World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

# In silico identification of neoantigens using integrated DNA-Seq, RNA-Seq and HLA typing for personalized cancer vaccine design

Akshay Raj C S * and Elamathi Natarajan

*Department of Bioinformatics, Biotecnika Info Labs Ltd, Bangalore, India.*

World Journal of Advanced Research and Reviews, 2026, 30(01), 1142-1153

Publication history: Received on 28 February 2026; revised on 06 April 2026; accepted on 08 April 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.30.1.0887

## Abstract

**Objective:** This study aimed to develop an integrated computational pipeline for the identification and prioritization of tumor-specific neoantigens using combined whole-exome sequencing (WES), RNA sequencing (RNA-seq), and HLA typing data for personalized cancer vaccine design.

**Methodology:** Publicly available sequencing data from the triple-negative breast cancer cell line HCC1395 and matched normal HCC1395BL were analyzed. Data preprocessing included quality control and trimming, followed by alignment to the GRCh38 reference genome. Somatic variants were identified and annotated, and protein-altering mutations were used to generate mutant peptides. HLA class I alleles were predicted, and peptide–HLA binding affinity was assessed. Gene expression and mutation clonality were evaluated, and candidate neoantigens were filtered based on binding affinity, expression level, clonality, and immunogenicity. Selected neoantigens were further used for multi-epitope vaccine design and population coverage analysis.

**Results:** A total of 561 somatic variants were identified, including 386 protein-altering mutations generating 13,452 mutant peptides. Binding predictions yielded 63,100 peptide–HLA combinations, which were refined to 55 high-confidence neoantigens. Seven top candidates demonstrated strong binding affinity, expression, and immunogenic potential. The predicted population coverage was 36.7% globally and 33.73% in the Indian population. The designed multi-epitope vaccine construct showed an antigenicity score of 0.4802 and was predicted to be non-allergenic.

**Conclusion:** The proposed pipeline effectively identifies and prioritizes biologically relevant neoantigens and supports the development of personalized cancer vaccines using integrated multi-omics and immunoinformatics approaches.

**Keywords** Neoantigen; Cancer vaccine; Immunoinformatics; HLA typing; RNA-seq; Personalized immunotherapy

## 1. Introduction

Cancer remains one of the leading causes of mortality worldwide, with an estimated 19.3 million new cases and nearly 10 million deaths reported in 2020 according to global cancer statistics [1]. Conventional treatment modalities such as chemotherapy and radiotherapy have improved patient survival; however, their effectiveness is frequently limited by tumor heterogeneity and the development of drug resistance [2,3]. In recent years, cancer immunotherapy has emerged as a transformative strategy that harnesses the host immune system to recognize and eliminate malignant cells, offering improved specificity and long-term therapeutic benefit [4,5]. Among the different immunotherapeutic approaches, personalized neoantigen-based vaccines have gained considerable attention due to their ability to induce highly specific anti-tumor immune responses with minimal off-target [6,7].

* Corresponding author: Akshay Raj C S

Neoantigens are tumor-specific peptides generated from somatic mutations that occur during cancer development. Unlike tumor-associated antigens, which may also be expressed in normal tissues, neoantigens are exclusively present in tumor cells, thereby reducing the risk of autoimmune toxicity and improving therapeutic precision [7,8]. These antigens can arise from multiple types of genomic alterations, including single nucleotide variants (SNVs), insertions and deletions (indels) that may cause frameshift mutations, gene fusions, and alternative splicing events [9]. Tumors with a high somatic mutational burden, such as melanoma, tend to generate more neoantigens and are associated with improved responses to immune checkpoint inhibitor therapy [10]. However, the personalized nature of neoantigen-based immunotherapy necessitates the development of patient-specific strategies for accurate neoantigen identification and vaccine design [11].

The identification of clinically relevant neoantigens is technically challenging because predicting true immunogenic peptides from large-scale multi-omics data is complex, and only a small fraction of tumor mutations results in peptides capable of binding MHC molecules and eliciting effective T-cell responses [12]. The analysis of tumor genomic and immunologic data generates large amounts of information; therefore, computational and informatics-based prediction pipelines have become essential tools in modern cancer immunotherapy research [13]. Current in silico workflows integrate genomic sequencing data, transcriptomic expression profiles, and immunological prediction algorithms to prioritize potential neoantigen candidates [14]. These pipelines typically include somatic variant calling, functional annotation, human leukocyte antigen (HLA) typing, and prediction of peptide–major histocompatibility complex (MHC) binding affinity using machine-learning-based tools such as MHCflurry and related algorithms [8,14-16]Incorporating RNA expression levels and variant allele frequency further improves the identification of clonal mutations that are actively transcribed, increasing the likelihood of selecting clinically relevant targets [14].

Accurate HLA typing is a critical requirement for reliable neoantigen prediction because peptide presentation is strictly dependent on the patient-specific HLA genotype [17]. High-resolution HLA typing tools, such as OptiType, enable precise identification of HLA class I alleles directly from next-generation sequencing data, allowing personalized prediction of peptide–MHC interactions [15,18]. In addition, population coverage analysis is necessary to evaluate whether selected epitopes can be presented by commonly occurring HLA alleles, thereby estimating the potential applicability of vaccine candidates across diverse populations [19].

