scientific reports

Check for updates

OPEN

Potential shared neoantigens from pan-cancer transcript isoforms

Jirapat Techachakrit[1,2], Aijaz Ahmad Malik[1], Trairak Pisitkun[2,3] & Sira Sriswasdi[1,3]✉

Isoform switching in cancer is a prevalent phenomenon with significant implications for immunotherapy, as actionable neoantigens derived from these cancer-specific events would be applicable to broad categories of patients, reducing the necessity for personalized treatments. By integrating five large-scale transcriptomic datasets comprising over 19,500 samples across 29 cancer and 54 normal tissue types, we identified cancer-associated isoform switching events common to multiple cancer types, several of which involve genes with established mechanistic roles in oncogenesis. The presence of neoantigen-containing peptides derived from these transcripts was confirmed in broad cancer and normal tissue proteome datasets and the binding affinity of predicted neoantigens to the human leukocyte antigen (HLA) complex via molecular dynamics simulations. The study presents strong evidence that isoform switching in cancer is a significant source of actionable neoantigens that have the capability to trigger an immune response. These findings suggest that isoform switching events could potentially be leveraged for broad immunotherapeutic strategies across various cancer types.

Keywords Shared pan-cancer neoantigens, Isoform switching, Immunotherapy targets

Cancer, characterized by various hallmark derivatives of dysfunction in healthy cells, presents a pervasive and multifaceted global health challenge[1]. In addition to accumulation of internal molecular changes, cancer cells survive by escaping external pressure from immune recognition[2–4]. As a result, immunotherapy techniques, including neoantigen vaccine[5,6], have been developed to activate the immune system to recognize and destroy cancer cells. Neoantigens are antigen peptides that are specifically presented on the surface of cancer cells. Neoantigens may be unique to an individual patient, also called personalized neoantigens, or shared across multiple patients[7,8]. Shared neoantigens can arise from many mechanisms, such as when multiple tumors accumulate the same mutations or activate the expression of the same transcripts that are not typically expressed in normal tissues[9–11]. Shared neoantigens are sought-after because a single investment in their production can benefit a large number of patients.

Initial discoveries of neoantigens focused on non-synonymous somatic mutations because their high tumor specificity minimizes the possibility of adverse effects on normal tissues[8]. Beyond mutational markers, the complex process of tumorigenesis also produces numerous unique transcriptomic signatures[12–14], including alternative splicing and isoform switching that can be exploited as potential sources of neoantigens[15–17]. Isoform switching occurs when the predominant transcript isoform expressed in normal tissue is replaced by a different isoform in cancer tissue. This process can be monitored at either the whole-isoform level or the local splicing junctions level. In this study, we focused on the switching of the whole isoforms. Mis-splicing of tumor-suppressor genes, such as BRCA1 and PTEN in a breast cancer study[18] and KRAS in lung cancer[19–21], and isoform switching of oncogenes, such as the CD44 cancer variant isoform[22,23], have been reported. Increased exon skipping[24] and isoform dysregulation[17,25] in cancer cells may also produce targetable pan-cancer neoantigens.

Over the past decade, the availability of large cancer and normal tissue transcriptomics datasets, including The Cancer Genome Atlas (TCGA), Pan-Cancer Analysis of Whole Genome (PCAWG), and Gene-Tissue Expression (GTEx), has enabled pan-cancer investigations of isoform switching events and their functional impacts[25–27]. For example, an analysis of 5,500 samples across 12 cancer types in the TCGA dataset revealed 4,446 cancer-specific isoform switching events across 2,352 genes[27], while an analysis of 1,209 samples from 27 cancer types in the PCAWG dataset revealed as many as 31,748 cancer-specific isoform switching events across 7,143 genes[25]. Previous pan-cancer analyses in TCGA and PCAWG identified transcript isoforms that were significantly more highly expressed in tumors compared to matched normal tissues. However, increased

[1]Center of Excellence in Computational Molecular Biology, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand. [2]Center of Excellence in Systems Biology, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand. [3]Center for Artificial Intelligence in Medicine, Research Affairs, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand. ✉email: sira.sr@chula.ac.th

expression alone does not imply cancer specificity. A cancer-specific transcript should also exhibit little to no expression in normal tissues. Among these studies, fewer than 0.2% of these cancer-specific isoform switching events were common among investigated cancer types, indicating a substantial discrepancy in isoform-level transcriptomic landscape across different cancer contexts. Nonetheless, the presence of shared cancer-specific isoforms still has considerable clinical applications.

This study combines five transcriptomic datasets, totaling 19,515 samples from 29 cancer types and 54 normal tissues, to identify cancer-associated isoforms and neoantigens that may be derived from them. Cancer datasets include TCGA (10 cancer types, 548 samples), PCAWG (28 cancer types with some overlap, 1,359 samples), and ONCOBOX Cancer (2 cancer types, 58 samples), while normal tissue data are drawn from GTEx (17,382 samples) and ONCOBOX-ANTE (168 samples). The impacts of dataset choice and *de novo* isoform assembly on the identification of pan-cancer signatures were also investigated. Characterization of HLA binding dynamics of candidate neoantigens indicate that even a small number of shared cancer-specific transcript isoforms can potentially produce many clinically useful neoantigens.

Methods

Transcriptomics data acquisition

Controlled-access raw total RNA-sequencing samples in Fast Access Quality (FASTQ) format and whole-exome sequencing alignment files in BAM format were downloaded from TCGA[28] via the Genomic Data Commons (GDC) Data Transfer Tool Client. Raw total RNA-sequencing in FASTQ format of the ONCOBOX-ANTE[29] and ONCOBOX Cancer[30] projects were downloaded from the Gene Expression Omnibus (GEO)[31] via the Sequence Read Archive (SRA) toolkit[32]. PCAWG[33] and GTEx[34] datasets were obtained as transcript expression tables in Tab-separated Value (TSV) format, downloaded from the International Cancer Genome Consortium (ICGC) data portal[35] and GTEx portal, respectively.

RNA-sequencing data processing

RNA-seq files underwent quality checks using FastQC and were then mapped to the GRCh38.d1.vd1 reference genome using Hierarchical Indexing for Spliced Alignment of Transcripts 2 (HISAT2) version 2.2.1[36]. Aligned reads were sorted and duplicates were marked using Samtools version 1.15.1[37]. Binary Alignment Map (BAM) files were processed using StringTie version 2.2.1[38], utilizing the Gencode v36 genome annotation as a reference. The resulting Gene Transfer Format (GTF) files from all samples were merged using StringTie's merge mode to create a consolidated master GTF file. A secondary StringTie assembly was performed on each BAM file with the master GTF file as reference. Known transcripts along with their expression values in Transcript per Million (TPM) units were extracted.

Detection of pan-cancer-specific transcript isoforms

The pan-cancer analyses were conducted at two levels: across all cancer types in the datasets and specifically focused on breast and lung cancers. The latter analysis was performed because the ONCOBOX Cancer dataset contains only breast and lung cancers. First, isoform switching events were broadly defined as the scenario where a minor isoform (isoform with relatively lower expression in normal tissues) became more highly expressed in cancer tissues. This was identified by comparing the average expression levels in TPM units, with no statistical testing involved because this was an initial filter. Then, among isoform switching events, cancer-specific isoforms were identified by further applying the following criteria: (i) consistent expression in at least 90% of the samples in each cancer type, (ii) an average expression of >1 TPM across cancer tissues, and (iii) an average expression of <1 TPM across normal tissues. These criteria ensure that the candidate isoforms are specific to cancer tissues while allowing for some heterogeneity among cancer samples and background expressions in normal tissues. Each cancer dataset (TCGA, PCAWG, and ONCOBOX Cancer) was also analyzed separately and compared to the pan-cancer results to examine dataset-specific biases.

Functional analyses of pan-cancer-specific transcript isoforms

The coding and non-coding status of each transcript isoform together with the list of all alternative isoforms of each gene was acquired from the ENSEMBL database. The difference in exon usage between the primary isoforms in normal tissues and the isoforms found in cancer tissues were analyzed to compare the changes in active protein domains between the two isoforms. Additionally, functional enrichment analyses utilizing Network Topology-based Analysis (NTA) and Over-representation Analysis (ORA) with *genome* as reference set were performed using WEB-based GEne SeT AnaLysis Toolkit (WebGestalt)[39] to uncover functional implications of pan-cancer-specific transcripts. However, at a false discovery rate cutoff of 5%, no significant enrichment was observed. Functional annotation of candidate isoforms was performed to assess their potential role in tumor biology. While neoantigen identification primarily depends on immunogenicity and tumor specificity, understanding the functional context of these isoforms can provide insights into their oncogenic significance and persistence in cancer cells, which could strengthen their suitability as therapeutic targets.

Neoantigen identification and evaluation

Open Reading Frames (ORFs) were identified from RNA sequences of selected transcript isoforms, which will be read only in the forward direction by the ribosome. Hence, only the three forward reading frames were considered. Each ORF was translated into an amino acid sequence and analyzed with NetMHCPan version 4.1[40] to identify the top 2.0% ranked 9-mers that can possibly bind to an HLA allele (default criteria for weak and strong binders). All available HLA class I alleles supported by NetMHCPan were considered. It should be noted that we focused on 9-mers, which are the most frequently presented by several HLA class I molecules including HLA-A*02:01, to streamline the downstream validation process. Each candidate 9-mer was then searched

against the reference human proteome (ENSEMBL, all isoforms, downloaded July 2023) using the Basic Local Alignment Search Tool for Proteins (BLASTP)'s *blastp-short* setting that has been optimized for short peptide inputs[41]. Only candidate 9-mers not present in the human proteome were retained. This approach highlights candidate neoantigens with novel amino acid sequences but may inadvertently lose those derived from relevant cancer-specific isoforms included in the human proteome database. Additionally, peptides from the human proteome that were similar to these candidate 9-mers (based on BLASTP results) were kept as reference for evaluating each candidate 9-mers' ability to interact with HLA proteins in downstream molecular dynamics (MD) simulations. BLASTP search was performed with a relaxed E-value cutoff of 1.0. For each candidate 9-mer, the top hit, sorted by E-value, that aligned to the whole 9-mer length without gap was selected.

The expressions of these candidate 9-mers in various cancer and normal tissue proteomics datasets were queried by searching for tryptic peptide precursors (up to 2 missed cleavages) of these 9-mers using PepQuery[42]. This step confirms that cancer-specific isoforms with the potential to produce the neoantigens were indeed translated into proteins in some cancer proteomes and are not transcriptional artefacts. All 20 cancer proteomes and 2 normal tissue proteomes (29 and 32 tissue types, and 201 and 50 biological samples, respectively) that are available on the web version of PepQuery were considered. GENe annotation COmpilation and DEscription (GENCODE) version 34 Human was set as the reference background. PepQuery hits where one of the provided tryptic peptides achieved the highest search scores (compared to other tryptic peptides from the reference database and random peptides) were extracted. This criterion corresponds to filtering entries with n_db = 0 and n_random = 0 from PepQuery's output.

