Title

TSAFinder: exhaustive tumor-specific antigen detection with RNAseq.

Permalink

https://escholarship.org/uc/item/67d9t4df

Journal

Computer applications in the biosciences : CABIOS, 38(9)

Authors

Sharpnack, Michael
Johnson, Travis
Han, Zhi
et al.

Publication Date

2022-04-28

DOI

10.1093/bioinformatics/btac116

Peer reviewed

Bioinformatics, 38(9), 2022, 2422–2427
https://doi.org/10.1093/bioinformatics/btac116
Advance Access Publication Date: 22 February 2022
Original Paper

Sequence analysis

TSAFinder: exhaustive tumor-specific antigen detection with RNAseq

Michael F. Sharpnack[1,†,‡], Travis S. Johnson[1,†,§], Robert Chalkley[2], Zhi Han[3], David Carbone[1], Kun Huang[3,*] and Kai He[1,*]

[1]Department of Internal Medicine, Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA, [2]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94158, USA and [3]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN 46202, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
[‡]Present address: Department of Pathology, University of California San Francisco, San Francisco, CA 94143, USA
[§]Present address: Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN 46202, USA

Associate Editor: Christina Kendziorski

Received on July 20, 2021; revised on January 10, 2022; editorial decision on January 13, 2022; accepted on February 19, 2022

Abstract

Motivation: Tumor-specific antigen (TSA) identification in human cancer predicts response to immunotherapy and provides targets for cancer vaccine and adoptive T-cell therapies with curative potential, and TSAs that are highly expressed at the RNA level are more likely to be presented on major histocompatibility complex (MHC)-I. Direct measurements of the RNA expression of peptides would allow for generalized prediction of TSAs. Human leukocyte antigen (HLA)-I genotypes were predicted with seq2HLA. RNA sequencing (RNAseq) fastq files were translated into all possible peptides of length 8–11, and peptides with high and low expressions in the tumor and control samples, respectively, were tested for their MHC-I binding potential with netMHCpan-4.0.

Results: A novel pipeline for TSA prediction from RNAseq was used to predict all possible unique peptides size 8–11 on previously published murine and human lung and lymphoma tumors and validated on matched tumor and control lung adenocarcinoma (LUAD) samples. We show that neoantigens predicted by exomeSeq are typically poorly expressed at the RNA level, and a fraction is expressed in matched normal samples. TSAs presented in the proteomics data have higher RNA abundance and lower MHC-I binding percentile, and these attributes are used to discover high confidence TSAs within the validation cohort. Finally, a subset of these high confidence TSAs is expressed in a majority of LUAD tumors and represents attractive vaccine targets.

Availability and implementation: The datasets were derived from sources in the public domain as follows: TSAFinder is open-source software written in python and R. It is licensed under CC-BY-NC-SA and can be downloaded at https://github.com/RNAseqTSA.

Contact: kunhuang@iu.edu or kai.he@osumc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Background

Neoantigen identification in human cancer predicts response to immunotherapy and provides targets for cancer vaccines and adoptive T-cell therapies (McGranahan *et al.*, 2016; Ott *et al.*, 2017; Sahin *et al.*, 2017). Traditional pipelines for identifying neoantigens rely on identifying somatic alterations from exome sequencing data and predicting the binding affinity of all possible resulting mutated peptides (Hundal *et al.*, 2016); however, there are several pitfalls to this approach. First, these methods are insensitive to post-transcriptional alterations that can produce antigens, such as intron retention, editing, and RNA fusions (Smart *et al.*, 2018; Yang *et al.*, 2019; Zhang *et al.*, 2018). Second, DNA sequencing-based methods may be of limited utility in tumors with low mutation burden (Löffler *et al.*, 2019) that do not contain a sufficient number of neoantigens for vaccine development. These tumors may particularly benefit from the inclusion of non-mutated tumor-specific antigens (TSAs), such as those utilized as vaccine targets to produce tumor regression in melanoma (Sahin *et al.*, 2020).

