From Cancer Sequencing Data to Neoantigen Prediction: A reusable pipeline using Snakemake

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

T

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 that targets only the protein-coding regions of a sample. By detecting genetic variants using both WES and RNA-Seq data, we can identify expressed somatic mutations, which can give rise to new cancer-induced antigens (neoantigens). Because neoantigens are unique to cancer cells, they can be used to target cancerous cells with immunotherapy. Predicting neoantigens using sequencing data can thus lead to personalized cancer immunotherapeutics.

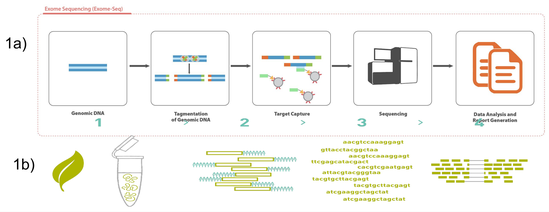


Figure 1:a) Whole Exome Sequencing workflow: Genomic DNA is fragmented, linked, and then tagged. The exomic regions were targeted for enrichment. **b)** RNA sequencing workflow: isolating the RNA, purifying the RNA and preparing cDNA, and sequencing.

Description of Pipeline

Our pipeline starts with raw sequencing data (RNA-Seq and WES data). It identifies expressed somatic variants using this data and generates neoantigen predictions. Open source bioinformatics tools were benchmarked for each step. GATK was selected for preprocessing and variant calling. NeoPredPipe was selected for neoantigen prediction.

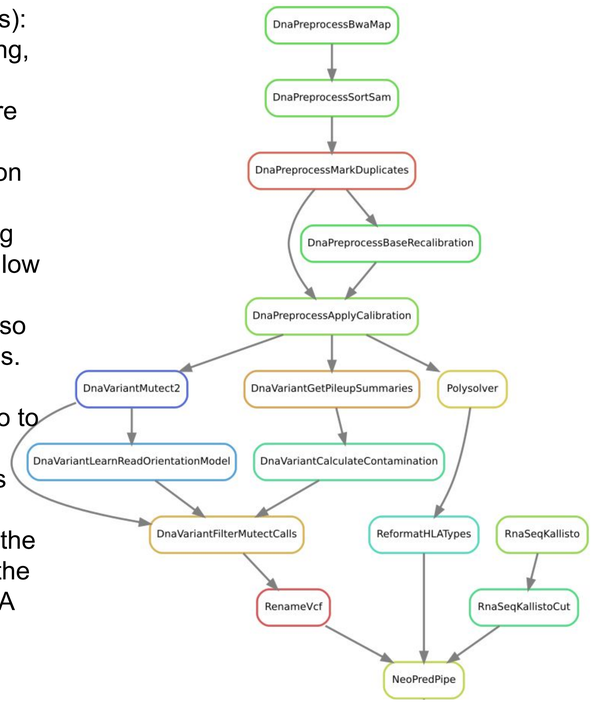
Snakemake Workflow Management

Snakemake is a lightweight workflow management system available through conda that greatly simplifies the process of putting individual samples through the long and complex pipeline involved in calling variants, quantifying RNA expression, and then using that information to predict neoantigens. After creating rules to produce outputs, snakemake intelligently dispatches slurm jobs that can vary on time, cpu usage, and number of nodes based on the complexity of the task, as well as the individual step, and then judge whether the job finished successfully, regardless of the slurm output code. It also features conda, singularity, & environmental module integrations.

Pipeline Steps

1. **Read Preprocessing Steps** (denoted by DnaPreprocess): Sequencing reads undergo preprocessing such as mapping, marking of duplicates, and base quality recalibration.
2. **Variant Calling Steps** (denoted by DnaVariant): Reads are input to GATK to call variants (SNVs and indels) using Mutect2. GetPileupSummaries and CalculateContamination are used to look for outside contamination and LearnReadOrientationModel is used to look for sequencing artifacts that bias one read strand over the other. These allow FilterMutectCalls to filter out many of the false positives generated by the initial mutation calling step. Filters are also imposed to only include somatic variants in further analysis.
3. **RNA-Seq quantification steps** (denoted by RnaSeqKallisto): RNA-Seq data is quantified using Kallisto to obtain expression values for each gene. This is used to identify expressed variants by only including those genes that are expressed in the RNA-Seq data.
4. **Neoantigen steps**: The neoantigen prediction tool uses the somatic mutations generated by the DnaVariant pipeline, the RNA expression levels generated by RnaSeq, and the HLA types generated by Polysolver to predict neoantigens.