Molecular dynamics analysis of neoantigen-HLA interaction

Structural docking of the peptide/MHC complex against HLA-A*02:01 was performed using homology-based PANDORA (Peptide ANchoreD mOdelling fRAmework)[43]. Briefly, PANDORA uses Molecular ORder DEtermination by Exemplar-based Learning in Evolutionary Relatedness (MODELLER)[44] for homology modeling of the Major histocompatibility complex (MHC) I and selects a single template from a custom-made database of known MHC structures. One of the key requirements for modeling with PANDORA is the designation of the anchor amino acid residues, which are used to restrain the loop modeling of the peptide within the MHC groove. A total of 20 candidate models were generated and scored using MODELLER's internal scoring functions, Modeller objective function (molpdf) and Discrete Optimized Protein Energy (DOPE). The best scoring model for each complex was used as the initial structure for further analysis.

Molecular dynamics was performed with the GROningen MAchine for Chemical Simulations (GROMACS) 2023.2 pancake[45] using Amber99sb-ildn force field[46] and a rhombic dodecahedral box with 2-nm minimum distance between each protein and the box boundaries. To reach a physiological salt concentration of 0.15 mol/L, the proteins were solvated in the Simple Point Charge (SPC) water model and neutralized by replacing the appropriate number of counter ions. In all the simulations, the temperature was maintained at 310 K using V-rescale, a temperature coupling algorithm using velocity rescaling with a stochastic term, with a time constant of 0.1 ps. The pressure was isotropically maintained at 1 atm using the Parrinello-Rahman barostat with a time constant of 2 ps. Long-range electrostatic interactions were modeled using Smooth Particle Mesh Ewald (SPME) electrostatics, and non-bonded interactions were modeled with the Verlet cutoff scheme. The cutoff distance for short-range electrostatic and van der Waals interactions was set to 1.2 nm. A timestep of 2 fs was used, with atomic coordinates recorded every 10 ps. The bonds containing hydrogen atoms were constrained using the LINear Constraint Solver (LINCS) algorithm[47]. Before each simulation, the initial structure was modified to minimize the energy via the steepest descent algorithm, with an initial step size of 0.01 nm and a tolerance of 10 kJ/(mol· nm^2. Positional restraints of 1000 kJ/(mol· nm^2 on all heavy protein atoms were enforced during both the 2 ns constant Number of particles, Volume, and Temperature (NVT) equilibration and 2 ns constant Number of particles, Pressure, and Temperature (NPT) equilibration for each system, and production MD simulations were conducted with positional restraints turned off. The simulation period of 200 ns was selected by monitoring the stability of the structures (e.g., Root Mean Square Deviation (RMSD) and Radius of gyration (Rg)).

The Molecular Mechanics/Poisson-Boltzmann (MM/PBSA) or Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) method was utilized to carry out the essential computations for calculating the binding free energy of the MHC/peptide complex. This was performed by running the gmx MM/PBSA program[48,49]. The last 20 ns of the molecular dynamics simulation trajectory were used for estimating the Total Gibbs free energy (G_{TOTAL}) for the chosen peptides. The MM/PB(GB)SA binding free energies of the MHC and peptides can be expressed as described in "End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design" by Wang and colleagues[50].

Results

Impact of the dataset choice on pan-cancer-specific isoform detection

The goal of this study is to identify common instances of transcript isoform switching across 41 cancer cohorts from three cancer databases, TCGA[28], PCAWG[33], and the Oncobox Atlas Cancer (ONCOBOX Cancer) database[30], and two panels of normal tissues, the GTEx[34], and the Oncobox Atlas of Normal Tissue Expression (ONCOBOX-ANTE) database[29]. Selected cancer-specific transcript isoforms were then analyzed to identify cancer-related mechanistic explanations and potential derived neoantigens that are compatible with HLA class I via homology modeling and molecular dynamics simulations (Fig. 1). Initial pan-cancer-specific isoform identification was performed separately for each cancer dataset to minimize batch effects, and the final selection of pan-cancer-specific isoforms was based on the intersection across datasets. To minimize technical bias from the various bioinformatics pipelines that were applied to these datasets, raw sequencing data and sequence alignment files were obtained from the two ONCOBOX databases and the TCGA database (see Materials and Methods) and re-analyzed using the same bioinformatics pipeline. The criteria for defining pan-cancer-specific

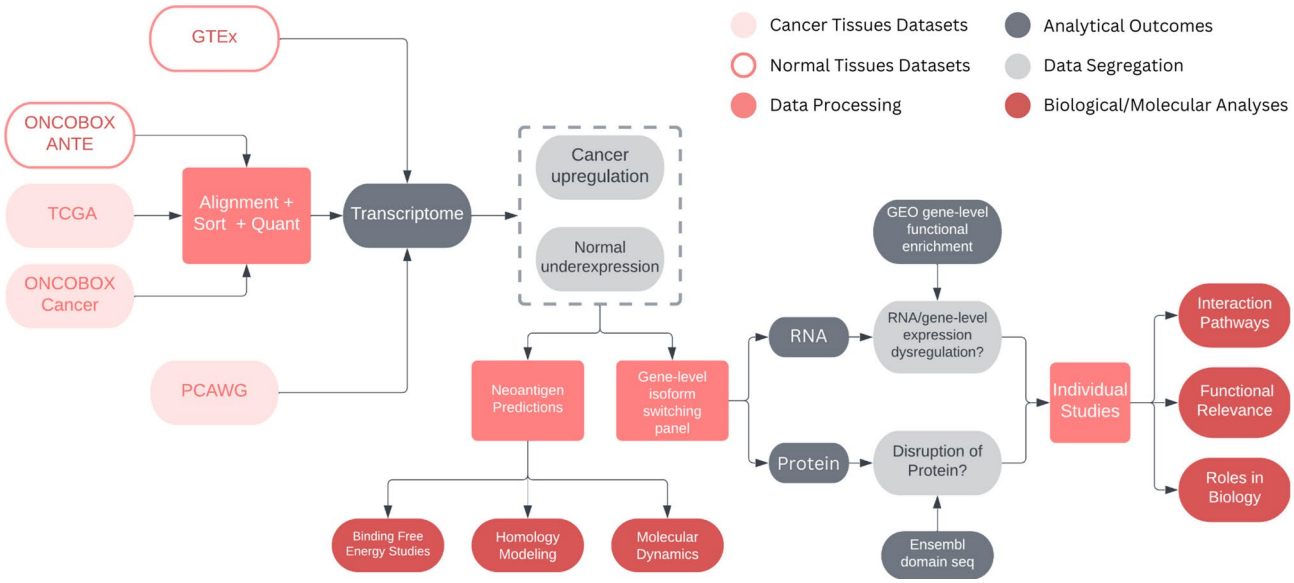

Fig. 1. Our pipeline for identifying pan-cancer-specific isoforms and neoantigens using a combination of standardized isoform calling and computational validation. FASTQ files (ONCOBOX Cancer and ONCOBOX ANTE) and BAM files (TCGA) are processed and merged with transcriptome profiling data in CSV format (GTEx and PCAWG). Transcript isoforms with consistent expression in at least 90% of samples with average TPM > 1 among cancer tissues and low expression with average TPM < 1 among normal tissues are designated as pan-cancer-specific. Selected transcripts then undergo neoantigen assessments, which include HLA binding prediction, alignment against normal peptides, search against cancer and normal tissue proteomes, and molecular dynamics simulation. Functional characteristics (lncRNA or protein-coding) and functional relevance in the context of caner are also evaluated.

isoforms are (i) consistent expression in at least 90% of samples for each cancer type (relaxed from 100% to allow for outliers and heterogeneity), (ii) an average expression of > 1 TPM across cancer tissues, and (iii) an average of < 1 TPM across normal tissues. When only one cancer type is considered, these criteria still apply for identifying cancer-specific isoforms for that cancer type.

Comparison between TCGA and GTEx identified many cancer-specific isoforms for each cancer type, including lymphoid neoplasm Diffuse Large B-Cell lymphoma (DLBC) (110 isoforms, $N = 48$), LUng Squamous Cell carcinoma (LUSC) (94 isoforms, $N = 100$), and REctum ADenocarcinoma (READ) (93 isoforms, $N = 50$) (Fig. 2a). On the other hand, KIdney cancer Renal Papillary cell carcinoma (KIRP) and KIdney cancer Renal Clear cell carcinoma (KIRC), exhibited the fewest cancer-specific isoforms (58 and 69 isoforms, $N = 100$ and 100, respectively). Many of these isoforms are very highly expressed in cancer tissues, with average expression levels well above 100–1000 TPM. When considering all 10 TCGA cancer types, there were 42 isoforms that were found to be pan-cancer-specific according to the three criteria above (Supplementary Table S1 and S2), 39%, 31%, and 16.5% of which involved pseudogenes, mitochondrial transfer RNA (mt-tRNA), and protein coding genes, respectively (Fig. 2b).

In total, we detected 42, 6, and 121 pan-cancer-specific isoforms in the TCGA, PCAWG, and ONCOBOX Cancer compared to GTEx, respectively (Fig. 2c). Additionally, 139 and 123 pan-cancer isoforms were detected in the TCGA and PCAWG compared to ONCOBOX ANTE, respectively. When both GTEx and ONCOBOX ANTE were combined as a single reference, only four pan-cancer-specific isoforms were detected, all from PCAWG. Unexpectedly, no pan-cancer-specific isoform was detected when comparing ONCOBOX Cancer to normal tissues from ONCOBOX ANTE, even though these two datasets were acquired by the same laboratory. The lack of cancer-specific isoforms in this comparison, together with the fact that 121 pan-cancer-specific isoforms were obtained when comparing ONCOBOX Cancer to GTEx (Fig. 2c), strongly indicates systematic differences between the two ONCOBOX datasets and the rest. Therefore, caution was exercised when interpreting results involving ONCOBOX datasets. Lastly, the difference in number of samples and tissue diversity (Fig. 2d) is also expected to influence the number of identified isoforms, with analysis of larger databases yielding lower numbers of isoforms. More discussion on these issues is provided below.

