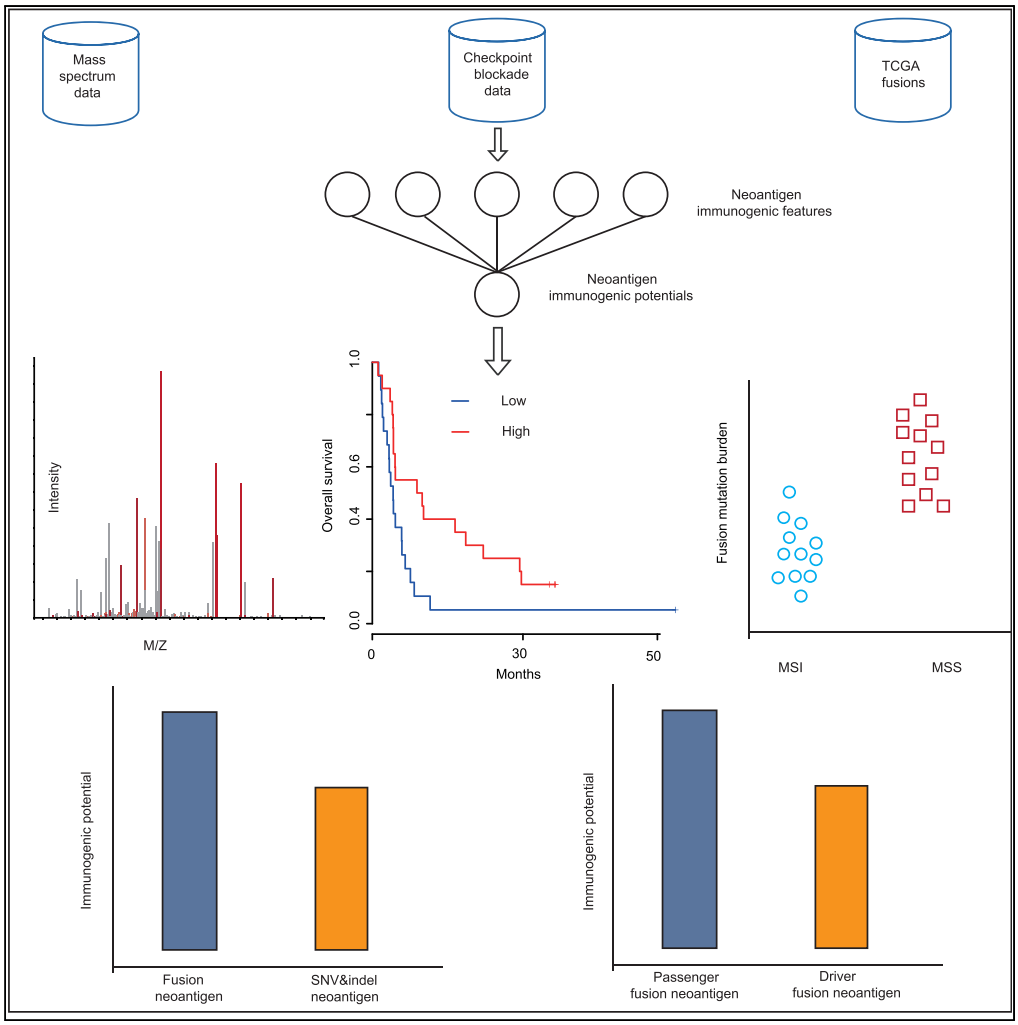
iScience

CellPress

Article

The Landscape of Tumor Fusion Neoantigens: A Pan-Cancer Analysis



Zhiting Wei, Chi Zhou, Zhanbing Zhang, Ming Guan, Chao Zhang, Zhongmin Liu, Qi Liu

zhangchao@tongji.edu.cn (C.Z.)
liu.zhongmin@tongji.edu.cn (Z.L.)
qiliu@tongji.edu.cn (Q.L.)

HIGHLIGHTS

A score scheme is presented to evaluate the immunogenicity of fusion neoantigen

Tumor fusion neoantigen burden is not related to the immunotherapy outcome

Compared with SNV&indel neoantigen, fusion neoantigen has higher immunogenicity

Oncogene fusion has lower immunogenicity than passenger gene fusion

Wei et al., iScience 21, 249–260
November 22, 2019
https://doi.org/10.1016/j.isci.2019.10.028

iScience

CellPress

Article

The Landscape of Tumor Fusion Neoantigens: A Pan-Cancer Analysis

Zhiting Wei,[1] Chi Zhou,[1] Zhanbing Zhang,[1] Ming Guan,[3] Chao Zhang,[1,*] Zhongmin Liu,[2,*] and Qi Liu[1,4,*]

SUMMARY

Compared with SNV&indel-based neoantigens, fusion-based neoantigens are not well characterized. In the present study, we performed a comprehensive analysis of the landscape of tumor fusion neoantigens in cancer and proposed a score scheme to quantitatively assess their immunogenic potentials. By analyzing three large-scale tumor datasets, we demonstrated that (1) the tumor fusion candidate neoantigen burden is not related to the immunotherapy outcome; (2) fusion neoantigens tend to have notably higher immunogenic potentials than SNV&indel-based candidate neoantigens, making them better candidates for cancer vaccines; (3) fusion candidate neoantigens distribute sparsely between individual patients. Although several recurrent candidate neoantigens exist, they usually have extremely low immunogenic potentials, suggesting that vaccination-based cancer immunotherapy must be personalized; (4) compared with fusion mutations involving tumor passenger genes, fusion mutations involving oncogenic genes have remarkably low immunogenic potentials, indicating that they undergo selection pressure during tumorigenesis.

INTRODUCTION

Vaccination therapy to fight cancer by boosting the response of the human immune system to cancer cells is a highly promising treatment strategy. Neoantigens are peptides generated from somatically mutated genes that play an important role in vaccination-based cancer immunotherapy (Liu and Mardis, 2017; Vitiello and Zanetti, 2017). It has been reported that predicted neoantigen load is strongly correlated with the clinical response to immunotherapy. Fusion is a hybrid gene formed from two previously separate genes, and it has long been known to play an important role in tumorigenesis (Mertens et al., 2015). Because of the ability to create new open reading frame (ORF) and produce plentiful neo-peptides, neoantigens can be generated from fusions (Yang et al., 2019). Current studies of neoantigen sources mainly focus on single nucleotide variants (SNV) and small insertions and deletions (indel), whereas fusion-based neoantigens across different cancers are not yet well characterized (Zhou et al., 2019). Therefore, in the present study, we comprehensively characterized the landscape of fusion candidate neoantigens presented by major histocompatibility complex class I (MHC I) in three cohort datasets, i.e., the dataset of cancer cell lines with MHC I mass spectrum (MS), the dataset of sequencing data from immune checkpoint blockade (ICB) trials, and The Cancer Genome Atlas dataset (TCGA). Specifically, we characterized the tumor fusion neoantigens by taking the T-cell receptor (TCR) recognition mechanisms into consideration. TCR is a molecule found on the surface of T cells, or lymphocytes, that is responsible for recognizing fragments of antigen as peptides bound to MHC molecules. To activate a T-cell response, peptide-MHC complex (pMHC) must be recognized by T-cell receptors. However, in the process of mature T-cell generation, negative selection mechanism removes T cells that are capable of strongly binding with self-peptides. As a result, the TCR repertoire has intrinsic biases in their generation probabilities (Murugan et al., 2012). Furthermore, due to the T-cell cross-reactivity (the ability of the T cell to recognize more than one pMHC) and negative selection, predicted neoantigens may differ greatly in their immunogenic potentials. Therefore, how to quantitatively and unbiasedly evaluate the neoantigen immunogenic potentials remains a challenge issue. We hereby proposed a score scheme to evaluate fusion neoantigen immunogenic potentials, taking two factors, i.e., the likelihood of peptide presentation by MHC (Bjerregaard et al., 2017) and pMHC subsequent recognition by T cells (Łuksza et al., 2017), into consideration. Through applying our score scheme to the MS, ICB, and TCGA cohort datasets, several findings are presented, providing useful clues for personalized cancer vaccination-based immunotherapy.

[1]Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, Shanghai Key Laboratory of Signaling and Disease Research, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China

[2]Department of Cardiac Surgery, Shanghai East Hospital, Tongji University, Shanghai 200092, China

[3]Department of Laboratory Medicine, Huashan Hospital, Fudan University, Shanghai 200433, China

[4]Lead Contact

*Correspondence: zhangchao@tongji.edu.cn (C.Z.), liu.zhongmin@tongji.edu.cn (Z.L.), qiliu@tongji.edu.cn (Q.L.)

https://doi.org/10.1016/j.isci.2019.10.028

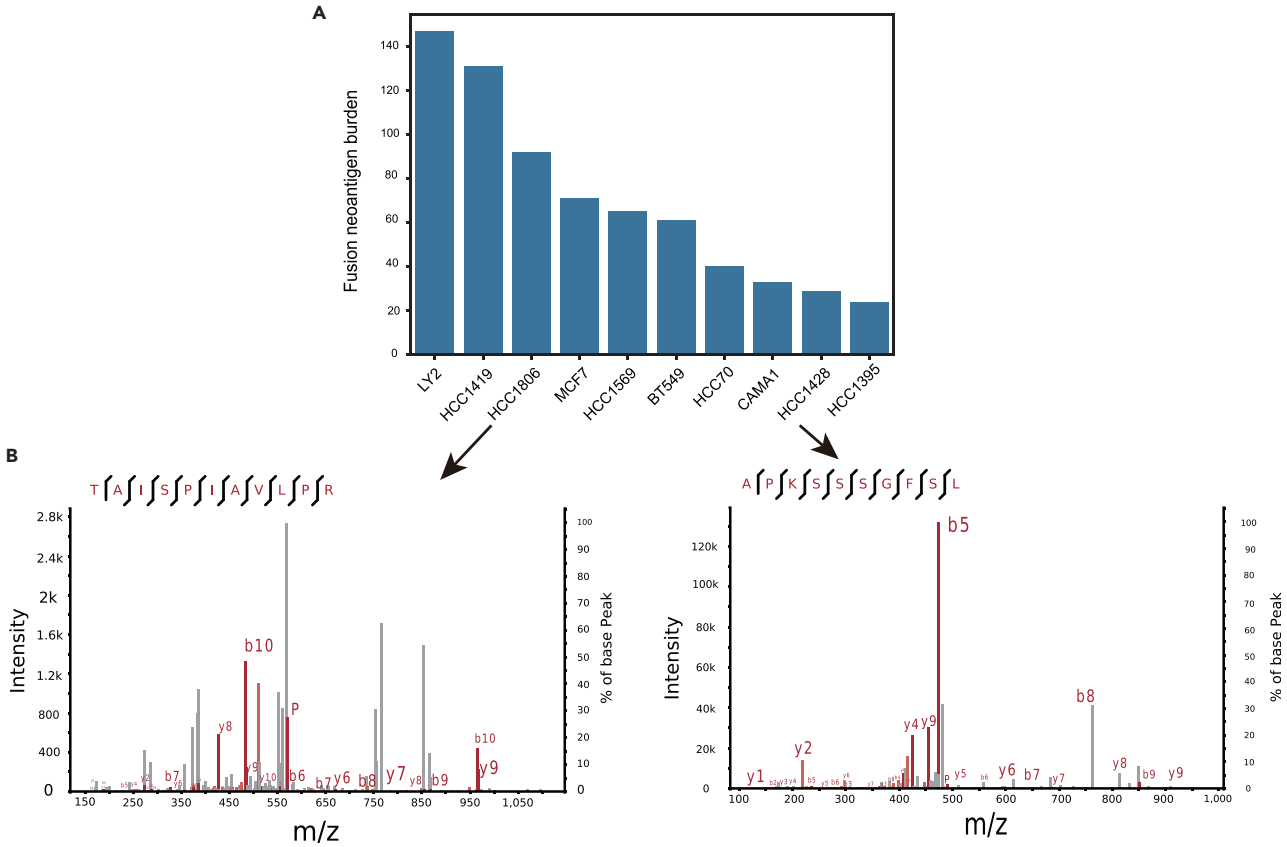


Figure 1. Analysis of the MS Cohort Dataset

(A) Ten breast cancer cell line RNA sequencing data were analyzed according to our computational pipeline *neoFusion*. Two fusion neo-peptides, i.e., TAISPIAVLPR produced by OTUB1-CDC20 in HCC1806 cell line and APKSSSGFSL produced by MTSS1L-RPS15A in HCC1428 cell line, were experimentally validated to be presented by MHC I using the MS data.

