OXFORD

Briefings in Bioinformatics, 2024, **25(2)**, 1–12
https://doi.org/10.1093/bib/bbae024
Problem Solving Protocol

ImmuneMirror: A machine learning-based integrative pipeline and web server for neoantigen prediction

Gulam Sarwar Chuwdhury†, Yunshan Guo†, Chi-Leung Chiang, Ka-On Lam, Ngar-Woon Kam, Zhonghua Liu and Wei Dai

Corresponding authors: Wei Dai, Department of Clinical Oncology, Center of Cancer Medicine, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong (SAR), P. R. China. Tel.: (852) 39176930; Email: weidai2@hku.hku; Zhonghua Liu, Department of Biostatistics, Columbia University, 722 West 168th Street, New York, NY 10032, USA. Tel.: (001) 212 305 6604; Email: zl2509@cumc.columbia.edu

†Gulam Sarwar Chuwdhury and Yunshan Guo contributed equally

Abstract

Neoantigens are derived from somatic mutations in the tumors but are absent in normal tissues. Emerging evidence suggests that neoantigens can stimulate tumor-specific T-cell-mediated antitumor immune responses, and therefore are potential immunotherapeutic targets. We developed *ImmuneMirror* as a stand-alone open-source pipeline and a web server incorporating a balanced random forest model for neoantigen prediction and prioritization. The prediction model was trained and tested using known immunogenic neopeptides collected from 19 published studies. The area under the curve of our trained model was 0.87 based on the testing data. We applied *ImmuneMirror* to the whole-exome sequencing and RNA sequencing data obtained from gastrointestinal tract cancers including 805 tumors from colorectal cancer (CRC), esophageal squamous cell carcinoma (ESCC) and hepatocellular carcinoma patients. We discovered a subgroup of microsatellite instability-high (MSI-H) CRC patients with a low neoantigen load but a high tumor mutation burden (> 10 mutations per Mbp). Although the efficacy of PD-1 blockade has been demonstrated in advanced MSI-H patients, almost half of such patients do not respond well. Our study identified a subset of MSI-H patients who may not benefit from this treatment with lower neoantigen load for major histocompatibility complex I ($P < 0.0001$) and II ($P = 0.0008$) molecules, respectively. Additionally, the neopeptide YMCNSSCMGV-TP53^{G245V}, derived from a hotspot mutation restricted by HLA-A02, was identified as a potential actionable target in ESCC. This is so far the largest study to comprehensively evaluate neoantigen prediction models using experimentally validated neopeptides. Our results demonstrate the reliability and effectiveness of *ImmuneMirror* for neoantigen prediction.

Keywords: neoantigen prediction; machine learning; multiomics; gastrointestinal tract cancer; immunotherapy; computational platform

INTRODUCTION

Immunotherapy uses the immune system to detect and fight against cancer cells. Accumulating evidence shows that the presence of neoantigens derived from somatic mutations in tumor cells elicits a potent immune response as a part of antitumor immunity through specific cytotoxic T cells [1, 2]. Previously, various methods have been proposed for neoantigen identification, such as MHCflurry [3], NetMHCpan [4–6] and NN-Align [7], which predict the binding affinity between peptides and their corresponding major histocompatibility complex (MHC) alleles. Binding affinity is a good reference to prioritize neoantigens because MHC classes I and II help the immune system bring the bonded complex to the surface of cancerous cells for recognition by T cells. Therefore, binding to MHC molecules is a prerequisite for immunogenicity. However, the actual variant expression, human leukocyte antigen (HLA) presentation, peptide processing and transportation, as well as the ultimate T-cell response to these neoantigens, have not been considered in these existing binding affinity-based tools, therefore, these previous methods may fail to provide reliable predictions in real-world scenarios. Recently, by integrating peptide features, Wells *et al.* developed a model of tumor epitope immunogenicity to filter out nonimmunogenic peptides, and the results improved the effectiveness of neoantigen prediction [8]. This model is based on stringent cutoffs for several selected features, including binding affinity, binding stability,

Gulam Sarwar Chuwdhury received his MPhil in Bioinformatics from Li Ka Shing Faculty of Medicine at The University of Hong Kong. His research interests include immunology, single-cell multi-omics sequencing analysis, machine learning, bioinformatics software and pipeline development.
Yunshan Guo is a biostatistics PhD student at Yale University working primarily with biomedical data.
Chi-Leung Chiang is a clinical assistant professor at the Department of Clinical Oncology of Li Ka Shing Faculty of Medicine at The University of Hong Kong. His research interests include stereotactic body radiotherapy of liver and pancreatic cancer, rectal brachytherapy, and immunotherapy in gastrointestinal track cancers.
Ka-On Lam is a clinical assistant professor in the Department of Clinical Oncology of Li Ka Shing Faculty of Medicine at The University of Hong Kong. His research interests are novel treatment strategies and circulating tumor cells in gastrointestinal cancers and nasopharyngeal cancer.
Ngar-Woon Kam is a senior scientist in Laboratory for Synthetic Chemistry and Chemical Biology Limited in Hong Kong. She is interested in understanding the resistance to immunotherapy in esophageal cancer.
Dr. Zhonghua Liu is currently an assistant professor in the Department of Biostatistics at Columbia University. His research interests include statistical genetics/genomics, causal inference, machine learning.
Dr. Wei Dai is currently an assistant professor in the Department of Clinical Oncology of Li Ka Shing Faculty of Medicine at The University of Hong Kong. Her research combines bioinformatics and clinical translational study for identifying therapeutic targets and biomarkers for immunotherapy.
Received: September 8, 2023. **Revised:** December 5, 2023. **Accepted:** January 16, 2024

tumor abundance, the ratio of binding affinity between mutant and wild-type peptides [9], and T-cell receptor recognition probability (foreignness); the model showed promising results with precision (true positive/(true positive + false positive)) above 0.7. However, Wells' study [8] used different criteria during the training and validation steps to filter neoantigens, making it difficult to implement with other data sets. Hence, continuous efforts are still needed to further improve the prediction accuracy for clinical application by incorporating more relevant biological features that are involved in the complicated biological processes.

In this study, we developed ImmuneMirror, an all-in-one bioinformatics pipeline using multiomics sequencing data, to access the key genomic and transcriptomic features associated with the cancer immunotherapy response. The ImmuneMirror pipeline and web server (version 1.0) incorporate a machine learning (ML) model to incorporate more predictive biological features for neoantigen prediction. With this advanced ML model trained by known neopeptides with T-cell immunogenicity, ImmuneMirror overcomes the issue of unbalanced neoantigen distribution, i.e. immunogenic mutation-derived neoantigens are relatively rare compared to the total number of mutations detected. We applied ImmuneMirror to real-world data sets to systematically investigate neoantigens in gastrointestinal tract (GIT) cancers using matched whole-exome sequencing (WES) and bulk RNA Sequencing (RNA-Seq) data; furthermore, we compared the results with putative neoantigens that are derived from hotspot mutations in cancer-related genes restricted by the following four common HLA alleles: HLA-A02:07, HLA-A24:02, HLA-A02:01 and HLA-A11:01. The top candidate neopeptide, YMCNSSCMGV-TP53^{G245V}, derived from a hotspot mutation restricted by HLA-A02, was evaluated experimentally.

MATERIALS AND METHODS

Selecting ML algorithms for neoantigen prediction

To build a prediction model for identifying neoantigens and incorporating more relevant genomic and transcriptomic features, we first gathered a list of neopeptides with experimentally confirmed T-cell responses as the training data for model construction.

The binding affinities of peptide amino acids were a key feature to be included and were predicted through pVACtools [10], which is a comprehensive tool to provide binding affinity scores calculated by various prediction algorithms for MHC class I. Additionally, considering that binding affinity is not the only feature governing tumor epitope immunogenicity, we added the following relevant features to improve prediction accuracy: 'agretopicity' [9, 11], 'foreignness' [12–14], hydrophobicity, binding stability, peptide processing and transportation scores. The final training data set included a total of 1199 peptides that were tested *in vitro*, 93 of which had positive T-cell responses. Of the 211 tested peptides, 10 were immunogenic. These neopeptides were identified from 19 published studies (Supplementary Tables S1 and S2).