Common cancer-specific isoforms in breast and lung cancer

We focused on breast and lung cancers because they are present in all three cancer databases considered (Fig. 2d). Based on the observation about a potential systematic bias in the ONCOBOX dataset above, GTEx and ONCOBOX ANTE were used separately as the reference normal dataset in these analyses. With GTEx as the reference dataset, when considering both breast and lung cancers, two cancer-specific isoforms were identified in all datasets, namely Immunoglobulin Kappa Joining 4 (IGKJ4) and Immunoglobulin Kappa Joining 1 (IGKJ1), both of which are components of the immunoglobulin protein and are involved in immune responses[51]. When

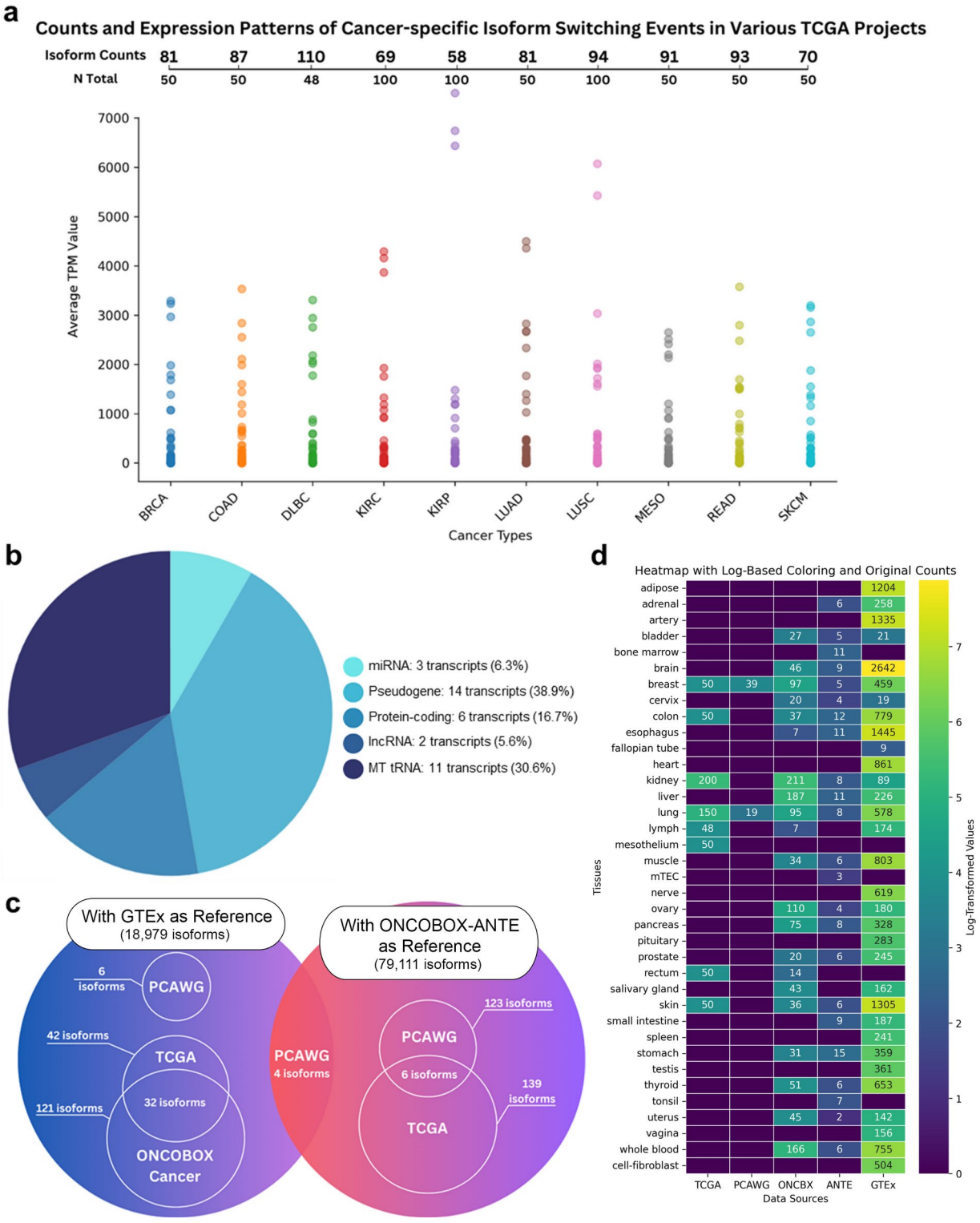

Fig. 2. Summary of the pan-cancer transcriptomics datasets and isoform analysis. (**a**) Counts and expression patterns of cancer-specific isoform events in across 10 TCGA projects grouped by cancer types. (**b**) Functional profile of pan-cancer-specific isoforms identified from the comparison of TCGA versus GTEx as normal reference. (**c**) Venn diagram showing the number and overlap of pan-cancer-specific identified using different cancer and normal tissue databases. Pan-cancer-specific transcripts are required to be expressed in at least 90% of cancer samples with average TPM > 1 and low expression with average TPM < 1 in normal samples. (**d**) The heatmap showing sample counts for each tissue type in each database.

focusing on breast cancer alone, three cancer-specific isoforms were identified in all three datasets and two additional isoforms were identified in both TCGA and PCAWG (Supplementary Table S3, S4, S5, and Fig. 3a). All five isoforms involve genes belonging to immunoglobulin proteins, namely the Immunoglobulin Heavy Joining 3, 4, 5 genes (IGHJ3, IGHJ4, IGHJ5) or Immunoglobulin Kappa Joining 1 and 4 genes (IGKJ1, IGKJ4). Similarly, when focusing on lung cancer alone, six cancer-specific isoforms were identified in all three datasets and three additional isoforms were identified in TCGA and PCAWG (Supplementary Table S3 and Fig. 3b). Again, all nine isoforms involved genes belonging to either immunoglobulin heavy chain (IGHJ1, IGHJ2, IGHJ3, IGHJ4, IGHJ5) or immunoglobulin kappa chain (IGKJ1, IGKJ3, IGKJ4, IGKJ5).

With ONCOBOX ANTE as the reference dataset, completely different results were obtained. When focusing on breast cancer alone, 16 cancer-specific isoforms were identified in both TCGA and PCAWG, namely ribosomal proteins (Ribosomal Protein S6 (RPS6), Ribosomal Protein L7a (RPL7A), Ribosomal Protein L14 (RPL14), Mitochondrial Ribosomal Protein L11 (MRPL11), and Proteasome 20 S Subunit Alpha 2 (PSMA2)),

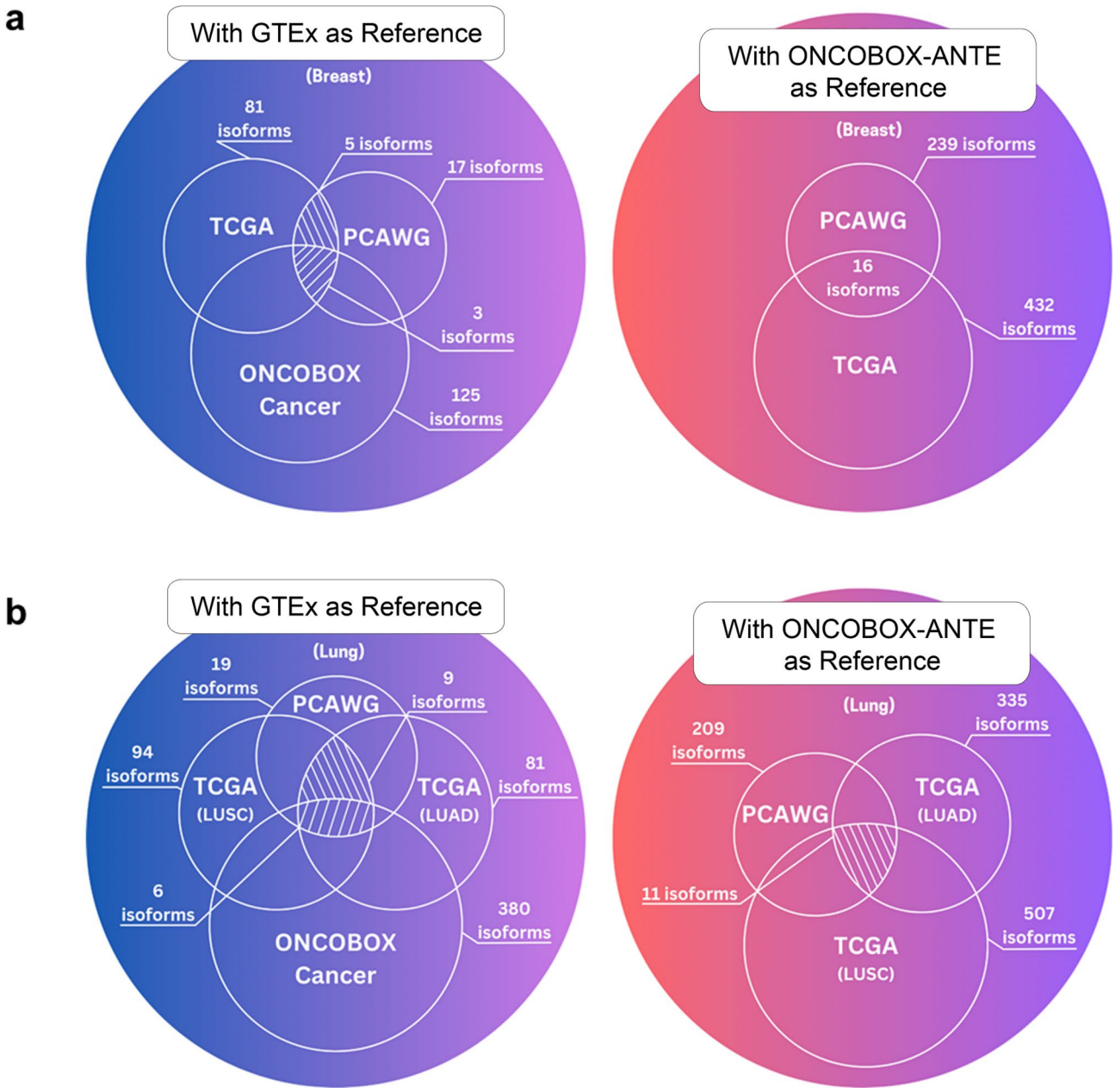

Fig. 3. Summary of isoform analysis in breast and lung cancer. (**a**) Venn diagram showing the number and overlap of pan-cancer-specific identified using different breast cancer and normal tissue databases. (**b**) Venn diagram showing the number and overlap of pan-cancer-specific identified using different lung cancer and normal tissue databases. TCGA-LUSC and TCGA-LUAD were shown separately. In both analyses, cancer-specific transcripts are required to be expressed in at least 90% of cancer samples with average TPM > 1 and low expression with average TPM < 1 in normal samples.

mitochondrial protein NADH Ubiquinone Oxidoreductase Subunit A3 (NDUFA3), protein processing and degradation (Pre-mRNA Processing Factor 31 (PRPF31), and Proteasome 20 S Subunit Beta 5 (PSMB5)), cellular signaling WD Repeat Domain 54 (WDR54), enzymes and chaperones (Peptidylglycine Alpha-Amidating Monooxygenase (PAM), NAD(P)H Quinone Dehydrogenase 1 (NQO1), Cystatin C (CST3)), RNA-binding DAZ Associated Protein 2 (DAZAP2), inflammation AE Binding Protein 1 (AEBP1), cellular migration Actin Gamma 1 (ACTG1), and Heat Shock Protein Family A Member A1 (HSPA1A). When focusing on lung cancer alone, 11 cancer-specific isoforms were identified in both TCGA and PCAWG, namely ribosomal proteins (RPS6, RPL7A, and RPL14), cellular signaling and structure (WDR54), transporter Solute Carrier Family 10 Member 3 (SLC10A3), protein processing and degradation (PSMB5, PSMA2), RNA-binding protein (DAZAP2), S100 Calcium Binding Protein A6 (S100A6), cellular migration (ACTG1), and heat shock stress response (HSPA1A).

