Technical Note

SCHOOL: Software for Clinical Health in Oncology for Omics Laboratories

Chelsea K. Raulerson1, Erika C. Villa1, Jeremy A. Mathews1, Benjamin Wakeland1, Yan Xu2, Jeffrey Gagan2, Brandi L. Cantarel1

1Bioinformatics Core Facility, University of Texas Southwestern Medical Center, Dallas, Texas, USA; 2Department of Pathology, University of Texas Southwestern Medical Center, Dallas, Texas, USA; 3Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, Texas, USA

Submitted: 18-March-2021 Revised: 01-September-2021 Accepted: 04-October-2021 Published: 05-January-2022

Abstract

Bioinformatics analysis is a key element in the development of in-house next-generation sequencing assays for tumor genetic profiling that can include both tumor DNA and RNA with comparisons to matched-normal DNA in select cases. Bioinformatics analysis encompasses a computationally heavy component that requires a high-performance computing component and an assay-dependent quality assessment, aggregation, and data cleaning component. Although there are free, open-source solutions and fee-for-use commercial services for the computationally heavy component, these solutions and services can lack the options commonly utilized in increasingly complex genomic assays. Additionally, the cost to purchase commercial solutions or implement and maintain open-source solutions can be out of reach for many small clinical laboratories. Here, we present Software for Clinical Health in Oncology for Omics Laboratories (SCHOOL), a collection of genomics analysis workflows that (i) can be easily installed on any platform; (ii) run on the cloud with a user-friendly interface; and (iii) include the detection of single nucleotide variants, insertions/deletions, copy number variants (CNVs), and translocations from RNA and DNA sequencing. These workflows contain elements for customization based on target panel and assay design, including somatic mutational analysis with a matched-normal, microsatellite stability analysis, and CNV analysis with a single nucleotide polymorphism backbone. All of the features of SCHOOL have been designed to run on any computer system, where software dependencies have been containerized. SCHOOL has been built into apps with workflows that can be run on a cloud platform such as DNANexus using their point-and-click graphical interface, which could be automated for high-throughput laboratories.

Keywords: Bioinformatics, cancer, NGS

INTRODUCTION

The rapid expansion and decreasing cost of next-generation sequencing (NGS) technology present an opportunity to improve the diagnosis and treatment of cancer through identifying tumor-specific mutations and enabling physicians to adapt treatment plans that suit the unique molecular profile of each patient.1,2 To address the evolving list of clinically actionable and prognostic biomarkers in the treatment of cancer, academic clinical laboratories have developed sequencing assays with varying size gene panels (100–1600 genes), with consistent quality to detect relevant genetic variants.

Bioinformatics analysis of sequence data includes two phases: (i) primary analysis, which converts the raw sequencing reads into predicted genetic variants and read abundances, and (ii) secondary analysis, which is customized for each clinical assay to maximize sensitivity and specificity by identifying artifact and poor-quality variant predictions. For a quick turn-around-time, the primary analysis is computationally demanding and requires computational resources with high memory (>32 GB) and multiple processors using a local high-performance cluster or cloud-computing resources. Furthermore, the primary analysis

Access this article online

Quick Response Code:  
Website: www.jpathinformatics.org  
DOI: 10.4103/jpi.jpi_20_21


This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: reprints@medknow.com

© 2022 Journal of Pathology Informatics | Published by Wolters Kluwer - Medknow
requires multiple elements for complete somatic variant detection including single nucleotide variants (SNVs), insertions/deletions (indels), copy number variants (CNVs), and structural variants such as translocations, large deletions, internal tandem duplications, and differences in microsatellite length. Because each assay might include different elements for detecting these different variant types, the primary analysis should be customizable for a variety of assays. Secondary analysis is much less resource-intensive, can be run on a desktop computer, and should be tailored to the needs of the specific assay.