The class distribution of the response variable is unbalanced due to the low proportion (about 7%) of neoantigens that activate T cells. Consequently, conventional ML classification algorithms are largely affected by the majority class (negative T cell response) and thus may give biased attention to the minority class (positive T cell response), leading to relatively poor prediction performance. To address this critical issue, we adapted the balanced random forest learning algorithm [15] to improve prediction accuracy that is evaluated using the area under the receiver operating characteristic curve (AUC) metric.

Random forests are ensemble ML algorithms that construct multiple base decision trees during the training process. More specifically, given a training dataset $(x_i, y_i) \in X \times Y, i = 1, \ldots, n$, where Y is the binary response variable with value 1 representing activation of T-cell response and 0 otherwise, and X represents all the predictive features (Supplementary Table S2). The random forests algorithm repeatedly selects a bootstrap random sample (B times) with replacement and a random subset of features from the training dataset, and then fits base decision trees f_b to each of those bootstrap random samples $X_b, Y_{b,}$. When building each base decision tree, a random number of m predictors are selected from the entire p predictor pool; typically, we set $m \approx \sqrt{p}$. The Gini index, defined as $G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$, where $\hat{p}_{mk}$ is the proportion of training observations in the mth region belonging to the kth class; the Gini index is used as a criterion to make the binary split when growing a tree. The final binary classification output is based on majority voting from all the base decision trees [16].

Random forests with the synthetic minority over-sampling technique

Conventional random forest uses the standard bootstrap resampling strategy with equal sample probability for each observation, and this strategy may not perform well for an imbalanced dataset. Therefore, we proposed to modify the random forest algorithm using a more advanced resampling technique that over-samples observations from the minority class and under-samples observations from the majority class to increase the minority-majority ratio from approximately 1:12 to 1:3. This under-sampling step can be achieved using the *smote_and_undersample* function in the R package *hyperSMURF* [17]. This function first generates synthetic examples based on the synthetic minority over-sampling technique (SMOTE), which retains each minority class sample and introduces synthetic examples along the line segments joining some of the k minority class nearest neighbors [18]. In our case, we set the multiplicative factor f_p to 2 and k to 5, so two neighbors from the five nearest neighbors were selected. Then, observations from the majority class were under-sampled to reach the preset class ratio. We then fit the conventional random forest on the resampled data. These processes were repeated 200 times to ensure that most of the data points are involved in the training process. To decide the optimal values of hyperparameter: 'number of tree' and 'tunelength', we used the 5-fold cross-validation. We tested the range of 'number of tree' from 60, 65, 70 to 100. And for 'tunelength', we set from 5, 10 to 15. We also tried different combinations, and for each iteration we evaluated the model's performance on the validation fold. Finally, we selected the model ('number of tree'=95, 'tunelength' = 5) with the best AUC (0.8294) based on the testing data (Figure 1).

Balanced random forest

To address the imbalanced data issue, Chen and Breiman [19] developed the balanced random forest algorithm, which substantially improved the performance of the random forest algorithm by replacing the equal-weight sampling strategy with random under-sampling of majority class for decision tree formation. More specifically, for each iteration in a random forest, we randomly drew a bootstrap sample from the minority class and obtained the same number of neoantigen candidates with negative T cell response from the majority class with replacement. Then, we formed a classification tree based on the resampled balanced data. We repeated the above steps 500 times

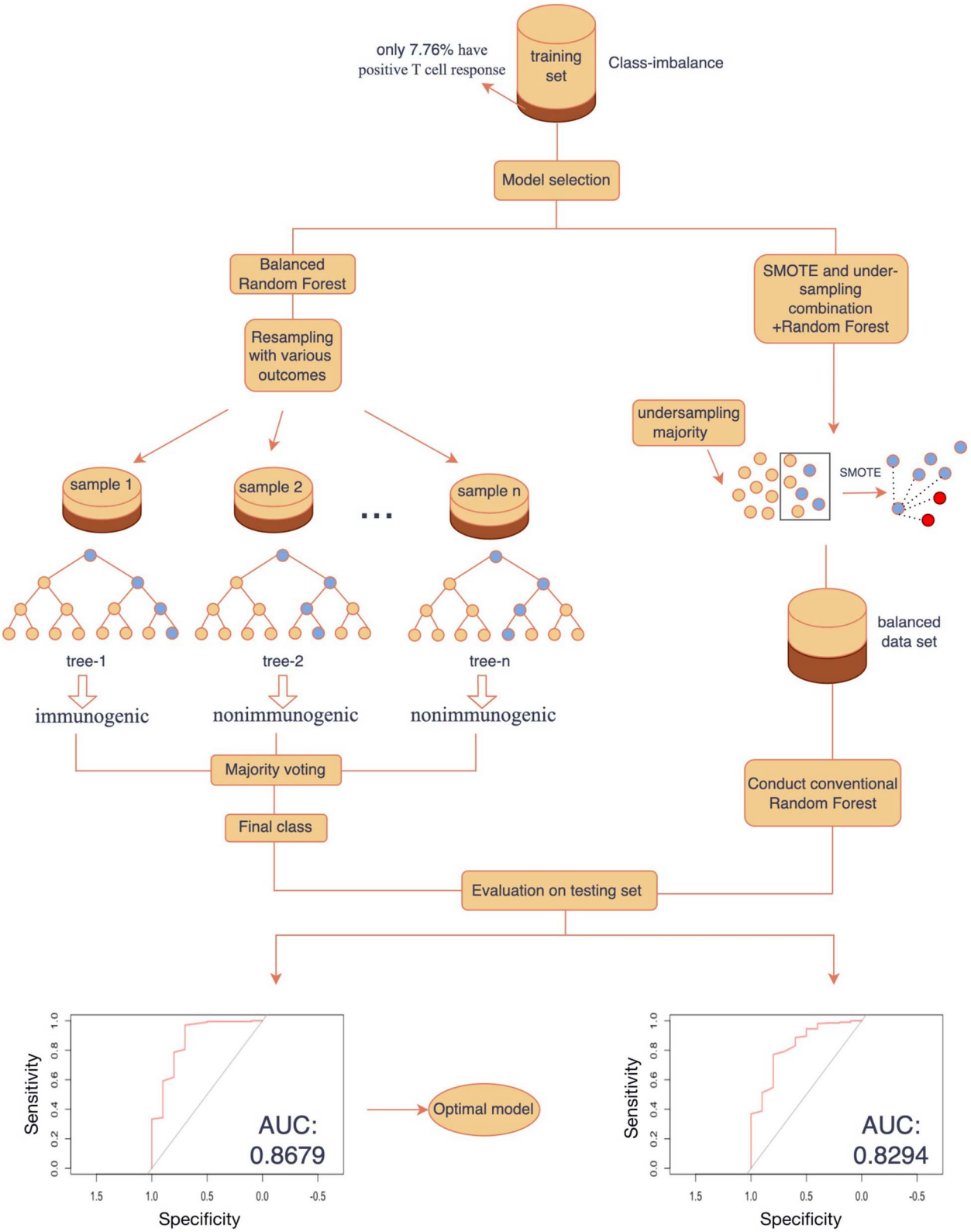

Figure 1. Evaluation of different random forest algorithms used for neoantigen prediction with AUC performance: Left panel: balanced random forest; right panel: random forest with under sampling.

and then determined the final prediction via majority voting [19]. The balanced random forest algorithm was implemented in the *train* function in the R package *caret* [20], and the optimal value of parameters was tuned by a 5-fold cross-validation method. The AUC on the testing set was 0.8679 (Figure 1). By evaluating the above two models, the balanced random forest model has better prediction accuracy for neoantigen prediction.

