Finding Expressed Mutations in Multiple Myeloma Cell Lines

Jensen Richardson*, Jafrin Pritha*, Wenxuan Jiang*, Rohit Prasad†, Dhivya Arasappan†, Jeanne Kowalski-Muegge‡

* Presenters, College of Natural Sciences, University of Texas at Austin; †Faculty Collaborator, College of Natural Sciences, University of Texas at Austin; ‡ Faculty Collaborator, Live**STRONG** Cancer Institute, Dell Medical School, University of Texas at Austin

Finding Expressed Somatic Mutations Can Lead to Neoantigen Prediction

RNA-Sequencing (RNA-Seq) is a sequencing technique to profile the expression levels of genes in a sample. Whole Exome Sequencing (WES) is a sequencing technique to genotype only the protein-coding regions of a sample. Finding expressed somatic mutations requires detection of variants using both WES and RNA-Seq data. Because a plethora of tools exist for this purpose, we compared several tools to develop a standardized pipeline for identifying expressed mutations in cancer. This can potentially lead to prediction of new cancer-induced antigens (neoantigens) and personalized immunotherapies for the treatment of cancers.

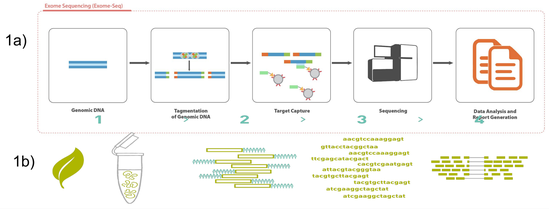


Figure 1:a) Whole Exome Sequencing workflow: Genomic DNA is fragmented, linked, and then tagged. The exomic regions were targeted for enrichment. **b)** RNA sequencing workflow: isolating the RNA, purifying the RNA and preparing cDNA, and sequencing.

71 Myeloma Cell Lines Were Analyzed

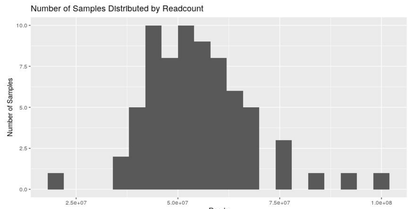


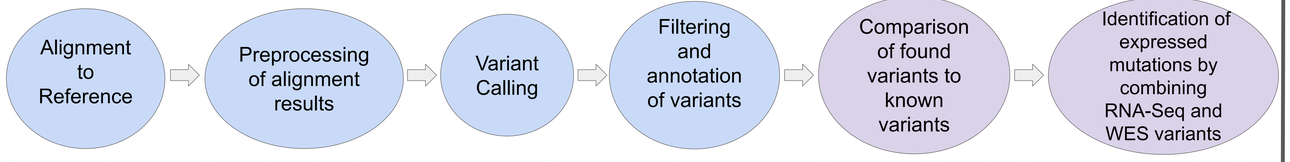
Figure 2: Distribution of read counts for 70 of the 71 samples (one was omitted due to the size of the dataset). The number of reads included in a sample is critical for calling variants. Higher read counts can increase variants called, but also increase processing time and false positives. File sizes for each sample ranged from 18-36gb each.

Using VarScan and GATK

The two tools used and compared were the Genome Analysis ToolKit (GATK) and VarScan. They are designed to call Single Nucleotide Variants (SNVs) and short insertions and deletions (indels), up to about 50 bases. Varscan is a Platform-independent mutation caller written in Java that works on both whole exome and whole genome sequencing data. It is able to work with individual samples or tumor-normal pair. GATK is a collection of Java tools for analyzing sequencing data. The tools themselves are very comprehensive and the pipeline starts with fastq files and ends with a vcf file of annotated variants.

General Workflow

Figure 3. A generalized workflow for benchmarking SNV tools. Completed steps in blue and future steps in purple.