It is interesting to note that regardless of the choice of reference normal tissue dataset, many cancer-specific isoforms were found in common between breast cancer and lung cancer. The enrichment of immunoglobulin (IG) transcripts in TCGA and PCAWG breast and lung cancer samples may be attributed to infiltrating antibody-producing plasma B cells in the tumor microenvironment, as bulk RNA sequencing data captures both tumor and immune cell signals. However, previous studies suggest that cancer cells themselves may express IG-like proteins, which have been implicated in tumor growth and immune evasion. In addition, the enrichment of ribosomal protein genes (RPS6, RPL7A, RPL14) and proteasome-related genes (PSMB5, PSMA2) in cancer samples when ONCOBOX-ANTE was used as the reference may be functionally relevant. Dysregulation of ribosomal proteins has been linked to altered protein synthesis rates, contributing to tumor progression and cellular proliferation. Similarly, proteasome components like PSMA2 and PSMB5 have been implicated in cancer cell survival and therapeutic resistance. These findings suggest that isoform-level changes in fundamental cellular machinery could play a role in cancer progression and warrant further exploration. However, the interpretation must be made carefully because such enrichment can also arise from the degradation or missed detection of certain transcripts in the normal tissue datasets.

Functional relevance of pan-cancer isoforms

To validate our findings, we annotated the functions of selected genes that consistently underwent isoform switching in multiple cancer types and explored their involvement in cancer progression. We first concentrated on Leucine Rich Repeat Transmembrane Neuronal 4 (LRRTM4), which displayed isoform switching in PCAWG samples when compared to both the GTEx and the ONCOBOX-ANTE databases. A study by Zhang and colleagues discovered a long non-coding RNA (lncRNA) (lnc-LRRTM4, LRRTM-207) located downstream of LRRTM4 locus that can promote proliferation, metastasis, and epithelial-mesenchymal transformation (EMT) in colorectal cancer by binding to and activating LRRTM4 promoter[52]. Furthermore, a knockdown of lnc-LRRTM4 inhibits both the gene and protein expressions of LRRTM4. Interestingly, the identified pan-cancer isoform of LRRTM4, LRRTM-206, overlaps the lnc-LRRTM4 locus and its RNA sequence is 70.3% identical to lnc-LRRTM4 (Fig. 4a,b). The two RNAs also possess many matched triplex-forming oligonucleotides (TFO) sequences which suggest that the cancer-specific LRRTM4 isoform can also bind and activate LRRTM4 promoter (Fig. 4c). Hence the LRRTM4-206 isoform (Fig. 4d,e) may play a similar regulatory role in modulating LRRTM4 gene expression in the cancer cohorts via similar mechanisms.

Among 32 pan-cancer isoforms identified in both TCGA and ONCOBOX Cancer datasets with GTEx as the reference (Supplementary Table S1 and Fig. 2c), there were two additional genes of interest, the heterogeneous nuclear ribonucleoprotein C1/C2 (HNRNPC) and the selenium binding protein 1 (SELENBP1). In normal tissues, HNRNPC regulates RNA metabolism by binding to pre-mRNA molecules and regulating splice site selection. HNRNPC serves as a biomarker for atherosclerosis and cervical cancer[53,54]. Elevated HNRNPC expression also correlates with poor overall survival and disease-free survival in several cancer types[55], while the suppression of HNRNPC has been reported to inhibit hepatocellular carcinoma proliferation, migration and invasion[56]. In our analysis, HNRNPC-210 was the only highly expressed isoform in normal tissues, whereas HNRNPC-203 and HNRNPC-224 were exclusively expressed in broad cancer types (Fig. 5a,b). However, as the HNRNPC-203 isoform is missing from GTEx pre-processed data (Fig. 5c), we denoted only HNRNPC-224 as the pan-cancer specific isoform. This isoform skips one exon near the N-terminus and gains one exon near the C-terminus of the protein, compared to the HNRNPC-210 isoform in normal tissues (Fig. 5b,d).

For SELENBP1, the highly expressed isoforms in normal tissues were annotated as retained introns (Supplementary Fig. S1, SELENBP1-214 and SELENBP1-218). On the other hand, isoforms SELENBP1-201 and SELENBP1-212, which were exclusively expressed in broad cancer types, include a protein-coding isoform (SELENBP1-201). However, as the SELENBP1-201 isoform is missing from GTEx pre-processed data, we denoted only SELENBP1-212 as the pan-cancer specific isoform. Overexpression of SELENBP1 is generally associated with suppression of cell proliferation, migration, and invasion of tumors. In a colorectal cancer study, suppression of SELENBP1 is correlated to increased tumor size and unfavorable patient prognosis[57]. Although the mechanism is not fully understood, it is believed that SELENBP1 inhibits EMT through the expression of E-cadherin and inhibition of N-cadherin. Furthermore, SELENBP1 expression is anti-correlated with those of S100A protein family which have close ties to EMT and tumor metastasis[57,58].

Exploring the actionable potential of pan-cancer isoform-derived neoantigens

To explore whether these pan-cancer-specific transcripts may give rise to actionable neoantigens, nine representative transcripts, namely ENST00000281428 (FLI1), ENST00000331035 (IL3RA), ENST00000396617 (MKL1), ENST00000463664.

(SELENBP1), ENST00000488161 (CYREN), ENST00000518824 (VEGFA), ENST00000522370 (TBCA), ENST00000556513 (HNRNPC), and ENST00000620660 (RYK) were analyzed in more depths. These representative transcripts were selected based on their frequent recurrence across cancer datasets and functional

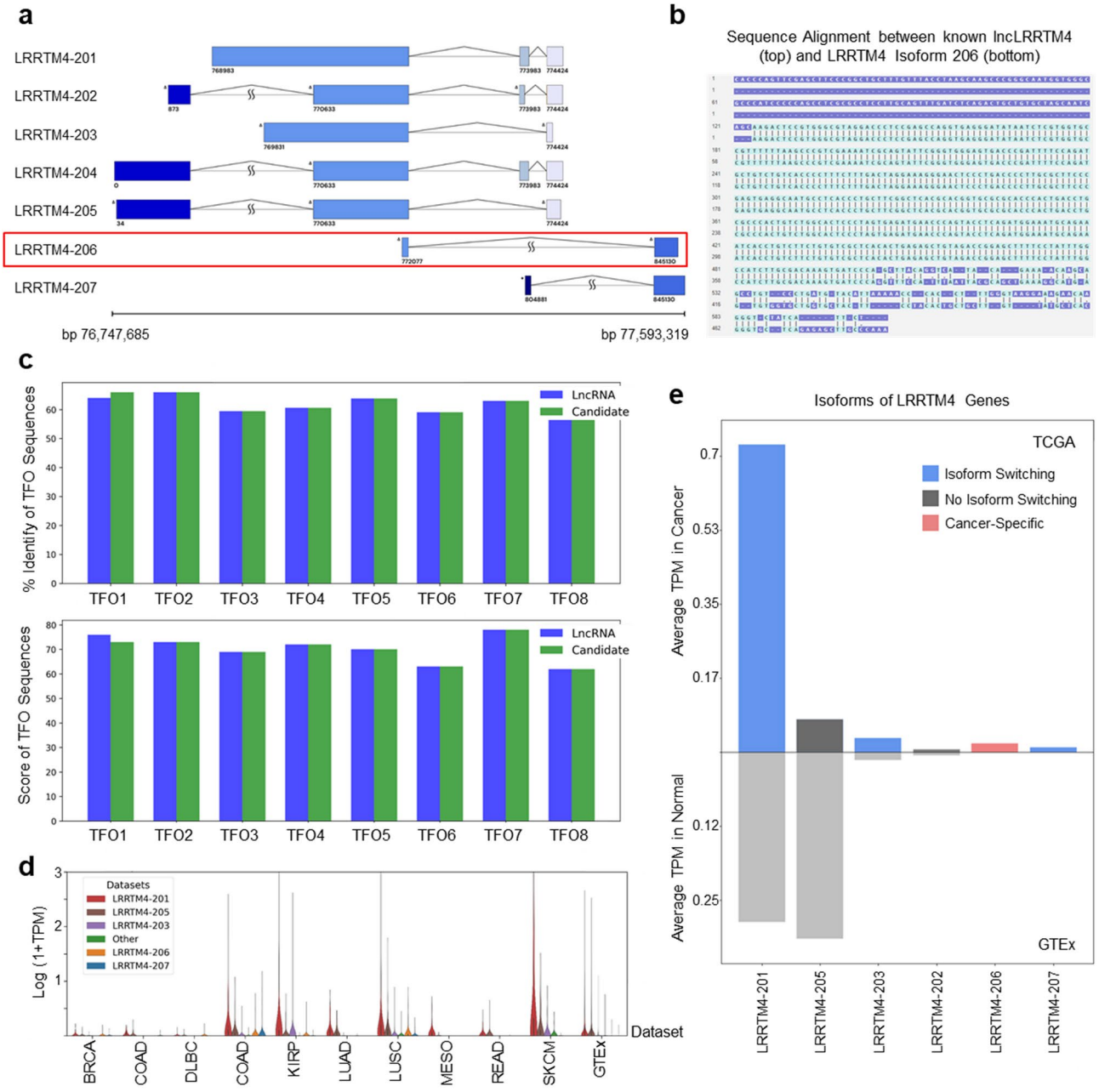

Fig. 4. In-depth characterization of isoform-switching in LRRTM4 as a known functional lncRNA. (**a**) Annotated isoform structure profiles for LRRTM4 gene. LRRTM4-206 is pan-cancer-specific. (**b**) Nucleotide sequence alignment between LRRTM4-206 isoform and lncLRRTM4, a known functional lncRNA. (**c**) Sequence identities and similarity scores between the promoter-binding triplex-forming oligonucleotide (TFO) domains of LRRTM4-206 and lncLRRTM4. (**d**) Expression profiles of LRRTM4 isoforms across cancer types in TCGA and GTEx datasets. (**e**) Side-by-side comparison of expression profile of LRRTM4 isoforms between TCGA and GTEx datasets. Candidate isoform-switching and cancer-specific isoforms are highlighted.

relevance in the context of cancer. Overall, from 96 non-redundant open reading frames (ORFs) on these transcripts, NetMHCPan[40] identified 268 unique 9-mer peptides that are not present in the reference human proteome and can bind strongly (% rank eluted ligand < 0.5) to some HLA alleles (Supplementary Table S6). An additional 383 weak binders (% rank eluted ligand between 0.5 and 2.0) were also found. The total number of potential peptide-HLA interactions is 1,114 (one 9-mer may be able to bind to multiple HLA alleles). It should be noted that while our approach of excluding all 9-mers found in the reference human proteome database helps highlight neoantigens with novel amino acid sequences, the reference proteome also contains annotated cancer-specific isoforms which will cause some valid neoantigens to be inadvertently removed. This issue may