Recent advances in immunoinformatics have facilitated the design of multi-epitope vaccines that combine multiple high-confidence neoantigens into a single construct to enhance immune activation [20]. Such multi-epitope approaches can induce broader and more robust immune responses compared to single-antigen vaccines [6]. Furthermore, in silico evaluation of antigenicity, allergenicity, and structural stability enables rapid screening of vaccine candidates, reducing both development time and cost [20,21]. Despite progress in machine learning-based neoantigen prediction, accurately prioritizing candidate epitopes remains challenging due to factors such as antigen processing, immunogenicity, and patient-specific HLA genotype. Integrating these parameters into a unified computational framework is essential to improve the reliability and clinical applicability of neoantigen-based vaccines [12,22].

Therefore, the present study aimed to develop an integrated computational pipeline for the identification and prioritization of tumor-specific neoantigens using combined DNA-Seq, RNA-Seq, and HLA typing data. Publicly available sequencing datasets from a triple-negative breast cancer model were analyzed to detect somatic mutations, predict peptide–HLA binding, evaluate gene expression and clonality, and prioritize immunogenic candidates. The selected neoantigens were further used for multi-epitope vaccine design and population coverage analysis. This study demonstrates a scalable in silico framework for personalized cancer vaccine development using multi-omics data and immunoinformatics tools.

## 2. Material and methods

This study employed a comprehensive computational pipeline to identify tumor-specific neoantigens from matched tumor-normal sequencing datasets. The methodology integrates data retrieval, preprocessing, alignment, variant calling, annotation, neoantigen prediction, immunogenicity assessment and multi-epitope vaccine design. All analyses were performed using the human reference genome GRCh38 (UCSC hg38 build).

### 2.1. Computational Tools and Software

A suite of computational tools was used for sequence processing, variant detection, gene expression quantification, HLA typing, peptide–MHC binding prediction, and immunogenicity analysis. The tools, versions and primary functions are summarized in Table 1. These tools were managed via Conda to ensure reproducibility.

**Table 1** Computational tools and software used in the neoantigen pipeline

| Step | Tool/Software | Version | Function |
|---|---|---|---|
| Data Retrieval | SRA Toolkit | v3.3.0 | Download and convert sequencing data from NCBI SRA |
| Quality Assessment | FastQC | v0.11.9 | Evaluate sequencing read quality |
| Adapter Trimming | Trimmomatic | v0.39 | Remove adapters and low-quality bases |
| RNA-seq Alignment | HISAT2 | v2.2.1 | Splice-aware alignment to GRCh38 |
| WES Alignment | BWA-MEM | v0.7.17-r1188 | Whole-exome sequencing alignment |
| BAM Processing | SAMtools | v1.13 | Sorting, indexing, format conversion |
| Duplicate Marking | GATK4 MarkDuplicates | v4.6.2.0 | Identify PCR duplicates |
| Variant Calling | GATK4 Mutect2 | v4.6.2.0 | Somatic SNV and indel detection |
| Variant Annotation | VEP | v115.2 | Functional annotation of protein-altering variants |
| Gene Expression | featureCounts | v2.0.3 | Exon-level read counting |
| Differential Expression | DESeq2 | v1.38.0 | Normalization and log2 fold-change calculation |
| HLA Typing | OptiType | v1.3.5 | Patient-specific HLA class I prediction |
| Peptide Generation | Biopython | v1.86 | Mutant and wild-type peptide extraction |
| MHC Binding | MHCflurry | v2.1.5 | Peptide–HLA binding affinity prediction |
| Immunogenicity | IEDB Tools | – | Prediction of immunogenicity and population coverage |
| Antigenicity | VaxiJen | v2.0 | Predict peptide antigenicity |
| Allergenicity | AllergenFP | v1.0 | Predict potential allergenicity |
| Toxicity | ToxinPred | – | Predict potential toxicity |

Version numbers and default parameters were used unless otherwise specified.

### *2.2.* Data Acquisition

Sequencing datasets for WES and RNA-seq were obtained from the NCBI SRA for the HCC1395 tumor cell line and matched HCC1395BL normal B-lymphoblastoid cell line. Matched tumor-normal pairs allowed accurate identification of somatic variants. A summary of sequencing metadata is provided in Table 2.

**Table 2** Sequencing sample metadata

| Sample | WES Tumor | WES Normal | RNA-seq Tumor | RNA-seq Normal |
|---|---|---|---|---|
| SRR Run | SRR7890844 | SRR7890845 | SRR9134727 | SRR9134696 |
| Library | WES_NC_T_1 | WES_NC_N_1 | bulk_RNA_HCC1395_seq_1_rep_1 | bulk_RNA_HCC1395BL_seq_2_rep_1 |
| Platform | Illumina HiSeq 2500 | Illumina HiSeq 2500 | Illumina NextSeq 550 | Illumina HiSeq 4000 |
| Strategy | WES | WES | RNA-seq | RNA-seq |
| Source | Genomic | Genomic | Transcriptomic | Transcriptomic |
| Layout | Paired | Paired | Paired | Paired |
| Spots | 54.5 M | 57.1 M | 30.2 M | 33.6 M |

| | | | | |
|---|---|---|---|---|
| Bases | 13.7 G | 14.4 G | 4.6 G | 6.7 G |
| Size | 4.5 GB | 4.7 GB | 1.7 GB | 2.3 GB |
| GC Content | 47.1% | 46.6% | 48.4% | 48.1% |

M = million; G = gigabases; TPM values calculated post-alignment.

The .sra files were converted to paired-end FASTQ format using fasterq-dump with --split-files and multi-threading. Previously trimmed high-quality reads were skipped to improve computational efficiency.