The reads were aligned to the human genome. Preprocessing steps such as marking duplicates, accounting for splicing events in RNA, and recalibrating base quality were performed. Preprocessed reads were input to GATK or Varscan to call variants (SNVs and indels). The tools tend to be very sensitive and variant calls contained false positives that were filtered out. Filters were also imposed to only include somatic variants in further analysis. Variants were annotated to indicate function and level of conservation and compared to known variants to measure accuracy for each tool. Variants found in WES and RNA-Seq data will be compared and combined to identify expressed mutations as a followup step (indicated in purple).

Comparing GATK and Varscan - GATK calls more SNVs and Indels

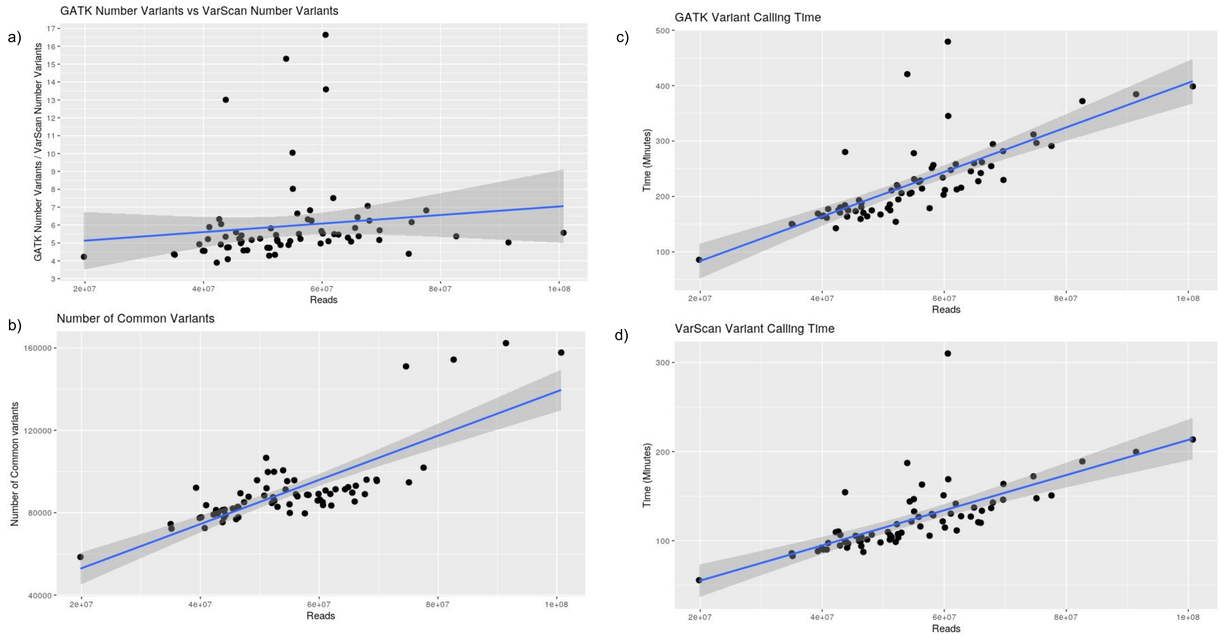


Figure 4: Number of variants (a), variants found by both tools (b) and processing time (c,d) for each sample. a) The ratio of variant detected by GATK to VarScan indicates that GATK calls on average 5.96 times more variants (indels + SNVs) than Varscan, and this is not highly correlated to read count. b) The number of common variants detected by GATK and VarScan increases roughly linearly with readcount (as expected) c&d) GATK generally takes on average 1.8 times longer than VarScan to detect variants. This time does not include preprocessing or post filtering.