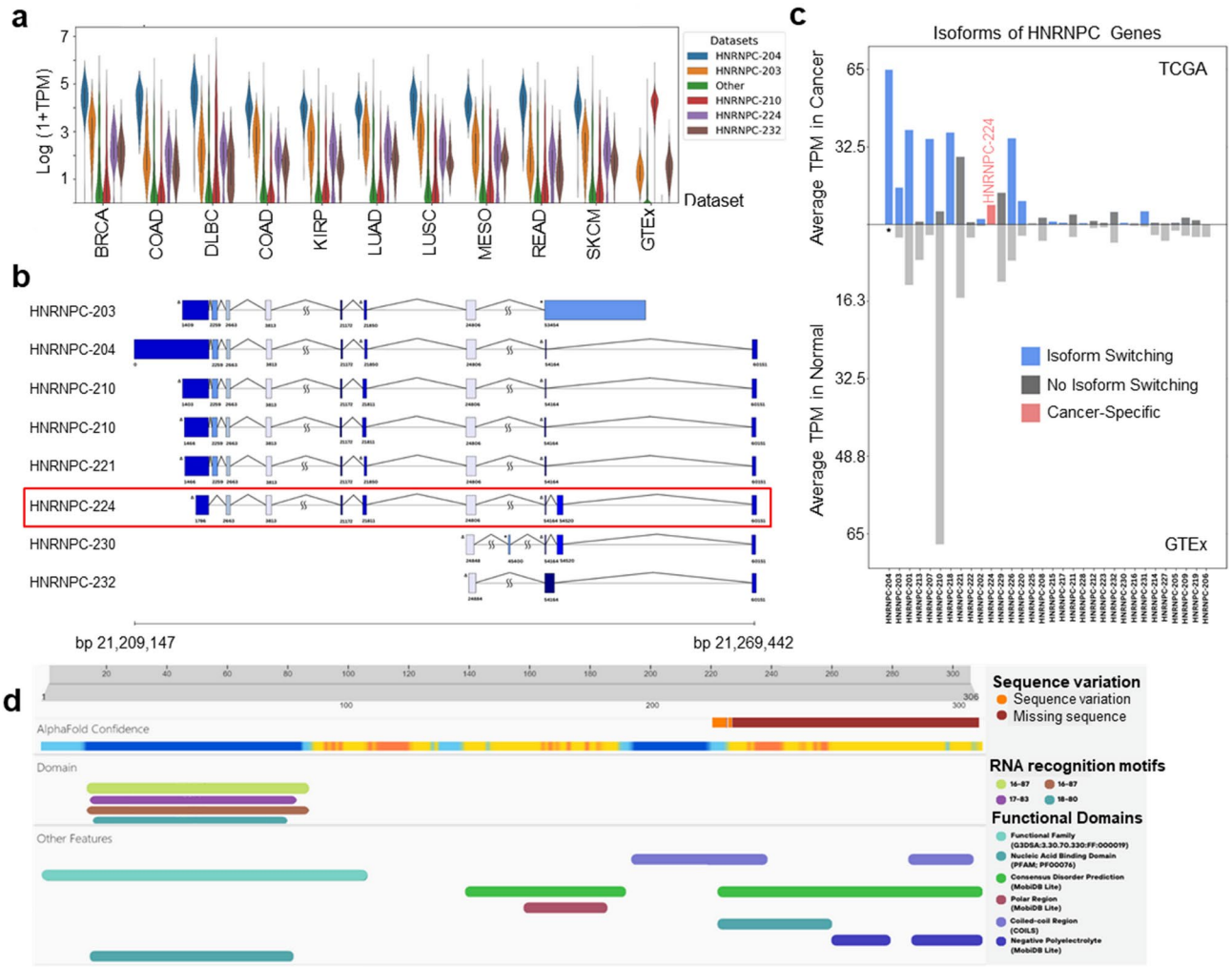

Fig. 5. In-depth characterization of isoform-switching in HNRNPC. (**a**) Expression profiles of HNRNPC isoforms across cancer types in TCGA and GTEx datasets. (**b**) Annotated isoform structure profiles for HNRNPC gene. HNRNPC-224 is pan-cancer-specific. (**c**) Side-by-side comparison of expression profile of HNRNPC isoforms between TCGA and GTEx datasets. Candidate isoform-switching and cancer-specific isoforms are highlighted. Asterisk indicates isoform HNRNPC-204 which is not reported by GTEx. (**d**) The Nightingale Plot, generated via InterPro, to compare protein sequence, structure, and function between HNRNPC-224 and the canonical isoforms of HNRNPC. Sequence variation indicates the differences in amino acid sequence and insertion/deletion. AlphaFold confidence shows the quality of predicted 3D protein structure for the HNRNPC-224 isoform. Positions of annotated functional domains are shown with different colors to indicate whether the regions that differ between the two isoforms contain an important function.

be mitigated by cross-referencing with human proteomics data of normal tissues and by incorporating a more nuanced selection of reference protein isoforms.

Next, PepQuery[42] was used to check whether pan-cancer-specific isoforms are translated into proteins in cancer, and not normal, tissue proteomes, and hence likely to produce cancer-specific neoantigens. As most bottom-up proteomics analyses used trypsin to digest proteins, all possible tryptic peptide precursors of these candidate 9-mers (with up to 2 missed cleavages) were searched against 20 cancer and 2 normal tissue proteomics datasets (Supplementary Table S7). It should be noted that the normal tissue datasets considered contain 29 and 32 distinct tissue types, with 201 and 50 biological samples, respectively, and should adequately serve as negative controls. Overall, tryptic peptide precursors for 155 candidate 9-mers were detected in some cancer proteomes but not in normal tissues. These tryptic peptides were observed in 5–6 cancer datasets on average (median = 5, 25-th percentile = 3, 75-th percentile = 7). Markedly, nine of these candidate 9-mers may be expressed in as many as 10–14 out of 20 cancer datasets investigated.

HLA-binding dynamics as a neoantigen screening tool

To further investigate the potential of putative pan-cancer neoantigens as immunotherapy targets, molecular docking to the HLA protein followed by a 200-nanosecond dynamics simulation were performed for five putative 9-mer neoantigens predicted to bind strongly to HLA-A*02:01. Although molecular dynamics simulations

are computationally costly, they offer valuable mechanistic insights, such as specific bonding interactions and the overall structural stability of molecules, which enhance confidence in predicted binding affinities. These simulations can also identify key amino acid residues in candidate neoantigens that contribute to binding and propose the surface of the HLA-antigen complex that will be accessible to the T cell receptor. These include NLIPLFLYL and YIFGFVRNL from ENST00000556513 (HNRNPC), MLTPSLPSI and SLFWALTSL from ENST00000463664 (SELENBP1), and FLFRGKFHA from ENST00000620660 (RYK). The PANDORA pipeline[43] was used to generate the initial docked structure and GROMACS with Amber force field was used to simulate the dynamics (see Materials and Methods). The root mean square deviation, radius of gyration, and solvent accessible surface area were monitored throughout the 200-ns period to ensure that the antigen-HLA complex has stabilized (Supplementary Fig. S2). Then, hydrogen bonding and binding free energy were calculated from the last 20 ns of the simulation to capture the characteristics of the antigen-HLA complex. Moreover, the HLA binding properties of the selected 9-mers were also compared to those of matched normal peptides with similar amino acid sequences as controls (Supplementary Table S8).

Among the selected putative neoantigens, NLIPLFLYL, MLTPSLPSI, and SLFWALTSL exhibited lower free energies and more frequent hydrogen bonding with the HLA protein compared to the matched normal peptides (Figs. 6 and 7, and Supplementary Fig. S3, respectively). For example, when comparing NLIPLFLYL to the matched normal peptide, WAIPLFLFL, there were considerably more hydrogen bonds between NLIPLFLYL and residues Y7, E63, Y84, Y99, T143, Y159, T163, and Y171 on the HLA protein (Fig. 6 and Supplementary Fig. S2G-H). This resulted in a much lower free energy for the neoantigen-HLA complex (ΔE = − 83.47 kJ/mol versus − 51.95 kJ/mol). Similarly, there were more hydrogen bonds between MLTPSLPSI and residues E63, K66, D77, Y84, Y99, T143, and K146 of the HLA protein compared to when the matched normal peptide MNTPLPSI was considered (Fig. 7). This resulted in much lower free energy for the neoantigen-HLA complex (ΔE = − 102.97 kJ/mol versus − 56.88 kJ/mol). Results for the two putative neoantigens that exhibit weaker binding affinity with HLA were provided in Supplementary Figs. S4 and S5. For these putative neoantigens, there were more hydrogen bonds formed between HLA protein and the matched normal peptides as well as structural distortions in the HLA alpha helices when accommodating putative neoantigens.