### 2.3. Read Preprocessing and Quality Control

Initial quality assessment was performed using FastQC v0.11.9. Adapter sequences and low-quality bases were removed using Trimmomatic v0.39 in paired-end mode with parameters: ILLUMINACLIP: TruSeq3-PE. fa: 2:30:10, LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, and MINLEN:36. post-trimming quality was performed using FastQC to confirm improvement.

### 2.4. Alignment to the Reference Genome

RNA-seq reads were aligned to GRCh38 using HISAT2 with the --dta option to support transcript-aware spliced alignment. WES reads were aligned using BWA-MEM v0.7.17-r1188. SAM files were converted to coordinate-sorted BAM files and indexed using SAMtools v1.13. Duplicates were marked using GATK MarkDuplicates (--CREATE_INDEX true), creating high-quality BAM files for variant calling and gene expression analysis.

### 2.5. Somatic Variant Calling

Tumor-normal pairs were analyzed using GATK Mutect2 v4.6.2.0. Germline variants were filtered using gnomAD reference datasets and a panel of normals. Orientation bias and contamination were corrected, and only PASS variants were retained.

### 2.6. Variant Annotation

PASS variants were annotated using VEP v115.2 with the --pick flag to retain the most biologically relevant transcript consequence for each variant. Functional annotations included gene symbols, protein consequences, SIFT and PolyPhen scores, and population allele frequencies. Only protein-altering variants were retained for neoantigen prediction.

### 2.7. HLA Typing and Neoantigen Peptide Generation

Patient-specific HLA class I alleles were determined using OptiType v1.3.5. Mutant and corresponding wild-type peptides (8–11 amino acids) were generated using Biopython v1.86, covering alternative transcript isoforms.

### 2.8. MHC Binding Prediction

Peptide–HLA binding affinities were predicted using MHCflurry v2.1.5. IC50 values (nM) were calculated for mutant and wild-type peptides. Peptides with stronger mutant binding were prioritized.

### 2.9. Gene Expression Quantification

Gene-level read counts were generated with featureCounts v2.0.3. TPM normalization and log2 fold-change calculation were performed using DESeq2 v1.38.0.

### 2.10. Mutation Clonality

Variant allele frequencies were extracted from VEP-annotated VCFs or calculated from allele depth:

$$\mathrm{VAF} = \frac{\mathrm{AD}_{alt}}{\mathrm{AD}_{ref} + \mathrm{AD}_{alt}}$$

Where:
$\mathbf{AD}_{alt}$ = alternate allele read depth
$\mathbf{AD}_{ref}$ = reference allele read depth

Mutations with VAF ≥ 0.25 were classified as clonal, <0.25 as subclonal, and missing values as Unknown.

### 2.11. Cancer Gene Annotation

Genes were mapped to the OncoKB database and classified as driver, cancer-associated, or unknown.

### 2.12. IEDB Immunogenicity Scoring

Candidate peptides were evaluated using the IEDB immunogenicity prediction tool to estimate their potential to elicit a T-cell response. Peptides were assigned immunogenicity scores, which were later used during integration and filtering to prioritize the most promising neoantigens.

### 2.13. Integration and Filtering of Neoantigen Data

Peptide–HLA binding, clonality, cancer relevance, expression, and immunogenicity scores were integrated. Filtering retained peptides with stronger mutant binding, VAF ≥ 0.25, TPM > 1, log2FC > 0, and positive IEDB immunogenicity scores. Duplicate peptides and excessive per-gene representation were removed.

### 2.14. Population Coverage Analysis

The curated neoantigen set was analyzed using IEDB Population Coverage tools to estimate global and regional HLA allele representation.

### 2.15. Multi-Epitope Vaccine Construction and Evaluation

Selected peptides were joined with **AAY linkers** for $CD8^+$ T-cell epitope presentation. Vaccine constructs were evaluated for antigenicity (VaxiJen v2.0), allergenicity (AllergenFP v1.0), and toxicity (ToxinPred)

## 3. Results and discussion

### 3.1. Whole-Exome Sequencing Quality and Variant Detection

Raw sequencing data for the tumor sample (SRR7890844) and matched normal sample (SRR7890845) were retrieved from the NCBI SRA database. The tumor sample generated 109,097,098 reads, whereas the normal sample produced 114,116,530 reads. All reads were successfully converted into paired-end FASTQ format and subjected to quality assessment using FastQC. The quality reports indicated high per-base sequence quality across most read positions, balanced GC content, and minimal adapter contamination, confirming the reliability of the sequencing output and suitability for downstream analysis. High-quality raw data are essential for accurate somatic mutation detection because sequencing artifacts may lead to false variant calls that can affect neoantigen prediction [23].

After trimming low-quality bases and adapter sequences using Trimmomatic, the filtered reads were aligned to the GRCh38 human reference genome using the BWA-MEM algorithm. Alignment results showed mapping rates of 99.5% for the tumor sample and 99.3% for the normal sample, with duplicate rates ranging from 3.02% to 3.33%. These values indicate excellent sequencing depth and low technical bias, both of which are important for reliable detection of somatic mutations. High mapping efficiency also reflects good library preparation and minimal contamination, which are critical factors for precision oncology workflows [24].