Both commercial and open-source solutions have been introduced to address primary analysis needs in cancer genomics. Commercially developed bioinformatics pipelines are proprietary, often have limited options for customization, and require licensing, which can increase computational costs. By contrast, open-source solutions are usually customizable and lack licensing costs. However, common best-practice tools for variant detection, such as BWA and GATK4, require computational programming expertise to run in a Linux command line environment. Some commonly used open-source tools for more complex variant detection lack thorough documentation, continued support for development, or the flexibility to process varied data types (tumor-only samples versus matched tumor/normal control). Additionally, many existing software tools are difficult to install and maintain, due to the sometimes difficult installation of software dependencies, which make them sensitive to updates and changes to default programs. Finally, these tools are not natively packaged as an end-to-end analysis pipeline, which starts with raw sequencing reads and results in predicted variants. A user-friendly interface is critical in a customizable, open-source bioinformatics pipeline that is easy to install and run without specialized computational training.

In order to address the need for an end-to-end customizable bioinformatics pipeline for the primary analysis of sequence data, we have developed a collection of analysis workflows for NGS data and the detection of genetic alterations in cancer called Software for Clinical Health in Oncology for Omics Laboratories (SCHOOL) [Figure 1]. SCHOOL: (i) can detect SNVs, indels, CNVs, and translocations from RNA and DNA sequencing, (ii) has tools for mutational profiling and omics integration, and (iii) is designed to be easy to run on local computing resources or the cloud, with all software packaged alongside dependencies, so that they can work on any system where singularity or docker programming packages are available. We have optimized each step to use a minimal amount of RAM and processors to reduce computation costs on the cloud. These workflows contain the steps necessary to complete primary NGS analysis, including variant detection and annotation. Furthermore, these workflows can execute on a local cluster using Nextflow, a command-line workflow manager, or on the cloud, https://platform.dnanexus.com/panx/projects/FvPKK200Y9g81KqkJ9X818/data/, using the DNANexus applets and workflows code.

**Technical Background**

**NGS analysis**

The primary analysis of sequence data for the detection of somatic variants in tumor samples requires five main steps including (i) alignment of the raw sequencing data to a reference genome, (ii) identification of SNVs and indels in DNA, (iii) identification of CNVs in DNA, (iv) identification of copy number and structural variants in RNA and DNA, and (v) annotation and the prediction of effect. For each step, there are many considerations for the bioinformatics workflow that can affect accuracy.

For alignments, the user should consider the genome reference, removal of duplicate reads, and the alignment program. There are currently two available versions of

---

**Figure 1:** Overview of SCHOOL workflow from sequencing through reporting. In SCHOOL, data flow from the sequencer into the primary analysis pipeline, which includes quality control, alignment, and variant calling appropriate for the sample type. Then, in secondary analysis, the variants can be annotated for eventual clinical reports.
the reference human genome: GRCh37 (hg19), released in 2009, and GRCh38 (hg38), released in 2013. GRCh38 has been shown to produce more accurate alignments. Sequence duplicates are caused when polymerase chain reaction (PCR) errors are amplified. Duplicates can be marked or removed, so that they are ignored in downstream steps. Picard (http://broadinstitute.github.io/picard/) can be used to mark duplicates, and Samtools can be used to remove duplicates. When unique molecular barcodes are included in the sequencing adapter, added during sample preparation, FGBio (see URLs) can be used to create consensus sequences of duplicates. BWA-MEM is the most used tool for sequence alignment of sequencing reads to the human reference genome. The output of BWA must be converted to BAM, sorted, and indexed using Samtools in order to be used in variant detection tools.

For SNVs and indels, there are many different open-source tools that can be used for the detection of these variant types. Somatic variants can be detected in somatic or tumor-only mode using the following tools: Strelka2, Freebayes, MuTect2, Pindel, BCFTools call, LoFreq, VarScan, Platypus, GATK4, and Scapel. Some tools such as GATK and BCFTools are designed to identify germline variants, whereas other tools, such as LoFreq and MuTect2, are designed to identify low-frequency variants common in tumor samples. A comparison of nine somatic variant calling programs found that Mutect2, Strelka2, and Virmid were among the most accurate.

