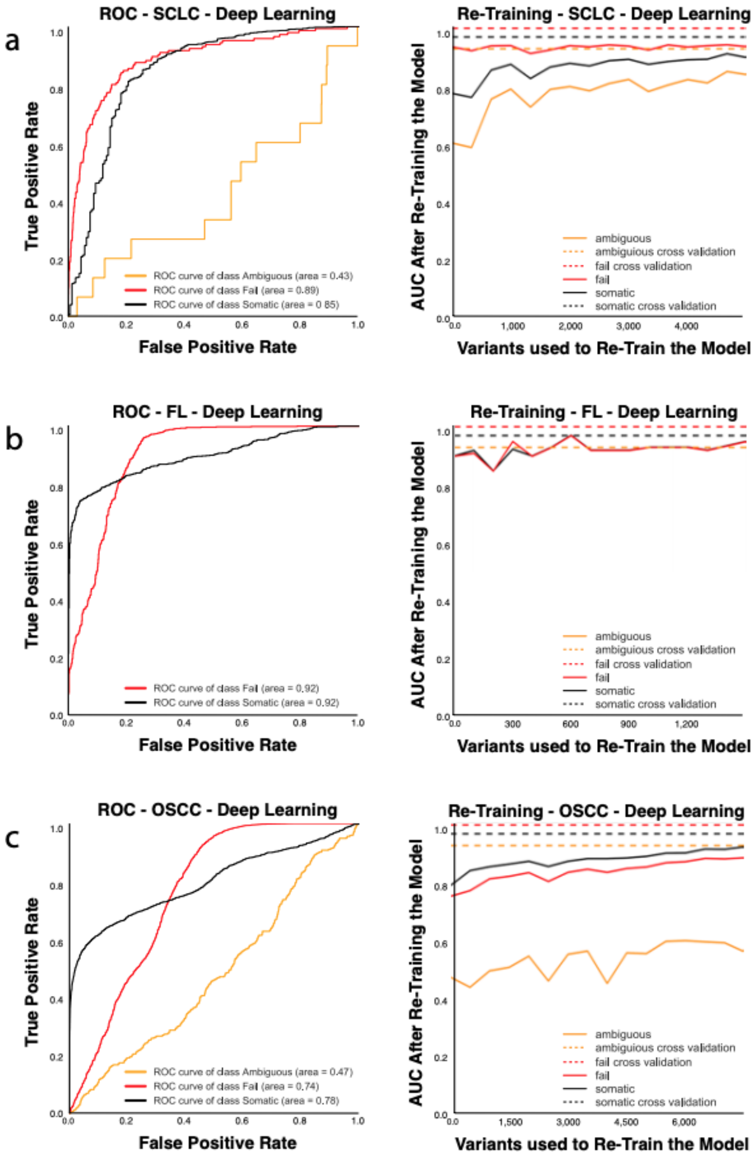
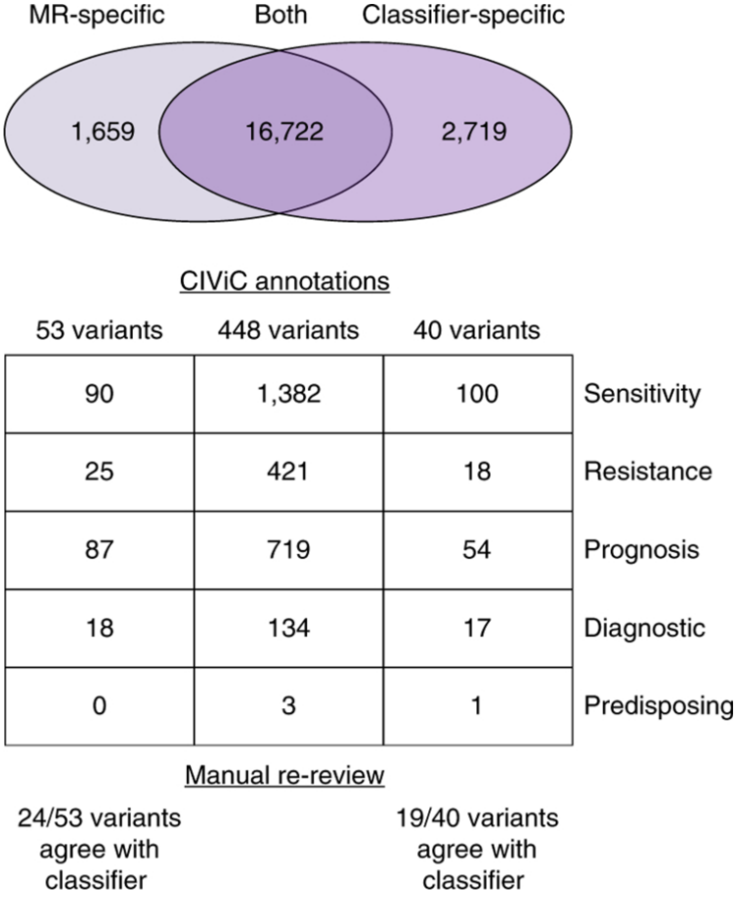


Figure 2.5 The deep learning model performance on three independent test sets. ***a*** *ROC curves outlining model performance on 4 small cell lung cancer (SCLC) cases with 2,686 variants and independent test set correction through model re-training to overcome batch effects associated with new manual reviewers, new sequencers and a new alignment strategy.* ***b*** *ROC curves outlining model performance on 14 follicular lymphoma (FL) cases with 1,723 variants and independent test set correction through model re-training to overcome batch effects associated with different frequencies of manual review classes.* ***c*** *ROC curves outlining model performance on 19 head and neck squamous cell carcinoma (HNSCC) cases with 9,170 variants and independent test set*

correction through model re-training to overcome batch effects associated with alignment to a different reference genome (GRCh38).

2.5.7 Analysis of clinically relevant variants

The deep learning model was used to assess whether machine learning algorithms for variant analysis could improve detection of clinically actionable variants mislabeled by manual refinement strategies. Of the 21,100 variants identified as somatic by either the deep learning model or by manual review, there were 16,722 variants that were called as somatic by both methods, 1,659 manual review (MR)-specific variants, and 2,719 classifier-specific variants (**Figure 2.6**).



53 variants	448 variants	40 variants	
90	1,382	100	Sensitivity
25	421	18	Resistance
87	719	54	Prognosis
18	134	17	Diagnostic
0	3	1	Predisposing

Figure 2.6 Manual review misclassifications recovered by the deep learning model. *The Venn diagram illustrates variants identified as somatic by manual review (MR-specific), by both pipelines (Both), and by the deep learning classifier (Classifier-specific). For these three groups, the number of variants that have direct overlap with CIViC annotations and the total number of evidence items associated with all variants within each group are shown. These evidence items*

are parsed by those that convey variant sensitivity to a drug, variant resistance to a drug, variant that confers better or worse prognosis, variant that confers disease diagnosis, and variant that shows predisposing evidence for disease. The manual re-review panel shows the number of clinically relevant variants that agreed with the classifier call upon re-review by seven individuals.

Discordant variants (MR- or classifier-specific) were evaluated for clinical relevance using the Clinical Interpretations of Variants in Cancer database (CIViC).[35] Each annotation within CIViC is based on evidence summaries that detail therapeutic, prognostic, predisposing, or diagnostic implications in cancer. After filtering extraneous evidence summaries (see **2.4 Methods and experimental procedures**), there were 425 clinically relevant CIViC annotations. Using these CIViC annotations, 40 classifier-specific variants were identified that were clinically actionable. These 40 variants were associated with 100 evidence items related to therapeutic sensitivity, 18 evidence items related to therapeutic resistance, 54 evidence items that detailed prognostic information, 17 evidence items that indicated diagnostic information, and one evidence item that supported predisposition to cancer. If we assume that the classifier more accurately predicts the true variant label, this would represent an 8.9% increase in detection of clinically relevant variants. Using relevant CIViC annotations, 53 manual review-specific variants were identified as clinically actionable. Of these manual review-specific variants, 90 evidence items related to therapeutic sensitivity, 25 evidence items related to therapeutic resistance, 87 evidence items detailed prognostic information, and 18 items illustrated diagnostic information. If we again assume that the classifier call is more accurate relative to the original manual review call, this would represent an 11.8% reduction in mislabeled, clinically relevant calls.

Blinded retrospective review of these mislabeled variants in IGV confirmed confidence in model predictions. Four examples of manual review miscalls that were originally labeled as somatic but were failed by the classifier are shown in **Appendix 2, Figure S4**. Two examples of manual review miscalls that were originally labeled as ambiguous or fail by manual reviewers but were identified as somatic by the classifier are shown in **Appendix 2, Figure S5**. In **Appendix 2, Figure S5a**, two clinically relevant PIK3CA variants were missed due to the manual reviewer assuming that two adjacent variants on the same strand were considered multiple mismatches. In **Appendix 2, Figure S5b**, a TP53 variant was missed in an AML case due to the manual reviewer's lack of awareness that hematologic cancers can have tumor cell contamination in normal tissue.

2.5.8 Analysis of discrepant calls

The classifier agrees with manual review in 89.3% (35,622/41,000) of calls; however, there were 4,378 variants (10.7%) for which the original manual review call was discrepant with the classifier call. To understand features influencing discrepant calls, unbiased manual re-review was performed on 179 discordant variants. Seven individuals proficient in manual review re-reviewed IGV snapshots of the 179 variants. For each variant, a consensus call was determined (see **2.4 Methods and experimental procedures**). When comparing the original manual review and classifier calls to the consensus call, 51 variants (28.5%) showed call-agreement between the consensus call and classifier call, and 53 variants (29.6%) showed call-agreement between consensus call and the original manual review. Additionally, 34 variants (19.0%) showed disagreement between the consensus call, the classifier call, and the original manual review call, and 41 variants (22.9%) had no consensus (**Appendix 2, Figure S6**). Of the 93 discrepant clinical variants evaluated during re-review, 50 variants were classified as either 'no agreement' or 'no consensus'. Therefore, we estimate that approximately 5.8% of all clinically relevant variant calls are fundamentally ambiguous, even on re-review.

2.6 Discussion

The random forest and deep learning models achieved high (average AUCs > 0.95) classification performance across all variant refinement classes (somatic, ambiguous, and fail), whereas the logistic regression model demonstrated reduced performance (average AUC = 0.89), particularly with the ambiguous class. High performance of model predictions confirms that an automated strategy can reduce the need for manual variant refinement. In addition, maintenance of performance after elimination of the manual reviewer feature further demonstrated that the trained model can be used on de novo data without reviewer information (**Appendix 2, Figure S1c**).

The deep learning and random forest models also showed high accuracy (AUCs > 0.95) when classifying independent sequencing data with orthogonal validation. The AML31 case outlined by Griffith et al.[22] had two unique features that made it optimal for assessing model performance. First, variants had manual review calls for both the original genome sequencing and the ultra-deep orthogonal sequencing. Second, the ultra-deep orthogonal sequencing was performed on all variants (false positives and true positives) called by automated somatic variant callers, allowing for quantification of both sensitivity and specificity.

With regards to the TCGA orthogonal validation datasets, the deep learning model showed initial reduction in average AUC (AUC = 0.78) relative to cross-validation performance. We hypothesized that reduction in accuracy was attributable to methods used for classifying TCGA false positives. Specifically, TCGA false positives were filtered by eliminating variants caused by 8-Oxoguanine (OxoG) DNA lesions using the DetOxoG tool,[84] eliminating variants with strand bias, and eliminating germline variants using a panel of normals. Since these features are not typically available to manual reviewers, they were not incorporated into the original model. Re-training with TCGA false positives allowed the model to learn new sequencing features that improve its ability to recognize false positives, ultimately restoring model accuracy (AUC = 0.93). Given these findings, we are hopeful that development of a model that incorporates these data will improve somatic variant refinement and eventually reduce the need for orthogonal validation sequencing.

For the independent sequencing data with manual review validation, we also observed a decrease in model performance. However, when re-training the model with as few as 250 calls, model performance was restored (**Figure 2.5**). Therefore, when employing the classifier on new datasets, we recommend manually reviewing or performing validation sequencing for a small subset of variants called via statistical variant callers (for example, 5% of all data) to re-train the classifier and improve performance. Our group has provided a command line interface to allow individuals to train a custom deep learning classifier, prepare data, and classify variants. The deep learning model was selected as the optimal method for somatic variants refinement due to its increased accuracy when employed on validation sets.

These results together show that a machine learning model can effectively automate somatic variant refinement. Standardization and systematization of this process decreases the human variability associated with manual refinement and increases the reproducibility of variant calling. In addition, automation of variant refinement eliminates a labor bottleneck, and its associated costs, allowing any number of somatic variant calls to be evaluated in a negligible amount of time. Finally, since the model offers probabilistic output, an economic framework can be used to set thresholds for confirmatory follow up testing, allowing investigators to optimize experimental design to improve accuracy within budgetary constraints.[85]

To illustrate the extent of this advance, we compared the manual review burden in a standard cancer genomics workflow with a workflow that utilizes the machine learning classifier. In a previously conducted breast cancer study,[24] 10,112 variants were identified via automated somatic variant callers. In this example, 1,066 variants were filtered using heuristic cutoffs, and 9,046 variants required manual review. Given that experienced reviewers can evaluate 70–100 variants per hour, manual review for this study would have taken ~90–130 hours. Using the machine learning approach, 5% of the data (~500 variants) would require manual review. This manual review data would be used to re-train the model and correct for associated batch effects. This manual review would require approximately 5 hours. In this example, the manual review burden would be reduced from ~100 hours to ~5 hours, detailing the considerable improvement in efficiency.

Through the re-review analysis, we showed that inter-reviewer variability affects variant detection, which can ultimately impact patient care. Many of the variants with high inter-reviewer variability and/or no consensus call had clinical significance. We believe that in these cases, an automated model can provide an unbiased and probabilistic output for variant classification, thereby eliminating reproducibility issues associated with manual refinement. In instances where the model makes an ambiguous call for a variant of clinical relevance, we recommend manually reviewing these variants to make a definitive call.

This model does have some limitations. Given the identified inter-reviewer variability associated with manual variant refinement calls, the training data likely contain a substantial amount of noise that might impact model performance. Moreover, there are sources of data that can be used to build better models. In an ideal scenario, highly accurate orthogonal validation sequencing would be performed to determine somatic variant status. Unfortunately, validation sequencing has a large monetary and tissue material expense, limiting our ability to use these types of data in the training set. Lastly, while the training data were produced using a varied array of capture strategies, libraries, Illumina sequencing instruments, and somatic variant callers, the model will likely require evaluation and some amount of retraining for non-Illumina sequencing instruments and divergent somatic variant analysis pipelines. It is also possible that the model has learned various other institutional batch-effects from our sequencing and analysis workflows. However, our results suggest that retraining with a small number of supplemental calls from an independent dataset may be sufficient to overcome these effects. We anticipate improving this model by adding genomic and sequencing features such as proximal sequence complexity (for example, presence of repeat regions), functional prediction (for example, conservation based variant impact scores), and other indicators associated with false positives.[86,87]

In conclusion, persistent weaknesses in variant calling pipelines remain, especially in an era of constantly changing and variable sequencing data quality. Sophisticated context-specific pattern matching abilities of humans are still needed to refine and confirm somatic variant calls, which is expensive and laborious. We show that with a relatively small amount of project-specific review for model retraining that most manual review can be replaced with an automated classifier approach, providing more reproducible and refined calls for clinically relevant variants.

2.7 Acknowledgements

We would like to thank A. Petti, G. Chang, T. Li, C. Miller, L. Trani, R. Lesurf, Z. Skidmore, K. Krysiak, A. Ramu, and F. Gomez for assisting in data assembly. We also acknowledge L. Trani for performing manual review and for valuable discussion on the project. We gratefully acknowledge L. Wartman, J. DiPersio, M. Jacoby, B. Van Tine, R. Fields, B. Tan, S. Chi, D.

Gutmann, and T. Ley for sharing genomic data that made this project possible. We also would like to thank the patients and their families for their selfless contribution to the advancement of science. Part of this work was performed as part of the Washington University School of Medicine Genomics Tumor Board, which was funded with private research support from the Division of Oncology and the McDonnell Genome Institute. E. Barnell was supported by the National Cancer Institute (T32GM007200 and U01CA209936). T. Rohan received support from NIH/NCI (R01CA142942) and the Breast Cancer Research Foundation. Select sample data was funded by the Genomics of AML PPG (T. Ley, PI, P01 CA101937). A. Wagner was supported by the National Cancer Institute (NIH NCI F32CA206247). B. Ainscough was supported by the Siteman Cancer Center. S. Swamidass is funded by the National Library of Medicine (NIH NLM R01LM012222 and NIH NLM R01LM012482) and acknowledges support from the Institute for Informatics at Washington University School of Medicine. M. Griffith is funded by the National Human Genome Research Institute (NIH NHGRI R00HG007940). O. Griffith is funded by the National Cancer Institute (NIH NCI K22CA188163 and NIH NCI U01CA209936).

2.8 Author Contributions

B.J.A. designed the study, assembled and cleaned training data, performed feature engineering, designed model architecture, tuned hyperparameters, performed model training and analysis, performed manual review, assembled validation data, wrote code, created figures, and wrote the manuscript. E.K.B. designed the study, performed manual review, performed model training and analysis, performed clinical data analysis, assembled validation data, wrote code, created figures, and wrote the manuscript. P.R. and K.M.C. wrote code, performed manual review, and edited the manuscript. A.H.W. wrote code. T.E.R., R.G., R.U., G.P.D, and T.A.F. shared genomic data that was used in training the model and revised the paper. M.G., E.R.M., S.J.S., and O.L.G. designed the study, supervised the project and revised the paper.

2.9 Data Availability

All analysis, preprocessing code, readcount training data, manual review calls, and trained deep learning and random forest models are available on the DeepSVR GitHub repository. The raw sequencing data are publicly available for most projects included in this study. Users can access the classifier command line interface via our open-sourced GitHub repository and can install the package through Bioconda49. After installation, the tool can be used to (1) train and save a deep learning classifier, (2) prepare data for training a classifier or classification, and (3) classify data using either the provided deep learning model or a custom model. A walkthrough of this process is available on the DeepSVR GitHub Wiki.

Chapter 3: Standard operating procedure for somatic variant refinement of sequencing data with paired tumor and normal samples

3.1 Preamble

The following chapter has been published:

Barnell E.K., Ronning P., Campbell K.M., Krysiak K., Ainscough B.J., Sheta L.M., Pema S.P., Schmidt A.D., Richters M., Cotto K.C., Danos A.M., Ramirez C., Skidmore Z.L., Spies N.C., Hundal J., Sediqzad M.S., Kunisaki J., Gomez F., Trani L., Matlock M., Wagner A.H., Swamidass S.J., Griffith M., Griffith O.L. Standard operating procedure for somatic variant refinement of sequencing data with paired tumor normal samples. Genetics in Medicine, October 5, 2018.

As an author of the published manuscript, and in compliance with the editorial policies at Genetics in Medicine, the cited publication is included in full in the following chapter. My role in this project was to develop the standard operating procedure (SOP), create figures for all variant annotations, design and execute the SOP efficacy study, and assemble the manuscript. A complete list of author contributions is included within the publication (**Chapter 3.8**).

3.2 Summary

Following automated variant calling, manual review of aligned read sequences is required to identify a high-quality list of somatic variants. Despite widespread use in analyzing sequence data, methods to standardize manual review have not been described, resulting in high inter- and intra-lab variability. The manual review standard operating procedure (SOP) presented here consists of methods to annotate variants with four different calls and 19 tags. The calls indicate a reviewer's confidence in each variant and the tags indicate commonly observed sequencing patterns and artifacts that inform the manual review call. Four individuals were asked to classify variants prior to, and after, reading the SOP and accuracy was assessed by comparing reviewer calls with orthogonal validation sequencing. After reading the SOP, average accuracy in somatic variant identification increased by 16.7% (p value = 0.0298) and average inter-reviewer agreement increased by 12.7% (p value < 0.001). Manual review conducted after reading the SOP did not significantly increase reviewer time. This SOP supports and enhances manual somatic variant detection by improving reviewer accuracy while reducing the inter-reviewer variability for variant calling and annotation.

3.3 Introduction

Large genome centers, such as the McDonnell Genome Institute, use a wide variety of sequencing workflows. Typically, extracted nucleic acid is subjected to fragmentation; size selection; KAPA (Wilmington, MA), Swift (Ann Arbor, MI), IDT (San Jose, CA), or Illumina (San Diego, CA) library preparation protocols (end-repair, tailing, ligation, amplification, etc.); NimbleGen (Basel, Switzerland) or IDT custom/exome capture; and subsequent sequencing via Illumina HiSeq 2500/4000 or Novaseq 6000. The sequencing workflow typically follows methods described by Griffith et al.[22] Subsequently, the bioinformatics pipeline requires alignment to the reference genome (GRCh37/38) via Burrows–Wheeler Aligner (BWA)[76] or BWA-MEM and postprocessing of aligned sequencing reads. Postprocessing requires deduplication of reads via Picard[78] and automated somatic variant calling using the intersection or union of Mutect,[18] SomaticSniper,[88] Strelka,[17] VarScan[19] or others. A multi-caller approach is used to identify a preliminary list of high-quality somatic variants from aligned sequence data.[89–91] The bioinformatics pipeline can be implemented using the Genome Modeling System.[5]

Automated pipelines can identify and filter many false variant calls that result from sequencing errors, misalignment of reads, and other factors; however, additional refinement of somatic variants is often required to eliminate variant caller inaccuracies. This additional refinement is critical because inaccurate identification of variants can lead to poor patient management and missed therapeutic opportunities, as outlined in the Association for Molecular Pathology (AMP) guidelines for interpretation and annotation of somatic variation.[38,52] Therefore, manual inspection of somatic variants identified by automated variant callers (i.e., manual review) is an important aspect of the sequencing analysis pipeline and is currently the standard for variant refinement. Manual review allows individuals to incorporate information not considered by automated variant callers. For example, a trained eye can discern misclassifications attributable to overlapping errors at the ends of sequence reads, preferential amplification of smaller fragments, or poor alignment in areas of low complexity. Due to computational limitations, automated methods for variant refinement are in early stages of development and manual review remains integral to variant identification workflows.[25]

Despite extensive use of manual review in clinical diagnostic and molecular pathology settings,[65,92,93] somatic variant refinement strategies are often unstated or only briefly mentioned in studies that report postprocessing of automated variant calls[53,54,93–95] Lack of formalized procedures for the sequencing pipeline, and specifically for somatic refinement, permits high levels of inter- and intra-lab variability and can hinder reproducibility of results.[95] Thus, development of a procedure to standardize and systematize somatic variant refinement would improve the overall quality of sequencing analysis pipelines.

Here we present a standard operating procedure (SOP) for manual review of paired tumor/normal samples to help standardize somatic variant refinement. We first detail instructions for downloading and using the publicly available Integrative Genomics Viewer (IGV)[23,96] and IGVNavigator (IGVNav) software to properly visualize somatic variants during manual review. We also show that adoption of a standardized method for somatic variant refinement through this manual review SOP improves the accuracy of somatic variant calls and reduces overall inter-reviewer variability.

3.4 Methods and experimental procedures

3.4.1 Setting up manual review using IGV

The Integrative Genomics Viewer (IGV) is a high-performance genomic data visualization tool. This SOP reviews IGV (v2.4.8) components that can be used to conduct manual review of variants identified by automated somatic variant callers. While we have chosen IGV to develop our SOP, many of the following concepts are applicable to other genomic viewers.[97–99] The IGV desktop application is available for all major operating systems.

The IGV interface is composed of three main panels: (1) Genome Ruler, (2) Data Tracks, and (3) Genome Features (**Figure 3.1**). The Genome Ruler provides navigation features to center a genomic locus of interest. A dropdown menu provides reference genome selection, the variant coordinates show the current field of view, the zoom buttons expand/contract the field of view, and other buttons provide additional display and navigation control. Within the Data Tracks section, each horizontal track represents one experiment, sample, or annotation. In **Figure 3.1**, a normal BAM track and a tumor BAM track are loaded. For BAM files, each data track consists of a coverage track and individual read alignments. Reads ideally represent a single originating molecule that was sequenced and aligned to a reference. In default settings, sequenced bases that disagree with the aligned reference sequence are highlighted. The Genome Features section provides reference information that can be used to supplement manual review. The reference DNA and protein sequence tracks are loaded by default. Optionally loaded tracks from the IGV server will typically appear in the Genome Features section.

IGV supports a variety of input files for sequence data visualization. The File dropdown menu details the various supported input files. Indexed BAMs can be efficiently accessed from a local file system. Alternatively, the Load from URL option permits direct URL input from a web service. The Load from Server option downloads tracks from supported data sets (e.g., the Cancer Genome Atlas, Ensembl, etc.).

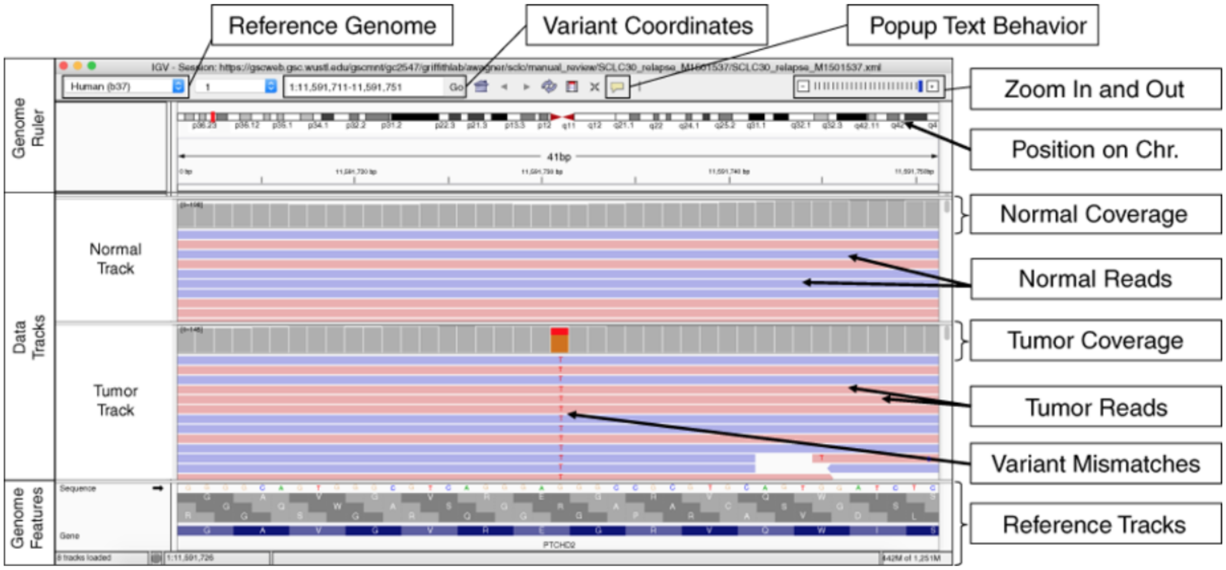
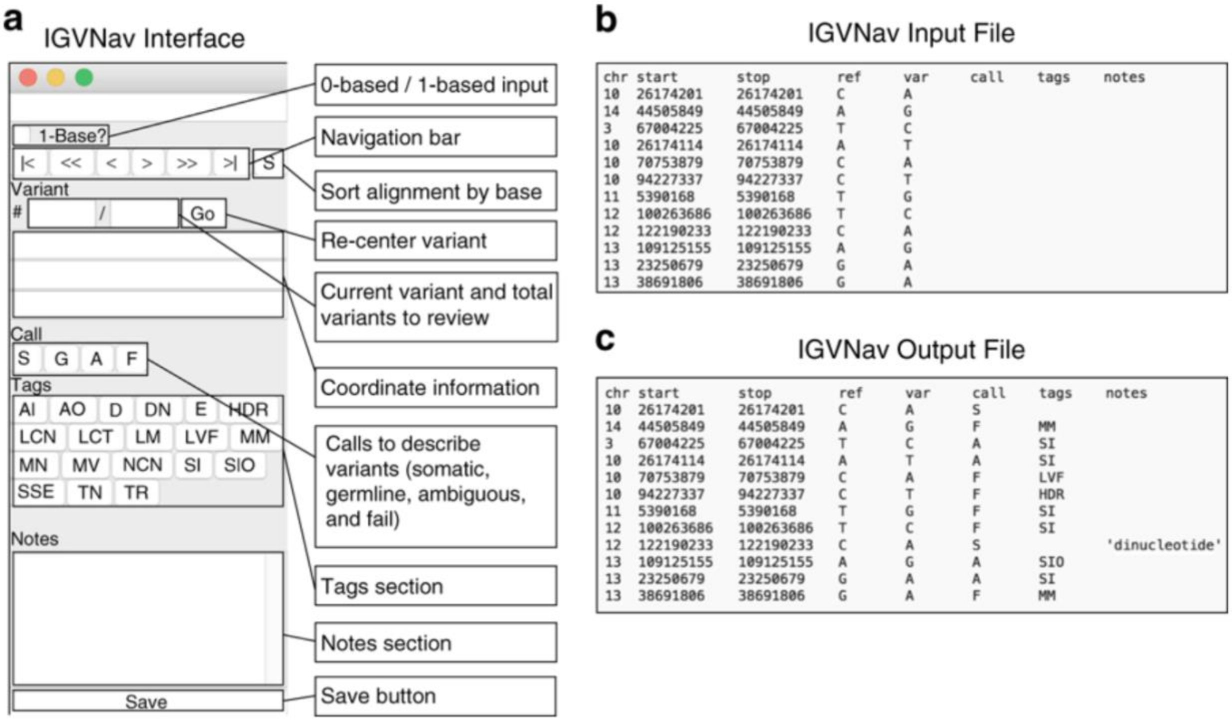


Figure 3.1 Example of the Integrative Genomics Viewer (IGV) interface with associated features. *The IGV interface is divided into three parts. The Genome Ruler details information about the genome assembly being visualized (Reference Genome), the coordinates currently being visualized (Variant Coordinates), and other navigation/display controls (e.g., Popup Text Behavior, Zoom In and Out, etc.). In this example, a portion of human chromosome 1 (build 37) is shown. The central section of IGV displays Data Tracks. In this case, short read DNA alignment data (e.g., BAM files) are shown for normal and tumor samples and are colored by read strand. Mismatches with the reference genome are highlighted by base: adenine (green), cytosine (blue), guanine (orange), and thymine (red). Coverage tracks summarize the total read depth at each base position. The Genome Features section shows the reference sequence itself, the amino acids for the three possible reading frames, and the gene associated with this locus (PTCHD2 in this example). The default gene track available with IGV is shown (RefSeq). Many other data formats and sources can be loaded as data tracks or genome features.*

3.4.2 Setting up manual review using IGVNav

IGVNav software (a Python applet/plugin for IGV), announced here, is available for download under an open access license (GNU) from GitHub (https://github.com/griffithlab/igvnav). When initiated, the user is prompted to open an input file for manual review. The input file is a tab delimited, 0- or 1-based BED-like file with the following columns: chromosome, start coordinate, stop coordinate, reference allele, variant allele, call, tags, and notes. For variants that have not yet been manually reviewed, the call, tags, and notes columns should be blank (**Figure 3.2b**). IGVNav features are shown in **Figure 3.2a**. The navigation bar permits movement through the input variant list. The "S" button sorts alignments by base so that variants appear at the tops of data tracks. Below the navigation bar is the current variant being visualized and the total number of variants in the input file. Editing this section and selecting the Go button will navigate to a specific variant of interest. The three horizontal bars display coordinate information for the current variant. The first bar details the chromosome, start, and stop position; the second bar shows the reference allele; and the third bar shows the variant allele. The Call section allows the manual reviewer to select one of the following: somatic (S) (**Appendix 3, Figure S1**), germline (G) (**Appendix 3, Figure S2**), ambiguous (A) (**Appendix 3, Figure S3**), or fail (F) (**Appendix 3, Figure S4**). The Tags section allows manual reviewers to annotate variants with commonly observed sequencing patterns. Tags can be used for any call (S, G, A, or F); however, they are especially important for ambiguous and fail calls to indicate the call rationale. Descriptions of calls and tags can be found in Table 1. The IGVNav interface also contains a Notes section, which allows for free text. At any point during a manual review session, the calls, tags, and notes can be saved to the original input file using the Save button (**Figure 3.2c**).



IGVNav Input File

```
chr start       stop        ref     var     call    tags    notes
10  26174201    26174201    C       A
14  44505849    44505849    A       G
3   67004225    67004225    T       C
10  26174114    26174114    A       T
10  70753879    70753879    C       A
10  94227337    94227337    C       T
11  5390168     5390168     T       G
12  100263686   100263686   T       C
12  122190233   122190233   C       A
13  109125155   109125155   A       G
13  23250679    23250679    G       A
13  38691806    38691806    G       A
```

IGVNav Output File

```
chr start       stop        ref     var     call    tags    notes
10  26174201    26174201    C       A       S
14  44505849    44505849    A       G       F       MM
3   67004225    67004225    T       C       A       SI
10  26174114    26174114    A       T       A       SI
10  70753879    70753879    C       A       F       LVF
10  94227337    94227337    C       T       F       HDR
11  5390168     5390168     T       G       F       SI
12  100263686   100263686   T       C       F       SI
12  122190233   122190233   C       A       S               'dinucleotide'
13  109125155   109125155   A       G       A       SIO
13  23250679    23250679    G       A       A       SI
13  38691806    38691806    G       A       F       MM
```

Figure 3.2 Example of the Integrative Genomics Viewer Navigator (IGVNav) interface with associated features. ***a*** *IGVNav is a simple plugin for IGV that provides a separate application window for recording results of manual review. The 1-Base? button can be selected for 1-base input files (default is 0-base). The "S" button will sort the read sequences in the data tracks so that mismatches appear at the top. The navigation bar displays variant information and allows for movement between variants. The Call, Tags, and Notes sections allow manual reviewers to annotate variants (**Table 3.1**), which is reflected in the output file. The Save button is used to update the output file.* ***b*** *An IGVNav input file consists of a header line and data for the first five columns (chromosome [chr], start coordinate [start], stop coordinate [stop], reference allele [ref], and variant allele [var]). Each line represents a variant that will be individually visualized using IGV. c During manual review, the input file is updated by clicking on the Save button. This will print the call, tags, and notes associated with individual variants to the original input file.*

3.4.3 Step-by-step guide: setting up IGV and IGVNav for manual review

Manual review setup involves six discrete steps (**Figure 3.3a**). First, an IGV session should be opened, and the appropriate reference genome should be selected/loaded. The reference genome species and build must match those used for alignment. Second, the IGV session should be populated with data tracks. When tumor DNA, normal DNA, and other DNA or RNA read alignments are available, they can all be loaded within a single IGV session. Step 3, optionally, allows for population of additional tracks that can assist in manual review. Step 4, also optional, recommends that tracks be colored by reads (right click on data track → Color alignments by → read strand) and the centered locus is visualized (View → Preferences → Alignments → Show center line). After initial setup of IGV, step 5 requires opening IGVNav and step 6 requires loading the manual review input file.

3.4.4 Step-by-step guide: performing manual review

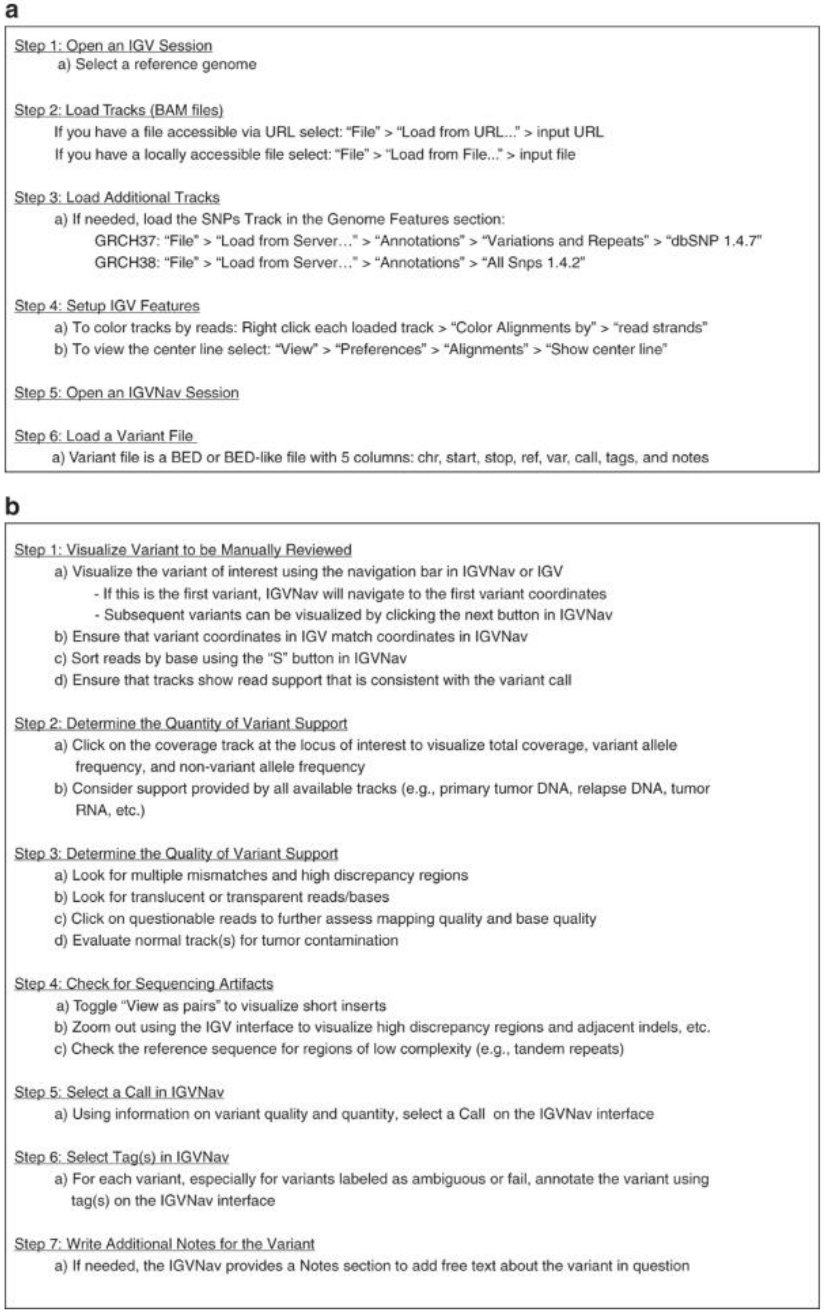
After initial setup, seven additional steps must be followed to properly review each variant (**Figure 3.3b**). First, the variant must be located by either using the navigation bar in IGVNav or by manually inserting coordinates into the IGV Genome Ruler. Variant-supporting reads can be visualized at the top of each data track by clicking the "S" button in IGVNav, or by using IGV options (right click on data track → Sort alignments by → base).

Step 2 evaluates the quantity of variant support. Selecting the locus of interest within the coverage track will ascertain strand direction, total coverage, and variant allele frequencies (VAFs). Strand direction might indicate a Directional (D) artifact (**Appendix 3, Figure S5**). Total coverage might indicate No Count Normal (NCN) (**Appendix 3, Figure S6**), Low Count Normal (LCN) (**Appendix 3, Figure S7**), or Low Count Tumor (LCT) (**Appendix 3, Figure S8**). VAFs might indicate Multiple Variants (MV) (**Appendix 3, Figure S9**) or Low Variant Frequency (LVF) (**Appendix 3, Figure S10**).

Step 3 evaluates the quality of variant support. Directly visualizing reads identifies Multiple Mismatches (MM) (**Appendix 3, Figure S11**) or High Discrepancy Regions (HDR) (**Appendix 3, Figure S12**). Reads that are translucent or transparent indicate Low Mapping (LM) quality (**Appendix 3, Figure S13**). Mapping quality information can be viewed by clicking on the read in question and viewing the Mapping section (e.g., Mapping = Primary @MAPQ 0). Base quality can also be evaluated in this popup in the Base section (e.g., Base = A @ QV 41). Similar to mapping quality, base quality is reflected by the transparency of the letter. The final part of step 3 is to ensure lack of variant support in normal track(s), (i.e., Tumor in Normal [TN] [**Appendix 3, Figure S14**]).

Step 4 requires identifying sequencing artifacts. First, toggle between View as pairs (right click each data track → View as pairs) to visualize Short Inserts (SI/SIO) (**Appendix 3, Figure S15**). Then use the zoom in ("+") and zoom out ("–") buttons on the Genome Ruler to identify Adjacent Indels (AI) (**Appendix 3, Figure S16**), High Discrepancy Regions (HDR) (**Appendix 3, Figure S12**), exclusive support from reads with Same Start/Ends (SSE) (**Appendix 3, Figure S17**), and support only at the Ends of reads (E) (**Appendix 3, Figure S18**). Finally, evaluating the reference sequence elucidates low complexity regions such as Mononucleotide repeats (MN) (**Appendix 3, Figure S19**), Dinucleotide repeats (DN) (**Appendix 3, Figure S20**), and Tandem Repeats (TR) (**Appendix 3, Figure S21**). If reviewer concerns cannot be described with previously defined tags, the reviewer can use the Ambiguous Other (AO) tag and comment in the Notes section (**Appendix 3, Figure S22**).

Steps 5 through 7 require synthesizing available information to manually review the variant. This involves selecting a call, tag(s), and optionally, providing free text in the Notes section of IGVNav.



a

Step 1: Open an IGV Session
- a) Select a reference genome

Step 2: Load Tracks (BAM files)
- If you have a file accessible via URL select: "File" > "Load from URL..." > input URL
- If you have a locally accessible file select: "File" > "Load from File..." > input file

Step 3: Load Additional Tracks
- a) If needed, load the SNPs Track in the Genome Features section:
  - GRCH37: "File" > "Load from Server…" > "Annotations" > "Variations and Repeats" > "dbSNP 1.4.7"
  - GRCH38: "File" > "Load from Server…" > "Annotations" > "All Snps 1.4.2"

Step 4: Setup IGV Features
- a) To color tracks by reads: Right click each loaded track > "Color Alignments by" > "read strands"
- b) To view the center line select: "View" > "Preferences" > "Alignments" > "Show center line"

Step 5: Open an IGVNav Session

Step 6: Load a Variant File
- a) Variant file is a BED or BED-like file with 5 columns: chr, start, stop, ref, var, call, tags, and notes

b

Step 1: Visualize Variant to be Manually Reviewed
- a) Visualize the variant of interest using the navigation bar in IGVNav or IGV
  - If this is the first variant, IGVNav will navigate to the first variant coordinates
  - Subsequent variants can be visualized by clicking the next button in IGVNav
- b) Ensure that variant coordinates in IGV match coordinates in IGVNav
- c) Sort reads by base using the "S" button in IGVNav
- d) Ensure that tracks show read support that is consistent with the variant call

Step 2: Determine the Quantity of Variant Support
- a) Click on the coverage track at the locus of interest to visualize total coverage, variant allele frequency, and non-variant allele frequency
- b) Consider support provided by all available tracks (e.g., primary tumor DNA, relapse DNA, tumor RNA, etc.)