Somatic variant calling was performed using MuTect2 with paired tumor–normal analysis to distinguish true somatic mutations from germline variants. A total of 561 high-confidence variants were identified, including 518 single-nucleotide variants (SNVs), 4 insertions, 32 deletions, and 7 substitutions. Functional annotation using the Variant Effect Predictor (VEP) revealed that 386 variants resulted in protein-altering consequences, including 341 missense mutations, 17 stop-gained mutations, 1 stop-lost mutation, 15 frameshift mutations, 2 in-frame insertions, 9 in-frame deletions, and 1 stop-retained mutation (Fig. 1). The predominance of missense and frameshift mutations is consistent with previous cancer genomics studies showing that non-synonymous SNVs and indels represent the major source of tumor-specific neoantigens and are therefore important targets for immunotherapy [25,26].

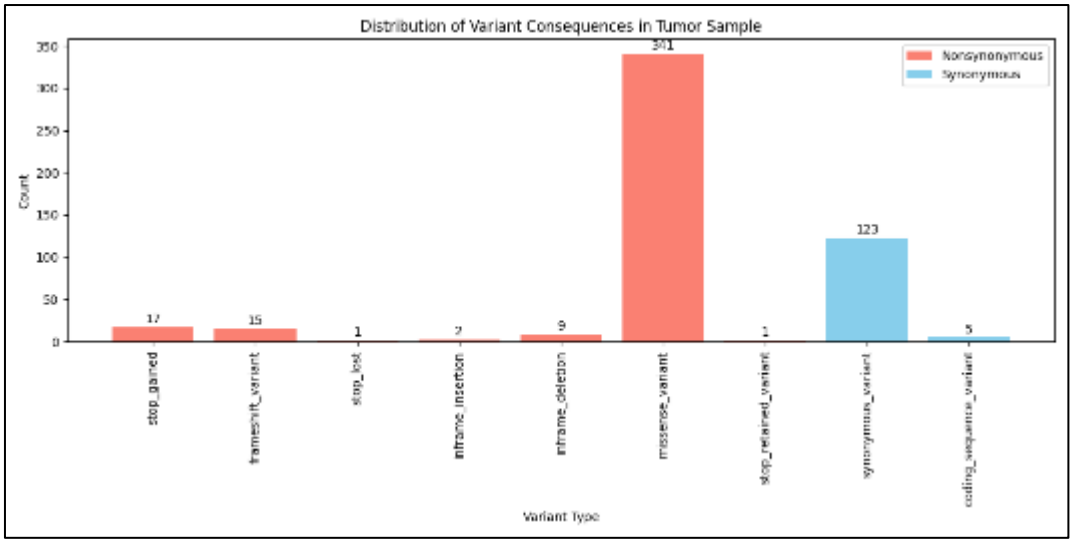


**Figure 1** Distribution of variant consequences in the tumor sample (SRR7890844). Nonsynonymous variants include missense, stop-gained, frameshift, in-frame insertions/deletions, and stop-retained variants. Synonymous variants include synonymous and coding sequence variants

To improve the reliability of neoantigen prediction, strict filtering criteria were applied based on read depth, allele frequency, and coding impact. Only variants capable of generating peptides of 8–11 amino acids at non-terminal positions were retained because these lengths are optimal for MHC class I binding. After filtering, 360 variants remained, generating 13,452 mutant peptide sequences along with corresponding wild-type peptides for comparative binding analysis. The use of paired tumor–normal sequencing together with stringent filtering greatly reduces false-positive calls and enriches for biologically relevant mutations [27,28]. Accurate identification of somatic mutations is a critical step in neoantigen discovery pipelines, as errors at this stage can propagate through all subsequent analyses.

### 3.2. RNA-seq Quality and Gene Expression Analysis

RNA-seq data for the tumor sample (SRR9134727) and matched normal sample (SRR9134696) were obtained from the NCBI SRA database. The tumor sample generated 60,401,134 reads, whereas the normal sample produced 67,106,258 reads. After adapter removal and quality trimming using Trimmomatic, high-quality reads were aligned to the GRCh38 reference genome using HISAT2. Alignment results showed mapping rates greater than 97% for both samples, with more than 94% of reads properly paired. These values indicate high sequencing quality and sufficient coverage for accurate transcriptome profiling. Previous studies have demonstrated that high mapping efficiency and proper pairing rates are necessary to minimize technical bias and ensure reliable quantification of gene expression [29].

Gene expression quantification was performed using featureCounts, which identified a total of 78,691 annotated gene features across the dataset. Among these, 24,130 genes were considered expressed in the tumor sample using a TPM threshold greater than 1. Differential expression analysis based on $\log_2$ fold-change identified 14,514 upregulated and 13,628 downregulated genes in the tumor sample using cutoffs of >1 and <−1, respectively. The presence of a large number of differentially expressed genes reflects the extensive transcriptional changes associated with tumor progression and supports the biological validity of the dataset.

Integration of RNA-seq expression data with whole-exome sequencing results is essential for accurate neoantigen prediction, as computational pipelines such as those developed by John Hundal et al. (2020) incorporate transcriptomic data to confirm the expression of mutated genes. Only mutations occurring in expressed genes can generate peptides that are processed and presented on MHC molecules, whereas variants located in transcriptionally silent regions may lead to false-positive predictions if expression data are not considered [30]. Therefore, combining genomic and transcriptomic information improves the biological relevance of predicted neoantigens and increases the likelihood of identifying clinically useful targets for personalized cancer immunotherapy [31].

### 3.3. HLA Typing and Neoantigen Prediction

*3.3.1. HLA Typing*

Patient-specific HLA class I alleles were predicted using OptiType, and the resulting HLA-A, HLA-B, and HLA-C alleles are listed in Table 3. Accurate HLA typing is a crucial step in neoantigen prediction because peptide presentation depends on allele-specific binding affinity. High-resolution HLA profiling allows precise identification of peptides that can be displayed on the tumor cell surface and recognized by cytotoxic T lymphocytes. Errors in HLA typing may lead to incorrect binding predictions and reduce the accuracy of neoantigen selection [32].