There are also many different methods for the detection of copy number and structural variation. Some methods use sequence depth of coverage, or the average number of reads overlapping each region of the genome, to determine copy number changes. Because biases in coverage differ in each region of the genome, this coverage is not uniform, but can be normalized with healthy control samples. Additionally, some methods can take into account the allele frequency of common polymorphisms, called b-allele frequency, to correct these biases in the depth of coverage. A comparison of four CNV detection tools found that CNVKit had high sensitivity but a lower specificity relative to other programs like Control-FREEC. Although there are also many methods for the detection of structural variants, the accuracy of detection of structural variants is low compared with SNV and indels for short-read data. The strategy that most laboratories employ is the validation of known clinically actionable structural variants such as the FLT3 internal tandem duplication or known gene fusion events.

**Computation interoperability and graphical interfaces**

A software container is a freestanding unit that comprises the software of interest and all of its dependencies. Containerization allows for better maintenance and stability of computer software because it supports: (i) the deployment of the same code on the same running environment on any computer system (on premises or cloud); (ii) the separation of software packages to their own environment to satisfy specific running conditions; and (iii) dependencies and easier implementation of new packages or tools. Two popular containerization tools are docker, which was designed to work in a cloud environment, and singularity, which was designed to work in a local high-performance computing environment. Fortunately, containers that are created using docker can be converted to run using singularity on a local cluster. Additionally, containers can be versioned and easily shared, making them ideal to distribute code for bioinformatics pipelines.

Whether it is running in a high-performance computing environment locally or on the cloud, a bioinformatics pipeline has several important elements including automated advancement from step to step and parallelization. Pipelines are designed to take the input of each step as it becomes available and run blocks of code to convert it into predetermined output files. Some steps are serialized, meaning that they are dependent on the successful completion of earlier steps, such as alignment being dependent on the completion of read trimming. Other processes, such as variant calling using different callers, can be run concurrently to save time in a process called parallelization. Pipelines can be controlled on local computing resources using languages such as Nextflow, WDL, and Snakemate. In the cloud, commercial cloud frameworks exist to allow biologists with limited computational expertise to run pipelines in a point-and-click environment, using platforms such as DNAnexus, Seven Bridges, and Illumina Connected Analytics.

**Approach**

To implement our pipelines, we (i) tested several software packages for analysis accuracy, (ii) created docker containers for each self-contained step in our workflow, (iii) implemented input options for customizing each step, and (iv) designed an end–end workflow for primary analysis that could be run on an internal high-performance computing cluster or run on the cloud using a point-and-click graphical interface.

**Tool testing**

We tested 10 variant calling methods, including Strelka2, Freebayes, MuTect2, Pindel, BCFTools call, LoFreq, VarScan, Platypus, GATK4, and Scapel, using data generated from an engineered cell line that was designed with variants with low allele frequency and validated by quantitative PCR (Table S1). Consistent with previous studies, we found high sensitivity with MuTect2 and Strelka2. Of the 21 SNVs and small indels (<50 bp) present, Freebayes detected all 21 variants, MuTect2 detected 20 and LoFreq detected 16; each caller was within 10% of the expected variants allele frequency.
effect.

Other sources of annotations include the database detection of common mutations and snpEff for gene HiSAT2,[28] and expression values are determined using assessing gene expression. Reads can be aligned with data, we have implemented the steps necessary for Because expression can also be assessed with the RNASeq workflow script and configuration were implemented for execution on a local high-performance computing cluster, end-to-end workflows

The bioinformatics workflow contains three elements: (i) a software container created using Docker, which contains all software dependencies for each step, (ii) scripts written in bash that contain software commands necessary to complete each step of the workflow, and (iii) the workflow script and configuration that defines the inputs and outputs of each step, the compute requirements, the bash script parameters, and the container used for each step. The workflow script and configuration were implemented for execution on a local high-performance computing cluster, using the workflow management program Nextflow, and on the cloud, using the DNANexus Toolkit.