Peptide synthesis and quality control

The selected peptides were synthesized. The CI resins were selected and deprotected in 20% piperidine dimethylformamide (DMF) solution. The resin was filtered off and rinsed with DMF three times to remove Fmoc residues. The completeness of amino deprotection was measured by taking a sample of the resin and mixing it with detection reagents A and B. If there was a color change, the Fmoc groups were removed successfully. The amino acid solution was added into a mixture of the resin and di-isopropyl carbodiimide in DMF, and the mixture was shaken at room temperature. The completeness of the coupling reaction was confirmed by taking a sample of the resin and mixing with detection reagents A and B, followed by resin washing with DMF three times. When all the amino acids were coupled onto the resins, the peptide chain was dissociated from the resins by treatment with TFA/DMF. The crude peptides were further purified by reversed-phase high-performance liquid chromatography and were frozen and dried under vacuum. The molecular weights of the selected peptides were analyzed by LC-MS. Endotoxin levels were detected using Horseshoe Crab Reagent. The peptide with an endotoxin level < 10 EU/mg was used for the MHC binding assays.

HLA-A02:01 peptide-binding assay

The QuickSwitch™ Quant HLA-A02:01 Tetramer Kit-PE was used to investigate the binding affinity of the selected neoepitope to MHC HLA-A02:01. The synthesized peptides were incubated with the MHC HLA-A02:01 complex, which already contained a control peptide. The tested peptide competed with the control peptide, and the exchange rate was used to identify the peptide-binding affinity. QuickSwitch™ Quant HLA-A02:01 Tetramer Kit-PE and flow cytometry were used to investigate the exchange rate of the MHC HLA-A02:01 control peptide. The tested peptide was mixed with the tetramer and peptide exchange factor for 4.5 h at room temperature. The peptide exchange rate was quantitated by flow cytometry [21, 22]. The reference positive peptide was provided by the QuickSwitch™ Quant HLA-A*02:01 Tetramer Kit-PE.

RESULTS

Overview

The overall workflow of this study is depicted in Figure 2. The ML model was developed using the balanced random forest algorithm for neoantigen prediction using multiple biological features relevant to neoantigen biogenesis, transportation, presentation, and T-cell recognition (agretopicity, foreignness, hydrophobicity, binding stability, peptide processing, and transportation scores). This ML model was incorporated into the ImmuneMirror bioinformatics pipeline, which is also a web server for neoantigen prediction and prioritization from multiomics sequencing data. The pipeline takes the raw FASTQ reads as input, while the web server takes variant call format (VCF) files containing the somatic mutations (Figure 3). We applied this pipeline to identify neoantigens derived from somatic mutations in cancer-related genes with common MHC class I subtypes in Pan-Cancer studies and from real-world WES and RNA-Seq data from GIT cancer patients. Experiments were carried out to confirm the binding affinity of the putative neoantigens with MHC class I HLA-A02:01.

Implementation of the ImmuneMirror pipeline and web server

We developed the ImmuneMirror pipeline for neoantigen prediction and prioritization based on multiple genomic and transcriptomic features. The workflow of the ImmuneMirror pipeline is depicted in Supplementary Figure S1. The pipeline was built as a docker container that can be run in any docker-supported operating system, such as Linux, Mac and Windows. The pipeline required FASTQ input of matched normal-tumor WES samples and tumor bulk RNA-Seq samples. The full list of packages and software that were used for ImmuneMirror pipeline development are listed in Supplementary Table S3. We implemented the prediction model as an R function for prioritizing the neoantigens restricted by HLA class I. The germline and somatic mutations, estimated tumor mutation burden (TMB), microsatellite instability (MSI) status [a condition of genetic hypermutability due to defective DNA mismatch repair (MMR)], HLA typing, neoantigen load for HLA class I and II, the top-ranked neoantigens with T-cell immunogenicity, and the expression of innate anti-PD1 resistance (IPRES) gene expression signature [23] are the final outputs of the pipeline. Users can download the Docker image and the relevant files (reference files and example samples) from http://immunemirror.hku.hk/ and clone the ImmuneMirror pipeline from the GitHub repository (https://github.com/weidai2/ImmuneMirror/) [24].

Apart from the development of the stand-alone pipeline, we also developed an ImmuneMirror web server (Figure 3) that takes a VCF file containing the somatic mutations detected by MuTect2 (GATK4) [25] as the input and identifies the potential neoantigens derived from somatic mutations for both HLA class I and class II molecules. Users can upload a VCF file, enter a set of alleles for both HLA class I and II, and select peptide lengths via the web interface. The uniform resource locator link for downloading the results will be sent to the user-provided e-mail automatically by the server upon job completion. The web server is freely available for users with detailed usage instructions at http://immunemirror.hku.hk/App/.

Graphical analysis report

With the advantages of our developed analysis database, ImmuneMirror produces a visual analysis report for each of the samples. The report, as illustrated in Supplementary Figure S2, includes TMB, HLA types, neoantigen load for HLA class I and II, MMR status, germline and somatic mutations, ImmuneMirror prediction score, and IPRES gene expression signature [23]. The TMB is shown as the number of mutations per Mb. The HLA typing of the sample is presented in a table for class I and class II. The number of neoantigens restricted by HLA class I and class II are illustrated as box plots and bar plots with indicators for high and low neoantigen loads, respectively. The MMR status of the sample is also reported. The cutoff for the MSI-high group was determined by the optimal value of the MSIsensor-pro score for distinguishing the MSI group from the other groups in CRC. The sample with a MSIsensor-pro score higher than the cutoff was defined as MMR deficient. ImmuneMirror prediction scores are presented as a boxplot (Supplementary Figure S2(I)). Moreover, both germline variants and somatic mutations of selected genes, such as *BRCA2*, *B2M*, *MLH1* and *MSH2*, and expression of the genes from the IPRES signature are included in the analysis report. It has been reported that these selected mutations and gene expression signatures are relevant to the immunotherapy response [23, 26].

Testing, computation speed evaluation and resources

We tested the pipeline on Linux operating systems (ubuntu 20.04). It took approximately 30 h to process one pair of samples with 13 threads. Moreover, we ran the pipeline with multiple pairs of samples from different cancer types, i.e. ESCC, HCC and CRC.