These challenges can be overcome by incorporating RNA sequencing (RNASeq) data, however, the incorporation of RNASeq data into neoantigen discovery algorithms has previously functioned to filter out neoantigens from lowly expressed genes (Hundal *et al.*, 2016). This is based on evidence that RNA expression of peptide-encoding transcripts can predict major histocompatibility complex (MHC) presentation. A limitation of this approach is that RNASeq reads encoding the actual mutated peptides are not quantified, and it is unclear how well gene and mutant peptide RNA expressions correlate (Löffler *et al.*, 2019). Laumont *et al.* (2018) proposed a method that bypassed exome sequencing to directly quantify TSAs from RNASeq experiments by matching RNA *k*-mers in tumor and control samples; however, this method does not account for the possibility that a peptide can be produced by multiple RNA sequences. A method by Chong *et al.* (2020) accurately detected non-canonical shared and personal antigens using a proteogenomics approach, but their method is not sensitive to the full range of somatic alterations seen in cancer.

Here, the authors describe TSAFinder, a pipeline to overcome these limitations by translating RNASeq reads into all possible small peptides and testing each peptide for tumor specificity and MHC-binding affinity. TSAFinder is sensitive to any genomic or post-transcriptional alteration and directly quantifies TSA RNA expression within a tumor. TSAFinder performed favorably to a prior method by Laumont *et al.* at discovering TSAs in a proteogenomics dataset, and we further implemented TSAFinder on The Cancer Genome Atlas (TCGA) lung adenocarcinoma (LUAD) dataset and discovered a small set of TSAs that is present in a majority of patients (Collisson *et al.*, 2014).

2 Materials and methods

2.1 Data collection

RNAseq fastq files associated with Laumont *et al.* (2018) were downloaded from NCBI GSE111092. RNAseq data for the TCGA LUAD cohort were downloaded from the secure TCGA data portal.

2.2 Detection of TSAs from RNAseq fastq files

The algorithm takes as input paired RNAseq fastq files for matched tumor and control samples. First, seq2HLA is used to impute human leukocyte antigen (HLA)-I genotypes on the tumor sample (Boegel *et al.*, 2012). Second, all fastq files are translated using three-frame translation into peptides of size, k. Here, $8 \le k \le 11$. Third, lists of peptides are filtered as follows:

$$D\frac{(L-k+1)}{L-8+1}\cdot\frac{P}{2\times10^{9}}, \tag{1}$$

where D is the prespecified tumor depth constant (here set to 10 000 for tumor and 2 for control samples), L is the number of peptides encoded in an entire read (RNAseq read length divided by 3, rounded down), k is the peptide size and P is the total number of peptides translated from the RNAseq sample. Fourth, the remaining peptides in the matched control sample are used to filter out non-tumor-specific TSAs from the list of highly expressed tumor peptides. Peptides containing stop codons are not considered for further analysis; however, upstream stop codons do not preclude peptide discovery by TSAfinder, allowing for the discovery of peptides produced by cryptic start sites. Finally, netMHCpan-4.0 is used to calculate binding affinities to yield a final list of TSAs with scores below a rank of 0.5%, corresponding to ‘strong’ binders in the netMHCpan output (Jurtz *et al.*, 2017). All results were produced using Python 2.7 and R 3.5 in a Linux-based operating system. The code was also tested using Python 3.6 and R 4.0.5.

2.3 Annotation of TSAs

The list of TSAs was interrogated for their coding location as follows. The RNAseq reads encoding each peptide are saved from the algorithm above. For each peptide, the encoding reads are assembled into contiguous segments (contigs). For bases in which there is disagreement between reads encoding the same peptide, the base with the most read coverage is used in the final contig sequence. Each contig is searched against the reference genome with BLAST+ (Camacho *et al.*, 2009). In this study, we allow for up to two misaligned bases and up to one alignment gap. For contigs that map incompletely against the reference genome, ends of length sufficient to include peptide-encoding bases are re-searched with BLAST+ to annotate contigs spanning splice junctions and fusion events. Mismatches and gaps in the BLAST+ output are annotated as variants and insertions/deletions, respectively. The variant sequence outputs can be submitted to variant annotation pipelines as a .vfc file with minimal alteration. Finally, gene and exon annotation were performed using the gencode v22 annotation gtf file. The gtf.load R function processes any gtf file for use in this pipeline. To validate the presence of a Ros1 fusion event in the luc2 sample, independent fusion detection was run with STAR-Fusion v1.9.1 with default parameters (Haas *et al.*, 2017).

