LENS
Landscape of Effective Neoantigens Software

Steven P. Vensko II*[1], Kelly Olsen[1], Dante S. Bortone[1], Christof C. Smith[2], Shengjie Chai[3], Alex Rubinsteyn[1], and Benjamin G. Vincent†[1]

[1]School of Medicine, University of North Carolina - Chapel Hill, Chapel Hill NC
[2]Brigham and Women's Hospital, Boston MA
[3]Uber, San Francisco CA

Abstract

Elimination of cancer cells by T cells is a critical mechanism of anti-tumor immunity and cancer immunotherapy response. T cells recognize cancer cells via engagement of T cell receptors with peptide epitopes presented by major histocompatibility complex (MHC) molecules on the cancer cell surface. Peptide epitopes can be derived from antigen proteins coded for by multiple genomic sources. Bioinformatics tools used to identify tumor-specific epitopes via analysis of DNA and RNA sequencing data have largely focused on epitopes derived from somatic variants, though a smaller number have evaluated potential antigens from other genomic sources. We report here an open-source workflow utilizing the Nextflow DSL2 workflow manager, Landscape of Effective Neoantigen Software (LENS), which predicts tumor-specific and tumor-associated antigens from single nucleotide variants (SNVs), insertions and deletions (InDels), fusion events, splice variants, cancer testis antigens (CTAs), overexpressed self-antigens, viruses, and human endogenous retroviruses (hERVs). The main advantage of LENS is that it extends the breadth of genomic sources of tumor antigens that may be discovered using genomics data. Other advantages include modularity, extensibility, ease of use, incorporation of phasing and germline variant information in epitope identification, and harmonization of relative expression level and immunogenicity prediction across multiple genomic sources. Current limitations include lack of support for class II MHC epitope predictions and advanced visualization features. To demonstrate the utility of LENS, we present an analysis of predicted antigen landscape in 115 acute myeloid leukemia

*steven_vensko@med.unc.edu
†benjamin_vincent@med.unc.edu

(AML) samples. We expect that LENS will be a valuable platform and resource for T cell epitope discovery bioinformatics, especially in cancers with few somatic variants and increased importance of tumor-specific epitopes from alternative genomic sources.

1 Introduction

Tumor-specific and tumor-associated antigens are of great interest for understanding cancer immunobiology and developing personalized immunotherapy approaches including vaccination or adoptive T cell therapy. Unfortunately, predicting which tumor antigens are immunogenic *in vivo* is challenging [1, 2]. These predictions are complicated by several factors influencing the appropriateness of candidate tumor antigens. Specifically, only a subset of tumor-specific or tumor-associated variants or sequences within a patient will be transcribed, translated, and processed by the proteosome. Only a subset of peptides generated through protein degradation will be presented on the cell surface by MHC and only a further subset of those will result in T-cell recognition, activation and cytotoxicity. Despite recent advances in understanding peptide attributes associated with *in vivo* immunogenicity, other factors including RNA editing, post-translational modifications, and peptide splicing may influence the effectiveness of computational predictions [3, 4, 5]. Improved cancer antigen prediction using genomics data could empower more detailed studies of anti-tumor T cell responses as well as better personalized combination immunotherapy strategies. [6].

Multiple workflows have been developed to predict tumor antigens from high throughput sequencing data. These include OpenVax, pVACTools, and nextNEOpi among others [7, 8, 9]. These tools have been successful in allowing users to study associations between tumor antigen burden and immunotherapy response, evaluate tumor-antigen specific T cell responses, and support clinical trials of personalized immunotherapy targeting predicted tumor antigens [10, 7, 11, 12, 13, 14, 6]. Here we present an extendable, modular, and open-source workflow, Landscape of Neoantigen Software (LENS), coupled with a Nextflow-based analysis platform, Reproducible Analyses Framework and Tools (RAFT, see Section 2.1), which addresses shortcomings of current neoantigen workflows, expands the repertoire of predicted tumor antigens, and serves as a springboard towards a community-driven advances in neoantigen prediction through iterative workflow improvements while encouraging reproducibility. LENS improves upon previous offerings by providing an end-to-end workflow utilizing Dockerized tools and modular workflows, expanding the types of tumor antigens predicted, providing a harmonized pan-tumor antigen abundance quantifier, and allowing full customization of the workflow.

2 Landscape of Effective Neoantigens Software

LENS, or Landscape of Effective Neoantigens Software, is a modular, extensible workflow which predicts tumor antigens from an array of sources including somatic single nucleotide variants (SNVs), conservative, disruptive, and frameshift insertions and deletions (InDels), splice variants, in-frame and frameshift fusion events, viruses, endogenous retroviruses, and cancer testis/self antigens (Figure 1). LENS includes phasing and germline variant information in epitope identification, and it harmonizes variant RNA expression across genomic sources to provide a more usable relative expression ranking each peptide epitope. More information about LENS, including workflow documentation and instructions for running LENS can be found at the LENS GitHub repository: https://gitlab.com/bgv-lens.

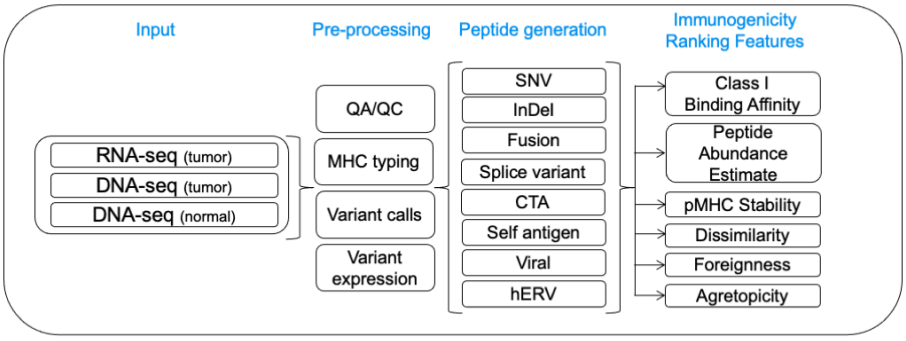

Figure 1: Landscape of Effective Neoantigens Software (LENS): A modular workflow for predicting Class I neoepitopes from SNVs, InDels, fusion events, splice variants, CTAs, self-antigens, viruses, and hERVs.

Here we provide an overview of the LENS workflow, compare it to other currently available workflows, describe technical aspects of LENS, and discuss the results of running LENS on 115 Acute Myeloid Leukemia patients in The Cancer Genome Atlas dataset (TCGA-LAML).

2.1 Implementation

The framework running LENS is a Nextflow wrapper called RAFT (Reproducible Analyses Framework and Tools) which enables our goals of modularity, extensibility, and improved collaboration. RAFT supplements the Nextflow DSL2 with a collection of tool-level modules, workflow modules, and automa-

tion to ease workflow creation and running (e.g. module dependency resolution, parameter capturing, etc.). A full description of RAFT's capabilities is beyond the scope of this text, but more information can be found at RAFT's GitHub: https://github.com/pirl-unc/RAFT

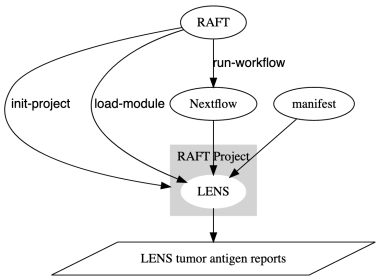

Figure 2: Relationship among RAFT, LENS, and Nextflow: LENS is a workflow constructed of Nextflow DSL2-compatible modules run through a heavily parameterized Nextflow2 instance directed by RAFT.

2.2 Flexibility

The modular design of LENS allows high flexibility in its execution. This presents itself in many forms including user-definable references files (FASTAs, GTF/GFFs, BED files, etc.) on either global or workflow-specific levels and with parameter inheritance among workflow scopes. As a result, users maintain fine-grain control over workflow and tool-level parameters. This allows manipulation of workflow behaviors to account for variability in sample type (e.g. fresh frozen vs. formalin fixed) or changing filtering criteria. Perhaps most importantly, the modular nature of LENS allows users to introduce novel tools or replace current tools a better suited tool. For example, LENS uses HLAProfiler as its default HLA typer – a tool which uses RNA sequencing data to perform HLA allele calls [15]. There may be situations where an exome sequencing-based tool is preferred, such as where DNA sequencing only is available, cancer cells have down-regulated MHC expression, or simply user preference based on performance. It's possible in this case to replace the HLAProfiler portion of LENS with a tool such as Optitype without requiring further modification to the rest of the workflow [16]. There may be other scenarios where swapping tools may be desirable or necessary. For example, there are a variety of variant callers, fusion callers, and copy number inference tools available with their individual benefits and disadvantages [17]. Replacing the LENS default tools with these alternatives or defining new ensemble approaches may produce richer and more meaningful results. Users may also want to change tools to be within compliance of any license requirements associated with their usage.

2.3 Extensibility

Properly leveraging novel technologies, protocols, and public datasets is crucial for progress towards meaningful and impactful insights in tumor antigen discovery. In light of this, we designed LENS such that it can be extended through creation of new modules and workflows. For example, the current report generated by LENS includes tumor type-specific relative abundance estimates for transcripts harboring SNV/InDel neoantigens from the corresponding TCGA patient cohort. LENS reports can further be supplemented by including published empirical immunogenicity data as more become available. Inclusion of single-cell sequencing modules would disentangle the genetic and transcriptional heterogeneity within a patient's tumor that cannot be observed in bulk sequencing data. Long read sequencing allows for detection of large structural variants and provided improved resolutions of haplotypes near tumor peptide-generating sequences of interest. Other technologies, such as Ribo-Seq, will not only improve current tumor antigen predictions by confirming reading frames, but can open new tumor antigen sources such as "genomic dark matter" arising from non-canonical reading frames [18]. We consider the version of LENS presented here to be a "snapshot" of the continuously improving workflow that will aid the immuno-oncology community towards improved therapeutics.

2.4 Comparison to Other Workflows

Neoantigen workflows have become increasingly popular as the value of personalized immunotherapy has been realized. There are currently dozens of neoantigen workflows publicly available including pVACtools, the OpenVax workflow, nextNEOpi, and MuPeXi [8, 7, 9, 19]. A comprehensive comparison among all available workflows is available in the literature [9]. Here, we seek to compare LENS to neoantigen workflows with documented usage within clinical trials as we expect LENS be applied to trials in the future. These comparisons focus on features rather than accuracy of neoantigen predictions, though we intend to explore these facets moving forward. A summary of comparisons can be found in Table 1.