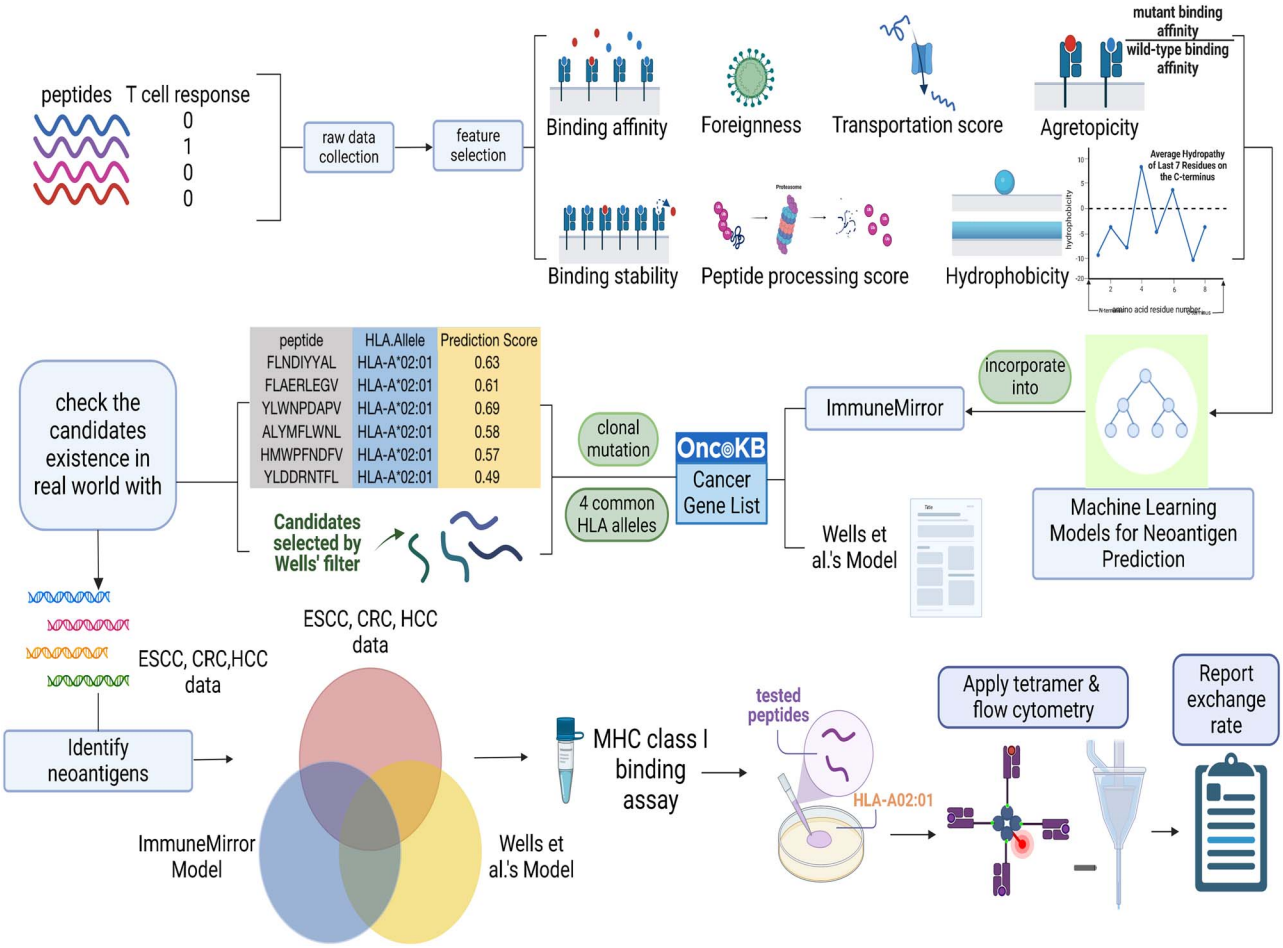

Figure 2. Overall study workflow. Neopeptides with experimentally confirmed T-cell responses were gathered as training data for model construction. Relevant features were selected through feature selection. The prediction model was established with the balanced random forest algorithm [19], and then the ImmuneMirror pipeline was developed by incorporating built prediction model. ImmuneMirror was subsequently applied to the hotspot mutations derived from the common cancer gene list from OncoKB [28] to predict potential neoantigens. Wells' criteria [8] were also applied to hotspot mutations for the selection of neoantigens. The publicly available data from ESCC, CRC and HCC patients were processed and analyzed by ImmuneMirror. We compared the results obtained from the two data resources and identified overlapping candidates that were then subject to experimental validation of binding affinity with HLA-A02. ESCC, esophageal squamous cell carcinoma; CRC, colorectal cancer; HCC hepatocellular carcinoma.

Users can run ImmuneMirror with a list of samples, and the actual run time depends on the computation speed and resources of their own devices. In general, we recommend a device with at least 64 GB of RAM and the necessary space for the pipeline, including docker image (79.6 GB), supporting files (483 GB) and analysis results (approximately 41 GB for one pair of samples), to successfully run the pipeline. The supporting files provide the necessary resources, such as the reference human genome (hg38), to run the pipeline; thus, no additional step is needed to download these files or to reconfigure the pipeline. The web server has been tested on Linux, macOS and Windows platforms with various web browsers (Supplementary Table S4). The format of the input/output files and detailed instructions are provided on the website and will be updated regularly.

Comparison of features for neoantigen prediction tools

We compared the bioinformatics tools available for neoantigen prediction (Supplementary Table S5). Compared to other existing pipelines, only ImmuneMirror has all the following seven unique features: methods used for prioritization, docker image, web server, neoantigen prediction for HLA class I and II, multiple prediction algorithms, open source. As a docker image, the ImmuneMirror pipeline takes the raw FASTQ files from both WES (matched normal-tumor pairs) and RNA-Seq (tumor, optional) data as the input. On the other hand, similar to pVAC-Seq [10], ImmuneMirror can be used for neoantigen prediction restricted by HLA class I and class II using multiple algorithms. Besides, ImmuneMirror provides a unique web server taking the VCF file that contains the somatic mutations detected by MuTect2 from GATK4 [25] as the input for neoantigen prediction, which makes ImmuneMirror more user-friendly.

Moreover, we performed a comparative analysis to identify neoantigens by OpenVax [27], a tool closely similar to ImmuneMirror pipeline. Both OpenVax and ImmuneMirror take raw FASTQ files from WES (normal-tumor) samples and bulk RNA-Seq data from tumor samples as their input, utilizing somatic mutations as the basis for neoantigen prediction. Compared to OpenVax, ImmuneMirror has an additional web server for neoantigen prediction from VCF input file, containing a list of somatic mutations. In this comparative analysis, we randomly selected three samples from distinct cancer types: CRC, HCC and ESCC for comparison of the neoantigen candidates. In total, we identified 44 and 52

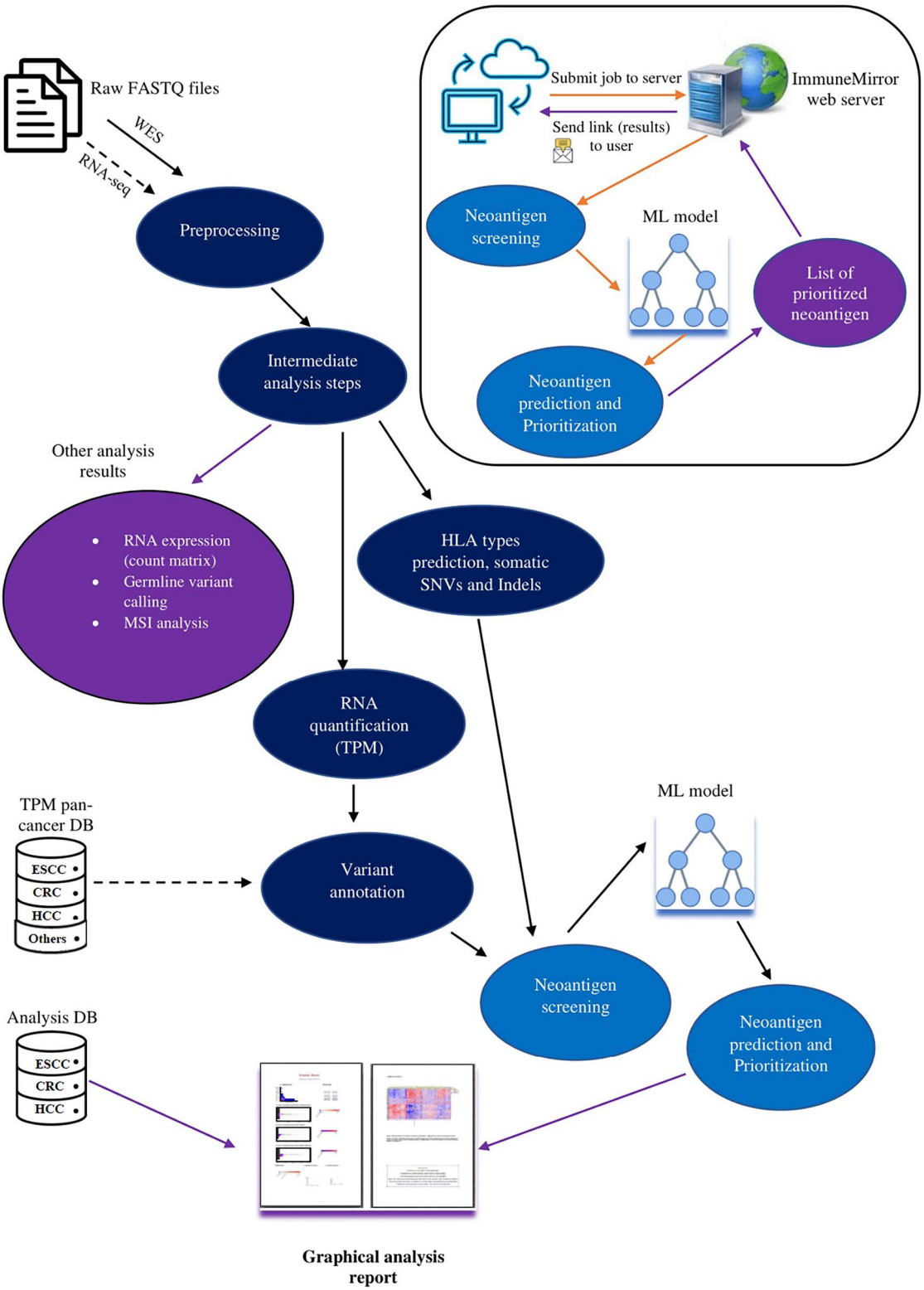

Figure 3. The overall workflow of ImmuneMirror, major analysis steps involved in the ImmuneMirror pipeline and the ImmuneMirror web server. The ImmuneMirror pipeline preprocesses raw FASTQ files, including multiple analysis steps (e.g. prediction of HLA subtypes, SNV and Indels detection, variant annotation, neoantigen prediction and prioritization), and generates a graphical analysis report for each sample. The input for the web server is a VCF file, and the analysis result (list of prioritized neoantigens) is sent as a web link to the email address of the end user. HLA, human leukocyte antigen; SNV, single nucleotide variant; Indel, insertion or deletion; VCF, variant call format.