For users with bioinformatics and computing expertise, Nextflow can be configured to run on a variety of platforms locally and on the cloud. Nextflow readily submits jobs to commonly used compute cluster scheduling software such as SGE, PBS, and SLURM but also can be configured to submit jobs to cloud systems on Amazon Web Services (AWS) and Google Cloud. Users can create a Nextflow configuration file to customize the workflow for their hardware. Additionally, Nextflow allows for users to resume failed jobs and has extensive logging of each step, making troubleshooting easy to document. Finally, these Nextflow workflows can be configured to run on individual tumor samples or tumor/normal sample pairs or in a batch mode for processing the data for an entire sequencing run.

In order to make these workflows accessible to users with limited computational expertise, we transformed the workflows to run on AWS resources on the DNANexus platform. DNANexus has a point-and-click user interface for running data analysis. Each step of the workflow was transformed into a DNANexus App, and apps were combined into a DNANexus Workflow. Users can run these pipelines with their raw sequence files in FastQ, a DNA reference tar gzip file and a gene panel reference tar gzip file. To reduce the price to run each step, every DNANexus app was run on test files to determine the minimal resources necessary, largely through monitoring memory and processor consumption and increasing resources incrementally when tools reached memory or disk usage limits (Table S2). We then set these resource requirements as the default settings for each app. Users can alter these settings to decrease computing time. The cost of analysis is highly dependent on the size of the data set and the machines chosen to do the analysis, where the user will want to balance cost and computational time.

**Conclusion**

We have developed SCHOOL, a set of bioinformatics analysis tools and pipelines for the analysis of NGS data in an academic clinical oncology laboratory (Figure S1), which has been in use at the CAP/CLIA laboratory at UT Southwestern Medical Center for four years. Additionally, SCHOOL pipelines have been used in over 20 research studies ranging from basic science to case reports.[19-43] SCHOOL includes tools and methods for primary analysis of sequencing data from raw reads to finished variant calls, accommodating germline and somatic DNA and tumor RNA. We further include tools for panel-specific customization, including extensions for: copy number analysis, microsatellite stability, integration between DNA and RNA data, and structural variant calling to detect gene fusions in DNA and ITDs.

(VAF). Callers such as Strelka2, GATK2, Platypus, and VarScan, using default parameters, detected the variants with >20% VAF, as designed (Table S1). Pindel detected the 300 bp internal tandem duplicate (ITD) in the FLT3 gene, along with two other indels at 35%–40% VAF.

**Tools implemented in docker and customization**

Our workflow allows users to choose for SNV and indel detection Strelka2,[10] Freebayes,[11] MuTect2,[12] and Pindel[10] or a combination of the four. For MuTect2,[12] alignments are first recalibrated using GATK4[13] with BaseRecalibrator and ApplyBQSR. If matched tumor/normal pairs are sequenced, users can ensure that these samples originate from the same patient using BCFtools[14] and NGS Checkmate.[22]

We implemented several tools for the detection of copy number and structural variants. For CNVs, we implemented CNVKit.[23] Because CNVKit works best with a panel of healthy normal control samples, we have also implemented a container and tool to generate this healthy control reference. To detect ITDs, users can use Pindel and ITDSeeker,[11,24] When microsatellite-specific baits are included, microsatellite stability can be estimated with an MSI-Sensor pro,[25] Gene fusions (translocations) can be detected using DNA- and RNA-specific tools. Star-Fusion is implemented for RNASeq data, and DELLY and SVABA are implemented for DNA[26,27] sequencing data.