**Table 3** Patient-specific HLA class I alleles determined using OptiType.

| sample ID | A1 | A2 | B1 | B2 | C1 | C2 | Reads | Objective |
|---|---|---|---|---|---|---|---|---|
| SRR9134727 | A*29:02 | A*29:02 | B*08:01 | B*45:01 | C*06:02 | C*07:01 | 2289 | 2247.798 |

*3.3.2. Peptide–MHC Binding Prediction*

Filtered somatic variants generated 63,100 mutant peptide–HLA combinations along with 67,260 corresponding wild-type peptide–HLA pairs. Binding affinity prediction was performed using MHCflurry, and the predicted affinities of mutant and wild-type peptides were compared to identify candidates showing improved binding in the mutant form. This comparison is important because true neoantigens should bind more strongly to MHC molecules than their normal counterparts, increasing the probability of T-cell recognition [33,45].

*3.3.3. Clonality Analysis*

Clonality analysis was performed using variant allele frequency (VAF) to evaluate the distribution of predicted neoantigens across tumor cell populations. A total of 60.5% of neoantigens were classified as clonal (VAF >0.25), 35% as subclonal, and 4.5% as unknown (Fig. 2). Clonal mutations are present in the dominant tumor population and are therefore preferred targets for immunotherapy because they are less likely to be lost during tumor evolution [34]. In contrast, subclonal mutations reflect intratumor heterogeneity and may lead to incomplete immune targeting, reducing therapeutic effectiveness [35].

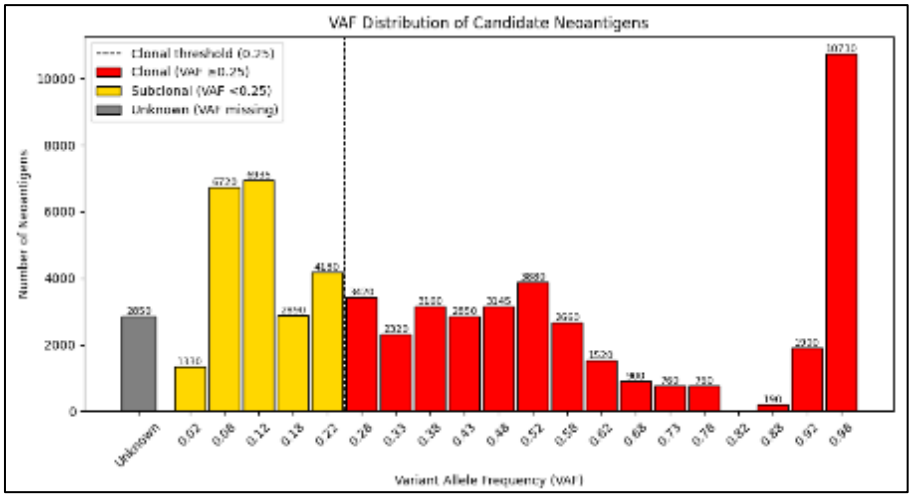


**Figure 2** Clonality distribution of candidate neoantigens based on variant allele frequency (VAF).

*3.3.4. Functional Annotation*

Functional annotation using the OncoKB database showed that 3.9% of neoantigens originated from driver genes, 0.9% from cancer-associated genes, and 95.2% from genes with unknown oncogenic relevance. Similar distributions have been reported in previous studies and may result from immunoediting, in which highly immunogenic mutations in essential genes are eliminated during tumor progression, leaving mutations in less critical genes [36,37].

*3.3.5. Multi-Parameter Filtering*

To obtain high-confidence neoantigens, multiple criteria including binding affinity, gene expression level, differential expression, clonality status, functional annotation, and immunogenicity score were integrated into a composite filtering strategy. Sequential application of these parameters reduced the initial 63,100 peptide candidates to 55 high-confidence neoantigens. Multi-parameter ranking has been shown to improve prediction accuracy compared with the use of binding affinity alone [38,11].

*3.3.6. Final Neoantigen Selection*

The highest-ranking neoantigens were derived from genes including TP53, DDX3X, PRDX5, TPM4, PCLO, and BCAR1, which are associated with tumor progression, stress response, and immune regulation. Composite scoring based on binding affinity, expression level, fold-change, and immunogenicity was used to select the top 10 neoantigens for vaccine design (Table 4). Targeting mutations in biologically relevant genes may improve therapeutic effectiveness because such mutations are less likely to be lost during tumor evolution [39,40].