(B) Peptide-spectrum matches with *q*-value < 0.01 were extracted by *mzR* R package and then visualized by *xiSPEC* software.

See also Table S1.

RESULTS

Analysis of the MS Cohort Dataset

Fusion Peptides Can Be Processed and Presented on the Tumor Cell Surface

High resolution mass-spectrometry enables the identification and quantification of MHC ligands that are naturally processed and presented. To demonstrate that gene fusion is able to generate neo-peptides that can be presented by the MHC I, fusion candidate neoantigens in 10 human breast cancer cell lines were predicted following our computational workflow (Methods). In our study, mutation burden is defined as the total number of somatic mutations detected in a tumor sample and neoantigen burden is defined as the total number of neoantigens produced by those mutations. The tumor fusion candidate neoantigen burden varies from 24 to 147 in cancer cell lines, with a median value of 63 (Figure 1A). Two fusion candidate neoantigens, i.e., TAISPIAVLPR produced by OTUB1-CDC20 (by CDC20 frameshift transcript) in HCC1806 cell line and APKSSSGFSL produced by MTSS1L-RPS15A (by RPS15A frameshift transcript) in HCC1428 cell line, were discovered in complex with MHC I via mass spectrometry with high confidences (Figure 1B, Percolator *q*-value<0.01, Table S1). These results provide direct evidences of the processing and presentation of fusion peptides through the MHC I, which were also recently experimentally demonstrated by Yang (Yang et al., 2019).

Fusion candidate neoantigens were further prioritized according to their score as defined in the Methods. It should be noted that when scoring those predicted fusion neoantigens, only the likelihood of peptide presented by MHC I (Methods) was calculated, as those peptides are eluted from pMHC complexes. The

fusion candidate neoantigens TAISPIAVLPR in HCC1806 cell line (92 fusion candidate neoantigens in total) and APKSSSGFSL in HCC1428 cell line (29 fusion candidate neoantigens in total) rank 6/92 and 2/29, respectively, suggesting that candidate neoantigens with higher ranks are presented by the MHC I with high priorities (p value<0.05, Methods).

Analysis of the ICB Cohort Dataset

Overview of the Landscape of Fusion Candidate Neoantigens in the ICB Dataset

Two melanoma ICB cohorts with whole-exome and RNA sequencing data were analyzed according to our computational workflow (Methods). In the Van Allen cohort, the total number of fusions per sample varied from 0 to 25, with a median value of 5. In the Hugo cohort, the total number of fusions per sample varied from 0 to 9, with a median value of 1. In these two cohorts, tumor fusion candidate neoantigens were notably lower than SNV&indel candidate neoantigens, in terms of both burden and score (Figure 2A, Table S2).

The Tumor Fusion Candidate Neoantigen Burden Is Not Associated with the Immunotherapy Outcome

Given that the tumor SNV&indel neoantigen burden closely correlates with the response to checkpoint inhibitors (Van Allen et al., 2015; Hugo et al., 2017; Lauss et al., 2017), we next examined whether the tumor fusion candidate neoantigen burden is similarly associated with the immunotherapy response. The tumor microenvironment, including the surrounding immune cells and fibroblasts, should be taken into consideration in predicting the immunotherapy outcome (Church and Galon, 2015). According to Rooney et al., cytolytic activity is a biomarker of the immune response (Rooney et al., 2015). Similarly, Balachandran et al. reported that tumors having both the highest neoantigen burden and the most abundant CD8+ T-cell infiltrates, but not either alone, stratified patients with the longest survival (Balachandran et al., 2017). In other words, high-quality neoantigens and sufficient T cells are needed simultaneously to elicit a T-cell response. To this end, the cytotoxic lymphocyte score (CTL, Methods) was incorporated to amplify the tumor candidate neoantigen score (score*CTL, Methods) for survival analysis. In both cohorts, the tumor fusion candidate neoantigen burden, neoantigen score, and the tumor fusion candidate neoantigen score*CTL were not associated with the checkpoint inhibitor response (Figures S1A, S1B, and 2B), suggesting that the tumor fusion candidate neoantigen burden is not a biomarker for the immunotherapy outcome.

*The Overall Tumor Candidate Neoantigen Score*CTL Is Associated with the Immunotherapy Outcome*

Since the response to checkpoint inhibitors is associated with the neoantigen burden, we reasoned that neoantigens generated by all mutations should be taken into account. Therefore, we summed the SNV&indel candidate neoantigen burden and fusion candidate neoantigen burden to obtain the overall tumor candidate neoantigen burden. Similarly, we summed the SNV&indel candidate neoantigen score and fusion candidate neoantigen score to obtain the overall tumor candidate neoantigen score. In the Van Allen cohort, the overall tumor neoantigen burden, score, and CTL, respectively were not related to the immunotherapy outcome (Figures S1C, S2A, and S2B). Survival, however, was significantly improved in patients with a higher overall tumor candidate neoantigen score*CTL (Figure 2C, log rank p value = 0.021) and overall tumor candidate neoantigen burden*CTL (Figure S2C, log rank p value = 0.032) in the Van Allen cohort. In the Hugo cohort, the overall tumor candidate neoantigen burden*CTL and CTL respectively were not related to the immunotherapy outcome (Figures S2C and S1C). Survival was significantly improved in patients with a higher overall tumor candidate neoantigen burden, score, and higher overall tumor candidate neoantigen score*CTL (Figures S2A, S2B, and 2C, log rank p value<0.05) in the Hugo cohort. Taking together, all the metrics except the overall tumor candidate neoantigen score*CTL have their limitations in immunotherapy response prediction in these two cohorts, indicating the rationality and effectiveness of our proposed score scheme.

It can be seen that in these two ICB cohorts, tumor fusion candidate neoantigen burden and score were notably lower than tumor SNV&indel candidate neoantigen burden and score respectively; as a result, adding the fusion candidate neoantigen burden and score to the overall tumor candidate neoantigen burden and score does not affect the p value in predicting the immunotherapy outcome.

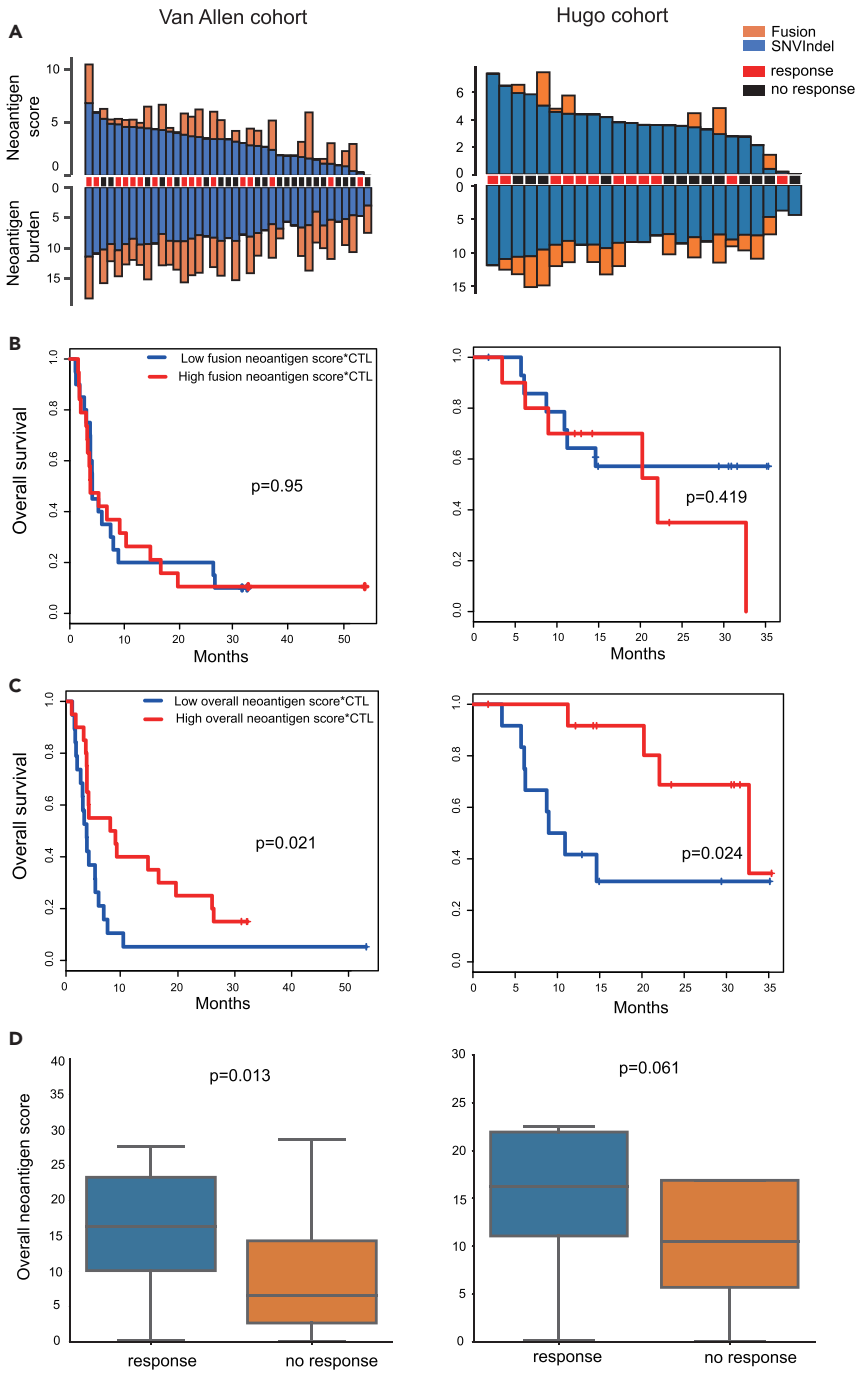


Figure 2. Analysis of the ICB Cohort Dataset

(A) The overview of the tumor candidate neoantigen burden and score in the two ICB melanoma cohorts. Y axis values were log2 transformed.

(B) The tumor fusion candidate neoantigen score*CTL could not separate patients in both cohorts.

(C) The overall tumor candidate neoantigen score*CTL significantly separated patients in both cohorts. Samples were split by the median value cutoff in C and D. See also Figures S1 and S2 and Table S2.

(D) In the Van Allen cohort, the overall tumor candidate neoantigen score of the response group was significantly higher than that of the no response group. In the Hugo cohort, there was no difference between the response group and no response group. Boxplots show the first, median, and third quartiles, and whiskers extend to 1.5X the interquartile range.

Moreover, we examined the overall tumor candidate neoantigen score*CTL, together with age and sex, in a multivariable model. Our results indicated that the overall tumor candidate neoantigen score*CTL was associated with the response to checkpoint inhibitors, independent of age and sex (Figure S1D, Van Allen cohort: hazard ratio [HR] 0.44, 95% confidence interval [CI] 0.22–0.89, log rank p value = 0.022; Hugo cohort: HR 0.24, 95% CI 0.064–0.92, log rank p value = 0.038).

It should be noted that patients in these two cohorts were previously stratified into two groups according to the RECIST criteria (Methods). In the Van Allen cohort, the overall tumor candidate neoantigen score of the response group was significantly higher than the no response group (Figure 2D, p value = 0.013, Mann-Whitney U test). In the Hugo cohort, however, there was no significant difference between the response and no response group (Figure 2D, p value = 0.061, Mann-Whitney U test), possibly due to the small sample size.

Incorporating Fitness Score into Our Score Scheme Boosts the Predictive Power in the Survival Analysis

Previously, Luksza et al. (Łuksza et al., 2017) presented a *fitness model* to calculate the likelihood of pMHC recognized by T-cell receptors (*fitness score*, Methods) by alignment neo-peptides with a set of epitopes retrieved from IEDB. To activate a T-cell response, pMHC must be recognized by T-cell receptors. Therefore, we evaluated whether incorporating the *fitness score* into our score scheme could boost the predictive power in the survival analysis. Our results showed that taking the *fitness score* into consideration significantly separated patients in both cohorts in predicting the immunotherapy outcome (Figure 2C). Although the patients could not be separated in the Hugo cohort without the incorporation of the *fitness score* (Figure S2D), indicating that *fitness model* does boost predictive power in the survival analysis.