Step 3: Determine the Quality of Variant Support
- a) Look for multiple mismatches and high discrepancy regions
- b) Look for translucent or transparent reads/bases
- c) Click on questionable reads to further assess mapping quality and base quality
- d) Evaluate normal track(s) for tumor contamination

Step 4: Check for Sequencing Artifacts
- a) Toggle "View as pairs" to visualize short inserts
- b) Zoom out using the IGV interface to visualize high discrepancy regions and adjacent indels, etc.
- c) Check the reference sequence for regions of low complexity (e.g., tandem repeats)

Step 5: Select a Call in IGVNav
- a) Using information on variant quality and quantity, select a Call on the IGVNav interface

Step 6: Select Tag(s) in IGVNav
- a) For each variant, especially for variants labeled as ambiguous or fail, annotate the variant using tag(s) on the IGVNav interface

Step 7: Write Additional Notes for the Variant
- a) If needed, the IGVNav provides a Notes section to add free text about the variant in question

Figure 3.3 Step-by-step instructions for setting up and executing somatic variant refinement via manual review. ***a*** *Method for setting up Integrative Genomics Viewer (IGV) and Integrative Genomics Viewer Navigator (IGVNav) for manual review.* ***b*** *Method for analyzing each variant during manual review.*

3.4.5 Validation of the manual review SOP

We assessed whether the manual review SOP improved accuracy of somatic variant refinement using an acute myeloid leukemia (AML) case with genome sequence data, extensive variant calling, and orthogonal validation (**Figure 3.4**).[22] To emulate normal conditions for genome sequencing manual review, we down-sampled the unaligned BAM files to 30× and 50× coverage for normal and tumor samples, respectively. Sequencing data was aligned to the reference genome (GRCh38) and variants were detected using the McDonnell Genome Institute's cancer genomics workflow.[100] Using the union of MuTect[18] and VarScan,[19] 143,042 potential variants were identified. A subset of these variants (n = 5,090) had orthogonal validation sequencing at ~1,000× coverage. Coordinates from the platinum variant list, published by Griffith et al., were lifted over to GRCh38 and used to label 1,186 variants as true positives (TPs). The remaining 3,904 variants were labeled as false positives (FPs). A random subset of 300 variants (150 TPs; 150 FPs) were selected for manual review. After receiving basic instruction on how to set up IGV and call variants using the required four classes (S, G, A, F), blinded novice reviewers manually reviewed 200 variants in two batches of 100 using the down-sampled genome sequencing BAM files. Subsequently, the reviewers read the SOP and reviewed two more batches of 100 variants. The final batch of 100 variants were among the 200 assessed prior to reading the SOP. Accuracy was assessed by comparing the manual review calls with the orthogonal validation labels. Inter-reviewer variability was calculated by developing a correlation matrix for all four calls across the four reviewers for each variant. Correlation for identical calls was 1, correlation for conflicting calls (e.g., fail and somatic) was 0, and correlation for semi-conflicting calls (e.g., fail and ambiguous) was 0.5 (**Appendix 4, Table S1**). The sum of the matrix was divided by the maximum possible score (i.e., 16 points) to create a relative metric for inter-reviewer agreement. The average agreement scores from before and after reading the SOP were compared. To determine if reviewers were using tags appropriately, tags assigned to false positives by novice reviewers were compared with gold standard tags created by expert reviewers for false positives reviewed after reading the SOP (**Figure 3.4a**).

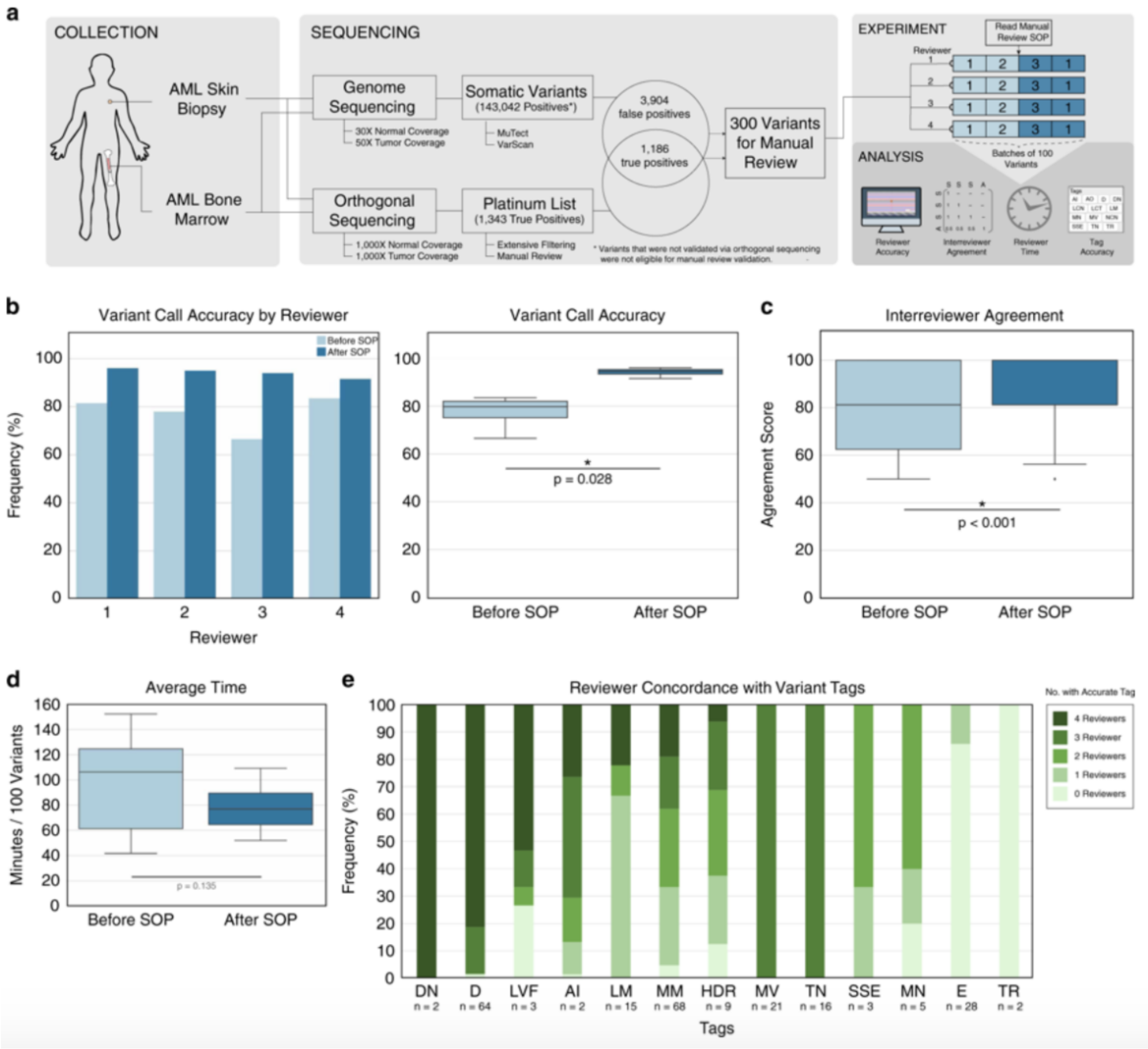


Figure 3.4 Validation of the manual review standard operating procedure (SOP). ***a*** *Sequencing data from an acute myeloid leukemia (AML) case was used to test the impact of the SOP on accurately identifying somatic variants. A total of 300 variants that had genome sequencing and orthogonal sequencing were identified for the experiment. Four novice reviewers assessed 200 variants prior to and after reading the SOP to determine improvement in accuracy, reduction in inter-reviewer variability, change in reviewer time per variant, and appropriate use*

of tags. ***b*** *Reviewer accuracy was assessed before and after reading the SOP. The bar plot shows accuracy stratified by reviewer and the box plot shows the reviewers' cumulative median accuracy.* ***c*** *Box plot showing the median inter-reviewer agreement before and after reading the SOP. Agreement for each variant was calculated by assessing the correlation between the four reviewer calls using a correlation matrix as described in the Methods.* ***d*** *Box plot showing the median time required to conduct manual review before and after reading the SOP.* ***e*** *Frequency diagram showing the number of reviewers that correctly annotated false positive variants with gold standard tags, parsed by tag. AI Adjacent Indel, D Directional, DN Dinucleotide repeat, E End of reads, HDR High Discrepancy Region, LM Low Mapping, LVF Low Variant Frequency, MM Multiple Mismatches, MN Mononucleotide repeat, MV Multiple Variants, SSE Same Start End, TN Tumor in Normal, TR Tandem Repeat.*

3.5 Results

3.5.1 Annotations observed during manual review

Screenshots were created for the 22 annotations used during manual review (**Appendix 3, Figure S1-S22**). The illustrations and comments emphasize IGV features that highlight sequencing patterns, describe cautions for challenging tumor types, and indicate deviations from standard protocol.

3.5.2 Analysis of four variant calls

This SOP and IGVNav software support four classes of variant calls: somatic (S), germline (G), ambiguous (A), and fail (F) (**Table 3.1**). For a call to be labeled as somatic, the variant must have sufficient read data support in the tumor with absence of obvious sequence artifacts (**Appendix 3, Figure S1**). Conversely, a germline variant is an alteration that has sufficient support in the normal, beyond what can be attributable to tumor contamination (**Appendix 3, Figure S2**). Barring inadequate sequencing depth and/or impact from copy-number alterations, the VAF for germline variants should be near 100% or 50% in both the normal and tumor tracks, indicative of homozygosity or heterozygosity, respectively. Ambiguous calls should be made when there is insufficient evidence to confidently label a variant with any other call class. The example in

Appendix 3, Figure S3 shows no support for the variant in the normal track and 14 reads of support in the tumor. However, most of the reads are on negative strands and some have multiple mismatches. If a reviewer has any residual doubt about failing a variant, then the variant should be labeled ambiguous. To fail a variant, the reviewer must confidently determine that the variant was called because of a sequencing or analysis artifact. For example, **Appendix 3, Figure S4** details a variant that was erroneously identified by an automated caller because reads had been aligned to a high discrepancy region.

Table 3.1 List and description of Integrative Genomics Viewer Navigator (IGVNav) features

Call Name	**Call**	**Description**
Somatic	S	Variant has sufficient support in the tumor with absence of obvious sequencing artifacts
Germline	G	Variant that has sufficient support in the normal sample beyond what is considered attributable to tumor contamination of the normal
Ambiguous	A	Variant does not meet acceptable criteria for any other label
Fail	F	Variant with low variant support and/or reads that indicate sequencing artifacts

Tag Name	**Tag**	**Description**
Adjacent Indel	AI	Variant is attributable to misalignment caused by a nearby insertion or deletion
Ambiguous Other	AO	Variant is surrounded by inconclusive genomic features that cannot be explained by other tags
Directional	D	Variant is only (or mostly) found on reads in the same direction (positive or negative)
Dinucleotide repeat	DN	Variant is adjacent to a region in the reference genome that has two alternating nucleotides (e.g., TGTGTG…)
End of reads	E	Variant is only seen close to the end (within 30 base pairs) of variant-supporting reads
High Discrepancy Region	HDR	Variant is supported by reads that have other recurrent mismatches across the track and in multiple tracks
Low Count Normal	LCN	Variant has inadequate coverage in the normal track, thus preventing effective comparison with the tumor track

Low Count Tumor	LCT	Variant has inadequate coverage in the tumor track, thus preventing effective comparison with the normal track
Low Mapping quality	LM	Variant is mostly supported by reads that have low mapping quality
Low Variant Frequency	LVF	Variant has low variant allele frequency (VAF) samples
Multiple Mismatches	MM	Variant is supported by reads that have other mismatched base pairs
Mononucleotide repeat	MN	Variant is adjacent to a region in the reference genome that has a single-nucleotide repeat (e.g., AAAAAA…)
Multiple Variants	MV	Variant locus has read support for three or more alleles
No Count Normal	NCN	Variant has no coverage in the normal track, thus preventing effective comparison with the tumor track
Short Inserts	SI	Variant is found mostly on small nucleic acid fragments whereby sequencing from each end results in overlapping reads
Short Inserts Only	SIO	Variant is exclusively found on small nucleic acid fragments such that sequencing from each end results in overlapping reads
Same Start End	SSE	Variant is only observed in reads that start and stop at the same positions
Tumor in Normal	TN	Variant has read support in the normal track
Tandem Repeat	TR	Variant is adjacent to a region in the reference genome that has three or more alternating nucleotides (e.g., GTGGTGGTG…)

3.5.3 Analysis of 19 variant tags

It is especially important to annotate fail and ambiguous calls with 1 or more of the 19 tags on the IGVNav interface (**Table 3.1**). Each tag represents a sequencing pattern or artifact that is commonly observed during manual review. These patterns can arise during DNA fragmentation, library construction, sequencing, read alignment, or variant calling. Alternatively, some concerns observed during manual review can be caused by simple structural aberrations or more complex issues intrinsic to the tumor being evaluated. Below, we describe how these concerning reads are created within the sequencing pipeline and detail the resulting pattern observed in IGV.

The tumor type and tissue origin can play a role in generating patterns observed during manual review. For example, hematologic tumors or highly metastatic tumors can cause Tumor in Normal (TN) patterns due to the presence of tumor cells in the normal biopsy (**Appendix 3, Figure S14**). Generally, it is important to characterize the average level of contamination across an individual sample to determine an acceptable threshold for TN. Tumor sample preparation can also impact manual review through sequencing of degraded nucleic acids (e.g., formalin-fixed, paraffin-embedded samples)[101] giving rise to Short Inserts (SI) or Short Inserts Only (SIO). When generating paired-end reads, degraded and/or short molecules will produce two sequences that have overlapping alignments. This can exaggerate variant support because most variant callers will consider the overlapping alignments as two independent pieces of evidence, despite representing a single originating DNA fragment (**Appendix 3, Figure S15**). Short inserts can be visualized in IGV by viewing reads as pairs and looking for horizontal gray bands (representing overlap) in the middle of the paired read alignments.

Additional errors can arise during fragmentation, library construction, and enrichment. DNA quality and quantity, capture reagent balance and efficiency, sample balance in multiplexed preparations, and other factors can impact the uniformity of coverage for a given sample. For example, a selection bias might skew which molecules are amplified/sequenced, resulting in an uneven distribution of sequencing (coverage) across the desired genome space.[102] These errors are labeled as No Count Normal (NCN) (**Appendix 3, Figure S6**), Low Count Normal (LCN) (**Appendix 3, Figure S7**), and Low Count Tumor (LCT) (**Appendix 3, Figure S8**). NCN and LCN are defined by no or few reads in the normal tracks and LCT is defined by few reads in the tumor track. Also, given that many real variants have a low VAF, due to tumor heterogeneity or low purity tumors, the combination of Low Variant Frequency (LVF) (**Appendix 3, Figure S10**) and

LCT can prevent a true variant from being confidently called. Our lab has often adopted a minimum VAF threshold of 5% and a coverage threshold of 20 reads for both the tumor and normal tracks. The rationale for the normal track coverage threshold is that if a sequencing artifact is present at a relatively low frequency (<5% occurrence), and if the normal track has <20 reads, it is difficult to confidently rule out the presence of a sequencing artifact. For experiments with higher average coverage, the minimum VAF threshold can be reduced accordingly.

After fragmentation and library preparation, nucleic acids are amplified using polymerase chain reaction (PCR), which can introduce Directional (D) and Same Start/End (SSE) artifacts. Directional artifacts occur when variant support is only apparent on reads in a specific direction (i.e., positive or negative). Typically, this occurs because the sequencing context affects the polymerase in one direction more than the reverse complement (**Appendix 3, Figure S5**).[103] SSE artifacts occur when a molecule is preferentially amplified and not removed through read deduplication programs.[104] This artifact can be confirmed when all variant support reads have the same (or very similar) start and end position after alignment (**Appendix 3, Figure S17**).

The next step in the pipeline is sequencing. Sequencing errors are defined as nucleotides misread by the sequencing instrument, which can be caused by inefficiencies in sequencing chemistry, technical errors made by the camera system, interference from neighboring clusters, instrument software errors, etc. One type of sequencing error, “dephasing,” occurs when a nucleotide without a proper 3’ -OH blocking group is incorporated or is not properly cleaved. The affected fragment(s) lose synchrony with the cluster, contributing to background noise.[105] Ends of reads (E), which occurs when variant support is exclusively found at the end of read sequences (within 30 base pairs), is indicative of a dephasing error (**Appendix 3, Figure S18**).[106] These errors occur with low probability; however, as the read length increases, the summation of errors can pollute the light signal. Because the light signal is used to calculate quality scores, the asynchronous signal should decrease sequence base quality, which may assist in elucidating artifacts caused by dephasing errors.

Many artifacts arise from incorrect alignment of sequence reads to a reference genome. These artifacts include Mononucleotide repeats (MN), Dinucleotide repeats (DN), Tandem Repeats (TR), High Discrepancy Regions (HDR), Low Mapping (LM), Multiple Mismatches (MM), Adjacent Indel (AI), and Multiple Variants (MV). MN (**Appendix 3, Figure S19**), DN (**Appendix 3, Figure S20**), and TR (**Appendix 3, Figure S21**) are attributable to regions of low complexity adjacent to the variant locus. They typically occur when there is a base pair deletion or insertion adjacent to one, two, or greater than two base pair repeats, respectively. HDR, LM, MM, and MV occur when single reads map to multiple and/or incorrect regions. This is typically caused by (1) homologous sequences at multiple loci, (2) highly variable regions between or within individuals (e.g., variable, diversity, and joining (VDJ) regions in immune cells), (3) high error rates in reads, and/or (4) errors in the reference genome. HDRs are apparent when multiple reads contain the same mismatches with the reference genome at various locations (**Appendix 3, Figure S12**). LM can be determined by looking for translucent reads (**Appendix 3, Figure S13**). MM is used when variants are supported by reads that disagree with the reference genome at multiple loci across the same read, indicating low sequencing quality or misalignment (**Appendix 3, Figure S11**). Similarly, MV is defined by read support for three or more different alleles at a given locus, which might indicate poor quality or misaligned reads (**Appendix 3, Figure S9**). AI is used when a structural variant or a small indel in a repetitive region causes local misalignment and creation of an apparent single-nucleotide variant (SNV)/indel (**Appendix 3, Figure S16**). Observing these artifacts requires careful scrutiny of the reference genome, base quality, and mapping quality.

In rare instances, if the pre-existing tags cannot adequately annotate a variant, it can be labeled as Ambiguous Other (AO). Given that this tag is nondescriptive, it is recommended to include free text in the Notes section to justify the tag and associated variant call. In the example provided (**Appendix 3, Figure S22**), the insertion variant shows a low complexity region with increased G/C content that is not contained within a tandem repeat region. This observation can be annotated using the AO tag.

3.5.4 Validation of the manual review SOP

Manual review performed by novice reviewers after reading the SOP improved identification of somatic variants by 16.7% (77.4% vs. 94.1%; p value = 0.0298) (**Figure 3.4b**) and increased the average inter-reviewer correlation score by 12.7% (80.7 points vs. 93.4 points; p value < 0.0001) (see Methods) (**Figure 3.4c**). The SOP did not significantly impact time required to conduct manual review (**Figure 3.4d**). Additionally, correct use of tags was observed for annotations made after reading the SOP. When evaluating 86 false positives that had 238 tags confirmed by expert reviewers, 143 tags were correctly identified by at least three novice reviewers and only 36 tags were missed by all reviewers (**Figure 3.4e**).

3.6 Discussion

Identification and interpretation of variants is crucial for conducting translational research and guiding clinical management of cancer patients.[38] In general, implementation of this SOP has improved variant identification consistency, limiting the total number of false positives requiring downstream analysis. Given that variant annotation remains a major bottleneck in translational and clinical research.[35,45] reduction in false positives should substantially improve the overall efficiency of lab operations. Therefore, we advocate that others adopt a standardized process for variant refinement such as the SOP presented here.

There are intrinsic limitations associated with manual review that will not be rectified by this SOP. First, manual reviewers have reported reviewer fatigue, especially when evaluating tumors with a high variant burden. Second, despite extensive training, some amount of inter-reviewer variability will likely remain, especially for ambiguous variants. Third, manual review of variants might change over time as an individual begins to recognize the idiosyncrasies associated with a particular tumor subtype or sequencing platform. Finally, the scope of this SOP is limited to the manual review of somatic SNVs/indels in situations where tumor/normal samples are available; although, many of the aspects of the protocol, including setup and assessment, can be directly applied to other analyses (e.g., structural variant assessment). It is our intent to continuously improve this protocol through subsequent revisions (https://doi.org/10.1101/266262). This will include developing an SOP for tumor-only samples, incorporating features that improve somatic variant refinement, and developing machine learning approaches to alleviate manual review burden.

Many of the existing limitations of manual review could be addressed by automating somatic variant refinement. This would further standardize the massively parallel sequencing pipeline and reduce the labor burden required to identify putative somatic variants. Advancements in computational approaches provide an opportunity for the development of such a process.

3.7 Acknowledgements

E.K.B. was supported by the National Cancer Institute (T32GM007200 and U01CA209936). B.J.A. and K.M.C. were supported by the Siteman Cancer Center (T32CA113275). S.J.S. is funded by the National Library of Medicine (NIH NLM R01LM012222 and NIH NLM R01LM012482). A.W. was supported by the National Cancer Institute (NIH NCI F32CA206247). M.G. is funded by the National Human Genome Research Institute (NIH NHGRI R00HG007940). O.L.G. is funded by the National Cancer Institute (NIH NCI K22CA188163 and NIH NCI U01CA209936). Additional funds were provided by the Washington University School of Medicine. This research would not have been possible without exemplary work performed by Michael McLellan, Heather Schmidt, Jennifer Hodges, Yat Tang, Li Ding, and others at the McDonnell Genome Institute to develop the original Medical Genomics/Manual Review Guidelines.

3.8 Author Contributions

E.K.B. wrote the manuscript, the manual review standard operating procedure, and conducted/analyzed all experiments. SPP, LMS, ADS, and MR participated in the manual review validation study. P.R., K.M.C., K.K., B.J.A., S.P.P., L.M.S., K.C.C., A.M.D., C.R., N.C.S., J.K., Z.L.S., J.H., M.S.S., F.G. and L.T. contributed to the manuscript. M.M. and A.W. wrote and updated the IGVNavigator tool. L.T., K.K., Z.L.S. and B.J.A. developed an initial version of the Manual Review Guidelines. S.J.S., M.G., and O.L.G. supervised the project and revised the paper.

Chapter 4: Open-sourced CIViC annotation pipeline (OpenCAP) to identify and annotate clinically relevant variants using single molecule molecular inversion probes

4.1 Preamble

The following chapter has been accepted as a manuscript for publication:

Barnell E.K., Waalkes A., Mosior M.C., Penewit K., Cotto K.C., Danos A.M., Sheta L.M., Campbell K.M., Krysiak K., Rieke D., Spies N.C., Skidmore Z.L., Pritchard C.C., Fehniger T.A., Uppaluri R., Govindan R.G., Griffith M., Salipante S.J., Griffith O.L. Open-sourced CIViC annotation pipeline (OpenCAP) to identify and annotate clinically relevant variants using single molecule molecular inversion probes. JCO Clinical Cancer Informatics, Accepted. September 3rd, 2019.

As an author of the published manuscript, and in compliance with the editorial policies at JCO Clinical Cancer Informatics, the cited publication is included in full in the following chapter. My role in this project was to develop the OpenCAP software, generate an exemplary panel for validation, and assemble the manuscript for publication. A complete list of author contributions is included within the publication (**Chapter 4.8**).

4.2 Summary

Clinical targeted sequencing panels are important for identifying actionable variants for cancer patients; however, existing approaches do not provide transparent and rationally-designed panels to accommodate the rapidly growing knowledge within oncology. We used the Clinical Interpretations of Variants in Cancer database (CIViC; https://civicdb.org) to develop an Open-sourced CIViC Annotation Pipeline (OpenCAP; https://opencap.org). OpenCAP provides methods to identify variants within the CIViC database, build probes for variant capture, employ probes on

prospective samples, and link somatic variants to CIViC clinical relevance statements. OpenCAP was tested using a single-molecule molecular inversion probe (smMIP) capture design on 27 cancer samples from 5 tumor types. In total, 2,027 smMIPs were designed to target 111 eligible CIViC variants (61.5 kb of genomic space). When compared to orthogonal sequencing, CIViC smMIP sequencing demonstrated a 95% sensitivity for variant detection (n=61/64 variants). Variant allele frequencies for variants identified on both sequencing platforms were highly concordant (Pearson correlation=0.885; n=61 variants). Moreover, for individuals with paired tumor/normal samples (n=12), 182 clinically relevant variants missed by orthogonal sequencing were discovered by CIViC smMIPs sequencing. The OpenCAP design paradigm demonstrates the utility of an open-source and open-access database built on attendant community contributions with peer-reviewed interpretations. Use of a public repository for variant identification, probe development, and variant interpretation provides a transparent approach to build dynamic next-generation sequencing–based oncology panels.

4.3 Introduction

Despite recognition that genomics plays an important role in tumor prognosis, diagnosis, and treatment, scaling genetic analysis for routine analysis of most tumor specimens has been unattainable.[5,107] Barriers preventing widespread incorporation of genomic analysis into treatment protocols include: costs associated with genomic sequencing and analysis,[25] computational limitations preventing timely identification of relevant variants,[25] and rapidly evolving knowledge of the clinical actionability of variants.[108] Technological improvements in sequencing and data analysis continue to reduce these first two limitations, however, less progress has been made in integrating dynamic genomic annotation into clinical workflows. Over 22% of oncologists have acknowledged limited confidence in their own understanding of how genomic knowledge applies to patients' treatment and 18% reported testing patients' genetics infrequently.[109] In the face of exponential growth in clinically relevant genomic findings, driven by precision oncology efforts, there will likely be increased inability for physicians to command the most current information, resulting in increasing delay between academic discovery and clinical utility. This information gap has been described as the "interpretation bottleneck".[45,108,109]

Alleviating the interpretation bottleneck will require co-development of targeted sequencing panels, bioinformatic tools, and variant knowledgebases that effectively elucidate and annotate clinically actionable variants from sequencing data.[110,111] These requirements each raise separate challenges. With regards to targeted panel development, commercial and academic pan-cancer clinical gene capture panels have now become commonplace, with at least two obtaining Food and Drug Administration (FDA) approval (FoundationONE CDx[34] [Cambridge, MA] and MSK-IMPACT[112] [New York, NY]). Even so, few panels indicate how genomic loci are selected for panel inclusion, and none have proposed a sustainable or scalable mechanism to allow for panel evolution over time in response to knowledge advances in molecular oncology. With regards to bioinformatic tool development, the OncoPaD[113] portal provides one of the only methods to create rational designed panels by linking clinically relevant variants to genomic loci based on a cohort of tumor samples, however, this database is not directly linked to actively updated clinical interpretations with detailed underlying evidence. The final challenge of building knowledgebases for variant interpretation perhaps poses even greater and more persistent challenges. Commercial entities typically rely on the manual curation and organization of research findings into structured databases, which are expensive to create and maintain, forcing companies to limit public access or to charge for use. The resulting lack of transparency creates inefficiencies in the field through unnecessary replication of curation effort and suboptimal communication with clinicians, ultimately hindering development of effective patient treatment plans. Separately, governmental and academic institutions have developed variant interpretation resources like COSMIC[29], ClinVar[27], and cBioPortal[114,115] that have drastically improved research efforts and academic discovery, however, these resources do not have well-supported (evidence-based) clinical relevance summaries for cancer variants that can be easily accessed and utilized by physicians. Several resources provide detailed clinical interpretation of cancer variants (e.g., oncoKB[31], JAX Clinical Knowledgebase[116], and others) but these databases are either limited by license restrictions or closed curation models.

To address these limitations, we developed a method to identify, capture, and annotate variants using the Clinical Interpretation of Variants in Cancer (CIViC) database (**http://www.civicdb.org/**). The CIViC database is a freely-accessible (public domain content), publicly curated, expert-moderated, repository of therapeutic, prognostic, predisposing, and diagnostic information in precision oncology.[35] The database provides a powerful platform for panel development and variant annotation for several reasons: 1) each variant within CIViC is described by clinical relevance summaries linked to medical literature, 2) history of curation within CIViC is stored and publicly available to all users, and 3) CIViC has an open-source, open-access applied programming interface (API) for external query. Using the CIViC database and API, we developed the Open-sourced CIViC Annotation Pipeline (OpenCAP) for creating custom capture panels, executing capture panel sequencing on prospective samples, identifying variants from sequencing data, and annotating variants for clinical relevance. An exemplary clinical capture panel was created using OpenCAP to demonstrate utility. Specifically, variants within the CIViC database were identified based on clinical relevance and single-molecule molecular inversion probes (smMIPs) were designed to target variants of interest. This panel was employed on cancer samples to evaluate design and identified somatic variants were compared to orthogonal sequencing. Variants identified via smMIPs capture were linked back to the CIViC database for clinical annotation (**Figure 4.1**). Ultimately, this method could be used to rapidly and efficiently link variants to clinical relevance summaries, enabling the development of custom capture panels for a variety of clinical and research scenarios.

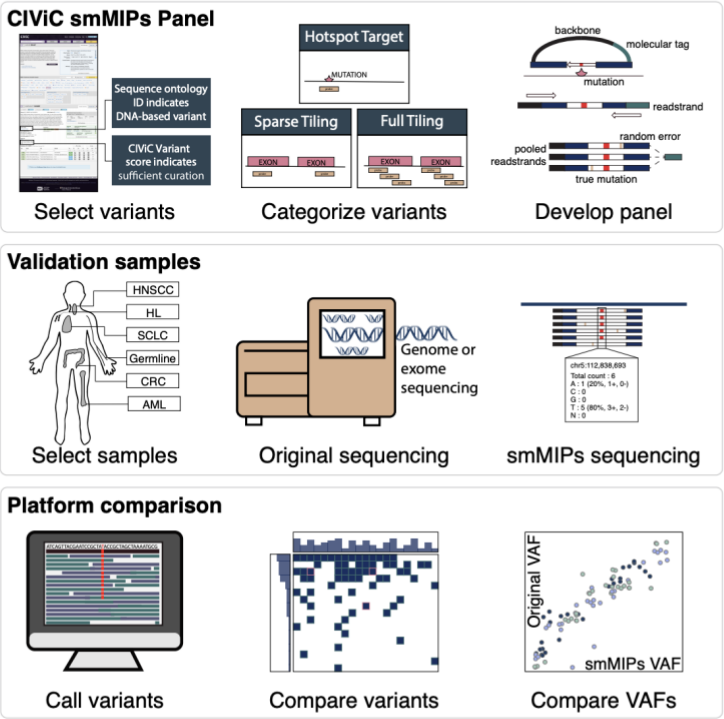


Figure 4.1 Methods for CIViC smMIPs development and validation using OpenCAP. *The first series describes CIViC smMIPs development. Variants were selected using sequence ontology IDs and the CIViC Variant Evidence Score. Subsequently, eligible variants were categorized based on length and smMIPs reagents were designed to target regions of interest. The second series describes sample selection and sequencing methods. In total, there were 22 tumor samples derived from 5 tumor subtypes. Of these 27 samples, 15 had tumor and paired normal samples and 7 were tumor-only samples. The third series shows the analysis used to validate the CIViC smMIPs design. Variants were called using the pipeline described in the methods, accuracy was attained by comparing variants observed on original sequencing to variants observed using the CIViC smMIPs capture panel. Variant allele frequencies across both platforms were also compared.*

4.4 Results

4.4.1 Identification of eligible CIViC variants for smMIP targeting

At the time of the CIViC smMIP capture panel design, there were 988 variants from 275 genes within the CIViC database with at least one evidence item. After filtering based on the Variant Evidence Score and the SOID (see **Appendix 5**), smMIPs were designed to cover all eligible CIViC variants. A set of 2,097 probes were developed and tested on control samples. Of these, 70 probes showed poor capture efficiency and were eliminated from the panel. Removal of the underperforming probes affected 32 variants across 16 genes. The final capture reagent targeted 111 CIViC variants spanning ~61.5 kb of genomic space (**Appendix 6, Table S1**). When compared to other pan-cancer panels, the CIViC capture panel showed high overlap with previously defined clinical variants. For example, the CIViC smMIPs capture panel covered 10 of the 13 well-defined variants on FoundationOne CDx (*EGFR* - Exon 19, L858R, and T790M; *BRAF* - V600E/K; *ERBB2* Amplification; KRAS G12/13; BRCA1; and BRCA2).[117] The 3 variants on FoundationOne CDx that were not originally covered by the smMIPs panel (KRAS wildtype, NRAS wildtype, and ALK rearrangements) have all since attained a Variant Evidence Score that would be sufficient for inclusion in a panel built today Of the 111 targeted variants, 71 required hotspot targeting, 14 variants required sparse exon tiling, and 26 required full exon tiling. The 111 variants covered by the CIViC smMIPs capture panel were based on 1,168 clinically relevant evidence items whereby 820 (70%) evidence items predicted response to a therapeutic, 232 (20%) detailed prognostic information, 52 (4%) indicated diagnostic information, and 64 (6%) evidence items supported predisposition to cancer (**Figure 4.2**).

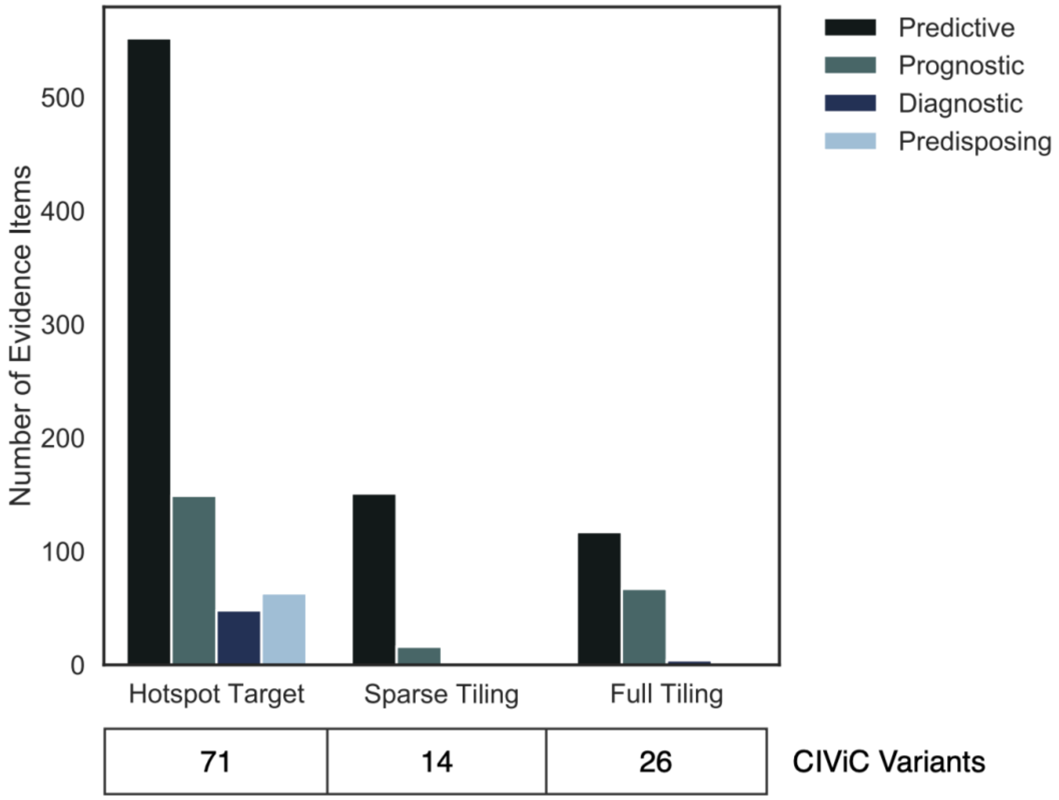


Figure 4.2 Regions targeted by the CIViC single molecule molecular inversion probes (smMIPs). *Variants that were eligible for CIViC smMIPs development were divided into various coverage methods based on sequence ontology identification number (SOID) and length. The bar graph shows the total number of evidence items used for each of the groups parsed by the evidence type.*

4.4.2 Tumor samples used to validate CIViC smMIPs design

Samples used to validate the CIViC smMIPs capture panel design were derived from 5 different cancer genomic studies (**Appendix 6, Table S2**). Tumor and paired normal samples were obtained from 5 individuals with head and neck squamous cell carcinoma (HNSCC), 9 individuals with small cell lung cancer (SCLC)[74], and 1 individual with Hodgkin's lymphoma (HL). Tumor-only samples were obtained from 1 individual with HL, 1 individual with acute myeloid leukemia (AML)[22], and 5 individuals with colorectal cancer (CRC). In total, 37 samples were evaluated from

22 individuals. Samples from the CRC cohort were formalin fixed paraffin embedded (FFPE) and all other samples were fresh frozen tissue.

Each of the 22 individuals had previously undergone whole exome or whole genome sequencing, somatic variant calling, and somatic variant refinement via manual review (see **Appendix 5**). Considering original sequencing there were 12,602 putative somatic variants called for these 22 samples. The average variant burden was 573 variants per sample with a range of 2 to 3,900 variants per sample. Variant coordinates from these samples were compared to the genomic region covered by the CIViC smMIP capture panel to determine potential validating variants. In total, there were 84 variants identified via original sequencing that overlapped with the CIViC smMIPs capture panel.

4.4.3 smMIP sequencing and data analysis

<u>Initial quality check</u>

The average number of tags captured for all samples was 5.4 million (standard deviation = 3.3 million tags). One HNSCC normal sample failed smMIPs capture, two HNSCC tumor samples had significantly fewer reads than the rest (i.e., >1 standard deviation), and one HL tumor sample had reduced tag complexity relative to the rest (i.e., <600,000 unique captured smMIPs). Sequencing failure for these four samples was attributable to poor template quality/quantity and not attributable to the capture reagents. All other samples passed sequencing quality checks. Post quality check, 31 samples derived from 19 individuals were eligible for reagent validation. These samples had 65 variants derived from orthogonal sequencing that had overlap with the CIViC smMIPs coverage (**Figure 4.3**). The average consensus read depth for these 65 variants was 2,942 reads (std = 4,697 reads).

Accuracy of CIViC smMIP variant identification compared to exome or genome variant identification

Of the 65 variants identified on exome sequencing, all but 4 were also identified using CIViC smMIP sequencing (**Figure 4.3**). One variant was missed due to lack of adequate coverage, two variants were missed due to low performing probes, and one variant was retrospectively considered ineligible due to smMIPs design (see **Appendix 5**). After removing this variant from the list of eligible variants, the CIVIC smMIP capture sequencing attained a 95% sensitivity for variant detection (n = 64 variants).

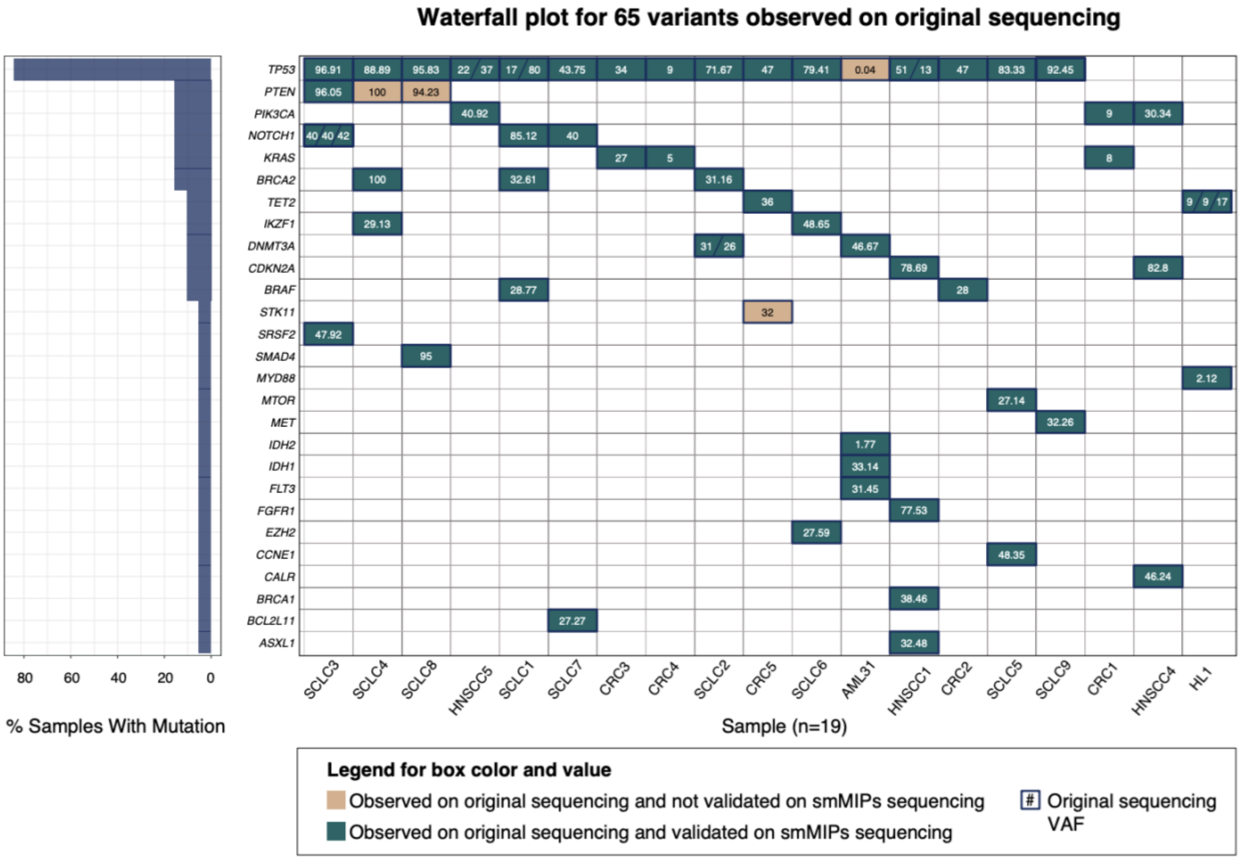
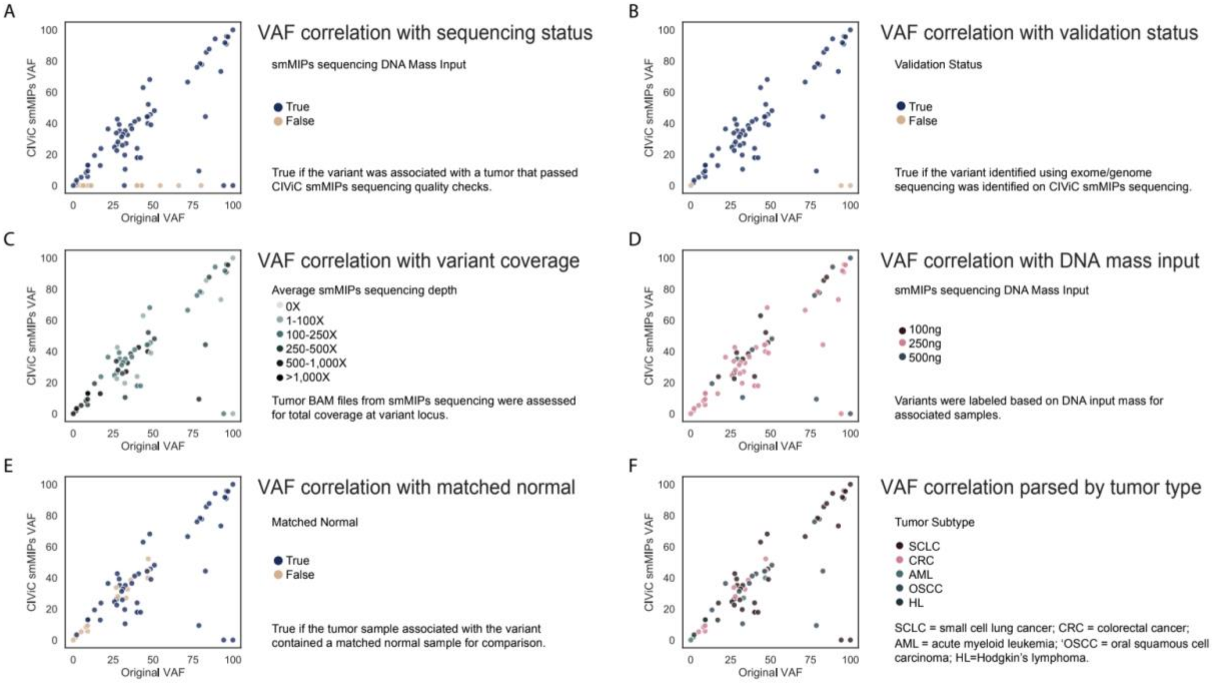


Figure 4.3 Waterfall plot showing extensive overlap between variants. *Each column represents a sample that had original exome or whole genome sequencing with subsequent orthogonal validation using the CIViC smMIPs sequencing. Rows represent mutated genes across all samples. Numbers within each box represent the variant allele frequency (VAF) observed on original exome*

or whole genome sequencing. Green boxes indicate that a variant was observed by CIViC smMIPs and validated with original exome or whole genome sequencing. Tan boxes indicate that the variant was observed on original exome or whole genome sequencing but not identified via the CIViC smMIPs capture panel. The left panel indicates the number of samples containing a mutation in the indicated gene.

