frontiers in Immunology

TECHNOLOGY AND CODE
published: 10 November 2020
doi: 10.3389/fimmu.2020.483296

ProTECT—Prediction of T-Cell Epitopes for Cancer Therapy

Arjun A. Rao[1,2,3], *Ada A. Madejska*[2,4], *Jacob Pfeil*[1,2,3], *Benedict Paten*[1,2,3], *Sofie R. Salama*[1,3,5*†] *and David Haussler*[1,2,3,5†]

[1] Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, United States, [2] Computational Genomics Lab, University of California, Santa Cruz, Santa Cruz, CA, United States, [3] UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, United States, [4] Department of Molecular, Cell, and Developmental Biology, University of California, Santa Cruz, Santa Cruz, CA, United States, [5] Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, CA, United States

Somatic mutations in cancers affecting protein coding genes can give rise to potentially therapeutic neoepitopes. These neoepitopes can guide Adoptive Cell Therapies and Peptide- and RNA-based Neoepitope Vaccines to selectively target tumor cells using autologous patient cytotoxic T-cells. Currently, researchers have to independently align their data, call somatic mutations and haplotype the patient's HLA to use existing neoepitope prediction tools. We present ProTECT, a fully automated, reproducible, scalable, and efficient end-to-end analysis pipeline to identify and rank therapeutically relevant tumor neoepitopes in terms of potential immunogenicity starting directly from raw patient sequencing data, or from pre-processed data. The ProTECT pipeline encompasses alignment, HLA haplotyping, mutation calling (single nucleotide variants, short insertions and deletions, and gene fusions), peptide:MHC binding prediction, and ranking of final candidates. We demonstrate the scalability, efficiency, and utility of ProTECT on 326 samples from the TCGA Prostate Adenocarcinoma cohort, identifying recurrent potential neoepitopes from TMPRSS2-ERG fusions, and from SNVs in SPOP. We also compare ProTECT with results from published tools. ProTECT can be run on a standalone computer, a local cluster, or on a compute cloud using a Mesos backend. ProTECT is highly scalable and can process TCGA data in under 30 min per sample (on average) when run in large batches. ProTECT is freely available at https://www.github.com/BD2KGenomics/protect.

Keywords: cancer, neoepitope, neoantigen, automated prediction, vaccine, cancer immunotherapy, adoptive cell therapy

OPEN ACCESS

Edited by:
*Mustafa Diken,
Johannes Gutenberg-Universität
Mainz, Germany*

Reviewed by:
*Jonas Ibn-Salem,
Johannes Gutenberg-Universität
Mainz, Germany
Zoltan Vereb,
University of Szeged, Hungary*

***Correspondence:**
*Sofie R. Salama
ssalama@ucsc.edu*

†These authors share senior authorship

Specialty section:
*This article was submitted to
Cancer Immunity
and Immunotherapy,
a section of the journal
Frontiers in Immunology*

Received: 06 July 2019
Accepted: 13 October 2020
Published: 10 November 2020

Citation:
*Rao AA, Madejska AA, Pfeil J, Paten B,
Salama SR and Haussler D (2020)
ProTECT—Prediction of T-Cell
Epitopes for Cancer Therapy.
Front. Immunol. 11:483296.
doi: 10.3389/fimmu.2020.483296*

INTRODUCTION

Tumor recognition by the adaptive immune system has been described in the literature as early as the 1980s. In 1987, Muul et al. described tumor infiltrating lymphocytes in a cohort of six melanoma samples that showed high cytotoxicity towards fresh, autologous melanoma tumor cells (1). However, at the time, T-cell responses were observed to be short lived, often lasting only a few days. Later studies showed that tumors were capable of suppressing immune responses *via* different mechanisms (2–5).

Checkpoint blockade therapy has seen a great increase in interest in the past few years with numerous drugs being approved by the FDA for clinical treatment (6–8). Prevention of PD-1:PD-L1 (9) and CTLA-4:B7.1/2 (10) binding *via* monoclonal antibodies re-enables the immune attack against the tumor; however, it can leave the patient open to development of autoimmunity or other toxicities associated with unchecked immune action (11, 12). The mutational load of a tumor (or Tumor Mutational Burden) is a good predictor of response to checkpoint therapy (13, 14). The observation that aberrations in DNA Mismatch repair genes impair tumor growth (15) suggests this effect is due to tumor "neoantigens" that act as markers for immune targeting.

Tumor infiltrating lymphocytes (TILs) from patient tumors can be activated and expanded *in-vitro* using minced autologous tumor (16). TILs can also be activated using autologous dendritic cells that are experimentally primed *in-vitro* with synthetically generated, neoepitope-bearing peptides or with RNA vaccines that contain coding transcripts for neoepitope-bearing peptides (17, 18). These cells selectively target cell-surface MHC-presented antigen produced by the tumor. Peptide vaccines attempt to produce the same result by stimulating dendritic cells *in-vivo via* synthetically produced peptides delivered subcutaneously to the patient. Experimentally primed dendritic cells and peptide vaccine therapies require prior knowledge of the mutations in the tumor in order to identify the potentially targetable sequence.

Bioinformatic analysis of tumor sequencing data can aid in the selection of neoepitopes to target in vaccine and adaptive immune system-based cancer therapies. pVAC-Seq (19) is an automated pipeline that identifies neoepitopes generated from a pre-computed, VEP-annotated (20) VCF file run with specialized plug-ins that incorporate wildtype and mutant protein sequence. Vaxrank (21) provides a ranked list of epitopes given an input mutation VCF, RNA-Seq BAMs and the patient MHC haplotype. Epidisco (22), the predecessor of Vaxrank, was capable of starting from input FASTQs. INTEGRATE-Neo (23) identifies neoepitopes from fusion genes provided in a pre-computed BEDPE file. NeoepitopePred (24) provided a workflow for epitope prediction from fusion genes and can be run through the applets on the DNAnexus cloud platform. These tools all require a user to previously align the sequencing data to a reference of choice and call variants before following the same logical paradigm of identifying mutant peptides and predicting peptide:MHC (pMHC) affinity binding [often *via* netMHC (25)]. The pipelines differ in their degree of automation, input mutation type and annotation, and presence or absence of a ranking schema. There is a clear need for a fully automated pipeline from end-to-end, beginning at the raw FASTQ files emitted by the sequencer from DNA and RNA sequencing. Recently, NeoFuse (26) was published, which automates fusion-gene-based neoepitope prediction from paired RNA-seq, but this tool does not include neoepitopes derived from single nucleotide variants (SNVs) or short insertions and deletions (INDELs).

We developed ProTECT, a fully automated tool for the Prediction of T-cell Epitopes for Cancer Therapy. We previously demonstrated the utility of ProTECT using an early version to analyze externally called SNVs in a neuroblastoma cohort (27). There we identified a potentially therapeutic neoepitope from the ALK:R1275Q hotspot mutation and proved that $CD8^+$ cytotoxic T-cells could recognize it using *in-vitro* MHC tetramer staining of peripheral blood mononuclear cells from two HLA-matched donors. The full ProTECT codebase, reported here, is completely self-contained. It accepts an input trio of sequencing data from a patient consisting of the paired tumor and normal DNA, and the tumor RNA reads in the FASTQ format and processes the data from end-to-end including alignment, *in-silico* HLA haplotyping, expression profiling, mutation calling, and neoepitope prediction.