Variants can be further analyzed and visualized using maftools

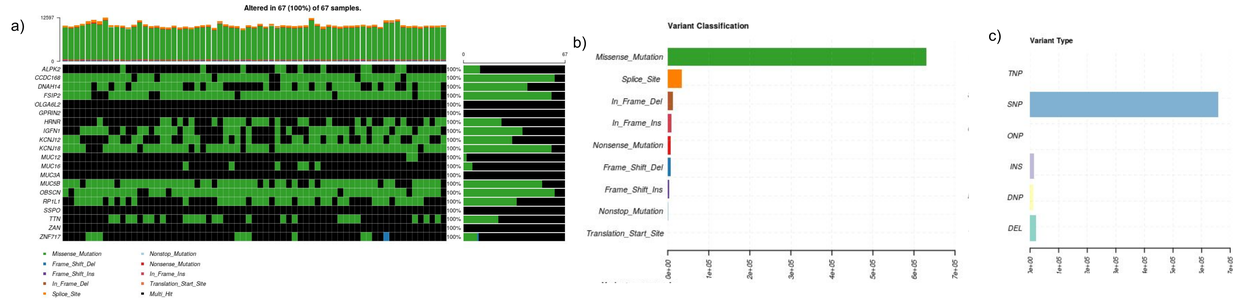


Figure 5: For 67 samples, The variants detected by GATK from the WES data was visuzalied using maftools to identify patterns. a) The 25 most mutated genes in the 67 samples along with the type mutation found in each sample. Black indicates multiple mutations, not the absence of mutations. b) The classes of mutations called indicates that the vast majority of the variants called were missense. c) The most common type of mutation called was SNP.

Preliminary Tool Comparison

	VarScan	GATK
Pros	• Adopts heuristic/statistic approach to call variants • Single command to run, easy to use	• Complete analysis toolkit from fastq to annotated variants • Exceptionally comprehensive set of tools • Active and helpful user base
Cons	• Unlike GATK, it is just a single variant caller. Users need to find compatible tools to do pre and post processing (such as annotation).	• Has lots of steps and is complicated to use which leads to complex and time intensive pipelines. • Under active development so the tools can change between versions.

Combining Workflows in Order to Find Expressed Mutations in Patient Data

In order to find expressed mutations, both whole exome data and RNA-seq data are needed, but there is no standardized method of combining those two datasets. The next step in our research would be developing a standardized method of doing so. Right now we are implementing these tools on cell lines in order to develop a workflow and gauge accuracy. Once that is done, the completed workflow will be applied to patient data.

Limitations

- No comprehensive evaluation of filtering parameters. Used default GATK and VarScan settings
- No evaluation of RNAseq variant calling tools and detected variants has been done yet, only evaluation of WES data.

Acknowledgements

We thank the Texas Advanced Computing Center (TACC), The University of Texas at Austin and the Biomedical Research Computing Facility (BRCF), The University of Texas at Austin for computational support. We also thank the Keats lab at the Translational Genomics Research Institute for the sequencing data.This work was supported by the TIDES FRI Summer Research Fellowship.

References

"Varscan - Variant Detection In Massively Parallel Sequencing Data". *Varscan.Sourceforge.Net*, 2020, http://varscan.sourceforge.net/. Accessed 13 July 2020.

Petti, Allegra A., et al. "A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing." *Nature communications* 10.1 (2019): 1-16

Team, GATK, and An Zheng. "RNAseq Short Variant Discovery (SNPs + Indels)." *GATK*, gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels-.

Van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler D, Gabriel S, DePristo M. "From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline." 2013. *Current Protocols In Bioinformatics*. 43:11.10.1-11.10.33

DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly M. "A framework for variation discovery and genotyping using next-generation DNA sequencing data." 2011. *Nature Genetics*. 43:491-498

Richters, Megan M., et al. "Best practices for bioinformatic characterization of neoantigens for clinical utility." *Genome medicine* 11.1 (2019): 56.

Mayakonda A, Lin D, Assenov Y, Plass C, Koeffler PH (2018). "Maftools: efficient and comprehensive analysis of somatic variants in cancer." Genome Research. doi: 10.1101/gr.239244.118.