2.4.1 LENS vs. pVACtools

pVACtools consists of several tools allowing various inputs which provides flexibility in its usage. Users may provide a pre-processed annotated, phased VCF with expression data to pVACseq or, alternatively, a FASTA file containing potential peptides to pVACbind to predict SNV- and InDel-derived Class I and Class II neoantigens. Neoantigens generated by fusion events are also available by providing an annotated (with AGFusion) fusion to pVACfuse. Binding affinity is predicted using either 8 Class I predictors or four class II predictors. The resulting peptides are filtered for binding affinity, coverage, and transcript support. pVACvector combines the filtered peptides into a DNA vector-based vaccine in such a way as to minimize the generation of junctional epitopes.

pVACtools emphasizes both visualization (to aid in interpretation) and neoantigen vaccine manufacturability for an improved user experience. It has been used in the cancer vaccine design of at least five clinical trials as of March 2020 [8].

pVACtools and LENS share a common goal, but differ in several notable ways. LENS considers six potential tumor antigen sources while pVACtools focuses on three (SNVs, InDels, and fusion). LENS employs an end-to-end approach in which raw patient FASTQs are inputs while pVACtools requires pre-processing prior to running (phased VCFs generation, HLA typing, and fusion calling). An adaption of pVACtools, nextNEOpi, discussed later, seeks to reduce the pre-processing burden by integrating those steps into the workflow. We feel the end-to-end nature of LENS provides several advantages: 1) the workflow generating the neoantigen predictions can be thoroughly interrogated for best practices and optimal parameterization, 2) all files and results have provenance such that all factors influencing the neoantigen predictions can be traced, and 3) results become easier to regenerate as bioinformatic processes explicitly documented within the code. A major disadvantage of an end-to-end approach is increased computational burden and longer run times involved with the more complex workflow.

pVACtools has some helpful features that are currently missing from LENS. For example, LENS currently has minimal visualization capabilities and performs minimal filtering, though its reports are well-suited for explorative analyses. LENS also does not include any manufacturing considerations, though the recent global usage of mRNA vaccines in response to the COVID pandemic suggest manufacturing limitations or constraints may be less burdensome for future vaccine development. pVACtools also provides support for class II peptides while LENS focuses solely on class I peptides. Nevertheless, we feel LENS is currently mature enough for us to focus on improving support for these aspects of the workflow and including these features in the future.

2.4.2 LENS vs. nextNEOpi

nextNEOpi (*next*flow *NEO*antigen prediction *pi*peline) is perhaps the most recently released neoantigen workflow [9]. It has not been directly utilized in clinical trials, but is likely the most closely related workflow to LENS due to its usage of the Nextflow language. nextNEOpi is an end-to-end workflow utilizing the pVACseq tool from the pVACtools along with NeoFuse to predict class I and class II SNV, InDel, and fusion neoantigens through a gradient-boosting machine approach[2, 9].

nextNEOpi and LENS both utilize Nextflow, but use differing Domain Specific Language, or DSL. LENS and its modules are written using the modular DSL2 syntax while nextNEOpi is written in DSL1. The DSL1 syntax requires the entirety of the workflow to be contained within a single, monolithic script with each step's inputs and outputs explicitly mapped. This syntax makes modifications to the script quite burdensome which may hamper workflow modification by end-users. Beyond implementation differences, LENS also supports more tumor antigen sources than nextNEOpi.

2.4.3 LENS vs. OpenVax

The OpenVax workflow is the neoantigen workflow released by the OpenVax group used in the PGV-001 clinical trial. The OpenVax workflow is managed through SnakeMake, a Python-based workflow manger,and is recommended to run within a Docker container. The pre-processing steps, including read alignment, alignment sanitization, and variant calling closely reflect those of LENS. OpenVax employs a set of custom tools for vaccine peptide selection, such as vaxrank (for ranking appropriateness of neoantigens), isovar (for accommodating neighboring variants during peptide generation), and varcode (for variant annotation and effect prediction). Running the OpenVax workflows generate a report containing ranked SNV- and InDel-derived Class I neoantigens that may serve as suitable targets for vaccination.

LENS and OpenVax are both end-to-end workflows and, unlike pVACtools and nextNEOpi, are both dependent on an RNA tumor sample to predict neoantigens. LENS further differs from OpenVax in the variety of tumor antigen sources considered and the degree of modularity as each component is controlled by separate Nextflow modules.

2.5 Workflow Overview

The LENS workflow orchestrates over two dozen separate tools to generate tumor antigen predictions. LENS currently supports tumor antigen detection from the following tumor antigen sources: SNVs, InDels, fusion events, splice variants, viruses, retroviruses, cancer/testis antigens, and self-antigens. LENS will be expanded to include support for single-cell data, long read data, additional bioinformatics tools, and the inclusion of appropriate external reference data.

2.6 Data Pre-processing, Read Alignment, and Transcript Quantification

LENS processes each sample through trim galore and alignment (BWA mem for DNA samples and STAR for RNA samples) against a reference genome with an appropriate corresponding annotation file. A GENCODE v25 GTF supplemented with gEVE endogenous retroviral elements is the default annotation file[20]. Sanitization steps are then performed (duplicate reads marked with picard, base qualities recalibrated with GATK4, and reads near insertions and deletions realigned with ABRA2) to ensure higher quality inputs for each downstream tumor antigen source-specific workflow. Transcripts contained within the GTF annotation file are quantified with salmon. The pre-processing workflow results in numerous intermediate files required for each tumor antigen workflow: normal DNA BAM, tumor DNA BAM, tumor RNA genomic BAM, tumor RNA transcriptomic BAM (created with –quantMode TranscriptomeSAM parameter), tumor RNA junctions file, and tumor RNA transcript counts. The pre-processing workflow is visualized in Figure 3.

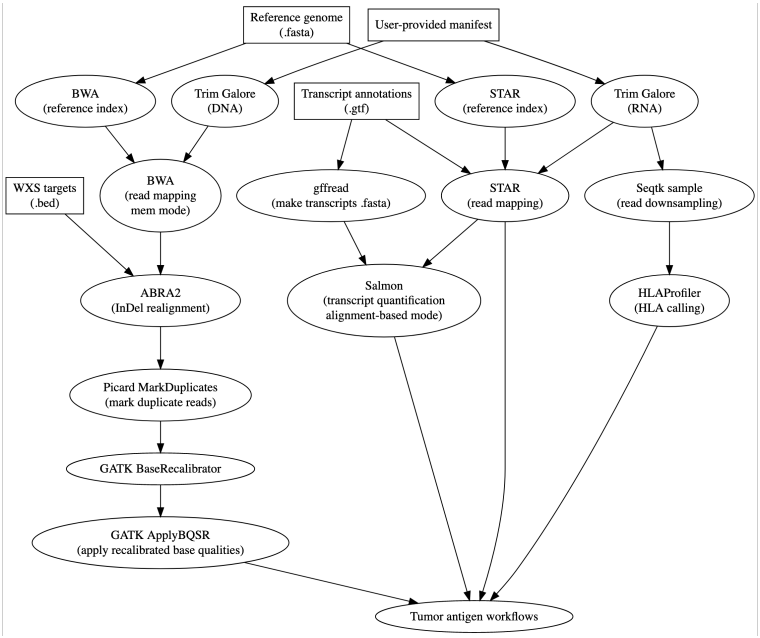

Figure 3: Pre-processing workflow: A patient's DNA tumor, RNA tumor, and DNA normal FASTQs are trimmed and aligned to a reference, sanitization steps are performed, and intermediate files are generated for downstream tumor antigen workflows.

2.7 Single Nucleotide Variants and Insertions/Deletions

Somatic single nucleotide variants (SNVs) and insertion/deletion variants (InDel) may result in tumor-specific peptides. These peptides may be a direct result of a somatic variant (missense SNVs, conservative in-frame InDels, or disruptive in-frame InDels) or may be a downstream consequence of the variant (e.g. alternative reading frames resulting from a frameshift InDel). Somatic SNVs cause non-synonymous changes within a protein that can be processed into a peptide that is recognized by T cells, triggering anti-tumor cytotoxicity [21]. These peptides are not expressed in non-tumor tissues, so T cells that target them could escape negative selection in the thymus, making them attractive targets for antigen-specific immunotherapy [21]. However, these targets are more prevalent in some cancers than others, as they are dependent upon tumor somatic mutation as well as immunogenicity of the individual SNVs present [22]. Vaccines targeting SNV-derived tumor-specific antigens have been used in clinical trials, with strategies including peptide vaccines [23, 24], dendritic cell vaccines [25, 26], DNA vaccines [27], and RNA vaccines [12]. Tumor-infiltrating lymphocytes cultures including T cells specific for SNV-derived tumor antigens have yielded clinical responses in multiple patients [28].

CAR T cell and other adoptive cell transfer (ACT) therapies are theoretically

applicable to SNV targeting, though these strategies are limited by the difficulty of generating single-chain variable fragments that specifically recognize neoepitope peptides bound to MHC molecules and low degree of SNV-derived neoepitope peptide sharing between patients [**imm2018imliu˙efficient˙2019**, 29]. Insertions and deletions (InDels) are another potential source of somatic variant-derived tumor antigens [30]. Some insertions and deletions lead to frameshifts, generating novel open reading frames. The sequences derived from these new open reading frames may be completely non-overlapping with germline sequences, and thus they may be more likely immunogenic than SNV-derived neoantigens. [21]. Impaired DNA mismatch repair pathways, in which insertions and deletions cannot be proofread and repaired, vastly increase frequency of InDel neoantigens. Fifteen percent of colon cancers are microsatellite instable due to hereditary or acquired mutations in mismatch repair genes [31]. Neoantigens derived from InDels showed higher affinity binding to antigen-presenting molecules than SNV-derived neoantigens and associated with greater T cell activation, showing that these antigens are highly immunogenic targets [30]. Bulk cytotoxic T lymphocyte cultures targeting frameshift-derived peptides have been shown to be capable of lysing cancer cells containing the frameshift [32].

The SNV and InDel LENS workflows initially filter variants using a consensus approach. Variants are called using three separate variant callers (Mutect2, Strelka2, and ABRA2), and only variants within the intersection of the appropriate callers (MuTect2 and Strelka2 for SNVs; all three callers for InDels) are considered for further investigation. This approach is similar to those used by other neoantigen prediction workflows [9]. Consistent detection of variants after this filtering culls potential false positive variants leaving a set of high confidence calls. Inclusion parameters are highly tunable by the user to select the level of stringency applied. The intersected VCF is filtered for coding variants (missense SNVs, conservative and disruptive InDels, and frameshift InDels) for further processing. Variants are filtered by two criteria at different points in the workflow: 1) their relative transcription abundance must exceed a user-specific percentile (e.g. 75%) of all transcripts and 2) sequence capable of coding variant-harboring peptides must be detectable in the patient's RNA sequencing data. Expression filtering creates a stable set of SNVs/InDels that are likely expressed, translated, and can be presented by the MHC on the tumor cell surface. Accurately generating tumor peptides from the patient's somatic variants requires phasing candidate somatic variants with neighboring germline variants (Figure 4). LENS calls germline variants with DeepVariant and performs phasing with WhatsHap using the patient's DNA- and RNA sequencing data. Phased heterozygous germline variants within the same haplotype block as a candidate somatic variant are incorporated into the resulting peptide where appropriate: 1) within the 10 amino acids upstream and downstream of the modified amino acid(s), 2) within the the matched wildtype sequence used to calculate agretopicity, and 3) within the 10 amino acids upstream and all amino acids downstream of frameshift InDels until the first stop codon. Homozygous germline variant are also integrated within each variant's respective contextual bounds while heterozygous germline missense variants are replaced by 'N's to

prevent their consideration during peptide prioritization. These germline variant integration steps ensure a more accurate representation of the patient's tumor peptides.