VAF correlation between CIViC smMIPs sequencing and exome or genome sequencing

Variant allele frequencies (VAF) obtained via original sequencing were compared to the VAF obtained using the CIViC smMIPs. To compare VAF quantitation across platforms, the 19 variants obtained from samples that failed CIViC smMIPs sequencing quality check were eliminated (**Figure 4.4A)**. Subsequently, we eliminated the four variants that were not validated using the CIViC smMIPs reagents (**Figure 4.4B**). When comparing original VAF to CIViC smMIPs VAFs, Pearson correlation for the remaining 61 variants was 0.885. There were several variants whereby the VAF observed by the CIViC smMIPs sequencing was lower than that observed by the original sequencing. These outliers were not associated with tumor type, sequencing mass input, average coverage, presence of matched normal, or sample type (**Figure 4.4C-F**).



A VAF correlation with sequencing status

True if the variant was associated with a tumor that passed CIViC smMIPs sequencing quality checks.

B VAF correlation with validation status

True if the variant identified using exome/genome sequencing was identified on CIViC smMIPs sequencing.

C VAF correlation with variant coverage

Tumor BAM files from smMIPs sequencing were assessed for total coverage at variant locus.

D VAF correlation with DNA mass input

Variants were labeled based on DNA input mass for associated samples.

E VAF correlation with matched normal

True if the tumor sample associated with the variant contained a matched normal sample for comparison.

F VAF correlation parsed by tumor type

SCLC = small cell lung cancer; CRC = colorectal cancer; AML = acute myeloid leukemia; 'OSCC = oral squamous cell carcinoma; HL=Hodgkin's lymphoma.

Figure 4.4 Variant allele frequencies (VAFs) observed using original sequencing compared to smMIPs sequencing. *Each column represents a sample that had original exome or whole genome sequencing with subsequent orthogonal validation using the CIViC smMIPs sequencing. Rows represent mutated genes across all samples. Numbers within each box represent the variant allele frequency (VAF) observed on original exome or whole genome sequencing. Green boxes indicate that a variant was observed by CIViC smMIPs and validated with original exome or whole genome sequencing. Tan boxes indicate that the variant was observed on original exome or whole genome sequencing but not identified via the CIViC smMIPs capture panel. The left panel indicates the number of samples containing a mutation in the indicated gene.*

4.4.4 Analysis of variants only identified using CIViC smMIP sequencing

Using samples that had sequencing data for both tumor and matched normal (n = 12 samples), we evaluated whether the targeted CIViC smMIP sequencing could identify clinically relevant variants that had not been observed by the original sequencing. There were 273 variants recovered by CIViC smMIP sequencing that were not identified using original sequencing. After manually reviewing these variants within the original exome or genome alignments, 55 variants (20.1%) were identified as germline mutations. smMIP sequencing VAF distribution at 50% and 100% further supported that these variants were germline polymorphisms (**Figure 4.5A**). An additional

36 variants (13.2%) were thought to be caused by pipeline artifacts and attributable to assumptions underlying automated callers or alignment problems. The majority of these artifacts were associated with nucleotide repeats in the reference sequence (**Figure 4.5B**). There were 171 (62.6%) variants called as somatic using CIViC smMIPs that did not have any variant support on the original sequencing. For these variants, we calculated the binomial probability that ≤3 reads would support the variant given the original coverage (number of chances to get a variant supporting read) and the observed smMIPs variant allele frequency (likelihood that a read would show variant support). If the binomial probability of ≤3 variant-supporting reads was >95%, then it was considered statistically unlikely that a variant would be called using original sequencing data. Using this calculation, 162 variants (94.7%) showed insufficient coverage in the original sequencing for detection (**Figure 4.5C**). Finally, 11 variants (4.2%) were not called as somatic on original sequencing but did show some variant support in those original sequencing data. The VAFs observed on original sequencing data were strongly correlated with the VAFs observed using CIViC smMIP sequencing (Pearson r = 0.92) (**Figure 4.5D**). Reviewing manual review files from the original sequencing, we observed that 6 of these variants failed manual review due to low VAF, 4 variants had not been called by automated somatic variant callers, and 1 variant failed manual review due to a perceived sequencing artifact. In summary, there were 182 potentially clinically relevant somatic variants missed by original sequencing, primarily due to insufficient coverage, that contained CIViC variant annotations.

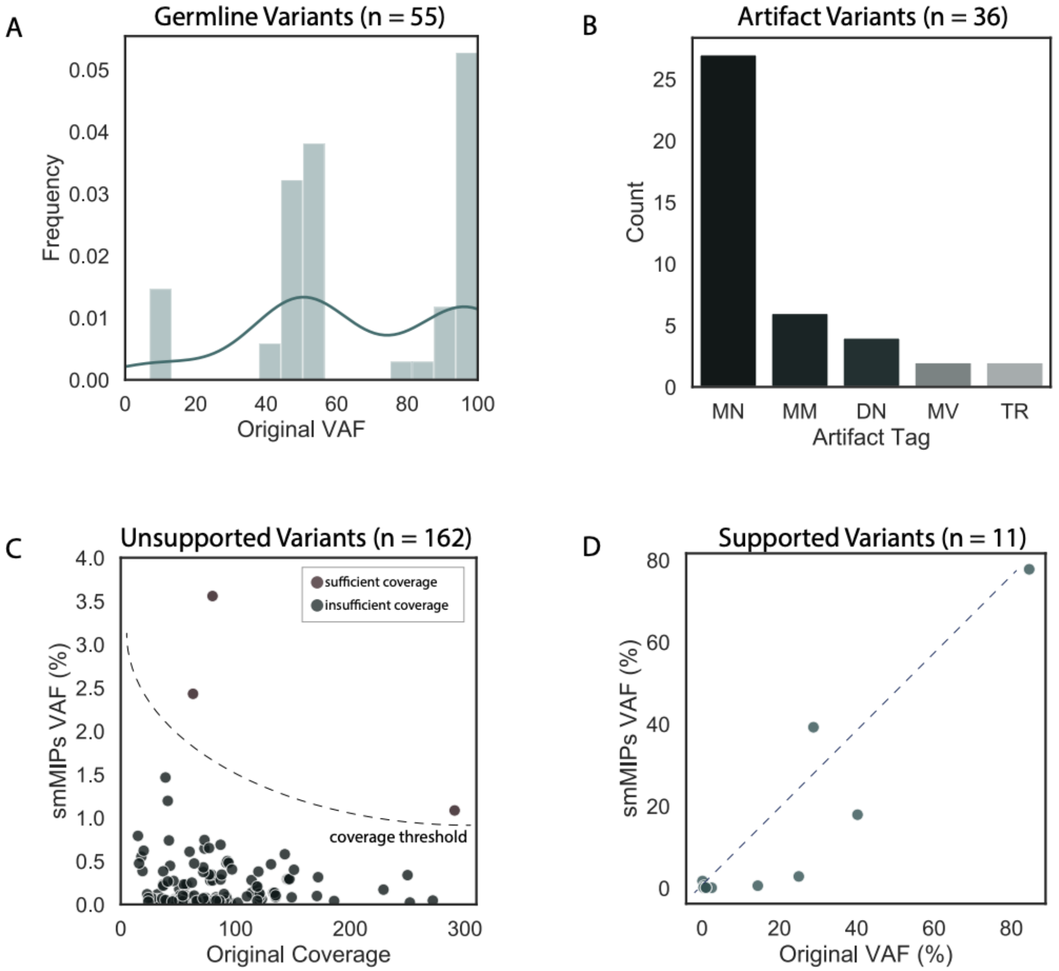


Figure 4.5 Analysis of variants rescued by CIViC smMIPs sequencing for samples. *There were 217 variants called as somatic by the CIViC smMIPs sequencing that were not identified by the original sequencing. All variants were manually reviewed using both CIViC smMIPs sequencing data and original sequencing data.* ***(A)*** *During manual review 55 variants were identified as germline. A histogram shows that the distribution of the smMIPs VAF for these germline variants are observed at 50% and 100% VAF, indicating heterozygosity and homozygosity, respectively.* ***(B)*** *An additional 36 variants were identified as sequencing artifacts. Most artifacts were either mononucleotide repeats (MN), dinucleotide repeats (DN), or tandem repeats (TR). Other artifacts include multiple mismatches (MM) or multiple variants (MV).* ***(C)*** *During manual review, 162 variants did not show any support in the original sequencing data. Most unsupported variants did not have sufficient coverage to be detected based on a binomial probability of at ≤3 variant-supporting reads (see Methods).* ***(D)*** *The remaining 11 variants had variant support in original sequencing but were not called as somatic in final original annotation. The scatter plot shows correlation between original VAF and CIViC smMIPs VAF for these variants.*

4.4.5 Annotation of CIViC smMIPs capture panel somatic variants using OpenCAP

Using the OpenCAP annotation software, we developed clinical interpretation reports for all variants observed using the CIViC smMIPs capture panel. In total, there were 1,340 variants observed across the 19 samples that passed smMIPs sequencing. Of the 1,340 variants observed, 127 had direct matches (chromosome, start, stop, reference, variant) with CIViC annotations (average = 6.7 variants/sample). An illustrative OpenCAP output report that displays most OpenCAP features, including CIViC Variant Descriptions and CIViC Assertions, was created using a previously-reported patient from the literature[118] (**Appendix 7, Figure S1**). For each identified clinical variant, links to external databases, CIViC Variant Descriptions, associated CIViC Assertions, and associated CIViC Evidence Items are provided. Associated Evidence items provide a brief description of the clinical relevance, links to CIViC Evidence Items (EIDs) and associated citations.

4.5 Discussion

The Open-sourced CIViC Annotation Pipeline (OpenCAP) (**https://opencap.org/**) is a resource for users to develop a custom capture panel that can be easily linked to actively maintained clinical relevance summaries. The methods described by OpenCAP to build a clinical capture panel offer several advantages relative to existing design paradigms. Use of an open-source database provides a systematic mechanism to survey existing literature within precision oncology to identify variants that are relevant for capture. Additionally, the public API permits rapid mapping of identified somatic and germline variants to CIViC clinical relevance summaries. Most importantly, the variants covered by CIViC and associated clinical summaries can be updated in real-time as knowledge is entered into the database to accommodate new information discovered within the field of precision oncology.

The smMIP capture method for sequencing provides inherent error correction capability, scalability to detect ultrasensitive variation, and cost effectiveness within a modular design. Combining the public access CIViC database with an ultrasensitive and versatile capture reagent provides an advantageous and principled method for building precision oncology capture reagents. This approach could enable a standardized framework for detecting and interpreting cancer-relevant genomic variation, lowering barriers to use of genomic analysis in the clinical practice of oncology. For maximal flexibility, OpenCAP describes methods for using both unique molecular identifiers (UMI) and non-UMI-based probes to capture variants of interest.

The CIViC smMIPs capture panel used Variant Evidence Scores and Sequence Ontology IDs to identify variants of interest for targeting. However, alternate filtering strategies have been outlined in OpenCAP documents. Regardless of variants targeted for capture, the presented research helped to show that CIViC variants and variant coordinates can be used for accurate capture panel design (95% detection accuracy with Pearson r^2 of 0.885 for VAFs). This finding helps to validate that the methods described in OpenCAP can be used to accurately interrogate desired variants of interest.

Like all targeted reagents, the preliminary CIViC smMIP design has limitations that can be addressed with future iterations. First, the reagent design is limited by the current knowledge within CIViC. Extensive curation from certain groups (e.g., the University Health Network curation of *VHL* variants) disproportionately increases representation for certain genes, cancers and variant types. Conversely, lack of curation in certain areas show a disproportionate decreased representation. To address existing curation disparities, CIVIC has joined the Variant

Interpretation for Cancer Consortium (VICC)[119] to integrate multiple variant interpretation knowledgebases into a single meta-knowledgebase. Successful execution of the aims outlined by the VICC would result in harmonization of information from CIViC, the Cancer Genome Interpreter[33], Clinical Knowledgebase[120], MolecularMatch, OncoKB[31], and Precision Medicine Knowledgebase[32], and others. This would allow users to leverage variant interpretations across multiple platforms for building custom capture panels that are linked to clinical relevance summaries.

In summary, the methods described here validate that community curated data on clinically relevant cancer variants can provide a systematic and dynamic method for capture reagent design. The curated coordinates in the database accurately map to desired variants, and probes designed using these coordinates show accurate recapitulation of the genomic landscape described by orthogonal sequencing. It is our hope that OpenCAP will provide the research community with a novel method to develop next-generation sequencing–based oncology panels.

4.6 Methods and experimental procedures

4.6.1 Development of operating procedure for OpenCAP

The Open-sourced CIViC Annotation Pipeline (OpenCAP) was built to guide users through the development of a custom capture panel linked to CIViC clinical relevance summaries (**www.opencap.org**). OpenCAP consists of five sections, each with examples and user tutorials. The first section describes CIViC (**www.civicdb.org**) and directs users through the CIViC web interface. The next section describes methods for building a custom capture panel, which includes identifying pertinent variants within the CIViC database and targeting those variants with probes using curated genomic coordinates. Subsequently, OpenCAP gives a high-level overview of the massively parallel sequencing pipeline, which includes brief summaries for sample procurement, nucleic acid generation, library preparation, and high-throughput sequencing. The final sections describe identifying variants from raw sequencing data and annotating those variants for clinical relevance.

4.6.2 Determining eligible CIViC variants for smMIP capture

Variants in CIViC were filtered using their Variant Evidence Score (required >20 points) and Sequence Ontology IDs (must be "DNA-based") (see **Appendix 5**). Variants were also filtered if: 1) all evidence supported only germline clinical relevance, 2) evidence was directly conflicting, or 3) a majority of evidence in a container variant (e.g., MUTATION) pointed to a hotspot that was already being covered. The remaining variants were eligible for the CIViC smMIPs capture panel.

4.6.3 Designing smMIPs for the CIViC capture reagents

Variants were further categorized by length. If the variant length was <250 base pairs, the variant was eligible for hotspot targeting. If the variant was >250 base pairs, the variant required either sparse or full tiling of the protein coding exons (see **Appendix 5**). For all variants, smMIPs were designed and synthesized as previously described[121] with the single alteration that the "-double_tile_strands_separately"[122] flag was used with the MIPgen tool to separately capture each strand of DNA surrounding the target.

4.6.4 Rescue and annotation of clinically relevant variants

Variants called using the CIViC smMIP capture panel were compared to variants called using original sequencing for samples that had matched tumor and normal sequencing. All genomic loci were manually reviewed[58] using both the smMIPs aligned BAM files and the original aligned BAM files. Variants only identified using smMIPs sequencing were grouped into four categories:

1) germline polymorphism, 2) pipeline artifact (low variant support or poor mapping), 3) variant support on smMIP sequencing but no support on original sequencing, or 4) variant support on both smMIP sequencing and original sequencing. For variants that showed support on smMIPs sequencing but no variant support on original sequencing, the binomial probability was used to assess if ≤3 variant-supporting reads would be detected with 95% confidence using the original coverage and the observed smMIPs VAF.

4.7 Acknowledgements

We gratefully acknowledge Tim Ley for sharing genomic data that made this project possible. We also would like to thank the patients and their families for their selfless contribution to the advancement of science. EKB was supported by the National Cancer Institute (T32GM007200 and U01CA209936). Select sample data was funded by the Genomics of AML PPG (T. Ley, PI, P01 CA101937). RU was funded by the National Comprehensive Cancer Network Oncology Research Program, general research support provided by Novartis Pharmaceutical Corporation (Novartis), the V Foundation for Cancer Research, and the National Institute of Dental and Craniofacial Research (NIH NIDCR R01DE024403). RG is funded by the National Cancer Institute (NIH NCI U01CA231844). MG is funded by the V Foundation for Cancer Research and the National Human Genome Research Institute (NIH NHGRI R00HG007940). SJS is funded by the National Cancer Institute (NIH NCI R33CA222344). OLG is funded by the V Foundation for Cancer Research, and the National Cancer Institute (NIH NCI U01CA209936, NIH NCI U24CA237719 and NIH NCI U01CA231844). This research was supported by a Cancer Moonshot funding opportunity, specifically, an Activities to Promote Technology Research Collaborations (APTRC) for Cancer

Research (Admin Supp) award to SJS and OLG under parent awards (NIH NCI R33CA222344 and NIH NCI U01CA209936).

4.8 Author Contributions

EKB wrote the manuscript. AW, KP, and SJS completed reagent development. EKB, KCC, KMC, and ALS, performed clinical data analysis. EKB, AMD, LMS, KCC, KMC, KK, DR, NCC, ZLS, MG and OLG contributed to CIViC database development and curation. EKB, AW, MCM, KMC, MG, SJS, and OLG edited the manuscript. AW, and KP developed and validated the smMIPs reagents. EKB, MCM, KCC, KMC, and ZLS generated figures. CCP, TAF, RU, and RG provided original sequencing data and tumor tissues. MG, SJS, and OLG supervised the project.

Chapter 5: Use of a clinical sequencing panel to influence treatment decisions for patients with AML

5.1 Preamble

The following chapter has been assembled as a manuscript for publication:

Barnell E.K., Skidmore Z.L, Krysiak K., Newcomer K.F., Anderson S.R., Wartman L.D., Oh S.T., Welch J.S., Stockerl-Goldstein K.E., Vij R., Cashen A.F., Pusic I., Westervelt P., Abboud C.N., Ghobadi A., Uy G.L., Schroeder M.A., Dipersio J.F., Spencer D., Duncavage E., Ley T.J., Griffith M., Jacoby M.A., Griffith O.L.. Use of a clinical sequencing panel to influence treatment decisions for patients with AML. Prepared for *Leukemia* on September 10th, 2019.

The MyeloSeq targeted capture panel is a laboratory developed test that is performed through the Pathology Department at the Washington University School of Medicine. This project evaluated how genomics impacted the development of treatment protocols for physicians. My role in this project was to collect data, analyze the results, and assemble the manuscript for publication. A complete list of author contributions is included within the manuscript (**Chapter 5.8**). The data presented here represents a work in progress and will be updated prior to final publication.

5.2 Summary

Targeted sequencing panels are being increasingly used within precision oncology by physicians to support clinical decisions. However, despite widespread use of sequencing for variant identification, the clinical use of such results is not well described. To elucidate the impact of genomic data on directing treatment decisions, we surveyed physicians who ordered a targeted panel (MyeloSeq®) for their patients. Specifically, we evaluated 346 MyeloSeq® reports that were generated for patients with hematologic malignancies. For the 122 cases with a definitive diagnosis of acute myeloid leukemia (AML), excluding cases with acute promyelocytic leukemia, a survey was sent to the treating physician to determine if the MyeloSeq® results altered the patient's

treatment plan. For the 114 cases in which physicians responded to the survey, 50 (44%) resulted in changes to treatment plan based on the MyeloSeq® results. Specifically, 38 new drugs were prescribed to target variants that were observed, 9 physicians altered consolidation therapy based on residual variant burden or high / low-risk variants, and 4 physicians used the MyeloSeq® panel to inform the patient's disease status. As observed here, the MyeloSeq® capture panel greatly influenced the physician decision and the ultimate treatment plan for the patient, which has direct implications in patient outcomes. Therefore, physicians should consider the use of targeted sequencing panels to inform treatment decision making in AML cases, although the benefit of testing has not yet been proven in prospective clinical trials.

5.3 Introduction

The integration of genomic information to individualize cancer treatment and improve patient outcomes is an area of interest within oncology.[123] Methods for obtaining genomic information includes in-house testing, which typically occurs at larger academic institutions, or outsourced assessment through commercial labs (e.g., Foundation Medicine, Guardant Health, etc). The genomic data can be obtained from single-gene testing, targeted sequencing panels, or comprehensive (whole-genome or whole-exome) sequencing approaches.

All sequencing data used for cancer assessment, regardless of source or size, requires variant identification and annotation. Many single-gene tests are FDA-approved and have associated companion diagnostics that designate a specific cancer subtype and indication. For example, LeukoStrat CDx® identifies *FLT3* variants to indicate response to Rydapt (midostaurin) or Xospata (gilterinib).[124] The vast majority of variant identification and annotation is performed by Laboratory Developed Tests (LDTs),[125] whereby results are provided to the ordering physician in the form of a clinical report. These reports typically contain highly complex data that requires a certain level of genomic literacy to understand.[126,127] Lack of report standardization, inadequate genomic training[42,128], limited infrastructure to support oncologists[123,129], and limited resources (e.g., low reimbursement rates and lack of access to clinical trials) have been shown to dramatically hinder the impact of sequencing data on clinical care.[130] The result is a discrepancy between the identification of clinically actionable variants[131] and implementation of change in treatment protocols.[123,132]

Use of sequencing data for the evaluation of acute myeloid leukemia (AML), is particularly complex due to its genetic heterogeneity. Over 250 recurrently mutated genes and specific hotspot variants have been described as clinically relevant within AML.[133] Additionally, the prognostic significance of an individual gene may be dependent on cooperative co-variants in multiple other genes,[134] thus currently available single gene variants tests (*FLT3*-ITD, NPM) may be insufficient to guide prognosis. Many pathogenic variants (e.g., within *FLT3*, *IDH2*, and *IDH1*) have been used as predictive markers for FDA-approved targeted therapeutics to supplement or augment treatment.[135–137] Other studies have used molecular markers, such as *TP53* variant status, to select the most appropriate chemotherapy for induction of remission.[71] Beyond the use of single molecular markers for evaluation of disease, multi-targeted approaches are being developed to further assess patient outcomes. Specifically, quantification of the variant allele frequency (VAF) of all tumor-associated variants has been used to detect measurable residual disease (MRD) and predict both relapse risk[68] and survival outcomes[138]. Other uses of multiple molecular markers for disease assessment includes evaluation of cytogenetic abnormalities to signify molecular remission and predict relapse-free survival.[139,140]

These new advancements in genomic understanding are slowly being integrated into standard treatment protocols for patients with AML. Current treatment recommendations for young (<60) and otherwise healthy AML patients include induction of remission via cytotoxic regimen of anthracycline and cytarabine.[141] Subsequently, patients receive consolidation therapy of either high-dose cytarabine (HiDAC) or upfront allogeneic stem cell transplant (SCT) for patients who achieve complete remission post-induction therapy. Recommendations for consolidation therapy are based on a variety of factors. Typically, patients with low relapse risk are treated with HiDAC, and patients with high relapse risk are more likely to be considered for SCT in first complete remission.[142–144] This is because SCT is associated with the lowest risk of relapse; however, it confers a considerable chance of morbidity, largely driven by graft-versus host disease, and treatment-related mortality.[142,143,145] Traditionally, stratification of relapse risk for directing consolidation includes assessment of pathological findings (i.e., blasts counts), cytogenetics, European LeukemiaNet (ELN) classification[146], NCCN-based risk assessment[147], and other demographic information.[148] As mentioned, genomic findings are now playing a larger role in governing selection of induction therapy, selecting targeted therapy to induce remission, and stratifying risk of relapse to dictate the method of consolidation therapy, and in some cases to justify deviation from the traditional treatment pathway.

Genomic data is undoubtedly impacting treatment decisions for cancer patients; however, the extent of influence is not well documented. This study attempts to quantify the impact of sequencing data on directing treatment protocols for patients with AML by observing a cohort of patient / physician pairs whereby the patient’s tumor is being assessed using a targeted clinical capture panel (MyeloSeq®). The MyeloSeq® capture panel evaluates 40 recurrently mutated genes or gene hotspots in AML and myelodysplastic syndrome (MDS). For each of 122 AML patients (excluding patients with promyelocytic leukemia) who received a MyeloSeq® panel assessment, we surveyed the treating physician on their use of MyeloSeq® results in making treatment decisions. Specifically, we queried for the use of targeted therapeutics based on observed variants, variant impact in directing consolidation therapy (stem cell transplant versus HiDAC consolidation), and other uses of the data that changed the original treatment plan. Using these surveys, we observed that 44% of all treatment protocols were altered based on the MyeloSeq® results. This includes 33 cases where one or more new targeted therapeutics was prescribed, 13 cases where consolidation therapy was altered based on high- or low-risk variants, and 4 cases where a definitive diagnosis was made based on observed variants. Additional non-indicated uses of the MyeloSeq® panel, including measurable residual disease testing and clonal / sub-clonal tracking, were also noted and analyzed in detail. Together these results indicate that a multi-targeted capture panel can be used to influence treatment plans for patients with AML.

5.4 Results

5.4.1 Identification of variants in myeloid malignancies by the MyeloSeq® Panel

Over an approximately 8-month period (August 17th, 2018 to April 9th, 2019) at the Washington University School of Medicine, 346 MyeloSeq® reports from 325 unique patients were generated with a median time to return of results of 15 days (range = 4 to 90 days). In total, there were 824 total variants observed in the 40 targeted genes across the 346 samples with a MyeloSeq® report (**Appendix 8, Figure S1**). The median number of failed genes across all reports was 2 genes (range = 0 to 24 genes). *WT1* (n = 211 cases), *CUX1* (n = 197 cases), and *CEBPA* (n = 111 cases) were genes that most frequently failed coverage requirements (**Appendix 9, S3**). Gene failure was attributable to specific gene regions that were recurrently difficult to target using existing reagents. The distribution of patient diagnoses for which a MyeloSeq® report was generated is shown in **Figure 5.1A-B**. The most common diagnoses were myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML), and the next most common diagnoses were essential thrombocytopenia (4.0%), clonal cytopenia of undetermined significance (3.7%), chronic myelomonocytic leukemia (3.5%), and primary myelofibrosis (3.5%). Across all samples, the most commonly mutated genes were *TET2* (n = 105 variants), *DNMT3A* (n = 72 variants), *TP53* (n = 68 variants), ASXL1 (n = 57 variants), *RUNX1* (n = 46 variants), *SRSF2* (n = 43 variants), and *FLT3* (n = 38 variants) (**Appendix 8, Figure S1**). There were 71 samples (20.5%) that had no variants detectable by MyeloSeq®, the majority of which (64.7%) were from patients with an ultimate diagnosis of MDS.

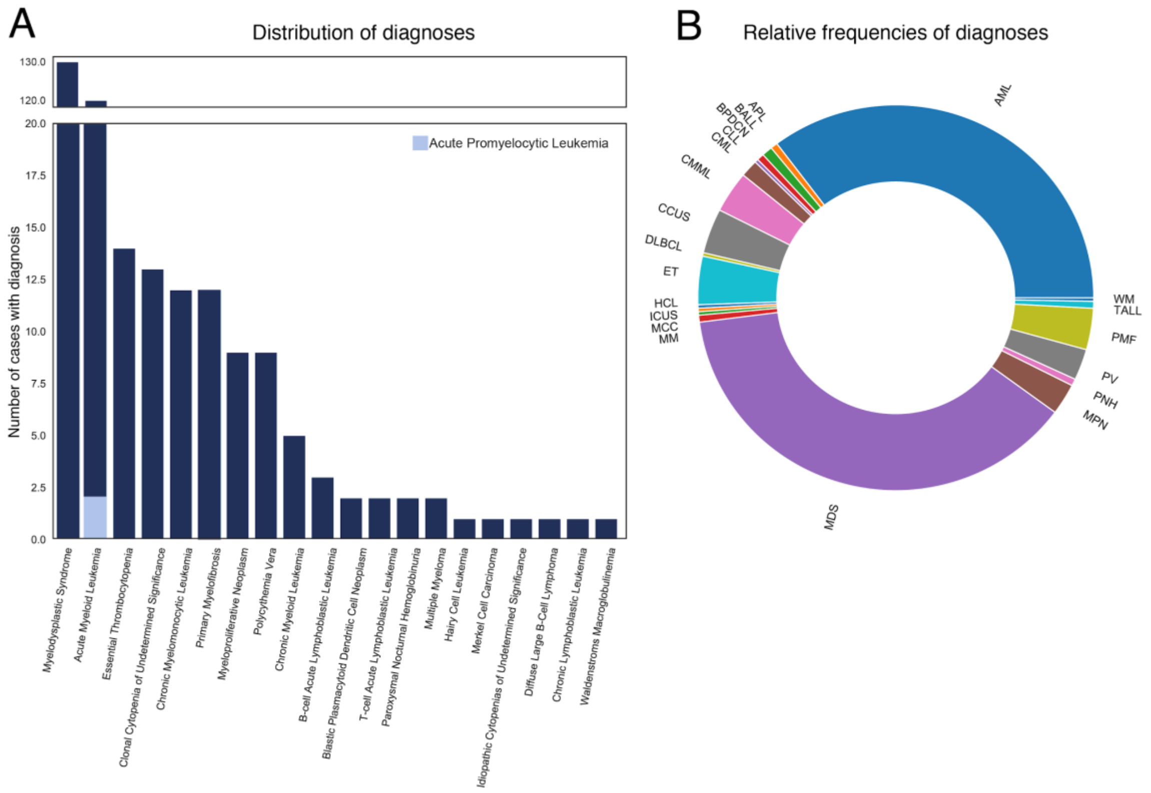


Figure 5.1 Distribution and relative frequency of patient diagnoses. ***(A)*** *Distribution and* ***(B)*** *relative frequency of hematologic disorders for the 346 samples evaluated with the MyeloSeq® panel. The diagnosis is based on the ultimate diagnosis made for the patient. AML with acute*

promyelocytic leukemia (APL) subtype (designated by a light blue bar under the AML distribution bar chart) were not eligible for analysis with a MyeloSeq® panel.

5.4.2 Patterns of MyeloSeq® panel usage in patients with acute myeloid leukemia

From all 346 MyeloSeq® reports generated, there were 124 samples from patients with a definitive diagnosis of AML, two of which were APL and were excluded from further study. Median duration for these MyeloSeq® reports to be generated was 15 days (range = 4 to 88) (**Table 5.1**). The non-APL AML samples (n = 122 samples) were derived from 109 unique patients who were under the care of 14 unique physicians in the Section of Leukemia and Bone Marrow Transplantation, Division of Oncology, Department of Medicine of Washington University. There were 11 patients for which 2 MyeloSeq® panels had been ordered for the same patient and 1 case for which 3 panels had been ordered for the same patient. The clinical characteristics of the 109 unique patients (**Table 5.1**) and the academic credentials of the treating physicians (**Table 5.1**) are also shown. Of the 109 unique patients, if multiple MyeloSeqs® were ordered, the earliest available test was used.

Table 5.1 Overview of demographics for patients diagnosed with acute myeloid leukemia

PATIENT CHARACTERISTICS (n = 109)	
Diagnosis of AML - Count, (%) *De-novo AML* *History of Myelodysplastic Syndrome (MDS)* *Therapy-related AML*	 *n = 59 (54.1%)* *n = 31 (28.4%)* *n = 19 (17.4%)*
Age Median - Median, (Range)	65 years (23 to 90 years)
Gender *Male* *Female*	 *n = 58 (53.2%)* *n = 51 (46.8%)*
REPORT CHARACTERISTICS (n = 122)	
Time for Report Generation - Median (Range)	15 days (4 to 88 days)

The gene variant frequencies for all AML samples in this cohort reflected the expected genomic landscape of AML (**Figure 5.2**).[133,134] Of the 122 samples, 33 showed *DNMT3A* variants, 31 had *FLT3* variants, 24 had *TET2* variants, and 22 showed *TP53* variants. In total, 37/40 genes interrogated by the MyeloSeq® panel were observed in at least one AML case. Only 13 samples showed no variants on MyeloSeq®, 8 of which were acquired from patients who were determined to be in clinical remission, confirmed by bone marrow biopsy.

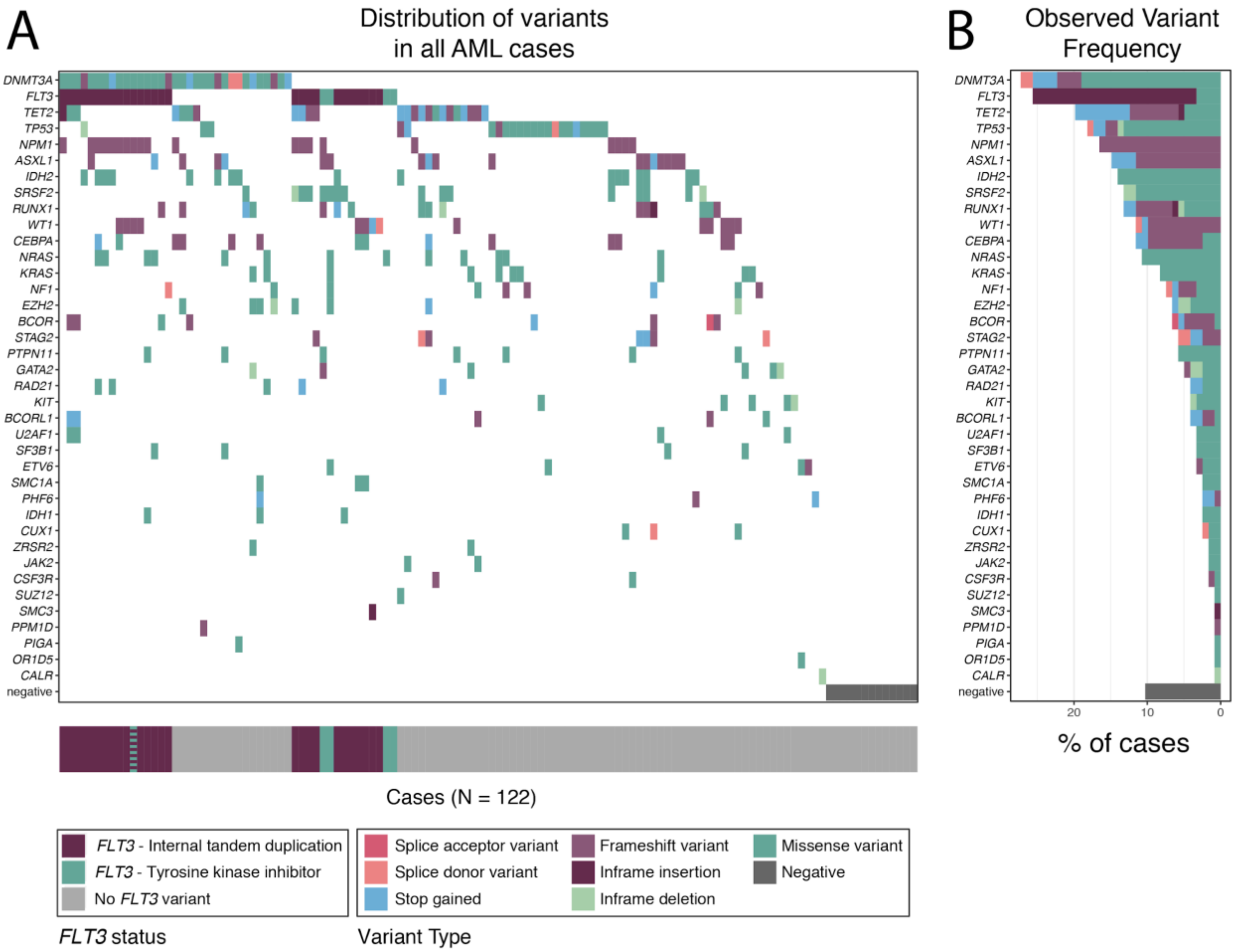


Figure 5.2 Distribution of variants in AML cases. ***(A)*** *The first panel displays a heatmap of the distribution of variants in all the AML cases reported. Each row represents a single gene, and each column represents a MyeloSeq case report (n = 122). Each colored square represents if a variant was observed in the designated gene. Colors indicate the variant type. If there was more*

than one variant observed per case, the most deleterious variant, based on the variant effect prediction (VEP)34 was listed. The bottom bar indicates the FLT3 status for all 122 patients. The bar color indicates the type of FLT3 variant (ITD vs. TKI). ***(B)*** *The second panel displays the percent of all cases whereby a variant was observed on the gene. Each row represents a single gene, and the color indicates the variant type. Again, if more than one variant was observed per case, the most deleterious variant was listed.*

MyeloSeq® panels were ordered at different timepoints during the disease course (**Figure 5.3**). For the first MyeloSeq® ordered, the most common time point was at diagnosis of AML (n = 65; 54%). The next most common time point was during induction therapy (n = 19; 15.7%). 7 were ordered during 7 + 3 induction, 10 were ordered during treatment with Decitabine, and 2 were ordered during treatment with Vyxeos. Additionally, 11 MyeloSeq® panels were ordered during or after consolidation chemotherapy and 4 were ordered after a stem cell transplant. Typically, if a second or third MyeloSeq® panel was ordered, it was ordered during consolidation / salvage therapy (n = 10 / 12; 83.3%). Of note, all patients (n = 9) that had non-traditional induction therapy (i.e., *Clinical Trial* or *Other Induction Therapy*) had a MyeloSeq® test performed at diagnosis of disease.

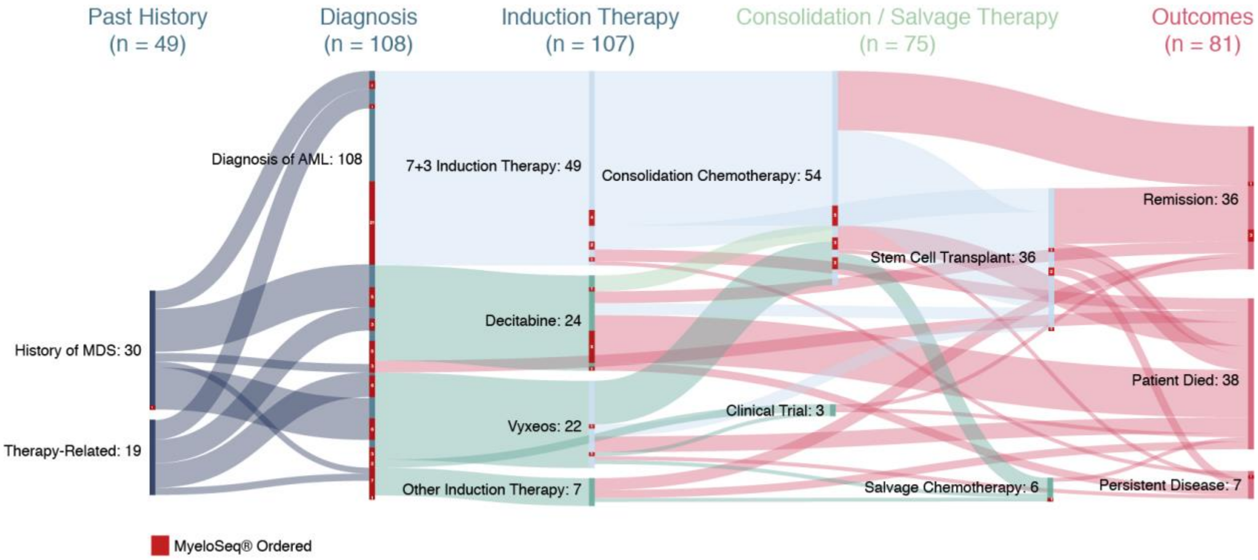


Figure 5.3 Alluvial plot demonstrating the disease course for all 108 patients evaluated with the MyeloSeq panel. *The disease course was split into 5 phases. Past history indicates if the patient had a history of myelodysplastic syndrome (MDS) or therapy-related AML. Diagnosis of AML indicates the 108 patients included in the study. Each patient (except for one) had an induction therapy (e.g., 7 + 3, decitabine, vyxeos, or other), and most patients (75 / 108) had*

consolidation or salvage therapy. These consolidation / salvage therapies included: 1) consolidation chemotherapy (i.e., HiDAC, Ventoclax, Decitabine), 2) stem cell transplant (Donor lymphocyte infusion, Allogeneic SCT, matched unrelated donor SCT, Cytokine-induced Memory-like NK Cell infusion, etc.), 3) enrollment into a clinical trial, or 4) salvage chemotherapy (e.g., CLAM, MEC, Targeted Therapeutic, DART). Outcomes for each patient included remission, persistent disease, or death. Lack of one-to-one mapping between sections is due to incomplete disease course. The red boxes within each section indicate if a MyeloSeq® was ordered for a patient. The value within the box indicates the number of patients that had a MyeloSeq® ordered at a specific time point given a specific disease course (e.g., 21 patients had a MyeloSeq® ordered at diagnosis with disease course of de-novo AML and 7+3 induction therapy). The majority of MyeloSeqs® were ordered at diagnosis and most patients that strayed from traditional therapy (e.g., other induction therapy or clinical trial recruitment) had a MyeloSeq ordered prior to enrollment.

