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

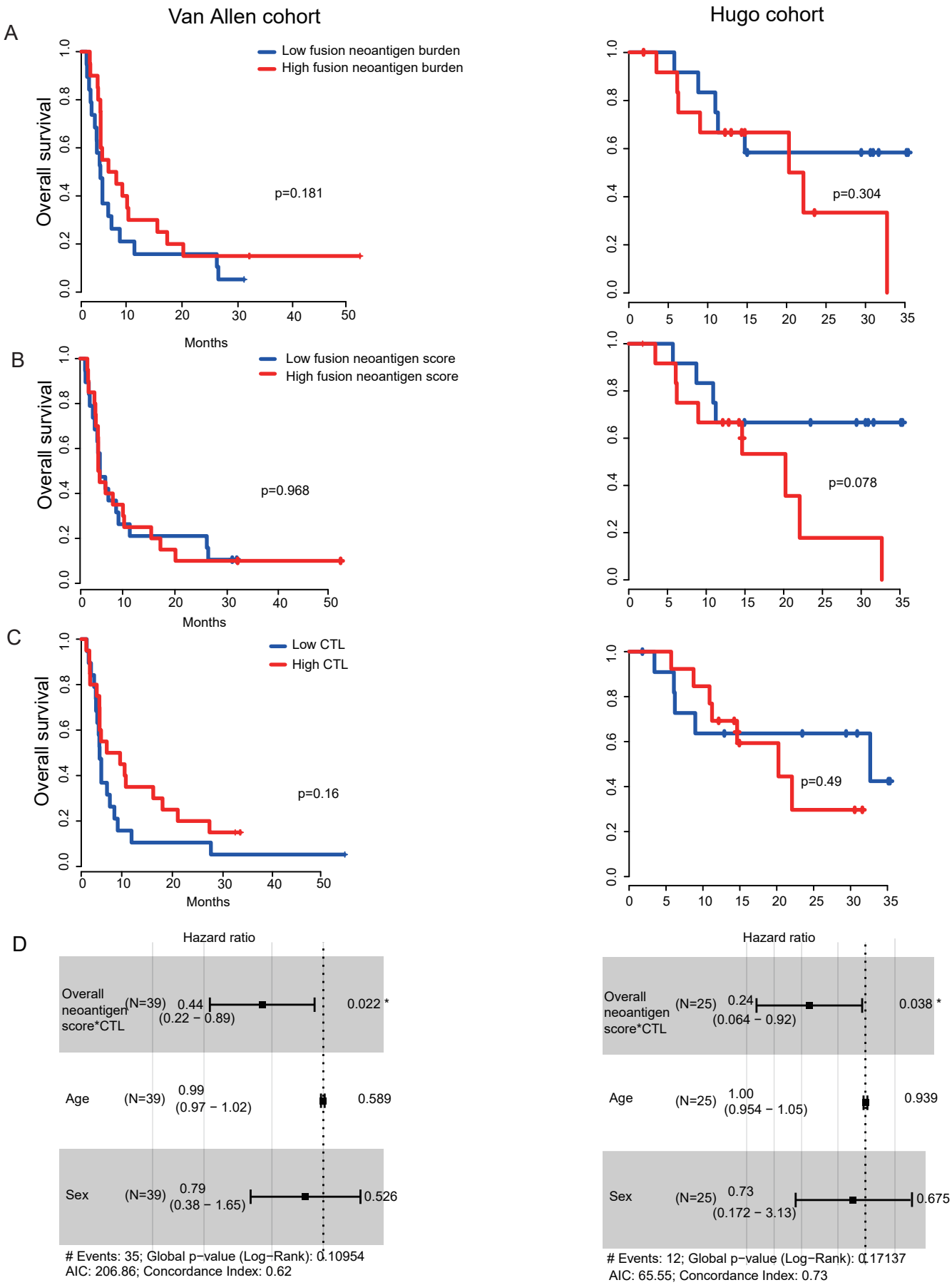

Figure S1. Survival analysis with different metrics. Related to Figure 2.
(A) The tumor fusion candidate neoantigen burden could not separate patients in both cohorts.
(B) The tumor fusion candidate neoantigen score could not separate patients in both cohorts.
(C) The CTL is not related to immunotherapy 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.