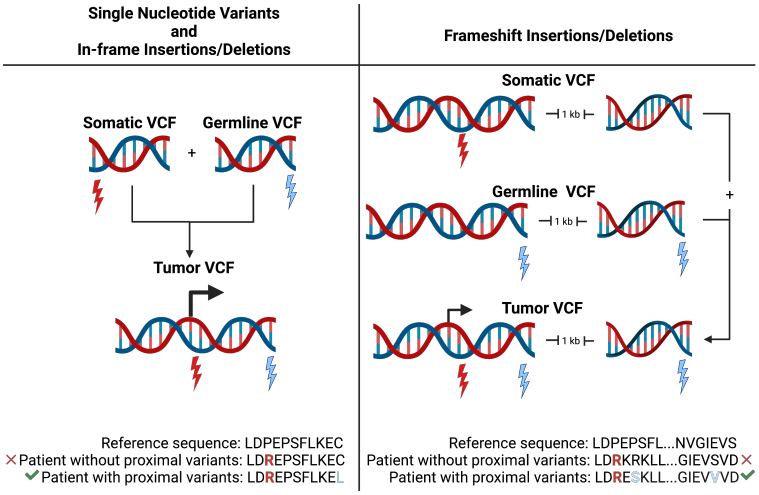

Figure 4: Combining of somatic variants with neighboring germline variants: Somatic variants of interest are combined with phased neighboring germline variants during peptide generation. Frameshift InDel variants are combined with phased neighboring gernline variants and downstream homozygous germline variants. This strategy maximizes the probability of predicted peptides reflecting those contained within a patient's tumors.

The report generated by the SNV workflow includes several metrics relevant to peptide prioritization such as the variant and wildtype binding affinities (as calculated by NetMHCpan-4.1b), agretopiticity (ratio of variant-to-wildtype binding affinities), the gene and transcript harboring the expressed variant, the transcript's relative abundance (in transcripts per million), as well as the number of RNA sequencing reads from the peptide's genomic origin containing a sequence which translates to the peptide. The last metric will be discussed in more detail later, but serves as an estimate of peptide abundance harmonized across tumor antigen workflows. Furthermore, LENS estimates variant cancer cell fraction with PyClone-VI by using copy number alternation (CNA) data from Sequenza along with MuTect2 allele frequencies. Understanding the resulting distribution of tumor sub-populations and estimating the clonality of variants will allow for improved prioritization of predicted neoantigens. The SNV/InDel workflow is visualized in Figure 5.

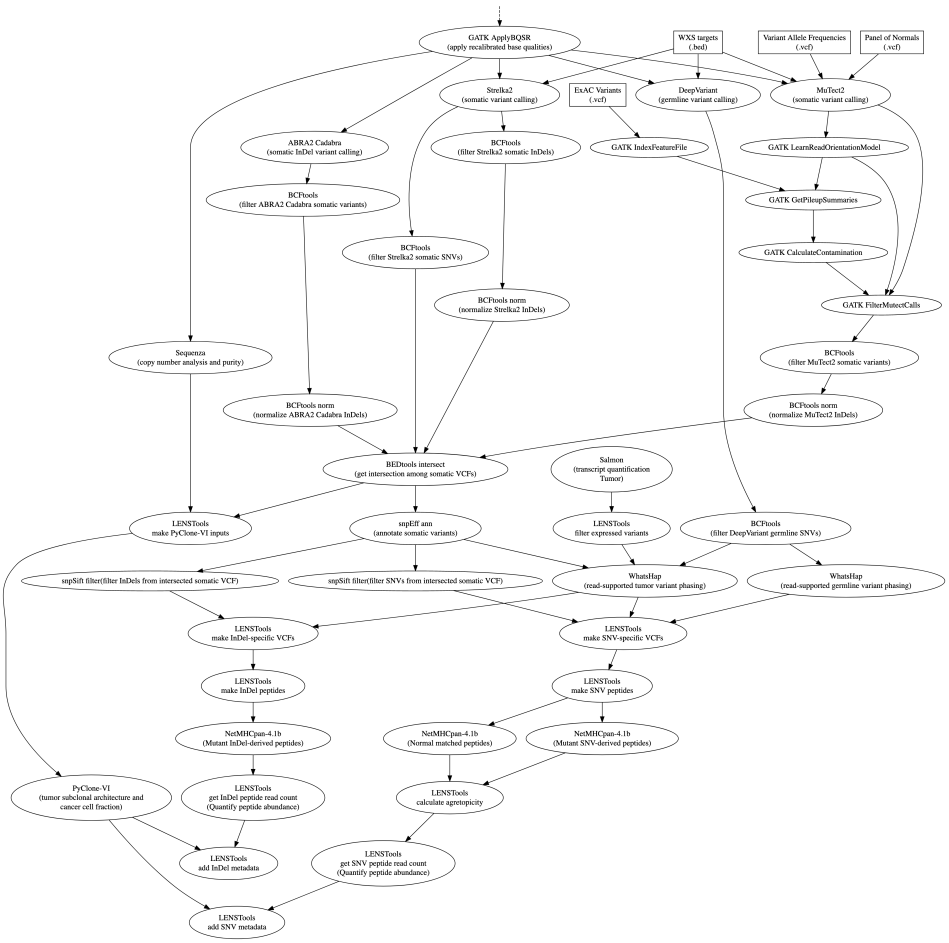

Figure 5: Single Nucleotide Variant, Insertion, and Deletion workflow: Somatic variants are called using three separate variant callers: MuTect2, Strelka2, and ABRA2. Germline variants are called using DeepVariant. Somatic variant harboring transcripts are filtered for expression, and the somatic variant and neighboring phased germline variants are incorporated into its coding sequence prior to peptide generation.

Many SNV and in-frame InDel neoantigens have shown limited utility for inducing a strong immune response [1]. SNV/InDel neoantigen therapies have focused on tumor types known to have high tumor mutational burden such as melanoma. These therapeutic approaches may not be appropriate for patients with lower tumor mutational burden (TMB) tumors, such as acute myeloid leukemia (AML). Furthermore, SNVs and InDel neoantigens tend to be private to individuals which limit "off the shelf" neoantigen vaccines. All of these factors suggest other tumor antigen sources beyond somatic variants are worthy of further consideration.

2.8 Splice Variants

Aberrant splicing events create tumor-specific transcripts through intron retention, exon skipping, or altered splice targeting (Figure 6). These transcripts may, in turn, be translated and processed into targetable tumor-specific peptides. Ninety-four percent of human genes have intronic regions [33] and most undergo alternative splicing to generate transcriptomic and proteomic diversity in order to fulfill a wide variety of protein functions [34]. Alternative splicing is influenced by RNA structure, chromatin structure, and transcription rate, and two major drivers of alternative splicing are cis-acting mutations which disrupt or generate splice sites and trans-acting mutations that alter splicing factors to cause aberrant splice variants in other genes [21]. Somatic mutations of splicing machinery are common in several tumor types and may lead to tumorigenesis by impacting genes that control proliferation, apoptosis, angiogenesis, and metabolism [33]. Shared spliceosome protein mutations have been observed across several types of hematologic malignancies, driving interest in alternative antigen types [21]. However, recent analysis of The Cancer Genome Atlas data shows alternative splicing events in many tumor types including solid tumors, demonstrating the relevance of these antigens in all tumor types [35]. Splice variants also cause about twice as many neoepitopes per event compared to non-synonymous SNVs, making them good neoantigen sources for targeting [36]. Targetability of splice variant antigens has been demonstrated *in vivo* in mice, with a CD20 splice variant upregulated in B cell lymphomas triggering CD4 and CD8 T cell response upon peptide vaccinated in HLA-humanized transgenic mice [37].

LENS utilizes NeoSplice, a graph-based k-mer searching algorithm, to detect tumor-specific splice variants and neoantigens derived from them. Specifically, NeoSplice considers both patient-specific splice variants derived from their tumor RNA sample relative to splice variants from a tissue-matched normal RNA sample. Differences in k-mer distributions between the tumor and normal splice variants allow for high confidence detection of tumor-specific splice variants. Peptide abundance is estimated by counting the number of reads containing the complete peptide's coding sequence mapping to the splice variant genomic origin. The splice variant workflow is visualized in Figure 7.

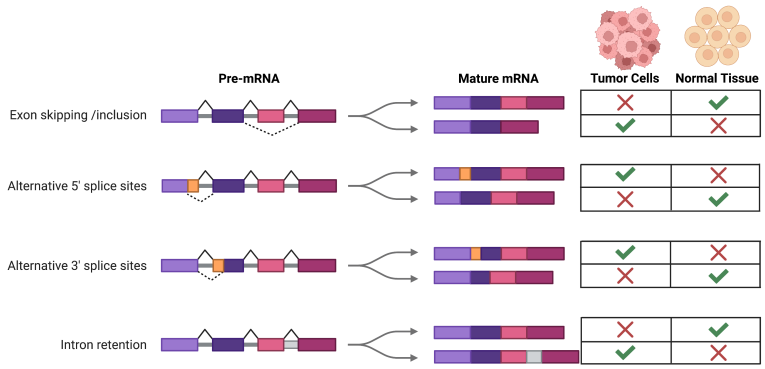

Figure 6: Examples of tumor splice variants: Splice variants can present as skipped exons, retained introns, or alternate splice sites. *Adapted from "mRNA Splicing Types", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates*

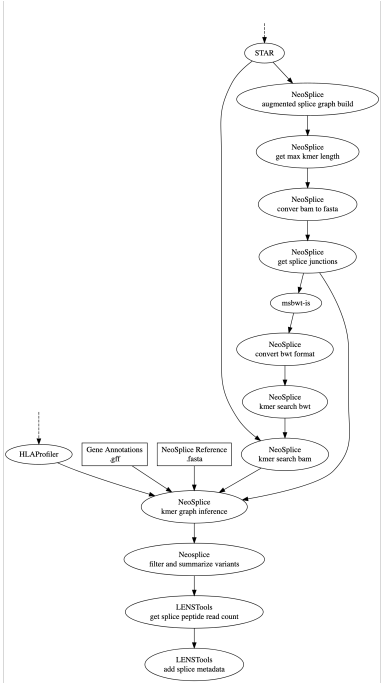

Figure 7: Splice variant tumor antigen workflow: The splice workflow relies upon NeoSplice for splice variant detection. NeoSplice considers k-mer distributions between the patient's RNA tumor sample and a tissue-matched control sample.

2.9 Fusion Events

Scenarios in which a translocation, deletion, or inversion causes two previously genomically distant loci to become neighboring sequences or "fuse" within the genome are also of interest in sources of tumor antigens. The combination of two naturally occurring intragenic coding sequences can create novel peptides that may serve as immunogenic epitopes. This effect may be amplified when fusion events result in reading frame frameshifts. Fusions can create novel open reading frames, and peptides that span the breakpoints between the two genes will be novel sequences that are similar to and as potentially as immunogenic as splice variants [38], theoretically moreso than SNV-derived neoantigens [39]. Fusion genes are often crucial tumorigenesis factors and therefore are often utilized as molecular targets for diagnosis, staging, and treatment [40]. Fusion-positive cancers overall have lower total tumor mutational burden, but some fusion products are known oncogenic drivers [39]. These oncogenic driver gene fusions are often conserved within certain cancer types, such as subtypes of leukemia and sarcoma. For example, t(11;22)(p13;q12) fusion mutations are characteristic for synovial sarcoma and are present in 100% of cases [38]. These highly conserved fusions are potential targets for shared neoantigen targeting strategies. The drug Imatinib which targets *BCR-ABL1* fusions was the first treatment targeting gene fusions, and was shown two decades ago to be effective for CML patients that failed interferon alfa treatment [41]. Many additional drugs have since been developed against *BCR-ABL1*, and novel drugs targeting *ALK*, *TRK*, *ROS*, and *FGFR* fusions have been added to the cancer therapeutic repertoire [38]. Clinical trials are also underway for many novel identified fusions. However, druggable fusions only account for 6 percent of cases, emphasizing the need for additional therapeutic development against fusion genes. Tumor vaccines for fusions have been tested for CML and sarcomas. CML patient given *BCR-ABL1* peptide vaccines developed antigen-specific T cell responses, but no clinical response, and many single agent fusion gene vaccine trials since have seen similar results [42]. Cancers with fusions are often very aggressive, so it is widely believed that tumor vaccines as single agents are insufficient to generate a clinical response in these cancers, but recent studies have sought to combine vaccines with anti-PD1 or anti-PDL1 antibodies to augment response [38]. Adoptive cell therapy is also a promising new approach for targeting fusion genes, with fusion peptide-specific TCR transduced CD8 T cells showing anti-leukemic activity *in vitro* [43].