5.4.3 Changes in treatment regimens in response to single variants observed on MyeloSeq®

Physicians provided survey responses for 120 of the 122 AML samples, which represents a 98.3% response rate. Of the 120 cases with survey responses, 6 were ineligible for further analysis. Specifically, 2 patients were lost to follow-up, 3 patients declined treatment, and 1 patient pursued care at an outside hospital. Of the remaining 114 cases that were eligible, physicians indicated that they changed their therapeutic plan based on the MyeloSeq® results for 43.8% of all cases (n = 50) (**Figure 5.4A**). In 33 of these 50 cases, physicians indicated that they recommended therapy based on the mutated gene identified by MyeloSeq® panel results (**Figure 5.4B**). A hypomethylating agent was started for 14 cases based on a *TP53*[71] or *TET2*[149,150] variant, kinase inhibitors (midostaurin or gilteritinib) were used in 12 patients with a *FLT3* variant[151] and 1 patient with a *KIT* variant[152], and *IDH1 / IDH2* inhibitors were used in 8 cases[153]. Of these six genes (TP53, TET2, FLT3, KIT, IDH1, IDH2) that resulted in a therapeutic intervention, only 3 have FDA-approved companion diagnostics.[154] There were 3 additional patients enrolled into clinical trials based on observed MyeloSeq® variants. Two patients had *ASXL1* variants, which supported enrollment into a cytokine induced memory-like (CIML) natural killer (NK) cells study, and one patient was enrolled in a clinical trial that uses residual disease to guide consolidation therapy based on observed MyeloSeq® variants.

The MyeloSeq® report was also used for prognostic indications. Specifically, 2 patients were offered a transplant based on *TP53* variant status, 1 patient went to transplant based on *DNMT3A* variant status, 1 patient was offered stem cell transplant based on observation of measurable residual disease (MRD+), and 1 patient received a Donor Lymphocyte Infusion (DLI) based on clonal tracking. Additionally, 5 cases had low-risk profiles determined by either bi-allelic *CEBPA* variants, *KIT* variants, or other methods (e.g., no high-risk variants observed). For these patients, chemotherapy was recommended for consolidation (**Figure 5.4B**).

Four physicians used the MyeloSeq® reports to inform diagnosis of disease. One physician used the report to confirm AML from a previous diagnosis of MDS, two physicians confirmed relapse in patients that had negative bone marrow biopsies, and one physician used the MyeloSeq® results to confirm remission (**Figure 5.4B**).

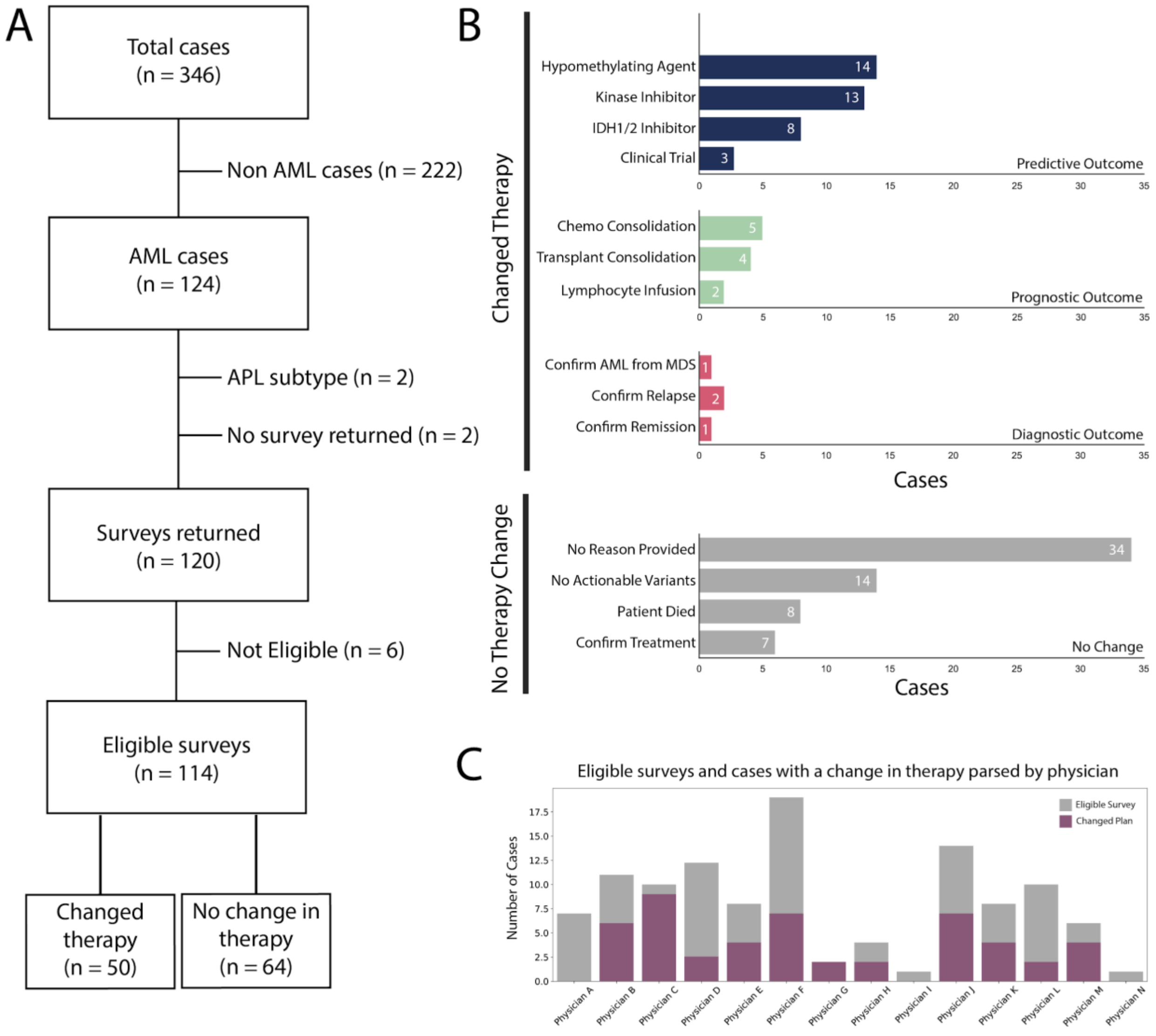


Figure 5.4 Impact of MyeloSeq® report in changing treatment protocols for patients. ***(A)*** *In total, 346 cases were enrolled as part of the study whereby 121 had a definitive diagnosis of AML. Of those 121 cases, 119 had a reported survey. Six cases were ineligible for analysis (e.g., Lost to Followup, Refused Treatment). Of all 113 eligible cases, 50 cases (44%) have a documented change in therapy based on the MyeloSeq® report.* ***(B)*** *Among all 50 cases whereby therapy was altered, it was reported that 39 new therapies were introduced, 13 MyeloSeq® reports provided prognostic information, and 4 MyeloSeq® reports were used to confirm a diagnosis. For physicians that did not use the MyeloSeq® results to inform decision making, 14 stated that there were no actionable variants, 8 physicians mentioned that the patient died before results were returned, 6 mentioned that the report did not alter their predetermined treatment path (i.e., confirmed current treatment protocol), and 34 did not provide a reason for their response.* ***(C)*** *There were 14 physicians who contributed at least one survey to this study. In general, there were no major outliers with regards to the number of eligible surveys and number of cases whereby the physician changed the therapeutic plan.*

5.4.4 Cases with no change in treatment regimens in response to MyeloSeq® reports

There were 64 cases where the physician noted that he or she did not change their therapeutic plan based on the MyeloSeq® report. The majority of cases (n = 34) did not state a reason for why the results were not used. Of the reasons cited for lack of use, the most common was that there were no actionable variants observed by MyeloSeq® (n = 14). 8 physicians mentioned that the report took too long, and that the patient expired prior to receiving results. Additionally, 7 physicians mentioned that the MyeloSeq® results did not provide additional information and merely confirmed what was already known about the patient. For example, the presence of high-risk variants detected by cytogenetics would have led the physician to recommend hematopoietic stem cell transplant regardless of the MyeloSeq® results (**Figure 5.4B**).

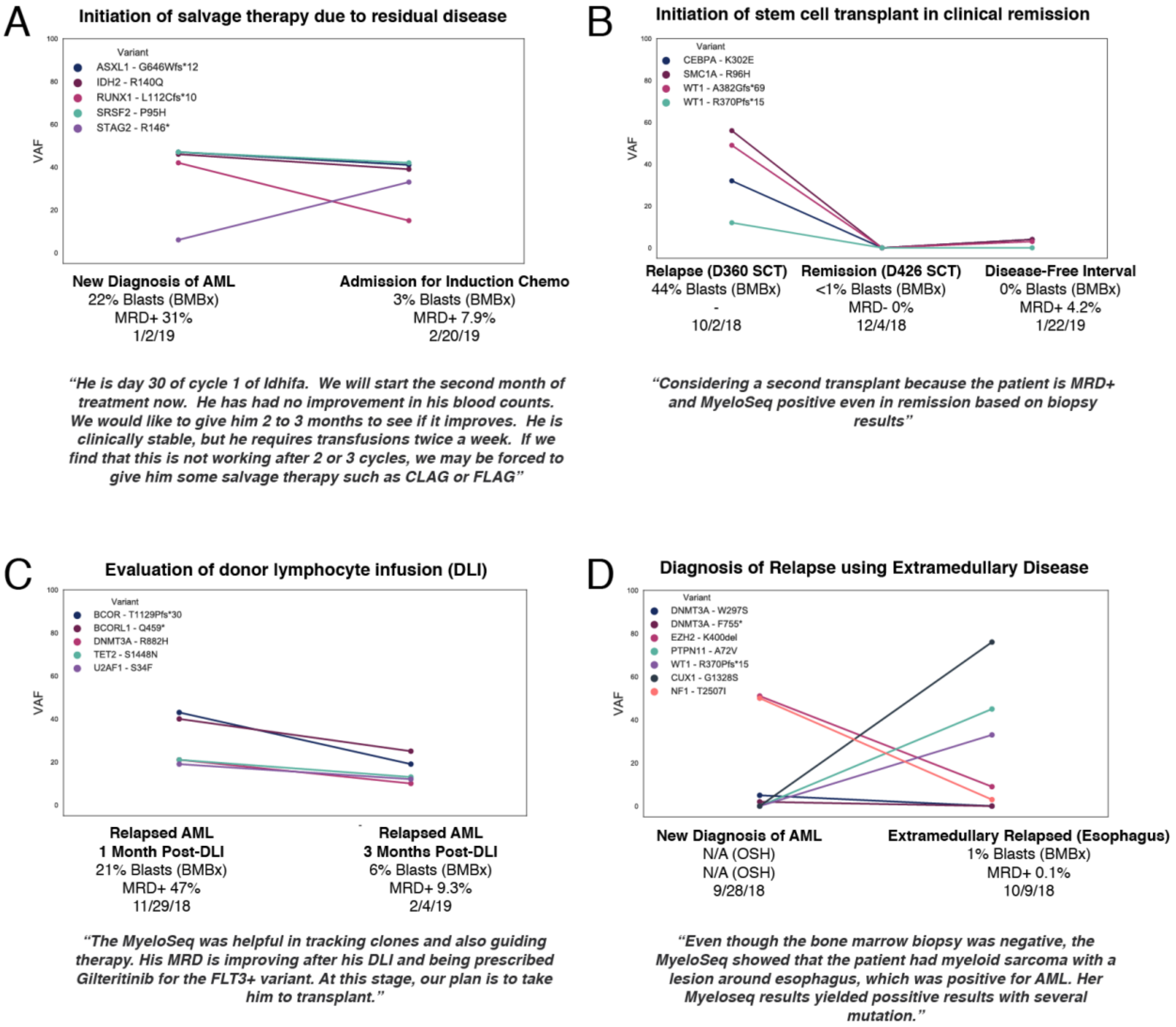
There were 14 physicians who provided at least one eligible survey for this study (**Appendix 9, Table S4**). The total number of eligible surveys completed by each physician and the total number of cases where the physician changed his or her plan is shown in **Figure 5.4C**. In general, there was high participation from all physicians with no noticeable differences between physician behaviors.

5.4.5 Multi-target and obscure uses of the MyeloSeq® capture panel results

In addition to the clear changes in therapy as described above, the interaction between variants observed on MyeloSeq® were also used to direct care. For example, in one case, the physician mentioned that the combination of *DNMT3, NPM1,* and *FLT3* ITD variants observed on Myloseq® implied that the patient had aggressive disease, which required immediate transplant. For another patient, a *GATA2* variant was observed at a 52% VAF, which required a subsequent skin biopsy to confirm lack of germline predisposition for disease. The physician noted that even though he did not change his treatment plan based on the MyeloSeq® report, there would have been treatment implication if there had been a germline polymorphism present. Finally, an additional physician noted that the results from 2 MyeloSeq® tests performed on the same patient, before and after consolidation therapy, showed that all variants cleared, except for a DNMT3A mutation. This variant had an original VAF of 42% and a persistent VAF of 43% after consolidation. MRD testing on her bone marrow at the same time was negative. The clearance of all previous variants and her cytogenetic abnormality, with the exception of the *DNMT3A* R882H variant with a VAF of 42%, suggested that the variant is evidence of age-related clonal hematopoiesis with an increased risk of relapse over time. The combination of data informed the plan to not transplant, despite the persistence of a single *DNMT3A* variant. This patient was closely monitored for relapse due to the noted increased risk.

5.4.6 Evaluation of patients who completed multiple MyeloSeq® panels at varied timepoints

There were 12 patients who had multiple MyeloSeq® panels ordered at different times during their disease course (**Figure 5.5, Appendix 8, Figure S2**). In many of these cases, it was observed that physicians used the MyeloSeq® panels for non-indicated use. For example, in **Figure 5.5A**, the physician ordered a MyeloSeq® panel at diagnosis and prior to induction chemotherapy. At this second time point, the physician stated that the patient had been on IDHIFA (enasidenib) to target the *IDH2* variant observed at diagnosis, however the second MyeloSeq® report showed residual disease with an *IDH2* variant allele frequency (VAF) of 39%. Given a lack of response to the targeted therapy, it was mentioned that a new approach might be used for future salvage. **Figure 5.5B** demonstrated a case where the physician obtained 3 MyeloSeq® panels for the patient. The first MyeloSeq® demonstrated relapse post-transplant with 4 detected variants. Subsequently, a MyeloSeq® test was ordered during remission after salvage therapy where there was no measurable residual disease. Two months later, during a disease-free interval, the physician ordered a third MyeloSeq® test, which showed recurrence of disease despite no excess blasts on the bone marrow biopsy (BMBx). The physician noted the requirement for a second stem cell transplant based on the MRD and MyeloSeq® results. **Figure 5.5C** showed how a physician used multiple MyeloSeqs® to follow the patient's response to a donor lymphocyte infusion (DLI). After 3 months, the MyeloSeq® report showed that VAFs for all observed variants were reduced by 7-23%, which indicated response to treatment. In this case, the physician noted that further reduction in the tumor burden would indicate success of the stem cell transplant. Finally, **Figure 5.5D** showed how the physician used the MyeloSeq® reports to demonstrate extramedullary relapse in a patient with negative BMBx findings. Specifically, the initial BMBx performed at an outside hospital was evaluated using the MyeloSeq® panel. Subsequently, the patient developed peri-esophageal lesions, which were biopsied and also evaluated using the MyeloSeq® panel. It was observed that the extramedullary disease showed presence of some original variants (*NF1* and *EZH2*), loss of other variants (*DNMT3A*) and three novel variants (*CUX1*, *PTPN11*, and *WT1*). These data confirmed relapse and demonstrated novel variants that were not originally present in the primary tumor, some of which have clinical implications.[155,156] The complete set of cases with multiple MyeloSeq® reports are provided in **Appendix 8, Figure S2**.



"He is day 30 of cycle 1 of Idhifa. We will start the second month of treatment now. He has had no improvement in his blood counts. We would like to give him 2 to 3 months to see if it improves. He is clinically stable, but he requires transfusions twice a week. If we find that this is not working after 2 or 3 cycles, we may be forced to give him some salvage therapy such as CLAG or FLAG"

"Considering a second transplant because the patient is MRD+ and MyeloSeq positive even in remission based on biopsy results"

"The MyeloSeq was helpful in tracking clones and also guiding therapy. His MRD is improving after his DLI and being prescribed Gilteritinib for the FLT3+ variant. At this stage, our plan is to take him to transplant."

"Even though the bone marrow biopsy was negative, the MyeloSeq showed that the patient had myeloid sarcoma with a lesion around esophagus, which was positive for AML. Her Myeloseq results yielded possitive results with several mutation."

Figure 5.5 Use of multiple MyeloSeq® reports to alter treatment plans for patients with AML. *Each panel represents a single patient where multiple MyeloSeq® panels were ordered. The plot indicates the variants observed with associated variant allele frequencies (VAFs). Each time point is labeled with associated bone marrow biopsy (BMBx) results and measurable residual disease (MRD) results, if available. Below each graph is a direct quote from physicians who ordered the report.* ***(A)*** *This example showed how the MyeloSeq® panel assessed for residual disease, to indicate require treatment with salvage therapy.* ***(B)*** *The second example demonstrated how MyeloSeq® results influenced the decision to initiate a second stem cell transplant (SCT) despite the patient being in clinical remission based on bone marrow biopsy results.* ***(C)*** *The third example showed how the MyeloSeq® panel was used to track efficacy of a donor lymphocyte infusion (DLI).* ***(D)*** *The final example demonstrates how the MyeloSeq® panel was used to diagnose extramedullary recurrence given a negative bone marrow biopsy.*

5.5 Discussion

This study reviews the utility of the MyeloSeq® clinical capture panel to inform treatment decisions for patients with AML. We showed that physicians changed their treatment plan in 44% of all eligible cases based on the MyeloSeq® results. Specifically, the genomic data influenced the prescription of targeted therapeutics, altered the consolidation therapy recommendations (HiDAC vs. stem cell transplant vs. other), and provided a definitive diagnosis of disease. Of the physicians that did not use the MyeloSeq® panel to change a specific treatment protocol (n = 64 cases), 21 stated that the panel was still informative in some manner (e.g., reaffirmed existing decision or no targetable variants were identified). This indicates that in this study, the genomic data was informing treatment decisions at a much higher rate than previously cited studies.[157]

It was observed that physicians used the MyeloSeq® panel in ways that were out of scope of the intention and indications claimed by the LDT. For example, there were 12 cases (10% of all AML cases) whereby the physician ordered at least two MyeloSeq® panels for the same patient at different times during the disease course. In these cases, physicians were typically monitoring measurable residual disease (MRD) and evaluating clonal / sub-clonal populations. MRD positivity and clonal evolution have known implications in patient outcomes and are important for optimizing patient care.[68,158] Additionally, although 65 MyeloSeq® reports were generated at initial diagnosis (54%), more than 40% of all reports were requested at relapse or during remission.

The results from this study also demonstrate that a multi-target assay that evaluated many recurrently mutated variants is preferred to single-variant diagnostics. Specifically, it was observed that the broad usage of the MyeloSeq® test supports the use of a multi-target approach for assessing disease. Specifically, only 21 cases (18%) revealed a variant that had an associated FDA-approved companion diagnostic (LeukoStrat CDx for *FLT3*, Abbott RealTime for *IDH1*, and Abbott RealTime for *IDH2*). However, in the majority of cases where therapy was changed, the physician either utilized a variant that does not have an FDA-approved diagnostic, or the physician assessed multiple variants to inform decisions.

The use of a tumor-only panel was observed to be a limitation of this approach. Specifically, lack of a germline / normal reference prevents variants from being definitively called as putative somatic and many variants being observed are potentially single nucleotide polymorphisms or *de-novo* germline variants. In one case, a possible germline variant (*GATA2*) was identified leading to further genetic testing for the patient to confirm somatic versus germline variant status. Introduction of a matched-normal sample could improve the variant calling pipeline and overall annotation of variants in the analysis.

The results from this study demonstrate that genomic data from targeted capture panels are providing physicians with information that is directly impacting patient care. On the surface it is clear that capture panels can identify variants that have associated targeted therapeutics or that provide prognostic indications, however, here we show that genomic data is being used more extensively by physicians. Multiple-analyte diagnostics like the MyeloSeq® panel have capabilities to monitor for residual disease or show expansion of tumor subclones over time. They can be used to improve sensitivity of remission status and indicate relapse prior to the onset of symptoms. Given the rapid turn-around time (~2 weeks) and the accuracy of the UMI-based sequencing approach, the MyeloSeq® capture panels, and other similar types of capture panels, have the potential to improve upon traditional assays (e.g., bone marrow biopsy or PCR-based diagnostics) for certain indications. Based on these initial data, it is clear that the measurement of genetically defined variants could provide an optimal and more accurate method to classify patients according to their risk of recurrence. This study provides justification for use of the MyeloSeq® LDT, or a similar UMI-based capture panel at all points in the disease course for patients with AML.

5.6 Methods and experimental procedures

5.6.1 Study design and patient eligibility

This study was conducted at the Washington University School of Medicine after approval by the institutional review board. Patients were eligible for assessment if a MyeloSeq® gene panel was ordered by a provider at any point in their treatment between August 17th, 2018 and April 9th, 2019. The MyeloSeq® panel is a targeted sequencing assay that evaluates 40 genes and gene hotspots that are recurrently mutated in myeloid malignancies (www.meyloseq.com; **Appendix 9, Table S1)**. For each patient found to have a definitive diagnosis of acute myeloid leukemia (AML), excluding those with acute promyelocytic leukemia (APL) subtype, a survey was sent to the treating physician. The physician was instructed to complete a 16-question survey, which queried how the physician used the MyeloSeq® panel, if at all, to inform treatment decisions (**Appendix 9, Table S2**). In instances where the provider indicated a change in therapy but did not indicate the reason for change, the rationale was identified and confirmed via independent chart review of the patient.

5.6.2 MyeloSeq® processing

The panel utilizes an amplicon capture-based enrichment with unique molecular identifier (UMI) for ultra-high variant sensitivity that targets an approximately 98,000 base-pair space. The MyeloSeq® panel is recommended for the following conditions: Acute Myeloid Leukemia (AML), Myelodysplastic Syndrome (MDS), Cytopenia, Clonal Cytopenia of Undetermined

Significance (CCUS), Clonal Hematopoiesis of Indeterminate Significance (CHIP), Myeloproliferative neoplasm (MPN), and Myeloid Disorder. Specimens were obtained from bone marrow aspirate, peripheral blood, or DNA extracted from fresh tissue. All steps of sample processing were performed in College of American Pathologists (CAP)-accredited, Clinical Laboratory Improvement Amendments (CLIA)-certified clinical diagnostic laboratories.

Target enrichment for the MyeloSeq® assay used a commercially available, targeted, next-generation sequencing approach (HaloplexHS, Agilent Technologies). Preparation consisted of 1) enzymatic fragmentation; 2) strand-specific ligation of sequencing primers, sample indexes, 10-bp degenerate molecular barcodes, and a biotin tag to single DNA molecules; 3) rapid liquid-phase enrichment of target loci using paramagnetic streptavidin-coated beads; and 4) on-bead PCR amplification. Sequencing was performed using the Illumina MiniSeq sequencing platform. Analysis of the sequence data was performed by aggregating individual reads with identical UMIs into consensus reads using customized GATK[16] software. The minimum read family size was 5 reads and consensus bases had to be in at least 2/3 of the reads in the read family.

Variant calling for the MyeloSeq® assay was performed using a computational pipeline that employs custom-built tools created at Washington University. Sequencer generated FASTQ files are first demultiplexed and aligned to the reference genome (GRCh37.2). Overall reads must exceed 1,000,000 total reads with >98% aligned. During this process, UMI sequences are added to the BAM file. Once aligned, a custom Java-based tool collapsed reads into read families and collapsed BAMs were used with standard variant calling tools, including Varscan2[19], Platypus[159], and Pindel[61]. Quality check (QC) required mean unique coverage to exceed 500 reads. Genes passed QC thresholds if >90% of positions within the targeted region had >50x coverage. These programs detected single nucleotide substitutions (SNVs), insertions or deletions (indels) up to

10bp, and *FLT3* internal tandem duplication (ITD) insertions between 21 and 108 bp. Initial variant annotation was performed using the Variant Effect Predictor tool (VEP).[160] Annotation was augmented by using a custom VCF file with variants identified in the AML The Cancer Genome Atlas data set as well as variants observed in 3 published sequencing reports that evaluated patients with MDS.[161–163] The final output from the VEP annotation is an annotated VCF file and a text file with variant information. Variant identification was subject to the following thresholds and cutoffs: 1) variants must be nonsynonymous, 2) 2% minimum VAF and 5 variant reads with support on each strand (*FLT3* ITD alleles require 1 read on each strand), 3) amplicons must have at least 5 reads assigned to it during consensus bam formation, and 4) 0.1% maximum population allele frequency (MAX_AF across all populations) for reporting as a potential somatic variants (1000 genomes, ExAC, gnomAD databases) OR presence in a custom MDS/AML variant database.

5.6.3 MyeloSeq® annotation and physician survey

MyeloSeq® reports were generated using annotated Variant Call Format (VCF) files, Binary Alignment Map (BAM) files, preliminary clinical annotation, and quality metrics. Briefly, tier 1 (variants with strong clinical significance) and tier 2 variants (variants with potential clinical significance),[126] filtered variants (i.e., variant allele frequency <2%), and variants of unknown significance were manually reviewed for clinical relevance. These variants were used to generate a clinical assertion that summarizes all relevant findings. Information in the assertions included interpretation from National Comprehensive Cancer Network (NCCN) guidelines, WHO recommendations, and outcome data from high-quality journals, as it relates to disease prognosis and therapeutic sensitivity / response. Reports were reviewed by faculty members from the Washington University Department of Pathology & Immunology and signed-out within the secure network. The final report was integrated into the patient’s electronic medical record (EMR) for review by the treating physician.

5.7 Acknowledgements

We would like to thank the patients and their families for their selfless contribution to the advancement of science. Additional regards are also given to the Siteman Cancer Center and the Barnes-Jewish Hospital Foundation. EKB was supported by the National Cancer Institute (U01CA209936). MG is funded by the V Foundation for Cancer Research and the National Human Genome Research Institute (NIH NHGRI R00HG007940). OLG is funded by the National Cancer Institute (NIH NCI U01CA209936 and NIH NCI U24CA237719).

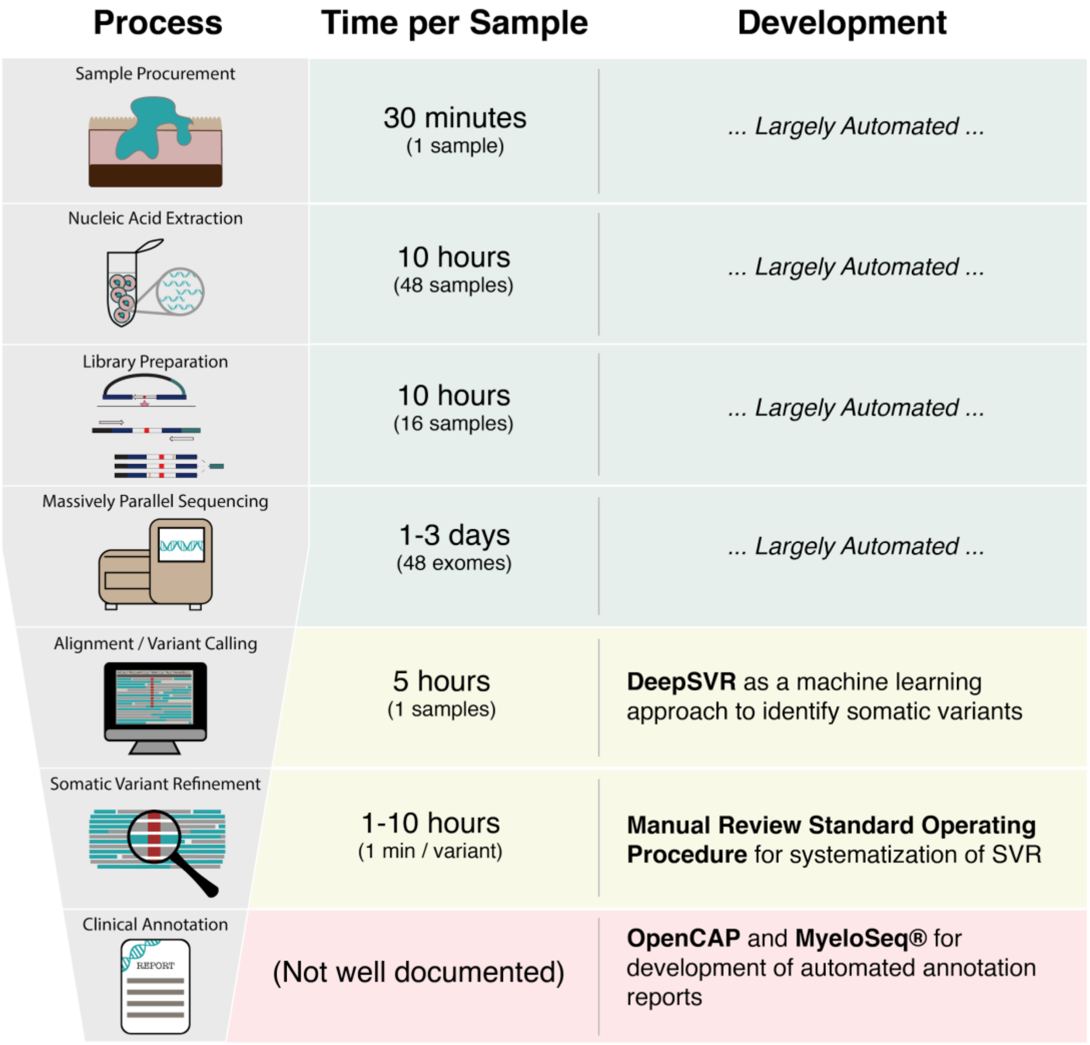
5.8 Author Contributions

EKB wrote the paper; ZLS, KK, KFN, SRA, TJL, MG, MAJ, OLG edited the paper; EKB, ZLS, SRA wrote code; EKB, ZLS, KFN, SRA analyzed the data; DS, ED developed and validated the MyeloSeq® capture panel; LDW, STO, JSW, KESG, RV, AFC, WP, PW, CNA, AG, GLU, MAS, JFD, MAJ provided surveys on the use of MyeloSeq® panels; EKB, ZLS created figures; DS, ED Performed lab analysis for MyeloSeq® processing; MAJ, OLG Supervised the project.

Chapter 6: Conclusion

6.1 Preamble

The preliminary steps in massively parallel sequencing (MPS) workflows (i.e., sample procurement, nucleic acid extraction, library preparation, and sequencing), have now been highly automated with platforms that permit high-throughput evaluation. However, the subsequent steps of alignment / variant calling, somatic variant refinement, and clinical annotation, are not standardized between and across laboratories and they require extensive manual labor to complete. Collectively, these steps are referred to as the annotation or analysis bottleneck.[25,45] The focus of this research was to develop bioinformatic tools that alleviate the analysis bottleneck within precision oncology (**Figure 6.1**). First, it was observed that automated somatic variant calling of aligned sequencing reads is deeply flawed and results in many false positives that are attributable to sequencing artifacts. As a response, DeepSVR, which is a deep learning somatic variant refinement algorithm, was built to eliminate false positives associated with automated somatic variant calling. Second, it was recognized that manual review of potential somatic variants is required within the MPS pipeline. However, methods for manual review are underreported and not standardized, which results in lack of reproducibility and highly variable manual review strategies between institutions. Therefore, a Standard Operating Procedure was developed and validated to optimize manual review of somatic variants. It was also observed that variant annotation and report generation was a highly manual process with high variability between and across institutions. To address existing issues with variant annotation, the Open-sourced Clinical Annotation Pipeline (OpenCAP) was developed to generate clinically relevant capture panels linked to automated clinical reports for physician use. Finally, many of these variant calling and reporting strategies described above were tested using the MyeloSeq® capture panel. This study demonstrated the utility of genomic data in driving change within treatment protocols to improve patient care. Collectively, we validated that these resources would reduce the manual labor required to execute the precision medicine pipeline to hopefully improve the ability for physicians to deliver optimal care to their patients.



Process	Time per Sample	Development
Sample Procurement	30 minutes (1 sample)	... Largely Automated ...
Nucleic Acid Extraction	10 hours (48 samples)	... Largely Automated ...
Library Preparation	10 hours (16 samples)	... Largely Automated ...
Massively Parallel Sequencing	1-3 days (48 exomes)	... Largely Automated ...
Alignment / Variant Calling	5 hours (1 samples)	**DeepSVR** as a machine learning approach to identify somatic variants
Somatic Variant Refinement	1-10 hours (1 min / variant)	**Manual Review Standard Operating Procedure** for systematization of SVR
Clinical Annotation	(Not well documented)	**OpenCAP** and **MyeloSeq®** for development of automated annotation reports

Figure 6.1 Tools to address the annotation bottleneck within precision oncology. *The precision oncology pipeline is composed of at least seven discrete steps. The first four steps (sample procurement, nucleic acid extraction, library preparation, and massively parallel sequencing) are all optimized and high-throughput processes. The three following steps sequencing (alignment / automated variant calling, somatic variant refinement, and clinical annotation) still require an immense amount of manual labor to properly execute. Combined, these steps are referred to as the annotation bottleneck. The research outlined here describe three tools that alleviate the annotation bottleneck. DeepSVR is a machine learning approach to improve*

automated somatic variant calling; the manual review standard operating procedure standardizes the manual review process required for somatic variant refinement; and the Open-sourced Clinical Annotation Pipeline (OpenCAP) develops automated clinical reports for physicians using the CIViC database. Execution of genomic report generation was validated using the MyeloSeq® clinical capture panel. These resources reduce the manual labor required to execute the precision medicine pipeline.

6.2 Summary and future direction for presented bioinformatic tools

6.2.1 DeepSVR is a machine learning approach that successfully recapitulate manual review of aligned sequencing reads

Automated somatic variant callers have known issues with identifying true somatic variants. Specifically, many variants called as somatic by traditional software are subsequently determined to be false positives by manual review or orthogonal sequencing. This occurs because traditional variant callers use simplistic algorithms that consider a minimal number of features for calling variants as somatic. To improve upon existing automated somatic variant calling software, we developed DeepSVR, which is a machine learning approach that recapitulates manual review labels. The algorithm was built using 41,000 manually reviewed variant calls derived from 405 tumors across 9 tumor subtypes. The model demonstrated high accuracy when performing internal cross-validation (ROC AUC = 0.96) and when evaluating a hold-out test set (ROC AUC = 0.96). The model also demonstrated high performance when comparing model labels to orthogonal sequencing data (n = 212,158; ROC AUC = 0.95). This model was packaged as a python application and made public via BioConda. Subsequently, a usage tutorial was made available on GitHub. The software could be integrated into the variant calling pipeline to identify the majority of false positives called by automated variant callers. This would dramatically reduce the labor required to define and annotate variants associated with an individual's tumor and would improve the accuracy of clinical reports generated from sequencing data.

6.2.2 Manual review standard operating procedure (SOP) improves accuracy and reproducibility of somatic variant refinement

After automated somatic variant calling, manual review is required for defining a putative list of true somatic variants associated with a patient's tumor. Although manual review is a very important step in the MPS pipeline, guidelines for proper execution are not well described. Therefore, we developed a manual review SOP to provide guidance on the manual review of aligned sequencing reads to determine if a variant called by automated somatic variant callers is a true somatic variant or a false positive. This SOP first describes data visualization setup using IGV and IGVNav. Subsequently, the SOP describes methods for review of individual variants. This includes making a variant call (somatic, ambiguous, fail, or germline) and, if needed, annotating variants using tags (e.g., low mapping, high discrepancy region, etc.) or notes (e.g., dinucleotides, SNP, etc.) to describe observations made during manual review. In this SOP, we also provide Supplemental Materials, which contain examples and descriptions of all calls and tags. The SOP was validated by evaluating manual review performance that was completed by novice reviewers. These reviewers assessed somatic variants before and after reading the SOP and manual review labels were compared to the true variant status, which was determined via orthogonal sequencing. This validation showed that the SOP improves somatic variant identification by 16.7% and increased inter-reviewer agreement by 12.7%. Therefore, this research demonstrated that the manual review SOP improves the somatic variant refinement step in the MPS pipeline.

6.2.3 The open-sourced CIViC Annotation Pipeline (OpenCAP) permits development of rationally designed clinical capture panels linked to clinical relevance summaries

Clinical annotation is a severe bottleneck within the precision oncology pipeline. For each cancer patient, there can be hundreds of true somatic variants that all require annotation and prioritization to optimize a custom treatment protocol for each individual. There are several issues with the existing methods used for clinical annotation. First, surveying the literature to obtain all information about a specific variant is time consuming and expensive. Typically, it is difficult for annotators to access and review all required data associated with a variant and they therefore miss critical information needed for evaluation. Second, condensing all information into a single actionability statement is convoluted and many variants have confounding or conflicting evidence that add complexity to variant annotation. Additionally, evidence from various sources should not be weighted equally and therefore cannot be combined into a single statement. Finally, generating a report in a timely fashion that is easily readable by physicians is not straightforward. Physician literacy with regards to genomic data can be limited and interpretation of reports is highly variable.[42,164] To address these issues, we presented the Open-sourced CIViC Annotation Pipeline (OpenCAP), which is an online tool that can build clinically relevant capture panels and generate clinical reports using variant coordinates. These reports contain information about the variant (e.g., MyVariant.info, coordinates, HGVS expressions, etc.), clinical information (Variant Descriptions, and Assertions), as well as evidence support linked to publications (i.e., PubMed IDs and ASCO IDs) in the form of Evidence Items. The OpenCAP software was tested using a validation panel and 27 individuals with known exome / genome sequencing. The panel built using OpenCAP showed 95% sensitivity in detecting variants observed on original sequencing and all variants were successfully analyzed by the software (**Appendix 7; Figure S1**). OpenCAP will serve as a tool that can automate variant annotation to eliminate the need for literature search and manual report generation.

6.2.4 The MyeloSeq® capture panel demonstrates clinical utility of genomic testing

The MyeloSeq® clinical capture panel is an LDT diagnostic that evaluates 40 genes recurrently mutated in acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS). Using the MyeloSeq® clinical capture panel, we demonstrated that successful genomic identification and annotation greatly impacted clinical care of cancer patients. Specifically, we surveyed physicians who used the MyeloSeq® panel for their patients and observed that 44% of physicians changed their treatment plan based on MyeloSeq® Results. Changes in treatment protocols included adding 39 targeted therapeutics to patients with a targetable variant, altering consolidation therapy in 10 cases based on prognostic information, and confirming AML relapse, or remission for 4 cases where a definitive diagnosis was not possible based on bone marrow biopsy alone. The reported response rate from MyeloSeq® panel was elevated relative to previously cited response rates for other types of genomic data.[123,165] These results provide evidence that alleviation of the annotation bottleneck improves ability for physicians to identify actionable variants and incorporate their implications into patient treatment protocols.

6.3 Future Direction of Existing Projects

6.3.1 Additional features required for the DeepSVR to improve usage and accuracy

Currently, DeepSVR is designed to assess variants that were called based on sequencing data from paired tumor and normal samples. Although several deviations from the testing set were assessed (e.g., different tumor types, alignment to an alternate reference genome, etc.), there are many situations that are not currently supported. For example, the algorithm cannot evaluate tumor-only variants and it cannot incorporate additional support from other sample types (e.g., sequencing data from a metastatic tumor or RNA sequencing data). To incorporate these situations, we will need to train the model with novel variant features such as population-level data (e.g., minor allele frequencies) and sequencing context (e.g., sequence conservation, or proximity to repeat elements). Additionally, the algorithm is exclusively provided on a python-based platform and is not available for other types of coding platforms such as R. These feature limitations reduce total usability for individuals who wish to incorporate the algorithm into their sequencing pipelines. Therefore, supplementing existing software with additional features that increase utility for these common use-cases would be important to optimize integration into workflows.

6.3.2 Improvements of the manual review standard operating procedure to increase utility

Although the manual review SOP addresses many issues associated with the existing refinement paradigm, there are several additional features that could be implemented to improve the existing version of the SOP. The current version provides instructions for visualizing sequencing data using IGV with paired tumor / normal samples. It is our hope to augment this SOP by adding instruction for tumor-only samples, samples with additional sequencing data (e.g., primary tumors from different timepoints, metastatic samples, RNA sequencing, circulating tumor cells, etc.), and samples with different sequencing approaches (e.g., targeted amplicon sequencing). Additionally, we could comment on other common genomic visualization software such as Savant, Trackster, BamView to assist individuals who do not use IGV for genomic visualization. Other developments could include integration of variant annotation into the SOP and into IGVNav. This would require variant annotation with software such as Variant Effect Predictor (VEP) and using the output from this software to inform manual review calls. For example, variants could be annotated with gnomAD values using VEP and we could provide guidance on how these values could be used to inform manual review calls. We could subsequently add features to the IGVNav software to indicate that these annotations informed the ultimate call of the variant.

6.3.3 Expanding Open-sourced CIViC Annotation Pipeline (OpenCAP) for development of rationally designed clinical capture panels linked to clinical relevance summaries

OpenCAP serves two main functions for users. The first allows for the generation of custom capture panels that are linked to clinically relevant variant knowledge. There are several limitations associated with this portion of the software. For example, OpenCAP only pulls variants from a single database (CIViC) and is limited by the features that are incorporated into CIViC variant curation. This could be improved by incorporating additional variant curation databases such as OncoKB, PMKB, and the CGI, and using variant data that are not available in CIViC (e.g., VEP annotations, or dbSNP frequencies). OpenCAP’s second function is to generate clinically actionable reports for physician use. There are also many limitations associated with the annotation feature of the software. Specifically, that output reports only generate annotation if the variant observed in the patient had a direct overlap with the CIViC database (chromosome, start, stop, reference allele, and variant allele). There is need to support categorical variants, which are a collection of variants that fit a named category (e.g., *KRAS* G12/G13, *EGFR* Exon 20 Insertion, and *PIK3CA* Mutation). Additionally, there have been requests to incorporate information about interaction between variants. For example, the presence of two specific variants has a different clinical implication relative to when variants are observed in isolation. This type of information should also be made available in OpenCAP reports.

6.3.4 Expanding the MyeloSeq® panel results to broader applications

The MyeloSeq® capture panel was used to determine how physicians use genomic data to change treatment protocols for their patients. Through surveys, we demonstrated that genomic data was used in the treatment of AML to inform prescription of targeted therapeutics, methods for consolidation therapy, and diagnosis of disease status. Beyond these direct clinical implications, the genomic data was also used for residual disease monitoring, confirmation of extramedullary relapse, and evaluating for efficacy of therapy. Although these preliminary results demonstrated how genomic data impacts treatment, there are several limitations and future directions that could be tested to bolster these claims. Specifically, this study evaluated a single disease (AML) at a single institution. Additionally, the test was being offered at a large academic institution with highly knowledgeable physicians and extensive genomic resources. Expanding to smaller sites, non-academic institutions, and local hospitals could inform the scalability, general applicability, and broader impact of this type of test. Additionally, the research focused exclusively on hematologic malignancy and specifically focused on AML. This disease has a well-defined genomic landscape with known implications in disease. AML has several FDA-approved targeted therapeutics, many on-going clinical trials based on observed variants, and described prognostic indications based on genomic data. Therefore, the results from this study might not be easily extrapolated to other disease states. It will be important to assess the utility of other types of genomic diagnostics in terms of changing treatment protocols for those cohorts of individuals.