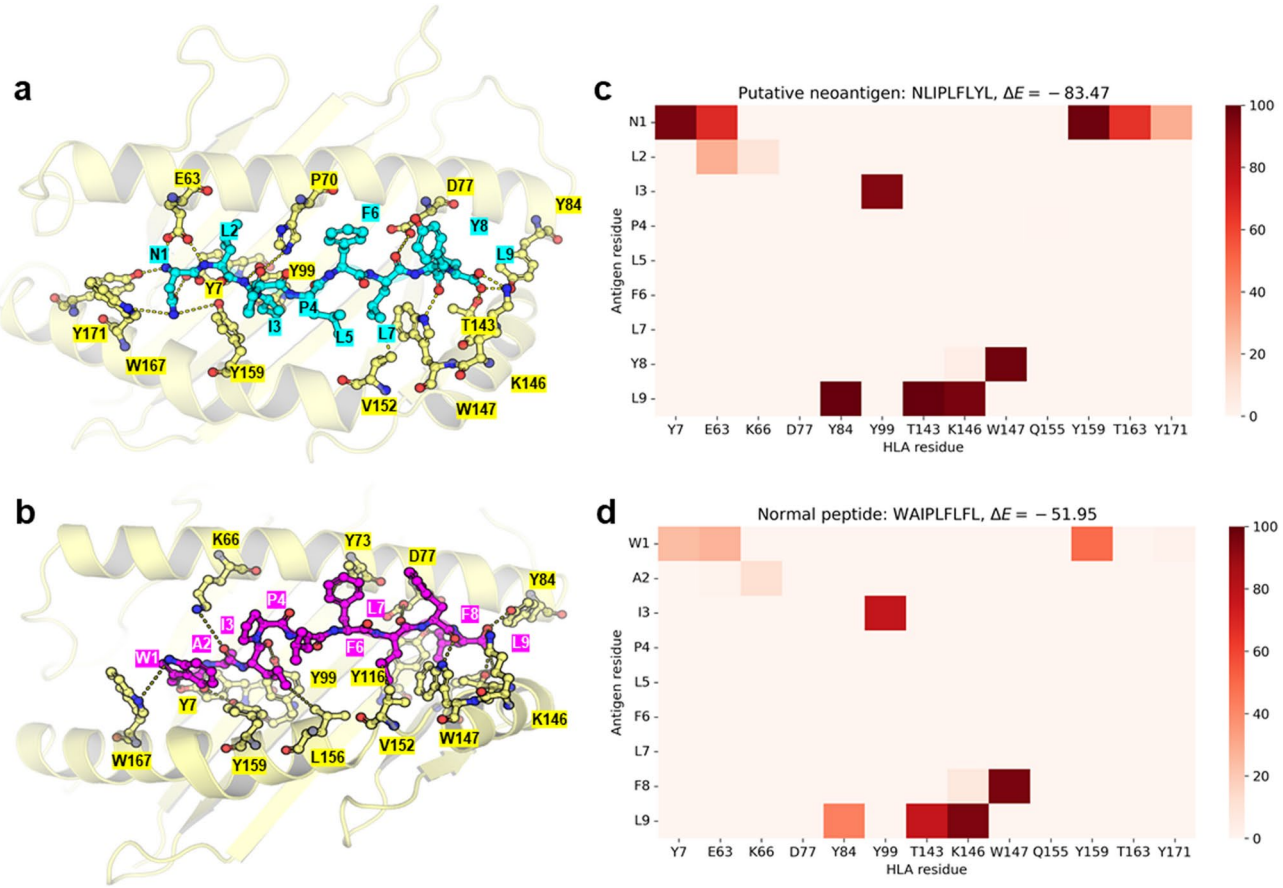

Fig. 6. Molecular dynamics comparison between the binding of putative neoantigen NLIPLFLYL and matched normal peptide WAIPLFLFL to HLA-A*02:01. (**a**) A final structure for putative neoantigen NLIPLFLYL derived from ENST00000556513 transcript. (**b**) A final structure for the matched normal peptide WAIPLFLFL. The side chains of the residues involved in hydrogen bonds (indicated by dashed lines) were highlighted. (**c**) Heatmap showing the percentage of hydrogen bond formation during the last 20 ns of the dynamics for the putative neoantigen NLIPLFLYL. (**d**) Heatmap for the matched normal peptide WAIPLFLFL.

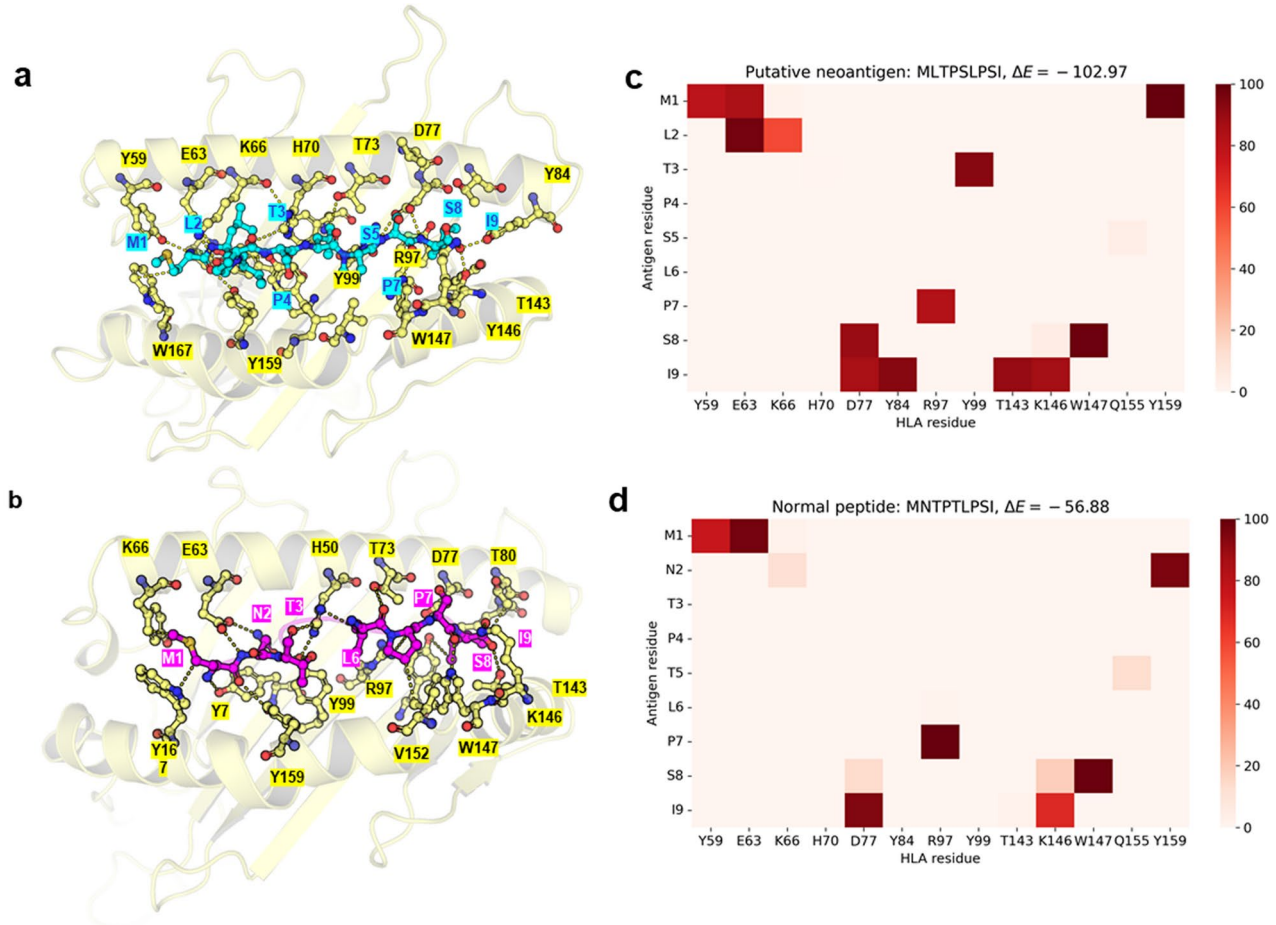

Fig. 7. Molecular dynamics comparison between the binding of putative neoantigen MLTPSLPSI and matched normal peptide MNTPTLPSI to HLA-A*02:01. (**a**) A final structure for putative neoantigen MLTPSLPSI derived from ENST00000463664 transcript. (**b**) A final structure for the matched normal peptide MNTPTLPSI. The side chains of the residues involved in hydrogen bonds (indicated by dashed lines) were highlighted. (**c**) Heatmap showing the percentage of hydrogen bond formation during the last 20 ns of the dynamics for the putative neoantigen MLTPSLPSI. (**d**) Heatmap for the matched normal peptide MNTPTLPSI.

Discussion

We conducted a comprehensive analysis of transcript isoform expression across 41 cancer cohorts from the TCGA, PCAWG, and ONCOBOX Cancer databases, alongside normal tissue panels from GTEx and ONCOBOX ANTE, and identified pan-cancer-specific transcript isoforms as well as those specific to breast and lung cancers. Several of these transcripts derive from genes with known functional roles in cancer progression, such as LRRTM4, HNRNPC, and SELENBP1. Although neoantigen selection is primarily based on immunogenic properties, the functional relevance of cancer-associated isoforms may contribute to their persistence in tumors. Isoforms with roles in oncogenesis or immune evasion may be more stable targets for immunotherapy, as they are less likely to be lost due to immune pressure. However, functional relevance alone does not determine tumor specificity, which remains a key criterion in candidate selection. Encouragingly, the cancer-specific segments of these transcripts, which harbor a number of promising 9-mer neoantigens predicted to bind strongly to the HLA proteins, were confirmed by PepQuery to be translated into proteins in multiple cancer tissues and thus could be captured by the antigen presentation pathway. PepQuery analysis also indicates that the protein precursors of these 9-mers were not detected in normal proteomes. Finally, the ability of predicted 9-mer neoantigens to bind to HLA proteins were demonstrated via molecular dynamics simulations, with similar peptides from the human proteomes as reference. Despite the scope of the study being limited to in silico analyses and restricted to 9-mers, our study provided strong evidence that pan-cancer transcript isoforms can generate actionable neoantigens. Inclusion of 8- to 11-mer peptides will produce a much broader neoantigen repertoire but will require more stringent screening criteria to minimize false positive.

It should be noted that the size of the transcriptomics dataset, in addition to the differences in patient populations, such as genetics background and cancer subtypes, can also influence the number of identified pan-cancer transcript isoforms. For example, the lower numbers isoforms derived from PCAWG ($N = 1,359$) compared to TCGA and ONCOBOX Cancer ($N = 648$ and 58, respectively) could very well be because it is more difficult for a transcript to meet the pan-cancer selection criteria over such large number of samples. On the

other hand, the use of ONCOBOX-ANTE ($N = 168$) as the normal background would naturally yield higher numbers of identified pan-cancer isoforms compared to when GTEx ($N = 17{,}382$) was selected. However, the use of a small normal tissue database as reference has the danger of producing false positive isoforms that are in fact expressed in normal tissues that are underrepresented in the database. These false positive isoforms are difficult to filter out in later stages because proteomics analysis has limited sensitivity and may not detect the corresponding translated proteins and peptides. Hence, it is advisable to include as many normal tissue samples as reference during the initial transcriptomics analysis stage. This still presents a challenge because there are clearly systematic differences in how isoforms are called across transcriptomics datasets (such as GTEx and ONCOBOX-ANTE) that necessitate a careful standardization of the bioinformatics pipeline. Therefore, additional biological information, such as the functional relevance of the detected isoforms in the context of cancer, is valuable in supplementing the screening process.

Most importantly, systematics differences between datasets can also arise due to technical reasons. As noted earlier, the lack of detected pan-cancer transcript isoforms when comparing ONCOBOX Cancer to ONCOBOX ANTE is highly unexpected because both datasets should be the most homogenous, given that they were produced by the same laboratory and were the smallest. As these datasets were re-analyzed by our bioinformatics pipeline, it may be possible that the isoform assignment was so homogenized that there was no dataset-specific isoform. This would then imply that the isoform differences between other datasets were due to technical biases in the bioinformatics analysis. However, we do not believe this to be the case because even though the TCGA dataset was also re-analyzed through our pipeline, the comparison between TCGA and ONCOBOX ANTE yielded 139 pan-cancer transcript isoforms, the largest among all comparisons. Another good validation of the impact of the bioinformatics pipeline standardization would be to also re-analyze PCAWG RNA-seq data to test whether more pan-cancer candidate isoforms could be detected in common with the results from the TCGA dataset. Altogether, our results raised an important caution for the re-analysis of transcriptomics data from multiple sources.

The enrichment of immunoglobulin kappa chain (IGK) and immunoglobulin heavy chain (IGH) gene families in TCGA and PCAWG breast and lung cancers (Supplementary Table S3, S4, and S5) may simply be due to the presence of infiltrating antibody-producing plasma B cells in the tumor microenvironment[59], since bulk tissue RNA sequencing data were utilized here. However, the expression of these IGK and IGH transcripts may also be indicative of the progression of malignancy itself[60,61]. Furthermore, several research revealed the existence of 'Ig-like proteins' at both transcript and protein levels in several epithelial cancer cell lines[62–64]. These Ig-like proteins, now also considered cancer-Ig, appear to play roles in promoting growth, invasion, and migration in various cancers, including lung cancer. On the other hand, the three immunoglobulin heavy chain gene's isoforms (IGHJ1, IGHJ2, and IGHJ4) were not consistently expressed in ONCOBOX lung cancer samples. This is not due to a low average gene expression or sequencing depth of the dataset but rather a lack of expression for these specific isoforms, which may indicate IgH-specific transcript degradation.