2.4 Search for neoantigens in a reference proteome

A unique peptide database for $8 \le k$-mer ≤ 11 sized peptides was created as follows. Protein sequences in fasta format were downloaded from protein databank (Berman *et al.*, 2000; rcsb.org). Duplicate protein sequences were removed. Each protein was then broken down into each possible peptide and unique peptides were added to the reference peptidome. ExomeSeq predicted neoantigens were downloaded from The Cancer Immunome Atlas (Charoentong *et al.*, 2017; tcia.at). To obtain RNA expression of each mutant peptide, the neoantigens were searched against the list of peptide expressions created for each tumor and control sample.

2.5 Custom proteomics database search

Proteomics peak list (.mgf) files from the Laumont *et al.* (2018) were searched against two types of databases: a custom sample-specific peptide database and either SwissProt human or mouse entries (depending on the sample) from a database downloaded on April 8, 2019 (human = 20 418 entries; mouse = 17 016 entries). In each case, sequence reversed entries were concatenated to the database to allow false discovery rate (FDR) estimation (Elias and Gygi, 2007). The custom sample-specific peptide data consists of all peptides in the tumor sample that met the threshold for ‘high’ expression in our RNAseq TSA algorithm (see above explanation of expression thresholding). Data were searched with no digestion enzyme specificity allowing for 8 ppm tolerance on precursor ions and 20 ppm tolerance for fragment ions. Variable modifications considered were oxidation of methionine, pyroglutamate formation from peptide N-terminal glutamine and protein N-terminal acetylation (SwissProt databases only). Results were thresholded at a 1% peptide FDR level based on target: decoy database searching results.

2.6 High confidence, high coverage TSA discovery

The TSA information for each of the 39 patients was imported into R. Duplicate TSAs were counted for each patient and the number of patients with each TSA was counted. Criteria were set for high confidence peptides by comparing the median expression and MHC-I binding percentile between presented and non-presented peptides in the Laumont *et al.* lung cancer samples. To generate a set of TSAs that maximized coverage of patients, we developed a greedy set coverage algorithm as follows. First, the list of TSAs is ranked by the number of patients expressing the TSA, and the top TSA is selected. The selected TSA is removed from the peptide list, and the patients covered by that TSA are removed from the patient pool. Next, TSAs are re-ranked by the number of remaining uncovered patients expressing each TSA, and the top TSA is selected. This process is repeated until a majority of patients are covered. The R function is included in the Supplementary Material. Using this simple algorithm, we identified a set of ‘optimal TSAs’ for our cohort of 39 LUAD patients. We ran this algorithm using various TSA inclusion criteria.

3 Results

3.1 Alignment-free detection of TSAs from RNAseq

To quantify sample-specific peptides directly from an RNAseq fastq file, each line of the fastq file was translated via three-frame translation into each possible peptide of a given size (8–11 length peptides for MHC-I binding prediction). The output of this step was a list of unique peptides and their frequencies (Fig. 1). To obtain tumor-specific peptides, the list of peptides produced from the tumor was compared with a list produced from a control sample (or samples) and filtered by expression. HLA A, B and C genotypes were predicted from the fastq file using the seq2HLA algorithm (Boegel *et al.*, 2012), and MHC-I binding predictions for the resulting list of peptides were predicted with netMHCpan4.0 (Jurtz *et al.*, 2017). The final output of these steps is a list of unique peptides, the abundance of the RNAseq reads that encode them, and their predicted MHC-I binding affinities.

To further understand the source of these peptides, the RNAseq reads encoding each TSA were assembled into contigs and matched to the reference genome with BLAST+ (see Section 2). For contigs with partial matches to the genome, the non-mapping ends were re-searched through BLAST+ to allow for mapping of contigs spanning splice junctions. This allowed the algorithm to annotate splice or fusion variants. Next, mismatches and gapped alignments are annotated to detect mutational, editing, insertion or deletion events. Finally, protein coding annotations are added. An example of the annotation output is included in Supplementary Table S1.