LENS utilizes STARFusion for detection of fusion events using recommended parameters as defaults [44]. Similarly to the SNV and InDel workflows, homozygous germline variants are incorporated into the fuse coding sequence prior to translation using phased germline VCFs. Proteins derived from in-frame fusion events are truncated upstream to a user-specified length and downstream of the fusion junction and processed through NetMHCpan-4.1b. Frameshift-derived proteins are truncated upstream of the fusion junction and include all downstream sequence until the first stop codon. Peptides are filtered and quantified by checking for their coding sequence within the reads that either map across

a junction point or are map to both sides of the junction point. The fusion workflow is visualized in Figure 8.

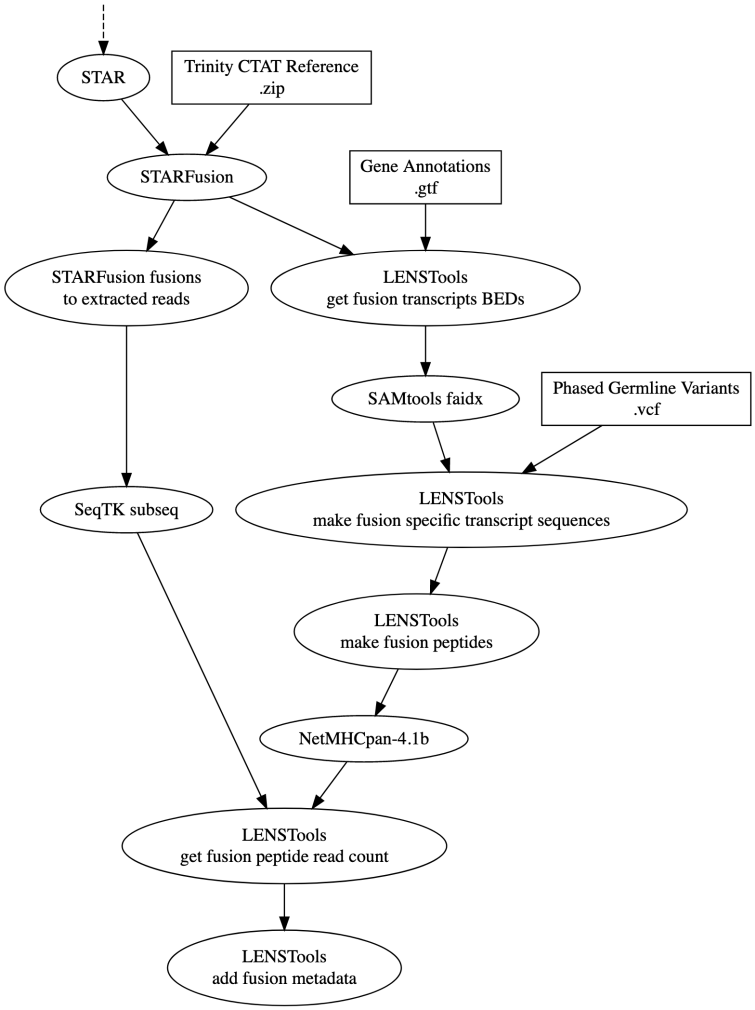

Figure 8: Fusion variant tumor antigen workflow: The fusion workflow utilizes STARFusion for fusion detection. Fused transcripts have homozygous germline variants integrated into their sequence prior to translatin.

2.10 Tumor-specific Viruses

Some viruses such as human papillomavirus (HPV), Epstein-Barr virus (EBV), Kaposi's sarcoma-associated herpesvirus (KSHV or HHV8), human T cell leukemia/lymphoma virus type 1 (HTLV1), hepatitis B virus (HBV), and hepatitis C virus (HCV) are

associated with development of cancers in the tissues they infect or reside within, earning them the label of oncogenic virus [45, 46, 47]. Viruses can drive oncogenesis through a variety of mechanisms, including disrupting host cell growth and survival, inducing DNA damage response causing host genome instability, causing chronic inflammation and tissue damage, or causing immune dysregulation creating a more permissive immune environment for tumorigenesis [48]. Antigens derived from the virus are regarded as tumor-associated antigens, yet are immunologically more foreign compared to self-derived peptides, making them good potential immunotherapeutic targets. While traditional cancer therapies cannot differentiate between infected and uninfected cells, immunotherapies can be targeted to the viral antigens in order to specifically kill virally infected cancer cells. One mechanism of immunotherapy for oncogenic viruses is co-opting a preventative vaccine for therapeutic use in the virus-associated cancer. For example, administration of HPV vaccines in patients with HPV-associated cancers activate CD8+ T cells to kill infected cancer cells [49]. Adoptive cell transfer is also being explored as a therapeutic strategy for cancers associated with oncogenic viruses. EBV-specific T cells adoptively transferred into patients with established EBV-associated tumors such as lymphomas and nasopharyngeal carcinoma have shown efficacy [48]. There is an ongoing clinical trial with a similar strategy for HPV, in which adoptive transfer of polyclonal HPV-specific T cells is being evaluated for treatment of HPV-associated malignancies of the head, neck, and anogenital regions [50]. Residual viruses present at elevated levels within tumors may also serve as vaccination or immunotherapy targets.

LENS detects viral-derived tumor antigens with a previously developed workflow, VirDetect, which was initially developed to detect viral contamination within bulk samples analyzed by RNA sequencing [51]. RNA reads are mapped against the reference genome using textsfSTAR, and any reads that cannot be mapped to the reference are diverted to the VirDetect workflow. VirDetect aligns these unmapped reads to coding sequences of over 1,900 vertebrate viruses contained within the VirDetect reference. Homozygous germline variants detected within the RNA sequencing data through BCFtools are incorporated into the viral sequence prior to translation. Peptide sequences resulting from translation of the viral coding sequences are considered as potential tumor antigens for downstream processing. Peptides are quantified by counting occurrences of peptide-associated coding sequences within the reads that map to expressed viral coding sequences. The viral workflow is visualized in Figure 9.

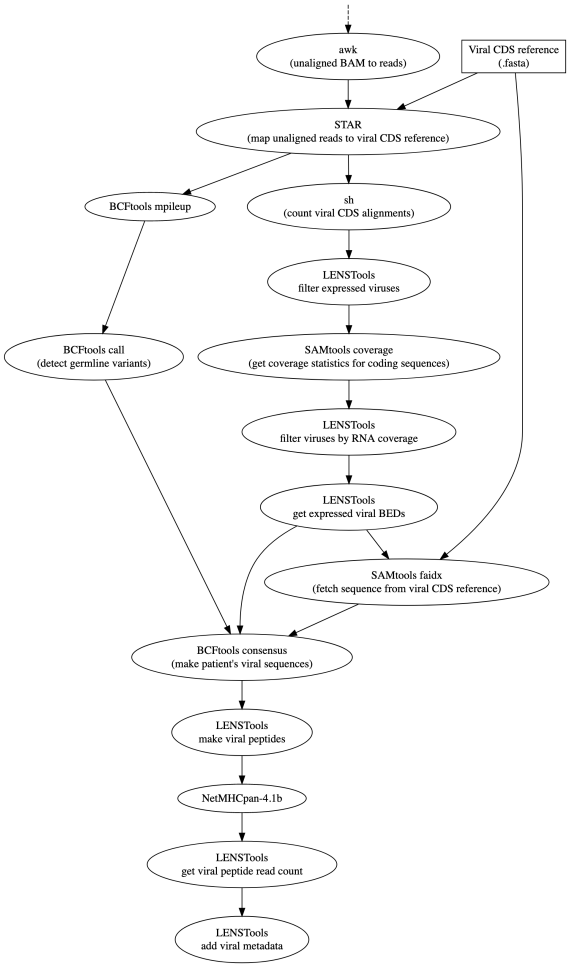

Figure 9: Viral tumor antigen workflow: The viral workflow uses the VirDetect reference and alignment strategy to detect viruses expressed within the patient's tumor. Homozygous germline variants are integrated into the viral sequence prior to translation.

2.11 Endogenous Retroviral Elements

hERVs (human endogenous retroviruses) are, unlike typical viruses, integral to the human genome and are believed to make up to 8% of the genome [52]. Some retroviral elements have retained intact open reading frames that may be transcribed, translated, processed, and presented by the MHC on the cell surface under abnormal transcriptional regulation within a tumor. Their lack of central tolerance, elevated abundance, and ability to be expressed under chaotic tumor conditions make endogenous retro viruses intriguing potential

sources of tumor antigens. Unsurprisingly, hERVs have been a focal point of tumor antigen research due to evidence suggesting CD8+ T cell recognition of ERV-derived peptides results in an immunogenic response [53, 54].

LENS detects ERV-derived candidate peptides through use of the gEVE (genome-based Endogenous Viral Element) database which includes tens of thousands of computationally predicted retroviral element open reading frames (ORFs) from which antigens may be generated [20]. Expressed ORFs have homozygous germline variants integrated into their sequences prior to translation. We assume some ERVs may have natural low levels of expression in some normal tissues, so differential expression filtering is used to narrow the considered pool (Figure 1). Specifically, tissue-matched normal control samples are processed through an ERV quantification workflow and the resulting raw counts are normalized to patient-specific ERV counts through EdgeR, and ERV ORFs that exceed a specific fold-change threshold are considered for downstream processing. ERV peptides are quantified by counting peptide-coding sequences among reads that map to expressed ORFs. The ERV workflow is visualized in Figure 11.