**Table 4** Top 10 neoantigen candidates after multi-parameter filtering and ranking

| Peptide (mutant) | Gene | HLA | Mutant affinity (nM) | Tumor_TPM | log2FC | IEDB_score | VAF | Clonality | Cancer relevance | RANK_score | VaxiJen Score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AETRAQFA | TPM4 | HLAB4501 | 92.89709684 | 254.9355116 | 0.932526401 | 0.09254 | 0.435 | Clonal | Unknown | 1.724451 | 0.3409 |
| LLADPTGALGK | PRDX5 | HLAC0602 | 287.70014 | 284.0877074 | 0.626545264 | 0.14229 | 0.502 | Clonal | Unknown | 1.554115 | 0.6541 |
| RAQFAERTV | TPM4 | HLAC0602 | 333.039155 | 254.9355116 | 0.932526401 | 0.27891 | 0.435 | Clonal | Unknown | 1.539927 | 0.1298 |
| HEDVPICIY | PTPN12 | HLAA2902, HLAB4501 | 396.3708115 | 186.6181958 | 2.695527148 | 0.19548 | 0.28 | Clonal | Unknown | 1.534905 | 0.4866 |
| QEHEDVPICIY | PTPN12 | HLAB4501 | 461.2973886 | 186.6181958 | 2.695527148 | 0.31979 | 0.28 | Clonal | Unknown | 1.533415 | 0.4552 |
| LLADPTGAL | PRDX5 | HLAC0602 | 351.9429938 | 284.0877074 | 0.626545264 | 0.11222 | 0.502 | Clonal | Unknown | 1.513088 | 0.6364 |
| ALGWEFLAF | PANX2 | HLAA2902 | 367.209397 | 49.64994743 | 5.506235491 | 0.45509 | 0.996 | Clonal | Unknown | 1.314423 | -0.1697 |
| STVRPCVVY | DDX3X | HLAA2902 | 59.58560552 | 78.23794322 | 0.254299066 | 0.06289 | 0.969 | Clonal | Driver | 1.220387 | 0.4365 |
| YRSTVRPCV | DDX3X | HLAC0602, HLAC0701 | 70.69349904 | 78.23794322 | 0.254299066 | 0.03342 | 0.969 | Clonal | Driver | 1.184798 | 0.4186 |
| LRREETDSF | PCLO | HLAC0701, HLA0602 | 57.61078504 | 35.11920422 | 2.336654739 | 0.17365 | 0.364 | Clonal | Unknown | 1.116602 | 0.7542 |

*3.3.7. Population Coverage*

Population coverage analysis using the IEDB tool showed 36.71% global coverage and 33.73% coverage for the Indian population. Moderate coverage values are expected due to high HLA polymorphism but support the feasibility of personalized neoantigen vaccines, where epitopes are selected based on patient-specific HLA alleles [41].

### 3.4. Multi-Epitope Vaccine Design

*3.4.1. Initial Construct*

A multi-epitope vaccine construct was designed by joining the top-ranked neoantigens using AAY linkers to facilitate antigen processing. The initial construct contained 10 peptides with a total length of 123 amino acids and showed an antigenicity score of 0.4002 using the VaxiJen server, indicating moderate immunogenic potential. Multi-epitope vaccines allow simultaneous targeting of multiple tumor antigens, increasing the probability of immune recognition in heterogeneous tumors [42].

*3.4.2. Optimized Construct*

To improve antigenicity and reduce low-rank candidates, the construct was redesigned using the top seven neoantigens. The final amino-acid sequence of the seven-epitope multi-epitope vaccine construct is shown in Fig. 3. The optimized construct contained 87 amino acids and showed increased antigenicity (0.4802). Shorter constructs may improve peptide processing efficiency and reduce the formation of non-functional junctional epitopes, thereby enhancing immune response [43].

```
>Multi_epitope_vaccine_7_peptides

LLADPTGALGKAAYHEDVPICIYAAYQEHEDVPICIYAAYLLADPTGALAAYSTVRPCVVYAAYYRSTVR
PCVAAYLRREETDSF
```

**Figure 3** FASTA sequence of the multi-epitope vaccine construct composed of seven selected neoantigen peptides joined using AAY linkers

*3.4.3. Antigenicity, Allergenicity, and Toxicity*

The optimized construct was predicted to be antigenic by VaxiJen, non-allergenic by AllergenFP, and non-toxic by ToxinPred. Evaluation of these parameters is essential in computational vaccine design because an ideal construct should have high immunogenicity while maintaining low toxicity and allergenicity [44,45].

### 3.5. Limitations

This study has several limitations that should be considered when interpreting the results. Only MHC class I-restricted epitopes of 8–11 amino acids were analyzed, whereas MHC class II-restricted and non-canonical neoantigens were not evaluated, although these have been reported to contribute significantly to anti-tumor immune responses [37,45,46].

In addition, all predictions were based on computational analysis, and experimental validation was not performed. Confirmation of predicted neoantigens requires immunopeptidomics analysis, peptide–HLA binding assays, and T-cell activation studies before clinical application [47,12].

## 4. Conclusion

This study developed an integrated immunogenomic workflow for the identification of tumor-specific neoantigens and the design of a multi-epitope vaccine candidate using combined whole-exome sequencing, RNA-seq analysis, HLA typing, peptide–MHC binding prediction, and multi-parameter filtering. High-quality sequencing data enabled reliable detection of somatic mutations, while integration of transcriptomic and immunoinformatics analyses allowed prioritization of biologically relevant neoantigens. Clonality analysis indicated that most predicted neoantigens originated from dominant tumor clones, supporting their suitability as immunotherapy targets.

Sequential filtering and composite scoring reduced the large candidate pool to a small number of high-confidence neoantigens, which were used to construct an optimized multi-epitope vaccine with improved antigenicity and favorable safety predictions. Population coverage analysis demonstrated moderate global and regional coverage, supporting the feasibility of multi-epitope strategies to address HLA diversity.

Although the results are based on computational predictions and require experimental validation, the proposed pipeline provides a systematic framework for neoantigen discovery and personalized cancer vaccine design. This approach may facilitate the development of patient-specific immunotherapies and supports the growing role of integrated genomics and immunoinformatics in precision oncology.