To benchmark TSAFinder's performance, we reanalyzed RNAseq samples from a previously published TSA detection method with available RNAseq and MHC-I purified proteomics data. Predicted TSAs were discovered in two murine and seven human tumor samples (Supplementary Table S2). Non-zero RNA expression of 40/40 (100%) of putative TSAs called by Laumont *et al.* were observed by TSAFinder; however, only 3/40 (7.5%) were called as TSAs by TSAFinder (Fig. 2A). The proposed TSAs rejected by TSAFinder were either observed in too few reads in the tumor samples (25/40, 62.5%) or too many reads in the control samples (12/40, 30%) based on thresholds used (see Section 2). For example, eight of the proposed TSAs described by Laumont *et al.* as having zero control expression had non-zero control expression by TSAFinder (Fig. 2B). The difference in expression is likely due to TSAFinder's detection of reads with varying sequences that encode identical peptides, whereas Laumont *et al.* required exactly matching RNA sequences.

To directly compare the performance of the two methods, we searched the proteomics data produced by Laumont *et al.* using custom peptide databases composed of predicted peptides that met the cutoff for high expression within the tumor RNAseq samples, as well as a reference peptide database (see Section 2 and

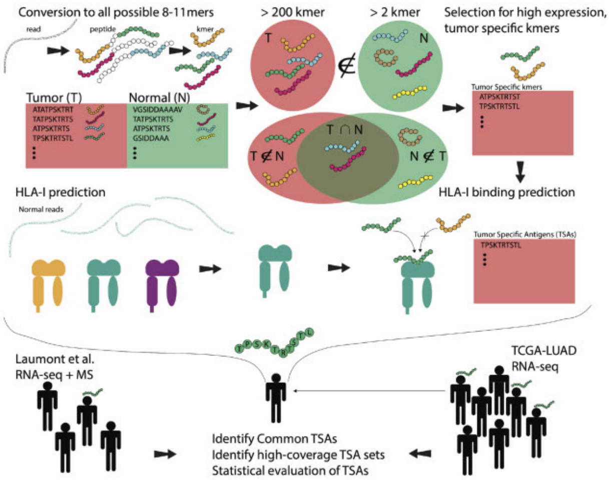

Fig. 1. Algorithm for exhaustive TSA detection from RNAseq. (top left) RNAseq reads are converted to all possible 8–11 mers by three-frame translation in both tumor and normal samples. (top middle) Peptides are compared between tumor and normal samples to find highly expressed tumor-specific *k*-mers. (middle left-to-right) HLA-I genotypes are called and peptides are tested for their ability to bind MHC-I proteins to find TSAs. (bottom) Common TSAs are identified in cohorts of patients and evaluated for their characteristics

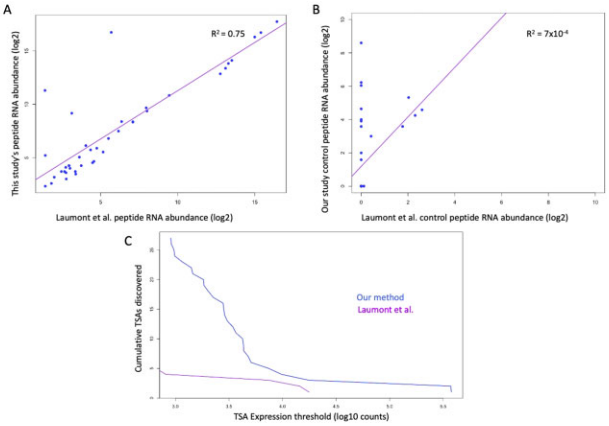

Fig. 2. Comparison with a previous study. (**A**) Peptide tumor RNA expression by Laumont *et al.* versus TSAFinder in 40 TSAs discovered by Laumont *et al.* (**B**) Peptide control RNA expression by Laumont *et al.* versus TSAFinder in 40 TSAs discovered by Laumont *et al.* (**C**) Number of TSAs discovered in the proteomics data at varying RNA expression thresholds. The purple line represents the linear regression in (A) and (B)