Analysis of the TCGA Cohort Dataset

Overview of the Landscape of Fusion Candidate Neoantigens in the TCGA Dataset

In total, there are 67,502 predicted fusion candidate neoantigens among the 6552 TCGA samples. The most common number of mismatches between the candidate neoantigen and the corresponding wild type peptide is 3, and frameshift fusion produces up to 145 candidate neoantigens. In our study, the tumor SNV&indel candidate neoantigen burden strongly correlated with the tumor SNV&indel mutation burden (pearson $R = 0.89$, p value$<2.2 \times 10^{-16}$). As for fusion, the correlation was slightly weaker (pearson $R = 0.74$, p value$<2.2 \times 10^{-16}$). Similar to the tumor candidate SNV&indel neoantigen burden (Thorsson et al., 2018), the overall tumor candidate neoantigen score*CTL is not a prognostic factor for overall survival (Figure S3A) except for TCGA BLCA. Previously, based on the mechanism of T-cell central tolerance, Turajlic defined an SNV&indel candidate neoantigen with a half maximal inhibitory concentration (IC50) of less than 50 nM and the corresponding wild-type peptide with an IC50 greater than 50 nM as the **specific candidate neoantigen** (Turajlic et al., 2017). Because of self-immune tolerance, compared with nonspecific candidate neoantigens, specific candidate neoantigens tend to have higher immunogenic potentials. Following this definition, we applied this idea to fusion candidate neoantigens. Because we used binding affinity percent rank metric to filter peptides in the present study, a neoantigen with binding affinity percent rank ≤ 2 and a corresponding wild-type peptide with a binding affinity percent rank >2 was defined as a **specific candidate neoantigen** (Jurtz et al., 2017). For different cancer types, the fusion-specific candidate neoantigen burden per sample varied from 0 to 205, with a median value of 0. In our study, fusion mutation burden is the total number of fusions detected in a sample. The fusion mutation burden per sample varied from 0 to 55, with a median value of 1. The fusion candidate neoantigen burden per sample varied from 0 to 360, with a median value of 0. Breast invasive carcinoma (BRCA) was the cancer type with the highest fusion mutation burden per sample. Kidney chromophobe (KICH), kidney renal cell carcinoma (KIRC), and kidney renal papillary cell carcinoma (KIRP) were the three cancer types with the lowest fusion mutation burden, fusion candidate neoantigen burden, and fusion-specific candidate neoantigen burden, each with a median of 0 (Figure 3A).

Fusion Candidate Neoantigens Tend to Have Higher Immunogenic Potentials Than SNV&Indel Candidate Neoantigens

For TCGA cohort data, the fusion mutation burden, candidate neoantigen burden, and specific candidate neoantigen burden were notably lower than the SNV&indel mutation, candidate neoantigen, and specific

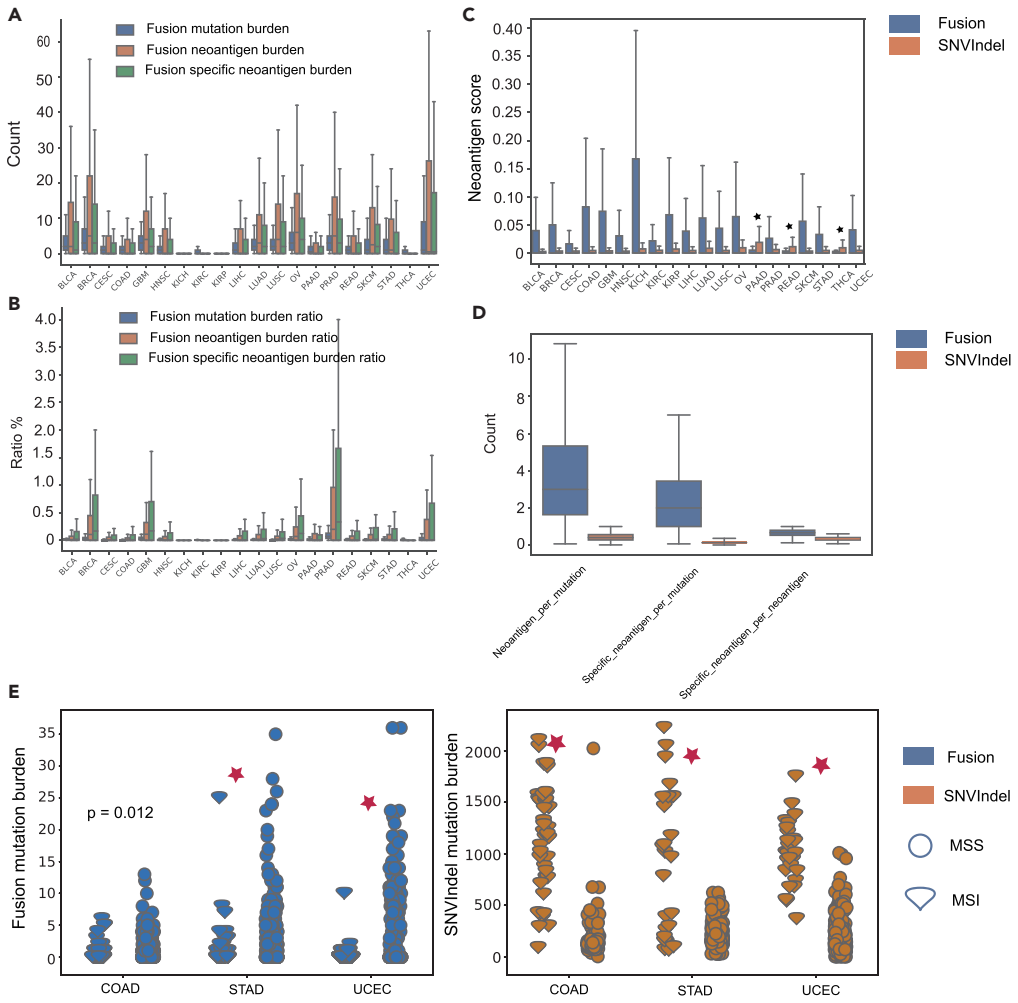


Figure 3. Analysis of the TCGA Cohort Dataset

(A) The landscape of the fusion mutation burden, fusion candidate neoantigen burden, and fusion-specific candidate neoantigen burden in the 20 cancer types.

(B) The landscape of the fusion mutation burden, fusion candidate neoantigen burden, and fusion-specific candidate neoantigen burden ratio in the 20 cancer types.

(C) The overview of the fusion and SNV&indel candidate neoantigen scores in the 20 cancer types. Except for PAAD, READ, and THCA, the fusion candidate neoantigen scores were significantly higher than that of the SNV&indel candidate neoantigen scores. Black star indicates p value>0.01.

(D) A fusion mutation was able to generate much more and high-quality candidate neoantigens than an SNV&indel mutation.

(E) In contrast to the SNV&indel mutation burden, the fusion mutation burden was significantly higher in microsatellite stable tumors.

Boxplots show the first, median, and third quartiles, and whiskers extend to 1.5X the interquartile range. Outlier points are not shown. See also Tables S3 and S4.

candidate neoantigen burden in all cancer types (Figure 3B). Except for pancreatic adenocarcinoma (PAAD), rectum adenocarcinoma (READ), and thyroid carcinoma (THCA), the score of fusion putative neoantigen was significantly higher than the score of SNV&indel candidate neoantigen in all cancer types (Figure 3C, Mann-Whitney U test, p value<0.01). Compared with SNV&indel, a fusion is able to produce more putative neoantigens (Figure 3D, p value<0.01, Mann-Whitney U test). Intriguingly, candidate neoantigens generated by fusion are much more likely to be specific candidate neoantigens (Figure 3D). Of 6,552 TCGA tumor samples, 3,161 simultaneously harbored tumor fusion and SNV&indel putative neoantigens. In 1,018 of 3,161 tumor samples, the putative neoantigen with highest immunogenic potentials were generated by

fusion. Given that the mean fusion candidate neoantigen burden per sample was notably lower than the SNV&indel candidate neoantigen burden, this result further supports the fact that the fusion candidate neoantigens have significantly higher immunogenic potentials (p value<0.001, binomial test).

The Fusion Mutation Burden Is Significantly Higher in Microsatellite Stable Tumors

Microsatellite instability (MSI), a pattern of hypermutation that occurs at genomic microsatellites, is caused by defects in the mismatch repair system. The US Food and Drug Administration approved the high MSI phenotype as a biomarker for immunotherapy (Le et al., 2015). MSI is highly positively correlated with the tumor SNV&indel mutation burden (Bonneville et al., 2017). In this study, we further investigated the relationship between MSI and the tumor fusion mutation burden in the TCGA cohort data. The proportion of the MSI sample, as measured by the MANTIS score (Reeser et al., 2016), varied substantially across 20 cancer types. Because only colon adenocarcinoma (COAD), stomach adenocarcinoma (STAD), and uterine corpus endometrial carcinoma (UCEC) had sufficient MSI samples for analysis (samples $\geq$15), we focused on these three cancer types. Consistent with previous studies, MSI tumors had a significantly higher SNV&indel mutation burden (Figure 3E). In contrast to the SNV&indel mutation burden, however, the fusion mutation burden was notably lower in MSI tumors (STAD, UCEC p value<0.01, COAD p value = 0.012, Mann-Whitney U test). Furthermore, we investigated the relationship between the category of fusion and the status of microsatellite. Fusion mutations were separated into two categories with respect to the gene involved, i.e., driver gene fusion and passenger gene fusion (Gao et al., 2018; Table S3). The proportion of MSS tumor harboring driver fusion is higher than that of MSI tumor (STAD, COAD 3.2% vs 1.2%, UCEC 5% vs 0.6%). The driver fusion mutation burden was higher in MSS tumor (STAD, UCEC p value<0.01, Mann-Whitney U test). These may be explained by the fact that as a result of mismatch repair system deficiency, MSI tumors harbored notably more SNV&indel mutations, and these tumors are primarily driven by SNV&indel mutations (Vaish and Mittal, 2002). By contrast, because fewer SNV&indel mutations exist, microsatellite stable (MSS) tumor cells are likely to rely on other mechanism to gain a growth advantage such as producing driver fusion mutations.

Frameshift Fusion Candidate Neoantigens Tend to Have Higher Immunogenic Potentials Than Inframe Fusion Candidate Neoantigens

In our study, fusions were also separated into three categories with respect to the frame of the 3′ gene, i.e., noframe fusions, inframe fusions, and frameshift fusions. Among the 25,664 TCGA fusions, there were 9,284 noframe fusions, 7,738 frameshift fusions, and 8,642 inframe fusions, respectively. Not surprisingly, frameshift fusions can produce more candidate neoantigens due to the ability to create new ORF (Figure 4A, p value<0.01, Mann-Whitney U test). On average, a frameshift fusion generates 6 candidate neoantigens, and an inframe fusion generates 2.46 candidate neoantigens. Moreover, frameshift fusion candidate neoantigens tend to have higher immunogenic potentials than inframe fusion candidate neoantigens (Figure 4B, p value<0.01, Mann-Whitney U test). It should be noted that compared with inframe fusion neoantigens, frameshift neoantigens could increase the nonsense-mediated decay (NMD) mechanism, thereby decreasing its own immunogenic potential. The main function of NMD is to reduce errors in gene expression by eliminating mRNA transcripts that contain premature stop codons. NMD will potentially reduce the expression level of frameshift fusion transcripts and thus their immunogenic potential. Our results indicated that compared with samples without frameshift fusion mutation, NMD activity in samples harboring frameshift fusion is slightly higher (Methods). Estimating NMD efficiency and taking expression level into consideration when evaluating fusion peptides immunogenic potentials should make our score scheme and conclusion more reliable. However, this factor was not incorporated in our score scheme in our present study, because the expression information of fusion genes is unavailable.

Furthermore, we examined the relationship between the fusion frame's status and the fusion's category. The TCGA fusions were separated into four categories with respect to the gene involved, i.e., oncogene (Onco) fusion, tumor suppressor gene (TSG) fusion, kinase fusion, and passenger gene fusion (Gao et al., 2018; Table S3). In total, 2,104 kinase fusions, 522 Onco fusions, 436 TSG fusions, and 23,115 passenger fusions were observed. Onco fusions and kinase fusions are more likely to be inframe than those of passenger fusions, as preserving the ORF is required to keep their oncogenic function (Figure 4C, p value<0.01, chi-squared test). TSG fusion during creating a new ORF will reduce or lose its function, which leads to tumorigenesis. In other words, TSG fusion is not required to maintain their function during tumorigenesis. Therefore, the inframe ratio does not differ between TSG fusion and passenger fusion.