Here we evaluate the scalability, utility, and performance of ProTECT using publicly available data. We use the 326 samples from The Cancer Genome Atlas (TCGA) Prostate Adenocarcinoma (PRAD) cohort (28) with trios of genomic data (tumor DNA, normal DNA, and tumor RNA), augmenting these data with eight previously published clinical melanoma samples (29). The TCGA PRAD cohort has an average of 21.5 exonic mutations per sample (30) and 31% of all samples are predicted to contain a fusion transcript (31), making it a good choice for detecting both SNV and fusion neoepitopes. Further, it was previously evaluated for fusion-gene-derived epitopes using INTEGRATE-Neo (23). The melanoma dataset was reported to have between 219 and 598 missense exonic mutations per sample and was previously analyzed by pVAC-Seq (19) as part of a clinical trial. We compared ProTECT's performance to the performance of these other tools.

MATERIALS AND METHODS

Procurement of Input Data

Genomic Trio (tumor DNA, normal DNA, and tumor RNA) BAM files containing sequences from 326 samples in the TCGA Prostate Adenocarcinoma (PRAD) cohort were downloaded from the Genomics Data Commons (GDC) at the National Cancer Institute using the GDC data transfer tool. The downloaded BAM files were converted back to FASTQ format as would be produced by direct sequencing using the SamToFastq module from Picard version 1.125[1]. MHCI haplotype calls using POLYSOLVER (32) for all samples were obtained externally and used for MHC haplotype prediction comparisons.

Genomic trios from three additional samples (Mel-21, Mel-38, Mel-218) were downloaded from the NCBI short read archive (SRA) (33) *via* Bioproject PRJNA278450/dbGaP accession phs001005. These patients were diagnosed with stage III resected cutaneous melanoma and had all previously received ipilimumab. Data from seven A*02:01 restricted vaccines tested for each patient were obtained from the supplementary information of the original manuscript (29).

[1] Obtained from http://broadinstitute.github.io/picard/

The input data for the INTEGRATE-Neo comparison included haplotype and fusion calls from 240 samples in the supplementary data of the INTEGRATE-Neo paper. The fusions from supplementary Excel sheet 1 were parsed into individual BEDPE format files and the epitopes from sheet 3 were extracted into individual haplotype list files with one MHC allele per line.

Indexes for the various tools were generated using the hg38 (GRCh38) reference sequence obtained from the UCSC genome browser (34). GENCODE (35) v25 was chosen as the reference annotation and was used in all relevant parts of the pipeline. Every generic hg38 index used in the analysis is available in our AWS S3 bucket 'protect-data' under the folders 'hg38_references'. These indexes can be pulled by any user to run ProTECT locally. A detailed list of commands used to create the various indexes is available in the same bucket in the 'README'.

Compute Resources Utilized

All TCGA-related analyses were conducted on a Mesos (36) cluster with one leader (12 cpus, 62 GB RAM, 500 GB Local disk) and eight identical agents (56 cpus, 250 GB RAM, 1.8 TB local disk).

The Melanoma data was analyzed on the Amazon Web Services EC2, and the data was stored securely using SSE-C encryption on S3.

326-Sample PRAD Compute

The 326 samples were run in batches of 1, 2, 5, 10, 20, or 50 samples in order to gauge the efficiency and scalability of the pipeline engine, Toil (37). Each batch size was run five times with unique samples to normalize the runtime information. The configuration file for each run was generated from a template containing all the required tool options and paths to the input reference files on the Network File System (NFS) storage server. Each batch was run once on the Mesos cluster using all nodes and an NFS-based Toil file job store to save the state of the pipeline. The five single-sample batches were also run separately without Mesos on individual nodes of the cluster using an NFS-based Toil file job store to document the time taken per sample on a single machine.

Comparison With pVAC-Seq

To compare our results with pVAC-Seq, we ran ProTECT on the input samples on AWS EC2 using an S3-based cloud job store. The input configuration for the run included paths to hg38-mapped reference files from our S3 bucket 'protect-data' and paths to the input FASTQ files in another secure bucket. The results were stored on S3 in the same bucket as the input. This analysis was conducted consistent with the mandatory cloud data use limitations on the input dataset.

Comparison With INTEGRATE-Neo

To compare our results with INTEGRATE-Neo, we parsed the data from the manuscript supplement into files acceptable by ProTECT *via* a python script. The initial input configuration file consisted of links to the fusion BEDPE format file for each of 240 samples, along with the haplotype and expression data called from the ProTECT 326 sample run. The final analysis included fusion and inferred haplotype calls for 83 samples from INTEGRATE-Neo along with ProTECT expression estimates. All ProTECT runs were conducted on the Mesos cluster.

PIPELINE SPECIFICS

ProTECT consists of eight major sections: sequence alignment, haplotyping, expression profiling, mutation calling, mutation translation, MHC:peptide binding prediction, neoepitope ranking, and reporting. **Figure 1** shows the schema for the run. Every tool used in the pipeline was hand-picked from industry-standard choices and literature reviews. Some aspects of the pipeline, notably TransGene and Rankboost, were developed in-house due to a lack of publicly available alternatives. Both tools are available as open-source repositories on github.

The entire analysis from end-to-end is built to process data against the same reference sequence and annotation. The user provides links to the properly generated indexes for each tool in the pipeline. We provide Gencode (35) version 19 annotated references for hg19 and Gencode version 25 annotated references for hg38 on our public AWS S3 bucket "protect-data"[2]. The input for a ProTECT run is a single configuration file that lists input files for each patient that will be processed and all the options and links to indexes that will be used during the run.

While ProTECT is built for end-to-end processing of sequencing trios per patient using our choice of software at each step, we understand that researchers have personal preferences for some software over others for mutation calling, gene expression estimation, *etc*. We have engineered ProTECT such that a user may run it with pre-computed SNVs, fusion calls, gene expression, and HLA haplotypes, provided they are formatted appropriately.

Sequence Alignment

DNA sequence alignment is carried out using the Burrows–Wheeler aligner (BWA) (38). The reads are aligned with BWA-mem to the provided BWA reference using default parameters. The SAM file produced upon alignment is processed to properly format the SAM header and is then converted to a coordinate-sorted BAM file with a corresponding index.

RNA sequence alignment is carried out using the ultra-fast aligner, STAR (39). The parameters for the run are optimized for fusion detection *via* STAR-fusion (40).

Alternatively, ProTECT accepts pre-aligned BAM files as an input if the MHC haplotype is provided as well. ProTECT assumes that the user has aligned the DNA and RNA using the same reference genome with the same genomic annotation.

Haplotyping

The HLA Haplotype of the patient is predicted using PHLAT (41). The haplotype is predicted using each input source of

[2] https://s3.console.aws.amazon.com/s3/buckets/protect-data/?region=us-west-2&tab=overview (Requires AWS account to access).