Table 1: The neoantigens identified from hotspot mutations in TCGA Pan-Cancer studies

Gene	Mutation	Neoepitope (mutation)	HLA restriction	ImmuneMirror score
BRAF	V600K	KIGDFGLATK	A∗11:01	0.7
PIK3CA	N345K	KILCATYVK	A∗11:01	0.9
PIK3CA	E542K	AISTRDPLSK	A∗11:01	0.6
PIK3CA	E545K	STRDPLSEITK	A∗11:01	0.7
PIK3CA	H1047L	ALHGGWTTK	A∗11:01	0.7
TP53	G245V	YMCNSSCMGV	A∗02:01	0.5
TP53	P250L	RLILTIITL	A∗02:01	0.6
TP53	C242F	HYNYMCNSSF	A∗24:02	0.7
TP53	S241F	HYNYMCNSF	A∗24:02	0.8

neoantigens from these three samples using ImmuneMirror and OpenVax, respectively. Notably, 24 neoantigens were found to be common between both tools (Supplementary Figure S3).

Application of ImmuneMirror to real-world data

Identification of clonal mutations as neoantigens from TCGA pan-cancer studies

The top 27 cancer-relevant genes with hotspot mutations (frequency > 0.1%) from the OncoKB [28] cancer gene list were selected for analysis. Each mutant is paired with four common HLA alleles, HLA-A02:07, HLA-A24:02, HLA-A02:01 and HLA-A11:01, derived from Asian populations. The prediction scores of the mutant-HLA combinations as well as many other biological features were calculated by ImmuneMirror. Wells *et al.* [8] developed selection criteria (binding affinity <34 nM; binding stability >1.4 hours; tumor abundance >33 TPM; agretopicity <0.1 or foreignness $>10^{-16}$) to select neoantigens based on several experimental validation results [8]. To present a more thorough analysis, we also applied Wells' criteria [8] to all the mutants identified from these 27 genes to compare with our prediction results. Neoantigen candidates were finalized if (1) the prediction score was greater than 0.515 (sensitivity: 0.851; specificity: 0.7, evaluated by the testing data set) and (2) they fulfilled Well's criteria [8] with an adapted gene expression cutoff of TPM >10. We finally identified a total of 9 neoantigens derived from the mutations of 27 genes with a mutation frequency > 0.1% from TCGA Pan-Cancer studies. The results included multiple potential neoantigens derived from $TP53^{P250L}$, $TP53C^{242F}$ and $TP53^{S241F}$ (Table 1).

In addition, our analysis also indicated that the hotspot mutations $BRAF^{V600K}$, $PIK3CA^{N345K}$, $PIK3CA^{E542K}$, $PIK3CA^{E545K}$ and $PIK3CA^{H1047L}$ are promising candidates for neoantigens derived from cancer-relevant genes. Nearly half of all cutaneous melanomas carry activating $BRAF^{V600}$ mutations, among which 10–30% contain the $BRFA^{V600K}$ mutation, making it the second most common genotype after $BRAF^{V600E}$ [29, 30]. $BRAF^{V600K}$ lead to a gain in BraF protein function, as demonstrated by increased kinase activity, increased downstream signaling, and the ability to transform cells *in vitro* [31, 32]. Clinically, $BRAF^{V600K}$ tumors cause patients to experience distant metastases sooner, and these patients have a higher risk of relapse and shorter survival than those with V600E tumors [33].

Identification of GIT cancer neoantigens

We further evaluated the genomic and transcriptomic data from colorectal cancer (CRC), esophageal squamous cell carcinoma (ESCC) and hepatocellular carcinoma (HCC) patients to further evaluate the putative neoantigens in these three types of cancers in the real world. We collected a total of 805 samples from different data sources (Supplementary Table S6). After quality checking, we analyzed a total of 691 samples, composed of 316 CRC samples, 290 ESCC samples and 85 HCC samples. On average, we identified 17 (0, 316), 5 (0, 76) and 6 (0, 64) neoantigens by ImmuneMirror for each CRC, ESCC and HCC patient, respectively. Noticeably, the neoantigen load was significantly correlated with favorable clinical outcome in terms of longer overall survival in ESCC samples (Supplementary Figure S4). CRC patients can be categorized as high MSI-high (MSI-H), low MSI-low (MSI-L) and microsatellite stability according to the status of the mismatch repair pathway [34]. MSI-H tumors respond well to immunotherapy, presumably due to a high TMB and neoantigen load [35, 36]. More interestingly, although the neoantigen load was not correlated with overall survival in CRC samples, we found that a subgroup of MSI-H CRC patients with MMR deficiency had a much lower neoantigen load for both HLA class I and II and a high TMB that was comparable to other MSI-H CRC patients (Figure 4). These patients were subject to advanced T stage (T4 versus others: 30.8% versus 0%, Fisher's exact test $P = 0.011$).

We identified a total of 12 putative neoepitopes that fulfilled Well's criteria [8] and had an ImmuneMirror prediction score > 0.5. These neoepitopes were derived from TP53, STAT3 and RAB35 with high affinity for the HLA-A∗02:01, HLA-A∗11:01, HLA-A∗33:03, HLA-A∗33:01, HLA-A∗03:01 and HLA-A∗02:06 HLA alleles (Table 2). More specifically, the neoepitope $TP53^{G245V}$ (YMCNSSCMGV) restricted by HLA-A∗02 was identified in the real-world data analysis of ESCC patient samples. This mutation affects the binding of p53 to DNA and interferes with the protein's transcription activity. The RNA-Seq data indicated that this mutant is widely expressed in the tumor tissues (Supplementary Figure S5).

Validation of HLA-A02 binding with TP53.pG245V

We evaluated HLA-A02 binding affinity with neoepitopes derived from multiple mutations at G245 in TP53 using the QuickSwitch Quant HLA-A∗02:01 Tretramer Kit-PE. The neoepitope $TP53^{G245V}$ (YMCNSSCMGV) had a higher reference peptide exchange rate of 97.03% than the wild-type peptide YMCNSSCMGG (80.8%) (Figure 5A). Among the five most common mutations at G245 of the gene TP53, the binding affinity of neoepitope-$TP53^{G245V}$ (YMCNSSCMGV) was the highest among $TP53^{G245R}$, $TP53^{G245D}$, $TP53^{G245C}$ and $TP53^{G245S}$ (Figure 5B), and the Pearson's correlation between the ImmuneMirror prediction scores and binding affinities was 0.897 (Figure 5C). This result confirmed the effectiveness and reliability of ImmuneMirror as an advanced tool for neoantigen prediction.