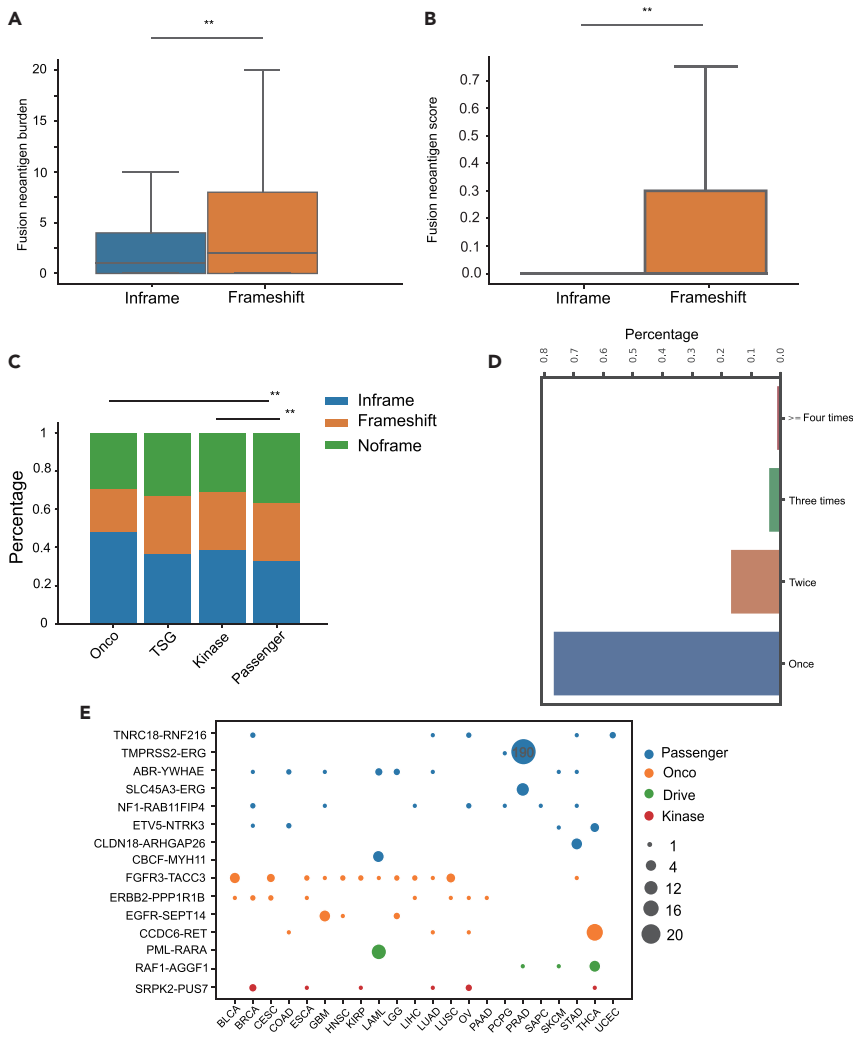


Figure 4. Analysis of the TCGA Cohort Dataset

(A) A frameshift fusion was able to generate much more candidate neoantigens than that of an inframe fusion.

(B) Frameshift fusion candidate neoantigens have notably higher immunogenic potentials. ** indicates p value<=0.01 and * indicates p value<=0.05.

(C) The inframe ratio of Onco and kinase fusion were significantly higher than that of passenger fusion. Onco, oncogenic; TSG, tumor suppressor gene.

(D) Candidate fusion neoantigens were extremely sparse.

(E) Several fusions occurred in most cancer types while several fusions only occurred in specific cancer types. Of 498 PRAD tumor samples, 190 harbored the TMPRSS2-ERG fusion. Only 15 most recurrent fusions were displayed. Size in the plot corresponds to the number of samples harbor that fusion.

Boxplots show the first, median and third quartiles and whiskers extend to 1.5X the interquartile range. Outlier points are not shown. See also Tables S3 and S4.

Onco Fusion Mutations Tend to Have Lower Immunogenic Potentials Than Passenger Fusion Mutations

Immunoediting, a dynamic process comprising immunosurveillance and tumor progression, describes the relation between tumor cells and the immune system (Schreiber et al., 2011). Although the immune system exerts negative selective pressure on tumors, it also helps to sculpt the tumor genotype. Mutations in the driver gene, while conferring cells a selective growth advantage, render cells vulnerable to the immune system as a result of generating neoantigens. In addition, driver mutations are necessary in the development of cancers. As a consequence, driver mutations detected in tumors should be biased toward with lower immunogenic potentials (Marty et al., 2018; Sun et al., 2017). The fusion score defined as the sum of

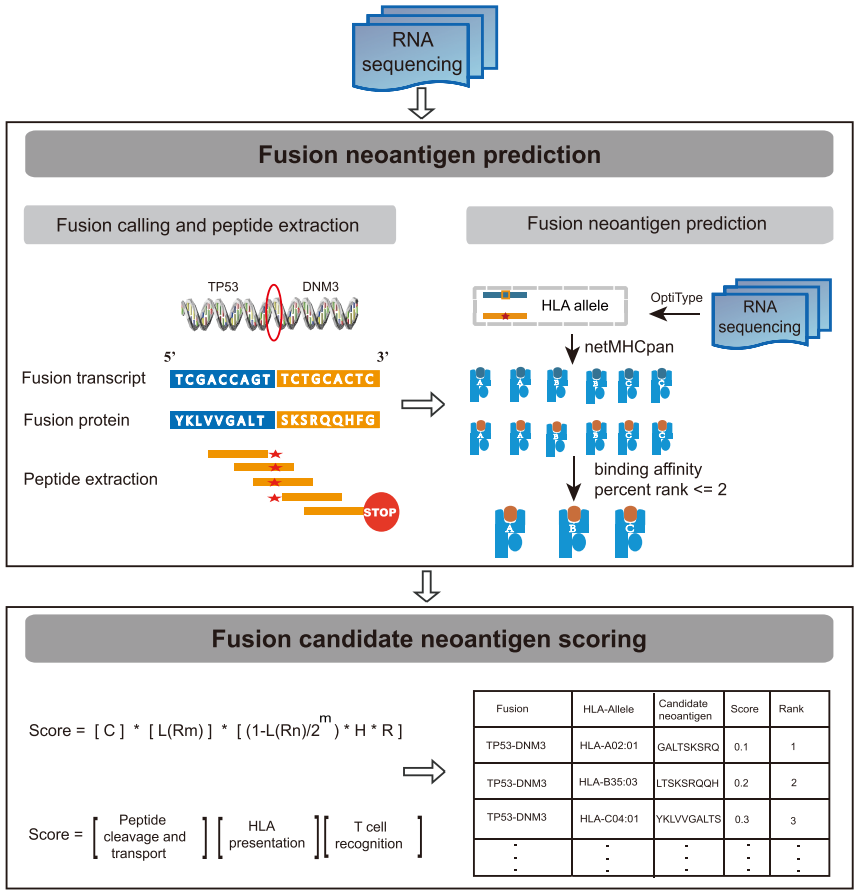


Figure 5. *neoFusion* Workflow Overview

Gene fusions were detected by *STAR-Fusion* with RNA sequencing data. Translated fusion proteins output by *STAR-Fusion* were chopped up into 9–11 kmers peptides until a stop codon. The pMHC binding affinity and binding affinity percent rank were determined by *NetMHCpan* in binding affinity mode. Peptides with binding affinity percent rank ≤ 2 were reported as candidate neoantigens. Fusion candidate neoantigens were scored according to our score scheme and ranked according to their scores. See also Figure S4.

candidate neoantigen scores generated by that fusion (Methods) showed that the Onco fusion score was significantly lower than the passenger fusion score but not others (Methods).

Highly Recurrent Candidate Neoantigens Tend to Have Extremely Low Immunogenic Potentials

Only 5.8% fusion candidate neoantigens in the TCGA cohort data were shared between patients (Figure 4D). We found that highly recurrent candidate neoantigens have extremely low immunogenic potentials. For example, the neoantigen score of KMALNSEAL, a candidate neoantigen generated by TMPRSS2-ERG, which presents in 38% PRAD, only ranks at the 92nd percentile (Figure 4E). The low immunogenic potential of highly recurrent fusion candidate neoantigens clearly suggests that neoantigen-based cancer vaccination immunotherapy must be personalized.

DISCUSSION

Fusion, which is an important class of somatic mutations, is an ideal source of tumor-derived neoantigens for creating an ORF. Compared with SNV&indel-based neoantigens, however, fusion-based neoantigens are not well characterized. A comprehensive literature review indicated that *INTEGRATE-neo* is the only existing in silico tool for fusion neoantigen prediction; however, it cannot assess their immunogenic potentials, which can be substantially different due to a single nucleotide mismatch (Bjerregaard et al., 2017). In this study, we propose an effective tool *neoFusion* for fusion neoantigen identification (Figure 5).

Furthermore, a rational score scheme to quantitatively assess the identified fusion neoantigen immunogenic potentials is presented (Figure 5).

By analyzing the ICB cohort dataset, we found that (1) neither the tumor fusion candidate neoantigen burden nor the tumor fusion candidate neoantigen score*CTL was associated with the immunotherapy outcome in the two melanoma ICB cohorts, indicating that the tumor fusion candidate neoantigen burden may not be a predictive biomarker for the immunotherapy response; (2) in the Van Allen cohort, only the overall tumor candidate neoantigen score*CTL and burden*CTL significantly separated patients. In the Hugo cohort, only the overall tumor candidate neoantigen score*CTL, score, and burden separated the patients. Taking together, all the metrics except the overall tumor candidate neoantigen score*CTL have their limitations in immunotherapy response prediction in these two cohorts, indicating the rationality and effectiveness of our score scheme; (3) so far, a higher PD-1 or PD-L1 expression (Garon et al., 2015), a higher neoantigen load, the microsatellite instability, and a higher peripheral baseline TCR diversity (Postow et al., 2015) are all reported to be associated with a better immunotherapy outcome; therefore, prediction of the response to immunotherapy is still an open question and a comprehensive model to accurately predict patient response is still lacking, likely requiring much more data to train and refine. We believe that the neoantigen score, tumor microenvironment such as CTL score, and other types of neoantigens besides SNV&indel based should be taken into consideration in predicting the immunotherapy outcome; (4) in these two ICB cohorts, tumor fusion candidate neoantigen burden and score were notably lower than tumor SNV&indel candidate neoantigen burden and score, respectively; therefore, adding them to the overall tumor candidate neoantigen burden and score does not improve the prediction accuracy of immunotherapy response. However, recently Yang et al. identified a patient exhibited complete response to anti-PD1 immunotherapy despite a low SNV&indel mutation burden and demonstrated that the patient elicited a T-cell response to neoantigen generated by fusion (Yang et al., 2019). Therefore, in certain cancer types such as BLCA, since the tumor fusion candidate neoantigens contribute to a relatively high proportion of the overall tumor candidate neoantigens, taking tumor fusion candidate neoantigens into consideration may improve the prediction of immunotherapy outcome.

Through comparing the TCGA fusion candidate neoantigens with the TCGA SNV&indel candidate neoantigens, we presented the following findings: (1) fusion, which is able to create novel ORF, generate 6-fold more candidate neoantigens and 11-fold more specific candidate neoantigens as SNV&indel. Compared with the SNV&indel candidate neoantigen burden, the fusion candidate neoantigen burden per sample was notably lower. Nevertheless, fusion candidate neoantigens tend to have notably higher immunogenic potentials. In 32.2% TCGA patients, candidate neoantigens with the highest immunogenic potentials were produced by fusion, making fusion neoantigens a better source for cancer vaccines; (2) similar to the SNV&indel candidate neoantigen burden, the fusion candidate neoantigen burden strongly correlated with the fusion mutation load. Furthermore, both types of candidate neoantigens were extremely sparse. Although several recurrent fusion candidate neoantigens exist, they usually have extremely low immunogenic potentials, further indicating that cancer vaccination strategies based on neoantigens must be personalized (Schreiber et al., 2011). To be recurrent, mutations must confer tumor cells a selective advantage. Producing neo-peptides that do not attract the attention of the human immune system confer such an advantage. Therefore, those highly recurrent fusion peptides such as KMALNSEAL in PRAD usually have low immunogenic potentials.