6.4 Remaining barriers and future horizon within precision oncology

In theory, precision oncology is an elegant and obvious method for optimizing treatment protocols for patients. In practice, however, the components and logistics required to effectively actualize precision oncology are extensive. In this research, we described many of the technological barriers preventing adoption of precision oncology and presented solutions to some of these barriers. However, there are additional challenges that must be overcome to further improve the depth and breadth of precision oncology. These include providing increased access to therapy, broader access to clinical trials, and improved education for patients and physicians.[166]

Treatment modalities that harness precision oncology are not widely accessible to all populations. Targeted therapeutics tend to be much more expensive than traditional cytotoxic chemotherapies and are therefore typically not first-line treatment unless patients can afford to pay out-of-pocket expenses.[167] Additionally, many of these medications are not be accepted by all insurance policies and are not available to people without insurance.[168] Beyond added expense, biologics or targeted therapies, such as monoclonal antibodies, typically require special drug delivery (e.g., injections), which requires regular access to larger hospitals or clinical care centers that have training and experience with these drugs.

In addition to medication access, clinical trial enrollment also results in differential care for underrepresented populations. Specifically, there are known issues with recruiting women and minorities into clinical trials.[169] Lack of women and minorities in these trials means that the clinical trial results do not reflect the ultimate performance of the tests for these patient populations. It also means that women and minorities do not have equal opportunity for enrollment in these clinical trials and therefore have a reduced access during drug development. Given that pharmaceuticals have a long path to market, this discrepancy in enrollment could significantly impact care for years or even decades. Both of these issues prevent the optimization of precision oncology for individuals with variants that might respond to the treatment modalities being tested in the clinics. In addition to the bias against women and minorities, rural populations are also underrepresented in clinical trial outcomes.[170] Similar to issues described above, rural populations have reduced access to large academic hospitals and an inability to regularly attend appointments, which makes them less desirable to enroll into clinical trials.[171] The outcome is a systemic bias against certain populations with reduced access to targeted therapy for cancer care.

Beyond known health disparities that hinder access to customized medicine, lack of health literacy among patients and the medical care team creates persistent obstacles in properly executing on precision oncology. There are three specific educational gaps that exist within precision oncology: 1) physician health literacy, 2) patient health literacy, and 3) communication between patients and their providers. It is recognized that there is a gap in knowledge with regards to genomic understanding at the physician level.[164] Specifically, provider specialty, location, years of practice, and the type of genomic services all impact the ability for physicians to understand genetic information. Finally, even when physicians understand and effectively act on the genomic data, their ability to convey this information to their patients is limited.[43,172,173] This ultimately limits the ability for patients to effectively understand their disease, which impacts the decision making process and appropriate path of care for the patient and their family.

In summary, beyond the technological barriers preventing implementation of precision oncology to clinical workflows, it is important to consider the social aspects that also hinder adoption. As we continue to improve sequencing and annotation technology, it will be necessary to build infrastructure that supports all patients, regardless of race, ethnicity, social status, or geographic location. This can potentially be accomplished through mandating equal representation within clinical trials, improving the educational tools for patients and physicians, and providing universal access to healthcare. Complete adoption of customized medicine will require a concerted effort by the entire community to embrace the intangible and more abstract challenges that prevent adoption of precision oncology to every clinical workflow.

References

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

2. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000).

3. The Cost of Sequencing a Human Genome. *National Human Genome Research Institute (NHGRI)* Available at: https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/. (Accessed: 15th April 2019)

4. Agyeman, A. A. & Ofori-Asenso, R. Perspective: Does personalized medicine hold the future for medicine? *J. Pharm. Bioallied Sci.* **7**, 239–244 (2015).

5. Griffith, M. *et al.* Genome Modeling System: A Knowledge Management Platform for Genomics. *PLoS Comput. Biol.* **11**, e1004274 (2015).

6. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).

7. De Paoli-Iseppi, R. *et al.* Comparison of whole-exome sequencing of matched fresh and formalin fixed paraffin embedded melanoma tumours: implications for clinical decision making. *Pathology* **48**, 261–266 (2016).

8. Binnewies, M. *et al.* Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat. Med.* **24**, 541–550 (2018).

9. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).

10. Stock, W. *et al.* Quantitative real-time RT-PCR monitoring of BCR-ABL in chronic myelogenous leukemia shows lack of agreement in blood and bone marrow samples. *Int. J. Oncol.* **28**, 1099–1103 (2006).

11. GRCh37 - hg19 - Genome - Assembly - NCBI. Available at:

https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/. (Accessed: 7th August 2019)

12. GRCh38 - hg38 - Genome - Assembly - NCBI. Available at: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/. (Accessed: 7th August 2019)

13. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197 (1981).

14. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).

15. M. Burrows, D. J. W. A block-sorting lossless data compression algorithm. (1994).

16. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

17. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).

18. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).

19. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).

20. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).

21. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).

22. Griffith, M. *et al.* Optimizing cancer genome sequencing and analysis. *Cell Syst* **1**, 210–223 (2015).

23. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant Review with the Integrative Genomics Viewer. *Cancer Res.* **77**, e31–e34 (2017).

24. Griffith, O. L. *et al.* Truncating Prolactin Receptor Mutations Promote Tumor Growth in Murine Estrogen Receptor-Alpha Mammary Carcinomas. *Cell Rep.* **17**, 249–260 (2016).

25. Mardis, E. R. The 1,000 genome, the 100,000 analysis? *Genome Med.* **2**, 84 (2010).

26. Dienstmann, R. *et al.* Standardized decision support in next generation sequencing reports of somatic cancer variants. *Mol. Oncol.* 859–873 (2019).

27. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–5 (2014).

28. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* **19**, A68–77 (2015).

29. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).

30. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271–281.e7 (2018).

31. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**, (2017).

32. Huang, L. *et al.* The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. *J. Am. Med. Inform. Assoc.* **24**, 513–519 (2017).

33. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).

34. Verma, A. *et al.* FoundationOne as a relevant tool for comprehensive genomic profiling and assessment of tumor mutation burden in the era of precision oncology in India. *J. Clin. Orthod.* **35**, e23096–e23096 (2017).

35. Griffith, M. *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.* **49**, 170–174 (2017).

36. Cotto, K. C. *et al.* DGIdb 3.0: a redesign and expansion of the drug–gene interaction database.

Nucleic Acids Res. **46**, D1068–D1073 (2018).

37. Xin, J. *et al.* High-performance web services for querying gene and variant annotation. *Genome Biol.* **17**, 91 (2016).

38. Li, M. M. *et al.* Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J. Mol. Diagn.* **19**, 4–23 (2017).

39. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

40. Tandy-Connor, S. *et al.* False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care. *Genet. Med.* **20**, 1515–1521 (2018).

41. Tsang, H., Addepalli, K. & Davis, S. R. Resources for Interpreting Variants in Precision Genomic Oncology Applications. *Front. Oncol.* **7**, 214 (2017).

42. Chow-White, P., Ha, D. & Laskin, J. Knowledge, attitudes, and values among physicians working with clinical genomics: a survey of medical oncologists. *Hum. Resour. Health* **15**, 42 (2017).

43. Arora, N. S. *et al.* Communication challenges for nongeneticist physicians relaying clinical genomic results. *Per. Med.* **14**, 423–431 (2016).

44. Kamps, R. *et al.* Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification. *Int. J. Mol. Sci.* **18**, (2017).

45. Good, B. M., Ainscough, B. J., McMichael, J. F., Su, A. I. & Griffith, O. L. Organizing knowledge to enable personalization of medicine in cancer. *Genome Biol.* **15**, 438 (2014).

46. The AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov.* **7**, 818–831 (2017).

47. Consortium, T. I. C. G. & The International Cancer Genome Consortium. Erratum: International network of cancer genome projects. *Nature* **465**, 966–966 (2010).

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

49. Hoskinson, D. C., Dubuc, A. M. & Mason-Suares, H. The current state of clinical interpretation of sequence variants. *Curr. Opin. Genet. Dev.* **42**, 33–39 (2017).

50. Yorczyk, A., Robinson, L. S. & Ross, T. S. Use of panel tests in place of single gene tests in the cancer genetics clinic. *Clin. Genet.* **88**, 278–282 (2015).

51. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

52. Roy, S. *et al.* Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J. Mol. Diagn.* (2017).

53. Rheinbay, E. *et al.* Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**, 55–60 (2017).

54. Ott, P. A. *et al.* Corrigendum: An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **555**, 402 (2018).

55. Ma, C. X. *et al.* A Phase I Trial of BKM120 (Buparlisib) in Combination with Fulvestrant in Postmenopausal Women with Estrogen Receptor-Positive Metastatic Breast Cancer. *Clin. Cancer Res.* **22**, 1583–1591 (2016).

56. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N. Engl. J. Med.* **372**, 2481–2498 (2015).

57. Rasche, L. *et al.* Spatial genomic heterogeneity in multiple myeloma revealed by multi-region sequencing. *Nat. Commun.* **8**, 268 (2017).

58. Barnell, E. K. *et al.* Standard operating procedure for somatic variant refinement of sequencing data with paired tumor and normal samples. *Genet. Med.* (2018). doi:10.1038/s41436-018-0278-z

59. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).

60. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

61. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).

62. Simola, D. F. & Kim, J. Sniper: improved SNP discovery by multiply mapping deep sequenced reads. *Genome Biol.* **12**, R55 (2011).

63. Ding, J. *et al.* Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data. *Bioinformatics* **28**, 167–175 (2012).

64. Spinella, J.-F. *et al.* SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. *BMC Genomics* **17**, 912 (2016).

65. Strom, S. P. Current practices and guidelines for clinical next-generation sequencing oncology testing. *Cancer Biol Med* **13**, 3–11 (2016).

66. Griffith, M. *et al.* Comprehensive genomic analysis reveals FLT3 activation and a therapeutic strategy for a patient with relapsed adult B-lymphoblastic leukemia. *Exp. Hematol.* **44**, 603–613 (2016).

67. Krysiak, K. *et al.* Recurrent somatic mutations affecting B-cell receptor signaling pathway genes in follicular lymphoma. *Blood* **129**, 473–483 (2017).

68. Klco, J. M. *et al.* Association Between Mutation Clearance After Induction Therapy and Outcomes in Acute Myeloid Leukemia. *JAMA* **314**, 811–822 (2015).

69. Uy, G. L. *et al.* Dynamic changes in the clonal structure of MDS and AML in response to epigenetic therapy. *Leukemia* **31**, 872–881 (2017).

70. Lesurf, R. *et al.* Genomic characterization of HER2-positive breast cancer and response to

neoadjuvant trastuzumab and chemotherapy-results from the ACOSOG Z1041 (Alliance) trial. *Ann. Oncol.* **28**, 1070–1077 (2017).

71. Welch, J. S. *et al.* TP53 and Decitabine in Acute Myeloid Leukemia and Myelodysplastic Syndromes. *N. Engl. J. Med.* **375**, 2023–2036 (2016).

72. Rohan, T. E. *et al.* Somatic mutations in benign breast disease tissue and risk of subsequent invasive breast cancer. *Br. J. Cancer* (2018). doi:10.1038/s41416-018-0089-7

73. Mahlokozera, T. *et al.* Biological and therapeutic implications of multisector sequencing in newly diagnosed glioblastoma. *Neuro. Oncol.* **20**, 472–483 (2018).

74. Wagner, A. H. *et al.* Recurrent WNT pathway alterations are frequent in relapsed small cell lung cancer. *Nat. Commun.* **9**, 3787 (2018).

75. Duncavage, E. J. *et al.* Mutation Clearance after Transplantation for Myelodysplastic Syndrome. *N. Engl. J. Med.* **379**, 1028–1041 (2018).

76. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

77. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).

78. Picard Tools - By Broad Institute. Available at: http://broadinstitute.github.io/picard/. (Accessed: 28th June 2018)

79. Varoquaux, G. *et al.* Scikit-learn: Machine Learning Without Learning the Machinery. *GetMobile: Mobile Comp. and Comm.* **19**, 29–33 (2015).

80. Oliphant, T. E. Python for Scientific Computing. *Computing in Science Engineering* **9**, 10–20 (2007).

81. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

82. Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* **6**, 224ra24 (2014).

83. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. Med.* **22**, 276–282 (2012).

84. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41**, e67 (2013).

85. Swamidass, S. J., Bittker, J. A., Bodycombe, N. E., Ryder, S. P. & Clemons, P. A. An economic framework to prioritize confirmatory tests after a high-throughput screen. *J. Biomol. Screen.* **15**, 680–686 (2010).

86. Settles, B. & Craven, M. An Analysis of Active Learning Strategies for Sequence Labeling Tasks. in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* 1070–1079 (Association for Computational Linguistics, 2008).

87. Settles, B. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* **6**, 1–114 (2012).

88. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).

89. Krøigård, A. B., Thomassen, M., Lænkholm, A.-V., Kruse, T. A. & Larsen, M. J. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. *PLoS One* **11**, e0151664 (2016).

90. Callari, M. *et al.* Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers. *Genome Med.* **9**, 35 (2017).

91. Cai, L., Yuan, W., Zhang, Z., He, L. & Chou, K.-C. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Sci. Rep.* **6**, 36540 (2016).

92. Kim, J. *et al.* Good Laboratory Standards for Clinical Next-Generation Sequencing Cancer Panel Tests. *J Pathol Transl Med* **51**, 191–204 (2017).

93. Govindan, R. *et al.* Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150**, 1121–1134 (2012).

94. Giannakis, M. *et al.* RNF43 is frequently mutated in colorectal and endometrial cancers. *Nat. Genet.* **46**, 1264–1266 (2014).

95. Sandmann, S. *et al.* Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Sci. Rep.* **7**, 43169 (2017).

96. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

97. Fiume, M., Williams, V., Brook, A. & Brudno, M. Savant: genome browser for high-throughput sequencing data. *Bioinformatics* **26**, 1938–1944 (2010).

98. Goecks, J., Coraor, N., Galaxy Team, Nekrutenko, A. & Taylor, J. NGS analyses by visualization with Trackster. *Nat. Biotechnol.* **30**, 1036–1039 (2012).

99. Carver, T. *et al.* BamView: visualizing and interpretation of next-generation sequencing read alignments. *Brief. Bioinform.* **14**, 203–212 (2013).

100. T. Mooney, J. Walker, S. Siebert, C. Miller, M. Griffith. *cancer-genomics-workflow*. (McDonnell Genome Institute).

101. Yost, S. E. *et al.* Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Res.* **40**, e107 (2012).

102. Walsh, P. S., Erlich, H. A. & Higuchi, R. Preferential PCR amplification of alleles: mechanisms and solutions. *PCR Methods Appl.* **1**, 241–250 (1992).

103. Potapov, V. & Ong, J. L. Correction: Examining Sources of Error in PCR by Single-Molecule Sequencing. *PLoS One* **12**, e0181128 (2017).

104. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).

105. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.*

39, e90 (2011).

106. Metzker, M. L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11**, 31 (2010).

107. Andrade Nunes, R., Nunes, R. A. & Harris, L. N. The HER2 Extracellular Domain as a Prognostic and Predictive Factor in Breast Cancer. *Clin. Breast Cancer* **3**, 125–135 (2002).

108. Collins, F. S. & Varmus, H. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).

109. Gray, S. W., Hicks-Courant, K., Cronin, A., Rollins, B. J. & Weeks, J. C. Physicians' Attitudes About Multiplex Tumor Genomic Testing. *J. Clin. Oncol.* **32**, 1317–1323 (2014).

110. Dorschner, M. O. *et al.* Actionable, Pathogenic Incidental Findings in 1,000 Participants' Exomes. *Am. J. Hum. Genet.* **93**, 631–640 (2013).

111. Green, R. C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **15**, 565–574 (2013).

112. Cheng, D. T. *et al.* Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J. Mol. Diagn.* **17**, 251–264 (2015).

113. Rubio-Perez, C., Deu-Pons, J., Tamborero, D., Lopez-Bigas, N. & Gonzalez-Perez, A. Rational design of cancer gene panels with OncoPaD. *Genome Med.* **8**, 98 (2016).

114. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, l1 (2013).

115. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).

116. Patterson, S. E. *et al.* The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. *Hum. Genomics* **10**, 4 (2016).

117. Administration, U. S. F. &. D. & U.S. Food & Drug Administration. FDA grants marketing approval to FoundationOne CDx in vitro diagnostic. *Case Medical Research* (2017).

doi:10.31525/fda1-ucm587387.htm

118. Lee, T. *et al.* Non-small Cell Lung Cancer with Concomitant EGFR, KRAS, and ALK Mutation: Clinicopathologic Features of 12 Cases. *J Pathol Transl Med* **50**, 197–203 (2016).

119. Wagner, A. H. *et al.* A harmonized meta-knowledgebase of clinical interpretations of cancer genomic variants. *bioRxiv* 366856 (2018). doi:10.1101/366856

120. Patterson, S., Statz, C., Yin, T. & Mockus, S. The JAX Clinical Knowledgebase: A Valuable Resource for Identifying Evidence Related to Complex Molecular Signatures in Different Types of Cancer. *Cancer Genet.* **214**, 33 (2017).

121. Waalkes, A., Penewit, K., Wood, B. L., Wu, D. & Salipante, S. J. Ultrasensitive detection of acute myeloid leukemia minimal residual disease using single molecule molecular inversion probes. *Haematologica* **102**, 1549–1557 (2017).

122. Boyle EA, E. al. MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/24867941. (Accessed: 20th March 2019)

123. Toner, B. Survey of Precision Oncology Programs Finds Agreement on Testing, Divergence in Care Delivery. *Precision Oncology News* (2019). Available at: https://www.precisiononcologynews.com/cancer/survey-precision-oncology-programs-finds-agreement-testing-divergence-care-delivery?utm_source=Sailthru&utm_medium=email&utm_campaign=PON%20Survey%20Mailing&utm_term=GW%20Registrants. (Accessed: 6th August 2019)

124. Stone, R. M. *et al.* Midostaurin plus Chemotherapy for Acute Myeloid Leukemia with a FLT3 Mutation. *N. Engl. J. Med.* **377**, 454–464 (2017).

125. Kim, A. S. *et al.* Comparison of Laboratory-Developed Tests and FDA-Approved Assays for BRAF, EGFR, and KRAS Testing. *JAMA Oncol* **4**, 838–841 (2018).

126. Li, M. M. *et al.* Standards and Guidelines for the Interpretation and Reporting of Sequence

Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J. Mol. Diagn.* **19**, 4–23 (2017).

127. Starlinger, J. *et al.* Variant information systems for precision oncology. *BMC Med. Inform. Decis. Mak.* **18**, 107 (2018).

128. Weipert, C. M. *et al.* Physician Experiences and Understanding of Genomic Sequencing in Oncology. *J. Genet. Couns.* **27**, 187–196 (2018).

129. Gornick, M. C. *et al.* Interpretations of the term 'actionable' when discussing genetic test results: what you mean is not what I heard. *J. Genet. Couns.* **28**, 334–342 (2019).

130. Statz, C. M., Patterson, S. E. & Mockus, S. M. Barriers preventing the adoption of comprehensive cancer genomic profiling in the clinic. *Expert Rev. Mol. Diagn.* **17**, 549–555 (2017).

131. Au, C. H., Wa, A., Ho, D. N., Chan, T. L. & Ma, E. S. K. Clinical evaluation of panel testing by next-generation sequencing (NGS) for gene mutations in myeloid neoplasms. *Diagn. Pathol.* **11**, 11 (2016).

132. Kurzrock, R. *et al.* NCCN Oncology Research Program's Investigator Steering Committee and NCCN Best Practices Committee Molecular Profiling Surveys. *J. Natl. Compr. Canc. Netw.* **13**, 1337–1346 (2015).

133. Cancer Genome Atlas Research Network *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).

134. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).

135. Wu, M., Li, C. & Zhu, X. FLT3 inhibitors in acute myeloid leukemia. *J. Hematol. Oncol.* **11**, 133 (2018).

136. Abou Dalle, I. & DiNardo, C. D. The role of enasidenib in the treatment of mutant IDH2 acute myeloid leukemia. *Ther. Adv. Hematol.* **9**, 163–173 (2018).

137. Nassereddine, S., Lap, C. J., Haroun, F. & Tabbara, I. The role of mutant IDH1 and IDH2 inhibitors in the treatment of acute myeloid leukemia. *Ann. Hematol.* **96**, 1983–1991 (2017).

138. Ivey, A. *et al.* Assessment of Minimal Residual Disease in Standard-Risk AML. *N. Engl. J. Med.* **374**, 422–433 (2016).

139. Marcucci, G. *et al.* Abnormal cytogenetics at date of morphologic complete remission predicts short overall and disease-free survival, and higher relapse rate in adult acute myeloid leukemia: results from cancer and leukemia group B study 8461. *J. Clin. Oncol.* **22**, 2410–2418 (2004).

140. Chen, Y. *et al.* Persistence of cytogenetic abnormalities at complete remission after induction in patients with acute myeloid leukemia: prognostic significance and the potential role of allogeneic stem-cell transplantation. *J. Clin. Oncol.* **29**, 2507–2513 (2011).

141. Ohtake, S. *et al.* Randomized study of induction therapy comparing standard-dose idarubicin with high-dose daunorubicin in adult patients with previously untreated acute myeloid leukemia: the JALSG AML201 Study. *Blood* **117**, 2358–2365 (2011).

142. Döhner, H. *et al.* Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood* **115**, 453–474 (2010).

143. Cornelissen, J. J. *et al.* The European LeukemiaNet AML Working Party consensus statement on allogeneic HSCT for patients with AML in remission: an integrated-risk adapted approach. *Nat. Rev. Clin. Oncol.* **9**, 579–590 (2012).

144. Bloomfield, C. D. *et al.* Frequency of prolonged remission duration after high-dose cytarabine intensification in acute myeloid leukemia varies by cytogenetic subtype. *Cancer Res.* **58**, 4173–4179 (1998).

145. Schlenk, R. F. *et al.* Mutations and Treatment Outcome in Cytogenetically Normal Acute Myeloid Leukemia. *N. Engl. J. Med.* **358**, 1909–1918 (2008).

146. Röllig, C. *et al.* Long-term prognosis of acute myeloid leukemia according to the new genetic risk

classification of the European LeukemiaNet recommendations: evaluation of the proposed reporting system. *J. Clin. Oncol.* **29**, 2758–2765 (2011).

147. O'Donnell, M. R. *et al.* Acute Myeloid Leukemia, Version 3.2017, NCCN Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Canc. Netw.* **15**, 926–957 (2017).

148. Weinberg, O. K., Sohani, A. R., Bhargava, P. & Nardi, V. Diagnostic work-up of acute myeloid leukemia. *Am. J. Hematol.* **92**, 317–321 (2017).

149. Bejar, R. *et al.* TET2 mutations predict response to hypomethylating agents in myelodysplastic syndrome patients. *Blood* **124**, 2705–2712 (2014).

150. Feng, Y., Li, X., Cassady, K., Zou, Z. & Zhang, X. TET2 Function in Hematopoietic Malignancies, Immune Regulation, and DNA Repair. *Front. Oncol.* **9**, 210 (2019).

151. Perl, A. E. Availability of FLT3 inhibitors--how do we use them? *Blood* blood.2019876821 (2019).

152. Stone, R. M., Manley, P. W., Larson, R. A. & Capdeville, R. Midostaurin: its odyssey from discovery to approval for treating acute myeloid leukemia and advanced systemic mastocytosis. *Blood Adv* **2**, 444–453 (2018).

153. Stein, E. M. *et al.* Enasidenib in mutant IDH2 relapsed or refractory acute myeloid leukemia. *Blood* **130**, 722–731 (2017).

154. Center for Devices & Radiological Health. List of Cleared or Approved Companion Diagnostic Devices. *U.S. Food and Drug Administration* (2019). Available at: https://www.fda.gov/medical-devices/vitro-diagnostics/list-cleared-or-approved-companion-diagnostic-devices-vitro-and-imaging-tools. (Accessed: 17th July 2019)

155. Hou, H.-A. *et al.* Characterization of acute myeloid leukemia with PTPN11 mutation: the mutation is closely associated with NPM1 mutation but inversely related to FLT3/ITD. *Leukemia* **22**, 1075–1078 (2008).

156. Gaidzik, V. I. *et al.* Prognostic impact of WT1 mutations in cytogenetically normal acute myeloid

leukemia: a study of the German-Austrian AML Study Group. *Blood* **113**, 4505–4511 (2009).

157. Perera-Bel, J. *et al.* From somatic variants towards precision oncology: Evidence-driven reporting of treatment options in molecular tumor boards. *Genome Med.* **10**, 18 (2018).

158. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–510 (2012).

159. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).

160. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

161. Haferlach, T. *et al.* Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia* **28**, 241–247 (2014).

162. Papaemmanuil, E. *et al.* Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* **122**, 3616–27; quiz 3699 (2013).

163. Walter, M. J. *et al.* Clonal diversity of recurrently mutated genes in myelodysplastic syndromes. *Leukemia* **27**, 1275–1282 (2013).

164. Ha, V. T. D., Frizzo-Barker, J. & Chow-White, P. Adopting clinical genomics: a systematic review of genomic literacy among physicians in cancer care. *BMC Med. Genomics* **11**, 18 (2018).

165. Morash, M., Mitchell, H., Beltran, H., Elemento, O. & Pathak, J. The Role of Next-Generation Sequencing in Precision Medicine: A Review of Outcomes in Oncology. *J Pers Med* **8**, (2018).

166. Kurtovic-Kozaric, A. *et al.* Lack of Access to Targeted Cancer Treatment Modalities in the Developing World in the Era of Precision Medicine: Real-Life Lessons From Bosnia. *J Glob Oncol* **4**, 1–5 (2018).

167. Shih, Y.-C. T., Smieliauskas, F., Geynisman, D. M., Kelly, R. J. & Smith, T. J. Trends in the Cost and Use of Targeted Cancer Therapies for the Privately Insured Nonelderly: 2001 to 2011. *J. Clin. Oncol.* **33**, 2190–2196 (2015).

168. Fang, P. *et al.* Rising and Falling Trends in the Use of Chemotherapy and Targeted Therapy Near

the End of Life in Older Patients With Cancer. *J. Clin. Oncol.* **37**, 1721–1731 (2019).

169. Chen, A. *et al.* Representation of Women and Minorities in Clinical Trials for New Molecular Entities and Original Therapeutic Biologics Approved by FDA CDER from 2013 to 2015. *J. Womens. Health* **27**, 418–429 (2018).

170. Virani, S., Burke, L., Remick, S. C. & Abraham, J. Barriers to recruitment of rural patients in cancer clinical trials. *J. Oncol. Pract.* **7**, 172–177 (2011).

171. Baquet, C. R., Commiskey, P., Daniel Mullins, C. & Mishra, S. I. Recruitment and participation in clinical trials: socio-demographic, rural/urban, and health care access predictors. *Cancer Detect. Prev.* **30**, 24–33 (2006).

172. Haga, S. B. *et al.* Developing patient-friendly genetic and genomic test reports: formats to promote patient engagement and understanding. *Genome Med.* **6**, 58 (2014).

173. Haga, S. B., Kim, E., Myers, R. A. & Ginsburg, G. S. Primary Care Physicians' Knowledge, Attitudes, and Experience with Personal Genetic Testing. *J Pers Med* **9**, (2019).

174. Eijkelenboom, A. *et al.* Reliable Next-Generation Sequencing of Formalin-Fixed, Paraffin-Embedded Tissue Using Single Molecule Tags. *J. Mol. Diagn.* **18**, 851–863 (2016).

175. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

176. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).

177. Larson, D. E., Abbott, T. E. & Wilson, R. K. Using SomaticSniper to Detect Somatic Single Nucleotide Variants. *Curr. Protoc. Bioinformatics* **45**, 15.5.1–8 (2014).

178. Reble, E., Castellani, C. A., Melka, M. G., O'Reilly, R. & Singh, S. M. VarScan2 analysis of de novo variants in monozygotic twins discordant for schizophrenia. *Psychiatr. Genet.* **27**, 62–70 (2017).

179. Huang, Z. GATK Test Protocol v1 (protocols.io.mhdc326). *protocols.io* (2018).

doi:10.17504/protocols.io.mhdc326

180. Hinrichs, A. S. *et al.* The UCSC genome browser database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).

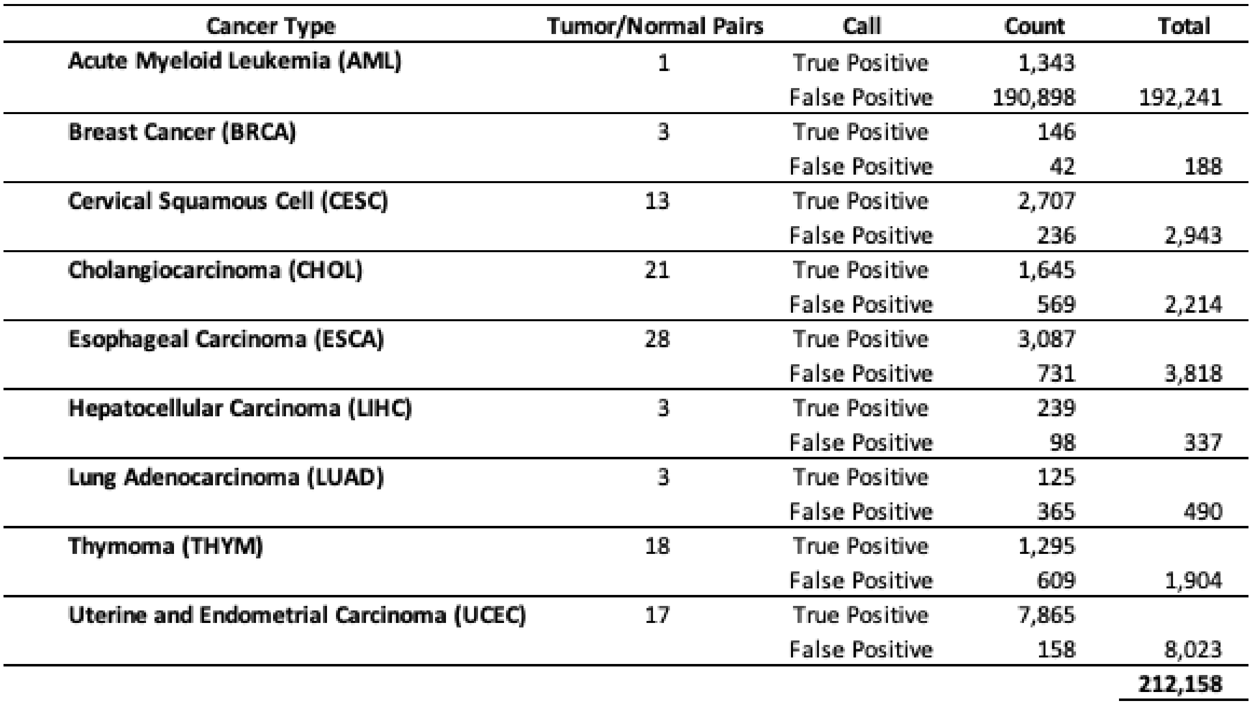
181. Pritchard, C. C. *et al.* Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. *J. Mol. Diagn.* **16**, 56–67 (2014).

Appendix 1. Chapter 2 Supplementary Tables

Table S1. Cross-tabulation performance on hold out test set parsed by reviewer, disease, normal depth, and tumor depth.

Feature		S	A	F	AUC
Reviewer	Reviewer 1	1282	871	1347	0.968
	Reviewer 2	4757	2511	2601	0.965
	Reviewer 3	62	59	9	0.926
	Reviewer 4	14	13	4	0.979
Disease	AML	872	581	1424	0.971
	GST	14	13	4	0.979
	MPNST	6	12	124	0.993
	SCLC	2465	1463	673	0.962
	Breast	2319	600	1401	0.956
	Colorectal	12	283	124	0.757
	GBM	150	205	57	0.908
	Lymphoma	242	238	148	0.964
	Melanoma	35	59	6	0.895
Normal Depth	X < 0.010	3101	1179	2360	0.960
	0.018 > X > 0.010	1465	856	661	0.970
	X > 0.018	1549	1419	940	0.967
Tumor Depth	X < 0.010	6042	3392	3885	0.966
	0.018 > X > 0.010	43	49	57	0.948
	X > 0.018	30	13	19	0.944

Table S2. Distribution of orthogonal validation calls from the AML31 case and the 106 The Cancer Genome Atlas (TCGA) tumor/normal pairs used to assess model performance.



Cancer Type	Tumor/Normal Pairs	Call	Count	Total
Acute Myeloid Leukemia (AML)	1	True Positive	1,343	
		False Positive	190,898	192,241
Breast Cancer (BRCA)	3	True Positive	146	
		False Positive	42	188
Cervical Squamous Cell (CESC)	13	True Positive	2,707	
		False Positive	236	2,943
Cholangiocarcinoma (CHOL)	21	True Positive	1,645	
		False Positive	569	2,214
Esophageal Carcinoma (ESCA)	28	True Positive	3,087	
		False Positive	731	3,818
Hepatocellular Carcinoma (LIHC)	3	True Positive	239	
		False Positive	98	337
Lung Adenocarcinoma (LUAD)	3	True Positive	125	
		False Positive	365	490
Thymoma (THYM)	18	True Positive	1,295	
		False Positive	609	1,904
Uterine and Endometrial Carcinoma (UCEC)	17	True Positive	7,865	
		False Positive	158	8,023
				212,158

Table S3. Distribution of manual review calls from the 37 cases used to assess model performance by independent sequencing data with manual review.

Cancer Type	Tumor/Normal pairs	Call	Count	Total
Small Cell Lung Carcinoma (SCLC)	4	Somatic	2,526	2,686
		Fail	145	
		Ambiguous	15	
Follicular Lymphoma (FL)	14	Somatic	865	1,723
		Fail	858	
		Ambiguous	0	
Head and Neck Squamous Cell Carcinoma (HNSCC)	19	Somatic	1,986	9,170
		Fail	6891	
		Ambiguous	293	
				13,579

Appendix 2. Chapter 2 Supplementary Figures

Figure S1. The deep learning model performs well on the hold out test set (n=13,530 variants), 10-fold cross validation with a simplified disease feature (n=27,470 variants), and 10-fold cross validation with the reviewer feature removed (n=27,470 variants). a) ROC curve and reliability diagram performance of the deep learning model on the hold out test set with all 71 described features. **b)** ROC curve and reliability diagram performance of the deep learning model 10-fold cross validation set with the cancer type simplified to solid versus liquid tumor status. c) ROC curve and reliability diagram performance of the deep learning model 10-fold cross validation set with the reviewer feature removed.

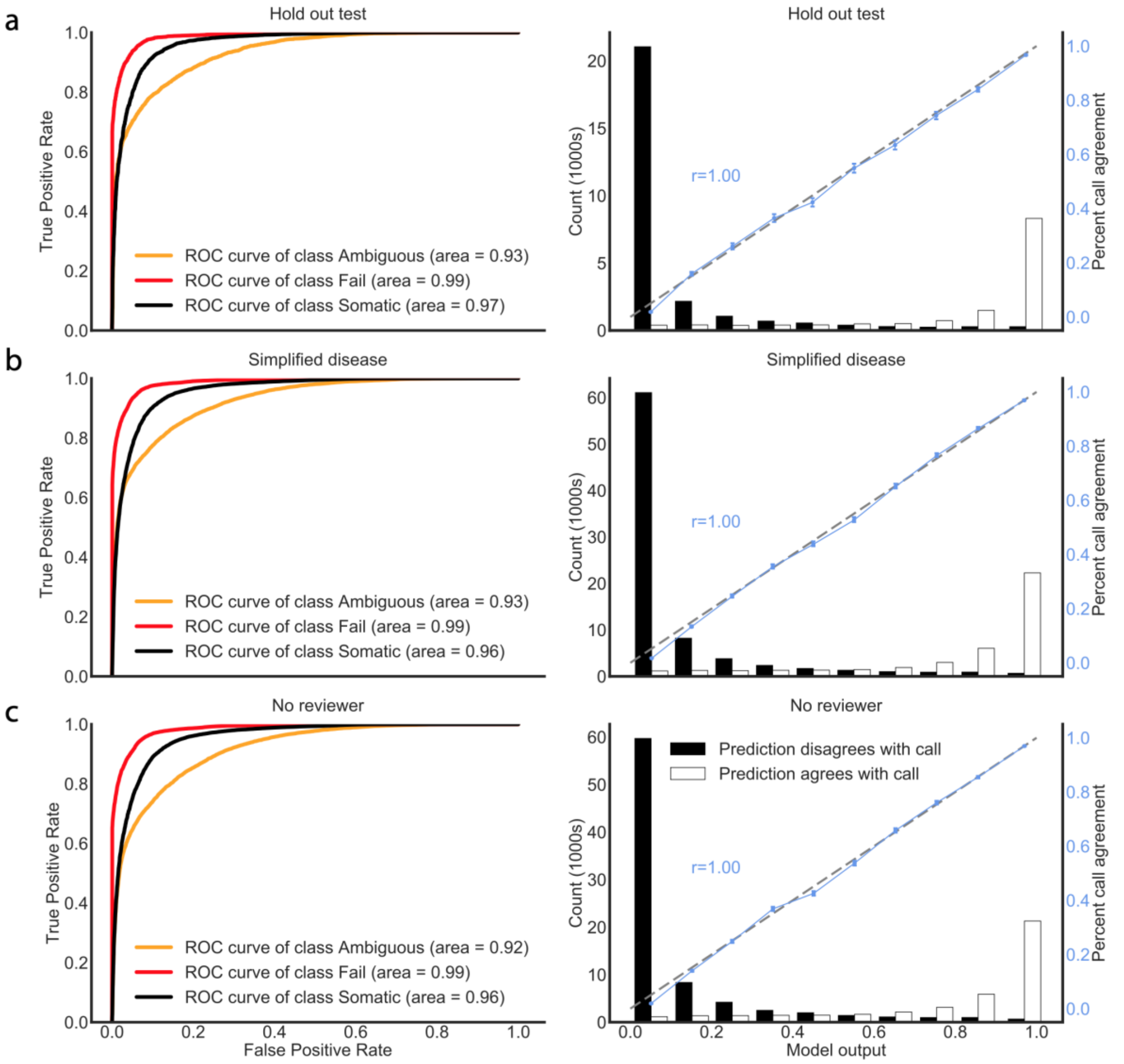


Figure S2. Deep learning model outputs from the hold out test set (n=13,530 variants) are well-scaled across all predicted classes (ambiguous, fail, and somatic). The correlation between the model output and the manual review call was assessed for all three different classes of calls (ambiguous, fail, and somatic). For each class, model outputs were binned into 10 groups ranging from 0.00-1.00. For each bin, the total number of manual review calls that agree and disagree with the individual class were plotted. The ratio of agreement to disagreement was plotted for each bin and compared to the identity line (x=y) using the Pearson's correlation coefficient (r).

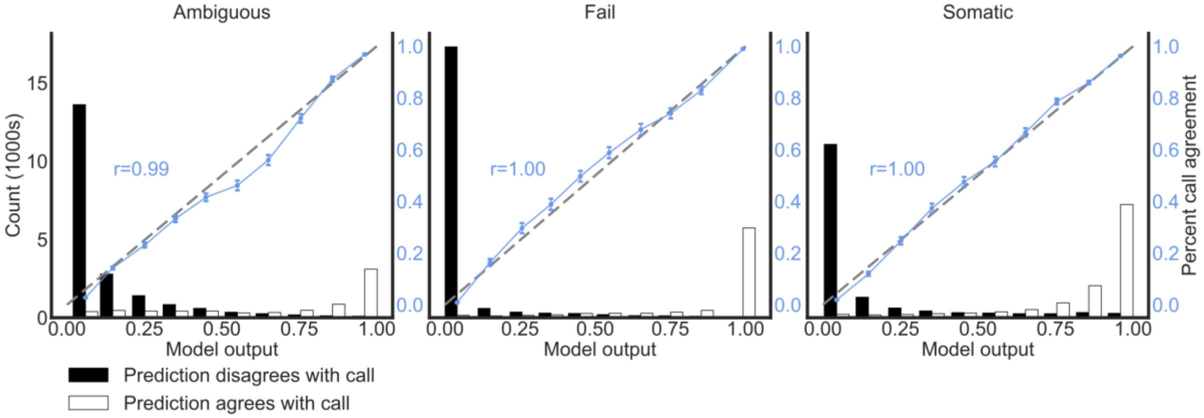


Figure S3. The deep learning model performs better than the random forest model on independent sequencing data with manual review labels (n=4 small cell lung cancer cases with 2,686 total variants). a) ROC curves outlining deep learning and random forest model performances on independent sequencing data with manual review labels (n=4 small cell lung cancer cases with 2,686 total variants). b) Curves showing batch effect correction after retraining machine learning models with incremental subsets of variants from the independent sequencing data. Independent sequencing data was partitioned in random stratified increments of 5% (from 0-75%) and used to train a new model (increments = 179 varians). The x-axis outlines the number of independent variants included in training. The y-axis plots the resulting model’s ROC AUC. The ambiguous class shows significant stochasticity due to low representation in the test dataset (n=15 variants).