Because expression can also be assessed with the RNASeq data, we have implemented the steps necessary for assessing gene expression. Reads can be aligned with HiSAT2,[28] and expression values are determined using FeatureCount and StringTie.[29,30] Variants in the RNA can be determined using Freebayes[11] and BamReadCt (see URLs).

Variants can be annotated using gnomAD[31] for the detection of common mutations and snpEff for gene effect.[32] Other sources of annotations include the database of oncoKB hotspots,[33] Encode repeat regions, [34] the database of non-synonymous of functional predictions (dbNSFP),[35] and variant databases dbSNP,[36] Clinvar,[37] and COSMIC.[38]
The SCHOOL pipeline can be optimized for the panel used in the assay. For example, when a panel of normal samples is included in assay development, a panel customization pipeline will align each sample and generate panel reference samples and the input BED files for copy number analysis. The aggregate normal sample VCF file will also be created to use when running MuTect2 to remove artifacts and rare variant sequences. Lastly, the normal samples can be used as a microsatellite reference for predicting microsatellite stability in the absence of a matched-normal sample.

For groups running an RNA Sequencing assay, there are several opportunities for data integrations, including comparison of RNA and DNA breakpoints in gene fusion events, comparison of splice site alteration using RNA data, and independent confirmation of variants by concurrent expression in RNA. The presence of variants in both the tumor RNA and tumor DNA provides enhanced confidence that a variant is not an artifact of the assay. Additionally, the presence of variants in RNA could indicate that the aberrant gene mutation is expressed in the tumor tissue. This could provide additional support of the importance of the variant, particularly for a suspected gain of function variants in oncogenic drivers or potential splice site variants that result in exon exclusion or intron inclusion. However, it is still important to note that variants in regulatory regions or gene deletions could prevent RNA expression.

These workflows represent a user-friendly, inexpensive, and flexible way to implement NGS bioinformatics for mutation detection and annotation in CAP/CLIA laboratories. SCHOOL can be easily installed on a local computing cluster that uses a queuing system such as SLURM or SGE with Nextflow with minimal dependency on pre-installed packages or runs on the cloud for laboratories that lack local resources and expertise with a user-friendly point-and-click graphic interface. We estimate the costs for analysis using cloud resources is a fraction of the costs of data generation and for some labs will reduce the need for additional computational expertise and resources on-site.

These pipelines implemented in SCHOOL perform computationally heavy analysis in variant detection, which we consider to be the primary analysis. Users should, in the course of their validation studies, determine the filtering parameters of these results to maximize sensitivity and specificity for their assay using a set of quality metrics for the variants based on variant type, including a number of alternate reads, percent of alternate reads, strand bias, and other quality scores. In this secondary analysis step, in addition to filtering variants and removing artifacts, the user can determine mutational profiling metrics like tumor mutational burden and distributions of SNV by codon change. These secondary analysis steps often need tuning based on the assay and are less computationally intensive, meaning they can be done locally on a PC or laptop computer.

**URLs**

- bcft2fastq (https://support.illumina.com/sequencing/sequencing_software/bcft2fastq-conversion-software.html)
- fgbio (https://github.com/fulcrumgenomics/fgbio)
- picard (http://broadinstitute.github.io/picard/)
- COSMIC (cancer.sanger.ac.uk)
- bamreadct (https://github.com/genome/bam-readcount)

**Data Availability**

All the code used in SCHOOL is available at the UTSW Clinical Laboratory github wiki site: https://medforomics.github.io/schoolwiki/ using the repos: school for pipelines, process_scripts used in each docker container, and dnanexus_applets for running on DNANexus.

**Financial support and sponsorship**

Nil.

**Conflicts of interest**

There are no conflicts of interest.