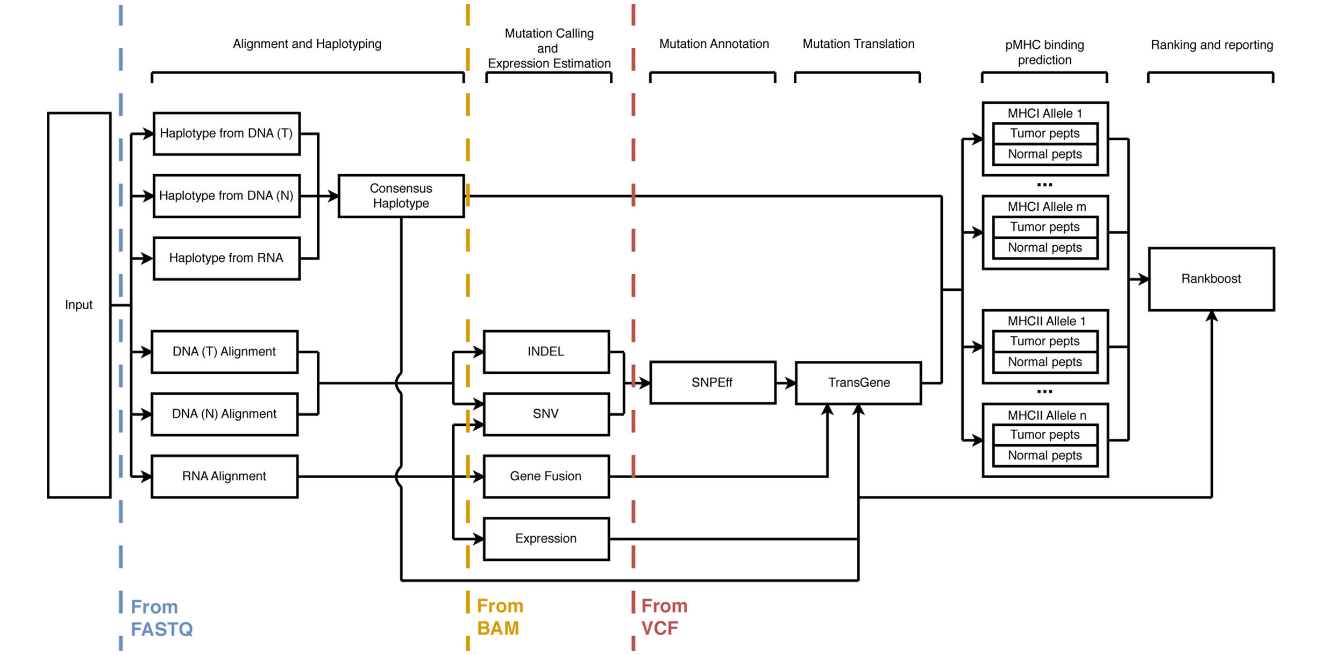

FIGURE 1 | A schematic description of the ProTECT workflow. ProTECT can process FASTQs all the way through the prediction of ImmunoActive Regions, including alignment, HLA haplotyping, variant calling, expression estimation, mutation translation, and pMHC binding affinity prediction. ProTECT also allows users to provide pre-computed inputs for various steps instead.

information (normal and tumor DNA, tumor RNA), and the consensus haplotype is generated based on agreement between any two of the three haplotype predictions. Due to limitations in the tool, we only proceed with HLA-A, HLA-B, and HLA-C for MHCI, and HLA-DPA/B and HLA-DRB for MHCII.

Expression Profiling

The gene-level and isoform-level expression is estimated using RSEM (42) with default parameters.

Mutation Calling

SNVs are predicted on a per-chromosome basis using five separate mutation prediction algorithms: MuTECT (43), MuSE (44), RADIA (45), Somatic Sniper (46), and Strelka (47). The choice of mutation callers was guided by the results from the ICGC DREAM mutation calling challenge (48). All called mutations are merged into a common file, and only events supported by two or more predictors advance to the translation step. Strelka additionally produces a callout for short insertions and deletions (INDELs). These are also used to identify neoepitopes.

Fusion calling occurs using STAR-Fusion (40) with default parameters. Candidate fusions are annotated using Fusion-Inspector[3] along with an optional assembly step using Trinity (49).

Mutation Translation

SNVs and INDELs are annotated using SNPEff (50). Mutations identified in coding regions of the genome are processed using an in-house translation tool, TransGene[4]. TransGene filters the input SNPEff-produced VCF file to exclude non-expressed calls based on the gene expression data obtained in the previous step. SNVs and in-frame INDELs are directly injected into the amino acid chain to produce the mutant sequence. Frameshift INDELs are translated downstream of the mutation event till a stop codon is found (or a user-defined threshold is reached). Events lying within 27, 30, and 45 bp of each other (for 9-mer-, 10-mer-, and 15-mer-containing peptides respectively) are chained together into an "immunoactive region" (IAR), or a region that will potentially produce an immunogenic peptide. Separate mutation events that are combined into a single immunoactive region are phased using the RNA-Seq data to ensure that they truly are co-expressed on the same haplotype.

Fusion IARs are generated using the breakpoints provided in an input BEDPE file. TransGene uses provided junction sequences or infers them from the input annotation file. The predicted IAR contains (n – 1)*3 bp on either side of the fusion junction from each donor for each n in 9-, 10-, and 15-mer. Fusion calls are optionally filtered at this stage to remove events arising from two mitochondrial genes or two immunoglobulin genes since these are usually false positive events arising from sequence similarity. Fusions can also be filtered for being

[3] Obtained from https://github.com/FusionInspector

[4] Hosted at https://github.com/arkal/transgene

potential transcriptional readthroughs (by default, two genes on the same chromosome within 500 kb of each other are rejected) or for having a 5′ lincRNA (under the assumption that these events are unlikely to be translated).

MHC:Peptide Binding Prediction

The predicted neoepitopes are assayed against each of the MHCI (9- and 10-mers) and MHCII (15-mers) predicted to be in the patient's HLA haplotype using the IEDB MHCI and MHCII binding predictions tools.

The IEDB tools run a panel of methods (51–57) on each input query (input peptide FASTQ + MHC allele) and provide a consensus "percentile rank" that describes on average, how well each peptide is predicted to bind against a background set of 100,000 UniPROT derived peptides. Calls predicted to bind within the top 5% of all binders are selected for further study. The normal, unmutated ("wildtype") counterpart peptide for each selected neoepitope is then also assayed against the MHC(s) identified to determine how well it binds, so that this can be compared to the binding affinity of the mutant version.

Neo-Epitope Ranking

Neoepitope: MHC calls are consolidated by the candidate IAR of origin. An in-house method, Rankboost[5], first arranges the IARs in descending order based on the best binding score of a contained neoepitope and then uses the boosting strategy described in Algorithm 1 to produce a final list of ranked IARs. Candidates satisfying certain biologically relevant criteria are boosted in rank based on user-specified weights. The features considered are the total number of calls originating from the IAR (npa) and ones with high predicted binding score (nph, percentile rank <=1.0), the promiscuity of the region (nmhc, *i.e.* the number of MHCs stimulated by peptides from the region), the combined expression of the isoforms displaying a neoepitope-generating mutation (TPM), the number of neoepitopes in the region predicted to bind to an MHC better than their wildtype counterpart, and the number of events where a 10-mer and 9-mer subsequence of it both bind well to an MHC (ovlp, this is only done for MHCI). Each candidate is assigned a score from 0-1 for each feature that is multiplied by a user-specified weight. The sum of the weighted score provides the boost received by the candidate. Feature score functions were generated based on empirical distributions of the features seen in IARS predicted in other TCGA and internal datasets. The algorithm iterates over the table of candidates three times and performs per-candidate boosting, resulting in a ranked list of epitopes in the sample. We ran our samples prioritizing overlap and promiscuity (0.68 and 0.32 respectively) for MHCI calls and set each covariate to 0.2 (equally important) for MHCII calls.