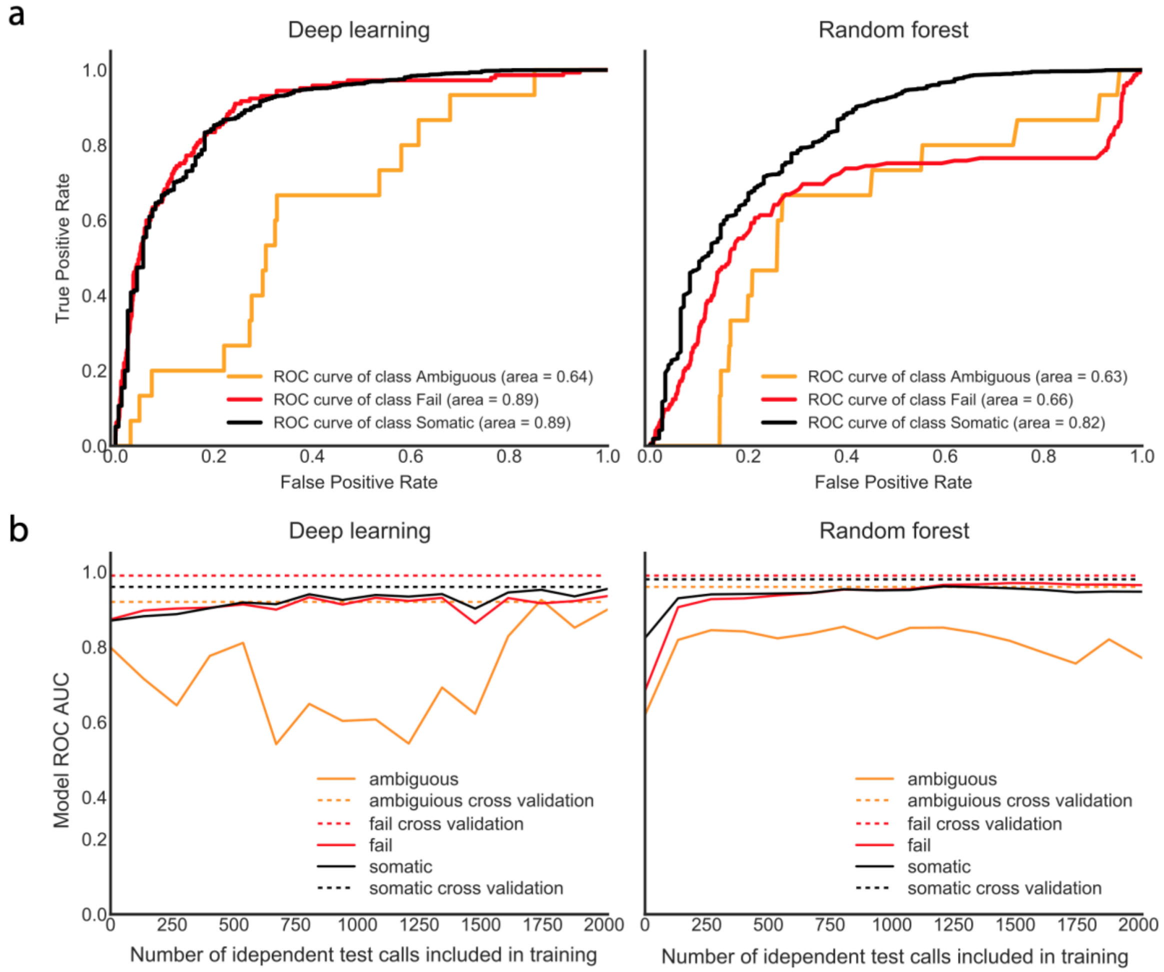


Figure S4. IGV snapshots of clinically relevant variants that were original labeled as somatic by manual reviewers but were subsequently identified as fail using the deep learning model and manual re-review. a) Failure due to short inserts and directional artifacts. b) Failure due to multiple variants artifacts. c) Failure due to multiple mismatches across variant-supporting reads. d) Failure due to ends of reads artifact.

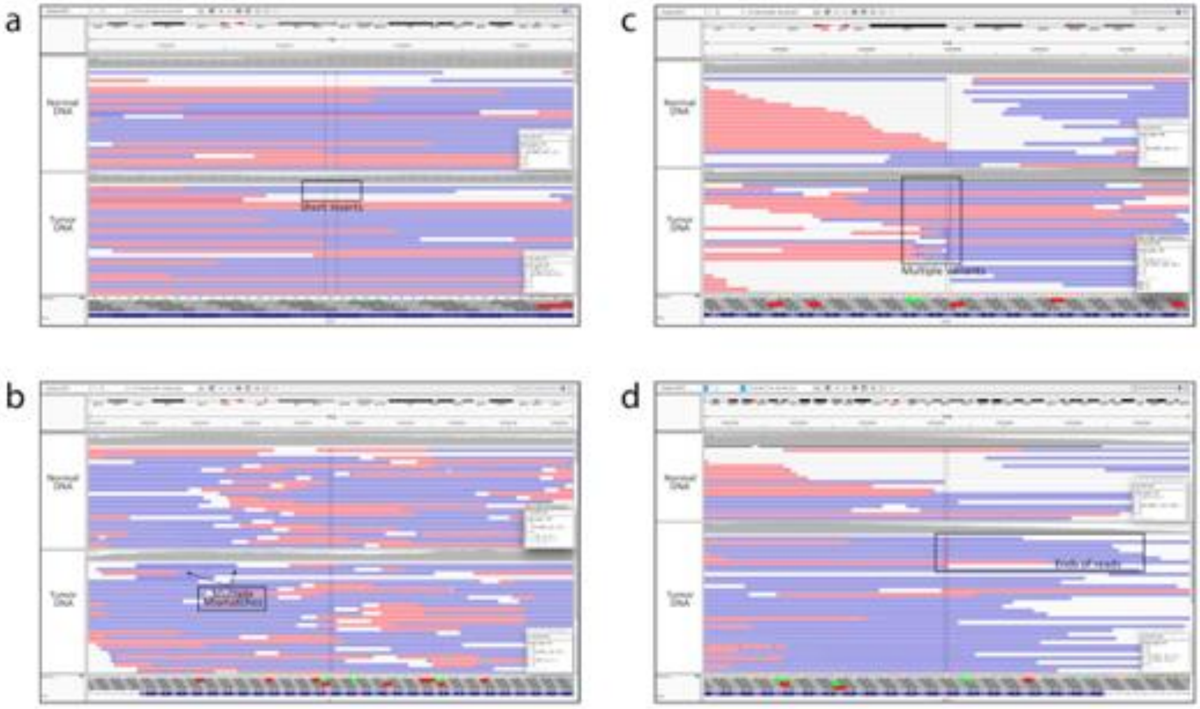
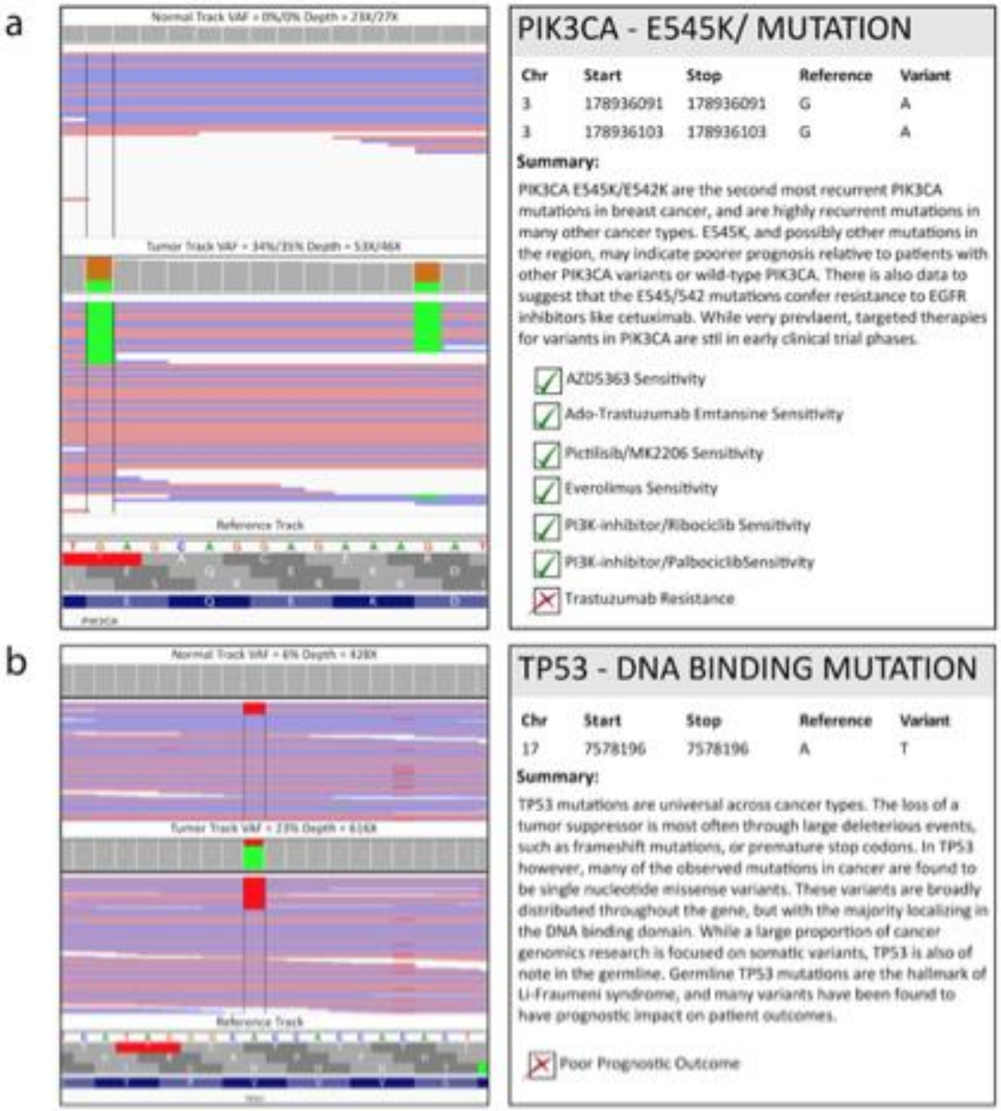


Figure S5. IGV snapshots of clinically relevant variants that were original labeled as fail or ambiguous by manual reviewers but were subsequently identified as somatic using the deep learning model and manual re-review. For each snapshot, the normal tracks and the tumor tracks show aligned reads that were obtained from normal tissue and the tumor tissue, respectively. Variant summaries obtained from CIViC show gene name, variant type, variant coordinates, clinical summary, and relevant clinical action items. **a)** Original reviewer conservatively labeled both PIK3CA variants as ambiguous due to multiple mismatches in reads, however, both variants appear to be somatic and occur at known cancer driver hotspots (E542K/E545K). **b)** Original reviewer failed this variant due to high levels of variant reads in the normal track, however, given that this variant was derived from a hematologic malignancy, this level of tumor in normal is permissible.


a
Normal Track VAF = 0%/0% Depth = 23X/27X
Tumor Track VAF = 34%/35% Depth = 53X/46X
Reference Track
PIK3CA
b
Normal Track VAF = 6% Depth = 428X
Tumor Track VAF = 23% Depth = 616X
Reference Track

PIK3CA - E545K/ MUTATION

Chr	Start	Stop	Reference	Variant
3	178936091	178936091	G	A
3	178936103	178936103	G	A

Summary:

PIK3CA E545K/E542K are the second most recurrent PIK3CA mutations in breast cancer, and are highly recurrent mutations in many other cancer types. E545K, and possibly other mutations in the region, may indicate poorer prognosis relative to patients with other PIK3CA variants or wild-type PIK3CA. There is also data to suggest that the E545/542 mutations confer resistance to EGFR inhibitors like cetuximab. While very previaent, targeted therapies for variants in PIK3CA are stil in early clinical trial phases.

☑ AZD5363 Sensitivity
☑ Ado-Trastuzumab Emtansine Sensitivity
☑ Pictilisib/MK2206 Sensitivity
☑ Everolimus Sensitivity
☑ PI3K-inhibitor/Ribociclib Sensitivity
☑ PI3K-inhibitor/PalbociclibSensitivity
☒ Trastuzumab Resistance

TP53 - DNA BINDING MUTATION

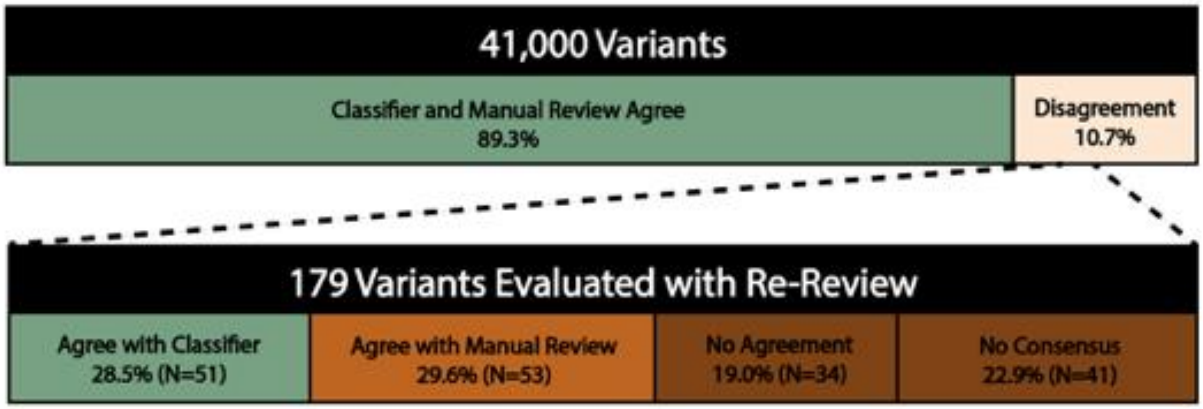
Chr	Start	Stop	Reference	Variant
17	7578196	7578196	A	T

Summary:

TP53 mutations are universal across cancer types. The loss of a tumor suppressor is most often through large deleterious events, such as frameshift mutations, or premature stop codons. In TP53 however, many of the observed mutations in cancer are found to be single nucleotide missense variants. These variants are broadly distributed throughout the gene, but with the majority localizing in the DNA binding domain. While a large proportion of cancer genomics research is focused on somatic variants, TP53 is also of note in the germline. Germline TP53 mutations are the hallmark of Li-Fraumeni syndrome, and many variants have been found to have prognostic impact on patient outcomes.

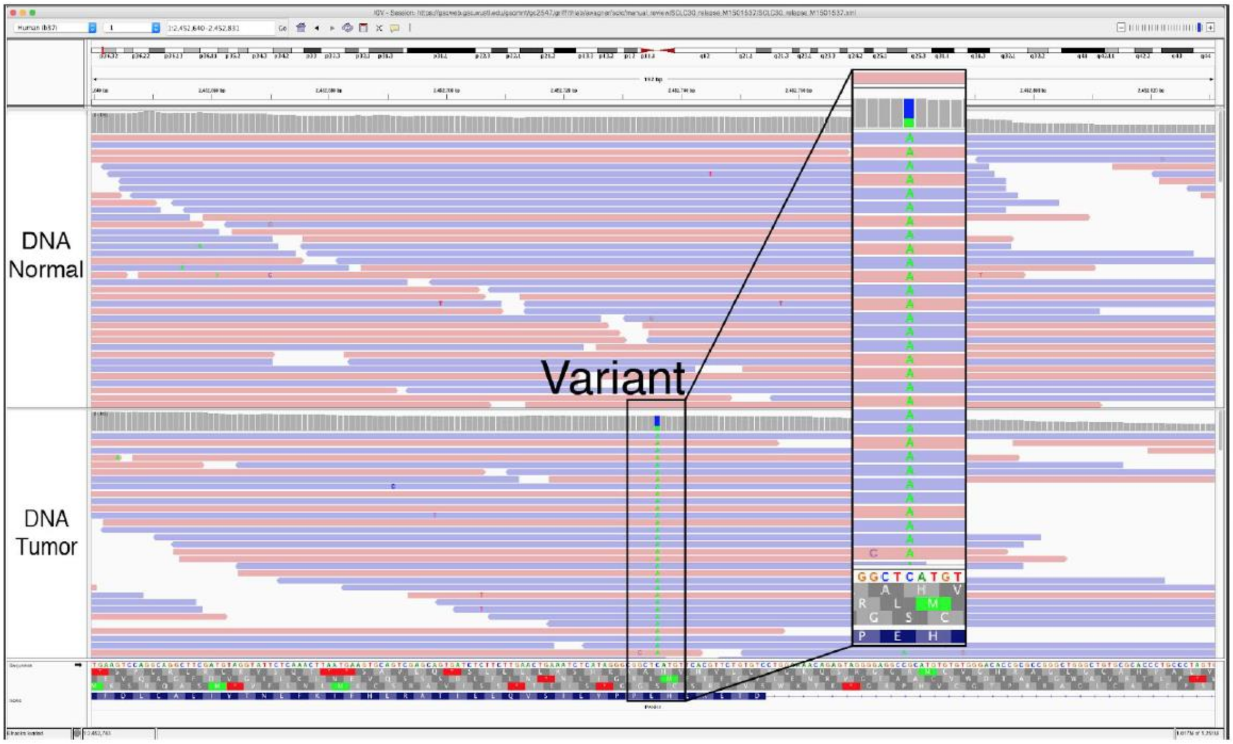
☒ Poor Prognostic Outcome

Figure S6. Variants that show disagreement between the classifier and original manual review demonstrates high levels of inter-reviewer variability. Of the 10.7% of variants that disagree with the original manual review, call, 179 variants were sampled to conduct manual re-review. When comparing the classifier call to the re-review consensus call, 42.9% of variants showed high inter-reviewer variability and/or inability to determine a consensus.



Appendix 3. Chapter 3 Supplementary Figures

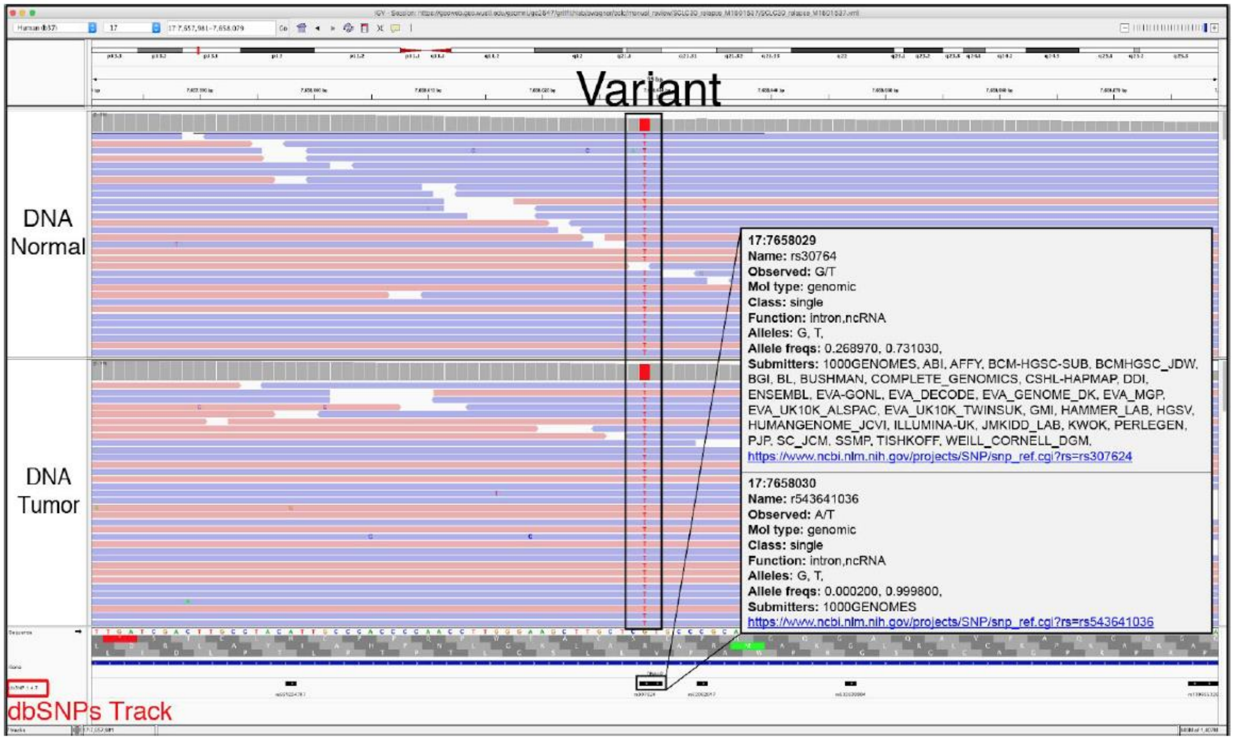
Figure S1. Example of a Somatic variant (S). Somatic calls are made when the variant has sufficient support in the tumor track with absence of obvious sequencing artifacts. In this example, the variant is presumed to be a real somatic variant. When evaluating the reference sequence in the Genome Features section, the reference allele is a cytosine (C). The alignments and coverage in the DNA tumor track show that approximately 20% of reads support a variant adenine (A) allele (green). Importantly, there are no reads supporting the variant in the normal sample, indicating that the variant is a somatic variant rather than a germline polymorphism. Using the gene annotation track, we can predict that this (C>A) base change would result in an ATG (M; Methionine) to ATT (I; Isoleucine) missense variant in the PANK4 gene (Note: this gene is transcribed on the negative strand).



Helpful Hints:

1. Somatic variants, due to impure tumor samples, will typically have VAF less than 50%. However, the latter is not a strict rule because random sampling, copy-number alterations, loss of heterozygosity, and other factors can sometimes produce somatic VAF at or above 50%.
2. If the expected variant is not visualized during manual review, it is possible that: 1) IGV is not focused on the correct coordinates, 2) the genome version is incorrect, or 3) the supporting reads have been lost due to down-sampling.

Figure S2. Example of a Germline variant (G). Germline calls are made when the variant has sufficient support in the normal track beyond what is considered attributable to tumor contamination of the normal. In this example, the variant is presumed to be a germline polymorphism. The reference allele is a guanine (G), however reads in the DNA normal/tumor tracks support a thymine (T) allele. This indicates that the variant is likely a homozygous germline polymorphism. The Single Nucleotide Polymorphisms (SNPs) Track provides further support that the variant in question is a common polymorphism.



17:7658029
Name: rs30764
Observed: G/T
Mol type: genomic
Class: single
Function: intron,ncRNA
Alleles: G, T,
Allele freqs: 0.268970, 0.731030,
Submitters: 1000GENOMES, ABI, AFFY, BCM-HGSC-SUB, BCMHGSC_JDW, BGI, BL, BUSHMAN, COMPLETE_GENOMICS, CSHL-HAPMAP, DDI, ENSEMBL, EVA-GONL, EVA_DECODE, EVA_GENOME_DK, EVA_MGP, EVA_UK10K_ALSPAC, EVA_UK10K_TWINSUK, GMI, HAMMER_LAB, HGSV, HUMANGENOME_JCVI, ILLUMINA-UK, JMKIDD_LAB, KWOK, PERLEGEN, PJP, SC_JCM, SSMP, TISHKOFF, WEILL_CORNELL_DGM,
https://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=rs307624

17:7658030
Name: r543641036
Observed: A/T
Mol type: genomic
Class: single
Function: intron,ncRNA
Alleles: G, T,
Allele freqs: 0.000200, 0.999800,
Submitters: 1000GENOMES
https://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=rs543641036

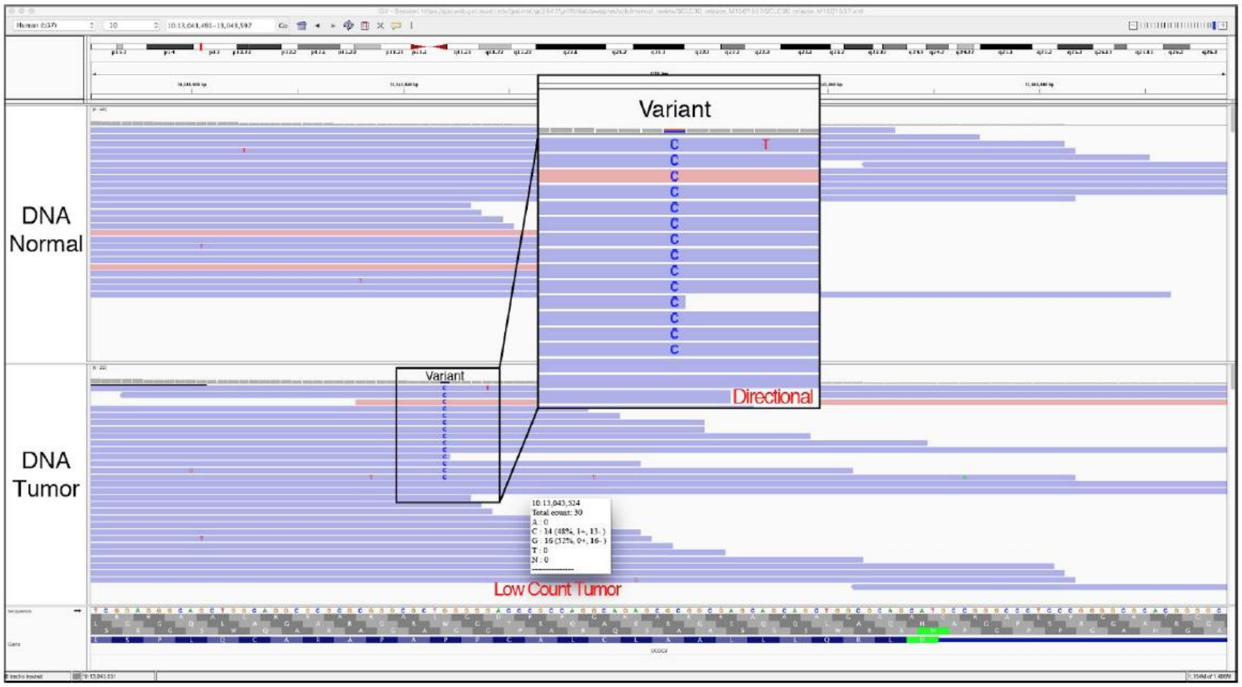
Helpful Hints:

1. Typically, germline variants present with a Variant Allele Frequency (VAF) near 50% or 100%,
2. indicating hetero- or homozygosity, respectively.
3. Bulk tumors typically contain some normal cells. Therefore, given adequate depth, 100% VAF in a non-purified tumor sample should be suspicious and is likely a homozygous germline polymorphism.
4. To view the SNPs Track in the Genome Features section, use the “Load from Server” feature in IGV. Examples for loading this track are shown below:

   GRCH37: “File” > “Load from Server...” > “Annotations” > “Variation and Repeats” > “dbsnps1.4.7”
   GRCH38: “File” > “Load from Server...” > “Annotations” > “Common Snps 1.4.2”

   If the variant in question is also in the SNPs Track, then it is most likely germline. Clicking on, or hovering over, the grey bar in the SNPs Track will create a popup with additional information about the germline SNP. 5) A germline call after somatic variant caller filtering is suspect and might reveal underlying issues with the analysis pipeline being used.

Figure S3. Example of an Ambiguous variant (A). Ambiguous calls are made when the variant in question could be a true somatic variant, but the reviewer is not confident due to sequencing features, genomic context, and/or, corresponding reads. In this example, the variant has support from fourteen reads, but most are on negative read strands (93%). Additionally, several of the supporting reads have multiple mismatches indicating potentially low-quality reads. More information would be required to call this variant somatic or fail, therefore, the correct label is ambiguous.

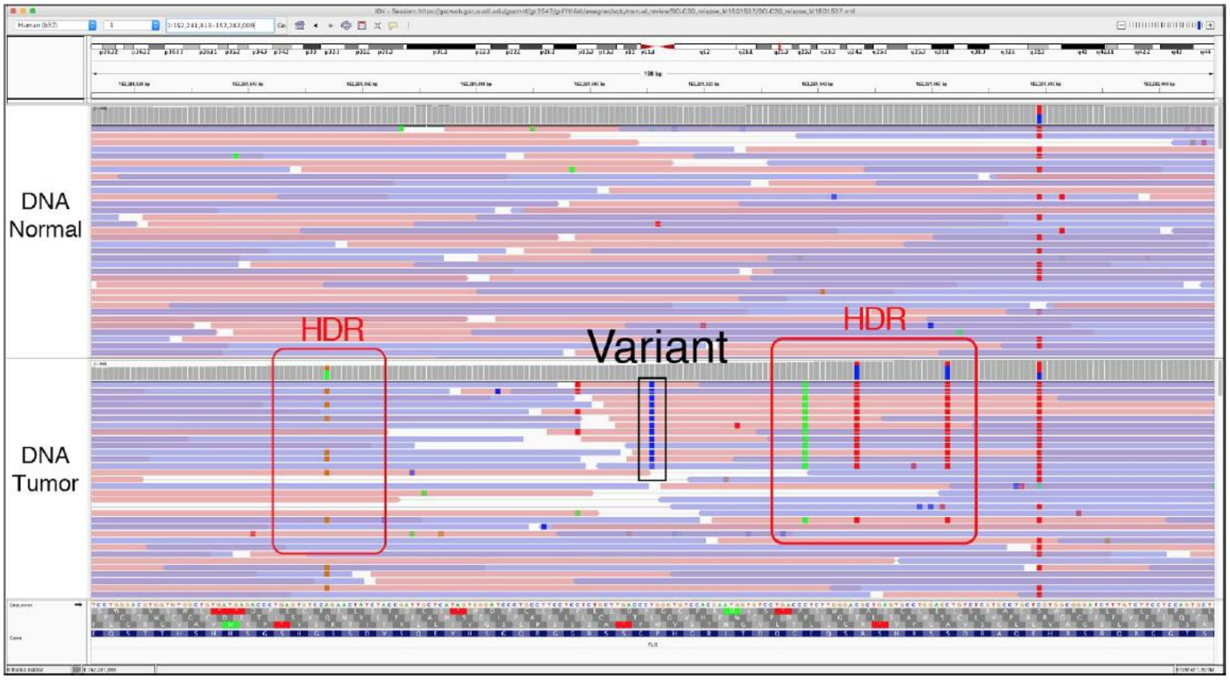


```
10:13,043,524
Total count: 30
A : 0
C : 14 (48%, 1+, 13- )
G : 16 (52%, 0+, 16- )
T : 0
N : 0
---------------
```

Helpful Hints:

1. Using Tags and Notes can help individuals understand why variants were labeled as ambiguous.

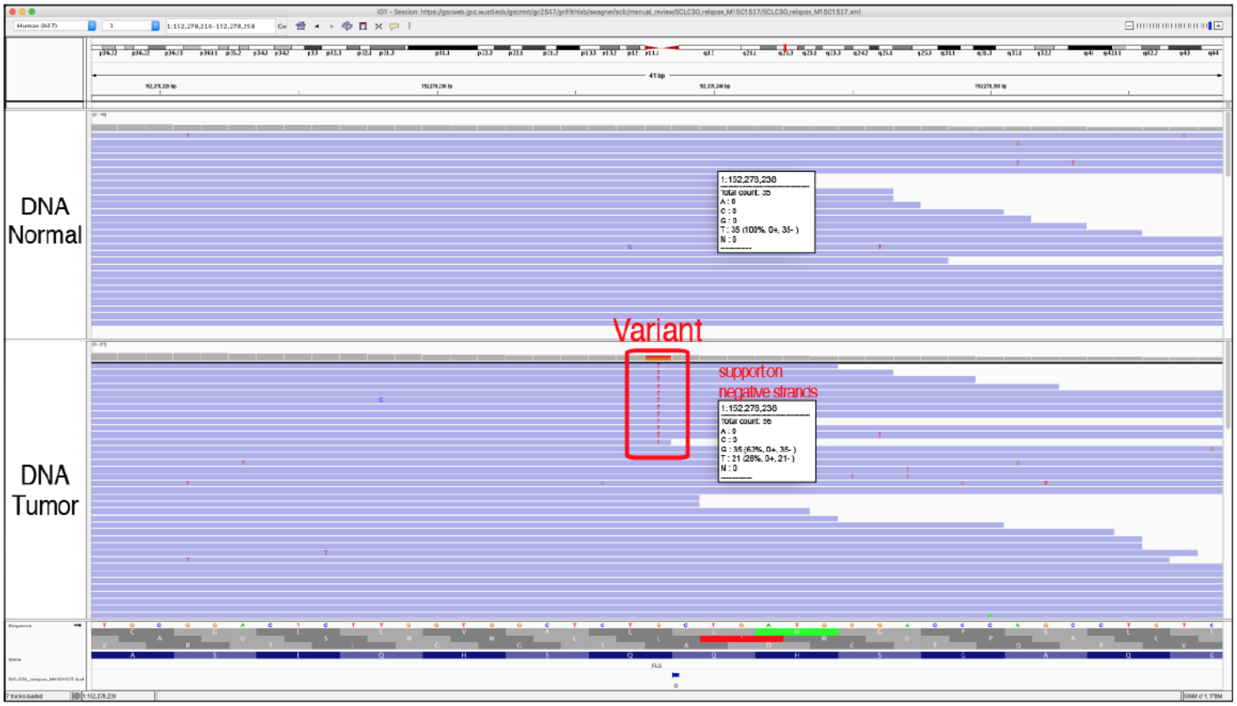
Figure S4. Example of a Failed variant (F). Failed calls are made when the variant has low variant support and/or reads that indicate a sequencing artifact. In this example, the variant in question is likely attributable to a pipeline artifact and is therefore not a true variant. When the IGV window is zoomed in, the variant appears to be somatic; however, in the provided zoomed-out window, we reveal a region of high discrepancy. High discrepancy regions (HDR) can suggest improper alignment in regions of high homology across the genome or errors in the reference assembly. Given the HDR pattern observed, this variant is most likely a false positive and should be failed during manual review.



Helpful Hints:

1. Using Tags and Notes can help individuals understand why variants were labeled as fail.

Figure S5. Example of Directional (D). The Directional tag is used when the variant in question can only be found on reads that are sequenced in either the positive or the negative direction. Typically, this is caused by strand bias during sequencing. To properly visualize the directional artifacts, IGV tracks must be colored by read strand.



1:152,278,238
Total count: 35
A : 0
C : 0
G : 0
T : 35 (100%, 0+, 35-)
N : 0

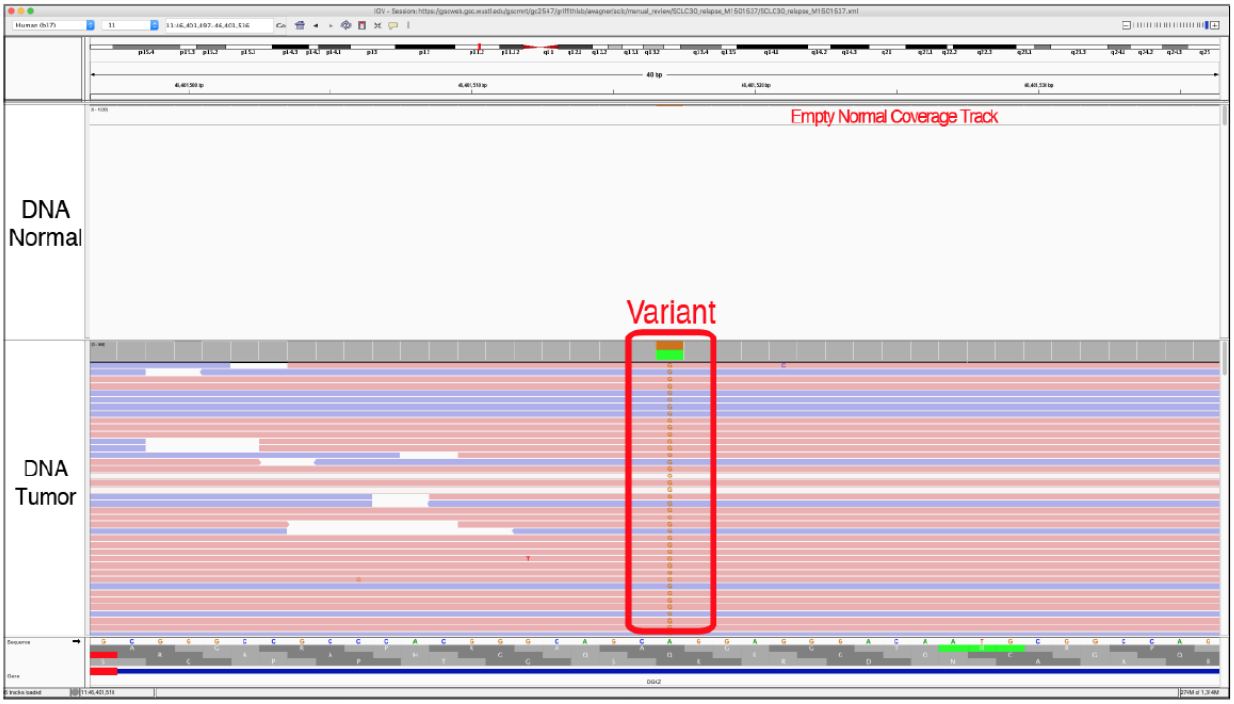
1:152,278,238
Total count: 56
A : 0
C : 0
G : 35 (63%, 0+, 35-)
T : 21 (28%, 0+, 21-)
N : 0

Helpful Hints:

1. This tag can best be assessed when the reads are not viewed as pairs. When viewing data tracks as pairs, the reads in both directions are overlaid and could possibly make the variant appear to be exclusively supported by read strands in a particular direction. 2) To observe this artifact, it is necessary to color the alignments by read strand:

   Right click on data track > "Color alignments by" > "read strand"

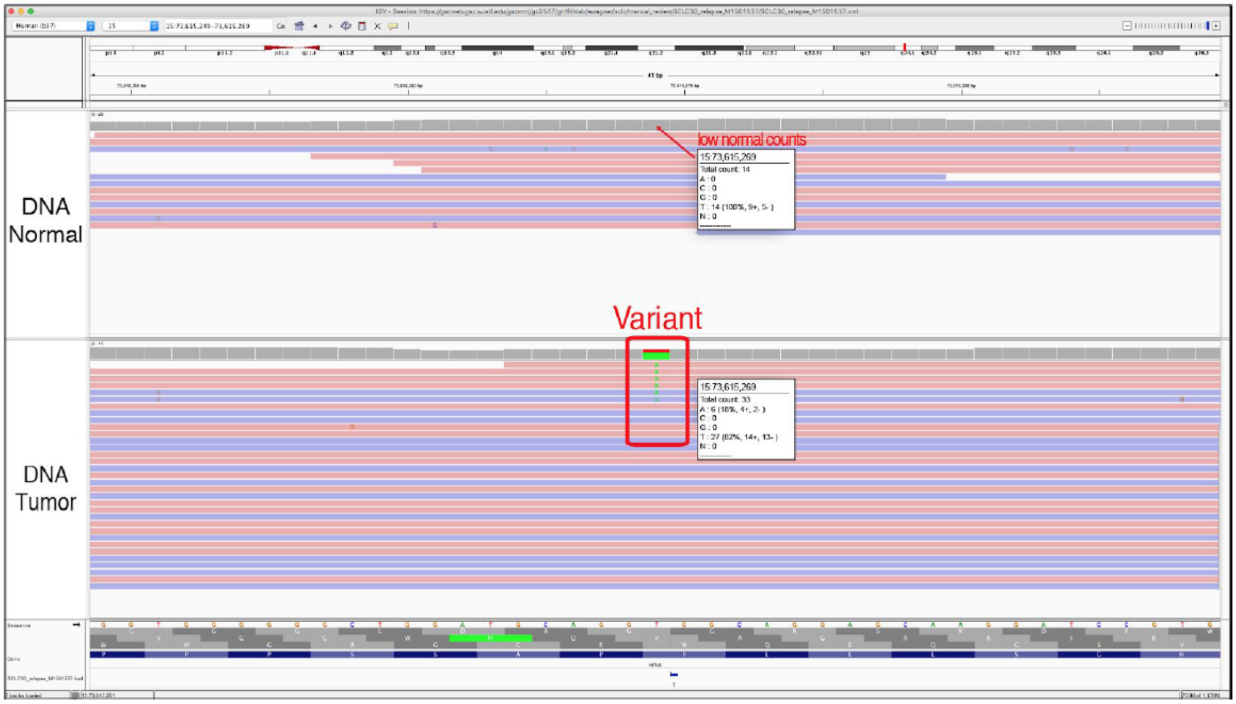
Figure S6. Example of No Count Normal (NCN). The No Count Normal tag is used when there is no coverage in the normal track, preventing adequate comparison to the tumor track. This can occur when there is no normal track available or if there is no coverage in the normal track at the locus in question. Typically, at least 20X coverage in both normal and tumor tracks is required to make accurate calls; however, this threshold is experiment-specific.



Helpful Hints:

1. If a variant has low coverage in the normal track, it can be treated like a tumor only sample. This might require populating the Genome Features section with a SNPs Track (e.g., dbSNP, 1000 genomes, ExAC, gnomAD, etc.) to ensure that the variant is not a polymorphism (see Step 3 in Figure 3A for setting up manual review).
2. Thresholds can be used to pre-filter variants with no coverage in tumor or normal to eliminate the need to evaluate these variants during manual review.

Figure S7. Example of Low Count Normal (LCN). The Low Count Normal tag is used when there is inadequate coverage in the normal track (coverage < 20X), preventing adequate comparison to the tumor track. A popup window with coverage information can be viewed by clicking on the locus position in the coverage track. Typically, at least 20X coverage in both normal and tumor tracks is required to make accurate calls; however, this threshold is experiment-specific.



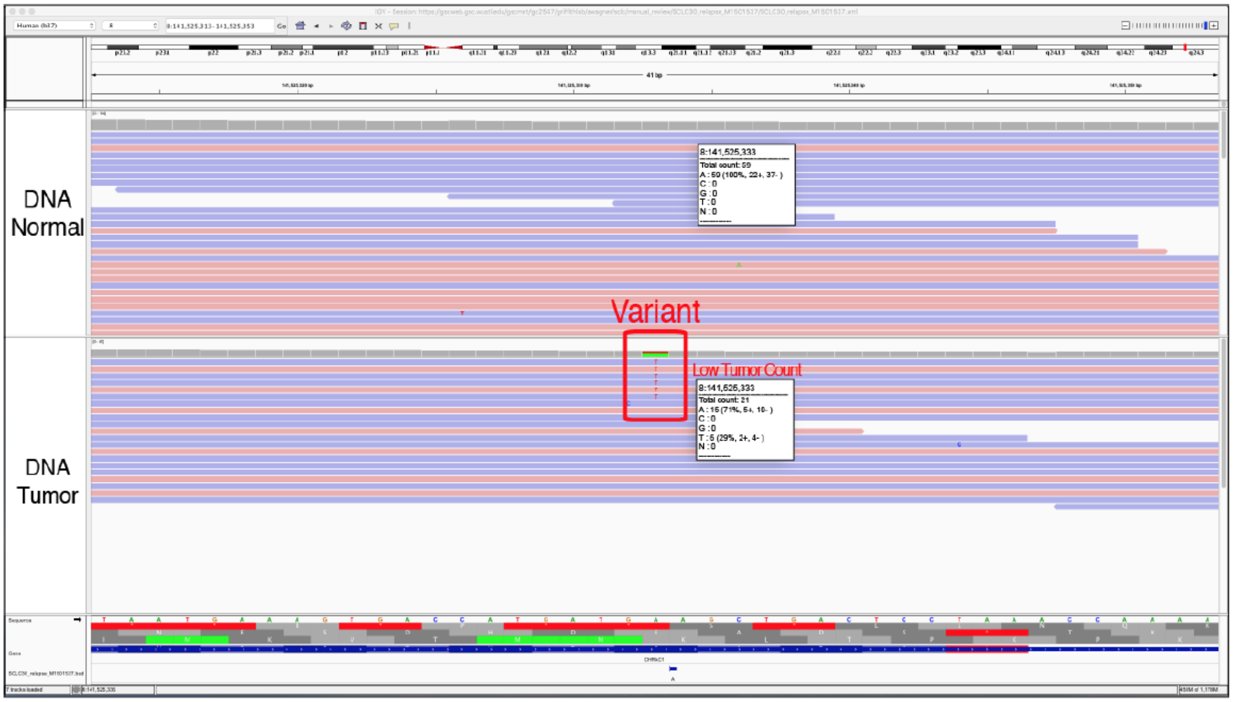
```
15:73,615,269
Total count: 14
A : 0
C : 0
G : 0
T : 14 (100%, 9+, 5- )
N : 0
```

```
15:73,615,269
Total count: 33
A : 6 (18%, 4+, 2- )
C : 0
G : 0
T : 27 (82%, 14+, 13- )
N : 0
```

Helpful Hints:

1. If a variant has low coverage in the normal track, it can be treated like a tumor only sample. This might require populating the Genome Features section with a SNPs Track (e.g., dbSNP, 1000 genomes, ExAC, gnomAD, etc.) to ensure that the variant is not a polymorphism (see Step 3 in Figure 3A for setting up manual review).
2. Thresholds can be used to pre-filter variants with low coverage in tumor or normal to eliminate the need to evaluate these variants during manual review.