The comparison between passenger fusion mutations and other types of fusion mutations indicated that (1) Onco fusion mutations tend to have lower immunogenic potentials than passenger fusion mutations. Onco fusion mutations, while conferring cells a selective growth advantage, render cells vulnerable to the immune system as a result of generating neoantigens. Cancer cells that harbor Onco fusion mutations poorly bound to MHC are thus positively selected during tumorigenesis. As tumor cells grow and activate mechanisms to evade the immune system, passenger mutations are acquired regardless of their affinities to the MHC complex (Marty et al., 2018). Therefore, Onco fusion mutations tend to have lower immunogenic potentials than passenger fusion mutations; (2) similar to Onco fusion mutations, TSG fusion mutations should have lower immunogenic potentials than passenger fusion mutations. However, the immunogenicity score did not differ between passenger fusion mutations and TSG fusion mutations. These may be explained by the fact that, in contrast to Onco fusions, TSG fusions tend to be under-expressed and thus insufficient to generate a T-cell response (Gao et al., 2018). In conclusion, neoantigens produced by Onco and TSG fusion mutations are less likely to induce a T-cell response, and passenger fusion neoantigens may have particular relevance for vaccine.

Limitations of the Study

Our study presents the first comprehensive profile of fusion neoantigens from a pan-cancer perspective, which provides useful clues for personalized cancer vaccination-based immunotherapy. Several limitations should be noted: (1) 8 kmer and 12 kmer and above peptides can also be displayed by MHC I; however, in the present study, only the most common 9–11 kmer peptides were considered; (2) the fusion expression level factor was not incorporated in our score scheme in the present study, because such information was absent; however, knowledge accumulated in immunotherapy community will make the accurate and objective evaluation of the peptides immunogenic potential feasible in the future.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

QUANTIFICATION AND STATISTICAL ANALYSIS

Survival analysis was performed using the Kaplan-Meier method provided by survival R (version 3.4.4) packages. The log rank test and Cox proportional hazard model were used to assess the correlation between metrics and overall survival. We used a one-sided nonparametric Mann-Whitney U test for non-normally-distributed variables to assess the difference in mean or median for a continuous variable between two groups. All statistical analyses were performed with the Python 3 SciPy, and NumPy libraries. The parameters of software used in our study were set as default without explicitly stated.

DATA AND CODE AVAILABILITY

neoFusion is available at https://github.com/bm2-lab/neoFusion, with a Docker version at https://hub.docker.com/r/bm2lab/neoFusion/.

The mass spectrum data: ten breast cancer cell line RNA-sequencing data were downloaded fromSequence Read Archive (NCBI: SRP026537). The corresponding MS proteomics data were downloaded from ProteomeXchange Consortium (proteomecentral.proteomexchange.org, PXD006406).

The immune checkpoint blockade cohort data: two cohorts of melanoma datasets were downloaded from the database of Genotypes and Phenotypes (dbGaP: phs000452.v2.p1) and SRA (NCBI: SRP070710), respectively. The overall survival and progression-free survival data and other data needed were retrieved from the original article supplementary.

The TCGA cohort data: TCGA fusion, oncogene, kinase gene, and tumor suppressor gene lists were retrieved from Gao et al. (Gao et al., 2018). The TCGA whole-exome sequencing (WES) VCFs and corresponding expression profile files were downloaded from TCGA website. HLA allele information was requested from The Cancer Immunome Atlas (https://tcia.at/home; Thorsson et al., 2018). The landscape of microsatellite instability of TCGA tumor samples were obtained from Bonneville et al.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.10.028.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (Grant No. 2017YFC0908500, 2016YFC1303205), National Natural Science Foundation of China (Grant No. 61572361), Shanghai Rising-Star Program (Grant No. 16QA1403900), Shanghai Municipal Health Commission Innovative integration for molecular oncology (Grant No. 2019CXJQ03) and Shanghai Natural Science Foundation Program (Grant No. 17ZR1449400).

AUTHOR CONTRIBUTIONS

Q.L., Z.M.L., and C.Z. conceived the study. Z.T.W., C.Z., Z.B.Z., and M.G. analyzed the tumor sample data. Z.T.W., Q.L., Z.M.L., and C.Z. wrote the manuscript with assistance from other authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 25, 2019
Revised: September 6, 2019
Accepted: October 15, 2019
Published: November 22, 2019

REFERENCES

Balachandran, V.P., Łuksza, M., Zhao, J.N., Makarov, V., Moral, J.A., Remark, R., Herbst, B., Askan, G., Bhanot, U., Senbabaoglu, Y., et al. (2017). Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. Nature *551*, S12–S16.

Bjerregaard, A.M., Nielsen, M., Hadrup, S.R., Szallasi, Z., and Eklund, A.C. (2017). MuPeXI: prediction of neo-epitopes from tumor sequencing data. Cancer Immunol. Immunother. *66*, 1123–1130.

Bonneville, R., Krook, M.A., Kautto, E.A., Miya, J., Wing, M.R., Chen, H.Z., Reeser, J.W., Yu, L., and Roychowdhury, S. (2017). Landscape of microsatellite instability across 39 cancer types. JCO Precision Oncol. *2017*, 1–15, https://doi.org/10.1200/po.17.00073.

Church, S.E., and Galon, J. (2015). Tumor Microenvironment and Immunotherapy: the whole picture is better than a glimpse. Immunity *43*, 631–633.

Gao, Q., Liang, W.W., Foltz, S.M., Mutharasu, G., Jayasinghe, R.G., Cao, S., Liao, W.W., Reynolds, S.M., Wyczalkowski, M.A., Yao, L., et al. (2018). Driver fusions and their implications in the development and treatment of human cancers. Cell Rep. *23*, 227–238.e3.

Garon, E.B., Rizvi, N.A., Hui, R., Leighl, N., Balmanoukian, A.S., Eder, J.P., Patnaik, A., Aggarwal, C., Gubens, M., Horn, L., et al. (2015). Pembrolizumab for the treatment of non-small-cell lung cancer. N. Engl. J. Med. *372*, 2018–2028.

Hugo, W., Zaretsky, J.M., Sun, L., Song, C., Moreno, B.H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., et al. (2017). Genomic and transcripomic features of response to anti-PD-1 therapy in metastatic melanoma. Cell *168*, 542.

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. J. Immunol. *199*, 3360–3368.

Lauss, M., Donia, M., Harbst, K., Andersen, R., Mitra, S., Rosengren, F., Salim, M., Vallon-Christersson, J., Törngren, T., Kvist, A., et al. (2017). Mutational and putative neoantigen load predict clinical benefit of adoptive T cell therapy in melanoma. Nat. Commun. *8*, 1–10.

Le, D.T., et al. (2015). PD-1 blockade in tumors with mismatch-repair deficiency. N. Engl. J. Med. *372*, 2509–2520.

Liu, X.S., and Mardis, E.R. (2017). Applications of immunogenomics to cancer. Cell *168*, 600–612.

Marty, R., Thompson, W.K., Salem, R.M., Font-Burgada, J., Zanetti, M., and Carter, H. (2018). Evolutionary pressure against MHC class II binding cancer mutations. Cell *175*, 416–428.e13.

Mertens, F., Johansson, B., Fioretos, T., and Mitelman, F. (2015). The emerging complexity of gene fusions in cancer. Nat. Rev. Cancer *15*, 371–381.

Murugan, A., et al. (2012). Statistical inference of the generation probability of T-cell receptors from sequence repertoires. Proc. Natl. Acad. Sci. U S A *109*, 16161–16166, https://doi.org/10.1073/pnas.1212755109.

Postow, M.A., Manuel, M., Wong, P., Yuan, J., Dong, Z., Liu, C., Perez, S., Tanneau, I., Noel, M., Courtier, A., et al. (2015). Peripheral T cell receptor diversity is associated with clinical outcomes following ipilimumab treatment in metastatic melanoma. J. ImmunoTherapy Cancer *3*, 23.

Reeser, J.W., et al. (2016). Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. Oncotarget *8*, 7452–7463, https://doi.org/10.18632/oncotarget.13918.

Rooney, M.S., Shukla, S.A., Wu, C.J., Getz, G., and Hacohen, N. (2015). Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell *160*, 48–61.

Schreiber, R.D., Old, L.J., and Smyth, M.J. (2011). Cancer immunoediting: integrating immunitys roles in cancer suppression and promotion. Science *331*, 1565–1570.

Sun, Z., Chen, F., Meng, F., Wei, J., and Liu, B. (2017). MHC class II restricted neoantigen: a promising target in tumor immunotherapy. Cancer Lett. *392*, 17–25.

Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.H., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., et al. (2018). The immune landscape of cancer. Immunity *48*, 812–830.e14.

Turajlic, S., Litchfield, K., Xu, H., Rosenthal, R., McGranahan, N., Reading, J.L., Wong, Y.N.S., Rowan, A., Kanu, N., Al Bakir, M., et al. (2017). Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol. *18*, 1009–1021.

Vaish, M., and Mittal, B. (2002). DNA mismatch repair, microsatellite instability and cancer. Indian J. Exp. Biol. *40*, 989–994.

Van Allen, E.M., Miao, D., Schilling, B., Shukla, S.A., Blank, C., Zimmer, L., Sucker, A., Hillen, U., Foppen, M.H.G., Goldinger, S.M., et al. (2015). Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science *350*, https://doi.org/10.1126/science.aad0095.

Vitiello, A., and Zanetti, M. (2017). Neoantigen prediction and the need for validation. Nat. Biotechnol. *35*, 815.

Yang, W., Lee, K.W., Srivastava, R.M., Kuo, F., Krishna, C., Chowell, D., Makarov, V., Hoen, D., Dalin, M.G., Wexler, L., et al. (2019). Immunogenic neoantigens derived from gene fusions stimulate T cell responses. Nat. Med. *25*, 767–775.

Zhou, C., Zhu, C., and Liu, Q. (2019). Toward in silico identification of tumor neoantigens in immunotherapy. Trends Mol. Med. https://doi.org/10.1016/j.molmed.2019.08.001.

Łuksza, M., Riaz, N., Makarov, V., Balachandran, V.P., Hellmann, M.D., Solovyov, A., Rizvi, N.A., Merghoub, T., Levine, A.J., Chan, T.A., et al. (2017). A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. Nature *551*, 517–520.

ISCI, Volume *21*

Supplemental Information

The Landscape of Tumor Fusion Neoantigens: A Pan-Cancer Analysis

Zhiting Wei, Chi Zhou, Zhanbing Zhang, Ming Guan, Chao Zhang, Zhongmin Liu, and Qi Liu