**REFERENCES**

**Supplementary Materials:**

### Table S1: Predicted variants allele frequencies (VAF) by variant calling tools for horizon discovery engineered cell line, HD829, qPCR variants validated

<table>
<thead>
<tr>
<th>Gene</th>
<th>Amino acid change</th>
<th>Variant type</th>
<th>Expected VAF</th>
<th>Freebayes (hotspot)</th>
<th>BCFtools</th>
<th>LoFreq</th>
<th>Platypus</th>
<th>GATK</th>
<th>MuTect2</th>
<th>Strelka2</th>
<th>Vscan</th>
<th>BCFtools</th>
<th>Scapel</th>
<th>Pindel</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLT3</td>
<td>ITD300</td>
<td>300bp INS</td>
<td>5%</td>
<td>Freebayes</td>
<td>BCFtools</td>
<td>LoFreq</td>
<td>Platypus</td>
<td>GATK</td>
<td>MuTect2</td>
<td>Strelka2</td>
<td>Vscan</td>
<td>BCFtools</td>
<td>Scapel</td>
<td>Pindel</td>
</tr>
<tr>
<td>NRAS</td>
<td>Q61L</td>
<td>SNP</td>
<td>10%</td>
<td>9.1%</td>
<td>8.4%</td>
<td>9.0%</td>
<td>9.2%</td>
<td>9.8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNMT3A</td>
<td>R882C</td>
<td>SNP</td>
<td>5%</td>
<td>4.4%</td>
<td>4.3%</td>
<td>4.4%</td>
<td>4.7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF3B1</td>
<td>G740E</td>
<td>SNP</td>
<td>5%</td>
<td>4.9%</td>
<td>4.7%</td>
<td>5.0%</td>
<td>4.7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDH1</td>
<td>R132C</td>
<td>SNP</td>
<td>5%</td>
<td>3.2%</td>
<td>3.1%</td>
<td>3.2%</td>
<td>4.4%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GATA2</td>
<td>G2006s*18</td>
<td>DEL</td>
<td>35%</td>
<td>32.8%</td>
<td>28.0%</td>
<td>34.2%</td>
<td>35.5%</td>
<td>34.2%</td>
<td>32.2%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TET2</td>
<td>R1261H</td>
<td>SNP</td>
<td>5%</td>
<td>4.3%</td>
<td>4.1%</td>
<td>4.4%</td>
<td>4.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPM1</td>
<td>W288fs*12</td>
<td>INS</td>
<td>5%</td>
<td>2.7%</td>
<td>1.8%</td>
<td>4.5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EZH2</td>
<td>R418Q</td>
<td>SNP</td>
<td>5%</td>
<td>3.6%</td>
<td>3.3%</td>
<td>3.6%</td>
<td>4.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JAK2</td>
<td>F537-K539&gt;l</td>
<td>DEL</td>
<td>5%</td>
<td>2.3%</td>
<td></td>
<td>3.4%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JAK2</td>
<td>V617F</td>
<td>SNP</td>
<td>5%</td>
<td>3.4%</td>
<td>3.3%</td>
<td>3.4%</td>
<td>3.9%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABL1</td>
<td>T315I</td>
<td>SNP</td>
<td>5%</td>
<td>4.0%</td>
<td>3.8%</td>
<td>3.9%</td>
<td>3.6%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBL</td>
<td>S403F</td>
<td>SNP</td>
<td>5%</td>
<td>4.3%</td>
<td>4.3%</td>
<td>4.3%</td>
<td>5.1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRAS</td>
<td>G13D</td>
<td>SNP</td>
<td>40%</td>
<td>32.7%</td>
<td>32.0%</td>
<td>32.8%</td>
<td>32.8%</td>
<td>35.9%</td>
<td>32.8%</td>
<td>31.3%</td>
<td>31.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLT3</td>
<td>D835Y</td>
<td>SNP</td>
<td>5%</td>
<td>3.7%</td>
<td>3.6%</td>
<td>3.8%</td>
<td>3.6%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDH2</td>
<td>R172K</td>
<td>SNP</td>
<td>5%</td>
<td>4.5%</td>
<td>4.4%</td>
<td>4.5%</td>
<td>5.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TP53</td>
<td>S241F</td>
<td>SNP</td>
<td>5%</td>
<td>5.3%</td>
<td>5.3%</td>
<td>5.4%</td>
<td>5.3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASXL1</td>
<td>G646fs*12</td>
<td>INS</td>
<td>40%</td>
<td>31.5%</td>
<td></td>
<td>31.1%</td>
<td>37.2%</td>
<td>5.3%</td>
<td>39.2%</td>
<td>32.0%</td>
<td>31.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASXL1</td>
<td>W796C</td>
<td>SNP</td>
<td>5%</td>
<td>4.9%</td>
<td>4.8%</td>
<td>5.1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUNX1</td>
<td>M267I</td>
<td>SNP</td>
<td>35%</td>
<td>33.5%</td>
<td>32.7%</td>
<td>33.4%</td>
<td>33.0%</td>
<td>32.4%</td>
<td>33.2%</td>
<td>32.3%</td>
<td>32.4%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCOR</td>
<td>Q11746s*8</td>
<td>INS</td>
<td>70%</td>
<td>63.4%</td>
<td></td>
<td>52.4%</td>
<td>65.1%</td>
<td>67.3%</td>
<td>67.2%</td>
<td>56.5%</td>
<td>47.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GATA1</td>
<td>Q119*</td>
<td>SNP</td>
<td>10%</td>
<td>9.1%</td>
<td>9.1%</td>
<td>9.0%</td>
<td>9.5%</td>
<td>9.9%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table S2: Computing resources needed for pipeline steps