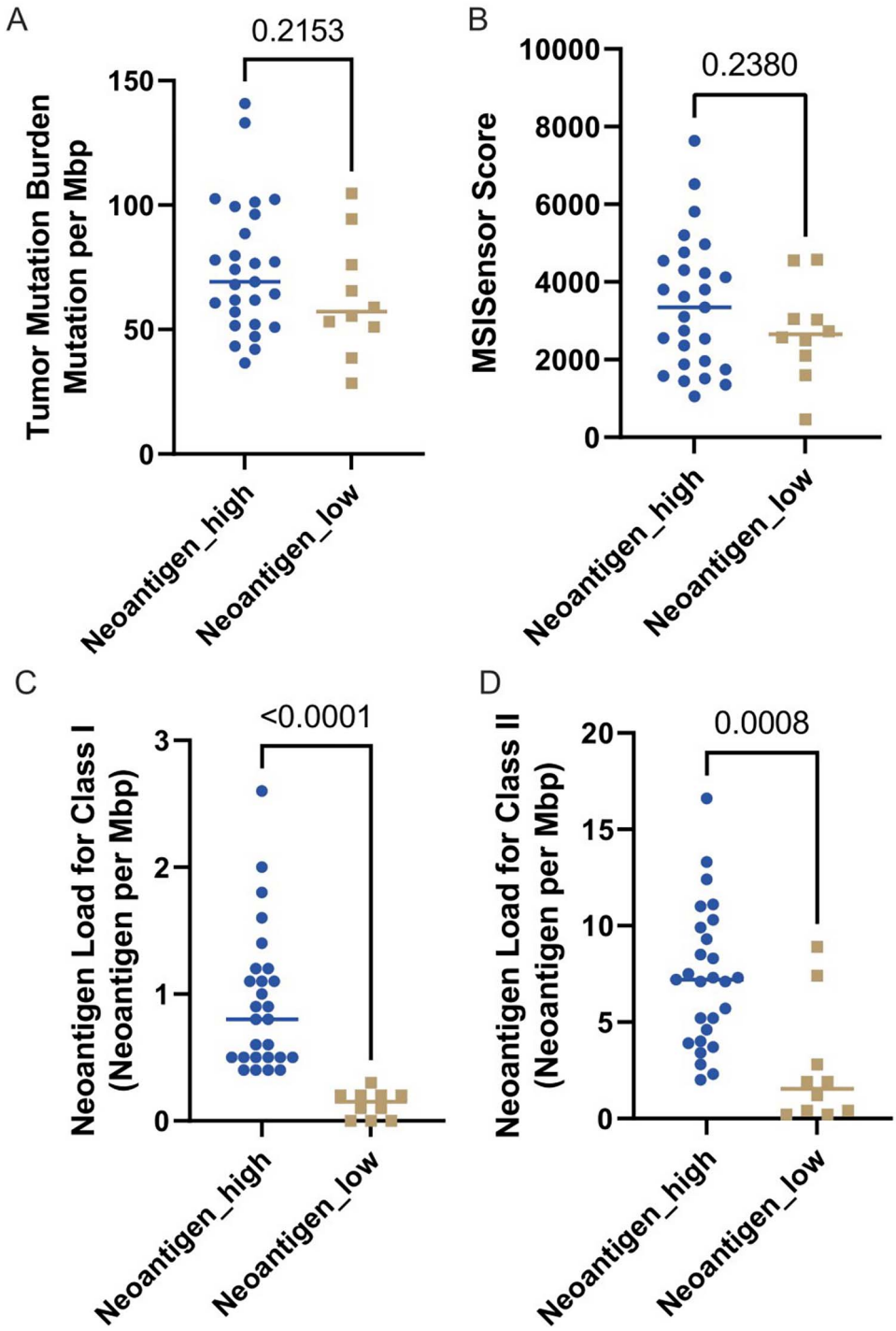

Figure 4. Lower neoantigen loads were detected in a subgroup of MSI-H CRC patients. (**A**) TMB. (**B**) MSISensor score for MSI status. (**C**) The neoantigen load for HLA class I. (**D**) The neoantigen load for HLA class II. HLA, human leukocyte antigen.

DISCUSSION

We developed ImmuneMirror as a self-standing open-source pipeline and a web server for neoantigen prediction and prioritization by integrating a balanced random forest model. ImmuneMirror was trained and tested using immunogenic neoantigens collected from 19 studies (Supplementary Tables S1 and S2). To the best of our knowledge, this is the largest study to date to comprehensively evaluate the neoantigen prediction model using experimentally validated neopeptides. Accurate neoantigen prediction depends on inclusion of the most predictive biological features that essentially govern epitope immunogenicity. Referring to published studies, our model integrates important biological features of immunogenic neoantigens. Then, we developed a prediction model based on the advanced balanced random forest algorithm [19]. The effectiveness and reliability of ImmuneMirror have been demonstrated by analyzing 805 samples of gastrointestinal tract cancers and experimental validation of selected neopeptide candidates.

Table 2: The neoantigens identified in GIT cancer samples

Sample_ID	Type	Gene	Mutation	Protein position	Neoepitope (mutation)	HLA restriction	ImmuneMirror score
WES_E09039T	ESCC	CREBBP	R/L	1408	LLTAVYHEI	HLA-A*02:01	0.6
TCGA-F4-6570-T	CRC	CTNNA1	E/K	529	HVNPVQALSK	HLA-A*11:01	0.5
TCGA-CA-6719-T	CRC	PCSK7	E/K	357	VTIGAVDEK	HLA-A*11:01	0.7
TCGA-CA-6718-T	CRC	POLE	P/R	286	TTKLPLKFR	HLA-A*33:03	0.7
TCGA-CK-5916-T	CRC	PPP6C	L/R	19	EIARLCKYR	HLA-A*33:01	0.7
TCGA-AY-6197-T	CRC	PRKAR1A	T/M	106	YMEEDAASYV	HLA-A*02:01	0.6
TCGA-CM-5349-T	CRC	RAB35	E/K	94	VVYDVTSAK	HLA-A*03:01	0.9
TCGA-BC-A3KF-T	HCC	STAT3	M/K	28	QLYSDSFPK	HLA-A*03:01	0.8
TCGA-D5-6922-T	CRC	TP53	R/L	213	YLDDRNTFL	HLA-A*02:01	0.5
TCGA-LN-A49Y-T	ESCC	TP53	G/V	245	YMCNSSCMGV	HLA-A*02:06	0.5
WES_E12230T	ESCC	TP53	H/R	179	EVVRRCPHR	HLA-A*33:03	0.7
TCGA-D5-6923-T	CRC	TSC2	E/K	134	KVIKDYPSNK	HLA-A*11:01	0.7