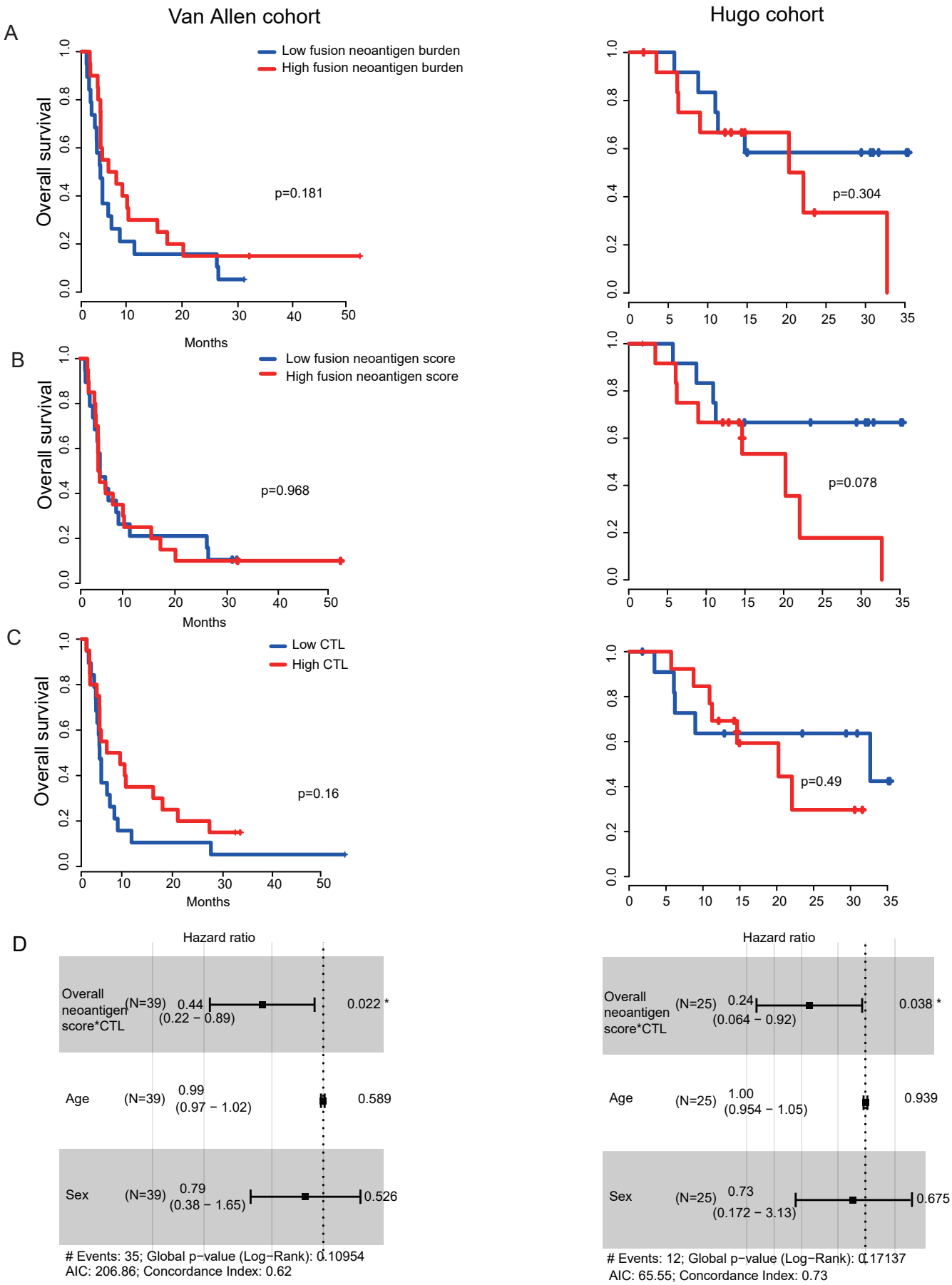


Figure S1. Survival analysis with different metrics. Related to Figure 2.
(A) The tumor fusion candidate neoantigen burden could not separate patients in both cohorts.
(B) The tumor fusion candidate neoantigen score could not separate patients in both cohorts.
(C) The CTL is not related to immunotherapy outcome in both cohorts.
(D) Multivariate cox regression showed that the overall tumor candidate neoantigen score*CTL was associate with checkpoint inhibitors outcome, independent of age and sex. Hazard ratio with 95% confidence interval was shown for overall neoantigen score*CTL, Age and Sex.

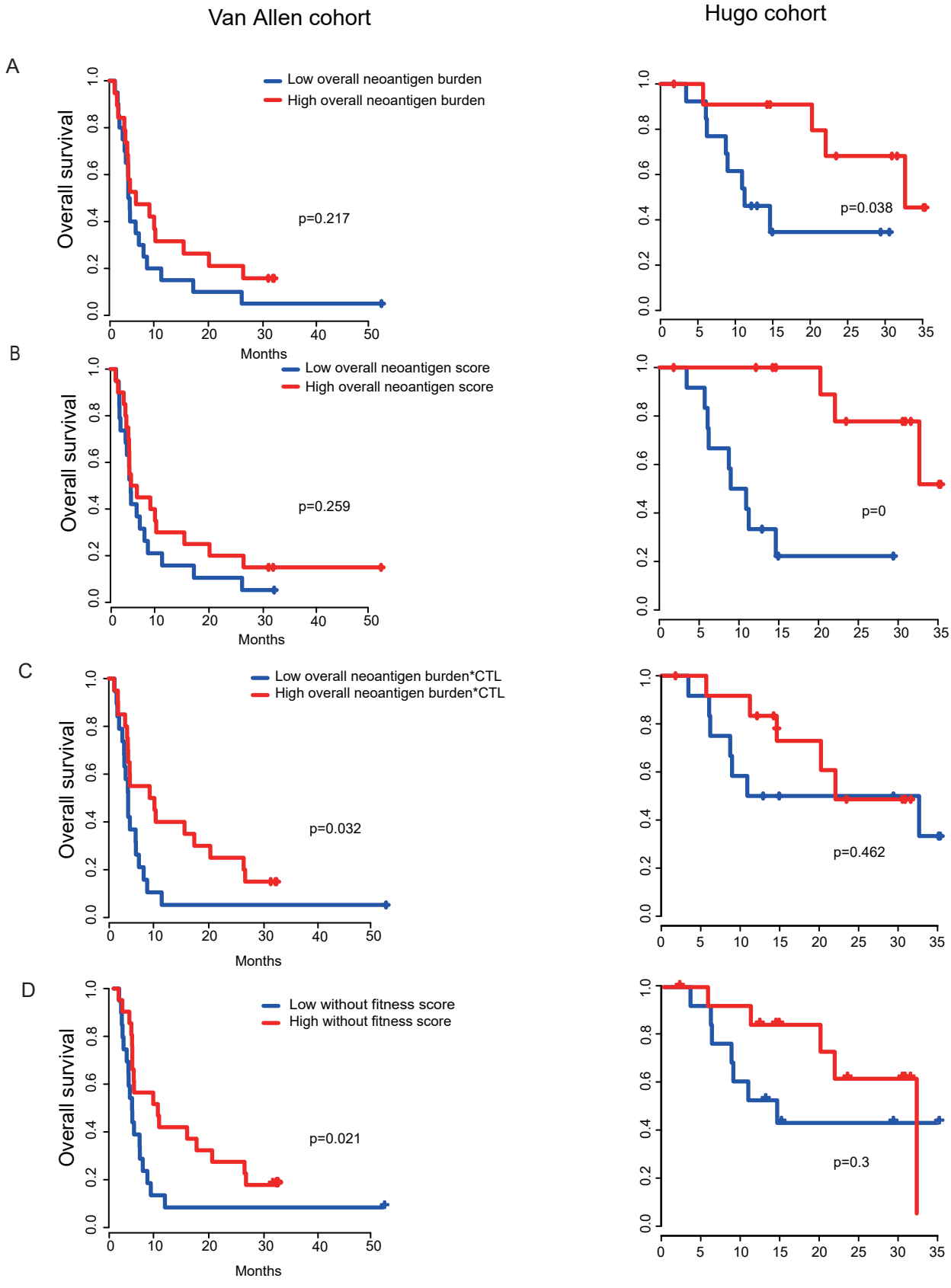


Figure S2. Survival analyses with different metrics. Related to Figure 2.
(A) The overall tumor neoantigen burden is related to immunotherapy outcome in the Hugo cohort, while not in the Van Allen cohort.
(B) The overall tumor neoantigen score is related to immunotherapy outcome in the Hugo cohort, while not in the Van Allen cohort.
(C) The overall neoantigen burden*CTL is related to immunotherapy outcome in theVan Allen cohort, while not in the Hugo cohort.
(D) Incorporating fitness score in our score scheme improves the accuracy in immunotherapy outcome prediction.

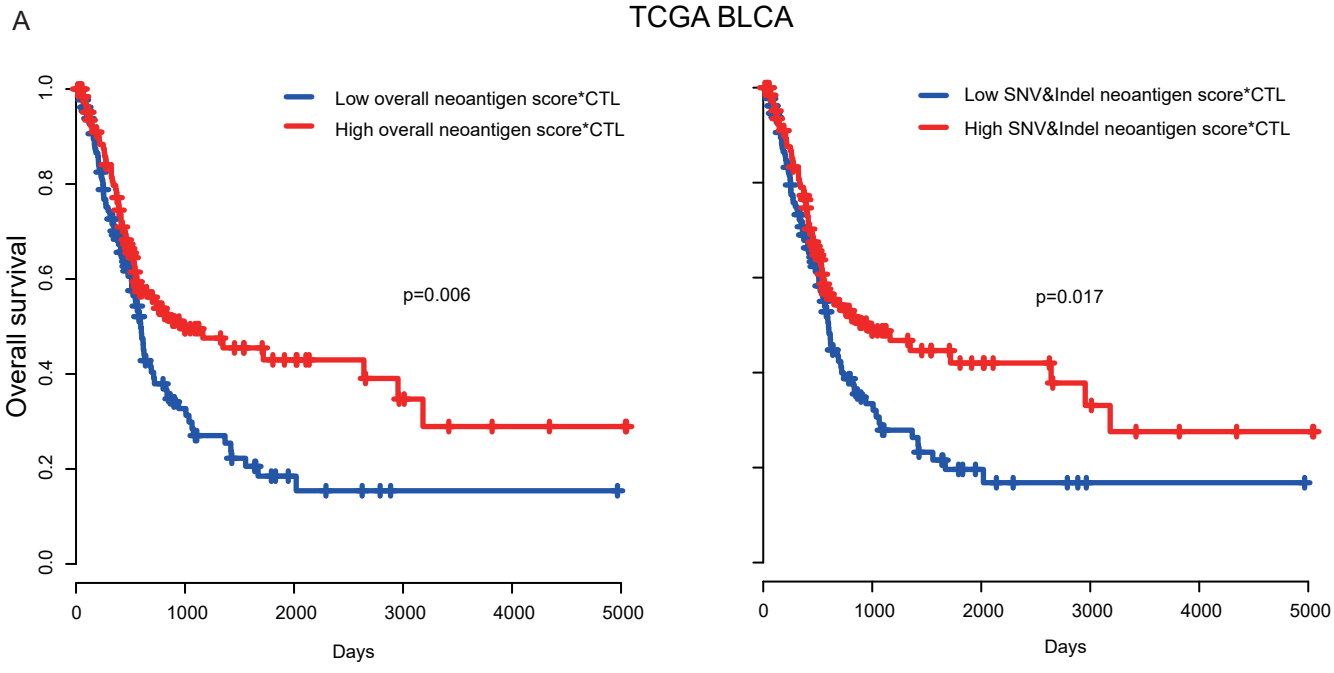


Figure S3. The overall tumor candidate neoantigen score*CTL is not a prognostic factor for overall survival. Related to Figure 2.

(A) The overall tumor candidate neoantigen score*CTL is not a prognostic factor for overall survival except for TCGA BLCA (20 cancer types were tested). Taking fusion candidate neoantigens into consideration improves the prediction accuracy of overall survival.