Figure S8. Example of Low Count Tumor (LCT). The Low Count Tumor tag is used when there is inadequate coverage in the tumor track (coverage < 20X), preventing adequate comparison to the normal track. A popup window with coverage information can be viewed by clicking on the locus position in the coverage track. Typically, at least 20X coverage in both normal and tumor tracks is required to make accurate calls; however, this threshold is experiment-specific.



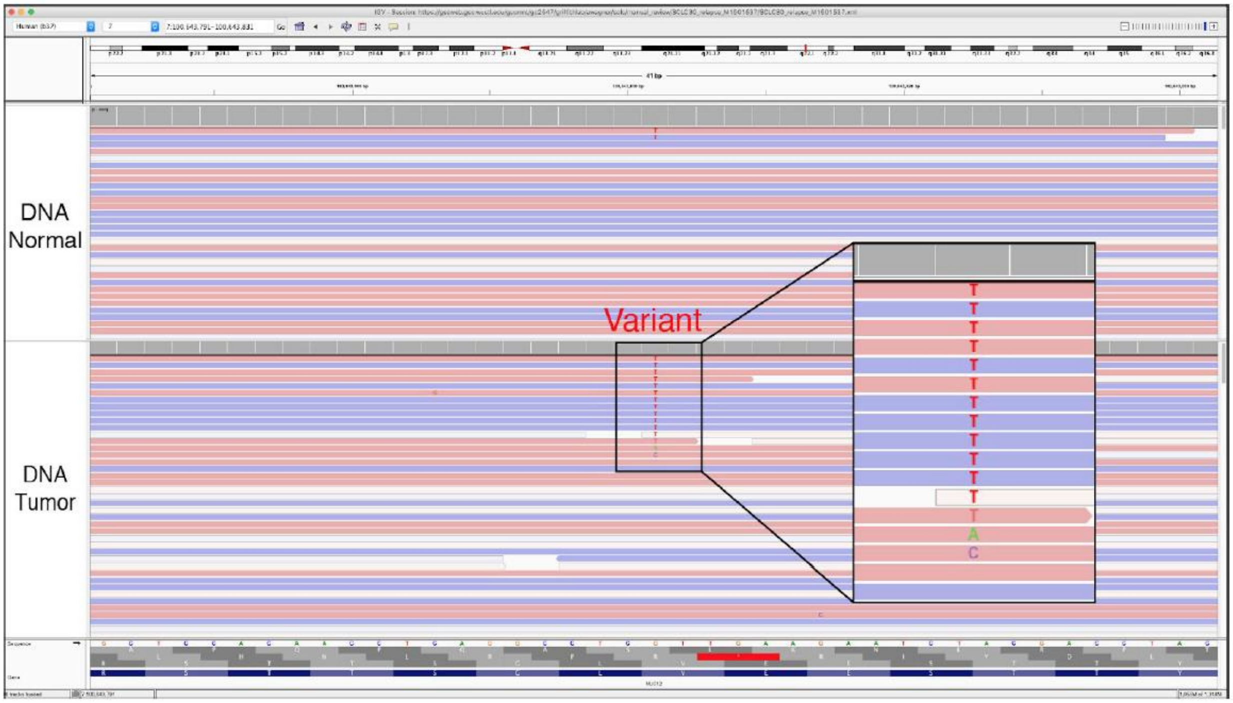
8:141,525,333
Total count: 59
A : 59 (100%, 22+, 37-)
C : 0
G : 0
T : 0
N : 0

8:141,525,333
Total count: 21
A : 15 (71%, 5+, 10-)
C : 0
G : 0
T : 6 (29%, 2+, 4-)
N : 0

Helpful Hints:

1. Calling a variant with low coverage has important downstream implications. When the tumor track has low coverage, variant allele frequency (VAF) estimates can be heavily influence by sequencing noise and sampling bias. This may result in a false negative with an underestimated VAF, a false positive due to over-estimation of the VAF, and/or a true positive call with inaccurate VAF.
2. The LCT tag acts as a bare minimum for tumor coverage but only in concert with a 5% VAF minimum with at least 4-5 reads of support (taking into account short inserts). Therefore, the LCT tag can denote that a variant was considered ambiguous or somatic in a rare sequencing context.
3. Thresholds can be used to pre-filter variants with low coverage in tumor or normal to eliminate the need to evaluate these variants during manual review.

Figure S9. Example of Multiple Variants (MV). The Multiple Variants tag is used if the variant's locus has reads supporting three or more different alleles. In the example shown, there is read support for all four nucleotides (A, C, G, and T) at the same locus. If the putative variant base co-occurs with multiple instances of other bases, it is less likely to be a true somatic variant.



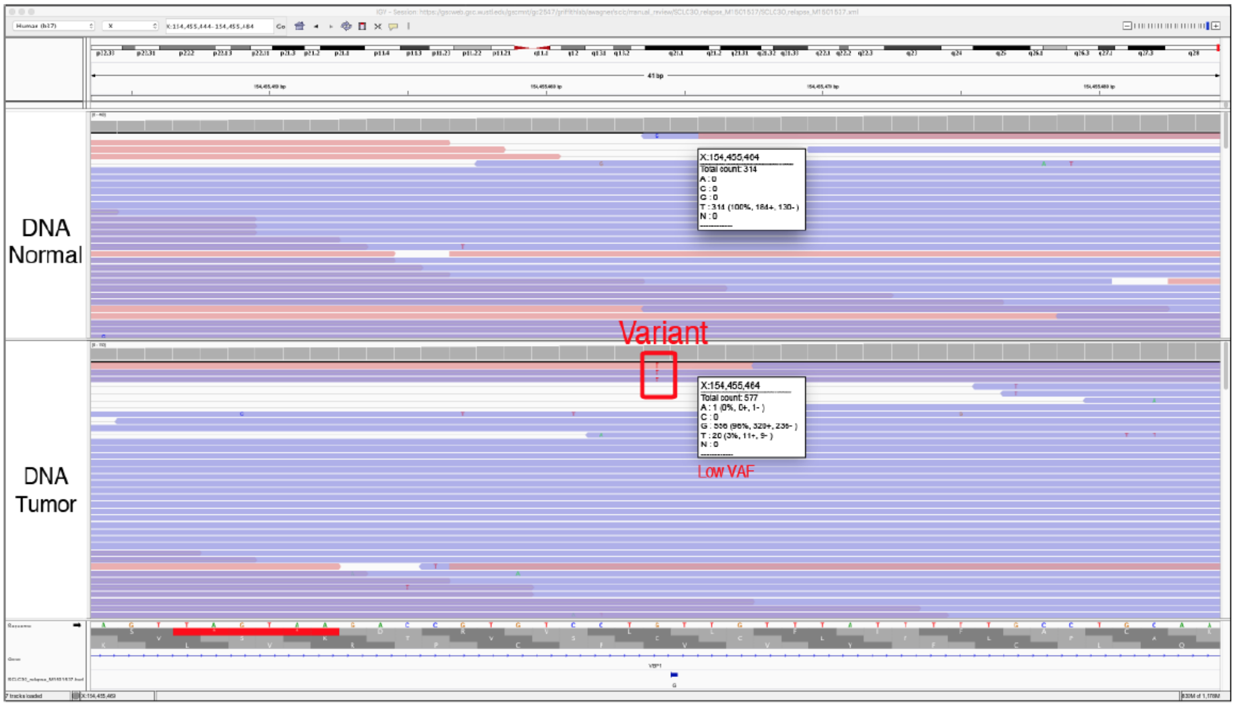
Helpful Hints:

1. Clicking on the coverage track will reveal a popup window with relative abundance of each base at the selected locus.
2. Do not rely on coverage track coloring as there might be multiple variants that have a variant allele frequency (VAF) too small to be represented in the coverage bar. The VAF threshold for coloring the coverage bar can be changed in the IGV preferences panel:

   "View" > "Preferences" > "Alignments" > "Coverage allele-fraction threshold" > insert threshold

3. For very deep data, multiple variants due to random error will start to accumulate. The relative abundance of each base should be considered in cases with deep coverage. 4) While rare, true multi-allelic somatic variants are possible in tumors.

Figure S10. Example of Low Variant Frequency (LVF). The Low Variant Frequency tag is used when there are some reads of support for the variant, but the variant allele frequency (VAF) is relatively low. A popup window with VAF information can be viewed by clicking on the locus position in the coverage track. Typically, at least 5% VAF is required to make confident calls (given 20X coverage); however, this threshold is experiment-specific.



```
X:154,455,464
-----------------------------
Total count: 314
A : 0
C : 0
G : 0
T : 314 (100%, 184+, 130- )
N : 0
-----------
```

```
X:154,455,464
-----------------------------
Total count: 577
A : 1 (0%, 0+, 1- )
C : 0
G : 556 (96%, 320+, 236- )
T : 20 (3%, 11+, 9- )
N : 0
-----------
```

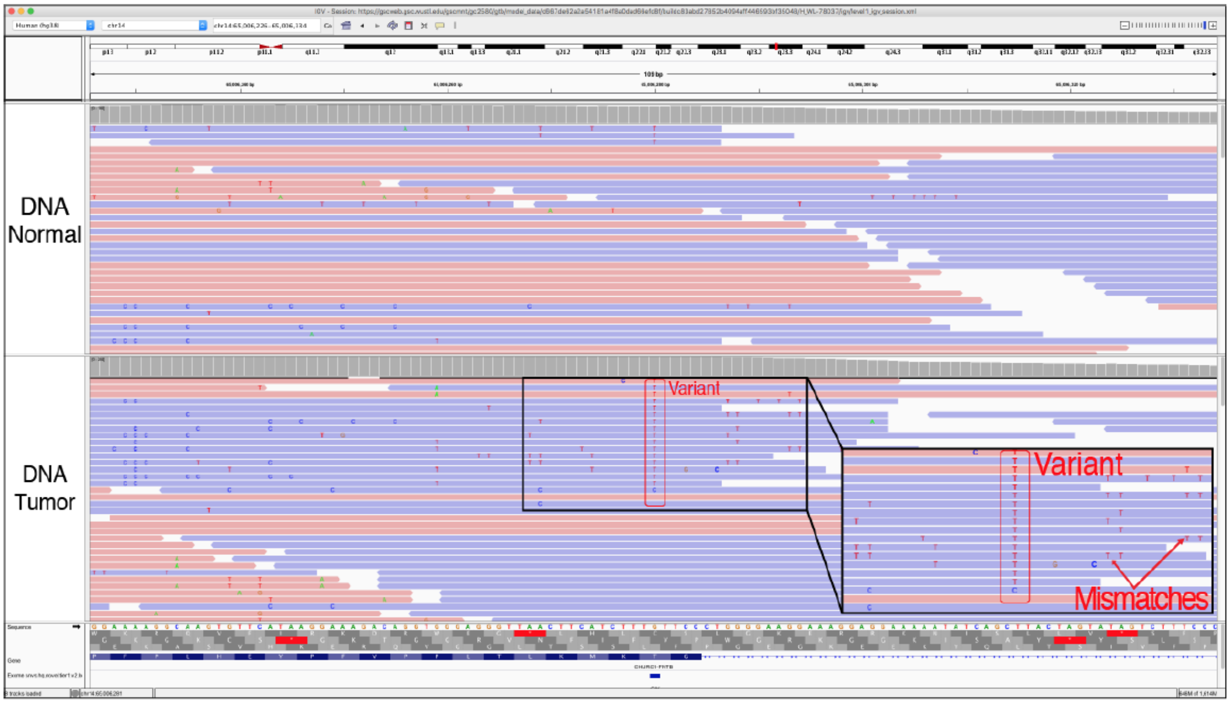
Helpful Hints:

1. The coverage track will be colored according to base when a variant is present at the default VAF. This threshold can be changed in the IGV preference panel:

   “View” > “Preferences” > “Alignments” > “Coverage allele-fraction threshold” > insert threshold

2. This can be particularly helpful for high depth samples and/or when low VAF (e.g., sub-clonal) variants are expected. With sufficient depth of coverage, the VAF threshold can be reduced.
3. Thresholds can be used to pre-filter variants with low tumor VAF to eliminate the need to evaluate these variants during manual review.

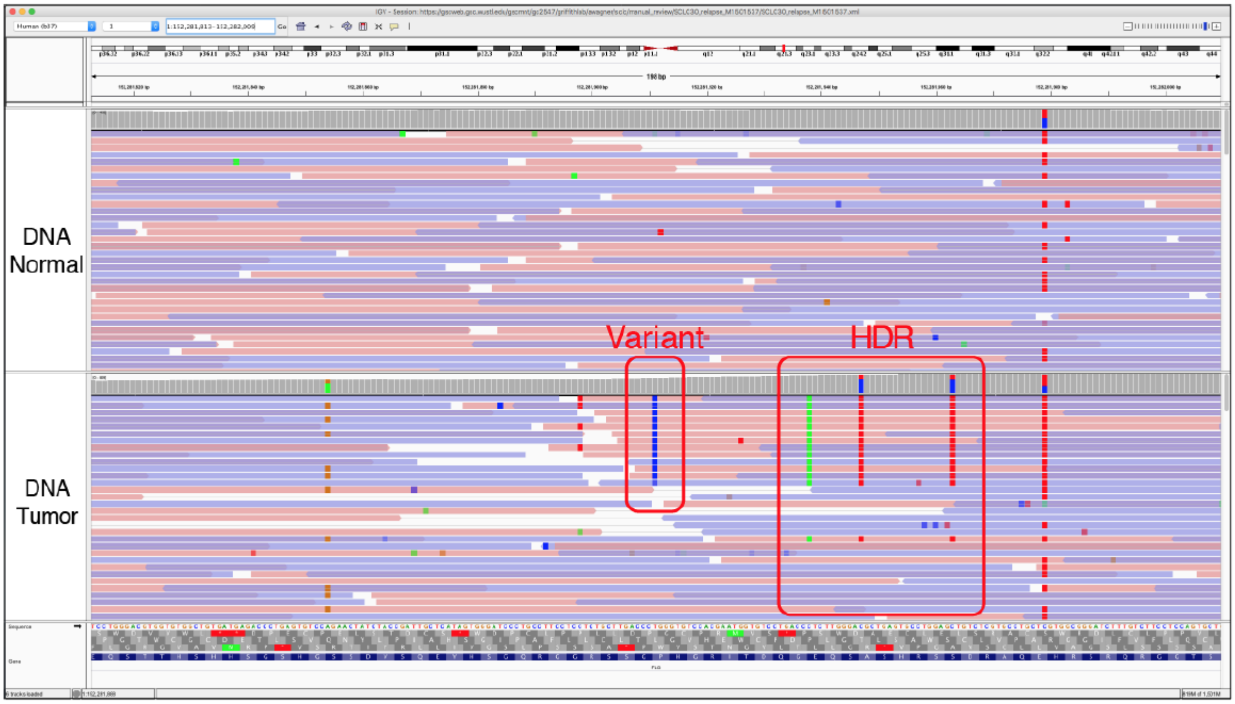
Figure S11. Example of Multiple Mismatches (MM). The Multiple Mismatches tag is used when the reads that contain the variant have other mismatched base pairs, which reduces the confidence in the read quality. Specifically, given a high error rate and a random distribution of errors, spurious variants can occur when the errors align across reads in the tumor sample but not in the normal sample. The MM and HDR tags are similar, in that both relate to mismatches in reads containing the variant; however, the MM tag is used when multiple mismatches are distributed unevenly (see **Appendix 3 - Figure S12**).



Helpful Hints:

1. If the mismatches are of high quality, this likely indicates that the read was properly sequenced. In this case, the mismatches occur due to misalignment. If the mismatches are of low quality, this likely indicates that the read was improperly sequenced. Both of these examples reduce confidence in the variant.
2. High densities of mismatches in the tumor track increase the probability that identical base substitution errors align across reads causing the VAF to surpass filtering thresholds. The higher the read depth, the less likely this situation is to arise, as low percentage VAF variants increase in plausibility with increased read depth.

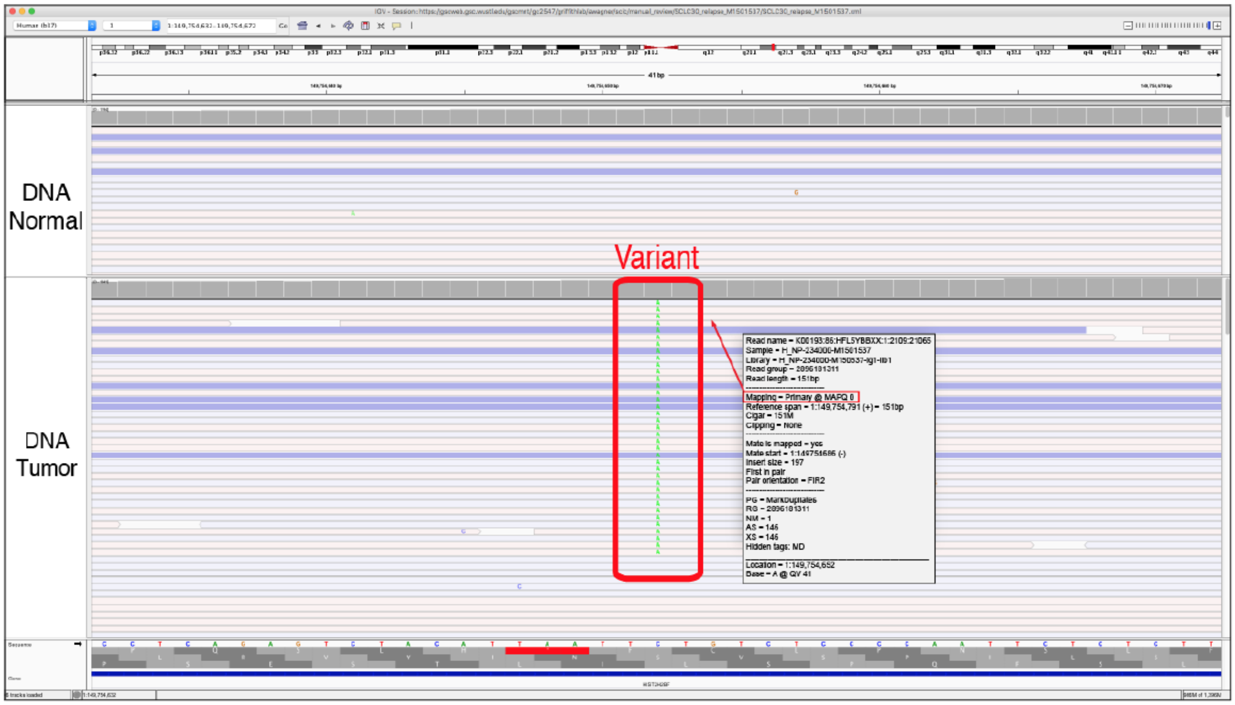
Figure S12. Example of High Discrepancy Region (HDR). The High Discrepancy Region tag is used when most reads containing the variant also contain other mismatches at the same locus. Typically, HDRs are observed when reads map to incorrect but homologous regions that contain localized differences, which are interpreted as variants. The HLA loci, duplicated loci, and other highly polymorphic regions are especially prone to this issue. These regions may require specialized alignment or assembly strategies for high quality variant calling.



Helpful Hints:

1. The presence of more than three identical mismatches within a 100-200 base-pair region is highly indicative of an HDR.
2. It is important to be sure that the variant being evaluated is not surrounded by a cluster of single nucleotide polymorphisms (SNPs). Sometimes, true variants can occur in close proximity to multiple SNPs and might be confused with an area of HDR. This is particularly true for individuals with haplotypes that are not well-represented by the reference sequence.

Figure S13. Example of Low Mapping (LM) quality. The Low Mapping tag is used to indicate variants that are mostly supported by reads that have low mapping quality. When reads are colored by readstrand, translucent/transparent reads indicate lower mapping quality and opaque reads indicate higher mapping quality. Mapping quality refers to a measure of confidence or probability that a read has been correctly aligned to the reference genome. Variants that are supported primarily or solely by low mapping quality reads are considered suspect.



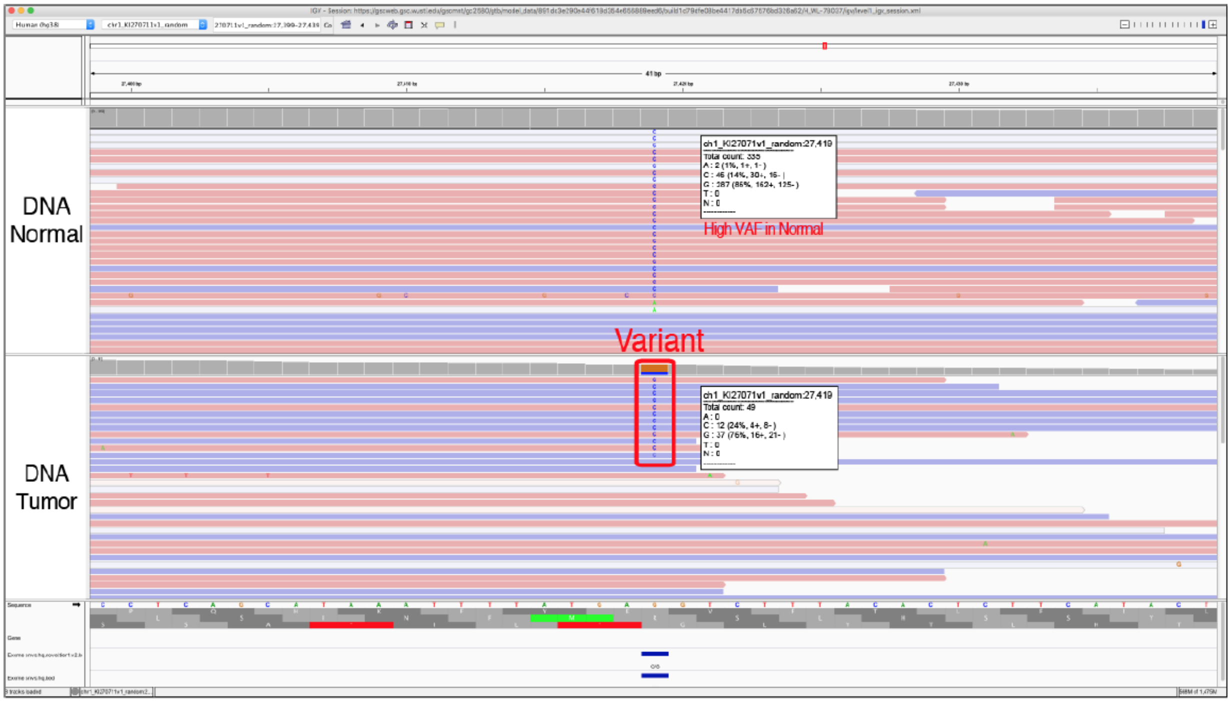
```
Read name = K00193:86:HFL5YBBXX:1:2109:21065
Sample = H_NP-234000-M1501537
Library = H_NP-234000-M150537-lg1-lib1
Read group = 2896181311
Read length = 151bp
----------------------------
Mapping = Primary @ MAPQ 0
Reference span = 1:149,754,791 (+) = 151bp
Cigar = 151M
Clipping = None
----------------------------
Mate is mapped = yes
Mate start = 1:149754686 (-)
Insert size = 197
First in pair
Pair orientation = F1R2
----------------------------
PG = MarkDuplicates
RG = 2896181311
NM = 1
AS = 146
XS = 146
Hidden tags: MD
________________________________
Location = 1:149,754,652
Base = A @ QV 41
```

Helpful Hints:

1. Mapping quality scores can be ascertained by clicking on the read.
2. In regions where numerous reads have a mapping quality of 0, the reads are often mapped to multiple locations across the genome. This results in low mapping quality reads in both the normal and tumor tracks. Alternate mapping locations can be ascertained by clicking on the read.
3. By default, all reads are shown in IGV, even if the mapping quality is 0. This threshold can be adjusted to eliminate low quality reads from IGV during manual review:

“View” > “Preferences” > “Alignments” > “Mapping quality threshold” > insert threshold

Figure S14. Example of Tumor in Normal (TN). The Tumor in Normal tag is used to indicate that the variant has reads of support in the normal track. This is a common occurrence in certain blood tumors (e.g., leukemia) as well as tumors that are highly metastatic. In some instances, TN might be a reason to fail the variant, whereas in other situation it can be used to denote ambiguity in the manual review call.



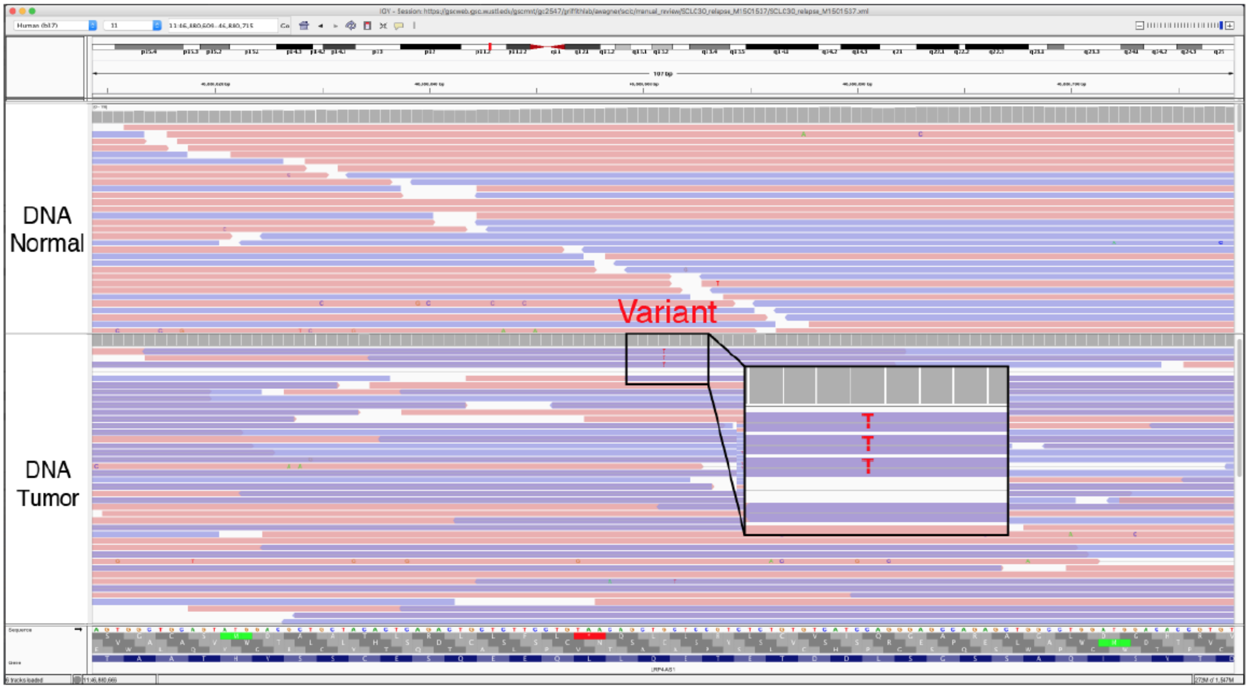
```
ch1_KI27071v1_random:27,419
Total count: 335
A : 2 (1%, 1+, 1- )
C : 46 (14%, 30+, 16- )
G : 287 (86%, 162+, 125- )
T : 0
N : 0
```

```
ch1_KI27071v1_random:27,419
Total count: 49
A : 0
C : 12 (24%, 4+, 8- )
G : 37 (76%, 16+, 21- )
T : 0
N : 0
```

Helpful Hints:

1. TN does not occur in all hematopoietic tumors but is likely when tumor cells are circulating in the bloodstream (e.g., acute myeloid leukemias with high blast counts).
2. Tumors that are metastatic may have tumor cells circulating in the bloodstream and thus can also have TN contamination.
3. Problems with sample barcoding (indexing) or cross contamination of samples can also lead to apparent support for a somatic variant in the normal.
4. Evaluating other normal samples from your cohort, or evaluating multiple variants within the same sample/experiment, can help set a relative acceptable TN threshold. This will help to differentiate sequencing and pipeline artifacts from tumor contamination of normal tracks.
5. Variants created by sequencing or alignment artifacts will also often occur in both the tumor and the normal tracks and can be labeled with TN.

Figure S15. Example of Short Inserts (SI) and Short Inserts Only (SIO). The Short Inserts tag is used when the variant is found on small nucleic acid fragments whereby sequencing from each end results in overlapping reads. In IGV, this is indicated as a grey bar through the middle of reads when reads are viewed as pairs. Variants supported by read pairs produced from these short fragments can result in the appearance of two independent reads supporting a variant when in reality, they represent only a single nucleic acid molecule. The SI tag is used when support for the called variant is primarily from short-insert read pairs but other read strands that are not short inserts also show variant support. The SIO tag is used when support for the called variant is exclusively present in paired reads from short inserts. This issue is prevalent in data derived from archival material (FFPE samples) or other source material with small/degraded DNA fragments (e.g., cell-free DNA).



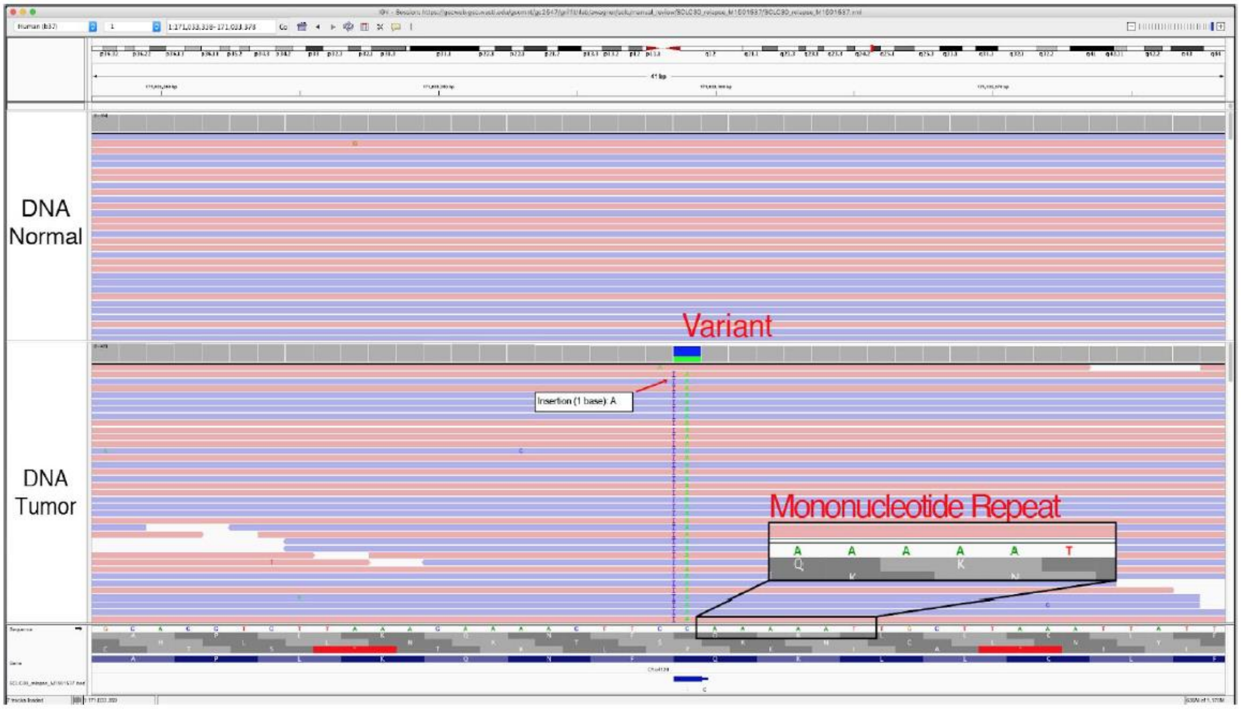
Helpful Hints:

1. To visualize short insert variants, you must view the tracks as pairs. Regions where the paired reads overlap will be dark purple and contain a horizontal grey line. At the ends, where there is no overlap, reads will remain blue or pink. Reads can be viewed as pairs using IGV commands:

   right click each data track > "View as pairs"

2. When viewing reads as pairs, short inserts can be observed; however, it will also overlay reads to reduce the total information available to the reviewer.
3. Short inserts are generally observed at lower variant frequencies and present in two or three read pairs (i.e., four to six reads in total).

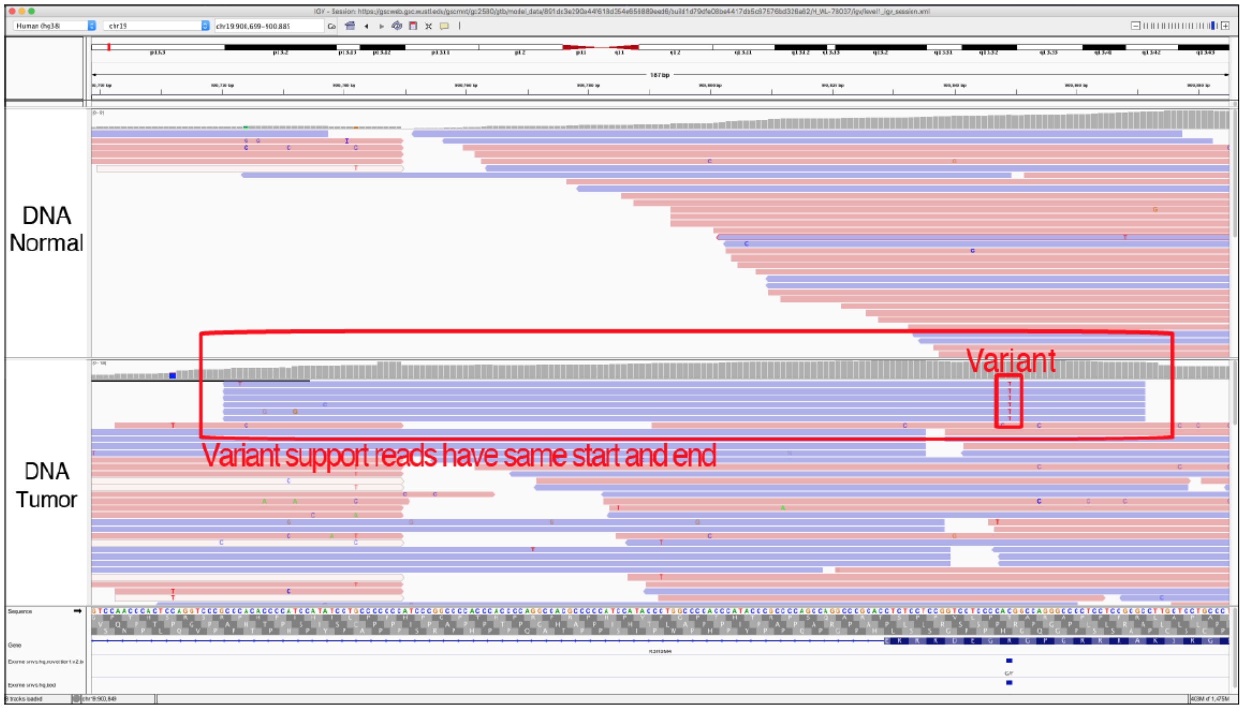
Figure S16. Example of Adjacent Indel (AI). The Adjacent Indel tag is used when a somatic variant was possibly caused by misalignment around a germline or somatic insertion/deletion (indel). In this example, it is likely that a real somatic variant is present, however, the variant is neither a simple 'A' insertion, nor a simple 'A' substitution. It is possible that the true variant is an ‘AA’ insertion that was miscalled by the automated somatic variant callers.



Helpful Hints:

1. To effectively visualize this pattern, it is necessary to zoom out using the IGV Genome Ruler.
2. It is important to evaluate the Genome Features section to visualize possible tandem repeats that might be implicated in the misalignment.
3. These cases can sometimes be resolved by correcting the nature of one or more called variants rather than failing the variant entirely. This is an instance where the IGV Notes section would be valuable.
4. This phenomenon is common with larger deletions where ends of reads will be misaligned within the deletion.

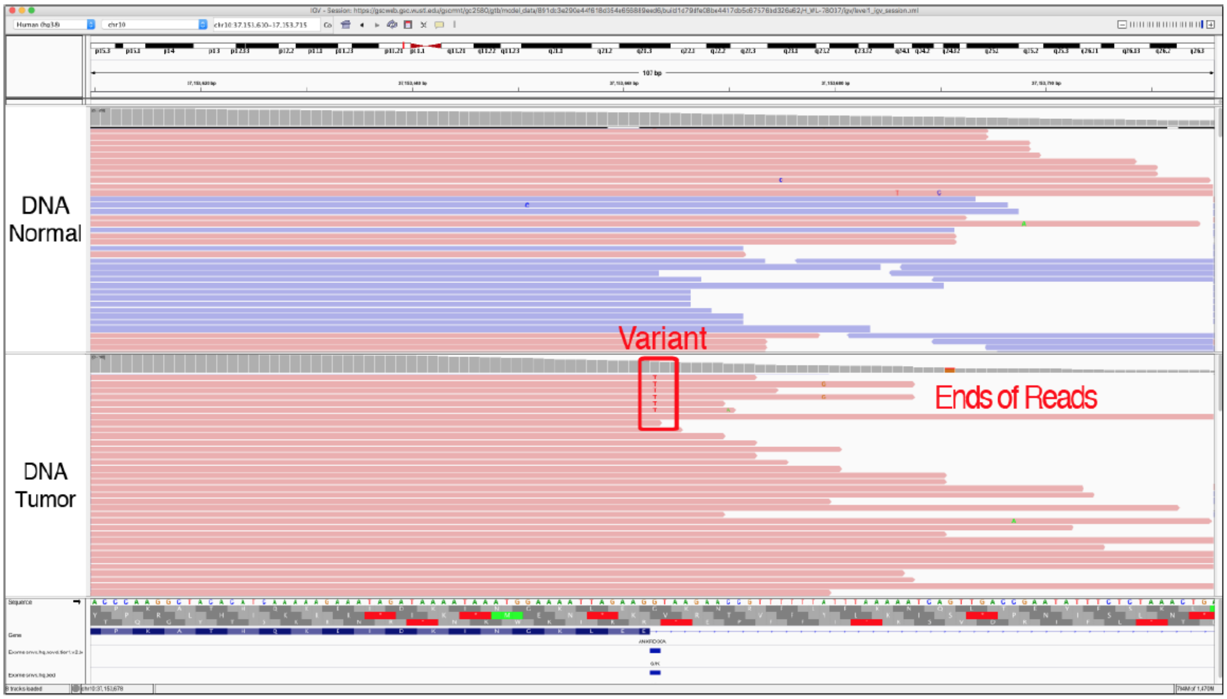
Figure S17. Example of Same Start/End (SSE). The Same Start/End tag is used when the variant is only contained by reads that start and stop at the same genomic loci. This is typically attributed to a variant called in multiple reads created from the same originating molecule during the library amplification process but erroneously not removed during read deduplication.



Helpful Hints:

1. Identifying SSE artifacts requires first sorting the reads by base and subsequently zooming out to view a larger genomic region. This allows for visualization of the ends of the reads.

Figure S18. Example of End of reads (E). The End of reads tag is used when the variant called is within 30 base pairs of the end of the variant-supporting reads. At read ends (especially the 3' end), there is an increased rate of error generation that can cause appearance of an erroneous variant.

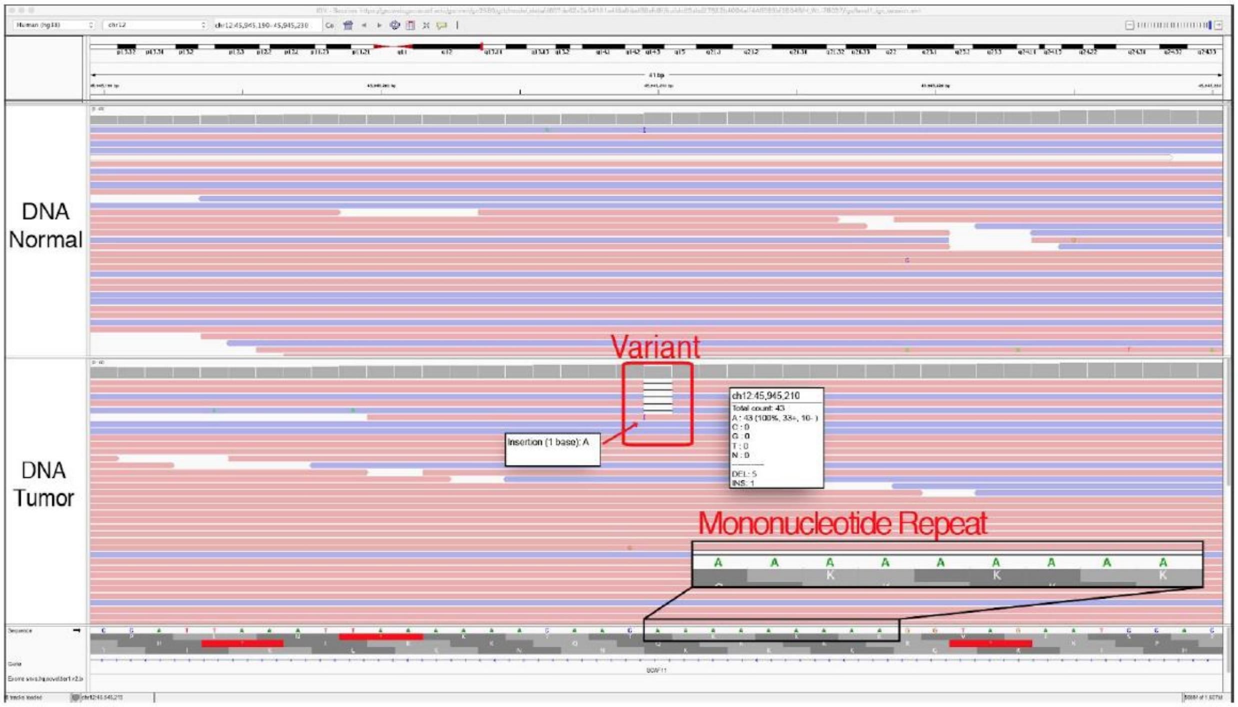


Helpful Hints:

1. Identifying End of reads artifacts requires first sorting the reads by base and subsequently zooming out to view a larger genomic region. This allows for visualization of the ends of the reads.
2. Additional mismatches downstream the called variant can increase confidence that the variant in question is a sequencing artifact.
3. This artifact is more easily evaluated by coloring the alignments by read strand:

Right click on data track > "Color alignments by" > "read strand"

Figure S19. Example of Mononucleotide repeat (MN). The Mononucleotide tag is used when a variant is called in proximity to a region of the reference sequence that contains a single nucleotide repeat (e.g., AAAAAAA...). In this instance, the called variant is most likely caused by misalignment of the reads to the reference genome. Some sequencers, particularly those dependent on the polymerase, are prone to making mistakes in repeat regions. However, it is important to note that mononucleotide repeats are also a common source of real human variation (inherited germline, de novo germline, or somatic) that arise due to errors produced by polymerase during DNA replication. Other factors, such as the size of the repeat, the VAF, or appearance in the normal, should be considered during manual review to confidently call the variant. The frequency in other samples processed in the same way (capture reagent, alignment algorithm, etc.) can be helpful in identifying common artifacts. Special alignment, assembly, or even additional sequencing technologies may be needed to validate short repeats of this nature.