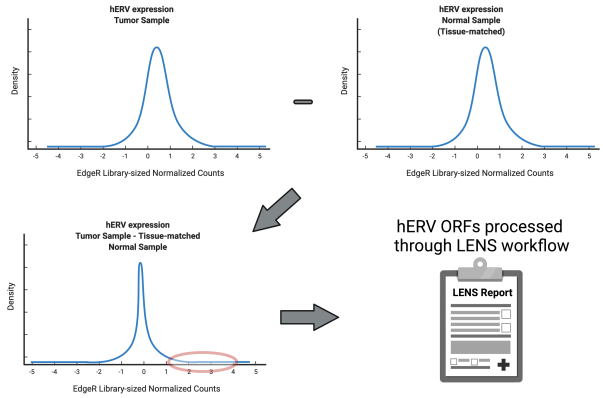

Figure 10: ERV expression filtering strategy: ERVs are filtered by comparing the tumoral expression to expression in a tissue-matched normal sample. Only ERVs that show higher expression relative to the normal tissue are considered for processing. *Created with BioRender.com.*

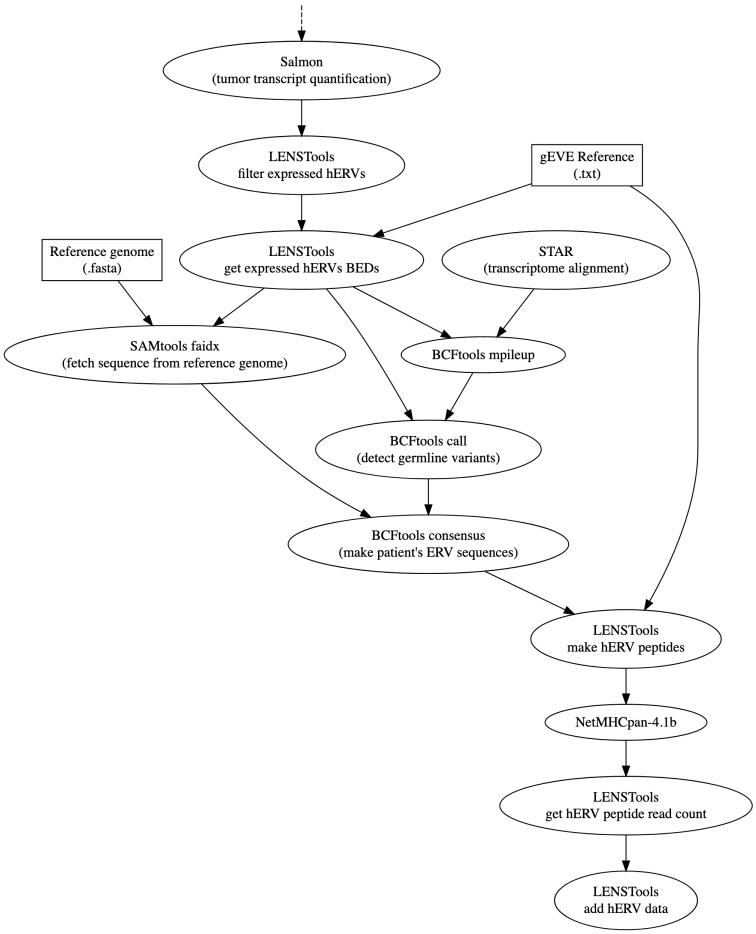

Figure 11: ERV tumor antigen workflow: The ERV workflow utilizes predicted ERV ORFs from the gEVE database. Homozygous germline variants are integrated into the coding sequences prior to translation.

2.12 Cancer Testis Antigens and Self-Antigens

Tumor self antigens are an additional potential target for immunotherapy. These antigens are non-mutated self antigens that may be overexpressed within a tumor. These antigens are traditionally viewed as suboptimal therapeutic targets for two reasons: (1) T cell recognizing self antigens would be expected to be deleted in the thymus, and (2) autoimmunity is difficult to avoid when targeting antigens expressed in both normal and tumor tissues. However, work has been done to identify methods of self antigen targeting while minimizing autoimmunity. Some endogenous immune response has been identified against self antigens in cancer patients, with one group identifying a self-antigen that

had higher epitope-specific CD4+ T cell responses within metastatic malignant melanoma patients compared to normal healthy individuals [55]. The pool of self antigens to select from for targeting is also large, as any given tumor will present many self antigens, making them potential vaccine targets [56, 57]. Some self antigens are prognostic markers and drug targets for a variety of cancers, such as HER-2/neu, an overexpressed antigen in breast, prostate, and lung cancer that is a frequent therapeutic target [57]. Early work shows that HER-2/neu could function as a cancer vaccine target, with generation of HER-2/neu specific T cells that did not induce autoimmunity [58]. Adoptive transfer of self antigen specific cells is another potential targeting strategy, with one mouse study showing that adoptive transfer of CD8+ T cells with TCRs modulated to have low affinity for self antigen mOVA induced transient regression of OVA positive tumors in mice without inducing autoimmunity [59]. Genes normally expressed in the immune privileged tissues, such as testis tissue, may be transcriptionally active in tumors. Antigens derived from these transcripts are commonly referred to as cancer testis antigens (CTAs) and are of particular interest due to their potential immunogenicity. These antigens are derived from normal genes that are usually expressed either during early development or in adulthood only within immune privileged tissues such as the testis. They may become overexpressed within tumors such as melanoma, breast cancer, or bladder cancer, and represent a promising target due to their lack of expression in normal tissues undergoing immune surveillance [60]. Care must be taken during selection of CTA and other self-antigen-derived peptides as even low expression in normal tissues can result in on-target, off-tumor effects causing a negative impact on patient health [61].

LENS accepts a user-provided list of CTA and self-antigen gene identifiers, but defaults to a set of cancer testis genes from the CTDatabase [62]. This list includes genes that are germline-biased and highly expressed in cancers, but have not necessarily to be shown initiate an immune response when targeted. The CTA and self-antigen transcripts are first filtered for transcription abundance exceeding a user-specified percentile threshold. Transcripts passing the filter then have germline homozygous germline variants incorporated and the resulting predicted peptides are run through NetMHCpan-4.1b. High binding affinity peptides are quantified by counting occurrences of peptide-coding sequence observed within reads mapping to the peptide's genomic origin. The CTA/Self-antigen workflow is visualized in Figure 12.

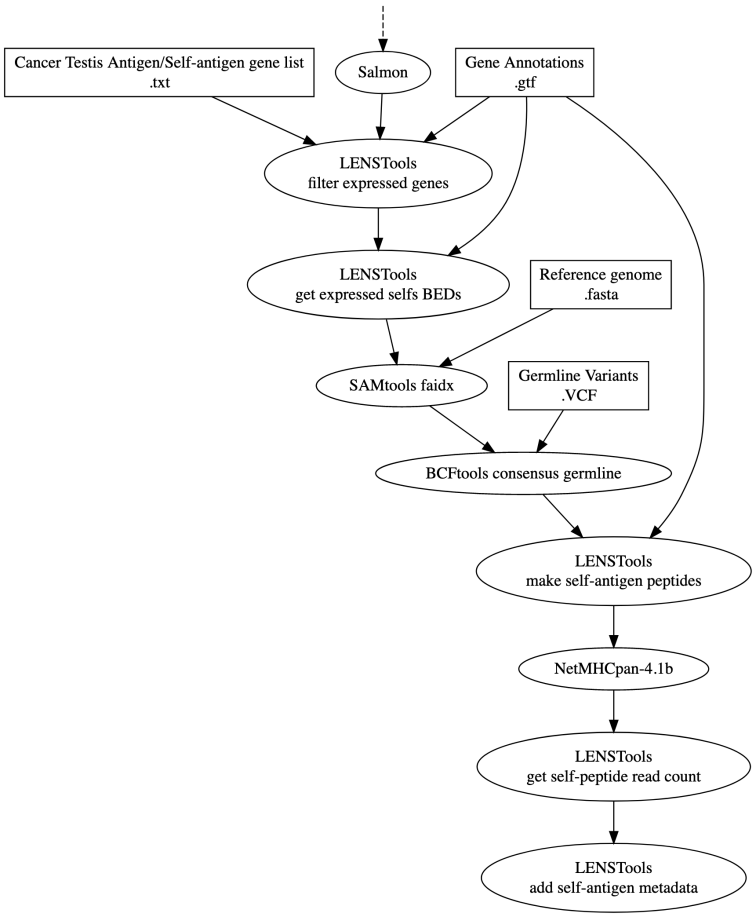

Figure 12: CTA/Self-antigen tumor antigen workflow: Cancer testis antigens and self-antigens are processed through the same workflow. Specifically, they are filtered for expression and have homozygous germline variants integrated into their coding sequences prior to further processing.

2.13 Harmonization of Peptide Abundance Estimates

The tumor antigen workflows within LENS have a variety of quantification metrics associated with the individual workflows. These metrics include Transcripts per Million (TPM) for transcript quantification in the SNV and InDel workflows, Fusion Fragments per Million (FFPM) for fusion quantification the fusion workflow, minimum observed expression for the splice variant quantification in the splice workflow, and read count for viral expression in the viral workflow. Each metric described may be broadly relevant to tumor antigen quantification, but most have at least two issues: 1) they conflate abundance of the peptide-

generating coding sequence with abundance of its neighboring context and 2) they do not allow meaningful comparisons of peptide abundance among workflows. As an example of the first issue, consider TPM, a metric commonly used to represent relative transcript abundance. TPMs have been used as a filtering criterion for SNV and InDel neoantigens, but they represent the abundance of the transcript (rather than just the peptide-generating subsequence) and may include alleles that do not code for the peptide of interest. To address this, we developed a novel quantification strategy utilizing the observed occurrences of the nucleotide sequence responsible for coding the peptide from the peptide's genomic origin from the patient's tumor RNA sequencing data. This metric resolves both of the limitations of currently used metrics and allow for prioritization by abundance across multiple tumor antigen types.

3 Results

3.1 Applying LENS to TCGA-LAML

The design principles of LENS allow it to avoid constraints around specific tumor types or tumor antigen sources. We demonstrate LENS by processing data from several patients of the widely available TCGA-LAML (Acute Myeloid Leukemia) dataset. Acute myeloid leukemia is suitable for demonstrating LENS due to its relatively low tumor mutational burden compared to other tumor types. This low TMB presents a situation where non-SNV- and InDel-derived neoantigens are crucial. The TCGA-LAML dataset also contains sufficient patient-level data including the three sample types (normal exome sequencing, abnormal (tumor) exome sequencing, abnormal (tumor) RNA sequencing) for 115 patients. Here, we discuss the results for each tumor antigen workflow currently available from LENS. This is not intended to be an exhaustive analysis of LENS outputs, but rather to provide an example of its utility in tumor antigen prediction using genomics data.

3.2 General Observations among Tumor Antigen Sources

TCGA-LAML patients show a range of predicted tumor antigen counts from 17 to 1,153 (median: 177) (Figure 13). Predicted antigen counts among patients vary by genomic source. Specifically, CTA/Self-antigen and ERV peptides made up over 20,000 total predicted peptides among patients (20,676/22,563 (92.7%)) while the remaining sources (SNVs, InDels, fusion events, splice variants, and viruses) included up 1,887 peptides (8.3%) (See Supplemental Table 1). This disparity is likely explained by two factors: 1) ERVs are both highly abundant within the genome and 2) the entirety of each CTA/Self-antigen's coding sequence is used for peptide generation, so the large amount of "raw material" translate to more potential peptides (even after binding affinity filtering, see Figure 14). Binding affinities among antigen sources had medians between 100 - 200 nanomolar with the exception of CTA/Self-antigens due to the manual

filtering to peptides with binding affinities under 25 nanomolar (Figure 14). Relative peptide abundance (as calculated by peptide occurrence within RNA sequencing reads) showed variability among antigen sources as well, with viruses and fusions having the lowest relative abundance (Figure 15).