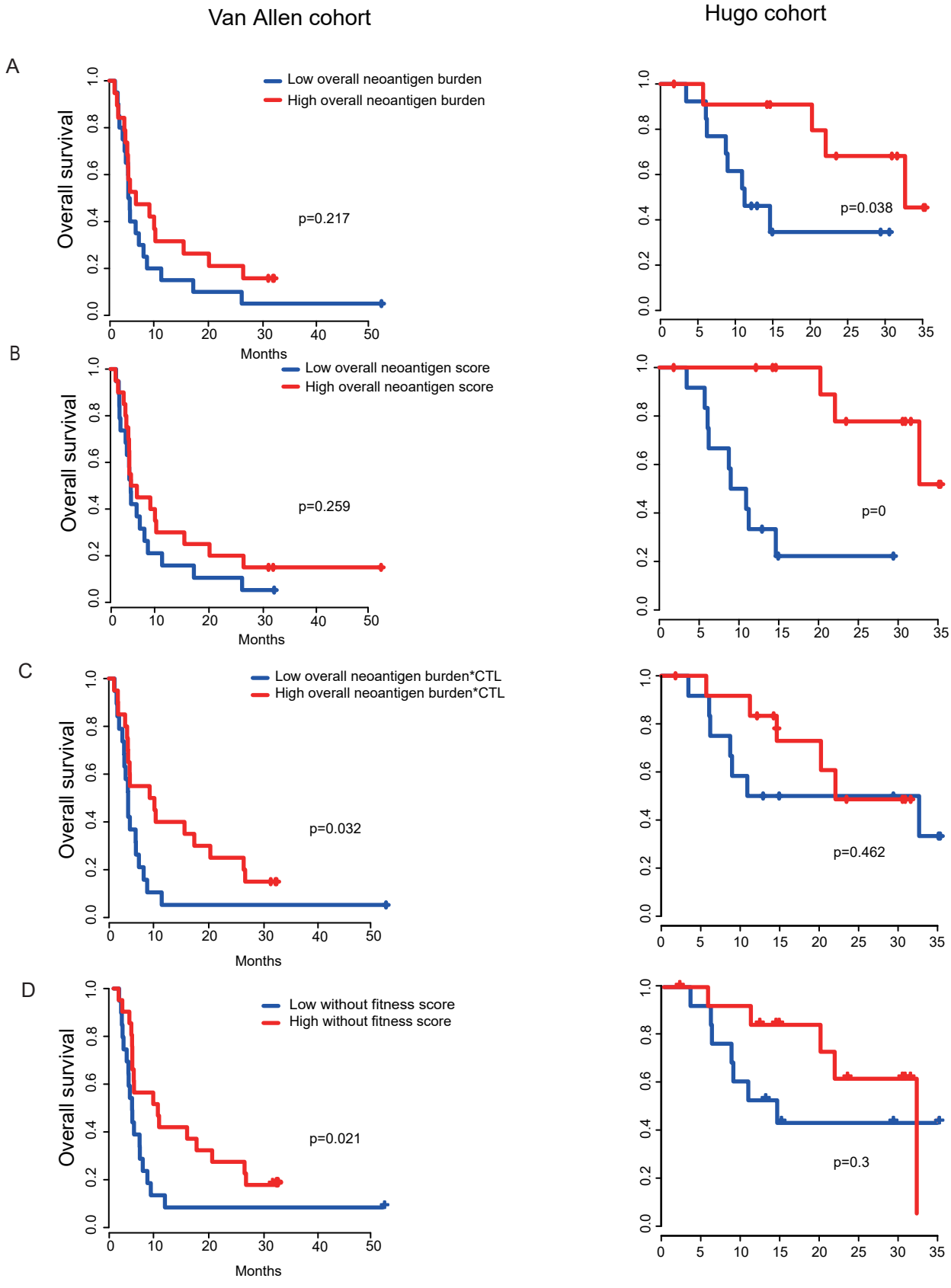

Figure S2. Survival analyses with different metrics. Related to Figure 2.
(A) The overall tumor neoantigen burden is related to immunotherapy outcome in the Hugo cohort, while not in the Van Allen cohort.
(B) The overall tumor neoantigen score is related to immunotherapy