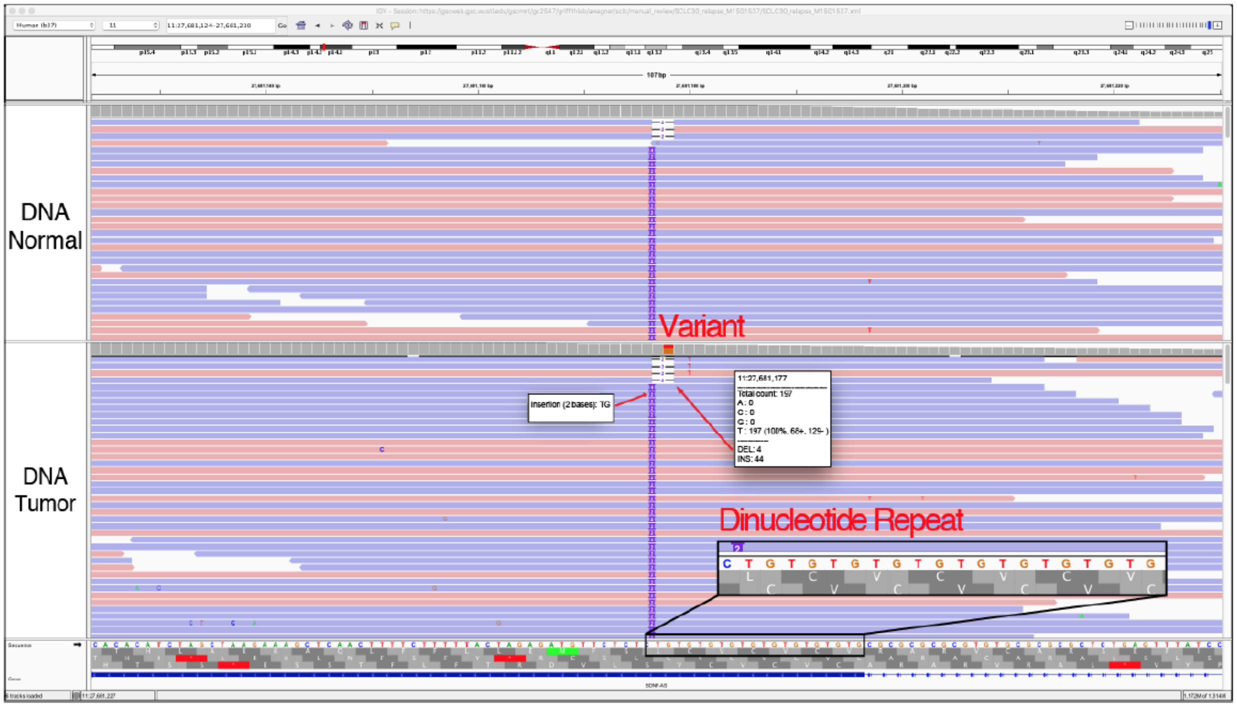
Insertion (1 base): A

ch12:45,945,210
Total count: 43
A : 43 (100%, 33+, 10-)
C : 0
G : 0
T : 0
N : 0
DEL: 5
INS: 1

Helpful Hints:

1. Typically, these variants are small deletions or insertions, and they are usually visualized in both the tumor and normal tracks.
2. Although the variant being evaluated here is a one base-pair deletion, other reads at the same locus typically have insertions and deletions of varying lengths.

Figure S20. Example of Dinucleotide repeat (DN). The Dinucleotide repeat tag is used when a variant is called in proximity to a region of the reference sequence that contains two alternating nucleotides (e.g., TGTGTG...). In this instance, the called variant is most likely caused by misalignment of the reads to the reference genome. Some sequencers, particularly those dependent on the polymerase, are prone to making mistakes in repeat regions. However, it is important to note that dinucleotide repeats are also a common source of normal human variation (inherited germline, de novo germline, or somatic) that arise due to errors produced by polymerase during DNA replication. Other factors, such as the size of the repeat, the VAF, or appearance in the normal, should be considered during manual review to confidently call the variant. The frequency in other samples processed in the same way (capture reagent, alignment algorithm, etc.) can be helpful in identifying common artifacts. Special alignment, assembly, or even additional sequencing technologies may be needed to validate short repeats of this nature.



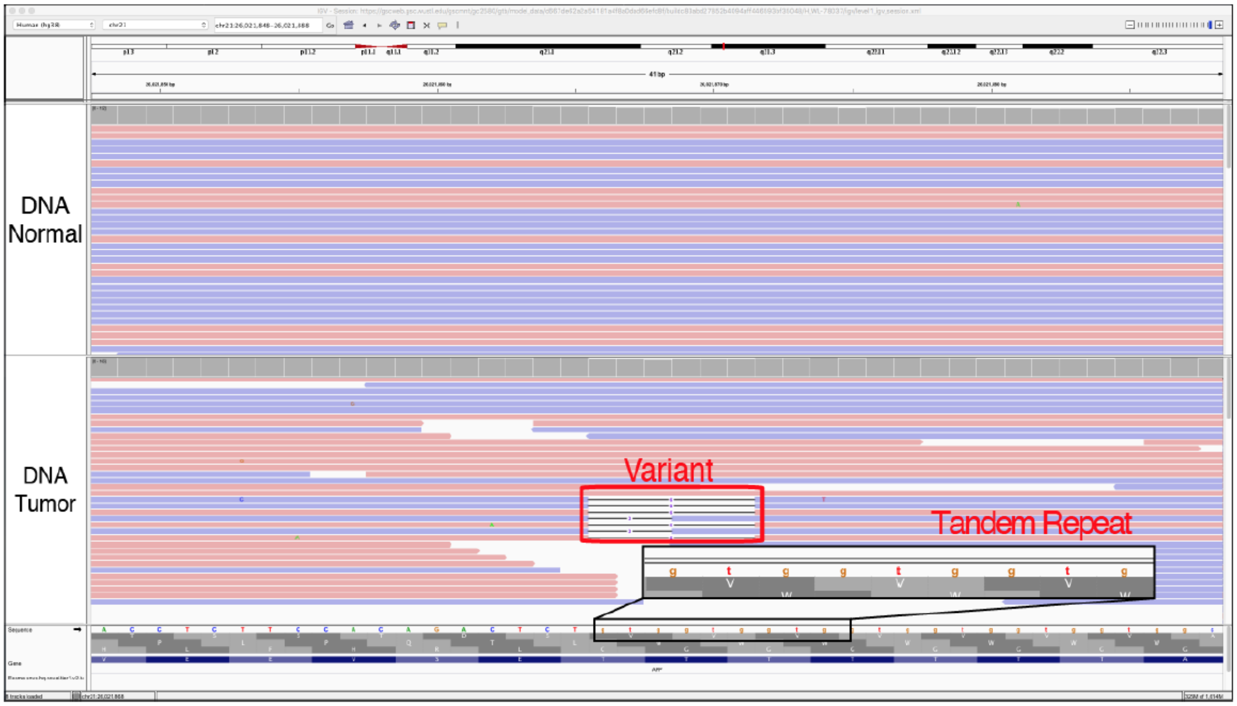
Insertion (2 bases): TG

```
11:27,681,177
-----------------------------
Total count: 197
A: 0
C: 0
G: 0
T: 197 (100%, 68+, 129- )
----------
DEL: 4
INS: 44
```

Helpful Hints:

1. Typically, these variants are small deletions or insertions, and they are usually visualized in both tumor and normal tracks.
2. Although the variant being evaluated is a two base-pair deletion, other reads at the same locus typically have insertions and deletions in multiples of two.

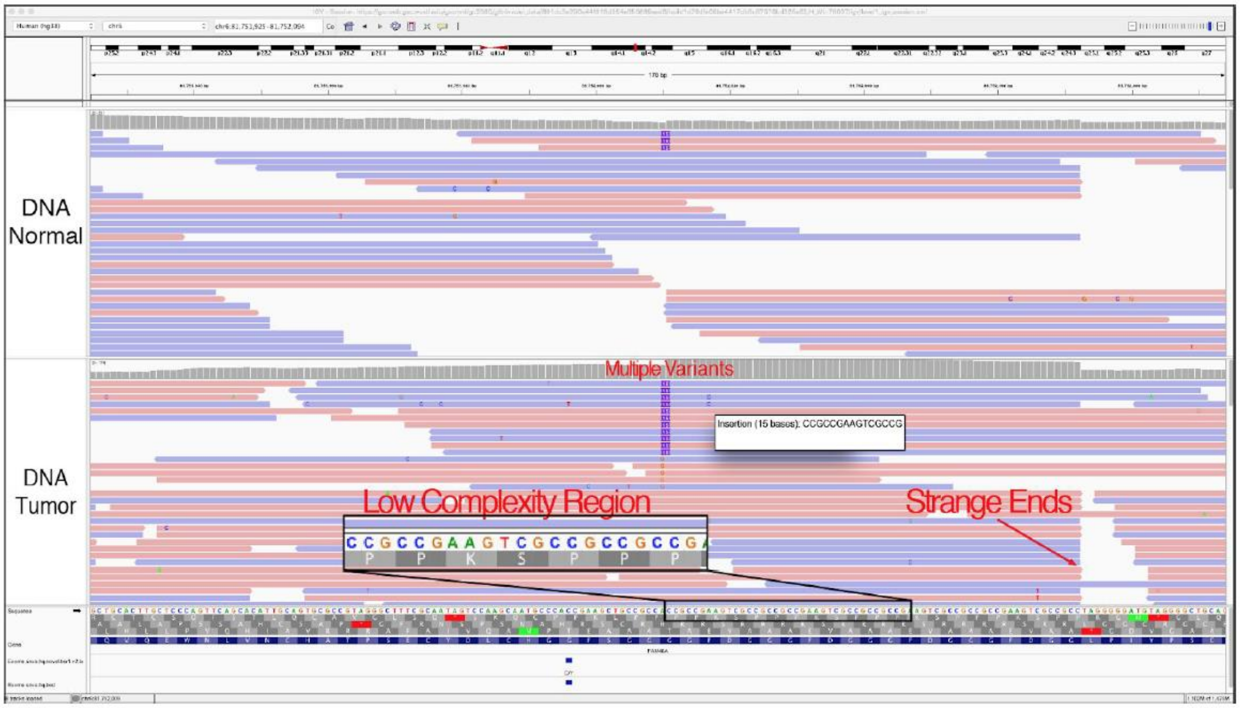
Figure S21. Example of Tandem Repeat (TR). The Tandem Repeat tag is used when a variant is called in proximity to a region of the reference sequence that contains some number of repeated nucleotides (e.g., GTGGTGGTG...). In this instance, the called variant is most likely caused by misalignment of the reads to the reference genome. Some sequencers, particularly those dependent on a polymerase, are prone to making mistakes in repeat regions. However, it is important to note that tandem repeats are also a common source of normal human variation (inherited germline, de novo germline, or somatic) that arise because of errors produced by polymerase during DNA replication. Other factors, such as the size of the repeat, the VAF, or appearance in the normal, should be considered during manual review to confidently call the variant. The frequency in other samples processed in the same way (capture reagent, alignment algorithm, etc.) can be helpful in identifying common artifacts. Special alignment, assembly, or even additional sequencing technologies may be needed to validate short repeats of this nature.



Helpful Hints:

1. Typically, these variants are small deletions or small insertions, and they are usually visualized in both the tumor tracks and the normal tracks.
2. In this example, the variant being evaluated is a three base-pair deletion, whereas other reads at the same locus have insertions and deletions in multiples of three, which reduces confidence in the called variant. This pattern can help distinguish a TR artifact from a true somatic variant.

Figure S22. Example of Ambiguous Other (AO). The Ambiguous Other tag is used to define a variant surrounded by inconclusive genomic features that cannot be explained by the other tags. In this example, we observe a low complexity region (e.g., genomic regions with increased A/T or G/C content), which can accurately be described with the AO tag.



Helpful Hints:

1. If the Ambiguous Other tag is used, it is highly recommended to include a short description in the notes section.

Appendix 4. Chapter 3 Supplementary Tables

Table S1. Values for inter-reviewer correlation matrix.

Example Call #1	Example Call #2	Score
S	S	1
A	A	1
F	F	1
G	G	1
F	G	0.5
F	A	0.5
A	S	0.5
G	A	0.5
S	F	0
G	S	0

Appendix 5. Chapter 4 Supplementary Methods

Determining eligible CIViC variants for smMIP capture

Filtering based on the Variant Evidence Score

All variants within the CIViC database are built on evidence statements that have been manually curated from the medical literature. Given that variants within the CIViC database have diverse quantity and quality of evidence support, the variant evidence score was developed to calculate the relative abundance of total available curated evidence for each variant. The variant evidence score reflects: 1) the strength of the evidence that was curated and 2) the total amount of curation that has been completed for each variant. To determine evidence strength, the Evidence Level Score and the Trust Rating Score were calculated. The Evidence Level Score is a 10-point scale that weighs the evidence strength based on category. Broadly, highest points are awarded to large clinical studies and lower points are awarded to case studies, *in vitro* studies, and inferential evidence. The Trust Rating Score is a 5-star scale that reflects the curator's confidence in the quality of the study. To determine the total level of curation for each variant, Evidence Level Scores were multiplied by Trust Rating Scores and summed across all Evidence Items. This final value (i.e., the CIViC Variant Evidence Score) was incorporated into the CIViC database and is now available for all variants in the CIVIC web interface, regular data releases, and application programming interface (API). Using the CIViC Variant Evidence Score, variants within the top 10% of total curation (corresponding to variant evidence score > 20 points) were selected to develop the CIViC smMIPs capture panel and were eligible for smMIP targeting. Of note, the CIViC Variant Evidence Score evaluates the total level of curation within the database and does not reflect the community consensus of clinical relevance.

Filtering based on the Sequence Ontology Identification Number

Variants were also filtered to only include variants that could be analyzed using a DNA-based sequencing platform. This required use of curated Sequence Ontology IDs (SOIDs). Within CIViC, SOIDs are manually classified as either: "DNA-based", "RNA-based", and/or "Protein-based". For example, variants with the Variant Type of "missense_variant" would be labeled as "DNA-based," whereas variants with the Variant Type of "transcript_variant" would be labeled as "RNA-based". Variants that had a "DNA-based" SOID were considered eligible for smMIP targeting and variants whose SOIDs were "RNA-based" and/or "Protein-based" were ineligible.

Categorization of variants based on length

Using CIViC curated coordinates, variant length was determined (i.e., variant start position minus variant stop position). This difference inferred the total number of smMIPs probes required to adequately assess each variant.

Hotspot targeting

If the variant length was <250 base pairs, the variant was eligible for hotspot targeting. For variants that required hotspot targeting, smMIPs probes were designed for the genomic region indicated in the CIViC database.

Sparse exon tiling and full exon tiling

If the variant was >250 base pairs, the variant required some or total tiling of the protein coding exons. For all variants that required sparse exon tiling or full exon tiling, the representative transcript from the CIViC database was used to obtain all possible exons associated with each Ensembl gene. The Ensembl gene was used to obtain all possible exons (biomart="ENSEMBL_MART_ENSEMBL", host="grch37.ensembl.org", dataset="hsapiens_gene_ensembl"). Exons were further filtered by Biotype to remove untranslated regions. Some large-scale copy number variants (i.e., “AMPLIFICATION”, “LOSS”, “DELETION”), were eligible for sparse tiling, wherein 10 probes distributed across the exons of the gene were retained to enable assessment of copy number state. Other variant types such as “MUTATION”, or “FRAMESHIFT MUTATION”, etc., required tiling of all protein coding exons. For variants that required full exon tiling, overlapping smMIPs (i.e., at least one basepair of overlap) were designed to tile across all protein coding exons in the gene that encompassed the variant. For variants that required sparse exon tiling, approximately 10 smMIPs were designed to cover a portion of the transcript.

smMIP sequencing and data analytics

Sequencing library construction and balancing of the probe pool were performed as described previously[121], and sequencing was performed using an Illumina NextSeq 500. Probes were excluded from the final reagent if they demonstrated poor hybridization to target sequence during initial quality checks.

Sequence data analysis was performed as previously described[121] with three enhancements. First, consensus reads were generated using the fgbiotools (http://fulcrumgenomics.github.io/fgbio/) CallMolecularConsensusReads utility with parameters “--error-rate-post-umi=30 --min-reads=2 --min-input-base-quality=20”. Second, a custom variant caller was utilized to identify all consensus calls at a site having at least 2 supporting reads with a minimum specified mapping quality (mapping quality score > 0). Third, variants were required to be detected on at least four DNA strands (at least 2 positive and at least 2 negative) in order to be considered real, rather than post-biological artifacts.[174] Collectively, these provisions require that at least two reads are derived from a common unique molecular identifier (UMID) to create a consensus read and that multiple consensus reads in both directions support the apparent variant. This helps to exclude pre-analytic artifacts reflecting DNA damage and stochastic errors that occur during library construction and sequencing. DNA input ranged from 100-500 ng across samples, however, any sample with an overlapping variant that had a VAF <5% utilized 500 ng to increase the number of template molecules interrogated.

Orthogonal sequencing and data analytics

Orthogonal sequencing data from previously conducted whole exome or genome sequencing was used to validate the CIViC smMIPs capture design. Sequencing alignment and somatic variant calling for the AML31 sample was performed according to *Griffith et al*.[22] Briefly, reads were aligned to GRCh37 using BWA v0.5.9[175] and variants were called using one of seven variant callers listed in the manuscript. Sequencing data from the SCLC cases, OSCC cases, and HL cases were analyzed using the Genome Modeling System[5] at the McDonnell Genome Institute. Reads from these studies were aligned to the reference genome (hg19/GRCh37 or hg38/GRCh38) using BWA-MEM v0.7.10[77] and duplicates were marked by Picard[78] and/or SAMBLASTER v0.1.22.[176]. For the SCLC cases, Single nucleotide variants (SNVs) were called using SomaticSniper[88,177], VarScan[19], and Strelka[17]; small insertions and deletions (indels) were called using GATK[16], Pindel[61], VarScan2[178], and Strelka. For OSCC cases, SNVs were detected using SomaticSniper v1.0.4, VarScan2 v2.3.6, Strelka v1.0.11, SAMtools r982[60], and Mutect v1.1.4[18]. Small indels were detected by GATK v5336[179], VarScan2, Strelka, and Mutect. For HL cases, SNVs were called using the intersection of SomaticSniper v1.0.4, VarScan v2.3.6, Strelka v1.0.11, and Mutect v1.1.4, and indels were called using GATK, Pindel v0.5, VarScan v2.3.6, and Strelka v1.0.11. For these three cohorts, variants identified by automated callers were subjected to heuristic filtering (removal of variants with low VAF [<5%] or low coverage [<20X in tumor or normal track]) and false positives were removed via manual somatic variant refinement.[58] If variant coordinates corresponded to GRCh38, their coordinates were converted to GRCh37 using LiftOver.[180] For the CRC cohort, sequencing, variant calling, and clinical annotation were performed according to methods highlighted in *Pritchard et a*l.[181] Briefly, sequencing was performed using Illumina next-generation sequencing (Illumina, San Diego, CA) and sequencing reads were aligned using BWA v0.6.1 and SAMtools v0.1.18. Indel realignment was then performed using GATK v1.6 and duplicate reads were removed using Picard v1.72. SNV and indel calling was performed using the GATK Universal Genotyper with default parameters and VarScan v2.3.2.

Assessment of variants missed using the CIViC smMIPs capture panel

Of the 65 variants identified on exome sequencing, all but 4 were also identified using CIViC smMIP sequencing. One variant was missed due to lack of adequate coverage, two variants were missed due to low performing probes, and one variant was retrospectively considered ineligible due to smMIPs design. The variant missed due to inadequate coverage was a *TP53* (p.G266R) variant identified in the AML31 tumor sample. Original sequencing indicated that this variant was present at 0.04% VAF, therefore, given smMIPs coverage of 2,388 reads at this site, there was only a 0.01% chance that this variant would have been detected (one-tailed probability of exactly, or greater than, 4 reads (K) out of 2,388 reads (n); p = 0.0046). However, this low-prevalence variant could have been recovered given additional sequence coverage. Additionally, there were two variants missed due to low MIP performance. The first variant that was missed (*chr10:g.89690805G>A* in the SCLC8 tumor sample at 94% VAF) was due to poor performance of the MIP covering the region of interest in the reverse direction. This MIP showed only 1 aligned read across all 36 samples and had no aligned reads in SCLC8. Despite the fact that there was extensive support from the forward MIP (95% VAF with 34 / 35 consensus reads), the requirement that both forward and reverse reads show support prevented this variant from being called. The second missed variant (*PTEN* - e8-1 in the SCLC4 tumor sample at 100% VAF) was due to low performance of MIPs in both directions. Even though both the forward and the reverse MIP showed variant support, the forward MIP only contained 2 consensus reads and the reverse MIP only contained 1 consensus read, preventing it from being called as somatic. The final variant (*chr17:g7577094C>T* in the CRC5 tumor sample at 32% VAF) was retrospectively considered ineligible because the original smMIPs developed to cover the eligible STK variant called for sparse tiling (i.e., identification of copy number change). As such, the variant was contained by a region that did not have full coverage in the forward direction. When evaluating the reverse MIP that contained this site, we observed a 34% VAF (402 / 1,184 reads), which was comparable to the original sequencing data. However, lack of a secondary probe designed against the complementary DNA strand prevented this variant from being called as somatic.

Appendix 6. Chapter 4 Supplementary Tables

Table S1. Variants and associated genes eligible for CIViC smMIPs design after filtering based on CIViC Variant Evidence Score and Sequence Ontology ID (SOID).

Gene	Variant
ABCB1	I1145I
ABL1	BCR-ABL F317L
AKT1	E17K
ALK	ALK FUSION I1171
ALK	F1174L
ALK	R1275Q
ASXL1	MUTATION
ATM	MUTATION
BAP1	MUTATION
BCL2L11	DELETION
BRAF	MUTATION
BRAF	V600
BRAF	V600D
BRAF	V600E
BRAF	V600K
BRCA1	LOSS-OF-FUNCTION
BRCA1	MUTATION
BRCA2	LOSS-OF-FUNCTION
BRCA2	MUTATION
BTK	C481S
CALR	EXON 9 FRAMESHIFT
CCND1	AMPLIFICATION
CCNE1	AMPLIFICATION
CDK4	AMPLIFICATION
CDKN2A	LOSS
CEBPA	INACTIVATION
CTNNB1	S45F
CTNNB1	S45P
DNMT3A	MUTATION
DNMT3A	R882
DPYD	DPYD*13 HOMOZYGOSITY
DPYD	DPYD*2A HOMOZYGOSITY
EGFR	AMPLIFICATION
EGFR	EXON 19 DELETION
EGFR	G719
EGFR	G719S
EGFR	L858R
EGFR	S492R
EGFR	T790M
ERBB2	AMPLIFICATION
ERCC2	K751Q
EZH2	MUTATION
EZH2	Y646
FCGR2A	H167R
FCGR3A	F212V
FGFR1	AMPLIFICATION
FGFR2	AMPLIFICATION
FLT3	D835
FLT3	ITD
FLT3	MUTATION
FLT3	TKD MUTATION
GNAS	T393C
HOXB13	G84E
IDH1	R132
IDH1	R132C
IDH2	MUTATION
IDH2	R140
IDH2	R172
IDH2	R172K
IKZF1	DELETION

Gene	Variant
JAK2	V617F
KIT	D816V
KIT	EXON 11 MUTATION
KIT	M541L
KRAS	A146T
KRAS	EXON 2 MUTATION
KRAS	RS61764370
MAP2K7	E116K
MET	AMPLIFICATION
MET	EXON 14 SKIPPING MUTATION
MTHFR	A222V
MTOR	MUTATION
MYCN	AMPLIFICATION
MYD88	L265P
NOTCH1	MUTATION
NOTCH1	P2514FS
NPM1	EXON 12 MUTATION
NRAS	MUTATION
NRAS	Q61
NT5C2	K359Q

Gene	Variant
PDGFRA	D842V
PIK3CA	AMPLIFICATION
PIK3CA	E542K
PIK3CA	E545K
PIK3CA	H1047R
PIK3CA	MUTATION
PTEN	DELETION
PTEN	LOSS
PTEN	MUTATION
RET	M918T
ROS1	CD74-ROS1 G2032R
SF3B1	MUTATION
SMAD4	MUTATION
SRSF2	MUTATION
STK11	LOSS
TERT	C228T
TERT	PROMOTER MUTATION
TET2	MUTATION
TP53	DNA BINDING DOMAIN MUTATION
TP53	P72R

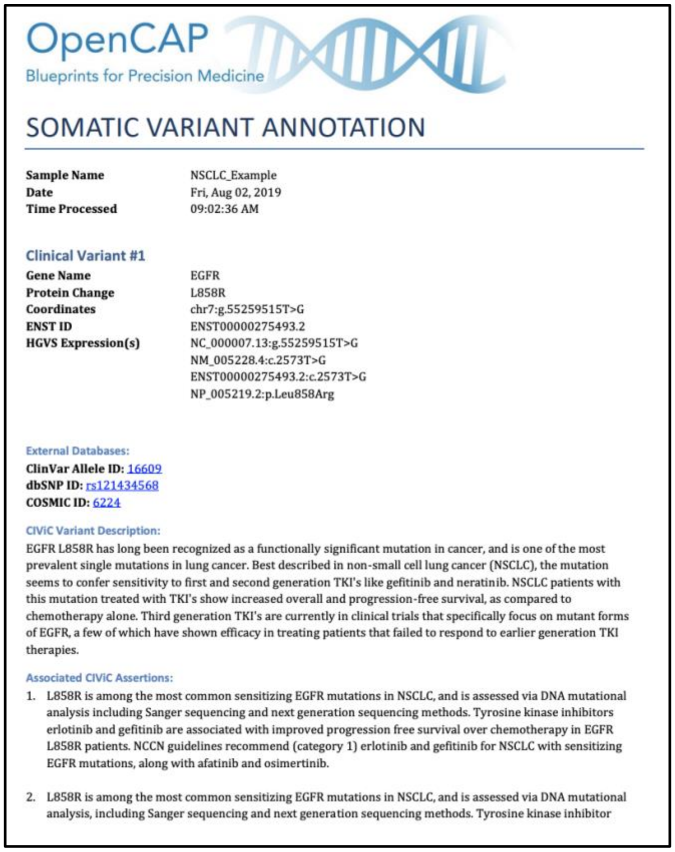
Gene	Variant
TP53	R273C
TP53	R273H
U2AF1	Q157P/R
U2AF1	S34Y/F
UGT1A1	UGT1A1*28
VHL	C162F (c.485G>T)
VHL	N131P (c.390
VHL	P86S (c.256C>T)
VHL	R167Q (c.500G>A)
WT1	EXON 7 MUTATION
XRCC1	Q399R

Table S2. Sequencing data availability for samples used in analysis.

Malignancy	**Abbreviation**	**Cases (n=22)**	**Publication status**	**Pubmed link**	**DBGaP**
Head and neck squamous cell carcinoma	HNSCC	5	In preparation	Not Available	phs001623
Small cell lung cancer	SCLC	9	Published	https://www.ncbi.nlm.nih.gov/pubmed/30224629	phs0001049
Hodgkin's lymphoma	HL	2	In preparation	Not Available	Not Available
Acute myeloid leukemia	AML	1	Published	https://www.ncbi.nlm.nih.gov/pubmed/26645048	phs000159
Colorectal Cancer	CRC	5	In preparation	Not Available	Not Available

Appendix 7. Chapter 4 Supplementary Figures

Figure S1. Exemplary OpenCAP interpretation report to highlight OpenCAP features. This report was generated using the OpenCAP annotation pipeline to showcase features of the software. In each report, the variant name, protein change, coordinates, ENST ID, and HGVS Expressions are shown. The report also links to external databases including ClinVar, dbSNP, and COSMIC. Finally, OpenCAP pulls data directly from the CIViC interface. Specifically, the report shows all CIViC Variant Descriptions, Associated CIViC Assertions, and Associated CIViC Evidence Items (EIDs). CIViC Evidence Items are only displayed if the EID is accepted and has an A- or B-level Evidence Item. If the EID is displayed, a link to the CIViC interface, and the supporting publication is displayed. Processing information, including total variants processed and those that had a clinical annotation, are also shown. This report was generated using variants associated with a non-small cell lung cancer described by Lee *et al*.[118] In this case, the patient had an EGFR - L858R variant and a KRAS - G12D variant.



OpenCAP
Blueprints for Precision Medicine

SOMATIC VARIANT ANNOTATION

Sample Name	NSCLC_Example
Date	Fri, Aug 02, 2019
Time Processed	09:02:36 AM

Clinical Variant #1

Gene Name	EGFR
Protein Change	L858R
Coordinates	chr7:g.55259515T>G
ENST ID	ENST00000275493.2
HGVS Expression(s)	NC_000007.13:g.55259515T>G NM_005228.4:c.2573T>G ENST00000275493.2:c.2573T>G NP_005219.2:p.Leu858Arg

External Databases:

ClinVar Allele ID: 16609
dbSNP ID: rs121434568
COSMIC ID: 6224

CIViC Variant Description:

EGFR L858R has long been recognized as a functionally significant mutation in cancer, and is one of the most prevalent single mutations in lung cancer. Best described in non-small cell lung cancer (NSCLC), the mutation seems to confer sensitivity to first and second generation TKI's like gefitinib and neratinib. NSCLC patients with this mutation treated with TKI's show increased overall and progression-free survival, as compared to chemotherapy alone. Third generation TKI's are currently in clinical trials that specifically focus on mutant forms of EGFR, a few of which have shown efficacy in treating patients that failed to respond to earlier generation TKI therapies.

Associated CIViC Assertions:

1. L858R is among the most common sensitizing EGFR mutations in NSCLC, and is assessed via DNA mutational analysis including Sanger sequencing and next generation sequencing methods. Tyrosine kinase inhibitors erlotinib and gefitinib are associated with improved progression free survival over chemotherapy in EGFR L858R patients. NCCN guidelines recommend (category 1) erlotinib and gefitinib for NSCLC with sensitizing EGFR mutations, along with afatinib and osimertinib.

2. L858R is among the most common sensitizing EGFR mutations in NSCLC, and is assessed via DNA mutational analysis, including Sanger sequencing and next generation sequencing methods. Tyrosine kinase inhibitor

Appendix 8. Chapter 5 Supplementary Figures

Figure S1. Distribution of variants identified in 346 MyeloSeq® reports. The number of variants identified in each of the 40 genes targeted by the MyeloSeq® panel, according to diagnosis.

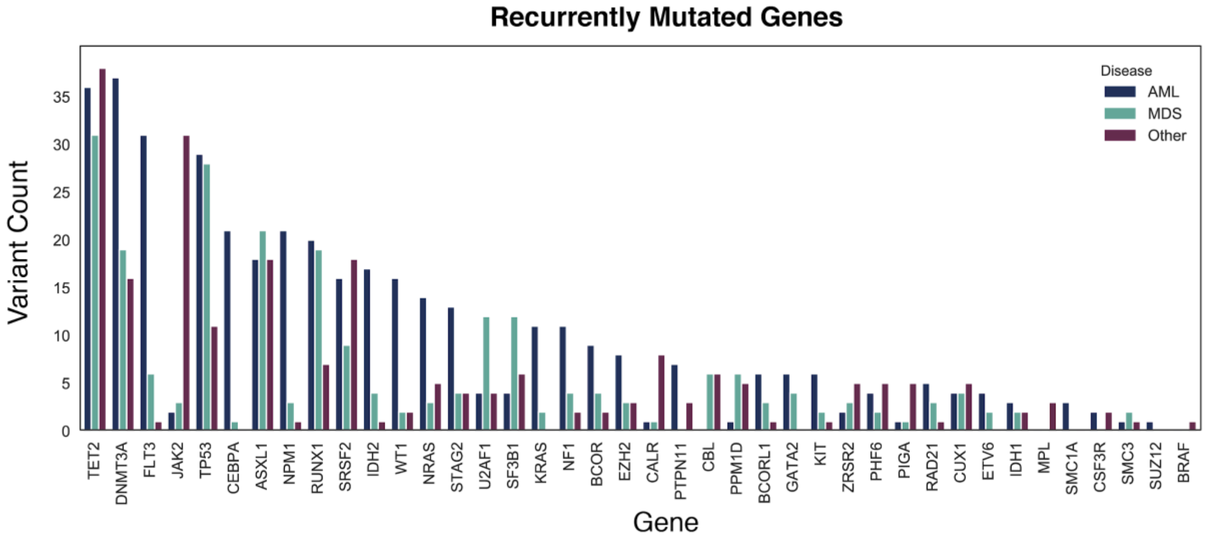
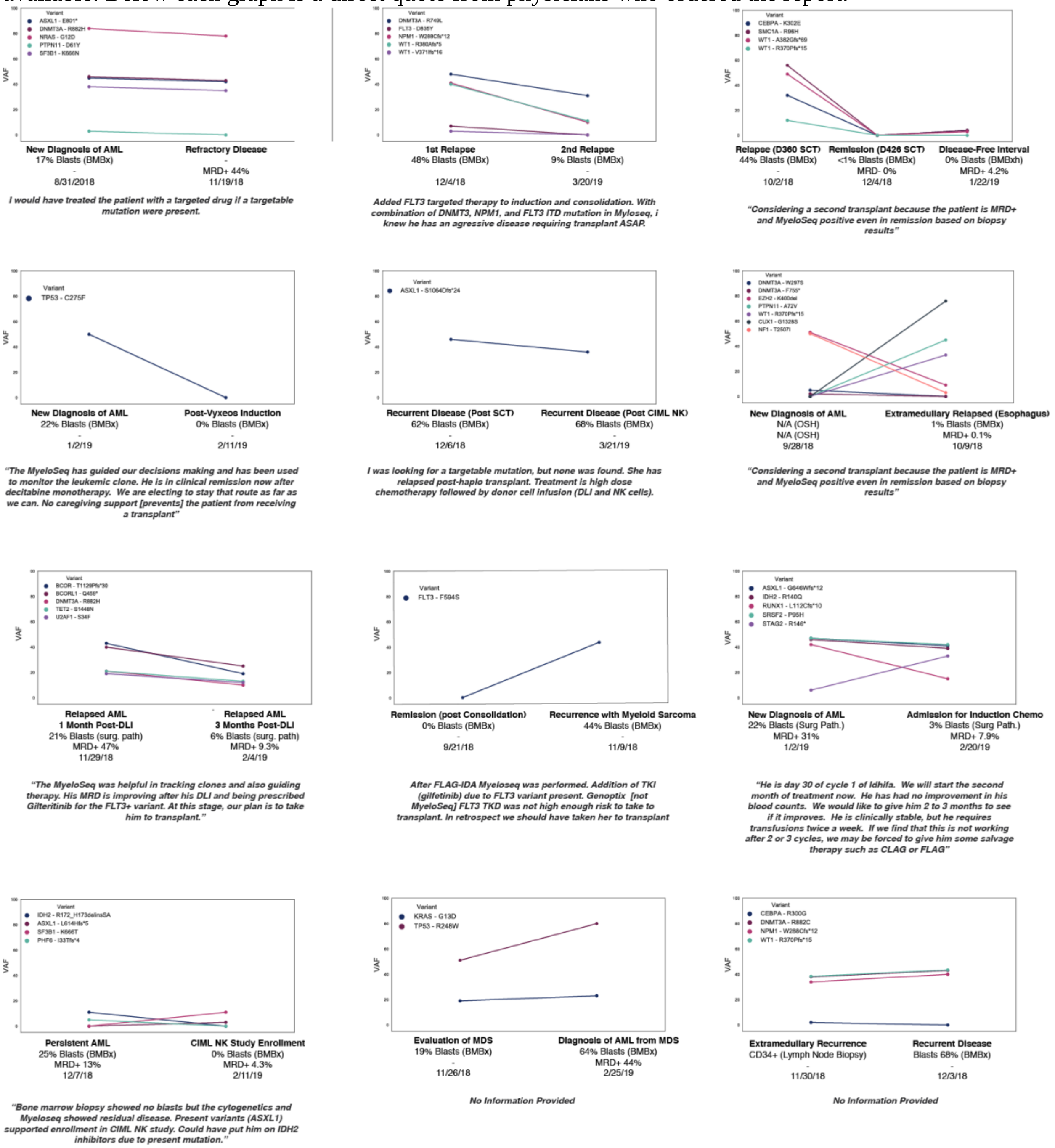


Figure S2. Evaluation of all AML patients that had multiple MyeloSeq® reports generated. Each panel represents a single patient where multiple MyeloSeq® panels were ordered. The plot indicates the variants observed with associated variant allele frequencies (VAFs). Each time point is labeled with associated bone marrow biopsy (BMBx) results and measurable residual disease (MRD) results, if available. Below each graph is a direct quote from physicians who ordered the report.



I would have treated the patient with a targeted drug if a targetable mutation were present.

Added FLT3 targeted therapy to induction and consolidation. With combination of DNMT3, NPM1, and FLT3 ITD mutation in Myloseq, i knew he has an agressive disease requiring transplant ASAP.

"Considering a second transplant because the patient is MRD+ and MyeloSeq positive even in remission based on biopsy results"

"The MyeloSeq has guided our decisions making and has been used to monitor the leukemic clone. He is in clinical remission now after decitabine monotherapy. We are electing to stay that route as far as we can. No caregiving support [prevents] the patient from receiving a transplant"

I was looking for a targetable mutation, but none was found. She has relapsed post-haplo transplant. Treatment is high dose chemotherapy followed by donor cell infusion (DLI and NK cells).

"Considering a second transplant because the patient is MRD+ and MyeloSeq positive even in remission based on biopsy results"

"The MyeloSeq was helpful in tracking clones and also guiding therapy. His MRD is improving after his DLI and being prescribed Gilteritinib for the FLT3+ variant. At this stage, our plan is to take him to transplant."

After FLAG-IDA Myeloseq was performed. Addition of TKI (gilfetinib) due to FLT3 variant present. Genoptix [not MyeloSeq] FLT3 TKD was not high enough risk to take to transplant. In retrospect we should have taken her to transplant

"He is day 30 of cycle 1 of Idhifa. We will start the second month of treatment now. He has had no improvement in his blood counts. We would like to give him 2 to 3 months to see if it improves. He is clinically stable, but he requires transfusions twice a week. If we find that this is not working after 2 or 3 cycles, we may be forced to give him some salvage therapy such as CLAG or FLAG"

"Bone marrow biopsy showed no blasts but the cytogenetics and Myeloseq showed residual disease. Present variants (ASXL1) supported enrollment in CIML NK study. Could have put him on IDH2 inhibitors due to present mutation."

No Information Provided

No Information Provided

Appendix 9. Chapter 5 Supplementary Tables

Table S1. Genes (and gene hotspots) targeted using the MyeloSeq® panel.

BRAF (V600E)	FLT3 (TKD and ITD)	JAK2 (V617, exon 12)	KIT (exons 2, 8-13, 17)
KRAS (G12, G13, Q61)	MPL (exon 10)	NF1	NRAS (G12, G13, Q61)
PTPN11 (exons 3, 13, 14)	GATA2	ASXL1	EZH2
SUZ12	CSF3R	WT1	RAD21
SMC1A	SMC3	STAG2	DNMT3A
IDH1 (R132	IDH2 (R140, R172)	TET2	CALR (exon 9)
CBL (exons 8, 9)	NPM1 (exon 11)	PIGA	PPM1D (exon 6)
CUX1	SF3B1	SRSF2 (exon 1)	U2AF1 (exons 2, 6)
ZRSR2	BCOR	BCORL1	CEBPA
ETV6	RUNX1	PHF6	TP53

Table S2. 16-question survey provided to presiding physicians to inquire about change in treatment protocol based on MyeloSeq® results.

	Question	Available Responses
1	What is the study number for the case?	[Free Text]
2	Please indicate if there is a reason why you are unable to complete a survey for the patient (e.g., lost to follow-up or Refused Treatment)	[Free Text]
3	Did you provide any of the following for the patient?	1) Hi-dose cytotoxic salvage chemotherapy 2) Targeted therapy 3) Clinical Trial Target Therapy 4) Transplant in Relapse 5) Other [Free Text]
4	Are you considering a potential transplant for this patient?	1) Yes 2) No
5	Do any of the following prevent the patient from being eligible for a transplant	1) N/A Considering a Transplant 2) Social Barriers 3) Too many comorbidities 3) No suitable donor 4) Clinical Trial would be better 5) Gene was targetable 6) Other [Free Text]
6	Did you change your therapeutic plan based on results from the MyeloSeq® assay?	1) Yes 2) No
7	Please describe how the MyeloSeq® changed your treatment plan.	[Free Text]
8	Did your patient accept your treatment plan recommendation?	1) Yes 2) No 3) Have not discuss with patient
9	If applicable, please provide any additional commentary on how the MyeloSeq® panel informed your treatment decision	[Free Text]

Table S3. Genes that recurrently failed sequencing based on a 50X coverage requirement across all protein coding exons contained by the gene.

Gene	Number of cases where coverage was <50X at some point in gene
WT1	212
CUX1	199
CEBPA	114
None	87
RUNX1	55
ZRSR2	45
SUZ12	36
MPL	20
NF1	18
TP53	16
PTPN11	11
BCOR	10
SF3B1	5
PIGA	5
PHF6	5
KIT	5
CALR	5
EZH2	4
NRAS	3
RAD21	3
SMC1A	2
SMC3	2
BCORL1	2
ASXL1	2

Table S4. Data describing the response rate for the 122 cases evaluated by the 14 physicians involved in the study. For each physician, the total number of patient cases (*Cases*), the total number of cases with a survey response (*Responses*), the total number of eligible surveys (*Eligible Surveys*), and the number of cases whereby the results from the MyeloSeq® panel changed the treatment plan (*Changed Plan*) are indicated.

Physician	Cases	Responses	Eligible Surveys	Changed Plan
A	7	7	7	0 (0%)
B	13	11	11	6 (54%)
C	10	10	10	9 (90%)
D	13	13	12	3 (25%)
E	8	8	8	4 (50%)
F	19	19	19	7 (37%)
G	2	2	2	2 (100%)
H	5	5	4	2 (50%)
I	1	1	1	0 (0%)
J	14	14	14	7 (50%)
K	8	8	8	4 (36%)
L	13	13	11	2 (20%)
M	8	8	6	4 (40%)
N	1	1	1	0 (0%)
Total	**121**	**119**	**113**	**50**