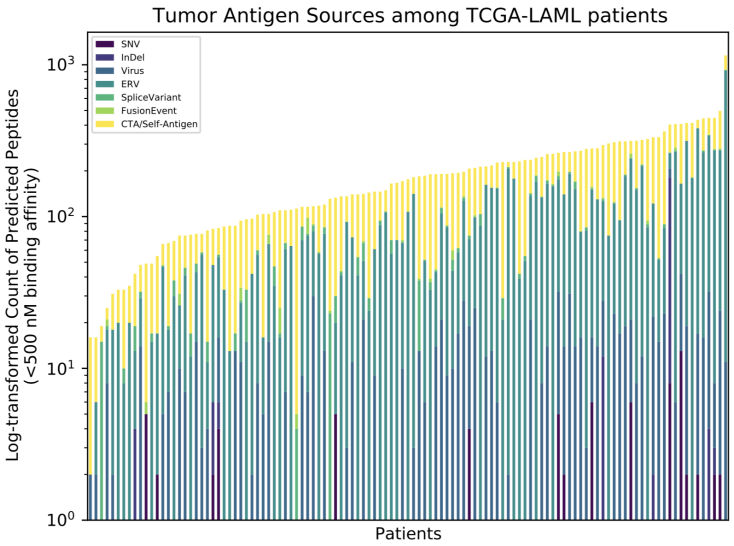

Figure 13: TCGA-LAML tumor antigen distribution by patient: Tumor antigen distributions across 115 TCGA-LAML patients. Predicted neoantigens range between 17 to 1,153 peptides with a median of 177 peptides per patient.

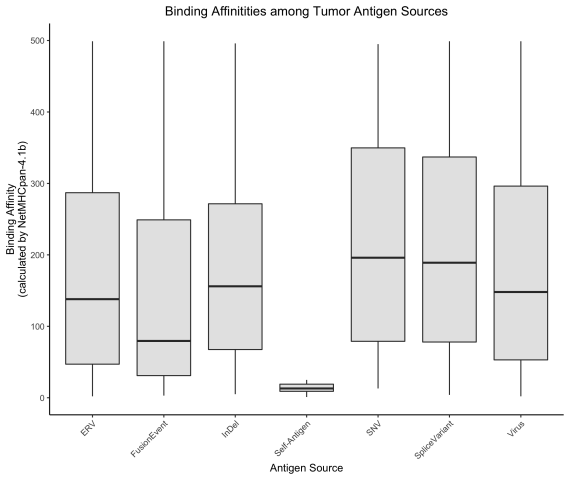

Figure 14: Binding affinities by tumor antigen source among TCGA-LAML patients: All tumor antigen sources showed binding affinity medians between 100 to 200 nanomolar. CTAs/Self-antigens shows markedly lower binding affinity due to the stringer filtering of affinities below 25 nanomolar to compensate for the larger number of peptides generated.

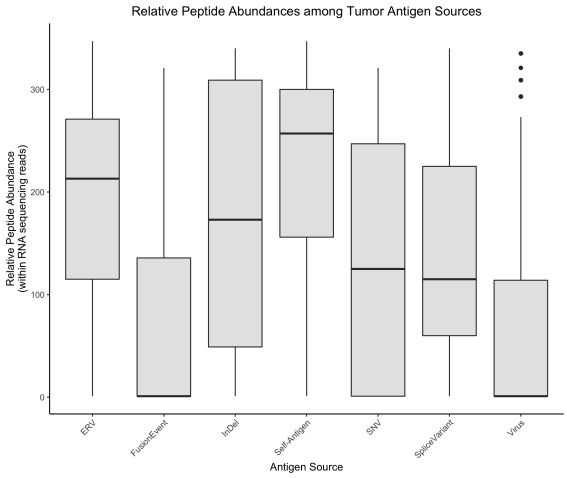

Figure 15: Relative peptide abundance by tumor antigen source among TCGA-LAML patients: Relative peptide abundance (measured by peptide coding sequence detection in RNA sequencing reads) varies across antigen source. Notably, both fusion events and viruses show low relative abundances compared to other antigen sources.

3.3 SNV and InDel Neoantigens in TCGA-LAML

Despite low tumor mutational burden in AML, some SNV neoantigens have been previously identified. Greif et al. reported five tumor-specific nonsynonymous mutations in a single AML sample via whole transcriptome sequencing [63]. InDels have also been identified within AML. For example, Lee et al. reported a novel InDel within the *KIT* gene, a commonly mutated gene in AML patients, in a clinical sample from a 35 year old female [64]. Two neoantigens derived from *NPM1* mutations have been identified [65, 66]. Variant calling was performed on TCGA-LAML samples using a consensus approach in which each variant had to be detected by at least two separate variant callers (for SNVs) or three separate variant callers (InDels) to be considered. Furthermore, variants had to exceed relative transcript abundance upper quartile and the nucleotide sequence surrounding the SNV or InDel variant of interest must also have been discoverable within the patient's corresponding RNA sequencing reads. As a result, a patient's SNV and InDel neoantigen peptide count is expected to the proportional to, but less than, the total number of high confidence somatic variant calls observed within the DNA data.

We discovered potentially targetable somatic SNVs in 97 of 115 patients with a median of 3 SNVs (min:1, max: 460). These SNVs translated into SNV neoantigen targets for 70 patients with a median of 1 SNV-derived neoantigen peptide (min: 1, max: 180). Expectedly, none of the genes harboring SNV neoantigens were shared among patients. These observations are not surprising given the previously observed low somatic mutation rate within AML patients [67].

InDel-derived neoantigens showed a similar pattern among patients with 63 or 115 patients having potentially targetable InDels with a median of 1 InDel per patient (min: 1, max: 115). Twelve of the 63 patients had InDel-derived peptides predicted with a median of 4 peptides per patient. Interestingly, one patient showed 230 InDel-derived peptides. Over 83% of these peptides appear to be a product of a frameshift mutation in *TNKS2*. This provides an example of the potency of frameshift mutations in the context of peptide generation. All loci that generated InDel-derived peptides were private with the exception of *KIT*, a locus previously shown to be mutated in AML, which was observed in two separate patients.

3.4 Splice Variant Neoantigens in TCGA-LAML

Given the relatively low mutational burden observed within AML samples, other potential tumor antigen sources are of great interest. Splice variants associated with mutations in splicing-associated proteins have previously been identified for AML, including possible driver mutations in genes such as *L-Myc* and *PTPN6* [68]. We tested patients for tumor-specific splice variants using NeoSplice coupled with a non-matched normal blood sample[**NEOSPLICE˙REF**].

Our NeoSplice analysis discovered splice-derived tumor antigens in 80 of the 115 patients with a median splice tumor antigen count of 5 (min: 2, max: 26).

Several loci harboring tumor splice antigens tended to be shared among patients, including *TRPM2* (12 patients), *ANKRD36* (9 patients), *STAB1* (6 patients), among others. Previous work showed one of our splice antigen-generating loci, *Myeloperoxidase* (*MPO*), can have a splice site substitution variant that results in the retention of intron 11 in some AML patients [69]. Another potential target, *Transient Receptor Potential Melastatin 2* (*TRPM2*), observed to have a single neoantigenic peptide, **QVAQTARAL**, observed in twelve separate patients. *TRPM2* has been shown to previously be highly expressed in AML, potentially in the context of allowing for increased tumor survival [70].

Several other genes, *SLCA11A1*, *GZMA*, and *NCF1*, were observed to have multiple tumor-associated splice variant peptides across multiple patients. Their cellular functions, however, suggest they may be associated with an immune response rather than being expressed by the tumor[71, 72].

3.5 Fusion Neoantigens in TCGA-LAML

The ability of fusion genes to create new, tumor-specific transcripts coupled with the potential for frameshifting effects on the protein product make them appealing targets for neoantigen predictions. A recent large-scale study of in-frame fusion events across 539 AML cases documented 296 fusion genes of which 57 shown to be recurrent. Chen et al.'s results focused on in-frame fusion events, but highlight two important factors for fusion-derived neoantigens: 1) fusion events are not rare within AML cases (82.2% of all acute leukemia cases considered had at least one high-confidence fusion events) and 2) recurrent fusion genes suggest the potential for vaccination of shared fusion events [40]. LENS can help with understanding the relationship between shared variants among patients and which neoantigens are worth pursuing.

Our TCGA-LAML fusion-derived neoantigen workflow identified a fusion gene observed among six patients (*RARA-PML*) and 2 fusion genes observed among two patients (*ABL1-BCR* and *UBE2Q2-FBX022*). The *RARA-PML* fusions resulted in a patient with 8 frameshift-derived peptides and 7 in-frame-derived peptides, 3 patients with between 1 and 3 in-frame-derived peptide, and a patient with 11 frameshift-derived peptides.

Eleven fusion genes (*SUGT1-NOL6*, *SLC45A3-ELK4*, *SEC63-FIG4*, *RUNX1-MECOM*, *PICALM-MLLT10*, *NF1-LRRC37B*, *KMT2A-ELL*, *KMT2A-AFDN*, *CHD1-MTOR*, *NUP98-NSD1*, *CDC73-FAM271A*, and *ZEB2-CXCR4*) were private to individual patients and previously observed within cancers [40]. Many fusion-derived peptides showed minimal read support, but several showed both elevated read support and Class I binding affinity. Caution must be taken when comparing peptide read support among patients as the metric is not normalized by total read count; however, a cursory analysis among patients highlights some potential candidates (Figure 16).

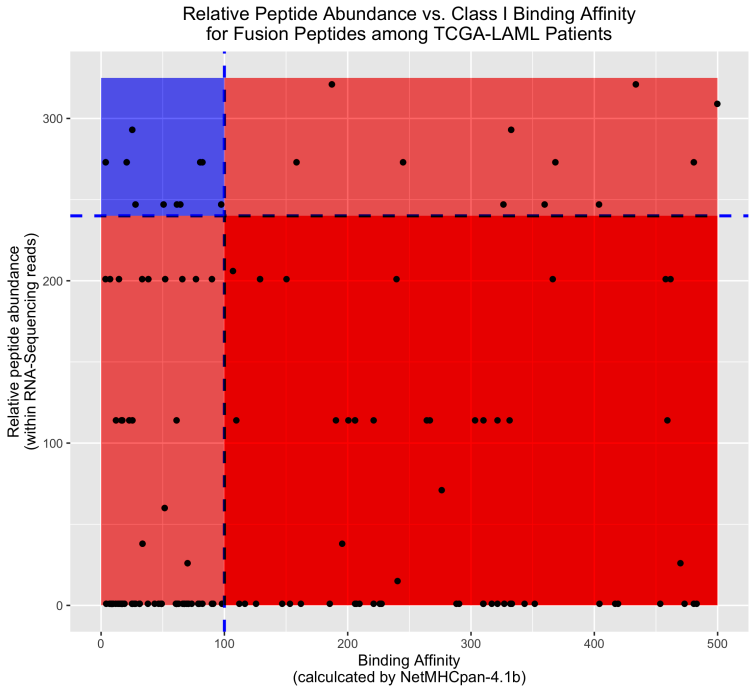

Figure 16: Fusion peptide read support vs. Class I binding affinity across patients: Several predicted fusion-derived peptides show both relatively high binding affinity and high relative peptide abundance (blue). Others show either high expression and low affinity or low expression and high affinity (light red), or both low expression and low affinity (red).