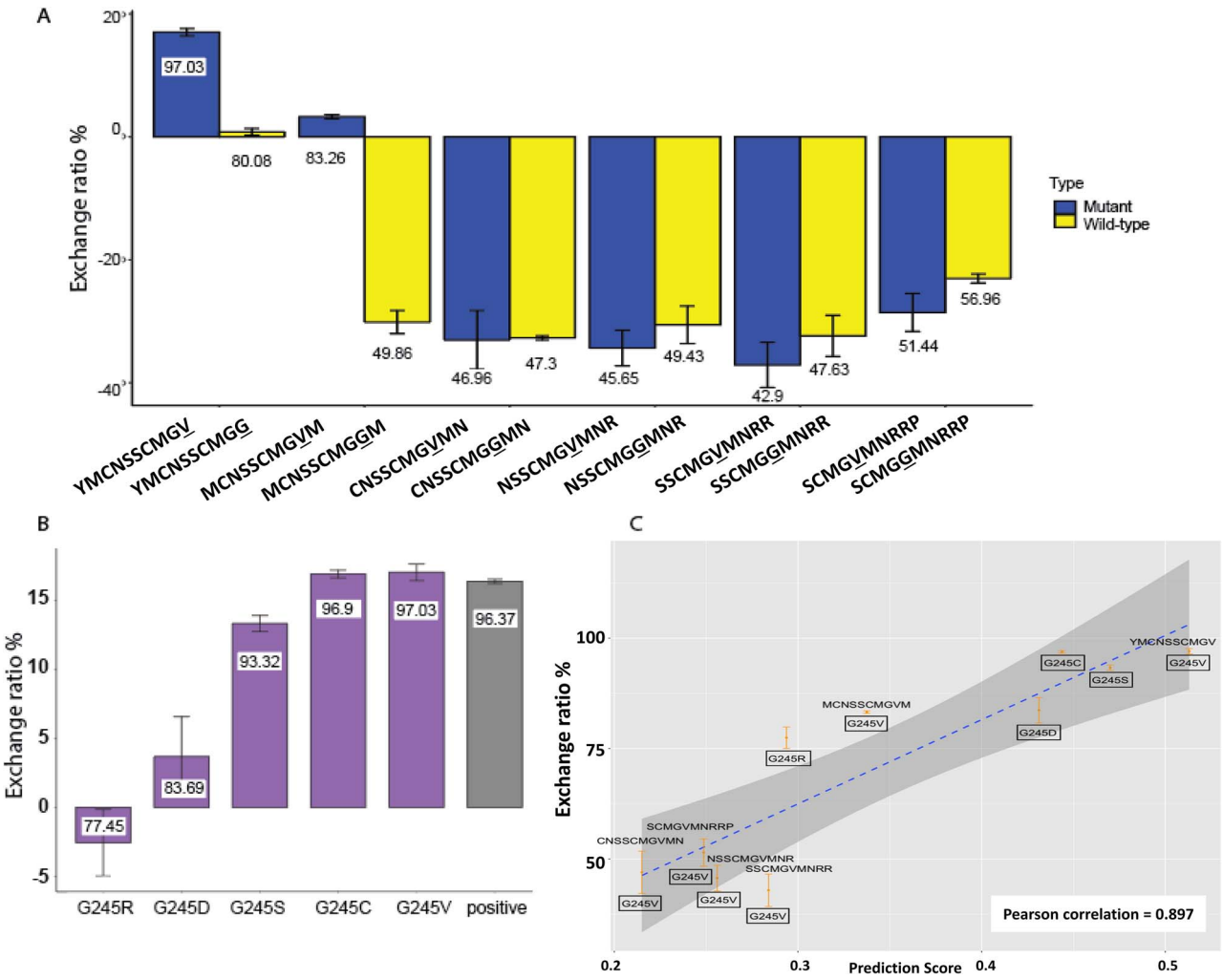

Figure 5. Validation of binding affinity between HLA-A02 and neoepitopes derived from TP53 mutations. (**A**) The exchange ratios (mean ± SD) of the $TP53^{G245V}$ mutant compared with the matched wild type (exchange ratio over 80% is used as the cutoff for positive and negative values). (**B**) The exchange ratio (mean ± SD) of $TP53^{G245}$ mutants compared with the positive control (an exchange ratio of 80% was considered biologically relevant). (**C**) Scatterplot of the prediction score versus the exchange ratio (mean ± SD) for $TP53^{G245}$ mutants.

Both the ImmuneMirror and Wells' study [8] indicate that neopeptides with strong MHC binding affinity, long half-life and low agretopicity are most likely to be neoantigens. From the feature importance plot in Supplementary Figure S6, we further found that agretopicity was the most important feature followed by stability rank and MHC binding affinity score. Moreover, in the ablation study, we fit a logistic regression model to predict the T-cell activity on test set (Supplementary Table S2) using only the MHC binding affinity. It was found that the area under the curve (AUC) of the model using only the MHC binding affinity

was 0.64, while the AUC of our proposed model is 0.87 (Supplementary Figure S7). However, when analyzing the same data set, Wells' criteria [8] tend to be very stringent about binding affinity, agretopicity and peptide stability to accommodate the needs of high specificity for clinical application, while the ImmuneMirror model offers an integrated approach by including more relevant biological features: binding affinity, 'agretopicity' [9, 11] (the ratio of binding affinity between neoantigen and wild-type counterpart), 'foreignness' of nonpolar substances to aggregate in ad aqueous solution and exclude water molecules), binding stability (half-life of peptide-HLA binding complex), peptide processing (proteasomal cleavage sites) and transportation (transporter associated with antigen processing) scores without cutoffs, which provides more potential candidates for downstream experimental validation. Neoantigen generation and T-cell activation is a complex biological process. Moreover, in the comparative analysis of predicting neoantigens between OpenVax [27] and ImmuneMirror, 20 neoantigens (45%) were identified solely by ImmuneMirror (Supplementary Figure S3). This finding indicates the limitations of current computational approaches and suggests that researchers should leverage diverse tools to improve accuracy in novel neoantigen prediction. Therefore, continuous efforts are still needed to collect more experimentally validated neoantigens, which will help better understand these complicated biological processes and select more relevant features for neoantigen prediction.

In our real-world data analysis, we found that neoantigen load was a predictor of good clinical outcomes in ESCC patients. Although it is known that MSI-H is an important molecular biomarker for selecting CRC patients who may benefit from anti-PD-1/PDL-1 therapy [35], we further identified a subgroup of MSI-H CRC patients enriched for advanced T stage that had relatively low neoantigen loads for HLA class I and II by ImmuneMirror. Promising results for immunotherapy have been demonstrated in a previous study that evaluated the efficacy of PD-11 blockade in advanced MSI-H patients across 12 different cancer types with an objective response rate in 53% of patients and complete response in 21% of patients [36]. Nevertheless, almost half of MSI-H cancer patients do not respond well to this treatment. This previous study also showed *in vivo* expansion of T-cell clones specifically activated by neoantigens in patient responses [36]. Our results suggest that further stratification of MSI-H cancer patients based on neoantigen loads may be necessary, and a more detailed evaluation of the objective response rate of this unique subset of MSI-H patients to anti-PD-1/PDL-1 therapy is needed in a clinical trial.

The $TP53^{G245V}$ mutation occurs at a total frequency of 0.13% in diverse cancers, such as diffuse glioma, non-small cell lung cancer, bladder urothelial carcinoma, endometrial carcinoma, head and neck squamous cell carcinoma, pancreatic adenocarcinoma and esophageal squamous cell carcinoma, according to the records in the cBio Cancer Genomic portal [37]. The discovery of the neoepitope $TP53^{G245V}$ (YMCNSSCMGV) derived from this mutation restricted by HLA-A*02, a common HLA class I type in Caucasians and Asians, showed the effectiveness and great potential of ImmuneMirror for detecting neoantigens. In addition to developing the neoantigen vaccine targeting this neoepitope, further identification of T cells that are specifically reactivated by this neoepitope is necessary for developing adoptive T-cell therapies for cancer patients carrying this specific mutation.

In summary, ImmuneMirror is an integrative analysis pipeline for neoantigen prediction and prioritization from a variety of cancer types. This powerful tool could assist biologists to systematically evaluate the genomic and transcriptomic features relevant to the immunotherapy response, including TMB, neoantigen load, MSI status, HLA typing and the expression of the IPRES. More importantly, ImmuneMirror is strategically useful as a guide for clinicians to tailor treatment strategies according to the genomic and transcriptomic profiles for precision medicine, and to facilitate patient stratification to select those who are more likely respond to immunotherapy in clinical trial design. In addition to GIT cancers, further evaluation of this tool in other cancer types could enhance its robustness and versatility and provide broader prospects for clinical applications. Additional experimental and clinical validation of the putative neoantigens identified in this study are warranted to determine their usefulness for more effective immunotherapy.

Key points

- We developed ImmuneMirror, as a stand-alone open-source and device independent (dockerized) pipeline and a web server for neoantigen prediction; source code and curated datasets are freely available to download and use.
- The balanced random forest model is integrated into ImmuneMirror for neoantigen prediction and prioritization; the prediction model was trained and tested using known immunogenic neopeptides collected from 19 published studies.
- We applied ImmuneMirror to the whole-exome sequencing and bulk RNA sequencing data obtained from gastrointestinal tract cancers including 805 tumors from colorectal cancer (CRC), esophageal squamous cell carcinoma (ESCC) and hepatocellular carcinoma (HCC) patients, and made novel discoveries.
- We experimentally validated HLA-A02 binding with TP53.pG245V, which demonstrated the effectiveness and reliability of ImmuneMirror as a robust tool for neoantigen prediction.

ACCESSION CODES

WES data: European Genome-phenome Archive (EGA): EGAS00001000932 [38]; NCBI Sequence Read Archive (SRA): SRP033394 [39], NCBI Bioproject: PRJNA399748 [40]; and TCGA ESCC, CRC and HCC samples from NCI Genomic Data Commons (https://portal.gdc.cancer.gov/). Data from a previous study carried out by Dai *et al.* [41].