## Compliance with ethical standards

*Acknowledgments*

The author acknowledges NCBI for providing publicly available sequencing data. The author also thanks the developers of tools such as BWA, GATK, HISAT2, VEP, MHCflurry, and IEDB.

*Disclosure of conflict of interest*

The author declares no conflict of interest.

*Statement of ethical approval*

This study used publicly available data and did not involve human or animal subjects. Therefore, ethical approval and consent were not required.

## References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.CA: A Cancer Journal for Clinicians. 2021;71(3):209–249.

[2] Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nature reviews Clinical oncology. 2018 Feb;15(2):81-94

[3] Longley DB, Johnston PG. Molecular mechanisms of drug resistance. The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland. 2005 Jan;205(2):275-92.

[4] Ribas A, Wolchok JD. Cancer immunotherapy using checkpoint blockade. Science. 2018 Mar 23;359(6382):1350-5.

[5] Sharma P, Allison JP. The future of immune checkpoint therapy. Science. 2015 Apr 3;348(6230):56-61.

[6] Aurisicchio L, Pallocca M, Ciliberto G, Palombo F. The perfect personalized cancer therapy: cancer vaccines against neoantigens. Journal of Experimental & Clinical Cancer Research. 2018 Apr 20; 37(1):86.

[7] Li X, You J, Hong L, Liu W, Guo P, Hao X. Neoantigen cancer vaccines: a new star on the horizon. Cancer Biology & Medicine. 2024 Apr 15; 21(4):274-311.

[8] Schumacher, T.N. and Schreiber, R.D., 2015. Neoantigens in cancer immunotherapy. *Science*, *348*(6230),pp.69-74.

[9] Capietto AH, Hoshyar R, Delamarre L. Sources of cancer neoantigens beyond single-nucleotide variants. International journal of molecular sciences. 2022 Sep 4; 23(17):10131.

[10] Nathanson T, Ahuja A, Rubinsteyn A, Aksoy BA, Hellmann MD, Miao D, et al. Somatic mutations and neoepitope homology in melanomas treated with CTLA-4 blockade. Cancer immunology research. 2017 Jan 1; 5(1):84-91.

[11] Bulashevska A, Nacsa Z, Lang F, Braun M, Machyna M, Diken M, et al. Artificial intelligence and neoantigens: paving the path for precision cancer immunotherapy. Frontiers in immunology. 2024 May 29; 15:1394003.

[12] Cai Y, Chen R, Gao S, Li W, Liu Y, Su G, et al. Artificial intelligence applied in neoantigen identification facilitates personalized cancer immunotherapy. Frontiers in oncology. 2023 Jan 9; 12:1054231.

[13] Hammerbacher J, Snyder A. Informatics for cancer immunotherapy. Annals of Oncology. 2017 Dec 1; 28: xii56-73.

[14] Hundal J, Kiwala S, McMichael J, Miller CA, Xia H, Wollam AT, et al. pVACtools: a computational toolkit to identify and visualize cancer neoantigens. Cancer immunology research. 2020 Mar 1; 8(3):409-20.

[15] Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics. 2014 Dec 1;30(23):3310-6.

[16] O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. Cell systems. 2020 Jul 22; 11(1):42-8.

[17] Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. Nucleic acids research. 2015 Jan 28;43(D1): D423-31.

[18] 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 research. 2020 Jul 2;48(W1): W449-54.

[19] Bui HH, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. BMC bioinformatics. 2006 Mar 17;7(1):153.

[20] Shey RA, Ghogomu SM, Nebangwa DN, Shintouo CM, Yaah NE, Yengo BN, et al. Rational design of a novel multi-epitope peptide-based vaccine against Onchocerca volvulus using transmembrane proteins. Frontiers in Tropical Diseases. 2022 Nov 3; 3:1046522.

[21] Dimitrov I, Naneva L, Doytchinova I, Bangov I. AllergenFP: allergenicity prediction by descriptor fingerprints. Bioinformatics. 2014 Mar 15; 30(6):846-51.

[22] Chowell D, Morris LG, Grigg CM, Weber JK, Samstein RM, Makarov V, et al. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. Science. 2018 Feb 2;359(6375):582-7.

[23] Zhao Y, Fang LT, Shen TW, Choudhari S, Talsania K, Chen X, et al. Whole genome and exome sequencing reference datasets from a multi-center and cross-platform benchmark study. Scientific data. 2021 Nov 9;8(1):296.

[24] Sosinsky A, Ambrose J, Cross W, Turnbull C, Henderson S, Jones L, et al. Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. Nature medicine. 2024 Jan; 30(1):279-89.

[25] Mercer TR, Xu J, Mason CE, Tong W, MAQC/SEQC2 Consortium. The Sequencing Quality Control 2 study: establishing community standards for sequencing in precision medicine. Genome Biology. 2021 Nov 8; 22(1):306.

[26] Tbeileh N, Timmerman L, Mattis AN, Toriguchi K, Kasai Y, Corvera C, et al. Metastatic colorectal adenocarcinoma tumor purity assessment from whole exome sequencing data. PLoS One. 2023 Apr 6;18(4): e0271354.

[27] Luo R, Chong W, Wei Q, Zhang Z, Wang C, Ye Z, et al. Whole-exome sequencing identifies somatic mutations and intratumor heterogeneity in inflammatory breast cancer. NPJ breast cancer. 2021 Jun 1; 7(1):72.

[28] Ptashkin RN, Ewalt MD, Jayakumaran G, Kiecka I, Bowman AS, Yao J, et al. Enhanced clinical assessment of hematologic malignancies through routine paired tumor and normal sequencing. Nature Communications. 2023 Oct 28;14(1):6895.