3.6 Viral and Endogenous Retroviral Tumor Antigens in TCGA-LAML

Oncogenic viruses that are highly expressed within a tumor may serve as a suitable target for tumor vaccination. AML is not canonically associated with viral oncogenesis, however viral and endogenous retroviral elements remain potential sources of tumor antigens in AML. Clinical studies investigating the relationship between hERVs and AML are ongoing (NCT04406207) and recent work using single-cell RNA sequencing discovered a variety of viral and retroviral elements expressed in AML tumors [73].

We discovered some level of retroviral expression in 105 patients. Of these, 22 patients expressed ORFs associated with single ERV-associated proteins, 31 patients expressed ORFs associated with two ERV-associated proteins, and 50 expressed ORFS from three or more ERV-associated proteins. The most abundant hERV type observed among patients was hERV-K10 followed by hERV-E.

Both the current hERV and viral workflows produce a multitude of predicted tumor antigen peptides due to the entirety of the viral or retroviral sequence

being included. Further work is required to more rigorously prioritize and validate both the expression of these elements and their potential contribute to tumor antigen vaccine strategies.

3.7 Cancer Testis Antigens and Self-antigens in TCGA-LAML

Multiple CTAs have been previously reported for AML, including *Cyclin A1*, *MAGE*, *PASD1*, *PRAME*, and *RAGE-1* [74]. The Cancer Testis Antigen Database (http://www.cta.lncc.br/) provides a variety of well-documented testis-specific or testis-elevated expressed transcripts that may serve as suitable vaccination targets within tumors expressing them. We considered the relatively expression value of all transcripts associated with genes in the Cancer Testes Database and required CTA transcripts to be relatively highly expressed (above 90th percentile) in a patient's tumor RNA data. Several CTA loci were highly expressed in a number of patients. This includes 113 patients expressing *KIAA0100*, a CTA expressed in hematological cancers and an association with an immune response in AML [75]. Another potential target, *DCAF12*, was observed to be expressed in 108 patients. *DCAF12* is associated with immune response in some tumor types, but is also involved in erythoid differentiation which suggests it may be a very poor target in AML [76]. This highlights the importance of considering tumor context when evaluating antigen sources. Two other intriguing targets, *PRAME* and *SPAG6*, were discovered in 20 patients and 7 patients respectively. *PRAME* expression has been shown to have an immunogenic response in AML [77] while *SPAG6* has been observed in AML previously [78]. Notably, CTAs and self-antigen peptides can be generated by the entirety of each transcript's coding sequence. As a result, patients typically have several magnitudes more CTA peptides than other types and will require more stringent filtering prior to prioritization.

The above rudimentary analysis emphasizes the importance of the considering a wide breadth of tumor antigen source when prioritizing antigens for vaccination. AML is an extreme case due to its relatively low TMB, but other tumor types would still require inclusion of tumor antigen sources beyond the standard SNV, InDel, and Fusion antigens to comprehensively map the expressed antigen landscape.

4 Discussion

Predicting the full suite of tumor antigens from genomics data remains a formidable challenge. We developed LENS, coupled with the RAFT framework, to improve upon current workflow offerings in a variety of ways: 1) LENS allows for modularity and extendibility through the Nextflow DSL2 language, 2) LENS explores more tumor antigen sources than previously considered by other workflows, and 3) LENS harmonizes peptide abundance estimates among various tumor antigen sources. LENS is open source, and we expect it to grow and expand through

both introduction of new modules, tools, and reference datasets, along with refinement of its reporting functionality and tumor antigen prioritization capabilities. There are several objectives discussed below concerning these goals.

A primary objective for the improvement of LENS is expanding the set of supporting sequencing technologies. LENS currently supports data generated through short-read technologies such as the Illumina platform . Illumina's sequencing-by-synthesis (SBS) chemistry sees widespread usage and support, both within the immuno-oncology community and more broadly throughout the life sciences research community. Despite its popularity, the short read and bulk cell approaches commonly used result in reduced information compared to long read and single cell approaches. We aim to address these limitations within LENS by supporting long read and single-cell sequencing technologies. Long read sequencing technologies allow improved resolution of structural variation, such as large insertion, deletions, or inversions, that may be relevant to neoantigen prediction. Additionally, single cell sequencing circumvents the confounding effects of bulk sequencing which allows improved understanding of tumor heterogeneity. Empirically deciphering this heterogeneity is crucial to properly prioritizing clonal tumor antigens while also mapping co-occurring subclonal neoantigen to optimize peptide targeting.

We plan to further expand LENS through inclusion of third-party reference data and additional bioinformatics tools to provide information about the support (or lack thereof) of a peptide's immunogenicity. The more important and immediate addition will be the inclusion of Class II binding affinity prediction. We also plan to extract relevant summarized data from large datasets (such as considering the relative expression of a SNV or InDel-harboring transcript within the appropriate TCGA cohort), additional tools or smaller scale dataset-specific observations in the literature.

Beyond the inclusion of technologies and data, there is also room for improvements within LENS in its current form. LENS supports a variety of tumor antigen sources, but currently it effectively treats these workflows independently. This independence among tumor antigen workflows in LENS also does not allow potentially useful "crosstalk" between workflows. For example, LENS may be able to provide higher confidence for potential splice antigen targets if there is evidence of somatic splice site variant in the corresponding variant data. LENS could also be extended in the future to report information about relevant tumor immune microenvironment features that can be computed from the input DNA and RNA sequencing data. While LENS does not solve all problems in the field of tumor antigen prediction, we hope that its breadth of features, flexibility, modularity, and ease of use will support wide adoption as a springboard towards iterative improvements as more data, tools, and an improved understanding of peptide presentation and immunogenicity become available.

Workflow	Inputs	MHC class support	Requires pre-processing	Workflow manager	Containerization	Harmonized Peptide Abundance Estimate	Modular (See caption)	Somatic SNVs	Somatic in-frame InDels	Somatic frameshift InDels	Fusion Events	Splice Variants	Endogenous Retroviruses	Viruses	CTAs/ Self-Antigens
LENS	DNA tumor FASTQs DNA normal FASTQs RNA tumor FASTQs	Class I	✗	RAFT/Nextflow DSL2	Process-level	✓	Supported through Nextflow DSL2	✓	✓	✓	✓	✓	✓	✓	✓
pVACtools	Annotated, phased VCFs (pVACseq) Peptide FASTA (pVACbind) Annotated Fusions (pVACfuse)	Class I & II	✓	N/A	✗	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗
OpenVax	DNA tumor FASTQs DNA normal FASTQs RNA tumor FASTQs	Class I	✗	SnakeMake (optionally within Docker container)	Workflow-level	✗	Supported through SnakeMake	✓	✓	✓	✓	✗	✗	✗	✗
nextNEOpi	DNA tumor FASTQs or BAM DNA normal FASTQs or BAM RNA tumor FASTQs or BAM (optional)	Class I & II	Optional (short read alignment)	Nextflow DSL1	Hybrid (See caption)	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗

Table 1: Comparison between LENS and other Neoantigen workflows LENS generally offers similar or improved functionality compared to other popular neoantigen workflows. Specifically, LENS supports more tumor antigen sources, does not require end-user pre-processing of input data, and is a both modular and extensible workflow. It is worth noting that all the included workflows are open-source and may subjectively be classified as modular and extensible as a result. We classify LENS as modular and extensible as these were desired features that drove design and development rather than consequences of code availability. The "Hybrid" classification for nextNEOpi's containerization metric is due to some components, like pVACtools and NeoFuse, having self-contained containers while other tools (e.g. variant callers) are all contained within a single container.

References

[1] Daniel K Wells et al. "Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction". In: *Cell* 183.3 (2020), pp. 818–834.

[2] Christof C Smith et al. "Machine-learning prediction of tumor antigen immunogenicity in the selection of therapeutic epitopes". In: *Cancer immunology research* 7.10 (2019), pp. 1591–1604.

[3] Juliane Liepe, Huib Ovaa, and Michele Mishto. "Why do proteases mess up with antigen presentation by re-shuffling antigen sequences?" In: *Current Opinion in Immunology* 52 (2018), pp. 81–86.

[4] Minying Zhang et al. "RNA editing derived epitopes function as cancer antigens to elicit immune responses". In: *Nature communications* 9.1 (2018), pp. 1–10.

[5] Zach Rolfs et al. "Global identification of post-translationally spliced peptides with neo-fusion". In: *Journal of proteome research* 18.1 (2018), pp. 349–358.

[6] Zhuting Hu et al. "Personal neoantigen vaccines induce persistent memory T cell responses and epitope spreading in patients with melanoma". In: *Nature medicine* 27.3 (2021), pp. 515–525.

[7] Julia Kodysh and Alex Rubinsteyn. "OpenVax: an open-source computational pipeline for cancer neoantigen prediction". In: *Bioinformatics for Cancer Immunotherapy*. Springer, 2020, pp. 147–160.

[8] Jasreet Hundal et al. "pVACtools: a computational toolkit to identify and visualize cancer neoantigens". In: *Cancer immunology research* 8.3 (2020), pp. 409–420.

[9] Dietmar Rieder et al. "nextNEOpi: a comprehensive pipeline for computational neoantigen prediction". In: *Bioinformatics* 38.4 (2022), pp. 1131–1132.

[10] Kevin Litchfield et al. "Meta-analysis of tumor-and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition". In: *Cell* 184.3 (2021), pp. 596–614.

[11] Patrick A Ott et al. "An immunogenic personal neoantigen vaccine for patients with melanoma". In: *Nature* 547.7662 (2017), pp. 217–221.

[12] Ugur Sahin et al. "Personalized RNA mutanome vaccines mobilize polyspecific therapeutic immunity against cancer". In: *Nature* 547.7662 (2017), pp. 222–226.

[13] Frank J Lowery et al. "Molecular signatures of antitumor neoantigen-reactive T cells from metastatic human cancers". In: *Science* 375.6583 (2022), pp. 877–884.

[14] Justina X Caushi et al. "Transcriptional programs of neoantigen-specific TIL in anti-PD-1-treated lung cancers". In: *Nature* 596.7870 (2021), pp. 126–132.

[15] Martin L Buchkovich et al. "HLAProfiler utilizes k-mer profiles to improve HLA calling accuracy for rare and common alleles in RNA-seq data". In: *Genome medicine* 9.1 (2017), pp. 1–15.

[16] András Szolek et al. "OptiType: precision HLA typing from next-generation sequencing data". In: *Bioinformatics* 30.23 (2014), pp. 3310–3316.

[17] Brian J Haas et al. "Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods". In: *Genome biology* 20.1 (2019), pp. 1–16.

[18] Ana Paula Delgado et al. "Open reading frames associated with cancer in the dark matter of the human genome". In: *Cancer genomics & proteomics* 11.4 (2014), pp. 201–213.

[19] Anne-Mette Bjerregaard et al. "MuPeXI: prediction of neo-epitopes from tumor sequencing data". In: *Cancer Immunology, Immunotherapy* 66.9 (2017), pp. 1123–1130.

[20] So Nakagawa and Mahoko Ueda Takahashi. "gEVE: a genome-based endogenous viral element database provides comprehensive viral protein-coding sequences in mammalian genomes". In: *Database* 2016 (2016).

[21] Christof C. Smith et al. "Alternative tumour-specific antigens". In: *Nature Reviews Cancer* 19.8 (2019), pp. 465–478.