Algorithm 1. Pseudocode for the rank boosting strategy. W_x describes the weight for covariate x, boost_x describes the score for the candidate x from 0 to 1, npa = number of peptides constituting an IAR, nph = number of strongly binding peptides constituting the IAR, nMHC = number of MHCs predicted to recognize a neoepitope from this IAR, TPM = expression of the transcript harboring the IAR, and ovlp = number of events where a 9-mer and 10-mer overlap and are predicted to bind to the same MHC (only valid for MHCI).

```
For i in 1, 2, 3
    ∀ candidate in candidates
        boost = W_npa * boost_npa + \
                W_nph * boost_nph + \
                W_nMHC * boost_nMHC + \
                W_TPM * boost_TPM + \
                W_ovlp * boost_ovlp
        new_rank = old_rank * (1-boost)
```

RESULTS AND DISCUSSION

We ran three experiments to demonstrate our pipeline. The first experiment was run on 326 samples from the TCGA PRAD cohort and highlights the scalability, efficiency, and utility of ProTECT. We also identify recurrent IARs in the cohort (containing mutations that occurred in more than one case) suggesting possible public neoepitopes for PRAD. The second experiment compares ProTECT to the published SNV- and INDEL-based neoepitope prediction pipeline, pVAC-Seq. The third experiment compares ProTECT to the published fusion-based neoepitope predictor, INTEGRATE-Neo. In all experiments, ProTECT was run using a consensus of two out of five mutation callers (as described above) and using all TransGene fusion filters to remove inter-mitochondrial, inter-immunoglobulin, 5′ lincRNA, and transcriptomic readthrough events. Results were tabulated using a mix of python scripts and manual curation on a local machine.

326 Sample Run

To describe the scalability, utility, and efficiency of ProTECT, we ran ProTECT on a total of 326 genomic trios from the TCGA PRAD cohort. We called a median of 79.5 SNVs and INDELs, and seven fusion genes per sample, and accepted 20 and three respectively for the production of IARs. We identified a median of 11 IARs per sample. Of the 326 samples, only three samples were predicted to have no IARs. These samples were observed to have no expressed non-synonymous mutations or filter-passing fusions. The entire metrics table is presented in **Supplementary Table 1**, and the results are submitted as **Supplementary File 1**.

Figure 2 shows the results from running ProTECT with different batch sizes on our local cluster (see section *Compute Resources Utilized*). As the number of samples increases, we see an expected increase in overall time, but the average time per sample decreases drastically because our pipeline engine maximizes resource utilization. We processed samples from end-to-end at a rate of 24.6 min per sample (calculated as total time divided by total number of samples) when run in a batch of 50 samples.

[5] Hosted at https://github.com/arkal/rankboost

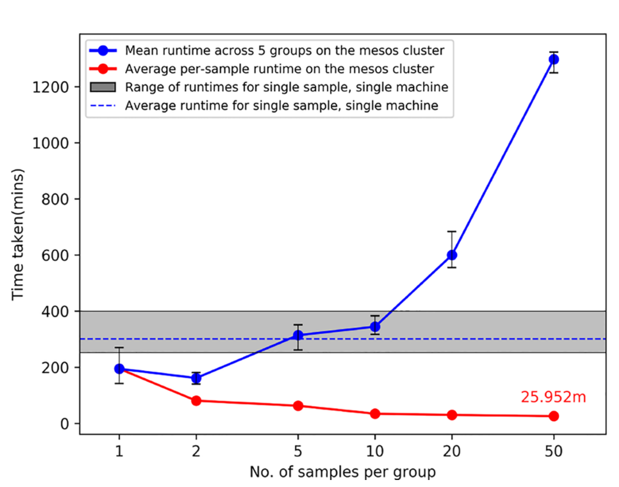

FIGURE 2 | Average runtimes on our cluster when ProTECT is run in a batch of 'n' samples. Each batch of size 'n' is run with five unique sample sets, and the range of runtimes is described by the whiskers at each datapoint. The gray bar describes the result of running ProTECT on a single sample on one machine. ProTECT takes considerably less time on average when run in a large group.

Recurrent Fusion-Gene-Derived IARs in PRAD

We detected the well-documented TMPRSS2-ERG fusion gene (58–60) in 131 samples. We predicted at least one IAR each arising from five of the 10 unique breakpoints called (**Table 1**). Of the five breakpoints that do not result in an IAR, four of these breakpoints are located in the 5′ UTR of TMPRSS2 and will not result in a neoepitope. The 5th breakpoint has a 5′ intronic breakpoint and a 3′ exonic one, and the resulting neoepitope should contain the translated product from the last few bases of TMPRRS2 Exon 1 and the first bases after the *de novo* splice acceptor is reached in ERG. This case is not handled by TransGene at this time, and so no neoepitope call was made. One IAR of particular interest is DNSKMALNSEALSVVSED from the junction chr21:41498119–chr21:38445621, which is found in 37 of the 48 unique samples harboring that junction (11% of the entire cohort). Peptides from this IAR are predicted to bind well to HLA-A*02:01 (Allele Freq: 0.26) and HLA-C*07:01 (allele Freq: 0.17), alleles frequently seen in Caucasian populations, which are highly represented in the TCGA cohort. Similarly, we predict SGCEERGAAGSLISCE from 22/35 samples with chr21:41507950–chr21:38445621, binding to C*07:01, C*04:01, B*44:02 (allele frequencies 0.14, 0.12, 0.08 respectively). The distributions of MHC alleles detected in patients harboring these events are shown in **Supplementary Figures 1** and **2**, respectively. These events are potentially viable candidates for public epitopes for patients with TMPRSS2-ERG and could be pursued as vaccines for these cancers.

Recurrent SNV-Derived IARs in PRAD

We detected a number of recurrent mutations in the SPOP gene concordant with previous reports (28, 61, 62). We detected seven unique recurrent variants across 19 samples that map to three different amino acid positions in the SPOP protein, p.F133C/V/I/L, p.F102C/V, and p.W131G (**Table 2**). The mutation at position 133 might be of immunological interest since Leucine, Isoleucine, and Valine have small hydrophobic side-chains and may stimulate the same TCR depending on pMHC binding. This hypothesis however, would require biological validation. Samples with SPOP mutations lack ETV family fusions, suggesting that vaccine therapies against SPOP and the TMPRSS2-ERG fusion would target different populations of PRAD patients.

Comparison of HLA Haplotypes Between PHLAT and POLYSOLVER

An important topic to highlight is HLA haplotypes called by PHLAT (41). We compared our results to the POLYSOLVER (32) calls, and consistent with prior work (63), we see that PHLAT miscalled HLA-A*02:01 as HLA-A*01:81 in 33 samples. However, 29 of these samples are predicted to be homozygous HLA-A*02:01 by POLYSOLVER so the effect of this miscall will be to add information to the final ranked IARS from one additional allele. Since most IARs contain peptides predicted to bind to more than one allele, the noise produced by this artifact should not adversely affect the scores generated *via* the signal from calls against the correct partners. The remaining four samples were predicted to be heterozygous HLA-A*02:01/HLA-A*01:01 *via* POLYSOLVER, and ProTECT identified these samples as HLA-*02:01/HLA-A*01:81. This is slightly worse

TABLE 1 | Recurrent TMPRSS2-ERG breakpoints in the cohort.