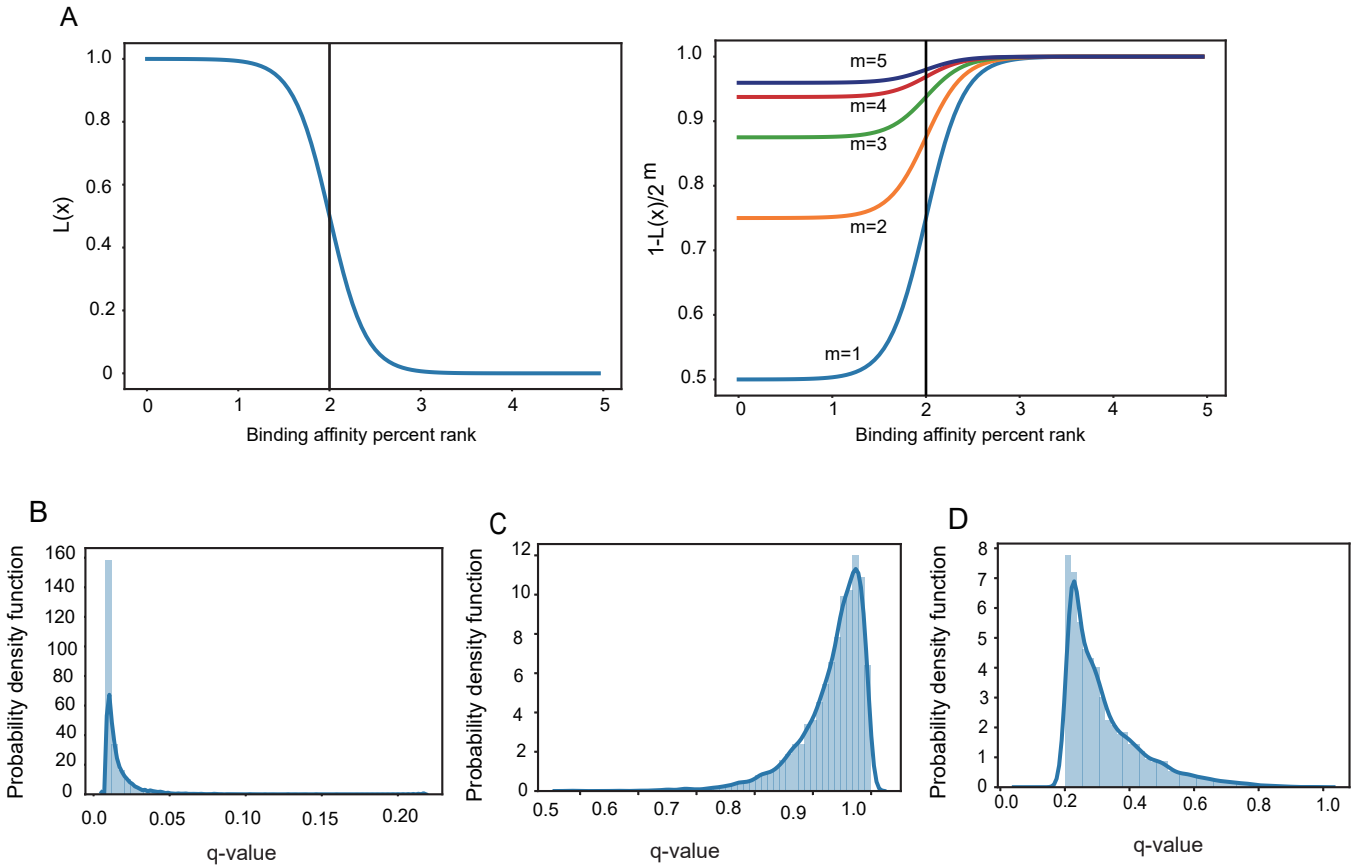


Figure S4. The negative logistic function and the distribution of q-values. Related to Figure 4 and 5.
(A) The negative logistic function of L(x), and m indicates the mismatch between the candidate neoantigen and the corresponding normal peptide.
(B) The distribution of q-values of the comparison between passenger fusion and Onco fusion.
(C) The distribution of q-vaues of the comparison between passenger fusion and TSG fusion.
(D) The distribution of q-values of the comparison between passenger fusion and kinase fusion.
TSG: tumor suppressor gene; Onco: oncogene

Transparent Methods

neoFusion for fusion neoantigen prediction and immunogenic potential assessment

A comprehensive literature review indicated that *INTEGRATE-neo* (Zhang, Mardis and Maher, 2017) is the only existing in silico tool for fusion neoantigen prediction. Several issues, however, remain to be overcome (1) In constructing peptides, *INTEGRATE-neo* only considered peptides spanning the fusion breakpoints. However, frameshift fusions can create new ORF, as a result all the downstream translated protein sequence alter and fusion candidate neoantigens may be missed by the *INTEGRATE-neo*; (2) *INTEGRATE-neo* do not assess the immunogenic potential of fusion candidate neoantigens. Here, we present *neoFusion*, a pipeline for fusion neoantigen prediction and prioritization with a quantitative score schema (**Figure 5**). Tools used in our fusion neoantigen prediction pipeline such as *STAR-Fusion* are based on literature survey and the state-of-the-art tools are chosen. For convenience, end users can take fusions or neo-peptides detected by themselves as input to *neoFusion* to assess neo-peptides immunogenic potentials. *neoFusion* outputs list of putative neoantigens generated by fusions and prioritizes these candidate neoantigens based on their immunogenicity scores. *neoFusion* is written by python, and it can be easily installed and deployed with Docker version. *neoFusion* comprises the following four main steps: data preprocessing, fusion detection and filtering, fusion neoantigen prediction, and fusion candidate neoantigen scoring and ranking (**Figure 5**).

Data preprocessing: Illumina adaptors, low quality (phred score below 20) and N bases of raw RNA sequencing data are removed by *Trimmomatic-0.36* (Bolger, Lohse and Usadel, 2014). Although single-end sequencing data is supported by *neoFusion*, paired-end data are highly recommended.

Fusion detection and filtering: Several bioinformatics methods and software have been developed to identify fusion transcripts from RNA-Seq. In our pipeline we employ *STAR-Fusion* to detect fusions as *STAR-Fusion* show a higher sensitivity in detecting the fusions reporting in previous TCGA studied (Gao *et al.*, 2018). Fastq files are mapped to the human reference genome (build hg38) followed by fusion calling using *STAR-Fusion* (parameters: --examine_coding_effect; Haas *et al.*, 2017). Fusions having FFPM less than 0.1 (fusion fragments per million total reads) or not supporting by LargeAnchor reads are filtered. Furthermore, fusions reported in normal samples were filtered, including the ones from GTEx tissues (The Genotype-Tissue Expression project) and non-cancer cell study. Fusions were separated into three categories with respect to the frame of the 3' gene, i.e., noframe fusions (breakpoint at UTR, intron or non-coding RNA. Those noframe fusions are not an obvious fusion protein based on the reference coding region annotations and they are filtered to reduce false positives in predicting fusion neoantigens; Haas *et al.*, 2017; Kim and Zhou, 2018), inframe fusions (fusion do not create transcript frameshift) and frameshift fusions.

Fusion neoantigen prediction: For each predicted fusion, we obtained the translated protein sequence output by *STAR-Fusion* and constructed 9-11 kmers (default parameter) peptides. Peptides existing in the human reference proteome were not likely to be neoantigens and they were filtered to reduce false positives. HLA alleles were determined (unless provided) from RNA sequencing data by *OptiType* (Szolek *et al.*, 2014), which, with the default setting, achieved ~97% accuracy. pMHC binding affinity and binding affinity percent rank were predicted by *NetMHCpan* version 4.0 (Jurtz *et al.*, 2017) in binding affinity mode with other parameters set as default. Peptides with binding affinity percent rank <=2 are reported as candidate neoantigens (Nielsen and Andreatta, 2016); The binding affinity percent rank was used for filtering as the authors of *NetMHCpan* demonstrated that different MHC molecules present epitopes at distinct binding thresholds. Specifically, for example, set 500nM binding affinity threshold to filter peptides that would not be presented by HLA-A02:02 is fine, however this threshold maybe not suitable for HLA-B07:02. Therefore, binding affinity percent rank was proposed for peptides filtering and proven to be more accurate: for each allele, *NetMHCpan* translated the predicted binding affinity values to a percentile score by comparing them to the predicted binding affinities of a set of 400000 random natural peptides.

Fusion candidate neoantigen scoring and ranking: We quantitatively assessed the immunogenic potential of candidate neoantigens by their candidate neoantigen scores and prioritized candidate neoantigens according to their scores. We aimed to prioritize neo-peptides that are likely to be presented by MHC I on the cell surface and recognized by T cells.

Candidate neoantigen score scheme

The following features were used to construct our candidate neoantigen score scheme based on our previous work (Zhou *et al*., 2019).

C: Combined score of binding affinity, proteasomal C' terminal cleavage, and TAP transport efficiency, as output by *NetCTLpan* (Stranzl *et al.*, 2010). One of the first steps involved in MHC I neoantigen presentation is the degradation of intracellular proteins by the proteasome. Only a subset of the peptides is transported by transporter associated with TAP complex into the endoplasmatic reticulum.

Rm: The binding affinity percent rank of the candidate neoantigen, as output by *NetMHCpan* 4.0.

Rn: The binding affinity percent rank of the candidate neoantigen corresponding wild type peptide. The wild type peptide, a single peptide as long as and most similar to the candidate neoantigen with up to 4 mismatches in the human reference proteome, was determined by *pepmatch_db_x86_64* program with default parameter (Bjerregaard *et al.*, 2017).

m: Mismatch between candidate neoantigen and the corresponding wild type peptide.

H: The hydrophobicity of amino acids at the TCR contact residues is a strong hallmark of CD8+ T cell-mediated immunity (Chowell *et al.*, 2015). In our previous work, three eXtreme Gradient Boosting (*XGBoost*) machine-learning models were trained to

predict the probability of pMHC recognized by T cells (Zhou *et al*., 2019). Briefly, immunogenic peptides (pMHCs with a T cell response) and non-immunogenic peptides (pMHCs without a T cell response) were collected from the Immune Epitope Database and Analysis Resource (Vita *et al.*, 2009). Then, the hydrophobicity of amino acid was used as the input feature to train the model.

***R* (*fitness score*)**: Recently, several methods measuring the T cell recognition probability of pMHC were proposed based on sequence comparison analysis. Here we used the neoantigen *fitness model* presented by Luksza et al. to calculate the T cell *fitness score* (Luksza *et al.*, 2017). Briefly, the model gives *R*, the likelihood that a neoantigen will be recognized by the TCR repertoire, by alignment with a set of peptides retrieved from IEDB. These peptides are linear epitopes from human infectious diseases that are positively recognized by T cells after class I MHC presentation. The model assumed that a neoantigen is more likely to be immunogenic if the neoantigen is more similar to those peptides. *R* was defined by a multistate thermodynamic model in which sequence similarity was treated as a proxy for binding energy. To assess the sequence similarity between a neoantigen with peptide sequence *s* and an IEDB epitope *e*, gapless alignment with a BLOSUM62 amino acid similarity matrix was computed and their alignment scores denoted as $|s,e|$. For a given neoantigen with peptide sequence *s*, the T cell recognition score was calculated as:

$$R = Z(k)^{-1} \sum_{e \in IEDB} exp\big(-k(a - |s,e|)\big) \quad (1)$$

where *a* represents the horizontal displacement of the binding curve, *k* sets the steepness of the curve at *a*, and

$$Z(k) = 1 + \sum_{e \in IEDB} exp\,(-k(a - |s,e|)) \quad (2)$$

Which represents the partition function over the unbound state and the all-bound state. Here, *k*=4.87 and *a*=26, which were determined in the original study.

The likelihood of peptide presented by MHC I is defined as:

$$A = C * L(Rm) \quad (3)$$

The likelihood of pMHC recognized by T cells is defined as:

$$B = H * R * (1 - 2^{-m} L(Rn)) \quad (4)$$

The candidate neoantigen score is defined as:

$$S = A * B \quad (5)$$

Where *L(x)* is a logistic function given by:

$$L(x) = \frac{1}{1+e^{5(x-2)}} \quad (6)$$

L(x) is a negative logistic function (Bjerregaard *et al.*, 2017; **Figure S4A**). This function gives a value approaching 0 for a high binding affinity percent rank, a midpoint at a binding affinity percent rank of 2, and a value of one for a lowbinding affinity percent rank. The constant 2 defines the inflection point and it was chosen since a binding affinity percent rank of 2 is the recommened cutoff for peptide binding. The equation $(1 - 2^{-m} L(Rn))$ is a penalized function when scoring the candidate neoantigens: If the candidate neoantigen corresponding wild type peptide has a low dissociation constant, tolerance mechanisms will remove TCRs that are specific to the wild type

peptide. Owing to cross-reactivity, candidate neoantigen specific TCRs could be reduced.

It should be note that (1) All the factors relevant to immunogenic potential in our score scheme is not fusion candidate neoantigen specific. Therefore, our score scheme can be employed to evaluate the immunogenic potential of the SNV&indel based candidate neoantigens as well as the fusion based candidate neoantigens; (2) The exact determinants of immunogenicity are not well understood, the score scheme is designed empirically based on current knowledge. Our score scheme can be updated when further knowledge related to immunogenicity becomes available.

Evaluation of the rationality and effectiveness of our proposed score scheme

To evaluate the rationality of our proposed score scheme, we applied it to five public peptides datasets with experimentally confirmed immunogenic and non-immunogenic peptides (**Table S5**). Of the five peptides datasets, four are SNV&indel mutation based neo-peptides, one is fusion mutation based neo-peptides recently validated by Yang (Robbins *et al.*, 2013; Rajasagi *et al.*, 2014; Carreno *et al.*, 2015; Gros *et al.*, 2016; Yang *et al.*, 2019). Furthermore, to evaluate the performance of our proposed score scheme, we compared it with other available tools, including the neoantigen *fitness model* (Luksza *et al.*, 2017), *MuPeXI* 1.2 (Bjerregaard *et al.*, 2017), *neopepsee* (Kim *et al.*, 2018) and a tool available at IEDB (Calis *et al.*, 2013) that were all developed for peptides immunogenic potential evaluation. Peptides were scored according to our score scheme and these tools (**Table S5**). Area under the precision-recall curve (PR-AUC) and area under the receiver operating characteristic curve (ROC-AUC) were used to benchmark the performance. In 2 of 5 peptides datasets, our score scheme presented the highest ROC-AUC and in 3 of 5 peptides datasets, our score scheme presented the highest PR-AUC, indicating its superiority and rationality.