<table>
<thead>
<tr>
<th>Step</th>
<th>Memory (GB)</th>
<th>Storage</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality Trim FastQ</td>
<td>3.75</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>DNA Alignment</td>
<td>30</td>
<td>340</td>
<td>16</td>
</tr>
<tr>
<td>Mark Duplicates</td>
<td>7.5</td>
<td>40</td>
<td>16</td>
</tr>
<tr>
<td>Sequence QC</td>
<td>61</td>
<td>160</td>
<td>8</td>
</tr>
<tr>
<td>SV Calling</td>
<td>30</td>
<td>340</td>
<td>8</td>
</tr>
<tr>
<td>Variant Profiling</td>
<td>7.5</td>
<td>40</td>
<td>8</td>
</tr>
<tr>
<td>Variant Calling (non-GATK)</td>
<td>30</td>
<td>340</td>
<td>8</td>
</tr>
<tr>
<td>GATK BQSR</td>
<td>17.1</td>
<td>420</td>
<td>2</td>
</tr>
<tr>
<td>Variant Calling (GATK)</td>
<td>30.5</td>
<td>80.4</td>
<td>4</td>
</tr>
<tr>
<td>VCF Union</td>
<td>3.75</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>Star-Fusion</td>
<td>61</td>
<td>160</td>
<td>8</td>
</tr>
<tr>
<td>RNA Alignment</td>
<td>15</td>
<td>160</td>
<td>8</td>
</tr>
<tr>
<td>BAM Read Ct</td>
<td>3.8</td>
<td>410</td>
<td>1</td>
</tr>
<tr>
<td>RNASeq BAM QC</td>
<td>3.75</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>Gene Abundances</td>
<td>3.75</td>
<td>40</td>
<td>2</td>
</tr>
</tbody>
</table>
**Supplemental Figure 1:** Overview of the RNA and DNA workflows. The RNA workflow (left) shows the tools used (adjacent to arrows) and files produced by the pipeline (boxes). The DNA workflow (right) shows the tools used to detect different kinds of variants, leading to a finalized “union” VCF. Output boxes shown in blue are recommended for all samples; those shown in green may require additional reference material, like a panel of normal samples, or are more dependent on the target panel, e.g., a panel with baits designed to capture microsatellite instability should use MSI Sensor.