Breakpoint	Count	5' breakpoint	3' breakpoint	Neoepitope Expected?	IAR	Count
21:41508081–21:38445621	122	5′ UTR	Exon 2	No	NA	
21:41498119–21:38445621	48	Exon 2	Exon 2	Yes	DNSKMALNS EALSVVSED	37
21:41507950–21:38445621	35	Exon 1	Exon 2	Yes***	SGCEERGAA GSLISCE	22
21:41508081–21:38474121	18	5′ UTR	Intron 1	No	NA	
21:41506445–21:38445621	18	Intron 1	Exon 2	Yes*	NA	
21:41508081–21:38584945	11	5′ UTR	5′UTR	No	NA	
21:41498119–21:38474121	7	Exon 2	Intron 1	Yes**	DNSKMALNS LNSIDDAQL	7
21:41508081–21:38423561	7	5′ UTR	Exon 3	No	NA	
21:41498119–21:38423561	4	Exon 2	Exon 3	Yes***	DNSKMALNS ELS	1
21:41494356–21:38445621	3	Exon 3	Exon 2	Yes***	SPSGTVCTS RSLISCE	3

*IARs from 21:41498119 to 21:38445621 and 21:41507950 to 21:38445621 are recurrent suggesting the viability of universal peptide vaccine candidates. We do not expect to see an IAR from fusions with 5′ UTR breakpoints. *TransGene cannot handle de novo splice acceptors. **An epitope will exist where the TMPRSS2 reads into the intron of ERG. ***A frameshift is seen on the ERG side of the fusion.*

TABLE 2 | Recurrent mutants in the SPOP gene target three codons.

Variant	Count	Gene	Mutant	IAR	Frequency
chr17:49619064A>C	5	SPOP	p.F133V	RFVQGKDWG **V** KKFIRRDFL	4
chr17:49619063A>C	3	SPOP	p.F133C	RFVQGKDWG **C** KKFIRRDFL	2
chr17:49619064A>T	2	SPOP	p.F133I	RFVQGKDWG **I** KKFIRRDFL	1
chr17:49619062G>T	2	SPOP	p.F133L	RFVQGKDWG **L** KKFIRRDFL	2
chr17:49619281A>C	2	SPOP	p.F102C	CPKSEVRAK **C** KFSILNAKG	2
chr17:49619282A>C	3	SPOP	p.F102V	CPKSEVRAK **V** KFSILNAKG	3
chr17:49619070A>C	2	SPOP	p.W131G	AYRFVQGKD **G** GFKKFIRRD	2

The F133V/C/I/L mutant may be of interest as a universal neoepitope due to the similar chemical properties of Leucine, Isoleucine and Valine.

than the first case since we're completely lacking HLA-A*01:01 peptide binding affinity predictions for all these samples. Overall, 67.5% of all samples had perfectly concordant haplotypes with POLYSOLVER, 28.8% differed by one allele and 3.7% differed by two (**Figure 3**). A large chunk of the second group consists of the miscalls mentioned above. ProTECT allows users to provide pre-computed MHC haplotype calls if they trust another external caller more than PHLAT, or if they have haplotype information from another source.

Comparison With Published Callers

Comparison With an SNV-Based Neoepitope Predictor

We ran ProTECT on the eight melanoma samples from three patients (one primary lymph node tumor each and multiple metachronous tumors in two samples) (29) that were used to benchmark pVAC-Seq (19). Carreno et al. predicted 11–28 expressed, HLA-A*02:01 binding candidate peptides per sample and synthesized seven unique peptide vaccines per patient based on presence of the mutants in the metachronous tumors and assessed binding of the predicted peptide to HLA-A*02:01 in T2 assays. Three peptides per patient were found to induce an immune reaction. ProTECT correctly identified the expected immunogenic mutations in every reported mutation: sample pair. In some cases, ProTECT even predicted the expected variant in a metachronous tumor where the original paper missed it (E.g. CDKN2A:E153K in the Lymph Node of Mel-21) (**Table 3**). Overall, ProTECT ranked IARs containing the validated variants relatively highly (in the top 15–20%, median absolute rank of 11) except in Mel218. We cannot definitively comment on the ranking in Mel218 since ProTECT considers every mutant and MHC allele in the MHC haplotype, while Carreno et al. only considered a curated list of peptides against HLA-A*02:01. In addition to the validated variants, we also provided a larger ranked set of possible candidates that broaden the spectrum of testable epitopes. The

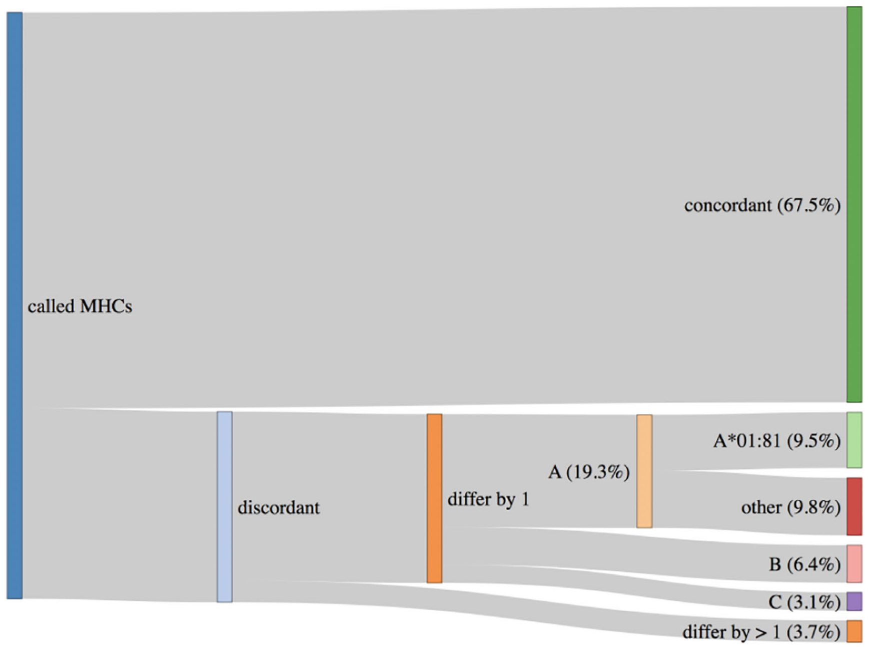

FIGURE 3 | HLA haplotypes called by ProTECT (using PHLAT) are fully concordant with POLYSOLVER haplotypes in only 67.5% of samples. 28.8% differ by one call and 3.7% by 2 calls. A majority of the miscalled HLA-A alleles are a documented PHLAT artifact.

TABLE 3 | ProTECT ranks on eight metachronous tumors across three MELANOMA patients.