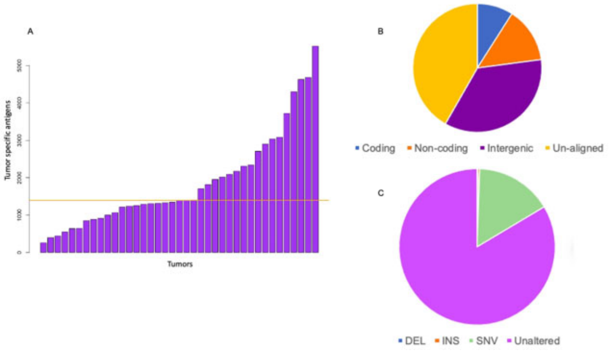

Fig. 3. Features of TSAs discovered in the TCGA LUAD cohort. (**A**) The number of TSAs discovered in each tumor sample. (**B**) The genomic location type of each TSA. (**C**) The mutation status of TSAs

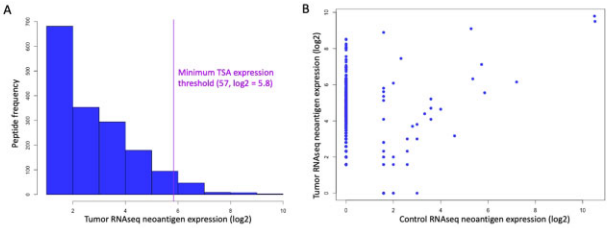

Fig. 4. RNAseq expression of neoantigens in tumor and control TCGA LUAD samples. (**A**) Histogram of neoantigenic peptides' RNAseq expression (log2). The lowest TSA threshold for any patient is shown by a vertical purple line. (**B**) Neoantigenic peptides' RNAseq expression in control versus tumor TCGA LUAD samples

Supplementary Tables S3–S17). After excluding TSAs discovered by Laumont *et al.* with non-zero expression in the control sample, we compared the ability of each algorithm to detect TSAs at a given tumor expression threshold (Fig. 2C). At high expression thresholds, both algorithms performed similarly; however, our algorithm begins to discover a larger number of TSAs with looser thresholding. At the lowest threshold used, TSAFinder discovered 30 peptides in 9 samples.

In the murine samples, only one peptide was called a TSA by both methods (VNYLHRNV in the el4 sample). Of note, Laumont *et al.* showed superior survival of mice immunized to VNYLHRNV than other predicted TSAs when challenged with el4 tumor cells. Additionally, we identified two peptides in the ct26 sample that were not identified in Laumont *et al.*, both of which mapped to viral proteins. Two peptides were identified in total in the human hematologic malignancy samples, SLTALVFHV in samples 07H103 and 12H018, and SQGPQVPPR in 12H018. In two of the hematologic malignancy samples, 10H080 and 10H118, no TSA candidates were identified in the proteomics data. Multiple TSAs were identified in three lung cancer samples, including TSAs from Mucin 5 in two of the samples. The luc6 sample expressed six TSAs that mapped to immunoglobulin heavy chain variable loci. Multiple studies have shown aberrant expression of immunoglobulin genes in cancer cells, and these may represent attractive vaccine targets given their high level of variability and limited expression to select cell types (Hu *et al.*, 2008). Of note, luc2 expressed a peptide from the Ros1 protein, which is frequently involved in an oncogenic fusion event in LUAD. The luc2 RNAseq sample was independently tested for fusion events and found to not have a Ros1 fusion event (see Section 2).

3.2 TSAs of TCGA LUADs

A total of 57 patients' matched tumor-control RNAseq fastq files were downloaded from the TCGA LUAD cohort. Of the 57 patients, 39 (68.4%) were able to be HLA typed successfully with seq2HLA and were followed up for further analysis. An average of 292 438 345 and 286 968 034 peptides were discovered in tumor and matched adjacent non-cancerous tissue control samples, respectively (Supplementary Fig. S1). To minimize the calling of TSAs that were expressed in the patient-specific adjacent lung tissue or human non-lung, non-cancerous tissue, peptides expressed in both the adjacent non-cancerous tissue and the human thymic epithelial cells from Laumont *et al.* were excluded from further analysis. Filtering TSAs with matched control and thymic epithelial cells has been shown to increase specificity (Ehx *et al.*, 2021). In total, 73 096 TSAs encoding 8–11 mers were identified, with an average of 1874 TSAs identified per patient (Fig. 3A and Supplementary Fig. S2). 42 586/73 096 (58%) unique TSAs were able to be aligned to the reference genome (see Section 2 and Supplementary Fig. S3). As in Laumont *et al.*, most of the TSAs aligned to non-protein-coding regions (35 996/42 586, 84.5%; Fig. 3B). Most TSAs were 9 or 10 mers, driven by higher likelihood of binding versus 8 and 11 mers (8.5% and 3.5% of 9 and 10 mers were strong binders, whereas 1.5% and 1.2% of 8 and 10 mers were strong binders; Supplementary Fig. S4). The total number of TSAs per patient did not correlate with the predicted neoantigen burden ($R^2 = 0.060$; Supplementary Fig. S5). We further tested genomic mappings of the peptide-encoding RNA sequences for single nucleotide variants, insertions, and deletions. Only a minority of the TSAs mapped to genomic regions imperfectly, (7003/42 586, 16.4%) the majority of which contained single nucleotide variants (6775/7003, 96.7%; Fig. 3C).