outcome in the Hugo cohort, while 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.

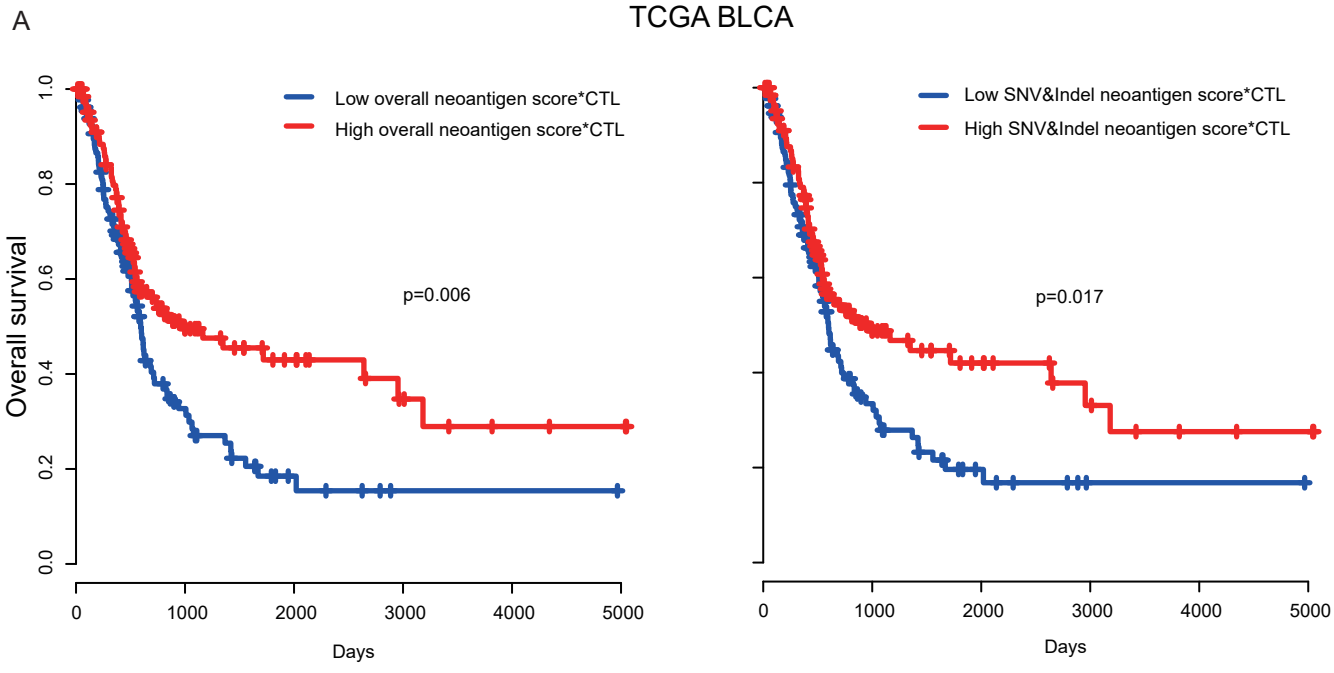

Figure S3. The overall tumor candidate neoantigen score*CTL is not a prognostic factor for overall survival. Related to Figure 2.

(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.