Sample and Source	Mel 21				Mel 38			Mel 218
	LN	Skin (2012)	Skin (2013)		Abdominal Wall	Axilla LN	Breast	LN
			RNA 1	RNA 2				
Collection date	1/30/2011	5/10/2012	6/6/2013	6/6/2013	4/16/2013	4/19/2012	2/14/2013	4/4/2005
Total Variants	1,532	2,140	1,681	1,679	1,213	1,121	1,259	2,176
Actionable Variants	332	400	393	391	219	216	224	449
Total IARs	105	137	114	116	73	80	86	155
Vaccine Candidates with ProTECT ranks	NDC1:F169L (Reported as TMEM48 F169L)				SEC24A:P469L			EXOC8:Q656P
	4	5	10	3	8	9	2	152
	TKT:R438W				AKAP13:Q285K			PABPC1:R520Q
	6	6	4	4	14	80	17	140
	CDKN2A:E153K				OR8B3:T190I			MRPS5:P59L
	21	–	18	17	11	13	14	26

Highlighted ranks describe instances where pVAC-Seq and ProTECT both call a neoepitope. Green: Dominant epitope (existing immunity, neoantigen processed from endogenous protein), Orange: Subdominant epitope (immunity after vaccination, neoantigen processed from endogenous protein), Red: Cryptic epitope (immunity after vaccination, neoantigen not processed from endogenous protein).

data for all seven tested peptides is provided in **Supplementary Table 2**, and all neoepitopes predicted by ProTECT in **Supplementary Table 3** and **Supplementary File 2**.

Comparison With a Fusion-Gene-Based Neoepitope Predictor

We compared our fusion prediction accuracy with INTEGRATE-Neo (23). INTEGRATE-Neo was demonstrated on 321 samples from the TCGA PRAD cohort, and at least one neoepitope was predicted from 161 samples. 240 of the 321 samples overlap with our 326 sample dataset, and this subset was used for this experiment. None of the predicted neoepitopes in this study have been validated using any biological experiments. We first attempted to compare our fusions (called using STAR-Fusion) with the fusion calls generated from INTEGRATE (64) as fusion callers are known for having varied performance across different datasets (65). As expected, the overlap between the ProTECT and INTEGRATE calls was relatively low (595/1519, with 120 unique calls in ProTECT), but a large chunk of the non-overlapping calls were from events with one spanning read support in INTEGRATE (**Supplementary Figure 3**). We see a better overlap when we increase the minimum support to two (an internal metric within ProTECT) and also find that 44 events rejected for having one read support in INTEGRATE were detected by STAR-Fusion (40) with >1 read support. Some of the INTEGRATE-specific calls were picked up by ProTECT but filtered out as low-read-evidence events. We further noticed that the concordance between MHC haplotypes called by HLA Miner (66) (used by INTEGRATE-Neo), PHLAT (41) (Used by ProTECT), and POLYSOLVER (32) was very low (**Supplementary Figure 4**). 61 of the unique HLAMiner predictions across the cohort did not match any of the other two callers, and 41 matched both. (Homozygous calls in a patient were treated as one call.) Two alleles were shared exclusively between ProTECT and INTEGRATE and only one between INTEGRATE and POLYSOLVER. In order to conduct a more comparable analysis, we reran ProTECT with the INTEGRATE fusion calls and the MHC haplotypes from the INTEGRATE-Neo manuscript (182 neoepitopes from 720 fusions over 83 samples, **Supplementary Figure 5**). ProTECT rejected 100 of the 720 provided fusion events as transcriptional readthroughs (92 events) or for having a 5′ non-coding RNA partner (eight events). ProTECT correctly identified 139/182 neoepitopes as IARs and rejected the remaining for being in a rejected fusion (23 neoepitopes), scoring below the 5% predicted binding score threshold (16 neoepitopes), having a 5′ breakpoint in the UTR (three neoepitopes), or for having a 5′ non-coding partner (one neoepitope) (**Supplementary Table 4**). On further inspection, we noticed that the three neoepitopes arising from the 5′ UTR breakpoints (TCGA-HC-7080, PRH1>>RP11-259O18.4 and PRH1>>M6PR) could have been detected if the 5′ partner had been annotated with a different gene (PRR4) at the same locus (**Supplementary Figure 6**), an issue arising due to the differing gene annotation GTFs used between the methods (Gencode v25 for ProTECT and Ensembl v85 for INTEGRATE). Interestingly, this type of event occurred in one other sample (TCGA-EJ-8474, C1QTNF3-AMACR>>NDUFAF2); however, INTEGRATE called the overlapping call as well (AMACR>>NDUFAF2), and since the epitopes were identical from both, ProTECT picked them up under the correct call (**Supplementary Figure 7**). The full set of results from running ProTECT on 83 INTEGRATE-Neo inputs is provided in **Supplementary File 3**. Easing ProTECT's 5% filter would increase the number of false positives called by too large a margin, so we stand by our decision to reject the 16 neoepitopes missed due to this filter. This experiment also highlights the modularity of ProTECT and its flexibility in accepting pre-computed inputs to run only the necessary steps to produce a ranked list of IARs.

Reproducibility

Every tool used the pipeline, from established aligners to the in-house script used to translate mutations, is wrapped in a Docker image (67) tagged with the appropriate tool version. Docker allows a developer to wrap a piece of code and any requirements into an image that can be instantiated into a container on any other machine. The tool within the container will run in the same manner on any machine, under the same environmental constraints, barring minor differences that may arise from asynchronous multiprocessing/multithreading. This way, results from ProTECT run on different machines with the same inputs will always be near-identical. The default versions of each tool used by ProTECT are mentioned in the repository, and users can containerize other versions of the same tools and specify the new version to ProTECT at runtime.

Automation, Scalability, and Efficiency

ProTECT is built to be run end-to-end without any user intervention. ProTECT is written in the Toil framework and will attempt to run the pipeline on the given input samples in a resource-efficient manner. The pluggable backend Toil APIs allow ProTECT to run on a single machine, a grid engine cluster, or a Mesos cluster setup on a local network or on AWS. Toil allows users to deploy scripts on Azure and Google cloud as well; however, ProTECT does not yet support these environments.

Users provide ProTECT a config file that details the input files and the various indexes and versions of tools to use during the run. ProTECT copies (or downloads) the files to a “file store” and then queues a graph of jobs for each input sample culminating in a ranked list of epitopes. The nodes in the graph are tuned to request an appropriate number of CPUs (for multithreaded jobs), memory, and disk space. Toil ensures that these queued jobs are spawned in a way that all available resources are utilized to the maximum extent. In practice, this means that smaller, low-compute, low-memory, short-duration jobs like variant calling, mutation translation, *etc.* from one sample can run parallel to higher-compute, high-memory, long-running jobs like alignment and haplotyping in another. The processing time of any single sample is strongly influenced by the long-running jobs but utilization of free compute to run queued short-jobs reduced the overall per-sample runtime.