3.3 RNAseq reads encoding neoantigenic peptides are present in control samples

The utility of neoantigens jointly depends on peptide expression and a lack of immune tolerance. To compare TSAs discovered by TSAFinder against neoantigens discovered from previous methods, we searched for neoantigens discovered from WES data by The Cancer Immunome Database (TCIA) within our peptide expression experiments. We first tested for RNA evidence of peptide expression by searching our tumor peptide lists. The average numbers of total neoantigens and those with non-zero expression were 177.6 and 42.8 (42.8/177.6, 28.1%), respectively. This is in agreement with the predicted non-zero gene-level RNA expression of 30–40% neoantigens in lung cancer (Rosenthal *et al.*, 2019). Despite frequent non-zero expression of neoantigens, the average non-zero expression was 15.6 reads per sample, suggesting an overall low expression, given that the minimum expressed TSA in this cohort was seen in 57 reads per sample. Only 8/6928 (0.11%) neoantigens were called TSAs in our dataset. We assessed the appropriateness for using gene

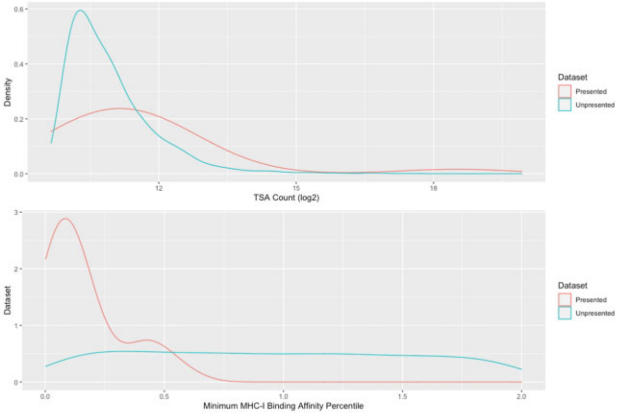

Fig. 5. RNA expression and MHC-I binding affinity in presented and unpresented TSAs. (**A**) Expression measured in peptide counts (log2) and (**B**) Minimum MHC-I binding affinity from netMHCpan

Table 1. High confidence, high coverage TSAs

TSA	Gene	TCGA patients ($n = 39$)
TRVPEVWIL	XAGE1A/B	15
FLDPHSHPF	CPS1	4
VLWQHPPLA	MUC5B	6
EPATRVPEVW	XAGE1A/B	14
IFFNIGENL	NT5DC1	4
SPSGPMRNF	GPX2 antisense	6

expression as a surrogate for TSA expression and found that for the 1555 neoantigens that have corresponding gene-level expressions quantified, the correlation between gene-level expression and peptide-encoding counts is low ($R^2 = 0.028$).

Immune tolerance to neoantigens can occur when the mutant peptide is in fact expressed in a non-mutated protein. Accordingly, the RNA expression of neoantigens in matched control tissue was quantified, and non-zero expression was found in 47/6928 (0.68%) neoantigens. No neoantigens were both found in control tissue and a reference *k*-mer database (see Section 2). Neoantigens with high expression in control tissue also tend to be highly expressed in tumor tissue, suggesting that the expression of these neoantigens may be driven by non-mutated peptide expression in tumor tissue, rather than true neoantigen expression (Fig. 4B).