The enrichment of ribosomal proteins, cellular signaling and structure, transporter proteins, protein processing and degradation, RNA-binding proteins, and heat shock proteins in TCGA and PCAWG breast and lung cancer samples may also be functionally relevant. For example, RPS6, RPL7A, and RPL14 are essential components of the ribosomes whose dysregulation can contribute to cancer by altering the rates of protein synthesis and impacting cellular growth and proliferation[65]. Changes in the expression levels of ribosomal protein transcripts are common in various cancer types and have been reported as prognostic predictors. A knockdown of PSMA2 and PSMB5, which are components of the proteasome, have been implicated in lung cancer responses[66], tumor proliferation, and poor prognosis in hepatocellular carcinoma[67]. WDR54, which is involved in signal transduction, RNA processing, vesicular trafficking, and cellular division, was reportedly associated with the promotion of tumorigenesis and metastasis in bladder cancer[68], by increasing the stability of mediator MEMO1 and altering its interaction with IRS1. This subsequently impairs the chemosensitivity of bladder cancer to cisplatin treatment. DAZAP2 was identified as a specifier of the p53 response to DNA damage[69], whose knockdown and deletion strongly potentiates cancer cell chemosensitivity in both cell culture and mouse xenograft models. However, despite these genes' clear roles in cancer, the isoform-specificity of their functions are still undetermined.

Conclusion

Under the limitation of an in silico study, our work strengthened the evidence that pan-cancer RNA transcripts can produce actionable neoantigens in two aspects: by detecting the presence and specificity of neoantigen-containing peptide precursors in proteomics datasets of cancer and normal tissues, and by simulating their binding dynamics with HLA protein. Our analysis focused on annotated transcript isoforms to survey the landscape of neoantigens derived from large-scale alterations such as different exon usages. However, an in-depth investigation of isoform switching at the splice junction level can reveal additional switching events involving unannotated isoforms as well as neoantigens resulting from these loci[70]. It is highly encouraging that as many as 24% (155 out of 651) of putative 9-mer neoantigens could be confirmed at protein level specifically in cancer tissues, and that nine of them were detectable in more than half (10–14 out of 20) of the cancer proteomes. A small-scale molecular dynamics simulation also supported the predicted high binding affinity between several putative neoantigens with HLA proteins by revealing the relative increase in hydrogen bonding, reduced structural variability, and lower binding free energy of the antigen-HLA complex. Compared to machine learning-based predictions of HLA binding affinity and stability, molecular dynamics simulation offers crucial explanations to support the predicted binding characteristics.

From a clinical standpoint, since the candidate transcript isoforms are expected to be present in at least 90% of tumors across all cancer types and some derived neoantigens were detectable in as many as 50–70% of the cancer proteomics dataset, we anticipated that a significant patient population could benefit from our approach.

The pool of potential pan-cancer neoantigens would also be significantly expanded once 8- to 11-mers are included. Importantly, the fact that some candidate transcripts are functionally relevant in the context of cancer also suggest that they will be persistent, specific, and therefore targetable in cancer cells. Nonetheless, there are still major downstream steps of experimentally confirming the HLA binding affinity and the immunogenicity of these candidate neoantigens.

Data availability

All transcriptomic data analyzed in this study were obtained from public databases as described in the Methods. Specifically, TCGA data were randomly subsampled in BAM alignment file format, including 50 samples each from TCGA-BRCA, TCGA-COAD, TCGA-DLBC, TCGA-LUAD, TCGA-MESO, TCGA-READ, and TCGA-SKCM, and 100 samples each from TCGA-KIRC, TCGA-KIRP, and TCGA-LUSC. Access to TCGA data was controlled through dbGaP, with details at https://portal.gdc.cancer.gov/. Pre-processed transcript expressio n data from the PCAWG project was obtained, open-access, via the International Cancer Genome Consortium (ICGC) Data Portal at https://dcc.icgc.org/. Although this portal was retired in June 2024, the provided URL instructs researchers to the updated access protocol. Pre-processed transcript expression data (read counts and TPM) from the Genotype-Tissue Expression (GTEx) project was obtained, open-access, via the GTEx Portal at https://gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression. For this project, we utilized the GTEX 2017-06-05_v8_RSEMv1.3.0 files. Raw RNA-seq sequencing reads from the ONCOBOX Atlas of Normal Tissue project and ONCOBOX Cancer resources were obtained as FASTQ files from the NCBI BioProjects PRJNA494560 (normal tissues) and PRJNA565016 and PRJNA578290 (cancer tissues), respectively. The lists of identified pan-cancer isoforms and candidate neoantigens are provided as Supplementary Materials. Transcript expression data in TPM units are provided on FigShare at 10.6084/m9.figshare.28324271 (TCGA), 10.6084/m9.figshare.28324229 (ONCOBOX ANTE), 10.6084/m9.figshare.28324232 (ONCOBOX Lung Cancer) and 10.6084/m9.figshare.28324235 (ONCOBOX Breast Cancer).

Received: 15 October 2024; Accepted: 30 April 2025
Published online: 07 May 2025

References

1. Hanahan, D. Hallmarks of cancer: new dimensions. *Cancer Discov.* **12**, 31–46. https://doi.org/10.1158/2159-8290.Cd-21-1059 (2022).
2. Lu, Y. C. & Wang, X. J. Harnessing the power of the immune system in cancer immunotherapy and cancer prevention. *Mol. Carcinog.* **59**, 675–678. https://doi.org/10.1002/mc.23211 (2020).
3. Sanchez-Danes, A. & Blanpain, C. Deciphering the cells of origin of squamous cell carcinomas. *Nat. Rev. Cancer.* **18**, 549–561. https://doi.org/10.1038/s41568-018-0024-5 (2018).
4. Sia, D., Villanueva, A., Friedman, S. L. & Llovet, J. M. Liver Cancer cell of origin, molecular class, and effects on?Patient prognosis. *Gastroenterology* **152**, 745–761. https://doi.org/10.1053/j.gastro.2016.11.048 (2017).
5. Shemesh, C. S. et al. Personalized Cancer vaccines: clinical landscape, challenges, and opportunities. *Mol. Ther.* **29** (2), 555–570. https://doi.org/10.1016/j.ymthe.2020.09.038 (2021).
6. Zamora, A. E., Crawford, J. C. & Thomas, P. G. Hitting the target: how T cells detect and eliminate tumors. *J. Immunol.* **200**, 392–399. https://doi.org/10.4049/jimmunol.1701413 (2018).
7. Sahin, U. & Tureci, O. Personalized vaccines for cancer immunotherapy. *Science* **359** (6382), 1355–1360 (2018).
8. Zhao, W., Wu, J., Chen, S. & Zhou, Z. Shared neoantigens: ideal targets for off-the-shelf cancer immunotherapy. *Pharmacogenomics* **21** (9), 637–645. https://doi.org/10.2217/pgs-2019-0184 (2020).
9. Klebanoff, C. A. & Wolchok, J. D. Shared cancer neoantigens: making private matters public. *J. Exp. Med.* **215** (1), 5–7. https://doi.org/10.1084/jem.20172188 (2018).
10. Naik, A., Lattab, B., Qasem, H. & Decock, J. Cancer testis antigens: emerging therapeutic targets leveraging genomic instability in cancer. *Mol. Ther. Oncol.* **32** (1), 200768. https://doi.org/10.1016/j.omton.2024.200768 (2024).
11. Xie, N. et al. Neoantigens: promising targets for cancer therapy. *Signal. Transduct. Target. Ther.* **8** (1), 9. https://doi.org/10.1038/s41392-022-01270-x (2023).
12. Dravis, C. et al. Epigenetic and transcriptomic profiling of mammary gland development and tumor models disclose regulators of cell state plasticity. *Cancer Cell.* **34** (3), 466–482e466. https://doi.org/10.1016/j.ccell.2018.08.001 (2018).
13. Malta, T. M. et al. Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* **173** (2), 338–354e315. https://doi.org/10.1016/j.cell.2018.03.034 (2018).
14. Yu, X. X. et al. Defining multistep cell fate decision pathways during pancreatic development at single-cell resolution. *EMBO J.* **38** (8). https://doi.org/10.15252/embj.2018100164 (2019).
15. Rathe, S. K. et al. Identification of candidate neoantigens produced by fusion transcripts in human osteosarcomas. *Sci. Rep.* **9** (1), 358. https://doi.org/10.1038/s41598-018-36840-z (2019).
16. Wang, G. et al. Y. An engineered oncolytic virus expressing PD-L1 inhibitors activates tumor neoantigen-specific T cell responses. *Nat. Commun.* **11** (1), 1395. https://doi.org/10.1038/s41467-020-15229-5 (2020).
17. Zhang, Y., Qian, J., Gu, C. & Yang, Y. Alternative splicing and cancer: a systematic review. *Signal. Transduct. Target. Ther.* **6** (1), 78. https://doi.org/10.1038/s41392-021-00486-7 (2021).
18. Okumura, N., Yoshida, H., Kitagishi, Y., Nishimura, Y. & Matsuda, S. Alternative splicings on p53, BRCA1 and PTEN genes involved in breast cancer. *Biochem. Biophys. Res. Commun.* **413**, 395–399. https://doi.org/10.1016/j.bbrc.2011.08.098 (2011).
19. Aran, V. et al. A cross-sectional study examining the expression of splice variants K-RAS4A and K-RAS4B in advanced non-small-cell lung cancer patients. *Lung Cancer.* **116**, 7–14. https://doi.org/10.1016/j.lungcan.2017.12.005 (2018).
20. Nussinov, R., Tsai, C. J. & Jang, H. Independent and core pathways in oncogenic KRAS signaling. *Expert Rev. Proteom.* **13**, 711–716. https://doi.org/10.1080/14789450.2016.1209417 (2016).
21. Pio, R. & Montuenga, L. M. Alternative splicing in lung Cancer. *J. Thorac. Oncol.* **4**, 674–678. https://doi.org/10.1097/JTO.0b013e3181a520dc (2009).
22. Brown, R. L. et al. CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. *J. Clin. Invest.* **121** (3), 1064–1074. https://doi.org/10.1172/JCI44540 (2011).
23. Chen, C., Zhao, S., Karnad, A. & Freeman, J. W. The biology and role of CD44 in cancer progression: therapeutic implications. *J. Hematol. Oncol.* **11**, 64–64. https://doi.org/10.1186/s13045-018-0605-5 (2018).
24. Kahles, A. et al. Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell.* **34** (2), 211–224e216. https://doi.org/10.1016/j.ccell.2018.07.001 (2018).