Figure 3. Snakemake rule graph for neoantigen prediction. (The color of the box is not significant). Prefixes DnaPreprocess, DnaVariant, & RnaSeq, indicate the general phase of the workflow. Those without a prefix are part of neoantigen prediction.



Comparing Variant Callers

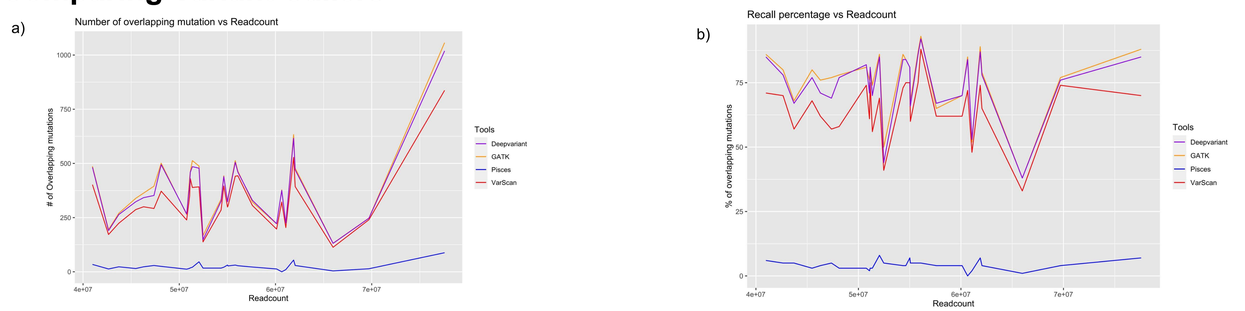


Figure 4: Comparison of Variant Callers by sample. Comparison of the number of mutations that overlapped with the accuracy database by sample, distributed by readcount. a) The raw number of overlapping mutations. Note that GATK and DeepVariant are almost identical, but GATK performs slightly better. b) Percent of overlapping mutations detected by each variant caller by sample, distributed by readgroup. GATK's slightly better accuracy is easier to see here. Note that Pisces performed far worse than any other variant caller, detecting less than 10% of the known mutations.

Neoantigen Prediction

Neoantigen prediction is a challenging task for several reasons, primarily because it relies on so many different factors. In addition to the mutation information produced by GATK, it requires knowledge of the variable HLA region of DNA to know which proteins the immune system will accept. We considered two tools to predict neoantigens: pVACseq and NeoPredPipe. pVACseq relies on a pre-existing database of predicted neoantigens and then measures them based on the HLA type and mutation.

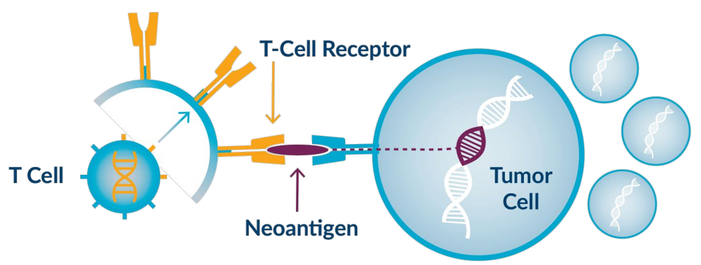


Figure 5: Diagram of a neoantigen allow a T cell to recognize a mutated surface protein on a cancer cell. The T Cell is able to recognize the HLA type because it has the correct HLA type and the tumor cell binds because neoantigen was customized to fit the mutation.

Tool Comparison

Before settling on GATK, we benchmarked 4 variant callers. They were GATK, DeepVariant, Varscan2, and Pisces in order of final accuracy. After preprocessing according to the industry standard GATK preprocessing pipeline, they were called according to their respective manuals. We measured accuracy by comparing detected variants to known variants derived from the DepMap and CCLE databases provided by the Broad Institute. (Figure 4)

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 they require completely different analysis pipelines that sometimes would be done together, and other times would be completely separate. Snakemake allows for the information of how to execute each pipeline to be stored in different git repositories that rely on each other for their dependencies. This is also indicated by the prefixes in Figure 3.

Limitations

- Lack of known neoantigens for the cell lines makes benchmarking challenging.
- 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) 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-.
Mölder F, Jablonski KP, Letcher B et al. Sustainable data analysis with Snakemake [version 2; peer review: 2 approved]. *F1000Research* 2021, 10:33 (https://doi.org/10.12688/f1000research.29032.2)
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.
Poplin, R., Chang, PC., Alexander, D. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* 36, 983–987 (2018). https://doi.org/10.1038/nbt.4235