The following definitions are also presented related to our evaluations:

specific candidate neoantigen: a candidate neoantigen with binding affinity percent rank <=2 and the corresponding wild type peptide with binding affinity percent rank >2. Due to self-immune tolerance, compared with non-specific candidate neoantigens, specific candidate neoantigens tend to have higher immunogenic potential (Turajlic *et al.*, 2017).

$$\textbf{fusion mutation burden ratio} = \frac{\text{fusion mutation burden}}{\text{SNV\&indel mutation burden}}$$

$$\textbf{fusion candidate neoantigen burden ratio} = \frac{\text{fusion candidate neoantigen burden}}{\text{SNV\&indel candidate neoantigen burden}}$$

$$\textbf{fusion specific candidate neoantigen burden ratio} = \frac{\text{fusion specific candidate neoantigen burden}}{\text{SNV\&indel specific candidate neoantigen burden}}$$

candidate neoantigen per mutation: candidate neoantigens a mutation can generate

specific candidate neoantigen per mutation: specific candidate neoantigens a mutation can generate

specific candidate neoantigen per candidate neoantigen = $\frac{\text{specific candidate neoantigen burden}}{\text{candidate neoantigen burden}}$, a metric to evaluate the likelihood that a candidate neoantigen is the specific candidate neoantigen

Analysis of the MS cohort dataset

We analyzed 10 breast cancer cell lines in the MS dataset obtained from Rozanov (Rozanov *et al.*, 2018). MHC I bound peptides were eluted by MHC I immunoprecipitation and the eluted peptides were analyzed by mass spectrometry. Fusion candidate neoantigens were predicted following our *neoFusion* pipeline with RNA sequencing data. We used *ProteoWizard* (Chambers *et al.*, 2012) to convert Raw MS data to mzML format. For each cancer cell line, MS data were searched against the human reference proteome downloaded from UniProt concatenated with fusion candidate neoantigens. MS data were searched with *Comet* (Eng, Jahan and Hoopmann, 2013) and filtered with *Percolator* (Käll *et al.*, 2007) to identify fusion peptides presented by MHC I at a false discovery rate of 1%. *Comet* software parameters were set as in the original article. Peptide-spectrum matches were visualized by *xiSPEC* (Kolbowski, Combe and Rappsilber, 2018), a web-based spectrum viewer.

In our study, all the predicted fusion candidate neoantigens were scored and prioritized according to our score scheme. It should be noted during scoring those predicted fusion neoantigens, only the likelihood of peptides presentation by MHC was calculated as those peptides were eluted from pMHC complexes. The fusion candidate neoantigen TAISPIAVLPR in HCC1806 (92 fusion candidate neoantigens in total) and APKSSSGFSL in HCC1428 (29 fusion candidate neoantigens in total) rank 6/92 and 2/29, respectively (**Table S1**). The probability of the co-occurrence of such two ranks or lower is equal to 0.0236.

Analysis of the ICB cohort dataset

Two ICB cohorts with whole-exome sequencing and RNA sequencing data were downloaded. Among 39 patients with melanoma treated by anti-CTLA-4 in the Van Allen cohort, 17 patients had responses, 22 patients had no responses. Among 25 patients with melanoma treated by anti-PD-1 in the Hugo cohort, 12 patients had responses, 13 patients had no responses. Fusion candidate neoantigens were predicted following our *neoFusion* pipeline. SNV&indel candidate neoantigens of the Van Allen cohort were determined by our inhouse pipeline. In brief, somatic SNV&indel VCFs were generated following *GATK* (Van der Auwera *et al.*, 2013) best practices workflow. Mutations should pass all the criteria described in the VCF file. Mutations with an allelic frequency less than 0.05, coverage less than 15X, or supported by fewer than 5 reads were filtered. SNV&indel VCFs of the Hugo cohort were obtained from the supplementary material of the original article. We utilized *StringTie* (Pertea *et al.*, 2015) to quantify the gene expression level in transcripts per million (TPM). HLA alleles of each sample were inferred from the RNA sequencing data by *OptiType*. VCFs and expression profile files were inputted to the *MuPeXI* program to predict SNV&indel

neoantigens (parameter, peptide length: 9,10,11; reference version: hg38). SNV&indel candidate neoantigen expression threshold was set to 1 TPM. Fusion and SNV&indel candidate neoantigen score were calculated according to our score scheme.

The tumor fusion candidate neoantigen score (*TFS*) was defined as the sum of the fusion candidate neoantigen score. The tumor SNV&indel candidate neoantigen score (*TSS*) was defined as the sum of the SNV&indel candidate neoantigen score. The overall tumor candidate neoantigen score was defined as: $TNS = TFS + TSS$. Like Luksza et al., the cytotoxic lymphocyte (CTL) fraction was used as the proxy for immune cytolytic activity (Luksza *et al.*, 2017). Gene expression profile files output by *StringTie* were inputted to *MCPcounter* (Becht *et al.*, 2016) to derive the CTL fraction.

Survival analysis was performed using the Kaplan-Meier method, with *p*-value determined by a log-rank test. Samples were split by the median value cutoff. Survival data were retrieved from the original study. The hazard ratio was determined through a Cox proportional hazards model. Multivariate Cox regression was performed using the overall tumor candidate neoantigen score*CTL, considering sex and age.

Analysis of the TCGA cohort dataset

Of 9624 tumor samples representing 33 tumor types, 25664 fusions were retrieved from Gao et al. (Gao *et al.*, 2018; **Table S3**). In addition, 7489 tumors SNV&indel VCFs from 20 solid tumor types were downloaded from TCGA. Finally, only 6552 samples possessed fusion mutation, SNV&indel mutation, and HLA allele information (Thorsson *et al.*, 2018). Fusion neoantigens were predicted following our *neoFusion* pipeline. Somatic SNV&indel VCFs and corresponding expression files were downloaded from TCGA and inputted to the *MuPeXI* program to predict SNV&indel neoantigens. Predicted fusion neoantigens and predicted SNV&indel neoantigens were scored using our score scheme (**Table S4**). The landscape of the microsatellite instability of TCGA tumor samples was obtained from Bonneville (Miya *et al.*, 2017). As suggested by Bonneville, for all cases, a threshold of 0.4 was set to differentiate samples with high microsatellite instability from those with microsatellite stability.

SMG1, SMG5, SMG6, SMG7, UPF1, UPF2, UPF3A and UPF3B genes were selected as the biomarkers of nonsense-mediated decay (Han *et al.*, 2018). The TCGA sample expression files were downloaded from TCGA website. Compared with samples without frameshift fusion mutation, except for the SMG6 and UPF3A genes, the expression level of other genes in samples harboring frameshift fusion are slightly higher (10%~20%, Student's t-test, p-value<0.01).

The fusion score was calculated as the sum of candidate neoantigen scores generated by that fusion. For the fusion that occurred multiple times, its median value was used to represent its fusion score. In total, there were 8634 passenger fusion scores, 844 kinase fusion scores, 204 Onco fusion scores and 172 TSG fusion scores. One-sided Mann-Whitney U hypothesis test might be affected by extremely different sample size.

To control sample size effect, we randomly sampled 600 passenger fusion scores and we compared them with fusion scores of other categories. We repeated random sampling procedure for 10000 times and we plotted the distribution of the corrected *p*-values to determine whether the passenger fusion scores are significantly different from other categories. It is shown that the Onco fusion score was significantly lower than the passenger fusion score, but not others (**Figure S4B-D**).

TCGA BLCA CTL fraction information was obtained from Thorsson (Thorsson *et al.*, 2018), and overall survival information was downloaded from TCGA website.

Supplemental References

Van der Auwera, G. A. *et al.* (2013) 'From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline', *Current protocols in bioinformatics*, 43(1110), p. 11.10.1-11.10.33. doi: 10.1002/0471250953.bi1110s43.

Becht, E. *et al.* (2016) 'Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression', *Genome Biology*, 17(1), p. 218. doi: 10.1186/s13059-016-1070-5.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Calis, J. J. A. *et al.* (2013) 'Properties of MHC Class I Presented Peptides That Enhance Immunogenicity', *PLoS Computational Biology*, 9(10). doi: 10.1371/journal.pcbi.1003266.

Carreno, B. M. *et al.* (2015) 'A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells', *Science*, 348(6236), p. 803 LP-808. doi: 10.1126/science.aaa3828.

Chambers, M. C. *et al.* (2012) 'A cross-platform toolkit for mass spectrometry and proteomics', *Nature Biotechnology*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 30, p. 918. Available at: https://doi.org/10.1038/nbt.2377.

Chowell, D. *et al.* (2015) 'TCR contact residue hydrophobicity is a hallmark of immunogenic CD8 + T cell epitopes', *Proceedings of the National Academy of Sciences*, 112(14), pp. E1754–E1762. doi: 10.1073/pnas.1500973112.

Eng, J. K., Jahan, T. A. and Hoopmann, M. R. (2013) 'Comet: an open-source MS/MS sequence database search tool.', *Proteomics*. Germany, 13(1), pp. 22–24. doi: 10.1002/pmic.201200439.

Gros, A. *et al.* (2016) 'Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients', *Nature Medicine*. Nature Publishing Group, 22(4), pp. 433–438. doi: 10.1038/nm.4051.

Haas, B. *et al.* (2017) 'STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq', *bioRxiv*, p. 120295. doi: 10.1101/120295.

Han, X. *et al.* (2018) 'Nonsense-mediated mRNA decay: A "nonsense" pathway makes sense in stem cell biology', *Nucleic Acids Research*. Oxford University Press,

46(3), pp. 1038–1051. doi: 10.1093/nar/gkx1272.
Käll, L. *et al.* (2007) 'Semi-supervised learning for peptide identification from shotgun proteomics datasets', *Nature Methods*. Nature Publishing Group, 4, p. 923. Available at: https://doi.org/10.1038/nmeth1113.
Kim, P. and Zhou, X. (2018) 'FusionGDB: fusion gene annotation DataBase', *Nucleic Acids Research*. Oxford University Press, pp. 1–11. doi: 10.1093/nar/gky1067.
Kim, S. *et al.* (2018) 'Neopepsee: Accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information', *Annals of Oncology*, 29(4), pp. 1030–1036. doi: 10.1093/annonc/mdy022.
Kolbowski, L., Combe, C. and Rappsilber, J. (2018) 'xiSPEC: web-based visualization, analysis and sharing of proteomics data.', *Nucleic acids research*. England, 46(W1), pp. W473–W478. doi: 10.1093/nar/gky353.
Nielsen, M. and Andreatta, M. (2016) 'NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets', *Genome Medicine*. Genome Medicine, 8(1), pp. 1–9. doi: 10.1186/s13073-016-0288-x.
Pertea, M. *et al.* (2015) 'StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.', *Nature biotechnology*. United States, 33(3), pp. 290–295. doi: 10.1038/nbt.3122.
Rajasagi, M. *et al.* (2014) 'Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia', *Blood*, 124(3), pp. 453–462. doi: 10.1182/blood-2014-04-567933.
Robbins, P. F. *et al.* (2013) 'Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells', *Nature Medicine*. Nature Publishing Group, 19(6), pp. 747–752. doi: 10.1038/nm.3161.
Rozanov, D. V. *et al.* (2018) 'MHC class I loaded ligands from breast cancer cell lines: A potential HLA-I-typed antigen collection', *Journal of Proteomics*, 176(September 2017), pp. 13–23. doi: 10.1016/j.jprot.2018.01.004.
Stranzl, T. *et al.* (2010) 'NetCTLpan: Pan-specific MHC class I pathway epitope predictions', *Immunogenetics*, 62(6), pp. 357–368. doi: 10.1007/s00251-010-0441-4.
Szolek, A. *et al.* (2014) 'OptiType: precision HLA typing from next-generation sequencing data', *Bioinformatics (Oxford, England)*. 2014/08/20. Oxford University Press, 30(23), pp. 3310–3316. doi: 10.1093/bioinformatics/btu548.
Vita, R. *et al.* (2009) 'The Immune Epitope Database 2.0', *Nucleic Acids Research*, 38(SUPPL.1). doi: 10.1093/nar/gkp1004.
Zhang, J., Mardis, E. R. and Maher, C. A. (2017) 'INTEGRATE-neo: A pipeline for personalized gene fusion neoantigen discovery', *Bioinformatics*, 33(4), pp. 555–557. doi: 10.1093/bioinformatics/btw674.
Zhou, C. *et al*. (2019) 'pTuneos: prioritizing Tumor neoantigens from next-generation sequencing data', *Genome Medicine*, Advance Access, 2019.