25. Kahraman, A., Karakulak, T., Szklarczyk, D. & von Mering, C. Pathogenic impact of transcript isoform switching in 1,209 cancer samples covering 27 cancer types using an isoform-specific interaction network. *Sci. Rep.* **10**, 14453–14453. https://doi.org/10.1038/s41598-020-71221-5 (2020).
26. Climente-Gonzalez, H., Porta-Pardo, E., Godzik, A. & Eyras, E. The functional impact of alternative splicing in Cancer. *Cell. Rep.* **20**, 2215–2226. https://doi.org/10.1016/j.celrep.2017.08.012 (2017).
27. Vitting-Seerup, K. & Sandelin, A. The landscape of isoform switches in human cancers. *Mol. Cancer Res.* **15**, 1206–1220. https://doi.org/10.1158/1541-7786.Mcr-16-0459 (2017).
28. Weinstein, J. N. et al. A. R. The Cancer genome atlas Pan-Cancer analysis project. *Nat. Genet.* **45** (10), 1113–1120. https://doi.org/10.1038/ng.2764 (2013).
29. Suntsova, M. et al. Atlas of RNA sequencing profiles for normal human tissues. *Sci. Data*. **6** (1), 36. https://doi.org/10.1038/s41597-019-0043-4 (2019).
30. Sorokin, M. et al. RNA sequencing in comparison to immunohistochemistry for measuring cancer biomarkers in breast cancer and lung cancer specimens. *Biomedicines* **8** (5). https://doi.org/10.3390/biomedicines8050114 (2020).
31. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* **41** (Database issue), D991–995. https://doi.org/10.1093/nar/gks1193 (2013).
32. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic Acids Res.* **39**, D19–21. https://doi.org/10.1093/nar/gkq1019 (2011).
33. Consortium, I. T. P. C. A. o. W. G. Pan-cancer analysis of whole genomes. *Nature* **578** (7793), 82–93. https://doi.org/10.1038/s41586-020-1969-6 (2020).
34. Consortium, G. The Genotype-Tissue expression (GTEx) project. *Nat. Genet.* **45** (6), 580–585. https://doi.org/10.1038/ng.2653 (2013).
35. Hudson, T. J. et al. International network of cancer genome projects. *Nature* **464** (7291), 993–998. https://doi.org/10.1038/nature08987 (2010).
36. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915. https://doi.org/10.1038/s41587-019-0201-4 (2019).
37. Li, H. et al. P. D. P. The sequence alignment/map format and samtools. *Bioinformatics* **25** (16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 (2009).
38. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33** (3), 290–295. https://doi.org/10.1038/nbt.3122 (2015).
39. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–w205. https://doi.org/10.1093/nar/gkz401 (2019).
40. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif Deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48** (W1), W449–W454. https://doi.org/10.1093/nar/gkaa379 (2020).
41. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421. https://doi.org/10.1186/1471-2105-10-421 (2009).
42. Wen, B., Wang, X. & Zhang, B. PepQuery enables fast, accurate, and convenient proteomic validation of novel genomic alterations. *Genome Res.* **29** (3), 485–493. https://doi.org/10.1101/gr.235028.118 (2019).
43. Marzella, D. F. et al. PANDORA: A fast, Anchor-Restrained modelling protocol for peptide: MHC complexes. *Front. Immunol.* **13**, 878762. https://doi.org/10.3389/fimmu.2022.878762 (2022).
44. Webb, B., Sali, A., Webb, B. & Sali, A. Protein structure modeling and modeller methods mol biol. *Methods Mol. Biol.* **1654**, 39–54. https://doi.org/10.1007/978-1-4939-7231-9_4 (2017).
45. Abraham, M. J. et al. High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX 1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001 (2015).
46. Lindorff-Larsen, K. et al. Improved side-chain torsion potentials for the amber ff99SB protein force field. *Proteins* **78** (8), 1950–1958. https://doi.org/10.1002/prot.22711 (2010).
47. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18** 1463–1472. https://doi.org/10.1002/(SICI)1096-987X(199709)18:123.0.CO;2-H (1997).
48. Valdes-Tresanco, M. S., Valdes-Tresanco, M. E., Valiente, P. A., Moreno, E. & gmx_MMPBSA A new tool to perform End-State free energy calculations with GROMACS. *J. Chem. Theory Comput.* **17**, 6281–6291. https://doi.org/10.1021/acs.jctc.1c00645 (2021).
49. Miller, B. R. et al. MMPBSA.py: an efficient program for End-State free energy calculations. *J. Chem. Theory Comput.* **8**, 3314–3321. https://doi.org/10.1021/ct300418h (2012).
50. Wang, E. et al. End-Point binding free energy calculation with MM/PBSA and MM/GBSA: strategies and applications in drug design. *Chem. Rev.* **119** (16), 9478–9508. https://doi.org/10.1021/acs.chemrev.9b00055 (2019).
51. Xia, M. et al. Next-Generation sequencing revealed a distinct Immunoglobulin repertoire with specific mutation hotspots in acute myeloid leukemia. *Biology (Basel)* **11** (2). https://doi.org/10.3390/biology11020161 (2022).
52. Zhang, J. et al. Lnc-LRRTM4 promotes proliferation, metastasis and EMT of colorectal cancer through activating LRRTM4 transcription. *Cancer Cell. Int.* **23** (1), 142. https://doi.org/10.1186/s12935-023-02986-8 (2023).
53. Gu, Z. et al. HNRNPC, a predictor of prognosis and immunotherapy response based on bioinformatics analysis, is related to proliferation and invasion of NSCLC cells. *Respir Res.* **23** (1), 362. https://doi.org/10.1186/s12931-022-02227-y (2022).
54. Martino, F. et al. The mechanical regulation of RNA binding protein HnRNPC in the failing heart. *Sci. Transl Med.* **14** (672), eabo5715. https://doi.org/10.1126/scitranslmed.abo5715 (2022).
55. Mo, L. et al. An analysis of the role of HnRNP C dysregulation in cancers. *Biomark. Res.* **10** (1), 19. https://doi.org/10.1186/s40364-022-00366-4 (2022).
56. Hu, J., Cai, D., Zhao, Z., Zhong, G. C. & Gong, J. Suppression of heterogeneous nuclear ribonucleoprotein C inhibit hepatocellular carcinoma proliferation, migration, and invasion via Ras/MAPK signaling pathway. *Front. Oncol.* **11**, 659676–659676. https://doi.org/10.3389/fonc.2021.659676 (2021).
57. Zhang, X. et al. SELENBP1 inhibits progression of colorectal cancer by suppressing epithelial-mesenchymal transition. *Open. Med. (Wars).* **17** (1), 1390–1404. https://doi.org/10.1515/med-2022-0532 (2022).
58. Li, S. et al. S100A8 promotes epithelial-mesenchymal transition and metastasis under TGF-β/USF2 axis in colorectal cancer. *Cancer Commun. (Lond).* **41** (2), 154–170. https://doi.org/10.1002/cac2.12130 (2021).
59. Sharonov, G. V., Serebrovskaya, E. O., Yuzhakova, D. V., Britanova, O. V. & Chudakov, D. M. B cells, plasma cells and antibody repertoires in the tumour microenvironment. *Nat. Rev. Immunol.* **20**, 294–307. https://doi.org/10.1038/s41577-019-0257-x (2020).
60. Cui, M. et al. Immunoglobulin expression in Cancer cells and its critical roles in tumorigenesis. *Front. Immunol.* **12**, 613530. https://doi.org/10.3389/fimmu.2021.613530 (2021).
61. Gercel-Taylor, C., Bazzett, L. B. & Taylor, D. D. Presence of aberrant tumor-reactive Immunoglobulins in the circulation of patients with ovarian cancer. *Gynecol. Oncol.* **81**, 71–76. https://doi.org/10.1006/gyno.2000.6102 (2001).
62. Chen, Z. & Gu, J. Immunoglobulin G expression in carcinomas and cancer cell lines. *Faseb J.* **21**, 2931–2938. https://doi.org/10.1096/fj.07-8073com (2007).
63. Qiu, X. et al. Human epithelial cancers secrete Immunoglobulin g with unidentified specificity to promote growth and survival of tumor cells. *Cancer Res.* **63** (19), 6488–6495 (2003).

64. Zhu, X. et al. Immunoglobulin mRNA and protein expression in human oral epithelial tumor cells. *Appl. Immunohistochem. Mol. Morphol.* **16** (3), 232–238. https://doi.org/10.1097/PAI.0b013e31814c915a (2008).
65. Goudarzi, K. M. & Lindstrom, M. S. Role of ribosomal protein mutations in tumor development (Review). *Int. J. Oncol.* **48**, 1313–1324. https://doi.org/10.3892/ijo.2016.3387 (2016).
66. Rashid, M. U., Lorzadeh, S., Gao, A., Ghavami, S. & Coombs, K. M. PSMA2 knockdown impacts expression of proteins involved in immune and cellular stress responses in human lung cells. *Biochim. Biophys. Acta Mol. Basis Dis.* **1869**, 166617–166617. https://doi.org/10.1016/j.bbadis.2022.166617 (2023).
67. Liu, J., Mi, J., Liu, S., Chen, H. & Jiang, L. PSMB5 overexpression is correlated with tumor proliferation and poor prognosis in hepatocellular carcinoma. *FEBS Open. Bio.* **12**, 2025–2041. https://doi.org/10.1002/2211-5463.13479 (2022).
68. Wei, X. et al. WD repeat protein 54-mediator of ErbB2-driven cell motility 1 axis promotes bladder cancer tumorigenesis and metastasis and impairs chemosensitivity. *Cancer Lett.* **556**, 216058. https://doi.org/10.1016/j.canlet.2023.216058 (2023).
69. Liebl, M. C. et al. DAZAP2 acts as specifier of the p53 response to DNA damage. *Nucleic Acids Res.* **49** (5), 2759–2776. https://doi.org/10.1093/nar/gkab084 (2021).
70. Wickland, D. P. et al. Comprehensive profiling of cancer neoantigens from aberrant RNA splicing?. *J. Immunother. Cancer* https://doi.org/10.1136/jitc-2024-008988 (2024).

Acknowledgements

This work was supported by Thailand Science, Research, and Innovation Fund, Chulalongkorn University [HEAF673000102 to J.T., T.P., S.S] and the Chulalongkorn University Second Century Fund [C2F, to A.A.M and S.S]. The results shown in this manuscript are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Author contributions

J.T. collected and processed the data. J.T and A.A.M analyzed the data. J.T., A.A.M, T.P., and S.S. interpreted the results. T.P. and S.S. conceived the project and supervised the research. J.T., A.A.M, and S.S. wrote the first draft. J.T., A.A.M., T.P., and S.S. revised the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

This study involved the analysis of publicly available transcriptomic and proteomic datasets, as well as controlled access data obtained from the Genomic Data Commons (GDC) through the authorized access of T.P. No new human participants, human data, or human tissue were directly involved in this research.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-00817-6.

Correspondence and requests for materials should be addressed to S.S.