[22] Miao Peng et al. "Neoantigen vaccine: an emerging tumor immunotherapy". In: *Molecular Cancer* 18.1 (2019), p. 128.

[23] Wolfgang Lilleby et al. "Phase I/IIa clinical trial of a novel hTERT peptide vaccine in men with metastatic hormone-naive prostate cancer". In: *Cancer Immunology, Immunotherapy* 66.7 (2017), pp. 891–901.

[24] W. Obara et al. "A phase I/II study of cancer peptide vaccine S-288310 in patients with advanced urothelial carcinoma of the bladder". In: *Annals of Oncology: Official Journal of the European Society for Medical Oncology* 28.4 (2017), pp. 798–803.

[25] Anna C. Filley and Mahua Dey. "Dendritic cell based vaccination strategy: an evolving paradigm". In: *Journal of Neuro-Oncology* 133.2 (2017), pp. 223–235.

[26] Beatriz M Carreno et al. "A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells". In: *Science* 348.6236 (2015), pp. 803–808.

[27] Lijin Li et al. "Optimized polyepitope neoantigen DNA vaccines elicit neoantigen-specific immune responses in preclinical models and in clinical translation". In: *Genome Medicine* 13.1 (2021), p. 56.

[28] Nikolaos Zacharakis et al. "Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer". In: *Nature medicine* 24.6 (2018), pp. 724–730.

[29] Vésteinn Thorsson et al. "The immune landscape of cancer". In: *Immunity* 48.4 (2018), pp. 812–830.

[30] Samra Turajlic et al. "Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis". In: *The Lancet Oncology* 18.8 (2017), pp. 1009–1021.

[31] Ide T. Spaanderman et al. "Framing the potential of public frameshift peptides as immunotherapy targets in colon cancer". In: *PLOS ONE* 16.6 (2021), e0251630.

[32] Yvette Garbe, Claudia Maletzki, and Michael Linnebacher. "An MSI tumor specific frameshift mutation in a coding microsatellite of MSH3 encodes for HLA-A0201-restricted CD8+ cytotoxic T cell epitopes". In: *PLoS One* 6.11 (2011), e26517.

[33] Yuanjiao Zhang et al. "Alternative splicing and cancer: a systematic review". In: *Signal Transduction and Targeted Therapy* 6.1 (2021), pp. 1–14.

[34] Luke Frankiw, David Baltimore, and Guideng Li. "Alternative mRNA splicing in cancer immunotherapy". In: *Nature Reviews. Immunology* 19.11 (2019), pp. 675–687.

[35] André Kahles et al. "Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients". In: *Cancer Cell* 34.2 (2018), 211–224.e6.

[36] Reyka G. Jayasinghe et al. "Systematic Analysis of Splice-Site-Creating Mutations in Cancer". In: *Cell Reports* 23.1 (2018), 270–281.e3.

[37] Charline Vauchy et al. "CD20 alternative splicing isoform generates immunogenic CD4 helper T epitopes". In: *International Journal of Cancer* 137.1 (2015), pp. 116–126.

[38] Yue Wang et al. "Gene fusion neoantigens: Emerging targets for cancer immunotherapy". In: *Cancer Letters* 506 (2021), pp. 45–54.

[39] Wei Yang et al. "Immunogenic neoantigens derived from gene fusions stimulate T cell responses". In: *Nature Medicine* 25.5 (2019), pp. 767–775.

[40] Xue Chen et al. "Fusion gene map of acute leukemia revealed by transcriptome sequencing of a consecutive cohort of 1000 cases in a single center". In: *Blood Cancer Journal* 11.6 (2021), pp. 1–10.

[41] B. J. Druker et al. "Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia". In: *The New England Journal of Medicine* 344.14 (2001), pp. 1031–1037.

[42] J. Pinilla-Ibarz et al. "Vaccination of patients with chronic myelogenous leukemia with bcr-abl oncogene breakpoint fusion peptides generates specific immune responses". In: *Blood* 95.5 (2000), pp. 1781–1787.

[43] Melinda A. Biernacki et al. "CBFB-MYH11 fusion neoantigen enables T cell recognition and killing of acute myeloid leukemia". In: *The Journal of Clinical Investigation* 130.10 (2020), pp. 5127–5141.

[44] Brian Haas et al. "STAR-Fusion: fast and accurate fusion transcript detection from RNA-Seq". In: *BioRxiv* (2017).

[45] Xuan Zhao et al. "Targeting neoantigens for cancer immunotherapy". In: *Biomarker Research* 9.1 (2021), p. 61.

[46] Joseph F. Perz et al. "The contributions of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide". In: *Journal of Hepatology* 45.4 (2006), pp. 529–538.

[47] Enrique A. Mesri, Ethel Cesarman, and Chris Boshoff. "Kaposi's sarcoma and its associated herpesvirus". In: *Nature Reviews. Cancer* 10.10 (2010), pp. 707–719.

[48] Haruko Tashiro and Malcolm K. Brenner. "Immunotherapy against cancer-related viruses". In: *Cell Research* 27.1 (2017), pp. 59–73.

[49] Anna Rosa Garbuglia et al. "The Use of Both Therapeutic and Prophylactic Vaccines in the Therapy of Papillomavirus Disease". In: *Frontiers in Immunology* 11 (2020), p. 188.

[50] Carlos A. Ramos et al. "Human papillomavirus type 16 E6/E7-specific cytotoxic T lymphocytes for adoptive immunotherapy of HPV-associated malignancies". In: *Journal of Immunotherapy* 36.1 (2013), pp. 66–76.

[51] Sara R Selitsky et al. "Virus expression detection reveals RNA-sequencing contamination in TCGA". In: *BMC genomics* 21.1 (2020), pp. 1–11.

[52] Nicole Grandi and Enzo Tramontano. "Human endogenous retroviruses are ancient acquired elements still shaping innate immune responses". In: *Frontiers in immunology* (2018), p. 2039.

[53] Christof C Smith et al. "Endogenous retroviral signatures predict immunotherapy response in clear cell renal cell carcinoma". In: *The Journal of clinical investigation* 128.11 (2019), pp. 4804–4820.

[54] Paola Bonaventura et al. "Identification of shared tumor epitopes from endogenous retroviruses inducing high-avidity cytotoxic T cells for cancer immunotherapy". In: *Science Advances* 8.4 (2022), eabj3671.

[55] Christopher E. Touloukian et al. "Expression of a "Self-" Antigen by Human Tumor Cells Enhances Tumor Antigen-specific CD4+ T-Cell Function". In: *Cancer research* 62.18 (2002), pp. 5144–5147.

[56] Martin A. Cheever et al. "The prioritization of cancer antigens: a national cancer institute pilot project for the acceleration of translational research". In: *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 15.17 (2009), pp. 5323–5337.

[57] Robert K Bright, Jennifer D Bright, and Jennifer A Byrne. "Overexpressed oncogenic tumor-self antigens". In: *Human Vaccines & Immunotherapeutics* 10.11 (2014), pp. 3297–3305.

[58] Constantin N. Baxevanis et al. "HER-2/neu as a target for cancer vaccines". In: *Immunotherapy* 2.2 (2010), pp. 213–226.

[59] Aaron M. Miller et al. "Leveraging TCR Affinity in Adoptive Immunotherapy against Shared Tumor/Self-Antigens". In: *Cancer Immunology Research* 7 (2019), pp. 40–49.

[60] Geoffrey Mitchell, Seth M. Pollack, and Michael J. Wagner. "Targeting cancer testis antigens in synovial sarcoma". In: *Journal for ImmunoTherapy of Cancer* 9.6 (2021), e002072.

[61] Richard A Morgan et al. "Cancer regression and neurologic toxicity following anti-MAGE-A3 TCR gene therapy". In: *Journal of immunotherapy (Hagerstown, Md.: 1997)* 36.2 (2013), p. 133.

[62] Luiz Gonzaga Almeida et al. "CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens". In: *Nucleic acids research* 37.suppl_1 (2009), pp. D816–D819.

[63] P. A. Greif et al. "Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing". In: *Leukemia* 25.5 (2011), pp. 821–827.

[64] Jun Hyung Lee et al. "A Novel KIT INDEL Mutation in Acute Myeloid Leukemia With t(8;21)(q22;q22); RUNX1-RUNX1T1". In: *Annals of Laboratory Medicine* 36.4 (2016), pp. 371–374.

[65] Dyantha I van der Lee et al. "Mutated nucleophosmin 1 as immunotherapy target in acute myeloid leukemia". In: *The Journal of clinical investigation* 129.2 (2019), pp. 774–785.

[66] Dyantha I van der Lee et al. "An HLA-A* 11: 01-Binding Neoantigen from Mutated NPM1 as Target for TCR Gene Therapy in AML". In: *Cancers* 13.21 (2021), p. 5390.

[67] Cyriac Kandoth et al. "Mutational landscape and significance across 12 major cancer types". In: *Nature* 502.7471 (2013), pp. 333–339.

[68] Jianbiao Zhou and Wee-Joo Chng. "Aberrant RNA splicing and mutations in spliceosome complex in acute myeloid leukemia". In: *Stem Cell Investigation* 4 (2017), p. 6.

[69] Samuel Li et al. "MPO as a Novel Susceptibility Gene in Myeloid Malignancies". In: *Blood* 134 (2019), p. 5402.

[70] Shu-jen Chen et al. "Transient receptor potential ion channel TRPM2 promotes AML proliferation and survival through modulation of mitochondrial function, ROS, and autophagy". In: *Cell death & disease* 11.4 (2020), pp. 1–17.

[71] Mathieu FM Cellier. "Cell-type specific determinants of NRAMP1 expression in professional phagocytes". In: *Biology* 2.1 (2013), pp. 233–283.

[72] Daniéle Masson, Moreno Zamai, and Jürg Tschopp. "Identification of granzyme A isolated from cytotoxic T-lymphocyte-granules as one of the proteases encoded by CTL-specific genes". In: *FEBS letters* 208.1 (1986), pp. 84–88.

[73] *Endogenous Retroviruses in Acute Myeloid Leukemia (ERVAL).* https://clinicaltrials.gov/ct2/show/NCT04406207. Accessed: 2022-03-22.

[74] S. Anguille, V. F. Van Tendeloo, and Z. N. Berneman. "Leukemia-associated antigens and their relevance to the immunotherapy of acute myeloid leukemia". In: *Leukemia* 26.10 (2012), pp. 2186–2196.

[75] Gang Chen et al. "Serological identification of immunogenic antigens in acute monocytic leukemia". In: *Leukemia research* 29.5 (2005), pp. 503–509.

[76] Xiuli An et al. "Global transcriptome analyses of human and murine terminal erythroid differentiation". In: *Blood, The Journal of the American Society of Hematology* 123.22 (2014), pp. 3466–3477.

[77] Concetta Quintarelli et al. "High-avidity cytotoxic T lymphocytes specific for a new PRAME-derived peptide can target leukemic and leukemic-precursor cells". In: *Blood, The Journal of the American Society of Hematology* 117.12 (2011), pp. 3353–3362.

[78] Daniel Steinbach et al. "Identification of a set of seven genes for the monitoring of minimal residual disease in pediatric acute myeloid leukemia". In: *Clinical cancer research* 12.8 (2006), pp. 2434–2441.