3.4 A small set of TSAs are present in a majority of LUAD tumors

TSAs across the 39 patients were mapped to 1711 unique genes, and 511 of these genes were observed in at least 2 patients (Supplementary Fig. S6). The most commonly observed genes were *XAGE1A* and *XAGE1B* (24/39, 61.5%). The most frequently observed TSAs come from antisense strands within the *XAGE1A/B* genes (Supplementary Fig. S7).

Next, the proteomics data were used as a training set to identify high-confidence TSAs. Twenty-four peptides were identified both as TSAs by our algorithm and MHC-I bound peptides in the proteomics validation data in the Laumont *et al.* lung cancer samples. High RNA expressions of peptides and high MHC-I binding affinities have been shown to be predictive of neoantigen presentation (Wells *et al.*, 2020), so we compared these values among presented TSAs and unpresented TSAs, all of which were included in the proteomics database search (Fig. 5A and B). Presented TSAs had nearly twice the median RNA expression of unpresented TSAs (2714.5 versus 1659.0 median expressions for presented and unpresented, respectively, Fig. 5A). Similarly, presented TSAs had a median binding affinity of 0.10 percentile (binders were defined as having a binding affinity less than second percentile), while unpresented TSAs had a median affinity of 0.94. Further, presented TSAs mapped with high fidelity to the reference peptidome, indicating that unaligned TSAs were unlikely to be presented.

A greedy set coverage algorithm was developed to identify small sets of peptides as potential vaccine targets (see Section 2). TSAFinder identified six peptides that were observed (satisfying the expression, alignment and MHC-I binding criteria) in 26/39 (66.7%) of the TCGA LUAD tumors (Table 1).

4 Discussion

Here, we presented a method to identify TSAs in matched disease-control RNAseq samples. In contrast to the RNA *k*-mer quantifying method presented by Laumont *et al.*, TSAFinder directly quantifies peptides, which has the added advantage of being able to compare peptides that are produced by non-identical RNA sequences. This could allow for enhanced detection of peptides from hypermutated regions, such as immunoglobulin variable regions, and may explain our discovery of several TSAs therein. This feature explains the existence of several TSAs with zero matched normal RNA *k*-mer expression by the method presented by Laumont *et al.* that had non-zero peptide expression by TSAFinder.

The strength of TSAFinder is directly proportional to the representativeness of the control RNAseq sample used to filter out immuno-tolerant peptides. To limit the false positive identification of TSAs produced by sequencing artifacts, we recommend a significant RNA expression threshold and a control RNAseq sample produced with the same sequencing library preparation. A patient-matched non-cancerous control sample allows for the filtration of patient or population-specific mutations, such as single nucleotide polymorphisms. Here, we found that a portion of exomeSeq predicted neoantigens were encoded by RNA in control tissues. In addition, the expression of many sequences is both temporally and geographically regulated, and therefore, while there was zero expression of a peptide-coding sequence in our matched lung tissue, there may be non-zero expression in other tissues and at different stages of development. To minimize false-positive TSA identifications, samples of human thymic epithelial cells were additionally used to filter out immuno-tolerant peptides. Previous studies have used gene expression in non-cancerous tissues to filter out immuno-tolerant peptides, but we showed that the number of RNAseq reads encoding peptides does not correlate well with gene abundance (Laumont *et al.*, 2018; Perna *et al.*, 2017). Other non-diseased tissue sample databases, such as the genotype-tissue expression project (gtexportal.org), could be used as a more representative control dataset. We recommend the use of both a non-diseased tissue sample database and a patient-matched non-cancerous tissue sample.

Although TSAFinder was developed to discover cancer vaccine targets, it fundamentally discovers differentially expressed sequences of RNA encoding a specific peptide. Therefore, it could be applied to any disease mediated by immunogenic peptides, such as autoimmune diseases driven by the recognition of tissue-specific MHC-I-bound peptides. For example, TSAFinder could discover peptides for inclusion in tolerance-inducing vaccines (Krienke *et al.*, 2021). Further, TSAFinder exhaustively discovers all potentially expressed peptides (prior to post-translational modifications) and could therefore be applied to catalogue the potential proteome of any given RNAseq sample—allowing for the discovery of peptides created from non-canonical reading frames and complex genomic rearrangements that are difficult to align. Finally, an important application of TSA vaccines may be in patients with low neoantigen burden. For example, a recent phase Ib neoantigen vaccine trial excluded all patients with fewer than 50 predicted non-synonymous somatic mutations, indicating the need for more inclusive vaccine generation protocols (Ott *et al.*, 2020).