CONCLUSION

We have described an efficient, automated, and portable workflow for the prediction of neoepitope candidates that can guide vaccine-based or adoptive T-cell therapies. We have shown that ProTECT scales well on a parallel processing environment and is highly efficient processing samples in large batches. On average, we processed a sample from end-to-end in 26.4 min when we ran 50 samples in a single batch on an eight-node cluster. We have shown that ProTECT is comparable to existing callers and improves on them by providing a ranked list of neoepitopes arising from SNVs, INDELs, and fusion genes. None of the currently published pipelines give results for all three types of mutations. Positive results from a clinical trial were ranked highly in our results, and we retrospectively identified additional events that could have been used in the trial. We identified recurrent epitopes arising from the well-documented TMPRSS2-ERG fusion, and these results suggest a peptide or RNA vaccine could be developed for one of the common breakpoints. While designed for use in the rapidly growing fields of cancer vaccines and Autologous T-cell therapies, ProTECT can also be used to understand the link between tumor mutational burden and response to checkpoint blockade therapies. It is our fervent hope that improvements in these fields will quickly establish neoepitope-targeted immunotherapies as standard-of-care for cancer treatment.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

AR, SS, and DH contributed to the conception and design of the study. AR developed the entire codebase available on github with significant contributions from JP. AM was involved with analyses on the 326 sample run. AR wrote the first draft of the manuscript; SS, DH, and AR contributed to manuscript revision. DH, SS, and BP acquired funding for this work. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Institutes of Health/National Cancer Institute [5U24CA143858], an Alex's Lemonade Stand Innovation Award, and the Howard Hughes Medical Institute.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions of Hannes Schmidt, Christopher Ketchum, Aisling O’Farrell, Swapnil Patil, Akul Goyal and Walter Shands towards the ProTECT source code, and Bala Ganeshan, Giridhar Basava, Hemant Trivedi of Stellus Technologies for the compute support that enabled the 326 TCGA sample run. We thank Dr. Catherine Wu for her permission to use the POLYSOLVER MHC calls for the TCGA PRAD study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2020.483296/full#supplementary-material

REFERENCES

1. Muul LM, Spiess PJ, Director EP, Rosenberg SA. Identification of specific cytolytic immune responses against autologous tumor in humans bearing malignant melanoma. *J Immunol* (1987) 138:989–95.
2. Facciabene A, Motz GT, Coukos G. T-Regulatory Cells: Key Players in Tumor Immune Escape and Angiogenesis. *Cancer Res* (2012) 72:2162–71. doi: 10.1158/0008-5472.CAN-11-3687
3. Ferris RL, Hunt JL, Ferrone S. Human leukocyte antigen (HLA) class I defects in head and neck cancer: molecular mechanisms and clinical significance. *Immunol Res* (2005) 33:113–33. doi: 10.1385/IR:33:2:113
4. Ghiringhelli F, Puig PE, Roux S, Parcellier A, Schmitt E, Solary E, et al. Tumor cells convert immature myeloid dendritic cells into TGF-β–secreting cells inducing CD4+CD25+ regulatory T cell proliferation. *J Exp Med* (2005) 202:919–29. doi: 10.1084/jem.20050463
5. Shevach EM. Fatal attraction: tumors beckon regulatory T cells. *Nat Med* (2004) 10:900–1. doi: 10.1038/nm0904-900
6. Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, et al. Pembrolizumab versus Chemotherapy for PD-L1–Positive Non–Small-Cell Lung Cancer. *N Engl J Med* (2016) 375:1823–33. doi: 10.1056/NEJMoa1606774
7. Carbone DP, Reck M, Paz-Ares L, Creelan B, Horn L, Steins M, et al. First-Line Nivolumab in Stage IV or Recurrent Non–Small-Cell Lung Cancer. *N Engl J Med* (2017) 376:2415–26. doi: 10.1056/NEJMoa1613493
8. Hodi FS, O'Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, et al. Improved Survival with Ipilimumab in Patients with Metastatic Melanoma. *N Engl J Med* (2010) 363:711–23. doi: 10.1056/NEJMoa1003466
9. Alsaab HO, Sau S, Alzhrani R, Tatiparti K, Bhise K, Kashaw SK, et al. PD-1 and PD-L1 Checkpoint Signaling Inhibition for Cancer Immunotherapy: Mechanism, Combinations, and Clinical Outcome. *Front Pharmacol* (2017) 8 (561). doi: 10.3389/fphar.2017.00561
10. Chambers CA, Kuhns MS, Egen JG, Allison JP. CTLA-4-Mediated Inhibition in Regulation of T Cell Responses: Mechanisms and Manipulation in Tumor Immunotherapy. *Annu Rev Immunol* (2001) 19:565–94. doi: 10.1146/annurev.immunol.19.1.565
11. Ryder M, Callahan M, Postow MA, Wolchok J, Fagin JA. Endocrine-related adverse events following ipilimumab in patients with advanced melanoma: a comprehensive retrospective review from a single institution. *Endocr Relat Cancer* (2014) 21:371–81. doi: 10.1530/ERC-13-0499
12. van Elsas A, Sutmuller RPM, Hurwitz AA, Ziskin J, Villasenor J, Medema J-P, et al. Elucidating the Autoimmune and Antitumor Effector Mechanisms of a Treatment Based on Cytotoxic T Lymphocyte Antigen-4 Blockade in Combination with a B16 Melanoma Vaccine: Comparison of Prophylaxis and Therapy. *J Exp Med* (2001) 194:481–90. doi: 10.1084/jem.194.4.481
13. Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, et al. Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol Cancer Ther* (2017) 16:2598–608. doi: 10.1158/1535-7163.MCT-17-0386
14. Samstein RM, Lee C-H, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet* (2019) 51:202. doi: 10.1038/s41588-018-0312-8
15. Germano G, Lamba S, Rospo G, Barault L, Magrì A, Maione F, et al. Inactivation of DNA repair triggers neoantigen generation and impairs tumour growth. *Nature* (2017) 552:116–20. doi: 10.1038/nature24673
16. Rosenberg SA, Yang JC, Sherry RM, Kammula US, Hughes MS, Phan GQ, et al. Durable Complete Responses in Heavily Pretreated Patients with Metastatic Melanoma Using T-Cell Transfer Immunotherapy. *Clin Cancer Res* (2011) 17:4550–7. doi: 10.1158/1078-0432.CCR-11-0116
17. Zacharakis N, Chinnasamy H, Black M, Xu H, Lu Y-C, Zheng Z, et al. Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer. *Nat Med* (2018) 24:724–30. doi: 10.1038/s41591-018-0040-8
18. McNamara MA, Nair SK, Holl EK. RNA-Based Vaccines in Cancer Immunotherapy. *J Immunol Res* (2015) 2015:794528. doi: 10.1155/2015/794528
19. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, et al. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med* (2016) 8:11. doi: 10.1186/s13073-016-0264-5
20. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol* (2016) 17:122. doi: 10.1186/s13059-016-0974-4
21. Rubinsteyn A, Hodes I, Kodysh J, Hammerbacher J. Vaxrank: A computational tool for designing personalized cancer vaccines. *bioRxiv* (2018) 142919. doi: 10.1101/142919
22. Rubinsteyn A, Kodysh J, Hodes I, Mondet S, Aksoy BA, Finnigan JP, et al. Computational Pipeline for the PGV-001 Neoantigen Vaccine Trial. *Front Immunol* (2018) 8:1807. doi: 10.3389/fimmu.2017.01807
23. Zhang J, Mardis ER, Maher CA. INTEGRATE-neo: a pipeline for personalized gene fusion neoantigen discovery. *Bioinformatics* (2017) 33:555–7. doi: 10.1093/bioinformatics/btw674
24. Chang T, Carter RA, Li Y, Li Y, Wang H, Edmonson MN, et al. The neoepitope landscape in pediatric cancers. *Genome Med* (2017) 9:78. doi: 10.1186/s13073-017-0468-3
25. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Res* (2008) 36:W509–12. doi: 10.1093/nar/gkn202
26. Fotakis G, Rieder D, Haider M, Trajanoski Z, Finotello F. NeoFuse: predicting fusion neoantigens from RNA sequencing data. *Bioinformatics* 36(7):2260–1. doi: 10.1093/bioinformatics/btz879
27. Toor JS, Rao AA, McShan AC, Yarmarkovich M, Nerli S, Yamaguchi K, et al. A Recurrent Mutation in Anaplastic Lymphoma Kinase with Distinct Neoepitope Conformations. *Front Immunol* (2018) 9:99. doi: 10.3389/fimmu.2018.00099
28. The Cancer Genome Atlas Research network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* (2015) 163:1011–25. doi: 10.1016/j.cell.2015.10.025
29. Carreno BM, Magrini V, Becker-Hapak M, Kaabinejadian S, Hundal J, Petti AA, et al. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science* (2015) 348(6236):803–8. doi: 10.1126/science.aaa3828
30. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* (2013) 499:214–8. doi: 10.1038/nature12213
31. Yoshihara K, Wang Q, Torres-Garcia W, Zheng S, Vegesna R, Kim H, et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* (2015) 34:4845–54. doi: 10.1038/onc.2014.406
32. Shukla SA, Rooney MS, Rajasagi M, Tiao G, Dixon PM, Lawrence MS, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat Biotechnol* (2015) 33:1152–8. doi: 10.1038/nbt.3344
33. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. *Nucleic Acids Res* (2011) 39:D19–21. doi: 10.1093/nar/gkq1019
34. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. *Genome Res* (2002) 12:996–1006. doi: 10.1101/gr.229102
35. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* (2012) 22:1760–74. doi: 10.1101/gr.135350.111
36. Hindman B, Konwinski A, Zaharia M, Ghodsi A, Joseph AD, Katz R, et al. *Mesos: A Platform for Fine-grained Resource Sharing in the Data Center. in Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation NSDI"11*. Berkeley, CA, USA: USENIX Association. Available at: http://dl.acm.org/citation.cfm?id=1972457.1972488 (Accessed June 7, 2019).
37. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, et al. Toil enables reproducible, open source, big biomedical data analyses. *Nat Biotech* (2017) 35:314–6. doi: 10.1038/nbt.3772
38. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* (2009) 25:1754–60. doi: 10.1093/bioinformatics/btp324
39. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* (2013) 29:15–21. doi: 10.1093/bioinformatics/bts635
40. Haas B, Dobin A, Stransky N, Li B, Yang X, Tickle T, et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv* (2017). doi: 10.1101/120295