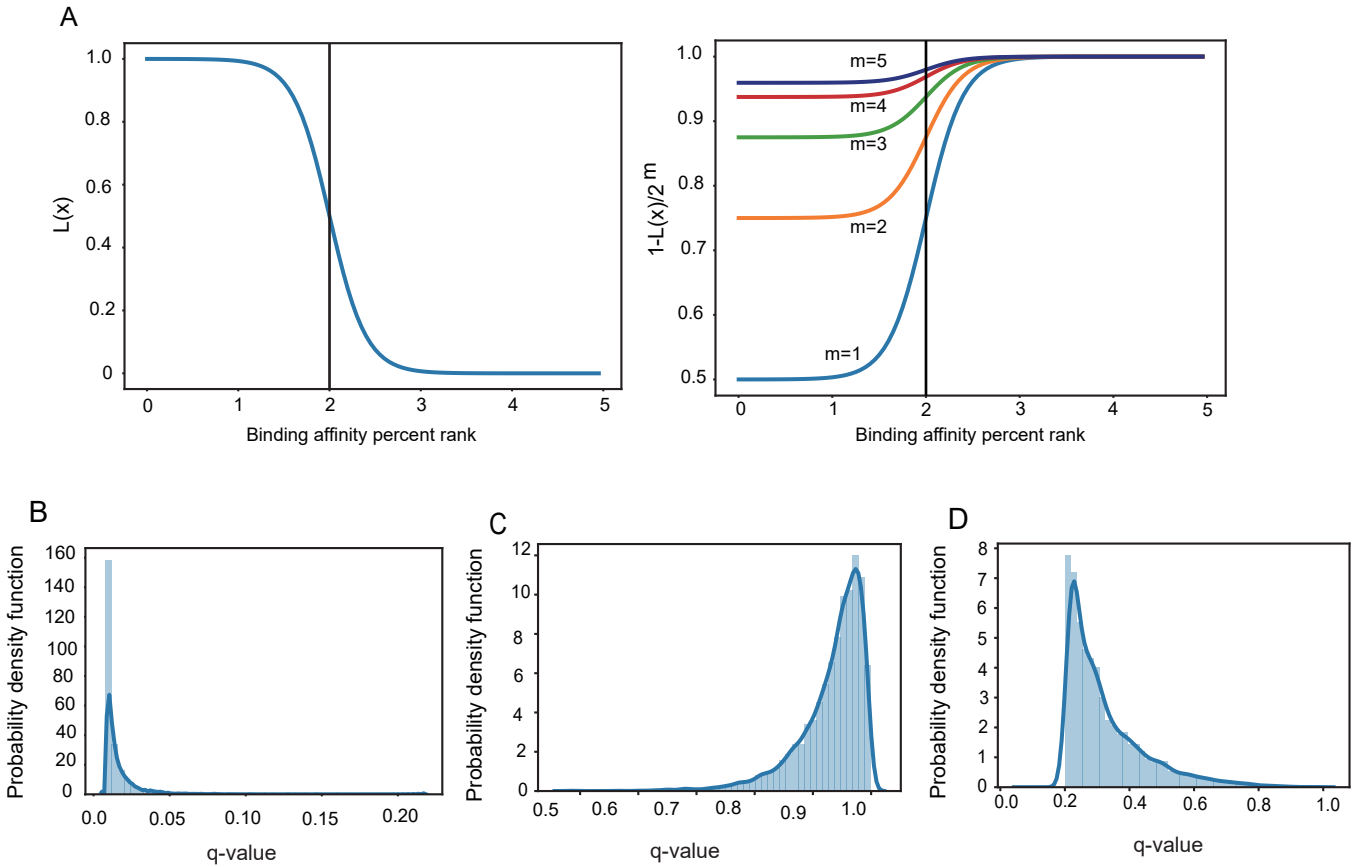

Figure S4. The negative logistic function and the distribution of q-values. Related to Figure 4 and 5.
(A) The negative logistic function of L(x), and m indicates the mismatch between the candidate neoantigen and the corresponding normal peptide.
(B) The distribution of q-values of the comparison between 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.