Consent for publication

All authors consent to the publication of this article.

Availability of data and materials

RNAseq fastq files associated with Laumont *et al.* were downloaded from NCBI GSE111092. RNAseq data for the TCGA LUAD cohort were downloaded from the secure TCGA data portal. All code is available at github.com/RNAseqTSA.

Authors' contributions

M.F.S., T.J.S., K.H. and K.He. conceived of the project. M.F.S. and T.J.S. wrote all code pertaining to the TSA discovery algorithm and the annotation algorithm and generated all of the figures. R.C. performed all proteomics analyses and interpretation. Z.H. downloaded, curated and preprocessed all TCGA sequencing data. M.F.S., T.J.S., R.C., D.P.C., K.H. and K.He. wrote and edited the article.

Financial Support: none declared.

Conflict of Interest: none declared.

References

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **28**, 899–907.

Boegel,S. *et al.* (2012) HLA typing from RNA-Seq sequence reads. *Genome Med.*, **4**, 102.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 1–9.

Charoentong,P. *et al.* (2017) Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.*, **18**, 248–262.

Chong,C. *et al.* (2020) Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.*, **11**, 1293.

Collisson,E. *et al.* (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, **511**, 543–550.

Ehx,G. *et al.* (2021) Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. *Immunity*, **54**, 737–752.e10.

Elias,J.E. and Gygi,S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, **4**, 207–214.

Haas,B. *et al.* (2017). STAR-Fusion: fast and accurate fusion transcript detection from RNA-Seq. BioRxiv. https://doi.org/10.1101/120295.

Hu,D. *et al.* (2008) Immunoglobulin expression and its biological significance in cancer cells. *Cell. Mol. Immunol.*, **5**, 319–324.

Hundal,J. *et al.* (2016) pVAC-Seq: a genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.*, **8**, 11–11.

Jurtz,V. *et al.* (2017) NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.*, **199**, 3360–3368.

Krienke,C. *et al.* (2021) A noninflammatory mRNA vaccine for treatment of experimental autoimmune encephalomyelitis. *Science*, **371**, 145–153.

Laumont,C.M. *et al.* (2018) Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.*, **10**, 1–12.

Löffler,M.W. *et al.*; HEPAVAC Consortium. (2019) Multi-omics discovery of exome-derived neoantigens in hepatocellular carcinoma. *Genome Med.*, **11**, 1–16.

Mcgranahan,N. *et al.* (2016) Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*, **351**, 1463–1470.

Ott,P.A. *et al.* (2017) An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, **547**, 217–221.

Ott,P.A. *et al.* (2020) A phase Ib trial of personalized neoantigen therapy plus anti-PD-1 in patients with advanced melanoma, non-small cell lung cancer, or bladder cancer. *Cell*, **183**, 347–362.e24.

Perna,F. *et al.* (2017) Integrating proteomics and transcriptomics for systematic combinatorial chimeric antigen receptor therapy of AML. *Cancer Cell*, **32**, 506–519.e5.

Rosenthal,R. *et al.*; TRACERx consortium. (2019) Neoantigen-directed immune escape in lung cancer evolution. *Nature*, **567**, 479–485.

Sahin,U. *et al.* (2020) An RNA vaccine drives immunity in checkpoint-inhibitor-treated melanoma. *Nature*, **585**, 107–112.

Sahin,U. *et al.* (2017) Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*, **547**, 222–226.

Smart,A.C. *et al.* (2018) Intron retention is a source of neoepitopes in cancer. *Nat. Biotechnol.*, **36**, 1056–1058.

Wells,D.K. *et al.*; Tumor Neoantigen Selection Alliance. (2020) Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. *Cell*, **183**, 818–834.e13.

Yang,W. *et al.* (2019) Immunogenic neoantigens derived from gene fusions stimulate T cell responses. *Nat. Med.*, **25**, 767–775.

Zhang,M. *et al.* (2018) RNA editing derived epitopes function as cancer antigens to elicit immune responses. *Nat. Commun.*, **9**, 1–10.