[29] Chen Q, Liu Y, Gao Y, Zhang R, Hou W, Cao Z, et al. A comprehensive genomic and transcriptomic dataset of triple-negative breast cancers. Scientific Data. 2022 Sep 24; 9(1):587.

[30] Petrizzo A, Tagliamonte M, Mauriello A, Costa V, Aprile M, Esposito R, et al. Unique true predicted neoantigens (TPNAs) correlates with anti-tumor immune control in HCC patients. Journal of Translational Medicine. 2018 Oct 19;16(1):286.

[31] Yu YJ, Shan N, Li LY, Zhu YS, Lin LM, Mao CC, et al. Preliminary clinical study of personalized neoantigen vaccine therapy for microsatellite stability (MSS)-advanced colorectal cancer. Cancer Immunology, Immunotherapy. 2023 Jul;72(7):2045-56.

[32] Kovacevic V, Milicevic OS, Ilic Raicevic NM, Kojicic M, Mijalkovic Lazic A, Skundric N, et al. INAEME: Integral Neoantigen Analysis with Entirety of Mutational Events. bioRxiv. 2023 Sep 29:2023-09.

[33] Routh ED, Van Swearingen AE, Sambade MJ, Vensko S, McClure MB, Woodcock MG, et al. Comprehensive analysis of the immunogenomics of triple-negative breast cancer brain metastases from LCCC1419. Frontiers in oncology. 2022 Jul 27;12:818693.

[34] Su. X, Jin H, Wang J, Lu H, Gu T, Gao Z, et al. Construction and validation of an immunoediting-based optimized neoantigen load (ioTNL) model to predict the response and prognosis of immune checkpoint therapy in various cancers. Aging (albany NY). 2022 May 25;14(10):4586.

[35] Qi T, Vincent BG, Cao Y. A multispecies framework for modeling adaptive immunity and immunotherapy in cancer. PLoS Computational Biology. 2023 Apr 21;19(4):e1010976.

[36] Nagel R, Pataskar A, Champagne J, Agami R. Boosting antitumor immunity with an expanded neoepitope landscape. Cancer research. 2022 Oct 17; 82(20):3637-49.

[37] Bedran G, Gasser HC, Weke K, Wang T, Bedran D, Laird A, et al. The immunopeptidome from a genomic perspective: establishing the noncanonical landscape of MHC class I–associated peptides. Cancer immunology research. 2023 Jun 2; 11(6):747-62.

[38] Boll LM, Perera-Bel J, Rodriguez-Vida A, Arpí O, Rovira A, Juanpere N, et al. The impact of mutational clonality in predicting the response to immune checkpoint inhibitors in advanced urothelial cancer. Scientific Reports. 2023 Sep 15; 13(1):15287.

[39] Tran NH, Qiao R, Xin L, Chen X, Shan B, Li M. Personalized deep learning of individual immunopeptidomes to identify neoantigens for cancer vaccines. Nature Machine Intelligence. 2020 Dec;2(12):764-71.

[40] Huber F, Arnaud M, Stevenson BJ, Michaux J, Benedetti F, Thevenet J, et al. A comprehensive proteogenomic pipeline for neoantigen discovery to advance personalized cancer immunotherapy. Nature biotechnology. 2025 Aug; 43(8):1360-72.

[41] Hao Q, Long Y, Yang Y, Deng Y, Ding Z, Yang L, et al. Development and clinical applications of therapeutic cancer vaccines with individualized and shared neoantigens. Vaccines. 2024 Jun 27;12(7):717.

[42] Imon RR, Samad A, Alam R, Alsaiari AA, Talukder ME, Almehmadi M, et al. Computational formulation of a multiepitope vaccine unveils an exceptional prophylactic candidate against Merkel cell polyomavirus. Frontiers in Immunology. 2023 Jun 27; 14:1160260.

[43] Biri-Kovács B, Bánóczi Z, Tummalapally A, Szabó I. Peptide vaccines in melanoma: chemical approaches towards improved immunotherapeutic efficacy. Pharmaceutics. 2023 Jan 30;15(2):452.

[44] Anzar I, Malone B, Samarakoon P, Vardaxis I, Simovski B, Fontenelle H, et al. The interplay between neoantigens and immune cells in sarcomas treated with checkpoint inhibition. Frontiers in immunology. 2023 Sep 20; 14:1226445.

[45] Wagutu G, Gitau J, Mwangi K, Murithi M, Melly E, Harris AR, et al. Whole exome-seq and RNA-seq data reveal unique neoantigen profiles in Kenyan breast cancer patients. Frontiers in Oncology. 2024 Dec 11; 14:1444327.

[46] Rospo G, Chilà R, Matafora V, Basso V, Lamba S, Bartolini A, et al. Non-canonical antigens are the largest fraction of peptides presented by MHC class I in mismatch repair deficient murine colorectal cancer. Genome Medicine. 2024 Jan 19; 16(1):15.

[47] Huff AL, Longway G, Mitchell JT, Andaloori L, Davis-Marcisak E, Chen F, et al. CD4 T cell–activating neoantigens enhance personalized cancer vaccine efficacy. JCI insight. 2023 Dec 8; 8(23): e174027.

[48] Tian H, Li G, Chiu CK, Li E, Chen Y, Zhu T, et al. Rapid and direct discovery of functional tumor specific neoantigens by high resolution mass spectrometry and novel algorithm prediction. Cell Insight. 2025 Jun 1;4(3):100251.