41. Bai Y, Ni M, Cooper B, Wei Y, Fury W. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC Genomics* (2014) 15:325. doi: 10.1186/1471-2164-15-325
42. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinf* (2011) 12:323. doi: 10.1186/1471-2105-12-323
43. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotech* (2013) 31:213–9. doi: 10.1038/nbt.2514
44. Fan Y, Xi L, Hughes DST, Zhang J, Zhang J, Futreal PA, et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol* (2016) 17:178. doi: 10.1186/s13059-016-1029-6
45. Radenbaugh AJ, Ma S, Ewing A, Stuart JM, Collisson EA, Zhu J, et al. RADIA: RNA and DNA Integrated Analysis for Somatic Mutation Detection. *PloS One* (2014) 9:e111516. doi: 10.1371/journal.pone.0111516
46. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* (2012) 28:311–7. doi: 10.1093/bioinformatics/btr665
47. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* (2012) 28:1811–7. doi: 10.1093/bioinformatics/bts271
48. Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods* (2015) 12:623–30. doi: 10.1038/nmeth.3407
49. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* (2013) 8:1494–512. doi: 10.1038/nprot.2013.084
50. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* (2012) 6:80–92. doi: 10.4161/fly.19695
51. Nielsen M, Lundegaard C, Worning P, Lauemøller SL, Lamberth K, Buus S, et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci* (2003) 12:1007–17. doi: 10.1110/ps.0239403
52. Bui H-H, Sidney J, Peters B, Sathiamurthy M, Sinichi A, Purton K-A, et al. Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics* (2005) 57:304–14. doi: 10.1007/s00251-005-0798-y
53. Peters B, Sette A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinf* (2005) 6:132. doi: 10.1186/1471-2105-6-132
54. Karosiene E, Lundegaard C, Lund O, Nielsen M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* (2012) 64:177–86. doi: 10.1007/s00251-011-0579-8
55. Zhang H, Lund O, Nielsen M. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* (2009) 25:1293–9. doi: 10.1093/bioinformatics/btp137
56. Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinf* (2007) 8:238. doi: 10.1186/1471-2105-8-238
57. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, et al. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol* (1999) 17:555–61. doi: 10.1038/9858
58. Tomlins SA, Laxman B, Varambally S, Cao X, Yu J, Helgeson BE, et al. Role of the TMPRSS2-ERG Gene Fusion in Prostate Cancer. *Neoplasia* (2008) 10:177–88. doi: 10.1593/neo.07822
59. Wang Z, Wang Y, Zhang J, Hu Q, Zhi F, Zhang S, et al. Significance of the TMPRSS2:ERG gene fusion in prostate cancer. *Mol Med Rep* (2017) 16:5450–8. doi: 10.3892/mmr.2017.7281
60. Kron KJ, Murison A, Zhou S, Huang V, Yamaguchi TN, Shiah Y-J, et al. TMPRSS2–ERG fusion co-opts master transcription factors and activates NOTCH signaling in primary prostate cancer. *Nat Genet* (2017) 49:1336–45. doi: 10.1038/ng.3930
61. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat J-P, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet* (2012) 44:685–9. doi: 10.1038/ng.2279
62. Blattner M, Lee DJ, O'Reilly C, Park K, MacDonald TY, Khani F, et al. SPOP Mutations in Prostate Cancer across Demographically Diverse Patient Cohorts. *Neoplasia* (2014) 16:14–W10. doi: 10.1593/neo.131704
63. Kiyotani K, Mai TH, Nakamura Y. Comparison of exome-based HLA class I genotyping tools: identification of platform-specific genotyping errors. *J Hum Genet* (2017) 62:397. doi: 10.1038/jhg.2016.141
64. Zhang J, White NM, Schmidt HK, Fulton RS, Tomlinson C, Warren WC, et al. INTEGRATE: gene fusion discovery using whole genome and transcriptome data. *Genome Res* (2016) 26:108–18. doi: 10.1101/gr.186114.114
65. Liu S, Tsai W-H, Ding Y, Chen R, Fang Z, Huo Z, et al. Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. *Nucleic Acids Res* (2016) 44:e47–7. doi: 10.1093/nar/gkv1234
66. Warren RL, Choe G, Freeman DJ, Castellarin M, Munro S, Moore R, et al. Derivation of HLA types from shotgun sequence datasets. *Genome Med* (2012) 4:95. doi: 10.1186/gm396
67. Merkel D. *Docker: Lightweight Linux Containers for Consistent Development and Deployment. Linux J* (2014). Available at: http://dl.acm.org/citation.cfm?id=2600239.2600241 (Accessed June 7, 2019).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer JI-S and handling editor declared their shared affiliation at the time of the review.