RNA-Seq data: TCGA ESCC, CRC and HCC samples from the NCI Genomic Data Commons (https://portal.gdc.cancer.gov/).

SUPPLEMENTARY DATA

Supplementary data are available online.

ACKNOWLEDGEMENTS

The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (IRB) of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (HKU/HA HKW IRB) (protocol code: UW21-421).

FUNDING

This work was supported by the Health Medical Research Fund (07182016) from the Research Fund Secretariat in Hong Kong, Theme-based Research Scheme (TBRS) from Hong Kong Research Grants Council (T12-703/22-R) and the 'Laboratory for Synthetic Chemistry and Chemical Biology' under the Health@InnoHK Program launched by the Innovation and Technology Commission, The Government of Hong Kong Special Administrative Region of the People's Republic of China. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

AUTHOR CONTRIBUTIONS

WD and ZL designed and supervised the study. GSC developed the ImmuneMirror pipeline, built the docker image, performed data analysis, tested the pipeline and web server, and compared ImmuneMirror with other similar pipelines. YG developed the machine learning model and performed the neoantigen analysis for frequent mutations derived from cancer-related genes. GSC and YG designed and developed the ImmuneMirror web server. GSC, YG, WD and ZL wrote the manuscript. CLC and KOL are clinicians who advised on the analysis of clinical specimens. NWK advised on experimental validation for the MHC class I binding assay. All the authors have read and approved the manuscript.

CODE AVAILABILITY

The open source ImmumeMirror pipeline and usage guide are available on GitHub (https://github.com/weidai2/ImmuneMirror) [24]. The source code is released under the GNU General Public License version 3 (GPL $\geq$ 3). The web server is freely available at http://immunemirror.hku.hk/App/ and does not have a login requirement.

DATA AVAILABILITY

The published article includes all data sets generated or analyzed during this study.

REFERENCES

1. Stevanović S, Pasetto A, Helman SR, *et al.* Landscape of immunogenic tumor antigens in successful immunotherapy of virally induced epithelial cancer. *Science* 2017;**356**(6334):200–5.
2. Li S, Simoni Y, Zhuang S, *et al.* Characterization of neoantigen-specific T cells in cancer resistant to immune checkpoint therapies. *Proc Natl Acad Sci U S A* 2021;**118**(30).
3. O'Donnell TJ, Rubinsteyn A, Bonsack M, *et al.* MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst* 2018;**7**(1):129–132.e4.
4. Hoof I, Peters B, Sidney J, *et al.* NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* 2009;**61**(1):1–13.
5. Jurtz V, Paul S, Andreatta M, *et al.* NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol* 2017;**199**(9):3360–8.
6. Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med* 2016;**8**(1):33.
7. Nielsen M, Lund O, NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinform* 2009;**10**:296.
8. Wells DK, van Buuren MM, Dang KK, *et al.* Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. *Cell* 2020;**183**(3):818–834.e13.
9. Ghorani E, Rosenthal R, McGranahan N, *et al.* Differential binding affinity of mutated peptides for MHC class I is a predictor of survival in advanced lung cancer and melanoma. *Ann Oncol* 2018;**29**(1):271–9.
10. Hundal J, Carreno BM, Petti AA, *et al.* pVAC-Seq: a genome-guided in silico approach to identifying tumor neoantigens. *Genome Med* 2016;**8**(1):11.
11. Duan F, Duitama J, al Seesi S, *et al.* Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *J Exp Med* 2014;**211**(11):2231–48.
12. Balachandran VP, Łuksza M, Zhao JN, *et al.* Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature* 2017;**551**(7681):512–6.
13. Łuksza M, Riaz N, Makarov V, *et al.* A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* 2017;**551**(7681):517–20.
14. Richman LP, Vonderheide RH, Rech AJ. Neoantigen dissimilarity to the self-proteome predicts immunogenicity and response to immune checkpoint blockade. *Cell Syst* 2019;**9**(4):375–382.e4.
15. Khnoshgoftaar TM, Golawala M, Hulse JV, An empirical study of learning from imbalanced data using random forest. In: *19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007)*. New York, NY: IEEE US, 2007, pp. 310–7.
16. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. New York: Springer, 2013, 311–2.
17. Schubach M, Re M, Robinson PN, Valentini G. Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. *Sci Rep* 2017;**7**(1):2959.
18. Chawla N, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res (JAIR)* 2002;**16**:321–57.
19. Chen C, Breiman L. *Using Random Forest to Learn Imbalanced Data*. Berkeley: University of California, 2004.
20. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw* 2008;**28**.
21. Hikichi T, Sakamoto M, Harada M, *et al.* Identification of cytotoxic T cells and their T cell receptor sequences targeting COVID-19 using MHC class I-binding peptides. *J Hum Genet* 2022;**67**(7):411–9.
22. Buchli R, VanGundy RS, Hickman-Miller HD, *et al.* Development and validation of a fluorescence polarization-based competitive peptide-binding assay for HLA-A*0201A new tool for epitope discovery. *Biochemistry* 2005;**44**(37):12491–507.
23. Hugo W, Zaretsky JM, Sun L, *et al.* Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell* 2016;**165**(1):35–44.
24. Gulam Sarwar C., Guo Y, Cheung C-L, *et al.* ImmuneMirror: a machine learning-based integrative pipeline and web server for neoantigen prediction. *GitHub* 2023; Available from: https://github.com/weidai2/ImmuneMirror/.
25. McKenna A, Hanna M, Banks E, *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**(9):1297–303.

26. Sharma P, Hu-Lieskovan S, Wargo JA, Ribas A. Primary, adaptive, and acquired resistance to cancer immunotherapy. *Cell* 2017;**168**(4):707–23.
27. Kodysh J, Rubinsteyn A. OpenVax: an open-source computational pipeline for cancer neoantigen prediction. In: Boegel S (ed). *Bioinformatics for Cancer Immunotherapy: Methods and Protocols*. New York, NY: Springer US, 2020, 147–60.
28. Chakravarty D, *et al.* OncoKB: a precision oncology knowledge base. *JCO Precis Oncol* 2017;**2017**.
29. The Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. *Cell* 2015;**161**(7):1681–96.
30. Sanchez-Vega F, Mina M, Armenia J, *et al.* Oncogenic signaling pathways in the cancer genome atlas. *Cell* 2018;**173**(2): 321–337.e10.
31. Wan PT, Garnett MJ, Roe SM, *et al.* Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell* 2004;**116**(6):855–67.
32. Ng PK, Li J, Jeong KJ, *et al.* Systematic functional annotation of somatic mutations in cancer. *Cancer Cell* 2018;**33**(3):450–462.e10.
33. Li Y, Anderson J, Kwan KY, Cai L. Single-cell transcriptome analysis of neural stem cells. *Curr Pharmacol Rep* 2017;**3**(2):68–76.
34. Bonneville R, Krook MA, Chen H-Z, *et al.* Detection of microsatellite instability biomarkers via next-generation sequencing. *Methods Mol Biol* 2020;**2055**:119–32.
35. Overman MJ, McDermott R, Leach JL, *et al.* Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): an open-label, multicentre, phase 2 study. *Lancet Oncol* 2017;**18**(9): 1182–91.
36. Le DT DJN, Smith KN, Wang H, *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* 2017;**357**(6349):409–13.
37. Cerami E, Gao J, Dogrusoz U, *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;**2**(5):401–4.
38. Gao YB, Chen ZL, Li JG, *et al.* Genetic landscape of esophageal squamous cell carcinoma. *Nat Genet* 2014;**46**(10): 1097–102.
39. Lin DC, Hao JJ, Nagata Y, *et al.* Genomic and molecular characterization of esophageal squamous cell carcinoma. *Nat Genet* 2014;**46**(5):467–73.
40. Deng J, Chen H, Zhou D, *et al.* Comparative genomic analysis of esophageal squamous cell carcinoma between Asian and Caucasian patient populations. *Nat Commun* 2017; **8**(1):1533.
41. Dai W, Ko JMY, Choi SSA, *et al.* Whole-exome sequencing reveals critical genes underlying metastasis in oesophageal squamous cell carcinoma. *J Pathol* 2017;**242**(4):500–10.